Bayesian 3D X-ray Computed Tomography with a Hierarchical Prior model for Sparsity in Haar Transform domain

Li Wang, Ali Mohammad-Djafari, Nicolas Gac, Mircea Dumitru

MDPI

Article

Bayesian 3D X-ray Computed Tomography with a Hierarchical Prior model for Sparsity in Haar Transform domain

Li Wang [1], Ali Mohammad-Djafari [1], Nicolas Gac [1] and Mircea Dumitru [1]

[1] Laboratoire des signaux et système, Centralesupelec, CNRS, 3 Rue Joliot Curie, 91192 Gif sur Yvette; firstname.surname@l2s.centralesupelec.fr

Academic Editor: name
Version December 3, 2018 submitted to Entropy; Typeset by LaTeX using class file mdpi.cls

Abstract: In this paper, a hierarchical prior model based on the Haar transformation and an appropriate Bayesian computational method for X-ray CT reconstruction are presented. Given the piece-wise continuous property of the object, a multilevel Haar transformation is used to associate a sparse representation for the object. The sparse structure is enforced via a generalized Student-t distribution ($\mathcal{S}t_g$), expressed as the marginal of a Normal-Inverse Gamma distribution. The proposed model and corresponding algorithm are designed to adapt to specific 3D data sizes and to be used in both medical and industrial Non Destructive Testing (NDT) applications. In the proposed Bayesian method, a hierarchical structured prior model is proposed, and the parameters are iteratively estimated. The initialization of the iterative algorithm uses the parameters of the prior distributions. A novel strategy for the initialization is presented and proved experimentally. We compare the proposed method with two state-of-the-art approaches, showing that our method has better reconstruction performance when fewer projections are considered and when projections are acquired from limited angles.

Keywords: X-ray Computed Tomography; inverse problem; sparsity; hierarchical structure; Generalized Student-t distribution; Haar transformation.

1. Introduction

Computed Tomography (CT) has been developed and widely used in medical diagnosis [1] and industrial Non Destructive Testing (NDT) [2] in recent decades. In CT, objects are observed using different techniques, for example X-rays [3], ultrasound [4], microwaves [5], infra-red [6]. X-ray CT employs the absorption of X-rays by the organs in a body or by the materials in industrial components to reconstruct the internal structure of the imaged object. When performing X-ray CT, a set of X-ray images of the measured parts are acquired. The intensity measured by the X-ray images corresponds to the intensity of the radiation passing through and attenuated by the object. CT reconstruction is typically treated as an inverse problem.

The conventional analytical techniques for CT reconstruction are based on the Radon Transform [7]. Filtered Back-Projection (FBP) [8] is the most frequently used analytical method in practical applications. FBP performs well when reconstructing from sufficient data with a high signal-to-noise ratio (SNR), but it suffers from artifacts when reconstructing from insufficient data or with noise.

Owing to considerations regarding patients' health in medical CT, and in order to reduce acquisition time in industrial applications, reconstruction with insufficient datasets is increasingly attracting the attention of researchers. Reconstruction from fewer projections is an ill-posed inverse problem [9,10]. In this case, conventional analytical reconstruction methods provide unsatisfactory results and iterative methods can be used to improve the reconstruction performance. The Algebraic

Reconstruction Technique (ART) [11,12], the Simultaneous algebraic reconstruction technique (SART) [13] and the Simultaneous Iterative Reconstruction Technique (SIRT) [14,15] are some of the iterative methods proposed initially. These methods consider the discretized forward system model: $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f}$, where $\boldsymbol{f} \in \mathbb{R}^{N\times 1}$ represents the object, $\boldsymbol{g} \in \mathbb{R}^{M\times 1}$ represents the observed dataset and matrix $\boldsymbol{H} \in \mathbb{R}^{M\times N}$ is the linear projection operator, mainly based on the geometry of acquisition (e.g. parallel beam, cone beam, etc) [16–18]. Typically, the system of equations is under-determined, i.e. $N > M$. In this context, regularization methods are frequently used and the forward system is modeled as:

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{M\times 1}$ represents the additive noise applied to the projection system. The regularization methods estimate the unknowns by minimizing a penalty criterion, which generally consists of two terms:

$$J(\boldsymbol{f}) = Q(\boldsymbol{g}, \boldsymbol{f}) + \lambda R(\boldsymbol{f}). \tag{2}$$

The loss function $Q(\boldsymbol{g}, \boldsymbol{f})$ describes discrepancies in the observed data, such as the quadratic ($L2$) loss $Q(\boldsymbol{g}, \boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2$ or L_q loss $Q(\boldsymbol{g}, \boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_q^q$ with $1 \leq q < 2$. Other expressions such as the Huber function are also reported. The regularization term $R(\boldsymbol{f})$ is a penalty on the complement criterion of $\boldsymbol{f}$, such as restriction for smoothness $\|\Phi(\boldsymbol{f})\|_2^2$ or for sparsity $\|\Phi(\boldsymbol{f})\|_1$, where $\Phi(\boldsymbol{f})$ represents a linear function of $\boldsymbol{f}$. The parameter λ is known as the regularization parameter, which controls the trade-off between the forward model discrepancy and the penalty term.

By choosing different regularization functions $R(\boldsymbol{f})$, different regularization methods can be implemented. $R(\boldsymbol{f}) = 0$ refers to the Least-Squares (LS) method [19], with the drawback that the reconstruction is sensitive to the noise due to the ill-posedness of the problem and the ill-conditioning of the operator $\boldsymbol{H}$. Quadratic Regularization (QR), also known as the Tikhonov method [20], is given by $R(\boldsymbol{f}) = \|\Phi(\boldsymbol{f})\|_2^2$, where the linear operator $\Phi(\cdot)$ is the derivation operator in most cases. The well-known Total Variation (TV) method [21–23] is defined by $R(\boldsymbol{f}) = \|\boldsymbol{D}_{TV}\boldsymbol{f}\|_{TV}$ where $\boldsymbol{D}_{TV}$ is the gradient operator. $\boldsymbol{D}_{TV}$ is equal to $\|\boldsymbol{D}_x\boldsymbol{f}\|_1 + \|\boldsymbol{D}_y\boldsymbol{f}\|_1 + \|\boldsymbol{D}_z\boldsymbol{f}\|_1$ for a 3D object in an anisotropic form, where $\boldsymbol{D}_x$, $\boldsymbol{D}_y$ and $\boldsymbol{D}_z$ are respectively the gradient operators in the x, y and z directions. The L_1 norm is used in TV for sparse estimations, which enforces the sparsity of $\boldsymbol{D}_{TV}\boldsymbol{f}$. The appearance of the non-differentiable L_1 term leads to difficulties for the implementation of optimization algorithms. Many optimization algorithms have been proposed to solve this L_1 norm optimization problem, for example the primal-dual method [24], the Split Bregman method [22], etc. In regularization optimization, due to the large projection data size and the great number of voxels, the explicite expression of the solution can not be used directly because of the impossibility of inversing the large size matrix such as $\left(\boldsymbol{H}^T\boldsymbol{H} + \lambda\boldsymbol{D}^T\boldsymbol{D}\right)^{-1}$. Hence, optimization algorithms such as gradient descent or conjugate gradient are often used.

More general regularization methods have been developed based on the constrained and dual-variable regularization method:

$$J(\boldsymbol{f}, \boldsymbol{z}) = Q_1(\boldsymbol{g}, \boldsymbol{f}) + \eta Q_2(\boldsymbol{f}, \boldsymbol{z}) + \lambda R(\boldsymbol{z}), \tag{3}$$

which corresponds to the maximum a posterior optimization of a hierarchical structured model where both $\boldsymbol{f}$ and $\boldsymbol{z}$ are unknown variables. In such a model, the penalty regularization term is set on $\boldsymbol{z}$, which is associated with $\boldsymbol{f}$ via a linear transformation. The loss functions $Q_1(\boldsymbol{g}, \boldsymbol{f})$ and $Q_2(\boldsymbol{f}, \boldsymbol{z})$ are for example quadratic ($L2$), i.e. $Q_1(\boldsymbol{g}, \boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2$ and $Q_2(\boldsymbol{f}, \boldsymbol{z}) = \|\boldsymbol{f} - \boldsymbol{D}\boldsymbol{z}\|_2^2$ where $\boldsymbol{D}$ is a linear transform operator such as a wavelet transformation.

Among the methods treating this type of regularization problem, we mention here the Alternating Direction Method of Multipliers (ADMM) [25]. It minimizes $\Phi(\boldsymbol{f}) + \Psi(\boldsymbol{z})$ subject to $\boldsymbol{A}\boldsymbol{f} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{C}$, and it covers a large number of estimation forms. One example is when $\Phi(\boldsymbol{f}) =$

$\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|^2$, $\Psi(\boldsymbol{z})=\boldsymbol{R}(\boldsymbol{z})$, $\boldsymbol{A}=\boldsymbol{I}$, $\boldsymbol{B}=-\boldsymbol{D}$ and $\boldsymbol{C}=\boldsymbol{0}$, and refers to the above-mentioned bi-variable regularization method corresponding to Eq. (3).

In the above-mentioned regularization methods, there is always a regularization parameter λ to be fixed. Sometimes the regularization term consists of more than one parts and each of them are weighted by a regularization parameter, for example the Elastic-net regularizer [26]. In these cases, two or even more regularization parameters are needed to be fixed. Cross Validation (CV) and the L-curve method [20,27,28] are conventional methods used to determine suitable values for these parameters. However, this work must be repeate for different simulated datasets and is therefore very costly. Statistical methods, therefore, have been developed and used to solve this problem.

From the probabilistic point of view, a Gaussian model for the additive noise in the forward model, Eq. (1), leads to the quadratic expression $\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|^2$ in the corresponding regularization criterion. But in some types of tomography, for example Positron Emission Tomography (PET) or X-ray tomography with very low number of phantoms, the noise is modeled by a Poisson distribution. In order to account for a more precise modeling of the noise and the other variables and parameters, statistical methods are used [29]. The Maximum Likelihood (ML) methods [30] and different estimation algorithms such as the Expectation Maximization (EM) algorithms [31], the stochastic EM (SEM) [32], the Ordered subsets-EM (OS-EM) [33] are commonly used in PET-CT reconstruction problems.

Another widely used type of probabilistic method for PET or X-ray CT reconstruction is the Bayesian inference [34–37]. The prior knowledge is translated by the prior probability model and is used to obtain the expression of the posterior distribution. The basic Bayesian formula is:

$$p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta})=\frac{p(\boldsymbol{g}|\boldsymbol{f},\theta_1)p(\boldsymbol{f}|\theta_2)}{p(\boldsymbol{g}|\boldsymbol{\theta})},\ \text{with}\ p(\boldsymbol{g}|\boldsymbol{\theta})=\int p(\boldsymbol{g}|\boldsymbol{f},\theta_1)p(\boldsymbol{f}|\theta_2)\,\mathrm{d}\boldsymbol{f}, \tag{4}$$

where $p(\boldsymbol{g}|\boldsymbol{f},\theta_1)$ is the likelihood, $p(\boldsymbol{f}|\theta_2)$ is the prior distribution, $p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta})$ is the posterior distribution, $\boldsymbol{\theta}=(\theta_1,\theta_2)$ are the parameters of these different distributions and $p(\boldsymbol{g}|\boldsymbol{\theta})$ is the evidence of the parameters in the data $\boldsymbol{g}$. By using Maximum A Posterior (MAP) estimator $\widehat{\boldsymbol{f}}=\arg\max_{\boldsymbol{f}}\{p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta})\}=\arg\min_{\boldsymbol{f}}\{-\ln p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta})\}$, links between the Bayesian method and almost all the regularization methods can be illustrated. A Gaussian prior for $p(\boldsymbol{f})$ in Eq. (4) leads to the Quadratic ($L2$) Regularization method, while a Laplacian prior in Eq. (4) leads to the $L1$ (LASSO or TV) regularization method. The regularization parameter can be related to θ_1 and θ_2. One advantage of the Bayesian method is having some explanation for the regularization parameter via its link with θ_1 and θ_2. For example when $p(\boldsymbol{g}|\boldsymbol{f},\theta_1)$ and $p(\boldsymbol{f}|\theta_2)$ are Gaussian with θ_1 and θ_2 respectively the variances of the noise and the variance of the prior, then the regularization parameter is $\lambda=\theta_1/\theta_2$. Another advantage of the Bayesian method is that these parameters can also be estimated to achieve unsupervised or semi-supervised methods. This is achieved by obtaining the expression of the joint posterior probability law:

$$p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})=\frac{p(\boldsymbol{g}|\boldsymbol{f},\theta_1)p(\boldsymbol{f}|\theta_2)p(\boldsymbol{\theta})}{p(\boldsymbol{g})}, \tag{5}$$

where $p(\boldsymbol{\theta})$ is an appropriate prior on $\boldsymbol{\theta}$. For a hierarchical structured model where a hidden variable $\boldsymbol{z}$ appears in the prior model, we have:

$$p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g})=\frac{p(\boldsymbol{g}|\boldsymbol{f},\theta_1)p(\boldsymbol{f}|\boldsymbol{z},\theta_2)p(\boldsymbol{z}|\theta_3)p(\boldsymbol{\theta})}{p(\boldsymbol{g})}, \tag{6}$$

where $\boldsymbol{\theta}=[\theta_1,\theta_2,\theta_3]$.

