Bayesian and Regularization Methods for Hyperparameter Estimation in Image Restoration

Rafael Molina[a], Aggelos K. Katsaggelos[b] and Javier Mateos[a] [1]

a)Departamento de Ciencias de la Computación e I.A. Universidad de Granada. 18071 Granada, España.

b) Department of Electrical and Computer Engineering, Northwestern University, Evanston, Illinois 60208-3118

Abstract

In this paper we propose the application of the hierarchical Bayesian paradigm to the image restoration problem. We derive expressions for the iterative evaluation of the two hyperparameters applying the evidence and MAP analysis within the hierarchical Bayesian paradigm. We show analytically that the analysis provided by the evidence approach is more realistic and appropriate than the MAP approach for the image restoration problem. We furthermore study the relationship between the evidence and an iterative approach resulting from the set theoretic regularization approach for estimating the two hyperparameters, or their ratio, defined as the regularization parameter. Finally the proposed algorithms are tested experimentally.

Keywords

Image restoration, regularization, parameter estimation, hierarchical Bayesian models.

I. Introduction

A standard formulation of the image degradation model is given in lexicographic form by [1]

$$\mathbf{g} = \mathbf{D}\mathbf{f} + \mathbf{w}, \tag{1}$$

where the $p \times 1$ vectors $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{w}$ represent respectively the original image, the available noisy and blurred image and the noise with independent elements of variance $\sigma_{\mathbf{w}}^2 = \beta^{-1}$, and $\mathbf{D}$ represents the known blurring matrix. The images are assumed to be of size $m \times n$, with $p = m \times n$. The restoration problem calls for finding an estimate of $\mathbf{f}$ given $\mathbf{g}$, $\mathbf{D}$ and knowledge about $\mathbf{w}$ and possibly $\mathbf{f}$. For a recent review and classification of the existing restoration techniques see Chapter 1 in [17].

The simplest way to approach the restoration problem is to use least squares estimation

[1]This work has been supported by the "Comisión Nacional de Ciencia y Tecnología" under contract PB93-1110.

and then select $\hat{\mathbf{f}}$, an estimate of the original image, as

$$\hat{\mathbf{f}} = \arg\{ \max_{\mathbf{f}} \frac{1}{Z_{noise}(\beta)} \exp[-\frac{1}{2}\beta \parallel \mathbf{g} - \mathbf{D}\mathbf{f} \parallel^2] \},$$

where $Z_{noise}(\beta) = (2\pi/\beta)^{p/2}$. However, as is well known, this approach does not lead to useful restorations [1], and we have to use our prior knowledge about the original image.

Smoothness constraints on the original image can be incorporated under the form of

$$p(\mathbf{f}|\alpha) \propto \alpha^{q/2} \exp\{-\frac{1}{2}\alpha S(\mathbf{f})\},$$

where $S(\mathbf{f})$ is a non negative quadratic form which usually corresponds to a conditional or simultaneous autoregressive model in the statistical community and to setting constraints on first or second differences in the engineering community (see [31]) and q is the number of positive eigenvalues of S and α is a constant. A form of $S(\mathbf{f})$ which has been used widely in the engineering community is

$$S(\mathbf{f}) = \parallel C\mathbf{f} \parallel^2,$$

where C is the Laplacian operator.

If the hyperparameters α and β were known, then following the Bayesian paradigm it is customary to select, as the restoration of $\mathbf{f}$, the image $\mathbf{f}_{(\alpha,\beta)}$ defined by

$$\mathbf{f}_{(\alpha,\beta)} = \arg\{ \min_{\mathbf{f}} [\alpha S(\mathbf{f}) + \beta \parallel \mathbf{g} - \mathbf{D}\mathbf{f} \parallel^2] \} = \arg\{ \max_{\mathbf{f}} p(\mathbf{f}|\alpha) p(\mathbf{g}|\mathbf{f},\beta) \}. \qquad (2)$$

An important problem arises when α and/or β are unknown. Much interest has centered on the question of how these parameters should be estimated in both the statistical and engineering communities. The ratio of the hyperparameters is typically called the regularization parameter.

To deal with estimation of the hyperparameters α and β, the hierarchical Bayesian paradigm introduces a second stage (the first stage consists of the formulation of $p(\mathbf{f}|\alpha)$ and $p(\mathbf{g}|\mathbf{f},\beta)$). In this stage hyperprior $p(\alpha,\beta)$ is also formulated, resulting in the distribution $p(\alpha,\beta,\mathbf{f},\mathbf{g})$. With the so called the evidence analysis, see [4], [25] for other possible names, $p(\alpha,\beta,\mathbf{f},\mathbf{g})$ is integrated over $\mathbf{f}$ to give the likelihood $p(\alpha,\beta|\mathbf{g})$; this likelihood is then maximized over the hyperparameters. Recently, an alternative procedure has been

suggested by Buntine and Weigend [8], Strauss, *et al.* [37], Wolpert [38], Molina [27] and commented by Archer and Titterington [3]. With this procedure, which is henceforth called the MAP analysis, one integrates $p(\alpha, \beta, \mathbf{f}, \mathbf{g})$ over α and β to obtain the true likelihood, and then maximizes the true posterior over $\mathbf{f}$.

There are a number of approaches from the engineering community to the restoration problem and also to the problem of estimating the regularization parameter ([1], [10], [12], [17]). Among such approaches is a set theoretic regularization approach [19]. According to it the prior knowledge constrains the solution to belong to both ellipsoids

$$Q_{\mathbf{f}} = \{\mathbf{f} | \parallel \mathbf{Cf} \parallel^2 \leq E^2\} \tag{3}$$

and

$$Q_{\mathbf{f/g}} = \{\mathbf{f} | \parallel \mathbf{g} - \mathbf{Df} \parallel^2 \leq \epsilon^2\}, \tag{4}$$

where $S(\mathbf{f}) = \parallel \mathbf{Cf} \parallel^2$ and $\mathbf{C}$, in this context, represents in general a high-pass filter, so that the energy of the restored signal at high frequencies, due primarily to the amplified broad-band noise is bounded. If the bounds ϵ^2 and E^2 are known, and the intersection of $Q_{\mathbf{f}}$ and $Q_{\mathbf{f/g}}$ is not empty, a solution to the problem can be found by solving

$$(\mathbf{D}^t\mathbf{D} + \lambda \mathbf{C}^t\mathbf{C})\mathbf{f} = \mathbf{D}^t\mathbf{g}, \tag{5}$$

where λ the regularization parameter is equal to $(\epsilon/E)^2$. A posterior test needs to be performed to verify that indeed the solution of Eq. (5) belongs to both ellipsoids in Eqs. (3) and (4). In [14], [15], [20], an iterative restoration algorithm is proposed according to which both E and ϵ and the estimation of the original image are updated at each iteration step based on the available partially restored image. It is noted here that when only one of the bounds E or ϵ is known then a constrained minimization problem results. That is, $||\mathbf{Cf}||^2$ is minimized subject to the constraint imposed by Eq. (4), or $||\mathbf{g} - \mathbf{Df}||^2$ is minimized subject to the constraint imposed by Eq. (3). In both cases the solution is found by solving Eq. (5). The regularization parameter λ is now a Lagrange multiplier which needs to be determined so that the constraint is satisfied. It is also noted that Eq. (5) provides also the maximizer of Eq. (2) with $\lambda = \alpha/\beta$.

In this paper we propose the application of the hierarchical Bayesian approach to the image restoration problem. We derive expressions for the iterative evaluation of the two

hyperparameters applying the evidence and MAP approaches. We show analytically that the analysis provided by the evidence approach is more realistic and appropriate than the MAP approach for the image restoration problem. We furthermore study the relationship between the evidence and an iterative approach resulting from the set theoretic regularization approach, for estimating the two hyperparameters, or their ratio, defined as the regularization parameter.

The paper is divided as follows. In section II we describe the hierarchical paradigm to hyperparameter and image estimation. Section III examines the evidence and MAP analyses. Section IV provides reasons for selecting between the evidence and the MAP approaches. Section V studies the set theoretic regularization approach. In section VI we justify the method described in section V from the Bayesian point of view, interpret the parameters involved in the iterative procedure and propose alternative iterative methods. Finally in section VII experimental results are shown and section VIII concludes the paper. The paper also contains an appendix where the posterior distribution of the hyperparameters is studied in order to characterized its local maxima.

II. Hierarchical Bayesian Paradigm

The hierarchical Bayesian paradigm is currently being applied to many areas of research related to image analysis. Buntine [6] has applied this theory to the construction of classification trees and Spiegelhalter and Lauritzen [36] to the problem of refining probabilistic networks. Buntine [7] and Cooper and Herkovsits [9] have used the same framework for constructing such networks. MacKay [23] and Buntine and Weigund [8] use the full Bayesian framework in backpropagation networks and MacKay [24], following Gull [13], applies this framework to interpolation problems.

In the hierarchical approach to image restoration we have at least two stages. In the first stage, knowledge about the structural form of the noise and the structural behavior of the restoration is used in forming $p(\mathbf{f}|\alpha)$ and $p(\mathbf{g}|\mathbf{f},\beta)$, respectively. These noise and image models depend on the unknown hyperparameters. In the second stage the hierarchical Bayesian paradigm defines a hyperprior on the hyperparameters, where information about these hyperparameters is included.

Although in some cases it would be possible to know, from previous experience, relations

between the hyperparameters we shall study here the model where the global probability is defined as

$$p(\alpha, \beta, \mathbf{f}, \mathbf{g}) = p(\alpha)p(\beta)p(\mathbf{f}|\alpha)p(\mathbf{g}|\mathbf{f}, \beta). \tag{6}$$

A. Components of the first stage

The components of the first stage are the noise and image models. We shall only discuss the image models, since the noise model defined by Eq. (1) will be used. Although the theory we shall develop here can be applied to any quadratic prior on $\mathbf{f}$, we shall study in full two particular priors, the ones we have been applying restoration of astronomical images [30], [31] .

Our prior knowledge about the smoothness of the object luminosity distribution makes it possible to model the distribution of $\mathbf{f}$ by a conditional autoregression (CAR) [35]. Thus,

$$p(\mathbf{f}|\alpha) \propto \exp\{-\frac{1}{2}\alpha\, \mathbf{f}^t(I - \phi N)\mathbf{f}\},$$

where the entries of the matrix N, N_{ij}, are equal to 1 if cells i and j are spatial neighbors (pixels at distance one) and zero otherwise and ϕ is equal to 0.25. The term $\mathbf{f}^t(I - \phi N)\mathbf{f}$ represents in matrix notation the sum of squares of the values $\mathbf{f}_i$ minus ϕ times the sum of $\mathbf{f}_i\mathbf{f}_j$ for neighboring pixels i and j (we denote by $\mathbf{f}_i$ the i-th entry of the vector $\mathbf{f}$).

The parameters can be interpreted by the following expressions describing the conditional distribution

$$\begin{aligned} E(\mathbf{f}_i \mid \mathbf{f}_j,\ j \neq i) &= \phi \sum_{j \text{ nhbr } i} \mathbf{f}_j, \\ \text{var}(\mathbf{f}_i \mid \mathbf{f}_j,\ j \neq i) &= \alpha^{-1}, \end{aligned}$$

where the subscript 'j nhbr i' denotes the four neighbor pixels at distance one from pixel i, and the parameter α measures the smoothness of the 'true' image. This model on log scale has been used in the galaxy deconvolution problem [30], [31] , a problem in which no more knowledge than the exponential decay of the luminosity can be incorporated, and on linear scale in planet deconvolution problems [32] .

Assuming a toroidal edge correction, the eigenvalues of the matrix $(I - \phi N)$ are $\lambda_{ij} = 1 - 2\phi(\cos(2\pi i/m) + \cos(2\pi j/n)), i = 1, 2, \ldots, m, j = 1, 2, \ldots, n$. So, for $\phi = 0.25$, $\mathbf{f}$ has a

singular multinormal distribution (see [26]). It can be shown that the density of $\mathbf{f}$ has the form

$$p(\mathbf{f}|\alpha) = \frac{1}{Z_{prior}(\alpha)} \exp\{-\frac{1}{2}\alpha\, \mathbf{f}^t \mathbf{C}\mathbf{f}\},$$

where $Z_{prior}(\alpha) = (\prod_{i,j\neq 0,0} \lambda_{ij})^{-1/2}(2\pi/\alpha)^{(p-1)/2}$, $\mathbf{C} = I - \phi N$ and $\mathbf{f}$ lies on a hyperplane of the form $\sum(\mathbf{f}_i - \mu) = 0$. It is important to note that μ is not determined at all in the model.

We shall also use the simultaneous autoregressive model (SAR) [35]. This model is characterized by

$$\mathbf{f}_i - \phi \sum_{j \text{ nhbr } i} \mathbf{f}_j = \epsilon_i,$$

where ϵ_i is independent and $\mathcal{N}(0, \alpha^{-1})$. Then, the corresponding distribution is given by

$$p(\mathbf{f}|\alpha) = \frac{1}{Z_{prior}(\alpha)} \exp\{-\frac{1}{2}\alpha\, \mathbf{f}^t \mathbf{C}^t\mathbf{C}\, \mathbf{f}\}$$

where, for this prior, we have $Z_{prior}(\alpha) = (\prod_{i,j\neq 0,0} \lambda_{ij}^2)^{-1/2}(2\pi/\alpha)^{(p-1)/2}$, $\mathbf{C} = I - \phi N$ and $\mathbf{f}$ lies again on an hyperplane of the form $\sum(\mathbf{f}_i - \mu) = 0$.

From the regularization point of view, the CAR model imposes constraints on the first differences and the SAR on the second differences of the image. Both of these models are related to autoregressive models which have been used for image modeling and restoration [18] . For notation simplicity we shall use here the SAR image model. Having defined the components of the first stage, we move now on to the second stage.

B. Components of the second stage

Because of the attractiveness of the Bayesian machinery to perform conditional analysis, Berger [4] describes the possibility of using the Bayesian approach when very little prior information is available. According to it, in situations without prior information what is needed is a non informative prior on the hyperparameters, α and β (the term "non informative" is meant to imply that no information about the hyperparameters is contained in the priors). For the problem at hand we can use improper non informative priors $p(\alpha) \propto const$ and $p(\beta) \propto const$ both over $[0, \infty)$.

However, it is also possible, as we shall see now, to incorporate precise prior knowledge about the value of the noise and prior variances. Let us examine the hyperpriors we shall

use.

