TECHNICAL REPORT NO. 612

July 1980

A NEW APPROACH TO THE NUMERICAL EVALUATION OF THE INVERSE RADON TRANSFORM WITH DISCRETE, NOISY DATA

by

Grace Wahba

This work was supported in part by the Lady Davis Trust and in part by the Office of Naval Research under Grant No. N00014-77-C-0675.
A NEW APPROACH TO THE NUMERICAL EVALUATION OF THE INVERSE RADON TRANSFORM WITH DISCRETE, NOISY DATA

Grace Wahba
Department of Statistics
University of Wisconsin
Madison, WI 53706 USA

Abstract
The inner (singular) integral in the inverse Radon transform for parallel beam computerized tomography devices can be integrated analytically if the Radon transform considered as a function of the ray position along the detector, is a cubic polynomial spline. Furthermore by using some spline identities, large terms that cancel can be eliminated analytically and the calculation of the resulting expression for the inner integral done in a numerically stable fashion. We suggest using smoothing splines to smooth each set of projection data and by so doing obtain the Radon transform in the above spline form. The resulting analytic expression for the inner integral in the inverse transform is then readily evaluated, and the outer (periodic) integral is replaced by a sum. The work involved to obtain the inverse transform appears to be within the capability of existing computing equipment for typical large data sets. In this regularized transform method the regularization is controlled by the smoothing parameter in the splines. The regularization is directed against data errors and not to prevent unstable numerical operations. Strip integral as well as line integral data can be handled by this method. The method is shown to be closely related to the Tikhonov form of regularization.

I. Introduction
Consider a thin "slice" of the human head. In modern computerized tomography (CT) with parallel beam geometry the equivalent of an array of 2N+1 X-ray beams is directed through the slice and the amount of attenuation of each beam is measured. This procedure is repeated as the array is rotated through M positions, s_1,...,s_M, about the head (see Fig.1) to give attenuation factors for a total of n=(2N+1)M beams through the slice. The log of the attenuation
factor for the $i$th beam when the array is in the $j$th position is given approximately by

$$L_{ij} = \int w_{ij}(x,y)f(x,y)dx
dxy$$

(1.1)

where $f(x,y)$ is the X-ray density of the head slice at the point $(x,y)$ and $w_{ij}$ is a non-negative weight function which is 0 outside a strip surrounding the $ij$th beam and represents the non-uniform effective distribution of the beam intensity across its narrow width. The formula makes the approximation that the X-ray attenuation coefficient is constant over the spread of energies present in the (nearly) monochromatic beam. In this report we model the data as

$$z_{ij} = L_{ij} + e_{ij}, \quad i=-N,...,N,$$
$$j=1,2,...,M,$$

where the $e_{ij}$ are independent zero mean random variables with approxi-
mately the same variance which model counting noise and any other (hopefully non-systematic) errors inherent in the measuring device and the approximations being made. The number \( n(2N+1)M \) of data points may be of the order of magnitude of \( 10^5 \).

In practice, various methods are used to estimate \( f \) from the data vector \( z = (z_{-N,1}, \ldots, z_{N,M}) \). The results are usually presented on a video display with different values of the estimate of \( f(x,y) \) represented by different levels on a gray scale. For more detailed discussions of the subject, see, for example Shepp and Logan (1974), Herman and Naparstek (1977).

In Section 2 we review briefly the estimation of a function \( f \) by Tikhonov regularization given data \( z_i = L_i f + c_i \), \( i = 1, 2, \ldots, n \), where the \( L_i \) are arbitrary continuous linear functionals on some appropriate Hilbert space. This approach is not usual in human head and body CT because of the apparent numerical difficulty and the computational convenience of transform methods. See, however Natterer (1980), Artzy, Elfving and Herman (1979). For our purposes, a close examination of this form of regularization will serve to clarify the resolution-noise sensitivity tradeoff common to most regularization methods for dealing with discrete, noisy data. The method is highly appealing in many mildly ill posed problems (as is the CT problem) whenever it is feasible to implement it.

