Refinement of Underdetermined Crystal Structures

Lynn F. Ten Eyck [*†]
San Diego Supercomputer Center
P.O. Box 85608, San Diego, California 92186-5608 USA
teneyckl@sdsc.edu

Abstract

The normal matrix for full matrix least squares contains a wealth of information about the behavior of the problem. In particular, the eigenvalues and eigenvectors contain all of the information concerning precision of parameter determination and correlation of parameters. This information is valid even if the refinement is underdetermined. Full-matrix analysis is likely to prove valuable for low resolution work as well as for high resolution work.

1 Introduction

Refinement of macromolecular structures as a mathematical problem is not different from refinement of small-molecule structures. Both are straightforward optimization problems. The difficulties arise because the macromolecular crystallographer rarely has sufficient data to answer questions at the same level of detail as the small molecule crystallographer. Nevertheless, he has a lot of data, and the temptation to over-interpret it is sometimes overwhelming. As a personal note, I began work on refinement of protein structures when I realized that despite having hundreds of thousands of observations I was unable to say with any certainty whether the heme group in deoxyhemoglobin was significantly domed. A quarter of a century later it is still not possible to put a direct measure of accuracy on estimates of the heme geometry in hemoglobin.

Structural questions about macromolecules can be posed on several levels, ranging from "What is the fold?" to "Is one of the bonds in the iron-sulfur cluster significantly different from the others?" The level of detail is widely variable, and structures which are adequate for the former purpose may not be adequate for the latter. Any of us who practice our craft for any length of time will see some beautiful-looking maps, which lead to structures in which we have high confidence - but we cannot put reliable numbers on that confidence. Similarly, we will see maps which are charitably described as obscure, where we can (perhaps) build the chain, but cannot be positive that the density we see is not a phase artifact. Sometimes we see both kinds of density in the same map. This presents us with a real problem. The accuracy of the structures we report is not uniform, but the methods for characterizing this information and reporting it to the users of the coordinates are very poor indeed.

There are a number of common practices in refinement of macromolecular structures which cause serious problems with the accuracy of the final structure. Some of these are omission of weak data, omission of low resolution data, improper treatment of solvent, and improper treatment of non-crystallographic symmetry restraints. Kleywegt and Jones [1, 2] discuss some of the problems that arise from these practices.

Many of these practices arise from the confusion of two distinct problems. The first problem is to solve the structure, which means to find the correct model. In this stage of the problem it is often highly appropriate to leave out weak data and concentrate on the strongest signal. It can also be appropriate to alter weights on restraints, relax non-crystallographic symmetry restraints, and generally let the molecule distort in order to fall into the best minimum. Once the model is determined, there is the second problem of finding the best values for the parameters of the model. This is a quite different problem and requires different treatment of the data. Much confusion arises because both problems are generally handled by the same software, and superficially appear the same.

The method of analysis presented here is directed at two problems. The first problem is to derive a reliable method of estimating the uncertainty of each individual parameter, which works for all resolutions and for all forms of parameterizing or restraining the model. It is shown below how to ascertain which parameters are determined precisely and which are not, by methods which are not limited by low resolution data. It is also shown how to determine the effect of different ways of parameterizing the problem on the accuracy of the parameters. Practical analysis according to these methods is not complete, but the results are almost certain to be pessimistic. In the words of Ecclesiastes 1:18, *For in*

[*]Supported in part by grant BIR 9223760 from the National Science Foundation.

[†]This manuscript is an expanded version of one prepared for the CCP4 study weekend on Macromolecular Refinement, January 4-6, 1996

much wisdom is much grief: and he that increaseth knowledge increaseth sorrow.

Another problem which can be addressed by the methods presented in this paper is to determine how the results of the crystallographic experiment can be improved. There are many open questions as to the "best practice" in any experimental field. For example, there are widely varying practices in the use of low resolution data, inclusion of weak reflections in refinement calculations, incorporation of non-crystallographic symmetry restraints, and the tradeoff between completeness and resolution in data collection. There are significantly different conceptual and mathematical descriptions of the models being refined. The mathematical and computational apparatus discussed in this paper provides a rigorous method for analysis of these questions. It appears that it may be possible to use these methods to determine optimal data collection protocols for answering specific questions about a particular structure at higher resolution, and will in general tell what must be done to achieve a specified level of accuracy in a structure determination. Due to space limitations this application will be given very short treatment.

2 Theory of least squares

The theory of least-squares analysis of poorly determined systems is well advanced mathematically, but seldom used extensively in practice. The books by Lawson and Hanson [3], and by Golub and Van Loan [4] are highly recommended to the reader. Excellent material is also found in Diamond's discussion of real-space refinement [5]. The following derivations are completely general for all least-squares problems, linear and non-linear. To avoid severe notational complexity, the specific language of the crystallographic problem is deferred until necessary.

