Progress in Scientific Computing
Vol. 2

Edited by
S. Abarbanel
R. Glowinski
G. Golub
H.-O. Kreiss

Birkhäuser
Boston · Basel · Stuttgart

Numerical Treatment of Inverse Problems in Differential and Integral Equations

Proceedings of an International Workshop,
Heidelberg, Fed. Rep. of Germany,
August 30 - September 3, 1982

P. Deuflhard
E. Hairer, editors

1983

Birkhäuser
Boston • Basel • Stuttgart

REGULARIZATION TECHNIQUES FOR INVERSE PROBLEMS IN MOLECULAR BIOLOGY

Stephen W. Provencher and Robert H. Vogel

1. Introduction

In molecular biology, as in most natural sciences, the number of indirect experiments involving ill-posed inverse problems is rapidly increasing. Three of the most important types of inverse problems involve either (a) severely ill-posed linear problems (e.g., Laplace transforms in relaxation or correlation experiments); (b) very large, and perhaps nonlinear, problems (e.g., estimation of three-dimensional structure from x-ray diffraction or electron microscopy); or (c) parameter estimation involving computationally complex models (e.g., multicomponent subnanosecond fluorescence decay strongly convoluted with the instrument response or excitation function).

In this paper we shall discuss four approaches to these problems. Two of these will only be outlined and two discussed in more detail. Common to all of these approaches are two general strategies, the principle of parsimony and the use of prior knowledge. Prior knowledge (e.g., nonnegativity) can be very useful at eliminating the vast majority of members from the (typically infinite) set of solutions that fit the data to within experimental error.

The principle of parsimony says, of all solutions not eliminated by prior knowledge, choose the simplest one, i.e., the one that reveals the least amount of detail or information that was not already known or expected. This is a standard strategy taken by statisticians and experimentalists building models. It is strictly to protect against artifacts and overinterpretation of the data. While the most parsimonious solution may not have all the detail of the true solution, the detail that it does have is necessary to fit the data and therefore less likely to be artifact. The definition of parsimony obviously depends on the problem and prior knowledge. Often smoothness of the solution in a particular space or minimum number of parameters in a model is an appropriate definition.

All four of these approaches maximize an approximate likelihood function, possibly modified by an additive regularizor term, which imposes parsimony or yields an optimal estimate (in the mean square error sense) using prior statistical knowledge of the mean and covariance matrix of the solution [22]. In this way, only the discrete, statistically weighted observed data points are used. This is important because in many cases the statistics of the noise are fairly well known and the noise is often strongly nonstationary. Furthermore, this eliminates the need to extrapolate or interpolate data to estimate (usually infinite) integrals in formal inversion formulas.

2. Constrained Regularization

Imposing prior knowledge of inequality constraints can greatly increase the resolution and stability of the solution. We have found this to be especially important when the solution has significant high frequency content, e.g., sharp edges or isolated peaks [15, 18]. In this section we mention two regularization approaches that can impose inequality constraints. They are described in detail elsewhere and will be only briefly summarized here.

A general-purpose regularization algorithm [16] and portable Fortran package [17] has been developed for linear operator equations subject to any linear equality or inequality constraints imposed by prior knowledge. With numerically stable orthogonal transformations [11] , the general quadratic programming problem is reduced to a ridge regression problem, whose statistical properties have been widely studied. The regularization parameter can then be chosen on the basis of classical confidence regions and F-tests [13, 15].

For the package mentioned above, part of the computation time is proportional to the cube of the number of parameters used to represent the solution. Computations with fewer than about 100 parameters can be done economically. This is usually more than adequate for solutions in one dimension, but not in two or three dimensions. Furthermore the operator equations must be linear. Neutron and x-ray diffraction experiments result in nonlinear operator equations, when the operator produces the absolute value squared of the Fourier transform of the desired electron density. In addition, the data often contains enough information so that $O(10^4)$ to $O(10^5)$ parameters are required to represent the solution. For these large, possibly nonlinear, problems a very efficient optimization algorithm [4] using Frieden's maximum entropy regularizor [7] has been developed. It has been applied to estimating the structure of the Pf1 virion down to a resolution of about 4 Å from x-ray fiber diffraction data [5]. Data from a heavy atom derivative and the native structure were simultaneously analyzed to help reduce the nonuniqueness due to the nonlinear operator (the so-called phase problem). The algorithm has also been applied to three-dimensional reconstruction from electron micrographs and could be applied to a wide variety of other large problems, particularly with missing data.

