united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international
centre
for theoretical
physics

preprint

TOPICS IN BAYESIAN STATISTICS AND MAXIMUM ENTROPY

R. Mutihac

C. Stănciulescu

A. Cicuttin

and

A. Cerdeira

30 - 06

IC/98/137

United Nations Educational Scientific and Cultural Organization
and
International Atomic Energy Agency

THE ABDUS SALAM INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS

TOPICS IN BAYESIAN STATISTICS AND MAXIMUM ENTROPY

R. Mutihac
ICTP-INFN Microprocessor Laboratory,
The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy,

C. Stănciulescu
University of Bucharest, Faculty of Physics, PO Box MG- H, Bucharest-Magurele, Romania

A. Cicuttin and A. Cerdeira
ICTP-INFN Microprocessor Laboratory,
The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy.

Abstract

Notions of Bayesian decision theory and maximum entropy methods are reviewed with particular emphasis on probabilistic inference and Bayesian modeling, The axiomatic approach is considered as the best justification of Bayesian analysis and maximum entropy principle applied in natural sciences. Particular emphasis is put on solving the inverse problem in digital image restoration and Bayesian modeling of neural networks. Further topics addressed briefly include language modeling, neutron scattering, multiuser detection and channel equalization in digital communications, genetic information, and Bayesian court decision-making.

MIRAMARE – TRIESTE
December 1998

I. Introduction

Probabilistic inference is an effective approach to *approximate reasoning* and empirical *learning* in artificial intelligence (AI) environments. Bayesian statistics and Maximum Entropy (ME) principle have been widely used for inferential problems in which corrupted or incomplete data were only available [Gull and Skilling 1984]. This paper highlights the potential power of entropic priors in Bayesian inference and briefly reviews some major fields of application.

Bayesian probability theory provides a unifying framework for data modeling that offers several benefits over classical approach [MacKay, 1998]. First, probabilistic modeling handles uncertainty in a natural and self-consistent manner. There is a unique prescription, namely *marginalization*, for incorporating uncertainty about parameters into predictions that yields better predictions. Secondly, the overfitting problem can be solved by using *Ockham's razor* to control the model complexity. Thirdly, Bayesian inference insures optimality in handling the prior knowledge. Any new piece of knowledge may be added in any stage of an experiment and more complex models may be developed that are capable of extracting more information from the data. Fourthly, Bayesian reasoning satisfies the likelihood principle [Berger 1985] in the sense that our inferences depend only on the probabilities assigned to the data that were received and not on the properties of other admissible data that did not actually occur.

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. Yet Bayesian inferences are *objective* since following the same assumptions on a model, or hypotheses, then identical inferences are drawn. Bayesian methods were simultaneously conceived by Bayes [1763] and subsequently enriched by Laplace [1814], then developed by Sir Harold Jeffreys [1939]. The logical basis for the Bayesian use of probabilities as measures of plausibility was subsequently established by Cox [1946], who proved that consistent inference in a closed hypothesis space can be mapped onto probabilities.

II. BAYESIAN METHODS

2.1 BAYES' THEOREM

The fundamental concept of Bayesian analysis is that the plausibility of alternative hypotheses are represented by probabilities and inference is carried out by evaluating these probabilities. Assume that we have a collection of admissible *models*, or *hypotheses* $\{H_1, H_2, ..., H_N\}$, competing to recount some data we collect. Our initial beliefs about the relative plausibility of these models are quantified by a list of probabilities

$p(H_1), p(H_2),..., p(H_N)$ that sum to *1*. Each model H_k makes predictions about how likely different data sets *D* are, if the given model H_k were true. These predictions are described by a probability distribution $p(D \mid H_k)$, which is actually the plausibility of *D* given the information H_k. When we observe the actual data *D*, Bayes' theorem describes the way of updating our beliefs in the model according to the newly acquired information *D*. If we denote by $p(H_k \mid D)$ the plausibility of the model H_k, given that we observed *D*, then

$$p(H_k \mid D) = \frac{p(H_k) \cdot p(D \mid H_k)}{p(D)} \tag{2.1}$$

where the *prior* probability $p(H_k)$ describes how plausible we considered H_k *before* acquiring the data *D*, the *likelihood* $p(D \mid H_k)$ expresses how much the model H_k *predicted* the data, and the *evidence* $p(D)$ is a normalizing factor that makes our final beliefs $p(H_k \mid D)$ add up to *1*. Bayes' theorem makes no explicit reference to any *sample* space or *hypothesis* space, and it can't determine the numerical value of any probability directly from our information either. A single application of Bayes' theorem gives us only a probability, not a probability distribution. The *likelihood* $p(D \mid H_k)$ expresses our confidence in the data *D* given that our hypothesis H_k is true. The *prior* probability $p(H_k)$ of H_k gets updated to the posterior probability $p(H_k \mid D)$ as a result of acquiring the data *D*. This includes parameter estimation, since H_k might be a model defined by a set of parameters, say w. Bayes' theorem explicitly shows that the measurements that give us $p(D \mid H_k)$ do not fully define our result $p(H_k \mid D)$ as far as we also need to assign the prior $p(H_k)$ to our model H_k. *Forward* probability uses probabilities and priors to evaluate the typical performance of a modeling method averaged over different data sets from a defined ensemble [Tishby 1989], without involving Bayes' theorem. Contrarily, Bayesian approach uses inverse probability to evaluate the relative plausibility of several alternative models in the light of a single data set that we actually observed.

However, Bayes' theorem follows directly from the standard axioms of probability relating the conditional probabilities of two events, say H_k and D, since we evidently have

$$p(H_k, D) = p(H_k) \cdot p(D \mid H_k) = p(D) \cdot p(H_k \mid D) \tag{2.2}$$

Yet the so-called "Bayesians" adopt a broader interpretation of probabilities than do the so-called "frequentists". To a Bayesian, $p(H_k \mid D)$ is a measure of the *degree of plausibility* of H_k, given *D*, on a scale ranging from *zero* to *one*. In this broader view H_k and *D* are not necessarily repeatable events, but they may be *propositions* or *hypotheses*. The equations of probability theory then become a set of consistent rules for conducting inference [Jaynes 1976]. Since plausibility

itself is always conditioned by some sort of prior set of assumptions, all Bayesian probabilities are regarded as *conditional* on some collective background information denoted hereafter by I. As for instance, there has to be some earlier expectation, or *belief*, so that a Bayesian could assign to H_k some degree of plausibility $p(H_k \mid I)$. For completeness, Bayes' theorem for the plausibility of H_k given the data D must be rewritten as

$$p(H_k \mid DI) = \frac{p(D \mid H_k I) \cdot p(H_k \mid I)}{p(D \mid I)} \tag{2.3}$$

If some additional information, say D', becomes available, we can further refine our estimate probability of H_k by simply substituting the background information I with DI in the above expression which, after elementary manipulations of the product rule of probabilities, leads to

$$p(H_k \mid D'DI) = \frac{p(D'D \mid H_k I) \cdot p(H_k \mid I)}{p(D'D \mid I)} \tag{2.4}$$

It comes out that the product data $D'D$ behaves like D *and* D' had been taken together. Bayesian analysis allows reporting additional information that characterizes the *model space*, the so-called "posterior bubble" in H_k [Jaynes 1976]. Therefore, the probability calculus shows us how to modify our preferences in the light of experience, but it does not suggest what our initial preferences should be. As the noiseless limit of the likelihood $p(D \mid H_k I)$ is approaching a *delta function*, the prior becomes irrelevant and so does the probability calculus, because data are forcing the *correct* result, almost regardless of theory. Yet the prior $p(H_k \mid I)$ plays a crucial role in the complex cases when the number of parameters that are to be estimated are exceeding the number of measurements. A nontrivial prior knowledge $p(H_k \mid I)$ is then necessary to resolve the degeneracy of the problem and to obtain a unique and stable solution.

2.2 DATA MODELING

There are three fundamental aspects related with scientific investigations of the physical world: (*i*) apparatus design, (*ii*) measuring techniques, and (*iii*) data processing from which conclusions are drawn. In a good experiment all these items have to be critically considered and carefully controlled.

Data processing is getting from *data space* to *map space*. The theory of the experiment can be regarded as an *operator* from map space to data space. The *forward problem* consists in applying this operator. It is always possible assuming that we can analyze the experiment. The *inverse problem* is related with the *inverse operator* that might take us from data space to map space. Yet the inverse operator might not exist at all since it usually operates on a space larger than

the data space. The approach is to consider all the solutions (models) in the map space that could give rise to the data. Once the forward problem is solved, we can check whether a solution is consistent with the data according to some criterion by taking into account the errors in the data. If all the consistent models are very similar, than there is no need to choose. However, if there are many different ones that comply, then a rule for choosing must be introduced to solve the inverse problem [Daniell 1994].

The process of elaborating a model on the basis of experimental data and drawing conclusions from it is an example of *inductive inference*. There are two levels of inference in any data modeling process. The first concerns fitting each model to the data, that is, to infer the values of the free parameters of each model might correspond to the data. The second performs model comparison that ranks how plausible the alternative models are with respect to the data. Bayesian methods deal consistently and quantitatively with both inductive inference levels and embody intrinsically Ockham's factor, which avoids data over-fitting and over-complex models.

Let us describe in Bayesian terms the two level of inference in data modeling processes. We assume that each model H_k has a vector of parameters $\boldsymbol{w}$. Then a model is defined by its functional form and two probability distributions: (*i*) a *prior* distribution $p(\boldsymbol{w} \mid H_k)$, which states what values the model's parameters might plausibly take, and (*ii*) a *posterior* distribution of the predictions $p(D \mid \boldsymbol{w}, H_k)$ that the model makes about the data D when its parameters have a particular value $\boldsymbol{w}$.

