Retrieval of Atmospheric Temperature and Composition From Remote Measurements of Thermal Radiation

C. D. RODGERS[1]

National Center for Atmospheric Research, Boulder, Colorado 80303

This paper reviews the methods which may be used to estimate the state of the atmosphere, i.e., the distribution of temperature and composition, from measurements of emitted thermal radiation such as are made by remote sounding instruments on satellites. The principles of estimation theory are applied to a linear version of the problem, and it is shown that many of the apparently different methods to be found in the literature are particular cases of the same general method. As an aid to understanding, the optimum linear solution is described in terms of the geometry of n dimensions, with $n = 3$ for illustration. In generalizing the approach to the nonlinear problem there are two stages: (1) finding any member of the infinite family of possible solutions, which may be done by any convenient iterative method, and (2) finding the optimum solution by satisfying appropriate constraints.

1. INTRODUCTION

Since remote sounding of the atmosphere temperature profile and composition from earth satellites was first suggested by *Kaplan* [1959], it has become a technique of major practical importance both for meteorological observing systems and for scientific observations of the atmosphere and will clearly continue to be of importance for the foreseeable future, both for the earth and for soundings of other planetary atmospheres. A large number of papers have appeared on the subject of retrieval theory, i.e., the derivation of atmospheric parameters from the measured radiation at the top of the atmosphere, proposing many seemingly different methods, and as a consequence a considerable mystique has developed around the subject. This is unfortunate because retrieval theory is essentially very simple, being hardly more than a generalization of an elementary process with which all scientists should be familiar, namely, combining independent measurements of a quantity by using the reciprocal of the square of its standard deviation as a weight.

The bibliography given below is almost certainly incomplete, but it should contain all the important papers published in the open literature on the subject of infrared remote sounding.

The purpose of this paper is to present a tutorial review of the basic methods which may be used in solving retrieval problems. No new methods are proposed, but it is hoped to provide a unified approach which will illuminate the nature of the possible solutions and aid the intuition and understanding. In its most general form the subject has a long and respectable history going back at least to Gauss. A useful modern textbook has been written by *Deutsch* [1965]. There is a large body of literature in many different fields, and the same methods have been independently described many times for different applications. However, literature in other fields is often overlooked, and when it appears in mathematical and numerical journals (which should not be overlooked), it is often couched in terms which make it difficult for the practical scientist to understand and apply.

Before indulging in wholesale retrievals of the atmospheric state from satellite measurements of radiance, it is important to decide whether this is appropriate to the problem at hand, for there may be more efficient means of achieving the goal. There are circumstances where retrieval is necessary, such as in providing data for numerical forecasting, but a lot of useful work has been done by studying the radiances as measured. For example, we can study wave motions in the stratosphere and the development of stratospheric warmings by examining the radiances or the equivalent brightness temperatures. Another similar approach is to perform wave number analyses on the temperature field by retrieving from a wave number analysis of the radiance field.

The problem that will be examined here is as follows: given a measurement or series of measurements of thermal radiation emitted by an atmosphere, the intensity and spectral distribution of which depend on the state of the atmosphere in a known way, deduce the best estimate of the state of the atmosphere. There are two distinct aspects to this problem which are not always clearly separated; they may be described as the 'inverse' problem and the 'estimation' problem. The inverse problem is the matter of inverting a known equation which expresses radiation as a function of the atmospheric state, so as to express atmospheric state in terms of radiation. This is usually an 'ill-posed' problem; i.e., it has no mathematically unique solution. We therefore have an estimation problem, that is, to find the appropriate criteria which determine the best solution from all the possible ones which are consistent with the observations. Needless to say, the estimation problem has received far less attention than it deserves in the meteorological literature.

One question that estimation theory can answer is that of the uniqueness of the solution and how to characterize it. Because there is noise in all physical measurements there is a nonuniqueness in the original observations which leads to a corresponding nonuniqueness in the solution. We must determine how the error bars of the measurement map onto the error bars of the solution. Another kind of nonuniqueness is more fundamental and is a way of saying that the problem is ill posed. There are components of the atmospheric profile which make no contribution to the quantities measured, so that in principle the size of these components could be infinite. The purpose of a priori constraints is to set bounds on them.

We will begin with a linearized version of the radiative transfer equation for which the inverse problem is trivial, so that we can concentrate on the estimation aspects. We will first describe some obvious but useless methods to highlight the difficulties and then go on to explain the principles of estima-

[1] On leave from Oxford University.

tion theory. Finally, applications to nonlinear problems will be discussed.

2. Simple Solutions

2a. *An Exact Solution to the Linear Problem*

The simplest solution to the radiative transfer equation is for the intensity of radiation $I(\nu)$ leaving the top of a nonscattering atmosphere in local thermodynamic equilibrium, when the optical depth is so great that the surface cannot be seen. In this case we may put

$$I(\nu) = \int_0^\infty B(\nu, z) \frac{dT}{dz}(\nu, z)\, dz \qquad (1)$$

where $B(\nu, z)$ is the Planck function at wave number ν and height z and $T(\nu, z)$ is the transmission from height z to space. It is convenient to use $z = -\log_e(p)$, where p is pressure, as our height coordinate, so that the unit of z is the local scale height. If we assume that transmission is independent of temperature and that we are making M measurements at a set of wave numbers $\nu_1, \nu_2, \cdots, \nu_M$, which are closely spaced, then we may put

$$I_i = I(\nu_i) = \int_0^\infty B(\bar{\nu}, z)\, K_i(z)\, dz \qquad i = 1 \cdots M \qquad (2)$$

where $\bar{\nu}$ is some representative wave number and $K_i(z) = dT(\nu_i, z)/dz$. The equation is now linear in $B(\bar{\nu}, z)$, which we may regard as the unknown. The intensity I_i to be measured is therefore a weighted mean of the Planck function profile with weighting function $K_i(z)$.

Solving equation (2) for $B(\bar{\nu}, z)$ is clearly an ill-posed or underconstrained problem, because the unknown profile is a continuous function of height and there are only a finite number of measurements. An obvious approach is to express $B(\bar{\nu}, z)$ as a linear function of M variables:

$$B(\bar{\nu}, z) = \sum_{j=1}^{M} b_j W_j(z) \qquad (3)$$

so that equation (2) becomes

$$I_i = \sum_{j=1}^{M} b_j \int_0^\infty W_j(z)\, K_i(z)\, dz = \sum_{1}^{M} A_{ij} b_j \qquad (4)$$

thus defining the known square matrix A. (We will denote matrices by uppercase boldface letters and column vectors by lowercase boldface letters. A matrix transpose will be denoted by superscript T.) We obtain M equations for the M unknowns b_j, which can in principle be solved exactly. This approach was explored by *Wark* [1960] and *Yamamoto* [1961] in two of the earliest papers on the retrieval problem of remote sounding, using linear interpolation between fixed levels and polynomials for the 'representation functions' $W_j(z)$. Unfortunately, this simple method has a major drawback in that for practical situations it is usually ill conditioned, so that any experimental error in the measurements is greatly amplified, making the solution virtually useless. This ill conditioning has been recognized for a long time in the mathematical and numerical literature as being a general feature of equations of this type, Fredholm integral equations of the first kind.

The solution obtained by solving equation (4) and substituting back in equation (3) may be written

$$B(\bar{\nu}, z) = \sum_{ij} W_j(z)\, A_{ji}^{-1} I_i = \sum_i D_i(z) I_i \qquad (5)$$

where A_{ji}^{-1} is the jith component of the inverse matrix $\mathbf{A}^{-1}$. Equation (5) defines the 'contribution function' $D_i(z)$ as the contribution to the solution profile due to the measured intensity I_i. For an exact solution, i.e., one that gives the same theoretical radiances as the original profile, the contribution function obviously must satisfy

$$\int_0^\infty D_i(z)\, K_j(z)\, dz = \delta_{ij} \qquad (6)$$

the Kroenecker delta. If there is an error ϵ_i in the measurement of I_i, then there will be an error $\epsilon_i D_i(z)$ in the solution, and ill conditioning is characterized by large values of $D_i(z)$. To illustrate this, we have constructed a synthetic set of weighting functions, as shown in Figure 1, and have calculated the contribution functions $D_i(z)$ for the exact solution where the representation $W_j(z)$ is a set of polynomials in z. The sensitivity to noise is shown by the large values of $D_i(z)$ in Figure 2. They are especially large outside the range of heights covered by the weighting functions and cannot be shown on the same scale.

We have chosen the representation quite arbitrarily so far. It is pertinent to ask whether there is a choice for which the noise sensitivity is a minimum, because this might provide a logical choice for an exact solution. The answer is of course yes, and it may be found by jointly minimizing every element of $D_i^2(z)$ subject to the constraint expressed by (6). The best value of $D_i(z)$ is easily found by using the calculus of variations to be

$$D_i(z) = \sum_j \left[\int_0^\infty K_j(z)\, K_i(z)\, dz \right]^{-1}_{ij} K_j(z) \qquad (7)$$

where the inverse is a matrix inverse. This solution may also be obtained by using the weighting functions, or any nonsingular linear combination of them, as a representation of the atmospheric profile. (By minimizing $D_i^2(z)$ in the derivation of (7) we have made the implicit assumption that the experimental error in each of the components of the measurement has the same variance and is independent of the others. However, the result may be generalized to an arbitrary error covariance matrix.) *Mateer* [1965] has studied this approach for the problem of Umkehr sounding of ozone, but it does not appear to have been used in the infrared remote sounding literature, although *Smith* [1970] has used weighting functions as a representation without commenting on the implications.

In Figure 3 we show how this choice has improved the contribution functions in comparison with the polynomials used for Figure 2. The extremely large values of $D_i(z)$ are no longer present, although $D_i(z)$ has increased in the middle range as a consequence of the constraint (equation (6)) that the solution be exact. The absolute value of the sensitivity to noise is still unacceptably large, however, and an error of ϵ in a measurement will lead to errors of up to 15ϵ in the solution. This noise sensitivity is a feature of all exact solutions to the inverse problem, whether linear or nonlinear, when there is significant overlap between the weighting functions. Qualitatively, it is difficult for the solution to follow noise in the observations without inducing oscillations, because if the solution has to be increased in the region of the peak of a weighting function in order to represent a positive error in the corresponding component of the measurements, there must be compensating decreases on either side so that components due to the neighboring weighting functions are unaffected. This means that the original increase must be larger than it otherwise would be and very much larger if the weighting functions are closely spaced. The compensating decreases will also in-

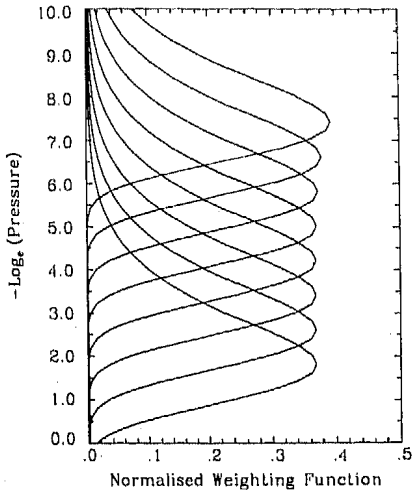

Fig. 1. A set of idealized weighting functions normalized to unit area.

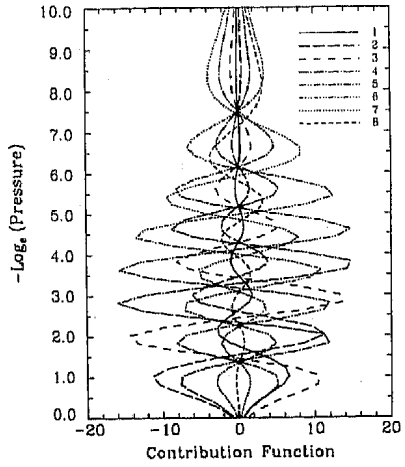

Fig. 3. The contribution functions for an exact solution in terms of the weighting functions.

duce compensating increases in the next component, and so on.

There is clearly no logical reason for us to try to find an exact solution. We can only reasonably require that the solution lie within experimental error of the measurements. This gives us more freedom in the choice of a solution, but it does raise the important question of what criteria we should use in this choice.

2b. *Least Squares Solutions*

When one is faced with an ill-conditioned problem, a common approach is to use a least squares solution, that is, to solve overconstrained equations so that the sum of the squares of the differences between the measurements and the values corresponding to the solution is minimized. In our case this may be done by representing the atmospheric profile by fewer functions than the number of components of the measurement

$$B(\bar{\nu}, z) = \sum_{i=1}^{N} b_i W_i(z) \qquad N < M \tag{8}$$

so that on substituting this into equation (2) the resulting set of equations for b_j are overconstrained:

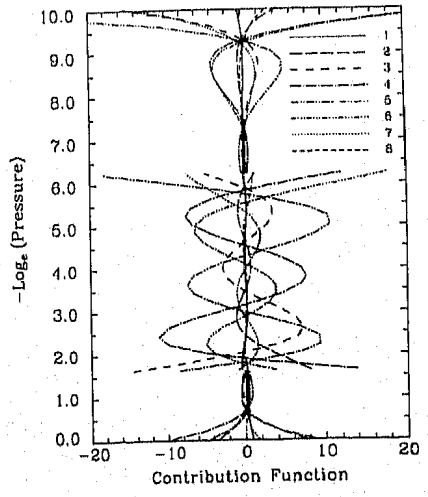

Fig. 2. The contribution functions for an exact solution in terms of polynomials in z. The scale on the abscissa corresponds to the middle section of the graph. The scale for the outer sections is −1000 to +1000.

$$I_i = \sum_{j=1}^{N} A_{ij} b_j \qquad i = 1, 2, \cdots, M \tag{9}$$

and must be solved by least squares. The solution $\hat{\mathrm{b}}$ for which $\sum_i (I_i + \sum_j A_{ij}\hat{b}_j)^2$ is a minimum is given by

$$\hat{\mathrm{b}} = (\mathrm{A}^T\mathrm{A})^{-1}\mathrm{A}^T\mathrm{I} \tag{10}$$

This solution is an improvement over an exact solution in that its noise sensitivity is smaller, but there are still no reasonable criteria to determine such things as how many terms to use and what form of representation is suitable. The least squares method is really only appropriate if the algebraic form of the solution is known from sound physical reasoning and the number of unknown parameters is considerably smaller than the number of independent measurements. It is not suitable for the atmospheric retrieval problem because the algebraic form is not known, and even if it were, there would be too many unknown parameters. In fact, the use of representations with a relatively small number of parameters is a way of pretending that we know the algebraic form. The atmospheric temperature profile is not a polynomial in height or a linear combination of weighting functions, and such representations have little hope of providing a good solution.

