Variational Bayesian Image Restoration with a Product of Spatially Weighted Total Variation Image Priors

Giannis Chantas, Nikolaos Galatsanos, Rafael Molina, and Aggelos Katsaggelos

***Abstract*—In this paper a new image prior is introduced and used in image restoration. This prior is based on products of spatially weighted Total Variations (TV). These spatial weights provide this prior with the flexibility to better capture local image features than previous TV based priors. Bayesian inference is used for image restoration with this prior via the variational approximation. The proposed restoration algorithm is fully automatic in the sense that all necessary parameters are estimated from the data and is faster than previous similar algorithms. Numerical experiments are shown which demonstrate that image restoration based on this prior compares favorably with previous state-of-the-art restoration algorithms.**

I. Introduction

Image restoration is a mature image processing topic with an over 30 year long history. This problem is well known to be ill-posed and consequently it requires regularization [1].

The field of image restoration is very broad. Thus an attempt to survey it and do justice to all its contributors is outside the scope of this paper. Therefore in what follows we reference only image restoration methods directly related to the proposed work.

Total Variation (TV) is a powerful concept for robust estimation [2]. It was first introduced as a regularizer for image restoration in [5]. Since then it has been used extensively and with great success for inverse problems because TV has the ability to smooth noise in flat areas of the image and at the same time preserve edges. For certain recent developments on TV based image recovery the interested reader is referred to [4] and [3].

Nevertheless, TV-based image restoration has certain shortcomings. One of them is the selection of the regularization parameter which to a large extent until recently has been ad-hoc. Rudin *et al.* [5] consider the minimization of the TV energy function constrained by the sum of the square of the observation errors being equal to $N\hat{\sigma}^2$, where $\hat{\sigma}^2$ represents an estimate of the noise variance, and N is the number of observations, and then proceed to estimate both the image and the associated Lagrange multiplier to this constrained optimization problem. Bertalmio *et al.* [24] make the Lagrange multiplier region dependent. Bioucas-Dias *et al.* [12], using their majorization-minimization approach [13], propose a Bayesian method to estimate the original image and regularization parameter assuming that an estimate of the noise variance is available. Recently, a Bayesian inference framework which requires the approximation of the prior partition function and is based on the variational approximation was proposed to handle the simultaneous parameter and image estimation [14].

An alternative image model has recently been proposed based on the combination of several image priors [11], [23] and [16]. It combines in product form multiple probabilistic models. Each individual model gives high probability to data vectors that satisfy just one constraint. Vectors that satisfy only this constraint but violate others are ruled out by their low probability under the other terms of the product model. Such priors were learned in [11] and [23] using a large training set of images and stochastic sampling methods, in contrast to the approach proposed in [16] where the product prior is learnt only from the observations. Each term in the product defining the prior in [16] corresponds to the output of a high-pass linear filter and is Student's-t distributed. The main contribution of [16] is the introduction of a Bayesian inference methodology based on the constrained variational approximation that bypasses the difficulty of evaluating the normalization constant of product type priors.

Variational based Bayesian inference using TV and Student't-t priors for the image has been used with success for blind image deconvolution (BID) problems also. In [17] a TV prior was used for the image while Gaussian priors were used for the point spread function (PSF). In [18] a Student's-t prior in product form was used for the image. However, the normalization constant for this prior was approximated. In [18] also a kernel based sparse Student's-t prior was used for the PSF. This prior provides a mechanism, for the first time, to estimate the spatial support of the PSF also.

In this paper we contribute to the field of prior image modeling by combining the advantages of TV image modeling in [14] and Student's-t Product of Experts (PoE) image modeling in [16]. The new image model has a number of

Giannis Chantas is with the Department of Computer Science, University of Ioannina, Greece, 45110. E-mail gchantas@yahoo.gr

Nikolaos Galatsanos (corresponding author) is with the Department of Electrical Engineering University of Patras Rio, Greece 26500. E-mail:ngalatsanos@upatras.gr

Rafael Molina is with Departamento de Ciencias de la Computacion e I. A., Universidad de Granada, 18071 Granada, Spain. E-mail: rms@decsai.ugr.es

Aggelos Katsaggelos is with the Department of Electrical Engineering and Computer Science, Northwestern University Evanston, IL 60208-3118, 2145 Sheridan Road. E-mail: aggk@eecs.northwestern.edu

This work was supported in part by the research project (PENED) which was supported in part by the E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%), the Comisión Nacional de Ciencia y Tecnología under contract TIC2007-65533, the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), and the Greek-Spanish collaboration program of Greek Ministry of Development-GSRT.

novel features. First, unlike [14] and [12], it uses a spatially weighted version of the TV. These spatial weights provide the prior with the flexibility to model explicitly the *local salient features* of the image. Second, as in [16], it is in product form, it is spatially weighted and it has the ability to enforce simultaneously a number of different properties on the image. This new prior can use arbitrary linear operators, not just first order differences as in TV and can combine them, in contrast to [16], in a non-linear manner. In order to avoid the over parameterization due to the spatially varying nature of the herein proposed prior, a model with two layers of hidden variables is proposed, which extends the one used in [15] and [16]. If the hidden variables of the second layer are marginalized the resulting density function *has similar form* to a Student's-t pdf thus we refer to it as *Modified Student's-t*.

Due to the complexity of this model we resort to the variational approximation for Bayesian inference [6]. Specifically, we develop several extensions of the methodologies used in [14] and [16]. First, to bypass the difficulties due to the non-quadratic nature of the new prior, we find a quadratic bound to the variational bound [14]. However, the bound used here, unlike the one in [14] is explicitly locally adaptive. Second, we use the *constrained variational* framework [16] (but tailored explicitly to locally adaptive filters) in order to bypass the problem of computing the partition function of the new prior.

Finally, we also propose a new iterative method to compute the diagonal elements of very large inverse matrices that are necessary for the herein proposed Bayesian inference algorithm. Similar computations are also required in [14]. However, in this work these matrices were approximated by circulants that can be explicitly inverted. The herein method is conjugate-gradient based and results in a significant improvement in the speed of the overall restoration algorithm as compared to the algorithm in [16]. Furthermore, the computation of the diagonal elements of the inverses of similar matrices required in sparse Bayesian models is a problem that has been known for a while in the machine learning community [7]. However, the recursive approach proposed in [8] cannot be applied to imaging problems due to their extremely large dimensionality [9].

The rest of this paper is organized as follows. In section II we present the imaging and image models. In section III we present the variational algorithm for Bayesian inference. In Section IV a brief discussion is presented of the mechanisms that introduce spatially adaptive regularization when TV and Student's-t based priors are used. In section V we present the details of the computational implementation of our algorithm. In section VI we present numerical experiments, and in section VII conclusions and future work.

II. Imaging and image model

In what follows we use one dimensional notation for simplicity. Let $\mathbf{f}$ be the original image represented as an $N \times 1$ vector, blurred by a convolutional operator $\mathbf{H}$, of size $N \times N$. The degraded observation is given by

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \mathbf{n}, \tag{1}$$

where $\mathbf{n}$ is the noise $N \times 1$ vector modeled as white Gaussian, i.e., $\mathbf{n} \sim N(\mathbf{0}, \beta^{-1}\mathbf{I})$, where $\mathbf{0}$ and $\mathbf{I}$ are the $N \times 1$ zero and $N \times N$ identity matrices, respectively, and β^{-1} represents the noise variance.

A. Modified Student's-t image prior

Image priors in product form are very attractive since they have the ability to enforce simultaneously many properties on an image; see for example [16]. For this purpose we propose herein a prior in product form for the image. To define such a prior we introduce P pairs of linear convolutional operators (filters) $(\mathbf{Q}_1, \mathbf{Q}_2,), (\mathbf{Q}_3, \mathbf{Q}_4), \ldots, (\mathbf{Q}_{2P-1}, \mathbf{Q}_{2P})$ of size $N \times N$ and assume that the filter outputs $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_{2P})$ are produced according to

$$\boldsymbol{\epsilon}_l = \mathbf{Q}_l \mathbf{f}, \ l = 1, \ldots, 2P. \tag{2}$$

Then, for each pixel location i, it is assumed that each pair $\epsilon_{2k}(i)$ and $\epsilon_{2k-1}(i)$ is jointly distributed with probability density function

$$p(\epsilon_{2k}(i), \epsilon_{2k-1}(i)|a_k(i)) = \frac{\lambda_k^2 a_k(i)^2}{2\pi} \exp\left(-\lambda_k a_k(i)\sqrt{\epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2}\right). \tag{3}$$

with $k = 1, \ldots, P$ and $i = 1, \ldots, N$ where N is the number of pixels in the image.

