Protein Structure Normal Mode Analysis on the Positive Semidefinite Matrix Manifold

Xiao-Bo Li*
Cheriton School of Computer Science
University of Waterloo, Canada
x22li@uwaterloo.ca

Forbes J. Burkowski
Cheriton School of Computer Science
University of Waterloo, Canada
fjburkowski@uwaterloo.ca

Henry Wolkowicz
Department of Combinatorics & Optimization
University of Waterloo, Canada
hwolkowicz@uwaterloo.ca

Abstract

M. Tirion's use of a simple Hookean potential energy in normal mode analysis (NMA) led to the study of protein dynamics using elastic network models (ENMs). Squaring the distance in Tirion's potential energy gives a new potential energy that is a function on the rank 3 positive semidefinite (PSD) matrix manifold. Fixed rank PSD matrix manifolds have received much attention within the context of optimization algorithms. This paper and our prior work suggests that these manifolds are an appropriate setting for studying protein dynamics using ENMs. Fully understanding the implications of this close relationship between ENMs and the rank 3 PSD matrix manifold is the subject of future research.

keywords: Euclidean distance matrices, elastic network models, matrix manifolds, normal mode analysis, positive semidefinite matrices, Riemannian manifold.

1 Introduction

Normal mode analysis (NMA) is a leading method for studying protein flexibility. Tirion [15] proposed to replace semi-empirical potentials with a simpler Hookean potential. This Hookean potential is a function of distance. By changing from distance to distance-squared, also known as *quadrance* [18], the Hookean potential becomes a function on the rank 3 positive semidefinite (PSD) matrix manifold. This same function has been used as an objective function in optimization [1, 14], and suggests a close relation between elastic network models (ENMs) and the rank 3 PSD matrix manifold.

This paper is structured as follows. In Section 2, we review some mathematical background on PSD matrices, and their closely related Euclidean distance matrices (EDMs). In Section 3, we review NMA as formulated using distances. We call this formulation *classical* NMA. In Section 4, we present NMA using quadrance. We refer to this formulation as *quadrance* NMA. In Section 5 we show that the root-mean-square (RMS) fluctuations and the eigenvalue histogram for classical NMA and quadrance NMA are very similar. This similarity further supports the conjecture that the rank 3 PSD matrix manifold is an appropriate setting for studying protein dynamics using ENMs.

2 The PSD Matrix Manifold

For a set of n atoms, labelled $1, \ldots, n$, in 3-dimensional space, an EDM matrix is the matrix where each entry is the quadrance between atoms i and j, $i, j \in \{1, \ldots, n\}$. Denote the Cartesian coordinates of atom i by $y_i \in \mathbb{R}^3$. Let Y be the $n \times 3$ matrix with each row being y_i^T, the transpose of y_i. Assume these points are centered at the origin because a protein structure is invariant to translation. The *centered* Gram matrix of these atoms is given by $X = YY^T$, it has rank 3. Given a fixed rank r, the set of fixed rank PSD matrices, denoted $\mathbf{S}_+^{n,r}$, is a Riemannian matrix manifold. The geometry of this manifold is not unique, see [17]. A protein structure is invariant to rotation. This invariance gives the set of rank 3 PSD matrices a quotient geometry, as seen in [7, 14].

2.1 The EDM Completion Problem

The EDM completion problem seeks to recover missing distances for a set of points, thereby recovering the coordinates of all points. For an $n \times n$ EDM, D, A

*Corresponding author.

linear mapping,

$$D = \mathcal{K}(YY^T) = \text{diag}(YY^T)\mathbf{1}^T + \mathbf{1}\text{diag}(YY^T)^T - 2YY^T \tag{1}$$

relates each centered Gram matrix to its corresponding EDM. Here, $\text{diag}(A) \in \mathbb{R}^n$ is a vector representing the diagonal of the $n \times n$ matrix A, and $\mathbf{1} \in \mathbb{R}^n$ is a vector of all 1's. The EDM completion problem thus provides a setting for studying the theory and applications of PSD matrix manifolds. Many of the mathemtical tools resulting from studying the EDM completion problem [1, 10, 14] is relevant to protein ENMs.

2.2 Classical Dynamics and Riemannian Manifolds

Recall that classical dynamics is formulated on Riemannian manifolds [2]. For a Riemannian manifold $\mathcal{M}$, Arnold [2] defines the *kinetic energy* as the quadratic form on the tangent space of each point $x \in \mathcal{M}$. He also defines the *potential energy* as any differentiable function $U : \mathcal{M} \to \mathbb{R}$. Diagonalizing the Hessian matrix of the quadratic approximation of the potential energy gives the normal modes of the system.