With the posterior distribution obtained from an unsupervised Bayesian inference as in Eq. (5), we distinguish three estimation methods. The first method consists of integrating out $\boldsymbol{\theta}$ from $p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})$ to obtain $p(\boldsymbol{f}|\boldsymbol{g})$ and then use $p(\boldsymbol{f}|\boldsymbol{g})$ to infer on $\boldsymbol{f}$. The second approach is firstly to marginalize $p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})$ with respect to $\boldsymbol{f}$ to obtain $p(\boldsymbol{\theta}|\boldsymbol{g})=\int p(\boldsymbol{f},\boldsymbol{\theta}|\boldsymbol{g})\,\mathrm{d}\boldsymbol{f}$ and estimate $\widehat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}\{p(\boldsymbol{\theta}|\boldsymbol{g})\}$,

then use $\widehat{\boldsymbol{\theta}}$ as it was known. Unfortunately, these approaches do not often give explicit expressions for $p(\boldsymbol{f}|\boldsymbol{g})$ or $p(\boldsymbol{\theta}|\boldsymbol{g})$. The third and easiest algorithm to implement is the joint optimization, which estimates variable $\boldsymbol{f}$ and parameter $\boldsymbol{\theta}$ iteratively and alternately. Bayesian point estimators such as joint maximum a posteriori (JMAP) [38] and Posterior Mean (PM) [39] via Variational Bayesian Approximation (VBA) methods [40–42] are often used.

In order to distinguish the details of a reconstructed object, a high-resolution image is expected. In industrial applications, especially for the NDT of a large-size object, the size of the projection (1000 images of 1000^2 pixels) and the number of the voxels (1000^3 voxels) become critical, and so does the projection and back-projection operators in CT. It is necessary to account for some computational aspects, for example the GPU processor [43,44].

In our previous work [45], we proposed to use the Bayesian method via a synthesis model, in which the multilevel Haar transformation coefficient $\boldsymbol{z}$ of the image is first estimated, and then the final image reconstruction result is obtained from post processing: $\widehat{\boldsymbol{f}} = \boldsymbol{D}\widehat{\boldsymbol{z}}$. In this case, when using a Laplacian prior model and the MAP estimator, the problem becomes equivalent to the optimization of $J(\boldsymbol{z}) = \|\boldsymbol{g} - \boldsymbol{HDz}\|^2 + \lambda \|\boldsymbol{z}\|_1$ which is a typical L_1 regularization method. The particularity of our work was to use a generalized Student-t (St_g) prior model [46] in place of the Laplacian model.

In this paper, we present a Hierarchical Haar transform-based Bayesian method (HHBM), first proposed in [47], in which the object to be reconstructed, $\boldsymbol{f}$, is related to the Haar transformation coefficient $\boldsymbol{z}$ by $\boldsymbol{f} = \boldsymbol{Dz} + \boldsymbol{\xi}$ where $\boldsymbol{\xi}$ represents the modelization uncertainties. $\boldsymbol{f}$ and $\boldsymbol{z}$ are estimated simultaneously. Wavelets provide an optimal representation for a piecewise continuous function consisting of homogeneous blocs separated by jump discontinuities (the contours), as the wavelet representation is sparse for such signals. Transformations used are, for example, the Haar transformation [48], the curvelet transformation (CVT) [49], the contourlet transformation (CT) [50], Dual-Tree Complex Wavelet Transform (DT-CWT) [51], etc. As long as the object under consideration, $\boldsymbol{f}$, is piecewise continuous or constant, the Haar transform is appropriate, with the advantage that: 1) the transform coefficients are sparse, 2) the transformation operator is orthogonal so that the inverse operator and the transpose are identical, and 3) the computation of this transformation consists of only additions and subtractions and the cost of computation is only $O(\sqrt{N})$ where N is the size of the object $\boldsymbol{f}$.

The sparsity of the transformation coefficient is generally defined by three classes of distributions: the Generalized Gaussian distributions [52], the mixture distributions [53] and the heavy-tailed distributions [54]. In this paper, we use a generalization of Student-t distribution (St_g), which belongs to the heavy-tailed family and has many advantages when enforcing the sparsity of variables [46].

In this paper, we extend extensively the previous work by: 1) adapting the forward model and prior models to the 3D case, which is more appropriate for real 3D large data size applications; 2) comparing the RMSE of the phantom reconstructed by HHBM method with those by conventional QR and TV methods, we show the advantages of the semi-supervised property of the HHBM method and that the HHBM method outperforms the TV method when infsufficient data is estimated; 3) proposing new ideas for fixing the hyper-parameters in the proposed model; and 4)evaluating the performance of the proposed method in the situations when the number or the angle distribution of the projections are limited.

The rest of this paper is organized as follows: section 2 presents the proposed hierarchically-structured Bayesian method; section 3 gives the details of the implementations and the choice of hyperparameters, as well as the simulation results; some points on the initialization of hyper-parameters are discussed in section 4; conclusions are drawn and prospective future research is presented in section 5.

2. The Semi-supervised Hierarchical Model

The Hierarchical Haar-based Bayesian Method (HHBM) is presented, in which the object $\boldsymbol{f}$ and its multilevel Haar transform coefficient $\boldsymbol{z}$ are considered jointly. A sparse enforcing prior is defined on $\boldsymbol{z}$. The wavelet transformation has been used for the reconstruction of tomography images in some state-of-the-art works [55–58], using both regularization and Bayesian methods. In these state-of-the-art methods, the phantom $\boldsymbol{f}$ is obtained by a post-processing from reconstructed coefficient $\boldsymbol{z}$. In this paper, the phantom $\boldsymbol{f}$ and the coefficient $\boldsymbol{z}$ are simultaneously estimated.

2.1. Forward system model and likelihood

In the proposed method, the forward model introduced in Eq. (1) is considered. Generally, the noise in tomography is modeled by Poisson distribution [59], but, in X-ray CT, the Gaussian approximation is often used. We adopt the Gaussian approximation and propose to use a zero mean and non-stationary model where the variance is considered to be unknown, belonging to an Inverse Gamma distribution given that this distribution provides a good adaptation of the positivity property of the variances v_{ϵ_i}:

$$p(\boldsymbol{\epsilon}|\boldsymbol{v}_\epsilon) = \mathcal{N}\left(\boldsymbol{\epsilon}|\boldsymbol{0}, \boldsymbol{V}_\epsilon\right), \ \boldsymbol{V}_\epsilon = \text{diag}\left[\boldsymbol{v}_\epsilon\right], \ \text{where} \ \boldsymbol{v}_\epsilon = \left[v_{\epsilon_1}, \ \cdots, \ v_{\epsilon_M}\right]' \in \mathbb{R}^{M\times 1} \tag{7}$$

$$p(\boldsymbol{v}_\epsilon|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) = \prod_{i=1}^{M} \mathcal{IG}(v_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) \tag{8}$$

The vector $\boldsymbol{v}_\epsilon$ is considered in order to account for the difference of sensitivity to noise for each detector in each projection direction.

According to the forward model of the linear system, Eq. (1), and the prior model of the noise, Eq. (7), the likelihood of this model system is:

$$p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{v}_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f}, \boldsymbol{V}_\epsilon). \tag{9}$$

In Bayesian inference, the likelihood is combined with the prior distributions to determine the posterior distribution.

2.2. Hierarchical prior model and prior distributions

Typically the objects considered in medical and industrial X-ray CT are piecewise continuous. In this paper, a hierarchical prior model is used to define the piecewise continuous property. In this hierarchical prior model, a sparsity enforcing model is defined for the wavelet transformation coefficients of the image. A large number of methods accounting for sparse structure of the solution have been proposed in the literature. Among them, the L1 regularization method is most frequently used, which minimizes the criterion $J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \lambda \|\boldsymbol{\Phi}\boldsymbol{f}\|_1$ where $\boldsymbol{\Phi}$ is a linear operator, for example the gradient in the TV method. Another class of methods, known as "synthesis" [60], minimizes $J(\boldsymbol{z}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{D}\boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{z}\|_1$ where $\boldsymbol{f} = \boldsymbol{D}\boldsymbol{z}$, and $\boldsymbol{z} = \boldsymbol{D}^{-1}\boldsymbol{f}$ is for example a wavelet transform.

In this paper, we propose to use the multilevel Haar transformation as the sparse dictionary. The transformation is modeled using a discretized forward model:

$$\boldsymbol{f} = \boldsymbol{D}\boldsymbol{z} + \boldsymbol{\xi}, \tag{10}$$

where $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ represents the inverse multilevel Haar transformation and $\boldsymbol{\xi} \in \mathbb{R}^{N\times 1}$ represents the uncertainties of the transformation, which is introduced to relax the exact relation of the transform operator $\boldsymbol{D}$. $\boldsymbol{\xi}$ is supposed to be sparse. Unlike the gradient operator used in the TV method, the multilevel Haar transformation is orthogonal, i.e. $\boldsymbol{D}^{-1} = \boldsymbol{D}^T$. This property provides certain

advantages during optimization, especially for the big data size problems, as the inversion and transpose of the operator are identical and can be replaced by each other for different types of $\boldsymbol{D}$.

$\boldsymbol{\xi}$ is modeled by a Gaussian distribution, $p(\boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\xi}|0, \boldsymbol{V}_{\xi})$, where $\boldsymbol{V}_{\xi} = \text{diag}\left[\boldsymbol{v}_{\xi}\right]$ and $\boldsymbol{v}_{\xi} = [v_{\xi_1}, \cdots, v_{\xi_N}]'$. $\boldsymbol{v}_{\xi}$ is considered an unknown variance. It is modeled in order to realize a semi-supervised system where the variance is estimated. Here $\boldsymbol{v}_{\xi}$ is modeled by an Inverse Gamma distribution, with the same consideration as the model of $\boldsymbol{v}_{\epsilon}$. The Gaussian model with an Inverse Gamma distributed variance, $p(\boldsymbol{v}_{\xi}|\alpha_{\xi_0}, \beta_{\xi_0}) = \prod_{j=1}^{N} \mathcal{IG}(v_{\xi_j}|\alpha_{\xi_0}, \beta_{\xi_0})$, leads to a generalized Student-t ($\mathcal{S}t_g$) distribution [46]. Consequently, a $\mathcal{S}t_g$ distribution is derived for $\boldsymbol{\xi}$, and the sparse property of $\boldsymbol{\xi}$ can be guaranteed. From Eq. (10), the conditional distribution $p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{v}_{\xi})$ is derived:

$$p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{v}_{\xi}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{D}\boldsymbol{z}, \boldsymbol{V}_{\xi}), \tag{11}$$

with

$$p(\boldsymbol{v}_{\xi}|\alpha_{\xi_0}, \beta_{\xi_0}) = \prod_{j}^{N} \mathcal{IG}(v_{\xi_j}|\alpha_{\xi_0}, \beta_{\xi_0}). \tag{12}$$

For practical applications where these parameters are not known or difficult to obtain, we use a semi-supervised method in which the variances of noises, $\boldsymbol{v}_{\epsilon}$ and $\boldsymbol{v}_{\xi}$, are unknown. In HHBM, Inverse Gamma distribution is used to model $\boldsymbol{v}_{\epsilon}$ and $\boldsymbol{v}_{\xi}$, $p(\boldsymbol{v}_{\epsilon}) = \mathcal{IG}(\boldsymbol{v}_{\epsilon}|\alpha_{\epsilon_0}, \beta_{\epsilon_0})$ and $p(\boldsymbol{v}_{\xi}) = \mathcal{IG}(\boldsymbol{v}_{\xi}|\alpha_{\xi_0}, \beta_{\xi_0})$. Consequently, both $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$ are modeled by a $\mathcal{S}t_g$ distribution.

Vector $\boldsymbol{z} = [z_1, \cdots, z_N]'$ represents the multilevel Haar transformation coefficient of piece-wise continuous $\boldsymbol{f}$. As mentioned above, $\boldsymbol{z}$ is sparse. In this paper, the generalized Student-t distribution ($\mathcal{S}t_g$) [46] is used to enforce the sparsity structure of $\boldsymbol{z}$. The $\mathcal{S}t_g$ distribution can be expressed as the marginal of a bivariate Normal-Inverse Gamma distribution:

$$\mathcal{S}t_g(\boldsymbol{z}|\alpha, \beta) = \int \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{v})\, \mathcal{IG}(\boldsymbol{v}|\alpha, \beta)\ \mathrm{d}\boldsymbol{v}. \tag{13}$$

Thanks to the fact that Normal and Inverse Gamma are conjugate distributions, the use of the $\mathcal{S}t_g$ via Eq. (13) simplifies the computations when using the Bayesian point estimators such as the Posterior Mean via the Variational Bayesian Approximation (VBA) method [42].

