# Deconvolution in the Presence of Noise Using the Maximum Entropy Principle* 

S. Steenstrup<br>Physics Laboratory II, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen ©, Denmark.

## Abstract

The main problem in deconvolution in the presence of noise is non-uniqueness. This problem is overcome in the present work by the application of the maximum entropy principle. The way in which noise enters the formulation of the problem is examined in some detail and the final equations are derived in such a way that the various assumptions are made explicit. Some examples of the use of maximum entropy deconvolution on both simulated and real X-ray diffraction data are given.

## 1. Introduction

The deconvolution problem in one or more dimensions is an old one, and has received much attention (see e.g. Jones and Misell 1970). Use of the maximum entropy principle (MEP) has produced remarkable results (Gull and Daniell 1978; Bryan and Skilling 1980; Burch et al. 1983; Skilling and Gull 1984), especially in the context of digital picture restoration.

The aim here is to present a slightly different viewpoint, bringing, it is believed, some of the assumptions and the underlying philosophy into a clearer light. The final equations to be solved turn out, fortunately, to be the same as those used earlier (see e.g. Burch et al. 1983).

The plan of the work is as follows: in Section 2 some general results concerning information theory are listed, followed in Section 3 by a formulation of the problem, from which the basic equations are derived in Section 4. In Section 5 a method of solution is presented and finally some one-dimensional examples are given using as a real case an X-ray energy dispersive diffraction spectrum.

## 2. Information Theory

In the application of information theory to data treatment at least two points need to be considered. One is the information content in a probability density function $P(x)$ on a space $E(P: E \rightarrow[0,1])$ relative to a function $P^{0}(x)$ given a priori on the same space. Kullback (1959), in extending the work of Shannon (1948) and Shannon

* Dedicated to Dr A. McL. Mathieson on the occasion of his 65th birthday.
and Weaver (1949), showed that, based on a few reasonable axioms, there exists a unique and consistent measure of the relative information content $I$ of the probability density $P(x)$ relative to $P^{0}(x)$ :

$$
\begin{equation*}
I\left[P, P^{0}\right]=\sum_{x \in E} P(x) \ln P(x) / P^{0}(x) \tag{1}
\end{equation*}
$$

where the sum is over the space $E$ (becoming an integral if the variable $x$ is continuous).

The second point is to identify the basic variables $x$ of the system under consideration, the space $E$ in which the variables $x$ take values and to express ones lack of knowledge as to the actual value of $x$ in terms of an a priori density $P^{0}(x)$. After an experiment, the remaining lack of knowledge about the system is expressed in terms of the density $P(x)$ and the information gain in the experiment is given by (1). Once such a density $P(x)$ is determined, an estimator of any quantity $g$ depending on the basic variables $x: g=G(x)$ is given by, say, the mean value of $G(x)$ :

$$
g=\langle G(x)\rangle=\sum_{x} G(x) P(x)
$$

Usually an experiment consists of the measurement, with random errors, of some quantities $A_{r}, r=1, \ldots, M$, related by some functional form to the basic variables:

$$
A_{r}=a_{r}(x)+e_{r}
$$

where $e_{r}$ is the (unknown) error. Moreover, the number of unknown quantities $x$ and $e_{r}$ usually exceeds the number known, and only a probability statement can be made:

$$
\begin{equation*}
A_{r}=\left\langle a_{r}(x)\right\rangle+e_{r}=\sum_{x \in E} a_{r}(x) P(x)+e_{r} \tag{2}
\end{equation*}
$$

Adoption of the MEP or the minimum information principle (MIP) for inference then involves the determination of $P(x)$ by minimizing (1), while satisfying (2) in accordance with assumed statistical properties of the errors. One possible statistical property, which is easily incorporated, is to assume that the errors are gaussian with known standard deviations $\sigma_{r^{\prime}}$ and that the square sum

$$
\chi^{2}=\Sigma\left(e_{r} / \sigma_{r}\right)^{2}
$$

follows a chi-squared distribution. However, other possibilities exist and have been used (see e.g. Bryan and Skilling 1980).