3. Fast Spline-Model Method for Certain Separable Least Squares Problems

Many commonly used methods for experimental data can be put in the form

$$y_k = \sum_{j=1}^{N_\lambda} \alpha_j f_k(\lambda_j) + \varepsilon_k, \quad k=1,\ldots,N_y, \tag{1}$$

where the y_k are experimental data with unknown zero-mean noise components, ε_k, with finite variances and the α_j and λ_j are to be estimated. The specified functions, $f_k(\lambda)$, are known, but can be expensive to compute. A common case is a convoluted exponential in fast luminescence decay processes,

$$f_k(\lambda) = \int_0^{t_k} \exp(-\lambda\tau)E(t_k-\tau)\, d\tau, \tag{2}$$

where $E(t)$ describes the impulse response of the instrument or the spread of the excitation.

A properly weighted least squares estimator is maximum likelihood when the ε_k are normally distributed and approximately so when the ε_k follow Poisson statistics [14]. However, because of the complexity of eq (2) and the fact that N_y, the number of data points, is typically $O(10^3)$, such an analysis can be expensive. Furthermore, several complete analyses starting from different points in parameter space

should be performed to have a better chance of finding the global optimum. Because of this, there has been considerable interest in transform methods that apply linear operators to the data and typically reduce the problem to solving a set of nonlinear equations. However, we have shown that this can result in serious losses in Fisher information and corresponding increases in the variances of the parameter estimates [19].

In this section we outline a method in which the computation time for a separable least squares analysis becomes independent of the complexity of the model, $f_k(\lambda)$, and of the number of data points, after some preliminary computations have been performed. The first step is to approximate the functions $f_k(\lambda)$ in an expansion of interpolating functions of small support,

$$\tilde{f}_k(\lambda) = \sum_{i=1}^{N_B} \beta_{ki} B_i(\lambda). \tag{3}$$

We use cubic B-spline interpolation at about 40 knots equally spaced on the $z=\ln\lambda$ axis. This approximates model functions of the type in eq (2) typically to within four significant figures. This is generally more than adequate, considering the fact that neither the data nor the model in eq (1) is so accurate anyway, since the model is only an approximation to the true state of nature. Note that this step is simply an interpolation of an exact analytic function. This is quite different and much faster and more reliable than the more common case of fitting splines to the noisy data.

The weighted least squares analysis involves finding the α_j and λ_j that minimize the weighted sum of squared residuals,

$$S \equiv \sum_{k=1}^{N_y} w_k[y_k - \sum_{j=1}^{N_\lambda} \alpha_j \tilde{f}_k(\lambda_j)]^2. \tag{4}$$

Newton and modified Gauss-Newton methods require many evaluations of S, the Hessian of S (or an approximation to it), and the gradient of S with respect to α_j and λ_j. This requires computation of terms like

$$\sum_{k=1}^{N_y} w_k \tilde{f}_k(\lambda_i)\tilde{f}_k(\lambda_j) = \sum_{k=1}^{N_y} w_k \sum_{n=1}^{N_B} \beta_{kn} B_n(\lambda_i) \sum_{m=1}^{N_B} \beta_{km} B_m(\lambda_j) \tag{5}$$

$$= \sum_{n=1}^{N_B} B_n(\lambda_i) \sum_{m=1}^{N_B} B_m(\lambda_j) C_{nm}, \tag{6}$$

where

$$C_{nm} \equiv \sum_{k=1}^{N_y} w_k \beta_{kn} \beta_{km} \tag{7}$$

Another type of term can be similarly evaluated

$$\sum_{k=1}^{N_y} w_k y_k \tilde{f}_k(\lambda_j) = \sum_{n=1}^{N_B} B_n(\lambda_j) d_n, \tag{8}$$

where

$$d_n = \sum_{k=1}^{N_y} w_k y_k \beta_{kn}. \tag{9}$$

The β_{ki} in eq (3) are independent of the data and weights. They depend only upon the model and the experimental design, e.g., the spacing of the t_k in eq (2). Very often for a particular series of experiments these are always the same, and the array β can be computed once and for all and stored. The complicated model functions in eq (2) do not have to be evaluated again. Similarly the array C in eq (7) need only be computed once if the w_k do not change. At worst it, together with the vector $\underset{\sim}{d}$ in eq (9), need only be evaluated once at the beginning of the analysis of a set of data.