The only reasonable requirement that we can put on the squared deviation is that

$$\sum_i (I_i - \sum_j A_{ij}\hat{b}_i)^2 \simeq M\sigma^2 \tag{11}$$

where σ^2 is the noise variance in the measurement. We should not require that it be a minimum, but we may require that it be approximately equal to the value expected from experimental error. It is possible for $M\sigma^2$ to be larger (or smaller) than the squared deviation found by a least squares process, although in the limit of a small number of unknowns and a large number of measurements the minimum value tends to $M\sigma^2$.

3. THE GENERAL LINEAR SOLUTION

3a. *Uniqueness and the Nature of a Priori Constraints*

As is mentioned in the introduction, there are components in the atmospheric profile which do not contribute anything to the measurements. For example, in the linear case we may add to the atmospheric profile any function which is orthogonal to all the weighting functions, without changing the radiances. Such components are unmeasurable, and information about

them must come from other sources, namely, our a priori constraints. This is true whether or not there is noise in the measurements. If there is noise, the situation may be worse, because components which would otherwise be measurable may be rendered unmeasurable.

'A priori constraint' is to be taken to refer to any constraint on the solution, not just statistical constraints, as is assumed by some authors. If no constraints are available, or if the available constraints are insufficient, then we cannot solve the problem as stated; i.e., we cannot find a solution profile whose error bounds are finite. All is not lost, however, because there is still information in the measurements which can be used. We must simply change the problem to one that we can solve. One approach in this case is that of *Backus and Gilbert* [1970], which will be discussed in section 7.

There are several papers on retrieval theory in which the authors claim to have found a method of solution which involves no a priori constraints. This can obviously never be true because there is not enough information to go around, and careful reading of the papers concerned will usually reveal the hidden constraint. For example, *Smith* [1970] has required that the solution be a linear combination of the weighting functions, and *Chahine* [1968, 1970] has used linear interpolation between temperatures at fixed altitudes.

It is a very common misconception that there are two types of solutions, namely, statistical and nonstatistical, and that statistical solutions require a priori information, whereas nonstatistical solutions do not. This is a gross oversimplification and can probably be construed as being untrue. The most enlightening approach to a priori constraints, whatever their source, is to regard them as 'virtual measurements,' that is, as being of the same nature as the direct measurements. They say something about the unknown profile just as the measurements do, and together with the measurements they determine whether we finally have a well-posed problem. This is the same philosophy as that employed by *Franklin* [1970] in a context of more general applicability.

We may define a linear constraint as one which takes the same mathematical form as a linear direct measurement; i.e., it gives a value for a known linear function of the unknown profile, together with an error covariance matrix for this value. If both the constraints and the direct measurements are linear, then we have a linear problem. If either or both are nonlinear, we have a nonlinear problem.

Great care must be exercised in selecting the constraints to be applied, so that they represent the best possible a priori information available. It is possible for a constraint to be too tight, too loose, or just plain wrong. A good example of the latter is using inappropriate statistics or an unsuitable representation, and an example of the other two possibilities is the use of a wrong size of smoothing parameter for the Twomey-Tikhonov solution (section 6*b*).

Many of the constraints that have been used in the literature fall into the class of linear constraints. Some of them are described below to help the reader recognize linearity.

Statistics. The climatological mean profile and its covariance may be regarded as a measurement and its uncertainty.

Forecast profile. The forecast profile and its error covariance may similarly be regarded as a measurement and its uncertainty. However, the error covariance is often not known very well, and so it must be overestimated.

Twomey-Tikhonov. The virtual measurement is zero or an a priori guessed profile. The error covariance is proportional to Twomey's constraint matrix H, often taken to be a unit matrix.

Linear representations. It is less obvious that a linear representation with a finite number of terms is a linear constraint as defined above. We may extend the representation to be a complete one with an infinite number of terms by including any infinite set of orthogonal functions which are all orthogonal to the original representation. The coefficients of the extended representation are then known linear functions of the unknown profile. The linear constraint states that the values and variances of the extra terms are all zero, while the values of the coefficients of the original representation are zero with infinite variance. Alternatively, they may be assigned some known value and covariance matrix.

Discretization. It is usual to express continuous functions such as the unknown profile and the weighting functions in a discrete form with a relatively large number of levels (typically 50–200) in order to simplify the numerical calculations and use the algebra of vectors and matrices instead of Hilbert space. This is of course a linear representation and may be regarded as a linear constraint exactly as above. The discretization must have high enough spatial resolution so that the implied assumptions are valid.

The general approach to the linear problem should now be clear. We have a set of direct measurements which alone lead to an ill-posed problem. We must supplement them with enough virtual measurements to make the problem well posed. This may be done in two stages, discretization and a priori constraints, although the distinction is not essential. Each measurement, direct or virtual, is in the form of an estimate of a linear function of the profile and a covariance. They are combined in the usual way of combining measurements, i.e., by weighting independent estimates with inverse covariances.

In the following sections we will normally use a discrete version of the problem for the reasons described above. Equation (2) will be replaced by

$$\mathbf{y} = \mathbf{K}\mathbf{x} \tag{12}$$

where $\mathbf{y}$ is a vector of quantities to be measured and $\mathbf{x}$ is the unknown profile. The matrix $\mathbf{K}$ is a discrete version of the weighting functions, and the integral becomes a matrix multiplication. The change from I to $\mathbf{y}$ and from B to $\mathbf{x}$ has been made for consistency with literature in other fields and is in accord with usual mathematical practice. We will not normally distinguish between $\mathbf{y}$ and a measurement of $\mathbf{y}$ whose value is $\mathbf{y} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is experimental error. It should be clear from context which is intended.

3*b*. *Combination of Observations*

The reader should be familiar with the usual way of combining two independent measurements x_1 and x_2 of a scalar quantity x by taking a weighted average, with the reciprocal of the square of the standard deviations as weights:

$$\hat{x} = (1/\sigma_1^2 + 1/\sigma_2^2)^{-1}(x_1/\sigma_1^2 + x_2/\sigma_2^2) \tag{13}$$

where $\hat{x}$ denotes the combined estimate of x. The variance of $\hat{x}$ is

$$\hat{\sigma}^2 = (1/\sigma_1^2 + 1/\sigma_2^2)^{-1} \tag{14}$$

This generalizes to vectors in a straightforward manner. If we have two measurements of the vector $\mathbf{x}$, namely, $\mathbf{x}_1$ and $\mathbf{x}_2$, with error covariances $\mathbf{S}_1$ and $\mathbf{S}_2$, then the estimate $\hat{\mathbf{x}}$ is

$$\hat{\mathbf{x}} = (\mathbf{S}_1^{-1} + \mathbf{S}_2^{-1})^{-1}(\mathbf{S}_1^{-1}\mathbf{x}_1 + \mathbf{S}_2^{-1}\mathbf{x}_2) \tag{15}$$

with covariance

$$\hat{\mathbf{S}} = (\mathbf{S}_1^{-1} + \mathbf{S}_2^{-1})^{-1} \tag{16}$$

The notation $\mathbf{S}$ is used rather than Σ for covariance matrices to avoid confusion with summation. In the case of the retrieval problem we often have a virtual measurement $\mathbf{x}_0$ with covariance $\mathbf{S}_x$ and a direct measurement of $\mathbf{y} = \mathbf{K}\mathbf{x}$ with error covariance $\mathbf{S}_\epsilon$. In this case the estimate is

$$\hat{\mathbf{x}} = (\mathbf{S}_x^{-1} + \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}(\mathbf{S}_x^{-1}\mathbf{x}_0 + \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{y}) \tag{17}$$

with covariance

$$\hat{\mathbf{S}} = (\mathbf{S}_x^{-1} + \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1} \tag{18}$$

These two equations are easily understood by noting that $\mathbf{D}\mathbf{y}$ is an estimate of $\mathbf{x}$ with inverse covariance $\mathbf{S}_D^{-1} = \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K}$, where $\mathbf{D}$ is any exact solution (e.g., (7)) such that $\mathbf{K}\mathbf{D} = \mathbf{1}$, the unit matrix. In these terms, (17) becomes

$$\hat{\mathbf{x}} = (\mathbf{S}_x^{-1} + \mathbf{S}_D^{-1})^{-1}(\mathbf{S}_x^{-1}\mathbf{x}_0 + \mathbf{S}_D^{-1}\mathbf{D}\mathbf{y}) \tag{19}$$

which is of the same form as (15). Note that $\mathbf{S}_D^{-1}$ is singular, so that the components of $\mathbf{S}_D$ are infinite.

Equations (17) and (18) for the best estimate of the profile and its covariance are useful for understanding the solution, but they are not useful for practical computation because they require the inversion of relatively large matrices, or equivalently, the solution of large systems of linear equations. By making use of the matrix identity

$$\mathbf{K}^T(\mathbf{1} + \mathbf{S}_\epsilon^{-1}\mathbf{K}\mathbf{S}_x\mathbf{K}^T) = \mathbf{K}^T + \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K}\mathbf{S}_x\mathbf{K}^T = (\mathbf{1} + \mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K}\mathbf{S}_x)\mathbf{K}^T \tag{20}$$

we can put (17) into the form

$$\hat{\mathbf{x}} = \mathbf{x}_0 + \mathbf{S}_x\mathbf{K}^T(\mathbf{K}\mathbf{S}_x\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}_0) \tag{21}$$

where there is only one relatively small matrix to be inverted. Also by using (20) the expression (18) for the covariance of $\mathbf{x}$ can be reduced to a computationally simpler form, although it is algebraically more complex:

$$\hat{\mathbf{S}} = \mathbf{S}_x - \mathbf{S}_x\mathbf{K}^T(\mathbf{K}\mathbf{S}_x\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}\mathbf{K}\mathbf{S}_x \tag{22}$$

Figure 4 gives an example of the contribution functions calculated according to (21), where $\mathbf{S}_\epsilon$ has been taken to be 0.11 and $\mathbf{S}_x$ is a covariance matrix computed from a sample of mid- and high-latitude rocket soundings. Diagonal elements of $\mathbf{S}_x$ are of the order of magnitude of 100 mW m^{-2} sr^{-1} (cm^{-1})$^{-1}$. It can be seen that the noise amplification problem is considerably reduced in comparison with Figures 2 and 3.

Equations equivalent to (21) have been derived or applied by *Westwater and Strand* [1968], *Turchin and Nozik* [1969], *Rodgers* [1970, 1971], and *DeLuisi and Mateer* [1971].

A derivation of these equations in a more general form should give some insight into the nature of the solution. The two basic ways of combining noisy measurements are to find the most likely value and to find the expected value. In general, they will lead to different solutions (the 'mode' or the 'mean'), but if the error statistics are Gaussian, the maximum likelihood and expected value combinations are identical.

