Bayesian Method With Sparsity Enforcing Prior of Dual-Tree Complex Wavelet Transform Coefficients for X-ray CT Image Reconstruction

Li Wang  
Laboratoire des signaux et système  
Université Paris Saclay  
Email: li.wang@lss.supelec.fr

Ali Mohammad-Djafari  
Laboratoire des signaux et système  
CNRS, France  
Email: djafari@lss.supelec.fr

Nicolas Gac  
Laboratoire des signaux et système  
Université Paris Saclay  
Email: nicolas.gac@lss.supelec.fr

Abstract—In this paper, a Bayesian method with a hierarchical sparsity enforcing prior model for Dual-Tree Complex Wavelet Transform (DT-CWT) coefficients is proposed. This model is used for X-ray Computed Tomography (CT) image reconstruction. A generalized Student-t distributed prior model is used to enforce the sparse structure of the DT-CWT coefficient of the image. The joint Maximum A Posterior algorithm (JMAP) is used in this Bayesian context. Comparisons with the conventional and other state-of-the-art methods are presented, showing that the proposed method gives more accurate and robust reconstruction results while the dataset is insufficient.

Index Terms—Computed Tomography (CT), Bayesian Approach, Hierarchical Model, Dual-Tree Complex Wavelet Transformation (DT-CWT), Generalized Student-t distribution, Joint Maximum A Posterior (JMAP).

I. INTRODUCTION

In X-ray CT, the intensity of X-ray radiation is attenuated when passing through the object, and the parameters to be reconstructed are the linear attenuation coefficients inside the object under the test. The Radon Transform (RT), presented in detail in [1], is the most commonly used forward modeling for X-ray CT. There have been different analytic methods to solve the reconstruction problem, for example the Back-Projection (BP) [2], the Filtered Back-Projection (FBP) [2], etc. Iterative methods are developed with a lot of efforts in order to solve the ill-posed CT reconstruction, concerning fewer projection numbers or less exposed time during the projection. Accounting for the modeling uncertainties and measurement noise, the forward system is modeled by:

\[ g = Hf + \epsilon \]  

where \( g \in \mathbb{R}^{M \times 1} \) represents the projection data, \( f \in \mathbb{R}^{N \times 1} \) the object and \( \epsilon \in \mathbb{R}^{M \times 1} \) the additive noise. Matrix \( H \in \mathbb{R}^{M \times N} \) corresponds to the linear projection system.

The regularization methods [3], [4] are often used for the ill-posed inverse problems. A general criterion of the regularization method is \( J(f) = \|g - Hf\|^2_2 + \lambda R(f) \) where \( R(f) \) is the regularization term and \( \lambda \) is called the regularization parameter. We list out some conventional regularization reconstruction methods: (a) the Least Square (LS) method [4] with \( R(f) = 0 \) where the obtained result minimizes the disparity but is not guaranteed to be precise because of the ill-posedness; (b) the Quadratic Regularization (QR) method [5] with \( R(f) = \|Df\|^2_2 \) which enforces the global smoothness of the estimated result and (c) the Total Variation (TV) method [6], where the difference of neighbour pixels are enforced to be sparse with \( R(f) = \|Df\|_1 \). By using \( l_1 \) norm, the sparsity of the penalty term is enforced. The appearance of the non-differentiable \( l_1 \) term leads to difficulties for the computation of the gradient of constraints. Many methods have been studied in order to solve this \( l_1 \) norm optimization problem, for example the Newton’s method [7] and the Split Bregman method [6]. Another class of the regularization method, called “Synthesis”, considers a linear sparse transformation \( f = D\tilde{z} \) and minimizes a criterion: \( J(z) = \|g - HD\tilde{z}\|^2_2 + \lambda R(z) \) for example \( R(z) = \|z\|_1 \) [8]. When \( \tilde{z} \) obtained the object is reconstructed by \( \hat{f} = D\tilde{z} \).

Note that in all the above mentioned regularization methods, there is a parameter \( \lambda \) controlling the trade-off of the disparity and the regularization penalty, and it needs to be fixed. There have been different methods to choose a suitable value for \( \lambda \), for example the Cross Validation (CV) and L-curve methods, with details presented in [9], [10]. However, the computation for choosing this parameter should be done for each situation and different dataset, which is very costly in computation. Bayesian methods [11], [12], therefore, are often used to estimate the parameters and variables simultaneously [13].