From Eq. (13), the $\mathcal{S}t_g$ prior distribution modeling $\boldsymbol{z}$ is expressed as the following model:

$$p(\boldsymbol{z}|\boldsymbol{v}_z) = \mathcal{N}(\boldsymbol{z}|0, \boldsymbol{V}_z), \ \text{ where } \boldsymbol{V}_z = \text{diag}\left[\boldsymbol{v}_z\right], \ \ \boldsymbol{v}_z = [v_{z_1}, v_{z_2}, \cdots, v_{z_N}] \tag{14}$$

$$p(\boldsymbol{v}_z|\alpha_{z_0}, \beta_{z_0}) = \prod_{j}^{N} \mathcal{IG}(v_{z_j}|\alpha_{z_0}, \beta_{z_0}), \tag{15}$$

where $v_{z_j}, \ \forall\, j = 1 : N$ are i.i.d. distributed. The difference between the standard $\mathcal{S}t$ distribution, $\mathcal{S}\boldsymbol{t}(z|\nu) = \int \mathcal{N}(z|0, v)\mathcal{IG}(v|\frac{\nu}{2}, \frac{\nu}{2})\, \mathrm{d}v$, and the generalized $\mathcal{S}t_g$, given in Equation (13), is that $\mathcal{S}\boldsymbol{t}(z|\nu)$ is governed by one parameter ν but $\mathcal{S}\boldsymbol{t}_g(z|\alpha, \beta)$ is governed by two parameters (α, β). With these two parameters, the $\mathcal{S}t_g$ does not only enforce the sparsity of the variable, but also control the sparsity rate [46]. By changing the values of the two hyper-parameters α_{z_0} and β_{z_0}, we can obtain either a heavy-tailed distribution with a narrow peak, or a distribution approaching a Gaussian distribution.

2.3. The HHBM method

The prior models of the proposed Bayesian method based on the forward model of Eq. (1) and the prior model of Eq. (10) are:

$$p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{v}_\epsilon) \propto |\boldsymbol{V}_\epsilon|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})^T \boldsymbol{V}_\epsilon^{-1}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})\right], \tag{16}$$

$$p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{v}_\xi) \propto |\boldsymbol{V}_\xi|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{D}\boldsymbol{z})^T \boldsymbol{V}_\xi^{-1}(\boldsymbol{f}-\boldsymbol{D}\boldsymbol{z})\right], \tag{17}$$

$$p(\boldsymbol{z}|\boldsymbol{v}_z) \propto |\boldsymbol{V}_z|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{V}_z^{-1}\boldsymbol{z}\right], \tag{18}$$

$$p(\boldsymbol{v}_z|\alpha_{z_0},\beta_{z_0}) \propto \prod_j^N v_{z_j}^{-(\alpha_{z_0}+1)} \exp\left[-\beta_{z_0} v_{z_j}^{-1}\right], \tag{19}$$

$$p(\boldsymbol{v}_\epsilon|\alpha_{\epsilon_0},\beta_{\epsilon_0}) \propto \prod_i^M v_{\epsilon_i}^{-(\alpha_{\epsilon_0}+1)} \exp\left[-\beta_{\epsilon_0} v_{\epsilon_i}^{-1}\right], \tag{20}$$

$$p(\boldsymbol{v}_\xi|\alpha_{\xi_0},\beta_{\xi_0}) \propto \prod_j^N v_{\xi_j}^{-(\alpha_{\xi_0}+1)} \exp\left[-\beta_{\xi_0} v_{\xi_j}^{-1}\right]. \tag{21}$$

Fig.1 shows the generative graph of the proposed model in which the hyperparameters in the rectangles need to be initialized:

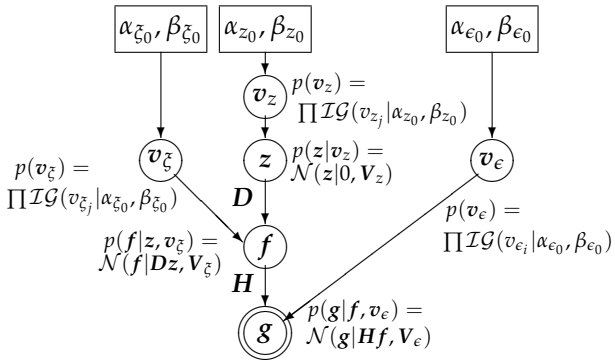


Figure 1. Generative graph of the proposed model illustrating all the unknowns (circles), hyperparameters (boxes) and data (double circles).

Via the Bayes rule, Eq. (6), the joint posterior distribution of all the unknowns given in the data is derived:

$$\begin{aligned} p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi,\boldsymbol{v}_z|\boldsymbol{g}) &= \frac{p(\boldsymbol{g},\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi,\boldsymbol{v}_z)}{p(\boldsymbol{g})} \\ &= \frac{p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{v}_\epsilon)p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{v}_\xi)p(\boldsymbol{z}|\boldsymbol{v}_z)p(\boldsymbol{v}_z)p(\boldsymbol{v}_\epsilon)p(\boldsymbol{v}_\xi)}{p(\boldsymbol{g})} \\ &\propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{v}_\epsilon)p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{v}_\xi)p(\boldsymbol{z}|\boldsymbol{v}_z)p(\boldsymbol{v}_z)p(\boldsymbol{v}_\epsilon)p(\boldsymbol{v}_\xi). \end{aligned} \tag{22}$$

Bayesian point estimators are often used for estimation via the a posteriori distribution. In this paper, we focus on the JMAP estimation, given that in the case of the large data size of the 3D object, the computational costs for the VBA algorithm is too expensive.

2.4. Joint Maximum A Posteriori (JMAP) Estimation

The negative logarithm of the posterior distribution is used as the criterion of optimization in order to simplify the exponential terms. The maximization of posterior distribution becomes a minimization of the criterion:

$$\begin{aligned}(\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_z,\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi) &= \arg\max\left\{p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi,\boldsymbol{v}_z|\boldsymbol{g})\right\}\\ &= \arg\min\left\{-\ln p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi,\boldsymbol{v}_z|\boldsymbol{g})\right\}\\ &= \arg\min J(\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_z,\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi).\end{aligned} \tag{23}$$

We substitute the distribution formulas and obtain:

$$\begin{aligned}&J(\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_z,\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi) = -\ln p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{v}_\epsilon,\boldsymbol{v}_\xi,\boldsymbol{v}_z|\boldsymbol{g})\\ =&\frac{1}{2}\sum_i^M \ln v_{\epsilon_i} + \frac{1}{2}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})^T\boldsymbol{V}_\epsilon^{-1}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})\\ &+\frac{1}{2}\sum_j^N \ln v_{\xi_j} + \frac{1}{2}(\boldsymbol{f}-\boldsymbol{D}\boldsymbol{z})^T\boldsymbol{V}_\xi^{-1}(\boldsymbol{f}-\boldsymbol{D}\boldsymbol{z})\\ &+\frac{1}{2}\sum_j^N \ln v_{z_j} + \frac{1}{2}\boldsymbol{z}^T\boldsymbol{V}_z^{-1}\boldsymbol{z} + (\alpha_{z_0}+1)\sum_j^N \ln v_{z_j}\\ &+\beta_{z_0}\sum_j^N v_{z_j}^{-1} + (\alpha_{\epsilon_0}+1)\sum_i^M \ln v_{\epsilon_i} + \beta_{\epsilon_0}\sum_i^M v_{\epsilon_i}^{-1}\\ &+(\alpha_{\xi_0}+1)\sum_j^N \ln v_{\xi_j} + \beta_{\xi_0}\sum_j^N v_{\xi_j}^{-1}.\end{aligned} \tag{24}$$

The unknown variables are determined by obtaining the expressions of the alternate minimization in Eq. (24):

$$\widehat{\boldsymbol{f}} = \left(\boldsymbol{H}^T\widehat{\boldsymbol{V}}_\epsilon^{-1}\boldsymbol{H} + \widehat{\boldsymbol{V}}_\xi^{-1}\right)^{-1}\left(\boldsymbol{H}^t\widehat{\boldsymbol{V}}_\epsilon^{-1}\boldsymbol{g} + \widehat{\boldsymbol{V}}_\xi^{-1}\boldsymbol{D}\widehat{\boldsymbol{z}}\right), \tag{25}$$

$$\widehat{\boldsymbol{z}} = \left(\boldsymbol{D}^T\widehat{\boldsymbol{V}}_\xi^{-1}\boldsymbol{D} + \widehat{\boldsymbol{V}}_z^{-1}\right)^{-1}\boldsymbol{D}^T\widehat{\boldsymbol{V}}_\xi^{-1}\widehat{\boldsymbol{f}}, \tag{26}$$

$$\widehat{v}_{z_j} = \left(\beta_{z_0} + \frac{1}{2}\widehat{z}_j^2\right)/\left(\alpha_{z_0}+3/2\right), \tag{27}$$

$$\widehat{v}_{\epsilon i} = \left(\beta_{\epsilon_0} + \frac{1}{2}\left(g_i - \left[\boldsymbol{H}\widehat{\boldsymbol{f}}\right]_i\right)^2\right)/\left(\alpha_{\epsilon_0}+3/2\right), \tag{28}$$

$$\widehat{v}_{\xi_j} = \left(\beta_{\xi_0} + \frac{1}{2}\left(\widehat{f}_j - [\boldsymbol{D}\widehat{\boldsymbol{z}}]_j\right)^2\right)/\left(\alpha_{\xi_0}+3/2\right), \tag{29}$$

$\forall i \in [1,M]$ and $\forall j \in [1,N]$.

In 3D X-ray CT, the inversion of matrix $\left(\boldsymbol{H}^T\widehat{\boldsymbol{V}}_\epsilon^{-1}\boldsymbol{H} + \widehat{\boldsymbol{V}}_\xi^{-1}\right)^{-1}$ and $\left(\boldsymbol{D}^T\widehat{\boldsymbol{V}}_\xi^{-1}\boldsymbol{D} + \widehat{\boldsymbol{V}}_z^{-1}\right)^{-1}$ in Eq.(25) and Eq.(26) are impossible due to the large data size. First-order optimization methods are generally used in this case. In this paper we use the gradient descent algorithm:

$$\text{for } k = 1 \to I_G:\ \widehat{\boldsymbol{f}}^{(k+1)} = \widehat{\boldsymbol{f}}^{(k)} - \widehat{\gamma}_f^{(k)}\nabla\boldsymbol{J}_f(\widehat{\boldsymbol{f}}^{(k)}), \tag{30}$$

$$\text{for } k = 1 \to I_G:\ \widehat{\boldsymbol{z}}^{(k+1)} = \widehat{\boldsymbol{z}}^{(k)} - \widehat{\gamma}_z^{(k)}\nabla\boldsymbol{J}_z(\widehat{\boldsymbol{z}}^{(k)}), \tag{31}$$

where I_G is the number of iterations for the gradient descent estimation, $\nabla \boldsymbol{J}_f(\cdot)$ and $\nabla \boldsymbol{J}_z(\cdot)$ are the derivatives of the criterion (24) regarding $\boldsymbol{f}$ and $\boldsymbol{z}$ respectively. $\widehat{\gamma}_f(\cdot)$ and $\widehat{\gamma}_z(\cdot)$ are the corresponding descent step lengths which are obtained by using an optimized step length strategy [61]:

$$\nabla \boldsymbol{J}(\boldsymbol{f}) = -\boldsymbol{H}^T \boldsymbol{V}_\epsilon^{-1} (\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}) + \boldsymbol{V}_\xi^{-1} (\boldsymbol{f} - \boldsymbol{D}\boldsymbol{z}), \tag{32}$$

$$\nabla \boldsymbol{J}(\boldsymbol{z}) = -\boldsymbol{D}^T \boldsymbol{V}_\xi^{-1} (\boldsymbol{f} - \boldsymbol{D}\boldsymbol{z}) + \boldsymbol{V}_z^{-1} \boldsymbol{z}, \tag{33}$$

$$\widehat{\gamma}_f^{(k)} = \frac{\left\| \nabla \boldsymbol{J}(\widehat{\boldsymbol{f}}^{(k)}) \right\|^2}{\left\| \widehat{\boldsymbol{Y}}_\epsilon \boldsymbol{H} \nabla \boldsymbol{J}(\widehat{\boldsymbol{f}}^{(k)}) \right\|^2 + \left\| \widehat{\boldsymbol{Y}}_\xi \nabla \boldsymbol{J}(\widehat{\boldsymbol{f}}^{(k)}) \right\|^2}, \tag{34}$$

$$\widehat{\gamma}_z^{(k)} = \frac{\left\| \nabla \boldsymbol{J}(\widehat{\boldsymbol{z}}^{(k)}) \right\|^2}{\left\| \widehat{\boldsymbol{Y}}_\xi \boldsymbol{D} \nabla \boldsymbol{J}(\widehat{\boldsymbol{z}}^{(k)}) \right\|^2 + \left\| \widehat{\boldsymbol{Y}}_z \nabla \boldsymbol{J}(\widehat{\boldsymbol{z}}^{(k)}) \right\|^2}, \tag{35}$$

where $\boldsymbol{Y}_\epsilon = \boldsymbol{V}_\epsilon^{-\frac{1}{2}}$, $\boldsymbol{Y}_\xi = \boldsymbol{V}_\xi^{-\frac{1}{2}}$ and $\boldsymbol{Y}_z = \boldsymbol{V}_z^{-\frac{1}{2}}$.

The algorithm concerning the optimization of all the unknowns is given in Algorithm 1.

