A&A 401, 951–958 (2003)
DOI: 10.1051/0004-6361:20030152

Astronomy & Astrophysics

A method for deriving stellar space densities

R. L. Branham Jr.⋆

Instituto Argentino de Nivología y Glaciología (IANIGLA), C.C. 330, 5500 Mendoza, Argentina

Received 16 July 2002 / Accepted 28 January 2003

Abstract. The fundamental integral equation of stellar statistics represents a direct, model-independent approach to calculating stellar densities. Many techniques exist for its solution, but some of these require assumptions, such as a Gaussian luminosity function or a specific form for the density function, that may be unrealistic. To solve the equation as an undeterdetermined total least squares system with Tikhanov regularization recognizes that the problem is ill-posed and generally ill-conditioned as well and offers decided advantages: it is unnecessary to assume a Gaussian luminosity function nor a specific form for the density function; discretization error in the kernel of the integral equation as well as the Poisson error in the star counts are accounted for; mean errors for the densities are calculated; the densities are constrained to be both continuous and positive. The greatest drawback to the method comes from the selection of the ridge parameter, but the drawback becomes surmountable. The method is first applied to three examples, general star counts, the distribution of K0 giants, and the distribution of M 2–M 4 dwarfs, and compared with densities calculated from methods such as Malmquist's and the (m, $\log\pi$) table. Regularized total least squares competes well with these methods. Then the method is applied to a new data set from the AC2000.2 catalog to calculate the densities of M giants and supergiants in the directions of the north and south galactic poles. The densities decrease exponentially to near zero at 2000 pc, with half-density points near 550 pc. No evidence for asymmetry between the two hemispheres can be seen.

Key words. galactic structure – methods – data reduction

1. Introduction

As Reed (1983) cogently remarks, "Star counts to faint limiting magnitudes remain the only direct, model-independent probe of the distribution of stars within the Milky Way". The adjectives "direct" and "model-independent" constitute the operative words in this quote: one obtains results by straight computation whose interpretation remains independent of an assumed model for the Galaxy. On the contrary, the models for the Galaxy will be constrained by the results of our computation. That these adjectives are appropriate becomes evident when we look at the fundamental integral equation for stellar statistics. Let $A(m)$ be the number of stars between apparent magnitude limits $m \pm \Delta m/2$, $\Phi(M)$ the assumed or derived luminosity function, the number of stars per cubic parsec per unit interval of absolute magnitude M, and $D(r)$ the density function, the number of stars per cubic parsec at distance r. Our fundamental equation is

$$A(m) = \omega \int_0^\infty \Phi(M) D(r) r^2 \mathrm{d}r, \tag{1}$$

where ω is the area subtended on the sky where the counts are made. Both $A(m)$ and $\Phi(M)$ are given, and Eq. (1) is to be solved for $D(r)$. If no correction has been made for interstellar absorption, then $D(r)$ represents the fictitious density function, which can be converted to the real density function in the presence of absorption. See Mihalas (1968) for details of how to do this.

Equation (1) represents an example of the Fredholm integral equation of the first kind. Although direct and model-independent, it is also deceptively simple. Deceptively, because if the solution were really simple there would be a standard method for calculating it, much as LU decomposition is the standard tool for solving linear systems. But in fact there are myriad ways to solve Eq. (1), each offering characteristic strengths and weaknesses.

Analytic solutions to Eq. (1) are possible given certain assumptions. Crowder (1959), for example, assumes a Gaussian shape for $\Phi(M)$ and a specific form for $D(r)$. He then calculates an analytic $D(r)$. But the analytic solution depends on our assumptions.

More commonly, however, one uses a discretized version of Eq. (1) rather than Eq. (1) itself. If m is discretized into k equal magnitude intervals and r into n equal distance intervals, then Eq. (1) can be replaced by the discretized equation

$$\begin{pmatrix} A(m_1) \\ A(m_2) \\ \vdots \\ A(m_k) \end{pmatrix} = \omega \sum_{j=1}^{n} \begin{pmatrix} \Phi(m_1 + 5 - 5\log r_j) D(r_j) r_j^2 \Delta r_j \\ \Phi(m_2 + 5 - 5\log r_j) D(r_j) r_j^2 \Delta r_j \\ \cdots \\ \Phi(m_k + 5 - 5\log r_j) D(r_j) r_j^2 \Delta r_j \end{pmatrix}, \tag{2}$$

⋆ e-mail: rlb@lanet.com.ar

or in matrix form

$$\boldsymbol{A} = K \cdot D, \tag{3}$$

where

$$\boldsymbol{A} = \left(A(m_1)\ A(m_2)\ \cdots\ A(m_k) \right)^T,$$

$$\boldsymbol{K} = \omega \begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & K_{22} & \cdots & K_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ K_{k1} & K_{k2} & \cdots & K_{km} \end{pmatrix}, K_{ij} = \Phi(m_i + 5 - 5\log r_j) r_j^2 \Delta r_j,$$

and

$$\boldsymbol{D} = \left(D(r_1)\ D(r_2)\ \cdots\ D(r_n) \right)^T.$$

$\boldsymbol{K}$ is referred to as the kernel of the discretized integral equation.

