# Image restoration with minimum a priori information 

V Yu Terebizh

## Contents

1. Introduction ..... 137
2. Formulation of the problem ..... 139
3. Model of deterministic image blurring ..... 1413.1 Basic assumptions; 3.2 Numerical examples; 3.3 Structure of the feasible estimate region; 3.4 Maximum likelihoodestimate; 3.5 Fisher's matrix and information inequality
4. Image randomness test (IRT)144
4.1 Formulation of the criteria; 4.2 Should the likelihood be maximum?; 4.3 Relation to Shannon's information theory;4.4 A numerical example; 4.5 Kullback-Leibler's distance and the structure of the feasible estimate region (FER);4.6 Insufficiency of the likelihood function; 4.7 Local nature of the inverse problem
5. Informational constraints150
5.1 General form of information inequality; 5.2 The Kolmogorov - Wiener optimal filter; 5.3 Information about theoriginal for Gaussian ensembles; 5.4 A numerical example
6. Occam's razor estimation
6.1 Principal components; 6.2 Analysis of the principal components of the maximum likelihood estimate;155
6.3 Basic features of the image restoration algorithm; 6.4 A numerical example
7. Superresolution
7.1 Rayleigh's problem; 7.2 Image restoration and pattern recognition; 7.3 Analytical results; 7.4 Examples; 7.5 Monte-Carlo simulations159
8. Concluding remarks ..... 165
References ..... 166


#### Abstract

A consistent approach to the image restoration problem is presented, which does not use Bayesian a priori information. Photon noise is taken into account. The unknown object is treated as a multidimensional set of parameters that have to be statistically estimated in an efficient way. The approach is based on an extended notion of feasible estimate (in the sense of information theory) and on Occam's razor rule of choosing the simplest object which is consistent with the data. Occam's rule is applied by transformation to principal components of the inverse (or maximum likelihood) estimate, which are generated by Fisher's information matrix. The same approach can also be applied to various other inverse problems.


There ain't any explanations. Not of anything. All you can do is point at the nature of things.

Robert Penn Warren All the King's Men

[^0]Received 20 October 1994
Uspekhi Fizicheskikh Nauk 165 (2) 143-176 (1995)
Translated by K A Postnov; edited by A Gelbtuch

## 1. Introduction

Image restoration problem consists of finding as complete as possible characteristics of the original object $S_{0}(x)$ by using the observed blurred image $y_{0}(x)$, as well as a given point spread function (PSF) $h\left(x, x^{\prime}\right)$ and statistical properties of stochastic background $\xi(x)$. One-dimensional model of image formation with a noncoherent source is usually described by the equation (see Frieden's review [1])

$$
\begin{equation*}
y_{0}(x)=\int_{a}^{b} h\left(x, x^{\prime}\right) S_{0}\left(x^{\prime}\right) \mathrm{d} x^{\prime}+\xi_{0}(x) \tag{1.1}
\end{equation*}
$$

where $\xi_{0}(x)$ is the background pattern, and the object $S_{0}(x)$ to be found is assumed to be nonnegative within the interval $[a, b]$ :

$$
\begin{equation*}
S_{0}(x) \geqslant 0 \tag{1.2}
\end{equation*}
$$

To keep the notation simple, we consider below only the one-dimensional version of the problem but this does not incur any loss of generality.

Although the problem formulation given above does not encompass all its aspects, it is representative enough to relate the image restoration problem to a broad class of inverse problems of mathematical physics. As the name itself implies, in inverse problems we have to determine the true properties of phenomena from their observed effects. The image restoration problem itself, in addition to being of
interest as such, is attractive by exposing very sharply the main difficulties of the theory of inverse problems, so that the results obtained in this field have general significance. To be specific, I shall deal here almost exclusively with image restoration.

As has been noted many times, every problem involving the interpretation of experimental data is essentially an inverse one. In particular, the image restoration theory is at the root of computer tomography, electron microscopy, radiography, radiolocation, and many branches of optics and geophysics. Inverse problems arise most frequently in astronomy, which has been dealing so far almost exclusively with the interpretation of passive experiments.

The first studies of the inverse problems in the modern context were done at the end of the XIX century by Lord Rayleigh [2], who suggested an iterative procedure for correcting spectral line profiles. Somewhat later Schuster $[3,4]$ introduced the periodogram of a time series as an estimate of its spectral density. Already then, specific difficulties appeared that were associated with attempts to invert the cause-and-effect relation: the Rayleigh iterative process turned out to be divergent, and the Schuster periodogram has an extremely 'jagged' shape familiar to all investigators in the field.

I should point out that in Eqns (1.1) and (1.2) photon noise is not taken into account, so that these equations do not adequately describe the image formation process. An unavoidable indeterminacy due to the photon noise is not only principally important, but also plays a special role because of the known instability of solutions of the majority of inverse problems, which manifests itself in the fact that substantially different originals correspond to almost indistinguishable images. Because of this, solutions obtained with and without allowing for photon noise are not necessarily close. As von Neumann said with reference to analogous matter, if one neglects viscosity in hydrodynamic equations, they will describe the properties of 'dry water' (cited in Ref. [5]).

Let me illustrate this by a simple imaginary experiment [6]. Let us assume that the required object is a point source with an intensity $F$ located at the origin of the coordinates, the PSF is the sum of two Dirac $\delta$-functions symmetrically arranged relative to the origin of the coordinates, and there is no background. This means that the image will appear as a double source with point components. The problem is as follows. Even when $S_{0}$ is not a determinate quantity, such as the average count number, but a random representation of the count number, then, according to Eqn (1.1), we shall always see the components as having equal brightness, although the total brightness can fluctuate. This prediction of Eqn (1.1) contradicts the experiment: in fact, the observed intensity of each component fluctuates randomly and independently of the intensity of the other component relative to the mean value $F / 2$ with a standard deviation close to $\sqrt{F / 2}$.

Nevertheless, even with the simplified formulation of the problem given by Eqns (1.1) and (1.2), the principal difficulty in finding its solution is clear; it is that the particular form $\xi_{0}(x)$ of the stochastic background in the observed image is unknown. Only statistical properties of the background are specified: in particular, its average value $\langle\xi(x)\rangle$, its power spectrum and, possibly, onedimensional distribution density. Attempts to use the average background $\langle\xi(x)\rangle$ or some typical form of back-
ground as $\xi_{0}(x)$ are unsuccessful because of the aforementioned fundamental problem of the instability of inverse solutions with respect to a small change in the parameters.

The instability of the solutions is, in turn, caused by inadequate information about the object contained in its image. 'Information' is here understood in the strict sense of the word as defined by Shannon [7, 8]. Most often it is principally impossible to determine in any detail characteristics of the object that gave rise to the observed image. For this reason, one can understand the attempts made by investigators to supplement the real information they have and obtain a better solution as a result.

Deterministic a priori information about the required object $S_{0}$ rarely goes beyond the condition of its nonnegativeness, so that in solving inverse problems investigators usually rely upon a priori stochastic information in the framework of the Bayesian approach [9, 10]. In general, this approach assumes that the object $S_{0}$, whose blurred image is being analysed, is randomly extracted from a set of objects with known properties. For example, the classical analysis by Kolmogorov [11] and Wiener [12] of the problem of filtration and prediction of time series starts from specifying the object as an outcome of a standard Gaussian random process with a known covariance function. Turchin et al. [13] clearly demonstrated that essentially the same assumption is used in the wellknown papers by Phillips [14], Twomey [15, 16], and Tikhonov [17, 18]. The maximum entropy principle [1922] assumes that the initial ensemble of objects is formed in accordance with the entropy of each of them. Obviously, specifying the initial ensemble determines to a large degree the solution to the inverse problem. If the investigator has no a priori information of any kind, the solution has almost always a characteristic 'oscillating' form, which testifies that it is unstable.

Using Bayesian a priori information seems to be quite natural for the problem of filtration and prediction of time series, for which the Kolmogorov-Wiener theory was formulated; earlier experience often gives grounds for asserting that the expected signal belongs to an ensemble with fully determinate properties. However, such information in image restoration, and in many other inverse problems, can be regarded as an exception to the rule. For this reason use of the Bayesian approach in those cases where the researcher does not know whether indeed a random choice from a given class of objects has been made gives only an illusion that the required solution has been obtained. This conclusion is difficult to avoid, if for no other reason that already many schools of 'Bayesians' exist, which a priori prefer different ensembles of objects (even followers of the maximum 'entropy' principle use different definitions for this function).

Thus, a situation arises when the inverse solution proves to be of little use without the availability of additional information, and substitution of the missing information by some plausible assumptions leads to an apparently acceptable solution, which in fact has a large and, in principle, unknown bias.

The purpose of the present review is to give as simple as possible systematic description of the approach evolved in the past few years that leads to objective results in the field of non-Bayesian approach to the inverse problems, including the image restoration problem [23-33]. The need to
evolve a unique point of view made it necessary to include some classical problems along with recently obtained results; some results presented here are new.

The following three premises are the starting ones.
(1) The unknown object is given as a set of parameters, for which it is necessary to find statistical estimates of minimum scattering for given observational results and a priori information. Consistent formulation of inverse problems in the framework of statistical estimation theory made it possible to classify a number of important aspects and, in particular, to prove the existence of a natural limit to the inverse solution accuracy.
(2) Use of the mean-square measure $L_{2}$ for the likeness of the images in selecting the solution does not nearly exhaust all the possibilities, and here more sophisticated methods of statistical analysis were required. Extension of the feasible solution notion as a statistical estimate satisfying a rigid image randomness test (IRT) leads to maximum reduction of the set of possible solutions. Nevertheless, for most inverse problems, this set is still too large.
(3) Final choice of the solution, is made by taking into account the aforementioned insufficiency of information: in the feasible estimates domain we choose the solution that is the 'simplest' in a certain sense. As might be expected, this choice is based on the principal components of the estimate of the object. The principal components, in turn, are generated by Fisher's information matrix.

We also note that, in order to determine natural limits of the efficiency of the inverse solution, it is useful to extend its a priori information to such an extent that the inverse problem is in fact reduced to the problem of testing statistical hypotheses.

Taken all together, the enumerated premises lead to an algorithm that ensures stability in the solution of the inverse problem without recourse to illusory Bayesian information (of course, after the solution has been obtained, one can interpret it from the point of view of the Bayesian approach for a specially selected ensemble).

The most important step in finding stable inverse solutions consists of the actual use of the condition of their maximum simplicity consistent with observational data. Apparently, W Occam was the first to clearly formulate the maximum simplicity requirement of a model for interpretating data, in the XIV century: "Plurality is not to be assumed without necessity". Thus, it is proper to call the implementation of the non-Bayesian approach considered here the Occam's razor approach.

## 2. Formulation of the problem

We introduce the following notations (Fig. 1). Let $S$ be an arbitrary element from the object space $\{S\} ; S_{0}$ the object to be found that belongs to the same space; $f(y \mid S)$ the conditional probability (defined by the image formation model) of obtaining a given set of counts for the original element $S$ in the image space $\{y\} ; q(S)=\langle Y f(Y \mid S)\rangle$ the average image that corresponds to determinate blurring $S$ and background averaging; and $y_{0}$ the observed image.

The sign $\langle\ldots\rangle$ means averaging over the probability ensemble. Object space elements $\{S\}$ are $n$-dimensional vector-columns, and image space elements $\{y\}$ are analogous $m$-dimensional vectors (usually $m \neq n$ ). The norm $\|x\|$ of vector $x$ is understood to be euclidian length. The fact


Figure 1. Schematic representation of the image formation model.
that a random variable $\xi$ obeys a distribution with density $f(x)$ is, for simplicity, written as $\xi \sim f(x)$. For example,

$$
\begin{equation*}
Y \sim f(y \mid S), \quad Y_{0} \sim f\left(y \mid S_{0}\right) \tag{2.1}
\end{equation*}
$$

and the observed image $y_{0}$ is the realisation of a random vector $Y_{0}$. For an arbitrary object $S$, the following normalisation condition is satisfied

$$
\begin{equation*}
\sum_{y} f(y \mid S)=1 \tag{2.2}
\end{equation*}
$$

Note that the random variable $\xi$ and the value $x$ it takes on are different notions, and their confusion has often led to an ambiguity in physical literature.

Of fundamental importance to us is discrete representation of the problem, and we regard continuous description as an approximation that sometimes simplifies computations.

To focus on more important problems in this context, we restrict our discussion to noncoherent sources of radiation. In this case, $S_{0}$ usually represents the set of mean radiation intensities that would be registered by individual pixels of a detector with an ideal forming system. It is obvious in this case that the counts $S_{0}$ are nonnegative. Moreover, part of the parameters can describe structural properties of the object, and then their nonnegativeness is not a necessary condition [34, 35, 25]. The theoretically most efficient way of recording an image, consisting of counting individual photoevents, is now most often used in experimental practice (see, for example, Refs [36-38]). Therefore, the components of vector $y$ in the image space are treated as nonnegative integers.

For a linear model of image formation model and for a noncoherent source we have

$$
\begin{equation*}
q(S)=H S+\gamma \tag{2.3}
\end{equation*}
$$

where the matrix $H=\left[h_{j k}\right]$ with nonnegative elements is the PSF, and the vector $\gamma$ with components $\gamma_{j}=\left\langle\xi_{j}\right\rangle \geqslant 0$ determines the mean value of the stochastic background. When exposure time significantly exceeds the radiation coherence time, the statistics of photoevents can be approximated with very good accuracy by the Poisson law [36-38], so that

$$
\begin{equation*}
f(y \mid S)=\prod_{j=1}^{m} \exp \left[-q_{j}(S)\right]\left[q_{j}(S)\right]^{y_{j}} \frac{1}{y_{j}!} \tag{2.4}
\end{equation*}
$$

The presence of photon noise is taken into account in this model by the probabilistic way of its representation; count fluctuations do not vanish even when the external background $\gamma$ is zero.

Generally speaking, some of the requirements listed above are not necessary, and one could choose a more general image formation model, in particular, take into account nonlinear effects and introduce the precise CoxMandel distribution [39-41] for the statistics of photoevents. The case when information about the statistics of photoevents is absent and some other models are discussed in Refs [25, 27]; a realistic model of registration with a CCD-detector that takes into account its inhomogeneity is investigated in Ref. [42]. However, we shall not complicate the presentation here, so as to be able to focus on the principal difficulties.

Our problem is to find an estimate $S^{*}\left(y_{0}\right)$ that differs from the original $S_{0}$ as little as possible, using the observed image $y_{0}$, a priori information about the original $S_{0}$, and the given image formation model $f(y \mid S)$. Since the rule of estimation (in other words, of image restoration) must apply not only to $y_{0}$, but also to any other image $Y_{0}$ generated by the original $S_{0}$, the estimate $S^{*}\left(Y_{0}\right)$ is a random variable, and its quality is determined by a degree of concentration of the distribution density $S^{*}$ around the point $S_{0}$. In particular, we would like to have an estimate without excessive systematic deviations from the original object (that is, with a small bias)

$$
\begin{equation*}
b\left(S_{0}\right)=\left\langle S^{*}\left(Y_{0}\right)\right\rangle-S_{0} \tag{2.5}
\end{equation*}
$$

and a small variance relative to the mean value. These requirements can be combined [43] by defining the meansquare measure of deviation of the components of $S^{*}\left(Y_{0}\right)$ from the original object $S_{0}$ as

$$
\begin{align*}
& \Omega_{i k}\left(S_{0}\right)=\left\langle\left(S_{i}^{*}-S_{0 i}\right)\left(S_{k}^{*}-S_{0 k}\right)\right\rangle, \quad i, k=1, \ldots, n \\
& \Omega\left(S_{0}\right)=\left[\Omega_{i k}\left(S_{0}\right)\right]=\left\langle\left(S^{*}-S_{0}\right)\left(S^{*}-S_{0}\right)^{\mathrm{T}}\right\rangle, \tag{2.6}
\end{align*}
$$

where T denotes transposition operation. The matrix $\Omega$ is known as the matrix of scatter of the estimate $S^{*}$. For unbiased estimates, $\Omega$ coincides with the variance matrix $D$.

Let a unit vector $a$ specify a direction in the object space. Then the dispersion of the estimate $S^{*}$ in this direction (a scalar value) can be represented by the quadratic expression:

$$
\begin{equation*}
\Omega_{a}=\left\langle\left[\left(S^{*}-S_{0}\right) a\right]^{2}\right\rangle=\sum_{i} \sum_{k} \Omega_{i k} a_{i} a_{k}=a^{\mathrm{T}} \Omega a \tag{2.7}
\end{equation*}
$$

The estimate $S^{*}\left(Y_{0}\right)$ is called an effective estimate in class $K$, if its dispersion in an arbitrary direction does not exceed the dispersion of any other estimate for all objects belonging to class $K$. In other words, the effective estimate is characterised by the most compact scattering ellipsoid around any object from class $K$.

The reference to a specific class of objects is important because without this specification, the fundamental notion of estimate efficiency introduced by Fisher [44] becomes meaningless. Indeed, by fixing some point $S_{\mathrm{c}}$ in the object space as an estimate for $S^{*}$, we get an inadmissibly large dispersion for almost all objects other than $S_{\mathrm{c}}$, but just for this object the dispersion of our estimate will be zero (a clock that has stopped is more precise than any other clock once or twice each day!). Usually, the evaluation is carried out in class $K_{b}$ of estimates with a given bias (2.5), in
particular in class $K_{0}$ of estimates with no bias. Despite the importance of these concepts for correct formulation of the inverse problem, we cannot go into this here in greater detail. A clear presentation of the related problems can be found in Borovkov's handbook [43].

Obviously, there are many ways of object evaluation (in other words, many ways of solving the inverse problem), but only those estimates are of real interest that either coincide with the efficient ones, or are close to them. Finding such estimates is the objective of solving any inverse problem. Because of the unavoidable stochasticity of description of internal and external noise, only by formulating the inverse problem in the framework of the statistical theory of parameter testing (and of the closely connected theory of decision making) can one approach the essence of the problem.

The meaning of the notion of 'estimate quality' must be defined by the observer, depending on specificity of the field the problem relates to. The dispersion $\Omega$ defined above is, in a more general context, the conditional risk of the estimate on the assumption of a quadratic loss function. Other definitions of estimate quality are also used. For example, according to the minimax approach, the optimal estimate is that for which the largest dispersion in the object space does not exceed the maximum dispersion for any other estimate (figuratively speaking, when looking for the shortest soldier, we choose the sub-unit in which the right-flank man is shorter than in all the other sub-units). As specific choice does not play a decisive role, we will keep the simple definition of quality based on the scattering matrix.

The kinds of a priori information are so diverse that sometimes it is hard to formalise them. Toraldo di Francia [45] described this state of affairs in an exhaustive way: 'The observer is always more or less relying on his past experience of what a real object can look like. Moreover, in the great majority of particular cases, he has at his disposal a much larger amount of a priori information about the object than he even realises. This information, if properly utilised, enables him to rule out some of the different objects which could correspond to the image. He thus may have the illusion that he can extract from the image more information than there is actually contained".

Part of the a priori information is included in the problem by constructing a stochastic model for image formation (usually this is the condition of nonnegativeness of the solution, an explicit form of the PSF, mean background level, and knowledge of event statistics). If the observer has at his disposal additional determinate or random a priori information, it should be included in the scheme at the appropriate stage. For example, information about 'smoothness' of the solution can be taken into consideration by narrowing down the domain of feasible estimates. If (1) the object $S_{0}$ belongs to a probability ensemble with a density $w(S)$, (2) it was chosen in a nonselective way from this ensemble and (3) the density $w(S)$ itself is known to the investigator, then one should turn to the scheme of Bayesian estimation of the parameters and introduce a two-dimensional density distribution $f(y, S)=w(S) f(y \mid S)$. Here we restrict ourselves to a discussion of the situation when a priori information has a deterministic character. Extension of the approach to the case with additional stochastic information is straightforward.

## 3. Model of deterministic image blurring

What has been said in Section 1 about the role of the photon noise means that the model of deterministic image blurring is of interest largely from the point of view of methodology: its analysis will provide a basis for further discussion of more realistic image formation models. In particular, the deterministic model enables one to trace the instability phenomenon under the simplest conditions. In this connection, it is necessary to draw attention to the widespread fallacy that the instability does not appear in problems with a finite (and all the more so with a small) number of estimated parameters. In fact, instability of solutions of the inverse problem fully manifests itself even when there are only two parameters to be determined.

### 3.1 Basic assumptions

If the randomness of data is due only to the background $\xi$, we get, instead of Eqns (2.1)-(2.4), a model in the form of a system of linear equations

$$
\begin{equation*}
Y_{0}=H S_{0}+\xi \tag{3.1}
\end{equation*}
$$

with the observed image $y_{0}$ being a sum of the deterministic term $H S_{0}$ and the stochastic background realisation $\xi_{0}$ (see Fig. 1):

$$
\begin{equation*}
y_{0}=H S_{0}+\xi_{0} . \tag{3.2}
\end{equation*}
$$

The matrix $H$, with the size $m \times n$, is assumed to be regular, i.e. $\operatorname{det} H \neq 0$. To simplify the model further, we shall ignore for a while the condition of the object, image, and background nonnegativeness, and assume only that $\left\langle\xi_{j}\right\rangle \equiv 0$ and the background counts in the individual pixels are uncorrelated and have an equal variance $\sigma^{2}$ which is known to the observer. The stochastic background is thus characterised by the following covariance matrix

$$
\begin{equation*}
c_{j s} \equiv\left\langle\xi_{j} \xi_{s}\right\rangle=\sigma^{2} \delta_{j s}, \quad C=\left[c_{j s}\right]=\sigma^{2} E_{m}, \tag{3.3}
\end{equation*}
$$

where $\delta_{j s}$ is the Kronecker symbol and $E_{m}$ is an $m-$ dimensional identity matrix.

