M5.21

ESTIMATING PRIORS IN MAXIMUM ENTROPY IMAGE PROCESSING

A. Mohammad-Djafari and G. Demoment

Laboratoire des Signaux et Systèmes (CNRS-ESE-UPS)
Plateau de Moulon, 91192 Gif-sur-Yvette Cédex, France.

ABSTRACT

In this paper we first propose a brief description of the Maximum *a posteriori* (MAP) Bayesian approach with Maximum Entropy (ME) priors to solve the linear system of equations which is obtained after the discretization of the integral equations which arises in various tomographic image restoration and reconstruction problems. We discuss then about the main problem which is to choose an *a priori* probability law for the image and to determine their parameters from the data. We propose then a method to estimate simultaneously the parameters of the ME *a priori* probability density function (pdf) and the pixel values of the image and give some simulated results which compare this method with some classical ones.

I. INTRODUCTION

We adress a class of discrete image reconstruction and restoration problems which can be described by the following problem: Estimate a positive vector $\mathbf{x}$ (representing the pixel intensities in an object) given a vector of measurements $\mathbf{y}$ (representing either a degraded image or the projections of the object) and a linear transformation $\mathbf{A}$ relating them by :

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{1}$$

where $\mathbf{b}$ represents the noise measurement which is supposed to be zero-mean and additive. Let us assume that we have only an approximate information about the noise variance σ^2 and some global information about the object.

We use the Bayesian approach and a *Maximum a posteriori* (MAP) estimation technique to solve this problem. Our estimator $\hat{\mathbf{x}}$ is the argument which maximizes the *a posteriori* distribution $p(\mathbf{x}|\mathbf{y})$ which is obtained by the Bayes' formula:

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}) \,/p(\mathbf{y}) \tag{2}$$

In this equation, $p(\mathbf{y})$ is independent of $\mathbf{x}$, $p(\mathbf{y}|\mathbf{x})$ is, in fact, related to the noise probability law, and $p(\mathbf{x})$ is an *a priori* law on $\mathbf{x}$.

We are not given directly $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$, and the main problem is how to determine them. To do this, we use the *Maximum Entropy* (ME) principle. The idea is that, if we have not enough information about a random process to assign it a probability law, we can choose the ME law which satisfies our *a priori* information.

The ME principle can be used if this knowledge can be stated as some constraints on $p(\mathbf{x})$. In general these only constraints are not sufficient to determine uniquely $p(\mathbf{x})$. Then, between all probability laws which satisfy these constraints, we choose the one which has maximum entropy [1-4].

Mathematically this leads to: given the constraints:

$$\mathrm{E}\{g_i(\mathbf{x})\} = \int g_i(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} = d_i \qquad i = 1, \ldots M \tag{3}$$

where $g_i(\mathbf{x})$ are known functions, determine $p(\mathbf{x})$ which maximizes the entropy:

$$H = -\int p(\mathbf{x}) \,\mathrm{Ln}\, p(\mathbf{x})\, d\mathbf{x} \tag{4}$$

The solution is classically given by:

$$p(\mathbf{x}) = \frac{1}{Z} \exp\Big[\sum_{i=1}^{M} \lambda_i\, g_i(\mathbf{x})\Big] \tag{5}$$

where Z is the partition function which is given by the normalization constraint:

$$Z(\lambda_1, \lambda_2, \ldots, \lambda_M) = \int \exp\Big[\sum_{i=1}^{M} \lambda_i\, g_i(\mathbf{x})\Big]\, d\mathbf{x} \tag{6}$$

and the Lagrange multipliers λ_i, $i = 1, \ldots, M$ are determined by the constraints (3) by solving the following system of equations:

$$\partial\, Z(\lambda_1, \lambda_2, \ldots, \lambda_M) \,/\, \partial\lambda_i = d_i \qquad i = 1, 2, \ldots, M \tag{7}$$

Now, if we can assign $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$, then the problem is solved by finding an algorithm which determines $\hat{\mathbf{x}}$ by:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}>\mathbf{0}}{\mathrm{Arg\,max}}\, p(\mathbf{x}|\mathbf{y}) = \underset{\mathbf{x}>\mathbf{0}}{\mathrm{Arg\,max}}\, \{p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x})\} \tag{8}$$

If we know only the variance σ^2 of the noise, then the ME principle will give us :

$$p(\mathbf{y}|\mathbf{x}) = \exp[-Q(\mathbf{x})] \quad \text{with} \quad Q(\mathbf{x}) = [\mathbf{y}-\mathbf{A}\mathbf{x}]^t[\mathbf{y}-\mathbf{A}\mathbf{x}] \,/\sigma^2 \tag{9}$$