In general, depending on $p(\alpha)$ and $p(\beta)$ used, $p(\alpha, \beta|\mathbf{g})$ may not be easily computable. A large part of the Bayesian literature is devoted to finding prior distributions for which $p(\alpha, \beta|\mathbf{g})$ can be easily calculated. These are the so called conjugate priors [4], which were developed extensively in Raiffa and Schlaifer [34].

Besides providing for easy calculation of $p(\alpha, \beta|\mathbf{g})$, conjugate priors have, as we shall see later, the intuitive feature of allowing one to begin with a certain functional form for the prior and end up with a posterior of the same functional form, but with parameters updated by the sample information. As described in [4], at least for initial analyses, conjugate priors like the one we are going to use in our problem can be quite useful in practice.

These attractive properties of conjugate priors are, however, only of secondary importance compared to the basic question of whether or not a conjugate prior can be chosen which gives a reasonable approximation to the true prior. This leads us to the problem of model selection, which will not be dealt with in this paper (see [23], [24]).

Taking into account the above considerations about conjugate priors, we use as hyperprior one such conjugate prior, the gamma distribution defined by

$$p(\omega) \propto \omega^{l/2-1} \exp[-a(l-2)\omega],$$

where ω denotes a hyperparameter, a is a constant whose meaning will be made precise later, and l is a non negative quantity. This distribution has the following properties

$$E[w] = \frac{l}{2a(l-2)} \qquad \text{and} \qquad Var[w] = \frac{l}{2a^2(l-2)^2}.$$

So, the mean of w, which represents the inverse of the prior or noise variance, is for l large, approximately equal to $1/2a$, and its variance decreases when l increases. This means that l can then be understood as a measure of the certainty on the knowledge about the prior or noise variances (see [28], [33]).

III. Hierarchical Bayesian Analysis

Having defined $p(\alpha, \beta, \mathbf{f}, \mathbf{g})$, the Bayesian analysis is performed. As mentioned earlier, in the evidence framework, $p(\alpha, \beta, \mathbf{f}, \mathbf{g})$ is integrated over $\mathbf{f}$ to give the evidence $p(\alpha, \beta|\mathbf{g})$

which is then maximized over the hyperparameters while the MAP framework $p(\alpha, \beta, \mathbf{f}, \mathbf{g})$ is integrated over α and β to obtain the true likelihood which is then maximized with respect to $\mathbf{f}$. Let us examine these analyses in details.

A. Evidence Analysis

In this approach $\hat{\alpha}$ and $\hat{\beta}$ are first selected as

$$\hat{\alpha}, \hat{\beta} = \arg\max_{\alpha,\beta} p(\alpha, \beta|\mathbf{g}), \tag{7}$$

and $\mathbf{f}_{(\hat{\alpha},\hat{\beta})}$, defined in Eq. (2), is then selected as the restored image.

A.1 Evidence Analysis with Flat Hyperpriors

With the improper hyperpriors (the gamma hyperprior case will be examined later), Eq. (7) amounts to selecting $\hat{\alpha}, \hat{\beta}$ as the maximum likelihood estimates (*mle*), of α, β from $p(\mathbf{g}|\alpha, \beta)$.

Let us examine the estimation process in detail. Fixing α and β and expanding the function $M(\mathbf{f}, \mathbf{g}|\alpha, \beta) = \alpha \parallel \mathbf{Cf} \parallel^2 + \beta \parallel \mathbf{g} - \mathbf{Df} \parallel^2$ around $\mathbf{f}_{(\alpha,\beta)}$, we have

$$M(\mathbf{f}, \mathbf{g}|\alpha, \beta) = M(\mathbf{f}_{(\alpha,\beta)}, \mathbf{g}|\alpha, \beta) + \frac{1}{2}(\mathbf{f} - \mathbf{f}_{(\alpha,\beta)})^t Q(\alpha, \beta)(\mathbf{f} - \mathbf{f}_{(\alpha,\beta)})$$

and therefore

$$\begin{aligned} p(\alpha, \beta|\mathbf{g}) &\propto p(\mathbf{g}|\alpha, \beta) && (8)\\ &= \frac{\exp\{-M(\mathbf{f}_{(\alpha,\beta)}, \mathbf{g}|\alpha, \beta)/2\}}{Z_{prior}(\alpha) Z_{noise}(\beta)} \int_{\mathbf{f}} \exp\{-\frac{1}{2}(\mathbf{f} - \mathbf{f}_{(\alpha,\beta)})^t Q(\alpha, \beta)(\mathbf{f} - \mathbf{f}_{(\alpha,\beta)})\}\, d\mathbf{f} && (9)\\ &= \frac{\exp\{-M(\mathbf{f}_{(\alpha,\beta)}, \mathbf{g}|\alpha, \beta)/2\}|Q(\alpha, \beta)|^{-1/2}}{Z_{prior}(\alpha) Z_{noise}(\beta)}, && (10) \end{aligned}$$

where $Q(\alpha, \beta) = \alpha \mathbf{C}^t\mathbf{C} + \beta \mathbf{D}^t\mathbf{D}$.

It is important to note that in cases where the image or the noise models are not Gaussian (for example entropy-based image models), the approximation of $p(\mathbf{g}|\alpha, \beta)$ obtained by expanding $\log p(\mathbf{f}, \mathbf{g}|\alpha, \beta)$ around $\mathbf{f}_{(\alpha,\beta)}$ and integrating on $\mathbf{f}$ is still usefull (see for instance [13]). Note, however, that for the image and noise models used in this paper the integration is exact.

We then have

$$\begin{aligned}\log p(\alpha,\beta|\mathbf{g}) &= -\alpha \parallel \mathbf{C}\mathbf{f}_{(\alpha,\beta)} \parallel^2 -\beta \parallel \mathbf{g}-\mathbf{D}\mathbf{f}_{(\alpha,\beta)}||^2 \\ &\quad -\log|Q(\alpha,\beta)|^{1/2} - \log Z_{prior}(\alpha) - \log Z_{noise}(\beta) + const.\end{aligned}$$

As MacKay [24] points out, the term $\parallel \mathbf{g}-\mathbf{D}\mathbf{f}_{(\alpha,\beta)}||^2$ represents the misfit of the restoration to the observation, the term $\alpha \parallel \mathbf{C}\mathbf{f}_{(\alpha,\beta)} \parallel^2$ measures how far $\mathbf{f}_{(\alpha,\beta)}$ is from its null value and $(2\pi)^{p/2}|Q(\alpha,\beta)|^{-1/2}/Z_{prior}(\alpha)$ is the ratio of the posterior accesible volume in the parameter space to the prior accesible volume.

We now differentiate $-2\log p(\alpha,\beta|\mathbf{g})$ with respect to α and β so as to find the conditions which are satisfied at the maxima. We have

$$\parallel \mathbf{C}\mathbf{f}_{(\alpha,\beta)} \parallel^2 + trace[Q(\alpha,\beta)^{-1}\mathbf{C}^t\mathbf{C}] = (p-1)/\alpha \tag{11}$$

$$\parallel \mathbf{g}-\mathbf{D}\mathbf{f}_{(\alpha,\beta)} \parallel^2 + trace[Q(\alpha,\beta)^{-1}\mathbf{D}^t\mathbf{D}] = p/\beta. \tag{12}$$

A study of these two equations in order to characterize the stationary points of $p(\alpha,\beta|\mathbf{g})$ and examine whether they are local maxima is found in appendix I. Let us now examine how to find those stationary points.

Let us use the EM-algorithm [11] with $\mathcal{X} = (\mathbf{f},\mathbf{g})^t$ and $\mathcal{Y} = \mathbf{g} = \mathcal{T}\mathcal{X}$ to iteratively increase $p(\alpha,\beta|\mathbf{g})$. We obtain the following iterative procedure for α and β (see [21] for details),

$$\alpha_{i+1}^{-1} = \{\parallel \mathbf{C}\mathbf{f}_{(\alpha_i,\beta_i)} \parallel^2 + trace[Q(\alpha_i,\beta_i)^{-1}\mathbf{C}^t\mathbf{C}]\}/(p-1), \tag{13}$$

$$\beta_{i+1}^{-1} = \{\parallel \mathbf{g}-\mathbf{D}\mathbf{f}_{(\alpha_i,\beta_i)} \parallel^2 + trace[Q(\alpha_i,\beta_i)^{-1}\mathbf{D}^t\mathbf{D}]\}/p. \tag{14}$$

This iterative scheme finds the stationary points of $p(\alpha,\beta|\mathbf{g})$, (see [5], [39] for the characteristics of these stationary points found by the EM-algorithm; the proofs in [11] are flawed) and can be seen as an iterative procedure derived from Eqs. (11) and (12).

A.2 Evidence Analysis with Gamma Hyperpriors

We now generalize the above result by considering the following gamma distributions as hyperpriors

$$p(\alpha) \propto \alpha^{l_1/2-1}\exp[-a_1(l_1-2)\alpha], \tag{15}$$

and

$$p(\beta) \propto \beta^{l_2/2-1} \exp[-a_2(l_2-2)\beta]. \tag{16}$$

Clearly improper hyperpriors correspond to $l_1 = l_2 = 2$. Differentiating the corresponding $-2\log p(\alpha, \beta|\mathbf{g})$ with respect to α and β, where $p(\alpha, \beta|\mathbf{g})$ can be easily obtained multiplying $p(\mathbf{g}|\alpha, \beta)$ in Eq. (10) by $p(\alpha)p(\beta)$, we have

$$\| \mathbf{Cf}_{(\alpha,\beta)} \|^2 + trace[Q(\alpha,\beta)^{-1}\mathbf{C}^t\mathbf{C}] + (l_1-2)2a_1 = (p-1+l_1-2)/\alpha, \tag{17}$$

$$\| \mathbf{g} - \mathbf{Df}_{(\alpha,\beta)} \|^2 + trace[Q(\alpha,\beta)^{-1}\mathbf{D}^t\mathbf{D}] + (l_2-2)2a_2 = (p+l_2-2)/\beta. \tag{18}$$

The application of the EM-algorithm results now in the following iterative procedure for α and β,

$$\alpha_{i+1}^{-1} = \omega_1\{\| \mathbf{Cf}_{(\alpha_i,\beta_i)} \|^2 + trace[Q(\alpha_i,\beta_i)^{-1}\mathbf{C}^t\mathbf{C}]\}/(p-1) + (1-\omega_1)2a_1, \tag{19}$$

$$\beta_{i+1}^{-1} = \omega_2\{\| \mathbf{g} - \mathbf{Df}_{(\alpha_i,\beta_i)} \|^2 + trace[Q(\alpha_i,\beta_i)^{-1}\mathbf{D}^t\mathbf{D}]\}/p + (1-\omega_2)2a_2, \tag{20}$$

where

$$\omega_1 = (p-1)/(p-1+l_1-2) \quad \text{and} \quad \omega_2 = p/(p+l_2-2).$$

There is a very nice and intuitive interpretation for this new iterative procedure. We are estimating the unknown hyperparameters by weighting their maximum likelihood estimates with our prior knowledge regarding their means and variances. So if we know, let us say from previous experience, the noise or image variances with some degree of certainty we can use this knowledge to guide the convergence of the iterative procedure.

As with flat priors for the hyperparameters, it is possible to study the characteristics of the stationary points of $p(\alpha, \beta|\mathbf{g})$ for gamma hyperpriors and, with the use of the DFT, it is also possible to see whether we have local maxima, see appendix I.

The values of α_{i+1} and β_{i+1} obtained from iterations (19) and (20) will be denoted by $\alpha_{(i+1)(EM,a\_\omega)}$ and $\beta_{(i+1)(EM,a\_\omega)}$, respectively, where a_ω is used to denote the vector $(2a_1, \omega_1, 2a_2, \omega_2)$. Clearly for $l_1 = l_2 = 2$, iterations (19) and (20) reduce respectively to iterations (13) and (14). We also use $\hat{\alpha}_{(EM,a\_\omega)}$ and $\hat{\beta}_{(EM,a\_\omega)}$ to denote the corresponding evidence solutions, i.e., the values of the hyperparameters $\alpha_{(i+1)(EM,a\_\omega)}$ and $\beta_{(i+1)(EM,a\_\omega)}$, respectively, at convergence.

B. MAP Analysis

We briefly describe next the MAP analysis approach (named *empirical* by Berger [4]), when the gamma distribution is use for the hyperpriors. We have (see [28] and [29] for details),

$$\begin{aligned} p(\mathbf{f},\mathbf{g}) \quad \propto \quad & \int_\alpha\int_\beta \alpha^{-l/2-1}\frac{1}{Z_{prior}(\alpha)}\exp\{-\frac{1}{2}\alpha[\| \mathbf{Cf} \|^2 +2a_1(l_1-2)]\} \\ & \times\beta^{-l_2/2-1}\frac{1}{Z_{noise}(\beta)}\exp\{-\frac{1}{2}\beta[\| \mathbf{g}-\mathbf{Df} \|^2 +2a_2(l_2-2)]\}\, d\alpha\, d\beta, \end{aligned} \tag{21}$$

where a_1, a_2, l_1 and l_2 are constants whose meaning has been made precise in the previous section.

Using the well known result from the gamma distribution

$$\int_0^\infty y^{u-1}e^{-ay}dy = \Gamma(u)a^{-u}$$

we have

$$p(\mathbf{f},\mathbf{g}) \propto [\| \mathbf{Cf} \|^2 +2a_1(l_1-2)]^{-(p-1)/2-l_1/2}\,[\,\| \mathbf{g}-D\mathbf{f} \|^2 \,+2a_2(l_2-2)]^{-p/2-l_2/2}.$$

Thus, to obtain a restored image we would just minimize

$$\frac{p-1+l_1}{2}\log[\| \mathbf{Cf} \|^2 +2a_1(l_1-2)] + \frac{p+l_2}{2}\log[\| \mathbf{g}-\mathbf{Df} \|^2 +2a_2(l_2-2)]. \tag{22}$$

This equation suggests the following iterative scheme to obtain a MAP solution (see Eq. (3) in [27] for details for flat hyperpriors). Starting at step 0 with $\mathbf{f}^0$, we will see in section VII how to choose this image, we use at step i of the iterative procedure

$$\alpha_{i-1}^{-1} = \omega_1 \| \mathbf{Cf}^{i-1} \|^2 /(p-1) + (1-\omega_1)2a_1\frac{l_1-2}{l_1}, \tag{23}$$

and

$$\beta_{i-1}^{-1} = \omega_2 \| \mathbf{g}-\mathbf{Df}^{i-1} \|^2 /p + (1-\omega_2)2a_2\frac{l_2-2}{l_2}, \tag{24}$$

and then define $\mathbf{f}^i$ as $\mathbf{f}_{(\alpha_{i-1},\beta_{i-1})}$, defined by Eq. (2), where

$$\omega_1 = \frac{p-1}{p-1+l_1} \text{ and } \omega_2 = \frac{p}{p+l_2}.$$

Note that we are not really concerned with the estimation α and β in the MAP analysis and that the above equations can be understood as intermediate steps to calculate $\hat{\mathbf{f}}$. However, these intermediate steps accept a nice interpretation, as it was the case with the evidence framework. That is, we are estimating the unknown hyperparameters by weighting the estimation of the noise and prior variances, which come fully from the restoration without involving the trace function, with our prior knowledge regarding their means and variances.