The principal divergence of the present problem from the classical least square scheme $[46,47]$ is that $\sigma^{2}$ is considered in the latter scheme as an unknown variable which is to be estimated along with the object $S_{0}$. The parameter $\xi$ is taken to be a random measurement error which takes care of the discrepancies between the model and the data that cannot be systematically explained. In contrast, in image restoration, $\boldsymbol{\xi}$ is a stochastic background whose properties can be readily studied in specially designed experiments. As we will see below, this fact substantially influences formulation of the problem and, of course, the results obtained.

### 3.2 Numerical examples

We consider a system of two simultaneous linear equations with two unknowns

$$
\begin{align*}
& x_{1}+2 x_{2}=4, \\
& 3 x_{1}+8 x_{2}=14, \tag{3.4}
\end{align*}
$$

the exact solution to which is the vector $[2,1]^{\mathrm{T}}$. The practical value of the exact solution obviously depends on its stability relative to unavoidable small errors in the coefficients and constant terms of the system. We restrict our discussion to uncertainties of the last kind. Equations (3.4) define the straight lines 1 and 2 in Fig. 2, which


Figure 2. Graphic interpretation of the system of equations (3.4).
are perpendicular to the vectors $[1,2]^{\mathrm{T}}$ and $[3,8]^{\mathrm{T}}$, respectively. As the angle between these vectors is small ( $\simeq 6^{\circ}$ ), the straight lines are nearly parallel to each other, and their intersection point, the solution to the system, proves to be very sensitive to possible errors in the representation of the right-hand sides of the equations.

Of main interest for us is the fact that the solution error strongly depends on the direction in the object space $\left(x_{1}, x_{2}\right)$. Indeed, by making a small change in the relative position of the lines, their intersection point shifts significantly along the lines but very little in the perpendicular direction. If one decomposes the solution radius-vector into two components, one along the average direction of the two straight lines and one normal to it, it becomes obvious that the longitudinal component of the solution can vary in a very wide range, without affecting the uncertainty of the given data. A more rigorous analysis in Sections 4 and 6 shows that the fixed overall uncertainty of the right-hand sides of the system Eqns (3.4) corresponds to an ellipse in the solution space with a ratio of the principal axes of 39 . For example, by substituting into (3.4) a test solution $[5.6 ;-0.4]^{\mathrm{T}}$ instead of the exact solution $[2.1]^{\mathrm{T}}$, we obtain $[4.8 ; 13.6]^{\mathrm{T}}$ as a constant term, which is very close to the values given in equations (3.4).

A natural description of the anisotropy of the error can be obtained by passing over to coordinate systems $\left(E_{1}, E_{2}\right)$ or $\left(P_{1}, P_{2}\right)$ tied to the principal axes of the ellipse mentioned above (the solution in the latter coordinate frame is given by the principal components).

Much more impressive examples of instability of the solution of the inverse problem relative to a small change in the data are known. Morris [48] gave an example of a system of equations $y=H x_{0}$ with
$y=\left[\begin{array}{l}23 \\ 32 \\ 33 \\ 31\end{array}\right], \quad H=\left[\begin{array}{rrrr}5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10\end{array}\right], \quad x_{0}=\left[\begin{array}{l}1 \\ 1 \\ 1 \\ 1\end{array}\right]$.

Substitution of the vector $[2.36,0.18,0.65,1.21]^{\mathrm{T}}$ for $x_{0}$ yields here a constant term with components the absolute values of which differ from the values in expression (3.5) by only 0.01 .

Such examples show that, from the practical point of view, solving a system of equations is more naturally associated with statistical theory of parameter estimation rather than with algebra.

We see that attempts to solve inverse problems that are 'equivalent to within noise' can lead to significantly different solutions even when there are only a few parameters to estimate. This applies to an even greater extent to a multidimensional case, for which the inverse solution 'profile' usually has an extremely irregular shape. This is due to the fact that the solution 'tries to explain' all details of the observations, including statistically insignificant noise fluctuations. For this purpose it can 'wander' into very remote localities of an extremely elongated feasible solution region, without affecting the accuracy of representation of the observational data. Indeed, as the operator $H$ strongly smoothes oscillations of the solution profile, one needs to introduce enormous oscillations into the solution to explain even small data fluctuations.

In situations like those described above, it is said that the system of equations is 'ill-conditioned' [49]. This property of the system is not a simple consequence of the smallness of the determinant of matrix $H$. For example, the determinant of system (3.4) is equal to 2 , and in the case of system (3.5) we have $\operatorname{det} H=1$. The point is that the determinant of matrix $H=\left[h_{1}, \ldots, h_{n}\right]$ depends not only on the degree of collinearity of vectors $\left\{h_{k}\right\}$, but also on their lengths. Therefore, a more correct idea about possible instability can be obtained by making obvious renormalisation of the Hadamard inequality [50, 51]. Finally, as was shown by Faddeev and Faddeeva [52, 53], a full description of the extent to which a system of equations is illconditioned is given by the spectrum of the matrix $H^{\mathrm{T}} H$. I shall return to this question in subsequent sections.

### 3.3 Structure of the feasible estimate region

Let $\Gamma$ be a space containing all possible deterministically blurred images (Fig. 3). In other words, $\Gamma=\{H S\}$ is the $H$-transformation of the object space $\{S\}$. In particular, element $H S_{0}$ belongs to $\Gamma$. If $\{S\}$ comprises a comparatively narrow class of functions, then $\Gamma$ represents a specific enough class as well. For example, if $\{S\}$ contains only 'smooth' objects, then $\Gamma$ consists of even 'smoother' functions, as the high spatial frequencies of the objects are cut off by blurring. Usually $\Gamma$ is a subspace of the entire object space. The observed image $y_{0}$ is obtained from $H S_{0}$ by adding the stochastic background realisation $\xi_{0}$, which need not have properties of the elements from $\Gamma$ (for example, need not be 'smooth' or negative). Therefore, $y_{0}$ generally does not belong to $\Gamma$.


Figure 3. Finding the least-square estimate $S_{\mathrm{m}}$.

To characterise acceptability of some test object $S$ for explaining the data $y_{0}$, we choose a so-called misfit (its
meaning is clear from Fig. 3):

$$
\begin{equation*}
\delta^{2}\left(y_{0}, S\right) \equiv\left\|y_{0}-H S\right\|^{2}=\left(y_{0}-H S\right)^{\mathrm{T}}\left(y_{0}-H S\right) \tag{3.6}
\end{equation*}
$$

The quantity $\delta\left(y_{0}, S\right)$ is a distance from $y_{0}$ to an arbitrary element $H S$ of the space $\Gamma$. In the framework of classical approach, the sought estimate of object $S_{0}$ is the element $S_{\mathrm{m}}$ that minimises this distance:

$$
\begin{equation*}
S_{\mathrm{m}}\left(y_{0}\right)=\arg \min _{S} \delta^{2}\left(y_{0}, S\right) \tag{3.7}
\end{equation*}
$$

The solution $S_{\mathrm{m}}$ is called the least-square estimate (LSE). Such formulation of the problem looks quite natural, as the observer, having no knowledge about the random error variance $\xi$, allots it the least possible role in data interpretation. It is easy to find an explicit representation of the LSE through the data vector $y_{0}$.

The discrepancy obviously reaches minimum for element $S_{\mathrm{m}}$ for which $H S_{\mathrm{m}}$ coincides with the projection of $y_{0}$ onto $\Gamma$. Thus we have $\left(y_{0}-H S_{\mathrm{m}}\right) \perp \Gamma$, and for each $S$ the following equality holds:

$$
\begin{equation*}
(H S)^{\mathrm{T}}\left(y_{0}-H S_{\mathrm{m}}\right)=0 \tag{3.8}
\end{equation*}
$$

The arbitrariness of $S$ implies that the operator $H^{\mathrm{T}}\left(y_{0}-H S_{\mathrm{m}}\right)$ is trivial, that is

$$
\begin{equation*}
H^{\mathrm{T}} H S_{\mathrm{m}}=H^{\mathrm{T}} y_{0} \tag{3.9}
\end{equation*}
$$

We have thus arrived at Euler's normal system of equations, which is characterised by a square matrix $H^{\mathrm{T}} H$. If the latter is regular, we finally find

$$
\begin{equation*}
S_{\mathrm{m}}\left(y_{0}\right)=H^{+} y_{0}, \quad H^{+} \equiv\left(H^{\mathrm{T}} H\right)^{-1} H^{\mathrm{T}} \tag{3.10}
\end{equation*}
$$

Matrix $H^{+}$is called pseudo-inverse with respect to $H$. It is defined by the general relationship $H H^{+}=H$, and when $H$ is a square regular matrix, the pseudo-inverse matrix coincides with the inverse matrix $H^{-1}$. Since the solution $S_{\mathrm{i}}=H^{-1} y_{0}$ follows directly from Eqn (3.2) at $\xi_{0} \equiv 0$, it is called the inverse solution. The LSE is thus a generalisation of the inverse estimate to the case when the number of equations does not equal the number of unknowns.

If the equation $H S=y_{0}$ has a solution, then $S_{\mathrm{m}}$ is the vector which has the minimum length. If there is no solution, then $S_{\mathrm{m}}$ minimises the sum of squares of deviations $y_{0}-H S$, and has the shortest length among all the vectors that have this property [54].

As can be easily seen, the LSE is almost always beyond the boundaries of the positive hypersquare in the object space. Indeed, even when the object $S_{0}$ is positive, the observed realisation of the image $y_{0}$ is almost sure to contain also negative components that are due to noise fluctuations. At the same time, it is easy to prove that the LSE applied to the whole set of feasible images $Y_{0}$ is an unbiased estimate, that is

$$
\begin{equation*}
\left\langle S_{\mathrm{m}}\left(Y_{0}\right)\right\rangle=S_{0} \tag{3.11}
\end{equation*}
$$

and its variance matrix is

$$
\begin{equation*}
D\left[S_{\mathrm{m}}\left(Y_{0}\right)\right]=\sigma^{2}\left(H^{\mathrm{T}} H\right)^{-1}=I^{-1} \tag{3.12}
\end{equation*}
$$

where

$$
\begin{equation*}
I \equiv \frac{1}{\sigma^{2}} H^{\mathrm{T}} H \tag{3.13}
\end{equation*}
$$

is an $n \times n$ matrix. We will see below that $I$ represents a particular expression of the familiar Fisher's information matrix [55] applied to the problem under consideration.

The importance of the LSE for the classical scheme stems from the fact that it has the minimum variance among the whole class of linear unbiased estimates, and thus is an efficient one in this class (Gauss-Markov theorem; its validity follows from the information inequality discussed below). The LSE efficiency holds for arbitrary unbiased estimates, on the additional assumption of normality of the background $\xi$. At the same time, the LSE variance is often inadmissibly large, so that one has to use a priori information to obtain a reliable solution.

From expressions (3.6) and (3.9) one can easily obtain the following representation for the misfit of any element $S$ :

$$
\begin{equation*}
\delta^{2}\left(y_{0}, S\right)=\delta^{2}\left(y_{0}, S_{\mathrm{m}}\right)+\left(S-S_{\mathrm{m}}\right)^{\mathrm{T}} H^{\mathrm{T}} H\left(S-S_{\mathrm{m}}\right),( \tag{3.14}
\end{equation*}
$$

where

$$
\begin{equation*}
\delta^{2}\left(y_{0}, S_{\mathrm{m}}\right)=y_{0}^{\mathrm{T}}\left(E_{\mathrm{m}}-P\right) y_{0} \tag{3.15}
\end{equation*}
$$

is the LSE misfit, and $P \equiv H\left(H^{\mathrm{T}} H\right)^{-1} H^{\mathrm{T}}=H H^{+}$is the operator of orthogonal projection onto $\Gamma$. As the second term in the right-hand side of Eqn (3.14) is nonnegative, we once more verify that the misfit reaches minimum when $S=S_{\mathrm{m}}$. Obviously, the degree of closeness of some estimate $S$ to $S_{\mathrm{m}}$ is determined not by the misfit for $S$ itself, but by an appropriately normalised difference of the misfits for $S$ and $S_{\mathrm{m}}$. A measure of the quality of estimate $S$ in the interpretation of image $y_{0}$ is therefore defined by the relation

$$
\begin{equation*}
\rho_{1}^{2}\left(S, y_{0}\right) \equiv \frac{1}{\sigma^{2}}\left[\delta^{2}\left(y_{0}, S\right)-\delta^{2}\left(y_{0}, S_{\mathrm{m}}\right)\right] \tag{3.16}
\end{equation*}
$$

It then follows from the expressions (3.13) and (3.14) that

$$
\begin{equation*}
\rho_{1}^{2}\left(S, y_{0}\right)=\left(S-S_{\mathrm{m}}\right)^{\mathrm{T}} I\left(S-S_{\mathrm{m}}\right) \tag{3.17}
\end{equation*}
$$

where $I$ is Fisher's matrix.
As we have already said, it is unwise to require too small a value of $\rho_{1}^{2}\left(S, y_{0}\right)$ for the desired $S$ estimates under given characteristics of the background $\xi$ : attempts to explain literally all, including statistically insignificant, image fluctua-tions only enhance the instability of the solution. Using the exact terminology, one can say that $\rho_{1}^{2}\left(S, Y_{0}\right)$ is a statistic that characterises $S$, and for acceptable estimates one must not require that the statistic reaches the most extreme of all possible values. It is more relevant to define the feasible estimates by the condition $A \leqslant \rho_{1}^{2}\left(S, y_{0}\right) \leqslant B$, where the constants $A$ and $B$ restrict the region of the most probable values of the statistic $\rho_{1}^{2}\left(S, Y_{0}\right)$. Using relation (3.17), we arrive at the inequality

$$
\begin{equation*}
A \leqslant\left[S-S_{\mathrm{m}}\left(y_{0}\right)\right]^{\mathrm{T}} I\left[S-S_{\mathrm{m}}\left(y_{0}\right)\right] \leqslant B \tag{3.18}
\end{equation*}
$$

which determines the feasible estimate region (FER) as a layer between two ellipsoidal surfaces resembling a hollow melon (Fig. 4). As mentioned earlier, the centre of the ellipsoid almost always lies outside the positive quadrant in $\{S\}$.

We studied the structure of FER for a fixed image $y_{0}$. As regards estimates that are admissible for describing the entire image ensemble $Y_{0} \sim f\left(y \mid S_{0}\right)$, it is useful to introduce a distance in the object space in analogy with expression (3.16) in the form of the relation

$$
\begin{equation*}
\rho^{2}\left(S_{0}, S\right) \equiv \frac{1}{\sigma^{2}}\left\langle\delta^{2}\left(Y_{0}, S\right)-\delta^{2}\left(Y_{0}, S_{0}\right)\right\rangle \tag{3.19}
\end{equation*}
$$



Figure 4. Feasible estimate region derived with the use of the Kullback-Leibler distance (single hatching) and of the image randomness test (double hatching).

Using the known results for the moments of quadratic forms [46], we readily obtain

$$
\begin{equation*}
\rho^{2}\left(S_{0}, S\right)=\left(S-S_{0}\right)^{\mathrm{T}} I\left(S-S_{0}\right) \tag{3.20}
\end{equation*}
$$

Here the extreme closeness to the object $S_{0}$ is not, of course, a defect of the estimate, so that the average for all images in FER is determined by the inequality

$$
\begin{equation*}
\left(S-S_{0}\right)^{\mathrm{T}} I\left(S-S_{0}\right) \leqslant C \tag{3.21}
\end{equation*}
$$

which defines an ellipsoid centred at $S_{0}$, which is quite understandable because the LSE is unbiased. The constant $C$ is determined by the adopted confidence level for feasible deviations.

The principal axes of the ellipsoids (3.18) and (3.12) are directed along the eigenvectors of the matrix $I$, and the lengths of their semiaxes are proportional to $\lambda_{k}^{-1 / 2}$, $k=1, \ldots, n$, where $\lambda_{k}$ are the eigenvalues of the matrix $I$. Thus, the form of FER is determined by the eigenvectors and the eigenvalues of positively defined Fisher's matrix I. Some of the $\lambda_{k}$ for typical inverse problems are extremely small, and then the FER very strongly stretches in the corresponding directions, assuming a filament-like form. Solution of the inverse problem is very inaccurate in these directions, and this extreme prolateness of the FER in fact indicates that the solution is unstable.

I shall discuss in more detail the form and properties of the FER, including the condition of nonnegativeness of the object in Sections 4 and 6.

### 3.4 Maximum likelihood estimate

So far we have made no assumptions about the shape of the background distribution function. If it is known, then in addition to the LSE one can construct other useful solutions to the inverse problem. The basic solution is the maximum likelihood estimate (MLE). By definition, the MLE is an object $\hat{S}\left(y_{0}\right)$ for which the probability $f\left(y_{0} \mid S\right)$ of getting the observed image $y_{0}$, treated as a function of $S$, reaches maximum (Fig. 1):

$$
\begin{equation*}
\hat{S}\left(y_{0}\right)=\arg \max _{S} f\left(y_{0} \mid S\right) \tag{3.22}
\end{equation*}
$$

It will be readily seen that in the case of a model with Gaussian density $f(y \mid S)$ and abandonment of the condition of nonnegativeness of the object, the MLE coincides with the LSE. These estimates remains close for other practically important cases as well, so one can expect the MLE to be
unstable like the LSE when attempts are made to explain statistically insignificant details of the image as precisely as possible.

### 3.5 Fisher's matrix and information inequality

Let us extend assumption (3.3) by considering the background counts $\xi$ to be normally distributed random variables. Then the distribution density of vector $Y_{0}$ in Eqns (3.1) assumes the form

$$
\begin{align*}
f\left(y \mid S_{0}\right)= & {\left[(2 \pi)^{m} \operatorname{det} C\right]^{-1 / 2} } \\
& \times \exp \left[-\frac{1}{2}\left(y-H S_{0}\right)^{\mathrm{T}} C^{-1}\left(y-H S_{0}\right)\right] \tag{3.23}
\end{align*}
$$

whence

$$
\begin{equation*}
-\frac{\partial^{2}}{\partial S_{0}^{2}} \ln f\left(y \mid S_{0}\right)=H^{\mathrm{T}} C^{-1} H \tag{3.24}
\end{equation*}
$$

Under these assumptions, Fisher's information matrix is treated as the average value

$$
\begin{equation*}
I=\left\langle-\frac{\partial^{2}}{\partial S_{0}^{2}} \ln f\left(Y_{0} \mid S_{0}\right)\right\rangle \tag{3.25}
\end{equation*}
$$

and coincides with the right-hand side of expression (3.24), or because of expression (3.3), with the representation (3.13). Accordingly, the inverse Fisher's matrix is $I^{-1}=\sigma^{2}\left(H^{\mathrm{T}} H\right)^{-1}$.

An important inequality for the variance of unbiased estimates arises in the parameter estimation theory, which was first put forward by Fisher [55] and later rigorously proved in Refs [57-61]:

$$
\begin{equation*}
D(S) \geqslant I^{-1} \tag{3.26}
\end{equation*}
$$

Here $S$ is taken to be an arbitrary unbiased estimate, and the matrix inequality $A \geqslant B$ means that the matrix $A-B$ is nonnegatively defined. Most frequently, the inequality (3.26) is called Rao-Kramer's inequality, or information inequality; we shall use the latter name. If for some estimate equality obtains in (3.26), we will call this estimate the boundary estimate. Obviously, any boundary estimate is also efficient, i.e. its variance is less than that for any other estimate; however, in many cases there are no boundary estimates, but efficient estimates are of main interest for the investigator as before. Problems related to information inequality are discussed in more detail in Section 5.

According to the variance matrix (3.12), the LSE is a boundary estimate in the class of unbiased estimates. This means that it is of no use to look for a better unbiased estimate; however, the LSE lies outside the region (3.18), so we have to look for an unbiased estimate with a larger variance, or admit a bias. The latter is preferable, since for a biased estimate one can achieve a smaller dispersion in the sense of expressions (2.6), than for the LSE. As the notion of estimate quality was defined precisely in this sense, such a search seems to be fully justified.

Usually, one can take as a biased estimate the so called ridge estimator $S_{\mathrm{r}}[62,63]$ which coincides with the optimal estimate according to Wiener [12], Phillips [14], Twomey $[15,16]$, and with a regularised solution according to Tikhonov [17, 18]:

$$
\begin{equation*}
S_{\mathrm{r}}=\left(H^{\mathrm{T}} H+\mu E_{n}\right)^{-1} H^{\mathrm{T}} y_{0} \tag{3.27}
\end{equation*}
$$

where $\mu$ is a parameter which is fitted in accordance with a given misfit. Comparison of expressions (3.27) and (3.10) shows that the addition $\mu E_{n}$ shifts the spectrum of the
reverse operator in the positive direction, and the scatter ellipsoid becomes rounder, which signifies increased stability of the problem. One can show [13, 47] that $S_{r}$ is a Bayesian estimate corresponding to the exact a priori knowledge of matrix $R$ in the linear relationship $R S_{\mathrm{r}}+\varphi=0$, where $\varphi$ is a random vector with zero mean value and independent components. Since we are interested in the case when the investigator has not at his disposal substantial a priori information about the required solution, we shall confine ourselves to a few, brief comments concerning the $S_{\mathrm{r}}$ estimate.