2.3 The Benefits of $\mathcal{M} = \mathbf{S}_+^{n,3}$

Tirion's potential energy is defined on $\mathcal{M} = \mathbb{R}^{3n}$. The benefits and limitations of studying protein dynamics using ENMs on $\mathbf{S}_+^{n,3}$ versus $\mathbb{R}^{3n}$ is still the subject of continued research. We state some known benefits.

(1) A protein structure's rotational invariance is encoded in the Gram matrix via the equivalence $YQ(YQ)^T = YQQ^TY^T = YY^T$, where Q is a 3×3 orthogaonal matrix, that is, $QQ^T = Q^TQ = I$. A protein structure's translational invariance is addressed by limiting to centered Gram matrices.

(2) The $n \times n$ EDM cone is convex for any number of points n. The convexity of the EDM cone gives it well understood mathematical properties. In contrast, the set of distance matrices is not convex when $n > 3$, see Section 6.3 of Dattorro [5], suggesting the EDM cone may be a more natural choice in applications.

(3) Protein dynamics is typically studied with coarse-grained models that approximate the original structure. When performing principal geodesic analysis on $\mathbf{S}_+^{n,3}$, facial reduction can be done in place of coarse-graining to reduce the size of the Gram matrix [12]. Facial reduction is *exact* in the sense the atomic coordinates are not changed. Facial reduction is not the focus of this paper.

(4) We generated transitional conformations between two protein conformations using $\mathbf{S}_+^{n,3}$ in [11]. We presented two example lattice structures which showed $\mathbf{S}_+^{n,3}$ preserved bond angles in the final transitional conformation better than $\mathbb{R}^{3n}$.

In the remaining of this paper, we show the potential energy defined on $\mathcal{M} = \mathbf{S}_+^{n,3}$ has similar properties to Tirion's potential energy. This similarity further supports the appropriateness of studying protein dynamics on $\mathbf{S}_+^{n,3}$.

3 Classical NMA

We now review how NMA is formulated using distances [8, 9, 15].

We will consider a coarse-grained network model with n residues whose positions are represented by their α-carbons. We will assume all atomic masses are 1. Let $\mathcal{D}$ denote the set of pairwise α-carbons within a given distance, or quadrance, threshold. Tirion proposed the Hookean potential energy:

$$\begin{aligned} U(\mathbf{y}) &= U(\mathbf{y}_0 + \delta) \\ &= \sum_{(i,j)\in\mathcal{D}} \frac{C}{2} \left(\|(y_i^0 + \delta_i) - (y_j^0 + \delta_j)\| - \|y_i^0 - y_j^0\| \right)^2 . \end{aligned} \tag{2}$$

C is a constant assumed to be the same for all interacting pairs [15]; without loss of generality, we will assume $C = 1$ in this paper. $\delta_i \in \mathbb{R}^3$ is a perturbation to the initial coordinate, y_i^0 of α-carbon i.

NMA requires the construction of the second order quadratic potential of the summand near $\mathbf{y}_0 \in \mathbb{R}^{3n}$, the vector containing all y_i^0's.

$$U(\mathbf{y}_0 + \delta) \approx \sum_{(i,j)\in\mathcal{D}} (\delta_i - \delta_j)^T G_{ij}(0)(\delta_i - \delta_j) . \tag{3}$$

The $G_{ij}(0)$ term in equation (3) is a 3×3 symmetric matrix. It is given by the Hessian of the summand in equation (2). When taking the derivative, the summand in equation (2) can be considered as a function of $r_{ij} = y_i - y_j$ evaluated at $r_{ij}^0 = y_i^0 - y_j^0$.

$$G_{ij}(0) = \frac{(y_i^0 - y_j^0)(y_i^0 - y_j^0)^T}{(y_i^0 - y_j^0)^T(y_i^0 - y_j^0)} . \tag{4}$$

Equation (3) can be expressed using matrices as $\delta^T G_0 \delta$ where G_0 has a Laplacian structure. Consider the case of just three α-carbons, $n = 3$, and the following special 9×9 Laplacian matrix of 3×3 blocks of G_{ij}:

$$G_0 = \begin{pmatrix} G_{01} + G_{02} & -G_{01} & -G_{02} \\ -G_{01} & G_{01} + G_{12} & -G_{12} \\ -G_{02} & -G_{12} & G_{02} + G_{12} \end{pmatrix} . \tag{5}$$

