

‘Bernoulli Cubes’: measuring effective dimensionality over a binary-partitioned space

Vincent Voelz

May 22, 2003

Abstract

In the case of a high-dimensional space that is described in terms of binary partitions, the N-Cube Analysis (NCA) method of Sullivan (2001) reduces to a formulation we will call Bernoulli-Cube Analysis (BCA). The use of BCA on a test problem is demonstrated, and implications for measuring the dimensionality of dynamics over a set of *contact states* are discussed.

1 Introduction

For trajectories in a continuous high-dimensional space, Sullivan (1999) proposed a method called N-Cube Analysis (NCA) to measure the effective dimensionality of a trajectory sampling. In this method, it is assumed that over the time scale of the trajectory, the trajectory is sufficiently ergodic to approximate a uniform sampling of some effective dimensional-space.

The NCA method works as follows. Suppose each point in the trajectory sample can be described as a ν -dimensional coordinate vector $\mathbf{x}_i = (x_1, x_2, \dots, x_\nu)_i$. For each point, there is a corresponding Euclidean squared-distance D_i defined as $\mathbf{x}_i \cdot \mathbf{x}_i$.

Suppose that the underlying distribution of \mathbf{x} is uniform over a hypercube of dimension N and edge length A . Then the distribution of D will have a characteristic mean M and variance V (Sullivan 1999):

$$M = \langle D \rangle = \frac{N(A^2)}{6} \quad (1)$$

$$V = \langle D^2 \rangle - \langle D \rangle^2 = \frac{7N(A^4)}{180} \quad (2)$$

These equations can be solved to find N and A in terms of a measured M and V :

$$N = \frac{7M^2}{5V} \quad (3)$$

$$A = \sqrt{\frac{30V}{7M}} \quad (4)$$

Thus, for values of M and V that have been measured from a trajectory sample, there is an effective ‘N-Cube Analysis’ (NCA) dimension N , and edge length, A to describe the size of the space sampled by the trajectory. The effective NCA ‘volume’ is A^N .

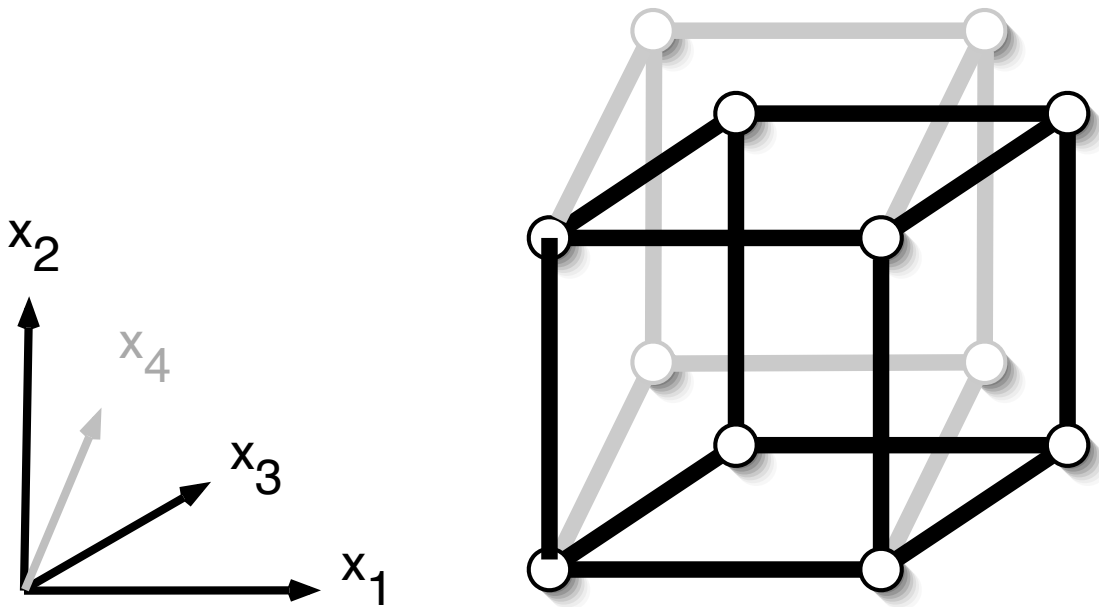


Figure 1: The space of *contact states* can be visualized as the corners of a unit hypercube.