In this paper we discuss first in detail how to choose the *a priori* $p(\mathbf{x})$. We will show that, with some global constraints on the image $\mathbf{x}$, $p(\mathbf{x})$ is in the form :

$$p(\mathbf{x}) = \frac{1}{Z} \exp\Big[-\lambda \sum_{i=1}^{n} H(x_i) - \mu \sum_{i=1}^{n} S(x_i)\Big] \tag{10}$$

where $S(x)=x$ and $H(x)$ can be either $-\mathrm{Ln}x$, $x\mathrm{Ln}x$ or x^h. The estimation problem (8) is then equivalent to :

$$\hat{\mathbf{x}} = \underset{\mathbf{x}>\mathbf{0}}{\mathrm{Arg\,min}}\, \{\, Q(\mathbf{x}) + \lambda\, \mathrm{H}(\mathbf{x}) + \mu\, \mathrm{S}(\mathbf{x})\} \tag{11}$$

which can also be considered as the solution of a regularization problem in which (λ, μ) are the regularization parameters (hyper-parameters).

Two main difficulties in real applications are :

i) When the hyper-parameters (λ, μ) are given how to solve (11) ? This can be achieved only by an iterative method.

ii) How to determine the hyper-parameter (λ, μ) values from the available data $\mathbf{y}$?

These two problems are the object of many authors' researches today [1, 2]. To solve the first one we used a conjugate gradient technique. The principal properties of this technique are now well established. The algorithmic details of our method is given in [3, 4]. The main contribution of this paper is that we present a joint method to estimate iteratively the hyper-parameters and the pixel values of the object.

CH2847-2/90/0000-2069 $1.00 © 1990 IEEE

The organisation of this paper is the following: in section II we give arguments to choose the form of the prior law $p(\mathbf{x})$. In section III, we give some relations to estimate the hyper-parameters of this law. In section IV we present a summary of the method and, finally, in section V we give some simulation results.

II. Determining the Form of $P(\mathbf{x})$

$p(\mathbf{x})$ is an *a priori* law on $\mathbf{x}$. It must be as general and noninformative as possible, *i.e.* it must only reflect our *a priori* knowledge about $\mathbf{x}$. In image reconstruction and restoration problems, we know for example that $x_i>0$, and may have some global *a priori* knowledge about the mean total intensity of the object we want to restore. The main problem is how to determine an *a priori* law to reflect this information. In the following we give two different viewpoints which give the same results.

II.1. A Statistical Viewpoint

We want to determine $p(\mathbf{x})$ from a finite set of statistical observations on $\mathbf{x}$. So we limit ourselves to a parametric representation of $p(\mathbf{x})$ with a few number of parameters.

Our main hypothesis is that we cannot have any *a priori* information about the correlations in $\mathbf{x}$. So we must not use any *a priori* information about the correlations to estimate the parameters of $p(\mathbf{x})$. The estimation is done from some finite scalar observation functionals $\phi_k(\mathbf{x})$, k=1, 2,..., m on the image. This hypothesis limits $\phi_k(\mathbf{x})$ to be in the form [5]:

$$\phi_k(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i) \quad k=1,\ 2,\ldots,\ m \tag{12}$$

Our next hypothesis is that we cannot *a priori* distinguish any region in the object which must be found. This means that the pixels are interchangeable so that $p(\mathbf{x})$ must be symmetric in x_i. This limits us to choose $f_i = f \ \ \forall i$. So we have:

$$\phi_k(\mathbf{x}) = \sum_{i=1}^{n} f(x_i) \quad k=1,\ 2,\ldots,\ m \tag{13}$$

Using the Lagrangian multiplier technique, given the m constraints $\phi_k(\mathbf{x})$, k=1, 2,..., m on the image $\mathbf{x}$, we find:

$$p(\mathbf{x}) = \exp\Big[\sum_{k=1}^{m}\lambda_k \sum_{i=1}^{n} f_i(x_i)\Big] = \prod_{i=1}^{n}\exp\Big[\sum_{k=1}^{m}\lambda_k f_i(x_i)\Big] = \prod_{i=1}^{n} p(x_i) \tag{14}$$