When cubic B-splines are used for the $B_i(\lambda)$, the second derivatives of $\tilde{f}_k(\lambda)$ are continuous. Terms analogous to eqs (5) and (8), but containing first or second derivatives of $\tilde{f}_k(\lambda)$ can be easily and rapidly evaluated by replacing the corresponding $B_n(\lambda)$ and $B_m(\lambda)$ with their derivatives. The shift invariance of B-splines with equally spaced knots further simplifies the computation. Because of the compact support of the cubic B-splines, at most only 16 of the N_B^2=O(1600) terms in the double sum in eq (7) are nonzero and need be evaluated.

Separable Gauss-Newton algorithms, in which the α_j are effectively treated as implicit functions of the λ_j and determined by the linear least squares conditions (e.g., with Algorithm I of [20]), have been implemented using formulas of the types above, as has the full Newton method. What cannot be straightforwardly implemented is separability using the numerically more stable differentiation of the pseudoinverse [8] . However, key parts have been coded in double precision, and numerous comparisons with conventional analyses without the spline-model method showed excellent agreement of the parameter estimates with both methods. Furthermore the computational priorities and strategies are now completely different. The evaluation of $f_k(\lambda)$ and its derivatives in eq (2), which are ordinarily the major burden, are now practically free. The main burden now is the matrix algebra. Therefore a procedure like differentiation of the pseudoinverse with a computational complexity proportional to N_y=O(10^3) would result in a major increase in computation.

Under typical conditions [25], our implementation of the separable modified Gauss-Newton algorithm using the spline-model method results in a speed increase of a factor of O(100). This permits an elaborate series of analyses to be performed from many different starting points in parameter space. This has been implemented in a portable user-oriented Fortran IV program [24] and will be available on request. It also permits a second term in eq (1),

$$\sum_{i=1}^{N_g} \gamma_i g_{ki}, \tag{10}$$

where the g_{ki} are known and the γ_i are to be estimated. This is important in allowing corrections for such things as background, and it can be easily handled using formulas similar to eq (8). There is also provision for simultaneously analyzing several sets of data, each having the same set of λ_j, but different α_j and γ_i. This can be very useful, e.g. when spectroscopic measurements at several wavelengths in kinetic studies are made to obtain more reliable estimates of the parameters [10].

The spline-model method is a general approach for very rapidly evaluating the objective function in eq (4), as well as its Hessian and gradient. It may therefore be useful in other optimization or parameter estimation procedures, such as homotopy methods, which can involve a very large number of evaluations of the objective function.

4. Three-Dimensional Reconstruction from Projections of Disordered Objects

4.1 Introduction

Under proper imaging conditions, an electron micrograph yields an estimate of the projection of the electron density of the object. The estimation of the three-dimensional (3-D) electron density from a series of images with the stage tilted to different angles is then formally the same as the inverse Radon transform problem in computer assisted tomography (CAT). However, there are two important additional difficulties. The first is limited data; the stage can only be tilted over a limited angular range, typically (-60^o, 60^o) rather than (-90^o, 90^o). This makes the problem even more ill-posed and seriously diminishes the practical applicability of standard Fourier reconstruction techniques. Second, and most important, is the poor quality data. The objects being studied typically have maximum linear dimensions of $O(10^{-6})$ cm rather than O(1) cm as in CAT. This means that the mass of the object is $O(10^{-18})$ times that in CAT. Thus, in general, by the time enough electrons would have interacted with the object to yield sufficient information, it has long since been completely destroyed.