If $k = n$, Eq. (3) becomes n linear equations in n unknowns solved by LU decomposition. Such solutions, however, are suspect. Because of the peakedness of the luminosity function, $\boldsymbol{K}$ is usually ill-conditioned. Figure 1 shows a surface plot of the kernel for the second example of Sect. 3, where the kernel's ill-conditioning becomes manifest. The ill-conditioning of the kernel coupled with the observational error in $\boldsymbol{A}$ plus the discretization error in $\boldsymbol{K}$ results in a $\boldsymbol{D}$ with wildly oscillating and even negative components. Given the errors in $\boldsymbol{A}$ and $\boldsymbol{K}$ Lucy (1974) has appropriately said, "...the problem under consideration is basically one of statistical estimation rather than an exercise in solving integral equations".

This "statistical estimation" has been implemented implicitly in a classical method for the solution of Eq. (3), the (m, $\log\pi$) table (Bok 1937). This method discretizes the logarithm of r rather than r itself and usually takes $n > k$; the table has more columns than rows. One also generally smooths, one way or another, the vector $\boldsymbol{A}$ to help dampen oscillations in $\boldsymbol{D}$. One assumes a preliminary $\boldsymbol{D}$, with positive components, and adjusts it until agreement with $\boldsymbol{A}$ is obtained. Reed (1985) has developed a computer algorithm for calculating an (m, $\log\pi$) table that replaces the tedium of a manual solution by a fast computer computation. Reed also shows how error estimates for the calculated densities, missing in the traditional exposition of the method and essential for any statistical estimation, can be obtained. The (m, $\log\pi$) table, however much praised, nevertheless suffers from two defects unmentioned in the literature. Because $k < n$ the linear system of Eq. (3) is underdetermined; it is hence impossible to calculate a unique solution. Thus, two users with identical input data will find two different density vectors. Even $k = n$, in theory permitting a unique solution, presents difficulties in practice. The (m, $\log\pi$) solution represents a technique for solving linear equations, relaxation, common in the pre-computer era. A preliminary, generally assumed, solution to the linear system of Eq. (3) results in a vector $\boldsymbol{\epsilon} = \boldsymbol{A} - \boldsymbol{K} \cdot \boldsymbol{D}$ of residuals. Adjusting $\boldsymbol{D}$ until $\boldsymbol{\epsilon}$ becomes small calculates the solution. Unfortunately, when $\boldsymbol{K}$ is ill-conditioned many vectors $\boldsymbol{D}$ yield small $\boldsymbol{\epsilon}$. Thus, the solution, once again, fails to be unique.

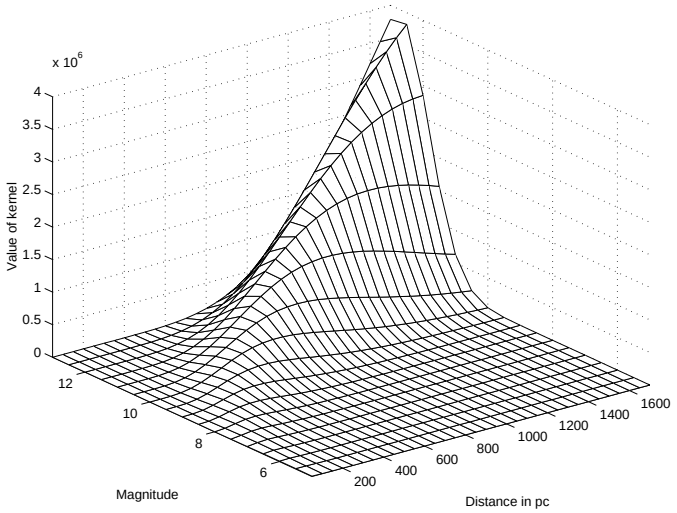

Fig. 1. Structure of Kernel of integral equation.

Cope & Rust (1979) have developed what is probably the most elegant solution to Eq. (3). They use linear programming to calculate upper and lower bounds on the solution vector, all of whose components are constrained to be non-negative, based on the given error in $\boldsymbol{A}$ and, should discretization error be important, also in $\boldsymbol{K}$. Despite its elegance their method suffers from the drawback that one must solve $2n$ linear programming problems. Moreover, and more importantly, should the constraints be inconsistent, then no solution can be found. Inconsistent constraints become likely when one uses real world data. In fact, I found inconsistent constraints for the second example presented later on.

This paper proposes as an alternative that Eq. (3) be solved as an underdetermined total least squares (TLS) problem to calculate a unique solution and with a modified form of Tikhanov regularization to handle the ill-conditioning in $\boldsymbol{K}$. This procedure offers substantial advantages, but with a notable disadvantage that, however, becomes superable.

2. Why total least squares? Why Tikhanov regularization?

Because we are interested in statistical estimation, that venerable workhorse, least squares, comes immediately to mind. If we take $k > n$ Eq. (3) transforms itself to an overdetermined linear system for the unknowns amenable to treatment by least squares that will calculate a unique solution. But $k > n$ becomes an unrealistic requirement. Our magnitude bins cannot be subdivided too finely if we want a reasonable number of stars in each bin. But we do want a fine subdivision for the density intervals, both to minimize discretization error and to provide good resolution for the density solution. This means that $k < n$, an underdetermined system just as we encounter with the (m, $\log\pi$) table. But least squares will nevertheless calculate a unique solution because it imposes the condition that of the infinity of possible solutions to Eq. (3), we chose the one that minimizes the Euclidean norm of the residuals $\boldsymbol{\epsilon}$.