2.3.1 Model fitting

Much of the Bayesian viewpoint can be argued as direct application of the theory of probability. Yet Bayesian methods, in contrast with the classic statistics (also known as "sampling theory"), assign objective preferences to the alternative models. Model comparison is a difficult task because a simple choice of the model that fits the data best would lead the *maximum likelihood* (ML) choice to implausible and impractical sophisticated models that can be poorly generalized. *Ockham's razor* is the principle that states the preference for simpler models. Bayes' theorem rewards models proportionally with their predictions of the data that occurred. These predictions are quantified by the *evidence* $p(D \mid H_k)$ for H_k, that is the normalized probability of the data D given the model H_k.

We consider that one model H_k is true and we infer the model's parameters $\boldsymbol{w}$ might correspond to the data D. Bayes' theorem provides the *posterior probability* of the parameters $\boldsymbol{w}$

$$p(\boldsymbol{w} \mid D, H_k) = \frac{p(D \mid \boldsymbol{w}, H_k) \cdot p(\boldsymbol{w} \mid H_k)}{p(D \mid H_k)} \tag{2.5}$$

The evidence $P(D \mid H_k)$ for H_k plays a normalizing role and it is generally ignored in the first level of inference, i.e., the choice of $\boldsymbol{w}$. Gradient-based variational methods are commonly used to find the maximum of the posterior probability, which defines the most probable value of the parameters $\boldsymbol{w}_{MP}$. Error bars on these best-fitting parameters are obtained from the curvature of the posterior distribution by Taylor-expanding the *log posterior* with $\Delta \boldsymbol{w} = \boldsymbol{w} - \boldsymbol{w}_{MP}$

$$p(\boldsymbol{w} \mid D, H_k) \cong p(\boldsymbol{w}_{MP} \mid D, H_k) \cdot exp\left(-\frac{1}{2} \Delta \boldsymbol{w}^T \mathbf{A} \Delta \boldsymbol{w}\right) \tag{2.6}$$

where $\mathbf{A} = -\nabla\nabla log\ p(\boldsymbol{w} \mid D, H_k)$ is the Hessian matrix of *log posterior* at $\boldsymbol{w}_{MP}$. It comes out that the posterior can be approximated as a Gaussian with covariance matrix $\mathbf{A}^{-1}$, which also provides the error bars (standard deviation for each component of $\boldsymbol{w}$).

2.4.2 Model comparison

We wish to infer which model is most plausible given the data. The posterior probability for each model H_k is

$$p(H_k \mid D) \propto p(D \mid H_k) \cdot p(H_k) \tag{2.7}$$

The data-dependent term $p(D \mid H_k)$ is the *evidence* for H_k, while $p(H_k)$ is a *subjective* prior over our hypothesis space, which expresses how plausible we believed the alternative models H_k, $k = 1,2,...,L$ were before data arrived. Typically, the subjective prior is overwhelmed by the objective one, i.e., the evidence. If we have no reason to assign sensibly unequal priors $p(H_k)$ to alternative models, then models $H_k, k = 1,2,...,L$ are ranked by evaluating the evidence [MacKay 1992]. However, to use only the evidence for model comparison is equivalent to using maximum likelihood for parameter estimation.

Model evidence

The evidence $p(D \mid H_k)$ for H_k can be evaluated for both parametric and non-parametric models by integrating over the whole parameter space

$$p(D \mid H_k) = \int_{\boldsymbol{w}} p(D \mid \boldsymbol{w}, H_k) \cdot p(\boldsymbol{w} \mid H_k) \cdot d\,\boldsymbol{w} \tag{2.8}$$

In many problems, the posterior $p(\boldsymbol{w} \mid D, H_k) \propto p(D \mid \boldsymbol{w}, H_k) \cdot p(\boldsymbol{w} \mid H_k)$ has a strong peak at the most probable parameters $\boldsymbol{w}_{MP}$. Then the evidence can be approximated by the height of the peak of the integrand $p(D \mid \boldsymbol{w}, H_k) \cdot p(\boldsymbol{w} \mid H_k)$ times its width $\Delta \boldsymbol{w}$ so that

$$p(D \mid H_k) \cong p(D \mid \boldsymbol{w}_{MP}, H_k) \cdot p(\boldsymbol{w}_{MP} \mid H_k) \cdot \Delta \boldsymbol{w} \tag{2.9}$$

The term $p(w_{MP} \mid H_k) \cdot \Delta w$ is the *Ockham's razor* that has a value less than unity and multiplies the *best fit likelihood* $P(D \mid w_{MP}, H_k)$. Thus the model H_k is penalized for having the parameters w. The posterior uncertainty in w is Δw.

Ockham's razor

Intuitively, Ockham's principle works as follows. Assume that a simple model H_1 and a complex one H_2 are making predictions, $p(D \mid H_1)$ and $p(D \mid H_2)$, respectively, over the possible data sets D. The simple model H_1 makes predictions over a limited range of data sets, say d, whereas the complex model H_2, which may 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.

We could conceive the concept of a *razor* as a natural measure of complexity of a parametric family of distributions relative to a given true distribution [Balasubramanian 1995]. Assume that the prior $p(w \mid H_k)$ is uniform on some large interval $\Delta_0 w$ that encompasses all values of w allowed by H_k before acquiring data. Then $p(w_{MP} \mid H_k) = 1/\Delta_0 w$ and the Ockham's razor, or factor, becomes $\Delta w / \Delta_0 w$, that is, the ratio of the posterior accessible volume of the model's parameter space to the prior accessible volume. Therefore, in the case of a single parameter, the evidence is

$$p(D \mid H_k) \cong p(D \mid w_{MP}, H_k) \cdot \frac{\Delta w}{\Delta_0 w} \tag{2.10}$$

If w is V-dimensional, and if the posterior $p(w \mid D, H_k)$ is well approximated by a Gaussian, then Ockham's factor is obtained from the determinant of the Gaussian's covariance matrix

$$p(D \mid H_k) \cong p(D \mid w_{MP}, H_k) \cdot p(w_{MP} \mid H_k) \cdot (2\pi)^{V/2} \cdot det^{-1/2} \mathbf{A} \tag{2.11}$$

As the amount of data collected M increases, the Gaussian approximation is expected to become increasingly accurate according to the central limit theorem. For the linear models with additive noise, the Gaussian expression is exact for any M.

The model H_k can be regarded, as being composed of a certain number of equivalent sub-models, out of which there remains only one after the data are collected. The Ockham's razor is the inverse of this number. Therefore, the *log* of the Ockham's factor can be interpreted as the amount of information we gain about the model when the data are collected. To conclude, Ockham's razor is a measure of the model complexity in the sense of the predictions that the model performs in data space so that it depends on the number of the data points and other properties of the data set.

The razor favors simpler, more robust models when the amount of data is small, but asymptotically picks the most accurate model in the relative entropy sense. The *log* of the Bayesian posterior probability of a model family given the data converges to the *log* of the razor rather strongly.

III MAXIMUM ENTROPY METHODS

3.1 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*)

$$S = k_B \cdot logW \tag{3.1}$$

where k_B stands for the Boltzmann's constant. Entropy was first perceived by Boltzmann, yet he never wrote it in the above form but M. Plank 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't be normally detected on a macroscopic scale.

In statistics, Shannon [1948, 1949] 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 gained 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

$$S(\{p_n\}) = -\sum_{n=1}^{N} p_n \cdot lnp_n \tag{3.2}$$

where $\{p_n\} = p(X = x)$ is the *probability distribution* of X. If $\{p_n\}$ and $\{q_n\}$ are two probability distributions of a random variable X, the *relative*, or *cross entropy* of $\{p_n\}$ and $\{q_n\}$ is defined as

$$H(\{p_n\}, \{q_n\}) = \sum_{n=1}^{N} p_n \cdot ln\left(\frac{p_n}{q_n}\right) \tag{3.3}$$

Evidently, minimizing $H(\{p_n\}, \{q_n\})$ is equivalent to maximizing $S(\{p_n\})$, assuming that $\{q_n\}$ is a uniform distribution. In both forms of entropy, 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.

3.2 MAXIMUM ENTROPY METHODS

Maximum entropy (ME) methods originate from physics (Boltzmann, Gibbs) and have been promoted as general methods of inference primarily by Kullback [1959] and Jaynes [1983]. As a prerequisite, to apply Bayes' theorem we must first use some other principle to translate the available information into numerical values. By *applying* the ME principle we mean *assigning* a

probability distribution $\{p_n\}=\{p_1, p_2,..., p_N\}$ on some *hypothesis space* $\{H_1, H_2,..., H_N\}$ by the criterion that it shall maximize some form of *entropy*, such as the informational entropy $S(\{p_n\})=-\sum_{n=1}^{N} p_n \cdot log\ p_n$ subject to constraints that express properties we wish the distribution to have, but are not sufficient to determine it. Entropy is used as the criterion for resolving the ambiguity remaining when we have stated all the conditions we are aware of. ME methods require that we specify in advance a definite hypothesis space $\{H_1, H_2,..., H_N\}$, which sets down the possibilities to be considered. They necessarily produce a probability distribution $\{p_1, p_2,..., p_N\}$, not just a probability. It does not make any sense to ask for the ME probability of an isolated hypothesis *H*, which is not embedded in a space of alternative hypotheses. ME methods do not require for input the numerical values of any probabilities on that space, rather they assign numerical values to our information as expressed by our choices of hypothesis space and constraints. Jaynes [1988] called this initial stage the "exploratory phase" of a problem.

In entropy-related contributing papers, ME methods are based on a variety of "entropy functionals" that differ from Shannon's informational entropy expression [Skilling 1988, Csiszar 1995]. In spectrum analysis, ME methods deals with maximizing $\int log\ p(x)\cdot dx$, that is called the Burg entropy of the power spectrum *p* or, if a non-constant default model *q* is given, minimizing $\int\left(log\frac{q(x)}{p(x)}+\frac{p(x)}{q(x)}-1\right)\cdot dx$, that is the Itakura-Saito distance of *p* from *q*. In a general sense, the *f-entropy* of a probability density function *p* ($p(x)\geq 0$) is formally defined as

$$S_f(p)=-\int f(p(x))\cdot\lambda(dx) \tag{3.4}$$

where $f(t)$ is any strictly convex differentiable function on the positive reals and λ denotes a given σ-finite measure on $\{H_1, H_2,..., H_N\}$. A corresponding measure of distance of *p* from *q* [Bregman 1967] is introduced

$$B_f(p,q)=\int[f(p(x))-f(q(x))--f'(q(x))\cdot(p(x)-q(x))]\cdot\lambda(dx) \tag{3.5}$$

Recall that the strict convexity implies $f(s)>f(t)+f'(t)\cdot(s-t)$ for any positive numbers $s\neq t$. Accordingly, if we take $t\cdot log\ t$, $-log\ t$, or t^2, then S_f will be the Shannon's entropy, Burg's entropy, or the negative square integral of *p*, respectively.