Most modern human CT devices use methods for estimating \( f \) based on an approximate numerical evaluation of a regularized inverse Radon transform. For a recent description of one such algorithm, see Herman Naparstek (1977), Chang and Herman (1978). In Section 3 we present a new approach for the approximate numerical integration of the inverse Radon transform from discrete, noisy data. The work was motivated by a study of, but is apparently quite different from the method described in the above two references. It is in fact quite close to the Tikhonov form of regularization with moment discretization. The method entails using a cubic smoothing spline to obtain a smooth function representing each set of projection data, that is, each set
\[
  z_{j} = (z_{-N,j}, z_{-(N-1),j}, \ldots, z_{N,j}) \quad \text{where } j \text{ is fixed.}
\]
Then the inner (singular) integral in the Radon transform can be evaluated analytically. After using certain relations between the coefficients in cubic splines, one obtains a computationally stable numerical inversion formula which
appears feasible to implement with data sets with $N$ and $M$ of the order order of $10^{2.5}$.

The smoothing parameter in the cubic smoothing splines controls the resolution noise sensitivity tradeoff. The suggested approach bypasses most of the usual discretization, quadrature and aliasing errors common to other methods. Unlike smoothing approaches which are, at least in part, directed against numerical problems connected with evaluating a singular integrand, the present approach directs the smoothing against the noisy data in such a way that the singularity can be integrated out analytically.

In Section 4 we indicate the relationship between the transform method proposed, and Tihonov regularization.

2. Tihonov regularization

Let $H$ be a Hilbert space of functions defined on some domain $\Omega$, let $f \in H$ and suppose one observes

$$z_i = L_if + \varepsilon_i, \quad i=1,2,\ldots,n \quad (2.1)$$

where the $L_i$ are continuous linear functionals on $H$, and the $\varepsilon_i$ are errors. It is supposed that the $\varepsilon_i$ are uncorrelated zero mean random variables with common variance.

Having chosen $H$, the (Tihonov) regularized estimate $f_{n,\lambda}$ of $f$ given the data $z=(z_1,\ldots,z_n)$ is the solution to the problem: Find $f \in H$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} (L_if - z_i)^2 + \lambda \|f\|^2. \quad (2.2)$$

The first term represents the "infidelity" of the solution to the data and, assuming $H$ is a space of "smooth" functions, $\|f_{n,\lambda}\|^2$ represents the "roughness" of the solution. The parameter $\lambda$ controls this tradeoff. Equivalently, $\lambda$ controls the tradeoff between sensitivity to noise, and resolution.

If $\lambda$ is large $\|f_{n,\lambda}\|$ will be small, and the solution will have low resolution but the sensitivity to noise will also be less, since

$$\frac{1}{n} \sum_{i=1}^{n} (L_if_{n,\lambda} - z_i)^2$$

can be larger. A small $\lambda$ will allow $\|f_{n,\lambda}\|$ to be large and correspondingly require $L_if_{n,\lambda}$ to match the data better in mean square.
Since the $L_i$ are continuous linear functionals on $H$, there exist representers $\eta_1, \ldots, \eta_n \in H$ such that $L_i f = \langle \eta_i, f \rangle$, where $\langle , , \rangle$ is the inner product in $H$. Then the minimizer of (2.2) can be shown to be given by

$$f_{n, \lambda} = (\eta_1, \ldots, \eta_n)(Q + n\lambda I)^{-1} z,$$

where $Q$ is the $n \times n$ gram matrix of the representers, with $ij$th entry $Q_{ij}$

$$Q_{ij} = \langle \eta_i, \eta_j \rangle.$$

Equivalently,

$$f_{n, \lambda} = K_n^* (K_nK_n^* + n\lambda I)^{-1} z,$$

where $K_n : H \rightarrow E_n$ is defined by $K_n f = (L_1 f, \ldots, L_n f)$, and $K_n^*$ is the adjoint of $K_n$ in the sense that $(z, K_n f) = \langle K_n^* z', f \rangle$, all $z \in E_n, f \in H$ where $(, , )$ is the Euclidean inner product. $(K_nK_n^*)$ is the operator of matrix multiplication by $Q$ Results are available concerning the convergence of $f_{n, \lambda}$ when $\lambda$ is chosen appropriately and are stated in a little more detail in Section 4.

We remark that if $H = L^2$ then $K_n^* (K_nK_n^* + n\lambda I)^{-1}$ is essentially a back projection operator, see Natterer (1980), however in this case $\lambda$ should be thought of as controlling the scale or dynamic range of the solution rather than its smoothness, and it is then not very important parameter for tumor detection.