The first is connected with Kolmogorov-Wiener's theory of optimal filtration of random processes. This theory postulates explicitly that the object belongs to a Gaussian ensemble of 'signals', the background to an analogical ensemble of 'noises', and the Lagrange multiplier $\mu$ ( $\alpha$ in Tikhonov's notations) is equal to the ratio of the background variance to the signal variance: $\mu=\sigma_{\xi}^{2} / \sigma_{\mathrm{s}}^{2}$. Thus, the stabilising role of a priori information manifests itself very clearly here.

Instead of the identity matrix $E_{n}$ in expression (3.27), Twomey's $[15,16]$ uses an arbitrary nonnegative definite matrix $G$ which minimises the scalar quantity $p \equiv S^{\mathrm{T}} G S$ for a given misfit value. When $G=E_{n}$, the quantity $p=S^{\mathrm{T}} S$ is the total power of the required estimate. In the general case, the solution that is being sought coincides with the point of contact of the scattering ellipsoid and the smallest centred ellipsoid, $S^{\mathrm{T}} G S=$ const.

We draw attention to the fact that the 'regularised' $S_{\mathrm{r}}$ estimate is biased not only in the direction that corresponds to the largest axes of the ellipsoid, but also in the direction of the smallest axes. The latter circumstance is very undesirable, as one should not degrade the combinations of estimated parameters that have been determined with the highest accuracy.

## 4. Image randomness test (IRT)

Two procedures basically underly the approach under discussion.

First, by using statistical considerations, we will narrow down the FER, an approximate form of which was found in Section 3. This is a sufficiently efficient way in the sense that a significant part of solutions that fall into the ellipsoidal layer (3.18) does not pass more sophisticated tests which any acceptable solution must satisfy. Nevertheless, the present statistical criteria insufficiently narrow down the FER, and in practice an undesirable freedom of choice between the different estimates of the object remains. The ultimate estimate choice in the FER is made in the framework of the second of the aforementioned procedures, which stems from information order considerations (Sections 5, 6).

The present section deals with the possibilities provided by the first procedure. We abstain here from using a definite model of data formation and consider $f(y \mid S)$ as being an arbitrary nonpathological function. The Poisson model will be frequently used for the sake of clarity.

### 4.1 Formulation of the criteria

An estimate $S$ has been considered earlier as being feasible when the distance (3.16) between its average projection $q(S)$ onto the image space and the observed image $y_{0}$ is within the limits recommended by statistics (see Fig. 1). In general,
this is quite a reasonable condition, but the concept of misfit (3.6) does not readily take into account all the possible types of differences between two images. For example, a long series of small-amplitude deviations of one sign can yield a misfit that does not stand out in any way, although the investigator, intuitively relying upon more powerful distinction criteria, may consider the test image, and thus the estimate $S$ as well, as being unacceptable.

For this reason one needs to specify explicitly the condition under which the test estimate $S$ is considered as explaining satisfactorily the observed image. This condition was formulated by Veklerov and Llacer [23, 24], and in a more general context, by Terebizh and Biryukov [29].

According to the terminology introduced above, the vector $q\left(S_{0}\right) \equiv\left[q_{j}\left(S_{0}\right)\right]$ is the mean projection of the object $S_{0}$ onto the image space; then the observed image $y_{0}$ itself can be considered as being one of the random projections of the object $S_{0}$. The notions of projections are applicable not only to the object $S_{0}$, but also to its arbitrary estimate $S$ (see Fig. 1).

In general form, the image randomness test (IRT) can be formulated as follows: only those estimates $S\left(y_{0}\right)$ of an unknown object $S_{0}$ are feasible, for which the observed image $y_{0}$ is statistically indistinguishable from a random projection ensemble generated by $S\left(y_{0}\right)$. A more general definition of the FER follows from this: this is the entire population of feasible estimates (in the aforementioned sense) in the object space.

Obviously, in the case of the specific model of image formation discussed in Section 2, the following particular formulation of the IR T applies: only those estimates $S\left(y_{0}\right)$ of an unknown object $S_{0}$ are feasible for which the observed counts $\left[y_{0 j}\right]$ can be considered as being realisations of independent in totality Poisson random variables with mean values equal to $\left[q_{j}(S)\right]$, respectively.

Subject to some conditions the IRT can be split into two requirements: (1) that random variables generating the observed intensity counts $\left[y_{0 j}\right]$ be independent not only mutually, but also in totality, and (2) that these random variables obey Poisson distribution with a mean value vector $\left[q_{j}(S)\right]$.

Some further elucidation of the IRT can be derived from Fig. 5, which shows in an arbitrary form the densities of intensity distribution $p\left[x_{j}, q_{j}(S)\right]$ in individual image pixels for some estimate $S$ (we neglect here the distribution discreteness). The feasible object estimate ensures that the density maxima $\left[p_{j}\right]$ are not too far removed from the observed counts $\left[y_{0 j}\right]$, which can be considered as a multidimensional realisation of Poisson random variables that are independent in totality.

As we shall see below, such a general and quite obvious requirement contained in the IRT allows us to find concrete ways of substantially improving the stability of the inverse problem. First, I shall show that in a multidimensional case an unconditional MLE almost always contradicts the IRT (I shall also show in Section 4.6 that the likelihood function itself does not provide an exhaustive description of solutions that satisfy the IR T).

### 4.2 Should the likelihood be maximum?

Since the maximum likelihood principle was introduced by Fisher [56] as a special method for parameter estimation (in an implicit form it had been used already at the end of the XVIII century), the MLEs have become not only the most


Figure 5. Intensity distribution densities in the image pixels.
popular, but also the most thoroughly theoretically studied. Under certain conditions the MLEs coincide with boundary estimates in the information inequality sense. Particularly good results are obtained in the asymptotic region when many independent realisations of the investigated random variable are available.

Following Fisher, it is accepted to call the conditional density $f(y \mid S)$ a probability if one means the dependence of this function upon $y$, and a likelihood when it is considered as a function of a second argument $S$. According to expressions (2.4) and (3.22), likelihood is a product of partial densities of the distribution

$$
\begin{equation*}
p(n, q)=\exp (-q) \frac{q^{n}}{n!} \tag{2.4a}
\end{equation*}
$$

and the $\operatorname{MLE} \hat{S}(y)$ is found by maximising the likelihood in the region $\{S\}$ selected by a priori information.

Fig. 5 clearly shows why $f\left(y_{0} \mid S\right)$ can be taken as a good measure of the closeness of the observed counts $\left[y_{0 j}\right]$ to the corresponding mean values $\left[q_{j}(S)\right]$ : if such closeness exists, then the product $L_{0} \equiv f\left(y_{0} \mid S\right)$ of partial probability densities (shown in Fig. 5 by thick vertical lines) is not too small compared to $L_{\max }$. The latter corresponds to the product of maximum probabilities near $\left[q_{j}(S)\right]$, which are arbitrarily shown by dashed lines.

Strictly speaking, in the case of image randomness $Y \sim f(y \mid S)$, the likelihood $L \equiv f(Y \mid S)$ is itself a random variable, or, as one says, a statistic, whose observed value $L_{0}$ allows one to judge the generic character of the experimental results. It is clear that small values of $L_{0}$,
$L_{0} \ll L_{\text {max }}$, indicate that the estimate $S$ is unacceptable owing to poor data interpretation. However, as is usual in statistics, the situation when $L_{0}$ is too close to $L_{\text {max }}$ is also inadmissible: it is improbable that a single experiment will yield the extreme value of the statistic $L$. In other words, 'too good' a data explanation conflicts with given statistical properties of the stochastic background and with the presence of photon noise.

The tendency of the MLE $\hat{S}$ to 'explain' literally all details of the observed image independently of their statistical significance leads to a substantial difference between $\hat{S}$ and the sought estimate. The object nonnegativeness condition almost always yields an estimate $\hat{S}_{+}$located at the boundary of the positive hyperquadrant, which differs from the inverse solution $S_{\mathrm{m}} \simeq \hat{S}$ (see Fig. 4); however, in typical inverse problems $\hat{S}_{+}$proves to be an unstable estimate. Therefore, the region $f\left(y_{0} \mid S\right)>C_{1}$ with the constant $C_{1}$ corresponding to a given confidence level $\alpha_{1}$ must be excluded from consideration, as well as the region $f\left(y_{0} \mid S\right)<C_{2}$ that corresponds to test estimates $S$ which are insufficiently close to $S_{0}$ at the confidence level $\alpha_{2}$. In fact, we are repeating here the arguments advanced in Section 3 for a more general approach that does not use the specific data formation model.

The MLE is unsatisfactory when there is obvious lack of information about the object; in contrast, when the number of estimated parameters is small and especially when samples are taken from a large set of data, the MLE often remains attractive.

One thus should expect that for a single image realisation the sample value $\ln L_{0}$ wil be close to the maximum of the statistics of distribution density $\ln L \equiv \ln f(Y \mid S)$. Considering this density as being not too asymmetric (in fact it is close to the $\chi^{2}$-distribution), one can state that the sample value of $\ln L_{0}$ will be close to the mean value $\langle\ln f(Y \mid S)\rangle$. For this reason it is natural to introduce, as suggested in Ref. [29], mean likelihood estimates $\bar{S}\left(y_{0}\right)$ determined by the condition

$$
\begin{equation*}
\ln f\left(y_{0} \mid \bar{S}\right) \simeq\langle\ln f(Y \mid \bar{S})\rangle . \tag{4.0}
\end{equation*}
$$

In the object space, they form a layer around $\hat{S}$ corresponding to the confidence level $\alpha \simeq 1 / 2$. I shall treat these ideas more rigorously in subsequent sections.

### 4.3 Relation to Shannon's information theory

The probability theory gives two definitions of information. The first was introduced by Edgeworth [64, 65] and Fisher [55] and characterises information about unknown deterministic parameters which is contained in a random sample. The second notion, introduced by Shannon [7, 8] in relation with the needs of the communication theory, characterises information about the realisation of one random variable which is contained in the realisation of another random variable stochastically connected with the first one. Under information theory one understands the group of concepts put forward by Shannon. In order not to confuse these notions, we will designate information defined by Fisher by $I$, and that defined by Shannon by $J$.

Let $\alpha$ be a random variable uniformly distributed within a segment $[0,1]$. The information $J$ contained in the statement that the realisation of $\alpha$ is within the interval $[x, x+\varepsilon]$ is equal to $-\log _{2} \varepsilon$ bits. This value corresponds to the number of first signs in the binary representation of $\alpha$
which must be communicated in order to determine its location with an accuracy $\varepsilon$ (see, e.g., Ref. [66]).

We consider now a discrete random variable $\xi$, which can be equal to $0,1, \ldots$ with probabilities $p(0), p(1), \ldots$, respectively. Added up, these probabilities give 1 , of course. Drawing the random variable $\xi$ can be imagined as sampling the uniform variable $\alpha$ within a unit interval composed of segments with lengths $p(0), p(1), \ldots$ If the obtained realisation $\alpha$ occurs in the segment $p(n)$, we assume that $\xi$ is equal to $n$. As was pointed out earlier, in this way we get the information $J_{n}=-\log _{2} p(n)$ about the realisation of $\alpha$. The same value characterises information about the accompanying realisation $\xi$. The mean information

$$
\begin{equation*}
\left\langle J_{n}\right\rangle \equiv H(\xi)=-\sum_{k=0}^{\infty} p(k) \log _{2} p(k) \tag{4.1}
\end{equation*}
$$

is described by Shannon as the entropy of the random variable $\xi$. In what follows, it is convenient to use natural logarithms instead of binary ones; the corresponding information unit is called nat (after natural digit). As is clearly seen, 1 nat $=\log _{2} \mathrm{e} \simeq 1.443$ bits.

An arbitrary image $Y$ is a set of independent onedimensional random variables $\left[Y_{j}\right]$ with partial densities [ $f\left(y_{j} \mid S\right)$ ]. To each of these random variables we can therefore attach the information

$$
\begin{equation*}
J\left(Y_{j} \mid S\right)=-\ln f\left(Y_{j} \mid S\right) \tag{4.2}
\end{equation*}
$$

Because of mutual independence of $\left[Y_{j}\right]$, the information connected with the whole sample $Y$ is
$J(Y \mid S)=\sum_{j=1}^{m} J\left(Y_{j} \mid S\right)=-\ln \prod_{j=1}^{m} f\left(Y_{j} \mid S\right)=-\ln f(Y \mid S)$.
Thus, the total information is equal to the logarithm of the likelihood function taken with the reverse sign. By definition, the mean information is the entropy of the object:

$$
\begin{equation*}
H(S)=\langle J(Y \mid S)\rangle=-\sum_{y} f(y \mid S) \ln f(y \mid S) \tag{4.4}
\end{equation*}
$$

To avoid misunderstanding, we note that the notion of entropy used here does not coincide with that used in the maximum entropy method.

It is now clear that the arguments in favour of some mean and not the maximum likelihood are equivalent to the fact that for a typical random realisation we do not expect the minimum Shannon's information, but that close to the most probable one. In particular, estimates of the mean likelihood $\bar{S}\left(y_{0}\right)$ coincide with mean information estimates which are determined by the condition

$$
\begin{equation*}
J\left(y_{0} \mid \bar{S}\right)=H(\bar{S}) \tag{4.5}
\end{equation*}
$$

One should require condition (4.5) to be fulfilled only within natural statistical fluctuation of the random variable $J(Y \mid S)$ which, as will be shown below, is $\sqrt{m / 2}$ ( $m$ is the number of pixels).

For the most important case in practice, that of Poisson distribution of counts, the entropy in one pixel of the image can be represented in the form
$H_{\mathrm{P}}(q)=q(1-\ln q)+\int_{0}^{1}[1-q x-\exp (-q x)] \frac{\mathrm{d} x}{x \ln (1-x)}$,


Figure 6. Dependence of the entropy of a Poisson random variable on its mean value.
where $q(S)$ is the mean intensity (Fig. 6). Summing contributions of the kind given by expression (4.6) over all pixels yields the total entropy of the image. The term outside the integral in Eqn (4.6) gives the first-order asymptotic when $q \rightarrow 0$. In the opposite case, when $q \gg 1$, we get the known expression for the entropy of a Poissonian random variable: $H_{\mathrm{P}}(q) \simeq \ln \sqrt{2 \pi \mathrm{e} q}$. Another asymptotic that is important for applications is that to Shannon's information variance of a Gaussian random variable:

$$
\begin{equation*}
D(J) \simeq \frac{1}{2}+\frac{1}{4 q}, \quad q \gg 1 . \tag{4.7}
\end{equation*}
$$

In fact, for the satisfactory fulfillment of the asymptotic relationships, only $q>2-3$ is required, so that the entropy of an image stretching over $m$ pixels is $H \simeq m \ln \sqrt{2 \pi \mathrm{e} q}$, and the standard deviation of the information is about $\sqrt{m / 2}$. One can find the proof for these relations in Ref. [29].

As for Shannon's information distribution function, it is easy to show that in the same approximation $q_{j} \gg 1$ the quantity

$$
\begin{equation*}
\theta(Y \mid S) \equiv m+2[J(Y \mid S)-H(S)] \tag{4.8}
\end{equation*}
$$

obeys a $\chi^{2}$-distribution with $m$ degrees of freedom. Agreement of formulas (4.4) and (4.7) with this result becomes quite obvious if one takes into account [61] that $\left\langle\chi_{m}^{2}\right\rangle=m, D\left(\chi_{m}^{2}\right)=2 m$.

Thus, the mean information estimate region is achieved in image restoration by requiring the equality $J \simeq H \pm \sqrt{m / 2}$ to be fulfilled. A more precise formulation of the criterion when to stop the iteration, based upon the accepted confidence level, can be readily indicated since the distribution of the random variable $\theta(Y \mid S)$ is known. Below we put forward arguments for the point of view that the equality $J \simeq H$ must be satisfied not only for the image as a whole (which is rather quickly achieved), but also for separate parts of the image with dimensions of the order of the PSF width (the local mean information principle).

In practice, in the expression for $\ln f(y \mid S)$ it is useful to retain all multipliers that are independent of the investigated object; this allows us not only to measure information in standard units, but also to compare images with each other.

### 4.4 A numerical example

We consider a model object t8 (Fig. 7a) taken as a Gaussian density with standard deviation $\sigma_{\mathrm{ob}}=2$ pixels and total intensity $F=10^{4}$ events [29]. Fig. 7b presents the result of a random blurring of the object with a Gaussian


Figure 7. Gaussian object (a), its image (b), conditional MLE (c), and two estimates of the mean information (d, e).

PSF at $\sigma=3$ pixels with subsequent addition of a stochastic Poisson background with a mean level of $\gamma=100$ counts/ pixel (realisation t8_13). If for the image restoration we make use of the maximum likelihood method with the object nonnegativeness as the only a priori information, we arrive at the estimates $\hat{S}_{+}$shown in Fig. 7c. Figs 7d and 7e show two global estimates of the mean information $\bar{S}$; their comparison with the MLE attests to a significant increase of the stability of the solution.

Fig. 8 shows the information and entropy for the restoration of two random realisations of object t8 by using the maximum and mean likelihood methods. In both cases the information calculated for object $S_{0}$ itself is within the limits of one standard deviation from the mean value, i.e. the entropy. The entropy of the estimates changes comparatively little during the subsequent restoration, since $H(q)$ is a slowly varying function (in general the entropy can be fixed at its sample value). Unlike the entropy, the information of the estimates changes significantly and it is on its final value that the result depends.


Figure 8. Change of the information ( - ) and entropy $(\bullet)$ for simulations t8-1 (a) and t8-13 (b) restored by different methods.

### 4.5 Kullback - Leibler's distance and the structure of the feasible estimate region (FER)

Using Shannon's information $J=-\ln f(Y \mid S)$ as the test statistic leads to quite definite conclusions about the structure of the FER, which I shall now briefly discuss.

In the framework of the pure significance test [67], we consider the null hypothesis that the true object $S_{0}$ coincides with its close estimate $S$, that is

$$
\begin{equation*}
\mathcal{H}_{0} z: \quad S_{0}=S \tag{4.9}
\end{equation*}
$$

If one takes $J(Y \mid S)$ as the test statistic, then as a measure of the 'distance' between the object $S_{0}$ and the test estimate $S$ it is natural to choose the information difference between these two elements averaged over $f\left(y \mid S_{0}\right)$ (see Fig. 1):

$$
\begin{align*}
\rho_{\mathrm{KL}}^{2}\left(S_{0}, S\right) & \equiv\left\langle J\left(Y_{0} \mid S\right)-J\left(Y_{0} \mid S_{0}\right)\right\rangle \\
& =\sum_{y} f\left(y \mid S_{0}\right) \ln \frac{f\left(y \mid S_{0}\right)}{f(y \mid S)} . \tag{4.10}
\end{align*}
$$

The quantity $\rho_{\mathrm{KL}}\left(S_{0}, S\right)$ is an information measure of the distance between the distributions $f\left(y \mid S_{0}\right)$ and $f(y \mid S)$ introduced by Kullback and Leibler [68].

Now let us find Kullback - Leibler's distance for close objects $S_{0}$ and $S$. For this purpose, we expand $J\left(Y_{0} \mid S\right)$ in Taylor's series in the vicinity of the point $S_{0}$; to the accuracy of second-order terms in $\varepsilon \equiv S-S_{0}$ we get
$J\left(Y_{0} \mid S\right) \simeq J\left(Y_{0} \mid S_{0}\right)+g\left(Y_{0} \mid S_{0}\right)^{\mathrm{T}} \varepsilon+\frac{1}{2} \varepsilon^{\mathrm{T}} \phi\left(Y_{0} \mid S_{0}\right) \varepsilon$,
where the components of the vector $g\left(Y_{0} \mid S_{0}\right)$ and the matrix $\phi\left(Y_{0} \mid S_{0}\right)$ are determined by the formulae

$$
\begin{align*}
g_{k} & =\left[-\frac{\partial}{\partial S_{k}} \ln f\left(Y_{0} \mid S\right)\right]_{S=S_{0}} \\
\phi_{i k} & =\left[-\frac{\partial^{2}}{\partial S_{i} \partial S_{k}} \ln f\left(Y_{0} \mid S\right)\right]_{S=S_{0}}, \quad i, k=1, \ldots, n . \tag{4.12}
\end{align*}
$$

When substituting expression (4.11) into formula (4.10) one should take into account the relations

$$
\begin{align*}
& \langle g(Y \mid S)\rangle \equiv 0  \tag{4.13}\\
& \langle\phi(Y \mid S)\rangle=\left\langle g(Y \mid S) g(Y \mid S)^{\mathrm{T}}\right\rangle \equiv I(S) \tag{4.14}
\end{align*}
$$

where $I(S)$ is Fisher's information matrix. The first of these can be derived from the normalisation condition (2.2) if one differentiates it with respect to $S_{k}$ and then multiplies and divides the expression under the summation sign by $f(y \mid S)$. The second equality can be obtained after simple transformations for the case under consideration of a 'good' enough distribution density (see, for example, $\operatorname{Refs}[43,69]$ ). As a result, we find

$$
\begin{equation*}
\rho_{\mathrm{KL}}^{2}\left(S_{0}, S\right) \simeq \frac{1}{2} \varepsilon^{\mathrm{T}} I\left(S_{0}\right) \varepsilon \tag{4.15}
\end{equation*}
$$

which differs from the distance $\rho^{2}\left(S_{0}, S\right)$ defined by formula (3.20) by the factor $1 / 2$, and, what is particularly important, by the dependence of Fisher's matrix on the unknown object $S_{0}$. It will be readily seen that this last circumstance indicates the presence of photon noise in this model, which is not there in the model of deterministic image blurring. This fact makes the image restoration problem essentially a local one, as it should be on simple intuitive grounds.