3.3 VARIANTS OF MAXIMUM ENTROPY

In many physical experiments, the observed data $g(s)$ from *data space* are transforms of the quantities of interest $f(t)$. The *linear transforms* [M.-Djafari 1995] encompass a large class of such experiments

$$g(s) = \int f(t) \cdot r(t,s) \cdot dt + b(s) \tag{3.6}$$

where $r(t,s)$ is the known *point spread function* (PSF) of the measuring equipment and $b(s)$ stands for the *measurement errors (noise)*. Experimentally, $g(s)$ is observed on a finite set of isolated points s_m, $m = 1,2,...,M$

$$g_m = g(s_m) = \int f(t) \cdot r(t,s_m) \cdot dt + b(s_m) = \int f(t) \cdot r_m(t) \cdot dt + b_m \ , \ m = 1,2,...,M \tag{3.7}$$

Assume each 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 $r_m(t)$ and with its own additive measuring error b_m. For a large number N of discrete points t_n, $n = 1,2,...,N$, which are sufficiently evenly spaced, neither $f(t)$ nor $r_m(t)$ vary significantly between t_{n-1} and t_{n+1}. Then quadrature-like expressions hold for the data

$$g_m = \sum_{n=1}^{N} R_{mn} f_n + b_m, \ m = 1,2,...,M \tag{3.8}$$

where we introduced the $M \times N$ matrix $\mathbf{R}$ with the components given by

$$R_{mn} = r_m(t_n) \cdot (t_{n+1} - t_{n-1})/2, \ m = 1,2,...,M; \ n = 1,2,...,N \tag{3.9}$$

Since the matrix $\mathbf{R} = \left\| R_{mn} \right\|_{\substack{m=1,2,...M \\ n=1,2,...,N}}$ is generally either *singular* or very *ill-conditioned* [Nashed 1981], the *well-posed inverse problem* of data modeling may be formulated by asking for some reliable estimate $\tilde{f}$ to the exact solution f, given the measured sample data g, the space-invariant PSF matrix $\mathbf{R}$, and some information about the errors b, such as their covariance matrix $\mathbf{C} = \left\| C_{ij} \right\|_{i,j=1,2,...,M}$.

There may be classified three different approaches of using the ME principle in solving the ill-posed inverse problems in data modeling [M.-Djafari 1992].

3.3.1 Classical ME

The function f actually is or has the properties of a probability density function. In the above discrete case, its components $\{f_n\}$, $n = 1,2,...,N$ are positive and, when normalized, can be considered as a probability distribution. The noiseless data $g = \mathbf{R} \cdot f$ are the exact linear

constraints on $f=\{f_n\}$. The task is that given **R**, g, and the reference (or template) $q=\{q_n\}$, then

$$maximize\ -\sum_{n=1}^{N} f_n \cdot ln\left(\frac{f_n}{q_n}\right)\ \ so\ that\ \ g=\mathbf{R}\cdot f\ \ and\ \ \sum_{n=1}^{N} f_n = 1 \tag{3.10}$$

The implicit solution of the problem if the constraints are given by $g=\mathbf{R}\cdot f$ is the following

$$f_n = \frac{1}{Z}\cdot q_n \cdot exp\left[-\sum_{m=1}^{M}\lambda_m \cdot R_{mn}\right] = \frac{1}{Z}\cdot q_n \cdot exp\left[-\left(\mathbf{R}^t\cdot\lambda\right)_n\right],\ n=1,2,...,N \tag{3.11}$$

where the Lagrange multipliers $\lambda=\{\lambda_1,\lambda_2,...,\lambda_M\}$ are obtained by solving

$$G_m(\lambda_1,\lambda_2,...,\lambda_M)=\sum_{n=1}^{N}\frac{1}{Z}\cdot q_n\cdot exp\left[-\left(\mathbf{R}^t\cdot\lambda\right)_n\right]=g_m,\ m=1,2,...,M \tag{3.12}$$

The main drawback is that this method does not account for the errors in the data.

3.3.2 ME in the mean

In this approach the function $f=\{f_n\}$ is assumed to be the outcome of a random vector for which we assume to be able to define the prior distribution $\{q_{nk}\}=\{p(f_n=\alpha_k)\}$ and the posterior distribution $\{p_{nk}\}=\{p(f_n=\alpha_k)\}$. The data are then supposed to be the expected values $\langle g_m\rangle$, $m=1,2,...,M$, where $\langle\cdot\rangle$ means the average over $\{p_{nk}\}$. That is, the data $g_m=\int_D E(f(t))\cdot r_m(t)\cdot dt=\mathbf{R}\cdot f$ are considered as the linear constraints on the *mean* $E(f(t))=\langle f\rangle$, or $g_m=E\left(\int_D f(t)\cdot r_m(t)\cdot dt\right)=\langle\mathbf{R}\cdot f\rangle$ are considered as the *mean* values of the linear constraints. In both cases we may write explicitly

$$g_m=\sum_{n=1}^{N}R_{mn}\left(\sum_{k=1}^{K}f_n\cdot p_{nk}\right),\ m=1,2,...,M\ .$$

The task is that given **R**, g, and the prior $q=\{q_{nk}\}$, find the probability distributions $\{p_{nk}\}$ which satisfy the constraints and minimize the cross entropy

$$maximize\ H(p,q)=\sum_{k=1}^{K}p_{nk}\cdot ln\left(\frac{p_{nk}}{q_{nk}}\right)\ \ so\ that\ \ g=R\cdot\langle f\rangle; \tag{3.13}$$

Therefore, we may construct an estimate $\tilde{f}_n$ for each $\langle f_n\rangle$, $n=1,2,...,N$ and the solution to the inverse problem is

$$\tilde{f}_n=\langle f_n\rangle=\sum_{k=1}^{K}f_n\cdot p_{nk},\ n=1,2,...,N \tag{3.14}$$

The drawback is that the errors on the data can't be accounted for directly since the measured data are considered to be mean values and the estimated solution is also a mean value.

If data are given in the form of measurement variances $\left\langle \left(g_m - \langle g_m \rangle\right)^2 \right\rangle$ and $\{q_{nk}\}$ is uniform or Gaussian, then $\{p_{nk}\}$ is also Gaussian and the ME in mean estimator is equivalent to the classical Least Squares estimators.

3.3.3 Bayesian ME approach

The function $f = \{f_n\}$ is considered to be as an outcome of a random vector on which we have the expected values but affected by noise (errors). It is assumed that we are able to define its prior probability distribution $p(f \mid I)$ according to ME principle, which translates our uncertainty or our incomplete prior knowledge I about f.

The data $g = \{g_m\}$ are also considered as an outcome of a random vector for which we are able to define the conditional distribution $p(g \mid f)$ representing our confidence in the data g, or the uncertainty about the errors (both measuring noise and modeling errors).

Bayes' theorem is used to combine these two knowledge states and to produce the posterior probability distribution law $p(f \mid g)$ representing our state of knowledge of the solution. The final step concerns the choice of an estimation procedure of the unknown function f by which the optimal estimator $\tilde{f}$ is to be determined.

The following issues have to be addressed for making an optimal choice of the method involving ME in inverse problems of data modeling. First, in order to select an *appropriate approach*, the next questions must be answered

- What physical quantity is represented by the unknown function ?
- What physical quantity is represented by the data ?
- What is our information concerning the errors and/or noise in the data ?
- What is the information we want about the unknown function?

Secondly, to make an appropriate option for an *effective method*, the following questions must be clarified

- What method considers the unknown function as it must be?
- What method takes into account the data as they really are?
- What method can yield what we are interested in about the unknown function?
- What method can better comply to our prior knowledge about the unknown function?
- What method can better comply to our prior knowledge about the errors and noise in our measurement system?

There is some evidence that the Bayesian approach answer fairly good to all the well-posed questions arisen by practical linear inverse problems, and satisfies all the consistency requirements

of the inference when we have to combine the prior information and the information contained in the data [Tikochinsky 1984, MacKay 1991, 1994].

IV. APPLICATIONS

4.1 INVERSE PROBLEMS AND BAYESIAN REGULARIZATION

Bayesian analysis is a fundamental statistical approach to physical system modeling. This approach is based on the assumption that all of the relevant information about the subject may be stated in probabilistic terms and prior probabilities are known. A highly widespread non-parametric method for interpolating various types of data is based on *regularization*, which looks for an interpolant that closely fits the data and it is also "smooth" in some sense to allow generalization. Formally, the interpolant is obtained by minimizing an error functional $M(f)$ defined for every admissible function f that is the weighted sum of a "fitting term" $D(f)$, which measures the distance of f from the given data set, and a "smoothness term" $S(f)$

$$M(f) = D(f) + \lambda S(f) \tag{4.1}$$

where the parameter $\lambda > 0$ decides the trade-off between fitting the data and smoothness. The solution is the function f that is minimizing the error, or cost, functional $M(f)$. As for instance, in the one-dimensional case, the objective function subject to minimization can be [Keren and Verman 1995]

$$M(f) = \sum_{i=1}^{n} \frac{[f(x_i) - y_i]^2}{2\sigma^2} + \lambda \int_0^1 f_{uu}^2 \cdot du \tag{4.2}$$

The Bayesian interpretation is straightforward: given the data D, we want to find the model f that maximizes the likelihood $p(f \mid D) \propto p(D \mid f) \cdot p(f)$. If we assume a Gaussian noise model with standard deviation σ, then the evidence for the data D is $p(D \mid f) \propto \frac{1}{\sigma^n} \cdot exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [f(x_i) - y_i]^2 \right)$. Adopting a physical model, it is common to define the evidence for the model f such as $p(f) \propto exp\left(-\int_0^1 f_{uu}^2 \cdot du \right)$. Hence $p(f \mid D) \propto exp(-M(f))$ and the function f that minimizes $M(f)$ maximizes the likelihood in this particular one-dimensional case of a Gaussian model.