2 Bernoulli Cube Analysis (BCA)

Let us now extend this idea to discrete binary variables, also known as Bernoulli variables. Such variables may be useful to describe *contact states* of proteins. If there are a total of nu possible pairwise inter-residue contacts, then a point in contact space can be represented by a ν -dimensional vector $\mathbf{x} = (x_1, x_2, \dots, x_\nu)_i$, where each component represents the presence of a pairwise contact. These components can either be 1 (present) or 0 (not present), i.e. a Bernoulli variable. If there are ν such contact variables, then the number of potential contact states is 2^ν .

The space of Bernoulli variables can be visualized as the corners of a unit hypercube. Whereas in NCA we assumed the cube had edge length A in all directions, here we assume that the corners of the unit hypercube are populated with probabilities $(1 - p)$ at 0, and p at 1, along every direction.

As before, we consider the mean and variance of the the Euclidean squared distance, D , in order to derive a relationship between the effective dimensionality N , and the effective probability, p .

Consider the Bernoulli variables x_1, x_2, \dots, x_N , where x_k is 1 with probability p , and 0 with probability $(1 - p)$. Then

$$D = \sum_k^\nu x_k^2 = \sum_k^\nu x_k$$

The mean and variance of D are the well-known expectation values from the binomial distribution:

$$M = E(D) = \sum_D D p(D) = \sum_D D \binom{N}{D} p^D p^{N-D} = Np$$

$$V = E((D - \langle D \rangle)^2) = \sum_D (D - \langle D \rangle)^2 \binom{N}{D} p^D p^{N-D} = Np(1 - p)$$

We can solve N and p in terms of M and V :

$$p = 1 - \frac{V}{M}$$

$$N = \frac{M}{p} = \frac{M^2}{M - V}$$

Unfortunately, we may obtain very high values of N and simultaneously values of p very close to 0 or 1. This is a rather unintuitive way to think about dimensionality, because we may have the situation that the dimensionality is large, yet most of the trajectory sample resides in one corner of the hypercube.

A more intuitive way to assess the dimensionality is by assuming there is a uniformly populated N -dimension subspace in the full ν -dimensional space, and asking what the effective value of N must be to explain the mean and variance. In this case, the subspace dimensions would have $p = (1 - p) = 1/2$, and

$$N = 4V$$

We call this N the ‘Bernoulli-Cube Analysis’ dimension, or BCA dimension.

3 Test Problem: a sampled elliptical trajectory

Here we present a simple test problem to test how well the BCA dimension captures effective dimensionality, compared to the NCA dimension. An elliptical trajectory

$$\begin{pmatrix} 2 \cos(2\pi t) \\ \sin(2\pi t) \end{pmatrix}$$

is sampled over different intervals $[0, t]$ for $t \in [0, 1]$. The NCA and BCA dimensions were calculated as described above, as a function of sampling time. The set of macrostates used for the BCA were the four quadrants of the 2-D plane.

The elliptical trajectory does not uniformly sample the 2-D plane, so varying amounts of normally-distributed random noise was added to the trajectory to emulate this. It was expected that NCA would perform better as the distribution of points in the 2-D became more uniform.

3.1 Results

Figures 1-3 show the elliptical trajectory sampled with varying levels of added noise ($\sigma = 0.0, 0.2, 0.8$). As the sampling time increases, the BCA dimension should go through 0 (a single macrostate ‘point’), through 1 (two macrostates forming a ‘line’), finally to 2.

Figure 4 shows the NCA dimension and BCA dimension calculated as a function of sampling time for various levels of noise. The observed BCA dimension reproduces what we should expect: transitioning between 0 to 1 to 2 as sampling increases. As noise is introduced, more macrostates are sampled, and the BCA dimension transitions between ~ 1 to 2.

Contrast this with the NCA dimension, which only reproduces the correct dimensionality in the limit of large amounts of noise.

4 The dimensionality of the microsecond villin headpiece trajectory over contact space

20 ps snapshots of the $1\mu s$ villin trajectory of Duan and Kollman (1999) were binned into hydrophobic contact states (using Chodera’s method). 107 types of pairwise contacts were identified, and the trajectory was translated into a 107-dimensional contact state trajectory.

Figure 6 shows the application of the BCA method, as well as the *naive* application of the NCA method, to the contact state trajectory, using a sampling window of 20 ns over time. Figure 7 shows the same analysis but for a 2 ns sampling window over time. Figure 8 shows a 0.2 ns sampling window.