Now if we limit ourselves to a solution with two parameters for $p(\mathbf{x})$ and choose two scalar observation functions:

$$\phi_1(\mathbf{x}) = S(\mathbf{x}) = \sum S(x_i) \tag{15}$$

$$\phi_2(\mathbf{x}) = H(\mathbf{x}) = \sum H(x_i) \tag{16}$$

then we have:

$$p(x_i) = \exp[\lambda\, H(x_i) + \mu\, S(x_i)] \tag{17}$$

or, equivalently:

$$p(\mathbf{x}) = \frac{1}{Z}\ \exp[\lambda\, H(\mathbf{x}) + \mu\, S(\mathbf{x})] \tag{18}$$

II.2. The Maximum Entropy viewpoint

In this case we use the ME principle to directly determine the form of $p(\mathbf{x})$. We suppose that the only *a priori* knowledge that we dispose about the object is in the form:

$$\begin{aligned} \mathrm{E}\{S(\mathbf{x})\} &= s \\ \mathrm{E}\{H(\mathbf{x})\} &= h \end{aligned} \tag{19}$$

where $S(\mathbf{x})$ and $H(\mathbf{x})$ are two known functions. With these two global constraints, the ME principle gives us an exponential pdf, the same as in (18):

$$p(\mathbf{x}) = \frac{1}{Z}\ \exp[\lambda\, H(\mathbf{x}) + \mu\, S(\mathbf{x})] \tag{20}$$

The parameters (λ, μ) are related to (s, h). They are obtained by calculating the partition function $Z(\lambda, \mu)$:

$$Z(\lambda, \mu) = \int \exp[\lambda\, H(\mathbf{x}) + \mu\, S(\mathbf{x})]\ \mathrm{d}\mathbf{x} \tag{21}$$

and by solving the following system of equations:

$$\begin{aligned} \partial Z(\lambda, \mu) / \partial \mu &= s \\ \partial Z(\lambda, \mu) / \partial \lambda &= h \end{aligned} \tag{22}$$

II.3. How to choose H and S.

Without any other restriction, $H(x_i)$ and $S(x_i)$ can be any function, but we want to be able to estimate the parameters λ and μ from some statistics on the data $\mathbf{y}$. The matrix $\mathbf{A}$ in equation (1) is, usually, singular or at best very ill-conditioned. This means that we can only relate accurately the mean value of $\mathbf{y}$, $(\mathrm{E}\{y_i\})$, to the mean value of $\mathbf{x}$, $(\mathrm{E}\{x_i\})$. So that we can only estimate accurately one statistical observation on $\mathbf{x}$ and can estimate only one parameter. However we will see that, with some conditions, it is possible to estimate another statistical observation on $\mathbf{x}$ which is var(x) from var(y) if the SNR is high enough. So we study only two cases:

a) the case with only one parameter (k = 1) where:

$$p(\mathbf{x}) = \frac{1}{Z}\ \exp[\lambda\, H(\mathbf{x})] \tag{23}$$

b) a special case of two parameters (k = 2) where:

$$p(\mathbf{x}) = \frac{1}{Z}\ \exp\ [\lambda\, H(\mathbf{x}) + \mu\, S(\mathbf{x})]$$

with: $S(\mathbf{x}) = \sum x_i$ and $H(\mathbf{x}) = \sum H(x_i)$ (24)

Now, to choose $H(x_i)$, we accept two axioms:

i) When we change the scale of the image $\mathbf{u}=k\,\mathbf{x}$, if we note by (λ_1, μ_1) the parameters of $p_X(\mathbf{x}; \lambda_1, \mu_1)$ and by (λ_k, μ_k) the parameters of $p_U(\mathbf{u}; \lambda_k, \mu_k)$, the parameters (λ_k, μ_k) must be obtained from the parameters (λ_1, μ_1) by a fixed transformation Ψ, $\forall k$.

ii) The solution of our problem $\mathbf{x}$ obtained by (11) must be independent of the scale of the measurement.

Mathematically these two axioms are:

i) $\exists\, \Psi$: $(\lambda_k, \mu_k) = \Psi(k\ ,\ (\lambda_1, \mu_1)) \quad \forall\ k.$

ii) $$\mathbf{x} = \operatorname*{Arg\,min}_{\mathbf{v}>\mathbf{0}} \Big[\ \frac{(\mathbf{A}\mathbf{v}-\mathbf{y})^{\mathrm{t}}(\mathbf{A}\mathbf{v}-\mathbf{y})}{\mathrm{var}(\mathbf{y})\, f(\sigma^2)} - \mathrm{Ln}\, p_X(\mathbf{v};\, \lambda_1, \mu_1)\ \Big]$$