The reason for using the MEP has been given very clearly and forcefully by Jaynes (1983). The usual information theoretic argument: 'in any honest data treatment only the known data should be used, and nothing more, hence the minimum information principle', has recently (Shore and Johnson 1980) been supplemented by a proof that the only consistent (in a sense given a specific meaning by Shore and Johnson) algorithm for treating data of the kind described here is equivalent to the MIP.

In principle then a function $P(x)$ in $N$ variables has to be evaluated. It has already been pointed out (Steenstrup and Wilkins 1984), that for certain special spaces $E$, certain prior densities $P^{0}(x)$ and certain forms of the constraint functions $a_{r}(x)$, considerable simplifications of the above procedure are possible. A specific case is now examined.


Fig. 1. Sketch of the experimental set-up.

## 3. Deconvolution

We consider for illustrative purposes a specific experimental set-up as shown in Fig. 1. A 'white' X-ray beam is diffracted by a sample and detected by an energy analysing counter, which counts the photons that are subsequently sorted according to energy by the multichannel analyser. Thus, in principle, this is an energy-dispersive X-ray diffractometer (Staun Olsen et al. 1981). As the resolution in energy-dispersive X-ray diffraction analysis is much lower than in the corresponding angle-dispersive analysis, it is used as an example for deconvolution. The quantity of interest in this experiment is the distribution in energy of the beam as it arrives at the counter. The variables $x$ thus consist of the set $\left\{x_{1}, \ldots, x_{n}\right\}$, with $x_{1}$ the energy of the first photon, $x_{2}$ that of the second and so on. It is actually sufficient, at least from the experimental point of view, to consider $x_{i}$ as a discrete variable in the sense that the energy axis is divided into intervals; $x_{i}$ is then the centre of interval $i$. Such a discretization process corresponds to the actual measurement and we will say that $x_{i}$ belongs to channel $i$. The basic variables then become a set of integers each in the interval $[1, N]$, where $N$ is the number of intervals, and the space $E$ is $[1, N]^{n}$. A simple prior density $P^{0}(x)$ on this space is $P^{0}(x)=N^{-n}$, corresponding to the belief that all $N^{n}$ possible spectra are equally likely. Such a choice does not really express the knowledge that the spectrum is an X-ray diffraction spectrum and not something else; for example, a Rutherford backscattering spectrum. However, the translation of this kind of information into mathematical form is not obvious.

In practice we are not interested in the order in which the photons arrive and a grouping into number of photons $n_{i}$ in channel $i$ is used. This grouping corresponds to a transformation of the variables $x$ into the set

$$
\left\{n_{1}, \ldots, n_{N}\right\}, \quad n_{i} \in[0, n], \quad \sum n_{i}=n
$$

and the prior $P^{0}(x)$, when expressed in terms of these new variables, becomes the multinomial distribution (Levine 1980)

$$
\begin{equation*}
P^{0}(n)=N^{-n} W \equiv N^{-n} n!/ n_{1}!\ldots n_{N}! \tag{3}
\end{equation*}
$$

It should be pointed out that the multinomial distribution is a result of our choice to ignore the order of arrival of the photons and is in no way related to the nature of the particles (fermions, bosons or classical Boltzmann particles). The prior (3), underlying the measuring process, can in fact be generalized in at least two ways:
(a) In (3) parameters $q_{1}, \ldots, q_{N}$ can be introduced with $\Sigma q_{i}=1$ and then the equation

$$
\begin{equation*}
\boldsymbol{P}^{0}(\boldsymbol{n})=W q_{1}^{n_{1}} \ldots q_{n}^{n_{N}} \tag{4}
\end{equation*}
$$

can be used corresponding to a non-uniform prior in the the $x$ variables.
(b) The multinomial distribution corresponds to the case where a fixed number of photons are counted. In practice the counting is usually for a fixed time, in which case the total number of counts $n$ becomes a variable and $P^{0}(n)$ a product of Poisson densities:

$$
\begin{equation*}
\boldsymbol{P}^{0}(\boldsymbol{n})=\prod_{j}^{N} \mathrm{e}^{-m_{j}}\left(m_{j}^{n_{j}} / n_{j}!\right) \tag{5}
\end{equation*}
$$

with $m_{j}$ parameters corresponding to $q_{j}$ above, and all equal for the simple case of a uniform $P^{0}(x)$.
For computational purposes (5) is often simpler to use than (4); there is for example no a priori correlation among the $n_{i}$ values if (5) is used, while they are weakly correlated if (4) is used.