The BCA dimension is much smaller than the full 107 contact dimensions, depending on what time scale is used to examine it. A ten-fold reduction in the size of the sampling window results in approximately halving the effective dimensionality. The number of contacts made and unmade on the time scale of the sampling window can be examined this way. It is interesting that on the time scale of 200 ps (Figure 8), it appears that about 5 to 15 contact states are accessible. The error in this calculation may be high, however, because at this time scale, only 10 points make up each time window's sample.

It is unclear how to interpret the negative effective Bernoulli probabilities at ~ 500 ns in Figures 6 and 7. For some reference to the original data, Figures 9 and 10 are provided from the Sullivan paper: the NCA dimension for the continuous trajectory coordinates, and the trajectory in a principal component (PCA) analysis.

5 References

1. David C. Sullivan and Irwin D. Kuntz. Conformational Spaces of Proteins. *PROTEINS: Structure, Function, and Genetics* 42:495-511 (2001)
2. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*. 1998 Oct 23;282(5389):740-4.

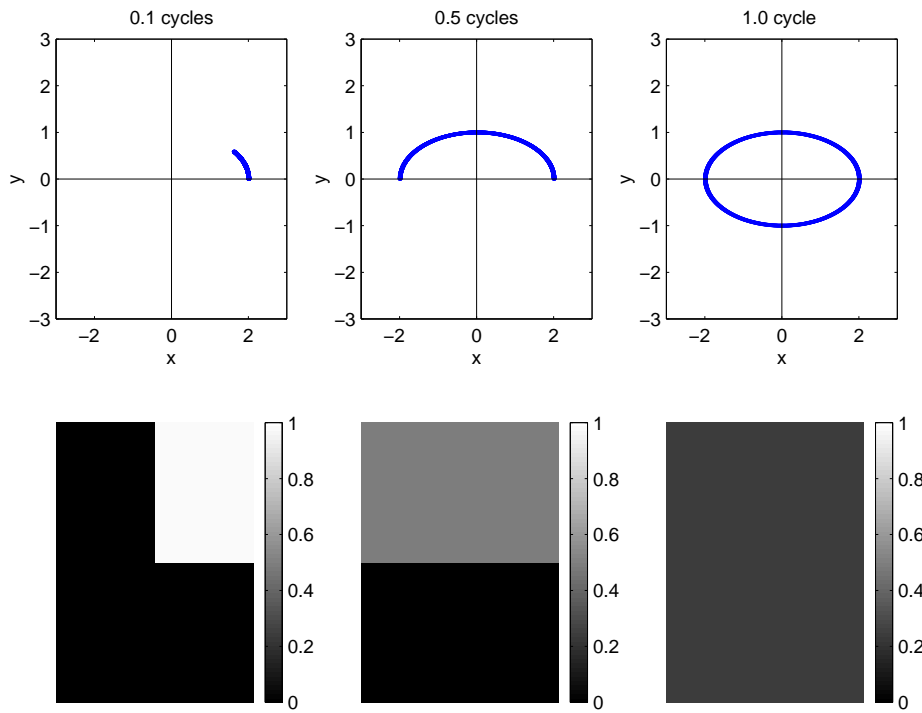


Figure 2: An elliptical trajectory sampled at different time scales. **Above:** The trajectory sample shown in a continuous variable representation. **Below:** The trajectory sample shown as a density of macrostates, where each macrostate is a quadrant of the graph.