But even with $k < n$ discretization error may be present. The quadrature rule that I use to discretize Eq. (1), eight-order

adoptable Newton-Cotes (Forsythe et al. 1977), assures that discretization error will not be large, but will still nevertheless be present. One should, therefore, solve the linear system by the precepts of total least squares, which allow for error in both the vector $\boldsymbol{A}$ of the data and also in the matrix $\boldsymbol{K}$. See Branham (2001) for a discussion of TLS. Like least squares, TLS calculates a unique solution for underdetermined systems.

Unique, but not necessarily good. No condition that the solution be positive has been enforced. Moreover, should $\boldsymbol{K}$ be poorly conditioned, the solution may still oscillate violently. This is where Tikhanov regularization, developed specifically for ill-posed problems like those presented by Fredholm's integral equation, comes in (Björck 1996). An ill-posed problem exhibits singular values from a singular value decomposition (SVD) of an ill-conditioned matrix such as $\boldsymbol{K}$ that decay gradually to zero with no gap in their spectrum. It thus becomes difficult or impossible to identify insignificant singular values that may be set to zero to remove the ill-conditioning of the matrix. (That singular values decay gradually to zero constitutes no condition sine qua non for an ill-posed problem. If the condition number of the matrix is low, the singular values may decay gradually for a well-posed problem; an orthogonal matrix has all unit singular values, which do not decay at all. But for an ill-conditioned matrix the gradual decay of the singular values becomes important to identify an ill-posed problem.)

Let $\| \cdot \|_2$ denote the Euclidean norm and $\| \cdot \|_F$ the Frobenius norm of a matrix or vector. We seek a solution

$$\| \boldsymbol{K} \cdot \boldsymbol{D} - \boldsymbol{A} \|_F = \min. \tag{4}$$

Tikhanov regularization places bounds on the norm of the solution by changing Eq. (4) to

$$\| \boldsymbol{K} \cdot \boldsymbol{D} - \boldsymbol{A} \|_F = \min. \text{ subject to } \| \boldsymbol{L} \cdot \boldsymbol{D} \|_2 \leq \tau, \tag{5}$$

where $\boldsymbol{L}$ is a nonnull matrix chosen to place a bound on $\boldsymbol{D}$ and τ is called the ridge parameter. $\boldsymbol{L}$, for reasons not immediately evident but based on sound statistics, is typically taken to be a discrete approximation to a finite difference operator. If, for example, we feel that the solution $\boldsymbol{D}$ should be approximately constant, we can take $\boldsymbol{L}$ as the $(n-1 \times n)$ matrix

$$\boldsymbol{L} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}. \tag{6}$$

It is unlikely, however, that the density function would be approximately constant. For various reasons, see Crowder (1959) for example, the function should possess some curvature, perhaps approximately parabolic. For this type of behavior the appropriate operator becomes the $(n-3 \times n)$ matrix

$$\boldsymbol{L} = \begin{pmatrix} 1 & -3 & 3 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -3 & 3 & -1 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -3 & 3 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -3 & 3 & -1 \end{pmatrix}. \tag{7}$$

Although Tikhonov regularization ensures a smooth function, it still does not ensure that the function be positive. To see why examine Eq. (5), equivalent to

$$(\boldsymbol{K}^T \cdot \boldsymbol{K} + \tau \boldsymbol{L}^T \cdot \boldsymbol{L}) \cdot \boldsymbol{D} = \boldsymbol{K}^T \cdot \boldsymbol{A}. \tag{8}$$

By its construction the matrix $\boldsymbol{K}$ has positive elements, see Eq. (3), and therefore $\boldsymbol{K}^T \cdot \boldsymbol{A}$ also has positive elements. $\boldsymbol{K}^T \cdot \boldsymbol{K}$, although with positive elements, is not necessarily diagonally dominant. Because $\boldsymbol{L}$ is not $n \times n$ the matrix $\boldsymbol{L}^T \cdot \boldsymbol{L}$ is subrank and, although nonnegative definite, also not positive definite: it has null singular values. The left-hand-side of Eq. (8), therefore, does not need to be diagonally dominant and thus does not demand that $\boldsymbol{D}$ be positive. But Hansen & O'Leary's extension of Tikhanov regularization to TLS, called regularized total least squares (RTLS), represents one way to obtain a positive solution (1996). Modify Eq. (8) to

$$(\boldsymbol{K}^T \cdot \boldsymbol{K} + \tau_1 \boldsymbol{I} + \tau_2 \boldsymbol{L}^T \cdot \boldsymbol{L}) \cdot \boldsymbol{D} = \boldsymbol{K}^T \cdot \boldsymbol{A}, \tag{9}$$

where $\boldsymbol{I}$ is the $n \times n$ unit matrix. τ_2 becomes a free parameter and

$$\tau_1 = -(\boldsymbol{K} \cdot \boldsymbol{D} - \boldsymbol{A})^T \cdot (\boldsymbol{K} \cdot \boldsymbol{D} - \boldsymbol{A}) / (1 + \boldsymbol{D}^T \cdot \boldsymbol{D}) \tag{10}$$

defines τ_1. Although not immediately obvious, $\tau_2 \gg \tau_1$ and for $\tau_2 \gg 0$ the matrix $\boldsymbol{L}^T \cdot \boldsymbol{L}$ is diagonally dominant, or nearly so. For sufficiently large τ_2 the left-hand-side of Eq. (9) becomes diagonally dominant. The solution, therefore, must of necessity be positive. To calculate a solution one selects arbitrary values for τ_1 and τ_2, solves Eq. (9), calculates τ_1 from Eq. (10), and iterates until the τ_1 converge. If $\boldsymbol{D}$ has negative components one keeps on increasing τ_2 until the solution becomes positive.