We make some remarks on choosing $\lambda$ and the space $H$. Natterer (1980) has suggested that for computerized tomography, $H$ should be chosen as the space $H^\alpha(\Omega)$,

$$H^\alpha(\Omega) = \{ f : \int (1 + |\xi|^2)^\alpha \hat{f}(\xi)|^2 d\xi < \infty, \text{ supp } f \subset \Omega \}$$

where $\hat{f}(\xi)$ is the Fourier transform of $f$ and $\alpha$ is close to $1/2$. Ideally, one should choose $H$ so that it "just" contains the true solutions. If one looks at the problem in "frequency Space" (see Craven and Wahba (1979)) or "eigenfunction space" (see Wahba (1979a)), one can see that the regularized estimate $f_{n, \lambda}$ can be thought of as passing the data through a "low pass filter" where $\lambda$ controls the half power point (or "bandwidth") of the filter and $\alpha$ controls the "shape", or steepness.
of the "roll off" of the filter. For $H$ fixed the method of generalized cross validation (GCV) can be used to estimate a good value of $\lambda$, or in the case of computerized tomography, to obtain good starting values for human "fine tuning". See Wahba (1979b) and references cited there. In the typical tomography problem it will be necessary to utilize the special structure of the problem and possibly to do GCV on a subset of the data. See the appendix.

Herman and Naparstek (1977) and Chang and Herman (1978) have recently studied regularized transform methods for CT reconstruction for a fan beam device. In this section we suggest a new numerical approach to the regularized inversion of the Radon transform for a parallel beam device. A similar but more involved analysis can be carried out for the fan beam inverse transform discussed by Herman and Naparstek but we do not do it here. The method to be given appears to have the advantage of introducing discretization errors and quadrature approximations relatively late in the numerical procedure, and, intuitively, the regularization parameter of the method appears to affect the resolution - noise sensitivity tradeoff in an appropriate manner. The noise suppression filtering acts directly on the raw data. The resulting smoothed data is in such a form that the singular integrand is evaluated analytically, and large terms which cancel are subtracted analytically.

Unstable numerical calculations and further discretization do not appear and thus their effect does not have to be suppressed with further filtering.

The object to be reconstructed is assumed to be within a circle of radius $D$. The device can be considered to be the equivalent of a raster of parallel rays, which are rotated about the axis. Let $\theta$ index the angular position of the raster, $\xi$ the distance from the axis to a parallel ray, and let the location of a point inside the circle be given in polar coordinates as $(r, \theta)$. (See Fig. 1). Then $f(r, \theta)$ is the X-ray absorption coefficient at the point $(r, \theta)$. Let $p(\xi, \theta)$ be the line integral over $f$ taken over the ray indexed by $(\xi, \theta)$.

We begin with the Radon inversion formula for parallel beams as quoted by Herman and Naparstek (1977), equ. (6).

$$f(r, \theta) = \lim_{\varepsilon \to 0} \frac{1}{\pi^2} \int_{-D}^{D} k_\varepsilon(\xi - \xi') \frac{d}{d\xi} p(\xi, \theta) d\xi d\theta$$  (3.1)
where \( k_\varepsilon(u) = \text{if } |u| > \varepsilon \text{ and } k_\varepsilon(u) = 0 \text{ otherwise, and} \)

\[ \ell' = r \cos(\Theta - \phi). \]

We first consider the idealized case where the beams are infinitely narrow. Then the data \( \{z_{ij}\} \) consist of noise contaminated measurements of \( p(\ell_i, \Theta_j) \), that is,

\[ z_{ij} = p(\ell_i, \Theta_j) + \varepsilon_{ij} \quad j = -N, \ldots, N \]

\[ j = 1, \ldots, M. \]

It is desired to estimate \( f(r, \phi) \) on a grid of points \( \{r_k, \phi_j\} \), from this data. It will simplify matters if we let \( \phi_j = 2\pi j/M \), \( j = 1, 2, \ldots, M \). Without loss of generality, we may derive our formulas by setting \( \phi = 0 \) since the formulas for \( \phi = \Theta \) may be obtained by relabeling the data.