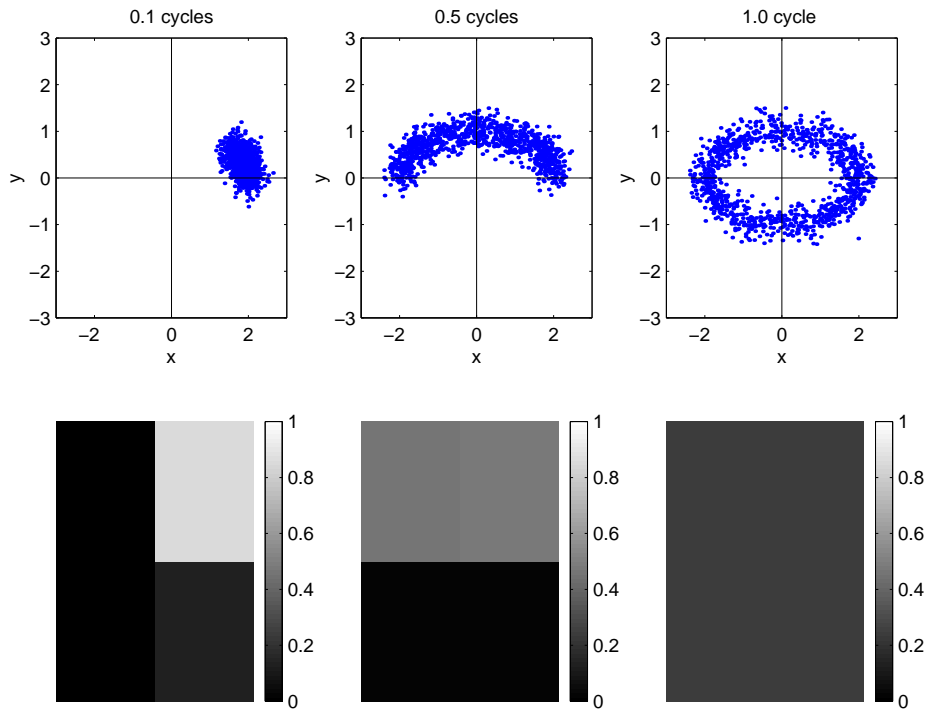


Figure 3: An elliptical trajectory with normally-distributed noise ($\sigma = 0.2$) sampled at different time scales. **Above:** The trajectory sample shown in a continuous variable representation. **Below:** The trajectory sample shown as a density of macrostates, where each macrostate is a quadrant of the graph.

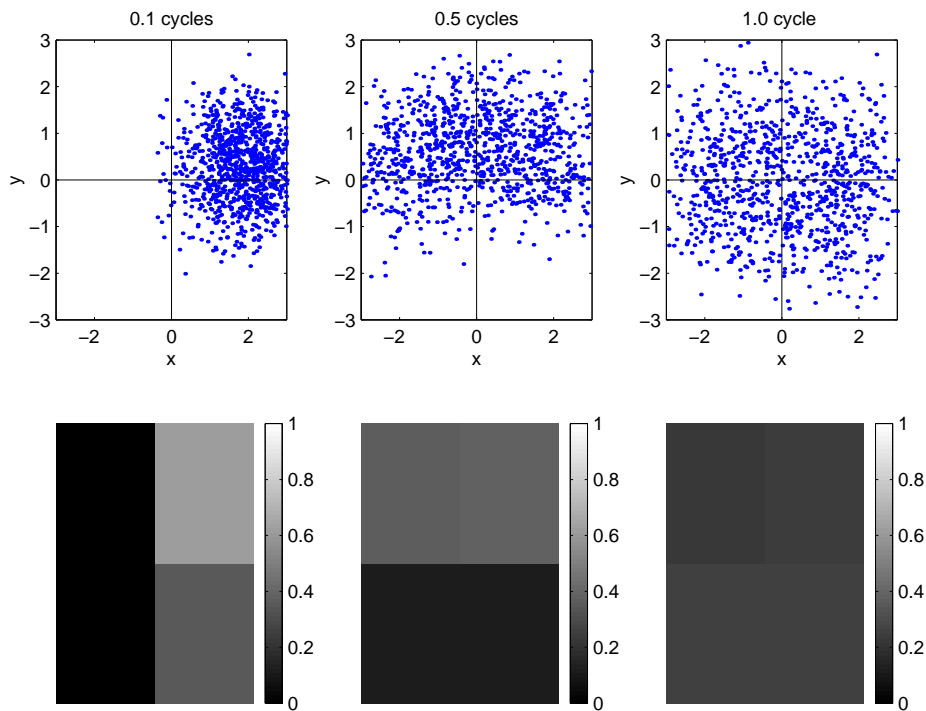


Figure 4: An elliptical trajectory with a *lot* of normally-distributed noise ($\sigma = 0.8$) sampled at different time scales. **Above:** The trajectory sample shown in a continuous variable representation. **Below:** The trajectory sample shown as a density of macrostates, where each macrostate is a quadrant of the graph.