Notice that for $P = 1$ and $\mathbf{Q}_1$ and $\mathbf{Q}_2$ the first order horizontal and vertical difference operators and $a_k(i) = a$, the prior becomes identical in form to the total-variation (TV) based prior proposed in [12] and [14]. However, the prior proposed herein is more general because it can use any linear operator not just first order differences. Notice also that $a_k(i)$ varies for every pixel i which makes it, in contrast to the model used in [14], [12], explicitly *spatially adaptive*.

In other words, in the herein prior the outputs $\epsilon_{2k}(i)$ and $\epsilon_{2k-1}(i)$ of the pairs of the operators $\mathbf{Q}_{2k}$ and $\mathbf{Q}_{2k-1}$ in Eq. (3) are assumed to be *differently distributed* per pixel i. This is captured by the spatially varying weights $a_k(i)$ which conceptually play the role of the precision (inverse variance) of the "local TV" given by $\sqrt{\epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2}$. In contrast previous TV priors in [14], [12] are of the form

$$p(\epsilon_{2k}(i), \epsilon_{2k-1}(i)|a_k) = \frac{\lambda_k^2 a_k^2}{2\pi} \exp\left(-\lambda_k a_k \sqrt{\epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2}\right), \tag{4}$$

with $k = 1, \ldots, P$ and $i = 1, \ldots, N$. Thus, $\epsilon_{2k}(i), \epsilon_{2k-1}(i)$ are assumed to be identically distributed *over the entire image*. Clearly, the herein spatially weighted TV prior in Eq. (3) provides more flexibility in capturing the local properties of the image than the previous ones in [12], [14].

Finally notice also that when the energy term in Eq. (3) is used without the square root and $\epsilon_{2k}(i) = \epsilon_{2k-1}(i)$, this prior simplifies to the one used in [16]. However, as it has been extensively reported in the literature non-quadratic energy priors produce better results than quadratic ones.

One drawback of the herein prior is the over-parameterization problem since PN unknowns $a_k(i)$ have to be estimated from N data points. In order to ameliorate this

problem we assume that each $a_k(i)$ is a Gamma distributed hidden random variable [15], [16] according to:

$$p(a_k(i)) = Gamma(a_k(i); \nu_k/2, \nu_k/2) \quad k = 1, \ldots, P,\ i = 1, \ldots, N. \tag{5}$$

Note that this distribution on $a_k(i)$ is flexible enough to provide a range of restrictions on $a_k(i)$: from very vague information, which would be modeled when $\nu_k \to 0$, to very precise information which is obtained when $\nu_k \to \infty$. Note also that from the definition in (5), for a given k, all the $a_k(i)$ coefficients come from the same distribution with variance moving from infinity to zero, as ν_k changes from zero to infinity.

The marginal distribution of $\epsilon_{2k}(i)$ and $\epsilon_{2k-1}(i)$ can be computed in *closed form* and is given by

$$\begin{aligned} &p(\epsilon_{2k}(i), \epsilon_{2k-1}(i)) \\ &= \int_{a_k(i)} p(\epsilon_{2k}(i), \epsilon_{2k-1}(i)|a_k(i))p(a_k(i))da_k(i) \\ &= \frac{\Gamma(\nu_k/2 + 1/2)}{\Gamma(\nu_k/2)} \left(\frac{\lambda_k}{\pi\nu_k}\right)^{1/2} \\ &\times \left(1 + \frac{\lambda_k\sqrt{\epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2}}{\nu_k}\right)^{-\nu_k/2-1/2} \end{aligned} \tag{6}$$

for $k = 1, 2, \ldots, P$ and $i = 1, 2, \ldots, N$.

This density function is very similar in form to the Student's-t pdf which is given by [6]

$$p(x) = \frac{\Gamma(\nu_/2 + 1/2)}{\Gamma(\nu_/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left(1 + \frac{\lambda x^2}{\nu},\right)^{-\nu/2-1/2} \tag{7}$$

thus, in the rest of this paper we label it as *Modified Student's-t*. It is important to note that this prior combines the advantages of both TV-based and Student's-t based priors. The former being the ability to suppress noise and maintain edges in an image beyond the capabilities of linear filters [12], [14] and the latter being the ability to explicitly introduce spatial adaptivity through the hidden random variables $a_k(i)$, which improves the prior modeling used in [15], [16].

At this point we note that we have not provided a prior for the image, $p(\mathbf{f})$. This was intentional, because we cannot compute it in closed form. More specifically, it is difficult to define a prior for the image $\mathbf{f}$ based on the prior in Eq. (3) because we cannot compute the partition function for such prior. First, the non-quadratic exponent in the pdf in Eq. (3) makes this calculation intractable even if our prior was not in product form ($P = 1$). Furthermore, if we want to use a prior in product form ($P > 1$) even with a quadratic exponent it is not possible to compute the partition function [16]. Consequently, in the next section we bypass the calculation of the partition function when the prior is defined on $\mathbf{f}$ by working in the domain of the filter outputs $\boldsymbol{\epsilon}$ [16], where the prior in (3) can be used directly and there is no need to define a prior for $\mathbf{f}$. The downside of this choice is that it is not obvious how to merge the estimates of all the $\boldsymbol{\epsilon}_l$, $l = 1, \ldots 2P$, to generate one estimate for $\mathbf{f}$. To handle this problem we will propose the use of the *constrained variational* approach in the next section.

In order to define the observation model in terms of $\epsilon_l,\ l = 1, \ldots, 2P$, let us examine in some detail the prior model we are using. Let us consider the image model in (3). If we remove one component from it, for instance $\epsilon_{2k-1}(i)$, we have a Laplace distribution with parameter $\lambda_k a_k(i)$. Since we consider jointly $\epsilon_{2k}(i)$ and $\epsilon_{2k-1}(i)$ its partition function is proportional to $1/(\lambda_k a_k(i))^2$. Consequently, given k we have two possible explanations for the data, one associated with ϵ_{2k} and the other with ϵ_{2k-1}.

We thus introduce an alternative observation model, which is derived by applying the operators $\mathbf{Q}_l$ to the original imaging model in (1). This yields:

$$\mathbf{y}_l = \mathbf{H}\boldsymbol{\epsilon}_l + \mathbf{n}_l,\ l = 1, \ldots, 2P, \tag{8}$$

where $\mathbf{y}_l = \mathbf{Q}_l\mathbf{g}$, $\mathbf{n}_l = \mathbf{Q}_l\mathbf{n}$ and thus $\mathbf{n}_l \sim N(0, \beta^{-1}\mathbf{Q}_l\mathbf{Q}_l^T)$.

We finally arrive at the Bayesian modeling of our problem, that is,

$$p(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}; \theta) = p(\mathbf{y}|\boldsymbol{\epsilon})p(\boldsymbol{\epsilon}|\mathbf{a}; \theta)p(\mathbf{a}; \theta), \tag{9}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{2P})$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_{2P})$, with $\boldsymbol{\epsilon}_l = (\epsilon_l(1), \ldots, \epsilon_l(N))$, $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_P)$, with $\mathbf{a}_k = (a_k(1), \ldots, a_k(N))$ $k = 1, \ldots, P$, and $\theta = (\lambda_1, \ldots, \lambda_P, \nu_1, \ldots, \nu_P)$, with the above probability distributions defined by

$$p(\mathbf{y}|\boldsymbol{\epsilon}) = \prod_{l=1}^{2P} p(\mathbf{y}_l|\boldsymbol{\epsilon}_l), \qquad p(\mathbf{y}_l|\boldsymbol{\epsilon}_l) = N(\boldsymbol{\epsilon}_l, \beta^{-1}\mathbf{Q}_l\mathbf{Q}_l^T), \tag{10}$$

$$p(\boldsymbol{\epsilon}|\mathbf{a}; \theta) = \prod_{k=1}^{P}\prod_{i=1}^{N} p(\epsilon_{2k}(i), \epsilon_{2k-1}(i)|a_k(i)), \tag{11}$$

and

$$p(\mathbf{a}; \theta) = \prod_{k=1}^{P}\prod_{i=1}^{N} p(a_k(i); \theta). \tag{12}$$

This Bayesian model will be used for inference in the next section where we treat $\boldsymbol{\epsilon}$ and $\mathbf{a}$ as *hidden variables* and θ as a parameter vector to be estimated. Observe that we use the notation $p(.; \theta)$ to denote that θ is a set of hyperparameters which are not treated as random variables. We could also have used $p_\theta(.)$. Notice also that we assume that the noise precision parameter β is assumed to have been previously estimated.