In terms of the variables $n$, equation (1) becomes

$$
I\left[P, P^{0}\right]=\sum_{n \in \Omega} P(n) \ln P(n) / P^{0}(n)
$$

where

$$
\begin{aligned}
\Omega & =\left\{\left(n_{1}, \ldots, n_{N}\right) \mid n_{i} \in N, \sum_{i} n_{i}=n\right\} & & \text { in case }(a) \\
& =\left\{\left(n_{1}, \ldots, n_{N}\right) \mid n_{i} \in N\right\} & & \text { in case }(b)
\end{aligned}
$$

The detector and the counting electronics introduce noise and a smearing over counting channels, such that the measured numbers $Y_{k}$ (number of counts in channel $k$ ) are related to the desired quantities $\left\langle n_{j}\right\rangle$ [the mean value of $n_{j}$ taken with respect to $P(n)$ ] by
or

$$
\begin{align*}
Y_{k} & =\sum_{j} R_{k j}\left\langle n_{j}\right\rangle+e_{k}  \tag{6a}\\
& =\sum_{j} R_{k j} \sum_{n \in \Omega} n_{j} P(n)+e_{k} \tag{6b}
\end{align*}
$$

with $R_{k j}$ the elements of a convolution matrix and $e_{k}$ the noise assumed to be independently normally distributed with zero mean and assumed variance $\sigma_{k}^{2}$. Hence, we have

$$
\begin{equation*}
\chi^{2}=\Sigma\left(e_{k} / \sigma_{k}\right)^{2} \leqslant \chi_{\alpha}^{2} \tag{7}
\end{equation*}
$$

with $\chi_{\alpha}^{2}$ the value at the $\alpha$ th significance level of the $\chi^{2}$ distribution.

## 4. Transformed Problem

Our desired result for the vector $\langle\boldsymbol{n}\rangle$ is obtained firstly, by minimizing (1a) with respect to $P(n)$ subject to constraints (6) and (7) and the normalization constraint and, secondly, by calculating the mean value of $n$ with respect to the $P(n)$ determined. It is now shown how these two steps can be combined into one involving $\langle\boldsymbol{n}\rangle$ only.

The constrained minimization follows the standard procedure of introducing Lagrange multipliers $\lambda_{k}$ for each of the constraints (6), one $\mu$ for constraint (7), and one for normalization, and then varying $P(\boldsymbol{n})$ and the errors. In the present context $P(\boldsymbol{n})$ and the errors are unconnected and the variation with respect to $P(n)$ is sufficient. The resulting minimum information density is given by (see e.g. Jaynes 1983)

$$
\begin{equation*}
P^{\mathrm{MI}}(n)=Z^{-1} P^{0}(n) \exp \left(-\sum_{k} \lambda_{k} \sum_{j} R_{k j} n_{j}\right) \tag{8}
\end{equation*}
$$

with

$$
\begin{equation*}
Z=\sum_{n \in \Omega} P^{0}(n) \exp \left(-\sum_{k} \sum_{j} \lambda_{k} R_{k j} n_{j}\right) \tag{9}
\end{equation*}
$$

where the Lagrange multipliers in fact depend on the errors. Now using the fact that $P^{0}(n)$ is multinomial (or Poisson) and that $n_{j}$ enters linearly, i.e.

$$
\exp \left(-\sum_{k} \sum_{j} \lambda R_{k j} n_{j}\right)=\prod_{j}\left\{\exp \left(-\sum_{k} \lambda R_{k j}\right)\right\}^{n_{j}}
$$

$Z$ can be evaluated in closed form and the mean values $\left\langle n_{j}\right\rangle$ can be expressed in terms of $\lambda_{1}, \ldots, \lambda_{N}$, i.e.

$$
\begin{equation*}
\left\langle n_{j}\right\rangle=Z^{-1} n q_{j} \exp \left(-\sum_{k} \lambda_{k} R_{k j}\right) \tag{10}
\end{equation*}
$$