The most successful strategy to reduce this problem has been to form regular two- (or three-) dimensional arrays of identical objects (particles), reduce the electron dose, and combine the information from the many particles using Fourier methods [23]. However, in general, as the size and complexity of the particle increases so does the difficulty of forming highly ordered regular arrays.

The general problem of combining the information from a number of identical disordered objects with unknown orientations is much more difficult because the relative orientations must be estimated, as well as the electron density. We outline a method for doing this with data from a relatively small number of tilt angles over the limited angular range available in electron microscopy.

4.2 Theory

The electron density, $y(r,\theta,\varphi)$, is expressed as a truncated expansion of a complete orthonormal set of functions,

$$y(r,\theta,\varphi) = \sum_{nlm} \gamma_{nlm}\Psi_{nlm}(r,\theta,\varphi),\ n=1,2,\ldots,N, \tag{11}$$

where

$$\Psi_{nlm}(r,\theta,\varphi) = K_{nl}S_{nl}(r)Y_l^m(\theta,\varphi),\ l=n-1,n-3,\ldots,1 \text{ or } 0,\quad m=-l,-l+1,\ldots,l, \tag{12}$$

$$K_{nl} = \{2\Gamma[(n-l+1)/2]/\Gamma[(n+l+2)/2]\}^{1/2}, \tag{13}$$

$$Y_l^m(\theta,\varphi) = N_{lm}P_l^{|m|}(\cos\theta)\exp(im\varphi), \tag{14a}$$

$$N_{lm} = \{(2l+1)(l-|m|)!/[4\pi(l+|m|)!]\}^{1/2} \tag{14b}$$

$$S_{nl}(r) = r^l\exp(-r^2/2)L^{l+1/2}_{(n-l-1)/2}(r^2), \tag{15}$$

$L_j^k(\cdot)$ are the generalized Laguerre polynomials and $P_l^m(\cdot)$ the associated Legendre polynomials defined in eqs (22.3.9) and (8.6.6) of [1], respectively, and the γ_{nlm} are to be estimated.

These basis functions in eq (11) are the eigenfunctions of the Schrödinger equation for the spherically symmetric harmonic oscillator (see p. 1663 of [12]). They have the following two useful properties:

(a) They are eigenfunctions of the Fourier transform, i.e.,

$$\int \exp(i\underset{\sim}{r}\cdot\underset{\sim}{R})\Psi_{nlm}(r,\theta,\varphi)d^3r = (2\pi)^{3/2}i^{n-1}\Psi_{nlm}(R,\theta,\varphi). \tag{16}$$

This is most easily evaluated by changing to Cartesian coordinates, in which the variables separate and Ψ_{nlm} is just a product of three one-dimensional harmonic oscillator wavefunctions (see p. 1679 of [12]). This makes the application of the projection-slice theorem [2], which says that the Fourier transform of a projection is a central slice (planar section) through the 3-D Fourier transform of the density, very easy. Thus we can compare the Fourier transform of the projection data directly with the 3-D electron density in eq (11) using eq (16).

(b) All of the angular dependence is in the spherical harmonics, $Y_l^m(\theta,\varphi)$, whose behavior upon rotation of the coordinate system can be easily expressed and rapidly computed using the rotation operators for spherical harmonics [3,9].

The Fourier transform of the projection data can then be modelled by transforming eq (11) and rotating the coordinate system through the known angle, τ, of tilt about the x-axis (arbitrarily defined to be the tilt axis) and through three (unknown) Euler angles, $\underset{\sim}{\omega}$, that reorient the particle's coordinate system to coincide with a reference system defined below,

$$\hat{F}(\rho,\Phi;\underset{\sim}{\omega},\tau) = (2\pi)^{3/2} \sum_{nlm}^{n=N} \gamma_{nlm} i^{n-1} \sum_{m'=-1}^{1} R^{1}_{m'm}(\underset{\sim}{\omega}) \times \sum_{m''=-1}^{1} R^{1}_{m''m'}(-\pi/2,\tau,\pi/2)\Psi_{nlm''}(R,\pi/2,\Phi), \quad (17)$$

where the Euler angles and rotation matrices $R^{1}_{m'm}(\cdot)$ are defined by Brink and Satchler [3], which is the only source we could find that was free of errors or inconsistencies. The variables ρ and Φ on the left-hand side are just the polar coordinates in the x-y plane and are numerically equal to R and Φ on the right-hand side.