The choice of prior distributions is crucial in Bayesian methods. In a previous work [14], a hierarchical structured prior distribution is defined via the Haar transformation, with the direct model \( f = Dz + \xi \), with coefficients presented by \( z \) and transform operator by \( D \), and \( \xi \) defines the uncertainties which relax the transformation relationship in this hierarchical model. \( \xi \) is then modelled through a heavy tailed distribution to enforce its sparsity. In particular a generalized form of Student-t distribution (details presented in Section II) has been used and can be presented with a hierarchical model by defining \( p(\xi|v_\xi) \) as a normal distribution and \( p(v_\xi) \) as an Inverse Gamma distribution, where \( v_\xi \) is the variance of noise \( \xi \). In this way, one can obtain \( p(f|z, v_\xi) \), \( p(z|v_\xi) \) and \( p(v_\xi) \) and by combining with the likelihood \( p(g|f, v_\xi) \) and
The sparse property of the DT-CWT transformation coefficient can be enforced by using three kinds of distributions: the Generalized Gaussian distributions, the Gaussian Mixture distributions and the heavy tailed distributions.

The standard Student-t distribution is heavy tailed, but from the definition of its variance we easily figure out that there is a lower limit of the variance value of Student-t distribution: \( \text{Var}[f] = \frac{\nu}{\nu - 2} > 1, \ (\nu > 2) \). This limit implies that this heavy-tailed distribution can’t have a small variance, therefore the sparsity couldn’t be intensively enforced. In this paper we use a generalization of Student-t distribution (Stg) which can be obtained by using the marginalization of the Normal-Inverse Gamma bi-variate distribution:

\[
S_{\nu}(f|\alpha, \beta) = \int \mathcal{N}(f|0, \nu)\mathcal{IG}(\nu|\alpha, \beta) \, dv. \quad (2)
\]

This generalization of Student-t distribution adds a supplementary parameter compared to the standard one, and hence more capable to control the level of sparsity of the prior distribution [13].

C. Hierarchical Model of the proposed HB-DTCWT

The additive noise is assumed zero mean but non-stationary Gaussian distributed with unknown variances: \( p(\epsilon|\nu_e) = \mathcal{N}(\epsilon|0, V_\epsilon) \), where \( V_\epsilon = \text{diag} [\nu_e] \) is a diagonal matrix and \( \nu_e = [\nu_e, \nu_e, \cdots]^{T} \). According to the linear forward model shown in Eq.(1), we obtain an expression for the likelihood: \( p(g|f, \nu_e) = \mathcal{N}(g|HF, V_\epsilon) \).

In the considered applications, generally the object to be reconstructed is piece-wise homogeneous, considering that it consists of several different materials. With this property, an information which can be considered as a prior knowledge is the sparseness of the DT-CWT coefficient.

As presented previously, we use the DT-CWT coefficient as the sparse representation of the piece-wise continuous image. \( z \) is used to present the \( l \)-level DT-CWT coefficients of the image \( f \). Prior distribution of \( f \) depending on \( z \) is: \( p(f|z, \nu_\xi) = \mathcal{N}(f|Dz, V_\xi) \) where \( D \) represents the inverse DT-CWT operator, and the additive noise \( \xi \) is considered to be zero mean Gaussian distributed and \( V_\xi = \text{diag}[\nu_\xi], \nu_\xi = [\nu_\xi, \nu_\xi, \cdots]^{T} \).

Vector \( z = [z_1, z_2, \cdots, z_{2^N}]^{T} \) is sparse. As mentioned above, the generalized Student-t distribution is used in order to enforce the sparsity. By using the Normal-Inverse Gamma
marginalization property given in Eq.(2), the prior distribution for \( z \) can be realized by:

\[
\begin{align*}
    p(z|v_z) &= \mathcal{N}(z|0, V_z) \quad \text{where} \quad V_z = \text{diag } [v_z], \\
    p(v_z|\alpha_{z0}, \beta_{z0}) &= \prod_{i=1}^{N} IG(v_{zi}|\alpha_{z0}, \beta_{z0}),
\end{align*}
\]

where the elements of vector \( v_z = [\cdots, v_{z_k}, \cdots]^T \) are supposed to be i.i.d.