III. Variational inference with the modified Student's-t prior

According to Bayesian inference we have to find the posterior distributions for the hidden variables ϵ and $\mathbf{a}$ given $\mathbf{y}$ and the parameter vector θ. However, the marginal of the observations which is required to find the posteriors of the hidden variables is hard to compute [6]. More specifically, the integral

$$p(\mathbf{y}; \theta) = \int_{\boldsymbol{\epsilon}, \mathbf{a}} p(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}; \theta)d\boldsymbol{\epsilon}d\mathbf{a} \tag{13}$$

is intractable.

The variational algorithm that we describe in what follows, bypasses this difficulty and maximizes a *lower bound* that can be found instead of the log-likelihood of the observations $\log p(\mathbf{y}; \theta)$ [6], [10]. This bound is obtained by subtracting

from $\log p(\mathbf{y};\theta)$ the Kullback-Leibler divergence, which is always positive, between an arbitrary $q(\boldsymbol{\epsilon},\mathbf{a})$ and $p(\boldsymbol{\epsilon},\mathbf{a}|\mathbf{y};\theta)$, that is,

$$L(q(\boldsymbol{\epsilon},\mathbf{a}),\theta) = \log p(\mathbf{y};\theta) - KL(q(\boldsymbol{\epsilon},\mathbf{a})||p(\boldsymbol{\epsilon},\mathbf{a}|\mathbf{y};\theta)), \quad (14)$$

and is equal to

$$\begin{aligned} L(q(\boldsymbol{\epsilon},\mathbf{a});\theta) &= \int_{\boldsymbol{\epsilon},\mathbf{a}} q(\boldsymbol{\epsilon},\mathbf{a}) \log p(\boldsymbol{\epsilon},\mathbf{a},\mathbf{y};\theta) d\boldsymbol{\epsilon} d\mathbf{a} \\ &- \int_{\boldsymbol{\epsilon},\mathbf{a}} q(\boldsymbol{\epsilon},\mathbf{a}) \log q(\boldsymbol{\epsilon},\mathbf{a}) d\boldsymbol{\epsilon} d\mathbf{a}. \end{aligned} \quad (15)$$

When $q(\boldsymbol{\epsilon},\mathbf{a}) = p(\boldsymbol{\epsilon},\mathbf{a}|\mathbf{y};\theta)$, this bound is maximized and $L(q(\boldsymbol{\epsilon},\mathbf{a});\theta) = \log p(\mathbf{y};\theta)$. Because the exact posterior $p(\boldsymbol{\epsilon},\mathbf{a}|\mathbf{y};\theta) = \frac{p(\boldsymbol{\epsilon},\mathbf{a},\mathbf{y};\theta)}{p(\mathbf{y};\theta)}$ cannot be found we use an approximation of the posterior. The mean-field approximation is a commonly used approach to maximize the variational bound w.r.t. $q(\boldsymbol{\epsilon},\mathbf{a});\theta$ [6], [10]. According to this approach the hidden variables are assumed to be independent, i.e., $q(\boldsymbol{\epsilon},\mathbf{a}) = q(\boldsymbol{\epsilon})q(\mathbf{a})$. However, for the herein model this is still not sufficient to obtain a closed form for $q(\boldsymbol{\epsilon})$ which is necessary for inference using this approach. More specifically, the square root in the joint $p(\boldsymbol{\epsilon},\mathbf{a},\mathbf{y};\theta)$ which originates from the prior $p(\boldsymbol{\epsilon}|\mathbf{a})$ makes the definition of $q(\boldsymbol{\epsilon})$ intractable.

A. A Lower Bound for $L(q(\boldsymbol{\epsilon},\mathbf{a}),\theta)$

For this purpose we also introduce a *lower bound* on L [14]. More specifically, we use the inequality

$$\sqrt{w} \leq \frac{w+u}{2\sqrt{u}}, \quad (16)$$

which holds for $w \geq 0$ and $u > 0$. Notice that equality holds when $w = u$. This inequality is used at every pixel i by setting $w_k(i) = \epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2$, for $k = 1, 2, \ldots, P$, where $u_k(i)$ are auxiliary variables used for this approximation. Using this and the prior in Eq. (3) we have

$$p(\epsilon_{2k}(i), \epsilon_{2k-1}(i)|a_k(i)) \geq M(\epsilon_{2k}(i), \epsilon_{2k-1}(i), u_k(i), a_k(i)) \quad (17)$$

where

$$M(\epsilon_{2k}(i), \epsilon_{2k-1}(i), u_k(i), a_k(i)) = \frac{\lambda_k^2 a_k(i)^2}{2\pi} \exp\left(-\frac{\lambda_k a_k(i)}{2} \frac{\epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2 + u_k(i)}{\sqrt{u_k(i)}}\right), \quad (18)$$

for $k = 1, \ldots, P$.

We also define $\mathbf{u}_k = (u_k(1), \ldots, u_k(N))$ and $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_P)$. Let us now define

$$L^b(q(\boldsymbol{\epsilon}), q(\mathbf{a}), \mathbf{u}, \theta) = \int_{\boldsymbol{\epsilon},\mathbf{a}} q(\boldsymbol{\epsilon}) q(\mathbf{a}) \log \frac{F(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}; \mathbf{u}, \theta)}{q(\boldsymbol{\epsilon}) q(\mathbf{a})} d\boldsymbol{\epsilon} d\mathbf{a}, \quad (19)$$

where

$$F(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}; \mathbf{u}, \theta) = p(\mathbf{y}|\boldsymbol{\epsilon}) \times \left[\prod_{k=1}^{P} \prod_{i=1}^{N} M(\epsilon_{2k}(i), \epsilon_{2k-1}(i), u_k(i), a_k(i))\right] p(\mathbf{a};\theta). \quad (20)$$

Then, since $F(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}; \mathbf{u}, \theta) \leq p(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a})$ we have

$$L^b(q(\boldsymbol{\epsilon}), q(\mathbf{a}), \mathbf{u}, \theta) \leq L(q(\boldsymbol{\epsilon}, \mathbf{a}), \theta), \quad (21)$$

and consequently the bound becomes tight when

$$\max_{\mathbf{u}} L^b(q(\boldsymbol{\epsilon}), q(\mathbf{a}), \mathbf{u}, \theta) \leq L(q(\boldsymbol{\epsilon}, \mathbf{a}), \theta). \quad (22)$$

Notice that the new lower bound L^b is quadratic in the hidden variables $\boldsymbol{\epsilon}$, thus it is possible to find $q(\boldsymbol{\epsilon})$ that maximizes it. In contrast, the original bound L was not quadratic in $\boldsymbol{\epsilon}$.

B. A Constrained Variational Inference Algorithm

As we have already explained, $\boldsymbol{\epsilon}_l$, $l = 1, \ldots, 2P$, are used instead of $\mathbf{f}$ to avoid the computation of the normalization constant of the prior on $\mathbf{f}$. Thus, a question that needs to be addressed is how one finds $\mathbf{f}$ given the different $q(\boldsymbol{\epsilon}_l)$.

Unconstrained maximization of the bound $L^b(q(\boldsymbol{\epsilon}), q(\mathbf{a}), \mathbf{u}, \theta)$ results in $q(\boldsymbol{\epsilon}_l) = N(\mathbf{m}_l, \mathbf{R}_l)$, where $\mathbf{R}_l = \mathbf{Q}_l \left(\beta \mathbf{H}^T \mathbf{H} + \mathbf{Q}_l^T \lambda_k \langle A \rangle_l \mathbf{U}_l^{-1/2} \mathbf{Q}_l\right)^{-1} \mathbf{Q}_l^T$, $\mathbf{m}_l = \beta \mathbf{R}_l {\mathbf{Q}_l}^{-T} \mathbf{H}^T \mathbf{g}$, with $l = 1, \ldots, 2P$, $\mathbf{U}_l = diag\{u_l(1), \ldots, u_l(N)\}$, $k = \lceil l/2 \rceil$, and $\langle \mathbf{A} \rangle_l = diag\{\langle a_l(1) \rangle_a, \ldots, \langle a_l(N) \rangle_a\}$, where $\langle . \rangle_a$ denotes the expectation w.r.t the distribution of $\mathbf{a}$.

Clearly each $\mathbf{m}_l$ suggests a different estimate for $\mathbf{f}$ given by $\hat{\mathbf{f}} = {\mathbf{Q}_l}^{-1} \mathbf{m}_l$. Thus, one needs to find a methodology to *merge* the information from all $q(\boldsymbol{\epsilon}_l)$ into one estimate of $\mathbf{f}$.