The term $\Psi_{nlm''}(R,\pi/2,\Phi)$ represents a two-dimensional slice through the x-y plane. Thus this uses the projection-slice theorem assuming that the projection is parallel to the z-axis of the coordinate system used in eq (11). For each particle, the unknown vector, $\underset{\sim}{\omega}$, of Euler angles rotates the coordinate system of that particle so that this is the case. If all particles had the same orientation, then no $\underset{\sim}{\omega}$ would be necessary since with large enough N the γ_{nlm} could represent the particle in any orientation. In practice we have always used this reasoning and arbitrarily fixed the coordinate system of one particle to be the reference to which all the others are rotated. Thus with N_p particles there are only (N_p-1) vectors $\underset{\sim}{\omega}$. However, with a relatively small N, it might be better to allow all the particles to rotate so that the limited number of terms in eq (11) can be most efficiently used. This would in any case be necessary if one were using only a

subset of the spherical harmonic terms to impose a particular symmetry, e.g., icosahedral symmetry [6], if the exact orientations of the symmetry axes were not known.

Despite the relatively compact and computationally efficient model in eq (17), the computational burden in a straightforward weighted least squares analysis would be overwhelming. This is mainly because of the number of data points and the five-fold sum in eq (17). A typical image is digitized to a 64×64 array and a discrete Fourier transform would yield the same number of values, i.e., $O(10^4)$. For 20 particles and 9 tilt angles, this would amount to a nonlinear least squares analysis with $O(10^6)$ rows. This amount of data can be reduced by a factor O(100) with almost no loss in information by applying an orthogonal transformation based on the orthogonality properties of the basis functions with respect to Φ and a polar coordinate sampling theorem [21].

All of the Φ dependence in eq (17) is in the term $\exp(im\Phi)$ in the spherical harmonics in eqs (16) and (12). Because of the orthogonality properties of this term, the circular transform,

$$\hat{F}_{\hat{m}}(\rho;\underset{\sim}{\omega},\tau) \equiv (1/2\pi)\int_0^{2\pi} \exp(-i\hat{m}\Phi)\hat{F}(\rho,\Phi;\underset{\sim}{\omega},\tau)d\Phi, \quad (18)$$

eliminates the innermost sum in eq (17) and reduces to

$$\hat{F}_{\hat{m}}(\rho;\underset{\sim}{\omega},\tau) = (2\pi)^{3/2} \sum_{nlm}^{n=N} \gamma_{nlm} i^{n-1} K_{nl} S_{nl}(\rho) N_{1\hat{m}} P_1^{|\hat{m}|}(0) \times \sum_{m'=-1}^{1} R^{1}_{m'm}(\underset{\sim}{\omega}) R^{1}_{\hat{m}m'}(-\pi/2,\tau,\pi/2) \quad (19)$$

Furthermore there are only nonzero terms when $|\hat{m}|<N$; i.e., there are only (2N-1) $\hat{m}$ values needed to represent all the information relevant to the model in eq (11).

The radial variable, ρ, can also be sampled. Although neither the model in eq (11) nor its Fourier transform in eq (16) are of compact support, they both can be considered to be approximately so. This is because the radial parts of both the function and its Fourier transform are strongly damped toward zero with increasing r or R by the Gaussian factor in eq (15). We denote by r_{max} and ρ_{max}, respectively, the values of r and ρ beyond which the model and its transform can be considered to be negligibly small compared to the maximum electron density (in the model) or the noise components (in its transform). These cutoff values depend weakly on the value of N in eq (1) and the signal-to-noise ratio in the data, but $\rho_{max}=4$ and $r_{max}=6$ have been found to be sufficiently large for $N\leq 13$. Space-limiting the model to $r\leq r_{max}$, the polar coordinate sampling theorem [21] says that all the information is obtained by sampling $\hat{F}_{\hat{m}}(\rho;\underset{\sim}{\omega},\tau)$ at ρ values given by

