Contents

1 Radial Density Function

1.1 Calculation of Distances with Periodicity

Suppose a large chemical structure has uncountably many atoms but the follow a periodic pattern of n atoms every p Angstroms. The atom locations within a period are given by $a_1, a_2, \ldots, a_n$ where $a_i \in \mathbb{R}^3$. The radial density function is the distribution of pairwise distances between these atoms.

The distances d between atoms a_i and a_j where $i \neq j$, atom a_i has been displaced by x, and atom a_j has been displaced by y per the periodicity is

$$
\begin{aligned}
d^2 &= \langle a_i + x - (a_j + y), a_i + x - (a_j + y) \rangle \\
&= \langle a_i - a_j, a_i - a_j \rangle + \langle x - y, x - y \rangle + 2\langle a_i - a_j, x - y \rangle
\end{aligned}
$$

where $x = (k_1 p, k_2 p, k_3 p)$ for $k_i \in \mathbb{Z}$ and $y = (l_1 p, l_2 p, l_3 p)$ for $l_i \in \mathbb{Z}$. Here $\langle x, y \rangle$ denotes the inner product between x and y.

Suppose D is a random variable that samples at random the distances, d, in the chemical structure. The radial density function is the probability density function of this random variable. This function can be estimated empirically via a histogram.

The histogram is then normalized by the volume a spherical shell.

$$
\begin{aligned}
\frac{4}{3}\pi(r + \Delta r)^3 &- \frac{4}{3}\pi r^3 \\
&= \frac{4}{3}(3r^2 \Delta r + 3r(\Delta r)^2 + (\Delta r)^3) \\
&\approx 4\pi r^2 \Delta r
\end{aligned}
$$

where Δr tends to zero.

For a histogram with frequency, f, for bin $[d_i, d_{i+1}]$, we replace f with f/d_i^2. And then normalize the histogram so that the sum over all bins is one.

1.2 Adding Noise For Atom Vibration

Due to the vibrations of the molecules, the radial density function will not be just the equilibrium positions. We can approximate this fluctuation in distances via a Gaussian filter or Weierstrass transform.

$$F(x) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} f(y) e^{-\frac{(x-y)^2}{4t}} dy$$

Given that the density function is only defined for a finite number of distances, we use a discrete version of the transform making sure to keep the sum of the weights equal to one.

$$F(d_k) = \frac{\sum_{d_i=d_0}^{d_n} f(d_i) \exp\left(-\frac{(d_k-d_i)^2}{4t}\right)}{\sum_{d_i=d_0}^{d_n} \exp\left(-\frac{(d_k-d_i)^2}{4t}\right)}$$

where d_0 is the minimum distance and d_n is the maximum distance.

1.3 Cubane Example

As an example of the above, below are the calculations for cubane (C_8H_8).