The multivariate Gaussian distribution may be written as

$$P(\mathbf{x}) = [(2\pi)^{n/2}\det(\mathbf{S})^{1/2}]^{-1} \cdot \exp[-\tfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T\mathbf{S}^{-1}(\mathbf{x} - \mathbf{x}_0)] \tag{23}$$

where $P(\mathbf{x})$ is the probability density function of the random vector $\mathbf{x}$ whose dimension is n, $\mathbf{x}_0$ is the expected value of $\mathbf{x}$, $E\{\mathbf{x}\}$, and $\mathbf{S}$ is the $n \times n$ covariance matrix, i.e.,

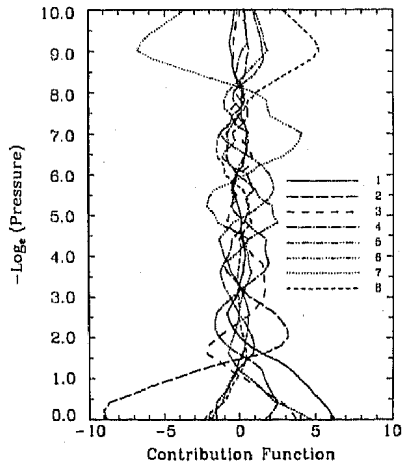

Fig. 4. The contribution functions for the linear solution with a statistical constraint.

$$\mathbf{S} = E\{(\mathbf{x} - \mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^T\} \tag{24}$$

Let us assume that we have l independent measurements $\mathbf{y}_i$, $i = 1 \cdots l$, of linear functions of the unknown profile $\mathbf{x}$:

$$\mathbf{y}_i = \mathbf{K}_i\mathbf{x} \tag{25}$$

each with error covariance $\mathbf{S}_i$. (So far we have considered the case where $l = 2$.) The most likely value of $\mathbf{x}$, given the set of data $\mathbf{y}_i$, is that which maximizes the conditional probability density function $P(\mathbf{x}|\mathbf{y}_1, \mathbf{y}_2 \cdots \mathbf{y}_l)$. By use of Bayes theorem we may write this as

$$P(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_l|\mathbf{x})P(\mathbf{x})/P(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_l) \tag{26}$$

If the measurements are independent, we may further expand this as

$$P(\mathbf{y}_1|\mathbf{x})P(\mathbf{y}_2|\mathbf{x}) \cdots P(\mathbf{y}_l|\mathbf{x})P(\mathbf{x})/P(\mathbf{y}_1) \cdot P(\mathbf{y}_2) \cdots P(\mathbf{y}_l) = P(\mathbf{x}) \prod_i P(\mathbf{y}_i|\mathbf{x})/P(\mathbf{y}_i) \tag{27}$$

If we know nothing about $\mathbf{x}$ other than the values of $\mathbf{y}_i$, then $P(\mathbf{x})$ is a constant. To find $\hat{\mathbf{x}}$, the most likely value of $\mathbf{x}$, we maximize with respect to $\mathbf{x}$, so that the terms $P(\mathbf{y}_i)$ are also constants. Thus we maximize $\Pi_i P(\mathbf{y}_i|\mathbf{x})$ or minimize minus its logarithm:

$$\frac{\partial}{\partial\hat{\mathbf{x}}}\sum_i(\mathbf{y}_i - \mathbf{K}_i\hat{\mathbf{x}})^T\mathbf{S}_i^{-1}(\mathbf{y}_i - \mathbf{K}_i\hat{\mathbf{x}}) = 0 \tag{28}$$

This may easily be solved to give a generalization of (17):

$$\hat{\mathbf{x}} = (\sum_i \mathbf{K}_i^T\mathbf{S}_i^{-1}\mathbf{K}_i)^{-1}(\sum_i \mathbf{K}_i^T\mathbf{S}_i^{-1}\mathbf{y}_i) \tag{29}$$

The covariance of the solution is

$$\hat{\mathbf{S}} = (\sum_i \mathbf{K}_i^T\mathbf{S}_i^{-1}\mathbf{K}_i)^{-1} \tag{30}$$

The solution only exists if the matrix $\sum_i \mathbf{K}_i^T\mathbf{S}_i^{-1}\mathbf{K}_i^T$, which is the inverse covariance matrix $\hat{\mathbf{S}}$, is nonsingular. This will always be the case if one of the measurements is of $\mathbf{x}$ directly, i.e., if one of the $\mathbf{K}_i$ is a unit matrix, although this is not a necessary condition. An a priori constraint on $\mathbf{x}$ ensures this.

It is left as an exercise to the reader to show that the conditional expectation

$$\hat{\mathbf{x}} = E\{\mathbf{x}|\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_l\} \tag{31}$$

leads to the same solution.

3c. *The Accuracy of Linear Solutions*

Uncertainty in the solution is a consequence of uncertainty in the measurements and is given formally by the solution covariance, (18), (22), or (30). To understand this equation, we must know how to interpret it in terms of possible deviations from the true value of the unknown profile. The diagonal of the covariance matrix contains the variances, or squares of standard deviations, of the individual components of the profile $\mathbf{x}$ and thus gives the simplest measure of accuracy. *Westwater and Strand* [1968] and *Rodgers* [1970] have used the residual variance, i.e., $\hat{S}_{jj}$, as a measure of the accuracy of the solution and have shown that the linear solution given by (21) is the one for which residual variance is a minimum. The off-diagonal elements of $\hat{\mathbf{S}}$ are nonzero, so that there are correlations between the errors at different levels, and merely stating that we know $x_j \pm \hat{S}_{jj}^{1/2}$ is an understatement of our knowledge of $\mathbf{x}$. Another approach is to find some way of expressing the error estimate as a sum of individual components such that the components are uncorrelated with each other. This can be done by diagonalizing $\hat{\mathbf{S}}$, i.e., by finding its eigenvectors and values

$$\hat{\mathbf{S}}\mathbf{L} = \mathbf{L}\Lambda \quad \text{or} \quad \hat{\mathbf{S}}l_i = \lambda_i l_i \tag{32}$$

where the eigenvectors l_i of $\hat{\mathbf{S}}$ form the columns of the matrix $\mathbf{L}$ and the eigenvalues λ_i form the diagonal elements of the diagonal matrix Λ. The eigenvectors may be normalized so that

$$\mathbf{L}^T\mathbf{L} = \mathbf{L}\mathbf{L}^T = 1 \tag{33}$$

Thus we may write

$$\mathbf{L}^T\hat{\mathbf{S}}\mathbf{L} = \Lambda \qquad \hat{\mathbf{S}} = \mathbf{L}\Lambda\mathbf{L}^T \tag{34}$$

If $\mathbf{x}$ has error covariance $\mathbf{S}$, then $\mathbf{L}^T\mathbf{x}$ has error covariance $\mathbf{L}^T\mathbf{S}\mathbf{L}$, i.e., Λ. Thus the errors in the quantities $l_i^T\mathbf{x}$ are independent, because Λ is diagonal. Now, $a_i = l_i^T\mathbf{x}$ is the coefficient of l_i in an expansion of $\mathbf{x}$ in terms of the eigenvectors; i.e.,

$$\mathbf{x} = \sum_i a_i l_i \qquad a_i = l_i^T\mathbf{x} \tag{35}$$

and the error variance in a_i is λ_i. Therefore we may say that the error in $\hat{\mathbf{x}}$ is of the form

$$\hat{\mathbf{x}} - \mathbf{x} = \sum_i \epsilon_i l_i \tag{36}$$

where the ϵ_i are independent random variables with variance λ_i.

As an illustration, we have computed the first eight eigenvectors and values for the case of the synthetic weighting functions of Figure 1 with experimental error covariance $\mathbf{S}_\epsilon = 0.11$ and the same covariance matrix $\mathbf{S}_x$ as was used for Figure 4. The eigenvalues are given in Table 1, and the eigenvectors in Figure 5.

TABLE 1. Eigenvectors of the Matrix $\hat{\mathbf{S}}$

n	λ_n	$\lambda_n^{1/2}$
1	469.1	21.7
2	274.1	16.6
3	83.9	9.2
4	47.0	6.9
5	39.5	6.3
6	35.5	6.0
7	21.2	4.6
8	18.8	4.3

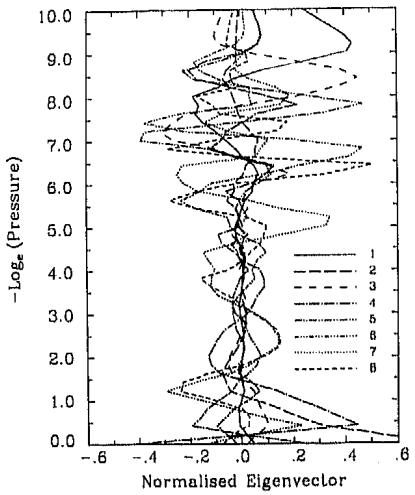

Fig. 5. Eigenvectors of the solution error covariance for the linear solution with a statistical constraint.

There is no very clear-cut relationship between the size of the eigenvalue and the shape of the eigenvector, except that the larger eigenvalues tend to be associated with uncertainties outside the range of the weighting functions. As a consequence, we would expect that the detailed shapes of the eigenvectors will depend on the range of heights under consideration.

4. A GEOMETRICAL INTERPRETATION

Now that we have described an optimum linear solution to the retrieval problem and discussed the nature of the nonuniqueness, we must study it further to gain some understanding of how it will behave in various situations and to see how it relates to other solutions that have been proposed.

Probably the best way of gaining insight into both the linear problem and its optimum solution is to interpret the algebra of matrices in terms of the geometry of finite dimensional Euclidean spaces. The heart of this approach is to find a 'basis,' or coordinate system, in which the problem is most simply stated. This was done in a purely algebraic manner in the previous section with regard to the solution error. To interpret that result geometrically, we may say qualitatively that the solution is contained in a hyperellipsoidal region centered on $\hat{\mathbf{x}}$. (This is the multivariate equivalent of error bars.) The covariance $\hat{\mathbf{S}}$ describes its orientation and size, because the quadratic equation

$$(\hat{\mathbf{x}} - \mathbf{x})^T\hat{\mathbf{S}}^{-1}(\hat{\mathbf{x}} - \mathbf{x}) = \text{const}$$

is a hypersurface of constant probability density of the solution. Transforming to the eigenvectors of $\hat{\mathbf{S}}$ is equivalent to finding its principal axes, whose lengths are the square roots of the corresponding eigenvalues.

There are two 'spaces' involved in the retrieval problem, which we will call measurement space and profile space. Measurement space is of M dimensions, and the radiances in the M spectral intervals form its original coordinate system. Profile space is of N dimensions, where N is usually much larger than M. Its coordinate system is formed by the values of the atmospheric profile at the N levels of the discretization. If no discretization is used, N is infinite, and we have a more general Hilbert space rather than a finite dimensional Euclidean space. The effect of the measurement is to 'map' profile space into measurement space, and the retrieval problem is to map measurement space back into profile space.

The weighting functions may be represented as vectors in profile space, and the radiances are then vector products of the profile vector with the weighting functions. It is obvious that the weighting functions cannot give any information about the profile outside the subspace 'spanned' by them. It would therefore be useful to find a coordinate system for this subspace.

The a priori constraints may be regarded as virtual measurements, and they define a region of profile space within which the solution must lie, according to our a priori knowledge. The virtual measurements will normally span the whole of profile space, but their covariance matrix will define a hyperellipsoidal region containing the solution. Figure 6 is an attempt to illustrate this in three dimensions, where the profile is defined by three numbers, and the measurement is of two different linear combinations of these numbers plus noise. The large ellipsoid represents the region of profile space determined by the a priori constraints. The axis of the infinite cylinder is the set of profiles which would all give the same measurements. Both of the weighting functions are perpendicular to this axis, and the subspace spanned by the weighting functions is any plane perpendicular to the axis, containing both weighting functions. The finite width of the cylinder represents experimental error. The region of intersection of the cylinder and the ellipsoid contains the solution. It is not a truncated cylinder, but an ellipsoid, because the surfaces indicated are really contours of a probability density function and not sharp boundaries. The small ellipsoid is the error ellipsoid, and its center is the optimum estimate $\hat{\mathbf{x}}$. Note that information about the position of $\hat{\mathbf{x}}$ along the length of the cylinder comes entirely from the a priori information and that $\hat{\mathbf{x}}$ does not lie on the axis of the cylinder.

4a. *The Weighting Function Subspace*

To identify the part of profile space which is spanned by the weighting functions, we must simply find a basis for this subspace. That is, we must find an orthonormal set of linear combinations of weighting functions. These basis vectors r will therefore be columns of the matrix **R** where

$$\mathbf{R} = \mathbf{K}^T\mathbf{L} \quad \text{or} \quad \mathbf{r} = \sum_i \mathbf{k}_i l_i \tag{37}$$

the nonsingular matrix L is the linear combination sought, and the vectors $\mathbf{k}_i$ are the individual weighting functions. For **R** to be orthonormal we must have

$$\mathbf{R}^T\mathbf{R} = \mathbf{L}^T\mathbf{K}\mathbf{K}^T\mathbf{L} = 1 \tag{38}$$

There is of course an infinite choice of basis vectors for a subspace, because we may rotate any coordinate system arbitrarily. We therefore need a criterion with which to choose a particular set. We will be interested in the accuracy with which we can measure the components of the profile in the directions of this new coordinate system, and so it is reasonable to choose it so that measurement errors are uncorrelated in the new system, i.e., so that

$$\mathbf{L}^T\mathbf{S}_\epsilon\mathbf{L} = \mathbf{\Lambda}' \tag{39}$$

where $\mathbf{\Lambda}'$ is diagonal. It is possible to find a transformation of this type (a similarity transform) which simultaneously diagonalizes the two positive definite matrices $\mathbf{S}_\epsilon$ and $\mathbf{K}\mathbf{K}^T$, but we will choose the easy way out and assume that $\mathbf{S}_\epsilon$ is already diagonal and is of the form $\sigma^2 1$. This requirement therefore becomes

$$\mathbf{L}^T\mathbf{L} = \mathbf{\Lambda} \tag{40}$$

where $\mathbf{\Lambda} = \mathbf{\Lambda}'/\sigma^2$. If we now premultiply and postmultiply both sides of (38) and (40) by $\mathbf{\Lambda}^{-1/2}$, we obtain

$$\mathbf{\Lambda}^{-1/2}\mathbf{L}^T\mathbf{K}\mathbf{K}^T\mathbf{L}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1}$$

$$(\mathbf{L}\mathbf{\Lambda}^{-1/2})^T(\mathbf{L}\mathbf{\Lambda}^{-1/2}) = 1 \tag{41}$$

so that $\mathbf{L}\mathbf{\Lambda}^{-1/2}$ is the matrix of eigenvectors of $\mathbf{K}\mathbf{K}^T$, with eigenvalues $\mathbf{\Lambda}^{-1}$. If we square both sides of the first of these two equations and use the second to do some simplification, we find

$$\mathbf{\Lambda}^{-1/2}\mathbf{L}^T\mathbf{K}\mathbf{K}^T\mathbf{K}\mathbf{K}^T\mathbf{L}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-2} \tag{42}$$