Algorithm 1 The JMAP algorithm for HHBM method

Fix parameters α_{z_0}, β_{z_0}, α_{ϵ_0}, β_{ϵ_0}, α_{ξ_0}, β_{ξ_0}, l
Input: $\boldsymbol{H}$, $\boldsymbol{D}$, $\boldsymbol{g}$
Output: $\widehat{\boldsymbol{f}}$, $\widehat{\boldsymbol{z}}$, $\widehat{\boldsymbol{v}}_z$, $\widehat{\boldsymbol{v}}_\epsilon$, $\widehat{\boldsymbol{v}}_\xi$
Initialization:
$\widehat{\boldsymbol{f}} \leftarrow$ normalized FBP
$\widehat{\boldsymbol{z}} \leftarrow l-$level Haar transformation of $\widehat{\boldsymbol{f}}$
for $k = 1$ to I_{max} **do**
    $\widehat{\boldsymbol{f}}^{(0)} = \widehat{\boldsymbol{f}}$
    for $k = 1$ to I_G **do**
        Calculate $\nabla \boldsymbol{J}(\widehat{\boldsymbol{f}}^{(k-1)})$ according to Eq. (32)
        Update $\widehat{\gamma}_f^{(k)}$ according to Eq. (34)
        Update $\widehat{\boldsymbol{f}}^{(k)} = \widehat{\boldsymbol{f}}^{(k-1)} - \widehat{\gamma}_f^{(k)} \nabla \boldsymbol{J}(\widehat{\boldsymbol{f}}^{(k-1)})$
    end for
    $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{f}}^{(I_G)}$
    $\widehat{\boldsymbol{z}}^{(0)} = \widehat{\boldsymbol{z}}$
    for $k = 1$ to I_G **do**
        Calculate $\nabla \boldsymbol{J}(\widehat{\boldsymbol{z}}^{(k-1)})$ according to Eq. (33)
        Update $\widehat{\gamma}_z^{(k)}$ according to Eq. (35)
        Update $\widehat{\boldsymbol{z}}^{(k)} = \widehat{\boldsymbol{z}}^{(k-1)} - \widehat{\gamma}_z^{(k)} \nabla \boldsymbol{J}(\widehat{\boldsymbol{z}}^{(k-1)})$
    end for
    $\widehat{\boldsymbol{z}} = \widehat{\boldsymbol{z}}^{(I_G)}$
    Optimize v_z according to Eq. (27)
    Optimize v_ϵ according to Eq. (28)
    Optimize v_ξ according to Eq. (29)
end for

3. Initialization and Experimental Results

For the simulations, the 3D simulated "Shepp-Logan" phantom, shown in Fig. 2 on left, Fig.3 on top and the 3D real "Head" object, shown in Fig.2 on right, Fig.3 on bottom, both of size 256^3, are used as the object of interest to compare the performance of the proposed method to the performance of the other state-of-the-art methods. Both Shepp-Logan and Head phantoms consist of several different homogeneous areas, so both are piecewise continuous. The voxel values of the original objects are normalized to $[0, 1]$. The projection directions are uniformly distributed, each projection consists of 256^2 detectors corresponding to a 256^2 size image.

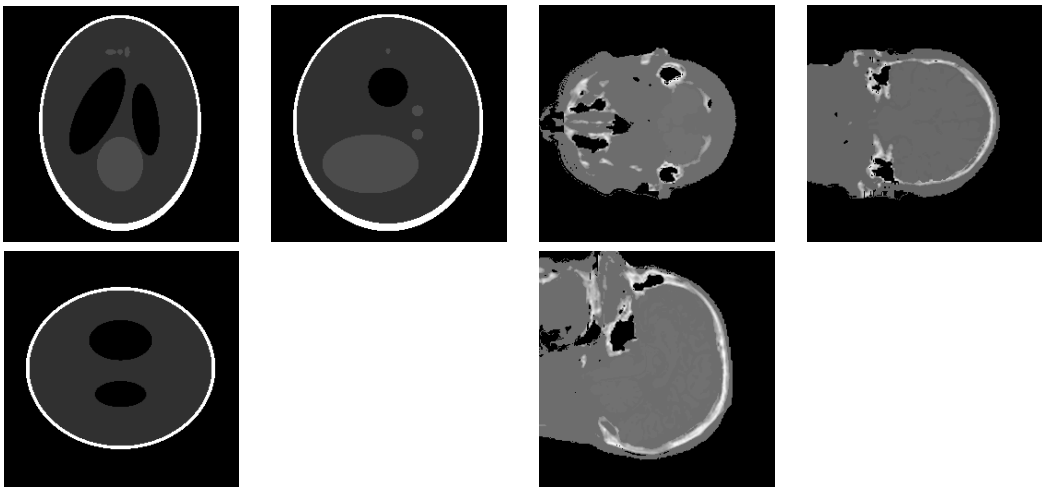


Figure 2. The three figure on left show three middle slice views of the 3 dimentional SheppLogan phantom, and the three figures on right show the 3 dementional Head phantom.

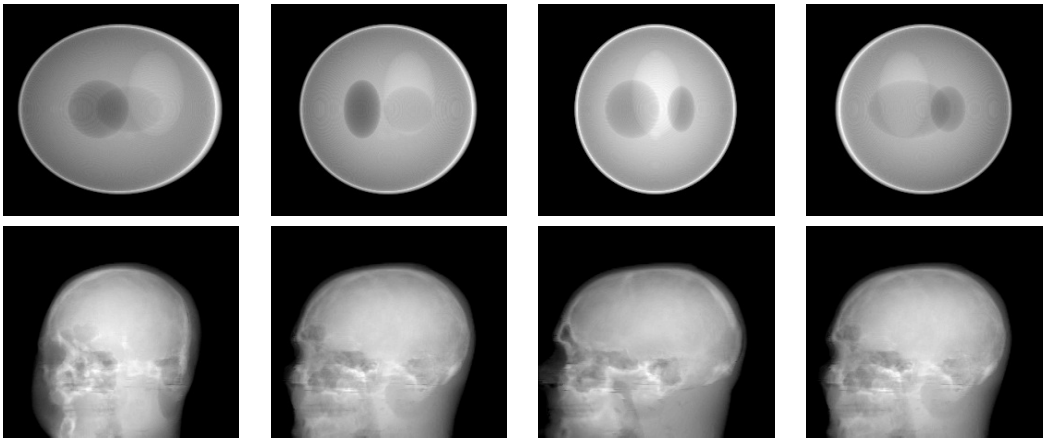


Figure 3. Top: four projections of the 3D SheppLogan phantom from 30, 60, 90, 120 degrees (left to right, respectively); bottom: four projections of the 3D Head phantom from 30, 60, 90, 120 degrees (left to right, respectively).

The proposed HHBM method is compared with the conventional Quadratic Regularization (QR) and Total Variation (TV) methods. For the QR method, the gradient descent algorithm is used for the 3D large data size problem. For the TV method, the Split-Bregman method [22] is used to solve the L_1 norm minimization problem.

To evaluate the proposed method and compare it with the state-of-the-art methods, four different metrics are used:

- the Relative Mean Square Error (RMSE), RMSE=$\left\|f-\widehat{f}\right\|^2 / \|f\|^2$ which shows a relative error of results;
- the Improvement of Signal-to-Noise Ratio (ISNR) which measures the improvement during iterations;
- the Peak Signal-to-Noise-Ratio (PSNR) which presents the SNR relative to the peak data value;
- the Structural Similarity of IMage (SSIM) [62] which evaluates the quality of result approaching human visualizing.

In 3D X-ray CT, the projection matrix $\boldsymbol{H}$ is very large and is not accessible. For the simulations, only the projection operator $\boldsymbol{Hf}$ and the back-projection operator $\boldsymbol{H}^T\boldsymbol{g}$ are used. Considering that the costly projection and back-projection operators are computed in every iteration, the GPU processor is used via the ASTRA toolbox [63] to accelerate the computation.

3.1. Initializations

The initialization for the variables f and z as well as the hyperparameters α_{ϵ_0}, β_{ϵ_0}, α_{ξ_0}, β_{ξ_0}, α_{z_0} and β_{z_0} are discussed in this section.

The reconstructed phantom obtained by using the Filtered Back Projection (FBP) method is considered to be initial value $\widehat{f}_{ini}$. The initialization of coefficient $\widehat{z}_{ini}$ is the multilevel Haar transformation of $\widehat{f}_{ini}$: $\widehat{z}_{ini} = \boldsymbol{D}^{-1}\widehat{f}_{ini}$. In this article, we choose the level of transformation such that z has a sparse structure. As shown in Fig.4, when the transform level is small, for example $2-$levels, the coefficient z is not sparse; when the transform level is sufficiently large, the coefficient is sparse. In this paper, we set z as a $5-$level Haar transform coefficient.

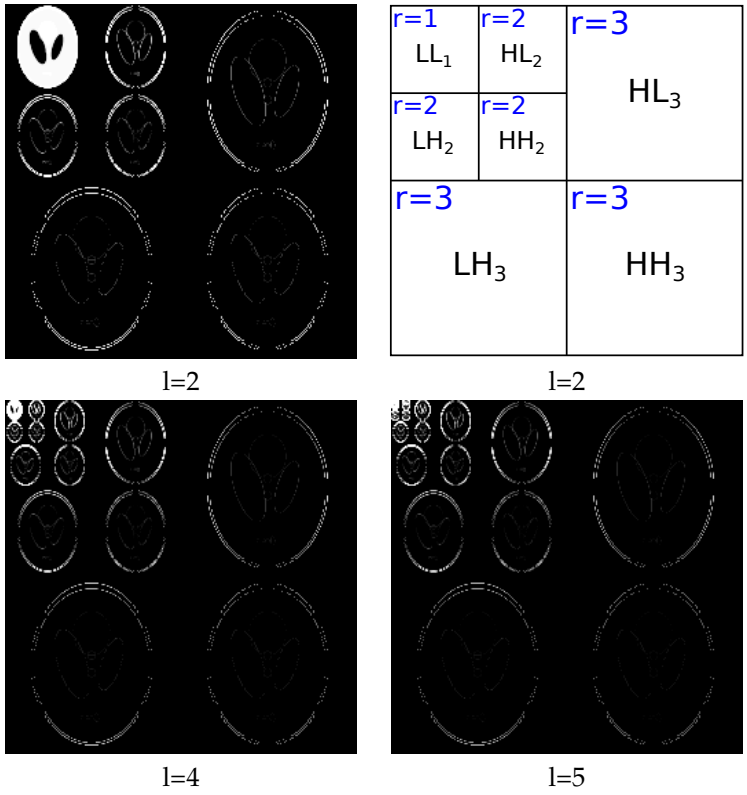


Figure 4. Slice of 3D multilevel Haar transformation coefficient z of the 256^3 Shepp-Logan phantom with: 2 levels (top-left), 4 levels (bottom-left) and 5 levels (bottom-right). The figure on the top-right shows the ranks of coefficients for a 2-level transformation.

The initialization for α_{z_0} and β_{z_0} is based on the sparse structure of the variable z. In Figure 4, we can see that the sparsity rate depends on the rank of transform coefficient r, where $r \in [1,\ l+1]$.

For example when $l = 2$, shown in Figure 4, the coefficient has 3 ranks, $r \in [1,\ 2,\ 3]$. The first rank $r = 1$ corresponds to the low frequency components in the transform coefficient.

The variable $\boldsymbol{z}$ is modeled by a $\mathcal{S}t_g$ distribution, with variance equal to $\text{Var}[z_j|\alpha_{z_0}, \beta_{z_0}] = \beta_{z_0}/(\alpha_{z_0} - 1),\ \forall j \in [0, N]$. In this article, we fix the value $\alpha_{z_0} = 2.1$ in order to have $\text{Var}[z_j|\alpha_{z_0} = 2.1, \beta_{z_0}] \approx \beta_{z_0},\ \forall j \in [0, N]$. The sparsity rate of $\boldsymbol{z}$ is defined by initializing different values for β_{z_0}, with a sparser structure when β_{z_0} is smaller. β_{z_0} is initialized as $\beta_{z_0} = 10^{-r+1}$. When $l = 5$, we have $r = [1, 2, 3, 4, 5, 6]$, and the hyperparameters $\beta_{z_0} = [10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$ respectively.

The initialization for the hyperparameters, $\alpha_{\epsilon_0},\ \beta_{\epsilon_0},\ \alpha_{\xi_0}$ and β_{ξ_0}, is based on the prior models of the variances $\boldsymbol{v}_\epsilon$ and $\boldsymbol{v}_\xi$ we have chosen. In the proposed method, we consider the background of the generalized Student-t distribution, in which both $\boldsymbol{\epsilon}$ and $\boldsymbol{\xi}$ are modeled by a Gaussian distribution with Inverse Gamma distributed variance, i.e. the $\mathcal{S}t_g$ distribution according to Eq. (13).

