MODELING THE PROPERTIES OF SILICATES

by

Kurt Lane Bartelmehs

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Geological Sciences

APPROVED:

G.V. Gibbs, Co-Chairman

M.B. Boisen, Jr., Co-Chairman

J.B. Birch

F.D. Bloss

S.B. Desu

MODELING THE PROPERTIES OF SILICATES

by

Kurt Lane Bartelmehs

Abstract

Assuming a simple force field involving only short range non-Coulombic molecular energy terms along with P1 symmetry, a variation of the SQLOO model (Boisen and Gibbs, 1993) successfully generates the structure types of both α and β quartz along with at least five alternative structure types of silica not yet observed to our knowledge. These structure types are identified by the existence of symmetry elements represented in the optimized atomic coordinates and cell parameters that define a minimizer in the model. A family of minimizers is discovered through the combined use of Monte Carlo simulated annealing followed by quasi–Newton minimization techniques. The results are in contrast with the assertion made by Tse et al. (1992) that new structure types of SiO_2 can only be arrived at by Molecular Dynamic methods. By varying the parameters used in the minimization process, different families of structure types are discovered. Several structure types were found to have high symmetries. These results are in contrast with the findings by Kramer et al. (1991) that the stability of high symmetry structures of silica are stabilized in part by ionicity. The results reported here are for calculations involving $Z = 3$ and 6 formula units. This strategy may be useful in the prediction of possible high silica zeolite structure types.

An examination of the atomic displacement parameters (ADPs) obtained for TO_4 tetrahedra (T = Si, Al) suggest rigid TO bonds are more common in non-framework than in framework silicates. Correlated motion is found among the ADPs that is consistent with TLS rigid body motion. For these data, the translational motion is represented by the ADPs of the central T

atom while both the librational and translational motion is contained in those of the surrounding O atoms. The libration angle for rigid tetrahedra is linearly dependent on the difference between the isotropic equivalent displacement parameter of the T and O atoms, B(T) and B(O), respectively. The value of B(O) is on average twice that of B(T) with a maximum value of $\sim$ 2.0Å^2. Variations in the SiO bond lengths of rigid tetrahedra in the silica polymorphs is related only to $f_s(O)$. Rigid TO and OO bonds are a necessary but not sufficient condition for rigid body motion. Nonrigid tetrahedra may represent crystals containing disorder or problems with the refinement.

The computer program EXCALIBR (Bloss and Riess, 1973; Bloss, 1981, p. 202) has been rewritten and markedly improved. Like EXCALIBR, EXCALIBR II solves optical extinction data, as determined with a spindle stage, and determines the optic axial angle 2V and the orientation of the crystal's optical indicatrix. EXCALIBR II uses a modification to Joel's equation as a means of obtaining the optic axes of a crystal. Furthermore, EXCALIBR II successfully solves extinction data where one optic axis of a biaxial crystal is 90° to the spindle axis, an orientation that had thwarted its predecessor. EXCALIBR II also accurately determines the optical indicatrix orientation for uniaxial crystals. After solving extinction data for several different wavelengths and/or temperatures, EXCALIBR II calculates the angular change of each optic direction with wavelength and/or temperature along with the error on the angle. Using a simple t-test, it then computes a p-value to aid in the decision as to whether the optical direction truly exhibits dispersion. This is a more valid and sensitive procedure than the χ^2 test used by EXCALIBR, particularly because the covariance in each optic vector's coefficients are taken into consideration and the results are invariant to the vector's orientation.

Acknowledgements

I would like to thank the National Science Foundation for supporting this work with Grant EAR-8803933. Jerry and Monte are especially thanked for writing the proposal to get the grant. I am grateful for their extraordinary efforts in providing me with intellectual, emotional and financial support throughout all the years I have attended Virginia Tech. I want to thank Jerry for inspiring me to become both an educator and a scientist, along with his guidance and encouragement to develop an interdisciplinary background. I am thankful to Monte for all that he has taught me about minimization and modeling and for his never-ending sense of humor. It was an extreme pleasure to work with Dr. Birch whom I am indebted to for teaching superb courses on statistics, along with his undivided attention and friendship during my many office visits. In fact, none of this would have happened had it not been for Dr. Bloss who inspired me to major in geology thirteen years ago. I want to thank Seshu for being one of the first to treat me as a peer. My thanks go out to the Department of Geological Sciences and its' staff for their support and assistance, especially Karen Hunt. I thank Lee Johnson for his nice work on the minimizer MADMAX.

The best part of this educational experience I owe to my best friend Bob Downs. The many exciting experiences and discoveries we have shared in together is what science is all about. I would also like to thank George Lager for his support and friendship during difficult times. I want to thank my dad and grandmother for their total support throughout my education. Most of all, I want to thank my wife Carol for her complete commitment to see our way through the past eight years. Her devotion and willingness to work so hard, along with her never-ending encouragement has provided me with the strength to complete this task. Finally, I want to dedicate this work to my late mother.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Generation of Stable and Meta-stable Structure Types of Silica

Introduction

The computer modeling of silicate structures has become an important topic in many areas of science ranging from mineralogy to materials science. Several different models have successfully reproduced a variety of observables such as crystal structure, volume compressibility, elastic and dielectric properties, and even phonon dispersion curves. (Catlow and Cormack, 1987; Stixrude and Bukowinski, 1988; Chelikowsky et at., 1990; Lazaraev and Migorodsky, 1991; Purton et al., 1993; Boisen and Gibbs, 1993) Originally developed to model phases in the earth's interior, the modeling of silicates has only recently become a powerful tool for the identification and generation of new structure types, including the generation of known zeolite structure types (Deem and Newsam, 1992) and a new high pressure form of silica (Tse et al., 1992).

In a recent study, Boisen and Gibbs (1993) developed the non–Coulombic potential energy function

$$\Delta E = C\Big(\frac{1}{2}[\mathbf{x}-\mathbf{x}_o]^t\mathbf{H}[\mathbf{x}-\mathbf{x}_o]\Big) + \sum_{O\in\mathcal{U}}\sum_{O\notin C} A\Big(e^{-br} - e^{-4b}\Big), \qquad (1)$$

where the constants A, b, and C are empirically determined from the compressibility curve of quartz, $\mathbf{H}$ is a 4×4 Hessian matrix and $\mathbf{x}_o$ is a 4×1 vector of equilibrium internal coordinates both calculated for the molecule $H_6Si_2O_7$ using molecular orbital methods. The diagonal elements of $\mathbf{H}$ contain the force constants used to describe the energy associated with two nonequivalent SiO bonds and $\angle$OSiO and $\angle$SiOSi, along with off–diagonal force constants describing the

R(SiO)·∠SiOSi and the R(SiO)·∠OSiO interactions. In their model, Boisen and Gibbs (1993) converted each ∠SiOSi term into a distance term, L, that measures the span of the angle from one unit along each of the two SiO bonds according to

$$L = \sqrt{2 - 2\cos(\angle \mathrm{SiOSi})}, \tag{2}$$

where the ∠SiOSi term is treated as a quadratic in L. The last term in Equation 1 represents an O·O repulsion term used only for non-codimer oxygen atoms. Collectively, Equation 1 is called the SQLOO model because of the scaled quadratic term which includes the L modeling of ∠SiOSi, used in conjunction with the O·O repulsion term.

The SQLOO model successfully reproduces the compressibility curve of quartz (up to 8 GPa) and cristobalite (up to about 1.5 GPa), along with equal frequencies of both right and left handed quartz as observed in nature (Boisen and Gibbs, 1993). Their model also reproduces the observed negative Poisson ratio of cristobalite, along with the relationship observed between SiO bond length and f_s(O) for coesite. Perhaps the most remarkable result of this study is that when the atomic coordinates and unit cell parameters of these minerals were optimized, starting near the constrained minimum and assuming $P1$ space group symmetry and triclinic cell dimensions, the resulting coordinates and cell parameters were found to exhibit the space group symmetry observed for the three silica structure types. Even though there are 150 degrees of freedom in the $P1$ structure of coesite, the resulting optimized coordinates of the atoms of the structure not only exhibited the observed $C2/c$ space group symmetry, but bond lengths and angles calculated for the structure are in excellent agreement with those observed (Boisen and Gibbs, 1993). Similar results were obtained for quartz and cristobalite using the SQLOO model by Downs (1992).

In the study, each of these structures was optimized using a quasi–Newton minimizer called MADMAX (Boisen and Gibbs, 1993). However, this type of minimizer is limited because it tends to locate local minima close to the original starting values. Because $P1$ space group symmetry can be assumed in the calculations, the phase space represented by Equation 1 can be explored for the existence of other minima. The exploration was carried out using a global minimization technique called Monte Carlo Simulated Annealing. This algorithm, unlike a quasi-Newton minimizer, accepts uphill steps that increase the value of the potential energy function. Uphill steps are allowed so that relatively high energy local minima will be escaped in search of lower energy minima. The atomic coordinates for each minimum located in the potential energy function are analyzed to determine whether the minimum represents a framework structure of SiO_4 tetrahedra. In addition, the atomic coordinates of each minima will be examined for the existence of symmetry elements. Just as X-ray diffraction techniques are used to identify materials, a theoretical X-ray powder diffraction pattern will be generated for each minimum energy structure to help identify similar results. Finally, the minima representing framework structure types for silica will be described.

Monte Carlo Simulated Annealing

In the early days of computer modeling, Metropolis et al. (1955) required an efficient method for calculating the properties of a substance or system composed of interacting molecules or particles at equilibrium. A statistical mechanic treatment (canonical ensemble) required that they compute a several hundred-dimensional integral. As such calculations are very time consuming, they developed a modified form of Monte Carlo simulation to solve the problem. For any system of particles, quantum mechanics suggests that the macroscopic system in

equilibrium is actually the ensemble average of many microscopic systems, each having a different configuration and energy level. Statistical mechanics shows that the overall distribution of microscopic configurations follows the Boltzmann distribution law where the probability for a given configuration r_i, $P(r_i)$, is given as

$$P(r_i) = e^{\frac{-E_i}{kT}}, \tag{3}$$

where E_i represents the energy of configuration i, k is the Boltzmann constant and T is the temperature of the system. Instead of randomly generating a possible configuration and weighting it according to Equation 3 in computing the average energy of the system, Metropolis et al. (1955) selected only random configurations based on the probability computed from Equation 3 and weighted each configuration equally.

Starting with a configuration of particles inside of a box, Metropolis et al. (1955) used an ionic model to calculate the energy, E, associated with that configuration. Then a small change was made in position of one particle and the resulting change in the energy, $\Delta E = E' - E$, was calculated where E' is energy of the new configuration. If $\Delta E < 0$, then the movement of the particle results in a lowering of the Coulombic energy and the new configuration is accepted and used in computing the average energy. On the other hand, if $\Delta E > 0$, then the probability, $P = e^{\frac{-\Delta E}{kT}}$, is compared with a random number, P_r, chosen from the interval between [0, 1]. If $P \geq P_r$, then the new configuration is rejected and the previous one used to compute the average. If $P < P_r$, then the new configuration is accepted and used in the calculation. Randomly displacing each particle in turn many times to compute the average energy allowed them to compute the pressure for a given volume. Using as few as 56 particles in their computer model, Metropolis et al. (1955) were able to accurately predict the equation of state for

a single-phase system.

It was not until 27 years later that Kirkpatrick et al. (1983) recognized a connection between the algorithm devised by Metropolis et al. (1955) and the problem of multivariate or combinatorial optimization. They recognized that the energy function used by Metropolis et al. (1955) can be replaced by any cost function and that a given configuration can be replaced by any discrete set of parameters. Using this procedure, a population of configurations for a given optimization problem can be generated at some effective temperature, T. Using T as a control parameter, Kirkpatrick et al. (1983) found an analogy could be made between the anealing of a material and global optimization or minimization. Their procedure, called simulated annealing, consists of choosing a high value for T in the same units as the cost function, $f(\mathbf{x})$, and a set of starting parameters, $\mathbf{x}_o$. Next, $\mathbf{x}_o$ is modified by a random amount $\Delta\mathbf{x}$ and the quantity,

$$\Delta F = f(\mathbf{x}_o + \Delta\mathbf{x}) - f(\mathbf{x}_o),$$

is computed. If $\Delta F \leq 0$, then the parameter modifications are accepted according to $\mathbf{x}' = \mathbf{x}_o + \Delta\mathbf{x}$. On the other hand, if $\Delta F > 0$, then the probability $P = e^{\frac{-\Delta F}{T}}$ is computed. As before, if $P \geq P_r$, then the modification $\Delta\mathbf{x}$ is rejected so that $\mathbf{x}' = \mathbf{x}_o$, and if $P < P_r$, then it is accepted. With the new set of parameters, $\mathbf{x}'$, a new random modification, $\Delta\mathbf{x}'$, is generated and a new ΔF computed and a decision is again made. This procedure is repeated M times, then T is lowered and the decision making process is repeated.

By applying the Metropolis procedure for high values of T, many acceptable choices for $\mathbf{x}$ can be generated because the condition $P < P_r$ will be often true. The analogy with annealing is made by considering a high T as representing the "melting" of the system. Slowly lowering the value of T during the minimization

process effectively reduces the flexibility in the parameters of the function and represents a "freezing" of the system. The sequence of temperature reductions and the number of parameter rearrangements or steps is called the annealing schedule. Kirkpatrick et al. (1983) point out that this procedure results in the possible location of the global minimizer because at these higher temperatures, the system can jump out of a local minima by accepting $\Delta\mathbf{x}$ that increases the value of the cost function, $f(\mathbf{x})$.

Whereas the global minimization method devised by Kirkpatrick et al.(1983) applies to combinatorial problems involving only discreet variables, Vanderbuilt and Louie (1984) modified the approach for optimization over continuous variables. Instead of blindly making a random change in the variables, they proposed a self-regulation mechanism to control the amount of change made in the variables. They proposed a natural way of making a step size according to

$$\Delta\mathbf{x} = \mathbf{Q} \cdot \mathbf{u},$$

where $\mathbf{Q}$ is a matrix designed to control the step size and $\mathbf{u}$ is a random vector where each element, u_i, is randomly chosen from the interval between $[-\sqrt{3}, \sqrt{3}]$. In the approach, they define the number of steps, M, generated at a value of T as the l^{th} random walk and suggest that $M = 15 \times \nu$, where ν is the number of parameters in the function. By calculating the second central moment matrix, $\mathbf{S}^l$, for the accepted parameter values for a random walk, Vanderbuilt and Louie (1984) suggest obtaining $\mathbf{Q}$ from the Choleski decomposition of $\mathbf{S}^l$. However, since this results in a lower diagonal matrix for $\mathbf{Q}$, the step control for each parameter would vary depending on the order in which they were specified in the problem. An alternative method suggested for the calculation of $\mathbf{Q}$ is

$$Q_{i,i}^{l+1} = \left(\frac{\chi_s}{\beta M} S_{i,i}^l\right)^{-\frac{1}{2}}, \tag{4}$$

where $Q_{i,i}^{l+1}$ is i^{th} diagonal element of $\mathbf{Q}$, χ_s is a growth factor typically chosen to be 3, β is calculated to be 0.11, $S_{i,i}^{l}$ is the i^{th} diagonal element of $\mathbf{S}$ calculated for the l^{th} random walk and $i = 1, \cdots, \nu$. For an annealing schedule beginning at $T = T_o$, Vanderbuilt and Louie (1984) suggest reducing T_o according to

$$T^l = \chi_T^{(l-1)} T_o, \tag{5}$$

where χ_T is the geometric temperature reduction factor chosen by trial and error from the interval [0, 1]. Finally, they suggest that the algorithm stop when

$$\frac{< E > - E_{min}}{< E >} < 0.001, \tag{6}$$

where $< E >$ is the average value of the cost function for the M steps and E_{min} is minimum value of the cost function during the l^{th} random walk segment.

Potential Energy Function

For this study, a modified version of the SQLOO potential energy function will be used. Whereas $\mathbf{x}_o$ in Equation 1 includes two nonequivalent equilibrium internal coordinates for R(SiO_1) and R(SiO_2), here the average of the two is used. Furthermore, an O·O repulsion energy term will be calculated for all pairs of oxygen atoms. Calculations have shown that this modification makes the function more well behaved in the general case. The potential energy function used in this study can then be written as

$$\Delta E = 0.67769[V(\mathrm{SiO}) + V(\mathrm{OSiO}) + V(\mathrm{SiOSi}) + V(\mathrm{SiO \cdot SiOSi})] + V(\mathrm{OO}) \tag{7}$$

where the first term is

$$V(\mathrm{SiO}) = \frac{1}{2} f_R \left[\sum_{\mathrm{Si}\in\mathcal{U}} \sum_{\mathrm{O}\in 4} (\mathrm{R(SiO)\text{-}R(SiO)}_o)^2/2 + \sum_{\mathrm{O}\in\mathcal{U}} \sum_{\mathrm{Si}\in 2} (\mathrm{R(SiO)\text{-}R(SiO)}_o)^2/2 \right],$$

where $f_R = 0.3951$ au/b^2 is the average calculated SiO force constant calculated for $H_6Si_2O_7$, R(SiO) is the SiO bond length (in bohrs, b) and $\mathrm{R(SiO)}_o =$

1.622Å/a_o is the average equilibrium bond length calculated for the molecule. Note that the second summation includes only the four closest O atoms to a given Si atom in the unit cell, while the fourth summation includes the two closest Si atoms to a given O atom in the unit cell. Each element of both summations is divided by two so as to be only counted once in calculating the total energy. The second term is

$$V(\mathrm{OSiO}) = \frac{1}{2} f_\theta \sum_{\mathrm{Si}\in\mathcal{U}} \sum_{\mathrm{O}\in 10} \sum_{j=1}^{45} \lambda(d)(\angle\mathrm{OSiO}_j - \angle\mathrm{OSiO}_o)^2, \tag{8}$$

where $f_\theta = 0.3452$ au/rad^2 is the calculated OSiO force constant, $\angle$OSiO is the OSiO bond angle (in radians, rad) and $\angle\mathrm{OSiO}_o$ is set to the ideal tetrahedral value of 0.955 rad (= 109.47°). The second summation includes only the ten closest O atoms to a given Si atom in the unit cell and the third summation is over the 45 unique $\angle$OSiO. The $\lambda(d)$ term is described below. The third term is

$$V(\mathrm{SiOSi}) = f_L \sum_{\mathrm{O}\in\mathcal{U}} \sum_{\mathrm{Si}}^{3.5\text{Å}} \lambda(d)\Big(L - L_o\Big)^2,$$

where $f_L = \frac{f_\phi}{1+\cos(137.6346°)}$, $f_\phi = 0.10029$ au/rad^2 is the calculated SiOSi force constant, L is the converted $\angle$SiOSi (Equation 2) and L_o is the converted $\angle\mathrm{SiOSi}_o$ which will be varied in the calculations ranging from 2.094 rad (= 120°) to 3.142 rad (= 180°). Note that this term can only equal zero when $\angle$SiOSi is equal to $\angle\mathrm{SiOSi}_o$, so that $\angle\mathrm{SiOSi}_o$ will be referred to as the target angle. The second summation includes all Si atoms within 3.5Å of a given O atom in the unit cell. The next term of the potential is

$$V(\mathrm{SiO}\cdot\mathrm{SiOSi}) = f_{RL} \sum_{\mathrm{O}\in\mathcal{U}} \sum_{\mathrm{Si}\in 2} \Big(\ \mathrm{R(SiO)\text{-}R(SiO)}_o\Big)\lambda(d)\Big(L - L_o\Big),$$

where $f_{RL} = \frac{f_{R\phi}\sqrt{2-2\cos(137.6346°)}}{\sin(137.6346°)}$, $f_{R\phi} = 0.04285$ au/(b·rad) is the average calculated SiO·$\angle$SiOSi force constant, L is the converted $\angle$SiOSi formed by the two

closest Si atoms to a given O atom. The final term in the potential energy function is

$$V(\mathrm{OO}) = \sum_{\mathrm{O}\in\mathcal{U}} A(e^{-br} - e^{-4b})/2,$$

where $A = 5.9073 \times 10^4$au, $b = 7.6919\text{Å}^{-1}$ and r is the separation between a given pair of O atoms (in bohrs).

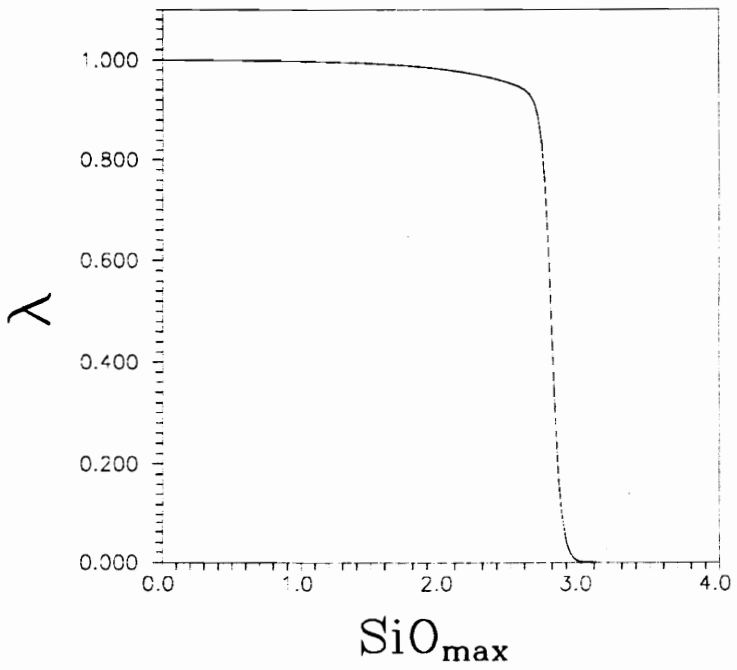


Figure 1–1. A plot showing how the angle factor, λ, varies with the longest SiO distance, $R(SiO)_{max}$, involved in the angle.

The function, $\lambda(d)$, is included in calculations involving the angle terms to help eliminate any discontinuities that may occur in the potential energy function caused by an abrupt change in the set of atoms used to compute the energies

associated with the angles. The appropriate form of $\lambda(d)$ found for this study is

$$\lambda(d) = \left((1 + e^{30(d-2.8)})(1 + e^{2(d-4)})\right)^{-1},$$

where d ($= \mathrm{R(SiO)}_{max}$) is the longest SiO bond length involved in the angle. Figure 1–1 shows a plot of λ versus $\mathrm{R(SiO)}_{max}$. In the case of Equation 8, the $\angle$OSiO are computed for the ten closest O atoms to a given Si atom. As indicated by the figure, the amount of energy that a particular angle contributes is weighted according to the proximity of the O atoms used to compute it. If at some point during the calculation the tenth and eleventh closest O atoms interchange, any abrupt change in the value of Equation 8 is removed because the λ associated with the new $\angle$OSiO is small.

Computational Details

Using the program PHOENIX, the simulated annealing minimization begins by randomly selecting the atomic coordinates of Z formula units of SiO_2 relative to a box representing the unit cell of a crystal. The random positioning of the atoms is accomplished by generating a random starting vector, $\mathbf{r}_o$, where each element r_i is independently and randomly chosen from the interval $[0, 1]$ where,

in the case of $Z = 3$, each element represents

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \\ r_{11} \\ r_{12} \\ r_{13} \\ r_{14} \\ r_{15} \\ r_{16} \\ r_{17} \\ r_{18} \\ r_{19} \\ r_{20} \\ r_{21} \\ r_{22} \\ r_{23} \\ r_{24} \\ r_{25} \\ r_{26} \\ r_{27} \end{bmatrix} = \begin{bmatrix} Si1_x \\ Si1_y \\ Si1_z \\ Si2_x \\ Si2_y \\ Si2_z \\ Si3_x \\ Si3_y \\ Si3_z \\ O1_x \\ O1_y \\ O1_z \\ O2_x \\ O2_y \\ O2_z \\ O3_x \\ O3_y \\ O3_z \\ O4_x \\ O4_y \\ O4_z \\ O5_x \\ O5_y \\ O5_z \\ O6_x \\ O6_y \\ O6_z \end{bmatrix}.$$

Beginning with a temperature of $T_o = 10000$, random steps are generated according to

$$\Delta\mathbf{r} = \mathbf{Q} \cdot \mathbf{u}.$$

The $\mathbf{Q}$ matrix used in the initial random walk, $l = 1$, is calculated using $\mathbf{S}$ set equal to the identity matrix (Equation 4). The annealing schedule used in this study is $M = 15\nu$ and $\chi_T = 0.9$. Following the Metropolis Monte Carlo method outlined above,

$$\Delta F = \Delta E(\mathbf{r} + \Delta\mathbf{r}) - \Delta E(\mathbf{r}),$$

is computed, where each ΔE term is computed according to Equation 7 using the atomic coordinates given by the vectors $(\mathbf{r} + \Delta\mathbf{r})$ and $\mathbf{r}$. If $\Delta F \leq 0$, the step is accepted and the position $\mathbf{r}$ is moved to $\mathbf{r}' = \mathbf{r} + \Delta\mathbf{r}$ in the phase space. On the other hand, if $\Delta F > 0$, then we accept $\Delta\mathbf{r}$ when

$$e^{-\frac{\Delta F}{T}} \leq P_r.$$

Otherwise, $\mathbf{r}$ is accepted and $\mathbf{r}'$ is set equal to $\mathbf{r}$. Following each random walk, a reduction in T is made according to Equation 5.

Recall that in addition to the temperature reduction at the end of each random walk, $\mathbf{Q}$ is calculated according to Equation 4. However, for this application of simulated annealing, it was necessary to place a cap on the maximum and minimum values of $Q_{i,i}$. This modification was necessary because the periodic boundary conditions imposed on the unit cell place limitations on the magnitudes of reasonable step sizes. Accordingly, each $Q_{i,i}$ is compared to both μ_L and μ_U which are initially set at $\mu_L = 0.01$ and $\mu_U = 0.5$. If $Q_{i,i} > \mu_U$ then $Q_{i,i}$ is set equal to μ_U, or if $Q_{i,i} < \mu_L$ then $Q_{i,i}$ is set equal to μ_L. As T is reduced, μ_L is reduced according to $\mu_L^{l+1} = 0.97\mu_L^l$, and their values compared with $Q_{i,i}^{l+1}$. The simulated annealing process is continued until the stopping condition given by Equation 6 is satisfied.

The value of $\mathbf{r}$ at the end of the simulated annealing optimization hopefully represents the approximate location of a well in the phase space. Using this value as a starting position, a quasi–Newton minimizer, MADMAX (Boisen and Gibbs, 1993), is used to adjust the values of $\mathbf{r}'$, where $\mathbf{r}'$ has the form

$$\mathbf{r}' = \begin{bmatrix} \mathbf{r} \\ a \\ b \\ c \\ \alpha \\ \beta \\ \gamma \end{bmatrix}$$

and contains the additional six unit cell parameters, until the gradient of Equation 7 is less than 1.0×10^{-8} (Appendix 1A). In the end, the optimized variables contained in $\mathbf{r}'$ represent a minimizer in the function space. At this point a single structure is generated. The procedure outlined above, Monte Carlo simulated annealing followed by quasi-Newton minimization are repeated many times, be-

ginning each structure generation from a completely independent random set of atomic coordinates, so as to generate all possible structure types.

Structure Types Generated Using Z=3 Starting with a Hexagonal Box

The first set of calculations was undertaken where the atoms comprising 3 formula units of SiO_2 were randomly placed inside a box having the same dimensions as the hexagonal unit cell of quartz (Kihara, 1990). The target angle used in these calculations is 137.6°. Figure 1–2 shows the calculated X-ray diffraction patterns, using the programs XPOW and XPOWPLOT (Downs et al., 1993), for the two lowest energy minima found. As is evident in Figure 1–2, these minima clearly represent two very different structure types for silica. The most frequently observed structure type, denoted structure type B, represents 25% of the structures generated, while the second most frequently generated structure type, denoted structure type C, represents 12%. While structure types B and C both consist of cornering-sharing SiO_4 tetrahedra, the remaining 63% of the generated structure types consist of one or more edge sharing tetrahedra.

Figure 1–3 shows that the powder pattern for B is very similar to that calculated for quartz (Kihara, 1990), despite the fact that the atomic coordinates were generated from a completely random set. In fact, an analysis of the atomic parameters obtained for all B structure types, using the program IDGROUP, indicates that 50 percent possess $P3_221$ symmetry while the other 50 percent possess $P3_121$ symmetry (Tables 1–1a and 1–1b). These are the symmetries observed for right and left handed quartz, respectively. The optimized cell parameters, even though they were allowed to vary during the quasi-Newton minimization step, resulted in an hexagonal unit cell with a density of 2.71 g/cm^3, in good agreement with that observed, 2.65 g/cm^3. As indicated in Figure 1–4, all bridging angles in the struc-

ture are the same (~137.0°), also in close agreement with that observed (~143°). Given that 3 formula units were used in the calculations, it was anticipated that the dominant structure type would represent that of quartz. Figure 1–5 shows a drawing of the B structure type viewed down [001].

On the other hand, an analysis of the atomic coordinates of C indicates that these coordinates possess $R3$ space group symmetry (Table 1–2). The optimized cell parameters define a rhombohedral unit cell with a density of 2.10 g/cm^3. The lower density is consistent with a more open framework structure SiOSi angles of 129.9° and 140.6° (Figure 1–6). It is noteworthy that two different bridging angles were generated for this structure in spite of the fact that a single target angle was used in the calculations. Examination of the crystal structure of C (Figure 1–7) indicates that it contains three membered rings of silicate tetrahedra. As this structure type has a slightly higher energy value than that of B, C possibly represents a new metastable structure type of silica.

The structure types B and C were generated by allowing the six unit parameters to vary only during the quasi-Newton minimization step. In order to assure that this strategy does not bias the results, a large number of additional calculations were completed in which all six unit cell parameters were allowed to vary during the simulated annealing minimization when T had been reduced to $T < 5$. At $T \geq 5$, the minimization was undertaken on the vector $\mathbf{r}$. However, when T was reduced to $T < 5$, the minimization will be completed on the vector $\mathbf{r}'$. During this step, the elements of $\mathbf{Q}$ that control the steps generated for the three cell lengths are constrained to the interval $0.1\mu_L \leq Q_{i,i} \leq 0.1\mu_U$, while those elements controlling the steps generated for the three interaxial angles are constrained to the interval $0.01\mu_L \leq Q_{i,i} \leq 0.01\mu_U$. Upon completion of the calculation, $\mathbf{r}'$ is used as the starting vector for the quasi-Newton minimization step. Allowing the

cell parameters to vary at higher temperatures led to computational difficulties and resulted in unrealistic results.

An analysis of the generated structures indicates that B and C were again generated using this modified procedure. In addition, the powder patterns shown in Figure 1–8 indicate that another structure type was generated. This new structure type, denoted D, has an orthorhombic unit cell with atomic coordinates that possess $C222$ space group symmetry (Table 1–3). The density of D, 2.74 g/cm^3, is slightly larger than B. In addition, the energy of the structure is higher suggesting that it is a new metastable structure that may be stabilized at high pressure. As expected, D contains two sets of bridging angles at 126° and 120° (Figure 1–9). A drawing of the crystal structure (Figure 1–10) shows that D possesses channels running parallel to [010].

In addition to the framework structure types B, C and D, the calculations generated several additional framework structures with $P1$ symmetry. The $P1$ structure types all have similar energies ($\sim$ 0.04) but are slightly lower than that calculated for D. However, all have energies that are significantly larger than B and C. The difference in energies can be related to the regularity of the silicate tetrahedra. For example, the known structure types of silica (quartz, cristobalite, coesite, etc.), typically possess regular tetrahedra with small tetrahedral angle variances ($\sim 1^{o^2}$). However, the tetrahedral angle variances calculated for the $P1$ ($12–14^{o^2}$) and D ($3–68^{o^2}$) structure types are much larger. As the angle variance for B and C are small (less than 1^{o^2}) and comparable with that exhibited by the known structure types, it is apparent that C is a more likely to be found in nature than the $P1$ and D structure types.

Structure Types Generated Starting with a Cubic Box

Given the successful generation of alternative forms of silica starting from a

hexagonal unit cell, additional calculations were undertaken for 3 formula units of SiO_2 that were randomly placed in a cubic box having the same volume as quartz. It was felt that by starting with a cubic rather than a hexagonal box any bias connected with the use of a hexagonal box might be removed. This resulted in the generation of new structure types plus all of the other structures obtained using a hexagonal cell (Figure 1–11). As before, in these calculations, the cubic cell parameters were only allowed to vary during the simulated annealing minimization at $T < 5$ and throughout the entire quasi-Newton minimization step.

This new structure type denoted F has a monoclinic unit cell and atomic coordinates that possess $C2$ space group symmetry (Table 1–4). Its energy is higher than that of B, and C and lower than D and the $P1$ structure types. The density of F, 2.58 g/cm^3, is lower than structure type B and D, suggesting a more open structure. Examination of Figure 1–12 indicates that the structure contains three nonequivalent SiOSi angles of 131.9°, 145.3°, and 152.2°. The crystal structure of structure type F is shown in Figure 1–13. In addition to these symmetrical structure types, the same $P1$ structure types described above were generated. As the tetrahedral angle variances calculated for the silicate tetrahedra in F are small ($\sim 1^{\circ 2}$), F is another phase that may occur in nature.

Structure Types Generated Starting with a Hexagonal Box with a Variety $SiOSi_o$ Target Angles

One additional difference in the structure types of silica generated is the SiOSi angles. In the calculations, the target angle was assumed to be 137.6°. To explore the role that this angle plays, a series of calculations were completed for a variety of target angles ranging between 120° and 180°. As before, each was initiated by placing a random arrangement of 3 formula units of SiO_2 inside of a hexagonal

unit cell. Like the early set of calculations, the six unit cell parameters were only allowed to vary during the quasi-Newton minimization step.

The calculated X-ray diffraction patterns for the framework structures generated using a target angle of 120° are displayed in Figure 1–14. The patterns show structure types B, C and D were generated. Because of the smaller target angle used in the calculations, the SiOSi angles generated for these structure types (and therefore the volumes) are smaller than those generated with a target angle of 137.6°(Table 1–8). However, in addition to B, C and D structure types, two additional structure types have been generated (Figure 1–14). The first, denoted A, has a monoclinic unit cell with atomic coordinates that possess $C2$ space group symmetry (Table 1–5). Figure 1–15 indicates that the structure contains two narrow SiOSi angles of 121.3° and 126.7° and a wide angle of 152.7°. The crystal structure of A is displayed in Figure 1–16. The other structure type, denoted E, has a lower energy than A but has $P1$ space group symmetry. This structure type has SiOSi angles ranging from 119° to 135° and a density of 2.61 g/cm^3.

The calculated X-ray diffraction patterns for the framework structures generated using a target angle of 130° are given in Figure 1–17. The patterns show structure types B, C, and D were generated. In addition to these, the patterns indicate that structure type F was also generated. However, a new $P1$ structure type denoted G with an energy lower than that of D and F was generated. This structure type has SiOSi angles ranging from 124.5° to 142.1° and a density of 2.32g/cm^3. The powder patterns (Figure 1–18) of the structures generated with a target angle of 140° indicate that no new structure types were generated. The powder patterns calculated for the structures generated using a target angle of 150° indicating that structure types B and D were generated (Figure 1–19). But, the pattern that resembles structure type C is actually a new structure type,

similar but with higher symmetry. This structure, denoted H also possesses a rhombohedral unit cell but an analysis of its atomic coordinates indicates they posses $R32$ space group symmetry (Table 1–6). H has a lower density, 1.98 g/cm^3, than that of structure type C because of its slightly larger SiOSi angles (Figure 20). Like C, Figure 1–21 shows that the crystal structure of H also comprised of three membered rings of SiO_4 tetrahedra. Moreover, a comparison of the linkage of the tetrahedra in both C and H shows that both structures are topologically equivalent (Figures 1–7 and 1–21).

The powder patterns calculated for the structures generated using a target angle of 160° are displayed in Figure 1–22. Examination of these patterns shows the generation of D and H structure types (Table 1–8). One of the patterns resembles that obtained for B but an analysis of its atomic coordinates showed it to possess the high quartz structure. This structure type denoted I was found 50% of the time to possess $P6_222$ space group symmetry and 50% of the time to possess $P6_422$ space group symmetry (Table 1–7a and 1–7b). These are the space group symmetries observed for right and left handed β-quartz, respectively. This is consistent with the α to β phase transition observed by Downs (1992) using the SQLOO model. The SiOSi bridging angle in structure type I is 155.7° which is consistent with the 153° angle observed in β-quartz (Kihara, 1990) (Figure 1–23). A drawing of structure type I is shown is Figure 1–24. The calculations completed using target angles of 170° and 180° resulted in the generation of structure types D, H and I (Table 1–8).

Structure Types Generated Using Z=6 Starting with an Orthorhombic Box

The question arises whether the short range force field represented in Equation 7 is capable of producing structures containing long range translational symmetry.

For example, in the case of quartz, the hexagonal unit cell can be transformed into an orthorhombic unit cell with twice the volume of the hexagonal unit cell with the cell containing six formula units. The lattice points of the hexagonal cell are related to the orthorhombic cell by C-centering translational isometries. In an exploration of this question, additional calculations were made, this time randomly placing 6 SiO_2 formula units in an orthorhombic box. The cell parameters were allowed to vary only during the quasi-Newton minimization step and a target angle of 137.6° was used.

A powder pattern calculated for the lowest energy structure optimized with 6 formula units is compared with that of B in Figure 1–25 where it is seen that the patterns are identical. Not only are the coordinates related by the symmetry operations of space group $P3_221$, but they also related by C-centering translations. This indicates that the model is capable of generating structures with translational as well as rotational symmetry. As expected, the energy of the $Z = 6$ structure type is exactly double that of the $Z = 3$ structure type. In addition to finding the $Z = 6$ quartz equivalent structure type, the Z = 6 equivalent to the rhombohedral structure type C was generated (Figure 1–26). By following the alternative minimization strategies described above, these results suggest that the $Z = 6$ equivalent to the other structure types can be generated along with possibly other structures.

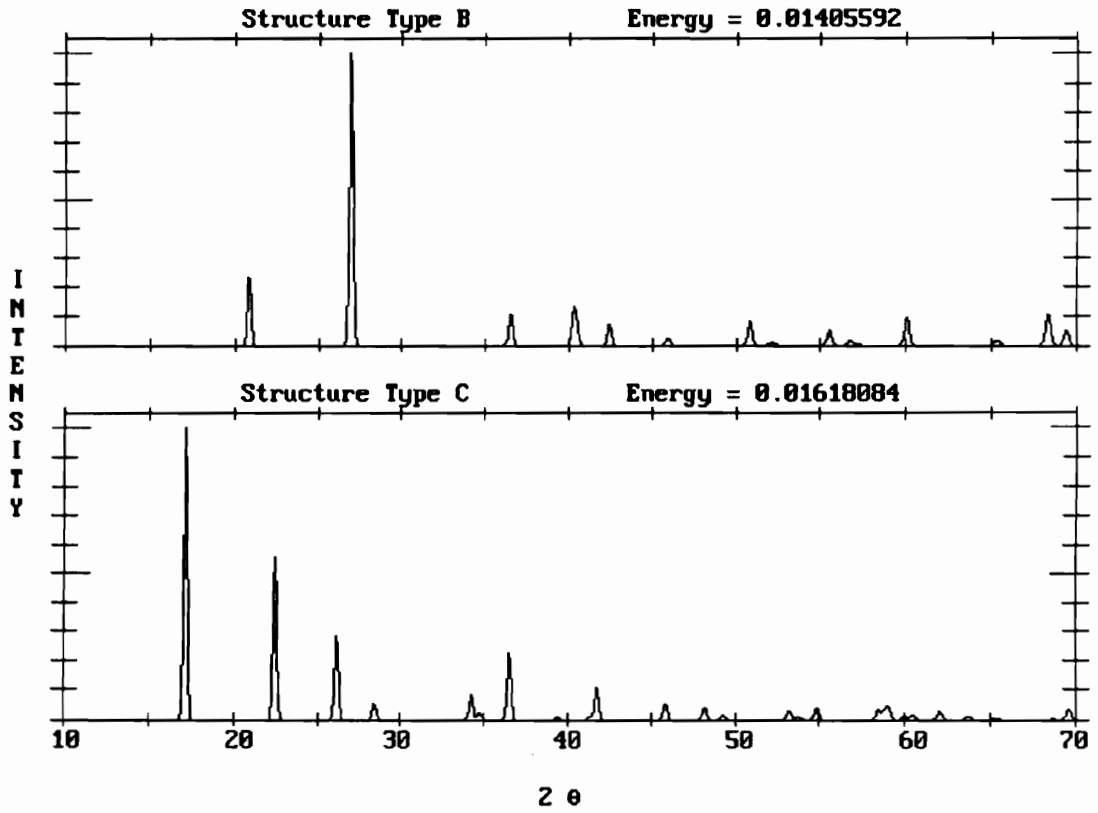


Figure 1–2. A comparison of the calculated X-ray powder diffraction patterns for structure types B and C.

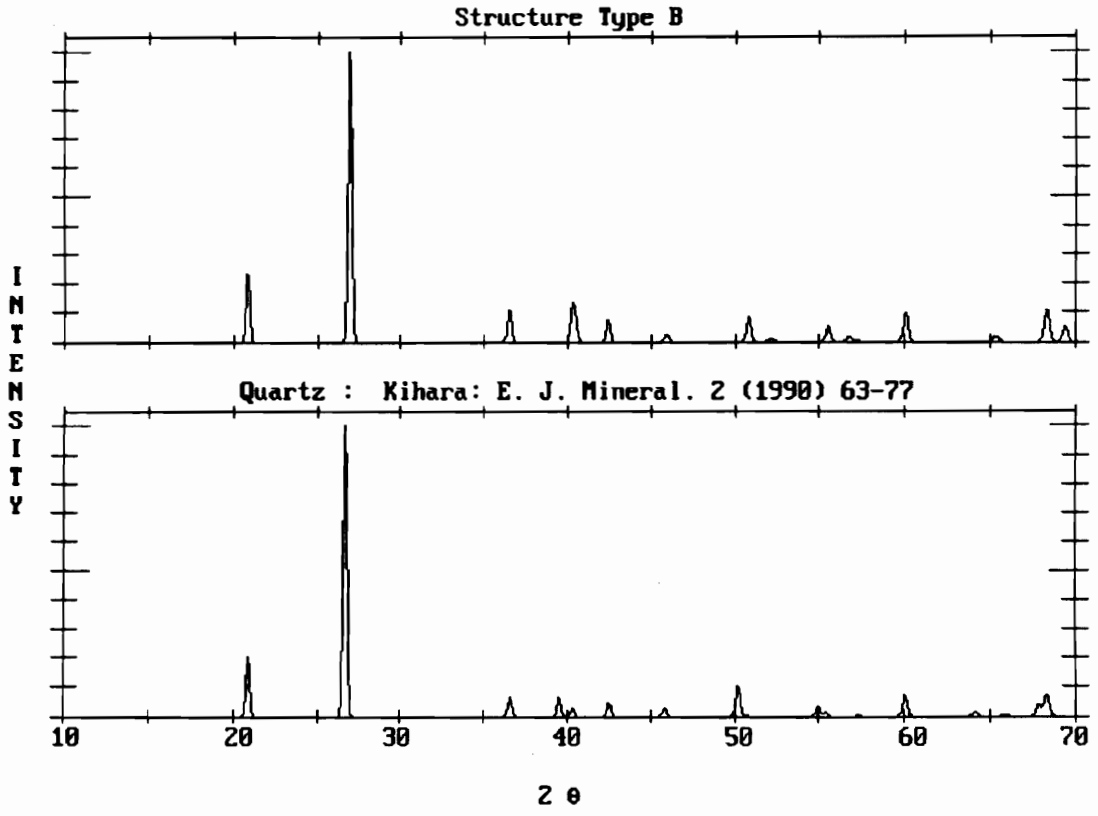


Figure 1–3. A comparison of the calculated X-ray powder diffraction patterns for structure type B and quartz (Kihara, 1990).

Table 1–1a. Optimized structural parameters for structure type B

Structure Type B			
Hexagonal: Space Group $P3_221$			
Cell Parameters: a = 4.92372Å, b = 4.92372Å, c = 5.26437Å α = 90.00000°, β = 90.00000°, γ = 120.00000°			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.54108	.00000	.66667
Si2	.00000	.54108	.33334
Si3	.45892	.45891	.00000
O1	.71217	.58918	.23344
O2	.28783	.87702	.43323
O3	.41081	.12297	.90011
O4	.87702	.28783	.56677
O5	.58919	.71216	.76657
O6	.12298	.41081	.09990
$P3_221$ Atomic Coordinates:			
atom	x	y	z
Si1	.54108	.00000	.66667
Si1	.00000	.54108	.33334
Si1	.45892	.45892	.00000
O1	.71217	.58918	.23344
O1	.28783	.87701	.43323
O1	.41082	.12299	.90011
O1	.87701	.28783	.56677
O1	.58918	.71217	.76656
O1	.12299	.41082	.09989

Table 1–1b. Optimized structural parameters for structure type B′

Structure Type B′			
Hexagonal: Space Group $P3_121$			
Cell Parameters: a = 4.92372Å, b = 4.92372Å, c = 5.26437Å α = 90.00000°, β = 90.00000°, γ = 120.00000°			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.54109	.54108	.00000
Si2	.00000	.45892	.66667
Si3	.45892	.00000	.33333
O1	.58919	.87702	.09990
O2	.12298	.71217	.43323
O3	.71217	.12298	.56677
O4	.28784	.41081	.76656
O5	.87703	.58919	.90010
O6	.41081	.28783	.23344
$P3_121$ Atomic Coordinates:			
atom	x	y	z
Si1	.54108	.54108	.00000
Si1	.00000	.45892	.66667
Si1	.45892	.00000	.33333
O1	.58919	.87702	.09990
O1	.12298	.71217	.43323
O1	.71217	.12298	.56677
O1	.28783	.41081	.76656
O1	.87702	.58919	.90010
O1	.41081	.28783	.23344

```
Space Group:  P3_221

Unit Cell Parameters =   4.9237  4.9237  5.2644  90.000  90.000 120.000
Unit Cell Volume =    110.526
Rcip Cell Parameters =  .234518 .234518 .189956  90.000  90.000  60.000
Rcip Cell Volume =    .009048

                        Distance     Angle          Atomic coordinates
Si1                                                .54108      .00000      .66667
       O1                1.6360                    .41082      .12299      .90011
       O1                1.6361                    .87701      .28783      .56677
       O1                1.6361                    .58918     -.28783      .76656
       O1                1.6361                    .28783     -.12299      .43323
   O1
       O1                2.6728     109.540
       O1                2.6708     109.422
       O1                2.6723     109.508
   O1
       O1                2.6705     109.399
       O1                2.6708     109.420
   O1
       O1                2.6728     109.539
Average bond length = 1.6361
Polyhedral volume =   2.2474
Tetrahedral angle variance =      .0042
Mean tetrahedral quadratic elongation =  1.0000

                        Distance     Angle          Atomic coordinates
O1                                                 .71217      .58918      .23344
       Si1               1.6361                   1.00000      .54108      .33334
       Si1               1.6360                    .45892      .45892      .00000
   Si1
       Si1               3.0435     136.909
Average bond length = 1.6361
```

Figure 1–4. Bond lengths and angles for structure type B.