There are various methods for determining the weights σ and λ that are consequently used to interpolate the function f, but this approach fails to actually compute the maximum likelihood (ML). The ML estimate should maximize the likelihood over the set of all admissible weights

$\int_{w} p(f \mid D, w) \cdot p(w \mid D) \cdot dw$, where we generically denoted the set of weights by the weight parameter vector w. A widely spread method for determining the smoothing parameter is Generalized Cross Validation (GCV), or *bootstrapping* [Craven and Wahba 1979]. The drawback of using a single set of weights is that the choice of the values of σ and λ is sometimes very sensitive to the data. Since these values are crucial to the shape of the fitted curve or surface, they may dramatically change the interpolant for small change in the data. Moreover, despite some nice asymptotic properties, the choice of the "optimal" values of σ and λ is essentially heuristic in nature.

Recent works deal with entropic smoothing [Strauss *et al.* 1993, Fischer *et al.* 1995] by marginalization over the regularization parameter in order to select the correct entropic prior. Comparison with the conventional ME method methods shows less overfitting of noise and demonstrates the residual ringing which is intrinsic to all ill-posed problems. Furthermore, the problem of computing the MAP solution in a Bayesian framework by integrating over the space of parameters and noise is addressed [Molina and Katsaggelos 1994]. For "integrating out" these two parameters, a uniform prior for them is assumed.

4.2 IMAGE RECONSTRUCTION

Many different types of images satisfy the axioms of positivity and additivity. Such images may be directly visual like *2D* distributions of light intensity, or can be *3D* densities of electrons in the free space, or *1D* sequences of numbers on magnetic tapes. If the image is represented as a sequence of positive numbers $\{f_1, f_2, ..., f_N\}$, then for the corresponding proportions $p_n = f_n \Big/ \sum_{n=1}^{N} f_n$, $n = 1,2,...,N$ we can write

$$
\begin{aligned}
&i)\ p_n \geq 0 \ , \ n = 1,2,...,N \qquad \text{(positivity)} \\
&ii)\ \sum_{n=1}^{N} p_n = 1 \quad \text{(normalization)} \\
&iii)\ p_{n \cup m \cup ...} = p_n + p_m + ... \ , n \neq m \neq ... \quad \text{(additivity)}
\end{aligned}
\tag{4.3}
$$

Therefore, the configurational structure of these images satisfies the axioms of probability theory and the concept of *image entropy* can be introduced. It follows that the image information content or image entropy may be defined such as $S(f) = -\sum_{n=1}^{N} p_n \cdot log\ p_n$, where $f = \{f_1, f_2, ..., f_N\}$ is the image and the distribution $\{p_n\} = \{p_1, p_2, ..., p_N\}$ corresponds to the light intensity in each pixel.

Now, if we consider a complete collection of images corresponding to all possible intensity distributions, then measurements act as a filter over the collection by restricting our attention to the images that satisfy the data with any conceivable constraints (noise). Among these, a natural choice may be the one that could have arisen in the maximum number of ways, depending on our counting rule. By maximizing the entropy, the smoothest and most uniform distribution among the set of all admissible distributions is selected, as the most structureless possible image. ME restoration allows to associate error bars on the retrieved images which provides means to assess quantitatively and objectively the reliability of the extracted features. Furthermore, we can assess different variants of ME and give quantitative comparisons [Skilling 1994].

4.2.1 Bayesian ME approach

We assume hereafter that there exists a scene that can be adequately represented by an orderly array of N pixels. Solving the inverse problem of image restoration within the framework of Bayesian statistics consists in selecting the *best statistical estimator* $\tilde{f}$ of an image f, assuming some hypotheses.

(*i*) The image is considered as an outcome of a *random vector* f given in the form of a *probability distribution* $f = \{f_1, f_2, ..., f_N\}$, $n = 1,2,...,N$.

(*ii*) A prior probability $p(f \mid I)$ is defined that reflects our uncertainty or incomplete knowledge about f. The prior probability should incorporate as much as possible of the known statistical characteristics of the ensemble of images from which the original image f is assumed to belong. Nevertheless, the prior probability must be as general and non-informative as possible. In general, this prior knowledge is not directly given in a probabilistic way, nor does it yield a unique prior law $p(f \mid I)$. However, the shape of any positive, additive image can be directly identified with a probability distribution. Consequently, whether it might be for spectral analysis of time series, radio astronomy, optical X-ray astronomy and tomography, or for any reconstruction of positive, additive, images, the ME principle assigns a prior probability to any given image f. The Bayesian approach does not specify any particular choice of prior. The *entropic* prior has been argued by Skilling [1990] to play a unique role for additive positive distributions. Whatever prior is used, it clearly affects the amount by which the reconstruction is offset from the true image [Hanson 1990, Myers and Hanson 1990]. The extent of this biasing effect depends on the relative weights of the *likelihood* and *prior* in the expression of the *posterior* probability in Bayes' theorem. Regardless of the prior chosen, it is recognized that the prior provides the regularization essential to solving *ill-posed* problems [Nashed 1981, Titterington 1985], which arise because **R** possesses a null-space [Hanson and Wecksung 1983, Hanson 1987]. A nontrivial prior knowledge $p(f \mid I)$ is necessary to resolve the degeneracy of the inverse problem by providing a unique and

stable solution, since the number of *parameters* (components of f) we have to estimate is comparable or larger than the number of *measured data* (components of g).

In the *ergodic* situations, i.e., where there is reason to believe that the *a priori* probabilities of the microscopic configurations are all the same, the *Bayesian ME prior* $p(f \mid I)$ for a *macroscopic* state f with entropy $S(f)$ is postulated as proportional to $exp\, S(f)$. Accordingly, if we have only partial information on a random process, we shall assume that it is distributed according to ME law that complies to our *a priori* information, that is

$$p(f \mid I) \propto exp\, S(f) = exp\left(-\sum_{n=1}^{N} f_n \cdot ln\, p_n \right) \tag{4.4}$$

It is important to emphasize that in Bayesian approach the concept of entropy is used *only* to express mathematically our prior knowledge about f. Invoking the ME principle requires that the prior knowledge has been stated as a set of constraints on $p(f \mid I)$ [Justice 1986]. Among the probability laws that satisfy these constraints, the ME principle leads to choose the one which maximizes the conformational entropy $S(f)$. There are nevertheless some other choices one may consider apart from the entropic one [Hanson 1991].

(*iii*) Data are also considered as an outcome of a *random vector* g, for which the conditional probability $p(g \mid fI)$ must be defined. It represents our confidence in the data and our incomplete knowledge about the errors and noise in the measurement system, which is included in the *noise random vector* b expressed also in the form of a *probability distribution*.

It is a common practice to express the probability of the observed data for any particular image f, i.e., the *likelihood* $p(g \mid fI)$, as a finite set of constraints [Mutihac 1997a,b]. For the widely spread case of additive Gaussian noise $b = k\sigma$ in linear experiments $g = \mathbf{R} \cdot f + b$, the likelihood is identical to the probability law of the noise. As usually, σ is the *standard deviation* of the noise, and k is drawn from the *unit normal distribution*

$$p(k) = (2\pi)^{-\frac{1}{2}} \cdot exp\left(\frac{k^2}{2} \right) \tag{4.5}$$

Within this framework, Gull and Daniell [1978] suggested a χ^2-constraint for handling the errors b_m, $m = 1,2,\ldots, M$. Accordingly, after a set of M measurements, the expression of likelihood is

$$p(g \mid fI) = \left[\prod_{m=1}^{M} \left(2\pi\sigma_m^{\,2} \right)^{-\frac{1}{2}} \right] \cdot exp\left[-\frac{1}{2}\chi^2(f) \right] \tag{4.6}$$

Instead of a Gaussian distribution assumed here, in some experiments the Poisson distribution is often a better model for the expected measurement fluctuations. Poisson data have larger absolute errors on the more intense data values varying as the square root, so that if noise amplification is to occur, it will be seen first on the bright regions. The entropy formula dominated

by the sum of terms like $f_n \cdot log\, f_n$ is relatively tolerant of bright spots corresponding to large f, so that oscillations are likely to develop in case of relatively extensive bright regions. Skilling [1994] demonstrates the generality of the probabilistic approach by deconvolving blurred data sets subject to Poisson statistics. In any case, the choice should be based on the statistical characteristics of the measurement noise, which we assume *a priori* known. The choice of *entropic prior* is nevertheless not affected by any of the *likelihood* functions we may consider.

(*iv*) The posterior probability $p(f \mid gI)$ properly combines the likelihood $p(g \mid fI)$, which is based on recently acquired measurements, with the prior probability $p(f \mid I)$, which subsumes all available information before the new data are acquired. By substitution of the prior probability $p(f \mid I)$, and the likelihood $p(g \mid fI)$, in Bayes' theorem the posterior probability, $p(f \mid gI)$ is estimated

$$p(f \mid gI) \propto (likelihood) \cdot exp\, S(f) \tag{4.7}$$

(*v*) Finally, the "best" statistical estimator, $\tilde{f}$, is selected by using the posterior probability, $p(f \mid gI)$. This requires to adopt an *estimation rule*, such as *posterior mean* or *maximum a posteriori* (MAP), in order to select an *optimal, unique* and *stable* solution.