First fix \( \Theta = \Theta_j \). The inner integral in (3.1) becomes

\[
\lim_{\varepsilon \to 0} \int_{-\varepsilon}^{\varepsilon} \frac{1}{\ell' - \ell} \frac{d}{d\ell} p(\ell, \Theta_j) \, d\ell + \int_{\ell' + \varepsilon}^{\ell' - \varepsilon} \frac{1}{\ell' - \ell} \frac{d}{d\ell} p(\ell, \Theta_j) \, d\ell
\]

(3.2)

The idea is as follows. One first obtains a cubic smoothing spline approximation, call it \( p_\lambda(\ell, \Theta_j) \) to \( p(\ell, \Theta_j) \). \( p_\lambda(\ell, \Theta_j) \) is the minimizer of

\[
\frac{1}{N} \sum_{i=-N}^{N} (f(\ell_i) - z_{ij})^2 + \lambda \int_{-D}^{D} (f''(\ell))^2 \, d\ell
\]

(3.3)

in \( W^2 [-D, D] \). \( p_\lambda(\ell, \Theta_j) \) has a representation

\[
p_\lambda(\ell, \Theta_j) = a_k + b_k \ell + c_k \ell^2 + d_k \ell^3 / 3, \quad \ell \in [\ell_k, \ell_{k+1}]
\]

where \( a_k, b_k, c_k \) and \( d_k \) are (for fixed \( \lambda \)), linear functions of the data \( z_{ij}, i = -N, \ldots, N \), and \( p_\lambda \) has two continuous derivatives in \( \ell \).

If \( \lambda \) is chosen well, then under circumstances that are likely to be satisfied here, \( p_\lambda \) is a good estimate of \( p(\ell, \Theta_j) \). See Craven and Wahba (1979). We estimate \( \frac{d}{d\ell} p(\ell, \Theta_j) \) by

\[
\frac{d}{d\ell} p_\lambda(\ell, \Theta_j) = b_k + c_k \ell + d_k \ell^2, \quad \ell \in [\ell_k, \ell_{k+1}].
\]

(3.4)

There exist coefficients \( w_{k+i} \), \( w_{k+i} \) and \( w_{k+i} \) independent of the data and depending on \( \lambda \) and \( \{\ell_i\} \) such that

\footnote{but see last paragraph below.}
\[ b_k = \sum_{i=-N}^{N} w_{ki}^b z_{ij} \]
\[ c_k = \sum_{i=-N}^{N} w_{ki}^c z_{ij} \]
\[ d_k = \sum_{i=-N}^{N} w_{ki}^d z_{ij} \]

These coefficients can be determined and stored once and for all (after \( \lambda \) is selected), requiring \( 3(2N)(2N+1) \) storage locations. The storage requirements can be reduced at the cost of time by exploiting recursion relations between the \( \{w_{ki}\} \), we omit the details.

By substituting (3.4) into (3.2) the inner integral can be evaluated analytically and the limit as \( e \to 0 \) taken.

First, let \( \lambda' = r\cos \Theta_j \) be in the interior of \([\lambda_m, \lambda_m+1]\). Then (3.2) becomes

\[
\sum_{i=-N}^{N-1} \int_{\lambda_i}^{\lambda_{i+1}} \frac{b_i + c_i \lambda + d_i \lambda^2}{\lambda^2 - \lambda} \, d\lambda
+ \lim_{e \to 0} \left[ \int_{\lambda_m}^{\lambda' - e} \frac{b_m + c_m \lambda + d_m \lambda^2}{\lambda^2 - \lambda} \, d\lambda + \int_{\lambda' + e}^{\lambda_{m+1}} \frac{b_m + c_m \lambda + d_m \lambda^2}{\lambda^2 - \lambda} \, d\lambda \right]
\]

\[ = J_r(\Theta_j), \text{ say}. \]

Upon carrying out the indicated integrations, one obtains

\[
J_r(\Theta_j) = \sum_{i=-N}^{N-1} (b_i + c_i \lambda + d_i \lambda' \lambda^2) \ln \left( \frac{\lambda' - \lambda_i}{\lambda' - \lambda_{i+1}} \right) + (b_m + c_m \lambda + d_m \lambda^2) \ln \left( \frac{\lambda' - \lambda_m}{\lambda_{m+1} - \lambda'} \right)
+ \sum_{i=-N}^{N-1} (c_i + 2d_i \lambda')(\lambda_{i+1} - \lambda_i) + \frac{1}{2} \sum_{i=-N}^{N-1} d_i ((\lambda' - \lambda_i)^2 - (\lambda' - \lambda_{i+1})^2).
\]

These calculations are all stable except possibly for the two terms involving \( \ln \left( \frac{\lambda' - \lambda_m}{\lambda_{m+1} - \lambda'} \right) \) and either \( \ln \left( \frac{\lambda' - \lambda_m - 1}{\lambda - \lambda_m} \right) \) (if \( \lambda' \) is near \( \lambda_m \) or
\( \ln \left( \frac{\varphi'_{m+1}}{\varphi'_{m} + \varphi_{m+2}} \right) \) (if \( \varphi' \) is near \( \varphi_{m+1} \)). We give the details for \( \varphi' \) close to \( \varphi_{m} \).