For a vector $\delta = (\delta_0^T, \delta_1^T, \delta_2^T)^T \in \mathbb{R}^9$, we have:

$$\delta^T G_0 \delta = \sum_{i<j} (\delta_i - \delta_j)^T G_{ij} (\delta_i - \delta_j) \tag{6}$$

For a protein with n α-carbons, the matrix G_0 of the quadratic form in equation (3) is thus given by a sparse $3n \times 3n$ matrix, consisting of 3×3 blocks. The (i,j)-th block, for $i \neq j$ is given by $-G_{ij}(0)$. The (i,i)-th diagonal block is given by,

$$\sum_{k=1}^{i-1} G_{ki}(0) + \sum_{k=i+1}^{n} G_{ik}(0) = \sum_{k:k\neq i} G_{ki}(0). \tag{7}$$

4 Quadrance NMA

The authors of [1, 14] formulated the Euclidean distance matrix completion problem using the objective function:

$$f(X) = f(YY^T) = \| H \circ (\mathcal{K}(YY^T) - D_0) \|_F^2 \ . \tag{8}$$

H is a symmetric matrix with binary entries; $H_{ij} = 1$ if $(i,j) \in \mathcal{D}$, 0 otherwise. D_0 is the partial EDM we wish to complete. In [13], the author gave the following equivalent expression:

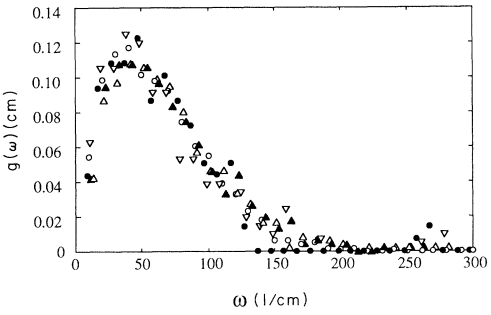
$$\begin{aligned} f(YY^T) &= \sum_{(i,j)\in\mathcal{D}} ((e_i - e_j)^T YY^T (e_i - e_j) - d_{ij}^0)^2 \\ &= \sum_{(i,j)\in\mathcal{D}} ((y_i - y_j)^T (y_i - y_j) - d_{ij}^0)^2 \ . \end{aligned} \tag{9}$$

Equation (9) is clearly the Hookean potential energy, equation (2), with distances replaced by quadrance; it is a matrix function on the rank 3 PSD matrix manifold. This is thus the potential energy for quadrance NMA. In the context of EDM completion, $d_{ij}^0 = \|y_i^0 - y_j^0\|^2 = (y_i^0 - y_j^0)^T (y_i^0 - y_j^0)$ is the (i,j)-th known entry. In the context of NMA, d_{ij}^0 is the initial EDM of the starting conformation, $e_i, e_j \in \mathbb{R}^n$ are canonical basis vectors, and the y_i's are perturbed by some amount δ_i, from the initial coordinates y_i^0's. Note that equation (2) cannot be expressed as a matrix function because the elementwise square-root function cannot be expressed as a matrix operation.

For the quotient geometry of $\mathbf{S}_+^{n,3}$ used in this paper, the Riemannian metric for matrices A, B on the tangent space is given by $\langle A, B \rangle = \text{Trace}(A^T B)$ [7, 14]. The kinetic energy for the initial conformation Y_0 at time $t = 0$ is thus:

$$T(\dot{Y}_0) = \frac{1}{2} \langle \dot{Y}_0, \dot{Y}_0 \rangle \ , \tag{10}$$

where $\dot{Y}_0$ is a tangent vector on the tanget plane at the initial conformation at Y_0.



(a)

Figure 1: Classical NMA found that the shape of the density of normal modes is similar for many proteins. Taken from [3].

We can find the quadratic approximation of the potential energy using a procedure analogous to that given in Section 3. The summand in equation (9) can be expanded to second order. When any constants are ignored we have:

$$G_{ij}(0) = (y_i^0 - y_j^0)(y_i^0 - y_j^0)^T \ . \tag{11}$$

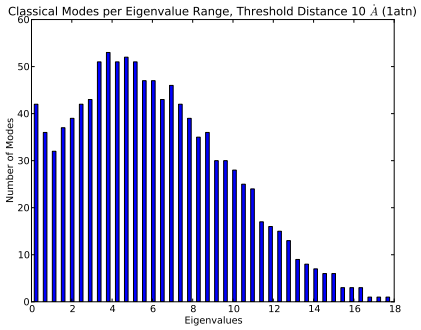
The Hessian matrix G_0 is given by the same structure as in Section 3. NMA requires eigendecomposition of G_0. From equation (4) and (11), we see that G_0 is similar for classical and quadrance NMA, except for a division done in equation (4). Therefore, the computational cost for classical and quadrance NMA is very similar.

