Algorithms **2009**, *2*, 850-878; doi:10.3390/a2020850

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Article

Bayesian Maximum Entropy Based Algorithm for Digital X-ray Mammogram Processing

Radu Mutihac

University of Bucharest, PO Box MG-11, Bucharest, Romania
E-mail: mutihac@netscape.net

Received: 5 December 2008; in revised form: 11 May 2009 / Accepted: 28 May 2009 / Published: 9 June 2009

Abstract: Basics of Bayesian statistics in inverse problems using the maximum entropy principle are summarized in connection with the restoration of positive, additive images from various types of data like X-ray digital mammograms. An efficient iterative algorithm for image restoration from large data sets based on the conjugate gradient method and Lagrange multipliers in nonlinear optimization of a specific potential function was developed. The point spread function of the imaging system was determined by numerical simulations of inhomogeneous breast-like tissue with microcalcification inclusions of various opacities. The processed digital and digitized mammograms resulted superior in comparison with their raw counterparts in terms of contrast, resolution, noise, and visibility of details.

Keywords: Bayesian inference; inverse problems; digital image restoration; X-ray mammography; maximum entropy methods

1. Introduction

Mammography is currently the best radiological technique for early detection of breast cancer. An efficient and robust algorithm has been developed to process digitized and/or digital X-ray mammographic images in order to detect subtle abnormalities, such as different kind of lesions (e.g., microcalcifications, opacities or stellate patterns of straight lines), which may indicate the presence of malignancies. The study is intended to improve the diagnosis of mammographic lesions by using computer image analysis methods to derive quantitative description of their visual characteristics like edge definition, extent, and texture. Though experienced radiologists may often establish the malignant or benign nature of a lesion on the basis of these characteristics, the image perception by human observers is subject to complex interpretation by the visual system. The consequence is that the perceived image does not always correspond to the actual data contained in the image, which might lead to inaccurate or incorrect conclusions [1].

Irrespective of their sophistication and diversity, the numerical methods commonly used to convert experimental data into interpretable images and spectra rely on straightforward transforms, such as the Fourier transform (FT) or quite elaborated emerging classes of transforms like wavelets [2, 3], wedgelets [4], ridgelets [5], and so forth. Implemented by its version known as fast Fourier transform (FFT), the widely spread FT is a an efficient and convenient means of converting time domain data into frequency domain spectra. Moreover, the FFT is linear, so that the relative intensities of resonances of different widths, shapes, and frequencies are not distorted, and it is model-free. Yet experimental data are incomplete and noisy due to the limiting constraints of digital data recording and the finite acquisition time. The pitfall of most transforms is that imperfect data are directly transferred into the transform domain along with the signals of interest. The traditional approach to data processing in the transform domain is to ignore any imperfections in data, set to zero any unmeasured data points, and then proceed as if data were perfect. Contrarily, the maximum entropy (ME) principle enforces data processing in space (time) domain [6]. In data analysis, the ME methods are primarily used to reconstruct positive distributions, such as images and spectra, from blurred, noisy, and/or corrupted data. The ME methods may be developed on axiomatic foundations based on the probability calculus that has a special status as the only internally consistent language of inference [7, 8]. Within its framework, positive distributions ought to be assigned probabilities derived from their entropy.

Measurements on a complete collection of images, which correspond to all possible intensity distributions, act as a filter by restricting the focus on the images that satisfy the acquired data with noise. Among these, a natural choice may be the one that could have arisen in the maximum number of ways, depending on some counting rule. Such an approach to statistical inference was first suggested by Bayes [9] and completed by Jaynes [10]. Bayesian statistics provides a unifying and self-consistent framework for data modeling. Bayesian inference naturally deals with uncertainty in data explained by marginalization in predictions of other variables. Data overfitting and poor generalization are alleviated by incorporating the principle of Occam's razor, which controls model complexity and sets the preference for simpler models [12]. Bayesian inference satisfies the likelihood principle [13] in the sense that inferences depend only on the probabilities assigned to data that were measured and not on the properties of some admissible data that were never acquired. As such, out of all images that satisfy the measurements, the entropic restoration corresponds to selecting the most probable individual one that maximizes a given measure of entropy. The entropy of an image is considered to be related with its prior probability in the Bayesian sense. Two features distinguish the Bayesian approach to learning models from data. First, beliefs derived from background knowledge are used to select a prior probability distribution for model parameters. Secondly, predictions of future observations are performed by integrating the model's predictions with respect to the posterior parameter distribution obtained by updating this prior with new data. Accordingly, it is possible to associate error bars on image restorations generated by the ME principle, which allows assessing quantitatively and objectively the reliability of the extracted features. In addition, different variants of ME can be assessed and quantitative comparisons are feasible [11].

2. Methods

There are three fundamental aspects related with scientific investigations of the physical world, namely apparatus design, measuring techniques, and data processing from which conclusions are drawn. In a serious experiment, all these items has to be critically considered and carefully controlled. The discussion hereafter is limited to data processing only, which provides the natural framework to introduce the inverse problem of image restoration.

2.1. Inverse problems

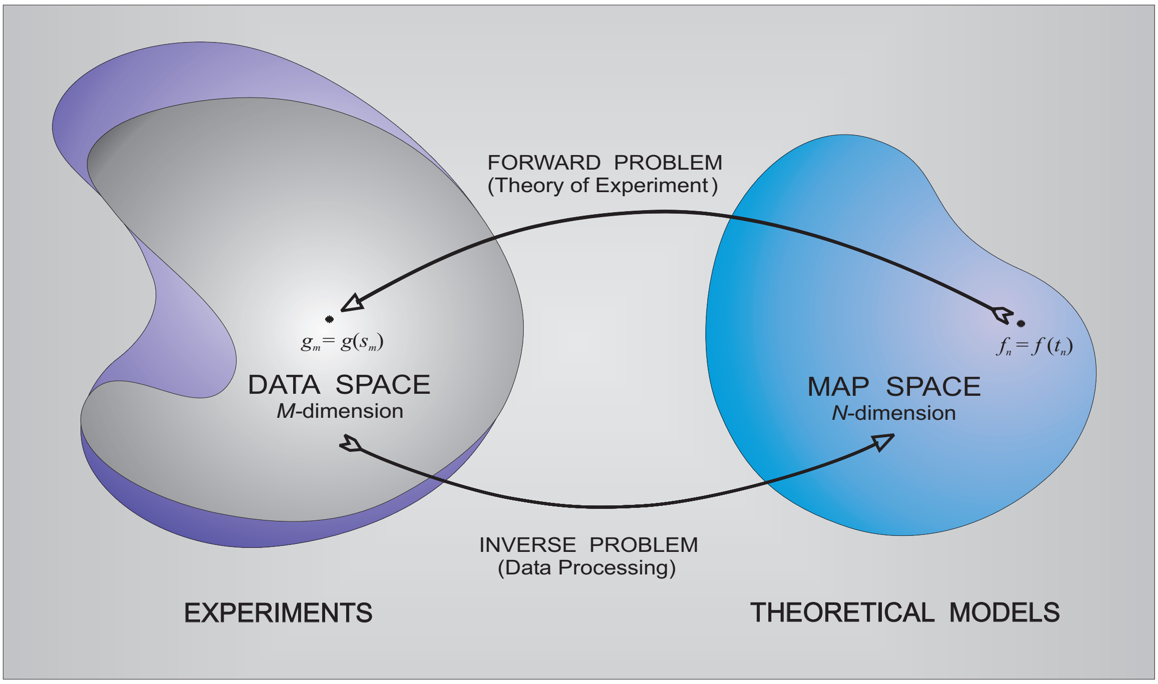
Data processing is getting from data space to map space. The theory of the experiment can be regarded as an operator from the map space to data space. The forward problem consists in applying this operator. It is always possible assuming that the experiment can be analyzed. An inverse problem refers to determining the causes for a desired or an observed effect or fitting the parameters of a mathematical model to reproduce observable data. Yet the inverse operator might not exist at all since it usually operates on a space larger than the data space (Figure 1). Typically, inverse problems are ill-posed, in contrast to the well-posed problems, which satisfy Hadamard's postulates of well-posedness: existence of solution(s) in the strict sense, uniqueness or continuous dependence on data, and stability [14]. An inverse problem is represented in the context of functional analysis by a mapping between some metric spaces. Most often, inverse problems are formulated in infinite dimensional spaces, whereas limitations imposed by a finite number of practical measurements and model parameters entail recasting inverse problems in a discrete form. Thus, inverse problems turn into ill-posed problems and special numerical methods (regularization) are introduced to relax the conditions imposed on solution(s) in order to circumvent data overfitting. When the stability of solution(s) is violated under data perturbations, regularization to cope with instabilities is needed, too. Several instances of regularized inverse problems may be interpreted in the sense of Bayesian inference [15].

The basic approach to resolving an inverse problem is to consider all solutions (models) in the map space that could give rise to measured data. Once the forward problem has been accomplished, solutions consistent with errors in data measurements and modeling according to some criterion(s) are searched. If all consistent models are very similar, than it is no need to make an option. However, if several distinct solutions exist that comply with the imposed criteria, then a rule for selecting the optimal one in some sense must be introduced in order to solve the inverse problem. In many practical experiments, the observed data are transforms of the quantities of interest. The linear transforms encompass a large class of such experiments. An integral form of a general linear relation between the observable data $g(s)$ and the unobserved interesting quantity $f(t)$ has the expression: [17]

$$g(s) = \int_D f(t)r(t,s)\mathrm{d}t + e(s) \tag{1}$$

where $r(t,s)$ is the transfer function or the response of the instrument, alternatively termed *blurring function* or *point spread function* (PSF) in image processing, which is considered to be perfectly known, while $e(s)$ stands for the measurement *errors* or *noise*. It is nevertheless assumed that the observed noise $e(s)$ is independent of the function $f(t)$.

Figure 1. Flow chart of the inverse and forward problem.



The correct formulation of problems involving incomplete and noisy data lies in the field of inverse theory. Previously, inverse problems have frequently been tackled by fitting empirical models, though this approach often leads to false confidence in the conclusions drawn. The inverse problem may be stated as to determine a unique and stable solution, say $\hat{f}(t)$, representing the unknown function $f(t)$, by using the measured values $g(s)$ of the available data. This is a typically ill-posed problem, in the sense that several solutions $f(t)$ may exist and produce the same data $g(s)$. Moreover, $g(s)$ in a physical experiment is observed on a finite set of isolated points s_m, $m = 1, 2, ..., M$ in the data space D, alternatively called the *transform space* when $f(t)$ and $g(s)$ are not in the same space [17]:

$$g_m = g(s_m) = \int_D f(t)r(t, s_m)\mathrm{d}t + e(s_m) = \int_D f(t)r_m(t)\mathrm{d}t + e_m,\ m = 1, 2, ..., M \tag{2}$$

The computations are effectively carried out after discretization of these equations. Further, the statement is correct only if discretization is adequately performed , i.e., each component g_m, $m = 1, 2, ..., M$ of $g(s)$ measures a distinct aspect f_n, $n = 1, 2, ..., N$ of $f(t)$ through its own linear response kernel $r_m(t)$, $m = 1, 2, ..., M$, and with its own additive measuring error e_m, $m = 1, 2, ..., M$. Actually, a large number N of discrete points t_n, $n = 1, 2, ..., N$ sufficiently evenly spaced is needed, so that neither $f(t)$ nor $r_m(t)$ vary significantly between t_{n-1} and t_{n+1}, $n = 1, 2, ..., N - 1$. For such a dense set of points, the integral expression for each g_m in eq. (2) can be replaced by a quadrature-like form:

$$g_m = \sum_{n=1}^{N} R_{mn} f_n + e_m,\ m = 1, 2, ..., M \tag{3}$$

or, in a more compact matrix form:

$$\boldsymbol{g} = \mathbf{R}\,\boldsymbol{f} + \boldsymbol{e} \tag{4}$$

where $\boldsymbol{g} = \{g_1, g_2, ..., g_M\}$ is the data vector, $\boldsymbol{f} = \{f_1, f_2, ..., f_M\}$ is the image vector, $\boldsymbol{e} = \{e_1, e_2, ..., e_M\}$ is the error vector, and the matrix R of size $M \times N$ has the elements:

$$R_{mn} = r_m(t_n)\,(t_{n+1} - t_{n+2})/2 \tag{5}$$

Each g_m in the data space may approximate the value of $f(t)$ at a certain location t_n in the image space, in which case $r(t_n, s_m) = R_{mn}$ should approximately have the form of a narrow instrumental response termed PSF and centered around $t = t_n$. This is the case of components g_m, $m = 1, 2, ..., M$ located in an entirely different function space from f_n, $n = 1, 2, ..., N$, such as measuring different Fourier components of $\boldsymbol{f}$.