The noise $\boldsymbol{\epsilon}$ depends on the SNR of the dataset. In order to exploit this information in the initialization, we express the biased dataset as the sum of uncontaminated dataset $\boldsymbol{g}_0$ and the additive noise $\boldsymbol{\epsilon}$:

$$\boldsymbol{g} = \boldsymbol{g}_0 + \boldsymbol{\epsilon}. \tag{36}$$

As the noise $\boldsymbol{\epsilon}$ and the uncontaminated data $\boldsymbol{g}_0$ are supposed to be independent, we have:

$$\|\boldsymbol{g}\|^2 = \|\boldsymbol{g}_0\|^2 + \|\boldsymbol{\epsilon}\|^2. \tag{37}$$

And the SNR of the dataset is:

$$\text{SNR} = 10\log\frac{\|\boldsymbol{g}_0\|^2}{\|\boldsymbol{\epsilon}\|^2} = 10\log\frac{\|\boldsymbol{g}\|^2 - \|\boldsymbol{\epsilon}\|^2}{\|\boldsymbol{\epsilon}\|^2}. \tag{38}$$

With $\mathbb{E}[\boldsymbol{\epsilon}] = 0$, we have:

$$v_\epsilon = \mathbb{E}\left[\epsilon^2\right] \approx \frac{\|\boldsymbol{\epsilon}\|^2}{M} = \frac{\|\boldsymbol{g}\|^2}{M} \times \frac{1}{1 + 10^{SNR/10}}. \tag{39}$$

The mean of variance $\boldsymbol{v}_\epsilon$ of the noise $\boldsymbol{\epsilon}$ is $\mathbb{E}[v_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0}] = \beta_{\epsilon_0}/(\alpha_{\epsilon_0} - 1),\ \forall i \in [0, M]$, so we obtain:

$$\beta_{\epsilon_0} = \frac{\|\boldsymbol{g}\|^2}{M} \times \frac{1}{1 + 10^{SNR/10}} \times (\alpha_{\epsilon_0} - 1). \tag{40}$$

The two hyperparameters α_{ϵ_0} and β_{ϵ_0} are combined according to Eq. (40); hence, initialization for one of them is sufficient. In real applications, the SNR of the dataset is unknown, but we can use the projection of an empty object, i.e. $\boldsymbol{f} = \boldsymbol{0}$, to obtain a rough value of the variance of noise $\boldsymbol{v}_\epsilon$.

Fig.5 shows the influence of the value of α_{ϵ_0} on the reconstruction. According to the results, a bigger value for α_{ϵ_0} results in a smaller value on RMSE for different numbers of projections and SNR of the dataset. This monotonous property facilitates the initialization of this hyperparameter, as a large value for α_{ϵ_0} satisfies all cases. When α_{ϵ_0} is greater than a threshold value, the RMSE does not change with different initialization values for α_{ϵ_0}.

For $\boldsymbol{\xi}$, both α_{ξ_0} and β_{ξ_0} are analyzed for the influence of the reconstruction results.

Fig.6 shows the influence of the hyperparameter α_{ξ_0}. Different colors represent an initialization with a different value of β_{ξ_0}. As here we focus on the analysis of α_{ξ_0} for all cases of β_{ξ_0} values, we don't show the corresponding β_{ξ_0} value for each different color. For different noise levels, different numbers of projections and different β_{ξ_0} values, the RMSE has an upward trend when the value of α_{ξ_0} becomes larger.

Fig.7 shows the influence of the hyperparameter β_{ξ_0}. Different colors represent an initialization with a different hyperparameter α_{ξ_0}. For different noise levels, different numbers of projections and different values for α_{ξ_0}, when the value of β_{ξ_0} increases, the RMSE decreases; then, after a slight increase, levels out.

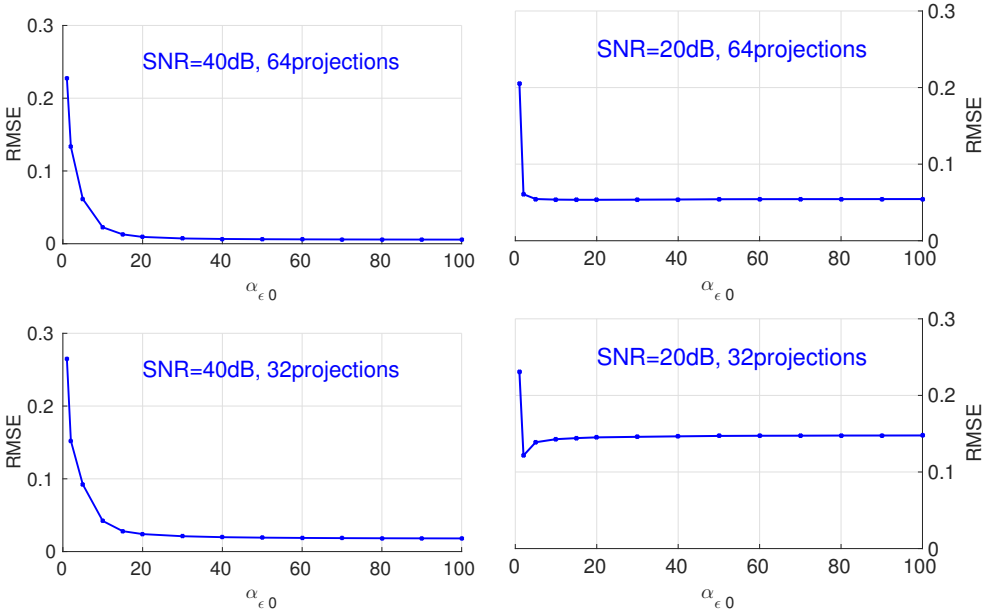


Figure 5. Influence of hyperparameter α_{ϵ_0} on RMSE of final reconstruction results for different numbers of projections and noise.

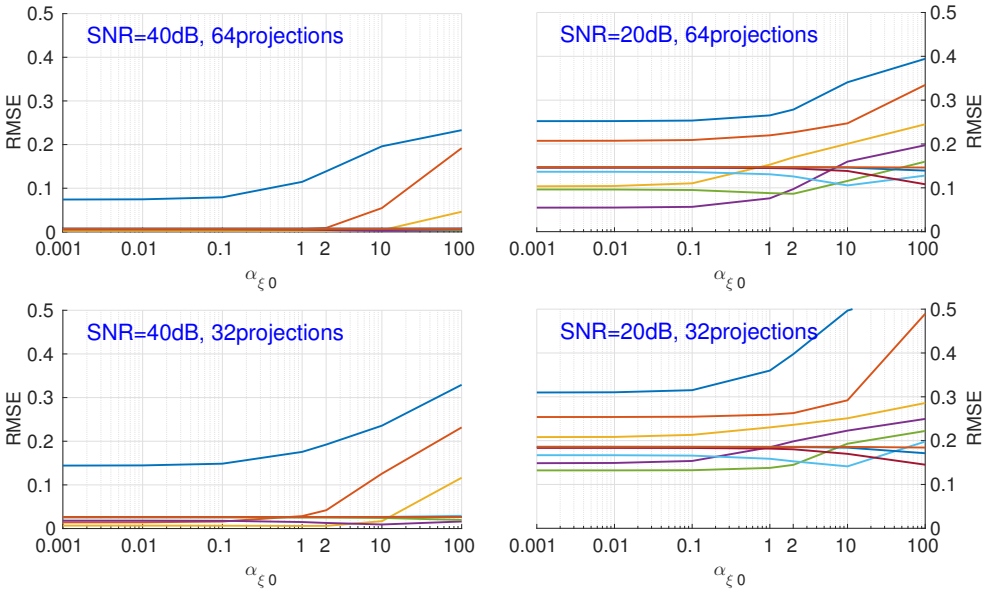


Figure 6. Influence of hyperparameter α_{ξ_0}, with different fixed values of β_{ξ_0}, on RMSE of reconstruction results for different number of projections and noise. Each color corresponds to a different initialization of β_{ξ_0}.

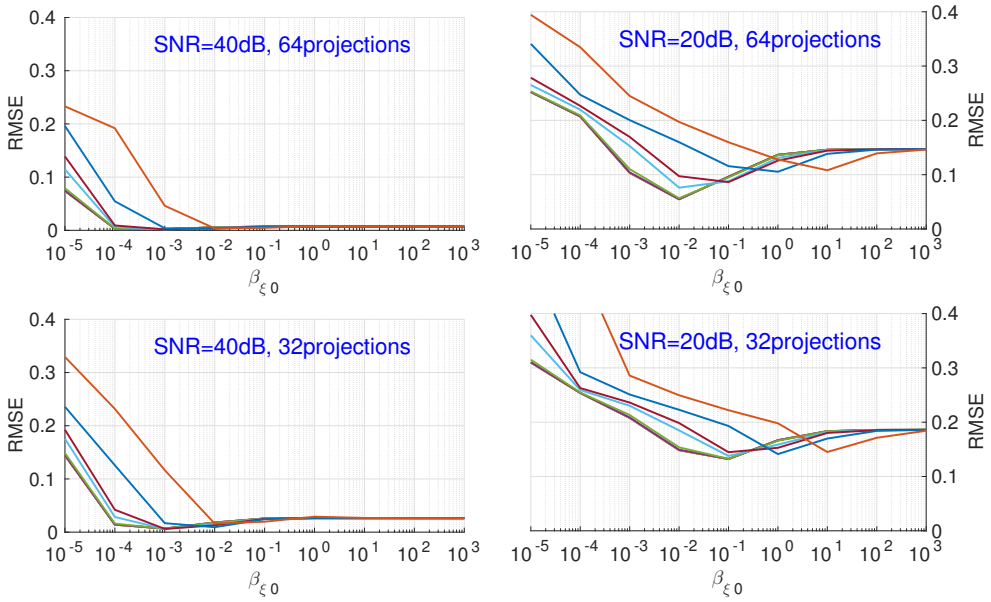


Figure 7. Influence of hyperparameter β_{ξ_0}, with different fixed value of α_{ξ_0}, on RMSE of reconstruction results for different number of projections and noise. Each color corresponds to a different initialization of α_{ξ_0}.

In [46], it is pointed out that when α and β of the $\mathcal{S}t_g$ distribution are both large, the $\mathcal{S}t_g$ distribution approaches a Gaussian distribution, which is the case for the additive noise $\boldsymbol{\epsilon}$. If α and β are very small (approaching 0), the $\mathcal{S}t_g$ distribution becomes an non-informative distribution (Jeffreys distribution); when α and β are both small, the $\mathcal{S}t_g$ has the sparsity enforcing property, which is the case for the sparse $\boldsymbol{\xi}$. Consequently, the initialization of the hyperparameters are theoretically supported, and they can be initialized with respect to these properties in other simulations.

Table 1. Comparison of RMSE, ISNR, PSNR and SSIM of reconstructed phantom with 50 global iterations (10 gradient descent iterations in each global iteration). The values of the regularization parameters are respectively $\lambda_{QR} = 10$ and $\lambda_{TV} = 50$ for SNR=40dB, $\lambda_{QR} = 600$ and $\lambda_{TV} = 100$ for SNR=20dB.

	$256 \times 256 \times 256$ Shepp-Logan phantom											
	180 projections						90 projections					
	40dB			20dB			40dB			20dB		
	QR	TV	HHBM	QR	TV	HHBM	QR	TV	HHBM	QR	TV	HHBM
RMSE	0.0236	0.0114	**0.0069**	0.1309	**0.0209**	0.0755	0.0401	0.0212	**0.0092**	0.1558	**0.0491**	0.1117
ISNR	5.5584	8.7217	**10.9346**	7.2024	**15.1775**	10.2162	6.6136	9.3832	**12.9973**	8.4583	**13.4765**	9.9056
PSNR	30.0675	33.2308	**35.4437**	22.6318	**30.6069**	25.0209	27.7743	30.5439	**34.1579**	21.8754	**26.8937**	23.3227
SSIM	0.9999	0.9999	**1.0000**	0.9992	**0.9999**	0.9995	0.9997	0.9999	**0.9999**	0.9990	**0.9997**	0.9993
	60 projections						45 projections					
	40dB			20dB			40dB			20dB		
	QR	TV	HHBM	QR	TV	HHBM	QR	TV	HHBM	QR	TV	HHBM
RMSE	0.0636	0.0321	**0.0107**	0.1656	**0.0753**	0.1293	0.0904	0.0474	**0.0132**	0.1854	0.0901	**0.1414**
ISNR	9.3826	12.3480	**17.1346**	9.1492	**12.5701**	10.2226	10.3301	13.1308	**18.6839**	10.0137	13.1476	**11.1916**
PSNR	25.7693	28.7347	**33.5214**	21.6116	**25.0325**	22.6849	24.2404	27.0412	**32.5942**	21.1195	24.2535	**22.2974**
SSIM	0.9996	0.9995	**0.9999**	0.9990	**0.9995**	0.9992	0.9994	0.9997	**0.9999**	0.9988	0.9994	**0.9991**
	36 projections						18 projections					
	40dB			20dB			40dB			20dB		
	QR	TV	HHBM	QR	TV	HHBM	QR	TV	HHBM	QR	TV	HHBM
RMSE	0.1177	0.0680	**0.0169**	0.1957	**0.1116**	0.1500	0.2581	0.2104	**0.0574**	0.2907	0.2313	**0.2014**
ISNR	10.6591	13.0424	**19.0933**	10.8633	**13.3032**	12.0187	10.7122	11.5992	**17.2373**	10.8088	11.8022	**12.4036**
PSNR	23.0949	25.4783	**31.5292**	20.8865	**23.3264**	22.0420	19.6263	20.5133	**26.1514**	19.1085	20.1020	**20.7033**
SSIM	0.9993	0.9996	**0.9999**	0.9988	**0.9993**	0.9990	0.9983	0.9987	**0.9996**	0.9981	0.9985	**0.9987**

3.2. Simulation results with limited number of projections

We apply 180, 90, 60, 45, 36 and 18 projections evenly distributed in $[0, 180]$ degrees for the reconstruction of the 3D Shepp-Logan phantom of size 256^3; each projection contains 256^2 detectors. The number of projections are chosen such that there is respectively one projection every 1, 2, 3, 4, 5 and 10 degrees.