We will show now that if $\mathbf{f}^i = (1-\mu)\mathbf{f}^{i-1} + \mu\mathbf{f}_{(\alpha_{i-1},\beta_{i-1})}$, where $\mathbf{f}^{i-1}$ is the current restoration at the $(i-1)$st iteration step and with μ close to zero the method is guaranteed to converge. We have used $\mu = 1$ in all the experiments and we have not observed any convergence problems.

Let us examine the convergence of the iterative procedure. Forming α_{i-1}^{-1} and β_{i-1}^{-1} as defined in Eqs. (23) and (24),expanding $M(\mathbf{f},\mathbf{g}|\alpha_{i-1},\beta_{i-1})$ around $\mathbf{f}^{i-1}$ and taking into account that $\mathbf{f}_{(\alpha_{i-1},\beta_{i-1})}$ minimizes the function $M(\mathbf{f},\mathbf{g}|\alpha_{i-1},\beta_{i-1})$ and $\alpha_{i-1}\mathbf{C}^t\mathbf{C}+\beta_{i-1}\mathbf{D}^t\mathbf{D}$ is non-negative definite, we have that

$$(\mathbf{f}_{(\alpha_{i-1},\beta_{i-1})} - \mathbf{f}^{i-1})^t(\alpha_{i-1}\mathbf{C}^t\mathbf{C}\mathbf{f}^{i-1} - \beta_{i-1}\mathbf{D}^t(\mathbf{g}-\mathbf{D}\mathbf{f}^{i-1})) < 0,$$

and so, for μ close enough to zero

$$\begin{aligned}
&(p-1+l_1)\log[\parallel \mathbf{C}\mathbf{f}^i \parallel + 2a_1\frac{l_1-2}{l_1}] + (p+l_2)2\log[\parallel \mathbf{g}-\mathbf{D}\mathbf{f}^i \parallel^2 + 2a_2l_2] \approx \\
&\quad (p-1+l_1)\log[\parallel \mathbf{C}\mathbf{f}^{i-1} \parallel^2 + 2a_1\frac{l_1-2}{l_1}] + (p+l_2)\log[\parallel \mathbf{g}-\mathbf{D}\mathbf{f}^{i-1} \parallel^2 + 2a_2\frac{l_2-2}{l_2}] \\
&\quad +\mu(\mathbf{f}_{(\alpha_{i-1},\beta_{i-1})} - \mathbf{f}^{i-1})^t(\alpha_{i-1}\mathbf{C}^t\mathbf{C}\mathbf{f}^{i-1} - \beta_{i-1}\mathbf{D}^t(\mathbf{g}-\mathbf{D}\mathbf{f}^{i-1})) \\
&\quad < \frac{p-1+l_1}{2}\log[\parallel \mathbf{C}\mathbf{f}^{i-1} \parallel^2 + 2a_1\frac{l_1-2}{l_1}] + \frac{p+l_2}{2}\log[\parallel \mathbf{g}-\mathbf{D}\mathbf{f}^{i-1} \parallel^2 + 2a_2\frac{l_2-2}{l_2}],
\end{aligned}$$

And so the method converges to a solution of

$$(p-1+l_1)\frac{\mathbf{C}^t\mathbf{C}\mathbf{f}}{\parallel \mathbf{C}\mathbf{f} \parallel^2 + 2a_1(l_1-2)} - (p+l_2)\frac{\mathbf{D}^t(\mathbf{g}-\mathbf{D}\mathbf{f})}{\parallel \mathbf{g}-\mathbf{D}\mathbf{f} \parallel^2 + 2a_2(l_2-2)} = 0.$$

IV. Relationship between the Evidence and MAP analyses

Let us examine how the evidence and MAP approaches are related. We have

$$p(\mathbf{f}|\mathbf{g}) = \int_\alpha\int_\beta p(\mathbf{f},\alpha,\beta|\mathbf{g})d\alpha d\beta = \int_\alpha\int_\beta p(\alpha,\beta|\mathbf{g})p(\mathbf{f}|\mathbf{g},\alpha,\beta)d\alpha d\beta.$$

Therefore, $p(\mathbf{f}|\mathbf{g})$ is close to $p(\mathbf{f}|\mathbf{g},\hat{\alpha},\hat{\beta})$ only when $p(\alpha,\beta|\mathbf{g})$ is sharply peaked at $\hat{\alpha},\hat{\beta}$. This is not the case for typical images, such as Lena (see also [13]).

However, the main difference between both analyses comes from the following facts. Following MacKay [25], consider the Gaussian distribution

$$p(\mathbf{f}) = (1/[\sqrt{2\pi}\sigma_{\mathbf{f}}])^k \exp[-\sum_1^k f_i^2/(2\sigma_{\mathbf{f}}^2)].$$

Nearly all the probability mass of this Gaussian is in a shell of radius $r = \sqrt{k}\sigma_{\mathbf{f}}$ and thickness $\propto r/\sqrt{k}$. For example, in 1000 dimensions, 90% of the mass of a Gaussian with $\sigma_{\mathbf{f}} = 1$ is in a shell of radius 31.6 and thickness 2.8. However, the probability density at the origin is $\exp[k/2]$ bigger than the density at this shell where most of the probability mass is, and where the main contribution to the estimation of the variance will come from.

In summary, probability density maxima often have very little associated probability mass, even though the value of the probability density there may be immense, because they have so little associated volume. So, the values provided by $\| \mathbf{g} - \mathbf{D}\mathbf{f}_{(\alpha,\beta)} \|^2$ and $\| \mathbf{C}\mathbf{f}_{(\alpha,\beta)} \|^2$ are not good estimators of β and α, respectively, since the estimator $\mathbf{f}_{(\alpha,\beta)}$ is not located where most of the mass of the distribution is. Furthermore, as a referee points out, while α and β were originally independent $\mathbf{f}_{(\alpha,\beta)}$ correlate them.

Finally, the following proposition tells us where the values of α and β should be.

Proposition. 1: Consider a sample $\mathbf{f}$ from $p(\mathbf{f}|\mathbf{g},\hat{\alpha}_{(a\_\omega)},\hat{\beta}_{(a\_\omega)})$. Then, the conditional average values of $\| \mathbf{C}\mathbf{f} \|^2$ and $\| \mathbf{g} - \mathbf{D}\mathbf{f} \|^2$ satisfy

$$\begin{aligned}
\hat{\alpha}_{(a_\omega)} &= \frac{p - 1 + l_1 - 2}{E[\| \mathbf{C}\mathbf{f} \|^2 \, |\mathbf{g},\hat{\alpha}_{(a_\omega)},\hat{\beta}_{(a_\omega)}] + (l_1 - 2)2a_1}, \\
\hat{\beta}_{(a_\omega)} &= \frac{p + l_2 - 2}{E[\| \mathbf{g} - \mathbf{D}\mathbf{f} \|^2 \, |\mathbf{g},\hat{\alpha}_{(a_\omega)},\hat{\beta}_{(a_\omega)}] + (l_2 - 2)2a_2}.
\end{aligned}$$

Proof.

Since

$$\mathbf{f}|\mathbf{g},\hat{\alpha}_{(a\_\omega)},\hat{\beta}_{(a\_\omega)} \sim \mathcal{N}(\mathbf{f}_{(\hat{\alpha}_{(a\_\omega)},\hat{\beta}_{(a\_\omega)})}, (\hat{\alpha}_{(a\_\omega)}\mathbf{C}^t\mathbf{C} + \hat{\beta}_{(a\_\omega)}\mathbf{D}^t\mathbf{D})^{-1})$$

and

$$E[\| \mathbf{g} - \mathbf{D}\mathbf{f} \|^2 | \, \mathbf{g},\hat{\alpha}_{(a\_\omega)},\hat{\beta}_{(a\_\omega)}] = \{\| \mathbf{C}\mathbf{f}_{(\hat{\alpha}_{(a\_\omega)},\hat{\beta}_{(a\_\omega)})} \|^2 + trace[Q(\hat{\alpha}_{(a\_\omega)},\hat{\beta}_{(a\_\omega)})^{-1}\mathbf{C}^t\mathbf{C}]\},$$

and

$$E[\| \mathbf{Cf} \|^2 | \mathbf{g}, \hat{\alpha}_{(a\_\omega)}, \hat{\beta}_{(a\_\omega)}] = \{\| \mathbf{g} - \mathbf{Df}_{(\hat{\alpha}_{(a\_\omega)}, \hat{\beta}_{(a\_\omega)})} \|^2 + trace[Q(\hat{\alpha}_{(a\_\omega)}, \hat{\beta}_{(a\_\omega)})^{-1}\mathbf{D}^t\mathbf{D}]\}$$

then using Eqs. (17) and (18) we obtain the given results.

Finally, let us examine and make even clearer the differences between the MAP and the evidence approaches with the following example. Let us use the simple image model,

$$p(\mathbf{f}|\alpha) \propto \alpha^p \exp[-\alpha \sum_i f_i^2/2]$$

and

$$p(\mathbf{g}|\beta) \propto \beta^p \exp[-\beta \sum_i (g_i - f_i)^2/2].$$

If we use improper priors for α and β and follow the MAP approach we have

$$p(\mathbf{f}, \mathbf{g}) \propto [\| \mathbf{f} \|^2]^{-p/2} [\| \mathbf{g} - \mathbf{f} \|^2]^{-p/2}.$$

Therefore, the solutions provided by the MAP analysis are $f_i = 0, \ \forall i$ or $f_i = g_i, \ \forall i$. This is a typical behavior of the MAP approach. In particular for the CAR and SAR image models we have found that it tends to find the flat solution, $f_i = const$, since that is a maximum of $p(\mathbf{f}, \mathbf{g})$. This can be avoided using a prior whose corresponding eigenvalues are all positive, as was done in [27] where we used $\phi < 0.25$. However, the problem of finding an accurate estimate of this parameter remains. This is a fact that has not been taken into account in some papers using the MAP approach, see for instance [3] .

Let us examine how the evidence analysis works in this simplified example. The corresponding equations to estimate α and β are

$$\begin{aligned} \| \mathbf{f}_{(\alpha,\beta)} \|^2 + p\frac{1}{\alpha+\beta} &= p/\alpha, \\ \| \mathbf{g} - \mathbf{f}_{(\alpha,\beta)} \|^2 + p\frac{1}{\alpha+\beta} &= p/\beta. \end{aligned}$$

The only maximum is for $\alpha = \beta$, which provides the solution $f_i = g_i/2, \forall i$. Therefore, the analysis provided by the evidence approach is much more realistic and appropriate for the image restoration problem.

V. Iterative Evaluation of the Regularization Parameter

In a series of papers [14], [15], [20] we have introduced and used a paradigm in which parameters required for obtaining a solution to the restoration problem, such as α and β, are evaluated iteratively based on the partially restored image. Using a set theoretic formulation of the restoration problem [19], it was proposed in [14], see also [2] for an alternative formulation, to obtain a restored image $\hat{\mathbf{f}}$ by solving

$$\Phi(\tilde{\mathbf{f}}) = \frac{\| \mathbf{g} - \mathbf{D}\tilde{\mathbf{f}} \|^2 + \delta_1}{\| \mathbf{C}\tilde{\mathbf{f}} \|^2 + \delta_2} \mathbf{C}^t\mathbf{C}\tilde{\mathbf{f}} - \mathbf{D}^t(\mathbf{g} - \mathbf{D}\tilde{\mathbf{f}}) = 0. \tag{25}$$

Such a solution is obtained using a successive approximations iteration, that is,

$$\mathbf{f}^i = \mathbf{f}^{i-1} - \epsilon\Phi(\mathbf{f}^{i-1}), \tag{26}$$

where $\hat{\mathbf{f}}^0 = \mathbf{g}$ and the relaxation parameter ϵ is chosen in such a way to insure convergence, see [16]. The regularization parameter $\lambda = \alpha/\beta$ is updated at each iteration step according to

$$\lambda_i = \frac{\| \mathbf{g} - \mathbf{D}\mathbf{f}^i \|^2 + \delta_1}{\| \mathbf{C}\mathbf{f}^i \|^2 + \delta_2}. \tag{27}$$

The quantities δ_1 and δ_2 are introduced in this context in order to keep the quantities in the numerator and denominator of λ away from zero. Their values are determined from the convergence analysis of Eq. (26). The convergence analysis, the derivation of the optimal range of λ and experimental results are shown in [14], [15], [20] . In [20] the value of δ_1 was set equal to zero, since it was shown experimentally to provide better results. The value of δ_2 on the other hand is considered to be a function of the iteration step, since it may be required to change along the iteration so that the sufficient conditions for convergence are satisfied. The steps in the convergence analysis are to first linearize the iteration and then determine the sufficient condition for convergence by forcing the difference between two iteration steps to go to zero. The relaxation parameter parameter used is set equal to one, and therefore an inequality (sufficient condition) for δ_2 results. In [15] the quantity in the numerator of λ is assumed to be known exactly, that is, it is replaced by a constant.

The iterative procedure takes the form

$$\mathbf{f}^{i+1} = (I - \epsilon\frac{\| \mathbf{g} - \mathbf{D}\mathbf{f}^i \|^2}{\| \mathbf{C}\mathbf{f}^i \|^2 + \delta_i} \mathbf{C}^t\mathbf{C})\mathbf{f}^i + \epsilon\mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}^i), \tag{28}$$

with $\mathbf{f}^0 = \mathbf{g}$. According to iteration (28) the regularization parameter is updated at each iteration, in the form

$$\lambda_i = \frac{\| \mathbf{g} - \mathbf{D}\mathbf{f}^i \|^2}{\| \mathbf{C}\mathbf{f}^i \|^2 + \delta_i}.$$

We henceforth refer to iteration (28) as the set theoretic (ST) iteration and δ_i is denoted by $\delta_{i(ST,0,1,0,1)}$. A lower bound for δ_i and an upper bound for ϵ are derived in [20] based on convergence analysis. An estimate of the noise variance is also obtained at each iteration step, equal to $\| \mathbf{g} - \mathbf{D}\mathbf{f}^i \|^2 / p$. This estimate will be improved in the coming section.