The general problem of fitting a non-linear model function to a set of observations can be written as a minimization of

$$\Phi(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{N} w_j^2 \left(f_j(\mathbf{x}) - y_j\right)^2 \tag{1}$$

where $\Phi(\mathbf{x})$ is the sum of squares of residuals, y_j is an observed value, w_j is a weighting factor based on the reliability of y_j, and $f_j(\mathbf{x})$ is the function which calculates the theoretical value of the observable quantity given the parameters $\mathbf{x}$ and the index j which specifies the conditions of the observation. There are a variety of methods for finding the parameters $\mathbf{x}$ which minimize $\Phi(\mathbf{x})$. The commonly used methods for the macromolecular crystallographic problem are simulated annealing [6], conjugate gradients applied directly to the non-linear function itself [7], and conjugate gradients applied to the linear approximation to $\Phi(\mathbf{x})$ [8, 9]. Refinement of parameters in small-molecule crystallography is normally done by directly solving successive linear approximations to $\Phi(\mathbf{x})$, a method known as full-matrix least squares [10, 11]. All of these methods work, some faster than others. Generally speaking, simulated annealing has the largest radius of convergence, conjugate gradients applied to the non-linear function (especially as modified by Tronrud [12]) is the fastest, and full-matrix least squares is the most accurate.

The simplest description of the linear approximation is to expand $\Phi(\mathbf{x})$ as a Taylor series about the minimum point $\phi_0 = \Phi(\mathbf{x}_0)$, where $\mathbf{x_0}$ is the set of parameters which minimize $\Phi(\mathbf{x})$. The expansion is

$$\begin{aligned}\Phi(\mathbf{x}) \quad \approx \quad & \phi_0 + \left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial \Phi}{\partial x_i}\right)_{\mathbf{x}_0} \right\rangle \\ & + \frac{1}{2}\left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial^2 \Phi}{\partial x_i x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle \\ & + \ldots \end{aligned} \tag{2}$$

where the Dirac bra-ket notation expresses a column vector as $|x_i\rangle$, a row vector as $\langle x_i|$, and a matrix as $|x_{ij}|$. $\langle x|y\rangle$ is thus the inner product of the vectors x and y. Since the expansion is about a minimum, the gradient at $\mathbf{x}_0$ vanishes for all parameters x_i. Thus we have the approximation that (to second order)

$$\Phi(\mathbf{x}) \approx \phi_0 + \frac{1}{2}\left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial^2 \Phi}{\partial x_i x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle \tag{3}$$

and, by differentiation,

$$\left| \left(\frac{\partial \Phi}{\partial x_i}\right)_{\mathbf{x}} \right\rangle \approx \left| \left(\frac{\partial^2 \Phi}{\partial x_i x_j}\right)_{\mathbf{x}_0} \right| (\mathbf{x}-\mathbf{x}_0) \Big\rangle . \tag{4}$$

Given the first and second derivatives of $\Phi(\mathbf{x})$, Equation (4) can be solved for the correction to $\mathbf{x}$ which brings it closer to $\mathbf{x}_0$. The approximation is the use of second derivatives evaluated at $\mathbf{x}$ instead of $\mathbf{x}_0$, and the neglect of higher order terms in the Taylor series. Neither condition is a problem for parameter estimates close to $\mathbf{x}_0$. Note that the assumption that the function can be approximated locally by a quadratic polynomial is equivalent to assuming that the matrix of second derivatives is constant.

An alternative is to expand the residuals in terms of the parameter shifts. In this formulation each weighted observation is expanded in a Taylor series as

$$w_j y_j = w_j f_j(\mathbf{x}) + w_j \left\langle \left(\frac{\partial f_j(\mathbf{x})}{\partial x_i}\right)_{\mathbf{x}} \middle| (\mathbf{x}_0 - \mathbf{x}) \right\rangle \tag{5}$$

where w_j is the weight associated with observation y_j. Writing the system of equations (5) as a matrix equation we have

$$\mathbf{A}\,(\mathbf{x}-\mathbf{x}_0) = \mathbf{r} \tag{6}$$

where $\mathbf{A}$ has m rows and n columns, $\mathbf{x}$ is a column vector of length n, and $\mathbf{r}$ is a column vector of length m, with elements $w_j\left(f_j(\mathbf{x})-y_j\right)$. $\Phi(\mathbf{x})$ is given by

$$\begin{aligned}\Phi(\mathbf{x}) &= \frac{1}{2}\langle r\ |r\rangle \\ &= \frac{1}{2}\mathbf{r}^T\mathbf{r} \\ &= \frac{1}{2}\left(\mathbf{x}-\mathbf{x}_0\right)^T\mathbf{A}^T\mathbf{A}\left(\mathbf{x}-\mathbf{x}_0\right) \qquad (7)\end{aligned}$$

where the superscript T denotes the transpose of a matrix or vector. It is well known [3] that the solution to (5) which minimizes $\|\mathbf{A}\left(\mathbf{x}-\mathbf{x}_0\right)-\mathbf{r}\|$ (and hence minimizes $\Phi(\mathbf{x})$) is the solution to the $n\times n$ matrix equation

$$\mathbf{A}^T\mathbf{A}\left(\mathbf{x}-\mathbf{x}_0\right)=\mathbf{A}^T\mathbf{r} \qquad (8)$$

The equivalence of the two approaches is readily demonstrated by expanding the terms in the two formulations. The elements g_i of $\mathbf{g}=\mathbf{A}^T\mathbf{r}$ and h_{ij} of $\mathbf{H}=\mathbf{A}^T\mathbf{A}$ are given by

$$g_i = \sum_{k=1}^{m} w_k^2\left(f_k\left(\mathbf{x}\right)-y_k\right)\left(\frac{\partial f_k\left(\mathbf{x}\right)}{\partial x_i}\right) \qquad (9)$$

$$h_{ij} = \sum_{k=1}^{m} w_k^2\left(\frac{\partial f_k\left(\mathbf{x}\right)}{\partial x_i}\right)\left(\frac{\partial f_k\left(\mathbf{x}\right)}{\partial x_j}\right). \qquad (10)$$