Remembering $\mathbf{R} = \mathbf{K}^T\mathbf{L}$ and canceling $\mathbf{\Lambda}^{-1/2}$, we find

$$\mathbf{R}^T\mathbf{K}^T\mathbf{K}\mathbf{R} = \mathbf{\Lambda}^{-1} \tag{43}$$

so that the required basis vectors are eigenvectors of $\mathbf{K}^T\mathbf{K}$ (actually the M eigenvectors with nonzero eigenvalues), with eigenvalues $\mathbf{\Lambda}^{-1}$. The measurement error variances in this representation are the diagonal elements of $\mathbf{\Lambda}'$, i.e., $\mathbf{\Lambda}\sigma^2$. The meaning of this analysis is illustrated in Figure 7 and Table 2, where we have used the weighting functions of Figure 1 to calculate **R** and $\mathbf{\Lambda}$. We can measure the component of the profile in the direction of the vector $\mathbf{r}_i$ with noise $\lambda_i^{1/2}\sigma^2$. The vectors with fine structure have large values of $\lambda_i^{1/2}$, so that we cannot measure these components very well, while the vectors with only large scale structure have small values of $\lambda_i^{1/2}$, so that we can measure them well. The values of $\lambda_n^{1/2}$ are all larger than unity in this case, which would seem to imply that

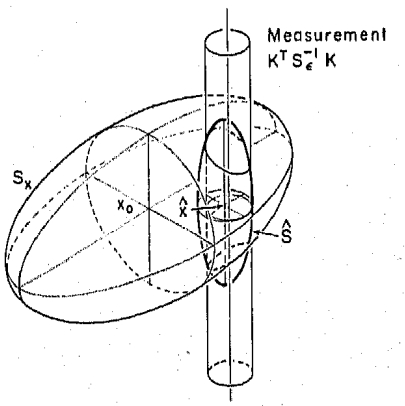

Fig. 6. An illustration of the relationship between the a priori covariance, the measurement covariance mapped into profile space, and the solution covariance.

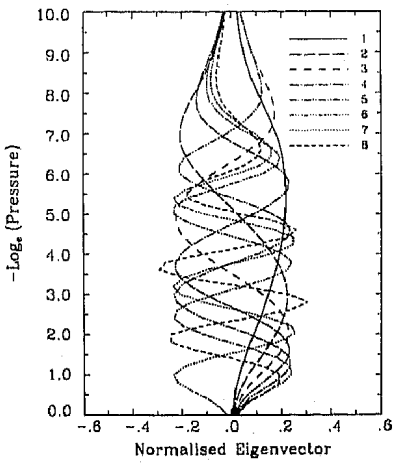

Fig. 7. Eigenvectors of $\mathbf{K}^T\mathbf{K}$, forming a basis for that part of profile space spanned by the weighting functions.

TABLE 2. Eigenvectors of the Matrices $\mathbf{K}^T\mathbf{K}$ and $\mathbf{K}\mathbf{K}^T$

n	λ_n^{-1}	$\lambda_n^{1/2}$	$0.22\lambda_n^{1/2}$
1	0.2081	2.2	0.49
2	0.1203	2.9	0.65
3	0.0536	4.3	0.97
4	0.0204	7.0	1.57
5	0.0069	12.0	2.7
6	0.0021	21.6	4.8
7	0.0006	41.0	9.2
8	0.0001	83.0	18.6

even the first component is known less accurately than the original radiances. This is not the case, however, because the original weighting functions were normalized so that $\sum k_z = 1$, not $\sum k_z^2 = 1$, as with the eigenvectors. The 'length' of a weighting function vector is typically 0.22, so that a comparative measure of accuracy is $0.22\lambda_n^{1/2}$.

The analysis given above can be taken further, but it is not immediately useful to the line of argument being developed. We will therefore leave it for the reader's amusement and recommend the Lanczos inverse [*Lanczos*, 1961, section 3.10] as a good starting point.

4b. *The Region of the a Priori Constraints*

So far, we have considered two useful sets of coordinates in profile space, that for the subspace spanned by the weight functions and that in which the solution error covariance is diagonal. A third useful coordinate system is that in which the a priori constraints are diagonal, e.g., the principal axis of the large ellipsoid in Figure 6.

If the a priori covariance is of the form $\gamma 1$, as in some applications of the Twomey-Tikhonov solution, then the ellipsoid is spherical, and any rotation of the original coordinates will be suitable. If it is a statistical, or other nondiagonal, covariance matrix $\mathbf{S}_x$, then we can diagonalize it by finding its eigenvectors

$$\mathbf{L}^T\mathbf{S}_x\mathbf{L} = \mathbf{\Lambda} \tag{44}$$

exactly as we did for the solution covariance above. The interpretation is similar. The profile can be expressed in the form

$$\mathbf{x} - \mathbf{x}_0 = \sum_i a_i l_i \qquad a_i = l_i^T(\mathbf{x} - \mathbf{x}_0) \tag{45}$$

where $\mathbf{x}_0$ is the a priori value of $\mathbf{x}$. The coefficients a_i of this representation are statistically independent random variables having variance λ_i, the eigenvalue of $\mathbf{S}_x$, and the vectors l_i are the corresponding eigenvectors of $\mathbf{S}_x$.

In the case where $\mathbf{S}_x$ is the covariance of a statistical population of profiles the eigenvectors are called variously 'principal components,' 'characteristic patterns,' and 'empirical orthogonal functions' and are a convenient way of representing members of any multivariate random population. They provide the most efficient means of approximating a random vector with fewer parameters than its dimension, in that for any given number of parameters the error is smaller than in any other representation. *Alishouse et al.* [1967] give a useful review of methods of computing empirical orthogonal functions. Smallness of an eigenvalue is interpreted as flatness of the ellipsoid in the direction of the corresponding eigenvector.

We may define 'total variance' of the population of profiles $\mathbf{x}$ as the sum of the variances of its individual components, i.e., as the sum of the diagonal elements of $\mathbf{S}_x$, or Tr $(\mathbf{S}_x)$. It can be shown that the trace of a matrix is unchanged under a similarity transform, so that the total variance is also the sum of

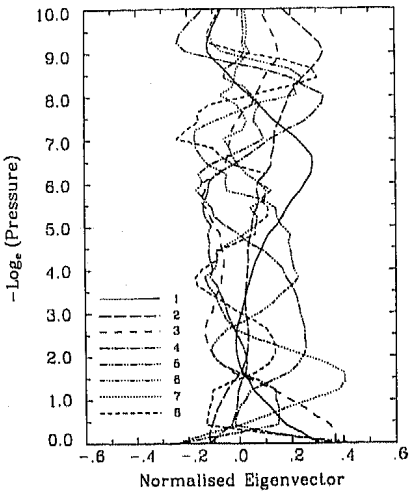

Fig. 8. Empirical orthogonal functions derived from the a priori constraint used for Figure 4. The very fine structure on some of these functions is due to the smallness of the statistical sample.

the eigenvalues. If we arrange the eigenvectors and values in order of decreasing size of eigenvalues, we can say that each vector 'explains' a certain fraction of the total variance, the first vector explaining most variance and the last vector explaining least variance. It is commonly found that the first few vectors explain most of the variance for samples of the real atmospheric profile. This is illustrated in Figure 8 and Table 3 which use the covariance matrix $\mathbf{S}_x$ that was used for Figure 4. We see that five vectors explain more than 90% of the variance in this case and that vectors which explain most variance have the largest scale structure.

4c. *Measurement Space*

The space in which the measurements lie is of smaller dimension than profile space and is related to it by

$$\mathbf{y} = \mathbf{K}\mathbf{x} \tag{46}$$

where $\mathbf{y}$ is a vector in measurement space and $\mathbf{x}$ is a vector in profile space.

Probably the most useful basis in measurement space is that consisting of the empirical orthogonal functions of a population of measurements, i.e., the eigenvectors of the covariance $\mathbf{S}_y$ of the population. We will present another derivation here that may also be applied to a population of profiles. We wish to find a transformation of the measurements so that the transformed measurements $\mathbf{z}$ are statistically independent or uncorrelated; i.e., we wish to find $\mathbf{L}$ such that

$$E\{(\mathbf{z} - \mathbf{z}_0)(\mathbf{z} - \mathbf{z}_0)^T\} = \mathbf{\Lambda} \tag{47}$$

where $\mathbf{z} = \mathbf{L}^T\mathbf{y}$ and $\mathbf{\Lambda}$ is diagonal. Therefore

$$E\{\mathbf{L}^T(\mathbf{y} - \mathbf{y}_0)(\mathbf{y} - \mathbf{y}_0)^T\mathbf{L}\} = \mathbf{\Lambda} \tag{48}$$

TABLE 3. Eigenvalues of the Constraint Covariance Matrix

n	λ_n	Percent Variance	Cumulative
1	3224	44.7	44.7
2	1905	26.4	71.1
3	808	11.2	82.3
4	488	6.8	89.1
5	185	2.6	91.6
6	126	1.8	93.4
7	116	1.6	95.0
8	75	1.0	96.0

or

$$\mathbf{L}^T\mathbf{S}_y\mathbf{L} = \mathbf{\Lambda} \tag{49}$$

We also require that experimental error $\sigma^2\mathbf{1}$ be unchanged:

$$\mathbf{L}^T \cdot \sigma^2\mathbf{1} \cdot \mathbf{L} = \sigma^2\mathbf{1} \tag{50}$$

Therefore $\mathbf{L}^T\mathbf{L} = \mathbf{1}$, and $\mathbf{L}$ must be the matrix of eigenvectors of $\mathbf{S}_y$ as expected. We can therefore represent $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{L}\mathbf{z} \quad \text{or} \quad \mathbf{y} = \sum_i z_i \mathbf{l}_i \tag{51}$$

where z_i are the coefficients of the representation and the eigenvalues λ_i are both the amount of variance explained by each vector and the variances of the coefficients z_i.

If we have a realistic covariance matrix for the profiles $\mathbf{S}_x$, then we can construct $\mathbf{S}_y$:

$$\mathbf{S}_y = \mathbf{K}\mathbf{S}_x\mathbf{K}^T \tag{52}$$

Table 4 gives the eigenvalues when $\mathbf{K}$ is the set of weighting functions in Figure 1 and $\mathbf{S}_x$ is that used for Figure 4. We can measure $\mathbf{y}$ or $\mathbf{z}$ with error variance σ^2, and the eigenvalues are the variances of $\mathbf{z}$, so that the only elements of $\mathbf{z}$ that we can really measure are those for which $\lambda_i > \sigma^2$. In this case, if our noise level is $\sigma^2 = 0.1$, then we only measure six independent quantities rather than the eight that might be expected.

If we take a measurement of $\mathbf{y}$, error covariance $\mathbf{S}_\epsilon$, and an a priori estimate $\mathbf{y}_0$, covariance $\mathbf{S}_y$, then the best estimate of the radiance is

$$\hat{\mathbf{y}} = (\mathbf{S}_y^{-1} + \mathbf{S}_\epsilon^{-1})^{-1}(\mathbf{S}_y^{-1}\mathbf{y}_0 + \mathbf{S}_\epsilon^{-1}\mathbf{y}) \tag{53}$$

$$\hat{\mathbf{y}} = \mathbf{y}_0 + \mathbf{S}_y(\mathbf{S}_\epsilon + \mathbf{S}_y)^{-1}(\mathbf{y} - \mathbf{y}_0) \tag{54}$$

using (15) and (21).

If we use the eigenvalues of $\mathbf{S}_y$ to transform these equations and assume that $\mathbf{S}_\epsilon = \sigma^2\mathbf{1}$, we obtain

$$\hat{z}_i = \left(\frac{z_{0i}}{\lambda_i} + \frac{z_i}{\sigma^2}\right)\left(\frac{1}{\lambda_i} + \frac{1}{\sigma^2}\right)^{-1}$$

$$= z_{0i} + \frac{\lambda_i}{\lambda_i + \sigma^2}(z_i - z_{0i}) \qquad i = 1 \cdots M \tag{55}$$

The z_i are independent and are estimated independently according to (13) for scalars. If $\lambda_i \ll \sigma^2$, the estimate z_i tends to the a priori value z_{0i}, and if $\lambda_i \gg \sigma^2$, the estimate tends to the measured value z_i.

5. Information Content of a Measurement

The name 'information content' has been used to cover a variety of concepts in the remote sounding literature. *Mateer* [1965] used it to describe the number of statistically independent quantities that a sounding system can measure, in the sense discussed above in section 4c. *Westwater and Strand* [1968] used 'information content' to describe the reduction in total variance due to a measurement, where total variance is the sum of the diagonal elements of a covariance matrix. *Twomey* [1970] used the name in a paper discussing the number of independently measurable quantities, in which he pointed out that increasing the number of spectral intervals does not necessarily increase the information content significantly unless there is a corresponding reduction of noise, because of the inherent resolution of the weighting functions. *Peckham* [1974] used the definition employed in information theory in a paper optimizing the configuration of a remote sounding radiometer.