In image reconstruction, we seek to estimate all pixel values in the original scene. The essence of the Bayesian approach lies in the steadfast use of the posterior probability, which is assumed to summarize the full state of knowledge concerning a given situation. An appropriate Bayesian solution to this problem is the image that maximizes the posterior probability. If we had to produce just one single image as the "best reconstruction", we would naturally give the *most probable* one which maximizes $p(f \mid gI)$, along with some *statement of reliability* derived from the spread of all reasonably probable images. In the particular case of Gaussian noise distribution, the Bayesian approach in inverse problems of image restoration deals with maximization of the posterior probability or, equivalently, with minimization of the $-log\, p(f \mid gI)$ that is

$$-log\, p(f \mid gI) = \frac{1}{2}\chi^2(f) + \sum_{n=1}^{N} f_n \cdot log\left(f_n \Big/ \sum_{n=1}^{N} f_n \right) \tag{4.8}$$

a task that falls in the field of constrained non-linear optimization problems [Agmon *et al.* 1979, Wilczek and Drapatz, 1985, Mutihac *et al.* 1998a].

4.3 NEURAL NETWORKS

Artificial neural models are currently receiving much attention due to newly developed topologies and algorithms [Neal 1992,1994, 1997], advanced very large scale integration (VLSI) implementation techniques [Mutihac 1992, 1995, 1996], and the belief that massive parallelism is essential in data processing for pattern recognition and emulating of the brain's functions. A

practical Bayesian framework for neural networks (NNs) modeling aims to develop probabilistic models that fit the data and perform optimal predictions.

The neural functional F establishes a correspondence between the input vector $\boldsymbol{x} = (x_1, x_2, ..., x_N)$ and the output one $\boldsymbol{y} = (y_1, y_2, ..., y_M)$, so that it could be represented as a *functional* or *mapping* from the input to the output space $\boldsymbol{y} = F(\boldsymbol{x})$. The non-linearity of the single neuron activity function is compulsory and provides greater computational flexibility than the standard linear regression models. By introducing additional non-linear processing between the input and output, the network could implicitly learn more complicated *prior probability* distributions, more complex noise models, and more about non-linear response functions. Threshold units are required for *universal approximation*, i.e., to assign *nonzero* output to *zero* input. The weights of the connections $w_{ij}^{I,O}$ and the biases $\theta_i^{I,O}$ altogether make up the *parameter vector* $\boldsymbol{w}$ of the network. A *network architecture* Ω is defined by specifying: (*i*) the number of layers, (*ii*) the number of units in each layer, (*iii*) the type of activation function performed by each unit, and (*iv*) the available connections between the units. Then a *neural model* is a non-linear parameterized mapping from an input activity $\boldsymbol{x}$ to an output activity $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}; \boldsymbol{w}, \Omega)$, where $\boldsymbol{w}$ stands for all the net parameters [Mutihac and Cicuttin 1998b].

Mathematically, *training*, or *learning*, refers to the adaptation of the function parameter set. The training set for the mapping to be learned consists of input-target pairs $D = \{\boldsymbol{x}^{(m)}, \boldsymbol{t}^{(m)}\}$ and the network is trained in the sense of fitting the mapping $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}; \boldsymbol{w}, \Omega)$ with the training set by minimizing an *error* or *objective* function, such as

$$E_D(D \mid \boldsymbol{w}, \Omega) = \sum_m \frac{1}{2}\left[\boldsymbol{y}(\boldsymbol{x}^{(m)}; \boldsymbol{w}, \Omega) - \boldsymbol{t}^{(m)}\right]^2 \tag{4.9}$$

Traditionally, the back-propagation (BP) algorithm [Rumelhart, Hinton and Williams 1986] is an effectively fairly widespread training method due to its relative simplicity to apply. The error function is essentially minimized by performing gradient descent on E_D in the $\boldsymbol{w}$-space. However, BP is slow and provides local minimum points only. Refinements of the plain BP include the addition of a momentum term and the inclusion of noise in the descent process. In order to decrease the tendency of a model to excessively fit the training set, extra regularizing terms are added to E_D, such as additive weight-dependent energies [Abu-Mostafa 1990]:

$$E_W(\boldsymbol{w} \mid \Omega, \mathfrak{R}) = \frac{1}{2}\left[\sum\left(w^{I,O}\right)^2 + \sum\left(\theta^{I,O}\right)^2\right] \tag{4.10}$$

The symbol $\mathfrak{R}$ denotes the *regularizer* using a particular energy function E_W. The summation is over the entire parameter space. It turns out that the objective function subject to minimization becomes after regularization (*weight decay*)

$$M(w) = \alpha E_W(w \mid \Omega, \Re) + \beta E_D(D \mid w, \Omega) \tag{4.11}$$

where α (the decay rate) and β are regularizing parameters. By introducing a probabilistic view of NN learning [Tishby *et al.* 1989], there is an assigned meaning to the functions and parameters already introduced. Then Bayesian methods may provide objective criteria for setting free parameters and comparing alternative solutions that depend only on the training data set.

4.3.1 NN learning as inference

A network defined by the control parameters Ω and w is making predictions about the targets as function of the input data such as [MacKay 1992a,b, Neal 1996]

$$p(t^{(m)} \mid x^{(m)}, w, \beta, \Omega, \aleph) = \frac{exp\left(-\beta E(t^{(m)} \mid x^{(m)}, w, \Omega)\right)}{Z_m(\beta)} \tag{4.12}$$

where $Z_m(\beta) = \int dt \cdot exp(-\beta E)$ is the normalization function, E is the error in a single datum, β is a measure of the presumed noise included in t, and $\aleph$ denotes the implicit noise model. It results that the error function E can be interpreted as the *log likelihood* for the noise model $\aleph$.

A *prior probability* distribution is assigned to alternative net connections

$$p(w \mid \alpha, \Omega, \aleph) = \frac{exp(-\alpha E_W(w \mid \Omega, \Re))}{Z_W(\alpha)} \tag{4.13}$$

where $Z_W = \int d^k w \cdot exp(-\alpha E_W)$ is the regularization and k denotes the number of free parameters, i.e., the dimension of the w-space. The interpretation of α is the measure of the expected connection magnitude.

The *posterior probability* of the network connections w is

$$p(w \mid D, \alpha, \beta, \Omega, \aleph, \Re) = \frac{exp(-\alpha E_W - \beta E_D)}{Z_M(\alpha, \beta)} = \frac{exp(-M(w))}{Z_M(\alpha, \beta)} \tag{4.14}$$

where $Z_M(\alpha, \beta) = \int d^k w \cdot exp(-\alpha E_W - \beta E_D) = \int d^k w \cdot exp(-M(w))$. Thus the minimization of the *misfit function* $M(w)$ corresponds to the *inference* on the parameters w, given the data D. The probability of the connections w is a measure of *plausibility* that the parameters of the probabilistic model should have a specified value w, and it is not related with the probability that a particular algorithm might converge to w. The term *model* refers to a set of three suppositions: (*i*) the network architecture Ω, (*ii*) the prior of the parameters, and (*iii*) the noise model $\aleph$; hence we may represent a NN model as a set $\Lambda = \{\Omega, \aleph, \Re\}$.

4.3.2 Model parameters

The search in model space may also be considered as an inference process that addresses the relative probability of alternative models, given the data [MacKay 1992a]. Prior distribution of

the model parameters captures our beliefs about the relationship that we are modeling with the net. Though parameter prior lacks any direct meaning, it is nevertheless more important the prior over functions computed by the net implied by the weight prior. Each model comprises a hypothesis with some free parameters that assign a probability density normalized so as to integrate to unity. The control parameters α and β determine the complexity of the network model $\Lambda = \{\Omega, \aleph, \Re\}$. Bayes' theorem indicates how to set these parameters by inference using the data. The posterior probability of the control parameters is

$$p(\alpha, \beta \mid D, \Lambda) = \frac{p(D \mid \alpha, \beta, \Lambda) \cdot p(\alpha, \beta \mid \Lambda)}{p(D \mid \Lambda)} \tag{4.15}$$

If we assign a uniform prior to α and β, which corresponds to the statement that we don't know what value α and β should have, then the normalizing factor of our previous inference that is actually the *evidence* for (α, β) becomes

$$p(D \mid \alpha, \beta, \Lambda) = \frac{Z_M(\alpha, \beta)}{Z_W(\alpha) \cdot Z_D(\beta)} \tag{4.16}$$

where $Z_D = \int d^N D \cdot exp(\beta E_D)$ and N is the number of degrees of freedom, i.e., the number of output units times the number of data pairs in D. Ockham's razor is implicitly contained because small values α penalize the large freedom in the prior range of possible values of $\boldsymbol{w}$ by the large value of Z_W. The optimum value of α adjudicates the compromise between fitting the data and simplicity of the model.

Note that it is commonly supposed that the most conservative prior on a parameter space is the uniform prior since that reflects complete ignorance. In fact this choice of prior suffers from a serious deficiency. The uniform priors relative to different parameterizations assign different probabilities to the same set of parameters [Jeffreys 1961, Lee 1989].

4.3.3 Model comparison

The Bayesian model comparison assesses, within a well-defined hypothesis space. how probable a set of alternative models is. Bayesian ranking is carried out by evaluating and comparing the evidence for alternative interpolations. Skilling *et al* [1990] defined the *evidence* of a model $\Lambda = \{\Omega, \aleph, \Re\}$ to be $p(D \mid \Lambda)$. Integrating the evidence for (α, β) in (4.15), we get the evidence for any admissible model Λ

$$p(D \mid \Lambda) = \int p(D \mid \alpha, \beta, \Lambda) \cdot p(\alpha, \beta \mid \Lambda) \cdot d\alpha \cdot d\beta \tag{4.17}$$

However, the misfit function $M(\boldsymbol{w})$ is not generally quadratic for NNs and the existence of multiple minima in NN parameter space complicates model comparison [MacKay 1992b].