By specifying the significance level of feasible fluctuations and the corresponding boundary value for Kullback-Leibler's distance of feasible estimates $S$, we arrive at the mean FER in the form

$$
\begin{equation*}
\left(S-S_{0}\right)^{\mathrm{T}} I\left(S_{0}\right)\left(S-S_{0}\right) \leqslant C \tag{4.16}
\end{equation*}
$$

Like the inequality (3.21), this last inequality describes an ellipsoid in the object space with the centre at $S_{0}$.

Since the object $S_{0}$ we are looking for is still unknown, inequality (4.16) is mainly of theoretical interest. In practice, we use information about the sample FER that corresponds to the observed image $y_{0}$. It is easy to show that in this case the sample FER is determined by inequalities similar to the inequality (3.18):

$$
\begin{equation*}
A \leqslant(S-\hat{S})^{\mathrm{T}} \phi\left(y_{0}, \hat{S}\right)(S-\hat{S}) \leqslant B \tag{4.17}
\end{equation*}
$$

where $\hat{S}\left(y_{0}\right)$ is an estimate that corresponds to the unconditional maximum of the likelihood function (see Fig. 4), and the constants $A$ and $B$ are determined by the significance levels of feasible fluctuations. In deriving
inequality (4.17), one should use a distance between $\hat{S}$ and $S$ for a fixed image $y_{0}$ given by the formula:

$$
\begin{equation*}
\rho^{2}\left(\hat{S}, S \mid y_{0}\right) \equiv J\left(y_{0} \mid S\right)-J\left(y_{0} \mid \hat{S}\right)=\ln \frac{f\left(y_{0} \mid \hat{S}\right)}{f\left(y_{0} \mid S\right)} \tag{4.18}
\end{equation*}
$$

make an expansion like (4.11) in the vicinity of $\hat{S}\left(y_{0}\right)$, and take into account that at the maximum point $g(y \mid \hat{S})=0$. The matrix $\phi\left[y_{0}, \hat{S}\left(y_{0}\right)\right]$ with elements

$$
\begin{equation*}
\phi_{i k}\left[y_{0}, \hat{S}\left(y_{0}\right)\right]=\left[-\frac{\partial^{2}}{\partial S_{i} \partial S_{k}} \ln f\left(y_{0} \mid S\right)\right]_{S=\hat{S}} \tag{4.19}
\end{equation*}
$$

can be considered as the sample Fisher's matrix for the given image.

For the Poisson model of image formation defined by the density (2.4), Fisher's matrix has the form [27]

$$
\begin{equation*}
I_{i k}(S)=\sum_{j=1}^{m} \frac{h_{j i} h_{j k}}{q_{j}(S)}, \quad I(S)=H^{\mathrm{T}} Q(S)^{-1} H \tag{4.20}
\end{equation*}
$$

where

$$
\begin{equation*}
Q(S) \equiv \operatorname{diag}\left[q_{1}(S), \ldots, q_{m}(S)\right] \tag{4.21}
\end{equation*}
$$

is a quadratic matrix with $q_{j}(S)$ located on the main diagonal and zero remaining elements. As has been pointed out earlier, a specific feature of inverse problems is that images corresponding to remote objects differ little from each other. For this reason we can replace, with an acceptable accuracy, in expression (4.20) the mean values $q_{j}\left(S_{0}\right)$ by really observed counts $y_{0 j}$, so that

$$
\begin{equation*}
I_{i k}\left(S_{0}\right) \simeq \sum_{j=1}^{m} \frac{h_{j i} h_{j k}}{y_{0 j}} \tag{4.22}
\end{equation*}
$$

Since the $\operatorname{MLE} \hat{S}\left(Y_{0}\right)$ is on average a feasible estimate, we can expect from relation (4.16) that $\hat{S}\left(Y_{0}\right)$ is distributed approximately normally with a variance matrix $I^{-1}\left(S_{0}\right)$. Indeed, let us as an approximation replace the random matrix $\phi\left(Y_{0} \mid S_{0}\right)$ by its mean value $I\left(S_{0}\right)$ in the second-order term of the expansion (4.11), and then differentiate the equality $\varepsilon \equiv S-S_{0}$. Then, taking into account the definition of MLE and the relationship

$$
\begin{equation*}
\frac{\partial}{\partial a} a^{\mathrm{T}} A a=2 A a \tag{4.22a}
\end{equation*}
$$

which is valid for an arbitrary vector $a$ and a symmetric matrix $A$ [54], we find

$$
\begin{equation*}
\hat{S}\left(Y_{0}\right)-S_{0} \simeq-I^{-1}\left(S_{0}\right) g\left(Y_{0} \mid S_{0}\right) \tag{4.23}
\end{equation*}
$$

It follows from relations (4.14) and (4.23) that

$$
\begin{equation*}
D\left[\hat{S}\left(Y_{0}\right)\right]=\left\langle\left[\hat{S}\left(Y_{0}\right)-S_{0}\right]\left[\hat{S}\left(Y_{0}\right)-S_{0}\right]^{\mathrm{T}}\right\rangle=I^{-1}\left(S_{0}\right) \tag{4.24}
\end{equation*}
$$

The last relation extends expression (3.12) to a more general data formation model.

### 4.6 Insufficiency of the likelihood function

As one would expect, by using a Gaussian approximation for the likelihood function, we arrived at a FER which locally has the same form of the ellipsoidal layer as that obtained by using misfit as a measure of distance between estimates (Section 3). It makes no sense to refer to higher orders of the likelihood function expansion, since both metrics (3.16) and (4.18) are insufficient for full description
of possible distinctions between random images generated by the objects that are being compared.

Let us assume, for example, that the original object $S_{0}$ is an extended source with a flat brightness distribution, and the mean background does not change within the format limits. If in a randomly blurred image $y_{0}$ the group of maximum counts over the entire field of view is transferred to adjacent pixels, the resulting picture $y_{0}^{\prime}$ will have the same probability as the real image. It is, however, clear that images $y_{0}$ and $y_{0}^{\prime}$ that are equivalent from the likelihood point of view cannot be considered as being really equivalent, as a statistically significant local brightness increase has to be interpreted by the corresponding peak in the brightness distribution of the object .

Thus, instead of the likelihood function (or in conjunction with it) in the inverse problem solution it is desirable to rely upon some noncommutative statistic that takes into account smooth systematic deviations of the compared images. An analogical need has been felt for a long time in the analysis of classical data [69], but the majority of the statistics used there have an artificial form. The same problem arises in stochastic dynamics and in studies of the quality of pseudo-random number generators.

One could achieve a natural choice of the test statistic by juxtaposing a powerful enough alternative to the null hypothesis (4.9); however, a subjectivity due to the choice of the alternative will then be introduced. For this reason, in comparing the images we shall use for the time being a number of known statistical criteria like those listed in Ref. [70], which are based on studying a sample distribution of series lengths, the power spectrum, uniformity of one- and two-dimensional sample distribution densities for specially transformed counts, etc. Taken together, these tests check a whole range of alternatives without specifying them explicitly. Experience shows that the corresponding FER contraction is quite noticeable, especially in the restoration of objects with an intense high frequency 'tail' in the power spectrum, i.e. with sharp details in their brightness distribution.

It is not appropriate here to consider in detail the related technically complex questions [29, 31], all the more so because no simple and effective statistic which naturally generalises the likelihood has as yet been found. Essentially, one uses now a particular realisation of the IRT which includes a sample of several tests complementing each other. The question of the existence of a unique general statistic that would allow one to test nonrandomness of a given finite number sequence most effectively still remains open.

The notion of complexity (entropy) $K_{v}$ of a sequence of length $v$ which was introduced by Kolmogorov [71, 72] (see also Refs [73-77]) appears to be very attractive in this context. In fact, the image randomness test requires maximum Kolmogorov's complexity of the image generated by a trial object (after standard 'uniformising' data transformation). Thus, by choosing $v$ image fragments, the FER is determined by the condition $K_{v}\left(y_{0} \mid S\right) \simeq v$ with the usual stipulation about the adopted significance level. Theoretical and model studies in this direction would be of great significance.

### 4.7 Local nature of the inverse problem

As the total PSF width $W \simeq 2 \Delta$ is finite, image segments with a length of the order of $W$ pixels can be considered as
'object-independent' image fragments. If several counts (approximately 2 ; the precise number depends on the PSF) are located within each interval with length $\Delta$, the wellknown counts theorem guarantees conservation of all information about the object contained in the image. Image discretisation will be considered in Section 7 in more detail, but now it is sufficient to assume that the minimum image fragments are chosen to satisfy the conditions of the sampling theorem.

Obviously, the preceding analysis can be equally applied to the entire image and its individual fragments. For example, we may try to achieve statistical closeness of information $J\left(y_{0} \mid S\right)$ and entropy $H(S)$ in separate parts of the image. However, statistical significance of local deviations depends on the extent of the whole image, so that the image restoration quality is also determined by the extent of the image.

Indeed, let us assume that a blurred image of a closely spaced binary star is in the field of view. If it occupies a significant part of the entire image format, the observed deviation from white noise is highly significant, and much work is required for restoration in order to establish binary nature of the object. Now let us suppose that the field of view is much larger than the image of the star. Then fluctuations are possible within the boundaries of the whole format with a size and amplitude comparable to those of the star image, and adequate statistical analysis may require no image restoration whatsoever. Essentially the same arguments have been used for a long time in engineering [78] and in astronomy, for example when estimating the limiting stellar magnitude under given observational conditions.

This implies that in restoration of an image with a large format one should specify the general significance level for admissible deviations and then try to make the localdeviation distribution function consistent with this significance level. At the same time, the observer has often at his disposal a priori information about individual object locations within the field of view, or is interested in a particular small region. Then the significance level should be locally specified, and the expected image restoration quality will be higher. An algorithmic implementation of this approach is discussed in Ref. [31].

## 5. Informational constraints

The main reason for the instability of inverse problem solutions became clear quite a long time ago (see, for example, Refs [14, 51, 79]): the investigator does not have the information needed for determining the original object with the required accuracy. However, correct diagnosis is insufficient for effective treatment, and a quantitative description of the situation is also required. This can be partially obtained in the framework of Shannon's information theory, corresponding conclusions of which are considered here in Sections 5.2-5.4. However, at first we need to discuss in greater detail the consequences of Fisher's information inequality briefly mentioned in Section 3.5.

### 5.1 General form of information inequality

Relationship (3.26) relates to arbitrary unbiased estimates, that is estimates which on average coincide with the original. Not disputing the importance of unbiased estimates, in Section 3.5 I presented arguments in favour
of the point of view that in multidimensional estimation problems one can achieve a smaller dispersion of the estimate (2.7), i.e. of the chosen measure of the quality of solution of the problem under consideration, by using an appropriate bias. Therefore let us consider a general form of information inequality for an arbitrary estimate $S^{*}(Y)$ from class $K_{b}$ with some bias vector $b(S)$ defined by formula (2.5). We shall consider an arbitrary element $S$ from the object space which we shall treat as the original.

Let $f(y \mid S)$ be the distribution function of an $m$ dimensional data vector $Y$ that depends on $n$ unknown parameters $S_{1}, \ldots, S_{n}$, and

$$
\begin{equation*}
I_{i k}(S) \equiv\left\langle\frac{\partial}{\partial S_{i}} \ln f(Y \mid S) \frac{\partial}{\partial S_{k}} \ln f(Y \mid S)\right\rangle \tag{5.1}
\end{equation*}
$$

be the elements of Fisher's symmetric matrix $I=\left[I_{i k}\right]$; $i, k=1, \ldots, n$. In the cases which interest us the density $f$ satisfies regularity conditions [43] which permit us to transform the initial definition (5.1) to a form known from expression (3.25)

$$
\begin{equation*}
I_{i k}(S)=\left\langle-\frac{\partial^{2}}{\partial S_{i} \partial S_{k}} \ln f(Y \mid S)\right\rangle \tag{5.2}
\end{equation*}
$$

Further, let $B(S)$ be a quadratic matrix with components

$$
\begin{equation*}
B_{i k}(S) \equiv \frac{\partial}{\partial S_{k}} b_{i}(S) \tag{5.3}
\end{equation*}
$$

The scatter matrix $\Omega(S)$ of the estimate $S^{*}$ is defined by relation (2.6), and matrix $\omega$ of size $n \times n$ is specified as follows:
$\omega(S) \equiv\left[E_{n}+B(S)\right] I^{-1}(S)\left[E_{n}+B(S)\right]^{\mathrm{T}}+b(S) b(S)^{\mathrm{T}}$.
Then for any estimate $S^{*}$ that belongs to class $K_{b}$ the following information inequality holds:

$$
\begin{equation*}
\Omega(S) \geqslant \omega(S) \tag{5.5}
\end{equation*}
$$

where the matrix inequality is equivalent to the matrix $\Omega-\omega$ being definitely nonnegative. Geometrical sense of inequality (5.5) is that the $n$-dimensional scattering ellipsoid for any estimate does not penetrate in any direction into the ellipsoid defined by matrix $\omega$. For unbiased estimates $b(S) \equiv 0, \omega(S)=I^{-1}(S)$, and $\Omega(S)$ coincides with the variance matrix $D\left[S^{*}\right]$, so that relation (5.5) takes the form (3.26). The information inequality is a consequence of the well-known Cauchy-Bunyakovsky-Shwartz inequality written in a matrix form [43].

From relation (5.5) we find the mean-square scatter of individual components of the estimate $S_{k}^{*}$ about the true values $S_{k}$

$$
\begin{equation*}
\Omega_{k k}(S)=\left\langle\left(S_{k}^{*}-S_{k}\right)^{2}\right\rangle \geqslant \omega_{k k}(S) . \tag{5.6}
\end{equation*}
$$

As has been said earlier, the investigator's aim is to find an estimate $S^{*}$ whose diagonal elements $\Omega_{k k}$ of the scatter matrix are as small as possible. The absolute values of nondiagonal elements of this matrix characterise the degree of linear dependence between the estimate components; obviously, the informativeness of individual components decreases when this dependence is strong (see Section 6).

It is possible to reach the lower boundary of the information inequality if and only if the distribution density $f(y \mid S)$ belongs to the so-called exponential family:

$$
\begin{equation*}
\ln f(y \mid S)=\sum_{i=1}^{n} S_{i}^{*}(y) \varphi_{i}(S)+\psi(S)+\chi(y) \tag{5.7}
\end{equation*}
$$

where $\psi(S)$ and $\chi(y)$ are scalar functions, and vector $\varphi(S)$ has a matrix of derivatives of the type

$$
\begin{equation*}
\left[\frac{\partial}{\partial S_{k}} \varphi_{i}(S)\right]=I(S)\left[E_{n}+B(S)\right]^{-1} . \tag{5.8}
\end{equation*}
$$

If the representation (5.7) is valid, then $S^{*}(Y)$ is a unique boundary, and hence, an effective estimate of the object $S$.

For the Poisson model, we find from relation (2.4) that

$$
\begin{equation*}
\ln f(y \mid S)=\sum_{j=1}^{m} y_{j} \ln q_{j}(S)-\sum_{j=1}^{m} q_{j}(S)-\sum_{j=1}^{m} \ln y_{j}! \tag{5.9}
\end{equation*}
$$

and in the general case when $m \neq n$, the first term on the right-hand side of Eqn (5.9) is not in a form required for the density to belong to the exponential family. This means that for the Poisson model of image formation no boundary estimate exists.

This fact by itself, being due only to the choice of the image formation model, introduces no principal difficulties into the problem: it was stressed in Section 2 that the problem consists of finding an effective estimate independently of the existence of the boundary estimate. Moreover, from the practical point of view, an efficient estimate may be close to the boundary given by the information inequality.

Fig. 9 shows Monte-Carlo simulations for a model object 'Iron'. The extents of the object and its image are $n=17$ and $m=21$ pixels, respectively, the object intensity is $F=12200$ events. In accordance with the Poisson distribution, 100 randomly blurred images of the object were simulated, and then a uniform Poisson background with a total mean brightness of 2100 events was added to every image. Each of the images obtained was restored by using the maximum likelihood estimate method, with object nonnegativeness being the only a priori information (in this case MLE instability does not manifest itself; the resons for this will be discussed in Section 6).

As with any method of inverse problem solution, the results of restoration of individual image realisations differ from each other; our aim must be to obtain object estimates that have the smallest possible bias relative to the true object and the smallest possible scatter around it. In this case the bias (Fig. 9a) proves to be much less than the standard deviation of the ensemble of the estimates. Fig. 9d compares the real estimates of brightness scatter in individual pixels with the minimum scatter dictated by the information inequality (in the first approximation we assumed that the estimate is unbiased). As is seen from the figure, the error corridor of restoration by the maximum likelihood method is close to the theoretically narrowest one under the conditions considered here. Thus, in the present case one can take the conditional MLE $\hat{S}_{+}$as an efficient estimate.

In practice, the situation can be often more complex because of a more pronounced instability that manifests itself when FER is very elongated. As mentioned earlier, this last circumstance arises when some eigenvalues of Fisher's matrix $I\left(S_{0}\right)$ are small, so that the corresponding diagonal elements of the inverse matrix $I^{-1}$ in the information inequality are very large. If under these conditions we are interested only in unbiased estimations, the information inequality sets the lower boundary for their variance at an unacceptably high level. The only way to reduce the dispersion is to take estimates from class $K_{b}$ with a bias $b(S)$ such that the terms $\left[E_{n}+B(S)\right]$ in


Figure 9. Object 'Iron' (a, solid curve), with examples of its blurred (b) and restored (c) images. The solid circles in panel (a) show mean object estimates, the dashed lines show the error corridor for individual estimates $( \pm \sigma)$. The solid circles and the dashed line in panel (d) correspond to real and theoretically attainable restoration accuracies, respectively.
identity (5.4) compensate $I^{-1}$. This does not guarantee by itself that the efficient estimate in fact approaches the theoretical lower boundary, but at least leaves a possibility of obtaining a practically acceptable solution. The difficulty is, however, that the investigator studying a particular image cannot specify the required bias class in advance.

The example from Section 2 with the stopped clock shows clearly enough that one cannot avoid specifying the class in which the object estimates are sought. At the same time, the aforesaid makes us look for possibilities of specifying such a nontrivial class without resorting to a fixed class of estimate bias. The questions connected with specifying such a class need further study.

### 5.2 The Kolmogorov- Wiener optimal filter

The notion of information introduced by Shannon [7, 8] relates to two realisations of random variables or random processes connected with each other. Since in the classical approach the object $S_{0}$ is not an element of a probabilistic ensemble, we cannot find directly Shannon's information about $S_{0}$ contained in a random realisation of its image $y_{0}$. For this reason, a particular image restoration method cannot rely upon some representation of Shannon's information. However, in discussing a typical situation that arises when solving inverse problems from some class, it is quite appropriate to introduce the corresponding probabilistic ensemble and to analyse the solution in the framework of the Bayesian approach. Consequences of this point of view are discussed in Sections 5.2-5.4.

As a probabilistic ensemble of objects (signals, after the terminology accepted in radio-physics) with properties which are assumed to be known to the investigator, we shall consider, as usual, a stationary Gaussian process $s(x)$ in an infinitely large interval with a spectral density $g_{s}(f)$, where $f$ is the spatial frequency. An additive background, $\xi(x)$, belongs to an analogical ensemble of noises with a spectral density $g_{\xi}(f)$. We neglect photon noise and assume, for the sake of simplicity, that the mean signal and the mean background are zero and the PSF depends only on the difference in the coordinates (i.e. is spatially invariant). Then the image

$$
\begin{equation*}
y(x)=\int h\left(x-x^{\prime}\right) s\left(x^{\prime}\right) \mathrm{d} x^{\prime}+\xi(x) \tag{5.10}
\end{equation*}
$$

also belongs to a Gaussian random functions ensemble. From now on we will omit the infinite integration limits.

As mentioned in the Introduction, the difficulty with the conversion of expression (5.10) is connected with the fact that one does not know the particular realisation of the additive noise. We use the script letter $\mathcal{F}$ to designate the Fourier transform $\mathcal{F}$ of an arbitrary function, so that, for example, $\mathcal{F}[s(x)] \equiv S(f)$. Then, by using Fourier transforms, model (5.10) can be written as

$$
\begin{equation*}
Y(f)=H(f) S(f)+\Xi(f) \tag{5.11}
\end{equation*}
$$

where $H(f) \equiv \mathcal{F}[h(x)]$ is the modulation transfer function (MTF). The MTF for any optical system vanishes above the cut-off frequency $f_{\mathrm{c}}$ [80]. This means that picture details with size less than $\simeq f_{c}^{-1}$ are not transmitted by electromagnetic radiation; they are said to be 'cut off' by the image formation system. In contrast, the noise power spectrum $q_{\xi}(f)$ usually extends far above the cut-off frequency into the high-frequency region. As seen from expression (5.11), under such conditions the image contains very little information about the high-frequency 'tail' of the signal $S(f)$ which is simply 'drowned' in the noise. In other words, one could vary the form of $S(f)$ near $f_{\mathrm{c}}$ in a wide range with practically no change in the image. Thus, instability of the inverse solution is most often connected with the presence of uncontrollable high-frequency intensity oscillations.

