SOLUTION TO THE BIVARIATE INTEGRAL INVERSION PROBLEM: THE DETERMINATION OF EMISSION MEASURES DIFFERENTIAL IN TEMPERATURE AND DENSITY

VERONIKA HUBENÝ
University of Maryland, College Park, MD 20742

AND

PHILIP G. JUDGE
High Altitude Observatory, National Center for Atmospheric Research, 2 P.O. Box 3000, Boulder, CO 80307-3000

Received 1995 March 31; accepted 1995 May 12

ABSTRACT

We present a general solution to the inversion problem of determining the “source function” \( f(t, n) \) from integral equations of the form

\[
g_t = \int \int K(t, n) f(t, n) \, dt \, dn.
\]

The function \( f(t, n) \) represents the most information that can be extracted from a set of observables \( \{ g_t \} \) for a source for which the kernel functions \( K(t, n) \), depending on the two independent variables \( t \) and \( n \), can be calculated a priori.

Our specific application is to the inversion of the equations for a set of optically thin emission line intensities \( g_t \) with kernel functions \( K(t, n) \) which depend on both the electron density \( n \) and the temperature \( t \), a problem defined by Jeffries and colleagues in the early 1970s. We determine “regularized” solutions [those for which derivatives of \( f(t, n) \) are minimized to constrain the allowed solutions] for \( f(t, n) \) from which the usual emission measure differential in temperature \( E(t) = \int f(t, n) \, dn \) can be obtained. Unlike some recent work, our solution is fully two-dimensional and is not restricted to cases where functional dependences are assumed to exist between \( t \) and \( n \) in \( f(t, n) \). We compare our solutions for the source terms (derived from inversions of calculated intensities from input source functions) with input source functions, for typical extreme-ultraviolet and UV lines formed in the solar transition region. Details, refinements, and applications are left to a later paper. This work is likely to be relevant to other areas of astrophysics, and can aid in planning observations with spacecraft such as the Hubble Space Telescope and the upcoming SOHO mission.

Subject headings: line: formation — methods: data analysis — Sun: UV radiation — techniques: spectroscopic

1. INTRODUCTION

Integral equations pervade astrophysics and other areas of science (e.g., Craig & Brown 1986; Menke 1989). The simplest case is for integral equations having just one independent variable \( t \):

\[
g_t = \int K(t) f(t) \, dt.
\]

(1)

The “inversion problem” can be stated as follows: given a set of measurements \( \{ g_t \} \), and a set of uncertainties \( \{ \sigma_t \} \), the aim is to determine the most information about the “source function” \( f(t) \) as a function of \( t \), given known kernel functions \( \{ K(t) \} \). [\( f(t) \) is here called the “source function” in analogy with, but more general than, the source function in radiative transfer.] This classical problem has received much attention, with particular emphasis on addressing the ill-posed nature of the problem (see the above references, for example). Our purpose here is not to look further into this problem (recent work specifically on inversion of integral equations for optically thin line intensities is documented by Harrison & Thompson 1991), but to extend existing one-dimensional methods to the bivariate (two-dimensional) case, namely,

\[
g_t = \int \int K(t, n) f(t, n) \, dt \, dn.
\]

(2)

Our motivation comes primarily from a problem posed by Jeffries, Orral, & Zirker (1972a, b). Given a set of frequency-integrated line-intensity measurements \( \{ g_t \} \), uncertainties \( \{ \sigma_t \} \), and kernels \( \{ K(t, n) \} \), computed with a given set of atomic parameters, these authors aimed to determine the source functions \( f(t, n) \) of coronal plasma. If the emitting gas is optically thin, is in statistical equilibrium, has constant element abundances and thermal electron velocity distributions, and hydrogen is fully ionized, the intensity integral can be written in the form of equation (2), since the kernels are calculable in terms of electron density \( n \) and temperature \( t \). Specifically,

\[
K(t, n) = \frac{h \nu}{4 \pi} \frac{n_e A_{\nu2}}{n^2};
\]

1 Visitor, High Altitude Observatory, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000.

2 The National Center for Atmospheric Research is sponsored by the National Science Foundation.
\( t \) and \( n \) are now the electron temperatures and densities, respectively; \( n_\kappa \) is the population density of level \( \kappa \); and \( h\nu \) and \( A_\kappa \) are the transition energy and Einstein-A-coefficient of the line transition labeled \( \kappa \) between atomic levels \( \kappa \) and \( \mu \).