Here are the coordinates of the elements in cubane in Angstroms.

```
Element, x, y, z
C, 1.2455, 0.5367,-0.0729
C, 0.9239,-0.9952, 0.0237
C,-0.1226,-0.7041, 1.1548
C, 0.1989, 0.8277, 1.0582
C, 0.1226, 0.7042,-1.1548
C,-0.9239, 0.9952,-0.0237
C,-1.2454,-0.5367, 0.0729
C,-0.1989,-0.8277,-1.0582
```

```
H, 2.2431, 0.9666,-0.1313

H, 1.6638,-1.7924, 0.0426

H,-0.2209,-1.2683, 2.0797

H, 0.3583, 1.4907, 1.9059

H, 0.2208, 1.2681,-2.0799

H,-1.6640, 1.7922,-0.0427

H,-2.2430,-0.9665, 0.1313

H,-0.3583,-1.4906,-1.9058
```
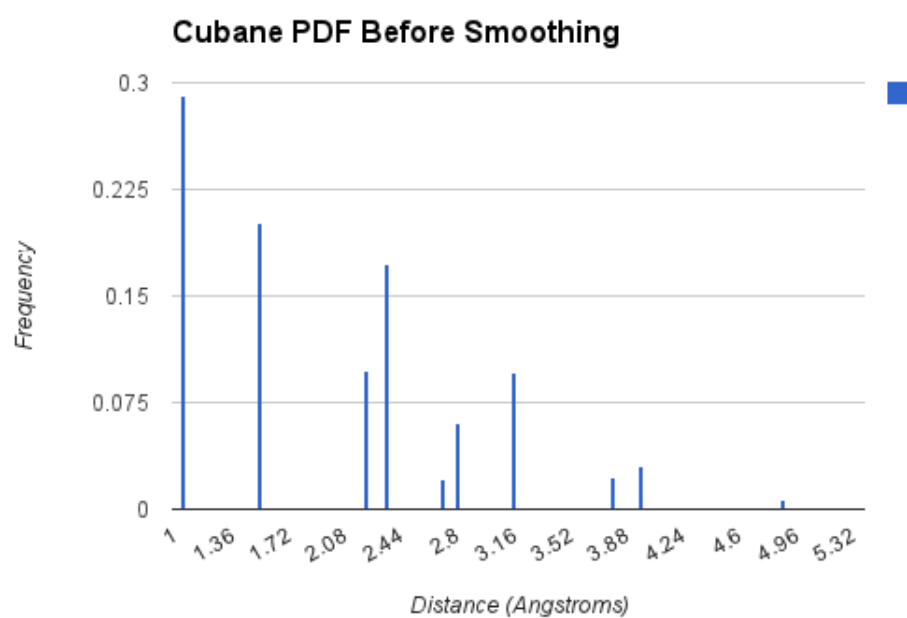
1.3.1   Cubane Radial Density Functions
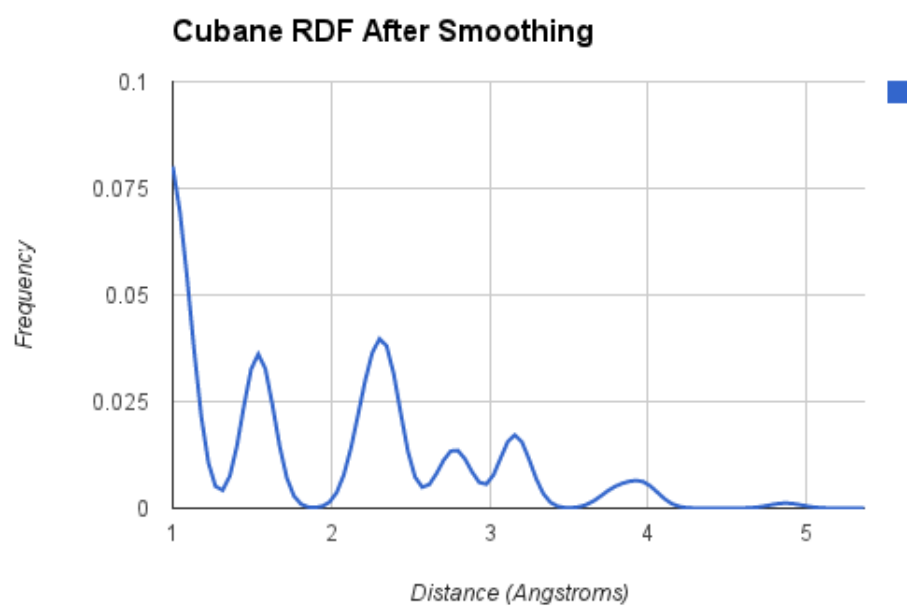

Figure 1: Before Smoothing

Figure 2: After Smoothing

1.4   Experimental and Theoretical RDFs for Known Structures

For some structures, we are able to theoretically calculate the RDF from atom locations and also have the experimental RDF from Xray scattering. These known matches provide some insight into understanding how the experiments and theory align. The RDF comparison are shown below.