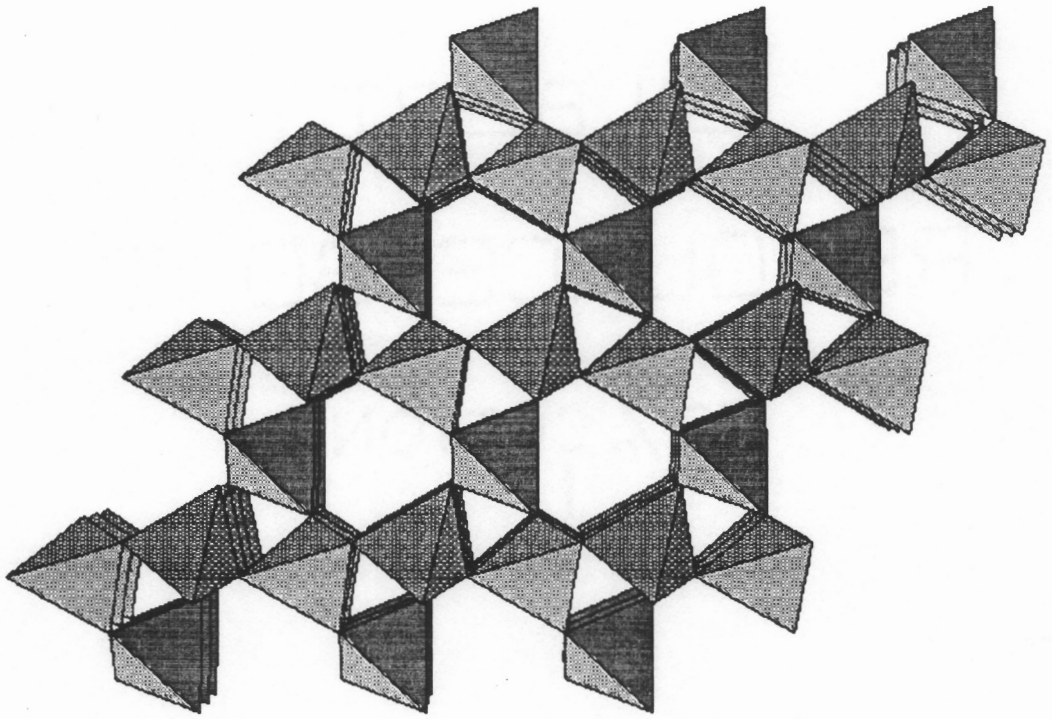


Figure 1–5. A drawing of structure type B viewed down [001].

Table 1–2. Optimized structural parameters for for structure type C

Structure Type C			
Rhombohedral: Space Group $R3$			
Cell Parameters: a = 5.26950Å, b = 5.26950Å, c = 5.26950Å α = 97.42986°, β = 97.42986°, γ = 97.42986°			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.12318	.32013	.55669
Si2	.55669	.12317	.32014
Si3	.32013	.55669	.12318
O1	.15505	.54227	.36633
O2	.12352	.45706	.85272
O3	.54227	.36632	.15505
O4	.85271	.12351	.45707
O5	.36633	.15504	.54228
O6	.45707	.85271	.12352
$R3$ Atomic Coordinates:			
atom	x	y	z
Si1	.12318	.32013	.55669
Si1	.55669	.12318	.32013
Si1	.32013	.55669	.12318
O1	.15505	.54227	.36633
O2	.12352	.45706	.85272
O1	.54227	.36633	.15505
O2	.85272	.12352	.45706
O1	.36633	.15505	.54227
O2	.45706	.85272	.12352

```
Space Group:  R3

Unit Cell Parameters =   5.2695  5.2695  5.2695  97.430  97.430  97.430
Unit Cell Volume =    142.279
Rcip Cell Parameters =  .193524 .193524 .193524  81.459  81.459  81.459
Rcip Cell Volume =    .007028

                        Distance     Angle          Atomic coordinates
Si1                                                .12318      .32013      .55669
        O1               1.6435                    .15505      .54227      .36633
        O1               1.6433                    .36633      .15505      .54227
        O2               1.6315                    .12352      .45706      .85272
        O2               1.6322                   -.14728      .12352      .45706
   O1
        O1               2.6594     108.018
        O2               2.6808     109.883
        O2               2.6826     109.958
   O1
        O2               2.6908     110.504
        O2               2.6654     108.923
   O2
        O2               2.6658     109.531
Average bond length = 1.6376
Polyhedral volume =   2.2533
Tetrahedral angle variance =      .7778
Mean tetrahedral quadratic elongation =  1.0002

                        Distance     Angle          Atomic coordinates
O1                                                 .15505      .54227      .36633
        Si1              1.6435                    .12318      .32013      .55669
        Si1              1.6433                    .32013      .55669      .12318
   Si1
        Si1              2.9773     129.873
Average bond length = 1.6434

                        Distance     Angle          Atomic coordinates
O2                                                 .12352      .45706      .85272
        Si1              1.6315                    .12318      .32013      .55669
        Si1              1.6322                    .32013      .55669     1.12318
   Si1
        Si1              3.0731     140.637
Average bond length = 1.6319
```

Figure 1–6. Bond lengths and angles for structure type C.

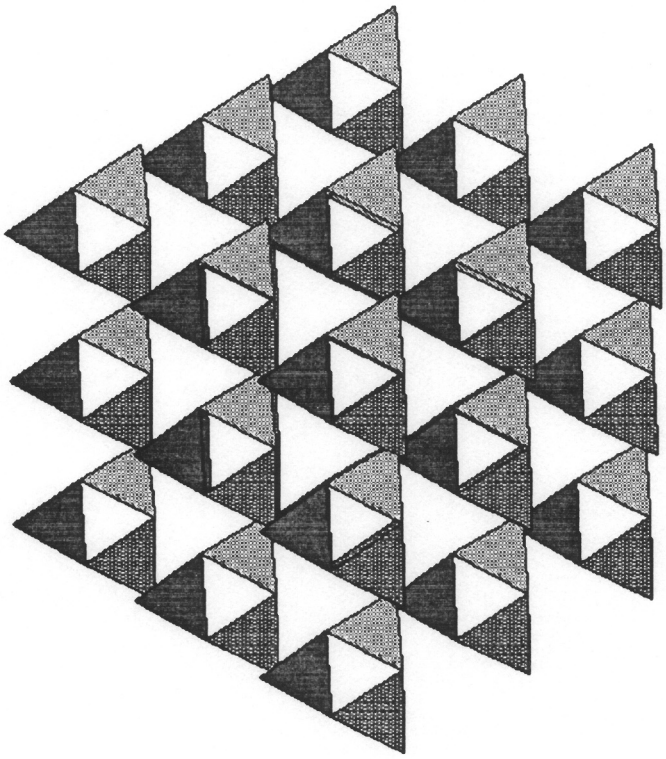


Figure 1–7. A drawing of structure type C viewed down [111].

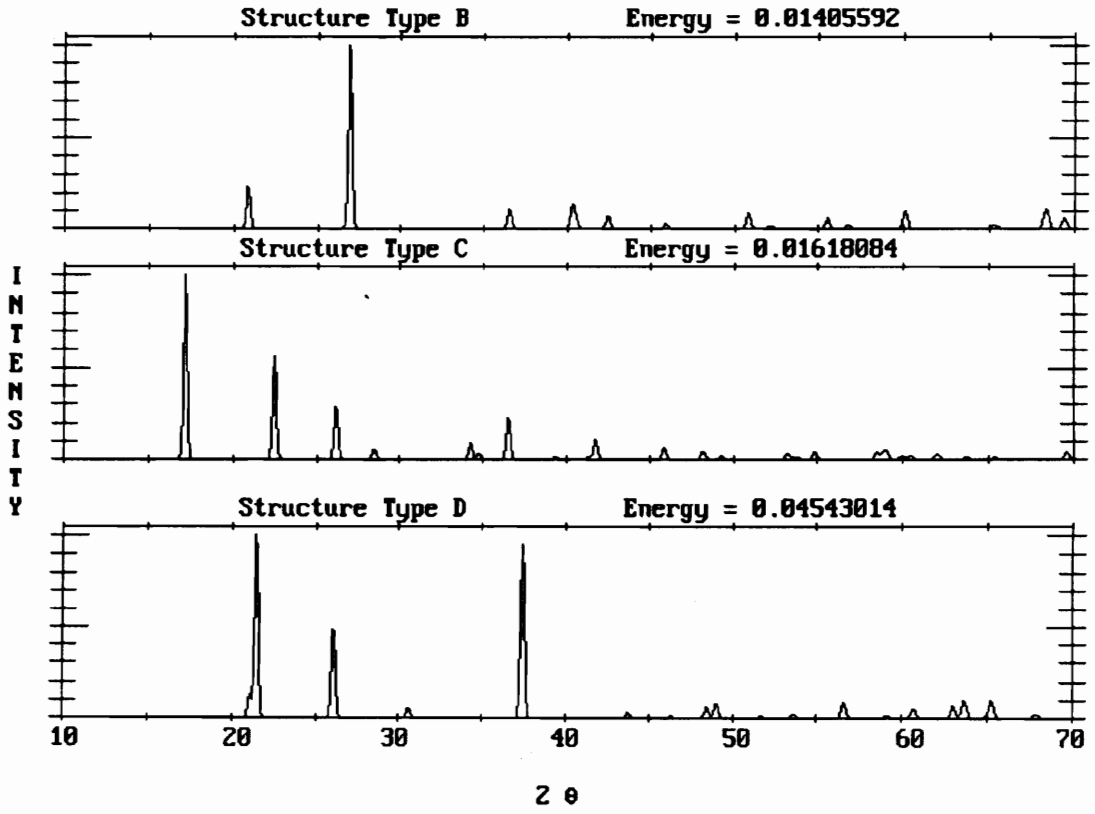


Figure 1–8. A comparison of the calculated X-ray powder diffraction patterns for structure types B, C and D. These structure types were generated by starting with a hexagonal unit cell then allowing the six cell parameters to vary at $T < 5$.

Table 1–3. Optimized structural parameters for for structure type D

Structure Type D			
Orthorhombic: Space Group $C222$			
Cell Parameters: a = 11.68764Å, b = 4.427591Å, c = 4.217210Å α = 90.00099°, β = 90.00034°, γ = 89.99889°			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.00000	.50000	.00001
Si2	.33248	.00000	.49999
Si3	.16752	.50001	.50000
O1	.25000	.25000	.67683
O2	.08607	.31541	.23963
O3	.58607	.81541	.23963
O4	.58607	.18461	.76037
O5	.25000	.75000	.32317
O6	.41392	.18459	.23963
Corresponding $C222$ Atomic Coordinates:			
atom	x	y	z
Si1	.00000	.50000	.00001
Si2	.33248	.00000	.49999
Si2	.16752	.50000	.49999
O1	.25000	.25000	.67683
O2	.08607	.31541	.23963
O2	.58607	.81541	.23963
O2	.58607	.18459	.76037
O1	.25000	.75000	.32317
O2	.41393	.18459	.23963

```
Space Group:  C222

Unit Cell Parameters =  11.6876  4.4276  4.2172  90.001  90.000  89.999
Unit Cell Volume =    218.233
Rcip Cell Parameters =  .085560 .225856 .237124  89.999  90.000  90.001
Rcip Cell Volume =    .004582

                        Distance     Angle          Atomic coordinates
Si1                                                .00000      .50000      .00001
        O2               1.6435                    .08607      .31541      .23963
        O2               1.6436                    .08607      .68459     -.23963
        O2               1.6435                   -.08607      .68459      .23963
        O2               1.6435                   -.08607      .31541     -.23963
   O2
        O2               2.5994     104.521
        O2               2.5922     104.115
        O2               2.8518     120.359
   O2
        O2               2.8518     120.360
        O2               2.5922     104.112
   O2
        O2               2.5994     104.520
Average bond length = 1.6435
Polyhedral volume =   2.2156
Tetrahedral angle variance =    68.7081
Mean tetrahedral quadratic elongation =  1.0188

                        Distance     Angle          Atomic coordinates
Si2                                               .33248      .00000      .49999
        O1               1.6464                   .25000      .25000      .67683
        O1               1.6464                   .25000     -.25000      .32317
        O2               1.6673                   .41393      .18459      .23963
        O2               1.6673                   .41393     -.18459      .76037
   O1
        O1               2.6693     108.320
        O2               2.6747     107.641
        O2               2.7381     111.440
   O1
        O2               2.7382     111.446
        O2               2.6747     107.640
   O2
        O2               2.7376     110.366
Average bond length = 1.6568
Polyhedral volume =   2.3309
Tetrahedral angle variance =     3.3211
Mean tetrahedral quadratic elongation =  1.0010
```

Figure 1–9. Bond lengths and angles for structure type D.

		Distance	Angle	Atomic coordinates		
O1				.25000	.25000	.67683
	Si2	1.6464		.33248	.00000	.49999
	Si2	1.6464		.16752	.50000	.49999
Si2						
	Si2	2.9356	126.131			
Average bond length = 1.6464						

		Distance	Angle	Atomic coordinates		
O2				.08607	.31541	.23963
	Si1	1.6435		.00000	.50000	.00001
	Si2	1.6673		.16752	.50000	.49999
Si1						
	Si2	2.8774	120.706			
Average bond length = 1.6554						

Figure 1–9 (continued). - Bond lengths and angles for structure type D

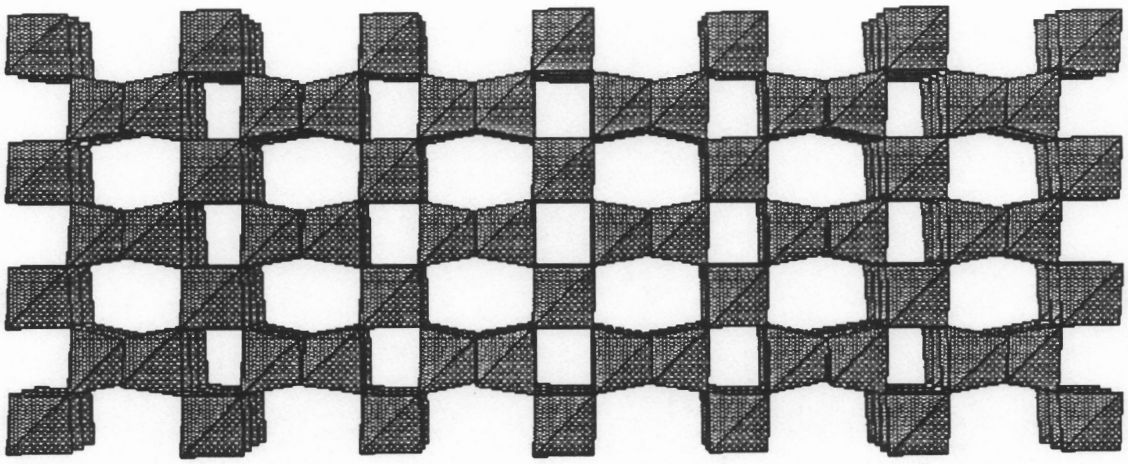


Figure 1–10. A drawing of structure type D viewed down [010].

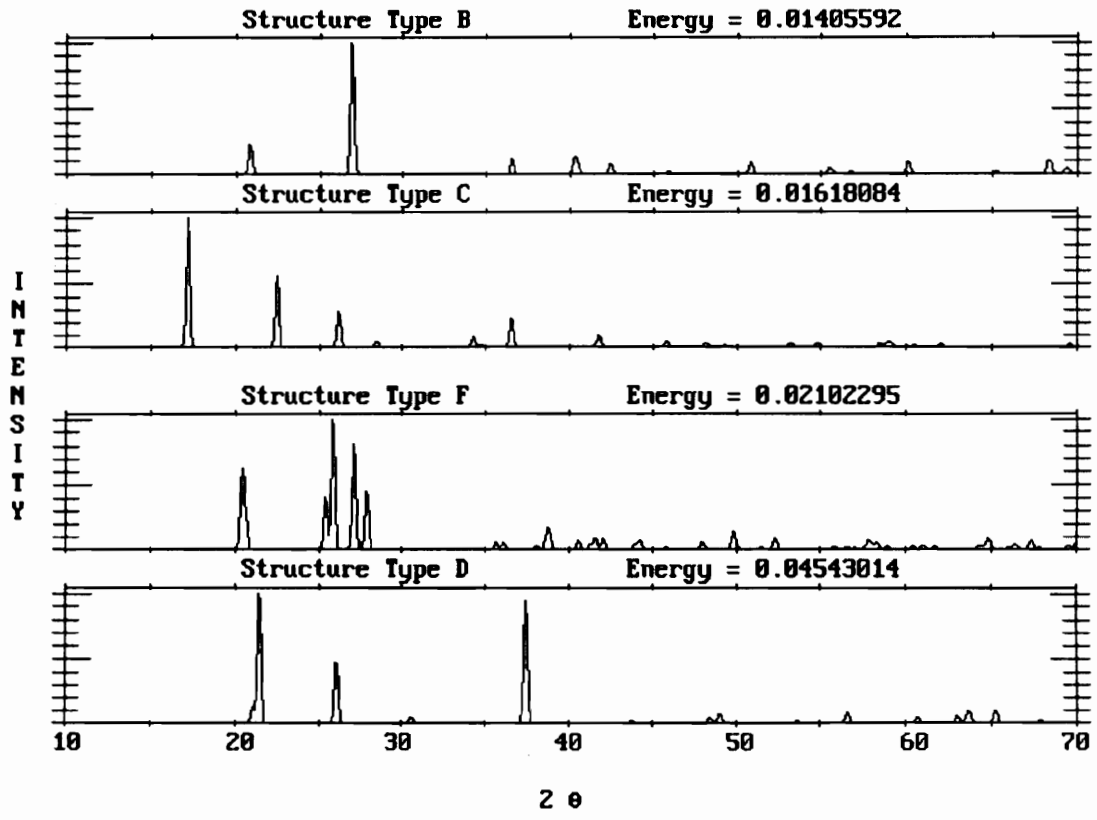


Figure 1–11. A comparison of the calculated X-ray powder diffraction patterns for structure types B, C, F and D. These structure types were generated by starting with a cubic unit cell then allowing the six cell parameters to vary at $T < 5$.

Table 1–4. Optimized structural parameters for for structure type F

Structure Type F			
Monoclinic: Space Group $C2$			
Cell Parameters: a = 8.63331Å, b = 5.04086Å, c = 5.34834Å α = 90.00000°, β = 95.32008°, γ = 89.99996°			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.99999	.71437	.99999
Si2	.77017	.50000	.33544
Si3	.22982	.50000	.66455
O1	.05978	.52834	.77642
O2	.21736	.29045	.43159
O3	.64369	.40130	.11062
O4	.14369	.90130	.11062
O5	.28263	.79045	.56840
O6	.35630	.40130	.88937
Corresponding $C2$ Atomic Coordinates:			
atom	x	y	z
Si1	.99999	.71437	.99999
Si2	.77017	.50000	.33544
Si2	.22983	.50000	.66456
O1	.05978	.52834	.77642
O2	.21736	.29045	.43159
O3	.64369	.40130	.11062
O3	.14369	.90130	.11062
O2	.28264	.79045	.56841
O3	.35631	.40130	.88938

```
Space Group:  C2

Unit Cell Parameters =   8.6333  5.0409  5.3483  90.000  95.320  90.000
Unit Cell Volume =    231.753
Rcip Cell Parameters =  .116332 .198379 .187783  90.000  84.680  90.000
Rcip Cell Volume =    .004315

                        Distance     Angle          Atomic coordinates
Si1                                                .99999      .71437      .99999
       O1                1.6401                  1.05978      .52834      .77642
       O1                1.6402                   .94022      .52834     1.22358
       O3                1.6251                  1.14369      .90130     1.11062
       O3                1.6250                   .85631      .90130      .88938
   O1
       O1                2.6912     110.255
       O3                2.6484     108.404
       O3                2.6799     110.326
   O1
       O3                2.6799     110.317
       O3                2.6484     108.411
   O3
       O3                2.6479     109.119
Average bond length = 1.6326
Polyhedral volume =   2.2323
Tetrahedral angle variance =      .8895
Mean tetrahedral quadratic elongation =  1.0003

                        Distance     Angle          Atomic coordinates
Si2                                                .77017      .50000      .33544
       O1                1.6421                   .94022      .52834      .22358
       O2                1.6294                   .78264      .29045      .56841
       O2                1.6308                   .71736      .79045      .43159
       O3                1.6255                   .64369      .40130      .11062
   O1
       O2                2.6750     109.699
       O2                2.6612     108.803
       O3                2.6537     108.607
   O2
       O2                2.6701     109.966
       O3                2.6835     111.067
   O2
       O3                2.6452     108.650
Average bond length = 1.6319
Polyhedral volume =   2.2298
Tetrahedral angle variance =      .9421
Mean tetrahedral quadratic elongation =  1.0002
```

Figure 1–12. Bond lengths and angles for structure type F.

		Distance	Angle	Atomic coordinates		
O1				.05978	.52834	.77642
	Si1	1.6401		-.00001	.71437	.99999
	Si2	1.6421		.22983	.50000	.66456
Si1						
	Si2	2.9973	131.897			
Average bond length = 1.6411						

		Distance	Angle	Atomic coordinates		
O2				.21736	.29045	.43159
	Si2	1.6294		.22983	.50000	.66456
	Si2	1.6308		.27017	.00000	.33544
Si2						
	Si2	3.1122	145.343			
Average bond length = 1.6301						

		Distance	Angle	Atomic coordinates		
O3				.64369	.40130	.11062
	Si1	1.6251		.49999	.21437	-.00001
	Si2	1.6255		.77017	.50000	.33544
Si1						
	Si2	3.1554	152.201			
Average bond length = 1.6253						

Figure 1–12 (continued). Bond lengths and angles for structure type F

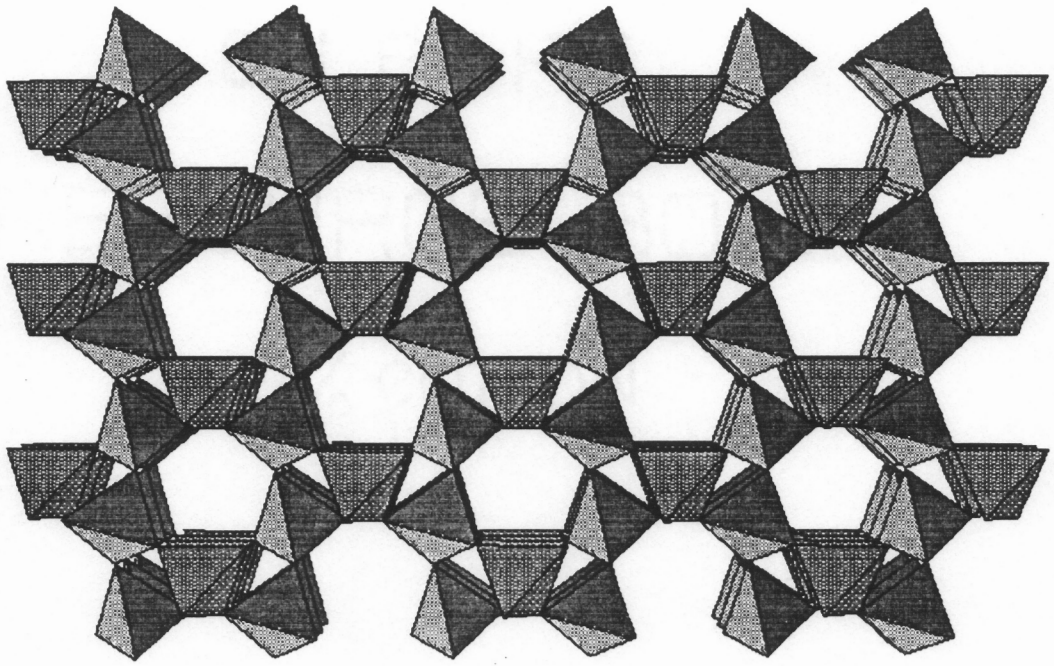


Figure 1–13. A drawing of structure type F viewed down [001].

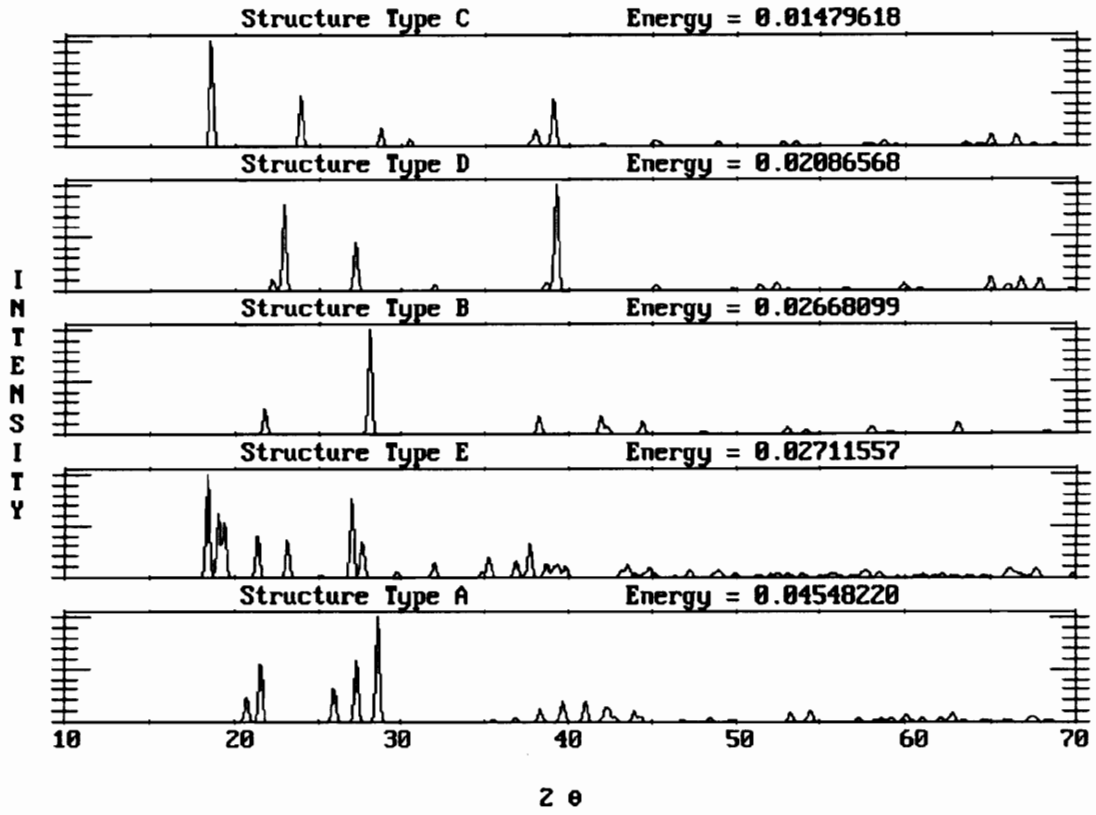


Figure 1–14. A comparison of the calculated X-ray powder diffraction patterns for structure types C, D, B, E and A. These structure types were generated using a target angle of 120°.

Table 1–5. Optimized structural parameters for for structure type A

Structure Type A			
Monoclinic: Space Group $C2$			
Cell Parameters: a = 12.67098Å, b = 4.70005Å, c = 5.07891Å α = 89.99821°, β = 42.53757°, γ = 89.99942°			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.00000	.17694	.49998
Si2	.23016	.99999	.70174
Si3	.73016	.49999	.70174
O1	.44463	.47187	.84740
O2	.73870	.33744	.40349
O3	.14333	.38206	.32176
O4	.94463	.97187	.84740
O5	.64333	.88206	.32176
O6	.55537	.47182	.15263
Corresponding $C2$ Atomic Coordinates:			
atom	x	y	z
Si1	.00000	.17694	.49998
Si2	.23016	.00001	.70174
Si2	.73016	.49999	.70174
O1	.44463	.47187	.84740
O2	.73870	.33744	.40349
O3	.14333	.38206	.32176
O1	.94463	.97187	.84740
O3	.64333	.88206	.32176
O1	.55537	.47187	.15260

```
Space Group:  C2

Unit Cell Parameters =  12.6710  4.7001  5.0789  89.998  42.538  89.999
Unit Cell Volume =    204.493
Rcip Cell Parameters =  .116734 .212764 .291230  90.002 137.462  89.999
Rcip Cell Volume =    .004890

                        Distance     Angle          Atomic coordinates
Si1                                                .00000      .17694      .49998
       O1                1.6463                   -.05537     -.02813      .84740
       O1                1.6462                    .05537     -.02813      .15260
       O3                1.6200                    .14333      .38206      .32176
       O3                1.6200                   -.14333      .38206      .67824
   O1
       O1                2.6692     108.326
       O3                2.6977     111.358
       O3                2.6661     109.420
   O1
       O3                2.6662     109.426
       O3                2.6978     111.371
   O3
       O3                2.6039     106.960
Average bond length = 1.6331
Polyhedral volume =   2.2326
Tetrahedral angle variance =     2.9577
Mean tetrahedral quadratic elongation =  1.0009

                        Distance     Angle          Atomic coordinates
Si2                                               .23016      .99999      .70174
       O1                1.6409                    .05537      .97187     1.15260
       O2                1.6267                    .26130     1.33744      .59651
       O2                1.6274                    .23870      .83744      .40349
       O3                1.6153                    .35667      .88206      .67824
   O1
       O2                2.6224     106.752
       O2                2.6919     110.905
       O3                2.6464     108.726
   O2
       O2                2.6419     108.558
       O3                2.6368     108.846
   O2
       O3                2.7018     112.854
Average bond length = 1.6276
Polyhedral volume =   2.2089
Tetrahedral angle variance =     4.5345
Mean tetrahedral quadratic elongation =  1.0011
```

Figure 1–15. Bond lengths and angles for structure type A.

		Distance	Angle	Atomic coordinates		
O1				.44463	.47187	.84740
	Si1	1.6463		.50000	.67694	.49998
	Si2	1.6409		.26984	.49999	1.29826
Si1						
	Si2	2.8653	121.306			
Average bond length = 1.6436						

		Distance	Angle	Atomic coordinates		
O2				.73870	.33744	.40349
	Si2	1.6267		.76984	-.00001	.29826
	Si2	1.6274		.73016	.49999	.70174
Si2						
	Si2	2.9080	126.672			
Average bond length = 1.6270						

		Distance	Angle	Atomic coordinates		
O3				.14333	.38206	.32176
	Si1	1.6200		.00000	.17694	.49998
	Si2	1.6153		.26984	.49999	.29826
Si1						
	Si2	3.1438	152.675			
Average bond length = 1.6177						

Figure 1–15 (continued). Bond lengths and angles for structure type A

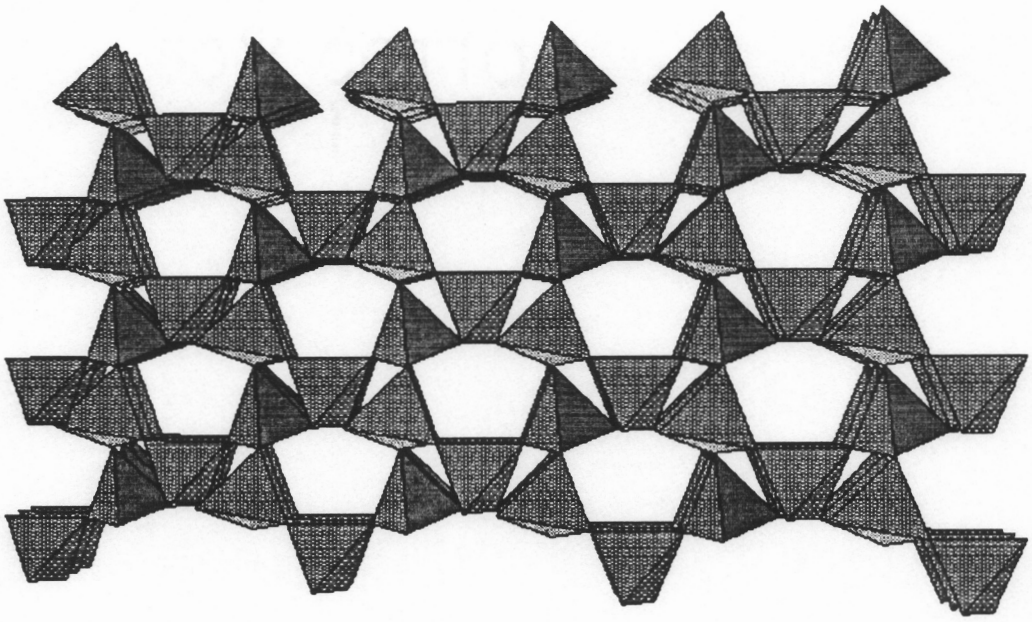


Figure 1–16. A drawing of structure type A viewed down [001].

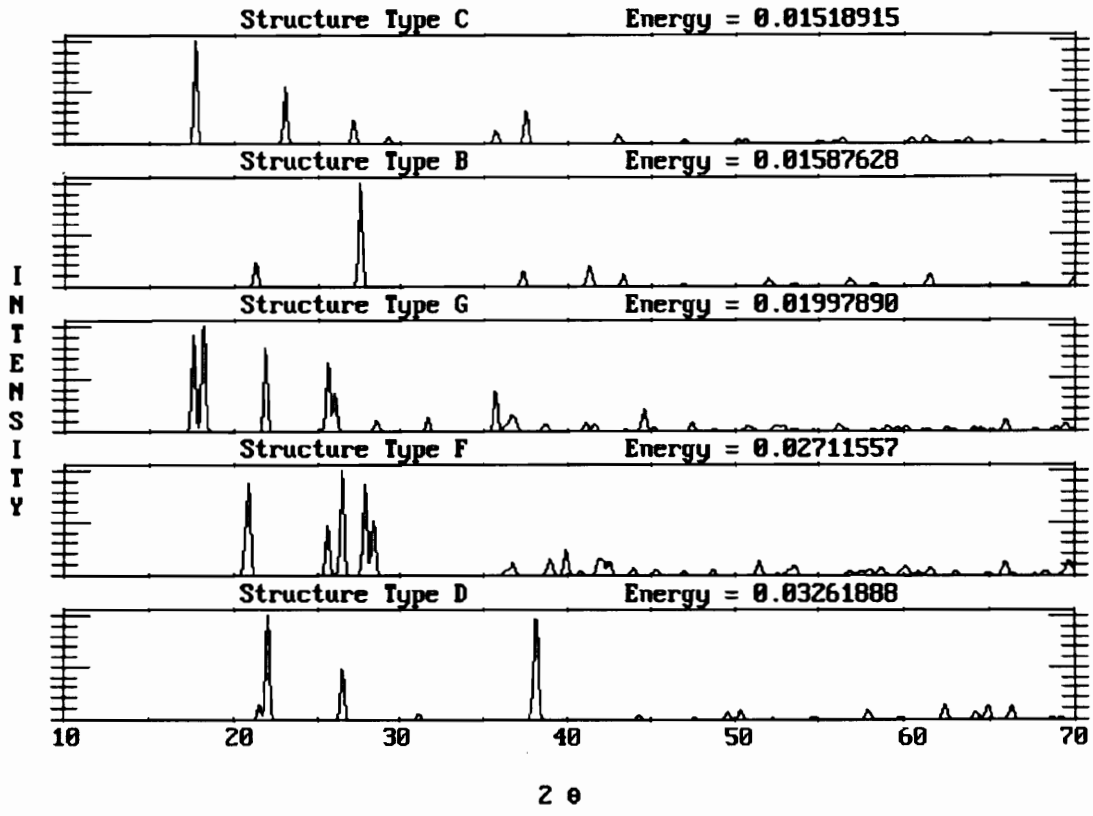


Figure 1–17. A comparison of the calculated X-ray powder diffraction patterns for structure types C, B, G, F and D. These structure types were generated using a target angle of 130°.

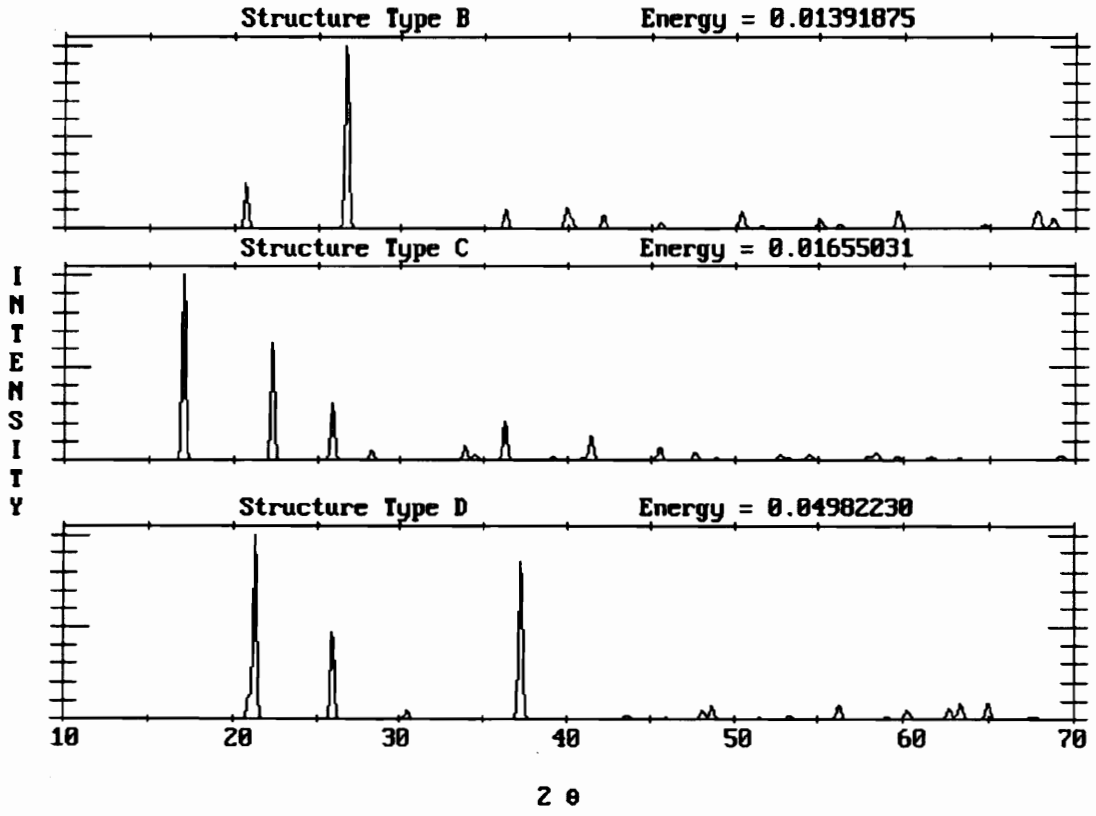


Figure 1-18. A comparison of the calculated X-ray powder diffraction patterns for structure types B, C and D. These structure types were generated using a target angle of 140°.

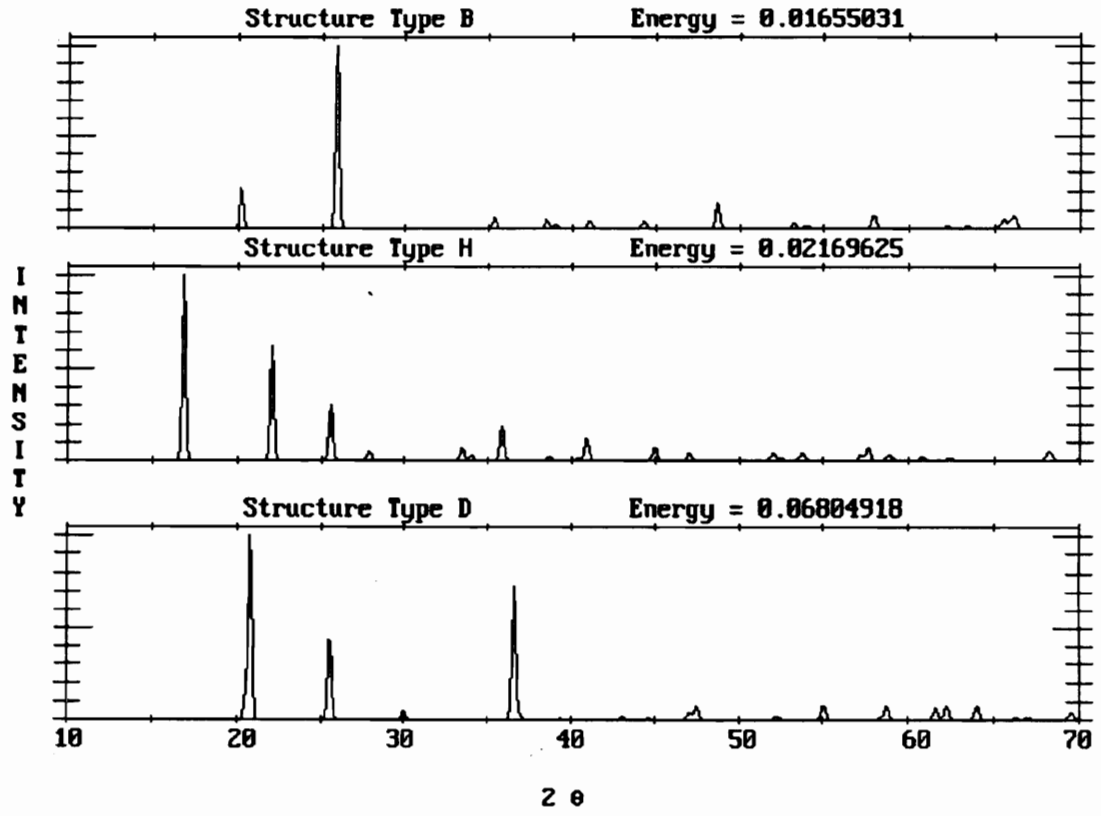


Figure 1–19. A comparison of the calculated X-ray powder diffraction patterns for structure types B, H and D. These structure types were generated using a target angle of 150°.

Table 1–6. Optimized structural parameters for structure type H

Structure Type H

Rhombohedral:
Space Group $R32$

Cell Parameters:

a = 5.37371Å, b = 5.37371Å, c = 5.37372Å
$\alpha = 97.25462°$, $\beta = 97.25474°$, $\gamma = 97.25460°$

Optimized $P1$ Atomic Coordinates:

atom	x	y	z
Si1	.00000	.21697	.78302
Si2	.21697	.78302	.00000
Si3	.78302	.00000	.21697
O1	.50000	.83380	.16620
O2	.99999	.81055	.18945
O3	.16619	.50000	.83380
O4	.81055	.18945	.99999
O5	.18944	.00000	.81055
O6	.83380	.16620	.49999

$R32$ Atomic Coordinates:

atom	x	y	z
Si1	.00000	.21697	.78302
Si1	.21697	.78302	.00000
Si1	.78302	.00000	.21697
O1	.50000	.83380	.16620
O2	.99999	.81055	.18945
O1	.16620	.50000	.83380
O2	.81055	.18945	.99999
O2	.18945	.99999	.81055
O1	.83380	.16620	.50000

```
Space Group:  R32

Unit Cell Parameters =   5.3737  5.3737  5.3737  97.255  97.255  97.255
Unit Cell Volume =    151.098
Rcip Cell Parameters =  .189583 .189583 .189583  81.690  81.690  81.690
Rcip Cell Volume =    .006618

                        Distance     Angle          Atomic coordinates
Si1                                                .00000      .21697      .78302
        O1               1.6358                    .16620      .50000      .83380
        O1               1.6358                   -.16620      .16620      .50000
        O2               1.6502                    .18945     -.00001      .81055
        O2               1.6502                   -.18945      .18945      .99999
   O1
        O1               2.6823     110.144
        O2               2.6934     110.103
        O2               2.6902     109.905
   O1
        O2               2.6902     109.908
        O2               2.6933     110.101
   O2
        O2               2.6465     106.619
Average bond length = 1.6430
Polyhedral volume =   2.2745
Tetrahedral angle variance =     1.9523
Mean tetrahedral quadratic elongation =  1.0005

                        Distance     Angle          Atomic coordinates
O1                                                 .50000      .83380      .16620
        Si1              1.6358                    .78302     1.00000      .21697
        Si1              1.6358                    .21697      .78302      .00000
   Si1
        Si1              3.1363     146.936
Average bond length = 1.6358

                        Distance     Angle          Atomic coordinates
O2                                                 .99999      .81055      .18945
        Si1              1.6502                    .78302     1.00000      .21697
        Si1              1.6502                   1.21697      .78302      .00000
   Si1
        Si1              3.0310     133.381
Average bond length = 1.6502
```

Figure 1–20. Bond lengths and angles for structure type H.