Though the measurements have reasonably been reduced to a finite number, the problem remains ill-posed and the matrix $\boldsymbol{R}$ is likely to be either *singular* or very *ill-conditioned*. In order to turn the inverse problem to well-posed, the goal is relaxed by asking for some reliable estimate only, say $\tilde{\boldsymbol{f}}$, which, in a practical sense, may be considered "very close" or "best approximation" to the exact image $\boldsymbol{f}$, given the measured sample data g_m, $m = 1, 2, ..., M$, the known response function $r_m(t)$, $m = 1, 2, ..., M$, of the measuring instrument, and some information about the noise components e_m, $m = 1, 2, ..., M$, such as their covariance matrix $\mathbf{C} = \{C_{ij}\}_{i,j=1,2,...,M}$. In this respect, Gull and Daniell [18] suggested the ME principle to assign a probability distribution to any image along with a χ^2-constraint for handling the errors e_m, $m = 1, 2, ..., M$. The χ^2-distribution measures how well a model $\tilde{\boldsymbol{f}}$ fits the measured data $\boldsymbol{g}$:

$$\chi^2\left(\tilde{\boldsymbol{f}}\right) = \sum_{i=1}^{M}\sum_{j=1}^{M}\left[g_i - \sum_{n=1}^{N} R_{in}\tilde{f}(t_n)\right] C_{ij}^{-1} \left[g_j - \sum_{n=1}^{N} R_{jn}\tilde{f}(t_n)\right] \tag{6}$$

or, in the form of a matrix product:

$$\chi^2\left(\tilde{\boldsymbol{f}}\right) = \left(\boldsymbol{g} - \mathbf{R}\tilde{\boldsymbol{f}}\right)^T \mathbf{C}^{-1} \left(\boldsymbol{g} - \mathbf{R}\tilde{\boldsymbol{f}}\right) \tag{7}$$

where $\mathbf{C}^{-1}$ is the inverse of the covariance matrix of noise. An approximate equality holds if neglecting the off-diagonal elements, which is strictly true when the noise is not correlated among pixels:

$$\chi^2\left(\tilde{\boldsymbol{f}}\right) \cong \sum_{m=1}^{M}\left[\frac{g_m - \sum_{n=1}^{N} R_{mn}\tilde{f}(t_n)}{\sigma_m}\right]^2 \tag{8}$$

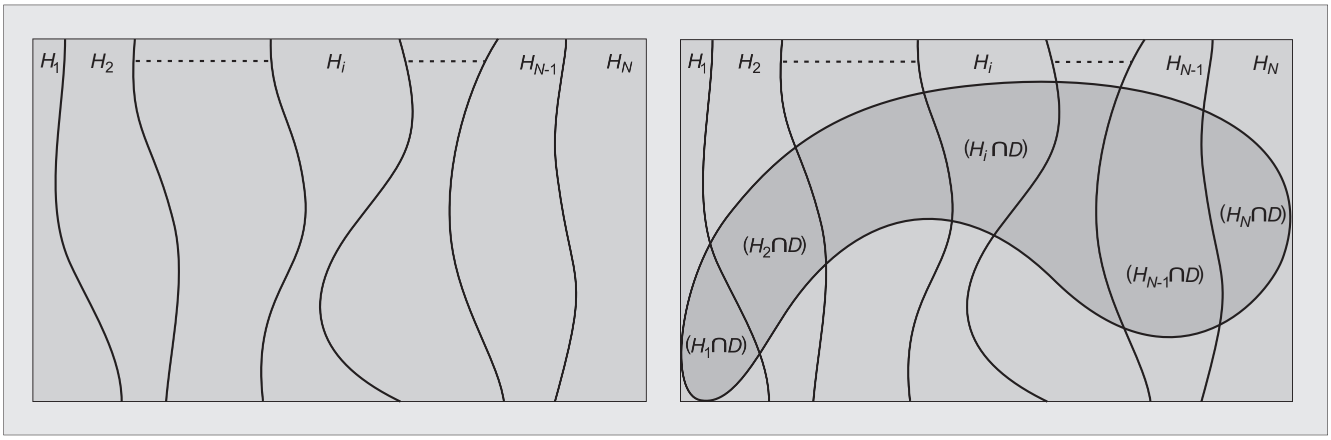
where the noise standard deviation, $\boldsymbol{\sigma}$, has the components:

$$\sigma_m = (C_{mm})^{\frac{1}{2}}\,,\ m = 1, 2, ..., M \tag{9}$$

2.2. Bayesian image modeling

Bayes' theorem follows directly from the standard axioms of probability relating the conditional probabilities of events. Probabilities may have two meanings: (i) to describe frequencies of outcomes in random experiments, or more generally, (ii) to describe degrees of belief in propositions (hypotheses) that do not involve random variables and repetitive events. The degrees of belief can be mapped onto probabilities if they satisfy simple consistency rules known as the Cox axioms [16], so that probabilities allow describing hypotheses as well. Bayesian methods deal with explicit assumptions and provide rules for reasoning consistently given those assumptions. Bayesian inferences are subjective in the sense that it is not plausible to reason about data without making assumptions based on some theoretical and/or experimental background. Yet Bayesian inferences are objective since following the same assumptions on a model, then identical inferences are drawn [12]. The fundamental concept of Bayesian analysis is that the plausibility of alternative hypotheses is represented by probabilities and inference is performed by evaluating these probabilities. Since plausibility is itself conditioned by some sort of prior assumptions, all Bayesian probabilities are regarded as *conditional* on some collective background information.

Figure 2. N mutually exclusive and exhaustive hypotheses $\{H_i\}_i$ (left); an event D in the hypothesis space (right).



A set of events is termed *exhaustive* if one or more of them must occur. Suppose that hypotheses $H_1, H_2, ..., H_N$ are mutually exclusive and exhaustive, then the rule of total probability [19] gives for any event D in the hypothesis space (Figure 2):

$$P(D) = \sum_{i=1}^{N} P(H_i)\, P(D|H_i) \tag{10}$$

Further, assume that the probabilities $\{P(H_i)\}_i$ and $\{P(D|H_i)\}_i$, $i = 1, 2, ..., N$ are known. Bayes' rule makes use of these probabilities to determine the conditional probabilities:

$$P(H_i|D) = \frac{P(H_i)\, P(D|H_i)}{\sum_{i=1}^{N} P(H_i)\, P(D|H_i)},\ i = 1, 2, ..., N \tag{11}$$

Bayes' theorem makes no reference to any sample or hypothesis space, nor it determines the numerical value of any probability directly from available information. As a prerequisite to apply Bayes' theorem, a principle to cast available information into numerical values is needed. It explicitly shows that the observations giving $\{P(D|H_i)\}$ are not enough to determine the results $\{P(H_i|D)\}$ since priors $\{P(H_i)\}$ need to be assigned.

In statistical restoration of gray-level digital images, the basic assumption is that a scene is adequately represented by an orderly array of N pixels. The task is to infer reliable statistical descriptions of images, which are gray-scale digitized pictures and stored as an array of integers representing the intensity of gray level in each pixel. Then the shape of any positive, additive image can be directly identified with a probability distribution. The image is conceived as an outcome of a random vector $\boldsymbol{f} = \{f_1, f_2, ..., f_N\}$ given in the form of a positive, additive probability density function (pdf). Likewise, the measured data $\boldsymbol{g} = \{g_1, g_2, ..., g_M\}$ are expressed in the form of a probability distribution. Further assumption refers to image data as a linear function of physical intensity, and that the noise $\boldsymbol{e}$ is data independent, additive, and Gaussian with zero mean and known standard deviation σ_m in each measured pixel m, $m = 1, 2, ..., M$.

A model is conceived as means to make inferences, predictions, and decisions. Any model depends on some free parameters. There are two levels of inference in any data modeling process. The first concerns fitting each model to the data, that is, to infer what the free parameters of each model might be given the data and some background information. The second refers to model comparison that ranks alternative models on the basis of their plausibility with respect to the data. Bayesian methods deal consistently and quantitatively with both inductive inference levels in data modeling (Figure 3).

Figure 3. Statistical inference levels in data modeling.

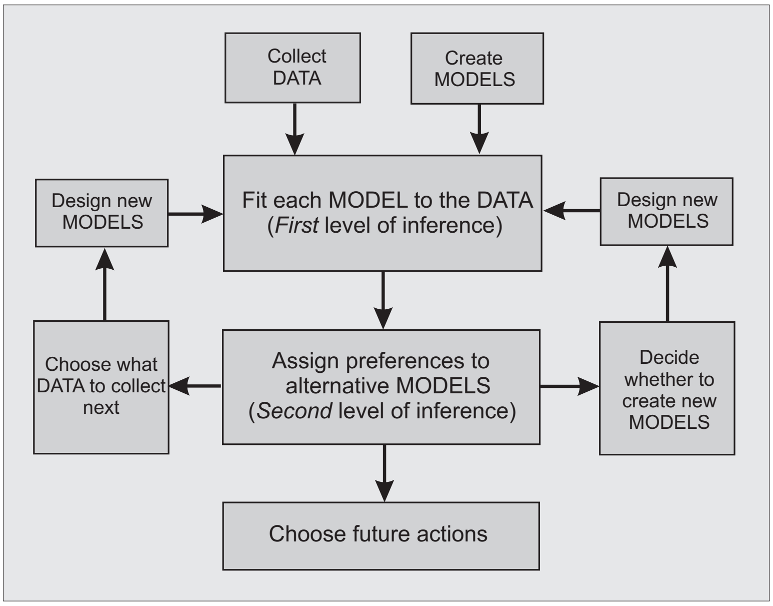

Collect DATA
Create MODELS
Design new MODELS
Fit each MODEL to the DATA (*First* level of inference)
Design new MODELS
Choose what DATA to collect next
Assign preferences to alternative MODELS (*Second* level of inference)
Decide whether to create new MODELS
Choose future actions

Image models are derived on the basis of intuitive ideas and observations of real images, and have to comply with certain criteria of invariance, that is, operations on images should not affect their likelihood. Each model comprises a hypothesis H with some free parameters $\boldsymbol{w} = \{\alpha, \beta, ...\}$ that assign a probability density $P(\boldsymbol{f}|\boldsymbol{w}, H)$ over the entire image space and normalized to integrate to unity. Prior beliefs about the validity of H before data acquisition are embedded in $P(H)$. Extreme choices for $P(H)$ only may exceed the evidence $P(\boldsymbol{f}|H)$, thus the plausibility $P(H|\boldsymbol{f})$ of H is given essentially by the evidence $P(\boldsymbol{f}|H)$ of the image $\boldsymbol{f}$.

Initially, the free parameters $\boldsymbol{w} = \{\alpha, \beta, ...\}$ are either unknown or they are assigned very wide prior distributions. The task is to search for the best fit parameter set $\boldsymbol{w}_{MP}$, which has the largest likelihood given the image. Following Bayes' theorem:

$$P(\boldsymbol{w}|\boldsymbol{f}, H) = \frac{P(\boldsymbol{f}|\boldsymbol{w}, H)\, P(\boldsymbol{w}|H)}{P(\boldsymbol{f}|H)} \tag{12}$$

where $P(\boldsymbol{f}|\boldsymbol{w}, H)$ is the likelihood of the image $\boldsymbol{f}$ given $\boldsymbol{w}$, $P(\boldsymbol{w}|H)$ is the prior distribution of $\boldsymbol{w}$, and

$P(\boldsymbol{f}|H)$ is the evidence for the image $\boldsymbol{f}$. A prior $P(\boldsymbol{w}|H)$ has to be assigned quite subjectively based on our beliefs about images.