On the other hand, when considering the variance of two noises: \( \nu \), \( n_t \), by knowing that the variance is positive, and the fact that the majority of the values are small, we choose the Inverse Gamma distribution to model them:

\[
\begin{align*}
    p(\nu|\alpha_{\nu0}, \beta_{\nu0}) &= \prod_{i=1}^{M} IG(\nu_{i}|\alpha_{\nu0}, \beta_{\nu0}), \\
    p(\nu|\alpha_{\nu0}, \beta_{\nu0}) &= \prod_{i=1}^{N} IG(\nu_{i}|\alpha_{\nu0}, \beta_{\nu0}).
\end{align*}
\]

With all the proposed prior distributions, the model containing all the variables, parameters and hyper-parameters is:

\[
\begin{align*}
    p(g|f, \nu) &\propto |V_e|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (g - Hf)^T V_e^{-1} (g - Hf) \right], \\
    p(f|z, \nu) &\propto |V_e|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} (f - Dz)^T V_e^{-1} (f - Dz) \right], \\
    p(z|\alpha_{\nu0}, \beta_{\nu0}) &\propto \prod_{i=1}^{M} v_{zi}^{-(\alpha_{\nu0}+1)} \exp \left[ -\beta_{\nu0} v_{zi} \right], \\
    p(\nu|\alpha_{\nu0}, \beta_{\nu0}) &\propto \prod_{i=1}^{N} v_{zi}^{-(\alpha_{\nu0}+1)} \exp \left[ -\beta_{\nu0} v_{zi} \right].
\end{align*}
\]

The corresponding directed acyclic graph (DAG) of the proposed model is shown in Fig.(2).

**E. Bayesian Point Estimation**

The JMAP computation aims at iteratively and alternately estimate the variables and parameters by maximizing the posterior distribution:

\[
\begin{align*}
    (\hat{f}, \hat{z}, \hat{\nu}, \hat{\nu}, \hat{\nu}) &= \arg \max_{f, z, \nu} \{ p(f, z, \nu, \nu, \nu | g) \} \quad (6)
\end{align*}
\]

Because of the huge size of the dataset, the descend gradient algorithm is used since the computation of big size matrix inversion is too expensive. The iterative updating rule is:

\[
\begin{align*}
    \text{iter:} & \quad f^{(n+1)} = f^{(n)} - \gamma_{f} \nabla J(f^{(n)}); \\
    \text{iter:} & \quad z^{(n+1)} = z^{(n)} - \gamma_{z} \nabla J(z^{(n)});
\end{align*}
\]

\[
\begin{align*}
    \gamma_{f} &= \left( \beta_{\nu0} + \frac{1}{2} \varepsilon_{z} \right) / \left( \alpha_{\nu0} + 3/2 \right), \quad \forall k \in [1, N]; \\
    \gamma_{z} &= \left( \beta_{\nu0} + \frac{1}{2} \varepsilon_{z} \right) / \left( \alpha_{\nu0} + 3/2 \right), \quad \forall j \in [1, M].
\end{align*}
\]

\[
\begin{align*}
    \hat{f}_{ik} &= \left( \beta_{\nu0} + \frac{1}{2} \varepsilon_{z} \right) (g_{i} - (Hf)^{2}) / \left( \alpha_{\nu0} + 3/2 \right), \quad \forall i \in [1, M]; \\
    \hat{z}_{kj} &= \left( \beta_{\nu0} + \frac{1}{2} \varepsilon_{z} \right) (\hat{f}_{j} - (Dz)^{2}) / \left( \alpha_{\nu0} + 3/2 \right), \quad \forall j \in [1, N],
\end{align*}
\]

where

\[
\begin{align*}
    J(f) &= \frac{1}{2} \| V_e^{-\frac{1}{2}} (g - Hf) \|_2^2 + \frac{1}{2} \| V_e^{-\frac{1}{2}} (f - Dz) \|_2^2; \\
    J(z) &= \frac{1}{2} \| V_e^{-\frac{1}{2}} (f - Dz) \|_2^2 + \frac{1}{2} \| V_e^{-\frac{1}{2}} z \|_2^2;
\end{align*}
\]

\[
\begin{align*}
    \gamma_{f} &= \left( \frac{\| \nabla J(f(n)) \|_2^2}{\left( \| \hat{Y}_e H \nabla J(f(n)) \|_2^2 + \| \hat{Y}_e \nabla J(f(n)) \|_2^2 \right)} \right)^{\frac{1}{2}}; \\
    \gamma_{z} &= \left( \frac{\| \nabla J(z(n)) \|_2^2}{\left( \| \hat{Y}_e D \nabla J(z(n)) \|_2^2 + \| \hat{Y}_e \nabla J(z(n)) \|_2^2 \right)} \right)^{\frac{1}{2}},
\end{align*}
\]

where \( \nabla J(\cdot) \) is the gradient of \( J(\cdot) \). \( \gamma_{f} \) and \( \gamma_{z} \) are obtained by using optimized step length strategy, see [22]. The initialization of the hyper parameters \( \alpha_{z0} \) and \( \beta_{z0} \) are:

\[
\begin{align*}
    \alpha_{z0} &= \alpha_{\nu0} = \alpha_{f0} = 2 + \varepsilon_{1}, \\
    \beta_{z0} &= \beta_{\nu0} = \beta_{\nu0} = \varepsilon_{2}, \quad \text{where} \quad \varepsilon_{1} \text{ and } \varepsilon_{2} \text{ are small values close to zero.}
\end{align*}
\]