For this purpose the *constrained variational* approximation first proposed in [16] is applied. According to this approach, each $q(\boldsymbol{\epsilon}_l)$ is constrained to have the form:

$$q(\boldsymbol{\epsilon}_l) = N(\mathbf{Q}_l \mathbf{m}, \mathbf{Q}_l \mathbf{R} \mathbf{Q}_l^T), \quad (23)$$

where $\mathbf{m}$ is a $N \times 1$ vector, taken as the mean of the image, and $\mathbf{R}$ the $N \times N$ image covariance matrix. This form is consistent with the equation $\boldsymbol{\epsilon}_l = \mathbf{Q}_l \mathbf{f}$ for which $\bar{\boldsymbol{\epsilon}}_l = E[\boldsymbol{\epsilon}_l] = \mathbf{Q}_l E[\mathbf{f}] = \mathbf{Q}_l \mathbf{m}$ and $E\left[(\boldsymbol{\epsilon}_l - \bar{\boldsymbol{\epsilon}}_l)(\boldsymbol{\epsilon}_l - \bar{\boldsymbol{\epsilon}}_l)^T\right] = \mathbf{Q}_l E\left[(\mathbf{f} - \mathbf{m})(\mathbf{f} - \mathbf{m})^T\right] \mathbf{Q}_l^T = \mathbf{Q}_l \mathbf{R} \mathbf{Q}_l^T$ with $\mathbf{R} = E\left[(\mathbf{f} - \mathbf{m})(\mathbf{f} - \mathbf{m})^T\right]$. Using this approximation the parameters $\mathbf{m}$ and $\mathbf{R}$ are learned instead of $q(\boldsymbol{\epsilon}_l)$ according to the herein constrained variational methodology.

We now present the maximization method by giving the updates for the variables of the bound L^b in the j-th iteration. In the VE-step, the maximization of $L^b(q(\boldsymbol{\epsilon}), q(\mathbf{a}), \mathbf{u}, \theta)$ is performed with respect to $q(\mathbf{a})$, $\mathbf{m}$ and $\mathbf{R}$ keeping $\mathbf{u}$ and θ fixed, while in the VM-step, the maximization of $L^b(q(\boldsymbol{\epsilon}), q(\mathbf{a}), \mathbf{u}, \theta)$ is performed with respect to $\mathbf{u}$ and θ keeping $q(\mathbf{a})$, $\mathbf{m}$, and $\mathbf{R}$ fixed. We have,

VE-step:

$$[\mathbf{m}^j, \mathbf{R}^j, q^j(\mathbf{a})] = \arg \max_{\mathbf{m},\mathbf{R},q(\mathbf{a})} L^b(q(\boldsymbol{\epsilon}), q(\mathbf{a}), \mathbf{u}^{j-1}, \theta^{j-1}) \quad (24)$$

VM-step:

$$[\mathbf{u}^j, \theta^j] = \operatorname*{argmax}_{\mathbf{u},\theta} L^b(q^j(\boldsymbol{\epsilon}), q^j(\mathbf{a}), \mathbf{u}, \theta) \quad (25)$$

The updates for the VE-Step are derived in the Appendix. These are

$$q^j(\boldsymbol{\epsilon}_l) = N(\mathbf{Q}_l \mathbf{m}^j, \mathbf{Q}_l \mathbf{R}^j \mathbf{Q}_l^T), \quad (26)$$

where

$$\mathbf{m}^j = \beta \mathbf{R}^j \mathbf{H}^T \mathbf{g}, \tag{27}$$

$$\mathbf{R}^j = (\beta \mathbf{H}^T \mathbf{H} + \frac{1}{2P} \sum_{k=1}^{P} \sum_{i=0}^{1} \lambda_k{}^{j-1} \mathbf{Q}_{2k-i}^T \langle \mathbf{A}_k \rangle^{j-1} (\mathbf{U}_k^{-1/2})^{j-1} \mathbf{Q}_{2k-i})^{-1}. \tag{28}$$

From the above equations it is clear that $\mathbf{m}$ merges information from all filters $\mathbf{Q}_l$ and is used as the estimate of $\mathbf{f}$.

Finally, the approximate posterior of $\mathbf{a}$ in the VE-step is given by

$$q^j(a_k(i)) = Gamma\left(a_k(i); \frac{\nu_k^{j-1}}{2} + 2, \frac{\nu_k^{j-1}}{2} + \lambda_k^{j-1} \sqrt{u_k^{j-1}(i)}\right)$$

for $i = 1, \ldots, N$ and $k = 1, 2, \ldots, P$. Thus, the expectation of $a_k(i)$ w.r.t $q^j(a_k(i))$ is

$$\langle a_k(i) \rangle_{q^j(\mathbf{a})} = \frac{\nu_k^{j-1} + 4}{\nu_k^{j-1} + 2\lambda_k^{j-1} \sqrt{u_k^{j-1}(i)}} \tag{29}$$

In the VM-step, the bound is maximized w.r.t to the parameters. To find $\mathbf{u}^j$ we have to solve

$$u_k^j(i) = \arg\min_{u_k(i)} \frac{\langle \epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2 \rangle_{q^j(\boldsymbol{\epsilon})} + u_k(i)}{\sqrt{u_k(i)}} \tag{30}$$

where $\langle . \rangle_{q^j(\boldsymbol{\epsilon})}$ represents the expectation w.r.t. $q^j(\boldsymbol{\epsilon})$, which produces

$$\begin{aligned} u_k^j(i) &= \langle \epsilon_{2k}(i)^2 + \epsilon_{2k-1}(i)^2 \rangle_{q^j(\boldsymbol{\epsilon})} \\ &= \sum_{r=0}^{1} ((\mathbf{m}_{2k-r}^j(i))^2 + \mathbf{C}_{2k-r}^j(i,i)) \end{aligned} \tag{31}$$

for $i = 1, \ldots, N$ and $k = 1, 2, \ldots, P$, where

$$\mathbf{m}_{2k-r}^j = \mathbf{Q}_{2k-r} \mathbf{m}^j, \;\; \mathbf{C}_{2k-r}^j = \mathbf{Q}_{2k-r} \mathbf{R}^j \mathbf{Q}_{2k-r}^T. \tag{32}$$

For λ_k we have that

$$\begin{aligned} L^b(q^j(\boldsymbol{\epsilon}), q^j(\mathbf{a}), \mathbf{u}, \theta) &= 2N \sum_{k=1}^{P} \log \lambda_k \\ &- \sum_{k=1}^{P} \sum_{i=1}^{N} \lambda_k \langle a_k(i) \rangle_{q^j(\mathbf{a})} \sqrt{u_k^j(i)} + \text{constant} \end{aligned} \tag{33}$$

when this function is considered to be as a function of λ_k only. Thus, the update formula is

$$\lambda_k^j = \frac{2N}{\sum_{i=1}^{N} \langle a_k(i) \rangle_{q^j(\mathbf{a})} \sqrt{u_k^j(i)}}. \tag{34}$$

Similarly, for ν_k, $k = 1, 2, \ldots, P$, we have that

$$\begin{aligned} L^b(q^j(\boldsymbol{\epsilon}), q^j(\mathbf{a}), \mathbf{u}, \theta) &= \frac{\nu_k}{2} \sum_{i=1}^{N} \langle \log a_k(i) \rangle_{q^j(\mathbf{a})} \\ &- \frac{\nu_k}{2} \sum_{i=1}^{N} \langle a_k(i) \rangle_{q^j(\mathbf{a})} - N\Gamma\left(\frac{\nu_k}{2}\right) \\ &+ N \frac{\nu_k}{2} \log\left(\frac{\nu_k}{2}\right) + \text{constant} \end{aligned} \tag{35}$$

when this function is considered as a function of ν_k only. Then ν_k^j is the root of the function ϕ which is proportional to the derivative of $L^b(q^j(\boldsymbol{\epsilon}), q^j(\mathbf{a}), \mathbf{u}, \theta)$ with respect to ν_k

$$\begin{aligned} \phi(\nu_k) &= \frac{1}{N} \sum_{i=1}^{N} \log \langle a_k(i) \rangle_{q^j(\mathbf{a})} - \frac{1}{N} \sum_{i=1}^{N} \langle a_k(i) \rangle_{q^j(\mathbf{a})} \\ &+ \psi\left(\frac{\nu_k^{j-1}}{2} + 2\right) - \log\left(\frac{\nu_k^{j-1}}{2} + 2\right) \\ &- \psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1, \end{aligned} \tag{36}$$

where ψ is the digamma function. We find $\phi(\nu_k^j) = 0$ numerically using the bisection method.