By using these expressions the information $I\left[P^{\mathrm{MI}}, P^{0}\right]$ for this specific case is expressible in terms of $\langle\boldsymbol{n}\rangle$ :

$$
\begin{equation*}
I_{(a)}\left[P^{\mathrm{MI}}, P^{0}\right]=\sum_{i=1}^{n}\left\langle n_{i}\right\rangle \ln \left\langle n_{i}\right\rangle / n q_{i} \tag{11a}
\end{equation*}
$$

The corresponding expression in case $(b)$ is

$$
\begin{equation*}
I_{(b)}\left[P^{\mathrm{MI}}, P^{0}\right]=\sum_{i=1}^{n}\left\{\left(m_{i}-\left\langle n_{i}\right\rangle\right)+\left\langle n_{i}\right\rangle \ln \left\langle n_{i}\right\rangle / m_{i}\right\} \tag{11b}
\end{equation*}
$$

Let us consider now the expressions

$$
\begin{gather*}
I_{(a)}^{\prime}(\boldsymbol{p}, \boldsymbol{q})=n \sum_{i=1}^{n} p_{i} \ln p_{i} / q_{i} ; \quad \Sigma p_{i}=1, \quad p_{i} \in R_{+}^{\prime},  \tag{12a}\\
I_{(b)}^{\prime}(f, \boldsymbol{m})=\boldsymbol{\Sigma}\left\{\left(m_{i}-f_{i}\right)+f_{i} \ln \left(f_{i} / m_{i}\right)\right\} ; \quad f_{i} \in R_{+}^{\prime} \tag{12b}
\end{gather*}
$$

In general $I_{(a)}^{\prime}$ or $I_{(b)}^{\prime}$ is not an information measure, but if $p_{i}=\left\langle n_{i}\right\rangle / n$ (or $f_{i}=\left\langle n_{i}\right\rangle$ ) its value equals $I\left[P^{\mathrm{MI}}, P^{0}\right]$ and it corresponds to a derived measure for information under the given particular conditions. A direct calculation now shows that minimizing
$I^{\prime}(p, q)\left[\right.$ or $\left.I^{\prime}(f, m)\right]$ subject to the constraints (6), where

$$
\boldsymbol{Y}_{k}=n \sum_{j} R_{k j} p_{j}+e_{k} \quad \text { or } \quad Y_{k}=\sum_{j} R_{k j} f_{j}+e_{k}
$$

and (7) yields $p_{i}=\left\langle n_{i}\right\rangle / n$ (or $f_{i}=\left\langle n_{i}\right\rangle$ ). It is thus sufficient to use this simpler minimization procedure.

This recasting is similar to the one introduced by Burg (1967), who showed that for quadratic constraints, for the entire real line as variable space and a uniform prior, i.e. for the time series problem, the maximization of entropy corresponds to maximization of the integral of the logarithm of the power spectrum.

It is usually claimed that the MEP yields the maximally non-committal answer or that it allows the maximal variability consistent with the data. This claim seems at first at variance with the fact that one particular solution is actually chosen. In the derivation given here this one particular solution is just the mean value of the maximum information probability density. It is this density which is actually the widest, consistent with the data.

The general consistency proof of Shore and Johnson (1980) has recently been specialized to the more specific case based on (12a), a multinomial prior and linear constraints (Tikochinsky et al. 1984).

The final equation ready for numerical solution now follows by minimizing (12) constrained by (6) and (7). Let us take case (b) for example, i.e. we minimize $Q$ with respect to $f_{1}, \ldots, f_{N}$ and $e_{1}, \ldots, e_{n}$ :

$$
\begin{equation*}
Q=I_{(b)}^{\prime}+\sum_{k} \lambda_{k}\left(\sum_{j} R_{k j} f_{j}+e_{k}\right)=\mu \Sigma\left(e_{k} / \sigma_{\cdot k}\right)^{2} . \tag{13}
\end{equation*}
$$