$$\rho_{\hat{m}k} \equiv Z_{\hat{m}k}/r_{max}, \tag{20}$$

when $Z_{\hat{m}k}$ is the kth zero of the Bessel function $J_{\hat{m}}(\rho)$. Band-limiting the model to Fourier components with $\rho\leq\rho_{max}$ (because the high-frequency Fourier components of the signal in the data become negligible compared with the components of the noise) yields a greatly reduced number of data points for each image, O(100) rather than $O(10^4)$,

$$\hat{F}_{\hat{m}k}(\underset{\sim}{\omega},\tau) \equiv \hat{F}_{\hat{m}}(\rho_{\hat{m}k};\underset{\sim}{\omega},\tau). \tag{21}$$

In order to fit the data to the model in eqs (21) and (19), the data must be transformed as follows:

$$F_{\hat{m}}(\rho;\underset{\sim}{\omega},\tau) = (1/2\pi)\int_0^{2\pi} d\Phi \exp(-i\hat{m}\Phi)\int d^2r \exp(i\underset{\sim}{r}\cdot\underset{\sim}{\rho})f(x,y), \tag{22}$$

where f(x,y) is used to represent the data because the images are scanned with a Cartesian grid. The integral over Φ can be performed analytically (see pp. 1678-1680 of [12]) to yield

$$F_{\hat{m}k}(\underset{\sim}{\omega},\tau) = i^{\hat{m}} \int_0^{r_{max}} dr\, r \int_0^{2\pi} d\varphi J_{\hat{m}}(r\rho_{\hat{m}k})\exp(-i\hat{m}\varphi)f(x,y). \tag{23}$$

This transform is orthogonal with respect to both of the indices $\hat{m}$ and k. The orthogonality with respect to $\hat{m}$ is clear from the orthogonality of $\exp(-i\hat{m}\varphi)$ over the interval $\varphi\in[0,2\pi]$. The orthogonality with respect to k follows from a change of variable to $t=r/r_{max}$, eq (20), and the standard orthogonality relation for integrals involving zeroes of Bessel functions, eq (11.4.5) of [1],

$$\int_0^1 tJ_{\hat{m}}(Z_{\hat{m}k}t)J_{\hat{m}}(Z_{\hat{m}k'}t)dt = 0,\ k\neq k'. \tag{24}$$

In practice, eq (23) is evaluated by numerical quadrature, and $\underset{\sim}{F}$, the vector of $F_{\hat{m}k}(\underset{\sim}{\omega},\tau)$ values is simply the linear transformation

$$\underset{\sim}{F} = C\underset{\sim}{f}, \tag{25}$$

where $\underset{\sim}{f}$ is the vector of image data points, f(x,y), and the matrix C (typically about $100\text{x}64^2$) accounts for the quadrature weights, the Cartesian grid of the data points, and the kernel of the transformation in eq (23). To within quadrature error, the matrix C is Hermitean because of the orthogonality of the integral transforms mentioned above. This is very convenient, because, if the covariance matrix of $\underset{\sim}{f}$ is an identity matrix, the covariance matrix of $\underset{\sim}{F}$ is diagonal. That is, uncorrelated stationary noise in the projection data $\underset{\sim}{f}$ remains uncorrelated in the reduced data $\underset{\sim}{F}$, and one can perform a simple weighted least squares fit of $\underset{\sim}{F}$ in eq (25) to the model in eqs (21) and (19). Otherwise one would have to work with a non-diagonal covariance matrix in the least squares analysis. The assumption of uncorrelated stationary noise is often not bad, even when the total signal is Poisson, because a large background must often be subtracted from the total signal to obtain $\underset{\sim}{f}$.

4.3 Practical Aspects

The matrix C in eq (25) typically takes about 20 min of CPU time to compute. However, it depends only on such things as r_{max} and ρ_{max} and the number of data points in the scanning grid for the images. These usually remain the same from one experiment to the next, and C can therefore be computed once and for all and stored. The reduction of a complete image to $\underset{\sim}{F}$ in eq (25) then takes only a few seconds of CPU time. From this point onwards, only the reduced data, $\underset{\sim}{F}$, is used.