Differentiation of Equation (1) gives

$$\left(\frac{\partial\Phi\left(\mathbf{x}\right)}{\partial x_i}\right) = \sum_{k=1}^{m} w_k^2\left(f_k\left(\mathbf{x}\right)-y_k\right)\left(\frac{\partial f_k\left(\mathbf{x}\right)}{\partial x_i}\right) \qquad (11)$$

$$\begin{aligned}\left(\frac{\partial^2\Phi\left(\mathbf{x}\right)}{\partial x_i\partial x_j}\right) = &\sum_{k=1}^{m} w_k^2\left(\frac{\partial f_k\left(\mathbf{x}\right)}{\partial x_i}\right)\left(\frac{\partial f_k\left(\mathbf{x}\right)}{\partial x_j}\right) \\ &+ \sum_{k=1}^{m} w_k^2\left(f_k\left(\mathbf{x}\right)-y_k\right)\left(\frac{\partial^2 f_k\left(\mathbf{x}\right)}{\partial x_i\partial x_j}\right) \qquad (12)\end{aligned}$$

The second derivative term on the right hand side of Equation (12) which is not found in Equation (10) vanishes as $\mathbf{x}\rightarrow\mathbf{x}_0$. Equations (4) and (8) thus converge to the same form.

The matrix equation

$$\mathbf{H}\left(\mathbf{x}-\mathbf{x}_0\right)=\mathbf{g} \qquad (13)$$

is the set of *normal equations* for the least squares problem. Since a protein refinement can easily have 10^4 parameters, the size of the normal matrices can become very large. Full matrix least squares is not normally applied to large proteins.

The normal equations can be solved by inverting $\mathbf{H}$,

$$\mathbf{H}^{-1}\mathbf{g}=\mathbf{H}^{-1}\mathbf{H}\left(\mathbf{x}-\mathbf{x}_0\right)=\left(\mathbf{x}-\mathbf{x}_0\right) \qquad (14)$$

$\mathbf{S}=\mathbf{H}^{-1}$ is the covariance matrix times the mean square residual, which means that, after scaling, the elements of $\mathbf{S}$ are

$$s_{ij}=c_{ij}\sigma_i\sigma_j \qquad (15)$$

where c_{ij} is the correlation coefficient between parameters i and j, and σ_i is the standard deviation of parameter i. Since the correlation of any parameter with itself is 1, the diagonal elements are the variances of the parameters determined by solving the normal equations. *The inverse of the normal matrix is the source of the detailed accuracy information from traditional small-molecule least squares analysis of X-ray diffraction data.*

The matrix of correlation coefficients is often used to detect dependencies between variables in a least-squares problem. Values of $|c_{ij}|$ close to 1 indicate dependencies. However, this is limited to the detection of pairwise dependencies. Higher order dependencies do not necessarily have pairwise components. Lawson and Hanson [3, page 72] give a 3×3 example of strongly interdependent variables in which the magnitude of the largest correlation is 0.49.

If there are insufficient observations to explicitly determine all parameters, the matrix $\mathbf{H}$ becomes singular and the inverse matrix is not defined. For crystallography this occurs if the resolution is low, which is a common case for macromolecules. All of the preceding analysis concerning the Taylor series expansions and normal equations is still valid up through Equation (13). Methods for minimizing $\Phi(\mathbf{x})$ which do not depend on inverting the matrix $\mathbf{H}$ (such as simulated annealing or conjugate gradients) will still find a minimum, but in a formal sense the variance of some of the parameters will be infinite. The minimum will not be unique.

Even singular normal equations can be solved by diagonalizing the matrix $\mathbf{H}$. The eigenvalues and eigenvectors of $\mathbf{H}$ are solutions to the matrix equation

$$\mathbf{H}\mathbf{v}=\lambda\mathbf{v} \qquad (16)$$

where λ is an eigenvalue of $\mathbf{H}$ and $\mathbf{v}$ is the corresponding eigenvector of $\mathbf{H}$.

For the case in which $\mathbf{H}$ is a normal matrix for a least squares problem, we have the interesting result from Equation (7) that

$$\Phi\left(\mathbf{x}_0+\mathbf{v}_i\right)=\frac{1}{2}\mathbf{v}_i^T\mathbf{H}\mathbf{v}_i=\frac{1}{2}\lambda_i\mathbf{v}_i^T\mathbf{v}_i=\frac{1}{2}\lambda_i \qquad (17)$$

when $\mathbf{v}_i$ is the i^{th} eigenvector of $\mathbf{H}$ and λ_i is the corresponding eigenvalue. The eigenvectors of $\mathbf{H}$ specify combinations of parameters which are statistically independent