Let \( \varphi_{m+1} - \varphi_{m} = h \) and let

\[ \varphi' = \varphi_{m} + \delta h \tag{3.8} \]

where \( 0 \leq \delta < 1/2 \). The possibly offending terms from (3.7) are

\[ (b_{m-1} + c_{m-1} \varphi'_{m-1} + d_{m-1} \varphi'^2_{m-1}) \ln \left( \frac{\varphi'_{m-1} - \varphi_{m-1}}{\varphi'_{m} - \varphi_{m}} \right) + (b_{m} + c_{m} \varphi'_{m-1} + d_{m} \varphi'^2_{m-1}) \ln \left( \frac{\varphi'_{m-1} - \varphi_{m-1}}{\varphi'_{m} - \varphi_{m}} \right) \ln \left( \frac{\varphi_{m+1} - \varphi_{m}}{\varphi_{m+1} - \varphi_{m}} \right) \tag{3.9} \]

Since the cubic smoothing spline has continuous first and second derivatives at \( \varphi_{m} \), we always have the relations

\[ (b_{m} - b_{m-1}) + (c_{m} - c_{m-1}) \varphi_{m} + (d_{m} - d_{m-1}) \varphi'^2_{m} = 0 \tag{3.10} \]

\[ (c_{m} - c_{m-1}) + 2(d_{m} - d_{m-1}) \varphi_{m} = 0 \]

Substituting (3.8) and (3.10) into (3.9) gives that (3.9) is equal to

\[ 2(d_{m} - d_{m-1}) h^2 \delta^2 \ln \left( \frac{\delta}{1+\delta} \right) + (b_{m} + c_{m} \varphi'_{m-1} + d_{m} \varphi'^2_{m-1}) \ln \left( \frac{\delta}{1+\delta} \right) \tag{3.11} \]

If \( 1/2 < \delta < 1 \), a similar expression may be obtained by summing the \( m \)th and \( m+1 \)st terms.

Substituting (3.11) into (3.7) one obtains, provided \( 0 \leq \delta < 1/2 \), and assuming the \( \varphi_{i} \) are equally spaced

\[ J(r) = \sum_{i=1}^{N-1} \left( b_{i} + (i+\delta)hc_{i} + (i+\delta)^2d_{i} \right) \ln \left( \frac{\delta + (m-i)}{\delta + (m-1-i)} \right) \]

\[ + h \sum_{i=-N}^{N-1} (c_{i} + 2(i+\delta)d_{i}) + \frac{1}{2} \sum_{i=-N}^{N-1} d_{i} [2(m+\delta)h - (2i-1)h^2] \]

\[ + 2(d_{m} - d_{m-1}) h^2 \delta^2 \ln \left( \frac{\delta}{1+\delta} \right) \]

\[ + b_{m} + c_{m}(m+\delta)h + d_{m}(m+\delta)^2 \ln \left( \frac{1+\delta}{1-\delta} \right) \tag{3.12} \]

Since \( \delta^2 \ln \left( \frac{\delta}{1+\delta} \right) \) and \( \ln \left( \frac{1+\delta}{1-\delta} \right) \to 0 \) as \( \delta \to 0 \), this expression is computed in a straightforward manner for \( 0 < \delta < 1/2 \), for some suitable
\( e_0 \), and set equal to 0 if \( 0 \leq \delta < e_0 \). A similar expression is obtained for \( 1/2 \leq \delta < 1 \).

Having evaluated \( J_p(\Theta_j) \), the estimate of \( f(r,0) \) is

\[
f(r,0) = \frac{1}{M} \sum_{j=1}^{M} J_p(\Theta_j).
\] (3.13)

Thus, one can process each set of projection data (i.e. the data for fixed \( \Theta_j \)) in parallel. For each \( j \) one collects \( z_j=(z_{-N,j}, \ldots, z_{N,j}) \) computes the \( \{b_k\}, \{c_k\} \) and \( \{d_k\} \) from (3.5), \( J_p(\Theta_j) \) from (3.12) or the corresponding expression for \( 1/2 \leq \delta < 1 \), and \( f(r,0) \) from (3.13). To obtain \( f(r,\Theta_p) \), \( \Theta_p \neq 0 \), one repeats the calculations with each data set \( z_j \) relabeled as \( z_{j-p} \). Note that the coefficients \( b_k, c_k \) and \( d_k \) depend only on \( z_j \). They can be computed in parallel once for each set of projection data and then the projection data discarded.