Although RTLS provides a positive solution, negative τ_1 entails inconveniences when calculating the TLS covariance matrix. See Branham (1999) for an algorithm for the covariance matrix, which requires the equations of condition. The equations of condition corresponding to Eq. (9) are

See Eq. (11) next page.

The presence of the imaginary unit in rows $n-2$ through $n-2+n$ assures that one must use complex arithmetic even though the final covariance matrix contains only real elements.

A modification of RTLS, however, obviates the need to use complex arithmetic. Use Eq. (8), but define $\boldsymbol{L}$ as an $n \times n$ matrix

$$\boldsymbol{L} = \begin{pmatrix} 1 & -3 & 3 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -3 & 3 & -1 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -3 & 3 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -3 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -3 & 3 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{12}$$

$\boldsymbol{L}^T \cdot \boldsymbol{L}$ is now a full-rank matrix, diagonally dominant, with all non-zero singular values. Taking τ sufficiently large assures that the solution is positive. The RTLS solution differs little if one uses this modification, as I have verified by experimentation. What experimentation? Three of the examples presented

$$\begin{pmatrix} \sqrt{\tau_2} & -3\sqrt{\tau_2} & 3\sqrt{\tau_2} & -\sqrt{\tau_2} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\tau_2} & -3\sqrt{\tau_2} & 3\sqrt{\tau_2} & -\sqrt{\tau_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \sqrt{\tau_2} & -3\sqrt{\tau_2} & 3\sqrt{\tau_2} & -\sqrt{\tau_2} \\ \sqrt{\tau_2}i & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\tau_2}i & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \sqrt{\tau_2}i \\ K_{11} & K_{12} & K_{13} & K_{14} & K_{15} & K_{16} & \cdots & K_{1n} \\ K_{21} & K_{22} & K_{23} & K_{24} & K_{25} & K_{26} & \cdots & K_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ K_{m1} & K_{m2} & K_{m3} & K_{m4} & K_{m5} & K_{m6} & \cdots & K_{mn} \end{pmatrix} \cdot \begin{pmatrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \\ D_6 \\ \vdots \\ D_{n-4} \\ D_{n-3} \\ D_{n-2} \\ D_{n-1} \\ D_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{pmatrix}. \quad (11)$$

later, the one for general star counts, the one for K0 giants, and the one for M dwarfs, used the original RTLS method and then my modification. The curves for the densities for these objects are virtually indistinguishable. Therefore, in the future when I refer to "RTLS" I mean my modification, which uses Eq. (12) rather than Eq. (7). The need to employ complex arithmetic is thus avoided. (If one prefers the original Hansen-O'Leary RTLS, one can calculate the solution from Eq. (9) and the covariance matrix from the equations of condition corresponding to Eq. (8).)

The fact that the density must be positive provides a constraint on the ridge parameter. That the fictitious density must approach zero as the distance from the Sun increases furnishes another constraint. Interstellar absorprtion assures that star counts made along the galactic plane incorporate densities that approach zero (the Kapeyn universe). Counts perpendicular to the plane, where absorption becomes negligible or at least a problem of minor magnitude, imply densities approaching zero because of the concentration of stars in the plane. The selection of the ridge parameter, therefore, represents less of a problem than imagined at first.

3. Some examples

3.1. General star counts

The mathematical exposition given so far is best illustrated with some concrete examples. It will also be useful to compare the method presented here with some alternatives. The first example comes from Mihalas (1968): calculate a density for stars of all spectral types using the van Rhijn luminosity function. Mihalas takes $\Delta m = 1$, with m running from 8 to 19; $\log \pi$ runs from -1.0 to -4.6 in intervals of 0.2; thus r varies from 10 pc to 4 kpc. Mihalas applies no smoothing, such as using reduced star counts, to the data. This is just as well because such smoothing can affect the results. Van Huffel & Vandewalle (1991) conclude that the TLS approach to solving Eq. (3) does not benefit from smoothing, whereas a direct solution by LU decomposition depends on how the data have been smoothed.

The condition number, defined as the ratio of the largest to the smallest singular value of the matrix, for this problem is 1.1×10^7, moderately but not excessively high. The singular value spectrum shows no gaps. The problem, therefore, is moderately ill-posed. Table 1 shows the vector $\boldsymbol{A}$ and the computed values for this vector from: 1) Mihalas; 2) as un undetermined ordinary least squares (OLS) system; 3) the algorithm from RTLS. At first glance it seems as if the OLS solution is the best; the agreement with observation is perfect. But first impressions are deceiving. Figure 2 shows the density calculated from each method, with maximum density normalized to

Table 1. Observed and calculated general star counts.

m	$A(m)$, observed	$A(m)$, Mihalas	$A(m)$, OLS	$A(m)$, RTLS
8	1.26	1.28	1.26	1.14
9	3.80	3.80	3.80	3.56
10	11.0	11.0	11.0	11.1
11	31.6	27.3	31.6	31.8
12	87.0	84.4	87.0	86.9
13	224	221	224	224
14	575	565	575	575
15	1410	1390	1410	1410
16	2880	3110	2880	2880
17	6910	6750	6910	6910
18	15800	14800	15800	15800
19	31662	⋯	31622	31622