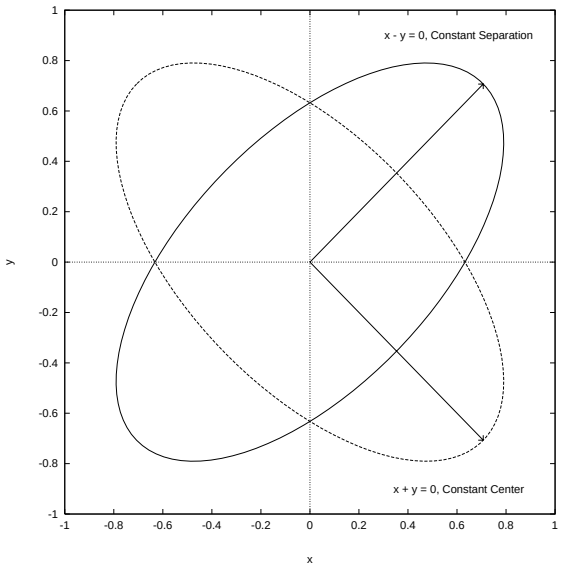

Figure 1: Both ellipses have the same eigenvectors, but the eigenvalues are swapped. The eigenvector in the $(+,+)$ quadrant corresponds to a shift of parameters which preserves the distance between two points. The eigenvector in the $(+,-)$ quadrant corresponds to a shift of parameters which preserves the center of mass of two points. The two ellipses reflect situations in which either the separation is more accurately known than the position, or in which the position is known more accurately than the separation.

of one another, and the eigenvalues are proportional to the reciprocal of the variance of those parameter combinations. Another way of expressing the same idea is that the eigenvectors which correspond to large eigenvalues are directions in which parameter shifts have a large effect on the sum of squares of the residuals, and thus are well determined. Eigenvectors which correspond to small eigenvalues have little effect on the sum of squares of the residuals and thus correspond to poorly determined combinations of parameter shifts. (In fact the reciprocals of the eigenvalues are proportional to the variances of the corresponding combinations of parameters.)

This situation is illustrated for a two-parameter case in Figure 1. The ellipses are contours of constant $\Phi(\mathbf{x})$ in the second order approximation. The principal axes of the ellipses correspond to the variances of the parameters. The short axis of the ellipse gives the direction in which $\Phi(\mathbf{x})$ has the most sharply determined minimum.

3 Poorly Determined Systems

The inverse of a matrix can be constructed from the eigenvectors and eigenvalues of the matrix. If $\mathbf{V}$ is the orthogonal matrix constructed so that the columns of $\mathbf{V}$ are the eigenvectors $\mathbf{v}$ of $\mathbf{H}$, it is easily shown that

$$\mathbf{H}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T \tag{18}$$

where $\mathbf{\Lambda}^{-1}$ is a diagonal matrix containing the reciprocals of the eigenvalues of $\mathbf{H}$. If $\mathbf{H}$ is singular some of the eigenvalues are zero, and $\mathbf{\Lambda}^{-1}$ is not defined. However, if we define the *pseudo-inverse* of $\mathbf{\Lambda}$ as

$$\mathbf{\Lambda}^{+} = \begin{cases} 1/\lambda_i & \lambda_i > 0 \text{ and } i = j \\ 0 & \lambda_i = 0 \text{ or } i \neq j \end{cases} \tag{19}$$

then

$$\mathbf{\Lambda}^{+}\mathbf{\Lambda} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where the matrix product yields an identity matrix of the same rank as $\mathbf{H}$, with the remainder of the product being 0. The equation corresponding to 18 is

$$\mathbf{H}^{+} = \mathbf{V}\mathbf{\Lambda}^{+}\mathbf{V}^{T}. \tag{20}$$

The elements of $\mathbf{H}^{+}$ contain the same correlation information and variance information as the elements of $\mathbf{H}^{-1}$, except that it applies only to the parameter combinations which are in fact still determined by the data. The pseudo-inverse is identical to the inverse if the matrix $\mathbf{H}$ is of full rank.

This apparatus provides a complete mechanism for determining which parameters of a model are actually determined by the least squares procedure. It also gives direct measures of the precision of the determinations of the parameters for those parameters which are actually derived from the data. Preliminary calculations on two small molecules and a protein have shown that even singular crystallographic systems contain a large number of large eigenvalues, and hence many accurately determined parameters.

4 Restraints and Diffraction Data

There are several different methods for applying restraints, and there are different degrees of approximation that can be used in computing the elements of the normal matrices. It is important that the effects of these different approaches be understood precisely. When the restraints are put into the least-squares calculation as additional observations to be fit, the matrix $\mathbf{A}$ of Equation (6) can be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \text{ and } \mathbf{A}^T = \begin{bmatrix} \mathbf{A}_1^T \mathbf{A}_2^T \end{bmatrix} \tag{21}$$

where all of the experimental observational equations are in $\mathbf{A}_1$ and all of the restraint equations are in $\mathbf{A}_2$. If we attach an explicit scale factor K_r to the equations of restraint, the matrix $\mathbf{H}$ becomes

$$\begin{aligned}\mathbf{H} &= \mathbf{A}^T\mathbf{A} \\ &= \mathbf{A}_1^T\mathbf{A}_1 + K_r^2\mathbf{A}_2^T\mathbf{A}_2 \\ &= \mathbf{H}_1 + K_r^2\mathbf{H}_2 \end{aligned} \quad (22)$$

This directly separates the contributions of the two portions of the problem to the solution and will at long last clarify the effects of different restraint schemes on the results of a crystal structure refinement.