Outside of these structures, there are not many other known matches. There are a few reasons for this. First, if a structures is already known at the atomic level then there is no need to run an xray diffraction experiment. Second, if a structure is periodic as in a lattice, the atomic structure can be determined by xray diffraction which is easier and cheaper than xray scattering.

1.4.1   Ga As

Experimental Data: Pair Distribution Functions Analysis, Valeri Petkov
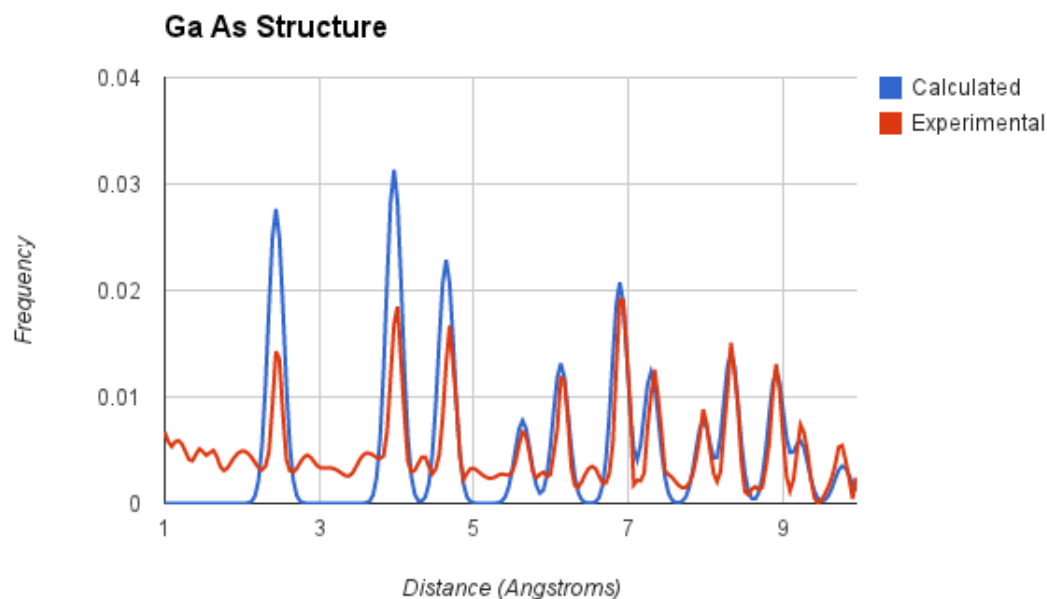Calculated Data: Maria Chan

Figure 3: Ga As

1.4.2   In As

Experimental Data: Pair Distribution Functions Analysis, Valeri Petkov
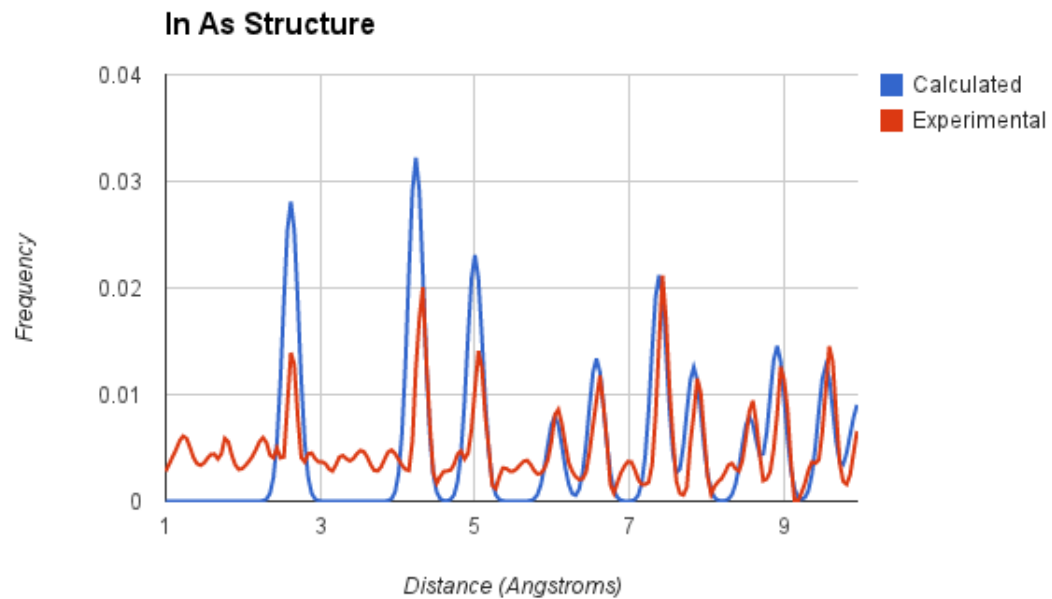
Calculated Data: Maria Chan