=>

$$k\,\mathbf{x} = \operatorname*{Arg\,min}_{\mathbf{v}>\mathbf{0}} \Big[\ \frac{(\mathbf{A}\mathbf{v}-k\,\mathbf{y})^{\mathrm{t}}(\mathbf{A}\mathbf{v}-k\,\mathbf{y})}{\mathrm{var}(k\,\mathbf{y})\, f(\sigma^2)} - \mathrm{Ln} p_U(\mathbf{v};\, \lambda_k, \mu_k)\ \Big]$$

Using these two axioms, we have showed [5] that, in the case when k=1, the only choices for H(x) are:

$$\{\mathrm{Ln}\ x,\ x^h\}$$

and, when k=2, the only choices for H are:

$$\{x\ \mathrm{Ln}\ x,\ \mathrm{Ln}\ x\ ,\ x^h\}.$$

In the following, we limit ourselves to the case k=2 with:

$$H(\mathbf{x}) = H_1(\mathbf{x}) = \sum \mathrm{Ln}\ x_i \tag{25}$$

We are going now to explain how one can estimate the parameters (λ, μ) from the data $\mathbf{y}$.

III. Parameter Estimation

If we knew the values of (s, h) and if we were able to

calculate the partition function $Z(\lambda, \mu)$, then the parameters (λ, μ) could be obtained by solving the system of equations (22). This is not the case, because, first it is not always possible to find an analytic solution for $Z(\lambda, \mu)$ and second, in practice (s, h) must be estimated from the data **y**. s is linear in **x**, it can be estimated from the data **y**, but h is not linear and it is not possible to estimate it directly from the data **y**. So we propose to estimate (λ, μ) by the method of moments, *i.e.* estimate the mean $e_x = E\{x_i\}$ and the variance $v_x = E\{(x_i - e_x)^2\}$ of the object pixels from the data **y** and relate them to the parameters (λ, μ). To do this, we show that if we can make the hypothesis:

$$\sum_i a_{ij} = \text{cte}, \quad \forall j, \tag{26}$$

then, (e_x, v_x) can be estimated from (e_y, v_y) by the following relations:

$$\begin{cases} e_x = m\, e_y / \Sigma\, \Sigma a_{ij} \\ v_x = 1/n\ (\mathbf{y} - e_y \mathbf{1})^t \mathbf{Q} (\mathbf{y} - e_y \mathbf{1}) \end{cases} \quad \text{with } \mathbf{Q} = \frac{\mathbf{A}\mathbf{A}^t}{(\mathbf{A}\mathbf{A}^t)^2 + \varepsilon \mathbf{I}} \tag{27}$$

where **Q** is the generalized inverse of $\mathbf{A}^t\mathbf{A}$. So if we can find a relationship between (λ, μ) and (e_x, v_x) the problem is then solved. To do this, note that we must be able to calculate the integrals:

$$I_\alpha(x) = \int_0^\infty x^{-\alpha}\, p(x)\, dx \qquad \text{for } \alpha = 0,\ 1,\ 2 \tag{28}$$

When $H(x)$ is in the form $\text{Ln}\, x$, $p(x)$ is in the form $p(x) = A\, x^{-\lambda} \exp[-\mu x]$ and we have analytic solutions to these integrals which converge for $\lambda < 1$, $\mu > 0$. It is then easy to show:

$$\begin{cases} e = -(\lambda+1)/\lambda \\ v = (\lambda+1)/\mu^2 \end{cases} \rightarrow \begin{cases} \lambda = (v - e^2)/v \\ \mu = e/v \end{cases} \tag{29}$$

When $H(x)$ is in the form $-x\text{Ln}\, x$, $p(x)$ is in the form $p(x) = A \exp[-\lambda\, x\, \text{Ln}\, x - \mu\, x]$ and we have no more analytic solutions to these integrals, so that it is impossible to establish an analytic relation between (λ, μ) and (e, v). However they converge for $\lambda < 0$, $\mu \in \mathbb{R}$, and it is possible to establish a numerical table which will give (λ, μ) *via* $(v/e^2, e)$. This will be explained in more details in a forthcoming paper.

IV. Summary of the Method

The method described above, and refered to as our **optimal method** is then the following:

i) Calculate e_y and v_y from the data **y**,
ii) Calculate e_x and v_x from e_y and v_y, using (27),
iii) Calculate (λ, μ) from e_x and v_x as described in the preceeding section, and
iv) Find the solution $\hat{\mathbf{x}}$ using (11).