IV. Spatial Adaptivity with TV and Student's-t Based Priors

At this point it is worth commenting on the spatial adaptivity properties of the restoration filter provided by the combination of the TV and Student's-t priors as compared to those that use only Student's-t and TV priors in [16] and [14], respectively.

When a TV prior is used within the Bayesian framework in [14] the prior introduces an automatic mechanism for spatially adaptive regularization in the restoration filter. This mechanism is manifested by the diagonal *spatial adaptivity* matrix $[\mathbf{W}(\mathbf{u}^k)]_{ii} = \frac{1}{\sqrt{u_i^k}}$ (Eq. (34) in [14]) which is used in the restoration filter defined in Eq. (48) in [14]. The elements of this matrix are inversely proportional to the square root of the local spatial activity of the pixels of the image. In other words, the value of $u_i^k = \langle (\Delta_i^h(\mathbf{x}))^2 + (\Delta_i^v(\mathbf{x}))^2 \rangle_{q(\mathbf{x})}$ in Eq. (36) of [14] captures the local activity at location i when the "traditional" TV prior model is used.

Very interestingly, this term is the same as the term computed herein as $u_k(i)$ in Eq. (31). The difference between the herein and the [14] approaches lies in the computation of the second term $\mathbf{C}_{2k-r}^j(i,i)$ in Eq. (31). This term *regularizes* the local spatial activity, obtained by the first term $((\mathbf{m}_{2k-r}^j(i))^2$ of Eq. (31) which in flat areas of the image can be zero yielding a spatial adaptivity matrix with infinite valued elements. The exact calculation of the term $\mathbf{C}_{2k-r}^j(i,i)$ is hard. As evidenced from Eq. (32), it requires the evaluation of the diagonal elements of a product of square $N \times N$ matrices which is of the form $\mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^t$ where N is the number of pixels of the image and $\mathbf{Q}$ a convolutional operator. However, $\mathbf{R}$ does not have a form which is amenable to easy inversion. In this work this term is computed by an iterative algorithm which converges to the exact result and gives a $\mathbf{C}_{2k-r}^j(i,i)$ which is spatial variant. This algorithm is explained in more detail in the next section. In [14] this term is approximated by assuming a block-circulant covariance for $q(\mathbf{x})$. This approximation yields a *regularization* term for the visibility weights which is constant for the entire image.

When a Student's-t prior is used in the Bayesian framework in [16] spatial adaptivity is again automatically introduced in the restoration filter. The diagonal matrices $\hat{\mathbf{A}}_l$ in Eq. (3.11) with elements $a_l(i)$ for $l = 1, 2, \ldots, 2P$ given by Eq. (3.13) (both equations in [16]) play this role. Specifically, the $a_l(i)$

are the local precisions of the filter outputs $\epsilon_l(i)$ according to $p(\epsilon_l(i)/a_l(i)) = N(0,(a_l(i))^{-1})$.

Notice that now the spatial adaptivity matrix does not contain a square root, it is just inversely proportional to the local spatial activity captured by the second term of the denominator of Eq. (3.13) in [16]. However, in this case the weights of the spatial adaptivity matrix contain *two regularization* terms that stabilize it in smooth areas of the image. The first one comes from the Gamma hyper-prior and is the ν_k term (degrees of freedom of Student's-t). The second one is identical to the one also found in the TV prior.

Observe, from the discussion above, that when the herein spatially weighted TV prior is used in combination with the marginalization of the spatial adaptivity weights $a_k(i)$ we obtain the Modified Student's-t used in this work, and so, with this combined prior modeling, *both* previously encountered spatial adaptivity mechanisms co-exist. Indeed, the restoration filter in Eq. (27) contains a *spatial adaptivity* diagonal matrix given by the *product* $\langle \mathbf{A}_k \rangle^{j-1} (\mathbf{U}_k^{-1/2})^{j-1}$. The first term of this product stems from the Student's-t nature of our prior while the second one is provided by the TV model. In other words, the spatial adaptivity matrix presented in this paper contains, as expected, both regularization mechanisms explained above. Furthermore, due to their TV origin the weights contain the term $\sqrt{u_k^{j-1}(i)}$ in Eq. (29) which unlike the "regular" (linear) Student's-t case contains a square root.

In section VI in order to visualize the nature of the above spatial adaptivity weights we show images of these weights (in logarithmic scale) for the herein restoration filter and compare them with the corresponding ones in [14] and [16].

V. Computational Implementation

Before analyzing the performance of the proposed image restoration algorithm, let us discuss important implementation issues. One iteration of the proposed algorithm consists of Eqs. (27)-(36). The image estimate is taken to be equal to $\mathbf{m}$ which is obtained by solving the linear system in Eq. (27). The dimensions of the matrices involved in Eq. (27) are $N \times N$, with N the number of pixels in the image. We solve this system iteratively using the conjugate-gradient algorithm [22]. We also utilized this method to evaluate the diagonal elements of matrix $\mathbf{C}_k$ in Eq. (29). More specifically, we utilized the $\mathbf{R}^{-1}$-conjugate vectors $\mathbf{p}_i,\ i = 1,\ldots,L,\ L < N,$ for which $(\mathbf{p}_i^T \mathbf{R}^{-1} \mathbf{p}_i = \delta_{ij})$. Then according to the conjugate-gradient algorithm the image estimate is updated at every iteration as

$$\mathbf{m}_i = \mathbf{m}_{i-1} + a\mathbf{p}_i,$$

where a is a scalar [22]. If the method is allowed to iterate N times we have $\mathbf{P}^T \mathbf{R}^{-1} \mathbf{P} = I$, where $\mathbf{P} = [\mathbf{p}_1 \ \ldots \ \mathbf{p}_N]$ with $\mathbf{p}_i,\ i = 1,\ldots,N$, all the $\mathbf{R}^{-1}$-conjugate vectors. Then, $\mathbf{R} = \mathbf{P}\mathbf{P}^T$ and the diagonal elements of $\mathbf{C}_k = \mathbf{Q}_k \mathbf{R} \mathbf{Q}_k^T$ can be computed by the formula

$$(\mathbf{Q}_k \mathbf{R} \mathbf{Q}_k^T)(i,i) = \sum_{j=1}^{N} \mathbf{p}_j'(i)^2 \approx \sum_{j=1}^{L} \mathbf{p}_j'(i)^2, \qquad (37)$$

where $\mathbf{p}_j' = \mathbf{Q}_k \mathbf{p}_j$. In practice the number of iterations L required for convergence of the conjugate-gradient method is much smaller than N ($L << N$). We found out that for 256×256 images $N = 65,536$, the conjugate-gradient algorithm gives satisfactory results in terms of image restoration with $L \approx 100 - 200$. An increase of $L \approx 1000 - 2000$ did not provide much benefit. At this point it is worth also noting that the Lanczos-based approach which was proposed in [16] required for similar size images $L \approx 1000 - 2000$ to provide similar restoration results. Thus, the herein proposed algorithm is faster than the algorithm in [16].

In [14] where the diagonal elements of matrices of similar structure appear in the computation of the restored image a circulant approximation was used. This approximation implies that all diagonal elements have the same value. In [14] this value was set in a trial and error manner. We found out that the obtained restoration results using the values of $\mathbf{C}_k(i,i)$ as computed by the herein proposed conjugate-gradient approach are noticeable better than using a circulant approximation for $\mathbf{R}$ or omitting the elements $\mathbf{C}_k(i,i)$ from Eq. (31).

The termination criterion we chose is given by

$$|(\mathbf{R}^j)^{-1}\mathbf{m}^j - \beta \mathbf{H}^T \mathbf{g}| > |(\mathbf{R}^{j-1})^{-1}\mathbf{m}^{j-1} - \beta \mathbf{H}^T \mathbf{g}|, \quad (38)$$

where $\mathbf{m}$ is the image estimate at the $j-th$ iteration and it is the solution of the linear system $(\mathbf{R}^j)^{-1}\mathbf{m} = \beta \mathbf{H}^T \mathbf{g}$ that the conjugate-gradients algorithm solves. This termination criterion is heuristic. We observed that as the residual of the conjugate gradient algorithm increased the restoration quality decreased.