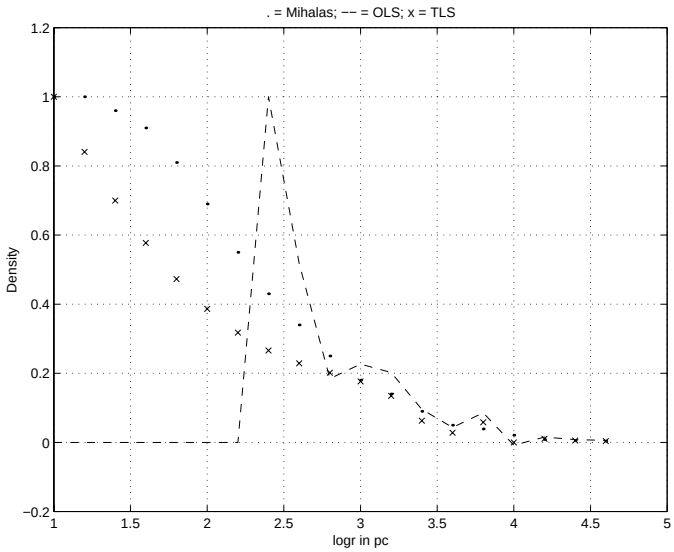

Fig. 2. Calculated stellar densities from three different methods.

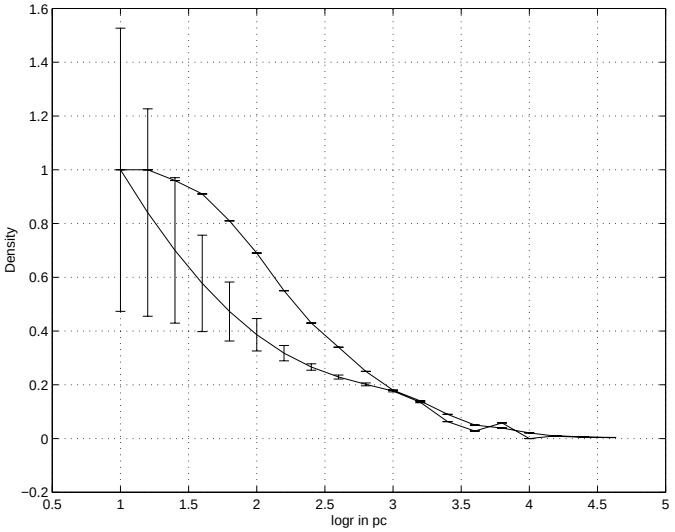

Fig. 3. Calculated stellar densities from RTLS and mlogpi table.

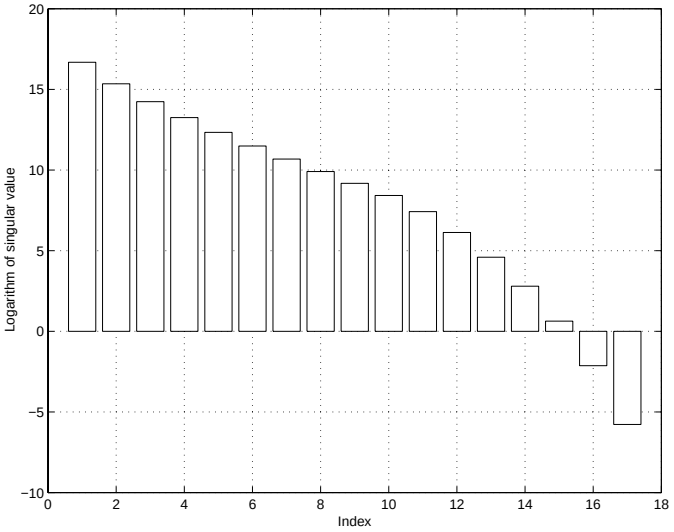

Fig. 4. Spectrum of singular values.

unity. The OLS solution appears totally unacceptable. A null density to 160 pc followed by a sharp rise to a maximum at 250 pc seems unlikely, and the negative density at 10 kpc, albeit only slight, is impossible. The RTLS solution is manifestly the best of the three. The agreement between the observed and the calculated $A(m)'s$ is perfect (within the decimals displayed) for seven of the twelve, and, although worse than OLS's agreement, the calculated density shows none of the objectionable features of the OLS density.

Figure 3 shows the density distribution with error bars and also plots the density that Mihalas calculates, but without errors because he gives none. The errors from the RTLS solution are generous; Reed's algorithm computes smaller errors. But the TLS errors are more realistic because they take into account not only the error in $A(m)$, as with Reed's algorithm, but also the discretization error in $\boldsymbol{K}$. Although Mihalas gives no errors, if we assign something like a twenty per cent error for his densities, which seems of roughly the same magnitude as the RTLS errors, then his density distribution and RTLS's agree within their respective errors.

3.2. Distribution of K0 giants

The second example involves calculating the density distribution of 597 K0 giants (luminosity class III) in the direction of the north galactic pole from data that Upgren (1962) provides. The star counts, with m ranging from 4.8 to 13.6 and $\Delta m = 1/2$, encompass an area of 396 square degrees on the sky. Upgren assumes a Gaussian distribution to represent the luminosity function, with mean absolute magnitude $M_0 = 1.8$ and dispersion $\sigma = 0.8$. Because the stars are counted in the direction of the galactic pole, interstellar absorption becomes negligible and one can take the fictitious star density as the real density. Upgren performs $(m, \log\pi)$ analyses of his data, which permit comparison with the density the RTLS method calculates. To make the comparison as valid as possible I use his reduced star counts, but repeat that smoothing the data is *not* required when one employs the TLS method.