If one takes the solution (5.10) at $\xi(x) \equiv 0$ as an estimate close to the original, then, as noted in Section 3, we get an inverse estimate $s_{\mathrm{i}}(x)=\mathcal{F}^{-1}\left[R_{\mathrm{i}}(f) Y(f)\right]$, where we introduced the inverse filter

$$
\begin{equation*}
R_{\mathrm{i}}(f) \equiv[H(f)]^{-1} \tag{5.12}
\end{equation*}
$$

Since at zero noise one ascribes all image fluctuations to the original, it is hard to expect a high quality for the
inverse solution. Indeed, owing to the smallness of $|H(f)|$ at high frequencies, filter (5.12) improperly amplifies the amplitudes of random fluctuations which are unavoidably present in the image.

The problem of noise filtration and of simultaneous signal deconvolution is posed by Kolmogorov [11] and Wiener [12] as follows. It is required to find a linear filter $r_{\mathrm{W}}(x)$ such that, when applied to experimental data in the form

$$
\begin{equation*}
s_{\mathrm{W}}(x)=\int r_{\mathrm{W}}\left(x-x^{\prime}\right) y\left(x^{\prime}\right) \mathrm{d} x^{\prime} \tag{5.13}
\end{equation*}
$$

it would give a signal estimate $s_{\mathrm{W}}(x)$ that is closer to the signal $s(x)$ in the sense of the mean-square deviation:

$$
\begin{equation*}
\varepsilon_{\mathrm{W}}^{2} \equiv\left\langle\left[s_{\mathrm{W}}(x)-s(x)\right]^{2}\right\rangle=\min \tag{5.14}
\end{equation*}
$$

In the case of spatially invariant PSF, the problem is solved easily (see for example Ref. [81]). The frequency characteristic of Wi ener's optimal filter $\mathcal{F}\left[r_{\mathrm{W}}(x)\right] \equiv R_{\mathrm{W}}(f)$ is found to have the form

$$
\begin{equation*}
R_{\mathrm{W}}(f)=\frac{H(-f)}{|H(f)|^{2}+g_{\xi}(f) / g_{s}(f)} \equiv \Phi_{\mathrm{W}}(f) R_{\mathrm{i}}(f) \tag{5.15}
\end{equation*}
$$

where we introduced the factor

$$
\begin{equation*}
\Phi_{\mathrm{W}}(f)=\frac{|H(f)|^{2}}{|H(f)|^{2}+g_{\xi}(f) / g_{s}(f)} \tag{5.16}
\end{equation*}
$$

It is the rapid decrease of $\Phi_{\mathrm{W}}(f)$ at high frequencies that prevents oscillations typical for inverse filtration from increasing. Please note that appearance of the factor $\Phi_{\mathrm{W}}(f)$ is due to the presence of a priori information about the ensembles to which the signal and noise belong.

It can be readily shown that if in expression (5.13) an arbitrary linear filter $r(x)$ is used instead of Wiener's one, the power spectrum for the resulting solution $s_{r}(x)$ will be

$$
\begin{equation*}
g_{r}(f)=|R(f)|^{2}\left[|H(f)|^{2} g_{s}(f)+g_{\xi}(f)\right], \tag{5.17}
\end{equation*}
$$

and the variance $s_{r}(x)$ is, as usual, equal to the integral of this expression over all frequencies. By substituting Eqn (5.15) into this expression we get the power spectrum of the optimal estimate:

$$
\begin{equation*}
g_{\mathrm{W}}(f)=\Phi_{\mathrm{W}}(f) g_{s}(f) \tag{5.18}
\end{equation*}
$$

Since $\Phi_{\mathrm{W}}(f) \leqslant 1$, the variance of Wiener's solution is less than the true signal variance

$$
\begin{equation*}
\sigma_{s}^{2}=\int g_{s}(f) \mathrm{d} f \tag{5.18a}
\end{equation*}
$$

that is, the optimal filter has apronounced smoothing effect on the estimate. The difference of variances, $\varepsilon_{\mathrm{W}}^{2} \equiv \sigma_{s}^{2}-\sigma_{\mathrm{W}}^{2}$, is equal to the mean-square error associated with the use of Wiener's filter. We have
$\varepsilon_{\mathrm{W}}^{2}=\int\left[1-\Phi_{\mathrm{W}}(f)\right] g_{s}(f) \mathrm{d} f=\int \frac{g_{\xi}(f)}{|H(f)|^{2}+g_{\xi}(f) / g_{s}(f)} \mathrm{d} f$.

K nowledge of the signal and noise power spectra is the price one has to pay for getting stable object and error estimates in the version of the Bayesian approach considered here. As seen from expression (5.15), the stability is
due to the presence of the ratio $g_{\xi}(f) / g_{s}(f)$ in the denominator, which does not allow the optimal filter to approach the sharply increasing inverse filter $R_{\mathrm{i}}(f)$ at high frequencies. If calculation of filtration errors is not required, it is enough to know only the signal-to-noise ratio, and not each of these functions separately. The methods developed by Phillips [14] and Tikhonov [17, 18] presume that the signal and noise belong to Gaussian ensembles with proportional spectral densities, so that $g_{\xi}(f) / g_{s}(f) \equiv \mu=$ const. The unknown parameter $\mu$, which is usually referred to as the regularising parameter, is chosen in such a way that the solution yields a smooth image which is consistent in the mean-square-deviation sense with the observational data.

Expression (5.15) for Wiener's filter is also easy to find by using straightforward maximisation of a posteriori probability density in the framework of standard Bayesian considerations.

Returning to expression (5.17), we note that the requirement $g_{r}(f)=g_{s}(f)$, that is that the signal power spectrum is equal to its linear estimate, leads to a homomorphous filter (first described by Cole [82]) with the frequency characteristic

$$
\begin{equation*}
\left|R_{\mathrm{h}}(f)\right|=\left[|H(f)|^{2}+\frac{g_{\xi}(f)}{g_{s}(f)}\right]^{-1 / 2} \tag{5.20}
\end{equation*}
$$

i.e. $\Phi_{\mathrm{h}}(f)=\sqrt{\Phi_{\mathrm{W}}(f)}$. This filter is used when a better restoration of high spatial frequencies is required than that possible with Wiener's filter.

### 5.3 Information about the original for Gaussian ensembles

Shannon's information theory allows one to look at the main concepts used in constructing the optimal filter in a new light.

According to Pinsker [83], the mean information about realisation of a Gaussian process $s(x)$ of unit length, which is contained in the realisation segment also of unit length for a Gaussian process $y(x)$ connected with the first one in a stationary mode, is

$$
\begin{equation*}
J(y, s)=J(s, y)=-\frac{1}{2} \int \ln \left[1-\frac{\left|g_{y s}(f)\right|^{2}}{g_{y}(f) g_{s}(f)}\right] \mathrm{d} f, \tag{5.21}
\end{equation*}
$$

where $g_{y s}(f)$ is the mutual spectral density of the processes, and $g_{y}(f)$ and $g_{s}(f)$ are their partial spectral densities. Here and below we use natural logarithms, so that information is measured in units of nat $\mathrm{cm}^{-1}$. The spectral densities for model (5.10) are known to have the form

$$
\begin{align*}
& g_{y}(f)=|H(f)|^{2} g_{s}(f)+g_{\xi}(f) \\
& g_{y s}(f)=H(f) g_{s}(f) \tag{5.22}
\end{align*}
$$

By substituting Eqn (5.22) into Eqn (5.21), we arrive at the expression

$$
\begin{equation*}
J(y, s)=\frac{1}{2} \int \ln \left[1+|H(f)|^{2} \frac{g_{s}(f)}{g_{\xi}(f)}\right] \mathrm{d} f, \tag{5.23}
\end{equation*}
$$

obtained by Fellgett and Linfoot [84] by optical image analysis.

We also take into account the fact that every optical system has a cut-off frequency $f_{\mathrm{c}}$, above which the MTF is zero [80]. By choosing the PSF width $\Delta \equiv f_{\mathrm{c}}^{-1}$ as a characteristic picture extent, we obtain the following
expression for the mean mutual information between the signal and image portions of length $\Delta$ :

$$
\begin{equation*}
J_{\Delta}(y, s)=\frac{1}{2} \int_{-f_{\mathrm{c}}}^{f_{\mathrm{c}}} \ln \left[1+|H(f)|^{2} \frac{g_{s}(f)}{g_{\xi}(f)}\right] \frac{\mathrm{d} f}{f_{\mathrm{c}}} \tag{5.24}
\end{equation*}
$$

Using the general formula (5.21), one can easily obtain an expression analogous to (5.23) for the mean mutual information between the image and the background:

$$
\begin{equation*}
J(y, \xi)=\frac{1}{2} \int \ln \left[1+\frac{g_{\xi}(f)}{|H(f)|^{2} g_{s}(f)}\right] \mathrm{d} f \tag{5.25}
\end{equation*}
$$

This information for real systems is infinitively large. If one restricts oneself to a spatial frequency range $|f| \leqslant f_{\mathrm{c}}$ and image and background portions of width $\Delta$, then

$$
\begin{equation*}
J_{\Delta}(y, \xi)=\frac{1}{2} \int_{-f_{\mathrm{c}}}^{f_{\mathrm{c}}} \ln \left[1+\frac{g_{\xi}(f)}{|H(f)|^{2} g_{s}(f)}\right] \frac{\mathrm{d} f}{f_{\mathrm{c}}} \tag{5.26}
\end{equation*}
$$

According to expression (5.23), the mutual information of the image and background portions is represented as a spatial frequency integral of some function $J_{\mathrm{f}}(y, s)$ that can be treated as the corresponding spectral information density. It is very significant that under typical conditions this function rapidly decreases with frequency. In contrast, the analogous spectral information density between the image and background $J_{\mathrm{f}}(y, \xi)$ in expression (5.25) rapidly increases with frequency, so that a critical frequency $f_{*}$ which depends on the signal-to-noise ratio exists, above which information about the signal 'drowns' in the information about the noise which is of no interest to us. This is the reason why it is difficult to estimate whether the highfrequency oscillations in the object brightness distribution, which are so typical for instability, are real not not.

The expressions given above are insufficient for further analysis, since they enable one to find mutual information only between finite parts of the image and the signal, whereas we need to estimate object brightness at individual points. The concept of mean partial signal $s(x)$ information contained in the entire image $y(x)$ should now be introduced. We denote this quantity by $J[y, s(x)]$. In model (5.10) it can be found as follows.

Let $y(x)$ be an arbitrary Gaussian stationary process in a possibly infinite interval, $\zeta$ be a Gaussian random variable with a variance $\sigma_{\zeta}^{2}$ connected with the process $y(x)$, and $\varepsilon_{\text {min }}^{2}$ be the least mean-square error of the linear approximation of $\zeta$ by the process $y(x)$. Then, as was shown by Gel'fand and Yaglom [85], the mean mutual information between $\zeta$ and the process $y(x)$ is

$$
\begin{equation*}
J(y, \zeta)=\frac{1}{2} \ln \left(\frac{\sigma_{\zeta}^{2}}{\varepsilon_{\min }^{2}}\right) \tag{5.27}
\end{equation*}
$$

Now we apply this general result to model (5.10). Consider the image as the process $y(x)$, and the signal at an arbitrary point $s(x)$ as $\zeta$. Then $\sigma_{\xi}^{2}=\sigma_{s}^{2}$, and the minimum error of the linear approximation is reached for the Kolmogorov-Wiener filter and is $\varepsilon_{\mathrm{W}}^{2}$. As a result, the information about a particular value of the object, which is contained in its blurred and noisy image, is represented as

$$
\begin{equation*}
J[y, s(x)]=\frac{1}{2} \ln \left(\frac{\sigma_{s}^{2}}{\varepsilon_{\mathrm{W}}^{2}}\right), \tag{5.28}
\end{equation*}
$$

where $\sigma_{s}^{2}=\int g_{s}(f) \mathrm{d} f$, and $\varepsilon_{\mathrm{W}}^{2}$ is determined by formula (5.19). It is easy to show that $J[y, s(x)]$ is nonnegative.

One can get a clearer concept of the information $J[y, s(x)]$ by equating it to the known expression $-\frac{1}{2} \ln \left(1-\rho^{2}\right)$ for the mean mutual information between two Gaussian random variables with a correlation coefficient $\rho$. This definition yields a correlation coefficient $\rho[y, s(x)]$ between a stationary Gaussian process $y$ and $a$ particular value $s(x)$ of another similar process, which, by using expressions (5.19) and (5.28), can be represented in the form

$$
\begin{align*}
\rho^{2}[y, s(x)]=1 & -\int \frac{g_{s}(f) \mathrm{d} f}{1+|H(f)|^{2} g_{s}(f) / g_{\xi}(f)} \\
& \times\left[\int g_{s}(f) \mathrm{d} f\right]^{-1} \tag{5.29}
\end{align*}
$$

The notion of mean information about a particular value of the original, which is contained in its blurred image, allows us to give a new interpretation of the Kolmogorov-Wiener optimal filter. We want to find such linear filter of type (5.13) with a kernel $r\left(x-x^{\prime}\right)$ that the estimate of object $s_{r}(x)$ obtained by using this filter has the same information about the particular value of the original $s(x)$ as that contained in the entire image $y$, i.e. that the following equality holds

$$
\begin{equation*}
J\left[s_{r}(x), s(x)\right]=J[y, s(x)] \tag{5.30}
\end{equation*}
$$

The filter we are looking for turns out to coincide with Kolmogorov-Wiener's one [86], with the above condition being necessary and sufficient. Thus, the optimal filter 'collects' all the information given by the image about the corresponding value of the original into a point estimate.

### 5.4 A numerical example

We illustrate the relations given above by a typical enough example when the signal correlation function has an exponential form so that the corresponding spectral density is

$$
\begin{equation*}
g_{s}(f)=\frac{g_{s}(0)}{1+\left(f / f_{s}\right)^{2}}, \quad g_{s}(0)=\frac{\sigma_{s}^{2}}{\pi f_{s}} \tag{5.31}
\end{equation*}
$$

where $f_{s}$ is a characteristic signal frequency. Let the background be a white noise, i.e. $g_{\xi}(f)=$ const, and the PSF have a diffractive form:

$$
\begin{equation*}
h(x)=\frac{1}{\Delta} \operatorname{sinc}^{2}\left(\frac{x}{\Delta}\right) \tag{5.32}
\end{equation*}
$$

where the function $\operatorname{sinc}(t)$ is defined by the relation

$$
\begin{equation*}
\operatorname{sinc}(t) \equiv \frac{\sin (\pi t)}{\pi t} \tag{5.33}
\end{equation*}
$$

In this case the width of the PSF $\Delta$ is equal to the distance from the central maximum to the first zero of $h(x)$. The corresponding MTF is a 'triangular' function:

$$
H(f)= \begin{cases}1-|f| / f_{\mathrm{c}}, & |f| \leqslant f_{\mathrm{c}}  \tag{5.34}\\ 0, & |f|>f_{\mathrm{c}}\end{cases}
$$

where $f_{\mathrm{c}} \equiv \Delta^{-1}$ is the cut-off frequency for the image formation system. By the way, expression (5.32) describes with good accuracy the known Airy solution [87] for a point-source image made by an ideal optical system with circular aperture. As was shown by O'Neill [88, 89], the precise MTF for this system differs little from that described by expression (5.34).

We introduce dimensionless parameters in the form of the ratio of the cut-off frequency to the characteristic frequency and the signal-to-noise ratio at zero frequency:

$$
\begin{equation*}
a=\frac{f_{\mathrm{c}}}{f_{s}}, \quad \mu=\frac{g_{s}(0)}{g_{\xi}} \tag{5.35}
\end{equation*}
$$

Then the formulas (5.24), (5.26), and (5.29) take on the following form:

$$
\begin{align*}
& J_{\Delta}(y, s)=\int_{0}^{1} \ln \left[1+\frac{\mu(1-x)^{2}}{1+a^{2} x^{2}}\right] \mathrm{d} x, \\
& J_{\Delta}(y, \xi)=\int_{0}^{1} \ln \left[1+\frac{1+a^{2} x^{2}}{\mu(1-x)^{2}}\right] \mathrm{d} x, \\
& \rho^{2}[y, s(x)]=\frac{2}{\pi}\{\arctan a \\
& \left.\quad-\frac{a}{R}\left[\arctan \left(\frac{a^{2}}{R}\right)+\arctan \left(\frac{\mu}{R}\right)\right]\right\}, \\
& R=\sqrt{\mu+(1+\mu) a^{2}} \tag{5.38}
\end{align*}
$$

It is seen from Fig. 10 that information about the signal $J_{\Delta}(y, s)$ increases with increasing signal-to-noise ratio approximately as $\log \mu$ for all cut-off frequencies. At the same time, information about the background dominates at low signal-to-noise ratios, so that in that region signal restoration is strongly hampered. We stress that this time we speak of integral information that relates to all frequencies below the cut-off frequency $f_{\mathrm{c}}$.


Figure 10. Information about the signal (solid line) and background (dashed line) as a function of the signal-to-noise ratio for different values of parameter $a$.

The function $\rho(\mu)$ calculated from expression (5.38) is shown in Fig. 11. We see that the correlation coefficient between the image and a particular count taken from the object reaches a significant value only for those cases when the characteristic signal frequency range is located within the system transmission band (i.e. at $a>1$ ). In practice, one usually comes across the opposite: the image formation system cuts off a significant part of the signal power


Figure 11. Correlation coefficient between the image and a particular value of the original as a function of signal-to-noise ratio for different values of parameter $a$.
spectrum, so that the information about an individual point of the object provided by the entire image is very small.

The aforesaid unequivocally leads to the conclusion that in the general case it is not individual counts from the original that will be restored, but only some combinations of them, the object's functionals, about which we have enough information. These functionals should as much as possible be independent of each other, in order that the estimate of each following functional would add more new information about the original. These properties are inherent to principal components introduced by Hotelling [90] in a general statistical context. However, I shall not discuss the corresponding approach in the framework of the Bayesian scheme, and prefer to use it in the next section, where I return to the analysis of the classical scheme with a deterministic sought object.

## 6. Occam's razor estimation

It has been often emphasised (see, for example, Ref. [91]) that a model constructed for any physical phenomenon is in principle not uniquely defined, and the choice of one of the models consistent with the data is based on the criteria of simplicity and predictive force. I shall try to apply the first of these criteria, which was explicitly formulated already by Occam, to solve the inverse problems in the same spirit as that used for the construction of physical models.

In the previous section we found that the observed image usually contains too little information to allow one to be able to restore the intensity distribution of the object at individual points with the required accuracy. The information we have allows us to estimate only some functions of the individual counts; for example, at a given mean background it is quite possible to estimate the total brightness of the object independently of the PSF. It is natural to expect that there exist other functions similar to the total brightness, more precisely the object's functionals, which provide solutions to the inverse problem. The stability of the latter means that the set of functionals
gives a sufficiently complete representation of the properties of the local object. Most often we deal with an unstable inverse problem, and then the set of functionals available for reasonable estimation becomes so narrow that it is consistent with a wide set of feasible objects with different degrees of complexity that are located in the FER.

Obviously, the set of functions described here can be determined by various means. We could, for example, estimate not the brightness distribution of the object, but the Fourier components corresponding to it. The part of the Fourier components that is estimated with the smallest errors contains the main information about the object, so that, according to the simplicity criterion, only this part should be taken into account in estimating the object.

With all the attractiveness of this approach, one needs to say that the use of Fourier transforms here appears artificial in many respects. With even more success we could choose a system of Haar's coefficients, which enables one to give not only local, but also uniformly convergent description of the object, or settle for another system of functionals. The shortcoming of any system selected in advance such as generalised Fourier coefficients is that the system is not adapted to a particular object; in other words, eigenfunctions of the operator that generates the given system (for example, a sinusoid), do not take into account the properties of a specific brightness distribution in a natural way.

Meanwhile, the structure of the FER itself, namely its extreme elongation in some directions, points towards a more natural way of choosing functionals for any particular case. As is easy to understand, in the framework of linear description the main functionals are linear combinations of point counts that are estimated with the highest accuracy, i.e. change in the direction of elongation of the FER. These combinations prove to be a part of the so-called principal components of the inverse estimate of the object, with the principal components being generated by the Fisher's sample information matrix. The last step in finding the inverse solution relies upon Occam's razor principle. This step consists of retaining the minimum number of principal components, which is sufficient for a satisfactory statistical description of data in the context of the image randomness test.

Now I turn to formal presentation of the procedure described in Refs [27, 30, 31].

### 6.1 Principal components

Let $\xi$ be an $n$-dimensional random vector with a known mean value $a$ and a variance matrix $\Omega=\left\langle(\xi-a)(\xi-a)^{\mathrm{T}}\right\rangle$. The problem is to find a random vector linearly connected with $\xi$ :

$$
\begin{equation*}
\eta \equiv\left[\eta_{1}, \ldots, \eta_{r}\right]^{\mathrm{T}}=A \xi, \quad r \leqslant n \tag{6.1}
\end{equation*}
$$

such that a linear combination

$$
\begin{equation*}
\tilde{\xi}=B+C \eta \tag{6.2}
\end{equation*}
$$

yields the best approximation in the mean-square sense to the original random vector. This requires that the quantity

$$
\begin{equation*}
\rho^{2}(\xi, \tilde{\xi}) \equiv\left\langle\|\xi-\tilde{\xi}\|^{2}\right\rangle \tag{6.3}
\end{equation*}
$$

be minimum. Thus, we look for an explicit representation of the matrices $A, B$, and $C$ through characteristics of the multidimensional random variable $\xi$. In the general case these matrices have the sizes $r \times n, n \times 1$, and $n \times r$, respectively.