Further motivation is provided by a fact sometimes ignored or glossed over in the analysis of solar and stellar spectra. Density sensitivity is important for the formation not only of intersystem lines but also of permitted lines (e.g., Vernazza & Raymond 1979), because of subtle but important density dependences in ionization and recombination rate coefficients.

As noted by Brown et al. (1991), the function \( f(t, n) \) (or variables equivalent to it) is the most we can ever learn spectroscopically about the plasma structure [from emission-line intensities] in the general case. It should be clear that the true meaning of the oft-used concepts of “temperature-diagnostic” and “density-diagnostic” line ratios, in which ratios of observed line intensities are used to determine some “mean” temperature \( \langle t \rangle \) (or density \( \langle n \rangle \), can be only seen through equation (2). For instance, to find the meaning of \( \langle t \rangle \) from a “good temperature-diagnostic line ratio,” several conditions must be met. First, kernels in both lines \( i \) and \( m \) must be (approximately) separable \([K(t, n) = k_i(t)k_m(n)]\), the dependences on \( n \) must differ \([k_i(t) \neq k_m(t)]\), the dependences on \( n \) must be the same \([k_i(n) = k_m(n)]\) or the integral over \( dn \) in equation (2) must (fortuitously) be the same for both lines, and, last, we must assume the plasma is isothermal \([f(t, n) = \delta(t - \langle t \rangle)h(n)]\). Then, from equation (2), \( \langle t \rangle \) is found from the solution to the equation \( g_i/g_m = \langle k_i(t)k_m(n) \rangle \). The quantity \( \langle t \rangle \) is then just the single temperature of an isothermal plasma that reproduces the observed intensity ratio \( g_i/g_m \). Unfortunately, the physical meaning and further interpretation of \( \langle t \rangle \) are unclear owing to the overly stringent conditions needed just to define it!

2. SOLUTION TO THE BIVARIATE INVERSION PROBLEM

The inversion of the full bivariate problem was not actually addressed by Jefferies and colleagues. Solutions to this had to wait until Almleaky, Brown, & Sweet (1989) and Brown et al. (1991) made pioneering efforts. Brown et al. (1991) formulated the problem, drew attention to its importance, and identified rigorously the meaning of the “source function” \( f(t, n) \). But in solving the problem, these authors had to resort essentially to one-dimensional inversions by assuming specific relationships between the distribution of source terms with respect to density and temperature, as did Jefferies et al. (1972a, b). There appears to be a common belief that the bivariate problem is not formally solvable in the same way that the one-dimensional problem is. For instance, Harrison & McWhirter (1991) state that “in order to proceed it is necessary to make some assumption about the relationship between density and temperature.”

We present here a solution to the bivariate integral inversion problem, solving for the source function \( f(t, n) \) which is the emission measure differential in density as well as in temperature. In our case we have chosen a regularization procedure (e.g. Craig & Brown 1986) to make the inversion well conditioned, but this is not intrinsic to our method. The solution is simple, and it avoids the special and ad hoc assumptions concerning the form of the sources made by Brown et al. (1991).

2.1. Overview of the One-dimensional Problem

Following Craig & Brown (1986, chap. 6), we write equation (1) in discretized form:

\[
J_i = \sum_{j=1}^{n} K_{ij} f_j,
\]

where we have implicitly incorporated integration weights \( w_j \) into the kernels term \( K_{ij} \) for notational convenience. In matrix notation equation (3) becomes \( J = Kf \). Straightforward inversion of the matrix \( K \) fails to produce acceptable solutions for \( f \) because of the well-documented ill-posed nature of the problem. One of many strategies to overcome this is to impose a smoothness constraint on the solution by minimizing the second derivative of \( f \) — the so-called second-order regularized solution for the source term (see Craig & Brown 1986). In one dimension this formulation takes the linear form