For completeness, Bayes' theorem should include the background knowledge denoted hereafter by I:

$$P(\boldsymbol{w}|\boldsymbol{f},H,I) = \frac{P(\boldsymbol{f}|\boldsymbol{w},H,I)\,P(\boldsymbol{w}|H,I)}{P(\boldsymbol{f}|H,I)} \tag{13}$$

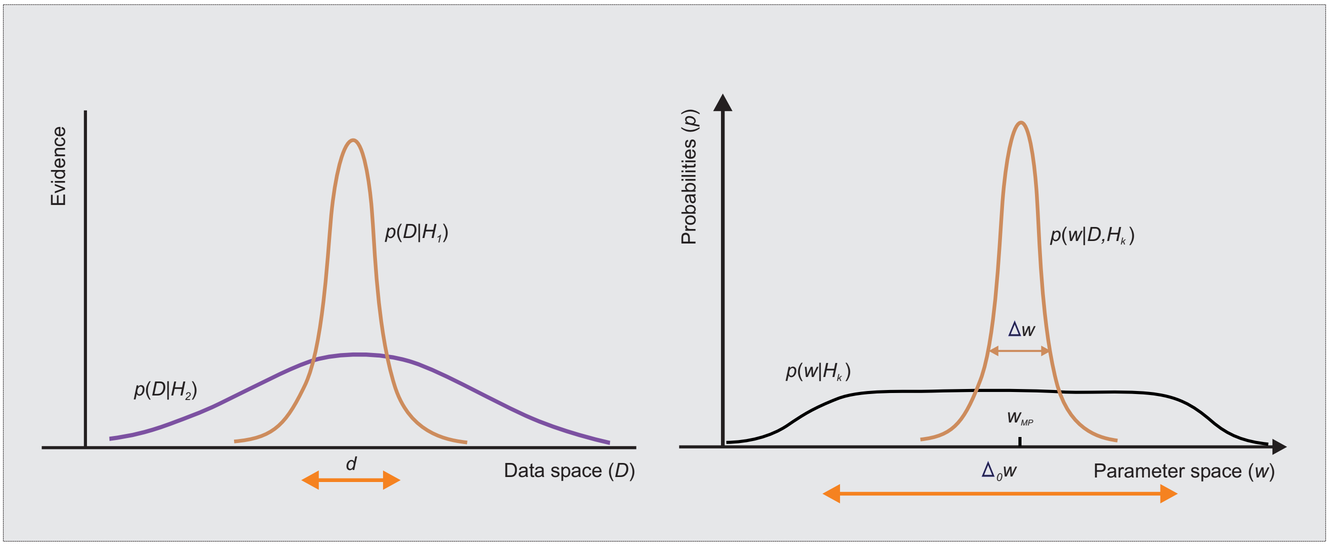
If some additional information, say H', becomes available, the estimate probability $P(\boldsymbol{w}|\boldsymbol{f},H,I)$ can be refined by simply substituting the background information I with $H'I$ in eq. (13). Elementary manipulations of the product rule of probabilities lead to:

$$P(\boldsymbol{w}|\boldsymbol{f},H'H,I) = \frac{P(\boldsymbol{f}|\boldsymbol{w},H'H,I)\,P(\boldsymbol{w}|H'H,I)}{P(\boldsymbol{f}|H'H,I)} \tag{14}$$

It turns out that the product $H'H$ behaves like H' and H were taken together. As such, the Bayesian analysis reports newly acquired information on the image space, the so-called ”posterior bubble”, but it does not suggest any prior. As the noiseless limit of the likelihood $P(\boldsymbol{f}|\boldsymbol{w},H)$ is approaching a delta function, the prior becomes irrelevant because data are forcing the correct result, irrespective of theory. Yet images are usually digitized on thousands or even million of cells, each of which has an intensity to be estimated. The prior $P(\boldsymbol{w}|H)$ plays a crucial role in such complex cases.

A nontrivial prior knowledge $P(\boldsymbol{w}|H'H,I)$ is necessary to resolve the degeneracy of the problem and to obtain a unique and stable solution, since the number of parameters (components of $\boldsymbol{f}$) to be estimated is comparable or larger than the number of measured data (components of $\boldsymbol{g}$). A reasonable choice for prior knowledge is that the image $\boldsymbol{f}$ is a positive-valued vector. The ME methods have historically been designed to make use of this prior knowledge in order to solve the inverse problem of image restoration.

Figure 4. Occam's razor.



Since the probability of the image $\boldsymbol{f}$ after observing the data $\boldsymbol{g}$, that is, $P(\boldsymbol{w}|\boldsymbol{f},H)$ is normalized to unity, then the denominator in eq. (12) ought to ensure $P(\boldsymbol{f}|H) = \int_{\boldsymbol{w}} P(\boldsymbol{f}|\boldsymbol{w},H)\,P(\boldsymbol{w}|H)\,\mathrm{d}\boldsymbol{w}$. The integral is often dominated by the likelihood in $\boldsymbol{w}_{MP}$, so that the evidence is approximated by [20]:

$$P(\boldsymbol{f}|H) \cong P(\boldsymbol{f}|\boldsymbol{w}_{MP},H)\,P(\boldsymbol{w}_{MP}|H)\,\Delta\boldsymbol{w} \tag{15}$$

Assuming uniform prior parameter distributions $P(\boldsymbol{w}|H)$ over all admissible parameter sets $\Delta_0\boldsymbol{w}$, then $P(\boldsymbol{w}_{MP}) = 1/\Delta_0\boldsymbol{w}$ and the evidence becomes:

$$P(\boldsymbol{f}|H) \cong P(\boldsymbol{f}|\boldsymbol{w}_{MP}, H)\,\frac{\Delta\boldsymbol{w}}{\Delta_0\boldsymbol{w}} \tag{16}$$

The ratio $\Delta\boldsymbol{w}/\Delta_0\boldsymbol{w}$ between the posterior accessible volume of the model parameter space and the prior accessible volume is the Occam's razor.

Data overfitting and poor generalization are alleviated in Bayesian methods by incorporating the principle of Occam's razor that may be explained in the following terms [21]. Assume that a simple model H_1 and a complex one H_2 are making predictions, $P(D|H_1)$ and $P(D|H_2)$, respectively, over the possible data sets D (Figure 4). The simple model H_1 makes predictions over a limited range of data sets, say d, whereas the complex and more powerful model H_2, which is likely to have more free parameters than H_1, is able to predict a greater variety of data sets. Therefore, H_2 does not predict the data sets in d as strongly as H_1. Now, if equal prior probabilities were assigned to both models, for any data set that happens to fall in the region d, the less powerful model H_1 is the most probable model. It turns out that Occam's razor can be conceived as a natural measure of complexity of a parametric family of distributions relative to the underlying distribution [22].

2.3. *Image entropy*

In physics, the entropy S of an isolated system in some macroscopic state is the logarithm of the number of microscopically distinct configurations, say W, that all are consistent with the observed macroscopic one (Boltzmann's principle):

$$S = k_B \log W \tag{17}$$

where k_B stands for Boltzmann's constant. It was first perceived by Boltzmann, yet he never wrote it in the above form but Planck did in 1906. It should be stressed that W denotes here the number of microstates of a single isolated system. Though W can only change by an integer, it is nevertheless very large and its property of being discrete can not be normally detected on a macroscopic scale.

In statistics, Shannon [23] defined the entropy of a system as a measure of uncertainty of its structure. Shannon's function is based on the concept that the information gain from an event is inversely related to its probability of occurrence. Accordingly, the entropy associated to a discrete random variable X with N possible outcomes $\{x_1, x_2, ..., x_N\}$ is defined as:

$$S\left(\{p_n\}\right) = -\sum_{n=1}^{N} p_n \, \log p_n \tag{18}$$

where $\{p_n\} = P\left(X = x\right)$ is the probability distribution of X. The base of the logarithm is irrelevant since the logarithmic function is selected on the basis of some positive argument and satisfying the property of additivity. Several authors have used Shannon's concept in image processing and pattern recognition problems. More definitions of entropy exist along with some justifications, such as considering the functional of exponential nature rather than logarithmic [24].

If a digital image made out of N pixels is represented as a sequence of positive numbers f_n, $n = 1, 2..., N$ with the corresponding proportions:

$$p_n = \frac{f_n}{\sum_{i=1}^{N} f_n},\ n = 1, 2, ..., N \tag{19}$$

then the following axioms are satisfied :

$$p_n \geq 0,\ n = 1, 2, ..., N \quad \text{(positivity)} \tag{20}$$

$$p_{n \cup m \cup ...} = p_n + p_m + ...\ ,\ n \neq m \quad \text{(additivity)} \tag{21}$$

$$\sum_{n=1}^{N} p_n = 1 \quad \text{(normalization)} \tag{22}$$

Whether it might be for spectral analysis of time series, radio or optical astronomy, X-ray medical imaging, and generally for any restoration/reconstruction of positive, additive images, the ME principle assigns a prior probability to any given image $\boldsymbol{f}$. Suppose that the luminance in each pixel is quantified (in some units) to an integer value. Denote by U the total number of quanta in the whole image. The conservation of the number of quanta equates to the following equality:

$$U = \sum_{m=1}^{M} g_m = \sum_{n=1}^{N} f_n \tag{23}$$

A prior can be formulated based on the fair reason that each quanta has an equal a priori chance of being in any pixel [25]. The number of ways for getting a particular image configuration $\boldsymbol{f} = \{f_1, f_2, ..., f_N\}$ out of U quanta is:

$$\frac{U!}{f_1! f_2! \cdots f_N!} \cong \exp\left[-\sum_{n=1}^{N} f_n \ln\left(\frac{f_n}{U}\right) + \frac{1}{2}\left(\ln U - \sum_{n=1}^{N} \ln f_n\right)\right] \tag{24}$$

where the symbol "!" denotes the factorial operator. Here, the left side represents the number of distinct orderings of all luminance quanta, divided by the numbers of equivalent reorderings within each pixel (the total number of permutations of U elements in which there are f_1 alike, f_1 alike, ... , f_N alike). The right side is Stirling's approximation of the factorial function for large U [26]. By neglecting terms of order $\ln\ U$ in the presence of terms of order U, the argument of the exponential becomes:

$$S(\boldsymbol{f}) = -\sum_{n=1}^{N} f_n \ln\left(\frac{f_n}{U}\right) \tag{25}$$

This expression stands for the *configurational entropy* of an image, which differs from the *thermodynamic entropy* of a beam of photons or the *informational entropy* in statistics! However, the most appropriate definition of image entropy and the particular mathematical function to describe it are still topics of debate [18].

2.4. Bayesian image restoration

The central task in Bayesian image restoration is to select the best statistical estimator $\tilde{\boldsymbol{f}}$ of an image $\boldsymbol{f}$, assuming that both the image $\boldsymbol{f}$ and the data $\boldsymbol{g}$ are outcomes of random vectors given in the form of probability distribution. The incomplete knowledge about the errors in measurements and modeling are included in a noise random vector $\boldsymbol{e}$, also expressed as probability distribution. Likewise, a prior probability has to be defined that reflects the uncertainty or partial knowledge on the image $\boldsymbol{f}$.

In the linear case $\boldsymbol{g} = \mathbf{R}\boldsymbol{f} + \boldsymbol{e}$, where the N-dimensional image vector $\boldsymbol{f}$ consists of the pixel values of a latent image, the M-dimensional data vector $\boldsymbol{g}$ consists of the pixel values of an observed image supposed to be a degraded version of $\boldsymbol{f}$, the matrix $\mathbf{R}$ is the PSF of the imaging system, and the errors $\boldsymbol{e}$ are assumed additive and zero-mean Gaussian. As such, $\boldsymbol{e} = k\boldsymbol{\sigma}$, which entails:

$$\boldsymbol{g} = \mathbf{R}\boldsymbol{f} + k\boldsymbol{\sigma} \tag{26}$$

where $\boldsymbol{\sigma}$ is the noise standard deviation and k is drawn from the standard normal distribution:

$$P(k) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{k^2}{2}\right) \tag{27}$$

Likelihood

After M measurements, the likelihood of the data $\boldsymbol{g}$ results:

$$P\left(\boldsymbol{g}|\boldsymbol{f}, \mathbf{C}, H\right) = \frac{1}{(2\pi)^{\frac{M}{2}} \cdot \det^{\frac{1}{2}} \mathbf{C}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{g} - \mathbf{R}\boldsymbol{f})^T \mathbf{C}^{-1}(\boldsymbol{g} - \mathbf{R}\boldsymbol{f})\right] \tag{28}$$

where $\mathbf{C}$ is the covariance matrix of the error vector $\boldsymbol{e}$. If the pixels are uncorrelated and each pixel has the standard deviation σ_m, $m = 1, 2, ..., M$, then the symmetric full rank covariance matrix $\mathbf{C}$ becomes diagonal with elements $C_{mm} = \sigma_m^2$, $m = 1, 2, ..., M$, that is, $\mathbf{C} = \sigma_m^2 \mathbf{I}$. Hence the probability of the data $\boldsymbol{g}$ given the image $\boldsymbol{f}$ can be written as:

$$P\left(\boldsymbol{g}|\boldsymbol{f}, \mathbf{C}, H\right) = \frac{1}{(2\pi)^{\frac{M}{2}} \cdot \prod_{m=1}^{M} \sigma_m} \cdot \exp\left[-\frac{1}{2}\sum_{m=1}^{M} \frac{\left(g_m - \sum_{n=1}^{N} R_{mn} f_n\right)^2}{\sigma_m^2}\right] \tag{29}$$

The argument of the exponential function amounts to $-\frac{1}{2}\chi^2(\boldsymbol{f})$ according to eq. (8), which gives a more compact relationship:

$$P\left(\boldsymbol{g}|\boldsymbol{f}, \mathbf{C}, H\right) = \frac{1}{(2\pi)^{\frac{M}{2}} \cdot \prod_{m=1}^{M} \sigma_m} \cdot \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{f})\right] \tag{30}$$

The Poisson distribution is often better modeling the fluctuations in data acquisition rather than the Gaussian distribution assumed here. Nevertheless, the choice depends on the statistical characteristics of the noise assumed *a priori* known.