However, step **ii)** needs a generalized inverse of $\mathbf{A}\mathbf{A}^t$. We present a **sub-optimal method** which does not need to do this. This is due to the fact that even when (λ, μ) are given, to determine $\hat{\mathbf{x}}$ we have to minimize (11) which is not quadratic in **x**. This can be done only by iteration. So at each iteration we have an estimate of the solution. So a sub-optimal estimate of (λ, μ) can be found from the current solution. This needs a good estimate in the first iteration.

The **sub-optimal** method works as follows:

i) The algorithm is initialized by either an approximate solution obtained previously or by: $\mathbf{A}^t\mathbf{y} / (\sum a_{1j})^2$
ii) A first approximation of the hyperparameters (λ, μ) is calculated using (29)
iii) A modified conjugate gradient algorithm is used to minimize (11) and to find a new estimate $\hat{\mathbf{x}}$.
iv) After some iterations, a new estimate of the hyperparameters (λ, μ) is calculated and we continue until some stop criterion is achieved.

More details about this algorithm, its theoretical foundations, its practical convergence and its use will appear shortly.

V. Some Simulation Results

In these simulations we call:

- **'Optimal 1'** the case when the actual hyperparameters (λ, μ) are known,
- **'Optimal 2'** the case when (λ, μ) are estimated from e_y and v_y using successively (27) and (29),
- **'Iterative 1'** the case when (λ, μ) are, at each step, estimated from the solution at that step using (e_x, v_x) and (29),
- **'Iterative 2'** the case when the noise variance σ^2 is also estimated at each step, by:

$\sigma^2 = \text{var}(\mathbf{b}) = \frac{1}{m}\sum (b_i - b)^2$ with $\mathbf{b} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}$ and $b = \frac{1}{m}\sum b_i$.

- **'ILSP'** (iterative least squares with positivity constraint) the case when we hold $\lambda = \mu = 0$ but apply the positivity constraint at each iteration.
- **'ILS'** (iterative least squares) the case when we hold $\lambda = \mu = 0$ and do not apply the positivity constraint at each iteration.

a) 1-D Image restoration

Figure 1 shows a 1-D object, the degraded data (blurred with a Gaussian impulse response and degraded by a zero-mean Gaussian noise with variance $\sigma^2 = 3.16 \times 10^{-2}$, which is equivalent, in this case, to a S/N ratio about 20 dB) and the different restorations obtained by the methods mentioned in the last paragraph.

The following table compares the different results. In this table $D = \|\hat{\mathbf{x}} - \mathbf{x}\|$ measures the misfit of the estimation, $Q = \|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|/\sigma^2$ measures the misfit of the data, H is the entropy of the estimation, S is the total intensity, $J = Q + \lambda H + \mu S$ is the achieved value of the criterion and, σ^2, λ and μ are the estimated parameters.

Method	Optimal1	Optimal2	Iterative 1	Iterative2
D	5.23×10^{-2}	5.19×10^{-2}	5.15×10^{-2}	5.15×10^{-2}
Q	107	110	109	127
H	-171	-174	-163	-170
S	67.5	66.6	67.5	67.5
J	430	509	511	541
σ^2	$\underline{3.16 \times 10^{-2}}$	$\underline{3.16 \times 10^{-2}}$	$\underline{3.16 \times 10^{-2}}$	3.25×10^{-2}
λ	$\underline{-0.65}$	- 0.96	- 0.95	- 0.97
μ	$\underline{3.13}$	3.49	3.66	3.69

What can be concluded from these preliminary results is that all the methods have an acceptable final distance D, but the

methods **Optimal 2**, **Iterative 1** and **Iterative 2** have over-estimated the parameter μ, but under-estimated the parameter λ. The noise variance in **Iterative 2** has been estimated correctly. We can see that the results obtained by **ILS** and **ILSP** have great variances and are not satisfactory. As mentioned before, these results are preliminary.