\[
(K^T K + \lambda \Lambda) f = K^T g,
\]

where \( K^T \) is the transpose of \( K \), and \( \lambda \) is a Lagrange underdetermined multiplier chosen so that the matrix in parentheses is now well conditioned (but not overconstrained). The second-order “smoothing matrix” \( \Lambda \) for the one-dimensional problem has the simple form (eq. [6.12] of Craig & Brown 1986):

\[
\begin{pmatrix}
1 & -2 & 1 \\
-2 & 5 & -4 & 1 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
1 & -4 & 5 & -2 & 0 & 0 & \ldots & \ldots \\
1 & -2 & 1 & & & & & \\
\end{pmatrix}
\]

The regularized solution for \( f \) is then obtained by inversion of the matrix on the left-hand side of equation (4), followed by multiplication of the right-hand side.

2.2. Extension to the Bivariate Problem

For notational simplicity we focus upon the unregularized formulation of the bivariate problem. Details of regularization and other issues will be presented by Hubeny & Judge (1995). The discretized form of equation (2) is

\[
J_i = \sum_{j=1}^{n} K_{ij} f_j, \quad i = 1, \ldots, k
\]

where the source term \( f_j \) now has two subscripts, \( j = 1, \ldots, \beta, k = 1, \ldots, \gamma \), and the kernel has an additional subscript \( i = 1, \ldots, \alpha \), representing the discretized form of the dependence of the intensities \( j_i \) on the two independent variables specified by \( j \) and \( k \). The inversion problem is formally analogous to the unregularized one-dimensional case. Note that, as in the one-dimensional case, if the \( K_{ik} \) are independent of one another, the defining set of equations (eqs. [6]) are linearly independent and therefore a unique solution for \( K_{ik} \) exists. In physical terms this means that the kernel functions in equation (2) must all be different from one another.

The crux of our method involves reducing the dimension of the array \( K_{ik} \) to a matrix \( K_{ik} \) of two dimensions, which is then amenable to inversion using standard numerical procedures. It
works simply in the following way. Define a new index $\kappa$ and a two-dimensional matrix $K_\kappa$, such that

$$\kappa = \gamma(f - 1) + k,$$

(7)

where $\gamma$ is the maximum value of $k$. The elements $K_\kappa$ are written into the two-dimensional matrix $K_\gamma$. Hence, with a suitable strategy to make the inverse problem well posed, the two-dimensional integral inversion problem can be solved. Once the appropriate regularization array $H$ is determined for the bivariate problem, it is decomposed in the same way as the kernel matrix through equation (7), and the resulting equation has the exact same form as equation (4). In this Letter, we adopt for simplicity an operator $H$ which is a linear combination of the zeroth- and second-order derivatives (cf. Tikhonov 1963), minimizing the weighted sum of the zeroth derivative and the discretized Laplacian. This choice of $\lambda$ is driven by its simple linear form. As we will show, improvements are needed before real applications can be considered.

We note that our solution is equivalent to the full two-dimensional inversion method discussed by Schou, Christensen-Dalsgaard, & Thompson (1994) in the context of inversion of 1.5-dimensional and two-dimensional helioseismic integrals.

2.3. Test Calculations

We have performed test calculations using specific forms for $f(t, n) = f^{\text{num}}(t, n)$, calculating model line intensities $g_\ell$ using representative kernel functions, adding random errors $\sigma$ to $g_\ell$, applying our inversion procedure, and comparing our results for $f^{\text{num}}(t, n)$ with input values. We are here concerned only with showing that the method works in potentially useful regimes. Therefore, we adopt trapezoidal integrations and assume that no errors exist in the calculated kernels. We set the Lagrange undetermined multiplier $\lambda$ such that norm ($H$)norm ($K'$) = $10^{-5}$/$g_\ell$ in the two-dimensional equivalent of equation (1), where “norm” is the 2-norm. This choice was found to give a reasonable balance between smoothness and details in the desired solutions.