Construction of $\mathbf{H}$ from Equation (22) has the advantage that the quality of the parameters and the goodness of fit can be studied as a function of K_r to determine the correct relative weight to assign to the restraints. The benefit of this approach over Brünger's [13] is that all of the data can be used while still avoiding overfitting. Brünger's cross-validation approach requires that a fraction of the data not be used in the refinement so that it can be used as an objective check on the progress of the refinement and on the validity of changes in parameters. In cases which are poorly determined it is not desirable to give up a fraction of the data if it can be avoided. (It should be noted that cross validation is good for testing other things besides K_r, such as the validity of basic changes in the model. It is not yet clear whether the methods being developed here could replace $R_{\rm free}$ for those purposes.)

5 Data Collection Protocol Analysis

Substitution of crystallographic variables into Equations (7) and (14) gives

$$h_{ij} = \sum_{hkl} w_{hkl} \left(\frac{\partial \left| \mathbf{F}_{hkl}^c \right|}{\partial x_i} \right) \left(\frac{\partial \left| \mathbf{F}_{hkl}^c \right|}{\partial x_j} \right) \quad (23)$$

which shows that the normal matrix *does not depend directly on the observed data*. The normal matrix depends on the model, the set of observations which are included in the calculation, and the statistical weight assigned to each observation, but does not depend on the values of the observations. The values of the parameters of the model do depend on the data, and this does affect the values of the elements of the normal matrix.

It is thus possible, given a model, to evaluate the effect of different data collection protocols on the accuracy with which the parameters will be determined. This formalism will decisively answer the question as to whether the omission of data observed at less than 2σ harms the accuracy of the model (it does), and settle the wars concerning the inclusion or omission of data inside the 6Å sphere during refinement. It will also tell specifically how the collection of additional data will improve the accuracy of the model, subject to the assumption that the model does not change dramatically in view of the new data. For example, just how good does your data have to get before you can tell if one bond in your iron-sulfur cluster is significantly different from the others? Are you better off getting more data, or improving the accuracy of the data you already have?

6 Sample Results

Exploratory calculations have been done on two systems using SHELXL-93 [14] to compute the matrices on the Cray C90 at the San Diego Supercomputer Center and on a Digital Equipment Corporation Alphastation 400 4/233. The first case examined was a small lactone, $\mathrm{C_7H_{11}NO_3}$, the data for which are distributed with the SHELXL-93 program as a test case. The data extend to 0.8 Å resolution. The model contains coordinates and anisotropic thermal parameters for each heavy atom; the hydrogen atoms ride at calculated positions with one exception, which has a free bond rotation parameter. In all, there are 105 parameters in the lactone model.

The second test case is a small protein, amicyanin [15, 16], which contains 105 amino acid residues and one copper atom. The model also contains 88 solvent molecules, for a total of 896 atoms and 3,856 parameters when using isotropic thermal parameters. (Two parameters are used for scaling.) The data extend to 1.07 Å resolution and are kindly provided by Professors F. Scott Mathews and N.-h. Xuong. The current model used in these preliminary calculations is not fully refined.

6.1 Small Molecule Test Case

The lactone refinement was done as described in the SHELXL-93 manual. Refinement is on $|\mathbf{F}|^2$, hydrogen atoms were given a riding model, and all heavy atoms were refined with anisotropic thermal paramters. The program was modified so that the matrix was written to disk before Marquadt damping was applied. The program ran in seconds on both computers, which is not surprising because it also ran in seconds for this case on an IBM PC.

Some results for the lactone are presented in the following figures. Figures 2 and 3 introduce the use of the power represented by a particular eigenvector component. The matrix $\mathbf{V}$ which contains the eigenvectors as columns has the property that

$$\sum_{k=1}^{N} v_{ik}^2 = 1$$

and partial sums of squares across the rows of $\mathbf{V}$ measure the projection of the parameter i into the subspace spanned by the corresponding columns of $\mathbf{V}$. Figures 2 and 3 show

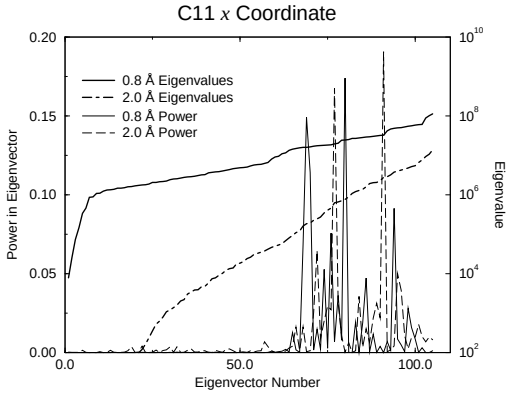

Figure 2: Contribution of each eigenvector to the x coordinate of atom $C11$ at 0.8 Å and 2.0 Å resolution. The eigenvalue spectra are shown on the same graph. At 2.0 Å the first 20 eigenvalues are zero, but the coordinate is still strongly determined by the data.