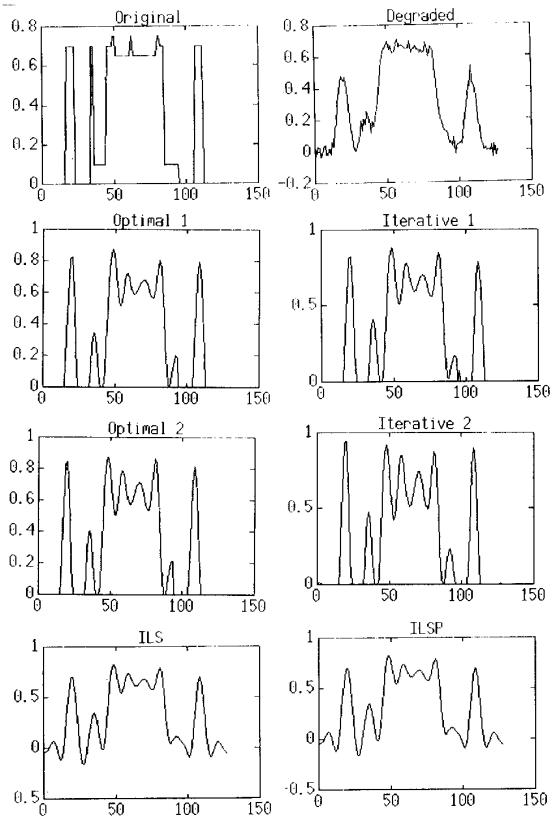

Figure 1: 1-D image restoration

2-D Image restoration

In these simulations we considered an image which is blurred by a Gaussian PSF and degraded by a Gaussian noise. Given then these data, we have restored the original image by **Optimal 1** and **Iterative 1** methods. Figure 3 shows these results. We can see that in this case, **Optimal 1** gave a more regularized result than **Iterative 1**.

VI. CONCLUSIONS

In this paper we have proposed a MAP Bayesian approach with Maximum Entropy (ME) priors to solve the integral equation which arises in various tomographic image restoration and reconstruction problems. A Bayesian approach is a coherent way for solving inverse problems because it allows us to take into account both the uncertainty on the data and the *a priori* information on the solution. One major difficulty, however, is the determination of *a priori* law of the image. The ME principle solves this difficulty in a coherent way.

When we know only the noise variance and some global constraints on the image, by applying the Bayesian approach

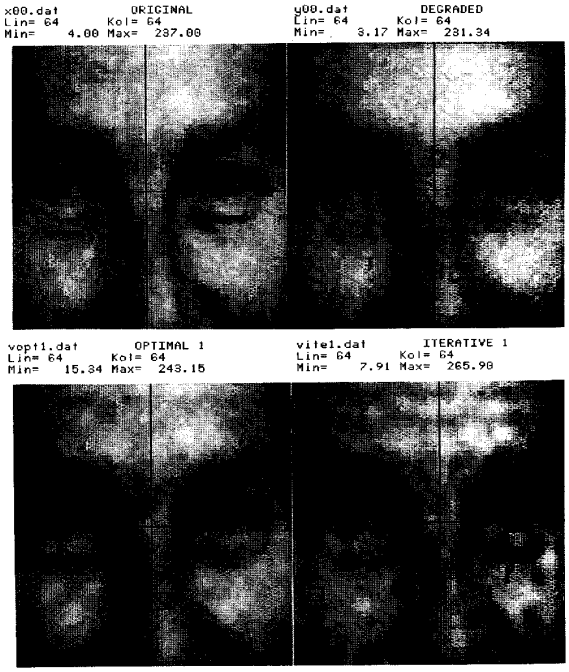

Figure 2: 2-D image restoration

and ME principle, we find a regularization problem in which the entropy of the image is used as a regularization functional.

In real applications two problems arise:

i) how to determine the hyper-parameters, i.e. the variance of the noise and the regularization parameters (λ, μ) from the data.

ii) how to minimize effectively the regularization criterion (12) which is not a quadratic form when the entropy is used as the regularization functional.

We proposed a method to determine iteratively the hyper-parameters and used a modified conjugate gradient method to solve the second.

REFERENCES

[1] J.H. Justice, *Maximum-Entropy and Bayesian Methods in Applied Statistics*, Cambridge: Cambridge University Press, 1986.

[2] J. Skilling, *Maximum-Entropy and Bayesian Methods*, J. Skilling ed., Dordrecht: *Kluwer Academic Publisher*, 1988.

[3] Mohammad-Djafari A. et Demoment G., "Utilisation de l'entropie dans les problèmes de restauration et de reconstruction d'images," Traitement du Signal, Vol. 5, No. 4, pp:235-248, (1988).

[4] Mohammad-Djafari A., "Bayesian Tomographic Image Processing with Maximum Entropy Priors," Statistics Earth and Space Sciences, Leuven, Belgium, August 22-26, 1989.

[5] Merle Ph., Marneffe Ch., Mohammad-Djafari A. and Demoment G., "Recherche d'une loi *a priori* en restauration d'images," internal report: LSS/89/023, CNRS-UPS, 1989.