The kernels we have used reflect the essential physics of the excitation of emission lines under the assumptions outlined in § 1: they are sharply peaked in electron temperature, by virtue of ionization balance in a plasma dominated by collisions with electrons, and they are flat in the dependence on electron density until a critical density is reached, when they vary with the inverse of the electron density. This reflects the physical situation that collisional de-excitation destroys photons that would otherwise have escaped when the density exceeds a critical density. Thus, the kernels are very broad in the density domain; this leads to solutions which are quite “blurred” in this variable.

The lines we have used are the following (wavelengths are given in angstroms): C II $\lambda$1334.53; C III $\lambda$1977.02, 1175.71, C IV $\lambda$1908.73; C IV $\lambda$1550.77; N III $\lambda$1748.49, 1752.00, 1749.52, N III $\lambda$1898.79, N IV $\lambda$1486.49, N IV $\lambda$765.15, N V $\lambda$1242.80; O I $\lambda$1666.15, O III $\lambda$703.86; O IV $\lambda$1399.78, 1404.81, 1401.16, O IV $\lambda$787.71, 609.83, 553.33; O V $\lambda$1218.34, O V $\lambda$629.73, 759.44; O VI $\lambda$1037.61; Si III $\lambda$1206.50, 1296.73; Si IV $\lambda$1402.77. As usual, the single closing bracket signifies a spin-changing transition with a low oscillator strength and therefore a low value of the critical electron density. The critical electron densities vary from $2 \times 10^{10}$ cm$^{-3}$ for C III] to over $10^{16}$ cm$^{-3}$ for the permitted transitions.

Our first test case was for a fully differentiable kernel and no input errors in the observed intensities (numerical round-off errors only). Figure 1 (Plate L2) shows that the agreement between input and inverted sources is, as expected, very good. But this extremely idealized case merely shows that we can construct a good solution from a well-behaved source term. The remaining badly behaved (nondifferentiable) test cases, in which small random errors of $\pm 1\%$ were added to the model intensities, included the following (see Fig. 1). Test 2: Two density components in an isothermal atmosphere. Test 3: As in test 2, using restricted sets of lines (those with wavelengths greater than 1200 Å). Test 4: As in test 2, but with more realistic random errors of $\pm 10\%$ (1 $\sigma$) introduced into the model intensities prior to inversion. Test 5: As in test 2, but using modified (sharply peaked) kernels in the density domain (see below). Test 6: Two temperature components in a constant-density atmosphere. Test 7: Modified (less steep slope between $\log T_{\text{eff}} - 4.2$ and 5.0 K) solar emission measure differential in temperature $\xi(t)$ at constant pressure (from Raymond & Doyle 1981). Test 8: As in test 7, but constant-density atmosphere assumed. The numerical grids shown included 15 points linearly spaced in $\log n_{\text{e}}$ ($n_{\text{e}}$) between 4.2 and 5.8, and 17 points between 9 and 12 in $\log n_{\text{e}}$ ($n_{\text{e}}$). Test cases 2–8 are quite tough inversion problems; the input source terms change dramatically (by up to two orders of magnitude—we set “zeros” of the sources to 0.01 of the maximum value) between grid points. Nevertheless, the inversion has clearly picked up the major features of the input distribution of sources in most cases. The solutions are much smoother than the input sources because of (1) the breadth of the kernels (2) the regularization matrix $H$. Some negative source solutions and “sidelobes” (especially in cases 2, 4, and 6) are encountered because of these effects and because we have imposed no positivity constraints. While these are problematic for real applications, and work is in progress to determine more physical constraints on the source terms, it is clear that the basic method provides much new information over existing analyses for cases potentially of interest in astrophysical applications.

Three additional points: First, given low enough errors in observations and in the kernels, it should be possible to discriminate between test cases 7 and 8, and thereby we could judge whether the solar transition region is at a constant pressure or closer to constant density (Laming & Feldman 1992). Second, test cases 2 and 5 show how the density dependence in the kernels qualitatively influences the solutions, since in test 5 we arbitrarily used an unphysical kernel with asymptotically a quadratic drop from the critical density to other densities. Third, test cases 2, 3, and 4 show quantitatively how much information on $f(t, n)$ is lost by examining realistically noisy data and fewer numbers of UV and extreme-ultraviolet lines. Thus, the inversion procedure can help to plan efficiently new instrumentation and observational sequences on existing or planned spacecraft such as the Hubble Space Telescope and the new SOHO mission.