The algorithm is initialized by the resulting image estimate of a Bayesian algorithm that uses a spatially invariant simultaneously autoregressive image prior [19]. In other words, we set the initial image estimate $\mathbf{m}^0$ equal to the restored image by this algorithm. The noise precision β is also estimated by the algorithm in [19] and we fix it to this value for the remaining of our algorithm. Thus, the overall algorithm can be summarized in the following steps:

- Initialize $\mathbf{m^0}$ and β with the algorithm using a stationary prior
- Until convergence do
  1. Update the parameters $\mathbf{u}$, λ and ν from equations (31), (34) and (36), respectively
  2. Update the image estimate $\mathbf{m^j}$ from equation (27) along with the diagonal elements of $(\mathbf{Q}_k \mathbf{R}^j \mathbf{Q}_k^T)$ in equation (37)
  3. Check for convergence using (38)

VI. Numerical Experiments

We demonstrate the value of the proposed restoration approach by testing it in experiments with three well known 256×256 input images: *Lena*, *Cameraman* and *Barbara* and one 512×512 image *USC-man*. Every image is blurred with two types of blur; the first blur has the shape of a Gaussian function with shape parameter 9, and the second is uniform with support a rectangular region of dimensions 9×9. The blurred signal to noise ratio ($BSNR$) was used to quantify the noise level:

$$BSNR = 10 \log_{10} \frac{\|\mathbf{Hf}\|_2^2}{N\sigma^2},$$

where σ^2 is the variance of the additive white Gaussian noise (AWGN). Three levels of AWGN were added to the blurred images resulting in $BSNR = 40$, 30 and 20 dB. Thus in total 24 image restoration experiments are presnted.

As performance metric, the improvement in Signal to Noise Ratio ($ISNR$) was used, given by

$$ISNR = 10 \log_{10} \frac{\|\mathbf{f} - \mathbf{g}\|_2^2}{\|\mathbf{f} - \hat{\mathbf{f}}\|_2^2},$$

where $\mathbf{f}$, $\mathbf{g}$ and $\hat{\mathbf{f}}$ are the original, observed degraded and restored images, respectively.

In the implementation of our proposed restoration algorithm $P = 2$ was used. In other words, four filter outputs were used for the prior and it is a product with two terms. The operators $\mathbf{Q}_1$ and $\mathbf{Q}_2$ correspond to the horizontal and vertical first order differences. Thus, these filters are used to model the vertical and horizontal image edge structure, respectively. The output of the Laplacian operator used frequently as a regularizer in stationary regularization approaches [1] and [19]is not appropriate in the context of Student's-t prior models for the image. Such models assume that local differences follow a zero mean Gaussian pdf with a different precision per pixel. When first order differences are used the explanation of this precision is very intuitive. When small (large variance) it implies a discontinuity (edge) between these two pixels. If higher order differences are used the explanation is not as clear since the difference contains contributions from a neighborhood of pixels and the location of the edge now is not obvious. We have verified this observation by extensive numerical experiments.

The other two operators $\mathbf{Q}_3$ and $\mathbf{Q}_4$ are used to model the diagonal edge component contained in the vertical and horizontal directions, respectively. These filters are obtained by convolving the previous horizontal and vertical first order differences filters with fan filters with vertical and horizontal pass-bands, respectively. In our experiments the fan filters in [21] were used. We show the magnitude of the frequency responses of filters $\mathbf{Q}_1, \ldots, \mathbf{Q}_4$ in Fig. 1. The fan filters combined with the difference filters were found empirically to provide better results than the use of the horizontal and vertical difference filters alone. To explain the choice of the fan filters we note that ideally we expect from a filter when applied to an image to produce outputs as close to zero as possible. The first order differences filters have to some extent this property, but at the edges of the image this property is canceled. Thus, more filters are needed that produce outputs closer to zero. The motivation to incorporate the fan filters to our algorithm is the use of them in the contourlet transform [21], which is shown to have more close to zero coefficients than the classical wavelet transform. Their ability to provide closer to zero outputs is interpreted as the ability to capture the correlations of the image edges. Hence, this renders the model more accurate. We must also note a key difference in our model with respect to [21]; in the contourlet transform the Laplacian pyramid is used as a first filter and the fan filters are applied on its output. Here, we have first order differences in the horizontal and vertical direction. For this reason, the filters $\mathbf{Q}_3$ and $\mathbf{Q}_4$ are the result of the vertical and horizontal fan filter applied to $\mathbf{Q}_1$ (horizontal) and $\mathbf{Q}_2$ (vertical), respectively.

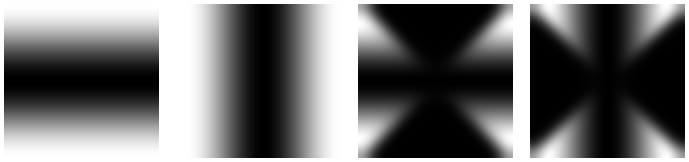
We compared the herein proposed restoration method, abbreviated as CGMK from the first letter of each author's last name, with the *Lena* and *Cameraman* images with four recent TV-based algorithms: the algorithms in [12] and [13] abbreviated by BFO1 and BFO2, respectively, and the algorithms in [14] abbreviated as BMK1 and BMK2. We also compared it with the variational Bayesian algorithm in [16] which is abbreviated as CGLS. This algorithm uses a product of Student's-t image prior and the same four $\mathbf{Q}_l$ as the ones described above.

The *ISNR* results of this comparison are shown in Tables I and II for the experiments with uniform, and Gaussian blurs, respectively. The *ISNR* results with algorithms abbreviated as BMK1, BMK2, BFO1 and BFO2 in Tables I and II are taken directly from [14]. This comparison is limited only to the *Lena* and the *Cameraman* images because in [14] the other two images (*Barbara*, *USC-man*) were not used. We also show an example of the restored images for these 2 experiments in Figure 2. Looking carefully at the restored image by the herein proposed algorithm and comparing them to the one by the approach in [16] we observe that it seem less "blurry". Specifically, it better preserves the edges of the image.

We also present *ISNR* results for similar experiments for the *Barbara* and the *USC man* images in Tables III and IV. These images were selected because they contain *large texture areas*. In these experiments the herein proposed methodology was tested with the approach in [16] and a Bayesian approach which uses a TV prior with *identical* spatial weights across the entire image. This approach was obtained by applying the herein proposed Bayesian inference via the constrained variational approximation methodology to a model that uses the prior in Eq. (4).This approach is not identical to the methodology used in [14] since it does not resort to an approximation of the normalization constant of the prior and is labeled as "Bayesian-TV". The purpose of this experiment was to test the proposed restoration method for images with large texture areas. We also show an example of the restored *USC-man, BSNR=20dB, uniform* 9×9 *blur* image for this experiments in Figure 3. Apart from the *ISNR* metric visually the image in Figure 3(b) (Bayesian-TV restored) is not as sharp as the image in Figure 3(d). Furthermore, the image is Figure 3(c) (restored using Student's-t prior) is "cartoon-like". Thus, the restored image by the herein algorithm in Figure 3(d) apart from being better in terms of *ISNR* is also visually more pleasing.

From the *ISNR* results in Tables I-IV one can say that the proposed method works consistently better for mid-low *BSNR*s (20 and 30 dB) for images with large texture areas. Also for certain images and experiments *USC-man,* 9×9 *blur and* $BSNR = 20$ *dB* the proposed approach provided up 0.44 dB improvement in *ISNR* as compared to its best predecessor. Overall out of the 24 experiments presented in this paper in 18 of them the herein proposed algorithm provided better *ISNR* from all other tested methods.

For 256×256 images the proposed algorithm implemented in Matlab requires 3-5 minutes on a Pentium 4 3.40GHz per-



(a) (b) (c) (d)

Fig. 1: Magnitude of frequency responses of the filters used in the prior: (a) and (b) the horizontal and vertical differences $\mathbf{Q}_1, \mathbf{Q}_2$, (c) $\mathbf{Q}_3$ and (d) $\mathbf{Q}_4$

sonal computer. This is 2-5 times faster than the algorithm in [16]. This difference in speed is attributed to the significantly smaller number of iterations required by the new conjugate-gradient algorithm used to compute the diagonal elements of matrix $\mathbf{C}_k$ in Eq. (31) introduced herein.

VII. Conclusions and Future Work

We presented a new promising image prior that is based on the Modified Student's-t pdf and a variational algorithm that estimates all the parameters of this model automatically and also the restored image. We compared this restoration approach with previous state-of-the-art methods and found out that it compares favorably to them. We also presented a new fast iterative conjugate-gradient-based algorithm to compute the diagonal elements of inverses of very large matrices that cannot be found explicitly. The computation of elements of such matrices is required in many sparse Bayesian models. The herein proposed prior can be used in a Bayesian setting for a variety of other image recovery problems, such as, super-resolution, blind-deconvolution, and tomographic reconstruction. Furthermore, it can be used in other imaging applications where a statistical model for the image is necessary, for example, detection of watermarks in images and image retrieval.