The regularization parameter here is \( \lambda \) (the choice of \( e_0 \), if reasonable, is secondary). If \( \lambda \) is fixed the \( w_{k,i} \) of (3.5) can be stored.

The ultimate choice of \( \lambda \) (or several values of \( \lambda \) to provide alternative pictures), should, of course be chosen by examining pictures with competitive values for their medical usefulness. Since \( \lambda \) controls the smoothness-fidelity tradeoff, varying \( \lambda \) is likely to have the visual effect of bringing the picture in and out of "focus". A too large \( \lambda \) should result in an oversmoothed, blurred picture while a too small \( \lambda \) should result in an overly grainy or "streaky" picture. A good set of candidate \( \lambda \)'s should be obtainable at the design stage by using the method of generalized cross validation (GCV), on data from typical subjects with the parameters (e.g. number of photons, number \( (2N+1) \) of rays, etc.) that will be used in practice. Transportable code is available for doing this (Merz (1979), Utreras (1979), Fleisher (1979)). Given \( \lambda \), the coefficients \( w_{k,i} \) may be obtained from standard spline theory (e.g. Reinsch (1967)). Numerical results on the estimation of the derivative from noisy data by this method may be found in Craven and Wahba (1979).

We now consider the case where a line integral approximation to the data is not adequate. Suppose it is more appropriate to assume

\[
z_{ij} = \int_{L_i} w_\lambda(x) p(x,\Theta_j) + e_{ij},
\]
say. Then \( p_\lambda(x,\Theta_j) \) is estimated by the minimizer of

\[
\min_{p_\lambda} \left\{ \int_{L_i} \left( p_\lambda(x) - f(x,\Theta_j) \right)^2 \right\}
\]
\[
\frac{1}{2N} \sum_{i=-N}^{N-1} w_i(x) f(x) dx \approx \int_{-D}^{D} (f''(x))^2 dx.
\]

If \( w_i(x) \) is taken as a constant, then \( p_\lambda(x, \Theta_j) \) has a representation

\[ p_\lambda(x, \Theta_j) = \hat{\alpha}_k + \hat{\beta}_k x + \hat{\gamma}_k x^2 + \hat{\delta}_k x^3 + \hat{\epsilon}_k x^4 / 4, \quad \lambda \in [\hat{\alpha}_k, \hat{\gamma}_k] \]

where \( \hat{\alpha}_k, \hat{\beta}_k, \hat{\gamma}_k, \hat{\delta}_k \) and \( \hat{\epsilon}_k \) are linear functions of the data and \( p_\lambda \) has 3 continuous derivatives. Expressions for the \( \hat{\alpha}_k, \hat{\beta}_k, \hat{\gamma}_k, \hat{\delta}_k \) and \( \hat{\epsilon}_k \) can be obtained, for example by using the representation for splines given in Wahba and Wendelberger (1979). An expression for \( J_\tau(\Theta_j) \) is obtained by adding terms corresponding to \( \hat{\epsilon}_k \) to (3.12). \( w_i(x) \) can also be modelled as, e.g. a trapezoid, which will still result in a piecewise polynomial representation for \( p_\lambda \), with a sufficient number of continuous derivatives to carry out a similar analysis. There will be more pieces to the piecewise polynomial, however.

4. On the relation between the spline transform method and Tikhonov regularization

In section 3, we have discussed a new method for the numerical inversion of the Radon transform which essentially consists of smoothing the data in the range space and then inverting the transform analytically. Due to the circular symmetry in \( \Theta \), if one obtains \( p_\lambda(x, \Theta) \) for \( \Theta \not\in \{, \ldots, \Theta_M \} \), by, e.g. any periodic spline interpolant in \( \Theta \) through \( p_\lambda(x, \Theta_k) \), \( k=1, 2, \ldots, M \), and then performs the integrations of (3.1) exactly, the result will be the same, namely (3.13).