VI. Relationships between the hyperpriors and regularization parameter

A. MAP analysis

We can rewrite the expression for λ_i in Eq. (27) as follows

$$\lambda_i = \frac{0.5 \| \mathbf{g} - \mathbf{D}\hat{\mathbf{f}}^i \|^2 / p + 0.5\delta_1/p}{0.5 \| \mathbf{C}\hat{\mathbf{f}}^i \|^2 / p + 0.5\delta_2/p}. \tag{29}$$

By comparing Eq. (29) to Eqs. (23) and (24), and by using the fact that $p - 1 \approx p$, we see that the form of the regularization parameter and the form of α/β in the MAP analysis are the same for $2a_2 \approx \delta_1/p$, $2a_1 \approx \delta_2/p$, and $\omega_1 = \omega_2 = 0.5$.

So, in principle, the MAP approach is more similar to the regularization approach using $a_2 = 0$ and $2a_1 = \delta_2/p$. However, if we observe the tests performed in [20], the extremely high values used for δ_2, in comparison to the corresponding values of $\| C\mathbf{f} \|^2$, ranging from exp[13.5] to exp[18], led us to try an alternative interpretation using the evidence analysis which seems to be more appropriate for the image restoration problem, as we will show next.

B. Evidence analysis

B.1 Hyperpriors with flat distribution

Let us use initially a flat improper hyperprior (the gamma hyperprior will be described in the next section). Equation (14) can be rewritten as,

$$\| \mathbf{g} - \mathbf{D}\mathbf{f}_{(\alpha_i,\beta_i)} \|^2 / l(\alpha_i, \beta_i) = \frac{p-1}{\beta_{(i+1)}}, \tag{30}$$

where

$$l(\alpha_i, \beta_i) = p/(p-1) - \beta_{i+1} trace[Q^{-1}(\alpha_i, \beta_i)\mathbf{D}^t\mathbf{D}]/(p-1),$$

with β_{i+1} obtained from Eq. (14).

Thus, the EM iterative procedure described in Eqs. (13) and (14) can be rewritten as follows. Starting at $\alpha_0, \beta_0, \mathbf{f}_{(\alpha_0,\beta_0)}$ define

$$\lambda_{i+1} = \frac{\alpha_{i+1}}{\beta_{i+1}} = \frac{\| \mathbf{g} - \mathbf{D}\mathbf{f}_{(\alpha_i,\beta_i)} \|^2}{\| \mathbf{C}\mathbf{f}_{(\alpha_i,\beta_i)} \|^2 + \delta_i} \tag{31}$$

where

$$\delta_i = l(\alpha_i,\beta_i) trace[Q^{-1}(\alpha_i,\beta_i)\mathbf{C}^t\mathbf{C}] - (1 - l(\alpha_i,\beta_i)) \| \mathbf{C}\mathbf{f}_{(\alpha_i,\beta_i)} \|^2 \tag{32}$$

and calculate its corresponding MAP solution from Eq. (5). We henceforth refer to this iteration as the EM iteration and δ_i in Eq. (32) is denoted by $\delta_{i(EM,0,1,0,1)}$.

It is clear that the analysis we have carried out is based on calculating the maximum a posteriori values of α and β iteratively. However, the iterative procedure described in the previous section (see [20]) is based on a gradient descendent method. Let us try to adapt this new iterative method to the form of the iteration (28).

Starting with $\mathbf{f}_0, \alpha_0, \beta_0$ define

$$\alpha_{i+1}^{-1} = \{\| \mathbf{C}\mathbf{f}^i \|^2 + trace[Q^{-1}(\alpha_i,\beta_i)\mathbf{C}^t\mathbf{C}]\}/(p-1), \tag{33}$$

$$\beta_{i+1}^{-1} = \{\| \mathbf{g} - \mathbf{D}\mathbf{f}^i \|^2 + trace[Q^{-1}(\alpha_i,\beta_i)\mathbf{D}^t\mathbf{D}]\}/p, \tag{34}$$

and use the iterative scheme

$$\mathbf{f}^{i+1} = \mathbf{f}^i - \rho\beta_{i+1}[\frac{\alpha_{i+1}}{\beta_{i+1}}\mathbf{C}^t\mathbf{C}\mathbf{f}^i - \mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}^i)]. \tag{35}$$

The important point is that α_{i+1}/β_{i+1} has the form

$$\frac{\alpha_{i+1}}{\beta_{i+1}} = \frac{\| \mathbf{g} - \mathbf{D}\mathbf{f}^i \|^2}{\| \mathbf{C}\mathbf{f}^i \|^2 + \delta_i},$$

where for this iterative scheme δ_i has the form

$$\delta_i = l(\alpha_i,\beta_i) trace[Q^{-1}(\alpha_i,\beta_i)\mathbf{C}^t\mathbf{C}] - (1 - l(\alpha_i,\beta_i)) \| \mathbf{C}\mathbf{f}^i \|^2 . \tag{36}$$

This iteration has the same form as the iterative scheme described in section V when we use a constant step

$$\epsilon = \rho\beta_{i+1}. \tag{37}$$

We shall use $\delta_{i(ST,0,1,0,1)}$ to denote the δ_i defined in Eq. (36).

The theory which was used to assure convergence of the algorithm described in section V can be applied to iteration (35). However, there is an important difference; there is no guarantee that the values of δ_i in Eq. (36) satisfy the convergence constraints defined in [20], at least during the first iterations. Another important difference is that the estimate of the noise variance we obtain from this iterative scheme is

$$\beta_{i+1}^{-1} = \{\| \mathbf{g} - \mathbf{D}\mathbf{f}^i \|^2 + trace[Q^{-1}(\alpha_i, \beta_i)\mathbf{D}^t\mathbf{D}]\}/p. \tag{38}$$

which is not the same as the one provided in [20].

B.2 A modified iteration to obtain the evidence solution

The fact that iteration (35), using Eqs. (33), (34) and(36) may not converge, leads us to the introduction of a new iteration in this section. Such an iteration is guaranteed to converge to a unique fixed point since it satisfies the contraction mapping theory.

Let us consider finding $\mathbf{f}_{(\alpha,\beta)}$ for a given α and β, in other words, solving

$$\alpha\mathbf{C}^t\mathbf{C}\mathbf{f} - \beta\mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}) = 0.$$

This equation can be rewritten as

$$(2\alpha + \beta)\mathbf{f} = 2\alpha\mathbf{W}\mathbf{f} + \beta[\mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}) + \mathbf{f}],$$

or

$$\mathbf{f} = \mu\mathbf{W}\mathbf{f} + (1 - \mu)[\mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}) + \mathbf{f}],$$

with $\mu = 2\alpha/(2\alpha + \beta)$, $\mathbf{C}^t\mathbf{C} = 2(\mathbf{I} - \mathbf{W})$, and $\mathbf{W} = \mathbf{I} - 0.5\mathbf{C}^t\mathbf{C}$.

This leads to a very simple iterative scheme to find $\mathbf{f}_{(\alpha,\beta)}$; starting at $\mathbf{f}^0$ define

$$\mathbf{f}^{i+1} = \mu\mathbf{W}\mathbf{f}^i + (1 - \mu)[\mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}^i) + \mathbf{f}^i],$$

which can be easily shown to be a contraction.

This iterative scheme can be used to find the evidence solution. Starting at $\mathbf{f}^0, \alpha_0, \beta_0$, calculate α_{i+1} and β_{i+1} using Eqs. (33) and (34) and then use

$$\begin{aligned}
\mathbf{f}^{i+1} &= \mu_{i+1}\mathbf{W}\mathbf{f}^i + (1 - \mu_{i+1})[\mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}^i) + \mathbf{f}^i] \\
&= \mathbf{f}^i - (1 - \mu_{i+1})[\frac{\mu_{i+1}}{1 - \mu_{i+1}}(I - \mathbf{W})\mathbf{f}^i - \mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}^i)] \\
&= \mathbf{f}^i - (1 - \mu_{i+1})[\frac{\alpha_{i+1}}{\beta_{i+1}}\mathbf{C}^t\mathbf{C}\mathbf{f}^i - \mathbf{D}^t(\mathbf{g} - \mathbf{D}\mathbf{f}^i)],
\end{aligned} \tag{39}$$

with $\mu_{i+1} = 2\alpha_{i+1}/(2\alpha_{i+1} + \beta_{i+1})$. We henceforth refer to iteration (39) as the contraction (CO) iteration.

B.3 Hyperpriors with gamma distribution

The results and iterative methods we have discussed have been developed for flat hyperpriors for α and β. The same results can be obtained for general gamma hyperpriors. In this case Eq. (31)

$$\lambda_{(i+1)(EM,a\_\omega)} = \frac{\| \mathbf{g} - \mathbf{D}\mathbf{f}_{(\alpha_{i(EM,a\_\omega)},\beta_{i(EM,a\_\omega)})} \|^2}{\| \mathbf{C}\mathbf{f}_{(\alpha_{i(EM,a\_\omega)},(\beta_{i(EM,a\_\omega)})} \|^2 + \delta_{i(EM,a\_\omega)}},$$

where

$$\begin{aligned} \delta_{i(EM,a\_\omega)} &= l(\alpha_i, \beta_i, a\_\omega)[trace[Q^{-1}(\alpha_i, \beta_i)\mathbf{C}^t\mathbf{C}] + (l_1 - 2)2a_1] \\ &\quad -(1 - l(\alpha_i, \beta_i, a\_\omega)) \\ &\quad \times \| \mathbf{C}\mathbf{f}_{(\alpha_{(i)EM,a\_\omega},\beta_{(i)EM,a\_\omega})} \|^2, \end{aligned}$$

and

$$\begin{aligned} l(\alpha_i, \beta_i, a\_\omega) &= (p + l_2 - 2)/(p - 1 + l_1 - 2) \\ &\quad -\beta_{i+1}\left\{trace[Q^{-1}(\alpha_i, \beta_i)\mathbf{D}^t\mathbf{D}]/(p - 1 + l_1 - 2) - (l_2 - 2)2a_2/(p - 1 + l_1 - 2)\right\}, \end{aligned}$$

where in these two last equations we have used, for simplicity, $\alpha_i = \alpha_{i(EM,a\_\omega)}$ and $\beta_i = \beta_{i(EM,a\_\omega)}$.

The same equations are valid for the ST (set theoretic) and CO (contraction) iterative methods replacing the corresponding MAP image $\mathbf{f}_{(\alpha,\beta)}$ we obtain in the iterative procedure by the image we obtain in the gradient method. It is clear that although the expressions for ST (Eq. (28)) and CO (Eq. (39)) are the same, the restorations we obtain are not the same because of the different steps used.

VII. Experimental Results

A number of experiments have been performed with the proposed algorithms, EM (Eqs. (35), (36) and (37)), ST (Eq. (28)), CO (Eq. (39)) and MAP (Eqs. (23) and (24) over the "Lena" image and a synthetic image obtained from a prior distribution. A synthetic

image generated from the prior was also use in order to see how well the methods estimate the prior and noise variance.

For both images, Lenna and the simulation, the blur was due to motion over 9 pixels, and the criteria $\| \mathbf{f}^{i+1} - \mathbf{f}^i \|^2 / \| \mathbf{f}^i \|^2 \leq 10^{-6}$ or $i = 200$, which ever was met first, were used for terminating the iteration. The performance of the restoration algorithms was evaluated by measuring the improvement in signal to noise ratio after i-iterations denoted by Δ^i_{SNR} and defined by

$$\Delta^i_{SNR} = 10 \log_{10} \frac{\| \mathbf{f} - \mathbf{g} \|^2}{\| \mathbf{f} - \mathbf{f}^i \|^2}.$$

Since we have four different iterative methods, we shall use $\Delta^i_{SNR}(EM)$, $\Delta^i_{SNR}(ST)$, $\Delta^i_{SNR}(CO)$ and $\Delta^i_{SNR}(MAP)$. Furthermore, we shall also use $\Delta^i_{SNR}(method, 2a_1, w_1, 2a_2, w_2)$ where $method = EM$, ST, CO or MAP and $2a_1$, w_1, $2a_2$, and w_2 are the corresponding parameters of the hyperpriors. We have used (see Eq. (37)) $\epsilon = \min_i(1 - \mu_i)$, where μ_i comes from the contraction method (see Eq. (39)).

First, we run the methods on the "Lena" image blurred and with additive Gaussian noise with variance $\beta^{-1} = 216$, $\beta^{-1} = 21.6$ and $\beta^{-1} = 2.16$, to obtain a SNR of 10dB, 20dB and 30dB, respectively. The selection of the starting image for the MAP algorithm was an additional problem; when starting with the observed image we noted that the algorithm converged to a very noisy image, $\beta^{-1} \approx 0$, we then followed the procedure described in [3] and started with the EM solution for the 10dB, 20dB and 30dB images, respectively. The values of Δ_{SNR} and the required number of iterations as well as the corresponding values of $\hat{\beta}^{-1}$ and $\hat{\alpha}^{-1}$ at convergence, for flat hyperpriors, are shown in table I. From this table it can be seen that the best results are produced by the EM iterative algorithm, followed by the one based on the set theoretical approach. However, a visual inspection of the three restored images shows that they are not very different. It is worth mentioning the speed of convergence of the EM iterative algorithm. The MAP algorithm always tends to a flat image, $\alpha^{-1} \approx 0$, for those initial parameters.

The estimate of α^{-1} provided by the EM method, the one that produced the best results, was always around 50 (see table I) so we decided to test our methods with two more values of $2a_1$, one greater and other less than 50 on the "Lena" image.

The results of the same experiments for $2a_1 = 10$, $w_1 = 0.5$, $w_2 = 1$ and $2a_2 = 0$ are

shown in table II. The use of this value for $2a_1$, which is inversely proportional for the mean of the hyperprior for α forces the restoration to be smoother than in the previous experiment (see table I). It can be observed that, in this case, the noise variance estimates are, as expected, greater than the corresponding to table I. The only exceptions are the ST method at 10dB, where the maximum number of iterations, 200, was used. In this case, the values of Δ_{SNR} are smaller than in table I except for the MAP method since we avoid the solution $\alpha^{-1} = 0$ by introducing the hyperprior on α.