4.4 COMPUTATIONAL PHYSICS

Many computational physics problems involve calculations with very large sparse Hamiltonian matrices. Finding all eigenvectors and eigenvalues requires CPU time scaling as $O(N^3)$ and memory scaling as $O(N^2)$ which is impractical. For ground or isolated eigenstates the preferred method is Lanczos diagonalization, which uses only matrix-vector multiply operations and requires CPU and memory scaling as $O(N)$. However, $O(N)$ methods are needed for properties involving many eigenstates such as the density of states (DOS) and spectral functions, and for quantities that can be derived from DOS such as thermodynamics, total energies of electronic structure and forces for molecular dynamics and Monte Carlo simulations. In such applications, limited energy resolution and statistical accuracy are often acceptable provided the uncertainties can be quantified. Maximum entropy [Mead and Papanicolaou 1984, Drabol and Sankey 1993] has been a popular approach to such problems, usually fitting power moments of a DOS or spectral function. Nevertheless, the non-linear convex optimization algorithms required to find ME solutions may be difficult to implement for a large number of power moments and for singular structures in DOS.

4.5 NEUTRON SCATTERING

Neutron scattering is a tool for the study of condensed matter, from superconductors to biological samples. The first applications of ME methods and Bayesian statistics were straight-forward deconvolutions, the work has been extended to make routine use of multi-channel entropy to additionally determine broad unknown backgrounds [Sivia 1989]. A more refined application involves the study of aggregation in a biological sample using Fourier-like data from small angle neuron scattering. Problems envisaged refer to the optimization of instrumental hardware, leading to the derivation of more efficient spectrometers and moderators, which may finally result in a far-reaching revision of ideas on the design of neutron scattering facilities. Practical considerations that arise in running ME methods in finding liquid structure factors from neutron scattering data are presented in a tutorial explanation by Daniell and Potton [1989].

4.6 LANGUAGE MODELING

Recently, ME methods have been successfully introduced to a variety of natural language applications. Language modeling attempts to identify regularities in a natural language and capture them in a statistical model. Language models are crucial ingredients in automatic speech recognition and statistical machine translation systems, where their use is naturally viewed in terms of the *noisy channel* model from information theory [Lafferty and Suhm 1995]. In this framework, an information source emits messages X from a distribution $p(X)$, then the messages enter into a noisy channel and come out transformed into the observable Y according to a

conditional probability distribution $p(Y \mid X)$. The problem of decoding is to determine the message (estimator) $\hat{X}$ having the largest posterior probability given the observation

$$\hat{X} = \underset{X \in \aleph}{arg\,max}\ p(X \mid Y) = \underset{X \in \aleph}{arg\,max}\ p(Y \mid X) \cdot p(X) \tag{4.18}$$

where $\aleph$ stands for the full collection of symbol strings that can be hypothesized by the decoder $arg\ max_{X \in \aleph}$. Thus, Bayesian decoding is carried out using a prior (probability) distribution $p(X)$ on messages, a channel model $p(Y \mid X)$, and a decoder $arg\ max_{X \in \aleph}$. For speech recognition and machine translation, the prior distribution is called a *language model*, and it must assign a probability to every string of symbols from $\aleph$. The most common language models used in today's speech systems are the *n*-gram models based on simple word frequencies. In ME approach, prior information, typically in the form of frequencies, is represented as a set of constraints that collectively determine a unique ME distribution.

However, in virtually all natural language applications, the pay-off for the power of ME method is a considerable increase in computational power and storage capacity requirements.

4.7 MULTIUSER DETECTION AND CHANNEL EQUALIZATION

Multiuser detection is the problem of detecting the data sequence from several simultaneous code division multiple access (CDMA) users and canceling the interference between users. Equalization of intersymbol interference (ISI) is a separate problem of data detection in the presence of interference generated by the channel's effect on the user's own data stream. Both problems can be solved by a maximum likelihood sequence estimation (MLSE) approach such as the Viterbi (VA) algorithm [Proakis 1989]. The code division multiple access communication system, in which many users modulated with special "signature waveforms" share the same transmission bandwidth, is presently considered for use in radio-based data networks. A well-known limitation of the basic system is the self-interference, or near-far effect, in which excessive bit-error-rate (BER) degradation can occur due to reception of "strong" signals from other users. It is the ideal multiuser detector that is free from any such near-far effect and thus does away with the need for power control or other ad-hoc remedies. The multiuser detector theoretically outperforms the conventional matched filter simply because the matched filter is optimal for the Gaussian channel, whereas the CDMA interference is not Gaussian [Verdu 1986].

The problem of multiuser detection is closely related to the problem of channel equalization. Consequently, the Abend and Fritchman (AF) algorithm [Abend and Fritchman 1970], which was originally developed for the equalization problem, may be applied to multiuser detection in order to compute the maximum *a posteriori* probabilities of each possible symbol in a recursive manner. This accords nicely with the Bayesian view of probability since the decision is in favor of the symbol in which we have the greatest belief in being correct [Mailander and Iltis

1993]. The widely spread VA algorithm is driven by likelihood conditioned on every possibility with length equal to the channel memory. It inherently incurs a probabilistic decoding delay and must maintain track of all survivor sequences through the trellis. Alternatively, a "symbol-to-symbol" approach like the AF algorithm, which is based on a Bayesian recursion of posterior probabilities, maximum *a posteriori* (MAP) symbol decisions are made on account of the entire history of past measurements. It offers a fixed decoding delay and no need to maintain survivor sequences. Given the state of modern digital signal processing technology, the additional operations required by the AF algorithm are no longer significant and therefore may be it preferable in some applications.

4.8 GENETIC INFORMATION

Living cells use a chemical code to store information about how to build the protein structural components of the cell and the enzymes that catalyze chemical reactions. The deoxyribonucleic acid (DNA) is a long polymer with 4 kinds of bases, say "a", "c", "g", and "t", strung together by sugar and phosphate groups. In the cell, the DNA is made of two strands twisted around each other so that every "a" on one strand is paired to a "t" on the other strand, and every "g" on one strand is paired to a "c" on the other. The strands have polarity and are read in opposite directions [Schneider 1988]. Along the DNA there are distinct regions called *genes* that have specific functions. The "binding sites" are *patterns* of the bases that indicate the starting and the stopping points for reading each gene. Their names come from the ability of special molecules, called "recognizers", to bind at.

There are many recognizers, such as RNA polymerase, which binds to a pattern called "promoter" near the starting point of genes. Then the RNA polymerase moves along the DNA and copies it by polymerizing bases together to form a strand of RNA. Likewise, the ribosome binds to a "ribosome binding site" on the RNA, and then moves along the RNA while translating the RNA into protein.

The protein that binds to some binding site sequences of DNA is called LexA. To evaluate the information content of the binding sites, a measure of uncertainty $Hs(L)$ of what base to expect in another LexA binding site is introduced as per base [Schneider 1988]

$$Hs(L) = -\sum_{B=a}^{t} f(B,L) \cdot log_2 f(B,L) \tag{4.19}$$

where L is the position in the site, $f(B,L)$ is the frequency of base B at position L and the sum is taken over all bases. As LexA finds sites, which is equivalent with bringing the sequences into alignment, the uncertainty drops for each position. According to Tribus and McIrvine [1971] this difference is a measure of the information contained in the pattern expressed in bits per base

$$Rsequence(L) = 2 - Hs(L) \tag{4.20}$$

By summing along the curve for these differences running across the site we obtain a measure of the total information in the site

$$Rsequence(L) = \sum_{L} (2 - Hs(L)) \tag{4.21}$$

expressed in bits per site. The minimum information needed to find a set of γ sites in the genetic material of an organism (genome) that each has G positions which a recognizer can bind and all positions are equally available for inspection is

$$Rfrequency = log_2 G - log_2 \gamma = -log_2 \frac{\gamma}{G} = -log_2 f \tag{4.22}$$

where f is the frequency of sites in the genome. In most of the cases, the ratio $Rsequence/Rfrequency$ is close to 1 [Schneider 1988]. This suggests that the amount of information in the binding site patterns is generally the amount one would expect given the size of the genome and the number of sites.

Mutations that might occur change the bases of the DNA and tend to increase the uncertainty of the pattern $Hs(L)$. If the uncertainty increases too much, then *Rsequence* will become smaller than *Rfrequency* and the genetic control system lacks full information to find the binding sites. The net effect is that the entropy of the binding sites increases to a maximum determined by the requirement of the organism to function. To conclude with, the entropy of patterns in genetic material tends to increase in the same way that the entropy of isolated physical systems tends to increase in accordance with the second law of thermodynamics. The second may indeed be applied here because there are only 4 possible bases in the DNA, whereas mutations are expressions of changes between these "states".

4.9 BAYESIAN COURT DECISION-MAKING

There have been some attempts to analyze evidential problems in court in a Bayesian manner [Vignaux and Robertson, 1993]. Despite controversial debates on whether Bayesian probabilistic analysis of legal disputes can improve court decision-making, the Bayesian approach is obviously right for analyzing those clearly definable and quantifiable problems that arise in court cases. Most of the objections against the general applicability of probabilistic methods in legal problems arise from the fact that these methods do not capture human thought processes even when these processes are rational. Moreover, many of these objections are based on a purely "frequentist" model of probability.

Typically there are two responses to the difficulties generated by the application of Bayesian methods to the legal form of evidence. The statisticians respond by redefining the question so that it can be answered using "orthodox" frequentist techniques. Alternatively, some lawyers respond that the evidence should be treated "holistically". Although Bayesian decision

theory offers an approach for analyzing and eventually overcoming the real life decision-making problems, it is nevertheless a complex and subtle task.

V. FURTHER TOPICS AND CONCLUSIONS

Any attempt to list the actual and potential fields of applications of Bayesian decision theory and ME principle would be at least exhaustive and questionable. Our review was intended to convey the flavor and basic ideas on reasoning with incomplete or corrupted information and still producing valuable results. Should a problem may be formulated in probabilistic terms, then it is our deep belief that Bayesian reasoning is the richest in inferences, and by using entropic priors it leads to the least committal conclusions. Area of application is limited by our fantasy only. Apart from the issues discussed to some extent above, common topics include economics [Zellner 1990], astronomy and astrophysics [Wilczek and Drapatz 1985], nuclear physics [Fröhner 1990], radar techniques [Heidbreder 1990], expert systems [Lippman 1988], geology and seismology [Pendock and Wedepohl 1991], Murphy's Law [Bordley 1991], and many more. We consider that the great interest on the Bayesian and ME methods is evidently expressed in the sequence of 15 up to now International Workshops on the topics with relevant contributions published in a series of books published by Kluwer Academic Publishers under the generic name of "Bayesian Methods and Maximum Entropy."