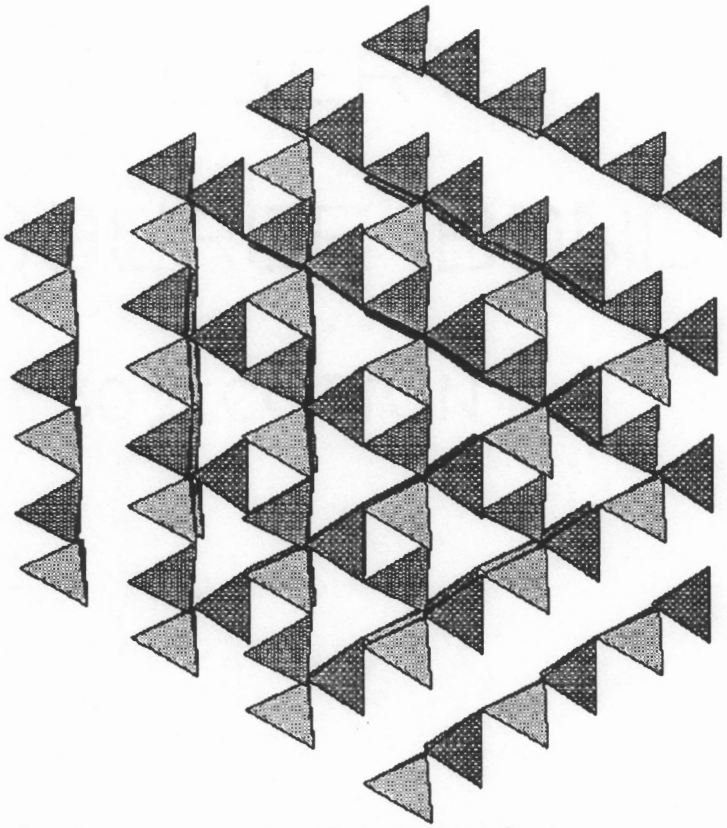


Figure 1–21. A drawing of structure type H viewed down [111].

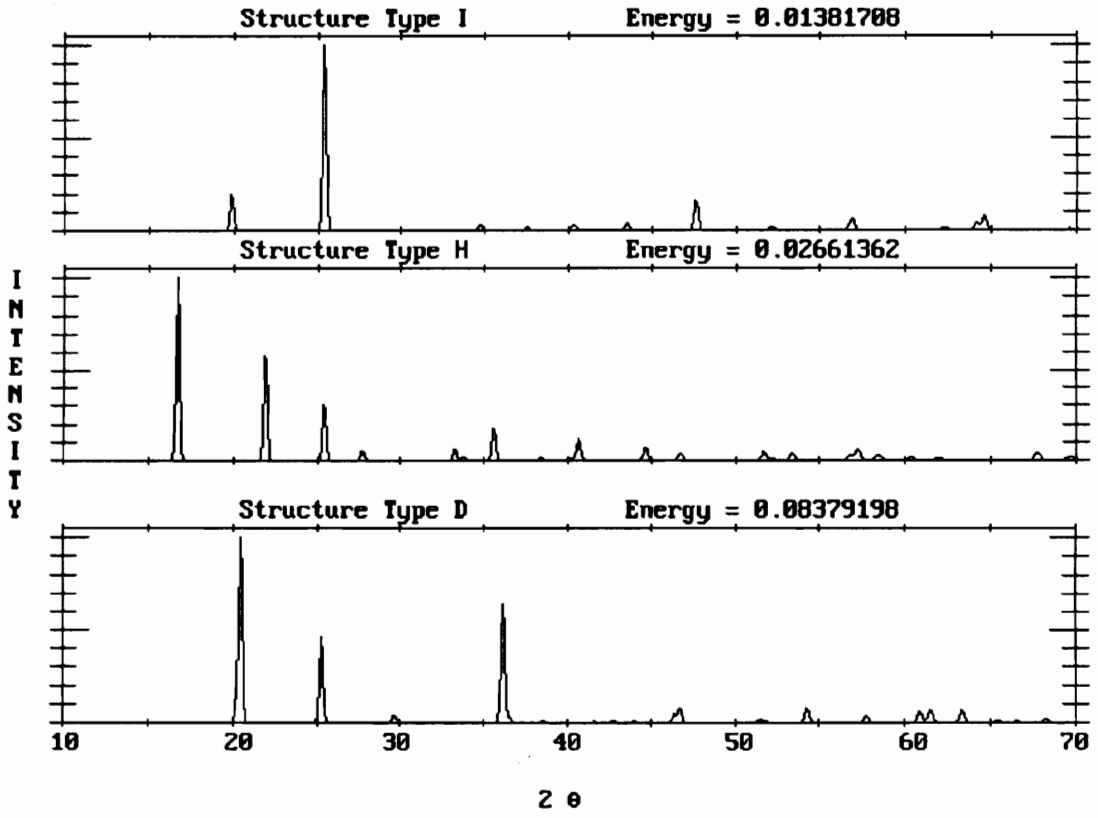


Figure 1–22. A comparison of the calculated X-ray powder diffraction patterns for structure types I, H and D. These structure types were generated using a target angle of 160°.

Table 1–7a. Optimized structural parameters for structure type I

Structure Type I			
Hexagonal: Space Group $P6_222$			
Cell Parameters: a = 5.16574Å, b = 5.16574Å, c = 5.67236Å α = 90.00001°, β = 89.99999°, γ = 120.00000°			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.50000	.00000	.50000
Si2	.00000	.50000	.16666
Si3	.50000	.50000	.83333
O1	.21144	.42289	.00000
O2	.78856	.57712	.00000
O3	.57712	.78856	.66666
O4	.42288	.21145	.66666
O5	.78856	.21145	.33333
O6	.21144	.78856	.33333
$P6_222$ Atomic Coordinates:			
atom	x	y	z
Si1	.50000	.00000	.50000
Si1	.00000	.50000	.16667
Si1	.50000	.50000	.83333
O1	.21144	.42289	.00000
O1	.78855	.21144	.33333
O1	.57711	.78855	.66667
O1	.42289	.21145	.66667
O1	.78856	.57711	.00000
O1	.21145	.78856	.33333

Table 1–7b. Optimized structural parameters for structure type I′

Structure Type I′			
Hexagonal: Space Group $P6_422$			
Cell Parameters: a = 5.16574Å, b = 5.16574Å, c = 5.67236Å $\alpha = 90.00000°$, $\beta = 90.00000°$, $\gamma = 120.00000°$			
Optimized $P1$ Atomic Coordinates:			
atom	x	y	z
Si1	.50000	.00000	.00000
Si2	.50000	.50000	.66667
Si3	.00000	.50000	.33333
O1	.21144	.42288	.50000
O2	.78856	.21144	.16667
O3	.42289	.21144	.83333
O4	.21144	.78855	.16667
O5	.78856	.57711	.50000
O6	.57712	.78855	.83333
$P6_422$ Atomic Coordinates:			
atom	x	y	z
Si1	.50000	.00000	.00000
Si1	.50000	.50000	.66667
Si1	.00000	.50000	.33333
O1	.21144	.42288	.50000
O1	.78856	.21144	.16667
O1	.42288	.21144	.83333
O1	.21144	.78856	.16667
O1	.78856	.57712	.50000
O1	.57712	.78856	.83333

```
Space Group:  P6_222

Unit Cell Parameters =   5.1657  5.1657  5.6724  90.000  90.000 120.000
Unit Cell Volume =    131.087
Rcip Cell Parameters =  .223531 .223531 .176293  90.000  90.000  60.000
Rcip Cell Volume =    .007629

                        Distance     Angle          Atomic coordinates
Si1                                                .50000      .00000      .50000
        O1               1.6372                    .78855      .21144      .33333
        O1               1.6373                    .57711     -.21145      .66667
        O1               1.6372                    .21145     -.21144      .33333
        O1               1.6373                    .42289      .21145      .66667
   O1
        O1               2.6747     109.540
        O1               2.6734     109.459
        O1               2.6727     109.414
   O1
        O1               2.6727     109.414
        O1               2.6735     109.461
   O1
        O1               2.6747     109.540
Average bond length = 1.6372
Polyhedral volume =   2.2523
Tetrahedral angle variance =      .0032
Mean tetrahedral quadratic elongation =  1.0000

                        Distance     Angle          Atomic coordinates
O1                                                 .21144      .42289      .00000
        Si1              1.6373                    .50000      .50000     -.16667
        Si1              1.6372                    .00000      .50000      .16667
   Si1
        Si1              3.2010     155.673
Average bond length = 1.6372
```

Figure 1–23. Bond lengths and angles for structure type I.

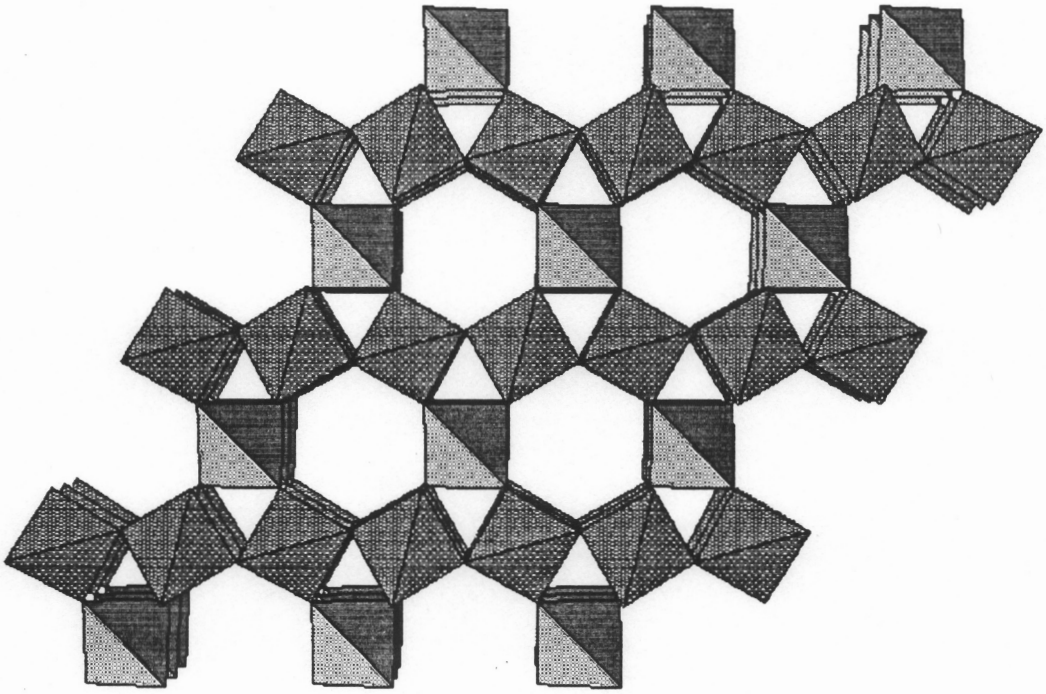


Figure 1–24. A drawing of structure type I viewed down [001].

Table 1–8. Summary of framework structure types generated using various target angles ($SiOSi_o$). A ([2]) denotes that the structure type was generated from a hexagonal box varied at $T < 5$. A ([3]) denotes that the structure type was generated starting with a cubic box.

Structure Type	Space Group	$SiOSi_o$ (°)	Energy	Volume (Å^3)	Density (g/cm^3)
B[2,3]	$P3_221/P3_121$	137.6	0.01405592	110.526	2.71
		120.0	0.02668099	97.590	3.07
		130.0	0.01587628	103.541	2.89
		140.0	0.01391874	113.099	2.65
		150.0	0.01655031	145.724	2.05
C[2,3]	$R3$	137.6	0.01618084	142.279	2.10
		120.0	0.01479618	113.887	2.63
		130.0	0.01518915	130.365	2.30
		140.0	0.01655031	145.724	2.05
D[2,3]	$C222$	137.6	0.04543014	218.233	2.74
		120.0	0.02086568	189.250	3.16
		130.0	0.03261888	206.179	2.90
		140.0	0.04982230	221.626	2.70
		150.0	0.06804918	233.961	2.56
		160.0	0.08379198	242.806	2.46
		170.0	0.09437132	248.103	2.41
		180.0	0.09809152	249.866	2.40
F[3]	$C2$	137.6	0.02012295	231.75	2.58
		130.0	0.02711557	218.38	2.74
A	$C2$	120.0	0.04548220	204.49	2.93
H	$R32$	150.0	0.02169625	151.098	1.98
		160.0	0.02661362	153.962	1.94
		170.0	0.03207569	155.712	1.92
		180.0	0.03416502	156.300	1.91
I	$P6_222/P6_422$	160.0	0.01381708	131.087	2.28
		170.0	0.01482477	132.067	2.27
		180.0	0.01546675	132.400	2.26

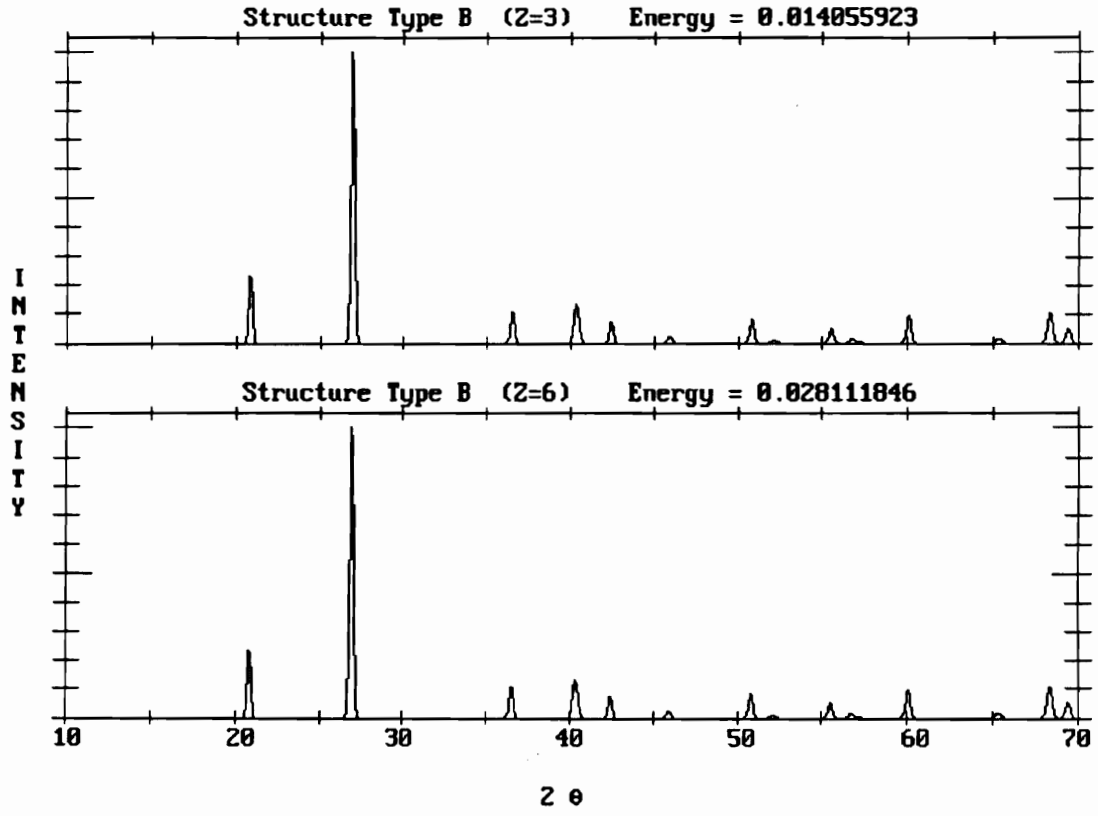


Figure 1–25. A comparison of the calculated X-ray powder diffraction patterns for structure type B generated using $Z = 3$ and $Z = 6$.

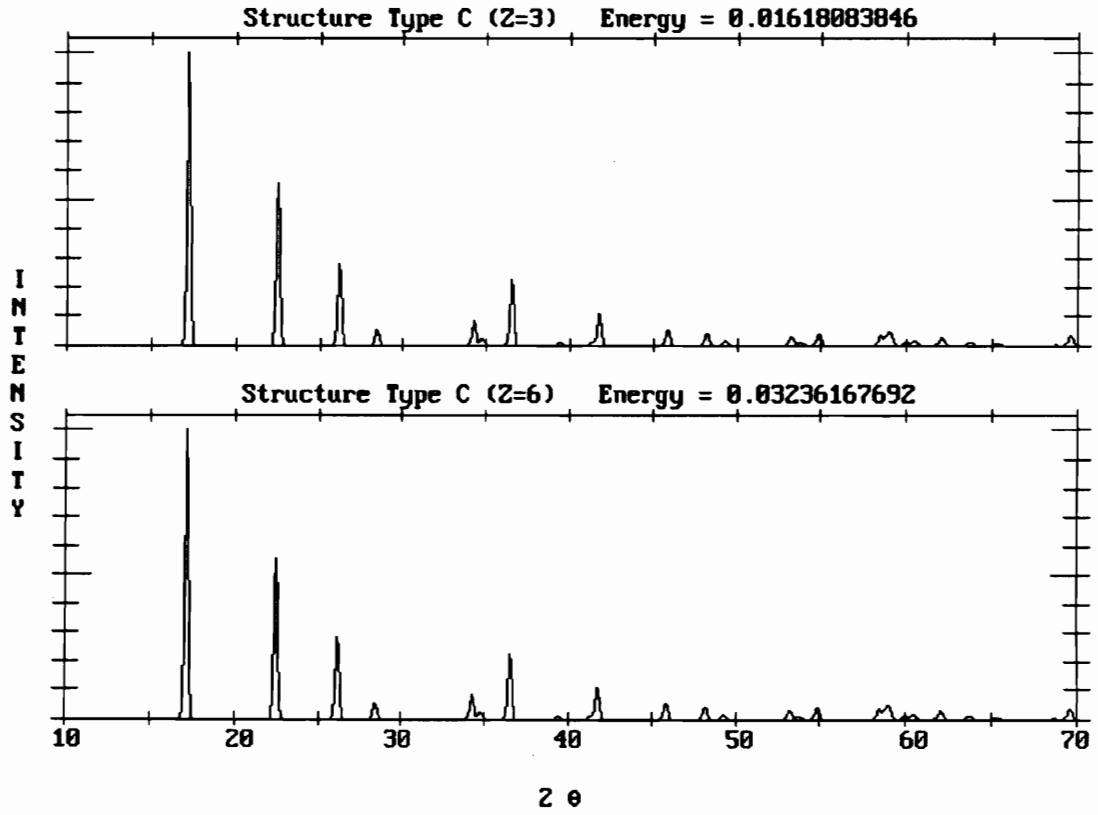


Figure 1–26. A comparison of the calculated X-ray powder diffraction patterns for structure type C generated using $Z = 3$ and $Z = 6$.

APPENDIX 1A

Methods of Minimization

Introduction

When atoms combine to form a material, they do so in manner such that the total energy of the resulting configuration is minimized. Consequently, if we have a function that yields energy as a function of atomic positions and if we minimize this function, we will have the analog to the configuration in nature. Therefore a basic understanding of the methods for finding the minimum value of a function is important in mineralogy. The process of minimization is also fundamental to numerous scientific fields from quantum mechanics to statistics.

Given some function, $f(\mathbf{x})$, where $\mathbf{x} \in \Re^n$ and $f(x) \in \Re$, find a point, $\mathbf{x}'$, where the gradient of the function is zero. Having arrived there, $\mathbf{x}'$ then represents the location of an extreme for the function. The extreme for $f(\mathbf{x})$ can be a maximum or a minimum. If $\mathbf{x}'$ corresponds to a minimum value of $f(\mathbf{x})$, then $\mathbf{x}'$ is called a **minimizer** of $f(\mathbf{x})$. Following the specification of some starting location, $\mathbf{x}_0$, the minimization techniques that will be discussed all attempt to get closer and closer to the minimizer by gathering some sort of information about $f(\mathbf{x})$ from which a **model** is formed. The model is then used in helping to locate the next choice for the approximation to the minimizer, $\mathbf{x}_{i+1}$. The process continues at each successive value of $\mathbf{x}_k$ until an acceptable minimizer has been found. All of the methods discussed here belong in the category of **local minimizers** in that they tend to locate a minimum near the original starting position, $\mathbf{x}_0$.

1A.1 Steepest Descent

The first minimization technique that will be discussed is called the method of steepest descent and will be presented through the use of an example. Therefore,

consider $\mathbf{x} \in \Re^2$ so that $\mathbf{x}^t = [\, x_1 \quad x_2 \,]$ and the function:

$$f(\mathbf{x}) = 2x_1^2 + 3x_1x_2 + 4x_2^2 + 5x_1 - 3x_2 + 14. \tag{1}$$

The task is to find the minimizer, $\mathbf{x}'$, where the function (Equation 1) has its lowest value (minimum). To help in finding this point, a model will be created (ie. choose a function) that hopefully mimics Equation 1 at the point, $\mathbf{x}_0$. The first type of model considered is a **linear model**. The model function, $f_m(\mathbf{x})$, represents a a linear approximation of $f(\mathbf{x})$ based at $\mathbf{x}_0$. However, in the beginning, a starting value for $\mathbf{x}_0$ must be chosen at which the initial linear model is built. For illustration purposes, choose $\mathbf{x}_0^t = [\, 1 \quad 2 \,]$. Substituting these values for x_1 and x_2 into Equation 1 gives $f(\mathbf{x}_0) = 37$.

At the **minimum** of any differentiable function the first derivative is zero. If the function includes more than one parameter, the first derivative involves taking the partial derivatives of the function with respect to each parameter. The first derivative now has the form of a vector which is called the **gradient**. For purposes of notation, the gradient of $f(\mathbf{x})$ at some point $\mathbf{x} = \mathbf{a}$ is denoted $\nabla_{\mathbf{a}} f$. Note that the gradient always points in the direction of greatest increase of the function. For Equation 1, the gradient is given as

$$\begin{aligned} \nabla_{\mathbf{a}} f &= \left[\begin{array}{c} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{array} \right] \Bigg|_{\mathbf{x}=\mathbf{a}} \\ &= \left[\begin{array}{c} 4x_1 + 3x_2 + 5 \\ 3x_1 + 8x_2 - 3 \end{array} \right] \Bigg|_{\mathbf{x}=\mathbf{a}}, \end{aligned} \tag{2}$$

where at the point $\mathbf{a} = \mathbf{x}_0$,

$$\nabla_{\mathbf{x}_0} f = \left[\begin{array}{c} 15 \\ 16 \end{array} \right].$$

In order to create a linear model, recall that a linear function has the form $f_m(z) = az + b$, where a is the slope, b is the intercept and z is a point in the

model subspace. Written in terms of gradients, the slope and intercept become

$$a = \nabla_{\mathbf{x}_0} f \text{ and } b = f(\mathbf{x}_o)$$

or $f_m(\mathbf{z}) = (\nabla_{\mathbf{x}_0} f)^t \mathbf{z} + f(\mathbf{x}_o)$, where $\mathbf{x}_o$ is the point in the function space where the model is constructed. By substituting $\nabla_{\mathbf{x}_0} f$ from Equation 2, a linear model can be written as

$$f_m(\mathbf{z}) = [\,15 \quad 16\,] \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + 37 \text{ based at } \mathbf{x}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}. \tag{4}$$

The task at hand is in choosing a value for $\mathbf{z}$. This point represents the step or shift we will make along the direction of the model. Since $-\nabla f$ is in the direction of steepest descent, it is natural to step in that direction. In other words, look for the minimizer in the direction that points downhill from $\mathbf{x}_0$. This is the basic idea behind the **method of steepest descent** (Dennis and Schnabel, 1983). As an example, two iterations of the steepest descent procedure will be presented (Table 1–10 and Table 1–11) using Equation 1 starting with the model represented by Equation 4.

In each table, the point $\mathbf{z}$ (column 2) represents the step vector chosen in the model subspace that lies along the $-\nabla_{\mathbf{x}} f$ (column 5) which is calculated according to Equation 2. The point $\mathbf{x}$ (column 3) is the new point in the function space resulting from the step, $\mathbf{z}$. The tables also include the functional value at this point, $f(\mathbf{x})$ (column 4), and the gradient at $\mathbf{x}$, $\nabla_{\mathbf{x}} f$ (column 5). Finally, the tables include the value predicted from the model (Equation 4) at the point $\mathbf{z}$, $f_m(\mathbf{z})$ (column 6).

Table 1–10. Illustration of a single iteration of the steepest descent method applied to Equation 1

Step 1	$\mathbf{z}^t$	$\mathbf{x}^t$	$f(\mathbf{x})$	$\nabla_{\mathbf{x}} f^t$	$f_m(\mathbf{z})$
a	—	[1 2]	37	[15 16]	—
b	[−15 −16]	[−14 −14]	1,750	[−93 −157]	-440
c	[−15/2 −16/2]	[−6.5 −6]	345	[−39 −70.5]	-199.5
d	[−15/11 −16/11]	[−.364 .546]	**11.4**	[5.18 .276]	-2.7

Observe from Table 1–10 that $f(\mathbf{x})$ did not decrease in value until Step 1d. Up until this point, smaller and smaller increments of $\mathbf{z}$ were chosen until $f(\mathbf{x})$ decreased. All of the choices for $\mathbf{z}$ were chosen along the $-\nabla_{\mathbf{x}_0} f$ direction. Instead of guessing at values for $\mathbf{z}$, line search methods could have been used to obtain more appropriate step values (Dennis and Schnabel, 1983). It is important to remember that the model remained fixed at $\mathbf{x}_0$ while searching for a reduction in the functional value for Equation 1. Notice how the values given in column 4 and column 6 differ. The difference between these columns is a measure of how good the model mimics the function. For this example there is not very good agreement.

Because the value of the function decreased in step 1d, the new point $\mathbf{x}^t = [-.364 \quad .546]$ is chosen whereby a new model is constructed as the search continues for the minimizer. In other words,

$$f_m^{new}(\mathbf{z}) = [5.18 \quad .276] \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + 11.4 \text{ based at } \mathbf{x}_1 = \begin{bmatrix} -.364 \\ .546 \end{bmatrix}.$$

Table 1–11. Illustration of a second iteration of the steepest descent applied to Equation 1

Step 2	$\mathbf{z}^t$	$\mathbf{x}^t$	$f(\mathbf{x})$	$\nabla_{\mathbf{x}} f^t$	$f_m(\mathbf{z})$
a	—	[−.364 .546]	11.4	[5.18 .276]	—
b	[−5.18 −.276]	[−5.55 .27]	42.8	[−16.37 −17.48]	-15.5
c	[−5.18/2 −.276/2]	[−2.96 .41]	12.51	[−5.596 −8.60]	-2.1
d	[−5.18/5 −.276/5]	[−1.40 .50]	**8.35**	[.87 −3.28]	6.02

As indicated in Table 1–11, the functional value of Equation 1 has again been reduced. Similarly, a new model is built at the point $\mathbf{x}$ where the reduction occurred. Using this new model,

$$f_m^{newer}(\mathbf{z}) = [\,.87 \quad -3.275\,]\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + 8.35 \text{ based at } \mathbf{x}_2 = \begin{bmatrix} -1.40 \\ .50 \end{bmatrix},$$

the search continues in a similar fashion until the $\nabla_{\mathbf{x}} f$ (column 5) is close to zero. The point $\mathbf{x}$ where this occurs will then represent a minimizer for Equation 1.

1A.2 Quadratic Functions

Re-inspection of Equation 1 shows that the function can be rewritten in the following form:

$$f(\mathbf{x}) = \frac{1}{2}[\,x_1 \quad x_2\,]\begin{bmatrix} 4 & 3 \\ 3 & 8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + [\,5 \quad -3\,]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 14. \tag{5}$$

In other words, Equation 1 represents a **quadratic function** in $\mathbf{x}$. The 2×2 symmetric matrix of the quadratic function shown in Equation 5 is called the **Hessian matrix** and is given by:

$$H_0 = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}|_{x=0} & \frac{\partial^2 f}{\partial x_1 \partial x_2}|_{x=0} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}|_{x=0} & \frac{\partial^2 f}{\partial x_2^2}|_{x=0} \end{bmatrix}.$$

Note that H_0 indicates that the Hessian is evaluated at the origin of the function. Furthermore, Equation 5 represents a quadratic function because H_0 is a constant. In addition to being symmetric, H_0 is a **Positive definite** matrix. A positive definite matrix is one where all eigenvalues are positive. A quick check that a symmetric matrix is positive definite is to evaluate the determinate of each of the principal minors of the matrix. The matrix is positive definite if each of these determinants is positive. For the 2×2 matrix given in Equation 5, the first principal minor ([4]) has a determinant of 4 and the second principal minor (the 2×2 matrix itself) has a determinant of 23. Therefore, the 2×2 matrix of Equation 5 is positive definite.

The last two terms (linear terms) of the quadratic function given in Equation 5 represent

$$\nabla_0 f^t = [5 \quad -3] \text{ and } f(0) = 14.$$

Thus, the general form of a quadratic function centered at the origin is given by

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^t H_0 \mathbf{x} + (\nabla_0 f)^t \mathbf{x} + f(0).$$

For quadratic functions centered at $\mathbf{x}_0$, the general form is

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^t H_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) + (\nabla_{\mathbf{x}_0} f)^t(\mathbf{x} - \mathbf{x}_0) + f(\mathbf{x}_0).$$

Like before, find the minimizer for the function where it has a minimum value and its' gradient is zero. Similarly, the gradient or derivative must be evaluated. Note that the following methods for obtaining the derivative apply to any quadratic function. Our task is to find:

$$\nabla_{\mathbf{a}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1}|_{\mathbf{a}} \\ \frac{\partial f}{\partial x_2}|_{\mathbf{a}} \end{bmatrix},$$

where **a** is a point in the function space.

To evaluate the gradient of a quadratic function, use will be made of the product rule. The derivative of the quadratic term involving H_0 is:

$$\begin{aligned} \frac{\partial\left(\frac{1}{2}[x_1 \quad x_2] H_0 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)}{\partial x_1} &= \frac{1}{2}[1 \quad 0] H_0 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{1}{2}[x_1 \quad x_2] H_0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= [1 \quad 0] H_0 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned}$$

because H_0 is symmetric. In an analogous manner:

$$\frac{\partial\left(\frac{1}{2}[x_1 \quad x_2] H_0 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)}{\partial x_2} = [0 \quad 1] H_0 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Combining the two results above gives:

$$\nabla\left(\tfrac{1}{2}\begin{bmatrix}x_1 & x_2\end{bmatrix}H_0\begin{bmatrix}x_1\\x_2\end{bmatrix}\right) = \begin{pmatrix}\begin{bmatrix}1 & 0\end{bmatrix}H_0\begin{bmatrix}x_1\\x_2\end{bmatrix}\\[2ex]\begin{bmatrix}0 & 1\end{bmatrix}H_0\begin{bmatrix}x_1\\x_2\end{bmatrix}\end{pmatrix}$$
$$= \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}H_0\begin{bmatrix}x_1\\x_2\end{bmatrix}$$
$$= H_0\begin{bmatrix}x_1\\x_2\end{bmatrix}.$$

Similarly,

$$\nabla\left((\nabla_0 f)^t\begin{bmatrix}x_1\\x_2\end{bmatrix}\right) = \begin{pmatrix}(\nabla_0 f)^t\begin{bmatrix}1\\0\end{bmatrix}\\[2ex](\nabla_0 f)^t\begin{bmatrix}0\\1\end{bmatrix}\end{pmatrix}$$
$$= (\nabla_0 f)^t\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$
$$= \nabla_0 f^t.$$

Thus, the **gradient of a quadratic function** at a point **a** is

$$\nabla_{\mathbf{a}} f = H_0\mathbf{a} + \nabla_0 f. \tag{6}$$

Note that the gradient at **a**, $\nabla_{\mathbf{a}} f$, differs from the gradient at the origin, $\nabla_0 f$, by $H_0\mathbf{a}$. To find the location of a minimum of the quadratic model, set Equation 6 equal to zero and solve for the minimizer, **a**, so that

$$\mathbf{a} = -H_0^{-1}\nabla_0 f. \tag{7}$$

For the example given by Equation 5

$$H_0^{-1} = \frac{1}{23}\begin{bmatrix}8 & -3\\-3 & 4\end{bmatrix},$$

Therefore, from Equation 7 the location for a minimizer of Equation 5 is

$$a = -\frac{1}{23}\begin{bmatrix}8 & -3\\-3 & 4\end{bmatrix}\begin{bmatrix}5\\-3\end{bmatrix}$$
$$= -\frac{1}{23}\begin{bmatrix}49\\-27\end{bmatrix}$$
$$\approx \begin{bmatrix}-2.13\\1.17\end{bmatrix},$$

with a minimum value of ~6.913.

1A.3 Newton-Raphson Method

If Equation 5 were not quadratic, solution for the minimum would not have been so straight forward. One type of minimization technique commonly used for this case is called the Newton-Raphson Method (Dennis and Schnabel, 1983). Even though Equation 5 is quadratic, the basic concepts of the Newton-Raphson method can still be illustrated. In a manner analogous to the steepest descent method, a model is constructed at some point $\mathbf{x}_0$ in the function space. However, the model now created at $\mathbf{x}_0$ is quadratic instead of linear. In turn, the minimum of the quadratic model is found where the minimizer is used as a step in finding the minimum of the function. The quadratic model for the example discussed in Section 1A.1 will have the form

$$\begin{aligned} f_m(\mathbf{z}) &= \frac{1}{2}\mathbf{z}^t H_{x_0}\mathbf{z} + (\nabla_{x_0} f)^t\mathbf{z} + f(x_0) \\ &= \frac{1}{2}\mathbf{z}^t H_{\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]}\mathbf{z} + (\nabla_{\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]} f)^t\mathbf{z} + f(\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]) \\ &= \frac{1}{2}\begin{bmatrix} z_1 & z_2 \end{bmatrix}\begin{bmatrix} 4 & 3 \\ 3 & 8 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} 15 & 16 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + 37. \end{aligned}$$

Because the model is quadratic, the location of the minimizer for the model at the starting point $x_0^t = [\,1 \quad 2\,]$ is given by Equation 7 as

$$\begin{aligned} \mathbf{z} &= -H_0^{-1}\nabla_0 f \\ &= -H_{x_0}^{-1}\nabla_{x_0} f \\ &= -H_{\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]}^{-1}\nabla_{\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]} f \\ &= -\frac{1}{23}\begin{bmatrix} 8 & -3 \\ -3 & 4 \end{bmatrix}\begin{bmatrix} 15 \\ 16 \end{bmatrix} \\ &= -\frac{1}{23}\begin{bmatrix} 72 \\ 19 \end{bmatrix}. \end{aligned}$$

The point $\mathbf{z}$ represents the location of a minimizer of the quadratic model and is called the **Newton step** of the model. Thus, the new point, $\mathbf{s}$, called the **Newton**

point is given by

$$\begin{aligned}
\mathbf{s} &= \mathbf{x}_0 + \mathbf{z} \\
\mathbf{s} &= \mathbf{x}_0 - H_0^{-1}\nabla_0 f \\
&= \begin{bmatrix} 1 \\ 2 \end{bmatrix} + -\frac{1}{23}\begin{bmatrix} 72 \\ 19 \end{bmatrix} \\
&= -\frac{1}{23}\begin{bmatrix} 49 \\ -27 \end{bmatrix} \\
&\approx \begin{bmatrix} -2.13 \\ 1.17 \end{bmatrix}.
\end{aligned}$$

Because the function (Equation 5) is quadratic, the Newton step is directly to the minimizer of the function. Even at an alternate starting value for $\mathbf{x}_0$, the Newton step would have directly lead to the minimizer of the function. Had the function not been quadratic, a new model would then be built at the Newton point, if it were downhill, and iterating would continue until the gradient of the function is approximately zero. When the gradient is "zero" and H is positive definite, a minimum is found.

1A.4 Quasi-Newton Method

Recall from the discussion above on the Newton-Raphson method of minimization that the hessian matrix, $H_{\mathbf{x}_o}$, computed according to

$$H_{\mathbf{x}_o} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}|_{\mathbf{x}=\mathbf{x}_o} & \frac{\partial^2 f}{\partial x_1 \partial x_2}|_{\mathbf{x}=\mathbf{x}_o} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}|_{\mathbf{x}=\mathbf{x}_o} & \frac{\partial^2 f}{\partial x_2^2}|_{\mathbf{x}=\mathbf{x}_o} \end{bmatrix},$$

requires that the second derivatives of $f(\mathbf{x})$ be evaluated. However, the second derivatives of a function such as Equation 7 in Chapter 1 are too complicated to evaluate. A function of this type therefore requires a modified approach be used in its minimization. Alternative methods for finding the minimum of a function without having to evaluate second derivatives are classified as **quasi-Newton methods**. Otherwise known as **variable metric** or **secant** methods, the quasi-Newton method builds an approximation to the Hessian gradually as the iterations

proceed. To develop the procedure, recall that Equation 6 gives the gradient of a quadratic function at some point **a**. For another point, **b**, in the function space:

$$\nabla_{\mathbf{a}} f = H_0 \mathbf{a} + \nabla_0 f$$

$$\nabla_{\mathbf{b}} f = H_0 \mathbf{b} + \nabla_0 f.$$

Taking the difference of the two gradients:

$$\nabla_{\mathbf{b}} f - \nabla_{\mathbf{a}} f = H_0 \mathbf{b} - H_0 \mathbf{a} = H_0(\mathbf{b} - \mathbf{a}) \tag{8}$$

If we define $\mathbf{s} = \mathbf{b} - \mathbf{a}$ and $\mathbf{y} = \nabla_{\mathbf{b}} f - \nabla_{\mathbf{a}} f$ leeds to the so called **Secant Equation** (Dennis and Schnabel, 1983):

$$\mathbf{y} = H\mathbf{s}. \tag{9}$$

From Equation 8, the gradient at two different points gives information about the Hessian matrix, H. The motivation for obtaining alternate information about the Hessian is to eliminate calculating the second derivatives of the function to be minimized. Approximations to the first derivatives are readily obtained by using difference methods that will be discussed later. The notation H_0 has been dropped because the discussion now addresses any general Hessian matrix, H.

A mental picture of the secant equation can be formed by imagining an xy plot of the derivative of some function, $f'(x)$. If a line is drawn that intersects this function at two points, a and b, then the slope of this secant line is

$$\frac{\Delta y}{\Delta x} = \frac{f'(b) - f'(a)}{b - a} \approx f''(a) \approx f''(b).$$

Rearranging the above equation results in

$$f'(b) - f'(a) = f''(a)(b - a),$$

which is of the same form as Equation 8.

Now, lets assume that we are unable or unwilling to calculate H for Equation 1 but that the gradient can be calculated according to,

$$\nabla_{\mathbf{a}} f = \begin{pmatrix} 4x_1 + 3x_2 + 5 \\ 3x_1 + 8x_2 - 3 \end{pmatrix} |_{\mathbf{x}=\mathbf{a}}$$

Lets choose three points

$$0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \ \mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \ \mathbf{a} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

so that the gradient at each of these points is as follows:

$$\nabla_0 f = \begin{bmatrix} 5 \\ -3 \end{bmatrix}; \ \nabla_{\mathbf{b}} f = \begin{bmatrix} 9 \\ 0 \end{bmatrix}; \ \nabla_{\mathbf{a}} f = \begin{bmatrix} 8 \\ 5 \end{bmatrix}.$$

By applying the secant equation (Equation 9):

$$H\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 9 \\ 0 \end{bmatrix} - \begin{bmatrix} 5 \\ -3 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$
$$H\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 8 \\ 5 \end{bmatrix} - \begin{bmatrix} 5 \\ -3 \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \end{bmatrix}$$

Recall that the columns of a matrix can be obtained according to

$$H \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix}, H \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{12} \\ h_{22} \end{bmatrix}, \text{ etc.},$$

so by applying the secant equation to the above three points, the Hessian matrix, H has effectively been obtained. In other words from before,

$$H \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix}$$
$$H \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \end{bmatrix} = \begin{bmatrix} h_{12} \\ h_{22} \end{bmatrix}$$

Therefore,

$$H = \begin{bmatrix} 4 & 3 \\ 3 & 8 \end{bmatrix}.$$

If the function is not quadratic, then the H matrix obtained by the above method is only an approximation because H is not constant. In this example,

H is constant because the function is quadratic. In other words, you would get a different H at different points depending on how "non-quadratic" the function behaves. This is why H must be updated as the search continues for a minimizer. As illustrated above, the points where the secant equation is applied should be linearly independent if possible. Keep in mind that the approximated Hessian is only good in the direction of the two points chosen for evaluation in the secant equation.

Note that approximations of H that are constructed using the secant method utilize the first derivatives of the function, or gradient. Recall that the approximation is necessary because the second derivatives of many functions cannot be evaluated. However, there are certain instances when even the first derivatives of a function are impossible to be empirically determined. These cases require using methods of numerically calculating the gradient. Recall the general form for the gradient of a function, f, at the point $\mathbf{a}$ is

$$\nabla_{\mathbf{a}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1}|_{\mathbf{a}} \\ \frac{\partial f}{\partial x_2}|_{\mathbf{a}} \\ \frac{\partial f}{\partial x_3}|_{\mathbf{a}} \\ \vdots \\ \frac{\partial f}{\partial x_n}|_{\mathbf{a}} \end{bmatrix}$$

where x_n represent the n^{th} variable of the function. Recall that for a function of a single variable, x, the derivative is obtained according to

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} \rightarrow f'(x) \qquad \text{as } \Delta x \rightarrow 0. \tag{10}$$

This is called the **Forward Difference Method** (Dennis and Schnabel, 1983). It turns out that a good choice for Δx for applications in computer programs is given by

$$\Delta x = 10^{-K/2} \times F_{usual},$$

where the precision is K digits and F_{usual} is a usual value of x.

Another method for numerically calculating the gradient of a function is given by the **Central Difference Method** (Dennis and Schnabel, 1983),

$$\frac{f(x+\Delta x)-f(x-\Delta x)}{2\Delta x} \rightarrow f'(x) \qquad \text{as } \Delta x \rightarrow 0$$

This method is slightly better in finding a minimum because the gradient is evaluated on two sides of the point x but is more costly in computation time than Equation 10 because the function must be evaluated twice.

As mentioned above, the value of the Hessian matrix, H, is dependent on the points chosen to evaluate the secant equation. Consequently, H must be modified or updated as the search continues for a minimizer. The various methods used in updating H will now be discussed. The conditions imposed on the updating methods fall under the category of **Secant Updating** (Dennis and Schnabel, 1983). Begin by defining

$$\mathbf{s} = \mathbf{x}_+ - \mathbf{x}_c \text{ and } \mathbf{y} = \nabla_+ - \nabla_c,$$

where $\mathbf{x}_+$ is a new point and $\mathbf{x}_c$ is the current point so that $\mathbf{s}$ represents the current step. Similarly, ∇_+ is the gradient of the function at the new point and ∇_c is the gradient at the current point and $\mathbf{y}$ is referred to as the yield of the current step (Dennis and Schnabel, 1983). There are two conditions that should be satisfied when H is updated. First, the new Hessian, H_+, should satisfy the secant equation. In other words, find a H_+ such that $H_+ s = y$. Second, H_+ should be a lot like H_c or "close" to H_c. An alternative statement is that the difference $H_+ - H_c$ should be "small".

First, consider the condition - "small". The smallest difference in something is of course zero. If A and B are $n \times n$ matrices and $\mathbf{x}_1 - \mathbf{x}_n$ are n linearly independent

vectors (ie. form basis of R^n) and if $A\mathbf{x}_i = B\mathbf{x}_i$ for all i, then $(A-B)\mathbf{x}_i = 0$ for all i or $A - B = 0$. To prove this, consider $\mathbf{x} \in R^n$, so that $\mathbf{x} = \sum a_i \mathbf{x}_i$ for some a_i, and therefore,

$$(A-B)\mathbf{x} = (A-B)\sum a_i \mathbf{x}_i = \sum a_i (A-B)\mathbf{x}_i = 0.$$

In other words,

$$(A-B)\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix} = 0 \rightarrow \text{ } 1^{st} \text{ column of } (A-B) \text{ is a zero column}$$

$$(A-B)\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} = 0 \rightarrow \text{ } 2^{nd} \text{ column of } (A-B) \text{ is a zero column}$$

$$(A-B)\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix} = 0 \rightarrow \text{ } 3^{rd} \text{ column of } (A-B) \text{ is a zero column}$$

$$\vdots$$

$$(A-B)\begin{pmatrix} \vdots \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = 0 \rightarrow \text{ last column of } (A-B) \text{ is a zero column}$$

Therefore, since $(A-B) = 0$, then $A = B$. It then follows that the assertion can be made that two matrices, A and B, are "close" if $A\mathbf{x}_i = B\mathbf{x}_i$ for all i except for one i !

As an illustration consider the vectors

$$\begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix},$$

that are linearly independent and thus form a basis for R^3, along with the matrices

$$A = \begin{bmatrix} 2 & 1 & 4 \\ 4 & 2 & 8 \\ 6 & 3 & 12 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

To examine whether A and B are "close", $A\mathbf{x}_i = B\mathbf{x}_i$ is computed for each vector,

$$\begin{bmatrix} 2 & 1 & 4 \\ 4 & 2 & 8 \\ 6 & 3 & 12 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 & 4 \\ 4 & 2 & 8 \\ 6 & 3 & 12 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 & 4 \\ 4 & 2 & 8 \\ 6 & 3 & 12 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \neq \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Thus, A and B are "close" because they agree on all but one of the chosen basis vectors. Alternatively stated, the same results are obtained in only two of the three dimensions. For two matrices to be "close" implies that the difference $A - B$ is "small." Closer examination of A indicates that both the columns and rows of the matrix are not linearly independent. The space generated by the rows of A is 1 dimensional so that the matrix is by definition a **rank 1 matrix**. Rank 1 matrices are small.

Rank 1 matrices can be written as the **outer product** of two vectors. For example, A can be written as

$$A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 2 & 1 & 4 \end{bmatrix}$$

or

$$A = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \begin{bmatrix} 1 & 1/2 & 2 \end{bmatrix}$$

illustrating that the representation is not unique. The outer product will be used to update H because it always results in a small matrix. In other words, the rank 1 update assures us that the difference between the current Hessian, H_c, and the updated Hessian, H_+ will be small.