All of the above are useful concepts, but we should reserve the name 'information content' for the meaning assigned to it by information theory. Qualitatively, it is reasonable to ask that the information content of a measurement should be a measure of the factor by which our uncertainty in the value of a quantity is decreased by the act of measuring it. In terms of Figure 6 a suitable definition might be the ratio of the volumes of the a priori ellipsoid and the error ellipsoid. In fact, the definition used in information theory [*Shannon and Weaver*, 1962] leads to the logarithm to base 2 of this quantity, which may be regarded roughly as the number of bits (binary digits) needed to represent the number of distinct measurements that could have been made.

Information is the change of something, the difference between before and after, and this something is called entropy. The entropy of a probability density function $P(\mathbf{x})$ is defined as

$$H(P) = -\int P(\mathbf{x}) \log_2 P(\mathbf{x})\, d\mathbf{x} \tag{56}$$

The information content of a measurement is defined as

$$H(P_1) - H(P_2) \tag{57}$$

where P_1 is the a priori probability density function of the unknown $\mathbf{x}$ and P_2 is its probability density function after the measurement has been made. It is easy to show that if $P(\mathbf{x})$ is a Gaussian distribution with covariance $\mathbf{S}$, then

$$H(P) = \tfrac{1}{2}\log_2 |\mathbf{S}| = \log_2 |\mathbf{S}|^{1/2} \tag{58}$$

The determinant of a matrix is unchanged by a similarity transform, so that $|\mathbf{S}|$ is equal to the product of the eigenvalues of $\mathbf{S}$, and the square root of the determinant is therefore equal to the product of the lengths of the axes of the ellipsoid, which is proportional to its 'volume.' Hence the information content of a measurement is in Shannon's definition equal to the logarithm of the ratio of the 'volumes of uncertainty' before and after the measurement.

For the case given in Table 4 the information content of a measurement is

$$H = \frac{1}{2}\sum \log_2\left(\frac{\lambda_n + \sigma^2}{\sigma^2}\right)$$

If $\sigma^2 = 0.1$, then $H = 20.1$ bits, implying that about 10^6 different atmospheres can be distinguished by measurement with this set of weighting functions.

6. A Handful of Linear Methods

6a. *Minimum Variance Method*

The minimum variance method is a different approach to the estimation problem that can only be applied when the a priori information is of a statistical nature. When the statistics

TABLE 4. Eigenvalues of the Covariance Matrix of the Measurements

n	λ_n	Percent Variance	Cumulative	$\lambda_n^{1/2}$
1	473.7	85.5	85.5	21.8
2	56.8	10.3	95.7	7.5
3	17.4	3.1	98.9	4.2
4	4.78	0.86	99.74	2.2
5	1.12	0.20	99.94	1.06
6	0.31	0.056	99.996	0.56
7	0.018	0.003	99.999	0.13
8	0.003	0.0005	100.0	0.05

are Gaussian, then the result is the same as the maximum likelihood estimator, as one might anticipate.

The principle behind this method is to find the linear predictor $\mathbf{D}$,

$$\hat{\mathbf{x}} = \mathbf{D}\mathbf{y} \tag{59}$$

such that the expected value of the variance of the error in the estimate is minimum; i.e., we minimize

$$E\{(\mathbf{x} - \hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}})\} \tag{60}$$

This is of course the classical problem of multiple regression, with the elements of $\mathbf{D}$ as the regression coefficients. The solution is straightforward and is left as an exercise for the reader. It results in the so-called 'normal equations,' which in this case may be written

$$\mathbf{D} = E\{\mathbf{x}\mathbf{y}^T\}(E\{\mathbf{y}\mathbf{y}^T\})^{-1} \tag{61}$$

Both the expected values in this expression are covariance matrices. They may be evaluated exactly as written, but it is more convenient to subtract means of $\mathbf{x}$ and $\mathbf{y}$ to obtain the covariance matrices that have been used in previous sections. If $\mathbf{y} = \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}$, then we can easily show that

$$E\{\mathbf{x}\mathbf{y}^T\} = \mathbf{S}_x\mathbf{K}^T \qquad E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{S}_y = \mathbf{K}\mathbf{S}_x\mathbf{K}^T + \mathbf{S}_\epsilon \tag{62}$$

Thus the minimum variance solution leads to

$$\hat{\mathbf{x}} = \mathbf{S}_x\mathbf{K}^T(\mathbf{K}\mathbf{S}_x\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}\mathbf{y} \tag{63}$$

which is identical to (21) for the maximum likelihood solution.

We may use the minimum variance solution in this form, or we may go back to (61) and use it in its multiple regression form. If we have a sufficiently large sample of cases where we have measured $\mathbf{y}$ with our remote sounder and also have independent direct measurements from radiosondes or rockets, then we can estimate $\mathbf{D}$ entirely from experiment, and we do not need to know the weighting functions. This method has been applied in practice to Sirs on Nimbus 3 [*Smith et al.*, 1970].

6b. *The Twomey-Tikhonov Method*

The first method applied to the retrieval problem in which due consideration was given to the estimation problem was that of *Twomey* [1963] and *Tikhonov* [1963]. The principle enunciated by Twomey in his first presentation of this method is to find an estimate $\hat{\mathbf{x}}$ such that the calculated measurement $\hat{\mathbf{y}} = \mathbf{K}\hat{\mathbf{x}}$ differs from the actual measurement by exactly the experimental error

$$(\mathbf{y} - \mathbf{K}\hat{\mathbf{x}})^T(\mathbf{y} - \mathbf{K}\hat{\mathbf{x}}) = M\sigma^2 \tag{64}$$

and such that $\mathbf{x}$ minimizes a given quadratic form

$$(\hat{\mathbf{x}} - \mathbf{x}_0)^T\mathbf{H}(\hat{\mathbf{x}} - \mathbf{x}_0) \tag{65}$$

where $\mathbf{x}_0$ is a 'first guess' value of $\mathbf{x}$ and $\mathbf{H}$ is chosen according to the problem at hand. Twomey suggests various possibilities including minimizing the variance of $\hat{\mathbf{x}}$ about $\mathbf{x}_0$ ($\mathbf{H} = 1$), minimizing the curvature, or second differences of $\hat{\mathbf{x}}$, etc.

When the minimization problem is formulated by using a Lagrangian multiplier,

$$\frac{\partial}{\partial\hat{\mathbf{x}}}\left[(\hat{\mathbf{x}} - \mathbf{x}_0)^T\mathbf{H}(\hat{\mathbf{x}} - \mathbf{x}_0) + \gamma(\mathbf{y} - \mathbf{K}\hat{\mathbf{x}})^T(\mathbf{y} - \mathbf{K}\hat{\mathbf{x}})\right] = 0 \tag{66}$$

we see that it is exactly the same form as the maximum likelihood solution if we interpret $\mathbf{x}_0$ and $\mathbf{H}^{-1}$ as the a priori value and its covariance and $\gamma 1$ as the inverse error covariance $\mathbf{S}_\epsilon^{-1}$.

Twomey finds that it is algebraically very difficult to determine the Lagrangian multiplier γ from his constraint, and so he proposes that it be determined empirically. The maximum likelihood method gives an interpretation which allows one to choose a reasonable value for γ.

This method was arrived at independently by Tikhonov, who called it 'regularization.' The case where $\mathbf{H}^{-1}$ is a statistical covariance matrix is often called 'statistical regularization' in the Russian literature.

The Twomey-Tikhonov solution with a unit constraint matrix $\mathbf{H} = 1$ is often referred to in the literature as the 'minimum information method.' This has nothing to do with information in Shannon's sense. It arises from the name given to a method due to *Foster* [1961] for solving a related problem. The unit matrix is the simplest estimate of a proper covariance matrix.

6c. *Truncated Orthogonal Expansion Methods*

Occasionally, methods have been proposed that involve finding an exact or least squares solution using some kind of orthogonal expansion. The commonest sets of functions are the eigenvectors of $\mathbf{K}^T\mathbf{K}$ and the empirical orthogonal functions, or eigenvectors of $\mathbf{S}_x$. For an example, see *Sellers and Yarger* [1969] and *Mateer* [1965].

Such methods are nonoptimum for the reasons discussed in section 2, although a method using empirical orthogonal functions may be nearly optimum, and it seems unnecessary to use this approach when the algebra and computation for an optimum method are so little different. However, a situation where a truncated empirical orthogonal function representation may be used to advantage is improving computational and storage efficiency on small computers. An expansion may be used instead of a high-resolution discretization to represent the profile, provided that the number of terms used is somewhat greater than the number of spectral intervals. For example, if the instrument has 10 spectral intervals, it may be found that a representation in terms of perhaps 20 empirical orthogonal functions is as good as a discretization at 100 levels.

6d. *Sequential Estimation*

The optimal estimation equations

$$\hat{\mathbf{x}} = \mathbf{x}_0 + \mathbf{S}_x\mathbf{K}^T(\mathbf{K}\mathbf{S}_x\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}_0) \tag{67}$$

$$\hat{\mathbf{S}} = \mathbf{S}_x - \mathbf{S}_x\mathbf{K}^T(\mathbf{K}\mathbf{S}_x\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}\mathbf{K}\mathbf{S}_x \tag{68}$$

are of general applicability, and there are many ways of using them.

For example, it is not necessary to perform a matrix inversion (or solve simultaneous linear equations explicitly) because we can use these equations sequentially. We may treat our measurement $\mathbf{y}$ as a set of scalars y_i, $i = 1 \cdots M$, and perform the following operations:

$$\begin{aligned} &\mathbf{S}_0 = \mathbf{S}_x \\ &\text{for} \quad i = 1 \cdots M \\ &\begin{bmatrix} \mathbf{x}_i = \mathbf{x}_{i-1} + \mathbf{S}_{i-1}\mathbf{k}_i(y_i - \mathbf{k}_i{}^T\mathbf{x}_{i-1})/(\mathbf{k}_i{}^T\mathbf{S}_{i-1}\mathbf{k}_i + \sigma^2) \\ \mathbf{S}_i = \mathbf{S}_{i-1} - \mathbf{S}_{i-1}\mathbf{k}_i\mathbf{k}_i{}^T\mathbf{S}_{i-1}/(\mathbf{k}_i{}^T\mathbf{S}_{i-1}\mathbf{k}_i + \sigma^2) \end{bmatrix} \\ &\hat{\mathbf{x}} = \mathbf{x}_M \qquad \hat{\mathbf{S}} = \mathbf{S}_M \end{aligned} \tag{69}$$

The column vector $\mathbf{k}_i$ is the ith weighting function, or the ith row of $\mathbf{K}$. We treat each element of $\mathbf{y}$ separately and at each stage make a new estimate of $\mathbf{x}$, using it as the a priori value for

the next stage. The matrix inverse becomes a scalar reciprocal, and the final value of $\hat{\mathbf{x}}$ is identical to that obtained from the original formula. In the linear case this has few advantages over the original method (the number of arithmetic operations is similar), although the computer program is shorter, and it is useful for very small machines. However, in the nonlinear problem (section 8) it can speed up convergence if the measurements are used in the right order, the more linear ones first.

The concept of sequential estimation allows one to use the continuity that exists along a subsatellite track to improve the accuracy of a retrieval. It is possible in principle to retrieve profiles from a whole orbit's data or even a whole day's data at one fell swoop, by using (17) and (18), where $\mathbf{x}$ now refers to the two- or three-dimensional distribution of temperature, but in practice it is difficult to estimate $\mathbf{S}_x$, and the size of the matrices involved is prohibitive. However, if we know something about how the profile changes in the horizontal, we can make a better a priori estimate using the retrieval previous in time and therefore close in space.

The simplest situation is to assume that the profiles follow a simple first-order stochastic process

$$\mathbf{x}_n = \alpha\mathbf{x}_{n-1} + \beta \tag{70}$$

where $\mathbf{x}_n$ is the profile at position n, α is a known matrix, and β is a random vector with zero mean and known covariance $\mathbf{S}_\beta$. A slightly better assumption might be

$$(\mathbf{x}_n - \bar{\mathbf{x}}_n) = \alpha(\mathbf{x}_{n-1} - \bar{\mathbf{x}}_{n-1}) + \beta \tag{71}$$

where $\bar{\mathbf{x}}_n$ is a climatology or forecast, i.e., the quantity that we would use for $\mathbf{x}_0$ if we were performing separate rather than sequential retrievals. We will defer discussion of the problems of determining α and $\mathbf{S}_\beta$ to a later stage.

If our retrieval at position $n-1$ has given an estimate $\hat{\mathbf{x}}_{n-1}$ with covariance $\hat{\mathbf{S}}_{n-1}$, then we may use the stochastic process to obtain an a priori estimate for $\mathbf{x}_n$,

$$\mathbf{x}_n = \bar{\mathbf{x}}_n + \alpha(\hat{\mathbf{x}}_{n-1} - \bar{\mathbf{x}}_{n-1}) + \beta \tag{72}$$

with covariance

$$\mathbf{S}_n = \alpha\hat{\mathbf{S}}_{n-1}\alpha^T + \mathbf{S}_\beta \tag{73}$$

These estimates are then combined with the observation at position n to obtain the estimate $\hat{\mathbf{x}}_n$ and its covariance $\hat{\mathbf{S}}$. This process gives a one-sided smoothing, but it can be improved by repeating the process along the orbit in the reverse direction and taking a proper weighted mean of the forward estimate $\hat{\mathbf{x}}_n$ and the backward a priori estimate $\mathbf{x}_n$. (Averaging the forward estimate and the backward estimate uses the measurement twice.)