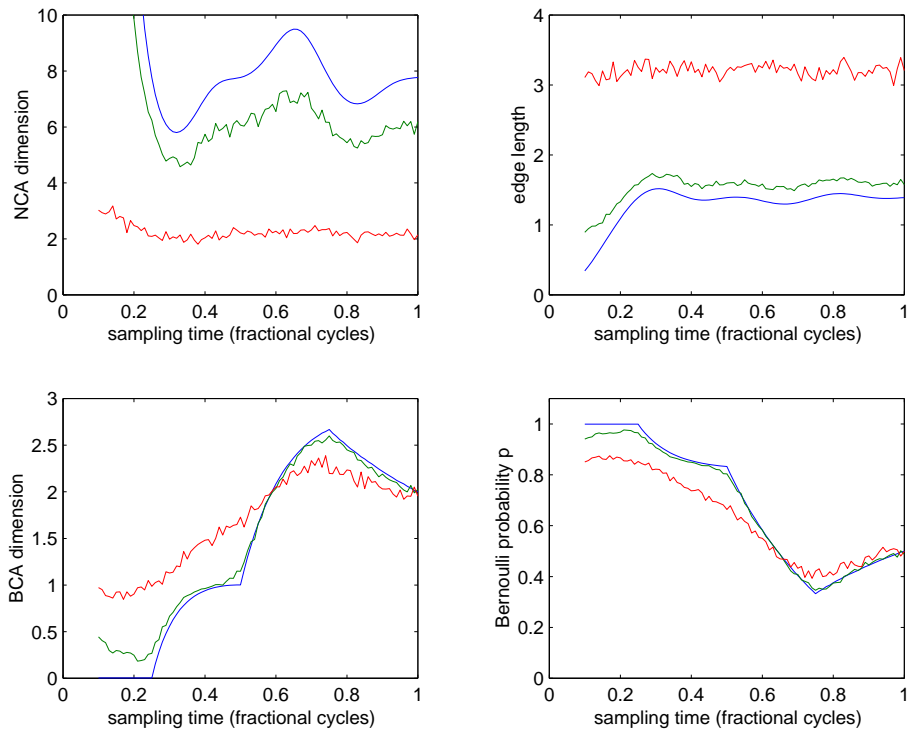


Figure 5: Dimensional analysis for trajectories as a function of sampling time. The blue, green, and red lines are for trajectories with noises $\sigma = 0.0, 0.2$, and 0.8 , respectively. **Above:** The NCA dimension and edge length calculated as a function of trajectory sampling time **Below:** The BCA dimension and average Bernoulli probability p calculated as a function of trajectory sample time.

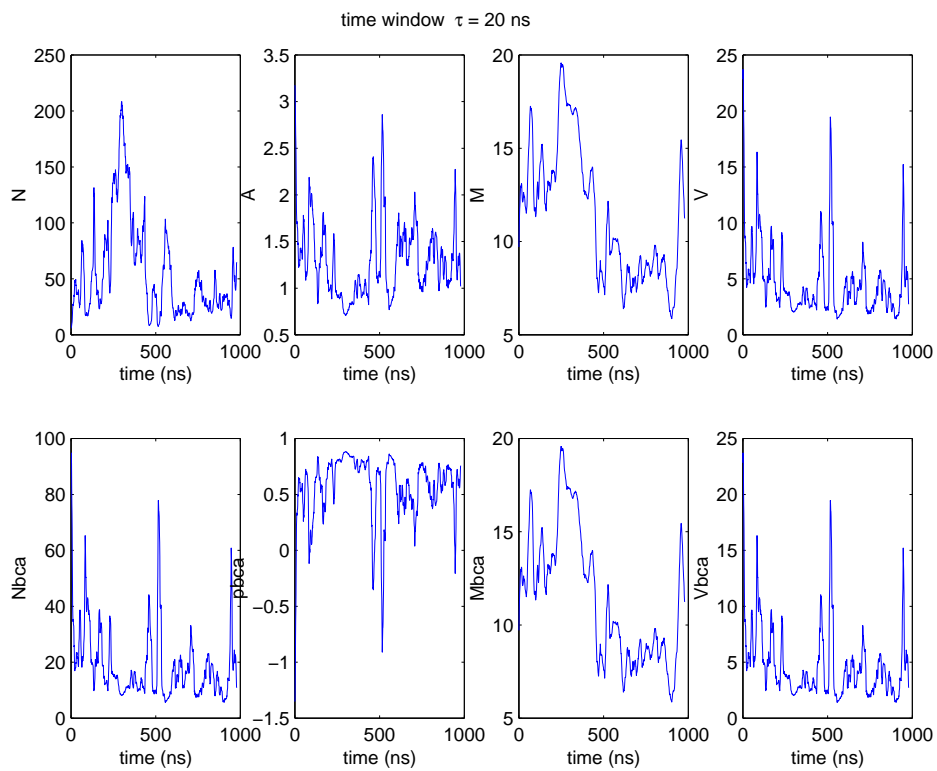


Figure 6: Dimensional analysis for the microsecond villin headpiece trajectory over the 107-dimension hydrophobic contact space, as a function of time, with a sampling window of 20 ns. **Top row:** The NCA dimension, edge length, mean and variance, naively calculated over over contact space. **Bottom row:** The BCA dimension, average Bernoulli probability, mean and variance, calculated as a function of trajectory sample time.