5 Comparison of Classical and Quadrance Modes

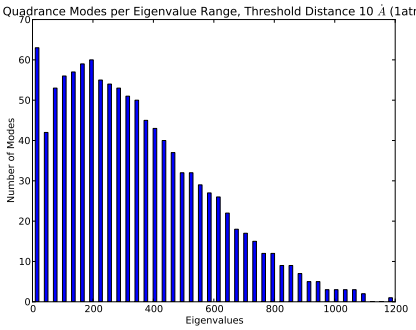
Tirion justified the appropriateness of the Hookean potential energy by showing the resulting density of normal modes (Figure 1 of [15]), and RMS fluctuations (Figures 2 and 3 of [15]) closely match the L79 potential. In this section, we present the close match of these graphs between classical and quadrance NMA to justify the appropriateness of using $\mathbf{S}_+^{n,3}$ to further study ENMs. Our graphs are generated using pyplot [6].

We begin with a discussion of G-Actin because this was the main protein used by Tirion [3, 4, 15, 16] to validate her methodology.

Ben-Avraham [3] observed the shape of the density of normal modes is similar for many globular proteins; he called this shape a "universal curve", see Figure 1. In Figure 2, we present the histograms generated by grouping eigenvalues into 40 bins, using pyplot's `hist` function. Both histograms exhibit the characteristic "universal curve" shape. They are not identical because the magnitude of the eigenvalues are different due to



(a) Classical modes.



(b) Quadrance modes.

Figure 2: Classical and quadrance mode histograms for 1ATN both follow the shape shown in Figure 1.

differences in equations (4) and (11).

We now examine the RMS fluctuations of G-actin. The formulas used have been described previously in for example [8, 16]. The RMS fluctuation of all α-carbons per normal mode k, σ_k, is given by:

$$\sigma_k = \left(\sum_{i=1}^{n} \frac{(\sigma_k^i)^2}{n} \right)^{\frac{1}{2}} , \tag{12}$$

where

$$\sigma_k^i = \left|\left| v_k^i \frac{\alpha_k}{\sqrt{2}} \right|\right| , \tag{13}$$

and $v_k = ((v_k^1)^T, \ldots, (v_k^n)^T)^T \in \mathbb{R}^{3n}$ is the eigenvector for mode k. The authors in [8,16] have used an α_k value of:

$$\alpha_k = \left(\frac{2k_BT}{\lambda_k} \right)^{\frac{1}{2}} , \tag{14}$$

where λ_k is the k-th eigenvalue, k_B is the Boltzmann constant, and T is temperature. However, since the constants do not affect the shape of the RMS plots, we have ignored them and will use an α_k value of:

$$\alpha_k = \frac{1}{\sqrt{\lambda_k}} . \tag{15}$$

Figure 3 plots these values for both classical and quadrance modes. Both graphs tapers off quickly. This is expected because, as discussed in [15], lower modes give large amplitude low frequency motions of atoms, while higher modes related to rapid small atomic oscillations. Quantum mechanical effects become important for such rapid small oscillations, and at this scale the Hookean and $\mathbf{S}_+^{n,3}$ potential energies are not appropriate. Figure 3 shows the potential energy on $\mathbf{S}_+^{n,3}$ has captured this drop in amplitude just as well as the classical Hookean potential energy.

Next, we consider σ^i, the RMS fluctuation of residue i due to *all* modes for each α-carbon, ignoring the first 6 which are rigid motions.

$$\sigma^i = \left(\sum_{k=7}^{3n} (\sigma_k^i)^2 \right)^{\frac{1}{2}} , \tag{16}$$