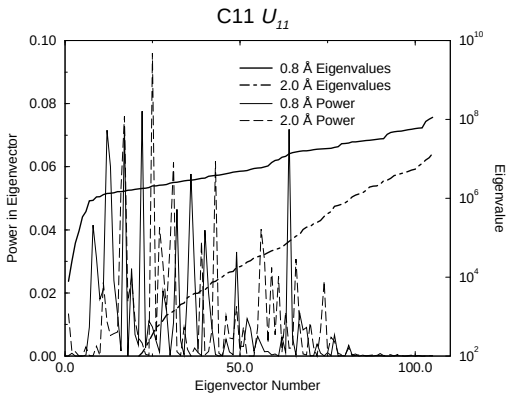

Figure 3: Contribution of each eigenvector to the U_{11} parameter of atom $C11$ at 0.8 Å and 2.0 Å resolution. This parameter is well determined at 0.8 Å resolution because the smallest eigenvalue contributing to the solution is greater than 10^6. At 2.0 Å resolution the parameter is essentially undetermined.

that the quality of the parameters does not decrease uniformly. Significant information persists at low resolution. These figures are also completely consistent with small-molecule refinement experience in that anisotropic thermal parameters cannot be refined at low resolution, and that they are clearly more strongly correlated with other parameters than the spatial coordinates are.

Figures 4 and 5 show the effect of resolution on the eigenvectors of the lactone. The high resolution vectors in Figure 4 show a small amount of correlation. The peak in vector 100 near 50 on the scale is the x, y, and z coordinates of atom N6, which is obviously making a concerted movement. The peaks at parameters 90-95 in vector 7 are the U_{ij} parameters of atom C10. Finally, the peak at 96 in vector 6 corresponds to the torsion angle of hydrogen H10.

The low resolution eigenvectors shown in Figure 5 are much more diffuse. Each eigenvector carries information about a number of parameters – but the number is not always large. For example, vector 100, which corresponds to an eigenvalue of 5.5×10^6, has four significant peaks, most of which span several adjacent parameters. These correspond to the x, y, and z coordinates of atoms O1, O2, and N6; and to the z coordinate of atom O3.

6.2 Macromolecular Test Case

Analysis of the amicyanin normal matrix is far more complex. Normal matrices were calculated for this protein at resolutions of 1.07, 2.0, 2.5, 3.0, and 3.5 Å. Refinement was done using $|\mathbf{F}|^2$ instead of the more conventional $|\mathbf{F}|$ because testing the differences between these two formulations of the refinement problem is part of the point of this proposal. Two refinements were calculated at each resolution – one which used the standard stereochemical restraints produced by the PDBINS program which is part of the SHELXL-93 distribution, and one which used *no* stereochemical restraints. Subtraction of the unrestrained normal matrix from the restrained matrix will give the matrix $\mathbf{H}_2$ of Equation 22 appropriate for this set of restraints, although this has not yet been done. The model was refined with isotropic thermal parameters.

Figures 6 and 7 show the eigenvalue spectra for the ten refinements. The restrained refinement contains no surprises; the matrices are non-singular at resolutions better than 3.0 Å. The unrestrained refinements shown in Figure 7 shows that even at 3.5 Å there are 657 eigenvalues greater than 10^5 and hence significant information determined by the X-ray diffraction data. The situation is made much clearer by Figure 8. For this figure thresholds of 10^6 and 10^3 were chosen for well determined and poorly determined parameters. The squares of the components of each parameter were summed in the subspaces determined by $\lambda > 10^6$ and $\lambda < 10^3$. The parameters are sorted so that the first two are

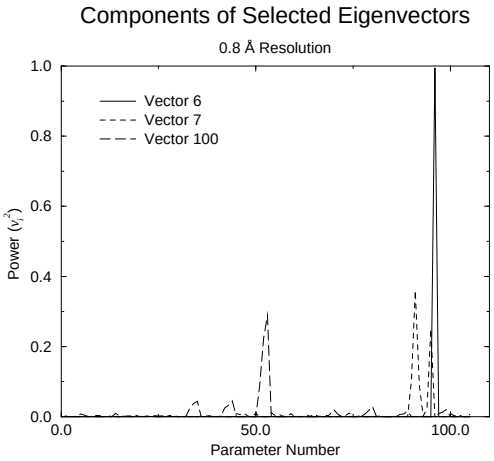

Figure 4: Selected eigenvectors for the lactone test case at 0.8 Å resolution. Each eigenvector has only a few significant components, showing very little correlation between these parameters at high resolution.

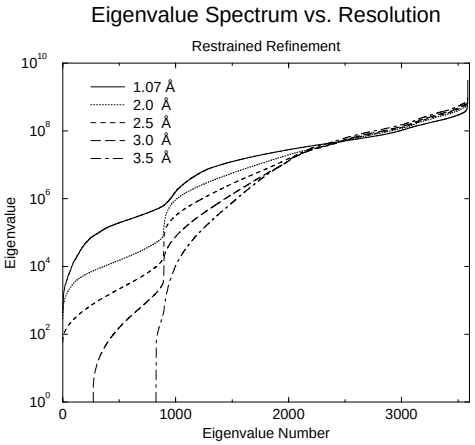

Figure 6: Eigenvalue spectra for restrained refinements of amicyanin show that the system becomes singular between 2.5 and 3.0 Å resolution. The pronounced “knee” in the spectra near 900 is due to the 896 thermal parameters.