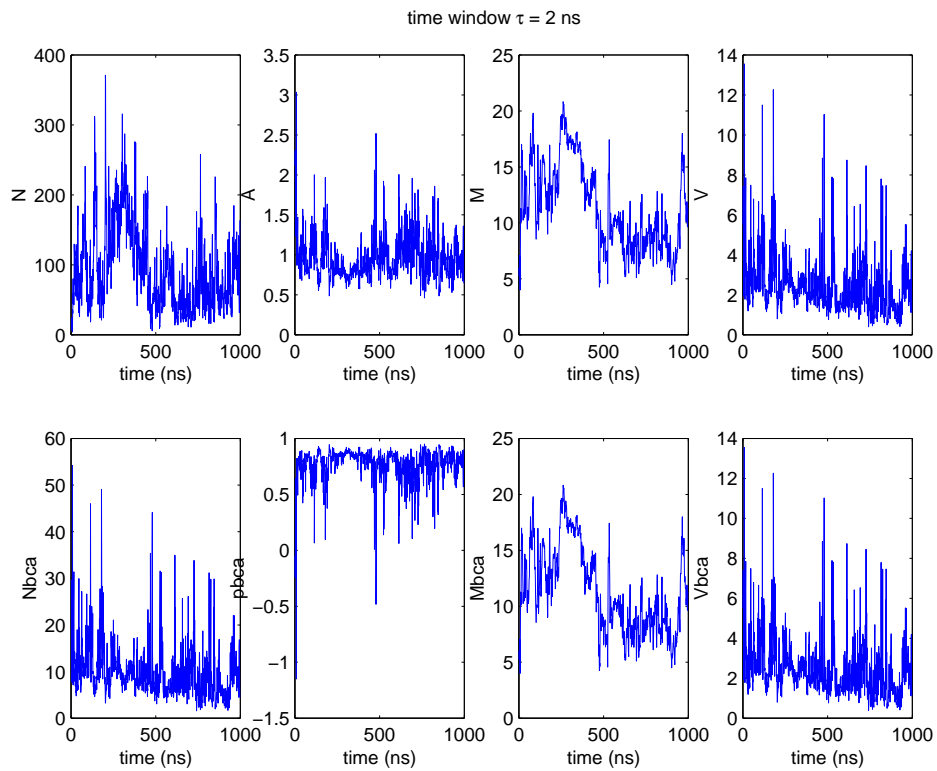


Figure 7: Dimensional analysis for the microsecond villin headpiece trajectory over the 107-dimension hydrophobic contact space, as a function of time, with a sampling window of 2 ns. **Top row:** The NCA dimension, edge length, mean and variance, naively calculated over over contact space. **Bottom row:** The BCA dimension, average Bernoulli probability, mean and variance, calculated as a function of trajectory sample time.