The idea of the principal components $\eta_{1}, \ldots, \eta_{r}$ introduced by Hotelling [90] is that at $r<n$ we can achieve a good approximation of $\xi$ in the form (6.2) by using a simpler variable $\eta$, which still contains the main statistical information comprised in $\xi$. In the general case such a possibility is due to the fact that the components of $\xi$ are correlated with each other, so that all the components of $\xi$ contain less information than they would if they had been independent. In contrast, the principal components represent statistically uncorrelated linear combinations of the components of $\xi$, so that by choosing an appropriate number of principal components we can make the information contained in $\xi$ more compact. The use of principal components is most effective when $\xi$ obeys Gaussian distribution or one close to it. Then the fact that the principal components are uncorrelated means that they are also independent. If $r=n$, then the formulas (6.1) and (6.2) yield the exact representation of the $n$-dimensional random variable $\xi$ through the system of uncorrelated random variables $\eta$ of the same dimension.

The variance matrix $\Omega$ represents a positive definite symmetric matrix with real components. Its eigenvalues $\mu_{k}$ are then positive and the eigenvectors $V_{k}$ relating to different eigenvalues are orthogonal. After normalisation of the eigenvectors we get

$$
\begin{align*}
& \Omega V_{k}=\mu_{k} V_{k}  \tag{6.4}\\
& V_{i}^{\mathrm{T}} V_{k}=\delta_{i k}, \quad i, k=1, \ldots, n \tag{6.5}
\end{align*}
$$

For convenience, we renumerate the eigenvalues in the order of their decreasing values, so that $\mu_{1} \geqslant \mu_{2} \geqslant \ldots \geqslant \mu_{n}>0$, and also we introduce square $n \times n$ matrices

$$
\begin{equation*}
V=\left[V_{1}, \ldots, V_{n}\right], \quad M=\operatorname{diag}\left[\mu_{1}, \ldots, \mu_{n}\right] \tag{6.6}
\end{equation*}
$$

and their 'shortened' variants

$$
\begin{equation*}
V_{*}=\left[V_{1}, \ldots, V_{r}\right], \quad M_{*}=\operatorname{diag}\left[\mu_{1}, \ldots, \mu_{r}\right], r \leqslant n \tag{6.6a}
\end{equation*}
$$

The columns of the matrices $V$ and $V_{*}$ consist of eigenvectors $\Omega$; the nonzero elements of $M$ and $M_{*}$ are eigenvalues of $\Omega$ located at the main diagonals of these matrices.

The solution to the problem stated above dates back to Refs [92-94] and can be found in Brillinger's monograph [95]. The extreme values of the sought matrices are

$$
\begin{equation*}
A=V_{*}^{\mathrm{T}}, \quad B=\left(E_{n}-V_{*} V_{*}^{\mathrm{T}}\right) a, \quad C=V_{*} \tag{6.7}
\end{equation*}
$$

where $E_{n}$ is the identity $n$-dimensional matrix. This means that the linear transformation of $\xi$ to be applied to the principal components $\eta$ and the best linear approximation $\tilde{\xi}$ are expressed through the eigenvectors of $\Omega$ as follows:

$$
\begin{align*}
& \eta=V_{*}^{\mathrm{T}} \xi, \quad\langle\eta\rangle \equiv b=V_{*}^{\mathrm{T}} a,  \tag{6.8a}\\
& \tilde{\xi}=a+V_{*}(\eta-b) \tag{6.8b}
\end{align*}
$$

This produces the minimum inaccuracy of the approximation (6.3), which is equal to the summary tail of the eigenvalues

$$
\begin{equation*}
\rho_{\min }^{2}=\sum_{k=r+1}^{n} \mu_{k} \tag{6.9}
\end{equation*}
$$

It follows from equalities (6.8a) that

$$
\begin{equation*}
D(\eta) \equiv\left\langle(\eta-b)(\eta-b)^{\mathrm{T}}\right\rangle=M_{*} \tag{6.10}
\end{equation*}
$$

that is, the variances of the principal components are equal to the eigenvalues of the variance matrix, and their covariations are zero.

In order to make more clear the geometrical sense of the principal components, let us consider the case $r=n$ and assume $\xi$ obeys a multidimensional Gaussian distribution: $\xi \sim \mathcal{N}(a, \Omega)$. Then the lines of constant density of probability form ellipsoids

$$
\begin{equation*}
(x-a)^{\mathrm{T}} \Omega^{-1}(x-a)=\mathrm{const} . \tag{6.11}
\end{equation*}
$$

The matrix $V$ is orthogonal, that is $V^{-1}=V^{\mathrm{T}}$, so that the direct and inverse linear transformations

$$
\begin{equation*}
x=V y, \quad y=V^{\mathrm{T}} x \tag{6.12}
\end{equation*}
$$

describe rotations of the coordinate system. By using the property (6.5) of orthonormalisation of the eigenvectors $V_{k}$ and the equality $\Omega^{-1} V=V M^{-1}$ following from expression (6.4), one can easily show that the transformation (6.12) brings Eqn (6.11) to the form

$$
\begin{equation*}
(y-b)^{\mathrm{T}} M^{-1}(y-b)=\mathrm{const} \tag{6.13}
\end{equation*}
$$

or, because of the diagonal position of $M$,

$$
\begin{equation*}
\sum_{k=1}^{n} \frac{\left(y_{k}-b_{k}\right)^{2}}{\mu_{k}}=\text { const } \tag{6.14}
\end{equation*}
$$

The last equations describe ellipsoid (6.11) in the Euclidean coordinate system $y_{1}, \ldots, y_{n}$, whose axes are parallel to the principal axes of the ellipsoid, with the length of the $k$ th axis of the ellipsoid being proportional to $\sqrt{\mu_{k}}$. This means that $\eta \sim \mathcal{N}(b, M)$, and components of this vector are independent of each other. This entire operation represents a standard reduction of the quadratic form (6.11) to the principal axes.

The case $r<n$, which is of main interest to us, corresponds to the coordinate transformation $x=V_{0} y$, instead of the pure rotation $x=V y$, where the matrix $V_{0}$ differs from $V$ in that the eigenvectors $V_{r+1}, \ldots, V_{n}$, which correspond to the minimum eigenvalues $\mu_{r+1}, \ldots, \mu_{n}$, are replaced by zero vectors. It is easy to see, that $V_{0}$ is no longer an orthogonal matrix, and the transformation $x=V_{0} y$ is equivalent to two operations: rotation of the coordinate system $x=V y$ and subsequent projection along the axes $y_{r+1}, \ldots, y_{n}$ onto the subspace $\left(y_{1}, \ldots, y_{r}\right)$. Thus, isodenses of the principal components $\eta$ are $r$-dimensional ellipsoids

$$
\begin{equation*}
\sum_{k=1}^{r} \frac{\left(y_{k}-b_{k}\right)^{2}}{\mu_{k}}=\text { const } \tag{6.15}
\end{equation*}
$$

which differ from ellipsoids (6.14). Naturally, one cannot return back to the latter by using a reverse rotation of the coordinate system and therefore the approximation (6.8b) includes an operation based on the knowledge of the mean value $\langle\xi\rangle=a$, consisting of a shift along those directions that were used for the projection. Because some information about $\xi$ is lost during the nonlinear operation of projection, the quantity $\tilde{\xi}$ given by expression ( 6.8 b ) is no longer equal to the initial quantity $\xi$ for $r<n$, but it still provides the best approximation to the original in the mean-square sense.

We illustrate the process of finding the principal components by the simplest example of solving the system of equations (3.4). If one writes down the corresponding estimation problem in the form (3.2), Fisher's matrix will, according to relation (3.13), be proportional to $H^{\mathrm{T}} H$, i.e.

$$
I \propto\left[\begin{array}{ll}
10 & 26  \tag{6.16}\\
26 & 68
\end{array}\right]
$$

The eigenvalues and eigenvectors $I$ are approximately

$$
\begin{align*}
& \lambda_{1}=77.949, \quad \lambda_{2}=0.051 \\
& V_{1}=\left[\begin{array}{l}
0.357 \\
0.934
\end{array}\right], \quad V_{2}=\left[\begin{array}{r}
-0.934 \\
0.357
\end{array}\right], \tag{6.17}
\end{align*}
$$

and the corresponding principal components of the solution are the components of the vector $P=V^{\mathrm{T}} X$, that is

$$
\begin{align*}
& P_{1}=0.357 x_{1}+0.934 x_{2} \\
& P_{2}=-0.934 x_{1}+0.357 x_{2} \tag{6.18}
\end{align*}
$$

As mentioned in Section 3.2, the accuracies of estimating $P_{1}$ and $P_{2}$ are different: the ratio of the standard deviations (semiaxes of the ellipse of feasible solutions in Fig. 2) is

$$
\begin{equation*}
\frac{\sigma\left(P_{2}\right)}{\sigma\left(P_{1}\right)}=\sqrt{\frac{\lambda_{1}}{\lambda_{2}}} \simeq 38.974 . \tag{6.18a}
\end{equation*}
$$

Since the accuracies of estimation of the principal components differ significantly, in the solution of the inverse problem we can restrict ourselves by showing only the estimate of $P_{1}$, as finding the estimate of $P_{2}$ adds almost nothing to the information about the solution.

### 6.2 Analysis of the principal components of the maximum likelihood estimate (MLE)

Let us return now to the discussion of restoration of the randomly blurred image. Putting aside for the time being the condition of nonnegativeness of the object $S_{0}$, we consider the maximum likelihood estimate $\hat{S}$ (see Fig. 4) as the random variable $\xi$ studied above. In principle, it contains all the available information about $S_{0}$; however this information, like a drop of honey in a barrel of tar, is almost entirely hidden by insignificant details of behaviour of the MLE. Thus, separate components of the inverse solution are so strongly correlated with each other that in order to clarify the matter one should add the variance matrix of component estimates with all its nondiagonal terms to the obtained brightness distribution. Our task is to extract real information about the object whilst suffering minimum information losses.

What has been said above provides a clear outline of a natural way we should move forward: if we restrict ourselves to linear analysis of the estimate $\hat{S}$, we need to find its principal components. According to expression (4.24), the variance matrix for $\hat{S}$ approximately equals the inverse Fisher's matrix:

$$
\begin{equation*}
\Omega \simeq I^{-1}\left(S_{0}\right) \tag{6.19}
\end{equation*}
$$

Let us take into account that the eigenvectors of any regular matrix $I$ and of the inverse matrix $I^{-1}$ coincide, and the eigenvalues of matrix $I$, which we denote by $\left[\lambda_{k}\right]$, $k=1, \ldots, n$, are equal to inverse eigenvalues of matrix $I^{-1}$. According to relation (6.19), for the latter we have
$\mu_{1} \geqslant \mu_{2} \geqslant \ldots \geqslant \mu_{n}>0 ;$ it is also convenient to number the eigenvalues of $I$ in the order of their decrease, so that

$$
\begin{align*}
& \lambda_{1}=\mu_{n}^{-1}, \ldots, \lambda_{n}=\mu_{1}^{-1}, \quad \lambda_{1} \geqslant \ldots \geqslant \lambda_{n}>0 \\
& I V_{k}=\lambda_{k} V_{k} \tag{6.20}
\end{align*}
$$

where the order of numbering of the eigenvectors now corresponds to that of the eigenvalues $\lambda_{k}$. The case of some eigenvalues of $I$ coinciding does not bring in any principal changes; since in practice this occurs extremely rarely, I shall not discuss it here.

Attention should be drawn to the significant difference between the ways the principal components are viewed in the standard approach and that used here. In the solution of typical statistical problems, the multidimensional random variable $\xi$ describes data sample for some set of objects, and the principal components corresponding to the maximum eigenvalues of the variance matrix $\Omega$ comprise the main part of statistical information contained in $\xi$. In solving inverse problems, in contrast, some statistical estimate of the object, say $\hat{S}$, is used as $\xi$, and its variance matrix $\Omega$ characterises estimation errors of the individual components of $S_{0}$. Obviously, the principal components of $\hat{S}$ with a relatively large variance contain little information about the sought object. Therefore in inverse problems we are interested in the principal components that correspond to the minimum eigenvalues of $\Omega$, that is, to the maximum eigenvalues $\left[\lambda_{k}\right]$ of Fisher's information matrix $I\left(S_{0}\right)$. For brevity, we shall call them senior principal components of $\hat{S}$. They correspond to the projection onto the space of the principal components of the original dispersion ellipsoid not across its elongation directions but along them (see Fig. 4).

The second principal difference of the discussed scheme from that used in typical statistical studies is connected with the fact that we do not know the mean value $a$ of the original vector $\xi$, or, in the context of image restoration, the original object $S_{0}$. Therefore, relation (6.8b) cannot be used to approximate $\hat{S}$ by applying a reverse shift of the feasible region from the space of the principal component along the projection directions; such approximation, however, is not required in the case considered here. Indeed, the whole idea of addressing Occam's principle consists of choosing the 'simplest' estimate of the object consistent with observational data, with simplicity being quite uniquely defined: the simplest estimate is that which has the minimum number of senior principal components. It is at this stage that we introduce a bias into the estimate we are seeking, whose stabilising role was noted in Section 3.

By projecting the original FER along the directions with maximum variances of the inverse solution, we come to a new, significantly 'rounder' FER. This means that the components of the biased and smooth estimate $\tilde{S}$ are significantly less mutually correlated than the components of $\hat{S}$, so that each of the components of $\tilde{S}$ contributes a significant piece of information about the original object. The restored brightness distribution becomes informative by itself, without additional specification of all the terms of the estimate variance matrix, and only now we can consider the image restoration task completed.

In order not to complicate the discussion, we abandoned above the condition of estimate nonnegativeness, which by itself stabilises the solution [96]. This condition can be included into consideration (see Fig. 4), by changing the
unconditional estimate $\hat{S}$ by its conditional analogue $\hat{S}_{+}-$ the maximum likelihood estimate in the region $S \geqslant 0$. In typical inverse problems the estimate $\hat{S}_{+}$is only slightly more stable than MLE, so all the considerations presented above relate to it to an equal extent. Hence, the pair of transformations connecting the estimate $\hat{S}_{+}\left(y_{0}\right)$ with its simplified version $\tilde{S}\left(y_{0}\right)$ has the form

$$
\begin{align*}
& P\left(y_{0}\right)=V_{*}^{\mathrm{T}} \hat{S}_{+}\left(y_{0}\right)  \tag{6.21a}\\
& \tilde{S}\left(y_{0}\right)=V_{*} P\left(y_{0}\right) \tag{6.21b}
\end{align*}
$$

where $P=\left[P_{1}, \ldots, P_{r}\right]^{\mathrm{T}}$ is the principal component vector ( $r \leqslant n$ ), and $V_{*}$ is an $n \times r$ matrix with columns formed by the eigenvectors $V_{k}$ of Fisher's information matrix $I\left(S_{0}\right)$ that correspond to its $r$ maximum eigenvalues $\left[\lambda_{k}\right]$. The two transformations can be combined into one:

$$
\begin{equation*}
\tilde{S}\left(y_{0}\right)=\mathcal{D} \hat{S}_{+}\left(y_{0}\right), \quad \mathcal{D} \equiv V_{*} V_{*}^{\mathrm{T}}, \tag{6.22}
\end{equation*}
$$

where $\mathcal{D}$ is a symmetric square $n \times n$ matrix.
The estimate $\hat{S}_{+}\left(y_{0}\right)$ is found by minimising information (4.3) in the positive hyperquadrant $\{S\}$. The corresponding numerical methods have been given, in particular, by Bertsekas [97]. To find eigenvectors $I$, one can use the transformations (4.19) or (4.22) as the first approximation. The last of them seems more reliable, but the final choice must be based on a special investigation. After a stable estimate of the object $\tilde{S}\left(y_{0}\right)$ has been found, the eigenvector and eigenvalue system for Fisher's matrix can be made more precise.

Of course, the experiment provides us only with a sample spectrum of matrix $I$, which is different, in general, from its real spectrum, and with the corresponding sample principal components. Fortunately, sample characteristics of the variance matrix have all the traditional properties of 'good' estimates, such as consistency, asymptotic efficiency, and normality [69, 98].

The equation $I V=V \Lambda$ that defines the spectrum of $I$, where $\Lambda=\operatorname{diag}\left[\lambda_{1}, \ldots, \lambda_{n}\right]$, can be rewritten because of the orthogonality of $V$, to give

$$
\begin{equation*}
I=V \Lambda V^{\mathrm{T}} \tag{6.23}
\end{equation*}
$$

which expresses the known spectral theorem. There are many numerical methods allowing one to give a spectral representation for any symmetric matrix. The method of singular value decomposition (SVD) proposed by Autonne [99] is especially effective; its algorithm is given, in particular, in Refs [100, 101].

We note the connection of the approach described here with Karunen-Loeve's expansion [102], known in the probability theory of a stationary random process $\xi(t)$, into a series in terms of the eigenfunction system of its autocovariance function $\psi(\tau)$. Like the principal components, the coefficients $\left[\xi_{k}\right]$ of this expansion are uncorrelated random variables, which enables one to achieve the maximum information compression in the linear approach framework [103]. It is easy to see that in the case of image restoration the transition to principal components is equivalent to the generalisation of Karunen-Loeve's expansion to the case of a nonstationary process which includes an additional noise source-the photon noise (which is reflected in the dependence of the information matrix on the sought object $S_{0}$ ).

For the same reasons, the approach described above can be considered as a generalisation of the known method of
pseudoinversion of a linear system of equations (see Refs [100, 101]), which was used for image restoration in Refs [104-108]. Formal differences are due to the fact that we use the spectral representation not of the PSF but of the sample Fisher's matrix depending on the object.

### 6.3 Basic features of the image restoration algorithm

The algorithm used by us relies upon a procedure connected with transformations (6.21) and (6.22) which, however, is more complex.

The point is that the ellipsoidal shape of FER, and, as a consequence, the possibility of using only Fisher's matrix to describe it, are caused by adopting the likelihood function $L$ as a statistic for the estimate feasibility test. As shown in Section 4, proximity to the mean value of $L\left(y_{0} \mid S\right)$ is not sufficient for considering $S$ to be a feasible estimate; the image randomness test (IRT) requires deeper statistical investigation of the correspondence of the generated estimates $S$ with the observed image. The true region FER, satisfying all the IRT requirements, constitutes only a part of the ellipsoidal FER (see Fig. 4) The study of model examples shows that it is insufficient to enter into the ellipsoidal FER in order to achieve a good quality image restoration. At the same time, ellipsoidal FER reflects quite well the general form of the true FER, and, what is most important, correctly indicates its excessive elongation direction. Therefore it is reasonable to use Fisher's information matrix only for finding these directions and for establishing the corresponding principal component space, and to remove the remaining estimate uncertainty by a thorough investigation, in the IRT framework, of the statistical properties of the image ensemble generated by the IRT.

The above considerations, together with the local approach discussed in Section 4 and adequate choice of pixel sizes in the image and object space (see Section 7), provide a general description of the principles underlying the image restoration algorithm. Of course, particular details of the algorithm are not uniquely determined, nor is its numerical implementation. As work on optimising this algorithm still remains to be done, we will not enter into detail here; those interested can find a fuller description in Ref. [31].

It is quite obvious, that the same principles are valid for many other problems related to information inversion.

### 6.4 A numerical example

In order to visualise more clearly special features of the problem under discussion, we shall consider an example when the object includes both low-frequency and highfrequency components.

A randomly blurred image (Fig. 12b) was computed with a diffractive PSF [Eqn (5.32)] with $\Delta=7$ pixels for the case of a uniform Poisson background with a mean value $\gamma_{j}=100$ counts $/$ pixel. The MLE, with the object nonnegativeness $\hat{S}_{+}$taken into account, is shown in Fig. 12c; we see a typical manifestation of instability when the smooth component of the object is strongly distorted by random oscillations. Finally, Fig. 12d shows the solution $\tilde{S}$ discussed above, obtained in the simplest form when the entire object was optimised as a whole.

It is interesting to examine which eigenvalues of the information matrix $I$ have been neglected in this example. The whole sequence $\left[\lambda_{k}\right]$ is shown in Fig. 13; as in any


Figure 12. Model object (a), its image (b), conditional estimate of the maximum likelihood (c), and Occam's estimate (d).


Figure 13. Eigenvalues of Fisher's matrix for the example shown in Fig. 12.
unstable problem, it covers an enormous range-twelve orders of magnitude in this case! Usually this diapason is characterised by the condition number $C \equiv \sqrt{\lambda_{\max } / \lambda_{\min }}$, which is large for unstable problems. The principal components of the true object and its estimates are shown in Fig. 14. Again, we see a typical high-frequency noise in the MLE, which, however, is fully suppressed in the stable estimate $\tilde{S}$ above the 40 th principal component. The corresponding inversion into the object space is shown in Fig. 12d. Even more precise restoration can be achieved in the local approach framework.

## 7. Superresolution

The image restoration term itself contains the assumption that the visibility of sharp details improves during data processing, that is in the course of restoration we progress into a high spatial frequency region. A number of questions arise in this connection. First, we would like to know the factors on which the degree of restoration of shallow details depends on, whether a natural limit of attainable resolution exists, and if does exist, what is this limit under typical conditions.