Today, the wide and expanding variety of subject matter to which the ME and Bayesian methods of inference are being applied constitutes the evidence of a simple truth. The goal is *inference from incomplete information* rather than *deduction*. The success of predictions based on our present state of knowledge is undoubtedly flattering, but we have learned not too much. It is only when our best predictions fail that we acquire new fundamental knowledge [Jaynes, 1990]. As for instance, recent results suggest that quantum mechanical phenomena may be interpreted as a failure of standard probability theory and may be described by a Bayesian complex probability theory [Youssef, 1995]. But all such subtleties are lost on those who do not comprehend the distinction between *deduction* and *inference*, and try to suppress any background information on the ground that is "subjective". As Jaynes excellently concluded, human information is all we have and we had better recognize the fact.

ACKNOWLEDGMENTS

The present work has been done in the ICTP-INFN Microprocessor Laboratory within the framework of the Associate Scheme supported by the *Abdus Salam International Centre for Theoretical Physics*. The encouragement and comments of Prof. Julian Chela-Flores are greatly appreciated. Special thanks to Ms. Stanka Tanaskovic for her exquisite support in collecting data, and to Mr. Italo Birri and Mr. Muhammad Iqbal for their competent and continuous technical support. We also thank to Ms D. Grilli for the carefully reading of the manuscript.

REFERENCES:

Abend, K. and Fritchman, B. (1970), Statistical detection for communication channel with intersymbol interference, *Proc. Of the IEEE*, Vol. **58**, pp. 779-785, May 1970

Abu-Mostafa, Y. S. (1990), Learning from hints in neural networks, *J. Complexity*, **6**, pp. 192-198.

Agmon, A., Alhassid, Y., Levine, R. D. (1979), An algorithm for finding the distribution of maximal entropy, *J. Comp. Phys.*, **30**, p. 250.

Balasubramanian, V. (1995), Occam's razor for parametric families and priors on the space of distributions, in *Maximum Entropy and Bayesian Methods*, Proc. of the 15th International Workshop, Santa Fe, New Mexico, USA, 1995 (K. M. Hanson and R. N. Silver, eds.), Dordrecht, Kluwer Academic Publishers, 1996, pp. 277-284.

Bayes, T. (1958), An essay towards solving a problem in the doctrine of chances, *Biometrika*, Vol. **45**, pp. 293-315, reprint of T. Bayes, Phil. Trans. R. Soc. London, 1763, Vol. **53**, pp. 330-418.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag.

Bordley, R. F. (1991), Murphy's Law and noninformative priors, in *Maximum-Entropy and Bayesian Methods*, Proc. of the 11th International Workshop, Seattle, USA, 1991 (C. R. Smith, G. J. Erickson, and P. O. Neudorfer, eds.), Kluwer Academic Publishers, 1992, pp. 445-448.

Bregman, L. M. (1967), The relaxation method of finding the common point of convex sets and its appliocation to the solution of problems in convex programming, *USSR Comput. Math. And Math. Phys.*, 7, pp. 200-217.

Cox, R. T. (1946), Probability, frequency, and resonable expectations, *Amer. J. Phys.*, **14**, pp. 1-13.

Craven, P. AND and Wahba, G. (1979), Optimal smoothing of noisy data with spline functions, *Numerische Mathematik*, **31**, pp. 377-403.

Csiszar, I (1995), Maxent, mathematics, and information theory, in *Maximum Entropy and Bayesian Methods*, Proc. of the 15th International Workshop, Santa Fe, New Mexico, USA, 1995, (K. M. Hanson and R. N. Silver, eds.), Dordrecht, Kluwer Academic Publishers, 1996, pp. 35-50.

Daniell, G. J. and Potton, J. A. (1989), Liquid structure factor determination by neutron scattering – Some dangers of maximum entropy, in *Maximum Entropy and Bayesian Methods*, Proc. of the 8th International Workshop, Cambridge, UK, 1989 (J. Skilling, ed.), Dordrecht, Kluwer Academic Publishers, 1989, pp. 151-162.

of the 8th International Workshop, Cambridge, UK, 1989 (J. Skilling, ed.), Dordrecht, Kluwer Academic Publishers, 1989, pp. 151-162.

Daniell, G. J. (1994), Of maps and monkeys: an introduction to the maximum entropy method, in *Maximum Entropy in Action* (B. Buck and V. A. Macaulay, eds.), Clarendon Press, Oxford, pp. 1-18.

Djafari, A. M.- (1992), Maximum entropy and linear inverse problems. A short review, in *Maximum Entropy and Bayesian Methods*, Proc. of the 12th International Workshop, Paris, France, 1992 (A. M.-Djafari and G. Demoment, eds.), Dordrecht, Kluwer Academic Publishers, 1993, pp. 253-264.

Djafari, A. M- (1995), A full Bayesian approach for inverse problems, in *Maximum Entropy and Bayesian Methods*, Proc. of the 15th International Workshop, Santa Fe, New Mexico, USA, 1995 (K. M. Hanson and R. N. Silver, eds.), Dordrecht, Kluwer Academic Publishers, 1996, pp. 135-144.

Drabol, D. A. and Sankey, O. F. (1993), Maximum entropy approach to linear scaling in the electronic structure problem, *Phys. Rev. Letts.*, **70**, p. 3631.

Fischer, R., W. Von Der Linden, and Dose, W. (1995), On the importance of α-marginalization in maximum entropy, in *Maximum Entropy and Bayesian Methods*, Proc. of the 15th International Workshop, Santa Fe, New Mexico, USA, 1995 (K. M. Hanson and R. N. Silver, eds.), Dordrecht, Kluwer Academic Publishers, 1996, pp. 229-236.

Fröhner, F. H. (1990), Entropy maximization in nuclear physics, in *Maximum-Entropy and Bayesian Methods*, Proc. of the 10th International Workshop, Laramie, Wyoming, USA, 1990 (Grandy, W. T. Jr. and Schick, L. H., eds.), Kluwer Academic Publishers, 1991, pp. 93-108.

Gull, S. F. (1988), Bayesian inductive inference and maximum entropy, in *Maximum Entropy and Bayesian Methods in Science and Engineering* (G. Erickson and C. Smith, eds.), Vol. 1, Foundations, Dordrecht, Kluwer Academic Publishers, pp. 53-74, 1988.

Gull, S. F. and Daniell, G. J. (1978) Image reconstruction from incomplete and noisy data, *Nature*, Vol. **272**, pp. 686-690.

Gull, S. F. and Skilling, J. (1984), Maximum entropy method in image processing, *IEE Proc.*, **131 F**, pp. 646-659.

Hanson, K. M. and Wecksung, G. W. (1983), Bayesian approach to limited-angle reconstruction in computed tomography, *J. Opt. Soc. Amer.*, **73**, pp. 1501-1509.

Hanson, K. M. (1987), Bayesian and related methods in image reconstruction from incomplete data, in *Image Recovery: Theory and Application* (H. Stark, ed.), Academic Press, Orlando, pp. 79-125.

Hanson, K. M. (1990), Object detection and amplitude estimation based on maximum a posteriori reconstructions, *Proc. SPIE*, **1231**, 164-175.

Hanson, K. M. (1991), Making binary decisions based on the posterior probability distribution associated with tomographic reconstructions, in *Maximum Entropy and Bayesian Methods* Proc. of the 11th International Workshop, Seattle, WA, 1991, (C. R. Smith, G. J. Erickson, and P. O. Neudorfer, eds.), Dordrecht, Kluwer Academic Publishers, 1992, pp. 313-326.

Heidbreder, G. (1990), Maximum entropy applications in radar, in *Maximum-Entropy and Bayesian Methods*, Proc. of the 10th International Workshop, Laramie, Wyoming, USA, 1990 (Grandy, W. T. Jr. and Schick, L. H., eds.), Kluwer Academic Publishers, 1991, pp. 127-136.

Jaynes, E. T. (1976) in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (W. L. Harper and C. A. Hooker, eds.), Reidel, Dordrecht.

Jaynes, E. T. (1983), *Papers on Probability, Statistics and Statistical Physics* (R. D. Rosenkrantz, ed.) Reidel, Dordrecht.

Jaynes, E. T. (1988), The relation of Bayesian and Maximum Entropy methods, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, Proc. of the 5th, 6th, and 7th International Workshops 1985, 1986, and 1987, (G. J. Erickson and C. R. Smith, eds.), Vol. **1**, Foundations, Dordrecht, Kluwer Academic Publishers, 1988, pp. 25-29.

Jaynes, E. T. (1990), Notes on present status and future prospects, in *Maximum-Entropy and Bayesian Methods*, Proc. of the 10th International Workshop, Laramie, Wyoming, USA, 1990 (Grandy, W. T. Jr. and Schick, L. H., eds.), Kluwer Academic Publishers, 1991, pp. 1-13.

Jeffreys, H. (1939), *Theory of Probability*, (2nd and 3rd editions: 1948, 1961) Oxford University Press, London.

Justice, J. H. (1986), *Maximum-Entropy and Bayesian Methods in Applied Statistics*, Cambridge University Press, Cambridge.

Keren, D. and Werman, M. (1995), Data-driven priors for hyperparameters in regularization, in *Maximum Entropy and Bayesian Methods*, Proc. of the 15th International Workshop, Santa Fe, New Mexico, USA, 1995, (K. M. Hanson and R. N. Silver, eds.), Dordrecht, Kluwer Academic Publishers, 1996, pp. 77-84.

Kullback, S. (1959), *Information Theory and Statistics*, John Wiley and Sons, New York.