If $\boldsymbol{g} = \mathbf{R}\boldsymbol{f} + k\boldsymbol{\sigma}$ and the noise standard error is the same for all pixels, $\sigma_m = \sigma$, $m = 1, 2, ..., M$, the likelihood of data $P\left(\boldsymbol{g}|\boldsymbol{f}, \mathbf{C}, H\right)$ can alternatively be written as:

$$P\left(\boldsymbol{g}|\boldsymbol{f}, \beta, H\right) = \frac{1}{Z_b(\beta)} \cdot \exp\left[-\sum_{m=1}^{M} \beta E_b\left(\boldsymbol{g}|\boldsymbol{f}, H\right)\right] \tag{31}$$

where $\beta = 1/\sigma^2$ is a measure of the error (noise) in each pixel, $Z_b(\beta) = (2\pi/\beta)^{\frac{M}{2}}$ is the noise partition function, and the error function is:

$$E_b\left(\boldsymbol{g}|\boldsymbol{f},H\right) = \frac{1}{2}\cdot\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2} = \frac{1}{2}\sum_{m=1}^{M}\frac{\left(g_m - \sum_{n=1}^{N} R_{mn}f_n\right)^2}{\sigma^2} = \frac{1}{2}\chi^2(\boldsymbol{f}) \tag{32}$$

Prior probability

Applying the ME principle amounts to assigning a distribution $\{p_i\}_{i=1,2,\dots}$ on some hypothesis space $\{H_i\}_{i=1,2,\dots}$ by the criterion that it shall maximize some form of entropy subject to constraints that express the desired properties of the distribution, but are not sufficient to determine it. The ME methods require specifying in advance a definite hypothesis space which sets down the possibilities to be taken into consideration. The result is a probability distribution, rather than a probability. The ME probability of a single hypothesis H that is not embedded in a space of alternative hypotheses does not make any sense. The ME approaches do not need numerical values of any probabilities on that space as input, rather they assign numerical values to available information as expressed by the choice of the hypothesis space and constraints. By all means, any set of constraints on $\boldsymbol{f}$ is affecting the amount by which the image restoration is offset from reality.

In the ergodic case, where reasons are to believe that all *a priori* probabilities of the microscopic configurations are the same, the Bayesian prior probability $P\left(\boldsymbol{f}|\alpha,H\right)$ for a macroscopic state $\boldsymbol{f}$ with entropy $S(\boldsymbol{f})$ is postulated [27] as given by a general potential function $\Phi\left(\boldsymbol{f}\right)$, such as:

$$P\left(\boldsymbol{f}|\alpha,H\right) = \frac{1}{Z(\alpha)}\,\exp\left[-\alpha\,\Phi\left(\boldsymbol{f}\right)\right] \tag{33}$$

where α is a positive parameter and $Z(\alpha)$ is a normalizing factor. When partial information on a random process is available only, the entropic prior is assumed to follow the ME law that complies with prior information. The entropic prior in the discrete case corresponds to potential functions like:

$$\Phi\left(\boldsymbol{f}\right) = -S(\boldsymbol{f}) = \sum_{n=1}^{N} f_n \ln\left(\frac{f_n}{U}\right) \tag{34}$$

resulting in:

$$P\left(\boldsymbol{f}|\alpha,H\right) = \frac{1}{Z(\alpha)}\,\exp\left[-\alpha\sum_{n=1}^{N} f_n \ln\left(\frac{f_n}{U}\right)\right] \tag{35}$$

The best suited prior probability distribution, $P\left(\boldsymbol{f}|\alpha,H\right)$, should incorporate as much as possible of the available statistical characteristics of the ensemble of images to which the original image, $\boldsymbol{f}$, is assumed to belong. Apart from the entropic one [28], there are more different choices to be considered and evaluated for prior probability distribution of an image.

Posterior probability

The full joint posterior probability $P\left(\boldsymbol{f},\theta|\boldsymbol{g},H\right)$ of an image $\boldsymbol{f}$ and the unknown PSF parameters denoted generically by θ need to be evaluated. Then the required inference about the posterior probability

$P\left(\boldsymbol{f}|\boldsymbol{g},H\right)$ is obtained as a marginal integral of this joint posterior over the PSF parameters:

$$P\left(\boldsymbol{f}|\boldsymbol{g},H\right)=\int P\left(\boldsymbol{f},\theta|\boldsymbol{g},H\right)\mathrm{d}\theta=\int P\left(\boldsymbol{f}|\theta,\boldsymbol{g},H\right)\cdot P\left(\theta|\boldsymbol{g},H\right)\mathrm{d}\theta \tag{36}$$

In this respect, Bayes' theorem applied for the parameters θ gives:

$$P\left(\theta|\boldsymbol{g},H\right)=\frac{P\left(\boldsymbol{g}|\theta,H\right)\cdot P(\theta|H)}{P(\boldsymbol{g}|H)} \tag{37}$$

which substituted in eq. (36) leads to:

$$\int P\left(\boldsymbol{f},\theta|\boldsymbol{g},H\right)\mathrm{d}\theta\propto\int P\left(\boldsymbol{f}|\theta,\boldsymbol{g},H\right)\cdot P\left(\boldsymbol{g}|\theta,H\right)\cdot\mathrm{d}\theta \tag{38}$$

If the evidence $P\left(\boldsymbol{g}|\theta,H\right)$ is sharply peaked around some value $\widehat{\theta}$ and the prior $P\left(\theta|H\right)$ is quite flat in that region, then $P\left(\boldsymbol{f}|\boldsymbol{g},H\right)\cong P\left(\boldsymbol{f}|\widehat{\theta},\boldsymbol{g},H\right)$. Otherwise, if the marginal integrant is not well approximated at the modal value of the evidence, then misleading narrow posterior probability densities may result.

The posterior probability of an image $\boldsymbol{f}$ drawn from some measured data $\boldsymbol{g}$ is given by Bayes' theorem:

$$P\left(\boldsymbol{f}|\boldsymbol{g},\alpha,\mathbf{C},H\right)\propto\exp\left[-\sum_{n=1}^{N}f_n\,\ln\left(\frac{f_n}{U}\right)\right]\cdot\exp\left[-\frac{1}{2}\sum_{m=1}^{M}\frac{\left(g_m-\sum_{n=1}^{N}R_{mn}f_n\right)}{\sigma_m^2}\right] \tag{39}$$

An estimation rule, such as posterior mean or maximum a posteriori (MAP), is required to choose an optimal, unique, and stable solution $\widetilde{\boldsymbol{f}}$ for the estimated image. The posterior probability is assumed to summarize the full state of knowledge on a given scene. Producing a single image as the best restoration naturally leads to the most likely one which maximizes the posterior probability $P\left(\boldsymbol{f}|\boldsymbol{g},\alpha,\mathbf{C},H\right)$, along with some statement of reliability derived from the spread of all admissible images. To conclude with, the Bayesian approach to inverse problem of image restoration equates to maximize the posterior probability in eq. (39), or, equivalently, to minimize the expression:

$$-\ln\left[P\left(\boldsymbol{f}|\boldsymbol{g},\alpha,\mathbf{C},H\right)\right]=\frac{1}{2}\chi^2(\boldsymbol{f})-S(\boldsymbol{f}) \tag{40}$$

where $\chi^2(\boldsymbol{f})$ is the chi-square function and $S(\boldsymbol{f})$ is the conformational entropy of the image $\boldsymbol{f}$. Its numerical minimization falls in the field of constrained nonlinear optimization problems.

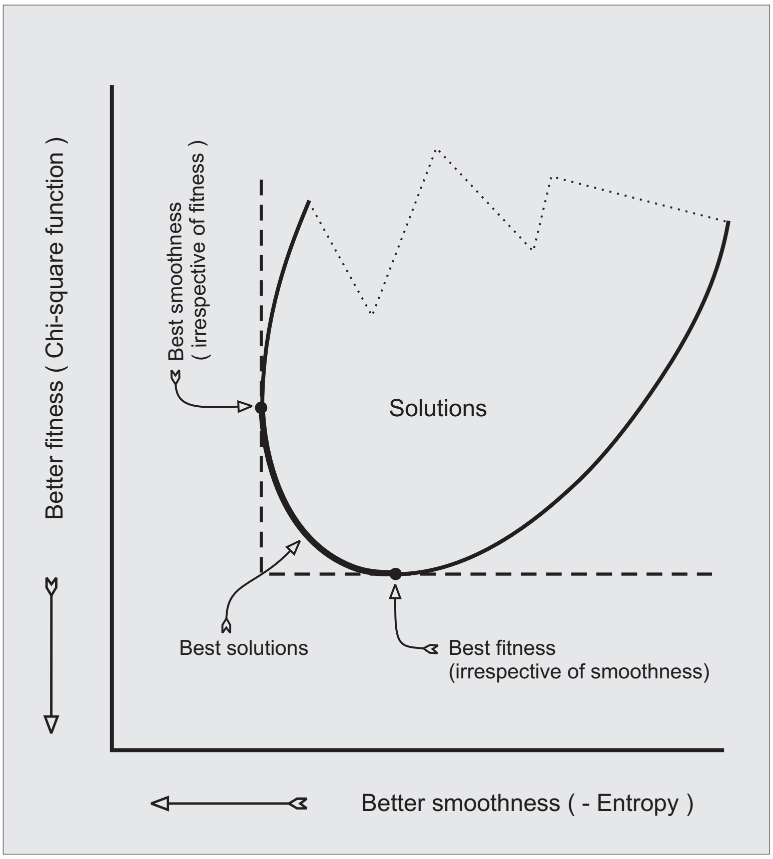
2.5. *Regularization of the inverse problem*

A set of admissible images $\{\boldsymbol{f}\}$ is convex if for any two elements, $\boldsymbol{f}_a$ and $\boldsymbol{f}_b$, all linearly interpolated combinations like $(1-\eta)\boldsymbol{f}_a+\eta\boldsymbol{f}_b$, $0\leq\eta\leq 1$, also belong to the set. Several deterministic constraints imposed on the solution $\widetilde{\boldsymbol{f}}$, such as positivity, compact support (i.e., zero value outside a specified region), or specified bounds (i.e., $\boldsymbol{f}_{min}\leq\boldsymbol{f}\leq\boldsymbol{f}_{max}$), define convex sets. Constraints can be formulated either in the image space or in the domain of any linear transform of the image $\boldsymbol{f}$.

Both the chi-square surface $\chi^2(\boldsymbol{f})=constant$ and the entropy surface $S(\boldsymbol{f})=constant$ are convex in the image space, so there is only one ME solution consistent with the data and the specified PSF. Consequently, the solution lies at the extremum of:

$$Q(\lambda,\boldsymbol{f})=\chi^2(\boldsymbol{f})-\lambda S(\boldsymbol{f}) \tag{41}$$

Figure 5. Trade-off between entropy maximization and chi-square minimization.


Better fitness (Chi-square function)
Best smoothness
(irrespective of fitness)
Solutions
Best solutions
Best fitness
(irrespective of smoothness)
Better smoothness (- Entropy)

The first functional, $\chi^2(\boldsymbol{f})$, measures the goodness or sharpness of fit between the model, $\boldsymbol{f}$, and the data, $\boldsymbol{g}$. By minimizing $\chi^2(\boldsymbol{f})$ alone, the agreement becomes unrealistically good, while the solution becomes unstable and oscillating. Thus, $\chi^2(\boldsymbol{f})$, by itself, defines a highly degenerate minimization problem. Reversely, $S(\boldsymbol{f})$ measures the smoothness of the solution with respect to the variations in data. By maximizing $S(\boldsymbol{f})$ alone results in a very smooth or stable solution, which is virtually not related with the measured data. Thus, $S(\boldsymbol{f})$, by itself, is termed the *stabilizing* functional or *regularizing* operator. While $\chi^2(\boldsymbol{f})$ is related with the *a posteriori* knowledge of a solution, $S(\boldsymbol{f})$ is related with *a priori* expectation. The adjustable parameter λ is setting the trade-off between the constraints imposed on the two functionals (Figure 5). In addition, λ is required by the necessity of invariance for the minimization principle to the units in which $\boldsymbol{f}$ is quantified (e.g., the change from 16-bit to 32-bit sampling) [26].

This is a typical problem that addresses Lagrange's method of undetermined multipliers. The expected output is to settle on a best value of λ based on some criterion, ranging from an objective statistical constraint for $\chi^2(\tilde{\boldsymbol{f}})$ to entirely subjective ones. As λ varies from 0 to ∞, the solution $\tilde{\boldsymbol{f}}$ varies along a trade-off curve between the task of minimizing $\chi^2(\boldsymbol{f})$ and the task of maximizing $S(\boldsymbol{f})$. For any linear combination of $\chi^2(\boldsymbol{f})$ with a nondegenerate quadratic form $S(\boldsymbol{f})$, such as $\chi^2(\boldsymbol{f}) - \lambda S(\boldsymbol{f})$, the minimization procedure yields a unique solution for $\boldsymbol{f}$. That is, when a quadratic minimization principle is combined with a quadratic constraint, and both are positive, only one of the two need to be nondegenerate for the overall problem to be well-posed [26].