The corresponding results for $2a_1 = 50, w_1 = 0.5, 2a_2 = 0, w_2 = 1$ are shown in table III. This test corresponds to roughly use the prior variance estimate provided by the EM method as one divided by the mean of the hyperprior for α. As expected, the results are very similar to the ones in table I except for the MAP method since we, again, avoid the maximum at $\alpha^{-1} = 0$.

The same experiments were performed for $2a_1 = 100, w_1 = 0.5, 2a_2 = 0, w_2 = 1$ and they are displayed in table IV. The EM, CO and MAP methods produce smaller values for the noise variance estimates. They also improve the SNR.

Using $\| \mathbf{g} - \mathbf{Df} \|^2 / p$ as the estimate for the noise variance, as proposed in [20] the noise variance estimates for SNR= 10, 20, 30dB are 163.2, 16.4 and 3.01, respectively. It can be seen that the noise variance estimates provided by the EM algorithm in table I, table III and table IV are closer to the true value.

Figures 1, 2 and 3 show the values of $\delta_{i(EM,0,1,0,1)}$, $\delta_{i(ST,0,1,0,1)}$ and $\delta_{i(CO,0,1,0,1)}$ for $\beta^{-1} = 216$, $\beta^{-1} = 21.6$ and $\beta^{-1} = 2.16$, respectively. As can be seen the behavior of these three plots is very similar to figure 1 in [20].

Figures 4, 5 and 6 depict the values of Δ_{SNR} for various hyperpriors including the flat hyperprior for α and β, for $\beta^{-1} = 216$, $\beta^{-1} = 21.6$ and $\beta^{-1} = 2.16$, respectively.

As expected the behavior of the methods $(EM, 0, 1, 0, 1)$ and $(EM, 50, w, 0, 1)$, $w = 0.1, \ldots, 0.9$, is very similar. This is because the mean of the gamma distribution used as hyperprior for α is approximately equal to one divided by the prior variance estimates obtained when using $(EM, 0, 1, 0, 1)$. Furthermore, the method $(EM, 10, w, 0, 1)$, $w = 0.1, \ldots, 0.9$, always produces a worse Δ_{SNR} than $(EM, 0, 1, 0, 1)$ while the method $(EM, 100, w, 0, 1)$, $w = 0.1, \ldots, 0.9$, sometimes increases Δ_{SNR}. This is reasonable since

the EM method for flat hyperprior has not been designed to maximize the Δ_{SNR}.

Finally, the original and noisy blurred images are shown in figure 7 and the corresponding best restorations in terms of Δ_{SNR}, according to figures 4, 5 and 6, are shown in figure 8.

In order to study how good the prior variance are estimated, we run the same experiments on a synthetic image obtained from a prior distribution based on a SAR model with prior variance 100 and mean 128. This image, shown in figure 9, was blurred and gaussian noise with variance $\beta^{-1} = 309.7$, $\beta^{-1} = 30.97$ and $\beta^{-1} = 3.097$ was added, to obtain a SNR of 10dB, 20dB and 30dB, respectively. The initial restoration for the MAP algorithm was the EM solution for the 10dB, 20dB and 30dB problems, respectively. The required number of iterations as well as the corresponding values of $\hat{\beta}^{-1}$ and $\hat{\alpha}^{-1}$ at convergence, for flat hyperpriors, are shown in table V. From this table it can be seen that the worst hyperparameter estimations are obtained by the MAP algorithm.

We also tried the methods on the synthetic image with different values of $2a_1$, one was the real variance, $2a_1 = 100, w_1 = 0.5, 2a_2 = 0, w_2 = 1$, one higher than the real variance, $2a_1 = 500, w_1 = 0.5, 2a_2 = 0, w_2 = 1$, and the other lower, $2a_1 = 10, w_1 = 0.5, 2a_2 = 0, w_2 = 1$. The MAP method was always outperformed by the other three methods except for the $2a_1 = 500, w_1 = 0.5, 2a_2 = 0, w_2 = 1$ case where using a very high value for α and combining it with the MAP tendency to $\alpha^{-1} = 0$ produces more sensible results.

VIII. Conclusions

In this paper we have proposed three iterative image restoration algorithms according to which a restored image and an estimate of the image and noise variances, or equivalently an estimate of the regularization parameter, are provided at each iteration step. For these methods no knowledge of the noise or prior variances is required, although this knowledge, if available, can be incorporated in the form of gamma hyperpriors in the iterative procedure. The iterative procedures proposed in this paper behave similarly to the algorithm described in [20]. The results in [20] can be used to guarantee the convergence of the step algorithm and also to test whether the stationary point of the MAP and contraction algorithms are indeed a local maximum.

The proposed algorithms allow the incorporation of vague knowledge about the noise and image variances. The algorithms can be used for solving other inverse problems such

as the motion estimation problem where the values of the local variance can be weighted with global estimates and estimates from previous frames can be used to calculate the corresponding variance estimates.

Appendix

I. A study of the stationary points and local maxima of $p(\alpha, \beta|\mathbf{g})$

Let us first start by studying Eqs. (11) and (12) in order to characterize the stationary points of $p(\alpha, \beta|\mathbf{g})$ when a flat hyperprior is used and examine whether they are local maxima.

Since both $\mathbf{C}$ and $\mathbf{D}$ are block circulant matrices, Eqs. (11) and (12) can be written in the discrete Fourier domain as,

$$\alpha \sum_{i=1}^{p} \frac{\beta^2 |c_i|^2 |d_i|^2 |G_i|^2}{(\alpha|c_i|^2 + \beta|d_i|^2)^2} = (p-1) - \alpha \sum_{i}^{p} \frac{|c_i|^2}{\alpha|c_i|^2 + \beta|d_i|^2} \quad (40)$$

$$\beta \sum_{i=1}^{p} \frac{\alpha^2 |c_i|^4 |G_i|^2}{(\alpha|c_i|^2 + \beta|d_i|^2)^2} = p - \beta \sum_{i}^{2} \frac{|d_i|^2}{\alpha|c_i|^2 + \beta|d_i|^2}, \quad (41)$$

where d_i and c_i are the eigenvalues of $\mathbf{D}$ and $\mathbf{C}$, respectively, and G_i is the ith components of the discrete Fourier transform (DFT) of $\mathbf{g}$.

We also note that

$$\alpha \sum_{i=1} i^p \frac{|c_i|^2}{\alpha|c_i|^2 + \beta|d_i|^2} + \beta \sum_{1=1} i^p \frac{|d_i|^2}{\alpha|c_i|^2 + \beta|d_i|^2} = p \quad , \quad F_i(\alpha, \beta) = \frac{\beta d_i^* G_i}{\alpha|c_i|^2 + \beta|d_i|^2}$$

and

$$\| \mathbf{g} - \mathbf{D}\mathbf{f}_{(\alpha,\beta)} \|^2 = \sum_{i=1}^{p} |G_i - \frac{\beta|d_i|^2}{\alpha|c_i|^2 + \beta|d_i|^2} G_i|^2, \quad (42)$$

where $F_i(\alpha, \beta)$ is the ith components of the DFT of $\mathbf{f}_{(\alpha,\beta)}$ and the superscript * denotes complex conjugate.

Following [24], we define

$$\gamma = \sum_{i=1}^{p} \frac{\beta|d_i|^2}{\alpha|c_i|^2 + \beta|d_i|^2},$$

which obtains values between 1 and p and observe that at the stationary points of $p(\alpha, \beta|\mathbf{g})$ we have

$$\alpha ||C\mathbf{f}_{(\alpha,\beta)}||^2 = \gamma - 1,$$

and

$$\beta \| \mathbf{g} - \mathbf{D}\mathbf{f}_{(\alpha,\beta)} \| = p - \gamma,$$

which implies that the total misfit at any stationary point satifies

$$\alpha||C\mathbf{f}_{(\alpha,\beta)}||^2 + \beta \parallel \mathbf{g} - \mathbf{D}\mathbf{f}_{(\alpha,\beta)} \parallel = p - 1.$$

Having characterized the stationary points of $p(\alpha, \beta|\mathbf{g})$, and by taking into account that it is not difficult to examine whether they are local maxima if we perform the calculations in the Fourier domain, let us now study the conditions for having a single maximum.

There is not much work reported on this problem; MacKay [25] conjectures about the unicity of the maximum and we will adapt his results to our problem.

If we accept the models for the image and degradation, then we believe there is a true value of $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$ for which

$$\mathbf{g} \,|\, \hat{\alpha}\, \hat{\beta} \sim \mathcal{N}(0, \mathcal{A}),$$

where $\mathcal{A} = \frac{1}{\alpha}\mathbf{D}[\mathbf{C}^t\mathbf{C}]^{-1}\mathbf{D}^t + \frac{1}{\beta}\mathbf{I}$. We now transform $\mathbf{g}$ to $\mathbf{z} = \mathcal{A}^{-1/2}\mathbf{g}$ and then expect, if the model is correct, that each z_i^2 is independently distributed like $\mathcal{X}^2$ with one degree of freedom.

Following Mackay [25], the observations g_i is grossly at variance with the model for given values of α and β at significance level τ, if z_i^2 is not in the interval $[e^{-\tau}, 1 + \tau]$. It is conjectured that if we find a value of $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$ which locally maximizes $p(\alpha, \beta|\mathbf{g})$, and with which all the observations are not grossly at variance, then there are no other maxima.

Let us illustrate the meaning of these concepts with an example and see why it is important that each observation is not grossly at variance with the model. Let us use the simple image model,

$$p(\mathbf{f}|\mathbf{f}, \alpha) \propto \alpha^p \exp[-\alpha \sum_{i=1}^{p} f_i^2/2],$$

and

$$p(\mathbf{g}|\beta) \propto \beta^p \exp[-\beta[\sum_{i=1}^{p-1}(g_i - f_i)^2 + (g_p - h_p f_p)^2]/2],$$

with $h_p > 0$ and β known. We want to estimate α. Then, Eq. (11) becomes

$$\alpha \left[\frac{\beta^2}{(\alpha+\beta)^2} \sum_{i=1}^{p-1} g_i^2 + \frac{h_p^2\beta^2}{(\alpha + h_p^2\beta)^2} g_p^2 + \frac{1}{\alpha+\beta}(p-1) + \frac{1}{\alpha + h_p^2\beta} \right] = p. \qquad (43)$$

We now note that if $\sum_{i=1}^{p-1} g_i^2$ is small we may have a solution for α for the above equation which is big, which suggests a small prior variance. Furthermore, if g_p^2 is big and h_p small, another possible solution is a small α which suggests a big prior variance. Let us illustrate this behavior with a simple numerical example.

Assume $\beta = 1, p = 10001$, $\sum_{i=1}^{p-1} g_i^2 = 10000$, $h_p = 0.1$, $g_p = 100000$. Figure 10 despicts

$$m(x) = p - x \left[\sum_{i=1}^{p-1} \frac{\beta^2}{(x+\beta)^2} g_i^2 + \frac{h_p^2 \beta^2}{(x + h_p^2 \beta)^2} g_p^2 + \frac{1}{x+\beta}(p-1) + \frac{1}{x + h_p^2 \beta} \right],$$

for several values of x. We note that in the range $[0, 0.2]$ there are two stationary points of $p(\alpha|\mathbf{g})$ (two solutions of $m(x) = 0$). The smallest one is a local maximum (see figure $10(a)$). Furthermore there is another stationary point, greater than 0.2, in the range shown in figure $10(b)$ which corresponds to another local maximum.

In order to have one maximum, it is therefore very important that the observations are not grossly at variance with the model.

As with flat priors for the hyperparameters, it is possible to study the characteristics of the stationary points of $p(\alpha, \beta|\mathbf{g})$ for gamma hyperpriors and, with the use of the DFT, it is also possible to see whether we have local maxima.

Let us now relate the solutions of Eqs. (17) and (18) to the solutions of Eqs. (11) and (12). We will assume that β is known and want to estimate α. Let α_{flat} be a solution of Eq. (11) and assume that $\alpha_{flat} > 1/2a_1$. Then, from Eq. (17) we have that $U(\alpha_{flat}) > 0$, where

$$U(\alpha) = \alpha \left(\| \mathbf{C}\mathbf{f}_{(\alpha,\beta)} \|^2 + trace[Q(\alpha,\beta)^{-1}\mathbf{C}^t\mathbf{C}] + (l_1 - 2)2a_1 \right) - (p - 1 + l_1 - 2).$$

Furthemore, if there is only one solution of Eq. (11), then $U(1/2a_1) \leq 0$, (note that $U(0) < 0$), and so there is a unique solution for Eq. (17) in the interval $[1/2a_1, \alpha_{flat}]$. A similar study can be carried out when $\alpha_{flat} < 1/2a_1$.

An interesting problem is to examine what happens when the solution of Eq. (11) is not unique. If $\alpha_{flat} > 1/2a_1$ and $U(1/2a_1) > 0$, then either α_{flat} is not a local maximum or there is a solution of Eq. (11) in $[1/2a_1, \alpha_{flat}]$ which is not a local maximum. It is important to note that in this case we will have a solution of Eq. (11) in $[0, 1/2a_1]$ which is now a local maximum, and that if we denote that maximum by $\hat{\alpha}_{flat}$ we will have a solution of Eq. (17) in $[\hat{\alpha}_{flat}, 1/2a_1]$.

References

[1] H.C. Andrews and B. R. Hunt, *Digital Image Restoration*, Prentice Hall, New York, 1977.

[2] G. Archer, *Some estimation and restoration techniques for statistical image analysis*, Unpublished Ph.D. thesis, University of Glasgow, 1994.

[3] G. Archer and D.M. Titterington, "On Some Bayesian/Regularization Methods for Image Restoration", *IEEE Trans. on Image Processing*, vol. 4, pp. 989-995, 1995.

[4] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, New York, Springer Verlag, Chapters 3 and 4, 1985.

[5] R.A. Boyles, "On the convergence of the EM algorithm",*J. Royal Statistical Society*, Ser. B, vol. 44, 1983.

[6] W.L. Buntine, "A Theory of Learning Classification Rules", Doctoral dissertation, University of Technology, Sydney, Australia, 1991.

[7] W.L. Buntine, "Theory Refinement on Bayesian Networks", in *Proc. of the Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 52-60, 1991.

[8] W.L. Buntine and A. Weigund, "Bayesian Back-propagation", *Complex Systems*, vol. 5, pp. 603-643, 1991.

[9] G.F. Cooper and E. Herkovsits, "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning*, vol. 9, pp. 309-347, 1992.

[10] G. Demoment, "Image Reconstruction and Restoration: Overview of Common Estimation Structure and Problems", *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. ASSP-37, no. 12, pp. 2024-2036, Dec. 1989.