Figure 4: In As

1.4.3   Si Lattice

Experimental Data: J. AM. CHEM. SOC. VOL. 133, NO. 3, 2011, P: 503-512
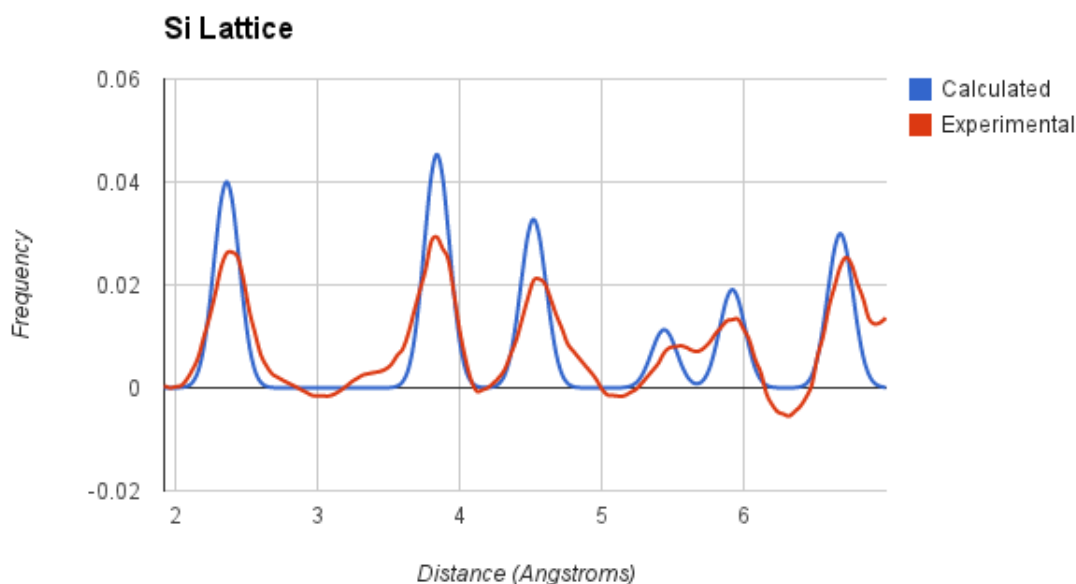
Calculated Data: http://materialsproject.org/materials/mp-149/

Figure 5: Si Lattice

2 Principal Component Analysis

2.1 Data Set

For the analysis following, a data set of radial density functions was used that contained 3,491 theoretical SiLi structures, 8 experimental SiLi structures, a pair of theoretical and calculated GaAs structures, and a pair of theoretical and calculated InAs structures. Each image had 128 evenly distances from 1.92 to 7 angstroms.