2.6. Derivation of the potential function

The space invariant linear system $\boldsymbol{g} = \mathbf{R}\boldsymbol{f} + \boldsymbol{e}$ with additive and zero-mean Gaussian noise assumed uncorrelated between pixels is considered hereafter. The task is to restore an image in the form of a probability distribution $\boldsymbol{f} = \{p_n\}$, where $p_n = f_n/U$, f_n is the number of quanta in the n-th pixel, U is the total number of quanta in the image, given the measured "blurred" data g_m, $m = 1, 2, ..., M$ affected by the errors e_m, $m = 1, 2, ..., M$, and which complies with a finite set of constraints:

1. Conservation of the total number of photons in the measured image, $\boldsymbol{g}$, and the model image $\boldsymbol{f}$:
$$\sum_{n=1}^{N} p_n = 1 \tag{42}$$

2. Linear transform between the model space and the data space:
$$g_m = \sum_{n=1}^{N} U \cdot R_{mn} \cdot p_n + e_m,\ m = 1, 2, ..., M \tag{43}$$

3. The errors e_m, $m = 1, 2, ..., M$, are normally distributed with zero mean, $\overline{e} = 0$, and variances σ_m^2, $m = 1, 2, ..., M$:
$$\sum_{m=1}^{M} \frac{e_m^2}{\sigma_m^2} = \Omega \tag{44}$$
where Ω denotes the expected value of the statistical goodness-of-fit χ^2.

Generally, the χ^2 statistic associated with e_m, $m = 1, 2, ..., M$ independent Gaussian random variables:
$$\chi^2 = \sum_{m=1}^{M} \frac{(e_m - \overline{e})^2}{\sigma_m^2} \tag{45}$$
has the expectation $M \pm \sqrt{M}$. Since the mean is fitted by setting $\overline{e} = \overline{\sum_{m=1}^{M} e_m} = 0$ inferred by the conservation of the number of quanta spread over the pixels, one degree of freedom has already been used up. Consequently, the expectation of χ^2 is reduced by 1 and set $\Omega = M - 1$. Though this distinction is commonly ignored, the difference is essential in cases where the number of free parameters is similar to the number of data points like interpolation [29].

In all cases, the potential function, Z, is useful if reliable estimates of the standard errors σ_m^2, $m = 1, 2, ..., M$ that determine the severity of blurring are *a priori* selected on some theoretical and/or experimental base. Noise degrades the overall quality of an image and makes it difficult to reveal certain details of interest. The major source of noise in X-ray imaging is the random distribution of photons over the surface of the image. The standard deviation of the photon distribution is the best quantitative estimator of the noise in an image. As such, the standard deviation σ_m of the photon distribution in each pixel $m = 1, 2, ..., M$ was set in agreement with Poisson's law as proportional with the square root of the recorded quanta g_m, such as:

$$\sigma_m = \alpha \cdot \sqrt{g_m},\ m = 1, 2, ..., M \tag{46}$$

The model parameter α sets down the proper scattering range of the signal-to-noise ratio (SNR) in each pixel around some common level and ensures convergence of the algorithm after a reasonable number of iterations. Assuming that the errors e_m, $m = 1, 2, ..., M$ are normally distributed with zero mean and variances σ_m^2, $m = 1, 2, ..., M$, then the expected value of their χ^2 distribution should lie within the interval $\left[M - \sqrt{M}, M + \sqrt{M}\right]$. In a similar approach, Myrheim and Rue assumed as convenient but not essential, that the covariance matrix of the errors being proportional to the unit matrix such as $\mathbf{C} = \sigma^2\mathbf{I}$ [30].

Among all admissible probability distributions of the model image, $\boldsymbol{f} = \{p_n\}$, which satisfy the imposed constraints, eqs. (42), (43), and (44), the one that complies with the ME principle is the most unbiased choice to estimate the true image. Accordingly, maximizing the image entropy:

$$S(\boldsymbol{f}) = -\sum_{n=1}^{N} f_n \ln\left(\frac{f_n}{U}\right) = -U\sum_{n=1}^{N} p_n \ln p_n \tag{47}$$

leads to the smoothest and uniform distribution, $\boldsymbol{f}$, among the admissible set of images. The least committal Lagrangian associated with the objective function, $S(\boldsymbol{f})$, including the specific constraints, has the following explicit form:

$$L\left(p_1, p_2, ..., p_N, e_1, e_2, ..., e_M, \lambda_0, \lambda_1, ..., \lambda_M, \rho\right) = \\ -\sum_{n=1}^{N} Up_n \ln p_n - \lambda_0\left(-\sum_{n=1}^{N} p_n - 1\right) - \sum_{m=1}^{M}\lambda_m\left(-\sum_{n=1}^{N} UR_{mn}p_n + e_m - g_m\right) - \rho\left(\sum_{m=1}^{M}\frac{e_m^2}{\sigma_m^2} - \Omega\right) \tag{48}$$

where $\lambda_0, \lambda_1, ..., \lambda_M$, and ρ are the Lagrange multipliers. The unknown errors, $e_1, e_2, ..., e_M$, were included in the argument. Applying Lagrange's method amounts to setting $\partial L/\partial p_n = 0$, $n = 1, 2, ..., N$ and $\partial L/\partial e_m = 0$, $m = 1, 2, ..., M$, leading to the following set of $N + M$ equations, respectively:

$$\ln p_n = -1 - \frac{\lambda_0}{U} - \sum_{m=1}^{M}\lambda_m R_{mn},\ n = 1, 2, ..., N \tag{49}$$

$$e_m = -\lambda_m\frac{\sigma_m^2}{2\rho},\ m = 1, 2, ..., M \tag{50}$$

There are $N{+}2M{+}2$ Lagrangian arguments, $p_1, p_2, ..., p_N, e_1, e_2, ..., e_M, \lambda_0, \lambda_1, ..., \lambda_M$, and ρ, exactly matched by the number of equations available (42), (43), (44), (49), and (50). By solving eq. (44) for p_n, $n = 1, 2, ..., N$ and using the total flux constraint in eq. (42) to eliminate λ_0 , each probability p_n can be expressed as a function of $\lambda_1, \lambda_2, ..., \lambda_M$:

$$p_n(\lambda_1, \lambda_2, ..., \lambda_M) = \frac{\exp\left(-\sum_{m=1}^{M}\lambda_m R_{mn}\right)}{\sum_{n=1}^{N}\exp\left(-\sum_{m=1}^{M}\lambda_m R_{mn}\right)},\ n = 1, 2, ..., N \tag{51}$$

Substituting eqs. (50) and (51) into the remaining eqs. (43) and (44), a nonlinear system of $M + 1$ equations with $M + 1$ variables, $\lambda_1, \lambda_2, ..., \lambda_M$, and ρ is obtained:

$$\sum_{n=1}^{N} UR_{mn}p_n(\lambda_1, \lambda_2, ..., \lambda_M) - \lambda_m\frac{\sigma_m^2}{2\rho} - g_m = 0,\ m = 1, 2, ..., M \tag{52}$$

$$\frac{1}{4\rho^2}\sum_{m=1}^{M}\sigma_m^2\lambda_m^2-\Omega=0 \tag{53}$$

The left hand side of the nonlinear system in eq. (52) is the gradient of the potential function:

$$Z(\lambda_1,\lambda_2,...,\lambda_M,\rho)=-U\ln\left[\sum_{n=1}^{N}\exp\left(-\sum_{m=1}^{M}\lambda_m R_{mn}\right)\right]-\frac{1}{4\rho}\sum_{m=1}^{M}\lambda_m^2\sigma_m^2-\sum_{m=1}^{M}\lambda_m g_m+\rho\Omega \tag{54}$$

Hence the problem of solving the nonlinear system formed by eqs. (52) and (53) can be formulated in terms of finding the extremum of the potential function Z in the $(M+1)$-dimensional space of the Lagrange multipliers by solving:

$$\nabla Z(\lambda_1,\lambda_2,...,\lambda_M,\rho)=0 \tag{55}$$

2.7. *Multidimensional optimization*

In variational problems with linear constraints, Agmon *et al.* [31] showed that the potential function associated to a positive, additive image is always concave for any set of Lagrange multipliers, and it possesses an unique minimum which coincides with the solution of the nonlinear system of constraints. As a prerequisite, the linear independence of the constraints is checked and then the necessary and sufficient conditions for a feasible solution are formulated. Wilczek and Drapatz [32] suggested Newton-Raphson's iteration method as offering high accuracy results. Though at each step of the iteration the Jacobian of the system formed by eqs. (52) and (53) has to be evaluated, the computing requirements may be significantly reduced by exploiting its symmetric structure. Ortega and Rheinboldt [33] adopted a continuation approach for the very few cases where Newton's method failed to converge. Basically, the χ^2 function is considered as a parameter and the iteration starts with such of its value that eq. (53) to be (almost) satisfied; subsequently, it is gradually adjusted to reach its originally intended value Ω. In practice, the techniques enumerated above are successful only for relatively small data sets and require a symmetric positive definite Hessian matrix of the potential function.

Consider the potential function $Z:\mathbb{R}^{M+1}\longrightarrow\mathbb{R}$, $\boldsymbol{P}\in\mathbb{R}^{M+1}$, and assume that both $Z(\boldsymbol{P})$ and $\triangle Z(\boldsymbol{P})$ can be calculated. The $(M+1)$-dimensional space corresponds to the Lagrange multipliers $\lambda_1,\lambda_2,...,\lambda_M$, and ρ. As such, Z is minimized following a conjugate gradient approach derived from Fletcher-Reeves, Polak-Ribière [34], and Cornwell-Evans [35] algorithms. Methods that iteratively renew the direction of minimization not down the new gradient but rather along the lines constructed to be conjugated to the old gradient and possibly to all previously traversed directions, are termed *conjugate gradient methods*. One pass of N line minimizations lands precisely on the minimum of a quadratic form, whereas for not-exactly quadratic functions, repeated cycles of N line minimizations will gradually converge quadratically to the minimum [26]. The sequence of conjugate gradient-based directions $\{\boldsymbol{v}_i\}$ is constructed by means of line minimizations only, contrarily to the *variable metric methods* employed in most similar approaches that explicitly need to evaluate the Hessian matrix $\mathbf{A}$ at each iteration.

3. Results and Discussion

3.1. Physics of X-ray imaging

The production of X-ray images is largely based on the assumption that the X-ray photons are passing through the imaged object (tissue) along rectilinear paths. As such, the refractive index of the X-rays is implicitly assumed practically one, which is reasonably true in a first approximation only. The interactions remove some of the photons from the beam in a process known as attenuation. The *linear attenuation coefficient* μ is defined as the relative decrease of the number of photons N in the incident beam while crossing the tissue by the rate $\mathrm{d}N/N = -\mu\,\mathrm{d}x$. Linear attenuation coefficient values indicate the rate at which photons interact as they move through the tissue and are inversely related to the average distance photons travel before interacting. Furthermore, only a fraction of the impinging photons are assumed to be absorbed while crossing the imaged object. The removal of photons from the beam is supposed to occur only because of the interactions with the atoms of the object (actually with the electrons).

The total attenuation depends on the individual rates of all interactions that may occur while the photons are passing through a sample. In the case of a photon beam with energy less than 100 keV, the main interactions are the photoelectric effect, Compton (or incoherent) scattering, and Rayleigh (or coherent) scattering. The breast has very low physical contrast being composed of soft tissues: essentially a background of fat surrounded by slightly more dense glandular structures. Even though calcium is somewhat more dense and has a higher atomic number, typical breast calcifications are quite small and produce low physical contrast. As such, relatively low energy photons are employed in mammography, most of the interactions of the X-ray photons with tissue being photoelectric processes. The main limiting contrast factors between soft tissues and between soft tissue and fluids are the small differences between their physical characteristics (density and atomic number) and the relatively low number of photoelectric interactions because of the tissue low atomic numbers. The rate of photoelectric interactions generally decreases with increasing energy of the individual photons. Since dose and contrast decrease with increasing photon energy with different rates, it is possible to adjust the X-ray beam spectrum for optimized imaging with respect to contrast and dose. The quantity that affects attenuation rate is not the total mass of an object but the *area mass*, which is the amount of material under the unit surface area. Assuming a homogeneous imaged object, the attenuation is obtained by integration:

$$\frac{N(E)}{N_0} = -\exp[-\mu(E)]\mathrm{d}x = \exp\left[-\frac{\mu}{\rho}(E)\,\rho\,\mathrm{d}x\right] \tag{56}$$

where N_0 is the number of photons in the incident beam and the dependence on the X-ray energy E was explicitly displayed. The *mass attenuation coefficient*, μ/ρ, which is defined as the rate of photons per unit area mass, depends on the material composition and the energy of the incident photons. A simple empirical approximation of the mass absorption coefficient for low Z matrices (tissue) was considered in the present experiments:

$$\frac{\mu}{\rho}(E) = 20.64 \cdot E^{-3.28} \cdot \overline{Z}^{4.62} + \sigma_{KN}(E) \cdot \overline{Z} + 2.8 \cdot E^{-2.02} \cdot \overline{Z}^{4.86} \tag{57}$$

where $\sigma_{KN}(E)$ is the Klein-Nishina cross section and $\overline{Z}$ is the effective atomic number of the sample.