Estimation of the matrices α and $\mathbf{S}_\beta$ is a difficult problem. In principle they should be obtainable from appropriately spaced in situ measurements, but in practice these are hard to obtain. It may be possible to construct $\mathbf{S}_\beta$ from statistics of the zonal wind in the case of sounding temperature from a polar orbiting satellite. A crude method which is nevertheless useful is to assume $\alpha = \gamma\mathbf{1}$, where γ is an empirically chosen constant close to unity, but a little smaller, in which case it is necessary to have

$$\mathbf{S}_\beta = (1 - \gamma^2)\mathbf{S}_x \tag{74}$$

for consistency with the global statistics.

7. Solvable Problems—The Backus-Gilbert Approach

If there are not enough measurements and constraints to make the problem well posed, or if those that exist do not reduce the solution error covariance $\hat{\mathbf{S}}$ sufficiently, then the original problem is not solvable, and we must look for other problems that are. This is the basic philosophy adopted by *Backus and Gilbert* [1970] in their approach to the problem of remotely sounding the structure of the solid earth with seismic waves, which was applied to the atmospheric remote sounding problem by *Conrath* [1972].

In the case of the linear problem, all we can reasonably do with our measurements is to take linear combinations of them. Every linear combination of measurements corresponds to a linear function of the unknown profile. The essence of the Backus-Gilbert method is to control the shape of this linear function so that it corresponds to some meaningful quantity.

A typical useful shape for the case of atmospheric sounding might be something like a 'boxcar' function, whereby the corresponding combination of measurements approximates mean temperatures over a layer, which is proportional to the thickness of the layer. However, most of the development of this method has been directed toward approximating a delta function, resulting in a singly peaked function with a width that may be specified by the user. The result is that we may represent the unknown profile at a range of spatial resolutions but not at infinite resolution. The highest resolution attainable depends on the set of weighting functions. The error in a profile seen at finite resolution is of a different kind from one seen at infinite resolution but with measurement error. For example, the difference between a finite resolution profile and the profile original will depend on the curvature of the original profile, while the errors in the infinite resolution profile depend only on the error covariances and not on the original profile.

Of course, experimental noise is important for the Backus-Gilbert method, and as one might expect, it is related to resolution. A high-resolution profile will have high noise, and a low-resolution profile will have low noise. One can trade off resolution against noise to obtain the best compromise for any particular application.

In order to approximate a delta function by a linear combination of weighting functions we need some parameter which measures how good the approximation is, which may be minimized for a best fit. Such a quantity is the spread, which is defined as follows: The spread $S(x)$ of a function of height $A(z)$ about a height x is

$$S(x) = 12 \int (x - z)^2 A^2(z)\, dz \tag{75}$$

The normalizing factor 12 is included so that a boxcar function centered at x will have a spread equal to its width. The function $A(z)$ must have unit area

$$\int A(z)\, dz = 1 \tag{76}$$

Backus and Gilbert tried several possible definitions of spread before choosing the one given above. Other definitions, such as the 'radius of gyration' of $A(z)$, produced significant negative excursions in the solution or had other similar disadvantages. In this section we will revert to our original notation in which quantities are treated as functions of height rather than in a discretized form as a vector.

We wish to find a solution profile $B_s(x)$ seen at finite resolution, which is given by a linear function of the measurements

$$B_s(x) = \sum_i D_i(x) I_i \tag{77}$$

If we substitute for I_i from (2), we obtain

$$B_s(x) = \int \sum_i D_i(x) K_i(z) B(z)\, dz \qquad (78)$$

showing that $B_s(x)$ is a smoothed version of $B(z)$ with a smoothing function

$$A_x(z) = \sum_i D_i(x) K_i(z) \qquad (79)$$

We wish to choose $D_i(x)$ so that $A_x(z)$ has the smallest possible spread. If we find the spread of this function by using (75), we obtain

$$S(x) = \sum_{jl} D_j(x) D_l(x) Q_{jl}(x) \qquad (80)$$

where

$$Q_{jl}(x) = 12 \int (x - z)^2 K_j(z) K_l(z)\, dz \qquad (81)$$

$$\sum_i D_i(x) = 1 \qquad (82)$$

because both $A(z)$ and $K(z)$ are normalized to unit area. We therefore minimize $S(x)$ with respect to $D_k(x)$ subject to the unit area constraint

$$\frac{\partial}{\partial D_k(x)} \left[\sum_{jl} D_j(x) D_l(x) Q_{jl}(x) + \lambda \sum_j D_j(x) u_j \right] = 0 \qquad (83)$$

where λ is an undetermined multiplier and $u_j = 1$ for mathematical convenience. This gives

$$2 \sum_j D_j(x) Q_{jk}(x) + \lambda u_k = 0 \qquad (84)$$

Therefore in matrix notation,

$$\mathbf{D}^T(x) = (-\lambda/2)\mathbf{u}^T\mathbf{Q}^{-1}(x) \qquad (85)$$

The effect of $\mathbf{u}^T$ here is to sum the rows of $\mathbf{Q}^{-1}$. To find λ, we substitute back into the unit area constraint, $\sum D_j(x) = 1$, or

$$\mathbf{D}^T\mathbf{u} = 1 \qquad (86)$$

Therefore

$$1 = (-\lambda/2)\mathbf{u}^T\mathbf{Q}^{-1}(x)\mathbf{u} \qquad (87)$$

$$\mathbf{D}^T(x) = \mathbf{u}^T\mathbf{Q}^{-1}/\mathbf{u}^T\mathbf{Q}^{-1}\mathbf{u} \qquad (88)$$

The effect of finding the best resolution without regard to noise is of course to produce considerable noise amplification. As before, the noise variances in the estimate of $B_s(x)$ must be

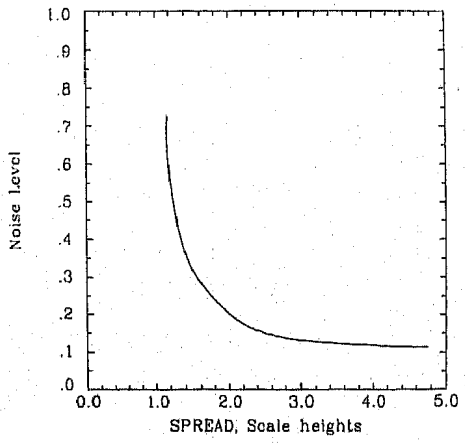

Fig. 9. Trade-off between noise and vertical resolution for the idealized set of weighting functions (Figure 1) at a height of $z = 5.0$.

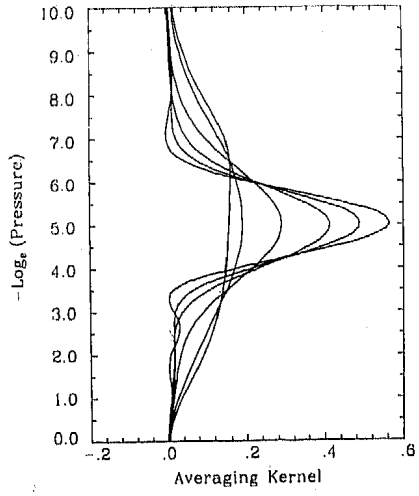

Fig. 10. Averaging kernels at $z = 0.5$ for $\mu = 0, 1, 4, 40, 400$, and ∞, $\mu = 0$ being the narrowest. These are normalized to unit area and may be directly compared with the original weighting functions in Figure 1.

$\mathbf{D}^T\mathbf{S}_\epsilon\mathbf{D}$. If we minimize a weighted sum of spread and noise variance

$$\frac{\partial}{\partial \mathbf{D}} (\mathbf{D}^T\mathbf{Q}\mathbf{D} + \lambda\mathbf{D}\mathbf{u} + \mu\mathbf{D}^T\mathbf{S}_\epsilon\mathbf{D}) = 0 \qquad (89)$$

in order to find $\mathbf{D}$, then we can control the solution noise variance, but at the expense of spread. The algebra is only changed by replacing $\mathbf{Q}$ by $\mathbf{Q} + \mu\mathbf{S}_\epsilon$, so that the solution is

$$\mathbf{D}^T = \mathbf{u}^T(\mathbf{Q} + \mu\mathbf{S}_\epsilon)^{-1}/\mathbf{u}^T(\mathbf{Q} + \mu\mathbf{S}_\epsilon)^{-1}\mathbf{u} \qquad (90)$$

We may regard μ as a 'trade-off' parameter, controlling the balance between noise and resolution. As $\mu \to 0$, we get the best resolution but poor noise. As $\mu \to \infty$, we get poor resolution but best noise. This trade-off between resolution and noise is physically very reasonable, and the value of μ must be chosen according to the application.

Figure 9 illustrates the trade-off between noise and spread that is obtained with the weighting functions of Figure 1 for a center height of five scale heights. A basic noise level of $\sigma^2 = 0.1$ or $\sigma = 0.316$ has been assumed. For $\mu = 0$ the best resolution is about 1.15 scale heights with a noise of 0.72, and for $\mu = \infty$, noise is about 0.1, and spread is 4.8 scale heights. Figure 10 shows the corresponding range of averaging kernels, the narrowest one being for $\mu = 0$ and the broadest for $\mu = \infty$.

Note that the Backus-Gilbert method can be applied to cases where there is a priori information, simply because the a priori information can be treated as virtual measurements [*Rodgers*, 1976].

8. NONLINEAR PROBLEMS

The general remote sounding retrieval problem is nonlinear. It is only by making simplifying assumptions that we can construct a linear problem. The main sources of nonlinearity in the equations are (1) temperature dependence of the atmospheric transmission, (2) wave number dependence of the Planck function across a spectral band, (3) Wave number dependence of the Planck function between spectral bands, (4) the dependence of transmission on absorber concentration in sounding for composition, (5) clouds, and (6) nonlinear constraints.

Sources 1 and 2 usually lead to relatively small nonlinearities, while the rest may cause large nonlinearities.

Formally, we may write the nonlinear problem as a general-

ization of the linear problem. We wish to know the value of an unknown vector $\mathbf{x}$. We can measure a vector $\mathbf{y}$, which is related to $\mathbf{x}$ in a known way:

$$\mathbf{y}_1 = \mathbf{F}_1(\mathbf{x}) \tag{91}$$

with an error covariance $\mathbf{S}_1$. We may have a priori information $\mathbf{y}_2$ about $\mathbf{x}$, which is similarly a known function of $\mathbf{x}$:

$$\mathbf{y}_2 = \mathbf{F}_2(\mathbf{x}) \tag{92}$$

with error covariance $\mathbf{S}_2$. The process of solution is in principle the same as for linear problems, although it may not be possible to find an explicit algebraic form for the best estimate $\hat{\mathbf{x}}$. We must first ensure that there are enough measurements and constraints to determine $\hat{\mathbf{x}}$ uniquely (i.e., with finite covariance) and then establish a set of equations which $\hat{\mathbf{x}}$ must satisfy, such as the maximum likelihood condition that $\hat{\mathbf{x}}$ minimizes

$$\sum_i [\mathbf{y}_i - \mathbf{F}_i(\mathbf{x})]^T \mathbf{S}_i^{-1} [\mathbf{y}_i - \mathbf{F}_i(\mathbf{x})] \tag{93}$$

The solution of this kind of equation is usually the most difficult part of a nonlinear estimate. When the solution has been found, we must characterize its error bounds by finding its error covariance matrix.

The degree of nonlinearity in a problem can be classified in terms of the approximations that may be used to solve it. Problems which are slightly nonlinear are best solved by some form of Newtonian iteration; that is, linearize the problems and use a linear solution method, iterating as required. A second class of moderately nonlinear problems is those which need other methods to find the solution efficiently but are sufficiently linear in the neighborhood of the solution for linearization to be used in the error analysis. A third class is composed of grossly nonlinear problems, which are the most difficult kind to solve and whose solutions are most difficult to understand. Solutions may always be found in principle simply by minimizing expression (93) by a brute force numerical method [e.g., *Barnett*, 1969], but this should be regarded as a last resort, since in any particular case it is usually possible to find some ad hoc method which exploits the algebraic form.

The correct optimum solution of nonlinear problems is hard. This is exemplified in the literature by a large variety of nonoptimum methods which use constraints designed to aid the iteration rather than to be realistic or which converge to exact (and therefore noise sensitive) solutions.