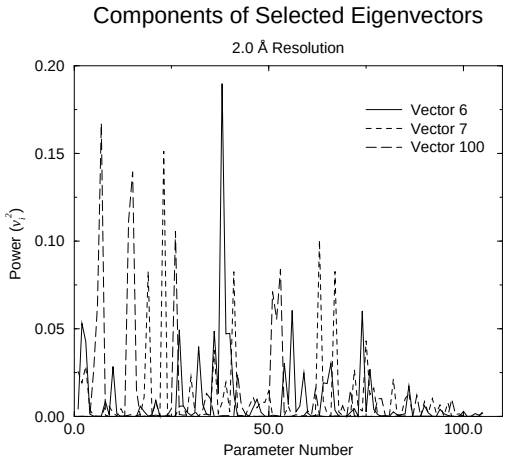

Figure 5: Selected eigenvectors for the lactone test case at 2.0 Å resolution. Note that the vertical scale is different from that in Figure 4. These eigenvectors show a great deal of multi-parameter correlation.

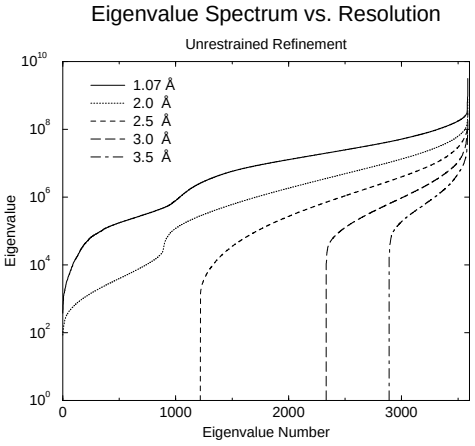

Figure 7: Eigenvalue spectra for unrestrained refinements of amicyanin show the importance of restraints in maintaining the conditioning of the refinement problem. The matrices become singular between 2.0 and 2.5 Å resolution.

the scale factors, the next 896 are the thermal parameters, and the remaining parameters are the spatial coordinates. In each case the parameters for the 88 solvent molecules are to the right. The well determined set is drawn in red; the poorly determined set is drawn in black. High values in red indicate well determined parameters, while high values in black indicate poorly determined parameters.

The first thing to note is the power of the method for picking out sections of suspect structure. The 1.07 Å restrained refinement contains a number of red spikes which clearly indicate sections which are not as well determined as the rest of the structure. Since this refinement is not complete, it is not yet determined whether this reflects disorder or error in the model. The spike in the black curve indicates a thermal parameter which is unstable on refinement. As the resolution decreases the behavior of the thermal parameters and the level of accuracy of the coordinates shows behavior consistent with current experience, except that the thermal parameters are perhaps more poorly determined at 2.5 Å than an optimist might hope. The vertical black bar close to 900 indicates that the thermal parameters for the solvent atoms are almost uniformly dubious at 2.0 Å.

Comparison of the restrained and unrestrained columns is particularly enlightening. Many non-solvent coordinates are well-determined at 3.5 Å in the restrained refinement, but are not determined at all without restraints. This shows how much of the information in the final structures is coming from the restraints as a function of resolution. This feature is further highlighted by the poor quality of the solvent coordinates in the restrained refinements; the only restraints which help them are the anti-bumping restraints. A particularly noteworthy feature is that the quality of the solvent coordinates in the 2.5 Å restrained structure is better than the quality of the same coordinates in the unrestrained structure even though there are few restraints on the solvent molecules. This indicates that the restraints help the overall structure, not just the portions restrained.

The calculation of the eigenvalues and vectors of the matrices for this problem required 3 hours on a DEC Alpha, or 20 minutes on a Cray C90. The calculations are clearly feasible on large workstations as well as on supercomputers.

References

[1] Gerard J. Kleywegt and T. Alwyn Jones, "Where freedom is given, liberties are taken," *Structure*, vol. 3, no. 6, pp. 535–540, 1995, Polemic on proper refinement procedures – very good.

[2] Gerard J. Kleywegt and T. Alwyn Jones, "Good model-building and refinement practice," in *Macromolecular Refinement*, R. M. Sweet and C. W. Carter Jr., Eds., Methods in Enzymology. Academic Press, Orlando, in press.

[3] Charles L. Lawson and Richard J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[4] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1989.

[5] R. Diamond, "A real-space refinement procedure for proteins," *Acta Cryst.*, vol. A27, pp. 436–452, 1971, asb.

[6] A.T. Brünger, J. Kuriyan, and M. Karplus, "Crystallographic R-factor refinement by molecular dynamics," *Science*, vol. 235, pp. 458–460, 1987, asb.

[7] D. E. Tronrud, L. F. Ten Eyck, and B. W. Matthews, "An efficient general-purpose least-squares refinement program for macromolecular structures," *Acta Cryst.*, vol. A43, pp. 489–501, 1987.

[8] John H. Konnert, "A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units," *Acta Cryst.*, vol. A32, pp. 614–617, 1976, asb.

[9] J. H. Konnert and W. A. Hendrickson, "A restrained-parameter thermal-factor refinement procedure," *Acta Cryst.*, vol. A36, pp. 344–350, 1980, asb.

[10] G. H. Stout and L. H. Jensen, *X-ray Structure Determination: A Practical Guide*, John Wiley and Sons, New York, 1989.

[11] G. M. Sheldrick, *SHELXL-93, a Program for the Refinement of Crystal Structures from Diffraction Data*, Institut fuer Anorg. Chemie, Goettingen, Germany, 1993.

[12] D.E. Tronrud, "Conjugate direction minimization - an improved method for the refinement of macromolecules," *Acta Crystallographica*, vol. A48, pp. 912–916, 1992.