Table 2. Observed and calculated star counts for K0 giants.

m	$A(m)$, observed	$A(m)$, TSVD	$A(m)$, RTLS
5.0	0.2	0.21	0.55
5.5	0.2	0.20	0.93
6.0	0.3	0.27	1.29
6.5	0.5	0.55	1.55
7.0	1.0	0.94	1.75
7.5	1.6	1.65	1.91
8.0	2.5	2.46	2.07
8.5	3.2	3.23	2.28
9.0	4.5	4.48	2.80
9.5	6.0	6.01	4.25
10.0	8.0	7.99	7.58
10.5	11.0	11.00	13.57
11.0	18.0	18.00	21.51
11.5	27.8	27.80	28.24
12.0	36.0	36.00	29.69
12.5	26.0	26.00	24.46
13.0	8.8	8.80	15.54

Because the luminosity function for the K0 giants is more peaked than the general luminosity function, the condition number of the matrix increases to 5.6×10^9. This problem, therefore, is more ill-posed than the previous one. Rather than discretize $\log r$ I opted to discretize r directly, in thirty-four intervals running from 50 pc to 1700 pc.

In addition to comparing the RTLS method with Upgren's solution for the density, I also decided to compare it with a method useful for ill-conditioned, but not ill-posed, problems, the truncated singular value decomposition (TSVD) (Björck 1996). TSVD works when the singular values show a noticeable gap. The smallest singular values can then be assumed to represent noise and suppressed from the solution. If s_i are the singular values, assumed ordered in decreasing order from 1 to n, and k of them are suppressed, the condition number of the matrix decreases from s_1/s_n to s_1/s_{n-k}. The improved conditioning of the matrix can markedly improve the fidelity of the solution. The singular values for the matrix $\boldsymbol{K}$ for this

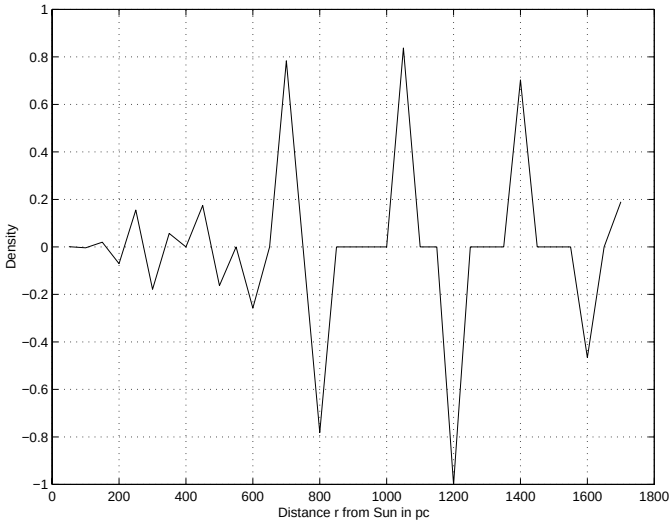

Fig. 5. Calculated stellar densities from truncated singular value decomposition.

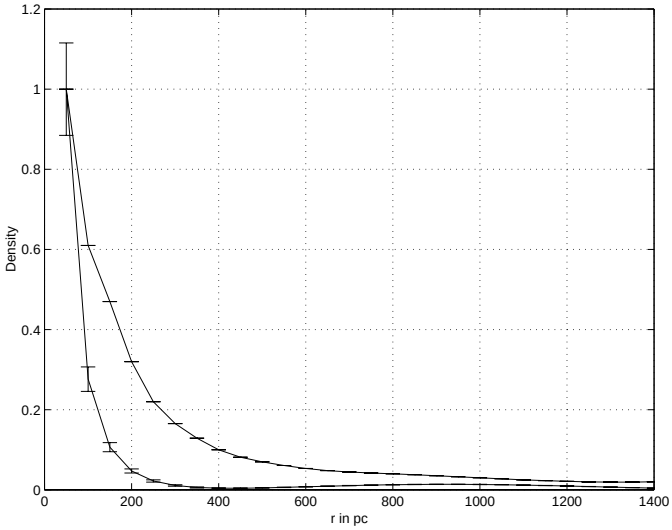

Fig. 6. Stellar densities for K0 giants.

problem, unfortunately, do not comply with the exigency of a spectrum with noticeable gap. They present rather a continuous spectrum, as Fig. 4 shows, with values running from $s_1 = 1.76 \times 10^7$ to $s_n = 0.0031$. Should we nevertheless eliminate the two smallest singular values, the condition number of the matrix decreases to 9.3×10^6, an improvement of nearly three orders of magnitude. Should one examine only the agreement with the observed $A(m)'s$, see Table 2, one might be inclined to say that the agreement is excellent. The solution for the density, however, see Fig. 5, is terrible, with wild oscillations and negative densities. TSVD does not work for problems such as those of solving the stellar density equation, problems not only ill-conditioned but also ill-posed.