Consider $\mathbf{uv}^t$ where $\mathbf{u}, \mathbf{v} \in R^n$, then $\mathbf{uv}^t$ is $n \times n$ and rank 1. Then

$$(\mathbf{uv}^t)\mathbf{x} = \mathbf{u}(\mathbf{v}^t\mathbf{x}),$$

where by inspection, $\mathbf{u}$ gives the direction of the resulting vector and $(\mathbf{v}^t\mathbf{x})$ affects only the length (scaler) of the vector. The above expression has the property of wiping out other dimensions thereby projecting space onto 1 dimension (recall that rank 1 matrices are 1 dimensional). Expansion of the above expression, according to

$$\begin{aligned}
(\mathbf{uv}^t)\mathbf{x} &= \left[\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 & \cdots & v_n \end{bmatrix} \right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \\
&= \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 & \cdots & u_1v_n \\ u_2v_1 & u_2v_2 & u_2v_3 & \cdots & u_2v_n \\ u_3v_1 & u_3v_2 & u_3v_3 & \cdots & u_3v_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_nv_1 & u_nv_2 & u_nv_3 & \cdots & u_nv_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \\
&= x_1v_1 \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} + x_2v_2 \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} + x_3v_3 \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} + \cdots + x_nv_n \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} \\
&= \mathbf{u}(\mathbf{v}^t\mathbf{x}),
\end{aligned}$$

illustrates that all space has been projected onto the 1 dimension represented by the vector $\mathbf{u}$.

Now consider the second desired condition that the updated Hessian matrix, H_+, must satisfy the secant equation. From Equation 9

$$H_+\mathbf{s} = \mathbf{y} \tag{11}$$

$$(H_c + \mathbf{uv}^t)\mathbf{s} = \mathbf{y}$$

$$H_c\mathbf{s} + (\mathbf{uv}^t)\mathbf{s} = \mathbf{y},$$

so that

$$(\mathbf{uv}^t)\mathbf{s} = \mathbf{y} - H_c\mathbf{s}.$$

From before, since $\mathbf{u}$ gives the direction of the resulting vector, requires that $\mathbf{u}$ must be parallel to $(\mathbf{y}-H_c\mathbf{s})$. If $\mathbf{u} = \mathbf{y}-H_c\mathbf{s}$, then $\mathbf{v}^t\mathbf{s}$ must equal 1. Consequently, if $\mathbf{v} = \frac{\mathbf{s}}{\mathbf{s}^t\mathbf{s}}$, then $\mathbf{v}^t\mathbf{s} = 1$. Therefore, one choice for updating H_c that satisfies Equation 11 (Secant Equation) is $\mathbf{u} = \mathbf{y} - H_c\mathbf{s}$ and $\mathbf{v} = \frac{\mathbf{s}}{\mathbf{s}^t\mathbf{s}}$ which yields

$$H_+ = H_c + \frac{(\mathbf{y} - H_c\mathbf{s})\mathbf{s}^t}{\mathbf{s}^t\mathbf{s}}. \tag{12}$$

This is known as the **Broyden Update** or **Secant Update** (Dennis and Schnabel, 1983). It has the property that it only updates in one direction and produces H_+ that satisfies the secant equation. It is worth noting that H_+ in Equation 12 satisfies Equation 9. This is verified by multiply $\mathbf{s}$ through Equation 12 and canceling the $\mathbf{s}^t$ term which results in the following:

$$\begin{aligned} H_+\mathbf{s} &= H_c\mathbf{s} + (\mathbf{y} - H_c\mathbf{s}) \\ &= \mathbf{y} + H_c\mathbf{s} - H_c\mathbf{s} \\ &= \mathbf{y}. \end{aligned}$$

As an example of the Broyden Update, consider the following:

$$H_c = \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix}, \qquad \mathbf{s} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}.$$

Therefore, according to Equation 12

$$\begin{aligned} H_+ &= \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} + \frac{\left(\begin{bmatrix} -3 \\ 4 \end{bmatrix} - \begin{bmatrix} 7 \\ 6 \end{bmatrix} \right) [2 \quad 1]}{5} \\ &= \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} + \frac{\begin{bmatrix} -20 & -10 \\ -4 & -2 \end{bmatrix}}{5} \\ &= \begin{bmatrix} -1 & -1 \\ 1/5 & 18/5 \end{bmatrix}. \end{aligned}$$

which in turn can be used to verify the secant condition (Equation 9),

$$\begin{bmatrix} -1 & -1 \\ 1/5 & 18/5 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}.$$

However, H_+ is not symmetric nor is it positive definite. This is a deficiency in the Broyden update (Dennis and Schnabel, 1983).

A **symmetric** rank one update can be obtained by defining the following:

$$H_+ = H_c + \frac{(\mathbf{y} - H_c\mathbf{s})(\mathbf{y} - H_c\mathbf{s})^t}{(\mathbf{y} - H_c\mathbf{s})^t\mathbf{s}}. \tag{13}$$

This is known as the **SR1 Update** (Dennis and Schnabel, 1983). Again, observe that Equation 13 will reduce to Equation 11 which is simply the secant equation (Equation 9). In other words,

$$\begin{aligned} H_+ &= H_c + \frac{(\mathbf{y} - H_c\mathbf{s})(\mathbf{y} - H_c\mathbf{s})^t}{(\mathbf{y} - H_c\mathbf{s})^t\mathbf{s}} \\ H_+\mathbf{s} &= H_c\mathbf{s} + (\mathbf{y} - H_c\mathbf{s}) \\ &= \mathbf{y} + H_c\mathbf{s} - H_c\mathbf{s} \\ &= \mathbf{y}. \end{aligned}$$

Consequently, both the Broyden update and the SR1 update, are nothing more than different solutions to the secant equation given by Equation 11.

Lets apply the SR1 update to the example given above for the Broyden update and examine the results:

$$\begin{aligned} H_+ &= \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} + \frac{\left(\begin{bmatrix} -10 \\ -2 \end{bmatrix} \begin{bmatrix} -10 & -2 \end{bmatrix} \right)}{22} \\ &= \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} + \frac{1}{22} \begin{bmatrix} -100 & 20 \\ 20 & 4 \end{bmatrix} \\ &= \frac{1}{11} \begin{bmatrix} 33 & 11 \\ 11 & 44 \end{bmatrix} - \frac{1}{11} \begin{bmatrix} 50 & 10 \\ 10 & 2 \end{bmatrix} \\ &= \frac{1}{11} \begin{bmatrix} -17 & 1 \\ 1 & 42 \end{bmatrix}. \end{aligned}$$

As before, the new symmetric H_+ satisfies the secant equation

$$\frac{1}{11} \begin{bmatrix} -17 & 1 \\ 1 & 42 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 4 \end{bmatrix},$$

but is still not positive definite.

Recall the requirements imposed on the updated Hessian matrix H_+:

1. $H_+\mathbf{s} = \mathbf{y}$
2. small change in H_c
3. H_+ is symmetric
4. H_+ is positive definite.

The Broyden update satisfies requirements 1 and 2. The SR1 update satisfies requirements 1, 2 and 3. Satisfaction of all four requirements is not possible using a rank 1 update. An obvious solution is to move to a **rank 2 update**. However, even this form of updating has its drawbacks because the updated Hessian will lose some information about where we have been in the parameter space (Dennis and Schnabel, 1983).

An updating procedure which satisfies all four requirements is known as the **BFGS Update**. This method is in practise the most successful secant update for the Hessian (Dennis and Schnabel, 1983) and is defined as follows:

$$\begin{aligned} H_+ &= H_c + \frac{\mathbf{y}\mathbf{y}^t}{\mathbf{y}^t\mathbf{s}} - \frac{(H_c\mathbf{s})(H_c\mathbf{s})^t}{(H_c\mathbf{s})^t\mathbf{s}} \\ &= H_c + \frac{\mathbf{y}\mathbf{y}^t}{\mathbf{y}^t\mathbf{s}} - \frac{H_c\mathbf{s}\mathbf{s}^tH_c}{\mathbf{s}^tH_c\mathbf{s}}. \end{aligned} \tag{14}$$

Algebraic manipulation will again show that Equation 14 is yet another form of the secant equation (Equation 9). A condition required for the BFGS updating procedure is that $y^ts > 0$ (Dennis and Schnabel, 1983). If for some reason this condition fails, then skip the update.

1A.5 Trust Region

From the secant equation, the quasi-Newton step is given as $\mathbf{s} = H_+^{-1}\mathbf{y}$. However, there may be times when a better choice can be made for both the

direction and the step length. This would result if the quadratic model does not adequately model the function in a region containing the full quasi–Newton step. For example, if the current point is $\mathbf{x}_c^t = [1 \quad 1]$ and the Newton step is $\mathbf{s}_n^t = [400 \quad 236]$, then it would be unreasonable to take this step because we don't "trust" the model over that distance. A method of keeping the step values in check is needed so that we do not jump around space hopelessly in search of the minimizer. The method discussed here for choosing a step involves the use of a trust region which will be implemented using a procedure called the double dogleg method (Dennis and Schnabel, 1983). The basic concept is that once a radius for a trust region centered at the current point, $\mathbf{x}_c$, has been chosen, determine the location of three different reference points. Connecting these points can be thought of as tracing out the shape of a "Dog leg". Where the dog leg intersects the boundary of this trust region is where we choose to take the step.

One of the points of interest, called the Newton point, has already been discussed. Likewise, the Newton step is given by the minimizer of the quadratic model based at the current point, in this case $\mathbf{x}_c$. Recall from Equation 7 that the minimizer of the quadratic model based at the point $\mathbf{x}_c$ is given by $\mathbf{z} = -H_{\mathbf{x}_c}^{-1}\nabla_{\mathbf{x}_c} f$, where $H_{\mathbf{x}_c}$ is the Hessian matrix at $\mathbf{x}_c$ and $\nabla_{\mathbf{x}_c} f$ is the gradient of the function at the point $\mathbf{x}_c$. Because the minimizer of the model is used as a step from the point $\mathbf{x}_c$ toward the minimum, the minimizer $\mathbf{z}$ is called the **Newton step** and for notation reasons is now represented by

$$\mathbf{s}^{np} = -H_{\mathbf{x}_c}^{-1}\nabla_{\mathbf{x}_c} f,$$

so that the **Newton point** is given by

$$n.p. = \mathbf{x}_c + \mathbf{s}^{np}.$$

The second point of interest is called the **Cauchy point**. The Cauchy point is

the step determined by the minimum of the model in the direction of $-\nabla_{\mathbf{x}_c} f$ (Dennis and Schnabel, 1983). This is a good direction to look for a step because this represents a downhill direction for the function, $f(\mathbf{x})$.

The Cauchy point is found by defining $-\lambda \nabla_{\mathbf{x}_c} f$ as a minimizer in the $-\nabla_{\mathbf{x}_c} f$ direction. To find this minimizer, again create a model. The model will be based at $\mathbf{x}_c$ and only in the direction of $-\nabla_{\mathbf{x}_c} f$. A quadratic model is chosen that has the form

$$\begin{aligned} M_c(-\lambda \nabla_{\mathbf{x}_c} f) &= \frac{1}{2}(-\lambda \nabla_{\mathbf{x}_c} f)^t H_{\mathbf{x}_c}(-\lambda \nabla_{\mathbf{x}_c} f) + (\nabla_{\mathbf{x}_c} f)^t(-\lambda \nabla_{\mathbf{x}_c} f) + f(\mathbf{x}_c) \\ &= \frac{\lambda^2}{2}(\nabla_{\mathbf{x}_c} f)^t H_{\mathbf{x}_c}(\nabla_{\mathbf{x}_c} f) - \lambda(\nabla_{\mathbf{x}_c} f)^t(\nabla_{\mathbf{x}_c} f) + f(\mathbf{x}_c) \end{aligned}$$

To find λ, the minimum of the model in the desired direction, set the first derivative equal to zero and solve for λ. In other words,

$$\frac{dM_c}{d\lambda} = \lambda(\nabla_{\mathbf{x}_c} f)^t H_{\mathbf{x}_c}(\nabla_{\mathbf{x}_c} f) - (\nabla_{\mathbf{x}_c} f)^t(\nabla_{\mathbf{x}_c} f) = 0,$$

so that

$$\begin{aligned} \lambda &= \frac{(\nabla_{\mathbf{x}_c} f)^t(\nabla_{\mathbf{x}_c} f)}{(\nabla_{\mathbf{x}_c} f)^t H_{\mathbf{x}_c}(\nabla_{\mathbf{x}_c} f)} \\ &= \frac{\| \nabla_{\mathbf{x}_c} f \|^2}{(\nabla_{\mathbf{x}_c} f)^t H_{\mathbf{x}_c}(\nabla_{\mathbf{x}_c} f)}. \end{aligned}$$

Finally the **Cauchy step** is given by

$$\mathbf{s}^{cp} = \left(\frac{- \| \nabla_{\mathbf{x}_c} f \|^2}{(\nabla_{\mathbf{x}_c} f)^t H_{\mathbf{x}_c}(\nabla_{\mathbf{x}_c} f)} \right) \nabla_{\mathbf{x}_c} f,$$

and the **Cauchy point** is given by

$$c.p. = \mathbf{x}_c + \mathbf{s}^{cp}.$$

Now there are two choices for possible steps to take, $\mathbf{s}^{np}$ and $\mathbf{s}^{cp}$. Because these steps are vectors, the magnitudes of each step can be calculated. Assume

an orthonormal basis so that the magnitudes of the steps, $|\mathbf{s}^{np}|$ and $|\mathbf{s}^{cp}|$, are given by

$$|\mathbf{s}| = \sqrt{\sum s_i^2}.$$

This leeds us to the development of the third point of interest for which a parameter γ is defined where

$$\gamma = \frac{|\mathbf{s}^{cp}|}{|\mathbf{s}^{np}|}.$$

In turn, γ will be used in an empirically determined formula for another parameter, η, suggested by Dennis and Mei (1979) as

$$\eta = 0.8\gamma + 0.2,$$

that is used to arrive at the third and final point of the "dogleg", called $\mathbf{s}^{n\hat{p}}$, which is given by

$$\mathbf{s}^{n\hat{p}} = \eta \mathbf{s}^{np}.$$

Note that $\mathbf{s}^{n\hat{p}}$ is colinear with $\mathbf{s}^{np}$.

Assuming an orthonormal basis, three points have been obtained and thus three magnitudes:

$$||\mathbf{s}^{np}||, \quad ||\mathbf{s}^{n\hat{p}}||, \quad ||\mathbf{s}^{cp}||.$$

These magnitudes are compared to the radius chosen for the trust region. The **trust region** is defined as a sphere of radius δ_c whose origin is the point $\mathbf{x}_c$. The acceptable step, $\mathbf{s}^+$, is given by the point where the double dogleg intersects the trust region. This is known as the **double dogleg method** (Dennis and Schnabel, 1983).

To visualize the so-called "dogleg," imagine the three vectors emanating from an origin $\mathbf{x}_c$, $\mathbf{s}^{cp}, \mathbf{s}^{n\hat{p}}$, and $\mathbf{s}^{np}$. Furthermore, the magnitude of these vectors is always in the order

$$||\mathbf{s}^{np}|| > ||\mathbf{s}^{n\hat{p}}|| > ||\mathbf{s}^{cp}||.$$

The "dogleg" is traced out by the following vectors:

$$a. \quad \mathbf{s}^{cp}+$$

$$b. \quad (\mathbf{s}^{n\hat{p}} - \mathbf{s}^{cp})+$$

$$c. \quad (\mathbf{s}^{np} - \mathbf{s}^{n\hat{p}}),$$

where the origin is at $\mathbf{x}_c$ and the endpoint is at $n.p.$. By knowing the magnitudes of the three steps and the radius of the trust region, the vector that intersects the trust region can be determined. For an illustration of the dogleg curve, see Figure 6.4.5 in Dennis and Schnabel, 1983.

For the first possibility, assume that the $\delta_c < |\mathbf{s}^{cp}|$. Therefore in this case, the trust region intersects the vector $\mathbf{s}^{cp}$. Consequently, the point along this vector that intersects the trust region must be determined. In other words, an acceptable step for this case is of the form

$$\mathbf{s}^{+} = \beta \mathbf{s}^{cp},$$

where β is the point where the trust region intersects $\mathbf{s}^{cp}$. This scaler is found by considering

$$||\mathbf{s}^{+}|| = \delta$$

$$\beta(\mathbf{s}^{cp})^t \beta(\mathbf{s}^{cp}) = \delta^2$$

$$\beta^2 ||\mathbf{s}^{cp}||^2 = \delta^2,$$

which implies that

$$\beta = \frac{\delta}{||\mathbf{s}^{cp}||}.$$

Therefore for this case, $\mathbf{s}^{+}$ is given by

$$\mathbf{s}^{+} = \delta \left(\frac{\mathbf{s}^{cp}}{||\mathbf{s}^{cp}||} \right).$$

The second possibility is that $\mathbf{s}^{cp} < \delta_c < \mathbf{s}^{\hat{np}}$. In this case, $\mathbf{s}^+$ is of the form

$$\mathbf{s}^+ = \mathbf{s}^{cp} + \beta(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp})$$

where β is obtained according to

$$(\mathbf{s}^{cp} + \beta(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp}))^t(\mathbf{s}^{cp} + \beta(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp})) = \delta^2$$

$$(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp})^t(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp})\beta^2 + 2(\mathbf{s}^{cp})^t(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp})\beta + ((\mathbf{s}^{np})^t\mathbf{s}^{cp} - \delta^2) = 0$$

Inspection of the above equation indicates that the function is quadratic whose roots, β, are obtained by

$$\beta = \frac{-2(\mathbf{s}^{cp})^t(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp}) \pm \sqrt{(2(\mathbf{s}^{cp})^t(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp}))^2 - 4((\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp})^t(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp}))((\mathbf{s}^{np})^t\mathbf{s}^{cp} - \delta^2)}}{2((\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp})^t(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp}))}$$

Note we are only interested in the positive root because this assures we are going in the downhill direction. Thus, in the case $\mathbf{s}^+$ is given by

$$\mathbf{s}^+ = \mathbf{s}^{cp} + \beta(\mathbf{s}^{\hat{np}} - \mathbf{s}^{cp}),$$

where β is defined above.

The third possibility is that $\mathbf{s}^{\hat{np}} < \delta < \mathbf{s}^{np}$. Because the vector $\mathbf{s}^{\hat{np}}$ is colinear with $\mathbf{s}^{np}$, then $\mathbf{s}^+$ is derived analogously to the first possibility discussed above and is given by

$$\mathbf{s}^+ = \delta\left(\frac{\mathbf{s}^{np}}{||\mathbf{s}^{np}||}\right).$$

The final possibility is that $\delta > \mathbf{s}^{np}$, in which case $\mathbf{s}^+$ is simply the Newton point. In other words,

$$\mathbf{s}^+ = -H_{\mathbf{x}_c}^{-1}\nabla_{\mathbf{x}_c}f.$$

Having determined an acceptable step to take, the new point

$$\mathbf{x}^+ = \mathbf{x}_c + \mathbf{s}^+$$

is used to calculate the value of the function being minimized. If $f(\mathbf{x}_+) \geq f(\mathbf{x}_c)$, then the trust region is shrunk (ie. reduce δ) and a new dogleg constructed. Otherwise, the Hessian matrix H_c is updated according to the BFGS method outlined above. The points used in constructing this update are

$$\mathbf{s} = \mathbf{x}_+ - \mathbf{x}_c$$

$$y = \nabla_+ - \nabla_c.$$

Proceed in this manner until the gradient for the function being minimized is within a given tolerance of being zero (Dennis and Schnabel, 1983).

References

Boisen, Jr., M.B. and Gibbs, G.V. (1993) A Modeling of the Structure and Compressibility of Quartz with a Molecular Potential and its Transferability to Cristobalite and Coesite. Physics and Chemistry of Minerals, 20, 123-135.

Catlow, C.R.A. and Cormack, A.N. (1987) Computer modeling of silicates. International Reviews in Physical Chemistry, 6, 227-250.

Chelikowsky, J.R.; King, Jr., H.E. and Glinnemann, J. (1990) Interatomic potentials and the structural properties of silicon dioxide under pressure. Physical Review B, 41, 10 866-10 869.

Deem, M.W. and Newsam, J.M. (1992) Framework Crystal Structure Solution by Simulated Annealing: Test Application to Known Zeolite Structures. The Journal of the Chemical Society, 114, 7189-7198.

Dennis, Jr., J.E. and Mei, H.H.W. (1979) Two new unconstrained optimization algorithms which use function and gradient values. Journal of Optimization Theory Applications, 28, 453-482.

Dennis, Jr., J.E. and Schnabel, R.B. (1983) Numerical Methods for Unconstrained Optimization and Nonlinear Equations. New Jersey, Prentice-Hall.

Downs, R.T. (1992) Librational displacements of silicate tetrahedra in response to temperature and pressure. Ph. D. Dissertation. Virginia Polytechnic Institute and State University. Blacksburg, Virginia.

Downs, R.T.; Bartelmehs, K.L.; Gibbs, G.V. and Boisen, Jr., M.B. (1993) Interactive Software for Calculating and Displaying X-ray or Neutron Powder Diffractometer Patterns of Crystalline Materials. American Mineralogist, 78, ??.

Kihara, K. (1990) An X-ray study of the temperature dependence of the quartz structure. European Journal of Mineralogy, 2, 63-77.

Kirkpatrick, S.; Gelatt, Jr., C.D. and Vecchi, M.P. (1983) Optimization by Simulated Annealing. Science, 220(4598), 671-680.

Kramer, G.J.; van Beest, B.W.H. and van Santen, R.A. (1991) Relation between crystal symmetry and ionicity in silica polymorphs. Nature, 351, 636-638.

Lazarev, A.N. and Mirgorodsky, A.P. (1991) Molecular Force Constants in Dynamical Model of α–Quartz. Physics and Chemistry of Minerals, 18, 231-243.

Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N. and Teller, A.H. (1953) Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, 21(6), 1087-1092.

Purton, J.; Jones, R.; Catlow, C.R.A. and Leslie, M. (1993) Ab initio Potentials for the Calculation of the Dynamical and Elastic Properties of α–Quartz. Physics and Chemistry of Minerals, 19, 392-400.

Stixrude, L. and Bukowinski, M.S.T. (1988) Simple Covalent Potential Models of Tetrahedral SiO_2: Application to α–quartz and Coesite at Pressure. Physics and Chemistry of Minerals, 16, 199-206.

Tse, J.S.; Klug, D.D. and Le Page, Y. (1992) Novel High Pressure Phase of Silica. Physical Review Letters, 69, 3647-3649.

Vanderbuilt, D. and Louie, S.G. (1984) A Monte Carlo Simulated Annealing Approach to Optimization over Continuous Variables. Journal of Computational Physics, 56, 259-271.

CHAPTER 2

TO_4 Rigid Body Motion in Silicates

Introduction

Anisotropic Gaussian displacement parameters (ADPs), routinely used as regressor variables in a refinement model to describe the time-and-space-averaged vibrational motion of an atom, have provided important physical information about a crystal (Appendix 2A). Several studies have focused on the difference displacement parameter evaluated along the vector between two adjacent atoms, A and B, Δ_{AB}, where

$$\begin{aligned}\Delta_{AB} &= z^2_{BA} - z^2_{AB} \\ &= [\mathbf{v}]^t_D(GU_BG)[\mathbf{v}]_D - [\mathbf{v}]^t_D(GU_AG)[\mathbf{v}]_D \\ &= \frac{1}{2\pi^2}[\mathbf{v}]^t_D(G\beta_BG)[\mathbf{v}]_D - \frac{1}{2\pi^2}[\mathbf{v}]^t_D(G\beta_AG)[\mathbf{v}]_D\end{aligned}$$

where z^2_{BA} and z^2_{AB} are the respective mean-square displacement amplitudes (MSDAs) of B toward A and of A toward B, $[\mathbf{v}]_D$ is an unit vector parallel to the direction between AB defined in terms of the direct basis $D = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, β_A (or U_A) and β_B (or U_B) is the ADP matrix (or temperature factor matrix) for A and B, respectively; and G is the metrical matrix. Not only does Δ_{AB} provide a measure of relative internal motion between atoms A and B (Dunitz et al., 1988), but it also provides a measure of the spin state for transition metals (Chandrasekhar and Bürgi, 1984) and the Jahn–Teller deformation of Cu complexes (Bürgi, 1984). It has also been used to detect errors in structure refinement models and in reported results (Kunz and Armbruster, 1990; Dunitz et al., 1988; Trueblood, 1978; Hirshfeld, 1976). In framework silicate and aluminosilicate crystals with TO_4 tetrahedra (T=Si,Al), the average Δ_{TO}, $\langle\Delta_{TO}\rangle$, provides a measure of positional

and structural disorder (Kunz and Armbruster, 1988; Kunz and Armbruster, 1990, Downs et al., 1990). Large values of Δ_{Si-O} recorded for plazolite and katoite have been taken as evidence for positional disorder of the oxygen atom (Armbruster and Lager, 1989). Downs et al. (1990) also found that the average vibrational motion between the T and O atoms in frameworks is consistent with rigid TO bond vibration.

The terms "rather rigid" and "quasi–rigid" have been used to describe the vibrational motion of SiO_4 tetrahedra in quartz (Liebau and Böhm, 1982), while various models describing the $\alpha - -\beta$ phase transition have assumed rigid tetrahedra (Megaw, 1973; Ghose et al., 1986; Boysen et al., 1980; Liebau and Böhm, 1982). Grimm and Dorner (1975) have used a bonding model based on sp^3 σ–orbitals in concluding that the tetrahedra in quartz are rigid. Rigid tetrahedra are also assumed in lattice dynamical calculations (Rao et al., 1988) and bond length corrections (Ghose, 1986; Downs et al., 1992). Computer models for silica have been constructed assuming rigid tetrahedra (Stixrude and Bukowinski, 1988). On the other hand, Kihara (1990), in a high temperature study of quartz, asserts that the tetrahedra are not rigid because of the 'considerable bending of the OSiO angles' encountered at high temperatures.

In this paper, the ADPs obtained for Si and O atoms in refinements of non–framework structure types will be examined. As a test of the assertion that SiO_4 tetrahedra behave as rigid bodies, the ADP data will be examined for experimental evidence consistent with rigid body vibrational motion. Because rigid SiO bonds are a necessary but not a sufficient condition for rigid body motion (Ghose et al., 1986; Hummel et al., 1990), the tetrahedral groups will be examined for correlated motion represented in the sizes, shapes, and orientations exhibited by the ADPs of the atoms in the group. An analysis of the motion will be completed by examining

the applicability of the rigid body hypotheses to the ADP data and the constraints placed on the data by rigid body motion.

Evidence of Rigid TO Bonds in Non–framework Silicates

In an examination of the relationship between z^2_{OT} and z^2_{TO} in framework silicate crystals, Downs et al. (1990) found that the ADPs determined for crystals free of static disorder are consistent with rigid TO bond vibrational motion. If the forces that govern the geometry of SiO_4 tetrahedra are short ranged (Gibbs, 1982), then the Si and O atoms of tetrahedra in all silicate structures should exhibit similar vibrational motion. Therefore, the ADPs determined for these atoms in non–framework silicates should represent vibrational motion consistent with that found within ordered framework silicates. For such structures, Downs et al. (1990) proposed a criteria to identify rigid TO bond behavior based on the distributional properties of $\langle\langle\Delta_{TO}\rangle\rangle$, the average of all the $\langle\Delta_{TO}\rangle$–values in a structure. They concluded that any framework structure that possesses rigid TO bonds satisfies the criteria that (1) $-\ 0.00125\text{Å}^2 \leq \langle\langle\Delta_{TO}\rangle\rangle \leq 0.002\text{Å}^2$ and (2) the esd of $\langle\langle\Delta_{TO}\rangle\rangle \leq 0.00125\text{Å}^2$. On the basis of this criteria, they concluded that about a third of the structures examined possess rigid TO bonds. As about 130 of the 670 individual Δ_{TO}–values exceed the maximum deviation (0.0015Å^2) from zero observed for quartz, cristobalite and coesite, a new criteria is proposed based on the $\langle\Delta_{TO}\rangle$–values for individual tetrahedra rather than for the ensemble of tetrahedra in a structure. The criteria used in this study is (3) $-\ 0.00125\text{Å}^2 \leq \langle\Delta_{TO}\rangle \leq 0.002\text{Å}^2$ and (4) the esd of $\langle\Delta_{TO}\rangle \leq 0.00125\text{Å}^2$. Non–framework tetrahedra that satisfy this criteria would then indicate rigid TO bond vibrational motion similar to that found in framework silicates. These criteria also provide a measure of both the relative perfection of the crystal and the physical acceptability of the results provided by the refinement (Downs et al., 1990; Kunz and Armbruster, 1990;

Hirshfeld, 1976; Dunitz et al., 1988).

In an examination of criteria 3 and 4, an extensive data set of non–framework silicates was obtained from the literature. The data were provided by refinements completed on data recorded at room pressure and at or below room temperature with refined ADPs being reported for all atoms (except hydrogen). A structure was accepted for study when reported bond lengths and angles, isotropic equivalent temperature factors could be reproduced and when the ADP matrix for each of its atoms was found to be positive definite, using the software METRIC. Of the 248 non–framework structures, 231 were accepted including 94 orthosilicates, 33 sorosilicates, 51 chain silicates, 29 ring silicates, and 41 sheet silicates. This resulted in 357 individual non–framework TO_4 with 99 occurring in orthosilicates, 62 in sorosilicates, 85 in chain silicates, 33 in ring silicates, and 78 occurring in sheet silicates.

The criteria was first applied to 469 silicate tetrahedra obtained for framework structures determined at standard conditions (Downs et al., 1990). The analysis indicates that 35% of the 469 tetrahedra satisfy both criteria and qualify as possessing rigid TO bonds. Of the 357 tetrahedra in non–framework structures, 50% satisfy both criteria, a significantly larger percentage than obtained for the frameworks. Thus, rigid TO bonds appear to be as common if not more so in non–framework than in the framework silicates. Of the 826 TO_4 groups examined, 352 satisfy (3) and (4). These include 166 from frameworks, 67 from orthosilicates, 42 from sorosilicates, 51 from chain silicates, 17 from ring silicates, and 9 from sheet silicates. The observation that nearly half (43%) of the data fail both criteria agrees with similar failure rates observed for a number of molecular compounds (Trueblood, 1978). The relatively low percentages of framework and sheet silicate tetrahedra (35% and 12%, respectively) that satisfy (3) and (4) suggests the pres-

ence of substitutional and structural disorder (Leibau, 1985). The ADPs for the 352 tetrahedra that satisfy both criteria appear to represent groups with little or no internal motion, positional disorder, and twinning, and thus provide a data set for the study of the vibrational motion of TO_4 tetrahedra (Hirshfeld, 1976).

Rigid Body Analysis

In our examination of the average vibrational motion of TO_4 tetrahedra, the observed ADPs for each T and O atom was compared with those calculated with the rigid body TLS model (Schomaker and Trueblood, 1968). On the one hand, it has been asserted by Hummel et al. (1990) and Chandrasekhar and Bürgi (1984), for example, that coordinated polyhedra behave as rigid bodies if a model based on rigid motion reproduces the experimental ADPs observed for the polyhedra to within experimental error. In fact, several methods based on differences between observed and calculated ADPs have been proposed to measure how well such calculations reproduce a set of observed ADPs (Hummel et al., 1990; Trueblood, 1978; Burns et al., 1967; Destro et al., 1977). On the other hand, as the TLS parameters obtained in a least squares refinement can incorporate the effects of internal motions of the individual atoms, Dunitz et al. (1988) has argued that such agreement does not necessarily mean that a coordinated polyhedron or a molecule behaves as a rigid body.

The agreement between observed and calculated ADPs was evaluated in this study using a strategy based on one devised by Burns et al., (1967). In their strategy, three ellipsoid agreement parameters were defined that relate to the differences in relative sizes, shapes, and orientations of a set of observed and calculated thermal ellipsoids for an atom (Appendix 2A). The shape and orientation ellipsoid agreement parameters, here denoted EAP2 and EAP3, respectively, are adopted as defined by Burns et al. (1967), but the size ellipsoid agreement pa-

rameter, here denoted EAP1, differs from that defined by them (Appendix 2B). They used a criterion based on the difference between the observed and calculated isotropic equivalent displacement parameters, $(U_{(obs)}) - (U_{(calc)})$. The sizes of the two ellipsoids were considered to be different when $(U_{(obs)}) - (U_{(calc)})$ exceeded $3\hat{\sigma}(U_{(obs)})$. However, reliable estimates of $\hat{\sigma}(U_{(obs)})$ would be difficult to obtain for almost all published data sets because the variance–covariance matrices are usually not included (Trueblood, 1978; Dunitz et al., 1988). Because of the large number of data that will be considered in this study and because such matrices are usually unavailable, we redefined the size parameter to be

$$\text{EAP1} = \frac{|U_{(obs)} - U_{(calc)}|}{U_{(obs)}}.$$

For calculated ADPs to be in satisfactory agreement with those observed, the three EAPs were required to simultaneously satisfy the criteria (A) EAP1 ≤ 0.1, (B) EAP2 ≤ 150, and (C) EAP3 $\leq 20°$. A cut–off of ≤ 0.1 was chosen for criterion (A) because it conforms with the results obtained by Burns et al. (1967) analysis of size. If the EAPs for all five atoms of the TO_4 tetrahedron satisfy all three criteria, then the thermal motion represented by the observed ADPs of the TO_4 tetrahedron is considered to be consistent with that predicted by rigid body motion.

Such motion implies that the vibrational behavior of all atoms of the group are highly correlated (Johnson, 1970; Ghose et al., 1986). Any correlated vibrational motion among the atoms comprising a rigid body will manifest itself in the relative sizes, shapes and orientations of the thermal ellipsoids. However, the assumption made in most crystal structure refinement models is that atomic coordinates of the nonequivalent atoms and their ADPs are independent (the IAM model, Appendix 2A). On the other hand, the ADPs calculated from the TLS rigid body model

give the simplest picture of correlated motion among the motions of the atoms comprising a TO_4 group (Johnson, 1970). By applying the criteria discussed above, the observed ADPs are examined for correlated motion by comparing their physical aspects (size, shape and orientation) with the calculated ADPs. The EAP method chosen here are believed to be a better measure of 'fit' than one involving the absolute values of observed ADPs which can absorb systematic errors in the diffraction data (Dunitz et al., 1988; Chandrasekhar and Bürgi, 1984; Bürgi, 1984; Armbruster et al., 1990) Furthermore, this method does not rely on statistical information frequently not or improperly provided with published refinements (Trueblood, 1978; Dunitz et al., 1988). The EAP method is similar to the visual methods used by Hummel et al. (1991) to study discrepancies between observed ADPs and ones predicted by various models.

The observed ADPs for 352 tetrahedra were each regressed against the 20 parameters required to define a general rigid body TLS model (Schomaker and Trueblood, 1968), using a FORTRAN77 program called TLS (Appendix 2C) and assuming C_1 point symmetry. All calculations were performed assuming a Cartesian basis centered at the central T atom (Boisen and Gibbs, 1985). Final parameter estimates were computed relative to the center of reaction (Johnson, 1970). Non–positive definite **T** or **L** matrices were calculated for 39 of the 352 tetrahedra (10 framework, 12 orthosilicate, 3 sorosilicate, 11 chain silicate, 2 ring silicate, and 1 sheet silicate). As these represent physically unrealistic results, only the remaining 313 tetrahedra were used for the rigid body analysis. Figure 2–1 shows a histogram of the coefficient of determination (R^2) values obtained from the TLS linear regression results of the tetrahedra (Appendix 2C). This narrow range of R^2 values between 0.960 to 1.000 suggests that the traditional R^2 statistic alone is inadequate to access the applicability of the TLS model.

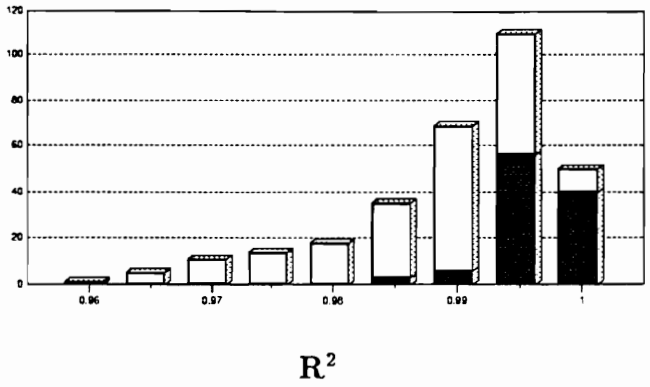


Figure 2–1. Histogram of R^2 values for the 313 tetrahedra used in the rigid body analysis. The black bars represent data that passes the EAP criteria whereas the white bars represent data that fail the criteria.

To appreciate the agreement between the thermal ellipsoids as they relate to the EAP criteria, drawings were made for the observed and calculated thermal ellipsoids for several TO_4 groups. For example, Figure 2–2 compares a set thermal ellipsoids for two tetrahedra taken from a refinement of triclinic bikitaite (Bissert and Liebau, 1986). The agreement between the size, shape and orientation of the observed and calculated thermal ellipsoids for the T4 tetrahedron (Figure 2–2a), which passes the criteria, is strikingly similar. On the other hand, the agreement for the Al5 tetrahedron, which fails the criteria, shows distinct differences between the observed and calculated thermal ellipsoids for some of the atoms (Figure 2–2b). As the tetrahedra in Figure 2–2a pass our EAP criteria, the observed thermal ellipsoids are concluded to represent correlated motion that is consistent with that predicted assuming rigid body motion. The EAP criteria calculated for the thermal ellipsoids in Figure 2–2b indicates unsatisfactory agreement between shape and orientation for three atoms of the Al5 tetrahedron. Here, we conclude that the thermal motion represented by the ADPs of the tetrahedron does not completely represent correlated motion and therefore is not consistent with rigid

body motion.

An application of the EAP criteria to the 313 tetrahedra indicates that the ADPs of 105 are consistent with rigid body motion of a TO_4 group. Of these, 61 occur in silicate and aluminosilicate frameworks, 20 occur in orthosilicates, 9 occur in sorosilicates, 8 occur in chain silicates, 5 occur in ring silicates and 2 occur in sheet silicates. The R^2 values for the tetrahedra that pass the EAP criteria are displayed in Figure 2–1 as solid bars whereas those that fail are displayed as open bars. Those that pass have R^2 values that range between 0.985 and 1.0 whereas those that fail have a larger range of values between 0.96 and $\sim$1.0. As observed above, these results indicate that the R^2 value for a TO_4 group is an unsatisfactory criterion for establishing whether a group is rigid. For example, more than half of the data that fail the EAP criteria have R^2 values of 0.99 or larger. There are several reasons why two thirds of the tetrahedra fail the EAP criteria: (1) Strict application of the criteria. In other words, rejection of a tetrahedron that has as few as one atom whose ADPs fail at least one of the criteria; (2) typos in the reported data (3) O atom and/or T atom positional disorder (4) problems in the refinement of the ADPs such as systematic errors or (5) nonrigid tetrahedra.

The 105 tetrahedra that pass the EAP criteria are listed in Table 2–1. Note that the table includes the TO_4 data of estatite used in the rigid body analysis by Ghose et al., (1986), the only silicate for which such an analysis has been completed. The table also includes the estimated libration angle (Θ) determined from the eigenvalues (or trace) of the **L** matrix determined for each tetrahedron (Appendix 2D). The range of Θs for all but one of these tetrahedra (2–7°) is in good agreement with the 6-9° upper limit for the applicability of the TLS model suggested by Trueblood (1978). For the Si2 tetrahedron of melanophogite (Gies, 1983), the Θ value is 16° . Because this structure is believed to exhibit structural

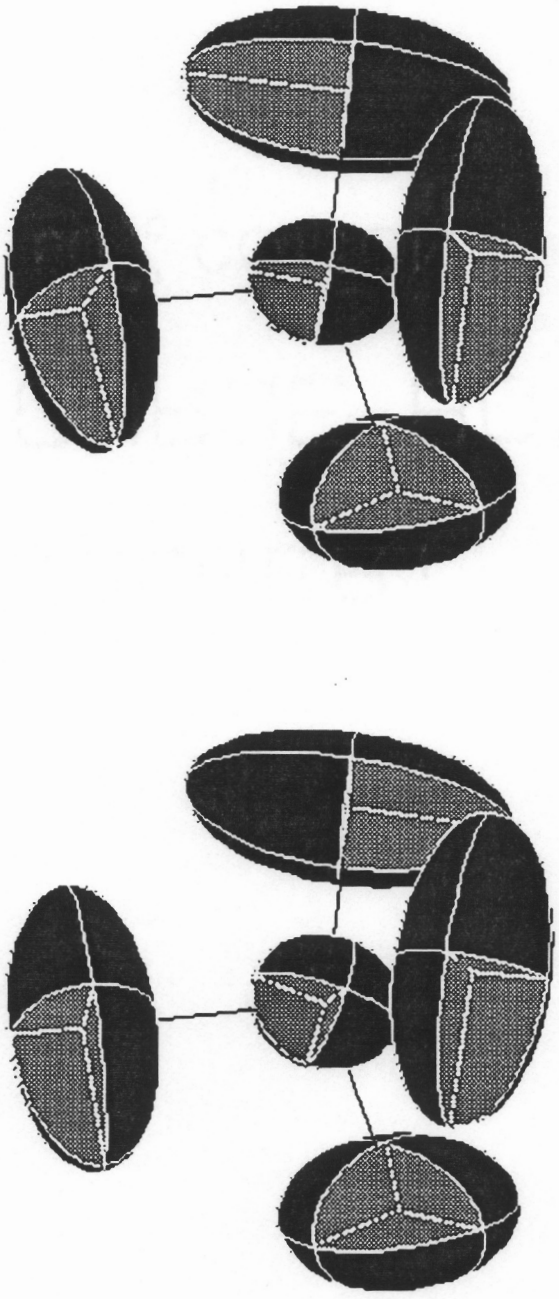


Figure 2–2a. A comparison of the observed thermal ellipsoids of the T4 tetrahedron taken from a refinement of triclinic bikitaite (top) (Bissert and Liebau, 1986) versus those calculated using the TLS rigid body model (bottom) (Schomaker and Trueblood, 1968). The EAP criteria indicate agreement between all five pairs of thermal ellipsoids.

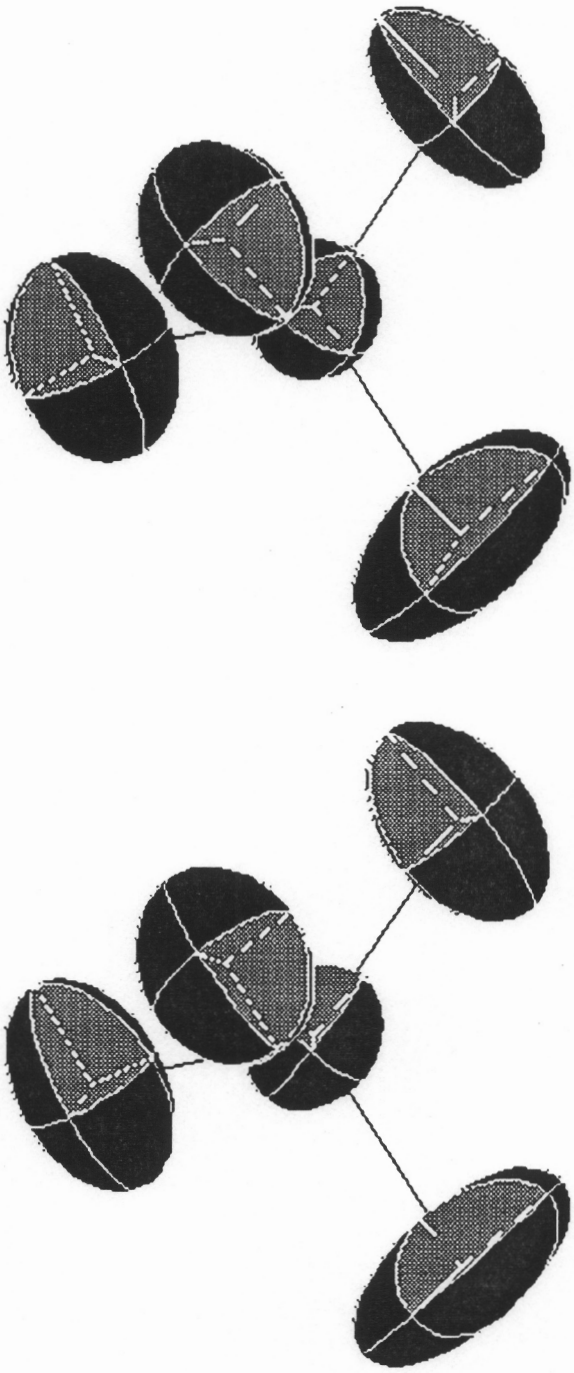


Figure 2–2b. A comparison of the observed thermal ellipsoids of the Al5 tetrahedron taken from a refinement of triclinic bikitaite (top) (Bissert and Liebau, 1986) versus those calculated using the TLS rigid body model (bottom) (Schomaker and Trueblood, 1968). The EAP criteria indicate disagreement in shape and orientation between three pairs of thermal ellipsoids.

disorder (Downs et al., 1990) and because its libration angle exceeds the limit of the applicability of the TLS model, this data set was excluded in discussing the systematics of rigid body tetrahedra. An additional reason for excluding this data set will be presented later.