All of the images must have the same origin. Fortunately, the center of gravity of a projection is independent of the orientation of the particle, and this is used as the origin. This is also a good choice in that it generally results in relatively rapid convergence of the expansion in eq (11). In practice the estimation of the center of gravity of each image must be done with care, after subtraction of the background.

A more general set of functions than $\Psi_{nlm}(r,\theta,\varphi)$ in eq (12) contains a scale factor multiplying r. We have arbitrarily set this to one. However, from section 4.2 it is clear that the r values in the input data must be scaled so that the maximum extent of the projections in the images is about r_{max}.

Regularization is imposed by the upper limit N. We start with a relatively small N, typically 5, and increase N until the decrease in the weighted sum of squared deviations of the fit to $\underset{\sim}{F}$ (in a plot versus N) does not seem to be significant. Classical F-tests might also be helpful. Experience so far indicates that imposing parsimony by truncation is not as bad here as truncating Fourier series or transforms, but slightly smoother solutions could probably be obtained by a more gradual tapering of the expansion in eq (11). However, the number of γ_{nlm} parameters is N(N+1)(N+2)/6, and such a gradual taper would be very expensive except for very low resolution solutions. Furthermore the stepwise increase in N results in a natural series of solutions with increasing detail. This can be seen from the natural hierarchy of increasing complexity of the spherical harmonics or the wave functions in eq (12) with increasing l or n. We have performed analyses with N as large as 13 (with 455 γ_{nlm} parameters), but N=9 (with 165 γ_{nlm}) is often sufficient to represent reasonably large structures to within the resolution attainable in electron microscopy.

There are generally far more linear (γ_{nlm}) parameters than nonlinear ($\underset{\sim}{\omega}$) ones. If N_p is the number of particles being analyzed, then there are either $3(N_p-1)$ or $3N_p$ nonlinear parameters, and N_p seldom exceeds 20. Therefore a least squares analysis exploiting separability should greatly improve the rate and region of convergence.

Because of the large size of the problem and the formulation, nonnegativity was not imposed. This is not as serious as in other cases because of the relatively low resolution attainable in electron microscopy.

The disorder of the particles prevents the problem from being reduced to a series of independent two-dimensional reconstructions of slices through the 3-D structure, as is often possible in CAT. The need for a direct 3-D reconstruction brings with it an unavoidable added computational burden, but it does have the advantage that

smoothness of the entire 3-D structure tends to be imposed, and this is generally not the case when a series of two-dimensional slices are reconstructed independently.

One of the main advantages of this formulation in terms of a straightforward problem in parameter estimation by weighted least squares is that the approximate covariance matrix of the parameter estimates is obtained. This gives a very clear indication of the reliability of the estimated structures and a warning when N is getting too large. It also permits general theoretical studies of the effects of such things as the number and range of tilt angles and the disorientation of the particles on the uncertainties in the estimates. It turns out that disordered particles can actually bring a benefit in that a wider range of views are obtained over the limited range of tilt angles available than if all particles had the same orientation. In fact simulations indicate that the Fisher information content for N_p randomly oriented particles can be almost as large as that for N_p^2 particles of identical orientation, even including the extra uncertainty due to the extra $\underset{\sim}{\omega}$ parameters for the disoriented case.

For a set of particles with identical orientations and an upper limit of N in eq (11), precisely N different tilt angles are needed; otherwise the parameter covariance becomes singular and the parameters indeterminate. With disoriented particles, this requirement can be relaxed, but it is still recommended. It is important to be able to use as few tilt angles as necessary. This permits the total dose of electrons that the particles can tolerate to be divided into larger doses for each tilt angle. This makes it easier and more reliable to subtract background, to locate the center of gravity, and to perform the rest of the steps in the analysis.

The information content deteriorates as the range of tilt angles is restricted, but more slowly for disordered than for ordered particles. In fact, simulations indicate that a tilt range of $(-45^{\circ},45^{\circ})$ is often sufficient with a set of disordered particles. It would be advantageous if the commonly used extremely oblique tilts in the range $(-60^{\circ},60^{\circ})$ could be avoided.