The basic unit of information in an imaging system is the detected photon. The physical image corresponds to the recorded number of photons impinging on some photosensitive device at specified locations. Photon emission is a Poissonian process so that the number of detected photons will be distributed according to Poisson statistics. Therefore, if in a unit time and per unit area an average number $\overline{n}$ of photons is recorded, the standard deviation of measurements will be $\sqrt{n}$. Such fluctuation in single measurements is intrinsic to the photon counting, so that one may speak of a *quantum* limited image.

An ideal imaging device can be thought as a spatial array of individual photon receptors and counters, each having identical properties. It is assumed that each receptor can receive light quanta (photons) independent of its neighboring receptors. The image state of a detector is completely determined by the number of quanta it receives and records, and each detector can be in one of a finite number of distinguishable image states. Since the possible number of states of a detector is finite, after a detector attains the final state, all other additional incident quanta will remain unrecorded, that is the detector gets saturated. In case of a digital image system, each pixel can be viewed as a receptor. The spatial resolution of the system depends on the spatial size of the pixel, whereas each pixel can have only a finite number of states until reaching its saturation level. The observed grey level of a pixel is the consequence of the received quanta by the corresponding receptor. Up to a certain level, the larger the number of the recorded quanta, the larger is the grey value.

The knowledge of the dose and the energy spectrum of the X-rays delivered to the subject during exposure allows in principle the computation of photon number per unit surface of the image system. Let us consider a sample to be imaged by means of a detector whose surface A is subdivided in identical square pixels of side d [42]. Let us denote by N_b the number of photons incident on a pixel looking at the ”background”, and N_d the number of photons incident on a pixel containing some texture to be imaged assuming that a percentage p of the original incident photons has been absorbed, that is $N_d = (1-p)N_b$. Further, define the *signal* as $N_b - N_d = pN_b$ and the noise as $\sqrt{N_b + N_d} = \sqrt{N_b(2-p)}$. Consequently, it appears natural to express the *contrast* C as:

$$C = \frac{N_b - N_d}{(N_b + N_d)/2} = \frac{2p}{2-p} \tag{58}$$

while the signal-to-noise ratio (SNR) to be given by:

$$SNR = \frac{pN_b}{\sqrt{N_b(2-p)}} \tag{59}$$

Hence the number of background photons required to reveal a certain target detail with a given contrast C is $N_b = 2(SNR)^2/pC$. If $p < 2$ then $C \approx p$ resulting in $N_b \approx 2(SNR/C)^2$. The result holds for an imaging system that is completely noise-free, i.e., the target is visible against a smooth background.

Image noise degrades the overall quality of an image making difficult to reveal certain details of interest. The major source of noise in X-ray and nuclear imaging is the random distribution of photons over the surface of the image. The amount of noise, or variation in photon concentration from area to area (e.g., from one detector cell to another) is inversely related to the number of photons used to form the image. Therefore, by increasing the total number of photons, the noise of an image can be reduced. It is important that the SNR to be large enough so that in a set of pixels the probability of any pixel giving accidentally a false signal to be negligible. In the case of $10^5 \div 10^6$ pixels, the condition is accomplished by setting $SNR = 5$ [36]. If N photons are incident on the surface A of the detector:

$$\frac{N}{N_b} = \frac{A}{d^2} \Rightarrow C = \frac{\sqrt{2}(SNR)}{d\sqrt{N/A}} \tag{60}$$

Inserting typical numerical values in eq. (60), such as $N/A = 4 \times 10^6$ photons/mm^2, $SNR = 5$, and the pixel side of the square detector $d = 0.2$ mm, leads to $C = 1.7 \times 10^{-2}$. Such a contrast value is extremely low since the practical rule is currently setting the minimum contrast perceivable by the human eye to be of about 2.5% over a smooth background. From the statistical point of view, the contrast value obtained according to eq. (60) entails that the number of photons per unit surface employed in a conventional radiological examination is quite excessive. Consequently, a detector with near-unity efficiency and capable of single photon counting would require less photons for the same quality of radiological images and, therefore, a lower radiation dose for the subject as well.

Table 1. Average recorded quanta around the incident point of the X-ray beam as resulted from simulations using GEANT for the most common X-ray energies.

X-ray beam energy (keV)	Central recorded photons	Total recorded photons
18	15,660 ± 125	18,390 ± 135
20	59,130 ± 240	70,900 ± 270
22	184,070 ± 430	226,290 ± 475
24	337,400 ± 580	424,340 ± 650
26	531,130 ± 730	683,330 ± 825
28	738,230 ± 860	976,540 ± 990
30	890,710 ± 940	1,204,780 ± 1100

The program GEANT4 [39], a toolkit originally designed for high energy physics experiments, was employed to simulate the passage and scattering of X-rays through various breast tissues in order to define the imaging system response. The SNR in the case of soft tissues like normal breast is a function of energy; Monte-Carlo simulations [37] indicated a maximum of the SNR in the range of 20 to 26 keV. Inhomogeneous breast-like tissue of 5 cm thickness with calcification inclusions and various opacities were considered in simulations, whereas the X-ray energy covered the range from 18 to 30 keV in steps of 2 keV [38]. The parameters used throughout in simulations were selected close to the corresponding values encountered in common radiological practice and the characteristics of our incoming silicon-based receptor: (i) tissue thickness of 5 cm (homogeneous and inhomogeneous); (ii) total number of events for each run: 10 millions; (iii) Dirac-distributed X-ray beam shape with orthogonal and central incidence; (iv) receptor area of 400×100 pixels, each square detector of 0.2 mm $\times 0.2$ mm; (v) recorded data for an area of 31×31 pixels around the incident point of the X-ray beam (Table 1).

The simulations of scattering at 24 keV suggested a symmetric space-invariant 5×5 pixel PSF matrix. The indexes of the matrix elements of the receptor are defined in Table 2 and blurring area produced by a simulated Dirac distribution of the X-ray beam centered on the *k*-th pixel is presented in Table 3. The convergence of the image restoration algorithm was ensured by simulations of various inhomogeneities embedded in breast-like tissue and setting the value of the model parameter, α, in such a way as to achieve the best average reproduction of the simulated opacities in terms of the mean energy of the reconstruction error [40]. The optimal experimental values of the model parameter were found to be within the range of $\alpha = 10^{-5} \div 10^{-4}$.

Table 2. Indexes of the receptor matrix elements: *ysize* stands for rows and *xsize* for columns.

Columns **Rows**	**(0)**	**(1)**	.....	**(xsize-2)**	**(xsize-1)**
(0)	0	1	.....	xsize-2	xsize-1
(1)	xsize	xsize+1	.....	2*xsize-2	2*xsize-1
(2)	2*xsize	2*xsize+1	.....	3*xsize-2	3*xsize-1
:	.....	.....	.....	.....	......
(ysize-2)	(ysize-2)*xsize	(ysize-2)*xsize+1	.....	(ysize-1)*xsize-2	(ysize-1)*xsize-1
(ysize-1)	(ysize-1)*xsize	(ysize-1)*xsize+1	.....	ysize*xsize-2	ysize*xsize-1

As for instance, an opacity of 2×2 pixel area embedded in a 31×31 pixel area of simulated inhomogeneous tissue is shown before and after processing (Figure 6). In addition, a test pattern designed to disclose any artifacts eventually introduced by entropic processing was designed. The pattern consisted of marble grains of various sizes embedded in organic materials with mammalian-like consistency. Neither spurious patterns nor any suspicious forms were detected (Figure 7) for the range of X-ray energy under study. These observations were sustained in a similar approach, by employing more elaborated phantoms in conjunction with MEM image processing to deblur mammographic images [41]. Simulations concluded that an X-ray beam is scattered around 20% at worst and about 98% of quanta are recorded on a 5×5 pixel area centered on its incident direction. Outside this area, the number of recorded quanta is of the order or less than the quantum noise in the central pixel of beam incidence. The digital and digitized mammograms under study were deconvolved in the spatial domain using the PSF as defined in Table 3, rather than running the discrete Fourier transform in the spectral domain. As such, errors introduced by truncation and aliasing were circumvented at the expense of some extra computational power.

Now assume that a detector has 100% efficiency and only the quantum noise is considered. Then eq. (60) holds for the smallest detail to be imaged down to the dimension of one pixel. For objects which are different from the pixel size, a more complex equation must be used:

$$C_m = \frac{1}{\sqrt{N/A}} \cdot \frac{k(a)}{\sqrt{a}} \tag{61}$$

where $k(a)$ is a rather complex function of the object area, and C_m is the minimum contrast which can be detected using a given number of photons per unit area and which could ideally be achieved with a 100% efficient noiseless detector [43]. Both of the above expressions, eqs. (60) and (61), show that contrast varies inversely with the square root of the superficial dose (which is proportional to the number of photons per unit surface impinging on the detector). Qualitatively, if a small object is to be detected, then it must have a large contrast, whereas large objects may be imaged with a relatively smaller contrast. However, practical detectors (with their electronic parts included) have an efficiency less than 100%, thereby reducing N/A and adding noise to the photon counters. If R is the ratio between the total noise (including the detector noise) and the intrinsic quantum noise, than the minimum detectable contrast becomes:

$$C = \frac{R}{\sqrt{\epsilon}} \cdot C_m \tag{62}$$

where ϵ is the detector efficiency. As eq. (62) indicates, a poor efficiency (small values of ϵ) as well as a noisy detector (small values of R) entail a minimum detectable contrast higher than the ideal case C_m and, consequently, degrades the performance of the detection system.

Table 3. PSF of the imaging system defined as a symmetrically space-invariant 5×5 matrix around the k-th pixel in the blurred image resulting from simulations of average soft breast-like tissue and typical hard calcification opacities for X-ray beam of 24 keV.

0.0030 (_k-2*xsize-2_)	0.0050 (_k-2*xsize-1_)	0.0070 (_k-2*xsize_)	0.0050 (_k-2*xsize+1_)	0.0030 (_k-2*xsize+2_)
0.0050 (_k-xsize-2_)	0.0120 (_k-xsize-1_)	0.0160 (_k-xsize_)	0.0120 (_k-xsize+1_)	0.0050 (_k-xsize+2_)
0.0070 (_k-2_)	0.0160 (_k-1_)	**0.7950** (**_k_**)	0.0160 (_k+1_)	0.0070 (_k+2_)
0.0050 (_k+xsize-2_)	0.0120 (_k+xsize-1_)	0.0160 (_k+xsize_)	0.0120 (_k+xsize+1_)	0.0050 (_k+xsize+2_)
0.0030 (_k+2*xsize-2_)	0.0050 (_k+2*xsize-1_)	0.0070 (_k+2*xsize_)	0.0050 (_k+2*xsize+1_)	0.0030 (_k+2*xsize+2_)

Several reports have pointed out the superiority of digital detectors over conventional films in three directions, at least. A valuable characteristic of most digital receptors is a constant sensitivity over a much wider range of exposures (latitude) than radiographic films. Secondly, the phantoms routinely used as test objects in mammography are already visible at a fluence with 3 order of magnitude lower by the digital recorders [42]. An increase in fluence corresponds to an increase in image contrast, which amounts to an increase in object visibility. Thirdly, once a digital detector came out with an image as a set of numerical data, the grey scale can be adapted as to enhance the contrast, and even more, tones of greys can be substituted by false colors.

Figure 6. Artificial-generated opacity of 2 × 2 pixel size (left) embedded in a simulated background breast-like tissue of 31 × 31 pixel area (middle) and subsequently restored by the entropic restoration algorithm.