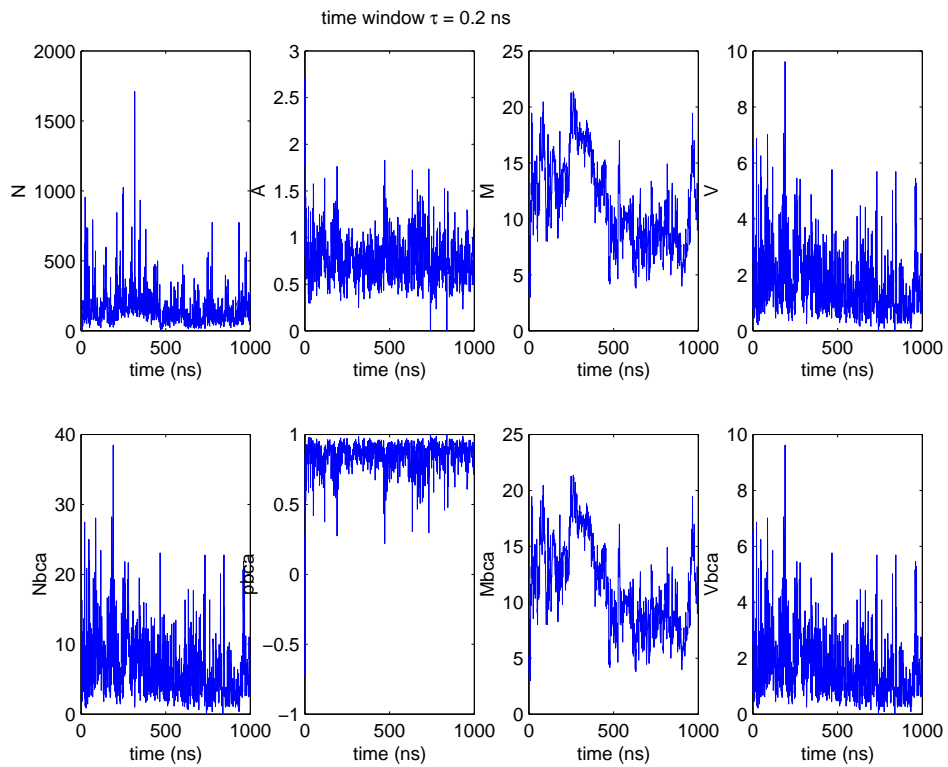


Figure 8: Dimensional analysis for the microsecond villin headpiece trajectory over the 107-dimension hydrophobic contact space, as a function of time, with a sampling window of 0.2 ns. **Top row:** The NCA dimension, edge length, mean and variance, naively calculated over over contact space. **Bottom row:** The BCA dimension, average Bernoulli probability, mean and variance, calculated as a function of trajectory sample time.