We now discuss the relation of such methods to Tikhonov regularization. Let \( H \) be any Hilbert space, let \( L_\tau, \tau \in T \) be a family of linearly independent continuous linear functionals on \( H \) and define the operator \( K \) by

\[
(Kf)(\tau) = g(\tau), \quad g(\tau) = L_\tau f, \quad \tau \in T.
\]

Letting \( X \) be the range of \( K \), we can make \( X \) a Hilbert space with the norm

\[
\|g\|_X = \inf_{f \in H, Kf = g} \|f\|_H
\]

Now consider the data smoothing problem: Find \( g \in X \) to minimize
\[ \frac{1}{n} \sum_{i=1}^{n} (g(t_i) - z_i)^2 + \lambda \| g \|_X^2. \quad (4.1) \]

Letting \( n_T \) be the reprenter of \( L_T \), it can be shown that \( X \) is a reproducing kernel space with reproducing kernel \( Q(s,t) = \langle n_s, n_t \rangle_H \). (See Nashed and Wahba (1974).) It then follows that the minimizer \( g_{n,\lambda} \) of (4.1) is given by

\[ g_{n,\lambda}(t) = (Q(t_1, t), \ldots, Q(t_n, t))(Q + n\lambda I)^{-1} z, \quad (4.2) \]

where \( Q \) is the \( n \times n \) matrix with \( ij \)th entry \( Q(t_i, t_j) = \langle n_{t_i}, n_{t_j} \rangle_H \). Now \( Q(t_i, t) = L_T n_{t_i} = Q(t_i)(t) \), say, so that

\[ Q(t_i) = n_{t_i}. \]

Letting \( L_T = L_{T_i} \), we have by inspection of (2.3) that \( f_{n,\lambda} \), the minimizer of

\[ \frac{1}{n} \sum_{i=1}^{n} (L_T f - z_i)^2 + \lambda \| f \|_H^2 \]

satisfies

\[ K f_{n,\lambda} = g_{n,\lambda}. \]

Now \( n_T, t \in T \) span the orthogonal compliment of the null space of \( K \), since \( \langle n_T, f \rangle = 0, t \in T \Rightarrow Kf = 0 \). Thus \( f_{n,\lambda} \in \mathcal{N}(K) \perp \), so that \( f_{n,\lambda} \) is the unique element of minimal norm satisfying \( Kf = g \) and so (by definition of the generalized inverse \( K^+ \)) \( f_{n,\lambda} = K^+ g_{n,\lambda} \). Thus minimal norm smoothing in the range space (endowed with the induced norm), with an exact inversion is equivalent to Tikhonov regularization in the domain space. Furthermore \( E \| g - g_{n,\lambda} \|_X^2 = E \| K^+ Kf - f_{n,\lambda} \|_H^2 \), where the expectation is taken over the \( \epsilon_i \) and convergence obtains as \( n \to \infty \) under general conditions and \( \lambda = \lambda(n) \) is chosen correctly. See Wahba (1977).

In the procedure we have discussed, smoothing in the \( \Box \) direction is not explicit and any periodic method will give the same result. Let \( p_{\lambda}(\ell, \Theta) \) be obtained by, say, cubic spline interpolation given \( p_{\lambda}(\ell, \Theta_k) \), \( k = 1, 2, \ldots, M \). If \( p_{\lambda}(\ell, \Theta) \) were the minimizer of, say

\[ \frac{1}{N(2M+1)} \sum_{i,j} (g(\ell_i, \Theta_j) - z_{ij})^2 + \int_0^\infty \int_0^\infty \left( \frac{\partial^4 g}{\partial \theta^2 \partial x^2} \right)^2 d\theta dx \quad (4.3) \]

in an appropriate space of functions periodic in \( \Theta \), then the method being proposed would be exactly equivalent to Tikhonov regularization.
The minimizer of (4.3) and \( p_\lambda(\ell, \Theta) \) do not appear to be exactly the same function, however, but the resulting inversion appears to be close.

As far as the choice of space is concerned, this method is appropriate for \( p(\cdot, \Theta) \in W^2_2 \),

\[
W^2_2 = \{ f : f, f', f'' \text{ continuous, } f'' \in L^2([-N, N]) \}.
\]