Due to the convex nature of the forms entering $Q$, constraint (7) is in fact an equality constraint (unless $f_{j}=m_{j}, j=1, \ldots, N$ happens to be an interior point in the allowed region). Setting $\partial Q / \partial f_{j}=0$ and $\partial Q / \partial e_{j}=0$, we get

$$
\begin{aligned}
& \partial Q / \partial f_{j}=\ln f_{j} / m_{j}+\sum \lambda_{k} R_{k j}=0, \\
& \partial Q / \partial e_{j}=\lambda_{j}+2 \mu e_{j} / \sigma_{j}^{2}=0 .
\end{aligned}
$$

Eliminating $\lambda_{j}$ and taking $e_{j}$ from (6) we obtain

$$
\begin{equation*}
\ln f_{j} / m_{j}-2 \mu \sum_{k} 1 / \sigma_{k}^{2}\left(Y_{k}-\sum_{l} R_{k l} f_{l}\right) R_{k j}=0 \tag{14}
\end{equation*}
$$

which is indeed identical to the equation used by Gull and Daniell (1978), but the derivation given here emphasizes that the constraints are linear in $f_{l}$.

## 5. Solution of Equation (14)

Since (14) corresponds to the minimization of a convex function, provided that the matrix $\mathbf{A}$ with elements

$$
A_{l j}=\sum_{k}\left(1 / \sigma_{k}^{2}\right) R_{k l} R_{k j}
$$

is positive definite, there exists a unique solution (see e.g. Ortega and Rheinboldt 1970). Two factors limit the useful types of algorithms. One is that the equation is
only defined for positive values of $f_{l}$, making it necessary, for any linearizing method (e.g. the Newton, the steepest descent, and so on), to take special precautions to avoid negative values. The other is the size of the problem; even for a one-dimensional case with, say, 1024 channels, a direct matrix inversion is ruled out, as discussed by Burch et al. (1983).

For one-dimensional cases of the size mentioned a successive over/under relaxation (SOR) type algorithm (Ortega and Rheinboldt 1970) has been found adequate. It is similar to the one described by Wilkins et al. (1983) and termed 'the single pixel approximation' (Wilkins 1983a, 1983b). The algorithm involves obtaining the new iterate $f^{(i+1)}$ from the old $f^{(i)}$ by considering (14) as a single equation in one unknown $f_{l}$, by taking $f_{j}, j \neq l, f_{j}^{(i+1)}, j<l$ and $f_{j}^{(i)}, j>l$, solving this single nonlinear equation 'exactly', and then setting

$$
\begin{equation*}
f_{l}^{(i+1)}=(1-\omega) f_{l}^{(i)}+\omega f_{l}, \tag{15}
\end{equation*}
$$

where $\omega$ is a relaxation factor. At each stage the Lagrange multiplier is updated by

$$
\mu=k\|\nabla I\| /\left\|\nabla \chi^{2}\right\| .
$$

Burch et al. (1983) used $k=1$, however, in the present context it was found that the convergence was faster if we put

$$
k^{2}=\chi^{(i) 2} / \chi_{(a)}^{2},
$$

with $\chi^{(i) 2}$ the actual value of $\chi^{2}$ at iteration number $i$ and $\chi_{(a)}^{2}$ the value aimed at.

## 6. Results

Results of deconvolution with a gaussian point spread function are shown in Fig. 2. Fig. $2 a$ presents the result of a test case. Five delta functions were convoluted with a gaussian, and noise was added producing the thin curve. This was treated as the original spectrum for unfolding with a gaussian with a slightly smaller standard deviation, yielding the thick and dotted curves such that $\chi^{2}$ was at the $95 \%$ and $5 \%$ significance limits, respectively. The positions of the peaks are precisely reproduced, the heights less so.

Fig. $2 b$ shows part of an energy-dispersive X-ray diffraction spectrum [the sample is a uranium sulfide powder at 14.2 GPa (Staun Olsen et al. 1984)]. The raw data yielded the thin curve while, after background subtraction (Steenstrup 1981), the unfolding yielded the thick curve. In this case it is of course not possible to tell whether the resulting structure is real. It is worth noting, however, that there is a phase change in uranium sulfide at around 15 GPa (Staun Olsen et al. 1984) from a f.c.c. structure to an orthorhombic structure. At the phase change the f.c.c. (220) reflection splits into the orthorhombic (200) and (114) reflections, and this splitting is definitely observed at a pressure slightly higher than the one shown. These peaks are the ones located at about channel number 1240. The peak at about 1330 is a merging of (311) and (222) f.c.c. into (204) and (212) orthorhombic. The result, using 500 channels, was reached in nine iterations or 16 s on a Sperry 1100/82 computer.