[11] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data", *J.Royal Statist. Soc. B*, vol. 39, pp. 1-38, 1972.

[12] 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, pp. 322-336, 1992.

[13] S.F. Gull, "Developments in Maximum Entropy Data Analysis", in *J. Skilling, editor, Maximum Entropy and Bayesian Methods, Cambridge*, Kluwer, pp. 53-71, 1989.

[14] M. G. Kang, *Adaptive Iterative Image Restoration Algorithms*, Ph.D. thesis, Northwestern University, Department of EECS, June 1994.

[15] M.G. Kang and A.K. Katsaggelos, "Simultaneous Iterative Image Restoration and Evaluation of the Regularization Parameter", *IEEE Trans. on Signal Processing*, vol. 40, pp. 2320-2334, 1992.

[16] A. K. Katsaggelos, "Iterative Image Restoration Algorithms," *Optical Engineering*, vol. 28, no. 7, pp. 735-748, July 1989.

[17] A.K. Katsaggelos, editor, *Digital Image Restoration*, Springer Series in Information Sciences, vol. 23, Springer-Verlag, 1991.

[18] A. K. Katsaggelos, J. Biemond, R. M. Mersereau and R. W. Schafer, "An Iterative Method for Restoring Noisy Blurred Images", *Circuits Systems and Signal Processing*, vol. 3, no. 2, pp. 139-160, 1984.

[19] A. K. Katsaggelos, J. Biemond, R. W. Schafer, and R. M. Mersereau, "A Regularized Iterative Image Restoration Algorithm", *IEEE Trans. Signal Processing,* vol. 39, pp. 914-929, 1991.

[20] A.K. Katsaggelos and M.G. Kang, "Iterative Evaluation of the Regularization Parameter in Regularized Image Restoration", *Journal of Visual Communication and Image Representation*, vol. 3, pp. 446-455, 1992.

[21] R. Lagendijk and J. Biemond, *Iterative Identification and Restoration of Images*, Kluwer Academic Press, pp. 31-32 and 124-129, 1991.

[22] K. T. Lay and A. K. Katsaggelos, "Image Identification and Restoration Based on the Expectation-Maximization Algorithm", *Optical Engineering*, vol. 29, pp. 436-445, 1990.

[23] D.J.C. MacKay, "A Practical Bayesian Framework for Backprop Networks", *Neural Computation*, vol. 4, pp. 448-472, 1992.

[24] D.J.C. MacKay, "Bayesian Interpolation", *Neural Computation*, vol. 4, pp. 415-447, 1992.

[25] D.J.C. MacKay, "Hyperparameters: Optimize Or Integrate Out", *Submitted to Neural Computation*, 1995.

[26] K.V. Mardia, J.T. Kent and J.M. Bibly, *Multivariate Analysis*, Academic Press, pp. 41-43, 1979.

[27] R. Molina, "On the Hierarchical Bayesian Approach to Image Restoration. Application to Astronomical Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. **PAMI**-16, no. 11, pp. 1222-1228, 1994.

[28] R. Molina and A.K. Katsaggelos, "On the Hierarchical Bayesian Approach to Image Restoration and the Iterative Evaluation of the Regularization Parameter", in *Proc. of the Visual Communication and Image Processing'94, Chicago*, pp. 244-251, 1994.

[29] R. Molina and A.K. Katsaggelos, "Prior Models and Methods of Choosing the Unknown Parameters in the Bayesian and Regularization Approaches to Image Restoration", in *Proc. of the IEEE Workshop on Nonlinear Signal and Image Processing*, Ed. I. Pitas, pp. 102-105, 1995.

[30] R. Molina, A. del Olmo, J. Perea and B.D. Ripley, "Bayesian Deconvolution in Optical Astronomy", *Astrom. J.* , vol. 103, pp. 666-675, 1992.

[31] R. Molina and B.D. Ripley, "Using Spatial Models as Priors in Astronomical Images Analysis", *J. Appl. Statist.*, vol. 16, pp. 193-206, 1989.

[32] R. Molina, B.D. Ripley, A. Molina , F. Moreno and J.L. Ortiz, "Bayesian Deconvolution with Prior Knowledge of Object Location. Applications to Ground-based Planetary Images", *Astrom. J.*, vol. 104, pp. 1662-1668, 1992.

[33] R.M. Neal, *Bayesian Learning for Neural Networks*, Ph.D. thesis, University of Toronto, Department of Computer Science, 1995.

[34] H. Raiffa and R. Schlaifer , *Applied Statistical Decision Theory*, Division of Research, Graduate School of Business, Administration, Harvard University, Boston, 1961.

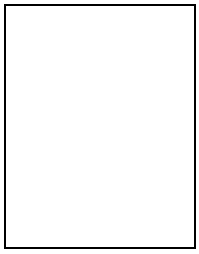
[35] B.D. Ripley, *Spatial Statistics*, John Wiley, New York, pp. 88-90, 1981.

[36] D. J. Spiegelhalter and S.L. Lauritzen, "Sequential Updating of Conditional Probabilities on Directed Graphical Structures", *Networks*, vol. 20, pp. 579-605, 1990.

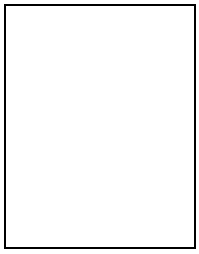
[37] C.E.M. Strauss, D.H. Wolpert and D.R. Wolf, "Alpha, Evidence and the Entropic Prior", in *A. Mohammed-Djafari editor, Maximum Entropy and Bayesian Methods, Paris*, Kluwer, pp. 53-71, 1992.

[38] D.H. Wolpert, "On the Use of Evidence in Neural Networks", in *C.L. Giles, S.J. Hanson and J.D. Cowan editors, Advances in Neural Information Processing Systems 5, San Mateo, California*, Morgan Kaufmann, pp. 539-546, 1993.

[39] C.F.J. Wu "On the convergence properties of the EM algorithm", *The Annals of Statistics*, no. 1, pp. 95-103, 1983.

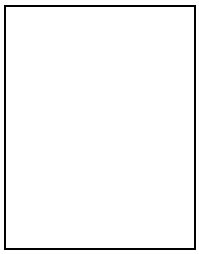


Rafael Molina was born in 1957 and received his degree in Mathematics (Statistics) in 1979. He completed his Ph. D. Thesis in 1983 in Optimal Design in Linear Models and became Associate Professor in Computer Science and Artificial Intelligence at the University of Granada in 1989. His areas of research interest are image restoration (applications to astronomy and medicine), parameter estimation, image compression, and blind deconvolution. He is a member of the IEEE, SPIE, Royal Statistical Society and AERFAI (Asociación Española de Reconocimento de Formas y Análisis de Imágenes) and currently the Dean of the Computer Engineering Faculty at the University of Granada.



Aggelos K. Katsaggelos received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1979 and the M.S. and Ph.D. degrees both in electrical engineering from the Georgia Institute of Technology, Atlanta, Georgia, in 1981 and 1985, respectively.

In 1985 he joined the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL, where he is currently professor, holding the Ameritech Chair of Information Technology. During the 1986-1987 academic year he was an assistant professor at Polytechnic University, Department of Electrical Engineering and Computer Science, Brooklyn, NY. His current research interests include image and video recovery, video compression, motion estimation, boundary encoding, computational vision, and multimedia signal processing. Dr. Katsaggelos is a Fellow of the IEEE, an Ameritech Fellow, a member of the Associate Staff, Department of Medicine, at Evanston Hospital, and a member of SPIE. He is a member of the Steering Committee of the *IEEE Transactions on Medical Imaging*, the IEEE Technical Committees on Visual Signal Processing and Communications, Image and Multi-Dimensional Signal Processing, and Multimedia Signal Processing, and the editor-in-chief of the *IEEE Signal Processing Magazine*. He has served as an Associate editor for the *IEEE Transcations on Signal Processing* (1990-1992), an area editor for the journal *Graphical Models and Image Processing* (1992-1995), a member of the Steering Committee of the *IEEE Transactions on Image Processing* (1992-1997). He is the editor of *Digital Image Restoration* (Springer-Verlag, Heidelberg, 1991), co-author of *Rate-Distortion Based Video Compression* (Kluwer Academic Publishers, 1997), and co-editor of *Recovery Techniques for Image and Video Compression,* (Kluwer Academic Publishers, 1998). He has served as the General Chairman of the 1994 Visual Communications and Image Processing Conference (Chicago, IL), and as technical program co-chair of the 1998 IEEE International Conference on Image Processing (Chicago, IL).



Javier Mateos was born in Granada, Spain, in 1968. He received the Diploma and M. S. degrees in Computer Science from the University of Granada in 1990 and 1991, respectively. He completed his Ph. D. in Computer Science at the University of Granada in July 1998.

Since September 1992, he has been Assistant Professor at the Department of Computer Science and Artificial Inteligence of the University of Granada. His research interest include image restoration and image and video recovery and compression. He is a member of the AERFAI (Asociación Española de Reconocimento de Formas y Análisis de Imágenes).

List of captions

TABLE I: Iterations required, noise and prior variance estimations and SNR improvement for flat hyperpriors for the "Lena" image.

TABLE II: Iterations required, noise variance estimation and SNR improvement for $2a_1 = 10, w_1 = 0.5, 2a_2 = 0, w_2 = 1$ hyperpriors for the "Lena" image.

TABLE III: Iterations required, noise variance estimation and SNR improvement for $2a_1 = 50, w_1 = 0.5, 2a_2 = 0, w_2 = 1$ hyperpriors for the "Lena" image.

TABLE IV: Iterations required, noise variance estimation and SNR improvement for $2a_1 = 100, w_1 = 0.5, 2a_2 = 0, w_2 = 1$ hyperpriors for the "Lena" image.

TABLE V: Iterations required, noise and prior variance estimation for flat hyperpriors for the "prior" image.

Fig. 1: Comparison of $\delta_{i(EM,0,1,0,1)}$, $\delta_{i(ST,0,1,0,1)}$ and $\delta_{i(CO,0,1,0,1)}$ for $\beta^{-1} = 216$.

Fig. 2: Comparison of $\delta_{i(EM,0,1,0,1)}$, $\delta_{i(ST,0,1,0,1)}$ and $\delta_{i(CO,0,1,0,1)}$ for $\beta^{-1} = 21.6$.

Fig. 3: Comparison of $\delta_{i(EM,0,1,0,1)}$, $\delta_{i(ST,0,1,0,1)}$ and $\delta_{i(CO,0,1,0,1)}$ for $\beta^{-1} = 2.16$.

Fig. 4: $\Delta^i_{SNR}(EM, 0, 1, 0, 1)$, $\Delta^i_{SNR}(EM, 10, w, 0, 1)$, $\Delta^i_{SNR}(EM, 50, w, 0, 1)$ and $\Delta^i_{SNR}(EM, 100, w, 0, 1)$ for $w = 0.1, 0.2, \ldots, 0.9$ for $\beta^{-1} = 216$.

Fig. 5: $\Delta^i_{SNR}(EM, 0, 1, 0, 1)$, $\Delta^i_{SNR}(EM, 10, w, 0, 1)$, $\Delta^i_{SNR}(EM, 50, w, 0, 1)$ and $\Delta^i_{SNR}(EM, 100, w, 0, 1)$ for $w = 0.1, 0.2, \ldots, 0.9$ for $\beta^{-1} = 21.6$.

Fig. 6: $\Delta^i_{SNR}(EM, 0, 1, 0, 1)$, $\Delta^i_{SNR}(EM, 10, w, 0, 1)$, $\Delta^i_{SNR}(EM, 50, w, 0, 1)$ and $\Delta^i_{SNR}(EM, 100, w, 0, 1)$ for $w = 0.1, 0.2, \ldots, 0.9$ for $\beta^{-1} = 2.16$.

Fig. 7: (a) Original image. Noisy-blurred images for 9-point motion blur (b) $\beta^{-1} = 216$, (c) $\beta^{-1} = 21.6$ and (d) $\beta^{-1} = 2.16$.

Fig. 8: Restorations. (a) $\beta^{-1} = 216$, $\Delta^{i=19}_{SNR}(EM, 100, 0.9, 0, 1) = 7.038$ (b) $\beta^{-1} = 21.6$, $\Delta^{i=6}_{SNR}(EM, 100, 0.4, 0, 1) = 8.547$ and (c) $\beta^{-1} = 2.16$, $\Delta^{i=10}_{SNR}(EM, 100, 0.1, 0, 1) = 11.494$.

Fig. 9: Synthetic image from a SAR prior distribution with variance 100.

Fig. 10: Finding the stationary points of $p(\alpha|\mathbf{g})$.