It is clear that how far one progresses into the high spatial frequency region depends primarily on existing a


Figure 14. The principal component of the object (a) and its two estimates (b, c) shown in Figs $12 \mathrm{a}, \mathrm{b}$, and c , respectively.
priori information about the object and only to a lesser extent on its shape, signal-to-noise ratio, shape of the PSF, and other factors. For this reason, in order to find the limiting relationships, we assume in this section that so much a priori information is provided that the image restoration problem is reduced, in fact, to a pattern recognition one. Of course, also under these conditions the inverse problem is of major independent interest. For example, we can obtain simple analytical expressions characterising an appropriately defined 'resolving power' for observational conditions usually encountered in practice. If the investigator has not got so much a priori information at his disposal, then the limiting relationships allow him to estimate correctly the theoretical possibilities both in designing the apparatus and interpreting the data.

### 7.1 Rayleigh's problem

Experience shows that the definition of resolving power accepted in some field of study may not be satisfactory under other conditions. The classical definition of resolution suggested by Lord Rayleigh comes from considering a situation when the observer tries to determine whether the observed blurred image is generated by a single point-like source or by a binary object with components of the same
total brightness. A problem posed in this way will be called here Rayleigh's problem in the narrow sense, whereas distinguishing objects with arbitrary shapes will be designated Rayleigh's problem in the broad sense. For the first of these problems the Rayleigh resolution limit is, in fact, taken to be equal to an appropriately chosen PSF width $\Delta$. If the image quality is limited only by diffraction of radiation on the aperture of the image formation system of diameter $D$, Rayleigh's limit coincides with the diffraction limit

$$
\begin{equation*}
\Delta_{\mathrm{d}} \simeq \frac{\lambda}{D} \tag{7.1}
\end{equation*}
$$

where $\lambda$ is the radiation wavelength and $\Delta_{\mathrm{d}}$ is measured in radians.

The arbitrary nature of that definition was clear to Rayleigh himself, who noted ([2], p. 420) that '"The rule is convenient on account of its simplicity; and it is sufficiently accurate in view of necessary uncertainty as to what is meant by resolution'. Indeed, if both types of noise photon and external - were entirely absent, one would be able to distinguish arbitrarily close sources; to do this, for example, one could expand the observed image into a Taylor series and trace sufficiently high terms of this series. Even more graphic can be Fourier-series expansion, in which the binarity appears in the form of deep minima in the power spectrum for harmonics with 'unsuitable' frequencies. The presence of noise due to both the external background and the quantum nature of light, principally complicates the problem in view of the instability of the inverse solution. Nevertheless, resolution of a source with the minimum distance between point components $\rho_{\text {min }}$ less than the width of PSF $\Delta$ is possible even in the presence of noise if one uses image restoration methods (Fig. 15). The aforesaid does not diminish the importance of Rayleigh's criterion; one must only bear in mind the limits of its field of application. For example, the use of Rayleigh's criterion is quite appropriate in visual data analysis.

We introduce for convenience a dimensionless resolution parameter

$$
\begin{equation*}
\mathcal{R} \equiv \frac{\rho_{\min }}{\Delta} \tag{7.2}
\end{equation*}
$$

and will speak of superresolution, if $\mathcal{R}<1$ is achieved. Of course, having one parameter is not sufficient for describing all aspects of the limiting resolution problem and even less so for dealing with the Rayleigh problem in the broad sense. The corresponding notions will be introduced later, as needed.

In the case of determinate blurring the superresolution phenomenon was discovered by Schelkunoff [109] and then studied in detail by others [45, 110-119]. An excellent description of the early investigations is given in Rautian's review [112]. The discussion that follows will be based upon papers published earlier [26, 28, 120], in the first of which the problem is solved for objects of arbitrary shape in the framework of general Neyman-Pearson theory of statistical hypothesis testing; the second paper generalises the results, taking into account the photon noise, and the third describes numerical simulations which complete the analytical consideration. A similar approach was used earlier by Kozlov [114], Harris [115, 116] and Snyder (see Ref. [119]). I shall refrain here from enumerating fairly obvious applications of the problem of limiting resolution to


Figure 15. (a) A double object with point components of equal brightness and one of its blurred images (the near-horizontal curve); (b) double object image on a larger scale; (c, d) examples of image restoration.
practical issues in optics, electron microscopy, tomography, and so on.

### 7.2 Image restoration and pattern recognition

No matter how the notion of limiting resolution power is defined, Rayleigh's setting of the problem involves a comparison of two or more alternative objects, one of which must be preferred on the grounds of a priori information and the observed image realisation. In such cases one speaks of pattern recognition, whereas when there is almost no a priori information available about the object generating the image, one speaks of image restoration. The line separating these notions is not quite uniquely defined, but this cannot lead to ambiguities. The essence is that object classification presumes estimating one or several parameters that describe subdivision of the object into classes, whereas construction of brightness distribution in the object requires estimating a large number of parameters
that can be represented, for example, by intensity in the individual pixels.

Obviously, the question of detection of a given object in the presence of noise cannot be solved in this setting: given the same signal-to-noise ratio $S / N$ and any selection test, we will prefer one or another alternative depending on the particular noise realisation. The problem is of necessity statistical, and we shall consider it from the point of view of statistical hypothesis testing theory developed by Neyman and Pearson (see [67, 69]).

Let us consider a basic case when two alternative objects are possible: $S_{0}$ and $S_{1}$ (say, a single star and a double star of the same brightness with a given separation between the components). Let $q_{0}=\left[q_{0 j}\right]$ and $q_{1}=\left[q_{1 j}\right]$ be the mean brightness distributions, like that given by expression (2.3), corresponding to the objects, and $y=\left[y_{j}\right]$ be the observed image of the unknown object. Naturally, two hypotheses about the nature of the object $S$ generating the image are possible:

$$
\begin{array}{ll}
\mathcal{H}_{0}: & S=S_{0}, \\
\mathcal{H}_{1}: & S=S_{1} \tag{7.3}
\end{array}
$$

One needs to construct a decisive rule (criterion) for choosing one of the hypotheses for the given image $y$, and to estimate its quality (that is the errors connected with its application). The most general formulation of the selection criterion is as follows: if $y$ belongs to some critical region $w$ in an $m$-dimensional image space, hypothesis $\mathcal{H}_{1}$ is accepted, and when $y$ lies outside the critical region, hypothesis $\mathcal{H}_{0}$ is accepted. The problem is thus reduced to the determination of the optimal, in some sense critical, region $w$ on the basis of available information about the background, image formation system, and the objects themselves.

In any choice of $w$ two kinds of errors are possible: (1) hypothesis $\mathcal{H}_{0}$ will be rejected when the image is generated by object $S_{0}$; (2) hypothesis $\mathcal{H}_{0}$ will be accepted when the image is generated by object $S_{1}$. We denote by $\alpha$ and $\beta$ the corresponding probabilities of errors of the first and the second kind, i.e.

$$
\begin{equation*}
\alpha=\operatorname{Pr}\left(\mathcal{H}_{1} \mid S_{0}\right), \quad \beta=\operatorname{Pr}\left(\mathcal{H}_{0} \mid S_{1}\right) \tag{7.4}
\end{equation*}
$$

Usually one calls $\alpha$ the test significance level, and $1-\beta$ its power.

Neyman-Pearson's approach to the hypothesis testing problem is that one should specify first the significance level $\alpha$ and then look for a such critical region $w_{\alpha}$ for which the probability of the second kind of error $\beta$ is minimum (that is, the test power $1-\beta$ is maximum). The choice of $\alpha$ takes into account the relative importance of the two types of error. For example, if we form a sample of quasars $\left(S_{0}\right)$ in a field of stars $\left(S_{1}\right)$, then the $\alpha$ is the probability of missing a quasar, and to obtain a more complete sample we should specify a comparatively small $\alpha$, say $\alpha=0.1$. If $\alpha$ is made too small, the sample will cover almost all the quasars, but too many stars will be included as well. Similar considerations usually lead to the investigator being interested in minimising $\beta$ at a fixed $\alpha$, thus posing the problem in the Neyman Pearson sense. The corresponding region $w_{\alpha}$ is called the best critical region, and the selection test obtained on its basis is the most powerful (theoretically not capable of being improved).

Obviously, availability of information about a comparatively small class of specified objects in pattern recognition simplifies solution of the inverse problem compared to direct image restoration when only the condition of nonnegativeness of the original is given. For this reason, the highest resolution power obtained in the framework of the pattern recognition theory must be considered as the upper limit of the resolution power that can be attained by the most efficient image restoration. Thus, by turning to the hypothesis testing theory, we not only obtain results that are of interest for their own sake, but also establish the limiting capabilities of the inverse problem theory.

### 7.3 Analytical results

Referring the reader to Refs [26, 28] for proofs, I present here explicit expressions for the limiting resolution obtained by juxtaposing two types of objects of arbitrary shape. For the Poisson distribution we use Gaussian approximation with the same mean value and variance; for close objects we assume $\ln \left(q_{1 j} / q_{0 j}\right) \simeq\left(q_{1 j}-q_{0 j}\right) / q_{0 j}$.

A 'distance' $\kappa\left(S_{0}, S_{1}\right)$ between the alternative objects defined as

$$
\begin{equation*}
\kappa\left(S_{0}, S_{1}\right)=\left[\sum_{j=1}^{m} \frac{\left(q_{1 j}-q_{0 j}\right)^{2}}{q_{0 j}}\right]^{1 / 2} \tag{7.5}
\end{equation*}
$$

naturally enters into the problem. Let, as above, $Y$ be a random vector whose realisation produces the observed image $y$. If hypothesis $\mathcal{H}_{0}$ is correct, the statistic

$$
\begin{equation*}
\tau\left(Y, S_{0}, S_{1}\right)=\kappa^{-1} \sum_{j=1}^{m} \frac{\left(Y_{j}-q_{0 j}\right)\left(q_{1 j}-q_{0 j}\right)}{q_{0 j}} \tag{7.6}
\end{equation*}
$$

obeys standard Gaussian distribution with zero mean and unit variance, that is $\left.\tau\right|_{\mathcal{H}_{0}} \sim \Phi(z)$, where

$$
\begin{equation*}
\Phi(z)=\frac{1}{\sqrt{2 \pi}} \int_{-\infty}^{z} \exp \left(-\frac{x^{2}}{2}\right) \mathrm{d} x \tag{7.7}
\end{equation*}
$$

It follows from this that the best critical region $w_{\alpha}$ is determined by the condition

$$
\begin{equation*}
\tau\left(y, S_{0}, S_{1}\right) \geqslant z_{\alpha} \tag{7.8}
\end{equation*}
$$

where $z_{\alpha}$ is a quantile of normal distribution of the order of $1-\alpha$, that is the root of the equation

$$
\begin{equation*}
\Phi(z)=1-\alpha \tag{7.9}
\end{equation*}
$$

When inequality (7.8) is valid, hypothesis $\mathcal{H}_{1}$ is accepted, and if the opposite inequality holds, one accepts hypothesis $\mathcal{H}_{0}$. We shall see below that the only significant parameter entering inequality (7.8) is the signal-to-noise ratio $\psi$.

The meaning of expression (7.6) is quite clear: its structure resembles the formula for the correlation coefficient between two random variables. If the differences $y_{j}-q_{0 j}$ have more often the same sign as $q_{1 j}-q_{0 j}$, the sum in expression (7.6) will be large, and inequality (7.8) requires that hypothesis $\mathcal{H}_{0}$ be rejected. But correlation of the signs of the aforementioned differences does in fact mean that object $S_{1}$ is preferable to $S_{0}$ for the explanation of image $y$.

As has been said before, the minimum probability of the second kind of error corresponds to the choice of the best critical region; this minimum value is

$$
\begin{equation*}
\beta=\Phi\left[z_{\alpha}-\kappa\left(S_{0}, S_{1}\right)\right] \tag{7.10}
\end{equation*}
$$

The formulas given above provide the complete solution to the problem about the two-alternative choice of the object, so that with given $\alpha, \kappa\left(S_{0}, S_{1}\right)$, and $\psi$, one can most efficiently choose one of the alternatives, and then estimate the acceptable probability of the second kind of error. However, to determine the limiting resolution under given conditions, we are more interested in the minimum distance $\kappa\left(S_{0}, S_{1}\right)$ which can be established for a fixed reliability and $\psi$. The corresponding expression is obtained by straightforward inversion of expression (7.10) with account taken of equality (7.9) and the relation $z_{1-\alpha}=-z_{\alpha}$ :

$$
\begin{equation*}
\kappa\left(S_{0}, S_{1}\right)=z_{\alpha}+z_{\beta} \tag{7.11}
\end{equation*}
$$

Formula (7.11) is the one we have been looking for; it permits calculation of the minimum distance $\kappa\left(S_{0}, S_{1}\right)$ at a fixed reliability level $(\alpha, \beta)$ for alternative objects of arbitrary shape and a given image formation model.

To understand more clearly the expressions given above, it is useful to consider their continuous analogue, by assuming, for the sake of simplicity, that the background is uniform ( $\gamma_{j} \equiv \gamma=$ const). Let $f_{\mathrm{c}}$ be the cut-off frequency of the image formation system, so that the MTF of the system $H(f) \equiv 0$ at $|f|>f_{\mathrm{c}}$, and $\Delta=f_{\mathrm{c}}^{-1}$ is a characteristic width of the PSF. We denote by $F$ the total brightness of each of the compared objects, and by $\psi$ the signal-to-noise ration in the image interval with a length of the order of the width of the PSF:

$$
\begin{equation*}
\psi=\frac{F}{\sqrt{\gamma \Delta}} \tag{7.12}
\end{equation*}
$$

Making use of Parseval's theorem, we find from Eqn (7.5)

$$
\begin{equation*}
\kappa\left(S_{0}, S_{1}\right)=\psi\left[\int_{-f_{\mathrm{c}}}^{f_{\mathrm{c}}}|H(f)|^{2}\left|\frac{\delta S(f)}{F}\right|^{2} \frac{\mathrm{~d} f}{f_{\mathrm{c}}}\right]^{1 / 2} \tag{7.13}
\end{equation*}
$$

where $\delta S(f)$ is the difference of Fourier transforms of brightness distributions in the considered objects. By substituting expression (7.13) into relation (7.11), we arrive at a general relationship

$$
\begin{equation*}
\left[\int_{-f_{\mathrm{c}}}^{f_{\mathrm{c}}}|H(f)|^{2}\left|\frac{\delta S(f)}{F}\right|^{2} \frac{\mathrm{~d} f}{f_{\mathrm{c}}}\right]^{1 / 2}=\frac{z_{\alpha}+z_{\beta}}{\psi} \tag{7.14}
\end{equation*}
$$

The functional in the left-hand side of this equality characterises the measure of closeness of the two objects that are being compared; in simple cases it depends on one parameter only, an appropriately defined limiting resolution $\mathcal{R}$ [see definition (7.2)]. It is indicative that all external conditions determining resolution are grouped together in the form of the combination

$$
\begin{equation*}
t \equiv \frac{z_{\alpha}+z_{\beta}}{\psi} \tag{7.15}
\end{equation*}
$$

This means [26] that the limiting resolution does not depend on many individual variables describing the observational conditions (object brightness, background level, PSF width, identification reliability, etc.), but depends only on their combination (7.15). One should also note that the probabilities $\alpha$ and $\beta$ enter into expression (7.15) symmetrically.

### 7.4 Examples

As usual, an analytical study of simple particular cases brings to light general dependences; such cases are almost always of interest for their own sake.
7.4.1 Shift of an arbitrary object. Let us suppose that the difference between the objects being compared consists only of shifting one relative to another by a distance $\rho$, so that $S_{1}(x)=S_{0}(x-\rho)$. Then formulae (7.2) and (7.14) at $\psi \gg 1$ yield

$$
\begin{equation*}
\mathcal{R} \propto \frac{z_{\alpha}+z_{\beta}}{\psi} \tag{7.16}
\end{equation*}
$$

where the proportionality constant is of the order of unity depending on the form of the PSF and the brightness distribution of object $S_{0}$. Please note that the limiting smallest shift which can be detected is proportional to the first power of the signal-to-noise ratio (Fig. 16).


Figure 16. Schematic representation of the relationship between the limiting resolution $\rho_{\min }$ in units of the PDF width $\Delta$ and the signal-tonoise ratio: 1 -shift of the object, 2 -separation of the double source with a single point object as an alternative, 3 -the same as 2 but for a Gaussian alternative object.
7.4.2 Separation of a source with point components. Consider the Rayleigh problem in the narrow sense for an object with point components of equal intensity $F / 2$, which is studied by a device with a diffractive PSF like that given by expression (5.32). Then the limiting resolution measured as a fraction of the width of the PSF is

$$
\begin{equation*}
\mathcal{R} \simeq\left(\frac{z_{\alpha}+z_{\beta}}{\psi}\right)^{5 / 8} \tag{7.17}
\end{equation*}
$$

where a more precise signal-to-noise definition than that given by expression (7.12) includes the photon noise as well:

$$
\begin{equation*}
\psi=\frac{F}{\sqrt{F+2 \gamma \Delta}} \tag{7.18}
\end{equation*}
$$

As seen from expression (7.17), the dependence of $\mathcal{R}$ on $\psi$ is in this case close to $\psi^{-1 / 2}$. In practice, $\psi$ can often be as high as $10^{4}-10^{6}$, so that under such conditions one can expect the resolution to exceed the Rayleigh limit by $10^{2}-10^{3}$ times (see Fig. 16). The reality of such values is confirmed by the simulations described below and by practical results.
7.4.3 A binary source with point components and a Gaussian object. Let the alternative objects be a binary star with a distance $\rho$ between the components and an object with a continuous brightness distribution described by Gaussian law with $\sigma=\rho / 2$. The two objects are assumed to have equal total brightness. Then it follows from expression (7.14) that

$$
\begin{equation*}
\mathcal{R} \simeq\left(\frac{z_{\alpha}+z_{\beta}}{\psi}\right)^{1 / 4} \tag{7.19}
\end{equation*}
$$

The decrease in the value of the exponent compared to expressions (7.16) and (7.17) testifies that in the given case one cannot reach resolutions as high as that obtained by shifting the object or by juxtaposing point sources (see Fig. 16). We note that in the absence of background we would have $\psi=\sqrt{F}$, and expression (7.19) would give $\mathcal{R} \propto F^{-1 / 8}$. The last relation was independently obtained by Lucy [121] by comparing the fourth moments of the observed images of the alternative sources.

### 7.5 Monte-Carlo simulations

To test the theory as well as to make a more profound study of some its aspects that cannot as yet be analytically examined (in particular, the form of the statistical error distribution function), it is necessary to perform numerical simulations of the image formation process and of subsequent image restoration. Here we shall consider part of the results reported in Ref. [120] for the classical Rayleigh problem.

The following information is supposed to be known: (1) a randomly blurred and noisy image was generated by an object consisting of two noncoherent point sources of equal brightness; (2) the total brightness of the object $F$ and the mean level of the uniform background $\gamma$ are specified;
(3) fluctuations of the count number obey Poisson distribution; (4) the PSF $h\left(x-x^{\prime}\right)$ has the diffractive form (5.32). On the grounds of a priori information and the observed image realisation, it is required to estimate the distance between the components of the double source as accurately as possible (the distance can also be zero).

We denote the coordinates of the components by $x_{1}$ and $x_{2}$, and their true relative distance by $\theta=\left|x_{1}-x_{2}\right| / \Delta$. In the course of the numerical simulations each of the components independently of each other was randomly blurred with 'photon after a photon', and then a random realisation of the background was added to the blurred image. The resulting image obviously obeys the multidimensional Poisson distribution (2.4) with mean counts
$q_{j}=\frac{F}{2}\left[h\left(j-x_{1}\right)+h\left(j-x_{2}\right)\right]+\gamma_{j}, \quad j=1, \ldots, m$.

Next, the coordinate estimates $\hat{x}_{1}$ and $\hat{x}_{2}$ are found and, finally an estimate of their relative distance $\hat{\theta}=\left|\hat{x}_{1}-\hat{x}_{2}\right| / \Delta$ is obtained. To construct sample distribution densities $p(t \mid \theta)$ for the estimates $\hat{\theta}$ the procedures described above were repeated tens to hundreds of thousands times for each $\theta$.

We note that here we consider a more general model than that discussed in Sections 7.2-7.4, where one had to choose one of only two types of alternative objects. At the same time, the present model appears to be closer to the practical conditions, since the precise distance between the components is very rarely known in advance.


Figure 17. Sample distribution densities for the estimate $\hat{\theta}$ versus the true component separation $\theta$.

In order to ensure that the high-frequency information about the original is retained and in the present case to have a possibility, if needed, of taking estimates of the coordinates of the components beyond the limits of one pixel, one should provide for a sufficiently small pixel size in the object space $p_{0}$. In contrast to that, in view of the well known sampling theorem it makes no sense to take the pixel size in the object space $p_{\mathrm{i}}$ much smaller than the PSF width
$\Delta$. The precise relationship between $p_{\mathrm{i}}$ and $p_{\mathrm{o}}$ depends on the signal-to-noise ratio. Indeed, relation (7.17) predicts that the minimum detectable component separation is approximately $\sqrt{\psi}$ times smaller than $\Delta$; hence one should take $p_{0}$ smaller than $\Delta$ by about the same factor.

Estimates of the component positions were obtained by using the maximum likelihood method. In Sections 3 and 4 it was mentioned that multidimensional MLEs are unstable; in
the present case, however, one needs to estimate only two parameters, so that there are grounds to expect good MLE efficiency. This is consistent with the results of test calculation.