RTLS, however, gives an acceptable solution, albeit the agreement with observation is not as good as with the first example, partly a consequence of the greater ill-conditioning of the matrix. Figure 6 shows the calculated density with mean errors and the density that Upgren gives. The error bars, shown

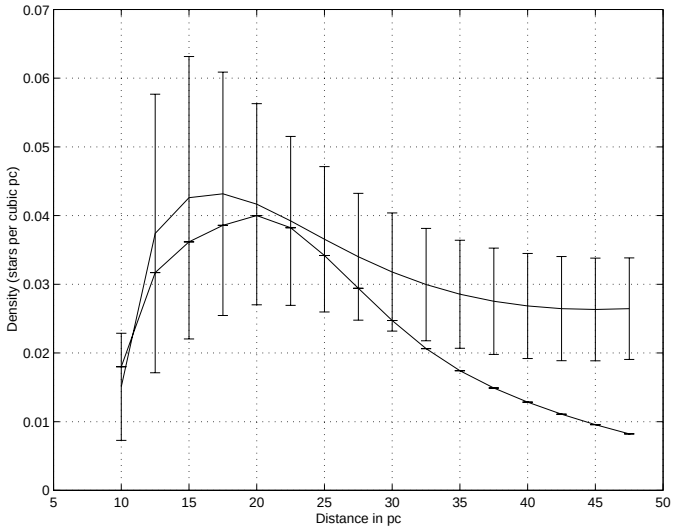

Fig. 7. Stellar densities for M 2–M 4 dwarfs.

Table 3. Observed and calculated star counts for M2–M4 Dwarfs.

m	$A(m)$, observed	$A(m)$, Reed	$A(m)$, RTLS
12.5	2	1.8	1.9
13.0	3	3.5	4.0
13.5	7	7.6	7.3
14.0	16	12.8	11.9
14.5	13	16.4	16.7
15.0	17	17.7	18.5
15.5	19	14.8	14.9
16.0	6	8.3	8.3

only for the RTLS solution because Upgren gives no error estimates, are smaller than those for the previous example because the discretization error is smaller, thirty-four shells instead of nineteen. The RTLS solution shows a density going to zero faster than Upgren's density, with a possible slight rise near 1000 pc. This rise, however, may well be questioned, not because of the formal mean errors, indicating that it seems real, but because it disappears in an independent calculation, Upgren's.

3.3. Distribution of M dwarfs

In both of the previous examples the density is initially high and decays somewhat like an exponential to zero. The next example shows that the RTLS method can handle situations where the initial density is low, rises, and then falls again. The data are taken from The & Staller's study of 84 M2 to M4 dwarfs in the direction of the South galactic pole (1974). Figure 7 graphs the density (this time shown as the actual number of stars per cubic parsec rather than normalized to unity), with its mean error, and compares it with the density that The & Staller calculate. Table 3 shows the agreement between the observed $A(m)'s$ and the computed $A(m)$ and, as a check, the $A(m)$ that Reed (1985) finds.

A comparison of the RTLS density with The & Staller's, computed from Malmquist's method (Mihalas 1968), shows good agreement out to 30 pc, but then the density decreases

more slowly. This behavior is interesting for two reasons. First, unlike the previous example, where the density for the K0 giants falls to zero more rapidly that Upgren's (m, $\log\pi$) calculated densities, here the decrease is slower than the (m, $\log\pi$) calculated densities. Second, the agreement out to 30 pc is good, and conflicts with Reed's higher densities (1985), with a maximum of over 0.08 stars per cubic parsec rather than the ≈0.04 calculated by RTLS. Dolan (1975), using a matrix method that treats the discretized problem as an $n \times n$ linear system (Dolan 1974), also finds a higher density, maximum of 0.09 stars per cubic parsec, see his Fig. 1, albeit with significant mean error, ±0.06, higher than Reed's, ±0.03, or mine, ±0.02. The & Staller feel that the density of M dwarfs near the south galactic pole is definitely lower than the density near the north galactic pole, which lies near 0.12 stars per cubic parsec. Dolan feels that, given the mean errors in the densities, the discrepancy may be more apparent than real. I feel that the matter of northern versus southern galactic pole M dwarf densities needs rediscussion by use of a consistent method that detects whether the discrepancy is real.

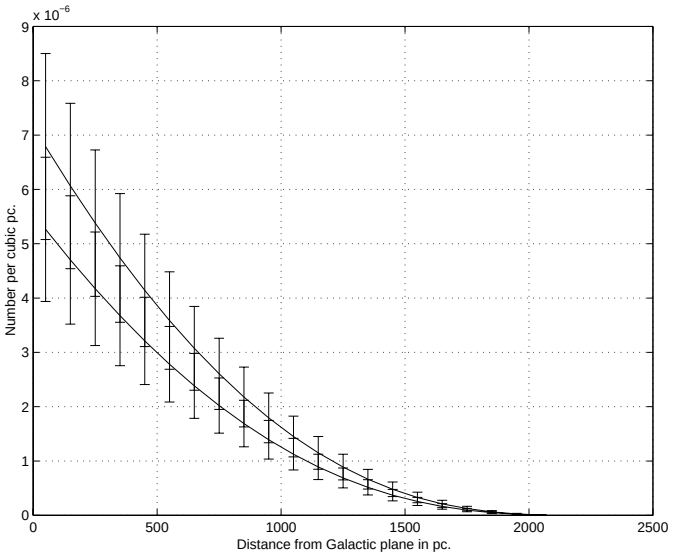

Fig. 8.