[13] Axel T. Brünger, "Free R-value - a novel statistical quantity for assessing the accuracy of crystal structures.," *Nature*, vol. 355, no. Jan. 30, pp. 472–475, 1992.

[14] G. M. Sheldrick and T. R. Schneider, "SHELXL: High-resolution refinement," in *Methods in Enzymology*, R. M. Sweet and C. W. Carter Jr., Eds. Academic Press, Orlando, in press.

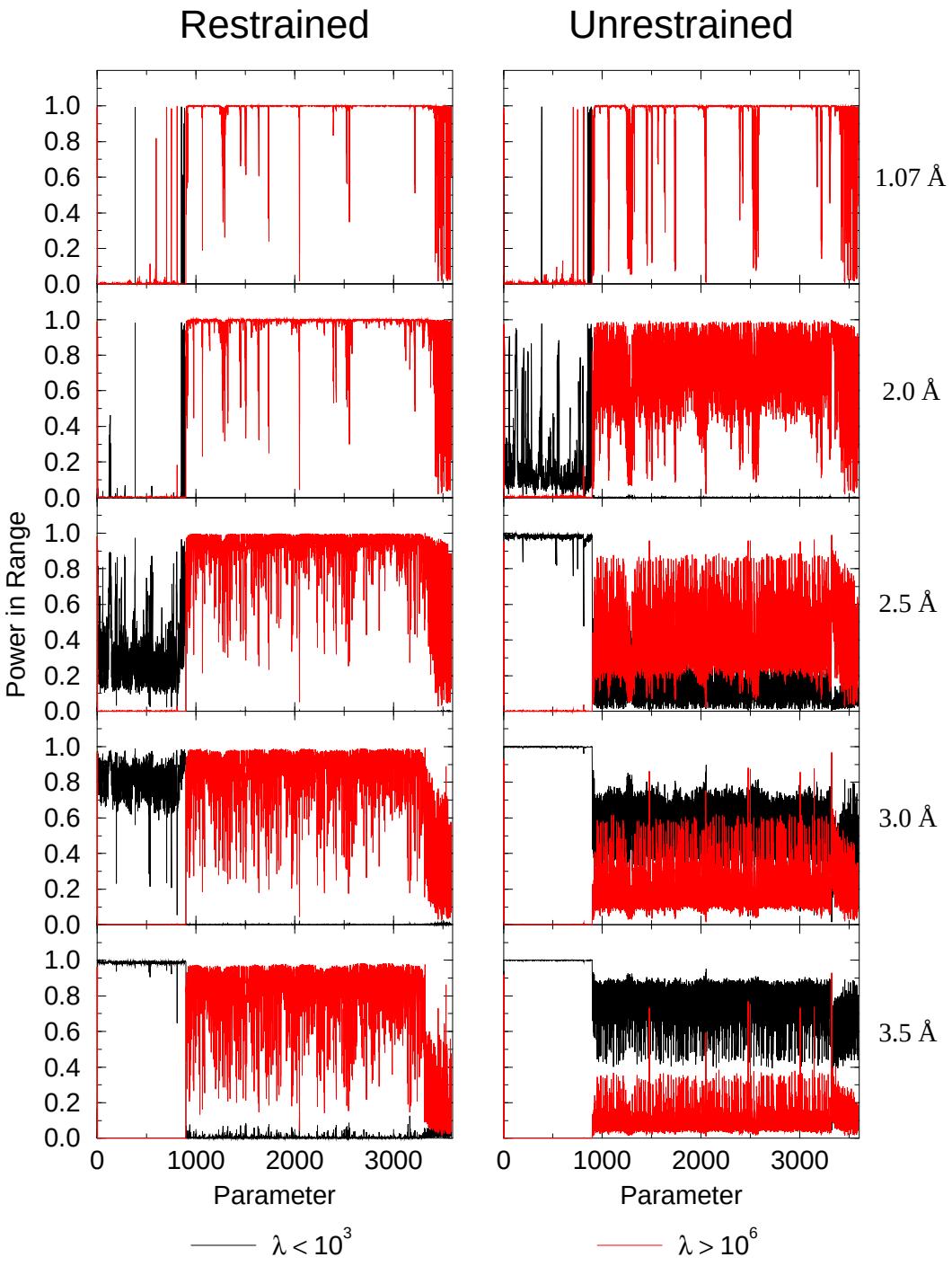

Figure 8: Fraction of the information for each parameter contained in eigenvectors either greater than 10^6 or less than 10^3 for the protein amicyanin at resolutions from 3.5 Å to 1.07 Å. The former are very well determined; the latter are badly determined. The ten graphs show the effects of resolution and of geometric restraints on the precision of the paramter determinations.

[15] R. Durley, L. Chen, L. W. Lim, F. S. Mathews, and V. L. Davidson, "Crystal structure analysis of amicyanin and apoamicyanin from paracoccus denitrificans at 2.0 ångstroms and 1.8 ångstroms resolution," *Protein Science*, vol. 2, pp. 739–752, 1993.

[16] Longyin Chen, Rosemary C. Durley, F. Scott Mathews, and Victor L. Davidson, "Structure of an electron transfer complex: Methylamine dehydrogenase, amicyanin, and cytochrome c551i.," *Science*, vol. 264, pp. 86–90, 1994.