Figure 2–3 shows the relationship between the libration angle and temperature at which the data set was recorded for the 104 tetrahedra that are consistent with TLS rigid body motion. The mean libration angle, $\langle\Theta\rangle$, of the seven lowest temperature framework data ($\langle\Theta\rangle = 3.0°$) is significantly smaller than the mean libration angle of the 94 room temperature data ($\langle\Theta\rangle = 4.4°$), suggesting that the librational motion of a rigid TO_4 tetrahedron tends to increase, as expected, with temperature. For the thirteen olivine structures (shown as open circles in Figure 2–3) $\langle\Theta\rangle = 3.1°$, which is similar to the mean low temperature libration angle found in framework structures. Restricted librational motion of a rigid TO_4 tetrahedron in olivine may be ascribed to the approximate *hcp* of the O atoms in this structure. It is noteworthy that the tetrahedra in the quartz exhibit larger $\langle\Theta\rangle$ values on average (5.6°), than those exhibited by coesite (4.9°). This is consistent with the more open, less dense framework structure of quartz which would allow a greater librational motion of a TO_4 tetrahedron.

It was recently suggested (Downs et al., 1992) for a tetrahedron assumed to be rigid that the translational motion of the group is contained in the ADPs of the central T atom. Using the EAP method presented here, a direct comparison can be made between the ADPs of the central T atom and the translational motion represented by the **T** matrix (representing translational motion) determined in the rigid body analysis. Figures 2–4a — 2–4c show histograms of EAPs used to make these comparisons for all 313 tetrahedra used in the rigid body analysis. For the 104 tetrahedra (consistent with TLS rigid body motion and represented

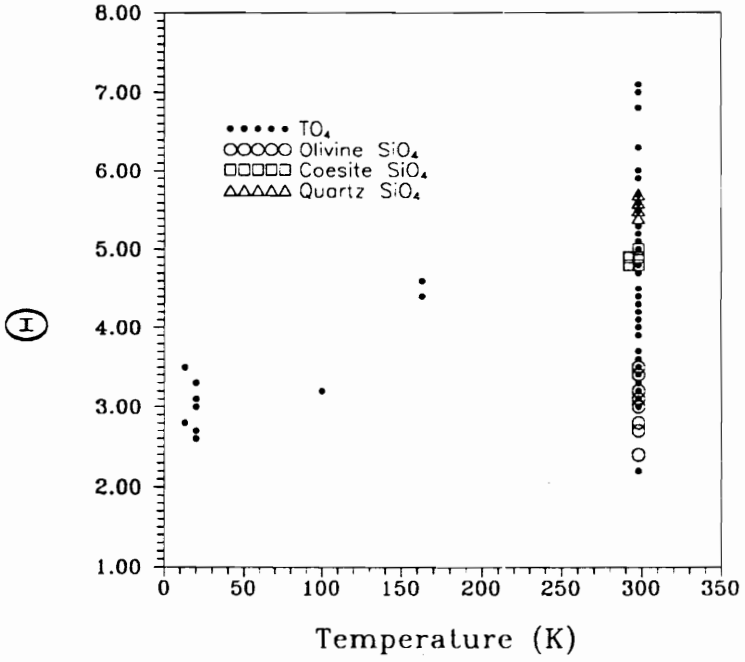
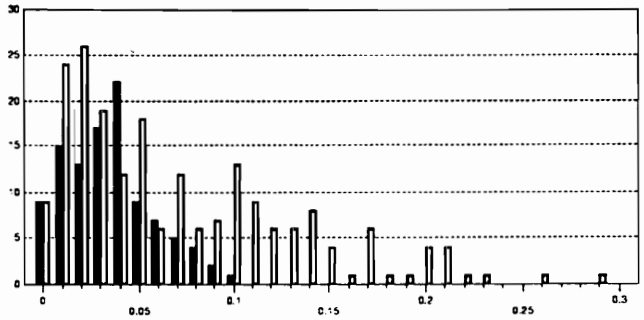


Figure 2–3. Libations angle, Θ, versus temperature at which the intensity data was collected for the 104 tetrahedra consistent with TLS rigid body motion.

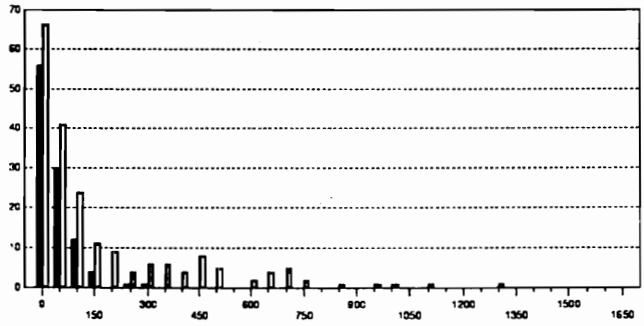
by black bars), a narrow ranges of the EAP values is recorded in the figures, indicating a general agreement between the estimated **T** matrix and ADP matrix for the T atom. On the other hand, the remaining 209 tetrahedra (represented by open bars) show a much wider range of EAP values. Consequently, for a TO_4 tetrahedron consistent with TLS rigid body motion, the ADPs of the central T atom represent, in large part, the translational motion of the tetrahedron with little or no librational motion.

If the ADPs of a central T atom embody mostly translational motion of the rigid TO_4 tetrahedron, then those of its coordinating oxygen atoms must contain



EAP1

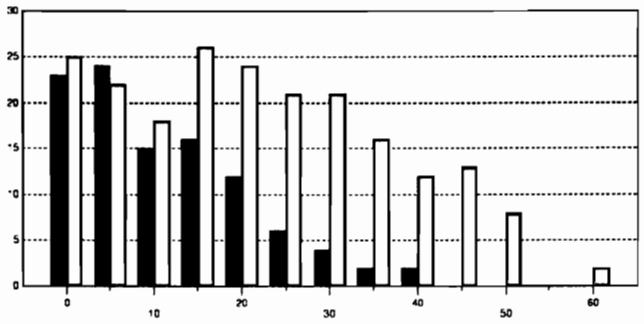
Figure 2–4a. Histogram of EAP1 (size) parameter values calculated for the ADPs of the central T atom and the **T** matrix determined in the TLS analysis. The black bars represent the data belonging to TO_4 groups that are consistent with TLS rigid body motion whereas the white bars represent data considered to be inconsistent.



EAP2

Figure 2–4b. Histogram of EAP2 (shape) parameter values calculated for the ADPs of the central T atom and the **T** matrix determined in the TLS analysis. The black bars represent the data belonging to TO_4 groups that are consistent with TLS rigid body motion whereas the white bars represent data considered to be inconsistent.

translational plus librational motion. This would suggest that the ADP ellipsoids for oxygen atoms of a TO_4 tetrahedron will be larger than that of the central T atom (Downs et al., 1992). Thus, by subtracting the isotropic equivalent displacement parameter of the central T atom, $B(\mathrm{T})$, from the average of the four oxygen isotropic equivalent displacement parameters, $B(O)$, then the difference, δ_B, should be positively correlated with the librational motion of the tetrahedron



EAP3

Figure 2–4c. Histogram of EAP3 (orientation) parameter values calculated for the ADPs of the central T atom and the **T** matrix determined in the TLS analysis. The black bars represent the data belonging to TO_4 groups that are consistent with TLS rigid body motion whereas the white bars represent data considered to be inconsistent.

as displayed in Figure 2–5. A linear regression analysis indicates that 94% of the variation of δ_B can be explained by a linear dependence on Θ. This result supports the observation that the refined ADP ellipsoids of the oxygen atoms are larger than those of the central T atom in a rigid TO_4 tetrahedron, not just because the O atoms are lighter than the T atom, but because they contain translational and librational motion of the tetrahedron (Downs et al., 1992).

Figures 2–6a and 2–7a show the distributions of Δ_{TO} and Δ_{OO} values denoted, Δ^r_{TO} and Δ^r_{OO}, respectively, observed for those tetrahedra that are consistent with TLS rigid body motion. Figures 2–6b and 2–7b show the corresponding distributions, Δ^f_{TO} and Δ^f_{OO}, respectively, for the nonrigid tetrahedra. Because there is no reference atom for a given pair of O atoms, O_1 and O_2, Δ_{OO} is computed as $\Delta_{OO} = |z^2_{O_1O_2} - z^2_{O_2O_1}|$. Note that all values equal to zero by symmetry have been removed from the data sets. The Δ^r_{TO} and Δ^r_{OO} distributions are both found to be statistically different from their corresponding Δ^f_{TO} and Δ^f_{OO} distributions, respectively. The separation of Δ_{TO} and Δ_{OO} into two sets of statistically different distributions by using the EAP criteria provides additional support for the

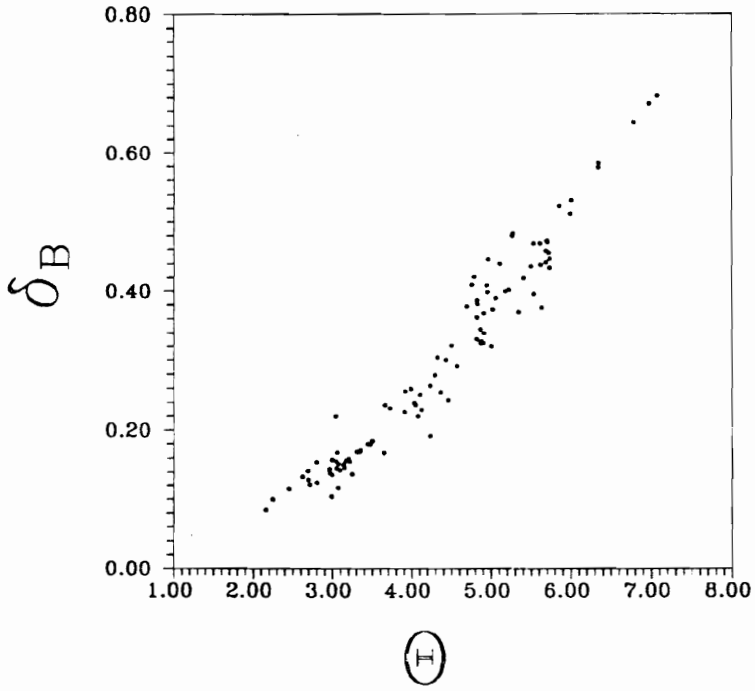
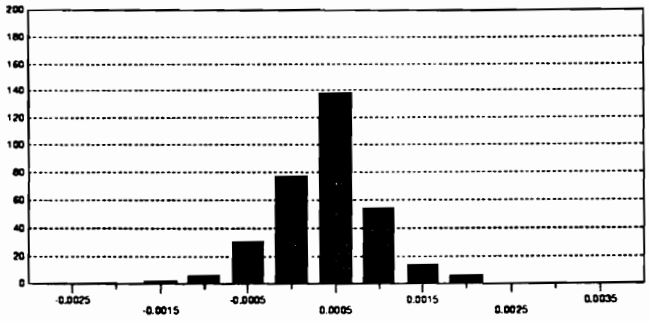


Figure 2–5. δ_B (= B(O)-B(T)) (Å^2) versus libration angle, Θ, for the 104 tetrahedra with vibrational motion consistent with TLS rigid body motion.

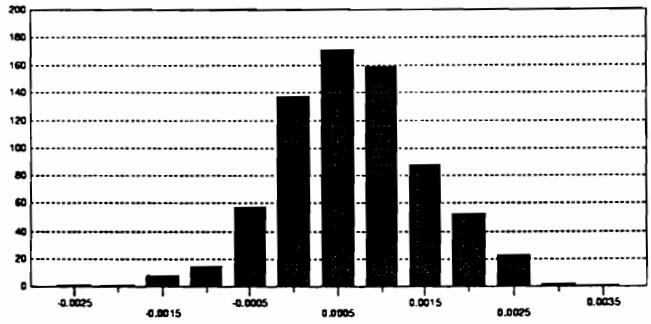
strategy used to define rigid tetrahedra.

The mean of Δ^r_{TO} (=0.00040Å^2) is smaller than the mean of Δ^f_{TO} (=0.00068Å^2), as expected for a distribution of experimental Δ_{TO} values consistent with rigid body motion. A positive mean for both of these distributions is also consistent with the notion that B(O) is larger that B(T). The mean of Δ^r_{TO} is very similar to the values, $\langle\Delta_{SiO}\rangle$ = 0.0004Å^2 and $\langle\Delta_{AlO}\rangle$ = 0.0005Å^2 reported by Kunz and Armbruster (1990) for completely ordered SiO_4 and AlO_4 tetrahedra in low albite. In addition, 97% of the Δ^r_{TO} data fall within ±0.0015Å^2, the range suggested by Downs et al. (1990) for rigid SiO bonds in quartz, cristobalite, and



Δ^r_{TO}

Figure 2–6a. Histogram of Δ^r_{TO} values (Å^2) taken from tetrahedra that pass the EAP criteria.

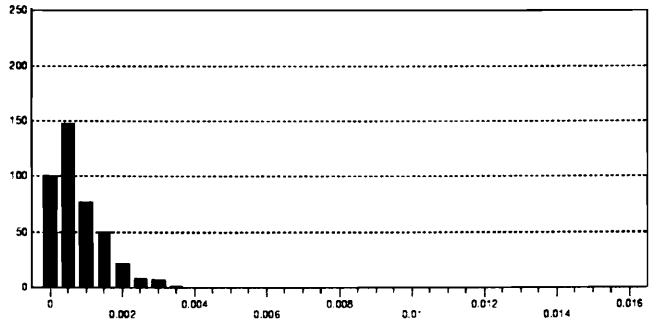


Δ^f_{TO}

Figure 2–6b. Histogram of Δ^f_{TO} values (Å^2) taken from tetrahedra that fail the EAP criteria.

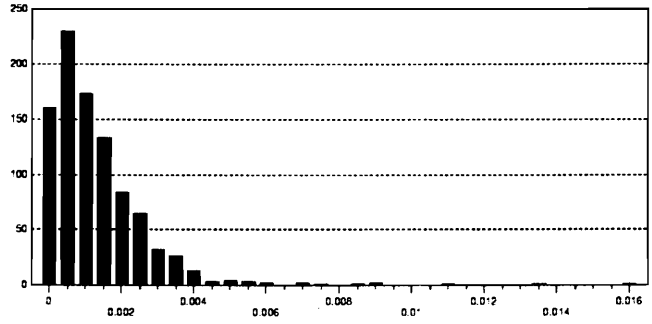
coesite. Note that 99% of the Δ^r_{OO} values fall within the 0.003Å^2 limit suggested by Bürgi (1984) for a rigid bond in transition–metal complexes.

Figure 2–8a shows considerable scatter in B(T) versus B(O) values for the 722 tetrahedra that fail criteria 1–4 and/or A–C. The wide scatter of B(O) values is consistent with that observed by Boisen et al. (1990). For the 104 tetrahedra considered to be consistent with TLS rigid body motion (represented by dots in Figure 2–8b), there is a more limited range of B(T) and B(O) values. The maximum value of B(O) is $\sim$ 2.0 Å^2 and is within the range suggested by Boisen et al. (1991) for B(O) values free of static disorder (B(O) $\leq$ 3.0Å^2). The maximum value of B(T) consistent with rigid TO bonds is about 1.0Å^2 which may be



Δ^r_{OO}

Figure 2–7a. Histogram of Δ^r_{OO} values (Å^2) taken from tetrahedra that pass the EAP criteria.



Δ^f_{OO}

Figure 2–7b. Histogram of Δ^f_{OO} values (Å^2) taken from tetrahedra that fail the EAP criteria.

taken as a limiting value for B(T) for structures free of static disorder. A linear regression indicates that B(T) is on average one–half of B(O), with 74 percent of the variation in B(T) explained in terms of a linear model containing B(O). The outlier point (open square) is for the Si2 tetrahedron of melanophogite which exhibits a 16° libration angle and which was excluded earlier from analysis.

The coesite data displayed in Table 2–1 constitute most of the room pressure data set used by Boisen et al. (1990) in the their study of SiO bond length (R(SiO)) variations in coesite. They found that 84% of the variation in R(SiO) can be explained by a linear regression model that included the parameters, f_s(O) = $1/(1-\sec\Phi)$ where $\Phi = \angle$ SiOSi, P, f_s(Si), B(O) and B(Si). Using only the coesite

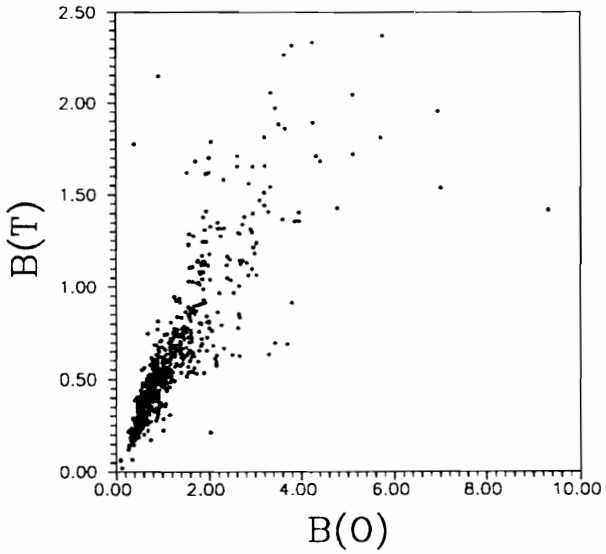


Figure 2–8a. B(T) versus B(O) for tetrahedra that fail rigid TO bond criteria (1–4) and/or EAP criteria (A–C).

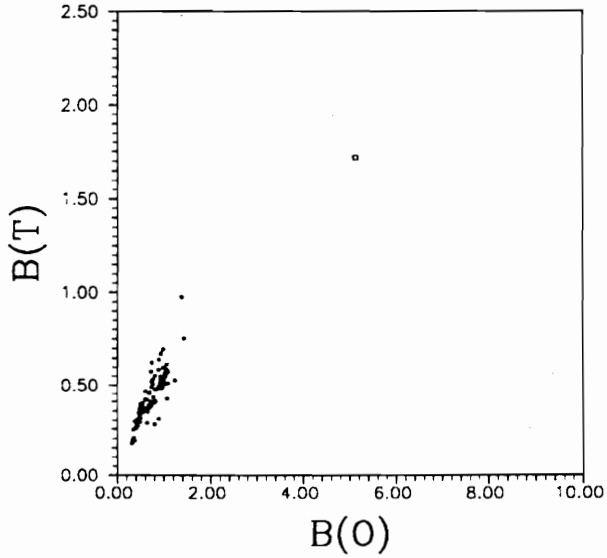


Figure 2–8b. B(T) versus B(O) for tetrahedra that pass rigid TO bond criteria (1–4) and EAP criteria (A–C).

refinement data where both tetrahedra satisfy the EAP criteria and excluding P as a regressor variable, a stepwise regression involving the remaining four parameters indicates that the "best" regression equation only involves f_s(O) (R(SiO)=1.75 –

$0.313f_s(O)$). In fact, 90% of the variation in the R(SiO) for the satisfactory coesite data is explained by the parameter $f_s(O)$ alone. Similarly, the "best" regression for all of the silica polymorph data listed in Table 2-1 in which all tetrahedra satisfy the EAP criteria also involves only $f_s(O)$ (R(SiO)=1.75 − $0.304f_s(O)$). Leibau (1985) asserts that a linear relationship between R(SiO) and $B(O)$ for the silica polymorphs is obtained when the $B(O)$ data contains static disorder. A linear regression indicates there is no linear relationship between R(SiO) and $B(O)$ for the silica polymorphs. This supports the assertion that tetrahedra consistent with rigid body motion are free of the effects of static disorder.

As mentioned above, rigid TO bonds are a necessary but not a sufficient condition for a TO_4 tetrahedron to qualify as a rigid body (Ghose et al., 1986; Hummel et al., 1990). This agrees with the fact that only 104 tetrahedra from the 313 tetrahedra that passed criteria (3) and (4) qualify as rigid body bodies. If a tetrahedron vibrates as a rigid body, then the distances between T and O atoms and those among the oxide ions are not expected to change during vibration (Dunitz et al., 1988). As indicated in figure 2-7b, even rigid O-O separations are not a sufficient condition for a TO_4 tetrahedron to qualify as a rigid body. Because the distributions of Δ^r_{TO} and $t(\Delta^r_{OO})$, where $t(\Delta^r_{OO}) = ln(\Delta^r_{OO} + 0.00015\text{Å}^2)$, represent standard-normal distributions along with large sample sizes, statistical $Z - -$ and $\chi^2 - -$ tests were formulated for identifying whether an observed sample of Δ_{TO}s and Δ_{OO}s are consistent with TLS rigid body motion. However, as this type of test is only a necessary condition for rigid body motion, sufficient evidence of rigid body motion may only be found through application of a rigid body model in conjunction with a fitting criteria like the EAP method used in this study.

Table 2–1. Tetrahedra that pass the EAP criteria

Mineral	Reference	Tetrahedron	Libration angle
microcline	Phillips, (1990), at 163K, $C\bar{1}$	$Si2_o$	4.4
		$Si2_m$	4.6
microcline	Phillips, (1990), $C\bar{1}$	Al1(0)	5.3
		Si1m	5.9
		Si2o	5.7
microcline	Blasi et al., (1984), sample 7813B, $C\bar{1}$	Si_1M	5.7
		Si_2O	5.7
microcline	Blasi et al., (1987), $C\bar{1}$	Si_2O	5.5
low albite	Smith et al., (1986), at 13K, $C\bar{1}$	Al_1O	2.8
		Si_1M	3.5
low albite	Harlow and Brown, (1980), neutron, $C\bar{1}$	Al_1O	5.0
		Si_1M	6.3
		Si_2M	5.7
low albite	Harlow and Brown, (1980), X-ray, $C\bar{1}$	Al_1O	5.1
low albite	Armbruster et al., (1990), $C\bar{1}$	Al_1O	5.0
		Si_2O	5.3
low cordierite	Armbruster, (1986), from Haddam, 100K,Cccm	T_16	3.2
low cordierite	Armbruster, (1986), from Haddam, Cccm	Al11	3.0
		T_16	4.0
low cordierite	Armbruster, (1986), from Kemiö, Cccm	T_16	3.9
		T_23	5.7
low cordierite	Armbruster, (1986), from Ferry, Cccm	Al11	3.1
		T16	4.5
low cordierite	Armbruster, (1986), from Sponda, Cccm	T_16	4.1
low cordierite	Cohen et al., (1977), neutron, Cccm	Si_16	3.1
low cordierite	Cohen et al., (1977), X-ray, Cccm	Si_16	4.0
cristobalite	Peacor, (1973), at 28°C, $P4_12_12$	Si	7.0
coesite	Levien and Prewitt, (1981), C2/c	Si1	4.9
		Si2	4.8
coesite	Kirfel and Will, (1984), C2/c	Si2	4.9
coesite	Geisinger et al., (1987), IAM refinement, C2/c	Si1	4.9
		Si2	4.8
coesite	Geisinger et al., (1987), IAM+ refinement, C2/c	Si1	5.0
		Si2	4.9
coesite	Smyth et al., (1987), at 292K, C2/c	Si1	4.9
		Si2	4.8
coesite	Gibbs et al., (1977), C2/c	Si1	4.9
quartz	Young and Post, (1962), $P3_212$	Si	5.6
quartz	Le Page and Donnay, (1976), $P3_212$	Si	5.6
quartz	Levien et al., (1980), $P3_212$	Si	5.4
quartz	Wright and Lehmann, (1981), at 25°C, $P3_212$	Si	5.5
quartz	Kihara, (1990), at 298K, $P3_212$	Si	5.7

Table 2–1 (continued). Tetrahedra that pass the EAP criteria

Mineral	Reference	Tetrahedron	Libration angle
natrolite	Artioli et al., (1984), at 20K, Fdd2	Al	2.7
		Si1	3.3
		Si2	3.1
scolecite	Kvick et al., (1985), at 20K, Cc	Al2	2.6
		Si3	3.0
scolecite	Joswig et al., (1984), Fd	Si_2	5.1
		Al10	4.8
edingtonite	Kvick and Smith, (1983), $C2_12_12$	Al	5.0
		Si1	5.6
thomsonite	Pluth et al., (1985b), Pncn	Al1	5.3
		Si1	5.7
		Si2	5.2
		Si3	5.5
mesolite	Artioli et al., (1986), Fdd2	Al1	4.8
		Al2	4.7
		Si3	5.7
bikitaite	Bissert and Liebau, (1986), P1	T4	7.1
anorthite	Kalus, (1978), $P\bar{1}$	T2ozi	4.4
melanophlogite	Gies, (1983), Pm3n	Si2	16.2
olivine	Miyaké et al., (1987), Pbnm, Co(03)	Si	3.4
olivine	Miyaké et al., (1987), Pbnm, Co(05)	Si	3.1
olivine	Miyaké et al., (1987), Pbnm, Co(18)	Si	3.5
olivine	Miyaké et al., (1987), Pbnm, Co(20)	Si	3.2
olivine	Nover & Will, (1981), Pmcn, Fe(10) P1	Si	3.4
olivine	Nover & Will, (1981), Pmcn, Fe(12) P2	Si	3.0
olivine	Nover & Will, (1981), Pmcn, Fe(12) P3	Si	3.0
olivine	Nover & Will, (1981), Pmcn, Fe(12) P4	Si	3.1
olivine	Boström, (1987), Pbnm, Ni=0.0	Si	3.0
olivine	Boström, (1987), Pbnm, Ni=0.51	Si	2.8
olivine	Boström, (1987), Pbnm, Ni=0.69	Si	2.4
olivine	Boström, (1987), Pbnm, Ni=1.00	Si	3.0
Co-garnet	Ohashi et al., (1981), Ia3d	Si	3.2
zircon	Hazen & Finger, (1979), $I4_1/amd$, P=1atm.	Si	4.2
braunite	Moore & Araki, (1976), $I4_1/acd$	Si	3.7
chondrodite	Fujino & Takéuchi, (1978), $P2_1/b$	Si	3.0
andalusite	Winter & Ghose, (1979), Pnnm, T=25° C	Si	3.5
sillimanite	Winter & Ghose, (1979), Pbmn, T=25° C	Si	3.9
forsterite	Francis & Ribbe, (1980), Pbnm, Fo(51)	Si	2.7
zunyite	Baur & Ohta, (1982), $F\bar{4}3m$, Arizona	Si1	6.3
		Al1	2.2
zunyite	Baur & Ohta, (1982), $F\bar{4}3m$, Colorado	Al1	2.2
helvite	Hassan and Grundy, (1985), #2, $P\bar{4}3n$	Si	4.2

Table 2–1 (continued). Tetrahedra that pass the EAP criteria

Mineral	Reference	Tetrahedron	Libration angle
rosenhanite	Wan et al., (1977), $P\bar{1}$	Si3	4.1
ilvaite	Takéuchi et al., (1983), $P2_1/a$, Tsumo	Si2	3.2
ilvaite	Finger & Hazen, (1987), $P2_1/a$, Seriphos	Si2	3.2
zoisite	Smith et al., (1987), Pnma, 298K X-ray	Si3	3.0
kilchoanite	Kimata, (1989), I2cm	Si1	4.1
epidote	Gabé et al., (1973), $P2_1/m$, HEP	Si3	3.7
F-richterite	Cameron et al., (1983), Na, I2/m, T=24° C	T1	4.5
$NaMnSi_2O_6$	Basso et al., (1989), C2/c	Si	3.4
enstatite	Ghose et al., (1986), Pbca	Si1	3.3
		Si2	3.0
jadeite	Cameron et al., (1973), C2/c, T=24° C	Si	3.5
acmite	Clark et al., (1969), C2/c	Si	3.6
$LiFeSi_2O_6$	Clark et al., (1969), C2/c	Si	5.0
bavenite	Cannillo & Coda, (1966), Cmcm	Al4	5.2
feruvite	Grice & Robinson, (1989), R3m	Si	4.8
milarite	Sandomirski et al., (1977), P6/mcc	Si	6.8
searlesite	Ghose & Wan, (1976), $P2_1$	Si1	6.0
$K_2Si^{VI}Si_3^{IV}O_9$	Swanson, (1983), T=25° C,$P6_3/m$	Si4	4.3
brannockite	Armbruster & Oberhänsli, (1988), P6/mcc,#1	Si	6.0
talc	Perdikatsis & Burzlaff, (1981), $C\bar{1}$	Si2	4.0
pyrophyllite	Lee & Guggenheim, (1981), $C\bar{1}$	Si1	4.3

APPENDIX 2A

Statistical Foundation of the Structure Factor Equation and the Interpretation of Temperature Factors

Introduction

There are two basic approaches to the analysis of bragg diffraction data for the effects of thermal motion: the statistical approach and the mechanistic approach. The major difference between the two approaches is that the statistical approach, the most commonly used method, assumes that all atomic motions are independent of one another (uncorrelated). On the other hand, the mechanical approach takes into account the correlated motion between neighboring atoms. An example of a mechanical approach would be the application of the TLS rigid body model used in Chapter 2 and described in Appendix 2C. Here the most fundamental aspects of the statistical approach are developed. The goal of this approach is to describe the time-averaged vibrational displacement of each atom in the asymmetric unit of the unit cell with a probability-density function and to disregard all correlations between the motions of different atoms (independence assumption) (Johnson, 1970).

The most commonly used model in crystallography to describe the instantaneous position of an atom in a crystal is based on the **trivariate-normal distribution** (Johnson, 1970). In a crystal structure refinement, the parameters of this distribution (the mean atomic coordinates (μ_x, μ_y, μ_z) and variance $(\sigma_x^2, \sigma_y^2, \sigma_z^2)$) are refined for each atom of the asymmetric unit. The variance terms are used to model atomic displacements due to thermal motion of the atom and are most commonly referred to as the principle mean square displacement amplitudes (MSDAs). Consequently, the surface enclosed by the resulting trivariate-normal distribution

for a given probability (ie. probability ellipsoid) is referred to as the atom's **thermal ellipsoid** (Figure 2–2). In addition to its size, the so-called 'anisotropic' refinement model includes parameters for the orientation of the thermal ellipsoid in reciprocal space. Collectively, the three principle MSDAs and their relative orientation form a symmetric second rank tensor called the anisotropic temperature factor matrix. The six unique elements of the temperature factor matrix, called the **anisotropic temperature factors**, are included as regressor variables in the refinement model for each atom of the asymmetric unit. It is an analysis of the parameter estimates for anisotropic temperature factors determined from numerous crystal structure refinements that formed the basis of the study undertaken in Chapter 2.

The discussion will begin with a univariate normal distribution from which its so-called characteristic function will be derived. Extended to three variables (the x, y and z coordinates of an atom), the trivariate characteristic function multiplied by the atoms' scattering factor (ability to scatter X-rays) and summed over the n atoms of the unit cell leads to the familiar structure factor equation containing the so-called temperature factors. The isotropic temperature factor and the isotropic equivalent to the anisotropic temperature factor will then be developed. Last, the notion of the mean square displacement of an atom will be discussed in terms of the variance of an atoms' mean position.

2A.1 Anisotropic Temperature Factors

The first assumption made is that the position of the atom is not fixed but exhibits a distribution of values around some mean position as a result of thermal motion. For simplicity, the discussion will begin by examining only the x coordinate of an atoms' location. Assume that the x coordinate, written with respect to some cartesian basis ($C = (\mathbf{i}, \mathbf{j}, \mathbf{k})$), is a normally distributed random variable

with mean, μ_x, and variance, $\mathrm{Var}(x) = \sigma_x^2$ or in statistical notation, $x \sim N(\mu_x, \sigma_x^2)$ The probability density function for x, $P(x)$, is written as

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2}.$$

The **characteristic function** is given as the Fourier transform of the probability density function, $\mathcal{F}(P(x))$ (Wilks, 1962). A method for obtaining the Fourier transform of the univariate normal distribution is presented in the following discussion and is valid for any known continuous probability density function. In this case, the function represents a normal (or Gaussian) distribution but could represent any continuous distribution such as an exponential distribution. The definition of the Fourier transform of a function (in this case $P(x)$):

$$\begin{aligned}
\mathcal{F}(P(x)) &= \int_{-\infty}^{+\infty} P(x) e^{2\pi i h x} dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2} e^{2\pi i h x} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2} e^{2\pi i h x} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2 + 2\pi i h x} dx, \qquad (1)
\end{aligned}$$

where h may be considered as a coefficient of some other vector in C space.

The method of trigonometric substitution will be used to evaluate the integral in Equation 1. This method basically involves systematically making substitutions for various terms in the integral, along with factoring out other terms, so as to simplify the integral. To begin let $y = x - \mu_x$. Therefore, $\frac{dy}{dx} = 1$ so that $dy = dx$. Recall that upon substitution the limits of the integral must also be changed. This is accomplished by substituting the limiting values of x into the expression chosen for substitution (in this case, our expression for y). If $x = -\infty$ then $y = -\infty$ and

if $x = \infty$ then $y = \infty$ so that Equation 1 can be rewritten as

$$\begin{aligned}\mathcal{F}(P(x)) &= \frac{1}{\sqrt{2\pi}\sigma_x}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(\frac{y}{\sigma_x})^2+2\pi ih(y+\mu_x)}dy \\ &= \frac{e^{2\pi ih\mu_x}}{\sqrt{2\pi}\sigma_x}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(\frac{y}{\sigma_x})^2+2\pi ihy}dy.\end{aligned}$$

For convenience let $y = x$ so that

$$\mathcal{F}(P(x)) = \frac{e^{2\pi ih\mu_x}}{\sqrt{2\pi}\sigma_x}\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\sigma_x^2}+2\pi ihx}dx. \tag{2}$$

Now, lets look at the term in the exponent of the integration term of Equation 2. In other words, consider

$$-\frac{x^2}{2\sigma_x^2} + 2\pi ihx. \tag{3}$$

Note that the first term involves only the square of certain quantities and that the last term contains no squares. This suggests Equation 3 may be rewritten in terms of some function squared. This may provide additional simplification of the integral given in Equation 2. This function is obtained by the method known as completing the square. In other words, if Q equals whatever needs to be subtracted from the right hand side of the following equality, then

$$\begin{aligned}(-\frac{x^2}{2\sigma_x^2} + 2\pi ihx) &= -(\frac{x}{\sqrt{2}\sigma_x} - \sqrt{2}\pi ih\sigma_x)(\frac{x}{\sqrt{2}\sigma_x} - \sqrt{2}\pi ih\sigma_x) - Q \\ &= -(\frac{x^2}{2\sigma_x^2} - \frac{x\sqrt{2}\pi ih\sigma_x}{\sqrt{2}\sigma_x} - \frac{x\sqrt{2}\pi ih\sigma_x}{\sqrt{2}\sigma_x} - 2\pi^2\sigma_x^2h^2) - Q \\ &= -(\frac{x^2}{2\sigma_x^2} - 2\pi ihx - 2\pi^2\sigma_x^2h^2) - Q \\ &= (-\frac{x^2}{2\sigma_x^2} + 2\pi ihx + 2\pi^2\sigma_x^2h^2) - Q.\end{aligned}$$

This implies that $Q = 2\pi^2\sigma_x^2h^2$. Thus, Equation 2 becomes

$$\mathcal{F}(P(x)) = \frac{e^{2\pi ih\mu_x}e^{-2\pi^2\sigma_x^2h^2}}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} e^{-(\frac{x}{\sqrt{2}\sigma_x}-\sqrt{2}\pi ih\sigma_x)^2}dx. \tag{4}$$

To continue simplification, let $y = \frac{x}{\sqrt{2}\sigma_x} - \sqrt{2}\pi i h \sigma_x$. Therefore, $\frac{dy}{dx} = 1/(\sqrt{2}\sigma_x)$ so that $dx = \sqrt{2}\sigma_x dy$. If $x = -\infty$ then $y = -\infty$ and if $x = \infty$ then $y = \infty$ so that Equation 4 can be rewritten as

$$\begin{aligned}
\mathcal{F}(P(x)) &= \frac{e^{2\pi i h \mu_x} e^{-2\pi^2 \sigma_x^2 h^2}}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{+\infty} e^{-y^2} \sqrt{2}\sigma_x dy \\
&= \frac{e^{2\pi i h \mu_x} e^{-2\pi^2 \sigma_x^2 h^2} \sqrt{2}\sigma_x}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{+\infty} e^{-y^2} dy \\
&= \frac{e^{2\pi i h \mu_x} e^{-2\pi^2 \sigma_x^2 h^2}}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-y^2} dy. \qquad (5)
\end{aligned}$$

At this point the integral $\int_0^{\infty} e^{-t^2} dt = \sqrt{\pi}/2$ will be used (Abramowitz and Stegun, 1975). In this case, the solution must be multiplied by 2 (allowed by symmetry) to account for the interval 0 through $-\infty$. Finally,

$$\begin{aligned}
\mathcal{F}(P(x)) &= \frac{e^{2\pi i h \mu_x} e^{-2\pi^2 \sigma_x^2 h^2}}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-y^2} dy \\
&= \frac{e^{2\pi i h \mu_x} e^{-2\pi^2 \sigma_x^2 h^2}}{\sqrt{\pi}} 2(\frac{\sqrt{\pi}}{2}) \\
&= e^{2\pi i h \mu_x} e^{-2\pi^2 \sigma_x^2 h^2} \\
&= e^{2\pi i h \mu_x - 2\pi^2 \sigma_x^2 h^2}. \qquad (6)
\end{aligned}$$

Equation 6 is called the **characteristic function** of a univariate normal probability density function for the normal random variable x (Wilks, 1962).

Using the characteristic function given in Equation 6, the parameters that appear in the definition of a normally distributed random variable (i.e. μ and σ^2) may be obtained (Milton and Arnold, 1990). Note that this is the general method of obtaining the mean and variance of a random variable following any known continuous distribution. For purposes of discussion, Equation 6 will be rewritten, substituting $t = 2\pi i h$, as the **moment generating function** for x,

$m_x(t)$, (Milton and Arnold, 1990)

$$\begin{aligned} m_x(t) &= e^{2\pi i h\mu_x - 2\pi^2\sigma_x^2 h^2} \\ &= e^{t\mu_x + \frac{t^2\sigma_x^2}{2}}. \end{aligned} \tag{7}$$

By definition, the k moments for a random variable x are given by

$$\frac{d^k m_x(t)}{dt^k}\Big|_{t=0} = E[x^k].$$

Also, by definition the **mean of a random variable** x is $E[x]$ (the first moment) and the **variance** of x is $\mathrm{Var}(x) = \sigma_x^2 = E[(x - E[x])^2] = E[x^2] - (E[x])^2$ (the second central moment) (Milton and Arnold, 1990). It then follows from the definitions above and Equation 7 that

$$\frac{dm_x(t)}{dt} = (\mu_x + t\sigma_x^2)e^{t\mu_x + \frac{t^2\sigma_x^2}{2}}$$

So that

$$E[x] = \frac{dm_x(t)}{dt}\Big|_{t=0} = \mu_x.$$

Similarly,

$$\begin{aligned} \frac{d^2 m_x(t)}{dt^2} &= \frac{d^2}{dt^2}[(\mu_x e^{t\mu_x + \frac{t^2\sigma_x^2}{2}}) + (t\sigma_x^2 e^{t\mu_x + \frac{t^2\sigma_x^2}{2}})] \\ &= (\mu_x^2 + t\mu_x\sigma_x^2)e^{t\mu_x + \frac{t^2\sigma_x^2}{2}} + \sigma_x^2 e^{t\mu_x + \frac{t^2\sigma_x^2}{2}} + (t\mu_x\sigma_x^2 + t^2\sigma_x^4)e^{t\mu_x + \frac{t^2\sigma_x^2}{2}}. \end{aligned}$$

So that

$$E[x^2] = \frac{d^2 m_x(t)}{dt^2}\Big|_{t=0} = \mu_x^2 + \sigma_x^2.$$

Making the appropriate substitutions, we have

$$\begin{aligned} \mathrm{Var}(x) &= E[(x - E[x])^2] \\ &= E[x^2] - (E[x])^2 \\ &= (\mu_x^2 + \sigma_x^2) - (\mu_x)^2 \\ &= \sigma_x^2. \end{aligned}$$

The mean and variance of a normally distributed random variable x is μ_x and σ_x^2, respectively (Milton and Arnold, 1990).

Assuming that each coordinate of the atom, x, y, and z, is an independent and normally distributed random variable, then the total probability density for all three random variables is given by the product of the individual probabilities. This gives rise to the trivariate normal distribution for independent random variables,

$$P(x,y,z) = P(x)P(y)P(z)$$
$$= \frac{1}{(2\pi)^{3/2}(\sigma_x\sigma_y\sigma_z)} e^{-\frac{1}{2}\begin{bmatrix} x-\mu_x & y-\mu_y & z-\mu_z \end{bmatrix}\begin{bmatrix} 1/\sigma_x^2 & 0 & 0 \\ 0 & 1/\sigma_y^2 & 0 \\ 0 & 0 & 1/\sigma_z^2 \end{bmatrix}\begin{bmatrix} x-\mu_x \\ y-\mu_y \\ z-\mu_z \end{bmatrix}}.$$

Note that the exponent term of the distribution is the equation for an ellipsoid centered at (μ_x, μ_y, μ_z) with the three principal axis lengths given by σ_x, σ_y and σ_z. It then follows from the independence assumption, that the trivariate characteristic function, $\mathcal{F}(P(x,y,z))$, is the product of the individual characteristic functions of each random variable. Therefore,

$$\mathcal{F}(P(xyz)) = (e^{2\pi i h\mu_x - 2\pi^2\sigma_x^2 h^2})(e^{2\pi i k\mu_y - 2\pi^2\sigma_y^2 k^2})(e^{2\pi i l\mu_z - 2\pi^2\sigma_z^2 l^2})$$

By defining $[\mathbf{s}]_C = [\,h\,k\,l\,]^t$ and $[\mathbf{v}]_C = [\mu_x\mu_y\mu_z]^t$ where $C = (\mathbf{i},\mathbf{j},\mathbf{k})$ and the origin at (μ_x, μ_y, μ_z), then $\mathcal{F}(P(xyz))$ can be written in matrix notation as

$$\mathcal{F}(P(x)P(y)P(z)) = e^{\left(2\pi i[\mathbf{s}]_C^t[\mathbf{v}]_C - 2\pi^2[\mathbf{s}]_C^t \begin{bmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{bmatrix}[\mathbf{s}]_C\right)}$$
$$= e^{2\pi i[\mathbf{s}]_C^t[\mathbf{v}]_C - 2\pi^2[\mathbf{s}]_C^t V[\mathbf{s}]_C}, \tag{8}$$

where the variance or dispersion matrix, V, is defined as

$$V = \begin{bmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{bmatrix}.$$

Equation 8 is the characteristic function of a trivariate normal distribution for independent random variables. Observe that in a crystal structure refinement, there is an equation of the form given in Equation 8 for each atom in the unit cell. Keep in mind that this implies that each atom possess a unique cartesian basis, C, centered at the atoms mean position represented by the vector, $\mathbf{v}$. Note that in addition to assuming that each atoms' coordinates are independent, the assumption is made that each atom is independent of the other atoms. If the characteristic function for each of the j atoms is multiplied by a scattering factor for that atom, $f(j)$, and sum over all n atoms of the unit cell, then the so-called **structure factor equation** is obtained. In other words.

$$\begin{aligned} F(hkl) &= \sum_{j=1}^{n} f(j)\mathcal{F}(P(x)P(y)P(z)) \\ &= \sum_{j=1}^{n} f(j)e^{2\pi i[\mathbf{s}]_C^t[\mathbf{v}]_C - 2\pi^2[\mathbf{s}]_C^t V[\mathbf{s}]_C}. \end{aligned} \tag{9}$$

However, note that the form of the structure factor given in Equation 9 is written with respect to a cartesian basis. For use in crystallography, it is desirable to rewrite the characteristic function (or structure factor) with respect to reciprocal space, $D^* = (\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*)$. The following circut diagram shows how to transform the above expression into D^*,

$$\begin{array}{ccc} [\mathbf{w}]_D & \xrightarrow{A} & [\mathbf{w}]_C \\ {\scriptstyle G}\downarrow & & \uparrow{\scriptstyle I_3}, \\ [\mathbf{w}]_{D^*} & \xrightarrow{A^{-t}} & [\mathbf{w}]_C \end{array}$$

where $D = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ is called the direct basis and G is the metrical matrix. Note that a vector $[\mathbf{w}]_C = A^{-t}[\mathbf{w}]_{D^*}$ and $G^{-1} = G^* = A^{-1}A^{-t}$ (Boisen and Gibbs, 1985). Remember there will be a unique A matrix for each atom because of the differing cartesian bases and that the origin of C and D is at $[\mathbf{v}]_D$.