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
10dB	216.1	37	EM	213.97	54.85	7.02
		52	ST	218.62	44.22	6.52
		32	CO	214.87	65.70	6.61
		12	MAP	2347.74	0.0	-3.52

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
20dB	21.61	5	EM	22.62	48.96	8.36
		36	ST	25.93	28.84	7.02
		60	CO	27.40	20.51	6.74
		20	MAP	1226.5	1.8×10^{-4}	-3.23

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
30dB	2.161	16	EM	2.54	47.72	11.10
		105	ST	3.37	75.72	9.21
		110	CO	4.59	25.19	8.37
		78	MAP	2130.0	0.0	-4.38

TABLE I

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
10dB	216.1	11	EM	226.25	10.57	6.43
		200	ST	216.01	43.37	6.75
		78	CO	295.64	10.10	4.39
		3	MAP	224.49	5.30	6.08

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
20dB	21.61	4	EM	25.99	13.50	7.38
		46	ST	31.06	10.67	6.22
		114	CO	32.29	10.53	6.08
		13	MAP	25.85	5.84	6.80

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
30dB	2.161	10	EM	3.66	14.00	9.50
		37	ST	6.36	11.80	7.39
		122	CO	6.21	11.64	7.47
		7	MAP	3.32	7.87	8.97

TABLE II

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
10dB	216.1	8	EM	214.57	50.00	7.00
		52	ST	223.31	48.36	6.30
		59	CO	223.75	48.34	6.28
		3	MAP	203.51	25.93	6.84

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
20dB	21.61	6	EM	22.40	48.68	8.36
		33	ST	24.34	47.36	7.44
		54	CO	24.34	47.31	7.46
		10	MAP	19.33	27.88	8.12

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
30dB	2.161	9	EM	2.53	50.17	11.14
		37	ST	4.09	45.36	8.78
		58	CO	3.98	45.55	8.87
		2	MAP	1.57	34.50	11.21

TABLE III

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
10dB	216.1	7	EM	210.45	97.37	7.02
		54	ST	220.59	94.44	6.34
		49	CO	216.11	94.72	6.62
		2	MAP	194.94	51.73	7.02

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
20dB	21.61	6	EM	21.39	91.94	8.55
		48	ST	28.98	87.74	6.63
		46	CO	22.99	89.97	7.85
		9	MAP	17.13	55.28	8.50

SNR	β^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$	Δ_{SNR}
30dB	2.161	10	EM	2.19	90.19	11.48
		40	ST	5.68	81.50	8.02
		50	CO	3.67	82.80	9.25
		7	MAP	0.97	71.65	11.33

TABLE IV

SNR	β^{-1}	α^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$
10dB	309.7	100	32	EM	303.71	59.18
			51	ST	297.07	64.40
			28	CO	295.00	90.09
			12	MAP	427.16	4.3×10^{-5}

SNR	β^{-1}	α^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$
20dB	30.97	100	3	EM	29.68	55.13
			19	ST	31.16	36.76
			11	CO	29.82	26.17
			6	MAP	32.50	7.5×10^{-5}

SNR	β^{-1}	α^{-1}	iterations	method	$\hat{\beta}^{-1}$	$\hat{\alpha}^{-1}$
30dB	3.097	100	4	EM	2.97	5.23
			14	ST	4.91	2.58
			11	CO	3.23	2.96
			6	MAP	3.12	8.3×10^{-5}

TABLE V

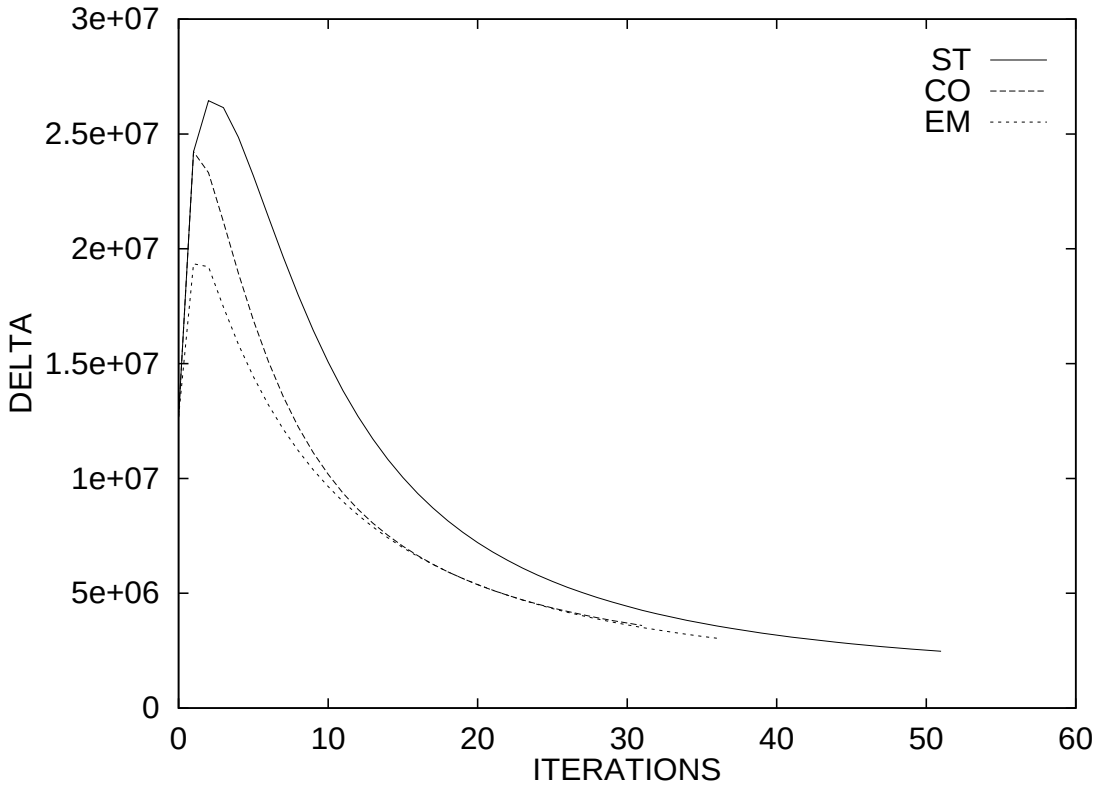

3e+07
2.5e+07
2e+07
1.5e+07
1e+07
5e+06
0
DELTA
0
10
20
30
40
50
60
ITERATIONS
ST
CO
EM

Fig. 1.

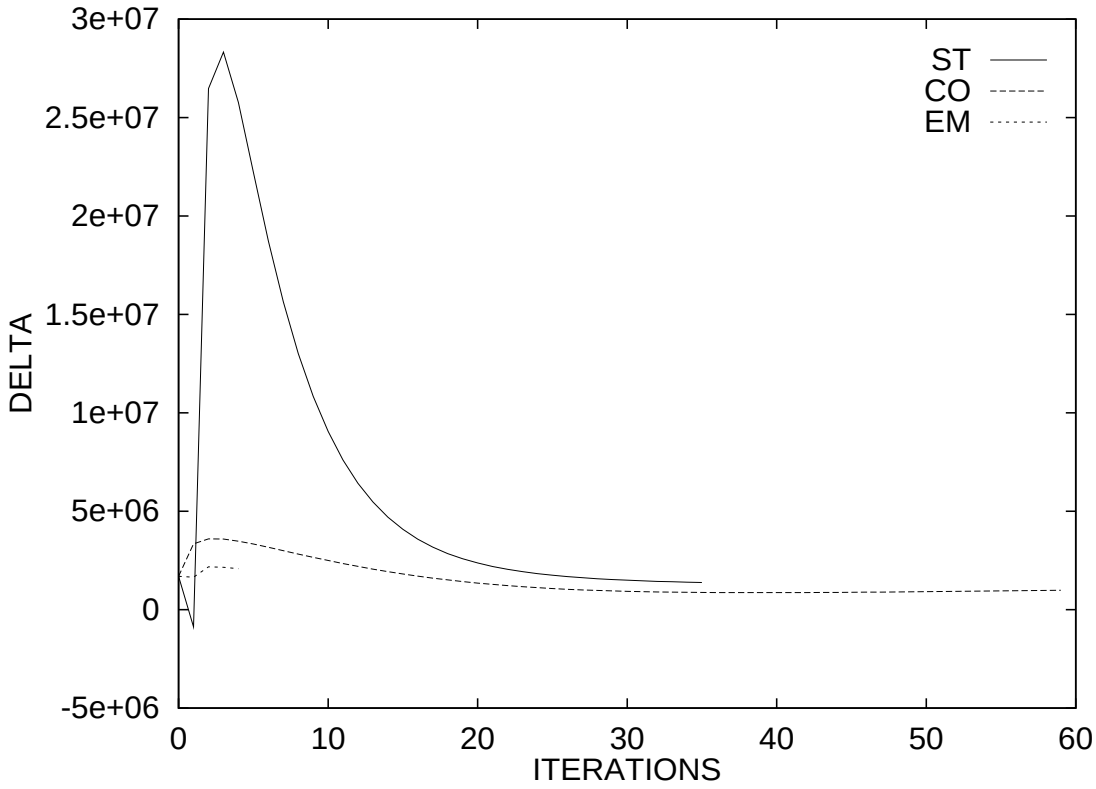

3e+07
2.5e+07
2e+07
1.5e+07
1e+07
5e+06
0
-5e+06
DELTA
0
10
20
30
40
50
60
ITERATIONS
ST
CO
EM

Fig. 2.

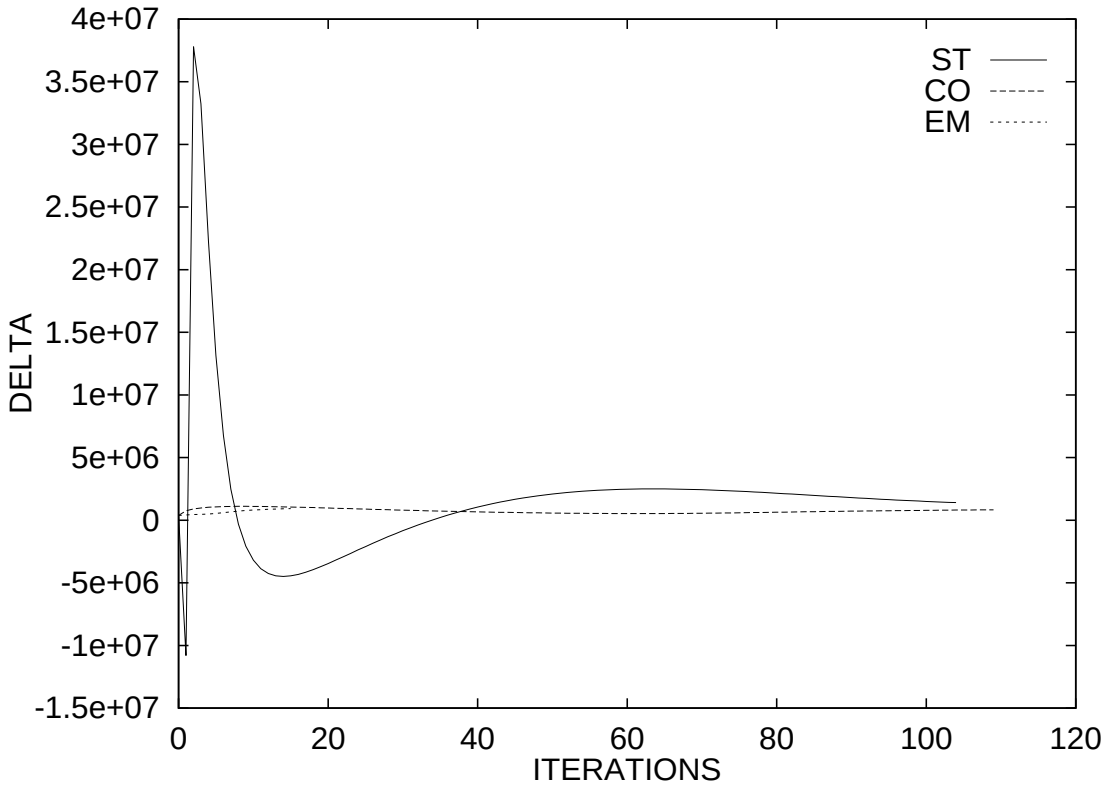

4e+07
3.5e+07
3e+07
2.5e+07
2e+07
1.5e+07
1e+07
5e+06
0
-5e+06
-1e+07
-1.5e+07
DELTA
0
20
40
60
80
100
120
ITERATIONS
ST
CO
EM

Fig. 3.

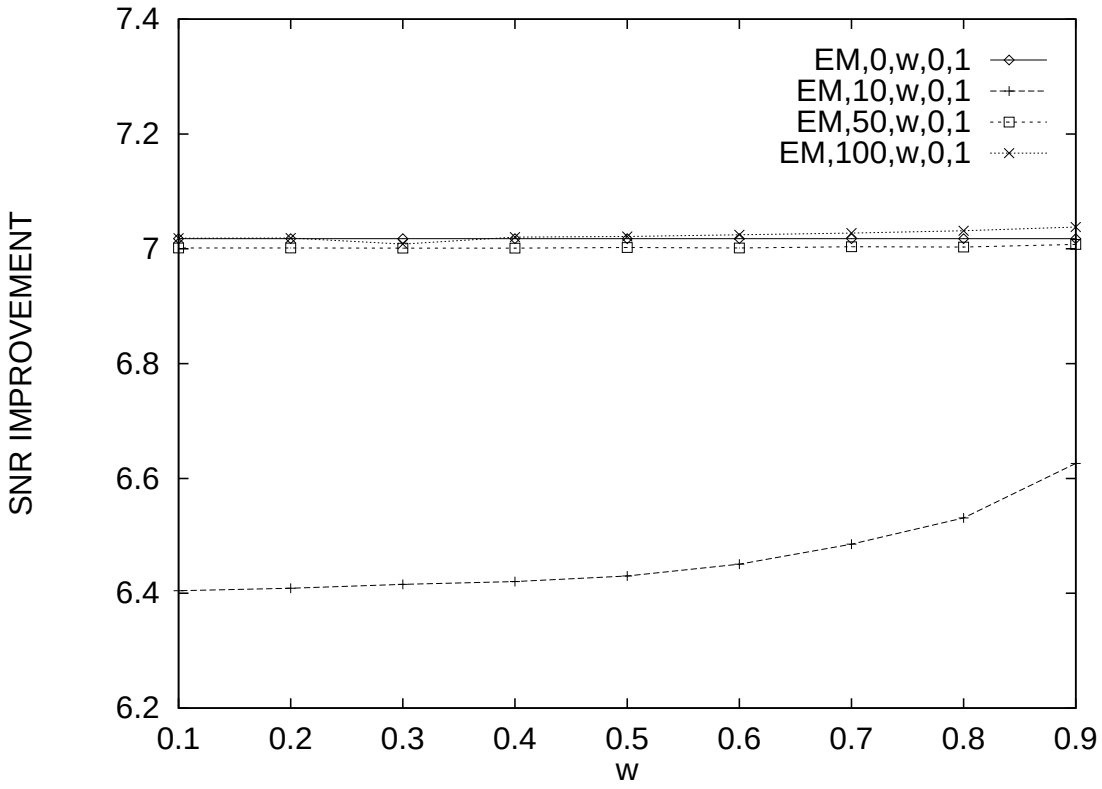

7.4
7.2
7
6.8
6.6
6.4
6.2
SNR IMPROVEMENT
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
w
EM,0,w,0,1
EM,10,w,0,1
EM,50,w,0,1
EM,100,w,0,1

Fig. 4.