Fig. 17 shows the results of simulations for the case when $F=10^{4}$ events; $\Delta=100$ pixels, $\gamma_{j} \equiv 10$ events/pixel. Such a broad PSF was chosen because of the danger of making the results of restoration unnecessarily coarse, as explained above. First, the calculations were done at $p_{\mathrm{i}}=p_{\mathrm{o}}=\Delta / 100$ (control computations with twice as fine discretisation do not change the results).

As one would expect, in the case of widely spaced pairs the calculated estimates $\hat{\theta}$ are closely grouped around the true values $\theta$. As the component separation $\theta$ decreases, the variance of estimates $\hat{\theta}$ increases and cases when the object is taken to be a single one occur more and more frequently. The peak of the sample distribution $p(t \mid \theta)$ at $t=0$ is due to the nonequivalence of the results for too tight and too open image realisations. After some critical value $\theta \simeq 0.06-0.08$ is reached, the density $p(t \mid \theta)$ remains practically unchanged, so that restoration of the shape of the object generating the observed image becomes impossible. This means that the aforementioned value of $\theta$ should be considered as being a limiting one under the given conditions. The existence of the limit is seen more clearly in Fig. 18, which shows how the mean value $\langle\hat{\theta}\rangle$ changes as the components become closer to each other. The limiting resolution obtained from the simulation is in good agreement with the theoretical limit (7.17).

As mentioned earlier, in virtue of the La Valle-Poussin -Kotel'nikov-Shannon sampling theorem (see Refs [122, $123,7,8]$, the choice of small image pixels $p_{\mathrm{i}}=0.01 \Delta$ is unjustified. According to this theorem, to preserve all the information about a determinate function with a limiting frequency in the spectrum $f_{\mathrm{c}}$, it is necessary to ensure a sampling rate equal to at least $2 f_{\mathrm{c}}$. In the case under consideration $f_{\mathrm{c}}=\Delta^{-1}$, and we must locate not less than two samples within the PSF 'radius' $\Delta$. The stochastic nature of the image and the departure of samples from


Figure 18. Relationship between the sample mean $\langle\hat{\theta}\rangle$ and the true component separation $\theta$.
being point-like strongly complicate analytical approach, so during the numerical simulations performed there was a good opportunity to determine the required ratio between the optimal size of the image pixel and the PSF width.

Fig. 19 shows sample distribution densities $\hat{\theta}$ for $F=10^{4}$ events, true component separation $\theta=0.10$, and a number of ratios $p_{\mathrm{i}} / \Delta$. We see that when $p_{\mathrm{i}} / \Delta \leqslant 0.5$, that is when the requirements of the sampling theorem are fulfilled, the restoration accuracy changes comparatively little, whereas when the image pixels are too coarse, when $p_{\mathrm{i}} / \Delta>0.5$, it rapidly decreases with increasing image pixel size.


Figure 19. Sample distribution densities for the estimate $\hat{\theta}$ for $\theta=0.10$ and a number of ratios $p_{\mathrm{i}} / \Delta$.

One should draw attention to the fact that when the distance between the components is reduced, the transition from reliable detection of the binarity to total indeterminacy occurs comparatively rapidly and covers only about $10 \%$ of the PSF width. For this reason one can expect to obtain useful results in a sufficiently broad region of resolutions between the Rayleigh limit and that given by relation (7.17).

## 8. Concluding remarks

The need to remove by some means excessive information contained in the inverse solution has been felt for quite a long time. For example, Snyder et al. [124, 125] suggested smoothing of the maximum likelihood estimate, and Mints and Prilepskii [32], and Pina and Puetter [33] first introduced into the object space a rough structure corresponding to the object averaged over specially selected areas. Both ways point in the right direction; the problem, however, is to avoid subjective considerations in the choice of a smoothing kernel or a large-granular structure. Moreover, this choice should not be final, but should depend on the studied object. Transition to the principal components of the sample Fisher's matrix (that is generated by a specific object) does in fact play the role of Occam's razor in the selection of useful information. One can show, in particular, that this transition is equivalent to considering the associated inverse problem with a coarser PSF depending on the object.

Thus, Fisher's information matrix plays a fundamental role in all inverse problems, irrespective of whether we deal
with solving a system of equations, a tomography problem, or are concerned with image restoration.

On the other hand, a significant difference in the variances of principal components in the original estimate naturally produces nonequivalence of solutions in the object space, which can be considered as some analogy of the a priori density $w(S)$ invoked in the Bayesian approach. However, the nonequivalence is in this case objectively caused by the conditions of the problem.

In the case of determinate smoothing and Toeplitz's PSF, the eigenvalues of the matrix $I$ coincide with the Fourier power spectrum $h(x)$, and the corresponding eigenvectors are equal to $\exp (\mathrm{i} 2 \pi f x)$, where $f$ is spatial frequency. Then the principal components of the inverse estimate are maximum estimates of the absolute values of Fourier coefficients. At first glance it would appear that to obtain a stable estimate of the object in the general case we can simply cut off in the restoration the high-frequency tail of the power spectrum of the object. The true nature of the proposed procedure goes far beyond this approach.

First, we take into account the unavoidable photon noise in the image, which makes the image restoration problem principally a local one (as is intuitively clear: bright parts of the object must be restored with higher accuracy). Formally, this property manifests itself in the PSF being 'weighted' by a factor $\sqrt{Q(S)}$. Second, instead of removing the high-frequency tail of the power spectrum, an optimal statistical estimation of the principal components is made. In this, not necessarily all those principal components that correspond to low spatial frequencies are retained, but only those having the maximum statistical weight. Finally, instead of the same trigonometrical functions being used for all objects, the eigenvectors of $I$ related to each particular problem are naturally employed.

In connection with the problem of discretisation in the object space $[126,28,127]$ we note that introduction of the principal components automatically preserves only the really available information even in the case when pixels have been chosen that are too small.

When searching for the optimal solution in the feasible estimates region, like the one a two-dimensional version of which is shown as an example in Fig. 4, it is useful to bear in mind some multidimensionality effects that conflict with intuition (see, e.g., Ref. [128]). For example, when the number of measurements $n \gg 1$, an arbitrarily thin 'surface layer' of a hypersphere with radius $r$ covers almost all its volume $V_{n}(r)$. Indeed, we have

$$
\begin{equation*}
\frac{V_{n}[(1-\varepsilon) r]}{V_{n}(r)}=(1-\varepsilon)^{n} \rightarrow 0 \tag{8.1}
\end{equation*}
$$

when $n \rightarrow \infty$ for any $\varepsilon \in(0,1)$. Therefore, undesirable FER regions where unstable estimates are located are much broader than it would appear by considering a twodimensional picture.

Since interpretation of experimental data belongs to the inverse class of problems, the difficulties involved in information inversion are encountered everywhere. One manages to avoid them only when the model chosen is too coarse, such that instability of the inverse solution is weak. If one tries to extract from data the information they contain, the problem almost always becomes more complicated, and we are forced to invoke the notions introduced above.

Applications of the approach presented here to the problem of compensating for atmospheric distortions of the image and to the restoration of ordinary and tomographic images can be found in Refs [127, 129].

In the field of numerical data analysis there has been very substantial progress during the last years in devising improved numerical algorithms and better hardware, so that this aspect of the image restoration problem deserves separate consideration.

Acknowledgements. The author is grateful to V V Biryukov (Moscow State University), E Veklerov (Lawrence Berkeley Laboratory, University of California), R E Gershberg (Crimean Astrophysical Laboratory), D L Snyder (Electronic Systems and Signals Research Laboratory, Washington University), O K Cheburnina and Yu G Cheburnin (Moscow State University) for useful discussions of the problems touched upon here.

## References

1. Frieden B R, in Picture Processing and Digital Filtering (Ed. T S Huang) (Berlin: Springer, 1979) p. 179
2. Lord Rayleigh (Strutt J W) Scientific Papers Vol. 1 (New York: Dover, 1964)
3. Schuster A Terrest. Magn. 313 (1898)
4. Schuster A Proc. R. Soc. London 77136 (1906)
5. Feynman R P, Leighton R B, Sands M The Feynman Lectures in Physics Vol. 2, Ch. 40 (Reading, MA: Addison-Wesley, 1964) [Translated into Russian (Moscow: Mir, 1966)]
6. Terebizh V Yu Proc. 3rd ESO/ST-ECF Data Analysis Workshop (Eds P J Grosbol, R H Warmels) (Garching: ESO Conf. and Workshop Proc. No. 38, 1991)
7. Shannon C Bell Syst. Techn. J. 27 379, 623 (1948)
8. Shannon C Proc. IRE 3710 (1949)
9. Bayes $T$ "An essay towards solving a problem in the doctrine of chances" (1763) [Reprinted in Biometrika 45293 (1958)]
10. Zacks S The Theory of Statistical Inference (New York: Wiley, 1971) [Translated into Russian (Moscow: Mir, 1975)]
11. Kolmogorov A N Izv. Akad. Nauk SS SR, Ser. Mat. 53 (1941)
12. Wiener N Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications (MIT NDRC Report, 1942; Reprinted by MIT Press, 1949)
13. Turchin V F, Kozlov V P, Malkevich M S Usp. Fiz. Nauk 102 345 (1970)
14. Phillips D L J. Assoc. Comput. Mach. 984 (1962)
15. Twomey S J. Assoc. Comp. Mach. 1097 (1963)
16. Twomey S J. Franklin Inst. 27995 (1965)
17. Tikhonov A N Dokl. Akad. Nauk SSSR 151501 (1963)
18. Tikhonov A N Dokl. Akad. Nauk SSSR 15349 (1963)
19. Jaynes E T Phys. Rev. 106620 (1957)
20. Jaynes E T Phys. Rev. 108171 (1957)
21. Burg J P, Paper presented at 37th Ann. Soc. of Expl. Geophys. Mtg. (Oklahoma City, OK: Stanford Univ. Geophys. Dept., 1967)
22. Frieden B R J. Opt. Soc. Am. 62511 (1972)
23. Veklerov E, Llacer J IEE E Trans. MI 6313 (1987)
24. Llacer J, Veklerov E IEE E Trans. MI 8186 (1989)
25. Terebizh V Yu Astrofizika 32327 (1990) [Astrophysics 32184 (1990)]
26. Terebizh V Yu Astrofizika 33409 (1990) [Astrophysics 33536 (1990)]
27. Terebizh V Yu Astron. Astrophys. Trans. 13 (1991)
28. Terebizh V Yu Astron. Astrophys. 270543 (1993)
29. Terebizh V Yu, Biryukov V V Ast ron. Astrophys. Trans. 6 (37) (1994)
30. Terebizh V Yu, Biryukov V V Dokl. Ross. Ak ad. Nauk (in press)
31. Terebizh V Yu, Biryukov V V Astrophys. Space Sci. 21865 (1994)
32. Mints M Ya, Prilepskii E D Opt. Spectrosk. 75696 (1993) [Opt. Spectrosc. 75413 (1993)]
33. Pina R K, Puetter R C PASP 105630 (1993)
34. Helstrom C W J. Opt. Soc. Am. 59164 (1969)
35. Helstrom C W J. Opt. Soc. Am. 60659 (1970)
36. Mehta C L, in Progress in Optics Vol. 8 (Ed. E Wolf) (Amsterdam: North-Holland, 1970) p. 373
37. Troup G G, in Progress in Quantum Electronics Vol. 2 (Eds J H Sanders, S Stenholm) (1972) p. 1
38. Terebizh V Yu Analiz Vremennykh Ryadov v Astrofizike (Time Series Analysis in Astrophysics) (Moscow: Nauka, 1992)
39. Cox D R J. R. Statist. Soc. B 17129 (1955)
40. Mandel L Proc. Phys. Soc. London 721037 (1958)
41. Mandel L Proc. Phys. Soc. London 74233 (1959)
42. Snyder D L, Hammoud A M, White R L J. Opt. Soc. Am. A 101014 (1993)
43. Borovkov A A Matematicheskaya Statistika (Mathematical Statistics) (Moscow: Nauka, 1994)
44. Fisher R A Philos. Trans. A 222309 (1921)
45. Toraldo di Francia G J. Opt. Soc. Am. 45497 (1955)
46. Seber G A F Linear Regression Analysis (New York: Wiley, 1977) [Translated into Russian (Moscow: Mir, 1980)]
47. Ermakov C M, Zhiglyavskii A A Mat ematicheskaya Teoriya Optimal'nogo Eks perimenta (Mathematical Theory of Optimal Experiment) (Moscow: Nauka, 1987)
48. Morris J Philos. Mag. 37 (7) 106 (1946)
49. Bertero M, in Inverse Problems (Ed. G Talenti) (Berlin: Springer, 1986) p. 52
50. Booth A D Numerical Methods (London: Butterworths, 1955) [Translated into Russian (Moscow: Fizmatgiz, 1959)]
51. Lanczos C Applied Analysis (New York: Prentice Hall, 1956) [Translated into Russian (Moscow: Fizmatgiz, 1961)]
52. Faddeev D K Tr. Mat. Inst. Akad. Nauk SS SR 53387 (1959)
53. Faddeev D K, Faddeeva V N Zh. Vychisl. Mat. Mat. Fiz. 1 412 (1961)
54. Bard Y Nonlinear Parameter Estimation (New York: Academic Press, 1974) [Translated into Russian (Moscow: Statistika, 1979)]
55. Fisher R A Philos. Trans. R. Soc. London A 222309 (1922)
56. Fisher R A Messenger of Mat hematics 41150 (1912)
57. Aitken A C, Silverstone H Proc. R. Soc. Edinburgh A 61186 (1942)
58. Frechet M Rev. Inst. Int. Stat. 182 (1943)
59. Darmois G Rev. Inst. Int. Stat. B 9 (1945)
60. Rao C R Bull. Calcutta Math. Soc. 3781 (1945)
61. Cramer H Mat hematical Methods of Statistics (New York: Princeton University Press, 1946) [Translated into Russian (Moscow: Mir, 1975)]
62. Hoerl A E, Kennard R W Technometrics 1255 (1970)
63. Hoerl A E, Kennard R W Technometrics 1269 (1970)
64. Edgeworth E Y J. R. Stat. Soc. 71 381, 499, 651 (1908)
65. Edgeworth E Y J. R. Stat. Soc. 7281 (1909)
66. Wiener N Cybernetics (New York: M.I.T. Press and Wiley, 1961) [Translated into Russian (Moscow: Nauka, 1983)]
67. Cox D R, Hinkley D V Theoretical Statistics (New York: Chapman and Hall, 1974) [Translated into Russian (Moscow: Mir, 1978)]
68. Kullback S, Leibler R A Ann. Math. Stat. 2279 (1951)
69. Kendall M G, Stuart A The Advanced Theory of Statistics Vols. 2, 3 (London: Griffin, 1969) [Translated into Russian (Moscow: Nauka, 1973)]
70. Bol'shev L N, Smirnov N V Tablitsy Matematicheskoi Statistiki (Tables of Mathematical Statistics) (Moscow: Nauka, 1965)
71. Kolmogorov A N Probl. Peredachi Inf. 13 (1965)
72. Kolmogorov A N Probl. Peredachi Inf. 53 (1969)
73. Von Mises R Wahrscheindlichkeit, Statistik und Wah rheit (Berlin: Springer, 1928)
74. Solomonoff R Inf. Control 71 (1964)
75. Martin-Lof P Inf. Control 9602 (1966)
76. Zvonkin A K, Levin L A Usp. Mat. Nauk 25(6) 85 (1970) [Russ. Math. Surv. 25 (6) (1970)]
77. Ford J Phys. Today 3640 (1983)
78. Rose A Vision: Human and Electronic (New York: Plenum Press, 1973) [Translated into Russian (Moscow: Mir, 1977)]
79. Bershad N J J. Opt. Soc. Am. 59157 (1969)
80. Born M, Wolf E Principles of Optics (Oxford: Pergamon Press, 1964) [Translated into Russian (Moscow: Nauka, 1970)]
81. Davenport W B, Root W L An Introduction to the Theory of Random Signals and Noise (New York: McGraw-Hill, 1958) [Translated into Russian (Moscow: Instr. Lit., 1960)]
82. Cole E R, PhD Dissertation, Department of Electrical Engineering, University of Utah, Salt Lake City, UT (1973)
83. Pinsker M S Dokl. Akad. Nauk SSSR 99213 (1956)
84. Fellgett P B, Linfoot E H Philos. Trans. R. Soc. London A 247 369 (1955)
85. Gel'fand I M, Yaglom A M Usp. Mat. Nauk 123 (1957)
86. Terebizh V Yu Astrofizika (in press)
87. Airy G B Trans. Cambr. Philos. Soc. 5283 (1835)
88. O’Neill E L J. Opt. Soc. Am. 46 285, 1096 (1956)
89. O'Neill E L Introduction to Statistical Optics (Reading, M A: Addison-Wesley, 1963) [Translated into Russian (Moscow: Mir, 1966)]
90. Hotelling H J. Educ. Psych. 24 417, 498 (1933)
91. Einstein A, Infeld L The Evolution of Physics (New York: Simon and Schuster, 1942) [Translated into Russian (Moscow: Nauka, 1967) Vol. 4, p. 357]
92. Kramer H P, Mathews M V IRE Trans. Inf. Theory IT-2 41 (1956)
93. Rao C R Sankhya A 26329 (1964)
94. Rao C R Linear Statistical Inference and Its Applications (New York: Wiley, 1965) [Translated into Russian (Moscow: Nauka, 1968)]
95. Brillinger D R Time Series Analysis (New York: Holt, Rinehart and Winston, 1975) [Translated into Russian (Moscow: Mir, 1980)]
96. Biraud Y Astron. Astrophys. 1124 (1969)
97. Bertsekas D P Constrained Optimization and Lagrange Multiplier Methods (New York: Academic Press, 1982) [Translated into Russian (Moscow: Radio i Svyaz', 1987)]
98. Aivazyan C A et al. Prikladnaya Statistika. Klassifikat siya i Snizhenie Razmernosti (Applied Statistics. Classification and Reduction of Dimensions) (Moscow: Fin. i Stat., 1989)
99. Autonne L Ann. Univ. Lion. 381 (1915)
100. Lawson C L, Hanson R J Solving Least Squares Problems (Englewood Cliffs, NJ: Prentice-Hall, 1974) [Translated into Russian (Moscow: Nauka, 1986)]
101. Press W H et al. Numerical Recipes (London: Cambridge University Press, 1986)
102. Loeve M Probability Theory (New York: Van Nostrand, 1963) [Translated into Russian (Moscow: Inostr. Lit., 1962)]
103. Jain A K Fundamentals of Digital Image Processing (Englewood Cliffs, NJ: Prentice-Hall, 1989)
104. Andrews H C, Patterson C L Am. Math. Mon. 11 (1975)
105. Andrews H C, Patterson C L IEE E Trans. Comput. C 25140 (1976)
106. Huang T S, Narendra P M Appl. Opt. 142213 (1975)
107. Andrews H C, Patterson C L IEE E Trans. Acoust. Speech Signal Process. 2426 (1976)
108. Pratt W K Digital Image Processing (New York: Wiley, 1978) [Translated into Russian (Moscow: Mir, 1982)]
109. Schelkunoff S A Bell Syst. Tech. J. 2280 (1943)
110. Gorelik G S Dokl. Akad. Nauk SS SR 83549 (1952)
111. Toraldo di Francia G J. Opt. Soc. Am. 59799 (1969)
112. Rautian S G Usp. Fiz. Nauk 66475 (1958) [Sov. Phys. Usp. 1 245 (1958)]
113. Wolter H, in Progress in Optics Vol. 1 (Ed. E Wolf) (Amsterdam: North-Holland, 1961) p. 155
114. Kozlov V P Opt. Spektrosk. 16501 (1964) [Opt. Spectrosc. (US SR) 16271 (1964)]
115. Harris J L J. Opt. Soc. Am. 54606 (1964)
116. Harris J L J. Opt. Soc. Am. 54931 (1964)
117. Frieden B R J. Opt. Soc. Am. 571013 (1967)
118. Rushforth C K, Harris R W J. Opt. Soc. Am. 58539 (1968)
119. Snyder D L, Miller M I Random Point Processes in Time and Space (Berlin: Springer, 1991)
120. Terebizh V Yu, Cheburnina O K Ast ron. Astrophys. Trans. (in press)
121. Lucy L Astron. J. 1041260 (1992)
122. Kolmogorov A N, Tikhomirov B M Usp. Mat. Nauk 14 (2) 3 (1959)
123. Kotel'nikov V A Mat erials to the 1st All-Union Me eting on Te chnical Reconstruction of Communications and We ak-Current Industry Development (Moscow, 1933)
124. Snyder D L, Miller M I IEE E Trans. Nucl. Sci. 323864 (1985)
125. Snyder D L et al. IEEE Trans. Med. Imaging 6228 (1987)
126. Snyder D L, in The Restoration of HST Images and Spectra, Proceedings of Wo rkshop STSc I (Eds R L White, R J Allen) (Baltimore: NASA, 1990) p. 56
127. Terebizh V Yu Astron. Astrophys. Trans. 3251 (1993)
128. Nicolis J S Dynamics of Hierarchical Systems. An Evolutionary Approach (Berlin: Springer, 1986) [Translated into Russian (Moscow: Mir, 1989)]
129. Terebizh V Yu, Cheburnina O K, Cheburnin Yu G Astrofizika 34419 (1991) [Astrophysics 34226 (1991)]

[^0]:    V Yu Terebíh Moscow State University, P K Shternberg Astronomical Institute, 334413 Nauchnii, Crimea, Ukraine
    Tel. (7-065-54) 713 82, 710 36, Fax (7-065-54) 40704
    E-mail: terebizh@crao.crimea.ua