Making these substitutions into the characteristic function (Equation 8), we have

$$\begin{aligned}
\mathcal{F}(P(xyz)) &= e^{2\pi i[\mathbf{s}]_C^t[\mathbf{v}]_C - 2\pi^2[\mathbf{s}]_C^t V[\mathbf{s}]_C} \\
&= e^{2\pi i[\mathbf{s}]_{D^*}^t(A^{-1}A^{-t})[\mathbf{v}]_{D^*} - 2\pi^2[\mathbf{s}]_{D^*}^t(A^{-1}VA^{-t})[\mathbf{s}]_{D^*}} \\
&= e^{2\pi i[\mathbf{s}]_{D^*}^t G^*[\mathbf{v}]_{D^*} - 2\pi^2[\mathbf{s}]_{D^*}^t(A^{-1}VA^{-t})[\mathbf{s}]_{D^*}} \\
&= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - 2\pi^2[\mathbf{s}]_{D^*}^t(A^{-1}VA^{-t})[\mathbf{s}]_{D^*}} \qquad \text{(10a)}
\end{aligned}$$

Now, observe from the circut diagram above that the form of A^{-t} must be

$$\begin{aligned}
A^{-t} &= [\,[\mathbf{a}^*]_C \quad [\mathbf{b}^*]_C \quad [\mathbf{c}^*]_C\,] \\
&= \begin{bmatrix} \mathbf{i}\cdot\mathbf{a}^* & \mathbf{i}\cdot\mathbf{b}^* & \mathbf{i}\cdot\mathbf{c}^* \\ \mathbf{j}\cdot\mathbf{a}^* & \mathbf{j}\cdot\mathbf{b}^* & \mathbf{j}\cdot\mathbf{c}^* \\ \mathbf{k}\cdot\mathbf{a}^* & \mathbf{k}\cdot\mathbf{b}^* & \mathbf{k}\cdot\mathbf{c}^* \end{bmatrix} \\
&= \begin{bmatrix} a^*\cos(\mathbf{a}^*\wedge\mathbf{i}) & b^*\cos(\mathbf{b}^*\wedge\mathbf{i}) & c^*\cos(\mathbf{c}^*\wedge\mathbf{i}) \\ a^*\cos(\mathbf{a}^*\wedge\mathbf{j}) & b^*\cos(\mathbf{b}^*\wedge\mathbf{j}) & c^*\cos(\mathbf{c}^*\wedge\mathbf{j}) \\ a^*\cos(\mathbf{a}^*\wedge\mathbf{k}) & b^*\cos(\mathbf{b}^*\wedge\mathbf{k}) & c^*\cos(\mathbf{c}^*\wedge\mathbf{k}) \end{bmatrix} \\
&= \begin{bmatrix} \cos(\mathbf{a}^*\wedge\mathbf{i}) & \cos(\mathbf{b}^*\wedge\mathbf{i}) & \cos(\mathbf{c}^*\wedge\mathbf{i}) \\ \cos(\mathbf{a}^*\wedge\mathbf{j}) & \cos(\mathbf{b}^*\wedge\mathbf{j}) & \cos(\mathbf{c}^*\wedge\mathbf{j}) \\ \cos(\mathbf{a}^*\wedge\mathbf{k}) & \cos(\mathbf{b}^*\wedge\mathbf{k}) & \cos(\mathbf{c}^*\wedge\mathbf{k}) \end{bmatrix} \begin{bmatrix} a^* & 0 & 0 \\ 0 & b^* & 0 \\ 0 & 0 & c^* \end{bmatrix} \\
&= CD
\end{aligned}$$

where C is a unitary matrix ($C^t = C^{-1}$) and D is a diagonal matrix. Making this substitution into Equation 10a,

$$\begin{aligned}
\mathcal{F}(P(xyz)) &= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - 2\pi^2[\mathbf{s}]_{D^*}^t(A^{-1}VA^{-t})[\mathbf{s}]_{D^*}} \\
&= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - 2\pi^2[\mathbf{s}]_{D^*}^t(DC^tVCD)[\mathbf{s}]_{D^*}} \\
&= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - 2\pi^2[\mathbf{s}]_{D^*}^t(DUD)[\mathbf{s}]_{D^*}}
\end{aligned}$$

$$= e^{2\pi i(h\mu_x + k\mu_y + l\mu_z) - 2\pi^2(U_{11}a^{*2}h^2 + U_{22}b^{*2}k^2 + U_{33}c^{*2}l^2 + 2U_{12}a^*b^*hk + 2U_{13}a^*c^*hl + 2U_{23}b^*c^*kl)}. \qquad \text{(10b)}$$

Equation 10b is the form of the characteristic equation commonly used in the structure factors of most refinements (Equation 9). Here the matrix U is defined

as $U = C^tVC$. U is a 3×3 symmetric matrix whose six unique elements are the so-called temperature factors of the atom. Thus, the temperature factor parameters used in the model include information about the variance of the mean position and its orientation. The individual elements of U can be written out in the following form:

$$
\begin{aligned}
U_{11} &= \sigma_1^2\cos^2(\mathbf{a}^* \wedge \mathbf{i}) + \sigma_2^2\cos^2(\mathbf{a}^* \wedge \mathbf{j}) + \sigma_3^2\cos^2(\mathbf{a}^* \wedge \mathbf{k}) \\
U_{22} &= \sigma_1^2\cos^2(\mathbf{b}^* \wedge \mathbf{i}) + \sigma_2^2\cos^2(\mathbf{b}^* \wedge \mathbf{j}) + \sigma_3^2\cos^2(\mathbf{b}^* \wedge \mathbf{k}) \\
U_{33} &= \sigma_1^2\cos^2(\mathbf{c}^* \wedge \mathbf{i}) + \sigma_2^2\cos^2(\mathbf{c}^* \wedge \mathbf{j}) + \sigma_3^2\cos^2(\mathbf{c}^* \wedge \mathbf{k}) \\
U_{12} = U_{21} &= \sigma_1^2\cos(\mathbf{a}^* \wedge \mathbf{i})\cos(\mathbf{b}^* \wedge \mathbf{i}) + \sigma_2^2\cos(\mathbf{a}^* \wedge \mathbf{j})\cos(\mathbf{b}^* \wedge \mathbf{j}) \\
&\quad + \sigma_3^2\cos(\mathbf{a}^* \wedge \mathbf{k})\cos(\mathbf{b}^* \wedge \mathbf{k}) \\
U_{13} = U_{31} &= \sigma_1^2\cos(\mathbf{a}^* \wedge \mathbf{i})\cos(\mathbf{c}^* \wedge \mathbf{i}) + \sigma_2^2\cos(\mathbf{a}^* \wedge \mathbf{j})\cos(\mathbf{c}^* \wedge \mathbf{j}) \\
&\quad + \sigma_3^2\cos(\mathbf{a}^* \wedge \mathbf{k})\cos(\mathbf{c}^* \wedge \mathbf{k}) \\
U_{23} = U_{32} &= \sigma_1^2\cos(\mathbf{b}^* \wedge \mathbf{i})\cos(\mathbf{c}^* \wedge \mathbf{i}) + \sigma_2^2\cos(\mathbf{b}^* \wedge \mathbf{j})\cos(\mathbf{c}^* \wedge \mathbf{j}) \\
&\quad + \sigma_3^2\cos(\mathbf{b}^* \wedge \mathbf{k})\cos(\mathbf{c}^* \wedge \mathbf{k}).
\end{aligned}
$$

An alternatively used form of the temperature factors reported in the literature may be obtained from Equation 10a as follows:

$$
\begin{aligned}
\mathcal{F}(P(xyz)) &= e^{2\pi i[\mathbf{s}]^t_{D^*}[\mathbf{v}]_D - 2\pi^2[\mathbf{s}]^t_{D^*}(A^{-1}VA^{-t})[\mathbf{s}]_{D^*}} \\
&= e^{2\pi i[\mathbf{s}]^t_{D^*}[\mathbf{v}]_D - [\mathbf{s}]^t_{D^*}\beta[\mathbf{s}]_{D^*}} \\
&= e^{2\pi i(h\mu_x + k\mu_y + l\mu_z) - (\beta_{11}h^2 + \beta_{22}k^2 + \beta_{33}l^2 + 2\beta_{12}hk + 2\beta_{13}hl + 2\beta_{23}kl)}.
\end{aligned}
\tag{10c}
$$

Here the matrix β is defined as $\beta = 2\pi^2(A^{-1}VA^{-t})$. β is a 3×3 symmetric matrix whose six unique elements represent an alternative form of the temperature

factors. The individual elements can be written out in the following form:

$$\beta_{11} = 2\pi^2 a^{*2}(\sigma_x^2 \cos^2(\mathbf{a}^* \wedge \mathbf{i}) + \sigma_y^2 \cos^2(\mathbf{a}^* \wedge \mathbf{j}) + \sigma_z^2 \cos^2(\mathbf{a}^* \wedge \mathbf{k}))$$

$$\beta_{22} = 2\pi^2 b^{*2}(\sigma_x^2 \cos^2(\mathbf{b}^* \wedge \mathbf{i}) + \sigma_y^2 \cos^2(\mathbf{b}^* \wedge \mathbf{j}) + \sigma_z^2 \cos^2(\mathbf{b}^* \wedge \mathbf{k}))$$

$$\beta_{33} = 2\pi^2 c^{*2}(\sigma_x^2 \cos^2(\mathbf{c}^* \wedge \mathbf{i}) + \sigma_y^2 \cos^2(\mathbf{c}^* \wedge \mathbf{j}) + \sigma_z^2 \cos^2(\mathbf{c}^* \wedge \mathbf{k}))$$

$$\beta_{12} = \beta_{21} = 2\pi^2 a^* b^*(\sigma_x^2 \cos(\mathbf{a}^* \wedge \mathbf{i})\cos(\mathbf{b}^* \wedge \mathbf{i}) + \sigma_y^2 \cos(\mathbf{a}^* \wedge \mathbf{j})\cos(\mathbf{b}^* \wedge \mathbf{j}) + \sigma_z^2 \cos(\mathbf{a}^* \wedge \mathbf{k})\cos(\mathbf{b}^* \wedge \mathbf{k}))$$

$$\beta_{13} = \beta_{31} = 2\pi^2 a^* c^*(\sigma_x^2 \cos(\mathbf{a}^* \wedge \mathbf{i})\cos(\mathbf{c}^* \wedge \mathbf{i}) + \sigma_y^2 \cos(\mathbf{a}^* \wedge \mathbf{j})\cos(\mathbf{c}^* \wedge \mathbf{j}) + \sigma_z^2 \cos(\mathbf{a}^* \wedge \mathbf{k})\cos(\mathbf{c}^* \wedge \mathbf{k}))$$

$$\beta_{23} = \beta_{32} = 2\pi^2 b^* c^*(\sigma_x^2 \cos(\mathbf{b}^* \wedge \mathbf{i})\cos(\mathbf{c}^* \wedge \mathbf{i}) + \sigma_y^2 \cos(\mathbf{b}^* \wedge \mathbf{j})\cos(\mathbf{c}^* \wedge \mathbf{j}) + \sigma_z^2 \cos(\mathbf{b}^* \wedge \mathbf{k})\cos(\mathbf{c}^* \wedge \mathbf{k})).$$

The relationship between these two commonly used forms of temperature factors (β_{ij} and U_{ij}) presented in Equations 10b and 10c is $\beta = 2\pi^2(DUD)$ or

$$\beta_{11} = 2\pi^2 U_{11} a^{*2}$$

$$\beta_{22} = 2\pi^2 U_{22} b^{*2}$$

$$\beta_{33} = 2\pi^2 U_{33} c^{*2}$$

$$\beta_{12} = \beta_{21} = 2\pi^2 U_{12} a^* b^*$$

$$\beta_{13} = \beta_{31} = 2\pi^2 U_{13} a^* c^*$$

$$\beta_{23} = \beta_{32} = 2\pi^2 U_{23} b^* c^*.$$

2A.2 Isotropic Temperature Factor

The reduced characteristic equation that contains the isotropic temperature factor can be obtained from Equation 8 by making the homogeneous (equal) vari-

ance assumption, In other words, if $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = \sigma^2$, then

$$\mathcal{F}(P(xyz)) = e^{2\pi i[\mathbf{s}]_C^t[\mathbf{v}]_C - 2\pi^2[\mathbf{s}]_C^t \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} [\mathbf{s}]_C}$$

$$= e^{2\pi i[\mathbf{s}]_C^t[\mathbf{v}]_C - 2\pi^2\sigma^2[\mathbf{s}]_C^t[\mathbf{s}]_C}$$

$$= e^{2\pi i[\mathbf{s}]_{D^*}^t(A^{-1}A^{-t})[\mathbf{v}]_{D^*} - 2\pi^2\sigma^2[\mathbf{s}]_{D^*}^t(A^{-1}A^{-t})[\mathbf{s}]_{D^*}}$$

$$= e^{2\pi i[\mathbf{s}]_{D^*}^t G^*[\mathbf{v}]_{D^*} - 2\pi^2\sigma^2[\mathbf{s}]_{D^*}^t G^*[\mathbf{s}]_{D^*}}$$

$$= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - 2\pi^2\sigma^2(1/d^2)}$$

$$= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - 2\pi^2\sigma^2(\frac{4\sin^2\theta}{\lambda^2})}$$

$$= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - 8\pi^2\sigma^2(\frac{\sin^2\theta}{\lambda^2})}$$

$$= e^{2\pi i[\mathbf{s}]_{D^*}^t[\mathbf{v}]_D - B\frac{\sin^2\theta}{\lambda^2}}.$$

Here $B = 8\pi^2\sigma^2$ is called the **isotropic temperature factor**. This isotropic model is often initially refined to obtain starting estimates for the full model presented in Equation 10b or 10c.

2A.3 Isotropic Equivalent to the Anisotropic Temperature Factor

For the purposes of discussion, consider the temperature factors only in the U_{ij} form. It will be shown later that the eigenvalues of the matrix $DUDG$ represent σ_x^2, σ_y^2, and σ_z^2 in D space. These eigenvalues written along the diagonal of a matrix represent the elements of the diagonalized $DUDG$ matrix, or the V matrix. Because $DUDG$ is by definition similar to V, the trace of the two matrices are equal. Recall that the trace of a matrix is invariant under a similarity transformation. As a result, the average variance of the atoms mean position, $\langle\sigma^2\rangle$, is obtained simply by examining the trace of $DUDG$. As discussed in the next section, $\langle\sigma^2\rangle$ represents the average mean square displacement of the atom

due to thermal motion. In other words,

$$\begin{aligned}\langle\sigma^2\rangle &= \frac{\sigma_x^2+\sigma_y^2+\sigma_z^2}{3}\\ &= \frac{tr(DUDG)}{3}\\ &= \frac{tr(\beta G)}{6\pi^2}\end{aligned}$$

Recall that the isotropic temperature factor, $B = 8\pi^2\sigma^2$. An estimate of B, called $\overline{B}$, can be computed by using $\langle\sigma^2\rangle$. $\overline{B}$ is called the **isotropic equivalent to the anisotropic temperature factor** and is given by

$$\begin{aligned}\overline{B} &= 8\pi^2\langle\sigma^2\rangle\\ &= \frac{8\pi^2}{3}tr(DUDG)\\ &= 8\pi^2\left[\frac{tr(\beta G)}{6\pi^2}\right]\\ &= \frac{4}{3}tr(\beta G).\end{aligned}$$

2A.4 Statistical Interpretation of Mean Square Displacement

Equation 10b shows that the temperature factor matrix, U, for an atom is nothing more than the variance matrix, V, transformed by the C^{-t} matrix into reciprocal space, D^* ($U = C^tVC$). Because each coordinate is independent (i.e. each covariance is zero and hence V is diagonal), the three directions upon which the variance of each coordinate occur lie along three orthogonal directions emanating from the atoms mean position, **v**. The variance of a normal random variable provides a measure of the width of its normal distribution. In crystallography, the variance of each coordinate, V, is interpreted as the mean square displacement of the atom from its mean position due to thermal motion. Crystallographers examine the amount of variance or mean square displacement along various directions in the crystal to study the thermal behavior of atoms in crystals.

The amount of variance (or thermal displacement) along some direction can be examined with the U or β matrix (i.e. temperature factor matrix) because this matrix contains information about the magnitude and orientation of the variance components (extreme in thermal displacements). In other words, the temperature factor matrix describes the variance space of the atom. To examine the variance space described by the U or β matrix, consider a vector in the direction of interest, $\mathbf{w}$, upon which the variance can be projected. The amount of variance in the direction of $\mathbf{w}$, called the mean square displacement of an atom along the unit vector $\mathbf{w}$, is given the symbol $\mathbf{z}_{\mathbf{w}}^2$. Computationally,

$$
\begin{aligned}
\mathbf{z}_{\mathbf{w}}^2 &= \sigma_{\mathbf{w}}^2 \\
&= \frac{[\mathbf{w}]_C^t V [\mathbf{w}]_C}{[\mathbf{w}]_C^t [\mathbf{w}]_C} \\
&= \frac{[\mathbf{w}]_{D^*}^t (A^{-1} V A^{-t}) [\mathbf{w}]_D^*}{[\mathbf{w}]_{D^*}^t (A^{-1} A^{-t}) [\mathbf{w}]_{D^*}} \\
&= \frac{[\mathbf{w}]_{D^*}^t (D C^t V C D) [\mathbf{w}]_D^*}{[\mathbf{w}]_{D^*}^t G^* [\mathbf{w}]_{D^*}} \\
&= \frac{[\mathbf{w}]_{D^*}^t (D U D) [\mathbf{w}]_D^*}{[\mathbf{w}]_{D^*}^t G^* [\mathbf{w}]_{D^*}} \\
&= \frac{[\mathbf{w}]_D^t (G D U D G) [\mathbf{w}]_D}{[\mathbf{w}]_D^t (G G^* G) [\mathbf{w}]_D} \\
&= \frac{[\mathbf{w}]_D^t (G D U D G) [\mathbf{w}]_D}{[\mathbf{w}]_D^t G [\mathbf{w}]_D} \\
&= \frac{[\mathbf{w}]_D^t (G \beta G) [\mathbf{w}]_D}{2\pi^2 [\mathbf{w}]_D^t G [\mathbf{w}]_D}
\end{aligned}
\tag{11}
$$

A plot of the surface of all $\mathbf{z}_{\mathbf{w}}^2$ given by Equation 11 for all possible $\mathbf{w}$ represents a three-dimensional surface resembling a peanut in shape in which three extreme occur along orthogonal directions, each representing a principle direction of variance. A two-dimensional example is given later in this appendix. The directions of these extreme can then used to formulate A^{-t} and subsequently diagonalize U

or β according to

$$\begin{aligned} V &= C^{-t}UC^{t} \\ &= CUC^{t} \\ &= \frac{D^{-1}\beta D^{-1}}{2\pi^2}. \end{aligned}$$

In order to find an extreme, Equation 11 must be differentiated with respect to $\mathbf{w}$ and the result set equal to zero to solve for $\mathbf{w}$. Each of the three extreme will have a unique $\mathbf{w}$ associated with it that will collectively form an orthogonal set. For derivation purposes, the U form of the temperature factor matrix in Equation 11 will be used so that

$$\begin{aligned} \frac{d\mathbf{z}_{\mathbf{w}}^2}{d\mathbf{w}} &= \frac{d}{d\mathbf{w}}\left(\frac{[\mathbf{w}]_D^t(GDUDG)[\mathbf{w}]_D}{[\mathbf{w}]_D^t G[\mathbf{w}]_D}\right) \\ &= \frac{(2GDUDG[\mathbf{w}]_D)([\mathbf{w}]_D^t G[\mathbf{w}]_D) - (2G[\mathbf{w}]_D)([\mathbf{w}]_D^t GDUDG[\mathbf{w}]_D)}{([\mathbf{w}]_D^t G[\mathbf{w}]_D)^2} \\ &= 0 \end{aligned}$$

Attempting to solve for $[\mathbf{w}]$ we have,

$$\begin{aligned} (GDUDG[\mathbf{w}]_D)([\mathbf{w}]_D^t G[\mathbf{w}]_D) &= (G[\mathbf{w}]_D)([\mathbf{w}]_D^t GDUDG[\mathbf{w}]_D) \\ (GDUDG[\mathbf{w}]_D) &= (G[\mathbf{w}]_D)\left(\frac{[\mathbf{w}]_D^t GDUDG[\mathbf{w}]_D}{[\mathbf{w}]_D^t G[\mathbf{w}]_D}\right) \\ &= \mathbf{z}_{\mathbf{w}}^2 G[\mathbf{w}]_D \\ &= \sigma_{\mathbf{w}}^2 G[\mathbf{w}]_D \\ DUDG[\mathbf{w}]_D &= \sigma_{\mathbf{w}}^2[\mathbf{w}]_D \\ \beta G[\mathbf{w}]_D &= 2\pi^2\sigma_{\mathbf{w}}^2[\mathbf{w}]_D, \end{aligned}$$

where an **eigenvalue problem** must be solved in order to find $\mathbf{w}$ (note that $\sigma_{\mathbf{w}}^2$

is also unknown). The eigenvalue problem is constructed as follows:

$$\begin{aligned} DUDG[\mathbf{w}]_D &= \sigma_{\mathbf{w}}^2[\mathbf{w}]_D \\ DUDG[\mathbf{w}]_D &= \lambda[\mathbf{w}]_D \\ DUDG[\mathbf{w}]_D - \lambda(I_3)[\mathbf{w}]_D &= 0 \\ (DUDG - \lambda(I_3))[\mathbf{w}]_D &= 0. \end{aligned} \tag{12}$$

Set the determinate or $|(DUDG - \lambda(I_3))| = 0$ and solve for λ. Because the form of the determinate will be cubic in λ, there will be $i = 3$ possible solutions, λ_i. The so-called eigenvalues (or roots), λ_i, represent σ_i^2 or $2\pi^2\sigma_i^2$ depending on the form of the temperature factor matrix used, and i represents x, y, z. Note that all eigenvalues must be positive because σ_i^2 is a positive quantity. This requires that the temperature factor matrix be positive definite. Once the eigenvalues are obtained, each λ_i is substituted back into Equation 12 to obtain $\mathbf{w}_i$, the direction of the extreme or eigenvector. Note that both the eigenvalue and eigenvector are defined in terms of the D basis. Also, observe that the matrix whose columns contain the eigenvectors (converted to a unit vector) is the C^{-1} or C^t matrix for that atom.

Upon completion of a structure refinement, estimates have been determined for μ_x, σ_x^2, μ_y, σ_y^2; and μ_z, σ_z^2 for the trivariate normal distribution for independent random variables for all atoms (i.e. mean position and temperature factors for each atom) To gain insight into the distribution of space that the atom occupies as a result of thermal motion, a picture is constructed of the probability density distribution of the atom by choosing some representative value of probability. The surface generated by the density function that encloses the probability represents an ellipsoid. In crystallography, this ellipsoid is called the thermal ellipsoid of the atom. From the exponential term of the trivariate normal distribution, assuming

the K^2 surface encloses a certain percentage of the total probability, compute

$$\begin{aligned} K^2 &= [\mathbf{w}]_C^t \begin{bmatrix} 1/\sigma_x^2 & 0 & 0 \\ 0 & 1/\sigma_y^2 & 0 \\ 0 & 0 & 1/\sigma_z^2 \end{bmatrix} [\mathbf{w}]_C \\ &= [\mathbf{w}]_C^t V^{-1} [\mathbf{w}]_C \\ &= [\mathbf{w}]_C^t (CU^{-1}C^t)[\mathbf{w}]_C \\ &= [\mathbf{w}]_C^t (CU^{-1}C^t)[\mathbf{w}]_C \\ &= [\mathbf{w}]_{D^*}^t (DC^tCU^{-1}C^tCD)[\mathbf{w}]_{D^*} \\ &= [\mathbf{w}]_{D^*}^t (DU^{-1}D)[\mathbf{w}]_{D^*} \\ &= [\mathbf{w}]_D^t (GDU^{-1}DG)[\mathbf{w}]_D \\ &= 2\pi^2 [\mathbf{w}]_D^t (GDD\beta^{-1}DDG)[\mathbf{w}]_D \end{aligned} \tag{13}$$

If $K = 1.5382$, then the surface generated by Equation 13 contains 50% of the total probability. Each eigenvector determined from Equation 12 is parallel to a principal axis of the ellipsoid whose length is given by the square root of the eigenvalue.

As a two-dimensional example, consider a Cartesian basis, $C = (\mathbf{i}, \mathbf{j})$, and the (2×2) symmetric matrix

$$\beta = \begin{bmatrix} 2 & -2 \\ -2 & 5 \end{bmatrix}.$$

The peanut shaped drawing in Figure 2–9 shows a plot of the mean square displacement, $\mathbf{z}_\mathbf{w}^2$, given by Equation 11 as

$$\begin{aligned} \mathbf{z}_\mathbf{w}^2 &= \sigma_\mathbf{w}^2 \\ &= \frac{[\mathbf{w}]_C^t \beta [\mathbf{w}]_C}{[\mathbf{w}]_C^t [\mathbf{w}]_C}. \end{aligned}$$

The eigenvalue matrix, V, and eigenvector matrix, C^t, of β are

$$V = \begin{bmatrix} 6 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$C^t = \begin{bmatrix} -1/\sqrt{5} & 2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}.$$

The ellipse shown in Figure 2–9 is given by Equation 13 as

$$1.0 = [\mathbf{w}]_C^t C V^{-1} C^t [\mathbf{w}]_C$$

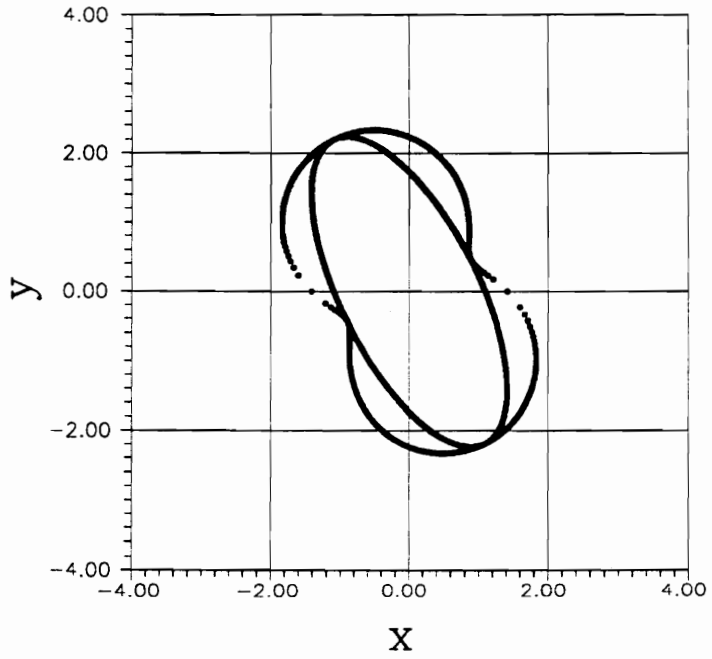


Figure 2–9. Plots of $\mathbf{z}_{\mathbf{w}}^2$ and $K^2 = 1$ for the β matrix given above.

APPENDIX 2B

Ellipsoid Agreement Parameters

Let M and M' represent two 3×3 real symmetric positive definite matrices written with respect to a Cartesian basis, $C = (\mathbf{i}, \mathbf{j}, \mathbf{k})$. Therefore, the graph of the function $[v]_C^t M [v]_C = 1$, where $[v]_C$ is the triple representative of any vector in C, represents an ellipsoid in C (For notation, see Boisen and Gibbs, 1985). A graph of the function $[v]_C^t M' [v]_C = 1$ represents a different ellipsoid in C if $M \neq M'$. The ellipsoid agreement parameters (EAPs) were constructed to provide a measure of the relative physical differences in size, shape, and orientation between two ellipsoids, centered at the same position. The first parameter, EAP1, measures the relative agreement between the sizes of two ellipsoids. A reasonable measure of the size of an ellipsoid is provided by the average of its three principle axis lengths (Appendix 2A). Computationally this is found by dividing the trace of M by three. This average can then be considered as the radius of a sphere that forms an isotropic equivalent to the ellipsoid, B_M. The size parameter, EAP1, is then computed as:

$$\text{EAP1} = \frac{|B_M - B_{M'}|}{B_M}.$$

The next parameter, EAP3, measures the difference in orientation between two ellipsoids. Let U_M and $U_{M'}$ represent unitary matrices consisting of the normalized eigenvectors of M and M', respectively, so that

$$U_M = \begin{bmatrix} [\mathbf{e}_1]_C & [\mathbf{e}_2]_C & [\mathbf{e}_3]_C \end{bmatrix}$$

and

$$U_{M'} = \begin{bmatrix} [\mathbf{e}'_1]_C & [\mathbf{e}'_2]_C & [\mathbf{e}'_3]_C \end{bmatrix}.$$

Note that the normalized eigenvectors of M and M' each represent an orthogonal basis, $E = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $E' = (\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$, respectively, with a common origin. Therefore, we can consider U_M and $U_{M'}$ as transformation matrices such that $U_M[\mathbf{v}]_E = [\mathbf{v}]_C$ and $U_{M'}[\mathbf{v}]_{E'} = [\mathbf{v}]_C$ (Boisen and Gibbs, 1985). Equating these transformations results in

$$\begin{aligned}[\mathbf{v}]_E &= U_M^{-1} U_{M'}[\mathbf{v}]_{E'} \\ &= U_M^t U_{M'}[\mathbf{v}]_{E'} \\ &= R[\mathbf{v}]_{E'},\end{aligned}$$

where $R = U_M^t U_{M'}$. Because R is the composition of two unitary matrices, R is also a unitary matrix and may be considered as a rotation matrix (Boisen and Gibbs, 1985). Because R represents the angular orientation between E and E', the turn angle of R provides a measure of the relative difference in orientation between the two ellipsoids represented by M and M' (Burns et al., 1969). The turn angle of R, ρ, is found by setting the trace of R equal to $1 \pm 2\cos\rho$ (Boisen and Gibbs, 1985) and represents the amount of rotation, about the vector $[\mathbf{l}]_C$, required to bring the basis E' into coincidence with the basis E. Note that because $\mathbf{R}$ is computed from normalized eigenvectors of M and M', ρ is dependent on the order in which they are listed by column in U_M and $U_{M'}$, respectively. Also note that if for example $[\mathbf{e}_1]_C$ is a normalized eigenvector of M, then by symmetry $-[\mathbf{e}_1]_C$ is also an normalized eigenvector of M. Thus, ρ is also dependent on the directions of the eigenvectors listed in U_M and $U_{M'}$. Systematically permuting the directions and order of the eigenvectors in U_M and $U_{M'}$, respectively, results in 42 unique R matrices and thus 42 unique ρ values each computed according to

$$\rho = \cos^{-1}[(trace(R) - 1)/2].$$

The smallest unique ρ was chosen as the parameter EAP3 thus providing a measure of the difference in orientation between two ellipsoids. If the two ellipsoids each contain a circular cross section, EAP3 is computed as the angle between the eigenvectors perpendicular to the circular section. Finally, if the two ellipsoids are spherical, then EAP3 is zero.

The final parameter, EAP2, measures the relative difference in shape between the two coincident ellipsoids (Burns et al., 1967). A measure of an ellipsoid's shape can be constructed by forming a unit vector from the eigenvalues of the ellipsoid. If we define $S = (\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3)$ as the orthonormal shape basis, then

$$\mathbf{s} = \frac{1}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}(\lambda_1\mathbf{s}_1 + \lambda_2\mathbf{s}_2 + \lambda_3\mathbf{s}_3)$$

and

$$\mathbf{s}' = \frac{1}{(\lambda_1'^2 + \lambda_2'^2 + \lambda_3'^2)}(\lambda_1'\mathbf{s}_1 + \lambda_2'\mathbf{s}_2 + \lambda_3'\mathbf{s}_3),$$

where $\mathbf{s}$ and $\mathbf{s}'$ are the shape vectors and $(\lambda_1, \lambda_2, \lambda_3)$ and $(\lambda_1', \lambda_2', \lambda_3')$ are the eigenvalues for M and M', respectively. If X is defined as the distance between the endpoints of $\mathbf{s}$ and $\mathbf{s}'$, then according to the law of cosines,

$$\begin{aligned} X^2 &= |\mathbf{s}|^2 + |\mathbf{s}'|^2 - 2(\mathbf{s} \cdot \mathbf{s}') \\ &= 2(1 - \mathbf{s} \cdot \mathbf{s}'). \end{aligned}$$

In accordance with Burns et al. (1967), the shape parameter EAP2 is computed as EAP2 $= (10,000 \times X^2)$. Note that the coefficients of $\mathbf{s}$, λ_1, λ_2, and λ_3, represent the ordered (smallest to largest) triad of eigenvalues of M. The coefficients of $\mathbf{s}'$, λ_1', λ_2', and λ_3', corresponded to the (1,1), (2,2), and (3,3) elements of Λ', respectively, where Λ' is computed according to

$$\begin{aligned} \Lambda' &= (U_{M'})^t B' U_{M'} \\ &= R^t (U_M)^t B' U_M R, \end{aligned}$$

where R is the rotation matrix representing EAP3.

As an example, consider the ADP matrix for the Al5 atom from the refinement of triclinic bikitaite (Bissert and Liebau, 1986). Transformed into a Cartesian basis, the ADP matrix becomes

$$M = \begin{bmatrix} .00529 & -.00048 & -.00020 \\ -.00048 & .00529 & .00012 \\ -.00020 & .00012 & .00585 \end{bmatrix}.$$

From a TLS analysis on the Al5 tetrahedron, the calculated ADP matrix is

$$M' = \begin{bmatrix} .00553 & -.00019 & -.00102 \\ -.00019 & .00583 & .00007 \\ -.00102 & .00007 & .00594 \end{bmatrix}.$$

Using the diagonal elements of both matrices, M and M', EAP1 is then computed as

$$\begin{aligned} EAP1 &= \frac{|((.00529 + .00529 + .00585)/3) - ((.00553 + .00583 + .00594)/3)|}{((.00529 + .00529 + .00585)/3)} \\ &= |.00548 - .00577|/.00548 \\ &= 0.053. \end{aligned}$$

The eigenvector matrices of M and M' are

$$U_M = \begin{bmatrix} .71816 & -.50491 & -.47886 \\ .69343 & .57695 & .43161 \\ .05836 & -.64202 & .76446 \end{bmatrix}$$

and

$$U_{M'} = \begin{bmatrix} .77592 & .04548 & -.62920 \\ .08634 & .98036 & .17734 \\ .62490 & -.19192 & .75675 \end{bmatrix},$$

respectively. By systematically permuting the columns of U_M and $U_{M'}$ and computing $\rho = \cos^{-1}[(trace(U_M^t U_{M'}) - 1)/2]$, results in the following possible values

of EAP3 (in units of degrees):

50.36512, 157.09911, 163.18183, 96.34870, 152.45227, 139.66569, 169.84140, 137.08715, 138.07657, 100.84512, 113.45986, 87.30535, 132.96488, 109.47860, 177.58958, 138.24508, 81.87923, 112.40431, 166.41430, 90.74347, 45.54239, 162.79760, 148.92892, 154.63437, 166.41430, 90.74347, 45.54239, 162.79760, 148.92892, 154.63437, 132.96488, 109.47860, 177.58958, 138.24508, 81.87923, 112.40431, 169.84140, 137.08715, 138.07657, 100.84512, 113.45986, 87.30535.

By definition we choose the smallest value, therefore EAP3 = 45.54239° which is computed from

$$R = -1 \times (U_M^t U_{M'}) = \begin{bmatrix} .70127 & .65357 & -.28473 \\ -.66587 & .74315 & .06585 \\ .25463 & .14342 & .95634 \end{bmatrix},$$

where $(U_M^t U_{M'})$ is multiplied by -1 because it represents an improper rotation.

Finally, the corresponding eigenvalues for U_M and $U_{M'}$ are

$$\lambda_1 = .00480$$

$$\lambda_2 = .00558$$

$$\lambda_3 = .00604$$

and

$$\lambda_1' = .00581$$

$$\lambda_2' = .00469$$

$$\lambda_3' = .00681,$$

respectively. We can the formulate $\mathbf{s}$ and $\mathbf{s}'$ where,

$$\mathbf{s} = .50429\mathbf{s}_1 + .58620\mathbf{s}_2 + .63408\mathbf{s}_3$$

and

$$\mathbf{s}' = .57472\mathbf{s}_1 + .46436\mathbf{s}_2 + .67384\mathbf{s}_3$$

We can then compute the distance between the endpoints as

$$\begin{aligned} X^2 &= 2(1 - \mathbf{s} \cdot \mathbf{s}') \\ &= 2(1 - .98931) \\ &= 0.021380, \end{aligned}$$

which results in EAP2 = 213.80.

APPENDIX 2C

Multiple Linear Regression

2C.1 Linear Models and Least Squares

The general form of the multiple linear regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i,$$

where $i = 1, ..., n$ observations (Myers, 1990). Recast in matrix notation the multiple linear regression model becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ 1 & x_{3,1} & x_{3,2} & \cdots & x_{3,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_k \end{bmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is the vector of **response variables** or dependent variables, $\mathbf{X}$ is the matrix of **regressor variables** or independent variables, $\boldsymbol{\beta}$ is the vector of **regression parameters** that must be estimated, k is the number of the number of regressor variables or **regressors**, and n is the sample size. The number of regression parameters with an intercept term β_0 is $p = k + 1$. Note that $n \geq p$ must be true.

As an example of a multiple linear regression model, consider the general form of the rigid body TLS model (Schomaker and Trueblood, 1968) used in Chapter 2. The TLS model is derived by approximating the displacement, $\mathbf{u}$, of an atom from its mean position, $\mathbf{r} = [x\ y\ z]^t$, as

$$\mathbf{u} = \mathbf{t} + \boldsymbol{\lambda} \times \mathbf{r} = \mathbf{t} + \mathbf{A}\boldsymbol{\lambda} \tag{2}$$

where $\mathbf{t}$ is a vector representing the translational component and $\boldsymbol{\lambda}$ is an axial vector representing the rotational component of the displacement. The rotational displacement takes place in the direction of $\boldsymbol{\lambda} \times \mathbf{r} = \mathbf{A}\boldsymbol{\lambda}$ where the elements of the matrix $\mathbf{A}$ are

$$\mathbf{A} = \begin{bmatrix} 0 & z & -y \\ -z & 0 & x \\ y & -x & 0 \end{bmatrix}.$$

According to Johnson (1970), the dispersion or covariance matrix of $\mathbf{r}$ is obtained by taking the direct products of Equation 2 and averaging each term

$$\begin{aligned} \langle \mathbf{u}\mathbf{u}^{\mathbf{t}} \rangle &= \langle (\mathbf{t} + \mathbf{A}\lambda)(\mathbf{t} + \mathbf{A}\lambda)^t \rangle \\ &= \langle \mathbf{t}\mathbf{t}^t \rangle + \langle \mathbf{t}(\mathbf{A}\lambda)^t \rangle + \langle (\mathbf{A}\lambda)\mathbf{t}^t \rangle + \langle \mathbf{A}\lambda(\mathbf{A}\lambda)^t \rangle \\ &= \langle \mathbf{t}\mathbf{t}^t \rangle + \langle \mathbf{t}\lambda^t\mathbf{A}^t \rangle + \langle \mathbf{A}\lambda\mathbf{t}^t \rangle + \langle \mathbf{A}\lambda\lambda^t\mathbf{A}^t \rangle \\ \mathbf{U} &= \mathbf{T} + \mathbf{S}^t\mathbf{A}^t + \mathbf{A}\mathbf{S} + \mathbf{A}\mathbf{L}\mathbf{A}^t \end{aligned} \tag{3}$$

where $\mathbf{U}, \mathbf{T}$, and $\mathbf{L}$ are symmetric matrices which represent the total displacement, translation and rotation, respectively, of the rigid body and $\langle \quad \rangle$ is used to represent the average. (Appendix 2D).

If Equation 3 is rearranged as $\mathbf{U} = \mathbf{T} + \mathbf{A}\mathbf{L}\mathbf{A}^t + \mathbf{A}\mathbf{S} + \mathbf{S}^t\mathbf{A}^t$, and similar terms are combined so that the TLS model for five atoms $(1-5)$ of a tetrahedron can be written in matrix form as

$$\mathbf{y} = \begin{bmatrix} U_{11}^1 \\ U_{22}^1 \\ U_{33}^1 \\ U_{12}^1 \\ U_{13}^1 \\ U_{23}^1 \\ U_{11}^2 \\ U_{22}^2 \\ U_{33}^2 \\ U_{12}^2 \\ U_{13}^2 \\ U_{23}^2 \\ U_{11}^3 \\ U_{22}^3 \\ U_{33}^3 \\ U_{12}^3 \\ U_{13}^3 \\ U_{23}^3 \\ U_{11}^4 \\ U_{22}^4 \\ U_{33}^4 \\ U_{12}^4 \\ U_{13}^4 \\ U_{23}^4 \\ U_{11}^5 \\ U_{22}^5 \\ U_{33}^5 \\ U_{12}^5 \\ U_{13}^5 \\ U_{23}^5 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & z_1^2 & y_1^2 & 0 & 0 & -2y_1z_1 & 0 & 0 & 0 & 2z_1 & 0 & 0 & -2y_1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & z_1^2 & 0 & x_1^2 & 0 & -2x_1z_1 & 0 & 0 & -2z_1 & 0 & 0 & 0 & 0 & 0 & 2x_1 \\
0 & 0 & 1 & 0 & 0 & 0 & y_1^2 & x_1^2 & 0 & -2x_1y_1 & 0 & 0 & 0 & 0 & 2y_1 & 0 & 0 & -2x_1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -x_1y_1 & -z_1^2 & y_1z_1 & x_1z_1 & -z_1 & 0 & 0 & 0 & z_1 & 0 & x_1 & -y_1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -x_1z_1 & 0 & y_1z_1 & -y_1^2 & x_1y_1 & 2y_1 & 0 & 0 & -x_1 & y_1 & z_1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -y_1z_1 & 0 & 0 & x_1z_1 & x_1y_1 & -x_1^2 & -x_1 & y_1 & -z_1 & 0 & -2x_1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & z_2^2 & y_2^2 & 0 & 0 & -2y_2z_2 & 0 & 0 & 0 & 2z_2 & 0 & 0 & -2y_2 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & z_2^2 & 0 & x_2^2 & 0 & -2x_2z_2 & 0 & 0 & -2z_2 & 0 & 0 & 0 & 0 & 0 & 2x_2 \\
0 & 0 & 1 & 0 & 0 & 0 & y_2^2 & x_2^2 & 0 & -2x_2y_2 & 0 & 0 & 0 & 0 & 2y_2 & 0 & 0 & -2x_2 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -x_2y_2 & -z_2^2 & y_2z_2 & x_2z_2 & -z_2 & 0 & 0 & 0 & z_2 & 0 & x_2 & -y_2 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -x_2z_2 & 0 & y_2z_2 & -y_2^2 & x_2y_2 & 2y_2 & 0 & 0 & -x_2 & y_2 & z_2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -y_2z_2 & 0 & 0 & x_2z_2 & x_2y_2 & -x_2^2 & -x_2 & y_2 & -z_2 & 0 & -2x_2 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & z_3^2 & y_3^2 & 0 & 0 & -2y_3z_3 & 0 & 0 & 0 & 2z_3 & 0 & 0 & -2y_3 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & z_3^2 & 0 & x_3^2 & 0 & -2x_3z_3 & 0 & 0 & -2z_3 & 0 & 0 & 0 & 0 & 0 & 2x_3 \\
0 & 0 & 1 & 0 & 0 & 0 & y_3^2 & x_3^2 & 0 & -2x_3y_3 & 0 & 0 & 0 & 0 & 2y_3 & 0 & 0 & -2x_3 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -x_3y_3 & -z_3^2 & y_3z_3 & x_3z_3 & -z_3 & 0 & 0 & 0 & z_3 & 0 & x_3 & -y_3 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -x_3z_3 & 0 & y_3z_3 & -y_3^2 & x_3y_3 & 2y_3 & 0 & 0 & -x_3 & y_3 & z_3 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -y_3z_3 & 0 & 0 & x_3z_3 & x_3y_3 & -x_3^2 & -x_3 & y_3 & -z_3 & 0 & -2x_3 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & z_4^2 & y_4^2 & 0 & 0 & -2y_4z_4 & 0 & 0 & 0 & 2z_4 & 0 & 0 & -2y_4 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & z_4^2 & 0 & x_4^2 & 0 & -2x_4z_4 & 0 & 0 & -2z_4 & 0 & 0 & 0 & 0 & 0 & 2x_4 \\
0 & 0 & 1 & 0 & 0 & 0 & y_4^2 & x_4^2 & 0 & -2x_4y_4 & 0 & 0 & 0 & 0 & 2y_4 & 0 & 0 & -2x_4 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -x_4y_4 & -z_4^2 & y_4z_4 & x_4z_4 & -z_4 & 0 & 0 & 0 & z_4 & 0 & x_4 & -y_4 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -x_4z_4 & 0 & y_4z_4 & -y_4^2 & x_4y_4 & 2y_4 & 0 & 0 & -x_4 & y_4 & z_4 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -y_4z_4 & 0 & 0 & x_4z_4 & x_4y_4 & -x_4^2 & -x_4 & y_4 & -z_4 & 0 & -2x_4 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & z_5^2 & y_5^2 & 0 & 0 & -2y_5z_5 & 0 & 0 & 0 & 2z_5 & 0 & 0 & -2y_5 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & z_5^2 & 0 & x_5^2 & 0 & -2x_5z_5 & 0 & 0 & -2z_5 & 0 & 0 & 0 & 0 & 0 & 2x_5 \\
0 & 0 & 1 & 0 & 0 & 0 & y_5^2 & x_5^2 & 0 & -2x_5y_5 & 0 & 0 & 0 & 0 & 2y_5 & 0 & 0 & -2x_5 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -x_5y_5 & -z_5^2 & y_5z_5 & x_5z_5 & -z_5 & 0 & 0 & 0 & z_5 & 0 & x_5 & -y_5 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -x_5z_5 & 0 & y_5z_5 & -y_5^2 & x_5y_5 & 2y_5 & 0 & 0 & -x_5 & y_5 & z_5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -y_5z_5 & 0 & 0 & x_5z_5 & x_5y_5 & -x_5^2 & -x_5 & y_5 & -z_5 & 0 & -2x_5 & 0 & 0 & 0
\end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} T_{11} \\ T_{22} \\ T_{33} \\ T_{12} \\ T_{13} \\ T_{23} \\ L_{11} \\ L_{22} \\ L_{33} \\ L_{12} \\ L_{13} \\ L_{23} \\ S_{11} \\ S_{12} \\ S_{13} \\ S_{21} \\ S_{22} \\ S_{23} \\ S_{31} \\ S_{32} \end{bmatrix}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{16} \\ \epsilon_{17} \\ \epsilon_{18} \\ \epsilon_{19} \\ \epsilon_{20} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{25} \\ \epsilon_{26} \\ \epsilon_{27} \\ \epsilon_{28} \\ \epsilon_{29} \\ \epsilon_{30} \end{bmatrix},$$

where $\mathbf{y}$ is composed of the estimated temperature factors determined in a structure refinement (Appendix 2A), $\mathbf{X}$ is composed of the atomic coordinates for each atom, $\boldsymbol{\beta}$ contains the unique elements of the $\mathbf{T}$, $\mathbf{L}$ and $\mathbf{S}$ matrices that are to be estimated and $\boldsymbol{\epsilon}$ is the error vector for the model. In this case of a tetrahedron, $n = 30$ observations and $p = 20$ parameters that must be estimated for the model.

For the model given in Equation 1, let $\mathbf{b}$, represent the estimate of $\boldsymbol{\beta}$ so that

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}, \tag{4}$$

where $\hat{\mathbf{y}}$ represents the model's estimated or fitted responses for all n observations.