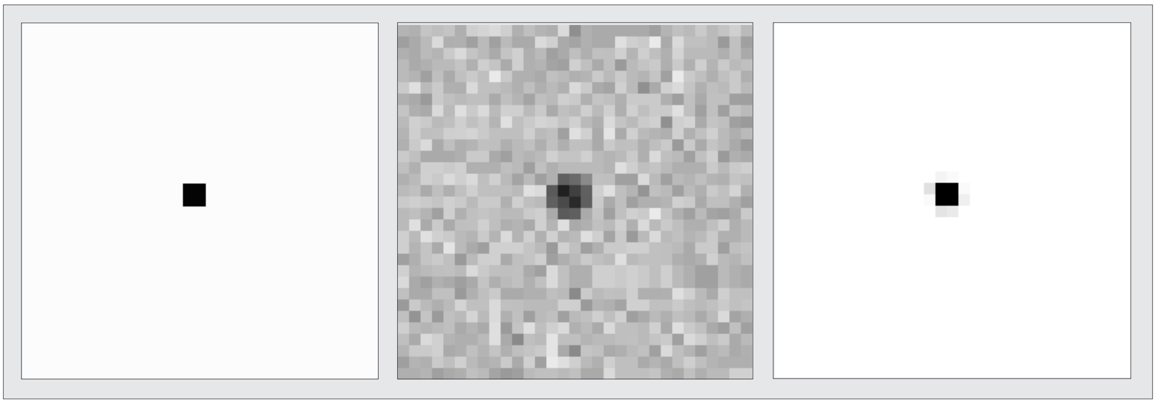
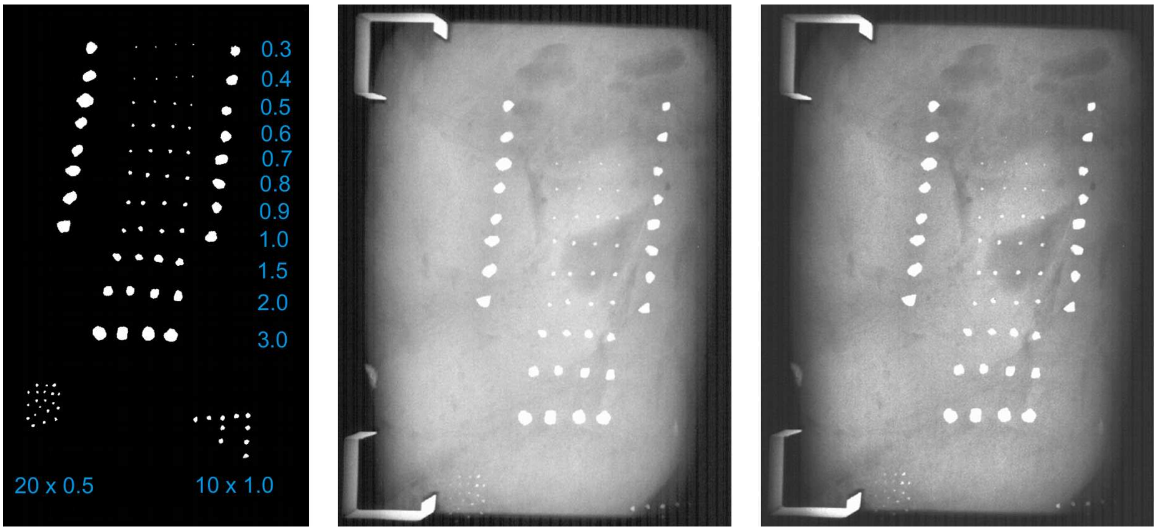


Figure 7. Arrangement of marble grains (left), raw radiography (middle), and processed radiography by the entropic algorithm (right). Average dimension of individual and cluster grains are specified in millimeters.



3.2. X-ray image quality assessment

The evaluation of the X-ray images was based on contrast-detail (CD) phantoms as fostered by the Department of Radiology, University Hospital Nijmegen [44]. The CDRAD phantoms consist of square plexiglas tablets of 265×265 mm with thickness of 10 mm. Each tablet contains cylindrical holes of exact diameter and depth with tolerances less than 0.02 mm arranged in 15×15 columns×rows small squares. First, CDRAD 2.0 phantoms were employed to plotting a CD-curve based on the threshold contrast as a function of object diameter. A better image should reveal smaller details or quite the same details at lower contrasts. All processed phantom images displayed a shift of the CD-curve to the lower left part of the phantom relatively to the CD-curve of the input images. Qualitatively and quantitatively, the meaning was a better overall visibility of details. Secondly, the image quality was expressed in a figure by the calculation of the ratio of the correctly identified hole-positions to the total number of the phantom squares. The correct observation ratios were higher for the processed images with $5 \div 12\%$. Thirdly, the CDRAD 2.0 phantoms were employed to quantify the visibility of details at various low-contrast values.

To estimate the restoration quality in case of $M = N$, the mean energy of reconstruction error:

$$D = \frac{1}{N} \sum_{n=1}^{N} \left(g_n - \widetilde{f}_n \right)^2 \tag{63}$$

where $\widetilde{\boldsymbol{f}}$ is the best statistical estimation of the correct solution $\boldsymbol{f}$, can be used as a factor of merit of the reconstruction algorithm. Yet too high a value for D may put the restored image too far away from the original one and raise questions about introducing spurious features for which there is no clear evidence in the acquired data.

A more realistic degradation measure of image blurring by additive noise can be formulated in terms of blurred signal-to-noise ratio (*BSNR*) metric redefined here by using the noise variance, σ, in each pixel:

$$BSNR = 10 \log_{10} \frac{1}{N} \sum_{n=1}^{N} \frac{(y_n - \overline{y}_n)^2}{\sigma^2} \tag{64}$$

where $\boldsymbol{y} = \boldsymbol{g} - \boldsymbol{e}$ is the difference between the measured data, $\boldsymbol{g}$, and the noise (error), $\boldsymbol{e}$.

In simulations, where the original image, $\boldsymbol{f}$, of the measured data $\boldsymbol{g}$ is available, the quality of image restoration algorithms may be assessed by the improvement in signal-to-noise metric:

$$ISNR = 10 \log_{10} \frac{\sum_{n=1}^{N} (f_n - g_n)^2}{\sum_{n=1}^{N} (f_n - \widetilde{f}_n)^2} \tag{65}$$

While mean squared error metrics like *ISNR* do not always reflect the perceptual properties of the human visual system, they may provide an objective standard in comparing different image processing techniques. It is nevertheless of crucial significance that various algorithms behavior should be analyzed from the point of view of ringing and noise amplification, which can be a key indicator of improvement in quality for subjective comparisons of restoration protocols [45].

3.3. Improvement of digital and digitized mammograms

There are two basic ways to create images with X-ray radiation, namely, projection imaging, that is to pass an X-ray through the sample section and project a shadow image over the receptor, and computed tomography (CT), that is, image reconstruction by computer from X-ray penetration data [46]. The focus here is on projection imaging as the basic process employed in common X-ray radiography and fluoroscopy practice. In the field of X-ray medical image processing, image restoration has been mostly used for filtering Poisson distributed film-grain noise in breast X-rays, mammograms, and digital angiographic images [45], removal of the additive noise in Magnetic Resonance Imaging (MRI) [47], and in the area of quantitative autoradiology (QAR) [48].

Entropic processing was carried out on optically scanned X-ray mammograms, as well as on digitized samples from mammography databases available on the Internet. Input images were full breast scanned mammograms (Figure 8), and details of clinical interest (Figure 9). The processed images were evaluated by expert radiologists from ”Fundeni” Hospital in Bucharest, Romania, and Casa di Cura ”Salus” and the Cancer Center, University of Trieste, Italy. Output images displayed overall superior quality as compared with their input counterparts in terms of contrast and visibility. The processed images restored existing blurred details in images that could not have been recovered by conventional analysis and facilitated their clinic interpretation by expert radiologists.

Benefits of digital mammography in early breast cancer detection include: (i) reduced radiation dose, (ii) post-acquisition image enhancement, (iii) rapid display of image, (iv) improved imaging in dense breasts, (v) improved sensitivity and specificity, (vi) simplified archival, transmission, and retrieval, (vii) potential usage in automatic diagnosis, and (viii) direct access to worldwide experts by telemammography. Greater benefits are to come due to the increasing incidence of breast cancer with age. Generally, it is easier to detect lesions in the older breast due to fatty involutional changes. X-raying the breast does present unique technical challenges. The contrast resolution of the image must be sufficient to discriminate between soft tissue and tumor. The spatial resolution must be adequate to distinguish fine calcifications well under 1 mm in size.

Figure 8. Full breast mammograms: Fundeni Hospital - normal breast (left), case mdb032, MIAS - benign ill-defined masses, fatty-glandular tissue (middle), case mdb005, MIAS - benign circumscribed masses - fatty tissue (right). Digitized X-ray images by optical scanning (top) and digitally restored images (bottom).

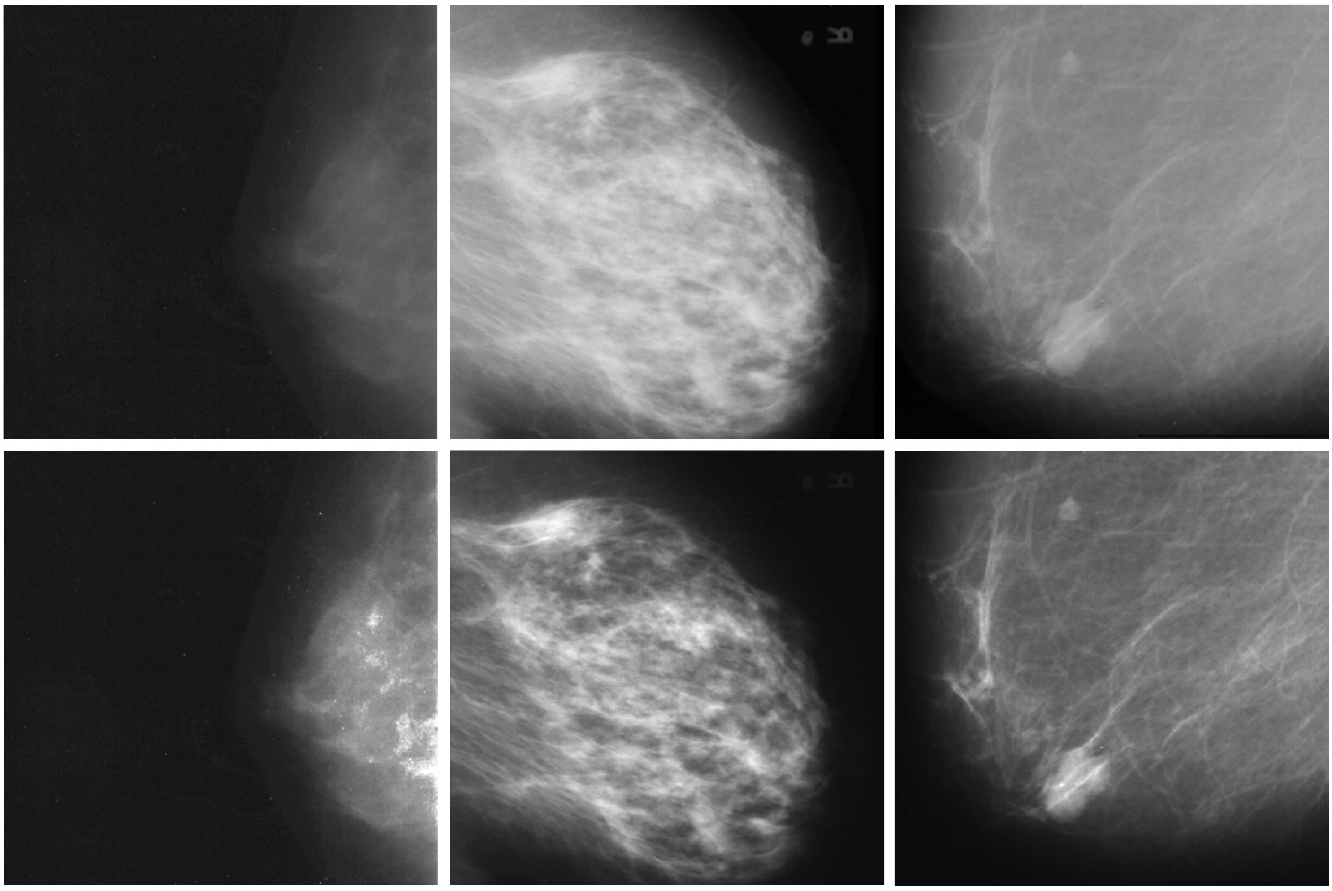
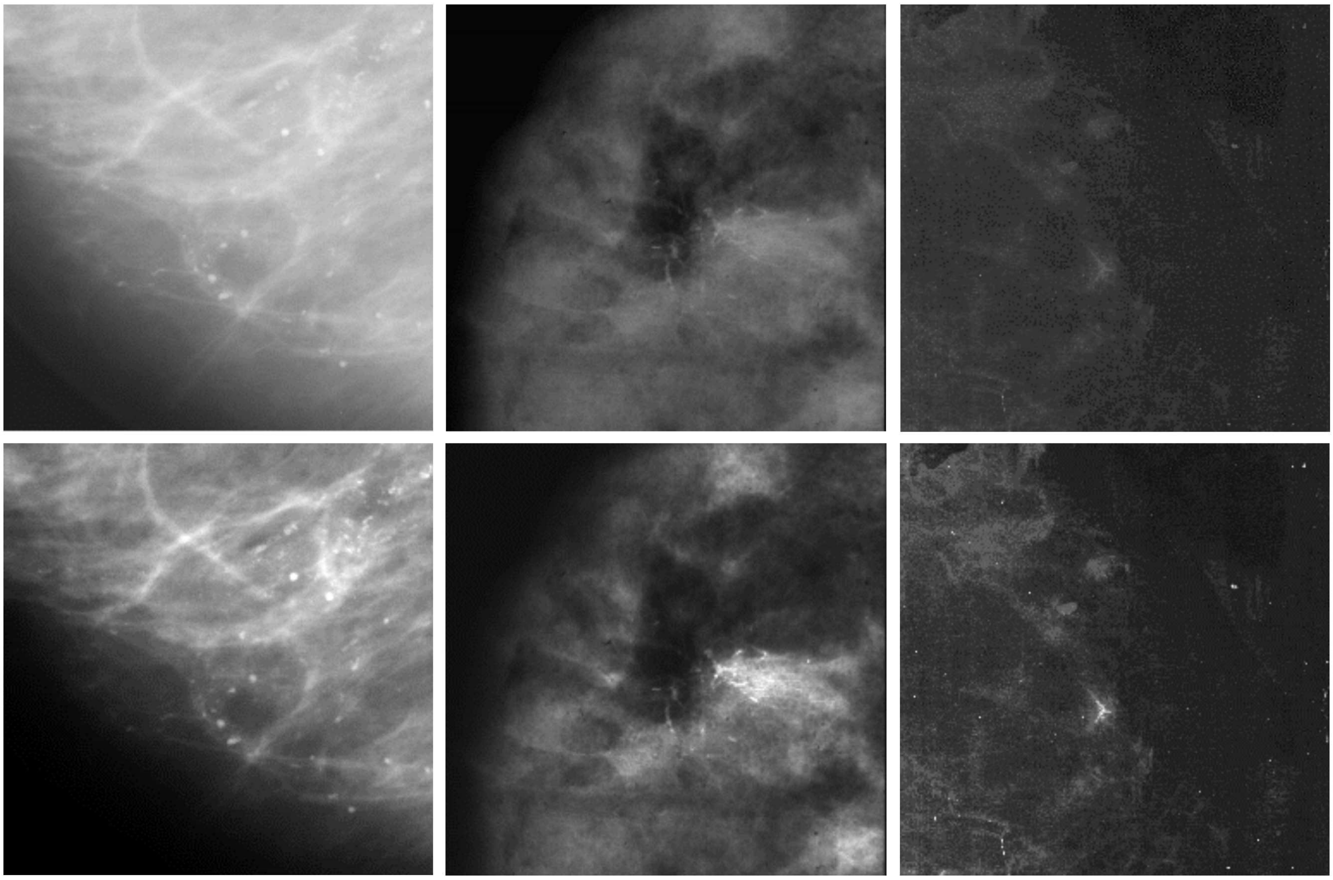