Fig. 2. Results of deconvolution with a gaussian point spread function for (a) a test case and (b) an X-ray energy-dispersive diffraction spectrum. In (a) the thin curve corresponds to 'measured' data obtained by adding gaussian noise to five gaussians with $\sigma=4$ channels at channel numbers $30,50,60,70$ and 100 with areas 4000 , 6000, 3000, 2500 and 4000 respectively; the thick and dotted curves correpond to MIP deconvolution with a gaussian point spread function with $\sigma=3.8$ channels and $\chi^{2}$ at the $95 \%$ and $5 \%$ limits respectively. In (b) the thin curve corresponds to raw data, while the thick curve is the MIP deconvolution after background subtraction.

## 7. Conclusions

It has been shown that the minimum information principle for a specific situation can be turned into a simpler optimization problem involving only the desired mean values. The usefulness of the MIP in yet another area, with a simple SOR algorithm for numerical solution, was also demonstrated.

## Acknowledgments

The main part of this work was done at the CSIRO Division of Chemical Physics, Australia, of which the hospitality and inspiration is greatly appreciated. Financial support from the Danish Science Foundation is also acknowledged.

## References

Bryan, R. K., and Skilling, J. (1980). Mon. Not. R. Astron. Soc. 191, 69-79.
Burch, S. F., Gull, S. F., and Skilling, J. (1983). Comput. Vision, Graphics Image Process. 23, 113-28.
Burg, J. P. (1967). Maximum entropy spectral analysis. Paper presented at 37th Ann. Int. SEG Meeting, Oklahoma City. [Reprinted (1979) in 'Modern Spectrum Analysis' (Ed. P. G. Childers) (Wiley: New York).]
Gull, S. F., and Daniell, G. J. (1978). Nature 272, 686-90.
Jaynes, E. T. (1983). In 'Papers on Probability, Statistics and Statistical Physics' (Ed. R. D. Rosenkrantz), p. 434 (Reidel: Dordrecht).
Jones, A. F., and Misell, D. L. (1970). J. Phys. A 3, 462-72.
Kullback, S. (1959). 'Information Theory and Statistics' (Wiley: New York).
Levine, R. D. (1980). J. Phys. A 13, 91-108.
Ortega, J. M., and Rheinboldt, W. C. (1970). 'Iterative Solution of Non-linear Equations in Several Variables' (Academic: New York).
Shannon, C. E. (1948). Bell Syst. Tech. J. 27, 379-423; 623-56.
Shannon, C. E., and Weaver, W. (1949). 'The Mathematical Theory of Communication' (Univ. of Illinois: Urbana).
Shore, J. E., and Johnson, R. W. (1980). IEEE Trans. Inf. Theory 26, 26-37.
Skilling, J., and Gull, S. F. (1984). SIAM-AMS Proc. 14, 167-89.
Staun Olsen, J., Buras, B., Gerward, L., and Steenstrup, S. (1981). J. Phys. E 14, 1154-8.
Staun Olsen, J., Steenstrup, S., Gerward, L., Benedict, U., Spirlet, J. C., and Andreetti, G. D. (1984). J. Less Common Met. 98, 291-300.

Steenstrup, S. (1981). J. Appl. Crystallogr. 14, 226-9.
Steenstrup, S., and Wilkins, S. W. (1984). Acta Crystallogr. A 40, 163-4.
Tikochinsky, Y., Tishby, N. Z., and Levine, R. D. (1984). Phys. Rev. Lett. 52, 1357-60.
Wilkins, S. W. (1983a). Acta Crystallogr. A 39, 892-6.
Wilkins, S. W. (1983 b). Acta Crystallogr. A 39, 896-9.
Wilkins, S. W., Varghese, J., and Lehmann, M. S. (1983). Acta Crystallogr. A 39, 47-60.