4. Distribution of M giants

All of the preceding examples have used data sets published previously to show the utility of the RTLS method and how it competes with alternatives. I would now like to present an example from new data and draw some conclusions from the analysis. The problem will be to study the distribution of M giants, and possibly some supergiants, in the directions of the north and south galactic poles to see how concentrated these stars are towards the galactic plane and if a north-south asymmetry exists in the distribution. The data used are taken from the AC2000.2 catalog (Urban et al. 2001), which contains positions and BV photometry of 4 621 751 stars covering the entire sky. Because U photometry is not included, it became necessary to restrict the study to stars that can be uniquely identified from BV photometry alone. Use of a $(B-V)$ color index greater than 1.65 will isolate M0 and greater giants, although some K4–K9 supergiants may slip in. Given the relative paucity of supergiants, the inclusion of a large number of K supergiants becomes improbable. Circular regions within 5° of the North galactic Pole (NGP) and the South galactic Pole (SGP) were used. Why 5° rather than some other limit? To include an acceptable number of stars and yet still affirm that we look in the directions of the poles. Seventy-eight stars in the NGP region and seventy in the SGP region were found. These numbers, while not large, are nevertheless sufficient to allow reliable density determinations. Because the interstellar absorption is much less of a problem in the directions of the galactic poles, the calculated densities may be taken as real.

For the luminosity function I used the one for M stars tabulated in Mihalas (1968) because I could find nothing better in the literature for M giants. (For M dwarfs one can definitely find something more recent, but my interest is M giants.) I also took, from the same source, a ratio of 1585 main sequence M stars for every giant or supergiant. Figure 8 shows the results of the calculation. We find, once again, an almost exponential decrease in density from the galactic plane to nearly zero at 2000 pc. The rate of decrease, however, is less than that for the K0 giants given previously. At about 550 pc the density is half of what it is at the galactic plane. Blanco (1965), using a list of fifty M giants in the direction of the NGP that Upgren provided, also found that about 500 pc represents the half-density point for M giants. But Blanco also found no significant difference between the density decrease of the K and M giants whereas I find a more pronounced decrease for the K giants; the half-density point occurs at about 150 pc. No evidence for an asymmetry between the NGP and the SGP densities can be asserted when we take into account the mean errors.

5. Conclusions

Although numerous methods exist for solving the fundamental integral equation of stellar statistics, regularized total least squares offers certain advantages. It is unnecessary to assume a Gaussian luminosity function nor a certain form for the density function. By treating the discretized equation as an undetermined linear system solved by total least squares, we calculate a unique solution, within the limits imposed by the ridge parameter and the specific form chosen for the matrix $\boldsymbol{L}$, that accounts for the discretization error as well as the Poisson error in the star counts. A least squares approach also allows one to calculate mean errors for the density, an important aspect not addressed in a classical (m, $\log\pi$) table. Treatment of the data, such as reducing the star counts, becomes unnecessary and even not recommendable. Tikhanov regularization assures that the solution will be continuous.

The greatest drawback to the RTLS approach comes from selection of the ridge parameter. This, however, represents a surmountable difficulty. One starts from an arbitrary, but low in magnitude, value and keeps incrementing it until the densities become not only continuous, but also positive.

The RTLS solution of the integral equation of stellar statistics, therefore, competes well with other methods. It may be recommended as a method to use for the calculation of stellar densities.

References

Björck, A. 1996, Numer. Meth. Least Squares Probl. 204 (Philadelphia: SIAM), 100
Blanco, V. M. 1965, in Galactic Structure, ed. A. Blaauw, & M. Schmidt (Chicago: U. Chicago), 241
Bok, B. J. 1937, The Distribution of the Stars in Space (Chicago: U. of Chicago Press), 26
Branham, R. L. Jr. 1999, AJ, 117, 1942
Branham, R. L. Jr. 2001, New Astron. Rev., 45, 649
Cope, J. E., & Rust, B. W. 1979, SIAM J. Numer. Anal., 16, 950
Crowder, H. K. 1959, AJ, 64, 22
Dolan, J. F. 1974, A&A, 35, 105
Dolan, J. F. 1975, A&A, 39, 463
Forsythe, G. E., Malcolm, M. A., & Moler, C. B. 1977, Computer Methods for Mathematical Computations (Englewood Cliffs, N.J.: Prentice-Hall), Sect. 5.5
Hansen, P. C., & O'Leary, D. P. 1997, in Recent Advances in Total Least Squares and Errors-in-Variables Modeling, ed. S. van Huffel (Philadelphia: SIAM), 127
Lucy, L. B. 1974, AJ, 79, 745
Mihalas, D. 1968, Galactic Astronomy (San Francisco: Freeman), Sect. 4.4, 77–80, 81–83
Reed, B. C. 1983, A&A, 118, 229
Reed, B. C. 1985, J. R. Astron. Soc. Can., 79, 294
The, P. S., & Staller, R. F. A. 1974, A&A, 36, 155
Upgren, A. R. Jr. 1962, AJ, 67, 37
Urban, S. E., Corbin, T. E., Wycoff, G. I., et al. 2001, The AC2000.2 Catalogue, CD_Rom version, U.S. Naval Observatory
van Huffel, S., & Vandewalle, J. 1991, The Total Least Squares Problem (Philadelphia: SIAM), 8