2.2 Dimensionality

To discover the minimal dimensionality of the RDF data, I charted the cumulative proportion of variance explained by adding successive principal components. We can see from Figure 6 that around 30 principal components explain 99% of the variance.
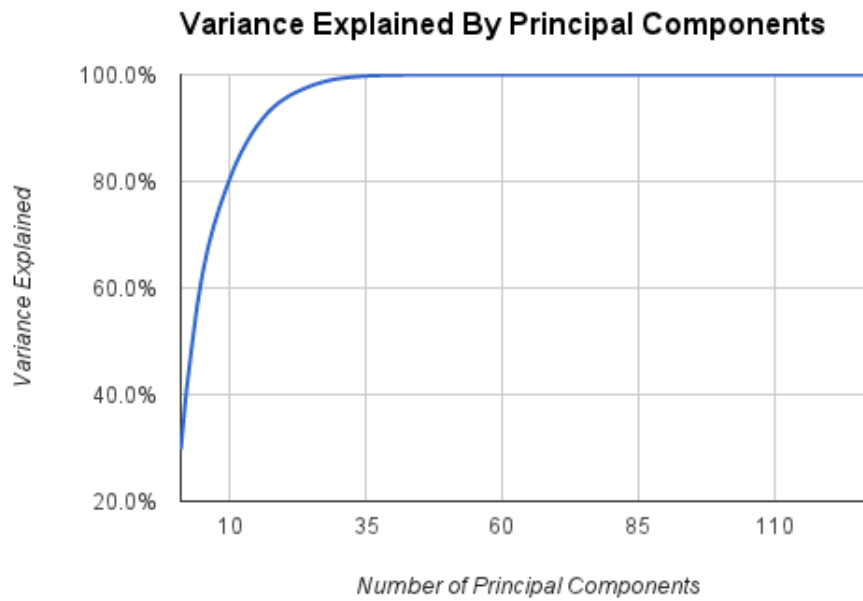
Figure 6: Proportion of Variance Explained

2.3   Basis Vectors

Sometimes PCA analysis gives intelligible basis vectors that identify a key characteristic in the data set. In the case of the RDF images, the basis vectors appear to be nonsensical. The first four principal component basis vectors are show in Figure 7.
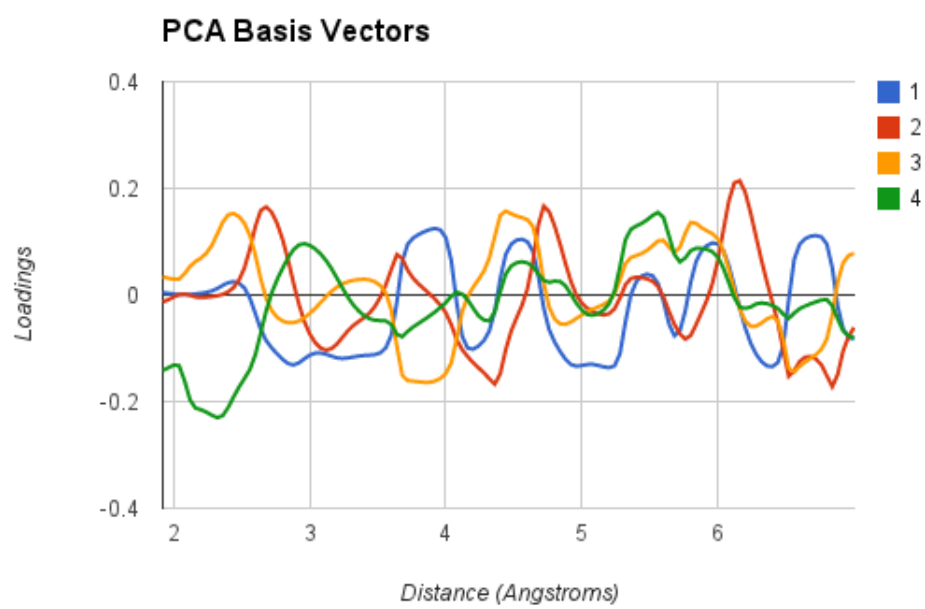
Figure 7: PCA Basis Vectors