3. DISCUSSION

We have presented a new algorithm which can provide solutions, given some necessary imposed constraints on the source functions, to integral equations where the observables...
Fig. 1.—Test calculations for a differentiable source (test 1) and nondifferentiable sources (tests 2–8). The left-hand panels correspond to input source functions $f(t, n)$, the right-hand panels to the inversions computed as described in the text. The ordinates list $\log_{10} n$ (cm$^{-3}$) and the abscissae $\log_{10} T$ (K). The “intensity” scale is linear for test case 1 and logarithmic for the others (as shown by the color bar). The inversions are generally substantially smoother than the input source terms, owing to the regularization imposed on the solutions and to the breadth of the kernel functions. This is especially true in the density domain, as can be seen by the essentially unconstrained solutions below $\log_{10} T = 4.5$ and above $\log_{10} T = 5.3$, where none of our chosen line’s kernels depends strongly on density.

Hubeny & Judge (see 448, L63)
depend explicitly on two independent variables. The algorithm is fast for small numbers of points in the domain of the desired solution; however, the computing timescales according to the time taken to invert the largest matrix [an array equivalent to $K'K$ in eq. (4) of dimension $(\beta \gamma) \times (\beta \gamma)$, where $\beta$ and $\gamma$ are the number of temperature and density grid points in the domain]. Inversion of the $255 \times 255$ matrices in the test cases took roughly 90 s of CPU time on a SPARCstation 10/30 using the Numerical Recipes inversion procedures released with IDL.

Some significant (but computationally minor) developments are needed before serious applications can be considered. These include introducing nonnegativity constraints, considering boundary conditions on $f(t, n)$ (here we conveniently adopted zeros and linear extrapolation in test cases 1–6 and in test cases 7 and 8, respectively), treatment of uncertainties in observations and especially in kernels, choice of the smoothing term $Hf$ in equation (1) (perhaps using maximum entropy or other nonlinear functionals), and refining the numerical integration scheme. But all of these fall within the realm of one-dimensional problems and so can be addressed with existing techniques. All are subjects of a future paper (Hubeny & Judge 1995).

One important and general area of concern is whether the kernels can be calculated accurately a priori for the plasmas of interest. The atomic emission coefficients needed for calculations of the kernels certainly contain errors not smaller than ±10%. But perhaps more important is the fact that evidence has been found for the breakdown of the assumption of statistical equilibrium in the Sun’s transition region (Judge, Woods, & Brekke 1995). If this proves to be a problem elsewhere, then no standard spectroscopic diagnostic technique relying on the calculated kernels can be applied without great care.

With this caveat, obvious applications of our method can include the planning of instruments and observations. For example, one can easily and without ambiguity determine an optimal set of lines to observe to answer a specific scientific problem, within the constraints imposed by the instruments and their operating modes, efficiencies, and limitations. Problems of interest to us include the determination of element abundances independently of temperature in the solar atmosphere as well as the determination of the emission measures differential in temperature and density from existing data sets.

In conclusion, we believe not only that a bivariate inversion technique is possible and easily executable but also that the method will prove to be an important tool for the planning of observations and for data analysis. Although we present the method in the context of the solar transition region, it is applicable to all situations described by an integral equation where the kernels are calculable and are linearly independent. Even more generally, the method may be expanded to higher dimensions in a similar way, but at the expense of performing inversions of larger matrices.

V. H. is grateful to the HAO Summer Students Program for funding during the summer of 1994. We are indebted to T. Bogdan, T. Brown, P. Charbonneau, and V. Hansteen for useful discussions, and an anonymous referee and A. Skumanich for helpful suggestions which greatly improved the paper. We also thank J. C. Brown for helping to clarify some points during proofreading.

REFERENCES

Craig, L. J. D., & Brown, J. C. 1986, Inverse Problems in Astronomy (Bristol: Hilger)

Hubeny, V., & Judge, P. G. 1995, in preparation
Tikhonov, A. N. 1963, Soviet Math.—Dokl., 4, 1035