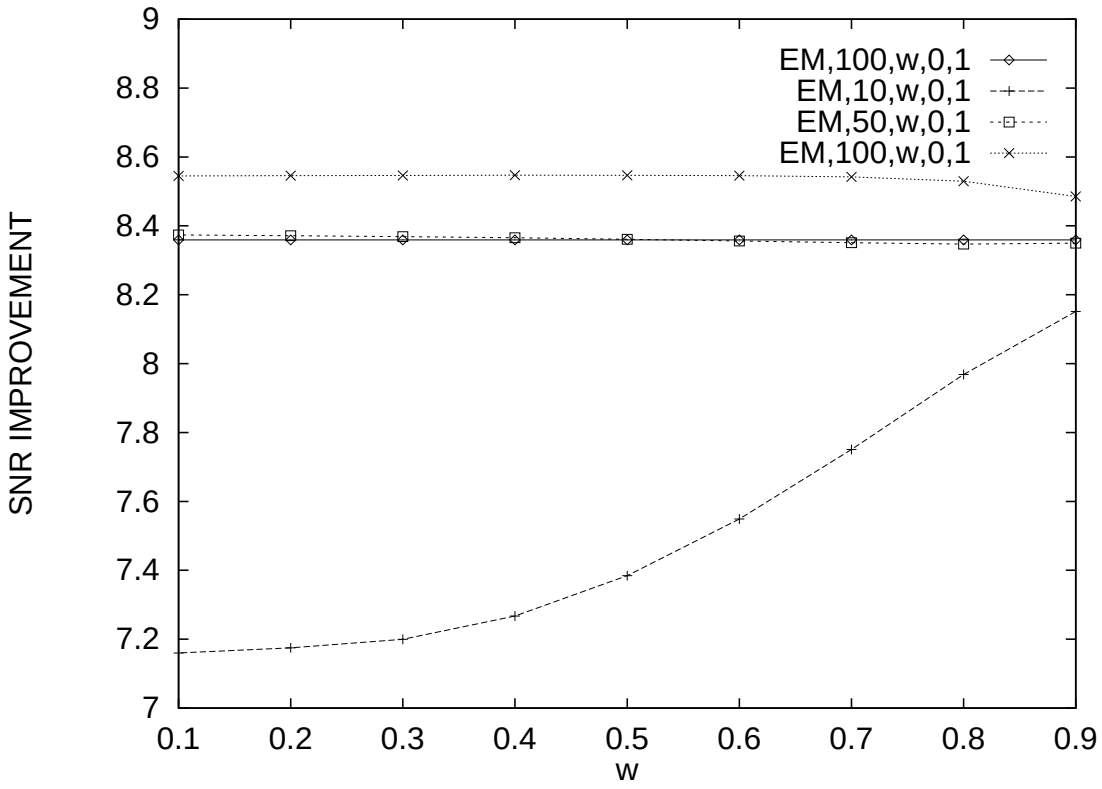

SNR IMPROVEMENT
w
EM,100,w,0,1
EM,10,w,0,1
EM,50,w,0,1
EM,100,w,0,1
9
8.8
8.6
8.4
8.2
8
7.8
7.6
7.4
7.2
7
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Fig. 5.

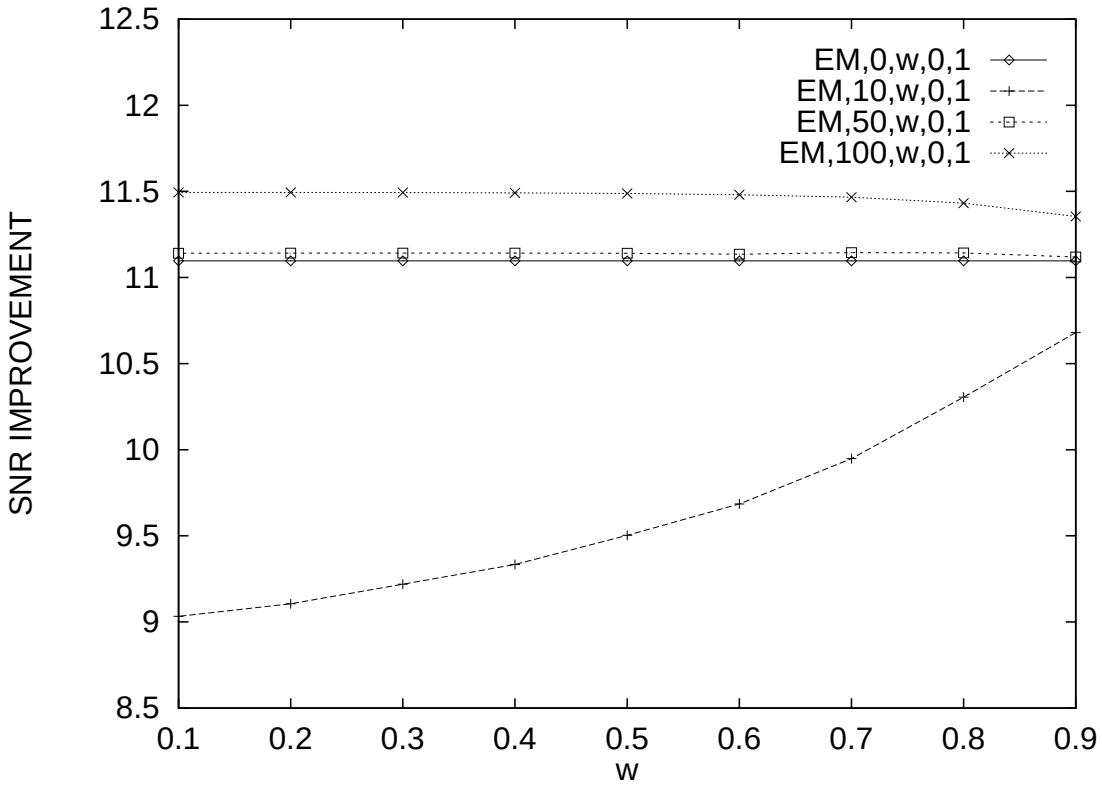

12.5
12
11.5
11
10.5
10
9.5
9
8.5
SNR IMPROVEMENT
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
w
EM,0,w,0,1
EM,10,w,0,1
EM,50,w,0,1
EM,100,w,0,1

Fig. 6.

(a) (b)

(c) (d)

Fig. 7.

(a)

(b)

(c)

Fig. 8.

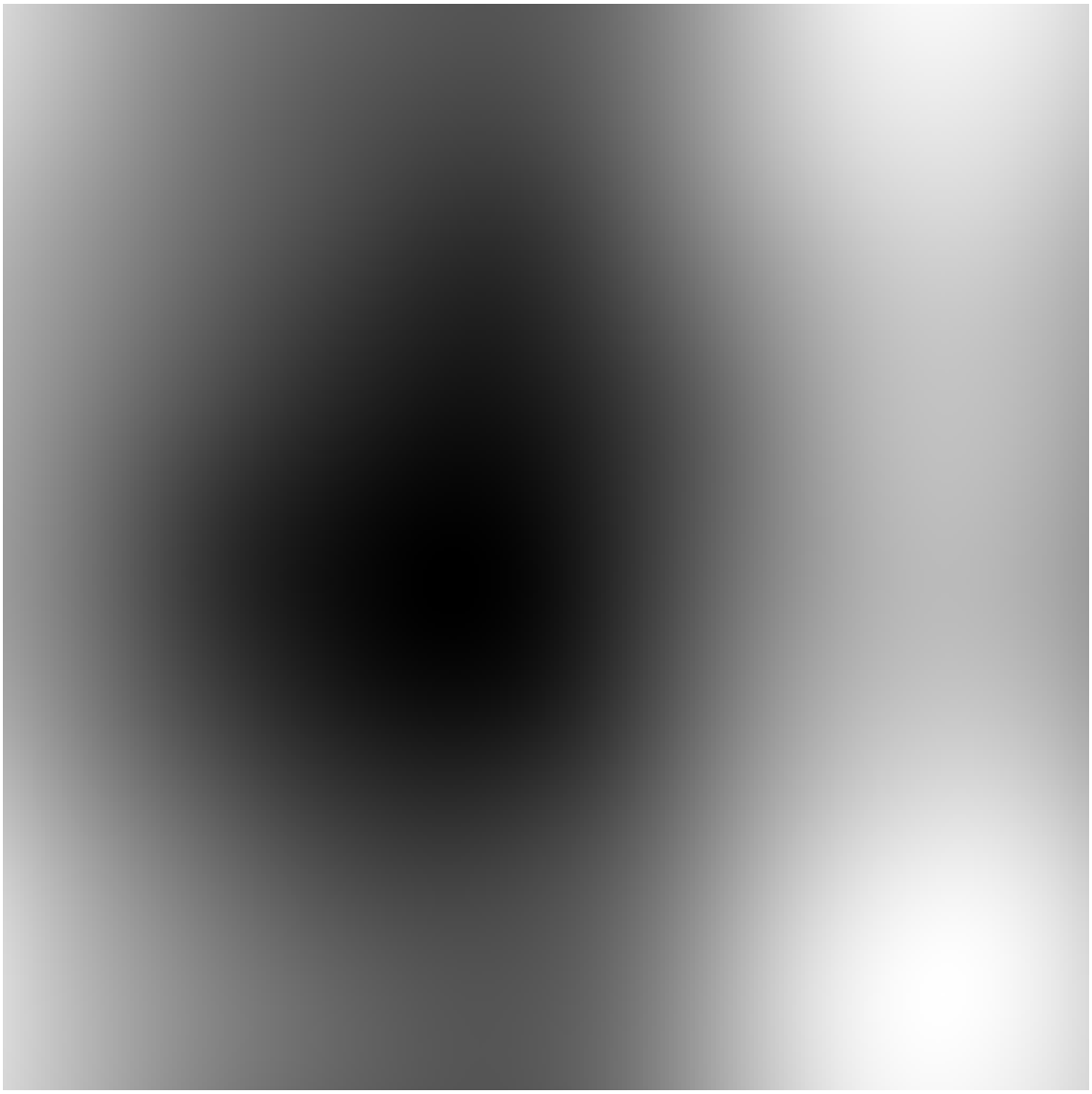
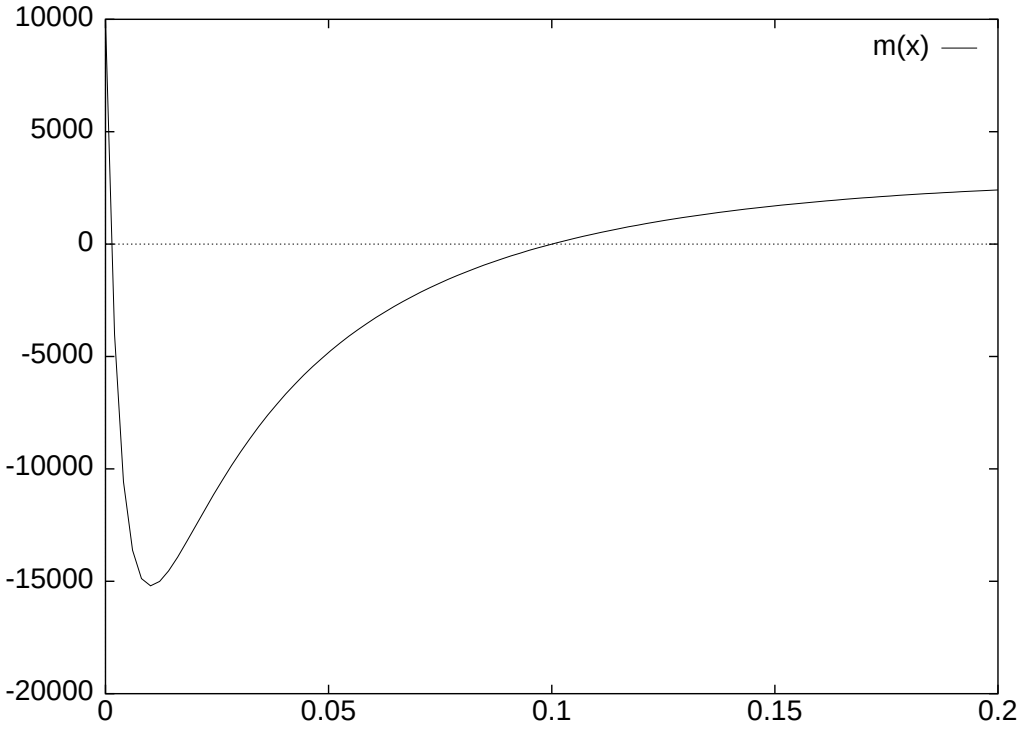
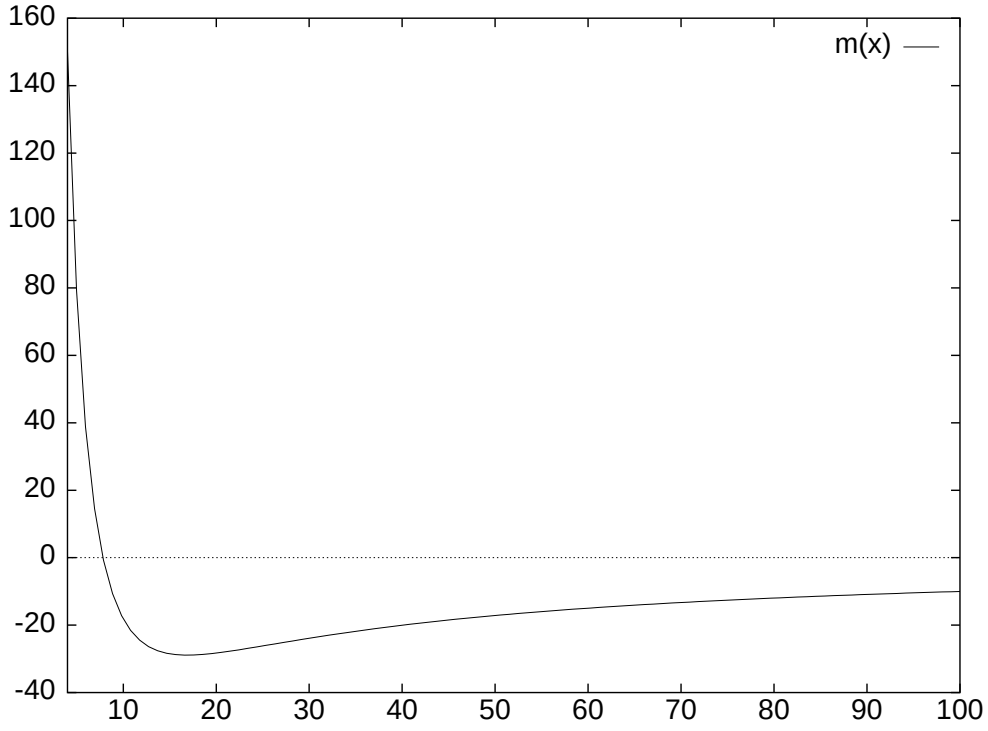


Fig. 9.


m(x)
10000
5000
0
-5000
-10000
-15000
-20000
0
0.05
0.1
0.15
0.2

(a)


m(x)
160
140
120
100
80
60
40
20
0
-20
-40
10
20
30
40
50
60
70
80
90
100

(b)

Fig. 10.