To arrive at a value of $\mathbf{b}$, we will use the method of **least squares**. To use least-squares to find estimates of the model parameters, $\boldsymbol{\beta}$, we must **minimize** L where

$$\begin{aligned} L &= \sum_{i=1}^{n} e_i^2 = \mathbf{e}^t\mathbf{e} \\ &= \sum_{i=1}^{n} (\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}}). \end{aligned} \tag{5}$$

where $\mathbf{e}$ is the estimator of the error vector $\boldsymbol{\epsilon}$ (Myers, 1990). The value of $\boldsymbol{\epsilon}$ is known only from the knowledge of the true parameters of the model.

Making the appropriate substitutions in Equation 5 results in

$$\begin{aligned} L &= [\mathbf{y} - \mathbf{X}\mathbf{b}]^t[\mathbf{y} - \mathbf{X}\mathbf{b}] \\ &= [\mathbf{y}^t - \mathbf{b}^t\mathbf{X}^t][\mathbf{y} - \mathbf{X}\mathbf{b}] \\ &= \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}\mathbf{b} - \mathbf{b}^t\mathbf{X}^t\mathbf{y} + \mathbf{b}^t\mathbf{X}^t\mathbf{X}\mathbf{b}. \end{aligned}$$

Further, since $\mathbf{y}^t\mathbf{X}\mathbf{b} = \mathbf{b}^t\mathbf{X}^t\mathbf{y}$, then

$$L = \mathbf{b}^t\mathbf{b} - 2\mathbf{y}^t\mathbf{X}\mathbf{b} + \mathbf{b}^t\mathbf{X}^t\mathbf{X}\mathbf{b}.$$

By rearranging the above equation into

$$L = \mathbf{b}^t\mathbf{X}^t\mathbf{X}\mathbf{b} - 2\mathbf{y}^t\mathbf{X}\mathbf{b} + \mathbf{y}^t\mathbf{y},$$

which shows that L has the form of a quadratic function in terms of $\mathbf{b}$.

Because L is quadratic, contours of constant L are ellipsoids centered at $\mathbf{b}$. The minimum is found by setting the gradient of L, ∇L, equal to zero and solving

for the location of the minimum. Evaluating ∇L, term by term, at the point $\mathbf{b}$, gives

$$\frac{\partial(\mathbf{b}^t(\mathbf{X}^t\mathbf{X})\mathbf{b})}{\partial \mathbf{b}} = 2(\mathbf{X}^t\mathbf{X})\mathbf{b}$$

because $(\mathbf{X}^t\mathbf{X})$ is a symmetric matrix. For the second term,

$$\frac{\partial(-2\mathbf{y}^t\mathbf{X}\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{y}^t\mathbf{X} = -2\mathbf{X}^t\mathbf{y}$$

and finally,

$$\frac{\partial(\mathbf{y}^t\mathbf{y})}{\partial \mathbf{b}} = 0.$$

Therefore

$$\frac{\partial L}{\partial \mathbf{b}} = \nabla_{\mathbf{b}} L = 2(\mathbf{X}^t\mathbf{X})\mathbf{b} - 2\mathbf{X}^t\mathbf{y}.$$

At the minimum, $\nabla_{\mathbf{b}} L = 0$, thus

$$(\mathbf{X}^t\mathbf{X})\mathbf{b} = \mathbf{X}^t\mathbf{y}. \tag{6}$$

The above equations written in matrix form are called the **least-squares normal equations** (Myers, 1990). If $(\mathbf{X}^t\mathbf{X})$ is non-singular, then the least-squares estimate of $\boldsymbol{\beta}$, is given as

$$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}. \tag{7}$$

Equation 7 is by definition the location of the minimizer in the $\mathbf{b}$ space. The value of $\mathbf{b}$ corresponds to a minimum in L because

$$\frac{\partial^2 L}{\partial \mathbf{b}^2} = 2\mathbf{X}^t\mathbf{X} = H$$

where H is a positive definite matrix called the **Hessian matrix** (Appendix 1A). Verification that $\mathbf{b}$ represents a minimum is also found by substitution back into the equation above for the gradient

$$\nabla_{\mathbf{b}} L = 2(\mathbf{X}^t\mathbf{X})(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} - 2\mathbf{X}^t\mathbf{y} = 0.$$

These two facts: a positive definite hessian matrix and the $\nabla_{\mathbf{b}} L = 0$ indicate a minimum in L has been located at $\mathbf{b}$.

2C.2 Properties of the Least-squares Estimators

In order to measure the precision of the least-square estimates, the first assumption made is that ϵ_i, the error of an i^{th} measurement, is a **random variable**. If the errors are random, each error can be assumed to be **normally and independently distributed** (NID) with mean (μ) equal to zero and variance σ^2, i.e. $\epsilon_i \sim NID(0, \sigma^2)$ (Appendix 2A). The assumption that each error has the same variance is called the **homogeneous variance assumption** (Myers, 1990). A desired property of point estimators is that they are **unbiased estimators** of the true parameters (Milton and Arnold, 1990). Because a parameter estimator is a statistic, like the errors, it too is a random variable with a mean, or expected, value. Consequently, if the mean is equal to the estimated parameter, then the parameter estimator is said to be unbiased (Myers, 1990). To learn if the parameter estimators, $\mathbf{b}$, represent unbiased estimators of the true parameters, $\boldsymbol{\beta}$, we must evaluate the expected value of $\mathbf{b}$ according to

$$\begin{aligned} E\{\mathbf{b}\} &= E\{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}\} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E\{\mathbf{y}\} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t E\{\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}E\{\boldsymbol{\beta}\} + E\{\boldsymbol{\epsilon}\} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + E\{\boldsymbol{\epsilon}\} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

because $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $E\{\boldsymbol{\epsilon}\} = 0$ or $\mu_{\boldsymbol{\epsilon}} = 0$. This result indicates that $\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$ under the assumption $E\{\boldsymbol{\epsilon}\} = 0$.

Obtaining estimates of population parameters is only half of the battle. What is the precision of these estimates and what is the interrelationship among the estimated parameters included in the model? Consequently, the remainder of the battle lies in the ability to provide valid estimates for their variances and covariances. The variances and covariances that characterize the uncertainties and interrelationships associated with our attempt to estimate $\boldsymbol{\beta}$ can be obtained by evaluating the variance-covariance matrix of $\mathbf{b}$. By definition, the variance of a random variable Z (Milton and Arnold, 1990) is

$$\begin{aligned} var(Z) &= \sigma_Z^2 \\ &= E\{(Z - \mu_Z)^2\} \\ &= E\{(Z - E\{Z\})^2\} \end{aligned}$$

Recall that for each random variable ϵ, the assumption $\mu_\epsilon = 0$ was made so that

$$\begin{aligned} var(\epsilon) &= \sigma_\epsilon^2 \\ &= E\{(\epsilon - E\{\epsilon\})^2\} \\ &= E\{(\epsilon - \mu_\epsilon)^2\} \\ &= E\{\epsilon^2\} \end{aligned}$$

Applied to a $(n \times 1)$ vector of $\boldsymbol{\epsilon}$, each having identical variance,

$$var(\boldsymbol{\epsilon}) = E\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \sigma^2 \mathbf{I}_n$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix with off-diagonal elements equal to zero because of the independence assumption

$$E\{\epsilon_i \epsilon_j\} = 0.$$

For $\mathbf{b}$, it follows that

$$\begin{aligned} var\{\mathbf{b}\} &= E\{(\mathbf{b} - E\{\mathbf{b}\})(\mathbf{b} - E\{\mathbf{b}\})^t\} \\ &= E\{(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^t\}. \end{aligned} \tag{8}$$

By rewriting Equation 7 as

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta}+(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon} \\ \mathbf{b}-\boldsymbol{\beta} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon},\end{aligned}$$

substitution back into Equation 8 results in

$$\begin{aligned}var\{\mathbf{b}\} &= E\{[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon}][(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon}]^t\} \\ &= [(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]E\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^t \\ &= \sigma^2\mathbf{I}_n(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} \\ &= \sigma^2\mathbf{C}\end{aligned}$$

because $E\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\} = \sigma^2\mathbf{I}_n$ and where $\mathbf{C}$ is defined as $(\mathbf{X}^t\mathbf{X})^{-1}$ (Myers, 1990). The elements of $\sigma^2\mathbf{C}$ provide the variances (diagonal elements) and covariances (off-diagonal elements) of the least-squares estimates $\mathbf{b}_j$ and have the general form

$$\mathbf{V} = \begin{bmatrix} var(b_1) & cov(b_1b_2) & \cdots & cov(b_1b_p) \\ & var(b_2) & \cdots & cov(b_2b_p) \\ & & \ddots & \vdots \\ & & & var(b_p) \end{bmatrix}.$$

However, before $\mathbf{V}$ can be evaluated, an unbiased estimate of the population variance, σ^2 must be found, in terms of the observational data. Fortunately, an estimate of the population variance, s^2, can be obtained from the results provided by the least squares refinement. From Equation 5,

$$\begin{aligned}\mathbf{e} &= \mathbf{y}-\hat{\mathbf{y}} \\ &= \mathbf{y}-\mathbf{X}\mathbf{b}.\end{aligned}$$

Consider the sum of squared residuals

$$\mathbf{e}^t\mathbf{e} = \mathbf{e}^t[\mathbf{y}-\mathbf{X}\mathbf{b}]. \tag{9}$$

Using Equations 1 and 7, Equation 9 becomes

$$\begin{aligned}
\mathbf{e}^t\mathbf{e} &= \mathbf{e}^t[(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})] \\
&= \mathbf{e}^t[(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon}] \\
&= \mathbf{e}^t[\boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon}] \\
&= [\boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon}]^t[\boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon}] \\
&= [\boldsymbol{\epsilon}^t - \boldsymbol{\epsilon}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t][\boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\epsilon}] \\
&= \boldsymbol{\epsilon}^t[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t][\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]\boldsymbol{\epsilon} \\
&= \boldsymbol{\epsilon}^t[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]^2\boldsymbol{\epsilon}. \qquad (10)
\end{aligned}$$

Let's take a look at the matrices inside the brackets. The matrix

$$\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \Omega,$$

where Ω is often referred to as the **Hat matrix** (Myers, 1990) and is a projection matrix. This matrix is a symmetric, idempotent ($\Omega^2 = \Omega$) matrix that transforms the observed response ($\mathbf{y}$) to the predicted or estimated response ($\hat{\mathbf{y}}$), i.e.

$$\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\
&= \mathbf{X}((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}) \\
&= \Omega\mathbf{y}.
\end{aligned}$$

It then follows that the matrices inside the brackets of Equation 10 have the similar property, i.e.

$$\mathbf{e} = [\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]\boldsymbol{\epsilon}$$

which shows that in reality $\mathbf{e}$ are correlated with non-constant variance because $[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t] \neq \mathbf{I}_n$. From Equation 10, the expected value of the estimated error is given by

$$E\{\mathbf{e}^t\mathbf{e}\} = E\{\boldsymbol{\epsilon}^t[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]\boldsymbol{\epsilon}\}$$

We can make use of a theorem on the expected value of a quadratic form (Graybill, 1976):

$$\begin{aligned} E\{\mathbf{e}^t\mathbf{A}\mathbf{e}\} &= \sigma^2 tr(\mathbf{A}) + \mu_\epsilon^t \mathbf{A}\mu_\epsilon \\ &= \sigma^2 tr(\mathbf{A}) \qquad because \quad \mu_\epsilon \quad assumed = 0 \\ &= \sigma^2 tr[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t] \\ &= \sigma^2 tr[\mathbf{I}_n - \Omega] \\ &= \sigma^2[tr(\mathbf{I}_n) - tr(\Omega)] \\ &= \sigma^2(n-p) \end{aligned}$$

because the $tr(\Omega) = p$, the number of model parameters. Thus, an unbiased estimator for σ^2, called s^2, can be formed by defining

$$\begin{aligned} s^2 &= \frac{\mathbf{e}^t\mathbf{e}}{n-p} \\ &= \frac{(\mathbf{y}-\mathbf{X}\mathbf{b})^t(\mathbf{y}-\mathbf{X}\mathbf{b})}{n-p} \\ &= \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-p}, \end{aligned}$$

where y_i and $\hat{y}_i$ represent the observed and estimated response for the i^th observation, respectively. We know s^2 is an unbiased estimator of σ^2 because

$$\begin{aligned} E\{s^2\} &= E\{\frac{\mathbf{e}^t\mathbf{e}}{n-p}\} \\ &= \frac{\sigma^2(n-p)}{n-p} \\ &= \sigma^2. \end{aligned}$$

It is important to note that s is not an unbiased estimator for σ. The estimator, s^2, is called **the mean square error** (MSE) and is a function of the observational data.

Written in matrix form, the **estimated covariance matrix** for the estimated regression parameters, b_0 through b_p, is

$$\hat{\mathbf{V}}_r = s^2\mathbf{C} = \begin{bmatrix} \hat{\sigma}^2(b_0) & \hat{\sigma}(b_0b_1) & \cdots & \hat{\sigma}(b_0b_p) \\ & \hat{\sigma}^2(b_1) & \cdots & \hat{\sigma}(b_1b_p) \\ & & \ddots & \vdots \\ & & & \hat{\sigma}^2(b_p) \end{bmatrix}$$

where $\hat{\sigma}^2(b_p)$ is the **estimated variance** and $\sqrt{\hat{\sigma}^2(b_p)}$ is the **estimated standard error** of the parameter estimate b_p.

Alternatively stated, the i^{th} diagonal element of $\mathbf{C}$ multiplied by s^2 estimates the variance of the i^{th} parameter, i.e.

$$var(b_i) = s^2C_{ii}.$$

Similarly, the estimated covariance of b_i and b_j is

$$cov(b_ib_j) = cov(b_ib_j) = s^2C_{ij}$$

where C_{ij} is the element at the i^{th} row and the j^{th} column of $\mathbf{C} = (\mathbf{X}^t\mathbf{X})^{-1}$ (Myers, 1990).

2C.3 More Regression Measurements

In addition to the MSE discussed above, another measure of the agreement between the model and data is provided by the **coefficient of determination**, R^2. It can be shown that the total variability of the data can be partitioned as

$$SST = SSM + SSE$$

where

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where SST is the total variability of the n responses, $\mathbf{y}$, SSE is the residual sum of squares, and $SST - SSE = SSM$ is the total variability accounted for by the model represented by Equation 1 (Myers, 1990). Computationally, the coefficient of determination is given by

$$R^2 = \frac{SSM}{SST}$$
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

R^2 is always positive and ≤ 1.0 with 1.0 indicating perfect agreement the model and the data.

Another quantity which is of interest in regression analysis is the **correlation matrix**, $\boldsymbol{\rho}$, which indicates the correlation between the estimated regression parameters (Myers, 1990). The elements of the correlation matrix are computed according to

$$\rho_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}.$$

The diagonal elements of $\boldsymbol{\rho}$ are always 1.0 because an estimated parameter is perfectly correlated with itself. The values of the off-diagonal elements ρ_{ij} occur in the range $-1.0 \leq \rho_{ij} \leq 1.0$. When ρ_{ij} is positive, this indicates that large values of the parameter estimate b_i are associated with large values of the parameter estimate b_j. When ρ_{ij} is negative, this indicates that large values of the parameter estimate b_i are associated with small values of the parameter estimate b_j or *visa versa*. Note that $\boldsymbol{\rho}$ is a symmetrical matrix. The eigenvectors of the correlation matrix are a measure of the orthogonality of the parameter space chosen for the model.

2C.4 Least Squares Versus Maximum Likelihood Estimators

For purposes of discussion consider a simple linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where β_0 is the slope and β_1 is the intercept. Again, the assumption is made that ϵ_i, the error of an i^{th} measurement, is a **random variable**. If the errors are random, each error can be assumed to be, $\epsilon_i \sim NID(0, \sigma^2)$. By assuming each error is normally distributed (i.e. follows a Gaussion distribution), the **density function** (or probability function) for ϵ_i can be written as

$$f(\epsilon_i) = \frac{e^{-\frac{(\epsilon_i - \mu^0)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}.$$

For a set of n observations, there will be a corresponding set of n errors, ϵ_i where $i = 1, 2, 3, ..., n$. Application of the independence assumption for ϵ_i results in the total probability of collectively observing these n errors is given by the product of the individual probabilities. In general, the probability of collectively observing a group of independent random variables is the product of the probabilities of observing each random variable. The resulting product of probabilities is called the **joint density** and the resulting product function is called the **likelihood function**, J,

$$\begin{aligned} J(\epsilon) &= \prod_{i=1}^{n} f(\epsilon_i) \\ &= \prod_{i=1}^{n} \frac{e^{-\frac{\epsilon_i^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2}. \end{aligned}$$

It is thus desired to maximize the function, $J(\epsilon)$, to obtain the highest probability of observing the errors that were randomly observed. Because $J(\epsilon)$ is a negative exponential, this is tantamount to **minimizing** the exponent or quantity

$$\sum_{i=1}^{n} \epsilon_i^2 = L.$$

This naturally leads to the justification for using the **least squares** procedure to estimate the parameters of a function. The previous form given for L results in the likelihood function

$$J(\epsilon) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}.$$

Again finding the maximum of this function is tantamount to minimizing the exponent. Stated this way, the estimates b_0 and b_1 must be found such that they maximize $J(\epsilon)$ and thus minimize the exponent. So out of all possible values for b_0 and b_1, we are finding the best estimates that most likely produce the observed measurements. This is of course identical to the least squares criterion, namely minimization of the residual (error) sum of squares. As a result, by assuming that the errors are normally distributed, $\epsilon_i \sim NID(0, \sigma^2)$, then the **maximum likelihood estimators** of the regression coefficients, b_0 and b_1, are also the least-squares estimators.

APPENDIX 2D

Interpretation of the L Matrix from a TLS Analysis

Using the observed temperature factors and coordinates of the atoms as the independent and dependent variables, respectively, a simple linear regression analysis permits estimates of the 20 parameters of the rigid body TLS model to be determined (Schomaker and Trueblood, 1968). These parameter estimates represent the translational, librational or rotational, and screw (libration coupled with translation) motion of the molecule under the quadratic approximation for rigid body motion. The disscussion here is concerned with the six parameter estimates that represent the librational motion of the molecule. These six parameters are the components of a 3×3 real symmetric matrix $\mathbf{L}$, called the libration matrix. According to Bürgi (1984), $\mathbf{L}$ contains information physically more meaningful than the $\mathbf{T}$ matrix. If the eigenvalues are positive then we can picture $\mathbf{L}$ as representing an ellipsoid. Let $\mathbf{U}$ represent an orthogonal matrix whose columns consist of the normalized eigenvectors of $\mathbf{L}$, so that

$$\mathbf{U} = \begin{bmatrix} [\mathbf{l}_1]_C & [\mathbf{l}_2]_C & [\mathbf{l}_3]_C \end{bmatrix},$$

where C represents a Cartesian basis ($C = \mathbf{i}, \mathbf{j}, \mathbf{k}$). Because $\mathbf{U}$ represents a linear transformation of L ($L = \mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3$) onto C, in other words

$$\mathbf{U}[\mathbf{v}]_L = [\mathbf{v}]_C$$

for all $\mathbf{v}$, $\mathbf{U}$ can be viewed as a general cartesian rotation matrix so that the

diagonalization of **L**, according to

$$\begin{aligned}\mathbf{U}^t\mathbf{L}\mathbf{U} &= \begin{bmatrix} [\mathbf{l}_1]_C^t \\ [\mathbf{l}_2]_C^t \\ [\mathbf{l}_3]_C^t \end{bmatrix} \mathbf{L} \begin{bmatrix} [\mathbf{l}_1]_C & [\mathbf{l}_2]_C & [\mathbf{l}_3]_C \end{bmatrix} \\ &= \begin{bmatrix} |[\mathbf{l}_1]_L|^2 & 0 & 0 \\ 0 & |[\mathbf{l}_1]_L|^2 & 0 \\ 0 & 0 & |[\mathbf{l}_1]_L|^2 \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1^2 & 0 & 0 \\ 0 & \lambda_2^2 & 0 \\ 0 & 0 & \lambda_3^2 \end{bmatrix},\end{aligned}$$

can be visualized as the rotation of L into coincidence with C. Note that λ_1^2, λ_2^2, and λ_3^2 represent the eigenvalues of **L**.

Recall that in the TLS model (Appendix 2C), any displacement **u** can be written as the sum of translational displacement **t** and rotational displacement in the direction of $\lambda \times \mathbf{r}$ according to

$$\mathbf{u} = \mathbf{t} + \lambda \times \mathbf{r} = \mathbf{t} + \mathbf{A}\boldsymbol{\lambda}.$$

In the development of the quadratic approximation for rigid body motion, the vector $\boldsymbol{\lambda} = \lambda_1\mathbf{l}_1 + \lambda_2\mathbf{l}_2 + \lambda_3\mathbf{l}_3$ is introduced as an **axial vector** (Johnson, 1970). Given a rotation, θ, with rotation axis parallel to the unit vector, $[\boldsymbol{l}]_L$, an axial vector can be constructed according to $[\boldsymbol{\lambda}]_L = \theta[\boldsymbol{l}]_L$. It then follows that the magnitude of $\boldsymbol{\lambda}$ is

$$\begin{aligned}([\boldsymbol{\lambda}]_L^t \cdot [\boldsymbol{\lambda}]_L)^{1/2} &= (\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{1/2} \\ &= (\theta[\boldsymbol{l}]_L \cdot \theta[\boldsymbol{l}]_L)^{1/2} \\ &= (\theta^2(l_1^2 + l_2^2 + l_3^2))^{1/2} \\ &= \theta,\end{aligned}$$

the angle of rotation (in radians) and the rotation axis $[\boldsymbol{l}]_L$ is given by

$$\begin{aligned}[\boldsymbol{l}]_L &= \frac{1}{\theta}[\boldsymbol{\lambda}]_L \\ &= \frac{1}{\theta}(\lambda_1 \mathbf{l}_1 + \lambda_2 \mathbf{l}_2 + \lambda_3 \mathbf{l}_3).\end{aligned}$$

Note that to obtain the rotation axis in cartesian space we must apply the transformation

$$det(\mathbf{U})\mathbf{U}[\boldsymbol{l}]_L = [\boldsymbol{l}]_C,$$

where we must include the $det(\mathbf{U})$ to account for improper rotations because $[\boldsymbol{l}]_L$ is axial (pseudo-vector) and will change sign for improper rotations.

When the components of $\boldsymbol{\lambda}$ $(\lambda_1, \lambda_2, \lambda_3)$ approach zero they are called infinitesimal rotation coordinates (in radians) and correspond to three independent rotations of α_1, α_2, and α_3 where the general cartesian rotation matrix is constructed for each rotation whereby

$\mathbf{M}_C(\alpha_1)$ represents a rotation of λ_1 about the vector $[\mathbf{l}_1]_C$

$\mathbf{M}_C(\alpha_2)$ represents a rotation of λ_2 about the vector $[\mathbf{l}_2]_C$

$\mathbf{M}_C(\alpha_3)$ represents a rotation of λ_3 about the vector $[\mathbf{l}_3]_C$

where $\mathbf{l}_1, \mathbf{l}_2$, and $\mathbf{l}_3$ are the basis vectors of L represented by the columns of $\mathbf{U}$. They are required to be infinitesimal coordinates for this description because successive rotations about different axes do not commute unless the rotations are infinitesimally small (Johnson, 1970). For the purposes of example let

$$\mathbf{M}_C(\gamma) = \mathbf{M}_C(\alpha_3) * \mathbf{M}_C(\alpha_1) * \mathbf{M}_C(\alpha_2).$$

An analysis of the rotation matrix, $\mathbf{M}_C(\gamma)$, in order to find its rotation angle, ρ, and unit vector parallel to the rotation axis, $[\mathbf{k}]_C$, results in

$$\rho = \hat{\theta}$$

and

$$[\mathbf{k}]_C = [\hat{\boldsymbol{l}}]_C,$$

where $\hat{}$ represents an estimate. Slightly different estimates are obtained if the order of composition is changed. The smaller the components of $\boldsymbol{\lambda}$, the more precise the estimates. Dunitz et al. (1988) suggests that this approximation holds for $\theta \leq 14°$ whereas Trueblood (1978) suggests that $\theta \leq 9°$ is the upper limit.

For a numerical example, consider a rigid body TLS fit to the $\mathrm{Si1}_m$ tetrahedra from the low albite refinement by Armbruster et al. (1990). The following are the results:

$$\mathbf{T} = \begin{bmatrix} .00597 & .00119 & .00020 \\ .00119 & .00691 & -.00079 \\ .00020 & -.00079 & .00558 \end{bmatrix}$$

$$\mathbf{L} = \begin{bmatrix} .00341 & .00008 & .00090 \\ .00008 & .00539 & -.00026 \\ .00090 & -.00026 & .00309 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} -.00018 & .00005 & -.00100 \\ .00005 & .00007 & .00029 \\ -.00100 & .00029 & .00011 \end{bmatrix}.$$

For the $\mathbf{L}$ matrix given above

$$\mathbf{U} = \begin{bmatrix} -.63753 & .77027 & -.01543 \\ .08183 & .08761 & .99279 \\ .76606 & .63167 & -.11889 \end{bmatrix}$$

$$\lambda_1^2 = .00231$$

$$\lambda_2^2 = .00416$$

$$\lambda_3^2 = .00542$$

It then follows that the libration angle θ is given by

$$\begin{aligned}\theta &= (trace(\mathbf{L}))^{1/2} \\ &= (.00341 + .00539 + .00309)^{1/2} \\ &= (\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{1/2} \\ &= (.00231 + .00416 + .00542)^{1/2} \\ &= (.109) * 180°/\pi \\ &= 6.247°\end{aligned}$$

and libration axis by

$$\begin{aligned}[\boldsymbol{l}]_C &= \begin{bmatrix} -.63753 & .77027 & -.01543 \\ .08183 & .08761 & .99279 \\ .76606 & .63167 & -.11889 \end{bmatrix} \begin{bmatrix} (.00231)^{1/2}/.109 \\ (.00416)^{1/2}/.109 \\ (.00542)^{1/2}/.109 \end{bmatrix} \\ &= \begin{bmatrix} .16404 \\ .75808 \\ .63119 \end{bmatrix}.\end{aligned}$$

Consider the composition of the three independent rotations

$$\mathbf{M}_C(\alpha_1) = \begin{bmatrix} .999314 & -.036879 & .003369 \\ .036758 & .998852 & .030714 \\ -.004497 & -.030569 & .999523 \end{bmatrix}$$

$$\mathbf{M}_C(\alpha_2) = \begin{bmatrix} .999155 & -.040568 & .006658 \\ .040849 & .997937 & -.049525 \\ -.004635 & .049755 & .998751 \end{bmatrix}$$

$$\mathbf{M}_C(\alpha_2) = \begin{bmatrix} .997293 & .008701 & .073009 \\ -.008784 & .999961 & .000815 \\ -.072999 & -.001454 & .997331 \end{bmatrix}$$

so that

$$\begin{aligned}\mathbf{M}_C(\gamma) &= \mathbf{M}_C(\alpha_1) * \mathbf{M}_C(\alpha_2) * \mathbf{M}_C(\alpha_3) \\ &= \begin{bmatrix} .994062 & -.068516 & .084535 \\ .069775 & .997490 & -.012037 \\ -.083498 & .017864 & .996348 \end{bmatrix}.\end{aligned}$$

An analysis of $\mathbf{M}_C(\gamma)$ indicates that

$$\begin{aligned}\rho &= \hat{\theta} \\ &= 6.306°\end{aligned}$$

and

$$[\mathbf{k}]_C = [\hat{\boldsymbol{\ell}}]_C = \begin{bmatrix} .13612 \\ .76495 \\ .62955 \end{bmatrix}$$

By composing all six possible permutations of α_1, α_2 and α_3 to form six $\mathbf{M}_C(\gamma)$ and computing the average $\langle\rho\rangle$ and $\langle[\mathbf{k}]_C\rangle$, along with its standard error (given in parenthesis) results in

$$\langle\rho\rangle = 6.246°(.060)$$

and

$$\langle[\mathbf{k}]_C\rangle = \begin{bmatrix} .16404(.023) \\ .75782(.013) \\ .63083(.017) \end{bmatrix}.$$

which is a very good estimate of

$$\theta = 6.247°$$

and

$$[\boldsymbol{\ell}]_C = \begin{bmatrix} .16404 \\ .75808 \\ .63119 \end{bmatrix}$$

determined directly from $\mathbf{L}$.

References

Abramowitz, M. and Stegun, I.A. (1972) Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York.

Armbruster, T. (1986) Crystal structure refinement and thermal expansion of a Li, Na, Be-cordierite between 100 and 550 K. Zeitschrift für Kristallographie, 174, 205-217.

Armbruster, T. and Oberhänsli, R. (1988) Crystal chemistry of double-ring silicates: Structures of sugilite and brannockite. American Mineralogist, 73, 595-600.

Armbruster, T. and Lager, G.A. (1989) Oxygen disorder and the hydrogen position in garnet-hydrogarnet solid solutions. European Journal of Mineralogy, 1, 363-369.

Armbruster, T., Bürgi, H.B., Kunz, M., Gnos, E., Brönnimann, S. and Lienert, C. (1990) Variation of displacement parameters in structure refinements of low albite. American Mineralogist, 75, 135-140.

Artioli, G., Smith, J.V. and Kvick, Å. (1984) Neutron diffraction study of Natrolite, $Na_2Al_2\ Si_3O_{10}\cdot\ 2H_2O$, at 20 K. Acta Crystallographica, C40, 1658-1662.

Artioli, G., Smith, J.V. and Pluth, J.J. (1986) X-ray structure refinement of mesolite. Acta Crystallographica, C42, 937-942.

Basso, R., Lucchetti, G. and Palenzona, A. (1989) Crystallographic and crystal chemical study on a natural C2/c ordered Na-Mn-clinopyroxene from Val di Vara (Northern Apennines, Italy). Neues Jahrbuch für Mineralogie Monatshefte, 1989, 59-68.

Baur, W.H. and Ohta T. (1982) The Si_5O_{16} pentamer in zunyite refined and empirical relations for individual silicon-oxygen bonds. Acta Crystallographica, B38, 390-401.

Bissert, G. and Liebau, F. (1986) The crystal structure of a triclinic bikitaite, $Li[AlSi_2O_6]\cdot H_2O$, with ordered Al/Si distribution. Neues Jahrbuch für Mineralogie Monatshefte, 1986, 241-252.

Blasi, A., Brajkovic, A., De Pol Blasi, C., Foord, E.E., Martin, R.F. and Zanazzi, P.F. (1984) Structure refinement and genetic aspects of a microcline overgrowth on amazonite from Pikes Peak batholith, Colorado, U.S.A.. Bulletin de Minéralogie, 107, 411-422.

Blasi, A., De Pol Blasi, C. and Zanazzi, P.F. (1987) A re-examination of the pellotsalo microcline: mineralogical implications and genetic considerations. Canadian Mineralogist, 25, 527-537.

Boisen, M.B., Jr. and Gibbs, G.V. (1985) Mathematical Crystallography, Volume 16, Reviews in Mineralogy. Mineralogical Society of America, Washington D.C..

Boisen, M.B., Jr., Gibbs, G.V., Downs, R.T. and D'Arco, P. (1990) The dependence of the SiO bond length on structural parameters in coesite, the silica polymorphs and the clathrasils. American Mineralogist, 75, 748-754.

Boström, D. (1987) Single-crystal X-ray diffraction studies of synthetic Ni-Mg olivine solid solutions. American Mineralogist, 72, 965-972.

Boysen, H., Dorner, B., Frey, F. and Grimm, H. (1980) Dynamic structure determination for two interacting modes at the M-point in α- and β-quartz by inelastic neutron scattering. The Journal of Physical Chemistry, 13, 6127-6146.

Burns, D.M., Ferrier, W.G. and McMullan, J.T. (1967) The Rigid-Body Vibrations of Molecules in Crystals. Acta Crystallographica, 22, 623-629.

Bürgi, H.B. (1984) Stereochemical lability in crystalline coordination compounds. Transactions of the American Crystallographic Association, 20, 61-71.

Cameron, M., Sueno, S., Prewitt, C.T. and Papike, J.J. (1973) High-temperature crystal chemistry of Acmite, Diopside, Hedenbergite, Jadeite, Spodumene, and Ureyite. American Mineralogist, 58, 594-618.

Cameron, M., Sueno, S., Papike, J.J. and Prewitt, C.T. (1983) High temperature crystal chemistry of K and Na fluor-richterites. American Mineralogist, 68, 924-943.

Cannillo E. and Coda, E. (1966) The crystal structure of bavenite. Acta Crystallographica, 20, 301-309.

Chandrasekhar, K and Bürgi, H.B. (1984) Dynamic Processes in Crystals Examined Through Difference Vibrational Parameters ΔU: The Low-Spin–High-Spin Transition in
Tris(dithiocarbamato)iron(III) Complexes. Acta Crystallographica, B40, 387-397.

Clark, J.R., Appleman, D.E. and Papike (1969) Crystal-chemical characterization of clinopyroxenes based on eight new structure refinements. Mineralogical Society of America Special Paper, 2, 31-50.

Cohen, J.P., Ross, F.K. and Gibbs, G.V. (1977) An x-ray and neutron diffraction study of hydrous low cordierite. American Mineralogist, 62, 67-78.

Destro, R., Pilati, T. and Massimo, S. (1977) The Structure and Electron Density of *sym*-Dibenzo-1,5-cyclooctadiene-3,7-diyne by X-ray Analysis of Three Different Temperatures. Acta Crystallographica, B33, 447-456.

Downs, R.T., Gibbs, G.V. and Boisen, M.B.Jr. (1990) A study of the mean-square displacement amplitudes of Si, Al and O atoms in framework structures: Evidence for rigid bonds, order, twinning and stacking faults. American Mineralogist, 75, 1253-1267.

Downs, R.T., Gibbs, G.V., Bartelmehs, K.L. and Boisen, M.B.Jr. (1992) Variations of bond lengths and volumes of silicate tetrahedra with temperature.

American Mineralogist, 77, 751-757.

Dunitz, J.D., Schomaker, V. and Trueblood, K.N. (1988) Interpretation of atomic displacement parameters from diffraction studies of crystals. The Journal of Physical Chemistry, 92, 856-867.

Finger, L.W. and Hazen, R.M. (1987) Crystal structure of monoclinic ilvaite and the nature of the monoclinic-orthorhombic transition at high pressure. Zeitschrift für Kristallographie, 179, 415-430.

Francis, C.A. and Ribbe, P.H. (1980) The forsterite-tephroite series: I. Crystal structure refinements. American Mineralogist, 65, 1263-1269.

Fujino, K. and Takéuchi Y. (1978) Crystal chemistry of titanian chondrodite and titanian clinohumite of high-pressure origin. American Mineralogist, 63, 535-543.

Gabe, Protheine and Whitlow (1973) A reinvestigation of the epidote structure: Confirmation of the iron location. American Mineralogist, 58, 218-223.

Geisinger, K.L., Spackman, M.A. and Gibbs, G.V. (1987) Exploration of structure, electron density distribution and bonding in coesite with Fourier and pseudoatom refinement methods using single-crystal x-ray diffraction data. The Journal of Physical Chemistry, 91, 3237-3244.

Ghose, S. and Wan, C. (1976) Structural chemistry of borosilicates, part II: Searlesite,
$NaBSi_2O_5(OH)$: Absolute configuration, hydrogen locations, and refinement of the structure. American Mineralogist, 61, 123-129.

Ghose, S., Schomaker, V. and McMullan R.K. (1986) Enstatite, $Mg_2Si_2O_6$: A neutron diffraction refinement of the crystal structure and a rigid-body analysis of the thermal vibration. Zeitschrift für Kristallographie, 176, 159-175.

Gibbs, G.V., Prewitt, C.T. and Baldwin, K.J. (1977) A study of the structural chemistry of coesite. Zeitschrift für Kristallographie, 145, 108-123.

Gibbs, G.V. (1982) Molecules as models for bonding in silicates. American Mineralogist, 67, 421-450.

Gies, H. (1983) Studies on clathrasils. III. Crystal structure of melanophlogite, a natural clathrate compound of silica. Zeitschrift für Kristallographie, 164, 247-257.

Graybill, F.A. (1976) Theory and Application of the Linear Model. Boston, Duxbury.

Grice, J.D. and Robinson, G.W. (1989) Feruvite, a new member of the tourmaline group, and its crystal structure. Canadian Mineralogist, 27, 199-203.

Grimm, H. and Dorner, B. (1975) On the mechanism of the $\alpha - \beta$ phase transformation of quartz. The Journal of the Physical Chemistry of Solids, 36, 407-413.

Harlow, G.E. and Brown, G.E., Jr. (1980) Low albite: an x-ray and neutron diffraction study. American Mineralogist, 65, 986-995.

Hassan, I. and Grundy, H.D. (1985) The crystal structures of helvite group minerals, $(Mn,Fe,Zn)_8(Be_6Si_6O_{24})S_2$. American Mineralogist, 70, 186-192.

Hazen, R.M. and Finger, L.W. (1979) Crystal structure and compressibility of zircon at high pressure. American Mineralogist, 64, 196-201.

Hirshfeld, F.L. (1976) Can x-ray data distinguish bonding effects from vibrational smearing?. Acta Crystallographica, A32, 239-244.

Hummel, W., Raselli, A. and Bürgi, H.B. (1990) Analysis of Atomic Displacement Parameters and Molecular Motion in Crystals. Acta Crystallographica, B46, 683-692.

Johnson, C.K. (1970) The effect of thermal motion on interatomic distances and angles. In Crystallographic Computing, editor F.R. Ahmed. Munksgaard, Copenhagen.

Joswig, W., Bartl, H. and Fuess, H. (1984) Structure refinement of scolecite by neutron diffraction. Zeitschrift für Kristallographie, 166, 219-223.

Kalus, C. (1978) Neue strukturbestimmung des anorthits unter berocksichtigung möglicher alternativen. Dissertation, Ludwig-Maximilians-universität zu München.

Kihara, K. (1990) An X-ray study of the temperature dependence of the quartz structure. European Journal of Mineralogy, 2, 63-77.

Kimata, M. (1989) The crystal structure of mangonaan kilochoanite, $Ca_{2.33}Mn_{0.67}Si_2O_7$: a site-preference rule for the substitution of Mn for Ca. The Mineralogical Magazine, 53, 625-631.

Kirfel, A. and Will, G. (1984) Ending the "$P2_1/a$ coesite" discussion. Zeitschrift für Kristallographie, 167, 287-291.

Kunz, M. and Armbruster, T. (1988) Static positional disorder studied by difference vibrational parameters: Na, K-feldspars with variable degree of Si/Al ordering. Zeitschrift für Kristallographie, 182, 166-168.

Kunz, M. and Armbruster, T. (1990) Difference displacement parameters in alkali feldspars: Effect of (Si,Al) order-disorder. American Mineralogist, 75, 141-149.

Kvick, Å. and Smith, J.V. (1983) A neutron diffraction study of the zeolite edingtonite. The Journal of Chemical Physics, 79, 2356-2362.

Kvick, Å., Ståhl, K. and Smith, J.V. (1985) A neutron diffraction study of the bonding of zeolitic water in scolecite at 20 K. Zeitschrift für Kristallographie, 171, 141-154.

Lee, J.H. and Guggenheim, S. (1981) Single crystal X-ray refinement of pyrophyl-

lite-1Tc. American Mineralogist, 66, 350-357.

Levien, L., Prewitt, C.T. and Weidner, D.J. (1980) Structure and elastic properties of quartz at pressure. American Mineralogist, 65, 920-930.

Levien, L. and Prewitt, C.T. (1981) High-pressure crystal structure and compressibility of coesite. American Mineralogist, 66, 324-333.

Le Page, Y. and Donnay G. (1976) Refinement of the crystal structure of low-quartz. Acta Crystallographica, B32, 2456-2459.

Liebau, F. and Böhm, H. (1982) On the Co-existence of Structurally Different Regions in the Low-High-Quartz and Other Dilative Phase Transformations. Acta Crystallographica, A38, 252-256.

Liebau, F. (1985) Structural chemistry of silicates: structure, bonding and classification. Springer-Verlag, Berlin.

Megaw, H.D. (1973) Crystal structures: A working approach. W.B. Saunders Co., Philadelphia, Pa.

Milton, J.S. and Arnold, J.C. (1990) Introduction to Probability and Statistics. New York, McGraw - Hill.

Miyake, M., Nakamura, H., Kojima, H. and Marumo, F. (1987) Cation ordering in Co-Mg olivine solid-solution series. American Mineralogist, 72, 594-598.

Moore, P.B. and Araki, T. (1976) Braunite: its structure and relationship to bixbyite, and some insights on the genealogy of fluorite derivative structures. American Mineralogist, 61, 1226-1240.

Myers, Raymond H. (1990) Classical and Modern Regression with Applications, Second Edition. Boston, PWS-Kent.

Nover, G. and Will, G. (1981) Structure refinements of seven natural olivine crystals and the influence of the oxygen partial pressure on the cation distribution. Zeitschrift für Kristallographie, 155, 27-45.

Ohasi, Fujita and Osawa (1981) Structure of $Co_3Al_2Si_3O_{12}$ garnet. Journal of the Japanese Association of Mineralogists, Petrologists, and Economic Geologists, 76, 58-60.

Peacor, D.R. (1973) High-temperature single-crystal study of the cristobalite inversion. Zeitschrift für Kristallographie, 138, 274-298.

Perdikatsis, B. and Burzlaff, H. (1981) Strukturverfeinerung am Talk $Mg_3[(OH)_2 Si_4O_{10}]$. Zeitschrift für Kristallographie, 156, 177-186.

Phillips, M. (1990) A Low Temperature Refinement of Maximum Microcline. Personal Communication to Bob Downs.

Pluth, J.J., Smith, J.V. and Kvick, Å. (1985) Neutron diffraction study of the zeolite thomsonite. Zeolites, 5, 74-80.

Rao, K.R., Chaplot, S.L., Choudhury, N., Ghose, S., Hasings, J.M., Corliss, L.M. and Price, D.L. (1988) Lattice Dynamics and Inelastic Neutron Scattering from Forsterite, Mg_2SiO_4: Phonon Dispersion Relation, Density of States and Specific Heat. Physics and Chemistry of Minerals, 16, 83-97.

Sandomirskii, P.A., Simonov, M.A. and Belov, N.V. (1977) Crystal structure of synthetic Mn-milarite $K_2Mn_5(Si_{12}O_{30})\cdot H_2O$. Soviet Physics Doklady, 22, 181-183.

Schomaker, V. and Trueblood, K.N. (1968) On the rigid-body motion of molecules in crystals. Acta Crystallographica, B24, 63-76.

Smith, J.V., Artioli, G. and Kvick, Å. (1986) Low albite, $NaAlSi_3O_8$: Neutron diffraction study of crystal structure at 13 K. American Mineralogist, 71, 727-733.

Smith, J.V., Pluth, J.J., Richardson, J.W. and Kvick A. (1987) Neutron diffraction study of zoisite at 15K and X-ray study at room temperature. Zeitschrift für Kristallographie, 179, 305-321.

Smyth, J.R., Smith, J.V., Artioli, G. and Kvick, Å. (1987) Crystal structure of coesite, a high-pressure form of SiO_2, at 15 and 298K from single-crystal neutron and x-ray diffraction data: test of bonding models. The Journal of Physical Chemistry, 91, 988-992.

Stixrude, L. and Bukowinski, M.S.T. (1988) Simple Covalent Potential Models of Tetrahedral SiO_2: Applications to α-quartz and Coesite at Pressure. Physics and Chemistry of Minerals, 16, 199-206.

Swanson, D.K. and Prewitt, C.T. (1983) The crystal structure of $K_2Si^{VI}Si_3^{IV}O_9$. American Mineralogist, 68, 581-585.

Takéuchi, Y., Haga, N. and Bunno, M. (1983) X-ray study on polymorphism of ilvaite,
$HCaFe_2^{2+}Fe^{3+}O_2[Si_2O_7]$. Zeitschrift für Kristallographie, 163, 267-283.

Trueblood, K.N. (1978) Analysis of Molecular Motion with Allowance for Intramolecular Torsion. Acta Crystallographica, A34, 950-954.

Wan, C., Ghose, S. and Gibbs, G.V. (1977) Rosenhahnite, $Ca_3Si_3O_8(OH)_2$: crystal structure and the stereochemical configuration of the hydroxylated trisilicate group, $[Si_3O_8(OH)_2]$. American Mineralogist, 62, 503-512.

Wilks, S.S. (1962) Mathematical Statistics. New York, Wiley.

Winter, J.K. and Ghose, S. (1979) Thermal expansion and high-temperature crystal chemistry of the Al_2SiO_5 polymorphs. American Mineralogist, 64, 573-586.

Wright, A.F. and Lehmann, M.S. (1981) The structure of quartz at 25 and 590° C determined by neutron diffraction. The Journal of Solid State Chemistry, 36, 371-380.

Young, R.A. and Post, B. (1962) Electron density and thermal effects in alpha

quartz. Acta Crystallographica, 15, 337-346.

CHAPTER 3

Excalibr II:

A Computer Program for Determining the Orientation of a Crystal's Optical Indicatrix

Introduction

Optical properties are a fundamental physical property of a mineral. Since its development in 1959, the spindle stage method for determining the refractive indices of crystals has led to a significant improvement in the measurement of optical data (Wilcox, 1959; Bloss, 1978; Bloss, 1981). Basically, the method involves first mounting a crystal on the needle tip of a spindle stage. Next, the spindle stage is mounted on the stage of a polarizing microscope where the crystal is submerged in an immersion oil. Using polarized monochromatic light, the crystal is then systematically rotated around the spindle axis and around the microscope axis until the crystal becomes extinct. These two degrees of rotation plus an analysis of a wide range of extinction positions permit the determination of the orientation of the optical indicatrix (Bloss and Riess, 1973; Bloss, 1981, p. 63).