In Table 1, different evaluation metrics of the reconstructed 256^3 Shepp-Logan phantom are compared. It is shown that the HHBM method does not always perform better than the TV method, especially when there are sufficient numbers of projections. But when there is insufficient projection data, the HHBM method is more robust than the TV method. On the other hand, as it is known that the choice of regularization parameter plays an important role in the regularization methods like QR or TV, and the value for the regularization parameter should be selected for each different system settings, the HHBM method is much more robust on the initialization of hyper-parameters. As we can see from Fig.5, Fig.6 and Fig.7, once we have chosen the hyperparameters in a certain interval, which is not difficult to fix according to the properties of the prior model, we can obtain the appropriate reconstruction results. More importantly, in the Bayesian approach, the prior model can be chosen from a variety of other suitable distributions, which gives more possibilities for the models than the conventional regularization methods. We may also choose different point estimators from the posterior distribution, for example the Posterior Mean, etc.

Figure 8 and Figure 9 show the reconstructed middle slice of the "Shepp-Logan" phantom and "Head" object by using the TV and HHBM methods from 36 projections with SNR=40dB and SNR=20dB. The red curve illustrates the profile of the blue line position. In the reconstructed Shepp-Logan phantom obtained using the TV method, the three small circles on the top of the slice are not evident. By using the HHBM method, we can distinguish these three small circles. By comparing the profiles of the slice of reconstructed Shepp Logan phantom, we can see that by using the HHBM method, the contour positions on the profile are closer to the original profile than those obtained using the TV method. In the reconstructed Head object, there are more details than the simulated Shepp-Logan phantom, especially in the zoom area in the second line in Figure 8. By comparing the results, we can see that for the type of object which contains some small details, the TV method derives a result with smoother homogeneous areas but with fewer details in the contour areas than the HHBM method. Some of the white material, which is dispersed into discontinuous small blocks in the Head object, is connected in the results of the TV method. From these images we conclude that with an insufficient number of projections the proposed method gives results with clearer contours and details.

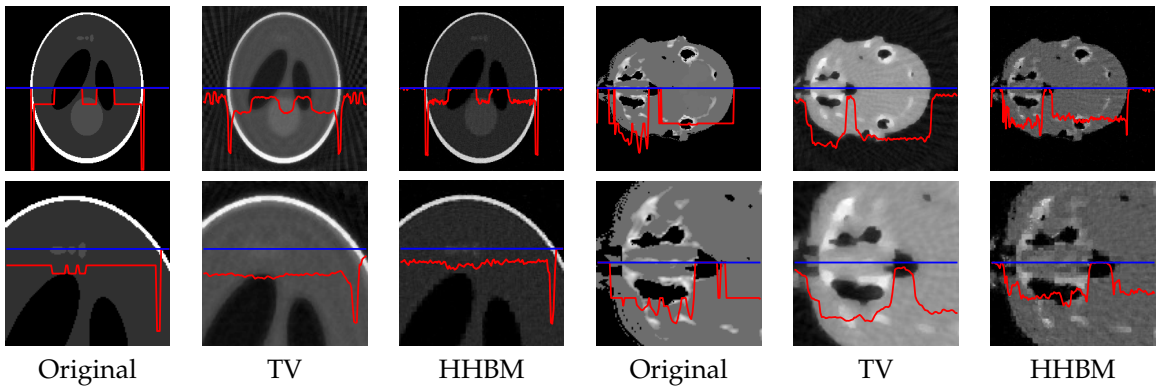


Figure 8. Reconstructed "Shepp-Logan" phantom and "Head" object of size 256^3, with a dataset of **36 projections** and **SNR=40dB**, by using the TV and HHBM methods. Bottom figures are zones of the corresponding top figures. The red curves are the profiles at the position of the corresponding blue lines.

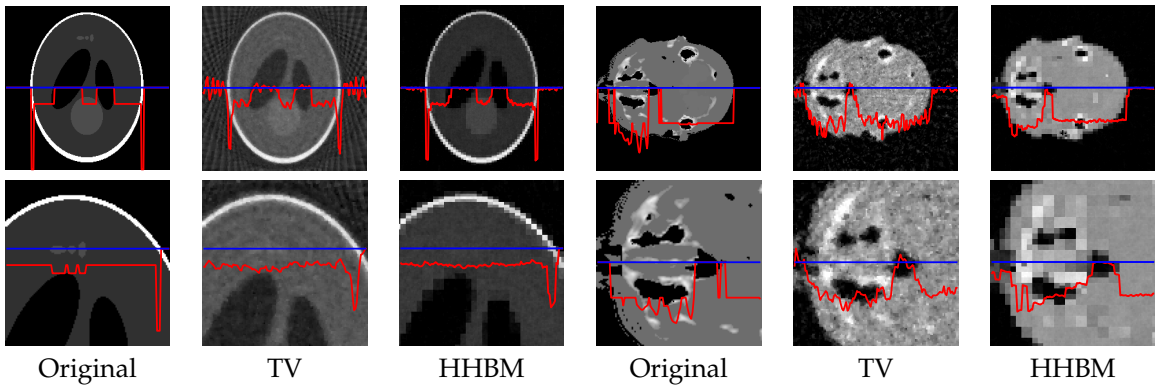


Figure 9. Reconstructed "Shepp-Logan" phantom and "Head" object of size 256^3, with a dataset of **36 projections** and **SNR=20 dB**, by using the TV and HHBM methods. Bottom figures are zones of the corresponding top figures. The red curves are the profiles at the position of the corresponding blue lines.

Fig.10 shows the reconstructed Shepp-Logan phantom from 18 projections of SNR=40dB and 20dB. In this very underdetermined case, the HHBM method can still obtain a result which is clear enough to distinguish the primary zones and contours of the object.

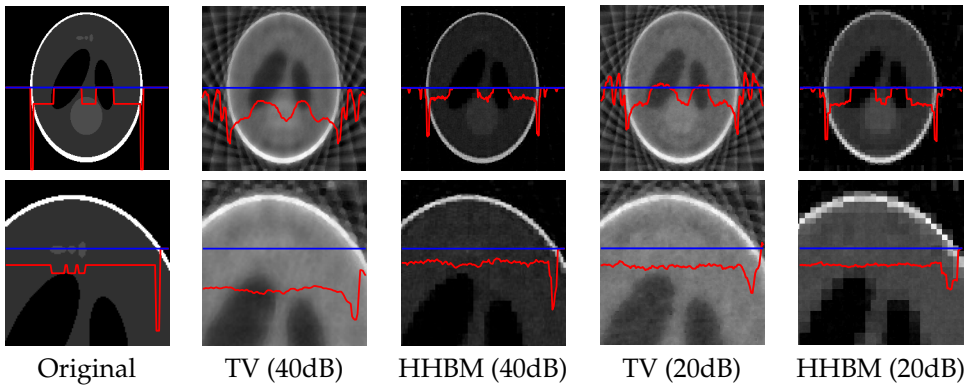


Figure 10. Reconstructed Shepp-Logan phantom of size 256^3, with a dataset of **18 projections** of **SNR=40dB and 20dB**, by using the TV and HHBM methods. The red curves are the profiles at the position of the corresponding blue lines.

Fig.11 and Fig.12 show the comparison between the QR, TV and HHBM methods with a high SNR=40dB and a low SNR=20dB dataset respectively. The abscissa corresponds to the number of projections evenly distributed from 0° to 180°, and the ordinate is the RMSE after 50 iterations. In the simulations we used a SNR=40dB to represent a weak noise case and SNR=20dB for a strong noise case. When SNR=40dB, the HHBM method outperforms both quadratic regularization and the TV method. When SNR=20dB, the TV method with the optimal regularization parameter outperforms the HHBM method. However, in Fig.12, we show another curve in light green (TV2) showing the TV reconstruction with some random regularization parameters, which are chosen as a value not far from the optimal regularization parameters. From these results we can see that the HHBM method is more robust than the TV reconstruction method with respect to the regularization parameter, the optimal value of which is, on the other hand, difficult to determine in the real applications where we cannot evaluate the estimation quality.

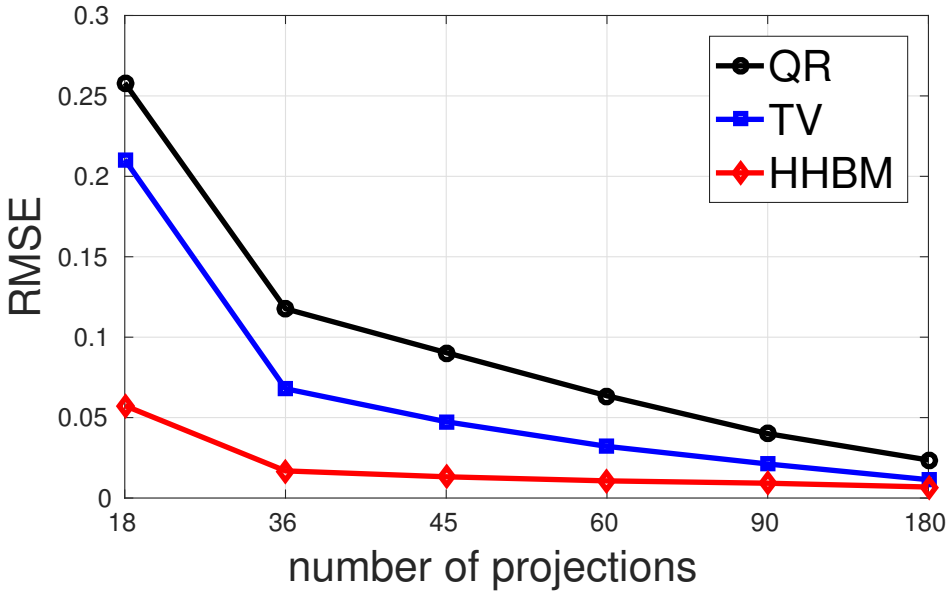


Figure 11. The performances of different methods for reconstructing Shepp-Logan phantom in terms of RMSE with different numbers of projections evenly distributed in $[0°, 180°]$ and a high **SNR=40dB**.

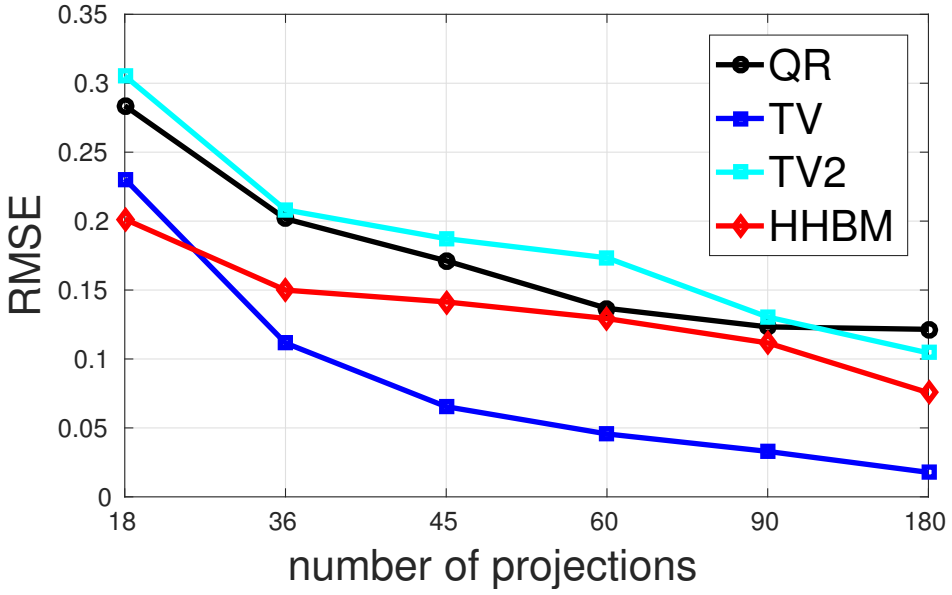


Figure 12. The performances of different methods for reconstructing Shepp-Logan phantom in terms of RMSE with different numbers of projections evenly distributed in $[0°, 180°]$ and a low **SNR=20dB**.

3.3. Simulation results with limited angle of projections

In both medical and industrial X-ray CT, another common challenge is the limit of projection angles. In this part of the simulation, we use evenly distributed projections in a limited range of angles for the simulated 3D "Shepp-Logan" phantom and the 3D "Head" object, both of which have a size of 256^3.

Figure 13 and Figure 14 show the middle slice of the reconstructed Shepp-Logan phantom and the Head object, from 90 projections distributed between 0° and 90°, with projection SNR of 40dB and 20dB. By using the TV method, the reconstructed object is blurry along the the diagonal direction for which there is no projection data, and there is a square corner where the object should have a rounded edge. By using the HHBM method, we get results that are more consistent with the true shape and clearer contours.