8a. *Newtonian Iteration*

The method of Newtonian iteration is simply a matter of expanding the direct model (equation (1) or (91)) as a Taylor series about a guessed value $\mathbf{x}_n$ of the solution

$$\mathbf{y} = \mathbf{F}(\mathbf{x}_n) + \partial \mathbf{F}/\partial \mathbf{x}\,(\mathbf{x} - \mathbf{x}_n) + 0(\mathbf{x} - \mathbf{x}_n)^2 \tag{94}$$

We may redefine some of the symbols in an obvious manner to give

$$\mathbf{y} = \mathbf{y}_n + \mathbf{K}_n(\mathbf{x} - \mathbf{x}_n) + 0(\mathbf{x} - \mathbf{x}_n)^2 \tag{95}$$

The higher-order terms are ignored, so that a set of underconstrained linear equations for $\mathbf{x}$ remains, which we solve in the usual way, using an a priori constraint $\mathbf{x}_0$ with covariance $\mathbf{S}_x$. We have

$$\mathbf{K}_n\mathbf{x} = \mathbf{y} - \mathbf{y}_n + \mathbf{K}_n\mathbf{x}_n \qquad \mathrm{Cov}\,(\mathbf{K}_n\mathbf{x}) = \mathbf{S}_\epsilon \tag{96}$$

$$\mathbf{x} = \mathbf{x}_0 \qquad \mathrm{Cov}\,(\mathbf{x}) = \mathbf{S}_x \tag{97}$$

Thus the solution is

$$\mathbf{x}_{n+1} = (\mathbf{S}_x^{-1} + \mathbf{K}_n^T\mathbf{S}_\epsilon^{-1}\mathbf{K}_n)^{-1}[\mathbf{K}_n^T\mathbf{S}_\epsilon^{-1}(\mathbf{y} - \mathbf{y}_n + \mathbf{K}_n\mathbf{x}_n) + \mathbf{S}_x^{-1}\mathbf{x}_0] \tag{98}$$

using (17). We have written $\mathbf{x}_{n+1}$ in place of $\hat{\mathbf{x}}$ because this is an iteration equation, and $\mathbf{x}_n \to \hat{\mathbf{x}}$ as $n \to \infty$. This may be rearranged to give $\mathbf{x}_{n+1}$ as a departure from either $\mathbf{x}_n$ or $\mathbf{x}_0$:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + (\mathbf{S}_x^{-1} + \mathbf{K}_n^{-1}\mathbf{S}_\epsilon^{-1}\mathbf{K}_n)^{-1}\{\mathbf{K}_n^T\mathbf{S}_\epsilon^{-1}(\mathbf{y} - \mathbf{y}_n) + \mathbf{S}_x^{-1}(\mathbf{x}_0 - \mathbf{x}_n)\} \tag{99}$$

$$\mathbf{x}_{n+1} = \mathbf{x}_0 + (\mathbf{S}_x^{-1} + \mathbf{K}_n^T\mathbf{S}_\epsilon^{-1}\mathbf{K}_n)^{-1}\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\{(\mathbf{y} - \mathbf{y}_n) - \mathbf{K}_n(\mathbf{x}_0 - \mathbf{x}_n)\} \tag{100}$$

However, the computationally efficient form can only be derived from the second of these:

$$\mathbf{x}_{n+1} = \mathbf{x}_0 + \mathbf{S}_x\mathbf{K}_n^T(\mathbf{K}_n\mathbf{S}_x\mathbf{K}_n^T + \mathbf{S}_\epsilon)^{-1}[\mathbf{y} - \mathbf{y}_n - \mathbf{K}_n(\mathbf{x}_0 - \mathbf{x}_n)] \tag{101}$$

This process has second-order convergence, because it gives the solution in one step in the linear case, and so we may use the value of $\mathbf{x}_{n+1} - \mathbf{x}_n$ as a convergence criterion. The covariance of the solution is of course

$$\hat{\mathbf{S}} = (\mathbf{S}_x^{-1} + \hat{\mathbf{K}}^T\mathbf{S}_\epsilon^{-1}\hat{\mathbf{K}})^{-1} \tag{102}$$

or its equivalent according to (22), where $\hat{\mathbf{K}}$ is $\mathbf{K}$ evaluated at $\hat{\mathbf{x}}$.

The main drawback of Newtonian iteration is that the 'Fréchet derivative' $\mathbf{K} = \partial\mathbf{F}/\partial\mathbf{x}$ must be evaluated at each stage, in contrast to the linear case where almost everything can be precomputed and the solution reduces to a matrix multiplication. However, in some nearly linear cases a constant can be used for $\mathbf{K}$. Convergence will be obtained, but the solution will not be optimum. The nonoptimum solution can then be used as a new starting point where a new value of $\mathbf{K}$ is computed. This approach will require more iterations but fewer evaluations of $\mathbf{K}_n$.

A trap that some authors have fallen into is to iterate the linearized equations by using the nth estimate $\mathbf{x}_n$ as the a priori information for the $(n + 1)$th stage, i.e., replacing $\mathbf{x}_0$ by $\mathbf{x}_n$ in the above equations. An example of this is given by *Smith et al.* [1972], who have used Twomey's method in this way. The effect is that the original a priori constraint is lost, and the solution will converge to the undesirable 'exact' solution if the iteration is not stopped.

Surmont and Chen [1973] have used a Newtonian iteration without apparent constraint, in such a way that it converges onto an exact solution. They found that in the absence of experimental noise a solution was obtainable for seven spectral intervals but was numerically unstable for 12 spectral intervals. The numerical instability was presumably due to the amplification of the computational truncation error by an inherently unstable method. The authors did not include experimental error.

8b. *Moderately Nonlinear Problems*

If a problem is too nonlinear for a linearization method to be efficient, then there are two courses of action open to us. We may solve the estimation equation by any of the general purpose routines to be found in computer program libraries, either minimizing expression (93) directly or solving the nonlinear equations obtained by differentiation of (93). This process certainly works but is likely to be inefficient, and there is

always the possibility of a nonunique solution. Provided the equations are sufficiently linear in the neighborhood of the solution, then the solution covariance is still given by (102).

A more practical approach is to make use of the nature of the equations to reduce the amount of work to be done, and to use some ad hoc method to find a solution. If necessary, the instrument should be designed with the solution of the retrieval problem in mind. Ad hoc solutions are usually nonoptimum, but they may lie within the population defined by the optimum solution and its covariance. If this is so, then a final stage of linearization will give an optimum solution.

There are many ad hoc approaches to be found in the literature, some of which are listed here: *Barnett* [1969], *Chahine* [1968, 1970, 1972, 1974], *Chow* [1974], *Conrath* [1969], *Gille and House* [1971], *King* [1964], *Scott and Chedin* [1971], and *Smith* [1967, 1968, 1970]. Most of these authors seem not to realize that their solution is nonoptimum and that they have found only one of a population of equally valid solutions. Consequently, the final step of linearization is not usually taken. In order to keep this article within reasonable bounds we will only describe some of the nonlinear ad hoc methods.

8c. *Chahine's Method*

Chahine's method relies on the weighting functions' having well-defined peaks, so that the radiance measured is closely related to the temperature in the region of the peak. The atmospheric temperature profile is represented by linear interpolation between temperatures at fixed levels defined by the peaks of the weighting functions. (In a variation due to *Smith* [1970] the temperature profile is represented by a linear combination of weighting functions.) At any stage of the iteration, radiances are computed for the current estimate of the temperature profile, and the profile is then adjusted by the iteration equation

$$B(\theta_j^{n+1}, \nu_j) = B(\theta_j^n, \nu_j) I_j^{obs} / I_j^{calc}(\mathbf{0}^n) \tag{103}$$

where θ_j^n is the temperature at level j (corresponding to weighting function j) at stage n of the iteration, $B(\theta, \nu)$ is the Planck function, I_j^{obs} is the observed radiance, and $I_j^{calc}(\mathbf{0})$ is the calculated radiance for the profile defined by $\mathbf{0} = (\theta_1, \theta_2, \cdots, \theta_M)$. The temperature θ_j^{n+1} is found by using the inverse Planck function. It can be seen intuitively that if the weighting functions are reasonably well peaked and separated, then the effect of this iteration is to produce a profile such that

$$\lim_{n \to \infty} I_i^{calc}(\theta^n) = I_i^{obs} \tag{104}$$

If the iteration is allowed to proceed indefinitely, then it will converge to an undesirable exact solution. However, a convergence analysis shows that broad scale features converge rapidly, and fine scale features converge slowly, so that a qualitatively 'reasonable' solution may be found if the iteration is stopped at the right point. In the case of a linear problem, Chahine's method will eventually converge to (5) for the appropriate form of the representation $W_j(z)$, and Smith's version will converge to (7). The main advantage of Chahine's method is that it is easy to understand and to program. However, the error analysis is difficult, and we do not know which of the infinite set of possible solutions will be produced. In a fairly strongly nonlinear problem it may converge faster than a linearization method, so that in this case an effective combination might be to use a Chahine solution as a first guess for an optimum linearization method.

8d. *Adjacent Fields of View*

Techniques exploiting the properties of adjacent fields of view have been described by *Smith* [1968], and some remote sounding instruments (VTPR, ITPR, HIRS) have been designed with these methods in mind, by scanning in such a way that there are many adjacent fields of view.

The main principle is that clouds have a considerable amount of spatial structure in the horizontal, in relation to other variables in the atmosphere, so that we may assume that in two sufficiently small adjacent fields of view the temperature profile is the same, and the cloud height (or distribution in height) is likely to be the same, but the cloud amount may be quite different.

In the simplest case of one layer of cloud at height c we may generalize (1) for the radiance to

$$I(\nu) = \int_0^\infty B(\nu, z) \frac{dT(\nu, z)}{dz}\, dz + n \int_0^c [B(\nu, c) - B(\nu, z)] \frac{dT(\nu, z)}{dz}\, dz \tag{105}$$

where n is the cloud amount. This may be written more simply as

$$I = I_0 + n\Delta I(c) \tag{106}$$

where I_0 is the 'clear column radiance' and $\Delta I(c)$ is the correction to I_0 for a complete cloud cover at height c.

We can make use of the form of this equation to eliminate $n\Delta I(c)$ and solve for I_0 in several ways. For example, if we have two adjacent fields of view with measured radiances I_1 and I_2 and cloud amounts n_1 and n_2, we obtain

$$I_1 = I_0 + n_1\Delta I(c) \qquad I_2 = I_0 + n_2\Delta I(c) \tag{107}$$

When $\Delta I(c)$ is eliminated,

$$I_0 = \frac{n_2 I_1 - n_1 I_2}{n_2 - n_1} = \frac{I_1 - n^* I_2}{1 - n^*} \tag{108}$$

where $n^* = n_1/n_2$. If we have an independent measurement of surface temperature and measurements of I_1 and I_2 in a spectral window, then n^* can be determined and used to solve for I_0 in different spectral intervals.

As described, this is not an optimum solution, but an optimum solution can be found, based on the same principles.

Another approach, due to *Smith and Woolf* [1976], reduces (108) to a nearly linear form,

$$I_1 = (1 - n^*)I_0 + n^* I_2$$

If we regard $(1 - n^*)I_0$ and n^* as the unknowns, then this is a linear equation. Unfortunately, this leads to more unknowns than equations, so that a priori constraints are required. Smith and Woolf apply these by representing the vector of measurements (19 in their case) by an empirical orthogonal function expansion with 10 terms, and they solve the resulting overconstrained linear equations by least squares. They appear to have found a linear solution to a nonlinear estimation problem. Unfortunately, they have not, because the least squares solution is not optimum. The optimum solution leads to a weighted least squares solution, where the weights depend on the solution. The problem has been reduced to a linear form, but the a priori constraints become nonlinear. Nevertheless, the basic idea is valuable, and the optimum version of the solution can be iterated.

8e. *Limb Sounding*

An important method of improving the vertical resolution of remote sounding is to use limb sounding. Here, one spectral interval and a narrow field of view are used, directed at the limb (Figure 11), so that a large proportion of the radiation reaching the instrument originates near the tangent height, resulting in narrow weighting functions. An idealized set is shown in Figure 12. Nonlinearities arise from three sources in this case. (1) Transmission is dependent on the temperature. (2) The hydrostatic equation gives a variable relationship between tangent height and pressure. (3) Spacecraft attitude is not well enough known, so that absolute tangent height must be derived from the measurements.

If temperature dependence of transmission were the only source of nonlinearity, we would have an ideal case for sequential estimation using (69), introducing the measurements in order, starting at the top, and calculating each weighting function as it is needed, on the basis of the retrieval temperatures above the level being considered. This has been described as the 'onion peeling' approach in its nonoptimal form [e.g., *Russel and Drayson*, 1972].

If the spacecraft attitude is known, then the sequential estimation method could be iterated to allow for the effect of redistribution of absorber amount with height due to changing the temperature profile, but it is not obvious that the process will converge, or converge rapidly enough.

However, the spacecraft attitude is not usually well enough known, so that it becomes necessary to design the experiment to measure attitude or to find its own reference level. This has been done in the case of the LRIR instrument [*Sissala*, 1975] by measuring limb radiance in two spectral intervals, a narrow band and a broad band. To a first approximation the ratio of the two signals is independent of the temperature profile and is equal to the ratio of the absorbtivities of the emitting gas (carbon dioxide), which is a known function of atmospheric pressure. Thus the direction of a known pressure level may be found and may be improved by iteration [*Gille and House*, 1971].

An interesting and efficient method has been developed by P. Bailey and J. C. Gille (personal communication, 1976) for retrieval of LRIR soundings. The quantity retrieved is not the temperature or Planck function profile, but an 'idealized limb radiance profile,' given by the product of the tangent height Planck function at a standard wavelength and an idealized limb emissivity which is a known simple function of the tangent height absorber concentration. The idealized limb radiance is related to the true limb radiance much more linearly than the temperature and composition profiles are, and it may be estimated from the measurements by means of a multiple regression method (section 6*a*). A measure of spacecraft attitude is also included in the multiple regression as an unknown. The temperature and composition profiles are then recovered by an exact Newtonian iteration method from the idealized limb radiance profile. The essence of Bailey and Gille's method is to find a simple nonlinear function of the required profile which is related to the measurements nearly linearly and to use it as an intermediate stage. It may be regarded as a generalization of the approach used in the case of clear sky nadir viewing, where the quantity retrieval is the Planck function profile rather than the temperature profile.