The program EXCALIBR was developed to computationally analyze these extinction data and determine the coordinates of a biaxial optical indicatrix (Bloss and Riess, 1973; Bloss, 1981, p. 202). These coordinates, in turn, indicate the precise orientation at which to measure the crystal's principle refractive indices α, β, and γ. The combination of spindle stage methods and EXCALIBR have been used to solve and discover numerous problems in optical mineralogy (Armbruster and Bloss, 1982; Gunter and Bloss, 1982; Su et al. 1984, Greiner and Bloss, 1987). However, the computational procedure used by EXCALIBR placed various

restrictions on both extinction data accuracy and crystal orientation (Bloss and Riess, 1973; Bloss, 1981, p. 217). In addition, there were problems with the estimated standard deviations of the coordinates and the statistical methods used for dispersion analysis (Bloss, 1981, p. 308).

Consequently, a new version of EXCALIBR has been written. Like EXCALIBR, the new program temporarily called EXCALIBR II (Bartelmehs et al., 1992), but eventually to assume the name of its predecessor, uses the equation

$$(\mathbf{q} \cdot \mathbf{a}_1)(\mathbf{q} \cdot \mathbf{a}_2) - (\mathbf{p} \cdot \mathbf{a}_1)(\mathbf{p} \cdot \mathbf{a}_2) = 0 \qquad (1)$$

introduced by Joel (Joel, 1965) in determining estimates ($\hat{}$) for the two normalized optic axes, $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$. This algorithm, as used by EXCALIBR, solves for six variables subject to the constraint that the two optic axis are normalized (Bloss and Riess, 1973). Using six variables leads to a ill-conditioned problem that requires accurate starting estimates. This was accomplished by organizing the data into groups of four whose spindle axis settings (S values) differed by at least 40°. For each group, EXCALIBR determines estimates for $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$ and uses their average as starting values for the least-squares problem (Bloss and Riess, 1973). The new algorithm (described below) used by EXCALIBR II only involves four variables which has significantly improved the conditioning of problem. As a consequence, the new program eliminates the need to organize the data into groups. This constitutes a major advantage. Now, even if optical extinctions can only be measured over a limited range of spindle settings, the optic axial angle 2V and the orientation of the indicatrix can be determined. In addition, the number of lines of code involved are about two thirds the original number and the run-time is reduced. EXCALIBR's computation time, operating at 10 megahertz for the Tiburon albite data (presented as an example), was approximately 1 minute,

37 seconds; EXCALIBR II's was 10 seconds. EXCALIBR II thus provides an approximate 90% reduction in computation time. The physical storage space required for EXCALIBR II has also been reduced by 40% to approximately 1K. In addition, data input has been significantly simplified over that of EXCALIBR primarily because of free format input.

Program Procedure

An example input data file, for the Tiburon albite discussed in the paper, is shown in Figure 3–1 to illustrate the simplicity of the input data structure. In general, the input file contains the spindle setting, S, followed by the microscope stage settings, M_s, that provided crystal extinction. Note that input of the reference azimuth, M_r, is now optional. M_r is the microscope stage setting that aligns the spindle axis precisely east-west (Bloss, 1981, p. 19).

Provided no M_r value was input, EXCALIBR II first calculates the average reference azimuth, $\overline{M}_r$, from all supplied M_s data whose S-values differ by 180°. For example, if extinction data are supplied for S values from 0°, 10°, ..., 350°, as in the so-called 360° option of EXCALIBR, EXCALIBR II calculates M_r for each of the 18 pairs of data. It then uses $\overline{M}_r$, the average of these 18 values of M_r in its calculations.

Next, the program uses the input M_r or the computed $\overline{M}_r$ to calculate, for all S settings, a corrected extinction angle E_s, where

$$E_s = M_s - \overline{M}_r.$$

Two equivibration directions **p** and **q** are then calculated by adding and subtracting 45° from each E_s angle, respectively. The angular coordinates for **p** and **q** are then converted to a Cartesian coordinate system ($C = \{\mathbf{i},\mathbf{j},\mathbf{k}\}$) defined relative to the microscope where $[\mathbf{p}]_C^t = [p_1\ p_2\ p_3]$ and $[\mathbf{q}]_C^t = [q_1\ q_2\ q_3]$. The x-axis is

```
* * * top of file * * *
Tiburon Albite - Wolfe
433 500 600 666/
/
0     217.9 218   218.1 218.2
10    218.5 218.3 218.3 218.6
20    219.1 219.3 219.2 219.5
30    219.6 219.7 219.5 219.3
40    219.4 219.6 219.3 219.2
50    218.6 218.6 218.9 219
60    217.3 217.3 216.8 216.9
70    213.6 213.6 213.8 213.7
80    205.9 205.9 206.4 206.4
90    189.9 189.9 191.3 191.6
100   171.7 171.7 172.8 172.5
110   159.2 159.2 159.5 159.8
120   153   153   153.6 154
130   149.8 149.8 149.7 150.4
140   147.1 147.4 147.7 147.8
150   145.8 145.7 145.9 146.1
160   144.4 144.6 144.6 144.5
170   143.5 143.4 143.4 143.5
180   142.4 142.3 142.2 142.1
* * * end of file * * *
```

Figure 3–1. Example input data file showing extinction data for the Tiburon Albite measured at wavelengths of 433, 500, 600, and 666 nm (Bloss, 1981, p. 210). A line by line description of the input data file is provided with the program.

defined to be east-west, the y-axis north-south, and the z-axis is defined to be perpendicular to the microscope stage.

After rewriting Joel's equation (Equation 1) in terms of four variables, a, b, c, and d,

$$[(q_1q_2 - p_1p_2) + (q_2^2 - p_2^2)c + (q_2q_3 - p_2p_3)d]a +$$

$$[(q_1q_3 - p_1p_3) + (q_3^2 - p_3^2)d + (q_2q_3 - p_2p_3)c]b +$$

$$[(q_1q_2 - p_1p_2)]c + [(q_1q_3 - p_1p_3)]d = -(q_1^2 - p_1^2),$$

the program employs the Gauss-Newton algorithm (GN) to obtain unbiased, minimum variance estimates: $\hat{a} = \hat{s}/\hat{r}$, $\hat{b} = \hat{t}/\hat{r}$, $\hat{c} = \hat{v}/\hat{u}$, and $\hat{d} = \hat{w}/\hat{u}$ (Appendix 3A).

The program is considered to converge when the parameter shifts, $\Delta\hat{a}, \Delta\hat{b}, \Delta\hat{c}$, and $\Delta\hat{d}$, are less than 1.0×10^{-14}. Estimated final coordinates for $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$, where $[\hat{\mathbf{a}}_1]_C^t = [\hat{r}\ \hat{s}\ \hat{t}]$ and $[\hat{\mathbf{a}}_2]_C^t = [\hat{u}\ \hat{v}\ \hat{w}]$, are obtained from the normalization condition. Given the optic axes, computation of estimates for the acute and obtuse bisectrix and the optic normal vectors is obtained by adding, subtracting, and forming the cross product, respectively.

In the case where the program is unable to converge to a set of parameter estimates after 100 iterations, it is likely that one of the optic axes is located in the yz plane, in other words, 90° from the spindle axis. This will cause the estimates of the regression parameters to increase to infinity since either $\hat{r}$ or $\hat{u}$ would equal zero. To obviate this difficulty, all **p** and **q** data are rotated 120° about the vector $[1\ 1\ 1]_C^t$. If convergence is again not achieved after 100 iterations (implying that $\hat{s}$ or $\hat{v}$ equal zero), then the optic axis must coincide with the z axis, that is, it is located in xz *and* yz planes. This requires all **p** and **q** data to be rotated another 120° about the vector $[1\ 1\ 1]_C^t$. Once convergence has been reached, the estimated optic axes are then transformed back to the original orientation.

Estimates for a single optic axis, ϵ, (uniaxial crystals) where $[\hat{\epsilon}]_C^t = [\hat{e}\ \hat{f}\ \hat{g}]$, are determined by rewriting Joel's Equation (Equation 1) in the following form

$$[2(q_1q_2 - p_1p_2) + (q_2^2 - p_2^2)a + (q_2q_3 - p_2p_3)b]a +$$
$$[2(q_1q_3 - p_1p_3) + (q_3^2 - p_3^2)b + (q_2q_3 - p_2p_3)a]b = -(q_1^2 - p_1^2).$$

Minimum variance estimates of a and b using the GN method (Seber and Wild, 1989) are computed where $\hat{a} = \hat{f}/\hat{e}$, $\hat{b} = \hat{g}/\hat{e}$. The program is considered to converge when the parameter shifts, $\Delta\hat{a}$ and $\Delta\hat{b}$ are less than 1.0×10^{-14}. Final coordinates for $\hat{\epsilon}$ are again obtained from the normalization condition.

An estimate of the covariance matrix, V_s, for the optic axes is obtained from the propagation of error equation (Kendal and Stuart, 1987)

$$\hat{V}_s = L^t \hat{V}_m L \tag{2}$$

where $\hat{V}_m$ is the estimated covariance matrix of the four (or two) estimated regression parameters (Milton and Arnold, 1990) and L^t is the transpose of the matrix L:

$$L = \begin{bmatrix} \partial r/\partial a & \partial s/\partial a & \partial t/\partial a & \partial u/\partial a & \partial v/\partial a & \partial w/\partial a \\ \partial r/\partial b & \partial s/\partial b & \partial t/\partial b & \partial u/\partial b & \partial v/\partial b & \partial w/\partial b \\ \partial r/\partial c & \partial s/\partial c & \partial t/\partial c & \partial u/\partial c & \partial v/\partial c & \partial w/\partial c \\ \partial r/\partial d & \partial s/\partial d & \partial t/\partial d & \partial u/\partial d & \partial v/\partial d & \partial w/\partial d \end{bmatrix}.$$

The estimated standard errors, given in the output for $\hat{r}$, $\hat{s}$, $\hat{t}$, and $\hat{u}$, $\hat{v}$, $\hat{w}$ are the square roots of the diagonal elements of $\hat{V}_s$, respectively. Similarly, estimated standard errors of other quantities are obtained by propagation of error from $\hat{V}_s$.

Bloss (1981, p. 210) presents EXCALIBR's solutions of extinction data determined at wavelengths 433, 500, 600, and 666 nm for an albite from Tiburon, California. The solutions for all wavelengths closely compare to the new solutions determined for the data by EXCALIBR II. For comparison Figure 3–2 shows the results for EXCALIBR II's 433 nm solution. The most significant differences occur in calculations of the estimated standard errors. Those calculated by EXCALIBR II are slightly larger than those calculated by EXCALIBR. It has been suspected that the estimated standard errors provided by EXCALIBR were too small (Bloss, 1981, p. 308).

Combined Optical and X-ray Studies

Following the determination of all solutions, the program calculates the upper and lower arc settings for the goniometer that will align a given optic vector along either the spindle axis (x-axis) or the light axis (z-axis). Arc settings are provided for both Type I and Type II goniometer heads (Bloss, 1981, p. 233). Of course

not all arc settings will be attainable because of the limited rotation range of most goniometer arcs. If an optic direction that coincides with a crystallographic axis can be brought parallel to x, the crystal becomes oriented for a Weissenberg, oscillation, or rotation photograph. If it can be brought parallel to z, the crystal will be oriented for a precession photograph with the optic direction as the precessing axis. For triclinic crystals, a spindle stage study will not help orient the crystal for an x-ray photograph unless the angles between the optic vectors and the crystallographic axes are, at least, approximately known.

When a spindle stage study is followed by an X-ray study of the crystal, EXCALIBR II uses the before and after goniometer arc settings to calculate the coordinates that the optic vectors assume after the goniometer arcs were re-set for the X-ray study. This is accomplished by computing a transformation matrix

$$T = M_C([010]\rho_L)M_C([001]\rho_U)$$

where M_C represents a general Cartesian rotation matrix (Boisen and Gibbs, 1985) and ρ_L and ρ_U are the turn angles about the y and z axes, respectively. These new optic vector coordinates permit the user to continue the optical study at the current goniometer arc settings. In addition, these new coordinates permit the calculation of the precise angles between the estimated optical vectors and the estimated crystallographic axes determined from the X-ray study.

Statistical Study of Optical Dispersion

Even with a simple detent spindle stage, the dispersion of the optical vectors or their change with temperature may be determined for a crystal. One simply needs to make careful spindle stage studies at two or more different conditions such as wavelengths or temperatures. For example, the wavelengths for which extinction positions were measured for the Tiburon albite sample are entered on

Tiburon Albite - Wolfe

Experimental Treatment ID number = 433.000
Average Reference Azimuth, Mr (esd) = 180.15 (.00)
based on 4 observations.

Biaxial Model
number of iterations(100 max.) = 7
R-squared = .99956
m.s.e. = .000425

S	Ms	Es	CALC(Es)	Es-CALC(Es)
.00	217.90	37.75	37.74	.01
10.00	218.50	38.35	38.49	-.14
20.00	219.10	38.95	39.04	-.09
30.00	219.60	39.45	39.31	.14
40.00	219.40	39.25	39.19	.06
50.00	218.60	38.45	38.50	-.05
60.00	217.30	37.15	36.85	.30
70.00	213.60	33.45	33.39	.06
80.00	205.90	25.75	26.01	-.26
90.00	189.90	9.75	10.56	-.81
100.00	171.70	171.55	170.77	.78
110.00	159.20	159.05	158.86	.19
120.00	153.00	152.85	152.93	-.08
130.00	149.80	149.65	149.55	.10
140.00	147.10	146.95	147.31	-.36
150.00	145.80	145.65	145.65	.00
160.00	144.40	144.25	144.31	-.06
170.00	143.50	143.35	143.20	.15
180.00	142.40	142.25	142.26	-.01

Optic Axial Angle, 2V (ese) = 77.567 (.368)

Computed Cartesian Coordinates

	x	(ese)	y	(ese)	z	(ese)
OA1	.9875	(.0006)	-.0205	(.0038)	.1565	(.0040)
OA2	.2304	(.0027)	.9719	(.0007)	.0494	(.0035)
AB	.7811	(.0006)	.6102	(.0008)	.1321	(.0019)
OB	.6044	(.0010)	-.7921	(.0006)	.0855	(.0055)
ON	-.1568	(.0039)	-.0131	(.0043)	.9875	(.0006)

Figure 3–2. Select portion of EXCALIBR II output file showing solution of the 433 nm data provided in Figure 1. For comparison with EXCALIBR, see Bloss 1981, p. 210.

line 2 (Figure 3–1). Measurement of a crystal's optical dispersion can provide additional insight into structural features such as phase transitions or even possi-

bly petrogenetic histories (Bloss, 1978). Statistics are necessary to help indicate whether the solutions determined for one treatment are really different from those solutions obtained from another treatment. By using a t-test to test the null hypothesis, H_o, of non-dispersion, EXCALIBR II provides an analysis of dispersion that is superior to that of its predecessor EXCALIBR.

When comparing two vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, for possible dispersion, the angle between the vectors, namely θ, is a natural parameter of interest. If θ equals zero, then obviously no dispersion occurred. However, because we can only estimate the optic vectors, $\hat{\mathbf{v}}_i$, we can only estimate the angle between vectors, $\hat{\theta}$. Thus, a finite estimated angle between estimated vectors does not necessarily indicate dispersion. Of equal importance is the estimated standard error of this angle, $\hat{\sigma}_{\hat{\theta}}$. Through propagation of error, EXCALIBR II computes a 3x3 estimated covariance matrix, $\hat{V}_{\hat{v}_i}$, for the coordinates of each estimated optic vector $\hat{\mathbf{v}}_i$. By combining $\hat{V}_{\hat{v}_i}$ for the two vectors under consideration, $\hat{\sigma}^2_{\hat{\theta}}$ is computed according to Equation 2 as

$$\hat{\sigma}^2_{\hat{\theta}} = L^t \begin{bmatrix} \hat{V}_{\hat{v}_1} & \mathbf{O} \\ \mathbf{O} & \hat{V}_{\hat{v}_2} \end{bmatrix} L \tag{3}$$

where

$$L = \begin{bmatrix} -\hat{x}_2/\sin\hat{\theta} \\ -\hat{y}_2/\sin\hat{\theta} \\ -\hat{z}_2/\sin\hat{\theta} \\ -\hat{x}_1/\sin\hat{\theta} \\ -\hat{y}_1/\sin\hat{\theta} \\ -\hat{z}_1/\sin\hat{\theta} \end{bmatrix},$$

$\mathbf{O}$ is a 3x3 matrix of zeros, $[\hat{\mathbf{v}}_1]^t_C = [\hat{x}_1 \ \hat{y}_1 \ \hat{z}_1]$, $[\hat{\mathbf{v}}_2]^t_C = [\hat{x}_2 \ \hat{y}_2 \ \hat{z}_2]$, and $\hat{\theta} = \cos^{-1}(\hat{x}_1\hat{x}_2 + \hat{y}_1\hat{y}_2 + \hat{z}_1\hat{z}_2)$.

EXCALIBR II uses the t-value (t= $\hat{\theta}/\hat{\sigma}_{\hat{\theta}}$) to determine a p-value from the appropriate t-distribution for the two vectors under consideration (Press et al., 1986). The p-value represents the probability of obtaining a statistic greater than

or equal to the observed value, assuming the null hypothesis, $H_o : \theta = 0$, is true. Small p-values (≤ 0.10) suggest rejection of the null hypothesis of non-dispersion. In other words, small p-values imply dispersion.

An equally valid test for indicating statistical differences resulting from experimental treatments can be found by utilizing the estimated differences in coefficients of the compared vectors. If $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$ are the vectors under consideration, then we define $\Delta\hat{\mathbf{v}}$ as

$$[\Delta\hat{\mathbf{v}}]_C = [\hat{\mathbf{v}}_1]_C - [\hat{\mathbf{v}}_2]_C = \begin{bmatrix} \Delta\hat{x} \\ \Delta\hat{y} \\ \Delta\hat{z} \end{bmatrix} = \begin{bmatrix} \hat{x}_1 - \hat{x}_2 \\ \hat{y}_1 - \hat{y}_2 \\ \hat{z}_1 - \hat{z}_2 \end{bmatrix}.$$

Multivariate theory suggests that a modified form of the Hotelling's T^2 statistic (Johnson and Wichern, 1982) be used to test the null hypothesis of nondispersion, $H_o : \Delta\mathbf{v} = 0$. The test statistic, T^2, is computed according to

$$T^2 = [\Delta\hat{\mathbf{v}}]_C^t \hat{S}_P^{-1} [\Delta\hat{\mathbf{v}}]_C \tag{4}$$

where $\hat{S}_P = (\hat{V}_{\hat{v}_1} + \hat{V}_{\hat{v}_2})/2$ is the pooled estimate of the covariance matrix for $\Delta\hat{\mathbf{v}}$. The null hypothesis is rejected if the p-value for T^2, determined from the appropriate F-distribution, is less than or equal to 0.10. As the number of observations or experimental measurements become very large (N > 200), then T^2 approximately follows a χ^2 distribution (Johnson and Wichern, 1982).

Assuming a χ^2 distribution with 2 degrees of freedom corresponding to the number of independent parameters, EXCALIBR used a χ^2-test statistic to test $H_o : \Delta\mathbf{v} = 0$. Computationally, the χ^2-test statistic is analogous to T^2 given in Equation 4 except it assumes all off-diagonal covariance terms in $\hat{S}_P$ are zero, a situation often not true in practice. EXCALIBR's use of a χ^2-test statistic is inappropriate for several reasons. First the χ^2-test only follows a χ^2 distribution when the standard errors of compared parameters ($\sigma_{\Delta\hat{x}}$, $\sigma_{\Delta\hat{y}}$, and $\sigma_{\Delta\hat{z}}$) are known.

In reality, however, these standard errors are unknown and we can only obtain estimates of them ($\hat{\sigma}_{\Delta\hat{x}}$, $\hat{\sigma}_{\Delta\hat{y}}$, and $\hat{\sigma}_{\Delta\hat{z}}$). Second, by using a χ^2-test statistic, EXCALIBR incorrectly ignored the inherent covariance between $\Delta\hat{x}$, $\Delta\hat{y}$, and $\Delta\hat{z}$. Recall that along with Joel's equation (Equation 1) are the constraints that $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$ are normalized. Consequently, this constraint introduces correlation within the coefficients for each optic axis, and therefore because these coefficients are correlated, any vector computed using $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$ must inherently have covariance between coefficients.

Ignoring the covariance between $\Delta\hat{x}$, $\Delta\hat{y}$, and $\Delta\hat{z}$ leads to a dependence of EXCALIBR's χ^2-test value on the orientation of $\Delta\hat{\mathbf{v}}$. EXCALIBR accounted for this orientation dependence essentially by rotating $\Delta\hat{\mathbf{v}}$ and computing a new χ^2-test. It then reported the harmonic mean of p-values determined from these various χ^2-test values (Bloss, 1981, p. 310). However like T^2, a true χ^2-test random variable (ie. zero covariance) should have no orientation dependence.

To illustrate the invariance of T^2 on the orientation of $\Delta\hat{\mathbf{v}}$, let M_C represent a Cartesian rotation matrix, then according to Equation 4

$$T^2 = [\Delta\hat{\mathbf{v}}]_C^t \hat{S}_P^{-1} [\Delta\hat{\mathbf{v}}]_C = [\Delta\hat{\mathbf{v}}_r]_C^t \hat{S}_{P_r}^{-1} [\Delta\hat{\mathbf{v}}_r]_C, \tag{5}$$

where $M_C \Delta\hat{\mathbf{v}} = \Delta\hat{\mathbf{v}}_r$ and $\hat{S}_{P_r} = M_C \hat{S}_P M_C^{-1}$ is the estimated covariance matrix of $\Delta\hat{\mathbf{v}}_r$, the rotated vector. When computing the χ^2-test value, EXCALIBR uses only the diagonal elements of $\hat{S}_P^{-1}$. By recomputing the χ^2-test value for $\Delta\hat{\mathbf{v}}_r$ again using only the diagonal elements of $\hat{S}_{P_r}^{-1}$, the equality in Equation 5 is violated.

Our studies indicate that the simple t-test provides a better measure of dispersion than use of the Hotelling's T^2 for the current application. In addition, use of the t-value as a test statistic has several advantages over the χ^2-test statistic used by EXCALIBR. First of all, the t random variable directly employs the notion

that we are only estimating the standard error, $\hat{\sigma}_{\hat{\theta}}$, of θ. The degrees of freedom used in the t-distribution now utilizes both the number of independent parameters under consideration and the number of experimental data used to compute $\hat{\theta}$. The degrees of freedom for the χ^2 distribution used by EXCALIBR was always two regardless of data size. Furthermore, the estimated covariance of the vector coefficients is incorporated into $\hat{\sigma}_{\hat{\theta}}$ according to Equation 3. In addition, the t random variable is independent of orientation because $\hat{\theta}$ and $\hat{\sigma}_{\hat{\theta}}$ is invariant under any rotation, M_C.

Table 3–1 shows the results of the dispersion analysis provided by both EXCALIBR and EXCALIBR II, respectively, using the data from Figure 3–1. The probability values given by EXCALIBR have been converted to p-values for comparison with p-values from EXCALIBR II. By accepting p-values less than or equal to 0.10 as evidence for rejection of either null hypothesis of non-dispersion, $H_o : \theta = 0$ for EXCALIBR II or $H_o : \Delta\mathbf{v} = 0$ for EXCALIBR, inspection of Table 3–1 shows that EXCALIBR may provide evidence for wrongly rejecting H_o. The p-values provided by EXCALIBR II are statistically valid and thus provide a better measure of dispersion. The results of EXCALIBR II more clearly indicate that wavelength differences of 100 nm or less are insufficient to produce verifiable changes in optic vector positions for the Tiburon albite (Bloss, 1978). Furthermore, EXCALIBR's erroneous rejection of H_o for the obtuse bisectrix between wavelengths 500 and 600 is absent from the dispersion analysis provided by EXCALIBR II. Thus, EXCALIBR II's dispersion analysis suggesting horizontal dispersion for the Tiburon albite supports the observations reported by Winchell and Winchell (1951).

Table 3–1 - Comparison of dispersion analysis provided by EXCALIBR II and EXCALIBR. The dispersion analysis is computed for pairs of wavelengths, λ_1 and λ_2, between 433 and 666 nm. The analysis provided by EXCALIBR II is listed first and includes the angle between optic vectors (Ang) along with its estimated standard error (ese) and the p-value (p). The analysis provided by EXCALIBR is listed below.

λ_1 λ_2	Optic axis 1			Optic axis 2			Acute bisectrix			Obtuse bisectrix			Optic normal		
	Ang	ese	p	Ang	ese	p	Ang	ese	p	Ang	ese	p	Ang	ese	p
433 500	.327	.347	.354	.144	.297	.632	.114	.153	.464	.356	.479	.462	.369	.471	.439
	.315	—	.364	.139	—	.787	.112	—	.493	.356	—	.459	.371	—	.368
433 600	.324	.336	.341	.654	.294	.033	.325	.166	.059	.439	.485	.372	.544	.385	.166
	.205	—	.744	.626	—	.002	.337	—	.004	.434	—	.265	.541	—	.016
433 666	.948	.352	.011	.852	.275	.004	.522	.168	.004	.055	.331	.870	.522	.178	.006
	.831	—	.008	.799	—	.000	.529	—	.000	.042	—	.990	.530	—	.005
500 600	.380	.366	.307	.753	.320	.025	.231	.172	.188	.795	.516	.132	.820	.491	.104
	.343	—	.258	.740	—	.000	.243	—	.069	.789	—	.015	.814	—	.001
500 666	.735	.369	.054	.895	.304	.006	.415	.180	.027	.320	.538	.555	.525	.333	.124
	.647	—	.055	.865	—	.000	.421	—	.000	.314	—	.493	.524	—	.003
600 666	.688	.379	.078	.392	.268	.153	.212	.166	.209	.477	.539	.382	.510	.518	.331
	.673	—	.020	.382	—	.067	.212	—	.027	.476	—	.244	.510	—	.126

APPENDIX 3A

Non-linear Regression:

The Gauss-Newton Method

Recall the general form of Joel's equation used in Chapter 3 involving four parameters, a, b, c, and d,

$$\begin{aligned}-(q_1^2 - p_1^2) = &[(q_1q_2 - p_1p_2) + (q_2^2 - p_2^2)c + (q_2q_3 - p_2p_3)d]a \\ &+ [(q_1q_3 - p_1p_3) + (q_3^2 - p_3^2)d + (q_2q_3 - p_2p_3)c]b \\ &+ [(q_1q_2 - p_1p_2)]c + [(q_1q_3 - p_1p_3)]d.\end{aligned}$$

Joel's equation is an example of a non-linear equation or model because the estimates determined for a and b are dependent on the estimates for c and d. In other words, unlike the multiple linear regression model given in Appendix 2A, the parameters of a non-linear model are not independent of one another.

Non-linear models require that we search for a set of parameter estimates that results in the minimization of

$$L = \sum \epsilon^2.$$

The following discussion presents one method, the so-called **Gauss-Newton method**, of setting up the least-squares quantity, L, and how to proceed to find its minimum value. For functions like Joel's equation, the non-linear function will approximated with a linear function whereby each regression coefficient only involves a single parameter. This linear approximation is made possible by the use of Taylor's formula.

For each measurement, i, the statistical model for Joel's equation can be written as

$$y_i = [x_{i,1} + x_{i,2}\theta_3 + x_{i,3}\theta_4]\theta_1 + [x_{i,4} + x_{i,5}\theta_4 + x_{i,6}\theta_3]\theta_2 + x_{i,7}\theta_3 + x_{i,8}\theta_4 + \epsilon_i$$

or in general for any non-linear model,

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon_i, \tag{1}$$

where y_i represents the response or dependent variable, $\mathbf{x}_i$ is the $1 \times k$ vector of regressor or independent variables, $[x_{i,1}\ x_{i,2} \cdots x_{i,k}]$, $\boldsymbol{\theta}$ represents the $p \times 1$ vector of model or regression parameters, $[\theta_1\ \theta_2, \cdots, \theta_p]^t$, and ϵ_i is the error or residual for the $i = 1, ..., n$ observations. For Joel's equation, $k = 8$ and $p = 4$.

For the general case, assume that $\boldsymbol{\theta}$ are the true parameters for some non-linear model, $f(\mathbf{x}_i, \boldsymbol{\theta})$. Note that $\boldsymbol{\theta}$ is never known from experiment because these are the parameters known only by observing the entire population of data in the universe. Because our experiment represents a finite sample taken from this population, we can only estimate the value of $\boldsymbol{\theta}$. Consequently, the accuracy of the estimates for $\boldsymbol{\theta}$ will depend on the size of this sample. Assume that $\hat{\boldsymbol{\theta}}_0$ is some chosen starting estimate that is close to $\boldsymbol{\theta}$, i.e.

$$\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0 + \boldsymbol{\gamma} \tag{2}$$

where $\boldsymbol{\gamma}$ is the difference vector. Because we know the value of $\hat{\boldsymbol{\theta}}_0$, it at this point where the linear approximation will be derived or polynomial expanded. According to **Taylor's formula** (Swokowski, 1983), the polynomial approximation to the non-linear function, $f(\mathbf{x}_i, \boldsymbol{\theta})$, in the neighborhood of $\hat{\boldsymbol{\theta}}_0$ is written as

$$f(\mathbf{x}_i, \boldsymbol{\theta}) \approx a_0 + a_1(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0) + a_2(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^2 + a_3(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^3 + \cdots + a_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^p \tag{3}$$

where $a_0, a_1, a_2, a_3, \ldots, a_p$, are constants. These constants are evaluated by equating the zero, first, second, ..., etc., derivatives of the polynomial approximation (Equation 3), to the zero, first, second , ..., etc., derivatives of the non-linear function $f(\mathbf{x}_i, \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0$. The first, second, and third derivatives of the polynomial represented by Equation 3 are

$$f'(\mathbf{x}_i, \boldsymbol{\theta}) = 1a_1 + 2a_2(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0) + 3a_3(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^2 + \cdots + pa_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^{(p-1)}$$

$$f''(\mathbf{x}_i,\boldsymbol{\theta}) = 1 \cdot 2a_2 + 2 \cdot 3a_3(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0) + \cdots + (p-1) \cdot pa_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^{(p-2)}$$

$$f'''(\mathbf{x}_i,\boldsymbol{\theta}) = 1 \cdot 2 \cdot 3a_3 + \cdots + (p-2) \cdot (p-1) \cdot pa_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^{(p-3)}.$$

If $\hat{\boldsymbol{\theta}}_0 \rightarrow \boldsymbol{\theta}$, then the two functions can be set equal resulting in

$$a_0 = \frac{f(\mathbf{x}_i,\boldsymbol{\theta})}{0!}; a_1 = \frac{f'(\mathbf{x}_i,\boldsymbol{\theta})}{1!}; a_2 = \frac{f''(\mathbf{x}_i,\boldsymbol{\theta})}{2!}; a_3 = \frac{f'''(\mathbf{x}_i,\boldsymbol{\theta})}{3!};\ \text{etc.}$$

Replacing the constants, $a_0, a_1, a_2, a_3, \ldots$, etc., on the right side of Equation 3 by these identities results in the following linear approximation:

$$\begin{aligned} f(\mathbf{x}_i,\boldsymbol{\theta}) \approx \frac{f(\mathbf{x}_i,\boldsymbol{\theta})}{0!} &+ \frac{f'(\mathbf{x}_i,\boldsymbol{\theta})}{1!}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0) + \frac{f''(\mathbf{x}_i,\boldsymbol{\theta})}{2!}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^2 \\ &+ \frac{f'''(\mathbf{x}_i,\boldsymbol{\theta})}{3!}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^3 + \cdots + \frac{f^p(\mathbf{x}_i,\boldsymbol{\theta})}{p!}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)^p \end{aligned} \tag{4}$$

Replacing $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)$ by $\boldsymbol{\gamma}$, Equation 4 can be rewritten as

$$\begin{aligned} f(\mathbf{x}_i,\boldsymbol{\theta}) \approx f(\mathbf{x}_i,\boldsymbol{\theta}) &+ \frac{f'(\mathbf{x}_i,\boldsymbol{\theta})}{1!}\boldsymbol{\gamma} + \frac{f''(\mathbf{x}_i,\boldsymbol{\theta})}{2!}\boldsymbol{\gamma}^2 \\ &+ \frac{f'''(\mathbf{x}_i,\boldsymbol{\theta})}{3!}\boldsymbol{\gamma}^3 + \cdots + \frac{f^p(\mathbf{x}_i,\boldsymbol{\theta})}{p!}\boldsymbol{\gamma}^p \end{aligned}$$

Furthermore, if $\boldsymbol{\gamma}$ is small, then to a **first approximation**

$$\begin{aligned} f(\mathbf{x}_i,\boldsymbol{\theta}) &\approx f(\mathbf{x}_i,\boldsymbol{\theta}) + \frac{f'(\mathbf{x}_i,\boldsymbol{\theta})}{1!}\boldsymbol{\gamma} \\ &\approx f(\mathbf{x}_i,\boldsymbol{\theta}) + \frac{\partial f(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{\gamma} \end{aligned} \tag{5}$$

Since we are only able to evaluate or calculate the function $f(\mathbf{x}_i,\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}_0$, Equation 5 will be rewritten as

$$\begin{aligned} f(\mathbf{x}_i,\boldsymbol{\theta}) &\approx f(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_0) + \left.\frac{\partial f(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0}\boldsymbol{\gamma} \\ &\approx \hat{y}_i + \left.\frac{\partial f(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0}\boldsymbol{\gamma}. \end{aligned}$$

It follows from Equation 1 that

$$y_i \approx \hat{y}_i + \frac{\partial f(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{\gamma} + \epsilon_i,$$

which leads to the **linear approximation** of the non-linear function

$$y_i^* = y_i - \hat{y}_i \approx \left(\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^t \boldsymbol{\gamma} + \epsilon_i, \tag{6}$$

for each $i = 1, \ldots, n$.

Equation 6 can be written in matrix notation as

$$\mathbf{y}^* \approx \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{7}$$

where

$$\mathbf{y}^* = \begin{bmatrix} y_1^* \\ y_2^* \\ y_3^* \\ \vdots \\ y_n^* \end{bmatrix}; \mathbf{W} = \begin{bmatrix} \frac{\partial f(\mathbf{x}_1, \boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(\mathbf{x}_1, \boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial f(\mathbf{x}_1, \boldsymbol{\theta})}{\partial \theta_3} & \cdots & \frac{\partial f(\mathbf{x}_1, \boldsymbol{\theta})}{\partial \theta_p} \\ \frac{\partial f(\mathbf{x}_2, \boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(\mathbf{x}_2, \boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial f(\mathbf{x}_2, \boldsymbol{\theta})}{\partial \theta_3} & \cdots & \frac{\partial f(\mathbf{x}_2, \boldsymbol{\theta})}{\partial \theta_p} \\ \frac{\partial f(\mathbf{x}_3, \boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(\mathbf{x}_3, \boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial f(\mathbf{x}_3, \boldsymbol{\theta})}{\partial \theta_3} & \cdots & \frac{\partial f(\mathbf{x}_3, \boldsymbol{\theta})}{\partial \theta_p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{x}_n, \boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(\mathbf{x}_n, \boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial f(\mathbf{x}_n, \boldsymbol{\theta})}{\partial \theta_3} & \cdots & \frac{\partial f(\mathbf{x}_n, \boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix};$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_p \end{bmatrix}; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

To find the least-squares estimates of the model parameters, $\boldsymbol{\theta}$, L must be **minimized** where

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^t \boldsymbol{\epsilon}.$$

Since according to Equation 7, $\mathbf{y}^* \approx \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, then $\boldsymbol{\epsilon} \approx \mathbf{y}^* - \mathbf{W}\boldsymbol{\gamma}$ so that

$$\begin{aligned} L = \boldsymbol{\epsilon}^t\boldsymbol{\epsilon} &\approx [\mathbf{y}^* - \mathbf{W}\boldsymbol{\gamma}]^t[\mathbf{y}^* - \mathbf{W}\boldsymbol{\gamma}] \\ &\approx [\mathbf{y}^{*t} - \boldsymbol{\gamma}^t\mathbf{W}^t][\mathbf{y}^* - \mathbf{W}\boldsymbol{\gamma}] \\ &\approx \mathbf{y}^{*t}\mathbf{y}^* - \mathbf{y}^{*t}\mathbf{W}\boldsymbol{\gamma} - \boldsymbol{\gamma}^t\mathbf{W}^t\mathbf{y}^* + \boldsymbol{\gamma}^t\mathbf{W}^t\mathbf{W}\boldsymbol{\gamma}. \end{aligned}$$

Further, since $\mathbf{y}^{*t}\mathbf{W}\boldsymbol{\gamma} = \boldsymbol{\gamma}^t\mathbf{W}^t\mathbf{y}^*$, then

$$L \approx \boldsymbol{\gamma}^t\boldsymbol{\gamma} - 2\mathbf{y}^{*t}\mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\gamma}^t\mathbf{W}^t\mathbf{W}\boldsymbol{\gamma}.$$

By rearranging the above equation into

$$L \approx \boldsymbol{\gamma}^t \mathbf{W}^t \mathbf{W} \boldsymbol{\gamma} - 2\mathbf{y}^{*t} \mathbf{W} \boldsymbol{\gamma} + \mathbf{y}^{*t} \mathbf{y}^*, \tag{8}$$

showing that L has the form of a quadratic function in terms of $\boldsymbol{\gamma}$ but is only approximately quadratic. Note that because the $\mathbf{W}$ matrix (derivative matrix) is evaluated at some point $\hat{\boldsymbol{\theta}}_0$, the Hessian matrix, $H = \mathbf{W}^t\mathbf{W}$, is also dependent on $\hat{\boldsymbol{\theta}}_0$. If L were truly quadratic (as in a linear model), then the elements of the Hessian would be constant and not depend on $\hat{\boldsymbol{\theta}}_0$. Furthermore, the value of $\mathbf{y}^*$ is also dependent on $\hat{\boldsymbol{\theta}}_0$ because some value of $\hat{\boldsymbol{\theta}}_0$ is needed to evaluate the function, $\hat{y}_i$, to subtract from y_i to obtain y_i^*. Consequently, finding the minimum value of L or the estimated location of the minimizer, $\hat{\boldsymbol{\gamma}}$, in a parameter space is dependent on $\hat{\boldsymbol{\theta}}_0$ and is thus an **iterative process**. When the value of $\hat{\boldsymbol{\gamma}}$ is close to zero, we have arrived at a acceptable value of $\hat{\boldsymbol{\theta}}$.

How do we find an improved value of $\hat{\boldsymbol{\theta}}_0$ that results in a lower minimum for L? For this we may employ the method of minimization referred to as the **Newton-Raphson method** (Appendix 1A). Basically, the method requires that a quadratic model be constructed at some point, $\hat{\boldsymbol{\gamma}}$, keeping in mind that the model is also dependent on $\hat{\boldsymbol{\theta}}_0$. The minimizer of the model, $\hat{\boldsymbol{\gamma}}$, is used as a step or shift in $\hat{\boldsymbol{\theta}}_0$, ie. compute

$$\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\gamma}}.$$

If $\hat{\boldsymbol{\gamma}} \approx 0.0$ (ie. $\hat{\boldsymbol{\theta}}_1 \approx \hat{\boldsymbol{\theta}}_0$), then $\hat{\boldsymbol{\theta}}_1$ is accepted as the best estimate of $\boldsymbol{\theta}$ for the given data set. Otherwise, $\hat{\boldsymbol{\theta}}_1$ is used to construct a new quadratic model resulting in a new value of $\hat{\boldsymbol{\gamma}}_1$ which hopefully leads to reduced value of Equation 8. The procedure continues t times until $\hat{\boldsymbol{\gamma}}_t$ is close to zero.

The formulation of the quadratic model, $M(\hat{\boldsymbol{\gamma}})$, occurs at each successive value that chosen for $\hat{\boldsymbol{\theta}}_0$. As far as the model is concerned, $\hat{\boldsymbol{\gamma}}$ is its origin. Therefore, the

form of the quadratic model having the same algebraic form as Equation 8 is

$$M(\hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{\gamma}}^t H_0 \hat{\boldsymbol{\gamma}} - 2(\nabla_0 M)^t \hat{\boldsymbol{\gamma}} + M(0)$$

where H_0 and $\nabla_0 M$ denote that the derivatives are evaluated at the origin (Appendix 1A). Recall that to find the minimizer of a quadratic function, the gradient is set equal to zero, i.e.

$$\frac{\partial M}{\partial \hat{\boldsymbol{\gamma}}} = \nabla_{\hat{\boldsymbol{\gamma}}} M = 2H_0 \hat{\boldsymbol{\gamma}} - 2\nabla_0 M = 0$$

therefore the estimated location of the minimizer of the quadratic model (called the **newton point**) is given by

$$\hat{\boldsymbol{\gamma}} = H_0^{-1} \nabla_0 M.$$

Because the model's origin is at the point $\hat{\boldsymbol{\gamma}}$, the Hessian of the model at its' origin, H_0, is really the Hessian of the function, L, at the point $\hat{\boldsymbol{\theta}}_0$, $H_{\hat{\boldsymbol{\theta}}_0}$. Furthermore, the gradient of the model at its origin, $\nabla_o M$, is really the gradient of the function, L, at $\hat{\boldsymbol{\theta}}_0$, $\nabla_{\hat{\boldsymbol{\theta}}_0} L$. Thus, the minimizer of the model, $\hat{\boldsymbol{\gamma}}$, can be written as

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= H_{\hat{\boldsymbol{\theta}}_0}^{-1} \nabla_{\hat{\boldsymbol{\theta}}_0} L \\ &= (\mathbf{W}^t \mathbf{W})^{-1} \mathbf{W}^t \mathbf{y}^*(\hat{\boldsymbol{\theta}}_0). \end{aligned}$$

A value of $\hat{\boldsymbol{\theta}}_t$ that further reduces L (Equation 8) is found by adding the location of the minimizer for the model, $\hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\theta}}_{(t-1)}}$, based on the value $\hat{\boldsymbol{\theta}}_{(t-1)}$. In other words,

$$\begin{aligned} \hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_{(t-1)} + (\mathbf{W}^t \mathbf{W})^{-1} \mathbf{W}^t \mathbf{y}^*_{\hat{\boldsymbol{\theta}}_{(t-1)}} \\ &= \hat{\boldsymbol{\theta}}_{(t-1)} + \hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\theta}}_{(t-1)}}. \end{aligned} \tag{23}$$

The above procedure is tantamount to resolving the normal equations at each successive point $\hat{\boldsymbol{\theta}}_{(t-1)}$ at each iteration t and is known as the **Gauss-Newton procedure** (Myers, 1986). Convergence is obtained when $\boldsymbol{\gamma}_{\hat{\boldsymbol{\theta}}^{(t-1)}}$ becomes very small which indicates that the gradient term $\mathbf{W}^t \mathbf{y}^*$ is close to zero.

References

Armbruster, T. and Bloss, F. D. (1982) Orientation and effects of channel H_2O and CO_2 in cordierite. American Mineralogist, 67, 284-291.

Bartelmehs, K.B.; Bloss, F.D.; Downs, R.T. and Birch, J.B. (1992) Excalibr II. Zeitschrift für Kristallographie, 199, 185-196.

Bloss, F. D. (1978) The spindle stage: a turning point for optical crystallography. American Mineralogist, 63, 433-447.

Bloss, F. D. (1981) The Spindle Stage: Principles and Practise. New York, Cambridge University Press.

Bloss, F. D. and Riess, D. (1973) Computer Determination of 2V and Indicatrix Orientation from Extinction Data. American Mineralogist, 58, 1052-1061.

Boisen, Jr., M.B. and Gibbs, G.V. (1985) Mathematical Crystallography. Washington, D.C., Mineralogical Society of America.

Greiner, D. J. and Bloss, F. D. (1987) Amblygonite-montebrasite optics: Response to (OH^-) orientation and rapid estimation of F from 2V. American Mineralogist, 72, 617-624.

Gunter, M. E. and Bloss, F. D. (1987) Andalusite-kanonaite: Lattice and optical parameters. American Mineralogist, 67, 1068-1087.

Joel, N. (1965) Determination of the optic axes and 2V: electronic computation from extinction data. The Mineralogical Magazine, 35, 412-417.

Johnson, R. A. and Wichern, D. W. (1982) Applied Multivariate Statistical Analysis. New Jersey, Prentice Hall.

Kendal, M., Stuart, A. and Ord, J.K. (1987) Kendall's Advanced Theory of Statistics, Volume 1. New York, Oxford University Press.

Myers, Raymond H. (1986) Classical and Modern Regression with Applications, Second Edition. Boston, PWS-KENT Publishing Company.

Milton, J.S. and Arnold, J.C. (1990) Introduction to Probability and Statistics. New York, McGraw-Hill Publishing Company.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1986) Numerical Recipes: The art of scientific computing. Melbourne, Cambridge University Press.

Seber, G.A.F. and Wild, C.J. (1989) Nonlinear Regression. Canada, John Wiley and Sons, Inc..

Su, S. C., Bloss, F. D., Ribbe, P. H., and Stewart, D. B. (1984) Optical axial angle, a precise measure of Al,Si ordering in T_1 tetrahedral site of K-rich alkali feldspars. American Mineralogist, 69, 440-448.

Swokowski, E.W. (1983) Calculus with analytic geometry, Alternate Edition.

Boston, PWS Publishing.

Wilcox, R. E. (1959) Use of spindle stage for determining refractive indices of crystal fragments. American Mineralogist, 44, 1272-1293.

Winchell, A.N. and Winchell, J. (1951) Elements of Optical Mineralogy, Part II: Descriptions of minerals. New York, Wiley.

VITA

Kurt Lane Bartelmehs was born in Pulaski, Virginia on March 30, 1962. After graduating from Pulaski County High School in June of 1980, he journeyed to Blacksburg to begin college. Following his stint in a rock and roll band and short job as a graphic arts photographer, he finally received his Bachelor of Science degree in geological sciences from Virginia Polytechnic Institute and State University, Blacksburg, Virginia in 1985. It was about this time that he married Carol. Together they worked so that he could receive his Master of Science degree in geological sciences from VPI & SU in 1987. Following a two year excursion in materials engineering science, he returned to geological sciences in 1989.