In the future we also plan to extend this prior in a number of ways. First more operators can be used in order to better capture directional dependencies at different scales of the image. Furthermore, we plan to investigate relaxing the independence assumption between the different filter outputs and the adjacent pixels in our image model. Although the relaxation of this assumption seems quite simple and natural it will increase the computational requirements of the proposed approach by at least an order of magnitude. Thus, new computational technologies have to be used such as the use of graphics processors in order to perform these iterative restoration algorithms [25].

Appendix A
Derivation of the VE-Step

In the VE-step the bound must be optimized with respect to $\mathbf{R}$, $\mathbf{m}$ and $q(\mathbf{a})$. The mean field approximation and Eq. (19) yield

TABLE I: $ISNR$'s for the *Lena* and *Cameraman* images. Experiments using uniform 9×9 blur

Uniform blur 9×9		Lena	Cameraman
$BSNR(dB)$	Method	ISNR(dB)	
$BSNR = 40$	$CGMK$	**8.52**	**9.61**
	$CGLS$	8.49	9.53
	$BMK1$	8.34	8.55
	$BMK2$	8.35	8.25
	$BFO1$	8.42	8.57
	$BFO2$	8.37	8.46
$BSNR = 30$	$CGMK$	**6.25**	**6.55**
	$CGLS$	6.10	6.29
	$BMK1$	6.08	5.68
	$BMK2$	5.64	4.65
	$BFO1$	5.89	5.41
	$BFO2$	5.58	4.38
$BSNR = 20$	$CGMK$	**4.24**	**3.55**
	$CGLS$	3.98	3.33
	$BMK1$	4.09	3.31
	$BMK2$	4.14	2.12
	$BFO1$	3.72	2.42
	$BFO2$	3.15	1.94

TABLE II: $ISNR$'s for the *Lena* and *Cameraman* images. Experiments using Gaussian blur (variance 9)

Gaussian blur variance 9		Lena	Cameraman
$BSNR(dB)$	Method	ISNR(dB)	
$BSNR = 40$	$CGMK$	4.64	3.49
	$CGLS$	**4.86**	3.45
	$BMK1$	4.72	**3.51**
	$BMK2$	4.50	3.27
	$BFO1$	4.78	3.39
	$BFO2$	4.49	3.26
$BSNR = 30$	$CGMK$	**4.08**	2.81
	$CGLS$	3.89	2.74
	$BMK1$	3.87	**2.89**
	$BMK2$	3.56	2.47
	$BFO1$	3.87	2.63
	$BFO2$	3.55	2.41
$BSNR = 20$	$CGMK$	**3.09**	2.07
	$CGLS$	2.76	1.86
	$BMK1$	3.02	2.13
	$BMK2$	2.47	**2.23**
	$BFO1$	2.87	1.72
	$BFO2$	2.42	1.42

TABLE III: $ISNR$'s for the *Barbara* and *USC-man* images. Experiments using Gaussian blur (variance 9)

Gaussian blur		Barbara	USC-man
$BSNR(dB)$	Method	*ISNR(dB)*	
$BSNR = 40$	$CGMK$	**1.59**	**4.15**
	$CGLS$	1.53	3.91
	$Bayesian - TV$	1.58	3.95
$BSNR = 30$	$CGMK$	**1.36**	**3.19**
	$CGLS$	1.30	2.95
	$Bayesian - TV$	1.33	2.91
$BSNR = 20$	$CGMK$	**1.16**	**2.20**
	$CGLS$	1.00	1.92
	$Bayesian - TV$	1.12	1.72

(a) (b)

(c) (d)

Fig. 2: Experiment on *Lena* image with Gaussian blur (variance 9) and $BSNR = 20dB$; *ISNR* comparison: (a) Degraded image, (b) Restored with spatially invariant prior, [19], $ISNR = 2.32dB$, (c) restored image with method in [16], $ISNR = 2.76dB$, (d) restored image with the proposed algorithm, $ISNR = 3.09dB$.

TABLE IV: $ISNR$'s for the *Barbara* and *USC-man* images. Experiments using uniform 9×9 blur

Uniform blur		Barbara	USC-man
$BSNR(dB)$	Method	*ISNR(dB)*	
$BSNR = 40$	$CGMK$	6.17	7.12
	$CGLS$	**6.23**	**7.70**
	$Bayesian - TV$	6.29	7.50
$BSNR = 30$	$CGMK$	**3.86**	**5.26**
	$CGLS$	3.75	4.86
	$Bayesian - TV$	3.75	4.89
$BSNR = 20$	$CGMK$	**1.37**	**3.13**
	$CGLS$	1.17	2.69
	$Bayesian - TV$	1.20	2.63

$$
\begin{aligned}
&L^b\left(q\left(\mathbf{a}\right), \theta_1, \theta_2\right) = \\
&\int \prod_{k=1}^{2P} q\left(\boldsymbol{\epsilon}_k; \theta_1\right) \prod_{k=1}^{P} q\left(\mathbf{a}_k\right) \log F\left(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}; \theta_2\right) d\boldsymbol{\epsilon} d\mathbf{a} \\
&- \int \prod_{k=1}^{2P} q\left(\boldsymbol{\epsilon}_k; \theta_1\right) \prod_{k=1}^{P} q\left(\mathbf{a}_k\right) \\
&\qquad \times \log \left(\prod_{k=1}^{2P} q\left(\boldsymbol{\epsilon}_k; \theta_1\right) \prod_{k=1}^{P} q\left(\mathbf{a}_k\right) \right) d\boldsymbol{\epsilon} d\mathbf{a}, \qquad (39)
\end{aligned}
$$

where $\theta_1 = [\mathbf{R}, \mathbf{m}]^T$ and $\theta_2 = [\mathbf{u}, \lambda_1, \ldots, \lambda_P, \nu_1, \ldots, \nu_P]^{\mathrm{T}}$.

(a) (b)

(c) (d)

Fig. 3: Experiment on *USC-man* image with uniform 9×9 blur and $BSNR = 20dB$; *ISNR* comparisons: (a) Degraded image, (b) Restored with Bayesian-TV using the prior in (4), $ISNR = 2.63dB$, (c) restored image with method in [16], $ISNR = 2.69dB$, (d) restored image with the proposed algorithm, $ISNR = 3.13dB$.

To optimize the above function with respect to θ_1, we operate on the function $L'(\theta_1)$, which includes the terms of $L^b\left(q\left(\mathbf{a}\right), \theta_1, \theta_2\right)$ that depend only on θ_1, that is,

$$L'\left(\theta_1\right) = A(\theta_1) - B(\theta_1)$$

where

$$A(\theta_1) = \sum_{k=1}^{P} \int q\left(\boldsymbol{\epsilon}_{2k-1}; \theta_1\right) q\left(\boldsymbol{\epsilon}_{2k}; \theta_1\right) q\left(\mathbf{a}_k\right) \times \log \left[\prod_{l=2k-1}^{2k} p\left(\mathbf{y}_l | \boldsymbol{\epsilon}_l; \theta_2\right) \prod_{i=1}^{N} M\left(\epsilon_{2k}(i), \epsilon_{2k-1}(i), a_k(i); \theta_2\right) \right] \prod_{l=2k-1}^{2k} d\boldsymbol{\epsilon}_l d\mathbf{a}_k$$

$$B(\theta_1) = \sum_{k=1}^{2P} \int q(\boldsymbol{\epsilon}_k;\theta_1) \log q(\boldsymbol{\epsilon}_k;\theta_1)\, d\boldsymbol{\epsilon}_k.$$