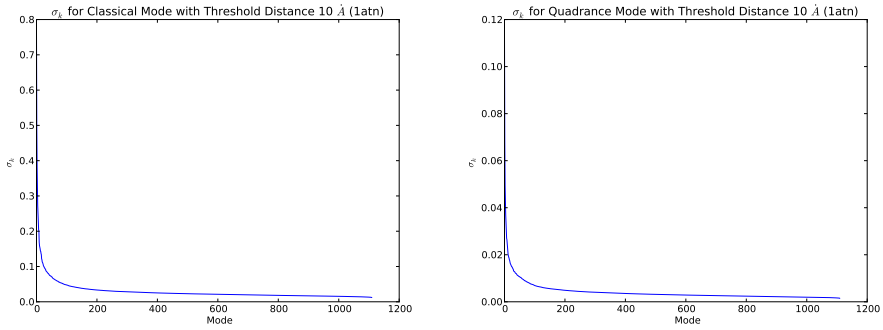
In Figure 4, we present the σ^i graph for numerous proteins. For these proteins, the classical and quadrance normal mode density histograms and the graph for σ_k have a similar discussion to 1ATN. That is, the density histograms all follow the shape in Figure 1, yet both are not identical, and the σ_k graph tapers off similar to Figure 3. Hence, we will not present those graphs. σ^i is more interesting because it is different for each protein. As Figure 4 shows, quadrance NMA reproduces the shape seen in classical NMA, implying that $\mathbf{S}_+^{n,3}$ is appropriate for studying protein dynamics.

6 Conclusion

In this paper, we presented NMA using quadrance. Theoretically, quadrance NMA is the formulation of NMA on the rank 3 PSD matrix manifold. This manifold is widely studied in optimization, but its use in studying protein dynamics is to date rare. We have observed the RMS fluctuations and eigenvalue histogram produced by quadrance modes show close match to their classical counterparts which in turn closely matches the L79 potential. These results suggest the rank 3 PSD matrix manifold should be further investigated as a setting for studying protein dynamics.

References

[1] Alfakih, A., Khandani, A., Wolkowicz, H.: Solving Euclidean distance matrix completion problems via semidefinite programming. Computational



(a) σ_k for classical mode.

(b) σ_k for quadrance mode.

Figure 3: Quadrance NMA captures the drop in RMS fluctuation as mode increases, just as seen in classical NMA. These graphs are for 1ATN.

Optimization and Applications 12(1-3), 13–30 (January 1999)

[2] Arnold, V.: Mathematical Methods of Classical Mechanics. Springer-Verlag (1978)

[3] ben-Avraham, D.: Vibrational normal-mode spectrum of globular proteins. Physical Review B 47(21), 14559 (1993)

[4] ben-Avraham, D., Tirion, M.M.: Normal modes analyses of macromolecules. Physica A: Statistical Mechanics and its Applications 249(1), 415–423 (1998)

[5] Dattorro, J.: Convex optimization and Euclidean distance geometry. MeBoo (2014)

[6] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing In Science & Engineering 9(3), 90–95 (2007)

[7] Journée, M., Bach, F., Absil, P.A., Sepulchre, R.: Low-rank optimization on the cone of positive semidefinite matrices. SIAM J. OPTIM 20(5), 2327–2351 (May 2010)

[8] Kim, M.K.: Elastic Network Models of Biomolecular Structure and Dynamics. Ph.D. thesis, The Johns Hopkins University (2004)

[9] Kim, M.K., Jang, Y., Jeong, J.I.: Using harmonic analysis and optimization to study macromolecular dynamics. International Journal of Control Automation and Systems 4(3), 382–393 (2006)

[10] Krislock, N.: Semidefinite facial reduction for Low-Rank Euclidean Distance Matrix Completion. Ph.D. thesis, School of Computer Science, University of Waterloo (2010)

[11] Li, X., Burkowski, F.: Generating conformational transitions using the Euclidean distance matrix. IEEE transactions on nanobioscience (2015)

[12] Li, X.B., Burkowski, F.J.: Conformational transitions and principal geodesic analysis on the positive semidefinite matrix manifold. In: Basu, M., Pan, Y., Wang, J. (eds.) Bioinformatics Research and Applications, Lecture Notes in Computer Science, vol. 8492, pp. 334–345. Springer International Publishing (2014)

[13] Meyer, G.: Geometric optimization algorithms for linear regression on fixed-rank matrices. Ph.D. thesis, University of Liège (2011)

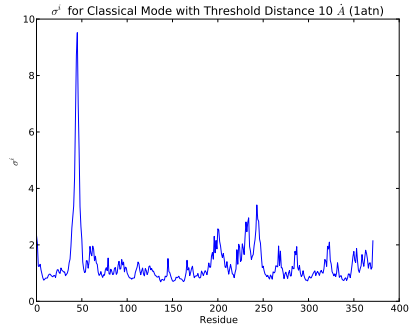
[14] Mishra, B., Meyer, G., Sepulchre, R.: Low-rank optimization for distance matrix completion. In: 2011 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC). pp. 4455–4460. Orlando, FL, USA (December 12-15 2011)