III. SIMULATIONS RESULTS

The simulated Shepp-Logan image of size 256\(^2\), shown in Fig.(3)(a), is used as the original image, and the value of the pixels are normalized to [0, 1]. Parallel projections are used as the dataset. For each projection, 256\(^2\) detectors are used, and an image of size 256 \times 256 is obtained. The red block is the zone that we choose to show the results in Fig.(4). The uncontaminated sinogram by applying 128 projections is shown in Fig.(3)(b). The sinogram with 128 projections and SNR=20dB is shown in Fig.(3)(c). The projections are distributed uniformly from 0 to \( \pi \). The signal-to-noise ratio of the dataset is defined as:

\[
\text{SNR} = \frac{\| g_{0} \|_2^2 / \| \epsilon \|_2^2}{\| g \|_2^2 / \| \epsilon \|_2^2},
\]

where \( g_{0} \) is the unbiased sinogram and \( g \) is the contaminated sinogram with the additive noise represented by \( \epsilon \) in Eq.(1). A lower level of SNR refers to the dataset with more critical additive noise \( \epsilon \).

The proposed BH-DTCWT method is used for the reconstructed and is compared with the conventional TV method and the HHBM method in our previous work [14]. In the
TV method, the Split-Bregman algorithm [6] is used. The regularization parameter is chosen empirically, which should have a bigger value when the data has a lower SNR value, namely the level of noise is higher. In both HHBM and BH-DTCWT methods, the multi-level transformations are chosen to have 5 levels.

The reconstructed images are shown in Fig.(4), represented by zones of them corresponding to Fig.(3)(a). The red curve represents the profile of the image at the position of the blue line. The first column shows the results reconstructed from dataset with 128 projections, by using TV, HHBM and BH-DTCWT methods respectively. Visually, the TV method give a better result with smoother homogeneous zones. In the right column, images are reconstructed from dataset with 64 projections, and they show the convenient of the HHBM and BH-DTCWT methods: they obtain a better result than the TV method with less number of projections.

In order to show more clearly the effect at the edges of different methods, we compare a zone of the profiles in Fig.(4), which is at the position of the thick green line shown in Fig.(3)(a). Comparisons are demonstrated in Fig.(5). In the figure on top are the profiles obtained with 64 projection, and in the bottom are those from 32 projections. As we can see, the proposed method stays more robust at the edges while there are insufficient number of distributions.

Both HHBM and BH-DTCWT methods have a good performance when dealing with the reconstruction with less number of projections. Fig.(6)-Fig.(8) show the comparison of the normalized Mean Square Error (NMSE), represent by $\delta_f$ in the ordinates in Fig.(6)-Fig.(8), with the definition:

$$\text{NMSE} = \frac{\| \hat{f} - f \|^2}{\| f \|^2}. \quad (10)$$

The performance of TV is very good when there are sufficient number of projections, but when the number of projections become insufficient, the HHBM and BH-DTCWT methods are more robust than the TV method. The BH-DTCWT method is not strictly better than the HHBM method, but is better in most of the cases in the reconstruction of Shepp Logan phantom.

**IV. Conclusion**

We proposed a Bayesian method based on a hierarchical prior model in which the Dual-Tree Complex Wavelet Transformation is used. Generalized Student-t distribution is used to enforce the sparsity of the transformation coefficient. Comparing with the state-of-the-art methods, the proposed method remain robust when dealing with reconstruction problems with less number of projections, namely insufficient dataset, in the CT reconstruction of the Shepp Logan image. Another advantage is that by using the Bayesian method we have no need to choose a value for the parameters, such as the regularization parameter in TV, and they are automatically estimated in the simulations. In the future work we will complete the comparisons by using different types of phantoms, for example phantoms with periodical textures, in order to analyse the influence of the transformation type.

**REFERENCES**


Fig. 5: Comparison of a zone of profile with TV method, from dataset of 64 projections on the top, and 32 projections on the bottom.

Fig. 6: NMSE of reconstructed image along iterations from sinogram with 128 projections and a high SNR=40dB.

Fig. 7: NMSE of reconstructed image along iterations from sinogram with 64 projections and a high SNR=40dB.

Fig. 8: NMSE of reconstructed image along iterations from sinogram with 32 projections and a high SNR=40dB.