Lafferty, J. D. and Suhm, B. (1995), Cluster expansions and iterative scaling for maximum entropy language models, in *Maximum Entropy and Bayesian Methods* Proc. of the 15th International Workshop, Santa Fe, New Mexico, USA, 1995, (K. M. Hanson and R. N. Silver, eds.), Dordrecht, Kluwer Academic Publishers, 1996, pp. 195-202.

Laplace, P. S. (1951), *A Philosophical Essay on Probabilities*, unabridged and unaltered reprint of Truscott and Emory translation, Dover Publications, Inc., New York, 1951, original publication data 1814.

Lee, P. (1989), *Bayesian Statistics: An Introduction*, Oxford University Press, London.

Lippman, A. (1988), A maximum entropy method for expert system construction, in *Maximum Entropy and Bayesian Methods in Science and Engineering* (G. J. Erickson and C. R. Smith, eds.), Vol. 2: Applications, Kluwer Academic Publishers, pp. 243-264.

MacKay, D. J. K. (1991), Bayesian interpolation, *Neural Computation*, **4**, pp. 415-447.

MacKay, D. J. K. (1992a), A practical Bayesian framework for backpropagation networks, *Neural Computation*, **4**, pp. 448-472, 1992.

MacKay, D. J. K. (1992b), Bayesian methods for adaptive models, *Thesis*, California Institute of Technology, Pasadena, California, 1992.

MacKay, D. J. K. (1994), Bayesian non-linear modelling for the prediction competition, *ASHRAE Transactions*, V. 100, Pt. 2, pp. 1053-1062, Atlanta, Georgia. ASHRAE, 1994.

MacKay, D. J. K. (1998), Bayesian methods for neural networks: Theory and applications, David MacKay Webpage http://131.111.48.24/mackay/abstracts/

Mailander, L. and Iltis, R. A. (1993), A common Bayesian approach to multiuser detection and channel equalization, in *Maximum Entropy and Bayesian Methods*, Proc. of the 13th International Workshop, Santa Barbara, California, USA, 1993 (Heidbreder, G. R., ed.), Kluwer Academic Publishers, 1996, pp. 365-374.

Mead, L. R. and Papanicolaou, N. (1984), Maximum entropy in the problem of moments, *J. Math. Phys.*, **25**, p. 2404.

Molina, R. and Katsaggelos, K. (1994), On the hierarchical Bayesian approach to image restoration and the iterative evaluation of the regularization parameter, *Visual Communications and Image Processing '94* (A. K. Katsaggelos, ed.), Proc. SPIE 2308, pp. 244-251.

Mutihac, R. (1992), Hardware Implementation of Neural Networks, *Proc. 4th Conference on Systems, Automatic Control and Measurements (SAUM)*, Kragujevac, Serbia, Yugoslavia, June 12-13, 1992, pp. 21-29.

Mutihac, R. (1995), VLSI Design of Neural Control Systems, *Proc. 5th Conference on Systems, Automatic Control and Measurements (SAUM)*, Novi-Sad, Serbia, Yugoslavia, October 2-3, 1995, pp. 121-128.

Mutihac, R. (1996), Advances in neural integrated circuits featuring parallel distributed processing, *Romanian Reports in Physics*, Vol. **48**, Nos. 9-10.

Mutihac, R. and Colavita, A. A. (1997a), Bayesian Maximum Entropy Approach to Image Reconstruction; Part I: The Inverse Problem of Image Reconstruction and the Bayesian MaxEnt Approach, *Romanian Reports in Physics*, Vol. **49**, 9-10.

Mutihac, R., Colavita, A. A., Cicuttin, A., Cerdeira, A. E., and Candusso, M. (1997b), X-Ray Image Improvement by Maximum Entropy, *Proc. of the 13th International Conference on Digital Signal Processing DSP97*, Santorini, Greece, July 2-4, 1997, Vol. II, Image Processing III, Session F4C.4, pp. 1149-1152.

Mutihac, R., Colavita, A. A., Cicuttin, A., and Cerdeira, A. E. (1998a), Maximum Entropy Improvement of X-ray Digital Mammograms, *Proc. of the 4th International Workshop on Digital Mammography IWDM'98*, The Nijmegen University, The Netherlands, June 7-10, 1998, p. 59.

Mutihac, R., and Cicuttin, A. (1998b), Bayesian modeling of neural networks, *Journal of Chaos Theory and Applications* (submitted).

Myers, K. J. and Hanson, K. M. (1990), Comparison of the algebraic reconstruction technique with the maximum entropy reconstruction technique for a variety of detection tasks, *Proc. SPIE*, **1231**, pp. 176-187.

Nashed, M. Z. (1981),Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory, *IEEE Trans. Antennas Propagat AP-29*, pp. 220-231.

Neal. R. M. (1992), Bayesian training of backpropagation networks by the hybrid Monte Carlo method, *Technical Report CRG-TR-92-1*, Dept. of Computer Science, University of Toronto.

Neal, R. M. (1994), Priors for infinite networks, *Technical Report* CRG-TR-94-1, Dept. of Computer Science, University of Toronto, 1994.

Neal, R. M. (1996), *Bayesian Learning for Neural Networks*, Springer-Verlag, NewYork, 1996.

Neal, R. M. (1997), Monte Carlo implementation of Gaussian process models for Bayesian regression and classification" *Technical Report* No. 9702, Dept. of Statistics, University of Toronto.

Pendock, N. and Wedepohl, E. (1991), Application of maximum entropy to radio imaging of geological features, in *Maximum-Entropy and Bayesian Methods*, Proc. of the 11th International Workshop, Seattle, USA, 1991 (C. R. Smith, G. J. Erickson, and P. O. Neudorfer, eds.), Kluwer Academic Publishers, 1992, pp. 397-402.

Proakis, J. G. (1989), *Digital Communications*, McGraw Hill.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986), Learning representations by back-propagation errors, *Nature*, **323**, pp. 533-536.

Schneider, T. D. (1988), Information and entropy of patterns in genetic switches, in *Maximum Entropy and Bayesian Methods in Science and Engineering* (G. J. Erickson and C. R. Smith, eds.), Vol. 2: Applications, Kluwer Academic Publishers, pp. 147-154.

Shannon, C. E. (1948), A mathematical theory of communication, *Bell. Syst. Tech. J.*, Vol. **27**, pp. 379-423, 623-656.

Shannon, C. E. and Weaver, W. (1949), *The Mathematical Theory of Communication*, Urbana, The University of Illinois Press.

Sivia, D. S. (1989), Applications of Maximum Entropy and Bayesian Methods in Neutron Scattering, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, Proc. of the 9th International Workshop, Dartmouth, USA, 1989 (P. F. Fougère, ed.), Dordrecht, Kluwer Academic Publishers, 1990, pp. 195-210.

Skilling, J. (1988), Classic maximum entropy, in *Maximum Entropy and Bayesian Methods*, Proc. of the 8th International Workshop, Cambridge, UK, 1988 (J. Skilling, ed.), Dordrech, Kluwer Academic Publishers, 1989, pp. 45-52.

Skilling, J., Robinson, D. R. T., and Gull, S. F. (1990), Probabilistic displays, in *Maximum Entropy and Bayesian Methods in Science and Engineering* Proc. of the 10th International Workshop, Laramie, Wyoming, USA, 1990 (W. T. Grandy and L. H. Schick, eds.), Dordrecht, Kluwer Academic Publishers, 1991, pp. 365-368.

Skilling, J. (1994), Fundamentals of MaxEnt in data analysis, in *Maximum Entropy in Action* (B. Buck and V. A. Macaulay, eds.), Oxford, Clarendon Press, 1994, pp. 19-40.

Strauss, C. E. M., Wolpert, D. H, and Wolf, D. R. (1993), Alpha, evidence, and the entropic prior, in *Maximum Entropy and Bayesian Methods*, Proc. of the 13th International Workshop,

Santa Barbara, California, USA, 1993 (G. Heidbreder, ed.), Dordrecht, Kluwer Academic Publishers, 1996.

Tikochinsky, Y., Tishby, N. Z. and Levine, R. D. (1984), Consistent inference probabilities for reproducible experiments, *Physics Review Letters*, Vol. **51**, pp. 1357-1360.

Tishby, N., Levin, E., and Solla, S. A. (1989), Consistent inference of probabilities in layered networks: predictions and generalization, *Proc. IJCNN*, Washington, 1989.

Titterington, D. M. (1985), General structure of regularization procedures in image reconstruction, *Astron. Astrophys.*, **144**, pp. 381-387.

Tribus, M. and McIrvine, E. C. (1971), Energy and information, *Sci. Am.*, Vol. **225**, pp. 179-188.

Verdu, S. (1986), Recent progress in multiuser detection, in *Advances in Communication and Signal Processing* (A. Porter and S. Kak, eds.), Springer-Verlag.

Vignaux, G. A. and Robertson, B. (1993), Lessons from the new evidence scholarship, in *Maximum Entropy and Bayesian Methods*, Proc. of the 13th International Workshop, Santa Barbara, California, USA, 1993 (Heidbreder, G. R., ed.), Kluwer Academic Publishers, 1996, pp. 391-399.

Wilczek, R. and Drapatz, S. (1985), A high accuracy algorithm for maximum entropy image restoration in the case of small data sets, *Astron. Astrophys.*, **142**, pp. 9-12.

Zellner, A. (1990), Bayesian methods and entropy in economics and econometrics, in in *Maximum-Entropy and Bayesian Methods*, Proc. of the 10th International Workshop, Laramie, Wyoming, USA, 1990 (Grandy, W. T. Jr. and Schick, L. H., eds.), Kluwer Academic Publishers, 1991, pp. 17-32.

Youssef, S. (1995), Quantum mechanics as an exotic probability theory, in *Maximum Entropy and Bayesian Methods*, (K. M. Hanson and R. N. Silver, eds.), Proc. of the 15th International Workshop, Santa Fe, New Mexico, USA, 1995, Dordrecht, Kluwer Academic Publishers, 1996, pp. 237-244.