1 Microsecond Villin Folding Trajectory

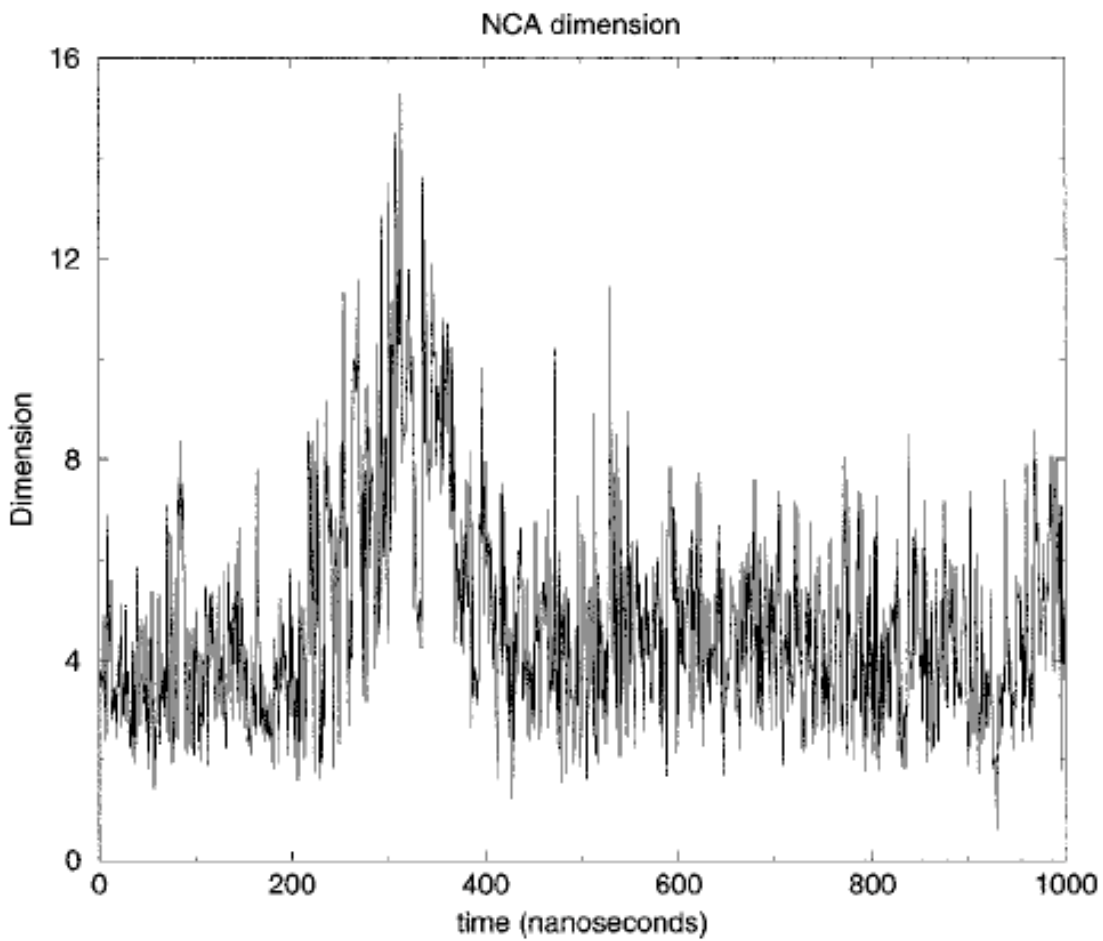


Figure 9: N-cube analysis for the villian headpiece trajectory calculated over continuous variables, over time with a 1 ns time window. (Figure taken from Sullivan 2001)

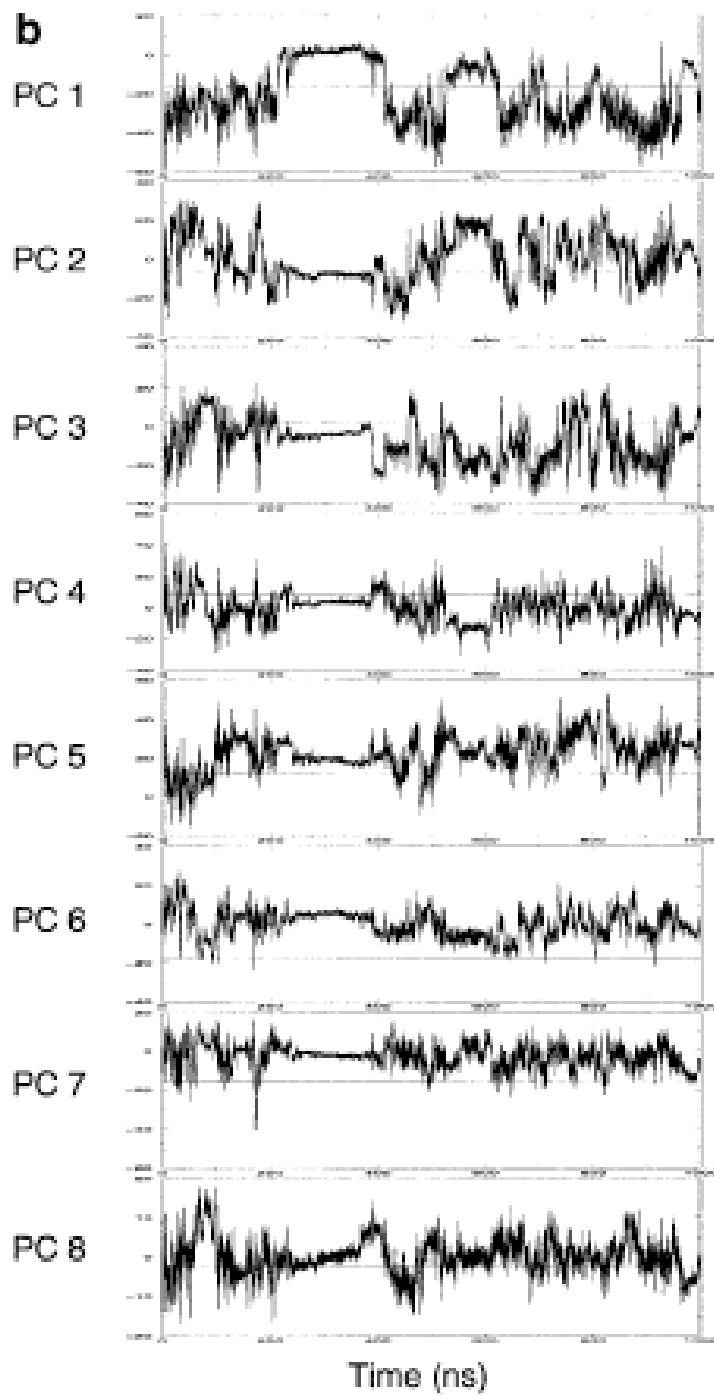


Figure 10: Projections of the villin trajectory onto the dominant principal components of motion for the entire trajectory. (Figure taken from Sullivan 2001).