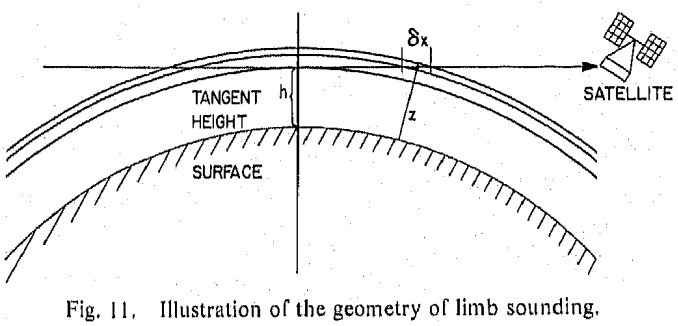

Fig. 11. Illustration of the geometry of limb sounding.

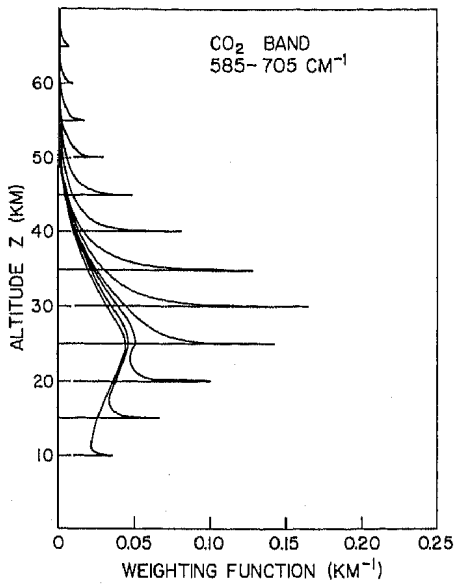

Fig. 12. A set of weighting functions for a realistic limb sounder [after *Gille and House*, 1971].

Acknowledgments. I would like to thank the Advanced Study Program and the Upper Atmosphere Project of NCAR for support and Paul Bailey, John Gille, Neall Strand, and Ed Westwater for reading the manuscript and making many useful suggestions. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

BIBLIOGRAPHY

Alishouse, J. C., et al., A discussion of empirical orthogonal functions and their application to vertical temperature profiles, *Tellus, 19*, 477, 1967.

Backus, G. E., and J. F. Gilbert, Uniqueness in the inversion of inaccurate gross earth data, *Phil. Trans. Roy. Soc. London, Ser. A., 266*, 123–192, 1970.

Barcilon, V., On Chahine's relaxation method for the radiative transfer equation, *J. Atmos. Sci., 32*, 1626, 1975.

Barnett, T. L., Application of a nonlinear least squares method to atmospheric temperature sounding, *J. Atmos. Sci., 26*, 457, 1969.

Chahine, M. T., Determination of the temperature profile in an atmosphere from its outgoing radiance, *J. Opt. Soc. Amer., 58*, 1634, 1968.

Chahine, M. T., Inverse problems in radiative transfer: A determination of atmospheric parameters, *J. Atmos. Sci., 27*, 960, 1970.

Chahine, M. T., A general relaxation method for inverse solution of the full radiative transfer equation, *J. Atmos. Sci., 29*, 741, 1972.

Chahine, M. T., Remote sounding of cloudy atmospheres, 1, The single cloud layer, *J. Atmos. Sci., 31*, 233–243, 1974.

Chow, M. D., An iterative scheme for determining sea surface temperatures, temperature profiles, and humidity profiles from satellite-measured infrared data, *J. Geophys. Res., 79*, 430–434, 1974.

Conrath, B. J., On the estimation of relative humidity profiles from medium resolution infrared spectra obtained from a satellite, *J. Geophys. Res., 74*, 3347, 1969.

Conrath, B. J., Vertical resolution of temperature profiles obtained from remote radiation measurements, *J. Atmos. Sci., 29*, 1262, 1972.

DeLuisi, F. J., and C. C. Mateer, On the application of the optimum statistical inversion technique to the evaluation of Umkehr observation, *J. Appl. Meteorol., 10*, 328, 1971.

Deutsch, R., *Estimation Theory*, Prentice Hall, Englewood Cliffs, N.J., 1965.

Foster, M., An application of Wiener-Kolmogorov smoothing theory to matrix inversion, *J. Soc. Ind. Appl. Math., 9*, 387–392, 1961.

Franklin, J. N., Well-posed stochastic extensions to ill-posed linear problems, *J. Math. Anal. Appl., 31,* 682, 1970.

Gautier, D., and I. Revah, Sounding of planetary atmospheres: A Fourier analysis of the radiative transfer equation, *J. Atmos. Sci., 32,* 881–892, 1975.

Gelman, M. E., A. J. Miller, and H. M. Woolf, Regression techniques for determining temperature profiles in the upper stratosphere from satellite-measured radiances, *Mon. Weather Rev., 100,* 542, 1972.

Gille, J. C., and F. B. House, On the inversion of limb radiance measurements, 1, Temperature and thickness, *J. Atmos. Sci., 28,* 1427, 1971.

Glasko, V. B., and Yu. M. Timofeyev, The solution of the thermal sounding problem for the atmosphere using the regularization method, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 4,* 170, 1968*a*.

Glasko, V. B., and Yu. M. Timofeyev, Possibilities of the regularization method in solving the problem of atmospheric thermal sounding, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 4,* 713, 1968*b*.

Herman, B. M., and D. N. Yarger, Estimating the vertical atmospheric ozone distribution by inverting the radiative transfer equation for pure molecular scattering, *J. Atmos. Sci., 26,* 133, 1969.

Jackson, D. D., Interpretation of inaccurate, insufficient and inconsistent data, *Geophys. J., 28,* 97–109, 1972.

Jackson, D. D., Marginal solutions to quasi-linear inverse problems in geophysics: The Edgehog method, *Geophys. J., 35,* 121–136, 1973.

Kaplan, L. D., Inference of atmospheric structure from remote radiation measurement, *J. Opt. Soc. Amer., 49,* 1004, 1959.

Kaplan, L. D., The spectroscope as a tool for atmospheric sounding by satellites, *J. Quant. Spectrosc. Radiat. Transfer, 1,* 89, 1961.

King, J. I. F., Inversion by slabs of varying thickness, *J. Atmos. Sci., 21,* 324, 1964.

Kondratiev, K. Y., A. A. Buznikov, V. P. Kozlov, and A. G. Pokrovskiy, On the information capability of spectral measurements of moisture in the stratosphere and mesosphere, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 10,* 638, 1974.

Kozlov, V. P., On the estimation of the vertical temperature profile from the outgoing radiation spectrum, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 2,* 80, 1966.

Lanczos, C., *Linear Differential Operators,* D. Van Nostrand, Princeton, N. J., 1961.

Mateer, C. L., On the information content of Umkehr observations, *J. Atmos. Sci., 22,* 370, 1965.

Peckham, G., The information content of remote measurements of the atmospheric temperature by satellite IR radiometry and optimum radiometer configurations, *Quart. J. Roy. Meteorol. Soc., 100,* 406, 1974.

Phillips, B. C., A technique for the numerical solution of certain integral equations of the first kind, *J. Ass. Comput. Mach., 9,* 84, 1962.

Pokrovskiy, O. M., Optimal conditions for indirect atmospheric sounding, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 5,* 766, 1969.

Pokrovskiy, O. M., Optimum conditions for indirect sounding of the atmosphere, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 8,* 634, 1972.

Pokrovskiy, O. M., and Yu. M. Timofeyev, The information yield in indirect sounding of various layers of the atmosphere, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 7,* 598, 1971.

Prabhakara, C., B. J. Conrath, R. A. Hanel, and E. J. Williamson, Remote sensing of atmospheric ozone using the 9.6μ band, *J. Atmos. Sci., 27,* 689, 1970.

Rodgers, C. D., Remote sounding of the atmospheric temperature profile in the presence of cloud, *Quart. J. Roy. Meteorol. Soc., 96,* 654, 1970.

Rodgers, C. D., Some theoretical aspects of remote sounding in the earth's atmosphere, *J. Quant. Spectrosc. Radiat. Transfer, 11,* 767, 1971.

Rodgers, C. D., The vertical resolution of remotely sounded temperature profiles with a priori statistics, *J. Atmos. Sci., 33,* 707, 1976.

Russel, J. M., and S. R. Drayson, The inference of atmospheric ozone using satellite horizon measurements in the 1042 cm^{-1} band, *J. Atmos. Sci., 29,* 376, 1972.

Scott, N. A., and A. Chedin, A least squares procedure applied to the determination of atmospheric temperature profiles from outgoing radiance, *J. Quant. Spectrosc. Radiat. Transfer, 11,* 405, 1971.

Sellers, W. D., and D. N. Yarger, The statistical prediction of the vertical ozone distribution, *J. Appl. Meteorol., 8,* 357, 1969.

Shannon, C. E., and W. Weaver, *The Mathematical Theory of Communication,* University of Illinois Press, Urbana, 1962.

Shaw, J. H., et al., Atmospheric and surface properties from spectral radiance observation in the 4.3μ region, *J. Atmos. Sci., 27,* 773, 1970.

Sissala, J. F., Ed., *The Nimbus 6 Users Guide,* Goddard Space Flight Center, Greenbelt, Maryland, 1975.

Smith, W. L., An iterative method for deducing tropospheric temperature and moisture profiles from satellite radiation measurements, *Mon. Weather Rev., 95,* 363, 1967.

Smith, W. L., An improved method for calculating tropospheric temperature and moisture from satellite radiometer measurements, *Mon. Weather Rev., 96,* 387, 1968.

Smith, W. L., Iterative solution of the radiation transfer equation for the temperature and absorbing gas profile of an atmosphere, *Appl. Opt., 9,* 1993, 1970.

Smith, W. L., and H. B. Howell, Vertical distribution of atmospheric water vapor from satellite infrared spectrometer measurements, *J. Appl. Meteorol., 10,* 1026, 1971.

Smith, W. L., and H. M. Woolf, The use of eigenvectors of statistical covariance matrices for interpreting satellite sounding radiometer observations, *J. Atmos. Sci., 35*(7), 1127–1140, 1976.

Smith, W. L., H. M. Woolf, and W. J. Jacob, A regression method for obtaining real time temperatures and geopotential height profiles from satellite spectrometer measurements and its application to Nimbus 3 SIRS observations, *Mon. Weather Rev., 98,* 582, 1970.

Smith, W. L., H. M. Woolf, and H. E. Fleming, Retrieval of atmospheric temperature profiles from satellite measurements for dynamical forecasting, *J. Appl. Meteorol., 11,* 113, 1972.

Staelin, D. H., Measurements and interpretation of the microwave spectrum of the terrestrial atmosphere near 1-cm wavelength, *J. Geophys. Res., 71,* 2875, 1966.

Strand, O. N., and E. R. Westwater, Statistical estimation of the numerical solution of a Fredholm integral equation of the first kind, *J. Ass. Comput. Mach., 15,* 100, 1968.

Surmont, J., and Y. M. Chen, Numerical solution of a nonlinear radiation transfer equation with inadequate data, *J. Comput. Phys., 13,* 288, 1973.

Tikhonov, A. N., On the solution of incorrectly stated problems and a method of regularization, *Dokl. Acad. Nauk SSSR, 151,* 501, 1963.

Turchin, V. F., and V. Z. Nozik, Statistical regularization of the solution of incorrectly posed problems, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 5,* 14, 1969.

Twomey, S., On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature, *J. Ass. Comput. Mach., 10,* 97, 1963.

Twomey, S., The application of numerical filtering to the solution of integral equations encountered in indirect sensing measurements, *J. Franklin Inst., 279,* 95, 1965.

Twomey, S., Indirect measurements of atmospheric temperature profiles from satellites, 2, Mathematical aspects of the inversion problem, *Mon. Weather Rev., 94,* 363, 1966.

Twomey, S., Information content and indirect sensing measurements, *J. Atmos. Sci., 27,* 515, 1970.

Twomey, S., Comparison of constrained linear inversion and an iterative nonlinear algorithm applied to the indirect estimation of particle size distributions, *J. Comput. Phys., 18,* 188, 1975.

Twomey, S., and H. B. Howell, A discussion of indirect sounding methods, *Mon. Weather Rev., 91,* 659, 1963.

Wang, J. Y., On the estimation of low altitude water profiles from ground-based infrared measurements, *J. Atmos. Sci., 31,* 513, 1974.

Wang, J. Y., and R. Goulard, Numerical solutions in remote sensing, *Appl. Opt., 14,* 862, 1975.

Wark, D. Q., On indirect temperature soundings of the stratosphere from satellites, *J. Geophys. Res., 66,* 77, 1961.

Wark, D. Q., and H. E. Fleming, Indirect measurements of atmospheric temperature profiles from satellites, 1, Introduction, *Mon. Weather Rev., 94,* 351, 1966.

Westwater, E. R., and O. N. Strand, Statistical information content of radiation measurements used in indirect sensing, *J. Atmos. Sci., 25,* 750, 1968.

Yakoulev, A. A., On the use of a priori models for solving inverse problems, *Izv. Acad. Sci. USSR Atmos. Oceanic Phys., 10,* 749, 1974.

Yarger, D. N., An evaluation of some methods of estimating the vertical atmospheric ozone distribution from the inversion of spectral UV radiation, *J. Appl. Meteorol., 9,* 721, 1970.

Yamamoto, G., Numerical method for estimating the stratospheric temperature distribution from satellite measurements in the CO_2 band, *J. Meteorol., 18,* 581, 1961.

(Received March 30, 1976;
accepted May 14, 1976.)