Figure 9. Close-up of malignant abnormalities in mammograms: Case mdb245, MIAS - microcalcification cluster, fatty tissue (left), case #9, UNC Radiology - breast carcinoma (middle), Salus Hospital - stellate patterns (right). Digitized X-ray images by optical scanning (top) and digitally restored images (bottom).



4. Conclusion

An intrinsic difficulty in Bayesian image restoration resides in determination of a prior law for images. The ME principle solves this problem and enforces the restored image to be positive, so that the spurious negative areas and complementary spurious positive areas are wiped off and the dynamic range of the restored image is substantially enhanced. Image restoration based on image entropy is effective even in the presence of significant noise, missing, or corrupted data. This is due to the appropriate regularization of the inverse problem of image restoration introduced in a coherent way by the ME principle. It satisfies all consistency requirements when combining the prior knowledge and the information contained in experimental data.

Bayesian ME approach is a statistical method which directly operates in the spatial domain, thus eliminating the inherent errors coming out from numerical Fourier direct and inverse transforms and from the truncation of signals in the transform domains. Theoretically, no artifacts should be introduced by processing, since the entropy maximization produces the most unbiased and featureless solution that is consistent with available data and complies with the errors in measurements and modeling.

In recent years, imaging equipment and techniques have significantly been improved that entailed the acceptance of mammography if performed with excellent positioning, firm compression of the breast, and dedicated processing facilities. The challenge today is to disclose early stages of breast cancer in the 35 to 50 age groups. Manufacturers exhibit an increasingly interest in diagnosing cancer in dense breast, which has substantial amounts of fibroglanduar tissue. Pattern recognition and classification methods will further be developed for computer assisted diagnosis ending up with expert systems that may perform all tasks required by digital biomedical data analysis and interpretation.

Acknowledgements

The present research was carried out in the Microprocessor Laboratory of the International Centre for Theoretical Physics, Trieste, Italy, within the framework of the X-ray ICTP-INFN-UNU Co-operation Project 95A2MC02C, which is gratefully acknowledged.

References and Notes

1. Claridge, E.; Richter, J.H. Characterisation of mammographic lesions. In *Digital Mammography*; Gale, A.G., Astley, S.M., Dance, D.R., Alistair, A.Y., Eds.; Elsevier: Amsterdam, Netherlands, 1994.
2. Meyer, Y. An introduction to wavelets and ten lectures on wavelets. *Bull. Amer. Math. Soc.* **1993**, *28*, 350-359.
3. Mallat, S. *Une Exploration des Signaux en Ondelettes*; Les Editions de l'Ecole Polytechnique, 2000.
4. Donoho, D.L. Unconditional bases and bit-level compression. *Applied and Computational Harmonic Analysis* **1996**, *1*, 100-105.
5. Candes, E.J. Ridgelets: Theory and applications. *PhD Thesis*, Department of Statistics, Standford University, CA, USA, 1998.
6. Mutihac, R.; Cicuttin,; Jansen, K.; Mutihac, R.C. An essay on Bayesian inference and maximum entropy. *Roumanian Biotechnological Letters* **2000**, *5*, 83-114.
7. Shore, J.E.; Johnson, R.W. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory* **1980**, *IT-26*, 26-39 and *IT-29*, 942-943.
8. Skilling, J. The axioms of maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering*; Erickson G.J., Smith C.R., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1988; Vol. I, pp. 173-187.
9. Bayes, T. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc. London* **1763**, *53*, 330-418.
10. Jaynes, E.T. Papers on Probability, Statistics and Statistical Physics. Rosenkrantz R.D., Ed.; Kluwer Academic Press: Dordrecht, Netherlands, 1983.
11. Skilling, J. Fundamentals of MaxEnt in data analysis. In *Maximum Entropy in Action*; Buck B., Macaulay V.A., Eds.; Clarendon Press: Oxford, UK, 1994; pp. 19-39.
12. MacKay, D.J.K. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.

13. Berger, J. *Statistical Decision Theory and Bayesian Analysis*; Springer-Verlag: New York, NY, USA, 1985.
14. Hadamard, J. *Lectures on the Cauchy Problem in Linear Partial Differential Equations*; Yale University Press: Yale, CT, USA, 1923.
15. Engle, H.; Hanke, M.; Neubauer, A. Regularization of inverse problems. In *Series - Mathematics And Its Applications*; *375*, Springer-Verlag: New York, NY, USA, 1996.
16. Cox, R. Probability, frequency, and reasonable expectation. *Am. J. Phys.* **1946**, *14*, 1-13.
17. Djafari, A.M. Maximum entropy and linear inverse problems. In *Maximum entropy and Bayesian methods.*; Djafari, A.M., Demoments, G., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1993; pp. 253-264.
18. Gull, S.F.; Daniell, G.J. Image reconstruction from incomplete and noisy data. *Nature* **1978**, *272*, 686-690.
19. Weiss, N.A. *Introductory Statistics*; Addison-Wesley: Boston, MA, USA, 2002.
20. MacKay, D.J.K. A practical Bayesian framework for backpropagation networks. *Neural Comput.* **1992**, *4*, 448-472.
21. Balasubramanian, V. Occam's razor for parametric families and priors on the space of distributions. In *Maximum Entropy and Bayesian Methods*, Proceedings of the 15th International Workshop, Santa Fe, 1995; Hanson, K.M., Silver, R.N., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1996; pp. 277-284.
22. Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **1997**, *9*, 2, 349-368.
23. Shannon, C.E. A mathematical theory of communication. *Bell. Syst. Tech. J.* **1948**, *27*, 379-423.
24. Pal N.R.; Pal S.K. Entropy: A new definition and its applications. *IEEE Trans. Syst., Man, Cybern.* **1991**, *21*, 1260-1270.
25. Skilling, J. Classic maximum entropy. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Kluwer Academic: Norwell, MA, USA, 1989; pp. 45-52.
26. Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T. *Numerical Recipes in C: The Art of Scientific Computing*. 3rd ed.; Cambridge University Press: Cambridge, UK, 2007.
27. Djafari, A.M. A full Bayesian approach for inverse problems. In *Maximum Entropy and Bayesian Methods*; Hanson, K.M., Silver R.N., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1995, pp. 135-144.
28. Hanson, M.K. Making binary decisions based on the posterior probability distribution associated with tomographic reconstructions. In *Maximum Entropy and Bayesian Methods*; Smith, C.R., Erickson, G.J., Neudorfer, P.O., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1992; pp. 313-326.
29. MacKay, D.J.C. (1992), Bayesian interpolation. In *Maximum Entropy and Bayesian Methods*; Smith, C.R., Erickson, G.J., Neudorfer, P.O., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1992.
30. Myrheim, J., Rue, H. New algorithms for maximum entropy image restoration. *Graphical Models and Image Processing Archive* **1992**, *54*, 223-238.
31. Agmon, N.; Alhassid, Y.; Levine, R.D. An algorithm for finding the distribution of maximal en-

tropy. *J. Comput. Phys.* **1979**, *30*, 250-258.

32. Wilczek, R.; Drapatz, S. A high accuracy algorithm for maximum entropy image restoration in the case of small data sets. *Astron. Astrophys.* **1985**, *142*, 9-12.
33. Ortega, J.M.; Rheinboldt, W.B. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press: New York, NY, USA, 1970.
34. Brodlie, K.W. Unconstrained minimization. In *The State of the Art in Numerical Analysis*; Jacobs, D.A.H., Ed.; Academic Press: London, UK, 1977; pp. 229-268.
35. Cornwell, T.J.; Evans, K.J. A simple maximum entropy deconvolution algorithm. *Astron. Astrophys.* **1985**, *143*, 77-83.
36. Evans, A.L. The evaluation of medical images. In *Medical Physics Handbooks*; Adam Hilger: Bristol, UK, 1981; Volume 10, pp. 45-46.
37. Benini, L. *et al.* Synchrotron radiation application to digital mammography. A proposal for the Trieste Project ”Elettra”. *Phys. Med.* **1990**, *VI*, 293.
38. Mutihac, R.; Colavita, A.A.; Cicuttin, A.; Cerdeira, A.E. Maximum entropy improvement of X-ray digital mammograms. In *Digital Mammography*; Karssemeijer, N., Thijssen, M., Hendriks, J., van Erning, L., Eds.; Kluwer Academic Publishers: Dordrecht, Netherlands, 1998; pp. 329-337.
39. Allison, J.; Amako, K.; Apostolakis, J.; Araujo, H.; etc. Geant4 developments and applications. *IEEE T. Nucl. Sci.* **2006**, *53*, 270-278.
40. Mutihac, R.; Colavita, A.A.; Cicuttin, A.; Cerdeira, A.E. X-Ray image improvement by maximum entropy. In Proceedings of the 13th IEEE & EURASIP International Conference on Digital Signal Processing, Santorini, Greece, 1997; Vol. II, pp. 1149-1152.
41. Jannetta, A.; Jackson, J.C.; Kotre, C.J.; Birch, I.P.; Robson, K.J.; Padgett, R. Mammographic image restoration using maximum entropy deconvolution. *Phys. Med. Biol.* **2004**, *49*, 4997-5010.
42. Arfelli, F. Silicon detectors for synchrotron radiation digital mammography. *Nucl. Instrum. Meth.* **1995**, *A 360*, 283-286.
43. Di Michiel, M. Un rivelatore di silicio a pixel per immagini in radiologia diagnostica. *PhD Thesis* (unpublished); Universita di Trieste, June, 1994.
44. Thijssen, M.A.O.; Bijkerk, K.R.; van der Burght, R.J.M. *Manual CDRAD-phantom type 2.0.* Department of Radiology, University Hospital Nijmegen, The Netherlands, 1988-1992.
45. Banham, M.R.; Katsaggelos, A.K. Digital image restoration. *IEEE Signal Proc. Mag.* **1997**, *3*, 24-41.
46. Sprawls, P., Jr. *Physical Principles of Medical Imaging*; Medical Physics Publishing: Madison, Wisconsin, USA, 1995; 2nd ed.; Ch. 12, pp. 171-172.
47. Zadeh, H.S-.; Windham, J.P.; Yagle, A.E. A multidimensional nonlinear edge-preserving for magnetic resonance image restoration. *IEEE T. Image Process.* **1995**, *4*, 141-161.
48. Goyette, J.A.; Lapin, G.D.; Kang, M.G.; Katsaggelos, A.K. Improving autoradiograph resolution using image restoration techniques. *IEEE Eng. Med. Biol.* **1994**, *8-9*, 571-574.