[15] Tirion, M.: Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Physical Review Letters 77(9), 1905–1908 (August 1996)

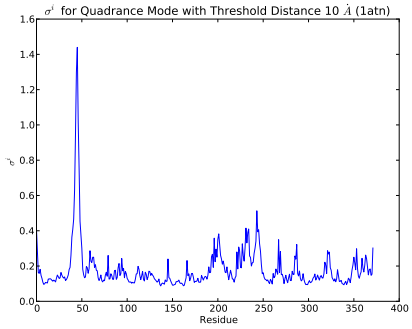
[16] Tirion, M.M., ben-Avraham, D.: Normal mode analysis of g-actin. Journal of molecular biology 230(1), 186–195 (1993)

[17] Vandereycken, B.: Riemannian and multilevel optimization for rank-constrained matrix problems. Ph.D. thesis, Department of Computer Science, KU Leuven (2010)

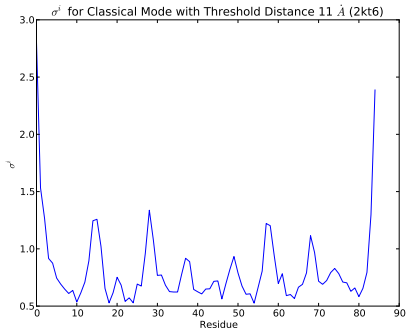
[18] Wildberger, N.J.: Divine Proportions: Rational Trigonometry to Universal Geometry. Wild Egg (2005)



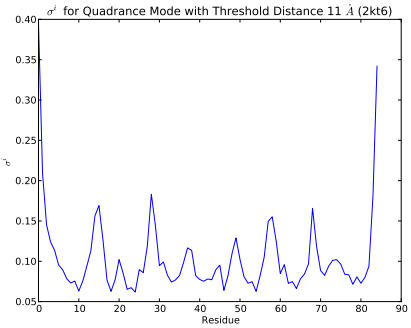
(a) 1ATN σ^i for classical mode.



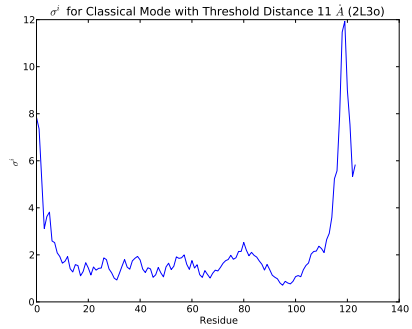
(b) 1ATN σ^i for quadrance mode.



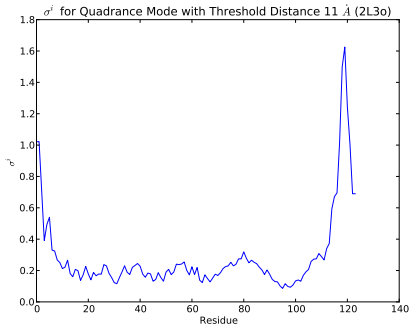
(c) 2KT6 σ^i for classical mode.



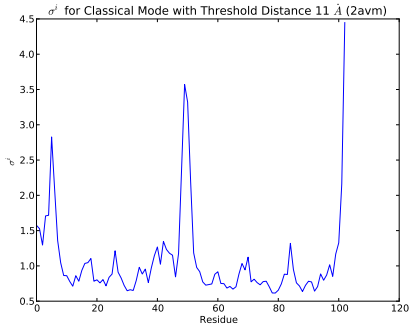
(d) 2KT6 σ^i for quadrance mode.



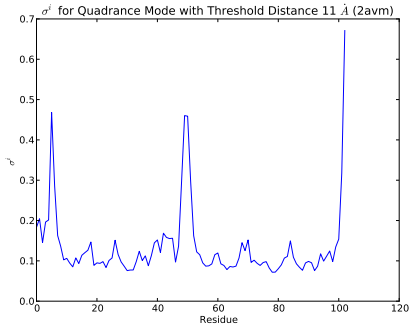
(e) 2L3o σ^i for classical mode.



(f) 2L3o σ^i for quadrance mode.



(g) 2AVM σ^i for classical mode.



(h) 2AVM σ^i for quadrance mode.

Figure 4: Quadrance NMA reproduces the σ^i seen in classical NMA