We now have

$$\begin{aligned} A(\theta_1) &= \sum_{k=1}^{2P} \Bigg\langle -\frac{\beta}{2} (\mathbf{H}\boldsymbol{\epsilon}_k - \mathbf{y}_k)^{\mathrm{T}} \mathbf{Q}_k^{-\mathrm{T}} \mathbf{Q}_k^{-1} (\mathbf{H}\boldsymbol{\epsilon}_k - \mathbf{y}_k) \\ &- \frac{\lambda_{\lceil \frac{k}{2} \rceil}}{2} \boldsymbol{\epsilon}_k^{\mathrm{T}} \boldsymbol{\Lambda}_{\lceil \frac{k}{2} \rceil} \boldsymbol{\epsilon}_k \Bigg\rangle_{q(\boldsymbol{\epsilon}_k;\theta_1) q(\mathbf{a}_k)} + \text{const} \\ &= -\beta P \|\mathbf{H}\mathbf{m} - \mathbf{g}\|_2^2 - \sum_{k=1}^{2P} \frac{\lambda_{\lceil \frac{k}{2} \rceil}}{2} \mathbf{m}^{\mathrm{T}} \mathbf{Q}_k^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{\lceil \frac{k}{2} \rceil} \mathbf{Q}_k \mathbf{m} \\ &- trace\left\{ \left(\beta P \mathbf{H}^{\mathrm{T}}\mathbf{H} + \sum_{k=1}^{2P} \frac{\lambda_{\lceil \frac{k}{2} \rceil}}{2} \mathbf{Q}_k^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{\lceil \frac{k}{2} \rceil} \mathbf{Q}_k \right) \mathbf{R} \right\} \\ &+ \text{const}, \end{aligned} \tag{40}$$

where $\lceil . \rceil$ denotes the 'ceiling' of a real number and $\boldsymbol{\Lambda}_k$, $\hat{\boldsymbol{\Lambda}}_k$ are diagonal matrices with elements

$$\boldsymbol{\Lambda}_k(i,i) = \frac{a_k(i)}{\sqrt{u_k(i)}}\ , \hat{\boldsymbol{\Lambda}}_k(i,i) = \frac{\langle a_k(i) \rangle_{q(a_k(i))}}{\sqrt{u_k(i)}}, i = 1, ..., N.$$

We also have

$$B(\theta_1) = -P \log\det |\mathbf{R}|\,. \tag{41}$$

Setting the derivative of L' w.r.t $\mathbf{R}$ and $\mathbf{m}$ equal to zero and using Eqs. (40) and (41) yields

$$\frac{\partial L'(\theta_1)}{\partial \mathbf{R}} = 0 \Rightarrow$$

$$\begin{aligned} 0 &= \frac{\partial trace\left\{ \beta P \mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{R} + \sum_{k=1}^{2P} \frac{\lambda_{\lceil \frac{k}{2} \rceil}}{2} \mathbf{Q}_k^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{\lceil \frac{k}{2} \rceil} \mathbf{Q}_k \mathbf{R} \right\}}{\partial \mathbf{R}} \\ &- \frac{P \partial \log\det |\mathbf{R}|}{\partial \mathbf{R}} \end{aligned}$$

$$\Rightarrow \beta P \mathbf{H}^{\mathrm{T}}\mathbf{H} + \sum_{k=1}^{2P} \frac{\lambda_{\lceil \frac{k}{2} \rceil}}{2} \mathbf{Q}_k^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{\lceil \frac{k}{2} \rceil} \mathbf{Q}_k - P\mathbf{R}^{-1} = 0 \Rightarrow$$

$$\mathbf{R} = \left(\beta \mathbf{H}^{\mathrm{T}}\mathbf{H} + \frac{1}{2P} \sum_{k=1}^{2P} \lambda_{\lceil \frac{k}{2} \rceil} \mathbf{Q}_k^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{\lceil \frac{k}{2} \rceil} \mathbf{Q}_k \right)^{-1}.$$

and

$$\frac{\partial L'(\theta_1)}{\partial \mathbf{m}} = 0 \Rightarrow \mathbf{m} = \beta \mathbf{R} \mathbf{H}^{\mathrm{T}} g.$$

The final part of the VE-step is the optimization w.r.t. the function $q(\mathbf{a})$. It is straightforward to verify that this is achieved when

$$q(\mathbf{a}) = \frac{\exp\left(\langle \log F(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}) \rangle_{q(\boldsymbol{\epsilon})} \right)}{\int \exp\left(\langle \log F(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}) \rangle_{q(\boldsymbol{\epsilon})} \right) d\mathbf{a}} = \prod_{\kappa=1}^{P} \prod_{i=1}^{N} q(\mathbf{a}_k(i))\,.$$

The product form is due to

$$\begin{aligned} \exp \langle \log F(\mathbf{y}, \boldsymbol{\epsilon}, \mathbf{a}) \rangle_{q(\boldsymbol{\epsilon})} &\propto \prod_{k=1}^{P} \prod_{i=1}^{N} (a_k(i))^{\frac{\nu_k}{2}+2-1} \\ &\times \exp\left\{ -\frac{\nu_k}{2} a_k(i) - \lambda_k \sqrt{u_k(i)} a_k(i) \right\}. \end{aligned}$$

Hence, each $q(a_k(i))$ is a Gamma distribution:

$$q(a_k(i)) = Gamma\left(a_k(i); \frac{\nu_k}{2} + 2, \frac{\nu_k}{2} + \lambda_k \sqrt{u_k(i)} \right).$$

REFERENCES

[1] N. P. Galatsanos and A. K. Katsaggelos, "Methods for Choosing the Regularization Parameter and Estimating the Noise Variance in Image Restoration and their Relation", *IEEE Trans. on Image Processing* , Vol. 1, No. 3, pp. 322-336, July 1992.

[2] P. J. Huber, *Robust Statistics*, Wiley, 2003.

[3] T. F. Chan, S. Esedoglu, F. Park, and M. H. Yip, Recent developments in total variation image restoration, *in Handbook of Mathematical Models in Computer Vision*, Editors: N. Paragios, Y. Chen and O Faugeras, Springer Verlag, 2005.

[4] T. F. Chan, J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, SIAM, 2005.

[5] L. I. Rudin, S. Osher and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *in Phys. D* Vol. 60, pp. 259-268, 1992.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Verlag, 2006.

[7] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine", *Journal of Machine Learning Research*, Vol. 1, 211-244, 2001.

[8] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models", *In Proceedings of Artificial Intelligence and Statistics* Key West, FL, Jan 3-6, 2003.

[9] D. Tzikas, A. Likas, and N. Galatsanos, "Large Scale Multikernel Relevance Vector Machine for Object Detection ", *International Journal on Artificial Intelligence Tools*, vol. 26, no. 12, pp. 1613-1622, December 2007.

[10] E. Sudderth and M. I. Jordan, "Shared segmentation of natural scenes using dependent Pitman-Yor processes ",*Advances in Neural Information Processing Systems (NIPS) 21*, to appear.

[11] S. Roth, M. J. Black, "Fields of Experts: A Framework for Learning Image Priors", " *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. II, pp. 860-867, June 2005.

[12] J. Bioucas-Dias, M. Figueiredo, J. Oliveira, "Adaptive Bayesian/total-variation image deconvolution: A majorization-minimization approach, " *in Proceedings of European Signal Processing Conference* , Florence, Italy, September 2006.

[13] J. Bioucas-Dias, M. Figueiredo, and J. Oliveira, "Total-variation image deconvolution: A majorization-minimization approach," *in Proceedings of International Conference on Acoustics and Speech and Signal Processing*, ICASSP 2006, May 2006.

[14] S. D. Babacan, R. Molina and A. K. Katsaggelos, "Parameter estimation in TV image restoration using variational distribution approximation," *IEEE Transactions on Image Processing*, vol. 17, no. 3, pp. 326-339, March 2008.

[15] G. Chantas, N. P. Galatsanos, and A. Likas, "Bayesian restoration using a new non-stationary edge-preserving image prior ," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2987-2997, October 2006.

[16] G. Chantas, N. P. Galatsanos, A. Likas and M. Saunders, "Variational Bayesian image restoration based on a product of t-distributions image prior" *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1795-1805, October 2008.

[17] S.D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational Bayesian Blind Deconvolution Using a Total Variation Prior", *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 12-26, Jan. 2009.

[18] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "Variational Bayesian Sparse Kernel-Based Blind Image Deconvolution With Student's-t Priors " *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 753-764, April 2009.

[19] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyper-parameter estimation in image restoration," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231-246, Feb. 1999.

[20] K. Lange, Optimization, *Spinger Texts in Statistics*, Springer-Verlag, 2004.
[21] A. L. Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3089-3101, Oct. 2006.
[22] Y. Saad, *Iterative Methods for Sparse Linear Systems*, Second Edition, Society for Industrial and Applied Mathematics, 2000
[23] D. Sun and W.-K. Cham, "Postprocessing of low bit-rate block DCT coded images based on a fields of experts prior, " *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2743-2751, Nov. 2007.
[24] M. Bertalmio and V. Caselles and B. Rougé and A. Solé, "TV Based Image Restoration with Local Constraints, "*J. Sci. Comput.*, vol. 19, no. 1, pp. 95-122, Dec., 2003.
[25] Fung, J.; Mann, S.; "Using graphics devices in reverse: GPU-based Image Processing and Computer Vision," *2008 IEEE International Conference on Multimedia and Expo* , Page(s): 9 - 12 June 23 -April 26 2008.