Another advantage of the statistical treatment of the problem is that the large data reduction in eq (25) is immediately demanded by the analysis. Sampling $\hat{F}(\rho,\Phi;\underset{\sim}{\omega},\tau)$ in eq (17) more finely resulted in a practically singular covariance matrix for the weighted least squares analysis, and it was apparent that a sampling theorem had to be used.

The method has been extensively tested with simulated projections and added noise of the level typically found in electron micrographs, and the results have been very encouraging. However, it is necessary to make tests with real micrographs containing the numerous systematic errors and artifacts that occur in electron microscopy, and this has been started.

5. References

[1] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions (Dover, New York, 1965).

[2] R.N. Bracewell, Strip integration in radioastronomy, Australian J. Phys. 9, 198 (1956).

[3] D.M. Brink and G.R. Satchler, Angular Momentum (Clarendon Press, Oxford, 1968).

[4] R.K. Bryan, Ph. D. Dissertation, University of Cambridge (1980).

[5] R.K. Bryan, M. Bansal, W. Folkhard, C. Nave, & D.A. Marvin (in preparation).

[6] N.V. Cohan, The spherical harmonics with the symmetry of the icosahedral group, Proc. Cambr. Phil. Soc. 54, 28 (1958).

[7] B.R. Frieden, Restoring with maximum likelihood and maximum entropy, J. Opt. Soc. Amer. 62, 511 (1972).

[8] G.H. Golub and V. Pereyra, The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, SIAM J. Numer. Anal. 10, 413 (1973).

[9] Z. Kam, The reconstruction of structure from electron micrographs of randomly oriented particles, J. Theor. Biol. 82, 15 (1980).

[10] H. Lachmann, New methods of parameter identification in kinetics of closed and open reaction systems, this volume.

[11] C.L. Lawson and R.J. Hanson, Solving Least Squares Problems (Prentice-Hall, Englewood Cliffs, 1974).

[12] P.M. Morse and H. Feshbach, Methods Of Theoretical Physics, Part II (McGraw-Hill, New York, 1953).

[13] R.L. Obenchain, Classical F-tests and confidence regions for ridge regression, Technometrics 19, 429 (1977).

[14] P.F. Price, A comparison of least-squares and maximum-likelihood estimators for counts of radiation quanta which follow a Poisson distribution, Acta Cryst. A35, 57 (1979).

[15] S.W. Provencher, Inverse problems in polymer characterization: Direct analysis of polydispersity with photon correlation spectroscopy, Makromol. Chem. 180, 201 (1979).

[16] S.W. Provencher, A constrained regularization method for inverting data represented by linear algebraic or integral equations, Comput. Phys. Commun. 27, 213 (1982).

[17] S.W. Provencher, CONTIN: A general purpose constrained regularization program for inverting noisy linear algebraic and integral equations, Comput. Phys. Commun. 27, 229 (1982).

[18] S.W. Provencher, in Photon Correlation Techniques, edited by E.O. Schulz-Dubois (Springer, Heidelberg, 1983).

[19] S.W. Provencher and R.H. Vogel, Information loss with transform methods in system identification: A new set of transforms with high information content, Math. Biosci. 50, 251 (1980).

[20] A. Ruhe and P.A. Wedin, Algorithms for separable nonlinear least squares problems, SIAM Rev. 22, 318 (1980).

[21] H. Stark, Sampling theorems in polar coordinates, J. Opt. Soc. Amer. 69, 1519 (1979).

[22] O.N. Strand and E.R. Westwater, On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature, J. Assoc. Comput. Mach. 15, 100 (1968).

[23] P.N.T. Unwin and R. Henderson, Molecular structure determination by electron microscopy of unstained crystalline specimens, J. Mol. Biol. 94, 425 (1975).

[24] R.H. Vogel, SPLMOD Users Manual, EMBL Technical Report DA06, European Molecular Biology Laboratory, Heidelberg, 1983.

[25] R.W. Wijnaendts van Resandt, R.H. Vogel, and S.W. Provencher, Double beam fluorescence lifetime spectrometer with subnanosecond resolution: Application to aqueous tryptophan, Rev. Sci. Instr. 53, 1392 (1982).