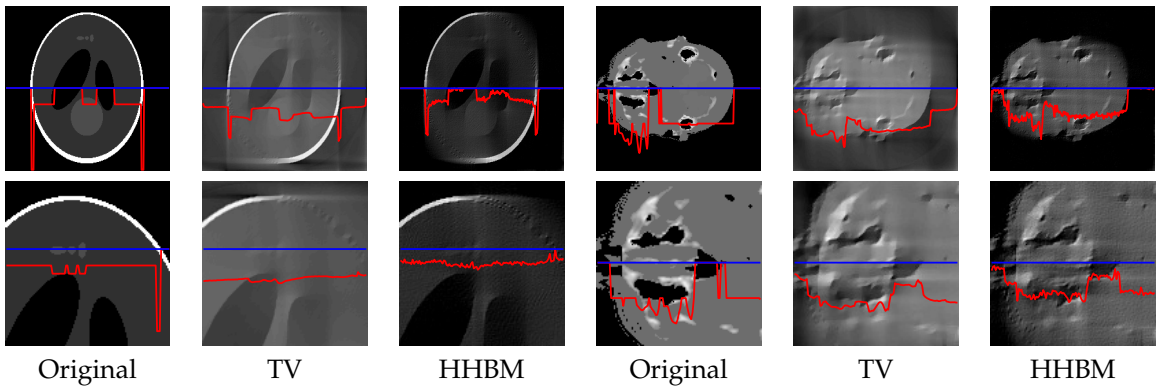


Figure 13. Slice of reconstructed 3D Shepp-Logan phantom and 3D Head object, with **90 projections** evenly distributed in $[0^\circ, 90^\circ]$, **SNR=40dB**. Bottom figures are parts of the corresponding top figures. The red curves are the profiles at the position of the corresponding blue lines.

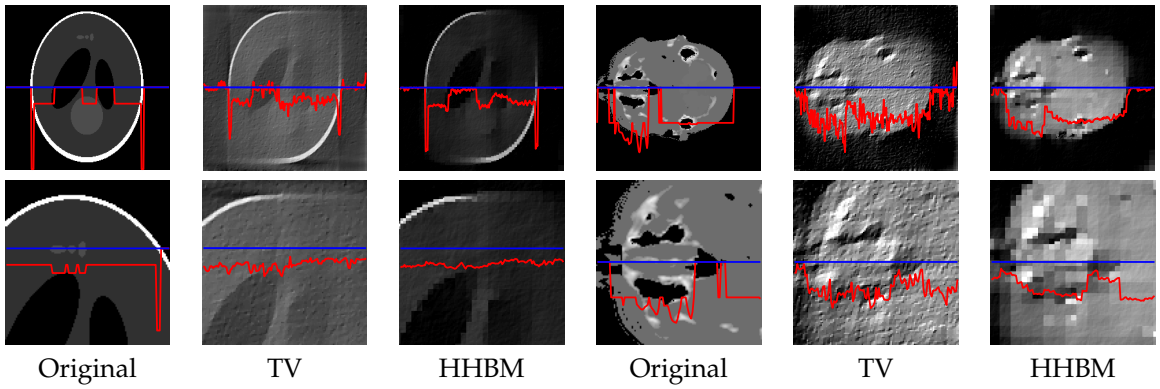


Figure 14. Slice of reconstructed 3D Shepp-Logan phantom and 3D Head object, with **90 projections** evenly distributed in $[0^\circ, 90^\circ]$, **SNR=20dB**. Bottom figures are parts of the corresponding top figures. The red curves are the profiles at the position of the corresponding blue lines.

Figure 15 and Figure 16 show the comparison of the performance in terms of RMSE of different methods with a high SNR=40dB and a low SNR=20dB. In this comparison, four cases of limited projection angles are considered and they are: 45, 90, 135 and 180 projections evenly distributed in $[0^\circ, 45^\circ]$, $[0^\circ, 90^\circ]$, $[0^\circ, 135^\circ]$ and $[0^\circ, 180^\circ]$ respectively. There is one projection every 1°. From these two figures, we conclude that the proposed HHBM method remains more robust than the other two conventional methods when there are limited projection angles.

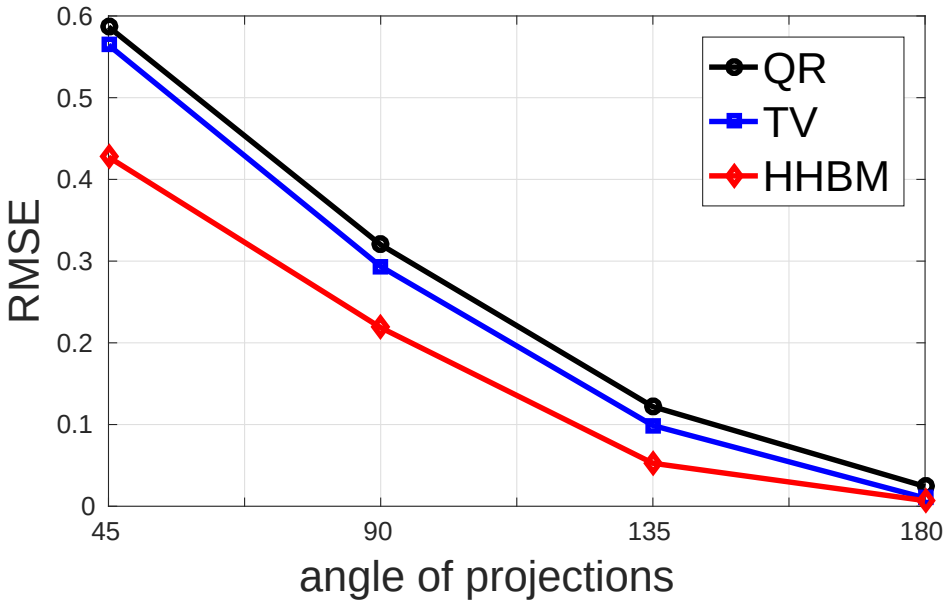


Figure 15. The performance of different methods for reconstructing Shepp-Logan phantom in terms of RMSE with different **limited projection angles** and a high **SNR=40dB**.

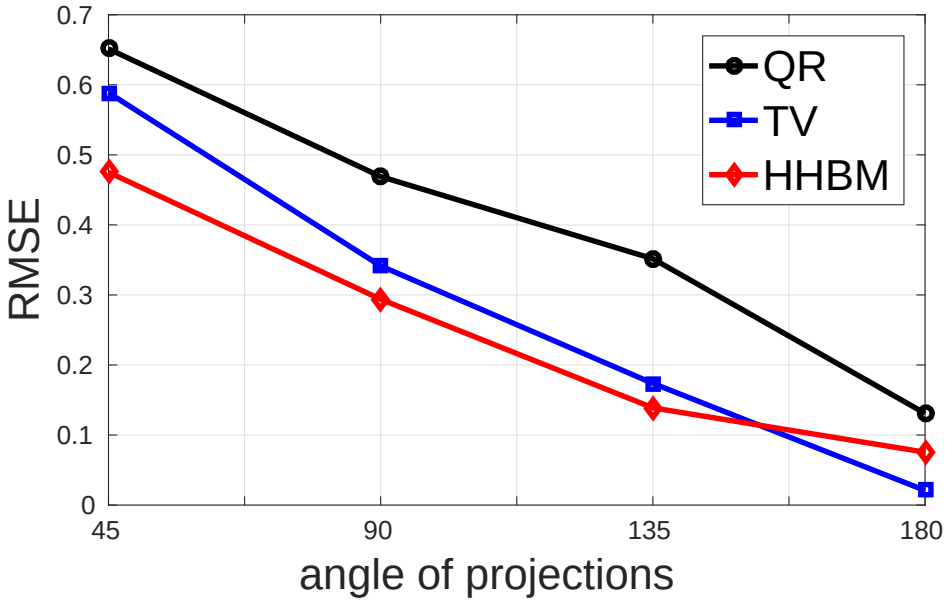
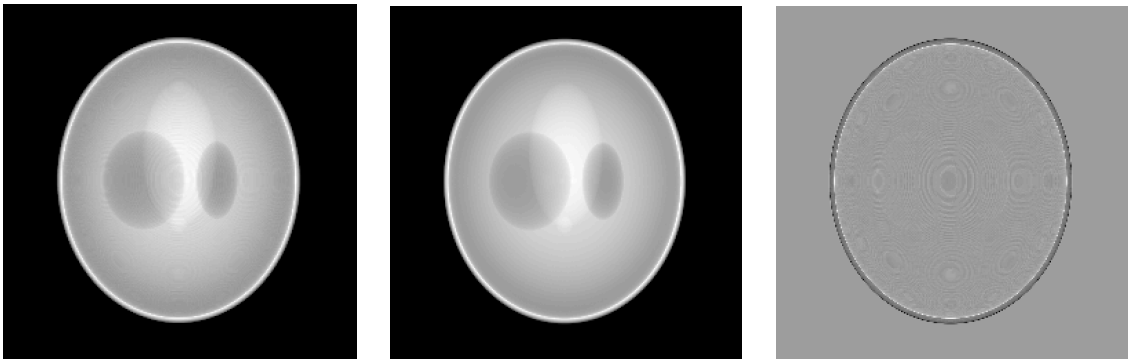


Figure 16. The performance of different methods for reconstructing Shepp-Logan phantom in terms of RMSE with different **limited projection angles** and a low **SNR=20dB**.

3.4. Simulation with a different forward model

In order to avoid the inverse crime, we apply a slightly different forward model. During the projection, the projector is applied on the Shepp-Logan phantom of size 1024^3, with the detector size of 256^2 for each projection direction.

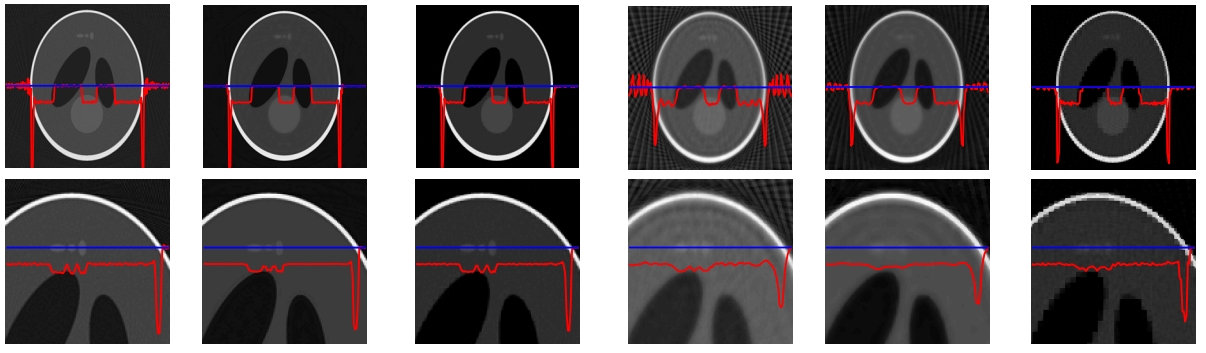
In Figure 17, we show the results of the two projection data obtained as well as their difference which can be considered as the forward modelling error. We then use the data obtained from 1024^3 phantom to reconstruct an object of size 256^3.



Proj from 256^3 phantom | Proj from 1024^3 phantom | difference

Figure 17. The projection image in angle 90 degree by using two different forward model: projection from **256^3 phantom** (left) and projection from **1024^3 phantom** (middle). The difference between them is on the right.

In Figure 18, the middle slice of the reconstructed Shepp-Logan phantom by using QR, TV and HHBM methods are presented, by using 180 and 36 projections respectively. From the figures, we can see that when there are 180 projections, all three methods performs well, and the TV method detects better the contours while HHBM method has more noise at the contour areas. When there are insufficient projection numbers (36 projections in this simulation), the HHBM method outperforms the QR and TV methods for reconstructing the details in the phantom, for example the three small circles in the top of the phantom.



QR (180proj) | TV (180proj) | HHBM (180proj) | QR (36proj) | TV (36proj) | HHBM (36proj)

Figure 18. Reconstructed phantom with different forward model and **180 projections** and **36 projections** by using QR (left), TV (middle) and HHBM (right) methods.

In Table 2, the RMSE of the reconstructed phantom by using the different methods are compared. We can conclude that, when the projector model is different than the reconstruction one, all these three methods (QR, TV and HHBM) have good performance when there is 180 projections. When the projection number decreases, TV and HHBM methods outperforms QR method. Comparing with

the TV method, HHBM method is more robust to the number of projections. When the projection number is smaller than 60, HHBM method outperforms TV method.

Table 2. The RMSE of reconstructed 256^3 Shepp-Logan phantom by using projection obtained from a 1024^3 Shepp-Logan phantom.

RMSE	QR	TV	HHBM
180 proj	0.0581	0.0540	0.0558
90 proj	0.0655	0.0573	0.0610
60 proj	0.0846	0.0675	0.0690
45 proj	0.1079	0.0830	0.0783
36 proj	0.1326	0.1027	0.0882

All the MATLAB codes for the simulations in this paper can be found on github [64].

4. Discussion

One advantage of the Bayesian approach is the estimation of the parameters along with the estimation of unknown variables of the forward model at each iteration. However, like in regularization methods, the hyper-parameters need to be initialized.