However, it is more natural to assume \( p(\cdot, \Theta) \in W^1_2 \cap \{ f : f \text{ continuous, } f' \in L^2([-N, N]) \} \) as follows: Consider, for example, head sections \( f(x, y) \) which are continuously differentiable functions of \( x \) and \( y \) plus a tumor which is the equivalent of adding a region of, say, constant higher density. If the boundary of this region is strictly convex and "smooth" then a little reflection will show that \( p(\ell, \Theta) \) is a continuous function of \( \ell \) and \( \frac{\partial}{\partial \ell} p(\ell, \Theta) \) is piecewise continuous, so that \( p(\cdot, \Theta) \in W^1_2 \). The preceding analysis with line integrals cannot be carried out to obtain a stable computational formula because the derivative of the linear spline is not continuous. However a similar analysis can be carried out with a double integral over a \( \Theta \)-increment, or, alternatively, doing spline smoothing assuming \( p(\cdot, \cdot) \in W^1_2 \cap \Theta \) \( W^1_2 \) (periodic). This will appear \( \lambda \) naturally. The ability of \( \frac{\partial}{\partial \ell} p_\lambda(\ell, \Theta) \) to approximate \( \frac{\partial}{\partial \ell} p(\ell, \Theta) \) in the \( W^2_2 \). \( \Theta \) has yet to be established in a practical sense but may be quite satisfactory for the present purpose if there is \( L^2 \) convergence.

Appendix

The GCV estimate of \( \lambda \) is the minimizer of \( V(\lambda) \) given by

\[
V(\lambda) = \frac{1}{n} \sum_{k=1}^{n} (\hat{z}_k(\lambda) - z_k)^2 / (1 - \frac{1}{n} \sum_{i=1}^{n} a_k(\lambda))^2
\]

where \( \hat{z}_k(\lambda) = L_k f_{i, n, \lambda} \), and \( a_k(\lambda) = \frac{\partial}{\partial \lambda} L_k f_{i, n, \lambda} \). See Wahba (1979b) and references cited there. Letting \( L_k = L_{1, j} \) where \( i \) indexes the ray number, and \( j \) indexes the rotational position of the detector \( (i=-N, \ldots, N, \ j=1, \ldots, M \) in the notation of Section 3), due to the rotational symmetry of the device one should have \( a_{i, j}(\lambda) = a_{i}(\lambda) \) independent of \( j \), thus

\[
V(\lambda) = \frac{1}{n} \sum_{i=-N}^{N} \sum_{j=1}^{M} (\hat{z}_{i, j}(\lambda) - z_{i, j})^2 / (1 - \frac{1}{2N+1} \sum_{i=-N}^{N} a_i(\lambda))^2
\]

(I thank F. Natterer for this observation).
One way of obtaining the denominator is to compute $a_i(\lambda)$ as $L_1$
$\delta_{n,\lambda}^{i_1j_1}$, where $\delta_{n,\lambda}^{i_1j_1}$ is the "picture" when the input is
$z=(0,\ldots,0,1,0,\ldots,0)$ with 1 in the $i$th position. As an approxima-
tion to $V(\lambda)$ one might consider

$$V(\lambda) = \frac{1}{(2N^2+1)M'} \sum_{k=-N}^{N'} \sum_{l=1}^{M'} \left( \hat{z}_{k,i_1j_1}^{k_1} - z_{k,i_1j_1}^{k_1} \right)^2 / (1 - \frac{1}{2N^2+1}) \sum_{k=-N}^{N'} a_i^2(\lambda)$$

where $\{i_k\}$ and $\{j_k\}$ are either representative or randomly selected
subsets of the indices.

We note that the use of GCV is appropriate to estimate the smoothing
parameter for low pass filtering methods other than Tikhonov regulariza-
tion, see Craven and Wahba (1979), Golub, Heath and Wahba (1979).

References

1. E. Artzy, T. Elfving, G.T. Herman (1979). Quadratic Optimization
   for image reconstruction, II. Computer Graphics and Image Process-
   sing 11, 242-261.
   convolution reconstruction algorithm. TR No. MIPG 6, Dept. of Com-
   puter Science, SUNY Buffalo.
   the 1110. Academic Computing Center, The University of Wiscon-
   sin-Madison.
   tion as a method for choosing a good ridge parameter. Technometrics
   21, 2, 215-223.
   on a Radon inversion formula appropriate for rapidly collected data.
   SIAM J. Applied Mathematics 33, 511-533.
7. P.H. Merz (1979). Spline smoothing by generalized cross-validation,
   a technique for data smoothing. Chevron Research Corp., Richmond,
   CA.
9. F. Natterer (1979). A Sobolev space analysis of picture reconstruc-
10. F. Natterer (1980). Efficient implementation of 'optimal' algorithms
    in computerized tomography, manuscript.
    177-183.


Acknowledgments

This work was supported in part by the Lady Davis Trust and in part by the Office of Naval Research (USA) under Grant No.N00014-77-C-0675. The author wishes to acknowledge the hospitality of the Mathematics Department of the Technion, Haifa, where the work was completed.