While the parameters in the regularization methods play an important role in the final results and they are costly to fix, the hyper-parameters in HHBM can be initialized based on the prior information (the sparse structure of z) and the prior model (the Student-t distribution). In this article, we have shown that once the hyper-parameters are fixed in a certain appropriate interval, which is not difficult to obtain, the corresponding algorithm is robust. In this work, the hyper-parameters are not fixed via the classical approach using non-informative prior laws (i.e. considering the Inverse Gamma corresponding parameters such that it approaches Jeffreys), [65].

5. Conclusions

In this paper, we propose a Bayesian method with a hierarchical structured prior model based on multilevel Haar transformation (HHBM) for 3D X-ray CT reconstruction. Simulation results indicate that for a limited number of projections or limited projection angles, the proposed method is more robust to noise and to regularization parameters than the classical QR and TV methods.

Indeed, we observe a relatively weak influence of the hyper-parameters in the behavior of the corresponding iterative algorithm. The interest of this weak dependency is that it offers a practical way to ensure the initialization of the algorithm which typically is not-trivial.

In this context, as future work, we are investigating the causes of the relative weak influence of the hyper-parameters and the theoretical foundation of the corresponding robust interval, extending the discussion to the same approach using sparsity enforcing priors expressed as Normal variance mixtures, but for other mixing distributions (Gamma, generalized inverse Gaussian), [66].

Another extension of this work is to consider the Posterior Mean as an estimator. This can be done via the Variational Bayesian Approach (VBA) but a practical implementation requires a method of accessing the diagonal elements of the large matrix $\boldsymbol{H}^T\boldsymbol{H}$, which is being studied by our group.

Author Contributions: This work has been done during the PhD preparation of the first author. Ali Mohammad-Djafari and Nicolas Gac have been the co-supervisors of the PhD work. Mircea Dumitru was a Post-Doc who also had finished his PhD under the supervision of Ali Mohammad-Djafari and was continuing to work on this project.

Conflicts of Interest: The authors declare no conflict of interest.

Bibliography

1. Arridge, S.R. Optical tomography in medical imaging. *Inverse Problems* **1999**, *15*, R41.

2. Hanke, R.; Fuchs, T.; Uhlmann, N. X-ray based methods for non-destructive testing and material characterization. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **2008**, *591*, 14–18.
3. Kalender, W.A. X-ray Computed Tomography. *Physics in medicine and biology* **2006**, *51*, R29.
4. Natterer, F.; Wubbeling, F. A propagation-backpropagation method for ultrasound tomography. *Inverse Problems* **1995**, *11*, 1225.
5. Bolomey, J.C.; Pichot, C. Microwave tomography: from theory to practical imaging systems. *International Journal of Imaging Systems and Technology* **1990**, *2*, 144–156.
6. Nelson, J.; Milner, T.; Tanenbaum, B.; Goodman, D.; Van Gemert, M. Infra-red tomography of port-wine-stain blood vessels in human skin. *Lasers in Medical Science* **1996**, *11*, 199–204.
7. Deans, S.R. *The Radon transform and some of its applications*; Courier Corporation, 2007.
8. Feldkamp, L.; Davis, L.; Kress, J. Practical cone-beam algorithm. *JOSA A* **1984**, *1*, 612–619.
9. Alifanov, O.M.; Artioukhine, E.A.; Rumyantsev, S.V. *Extreme methods for solving ill-posed problems with applications to inverse heat transfer problems*; Begell house New York, 1995.
10. O'Sullivan, F. A statistical perspective on ill-posed inverse problems. *Statistical Science* **1986**, pp. 502–518.
11. Gordon, R. A tutorial on ART (algebraic reconstruction techniques). *IEEE Transactions on Nuclear Science* **1974**, *21*, 78–93.
12. Gordon, R.; Bender, R.; Herman, G.T. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *Journal of Theoretical Biology* **1970**, *29*, 471–481.
13. Andersen, A.H.; Kak, A.C. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrasonic Imaging* **1984**, *6*, 81–94.
14. Trampert, J.; Leveque, J.J. Simultaneous iterative reconstruction technique: physical interpretation based on the generalized least squares solution. *J. geophys. Res* **1990**, *95*, 553–9.
15. Bangliang, S.; Yiheng, Z.; Lihui, P.; Danya, Y.; Baofen, Z. The use of simultaneous iterative reconstruction technique for electrical capacitance tomography. *Chemical Engineering Journal* **2000**, *77*, 37–41.
16. Coric, S.; Leeser, M.; Miller, E.; Trepanier, M. Parallel-beam backprojection: an FPGA implementation optimized for medical imaging. Proceedings of the 2002 ACM/SIGDA tenth international symposium on Field-programmable gate arrays. ACM, 2002, pp. 217–226.
17. Ay, M.R.; Zaidi, H. Development and validation of MCNP4C-based Monte Carlo simulator for fan-and cone-beam x-ray CT. *Physics in Medicine and Biology* **2005**, *50*, 4863.
18. Mozzo, P.; Procacci, C.; Tacconi, A.; Tinazzi Martini, P.; Bergamo Andreis, I. A new volumetric CT machine for dental imaging based on the cone-beam technique: preliminary results. *European Radiology* **1998**, *8*, 1558–1564.
19. Ben-Israel, A.; Greville, T.N. *Generalized inverses: theory and applications*; Vol. 15, Springer Science & Business Media, 2003.
20. Calvetti, D.; Morigi, S.; Reichel, L.; Sgallari, F. Tikhonov regularization and the L-curve for large discrete ill-posed problems. *Journal of Computational and Applied Mathematics* **2000**, *123*, 423–446.
21. Chambolle, A.; Lions, P.L. Image recovery via total variation minimization and related problems. *Numerische Mathematik* **1997**, *76*, 167–188.
22. Goldstein, T.; Osher, S. The split Bregman method for L1-regularized problems. *SIAM Journal on Imaging Sciences* **2009**, *2*, 323–343.
23. Sidky, E.Y.; Pan, X. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine and Biology* **2008**, *53*, 4777.
24. Chan, T.F.; Golub, G.H.; Mulet, P. A nonlinear primal-dual method for total variation-based image restoration. *SIAM journal on scientific computing* **1999**, *20*, 1964–1977.
25. Wahlberg, B.; Boyd, S.; Annergren, M.; Wang, Y. An ADMM algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes* **2012**, *45*, 83–88.
26. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2005**, *67*, 301–320.
27. Ramani, S.; Liu, Z.; Rosen, J.; Nielsen, J.F.; Fessler, J.A. Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods. *IEEE Transactions on Image Processing* **2012**, *21*, 3659–3672.

28. Galatsanos, N.P.; Katsaggelos, A.K. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Transactions on Image Processing* **1992**, *1*, 322–336.
29. Wang, G.; Schultz, L.; Qi, J. Statistical image reconstruction for muon tomography using a Gaussian scale mixture model. *IEEE Transactions on Nuclear Science* **2009**, *56*, 2480–2486.
30. Redner, R.A.; Walker, H.F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **1984**, *26*, 195–239.
31. Moon, T.K. The expectation-maximization algorithm. *IEEE Signal Processing Magazine* **1996**, *13*, 47–60.
32. Tregouet, D.; Escolano, S.; Tiret, L.; Mallet, A.; Golmard, J. A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. *Annals of human genetics* **2004**, *68*, 165–177.
33. Hudson, H.M.; Larkin, R.S. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Transactions on Medical Imaging* **1994**, *13*, 601–609.
34. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian data analysis*; Vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA, 2014.
35. Bali, N.; Mohammad-Djafari, A. Bayesian approach with hidden Markov modeling and mean field approximation for hyperspectral data analysis. *IEEE Transactions on Image Processing* **2008**, *17*, 217–225.
36. Mohammad-Djafari, A. Joint estimation of parameters and hyperparameters in a Bayesian approach of solving inverse problems. Image Processing, 1996. Proceedings., International Conference on. IEEE, 1996, Vol. 2, pp. 473–476.
37. Kolehmainen, V.; Vanne, A.; Siltanen, S.; Jarvenpaa, S.; Kaipio, J.P.; Lassas, M.; Kalke, M. Parallelized Bayesian inversion for three-dimensional dental X-ray imaging. *IEEE Transactions on Medical Imaging* **2006**, *25*, 218–228.
38. Qi, J.; Leahy, R.M. Resolution and noise properties of MAP reconstruction for fully 3-D PET. *IEEE Transactions on Medical Imaging* **2000**, *19*, 493–506.
39. Ericson, W.A. A note on the posterior mean of a population mean. *Journal of the Royal Statistical Society. Series B (Methodological)* **1969**, pp. 332–334.
40. Fox, C.W.; Roberts, S.J. A tutorial on variational Bayesian inference. *Artificial intelligence review* **2012**, *38*, 85–95.
41. Tzikas, D.G.; Likas, A.C.; Galatsanos, N.P. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine* **2008**, *25*, 131–146.
42. Ayasso, H.; Mohammad-Djafari, A. Joint NDT image restoration and segmentation using Gauss–Markov–Potts prior models and variational Bayesian computation. *IEEE Transactions on Image Processing* **2010**, *19*, 2265–2277.
43. Noël, P.B.; Walczak, A.M.; Xu, J.; Corso, J.J.; Hoffmann, K.R.; Schafer, S. GPU-based cone beam computed tomography. *Computer methods and programs in biomedicine* **2010**, *98*, 271–277.
44. Gac, N.; Mancini, S.; Desvignes, M.; Houzet, D. High speed 3D tomography on CPU, GPU, and FPGA. *EURASIP Journal on Embedded Systems* **2008**, *2008*, 5.
45. Wang, L.; Mohammad-Djafari, A.; Gac, N.; Dumitru, M. Computed tomography reconstruction based on a hierarchical model and variational Bayesian method. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 883–887.
46. Dumitru, M. A Bayesian approach for periodic components estimation for chronobiological signals. PhD thesis, Paris Saclay, 2016.
47. Wang, L.; Mohammad-Djafari, A.; Gac, N. X-ray Computed Tomography using a sparsity enforcing prior model based on Haar transformation in a Bayesian framework. *Fundamenta Informaticae* **2017**, *155*, 449–480.
48. Stanković, R.S.; Falkowski, B.J. The Haar wavelet transform: its status and achievements. *Computers & Electrical Engineering* **2003**, *29*, 25–44.
49. Starck, J.L.; Candès, E.J.; Donoho, D.L. The curvelet transform for image denoising. *IEEE Transactions on Image Processing* **2002**, *11*, 670–684.
50. Do, M.N.; Vetterli, M. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing* **2005**, *14*, 2091–2106.
51. Selesnick, I.W.; Baraniuk, R.G.; Kingsbury, N.C. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine* **2005**, *22*, 123–151.

52. Nadarajah, S. A generalized normal distribution. *Journal of Applied Statistics* **2005**, *32*, 685–694.
53. Rasmussen, C.E. The infinite Gaussian mixture model. NIPS, 1999, Vol. 12, pp. 554–560.
54. Klebanov, L.B. *Heavy tailed distributions*; Matfyzpress, 2003.
55. Frese, T.; Bouman, C.A.; Sauer, K. Adaptive wavelet graph model for Bayesian tomographic reconstruction. *IEEE Transactions on Image Processing* **2002**, *11*, 756–770.
56. Delaney, A.H.; Bresler, Y. Multiresolution tomographic reconstruction using wavelets. *IEEE Transactions on Image Processing* **1995**, *4*, 799–813.
57. Rantala, M.; Vanska, S.; Jarvenpaa, S.; Kalke, M.; Lassas, M.; Moberg, J.; Siltanen, S. Wavelet-based reconstruction for limited-angle X-ray tomography. *IEEE Transactions on Medical Imaging* **2006**, *25*, 210–217.
58. Loris, I.; Nolet, G.; Daubechies, I.; Dahlen, F. Tomographic inversion using l1-norm regularization of wavelet coefficients. *Geophysical Journal International* **2007**, *170*, 359–370.
59. Sauer, K.; Bouman, C. A local update strategy for iterative reconstruction from projections. *IEEE Transactions on Signal Processing* **1993**, *41*, 534–548.
60. Elad, M.; Milanfar, P.; Rubinstein, R. Analysis versus synthesis in signal priors. *Inverse Problems* **2007**, *23*, 947.
61. Boyd, S.; Vandenberghe, L. *Convex optimization*; Cambridge university press, 2004.
62. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **2004**, *13*, 600–612.
63. Palenstijn, W.J.; Batenburg, K.J.; Sijbers, J. The ASTRA tomography toolbox. 13th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE, 2013, Vol. 2013.
64. Dumitru, M.; Wang, L.; Gac, N.; Mohammad-Djafari, A. iterTomoGPI-GPL, Gitlab repository, Apr2018. [Online; https://github.com/nicolasgac/TomoGPI_for_Astra].
65. Figueiredo, M.A.; Nowak, R.D. Wavelet-based image estimation: an empirical Bayes approach using Jeffrey's noninformative prior. *IEEE Transactions on Image Processing* **2001**, *10*, 1322–1331.
66. Dumitru, M.; Wang, L.; Gac, N.; Mohammad-Djafari, A. Performance Comparison Of Bayesian Iterative Algorithms For Three Classes Of Sparsity Enforcing Priors With Application In Computed Tomography. 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017.