2.4 Projected RDFs

To further assess the capacity of PCA to reduce the dimensionality of the data, I sampled a few images and projected them onto PCA space with decreasing dimensions.
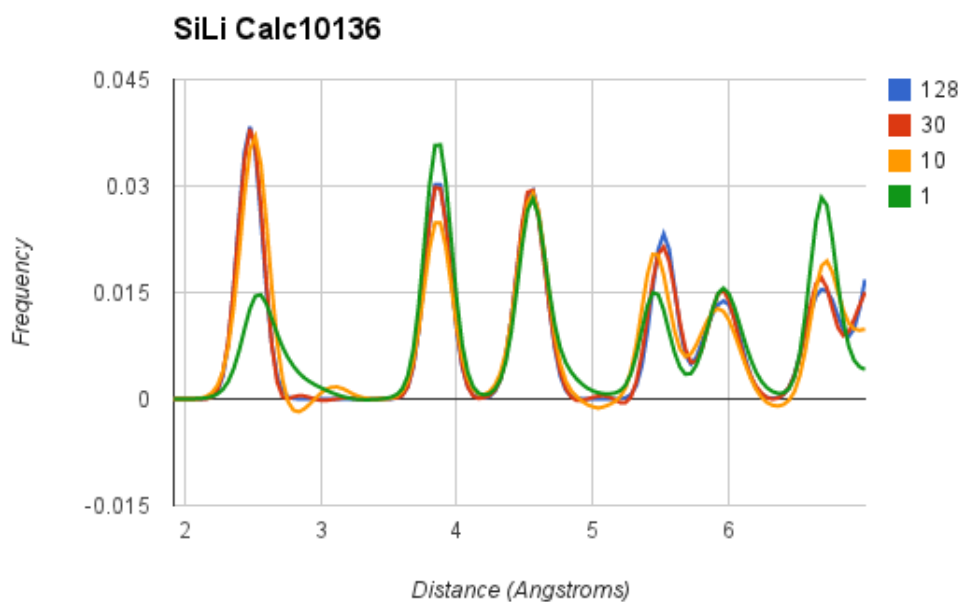
2.4.1 SiLi Calc10136

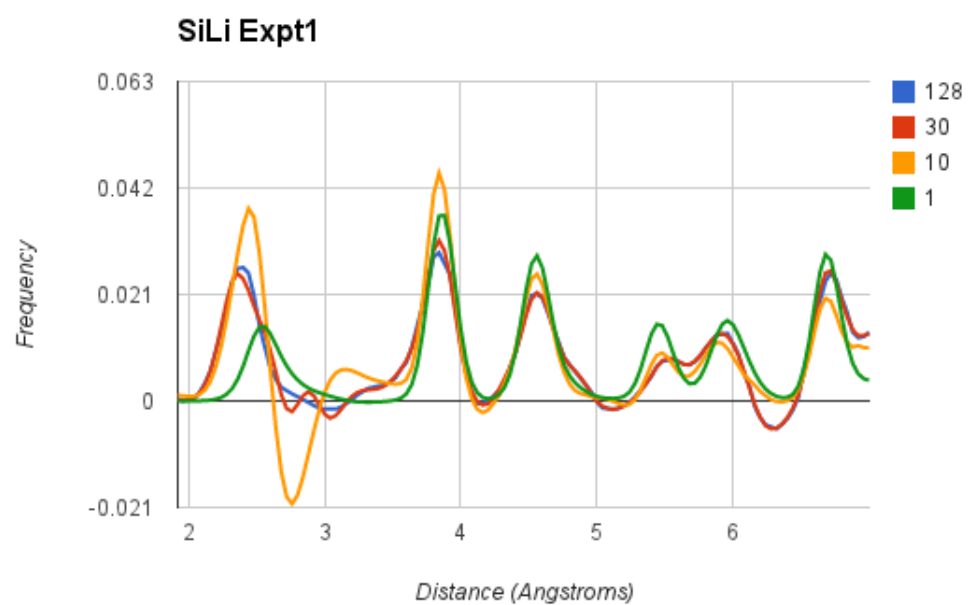Figure 8: SiLi Calc10136 Projections

2.4.2 SiLi Expt1

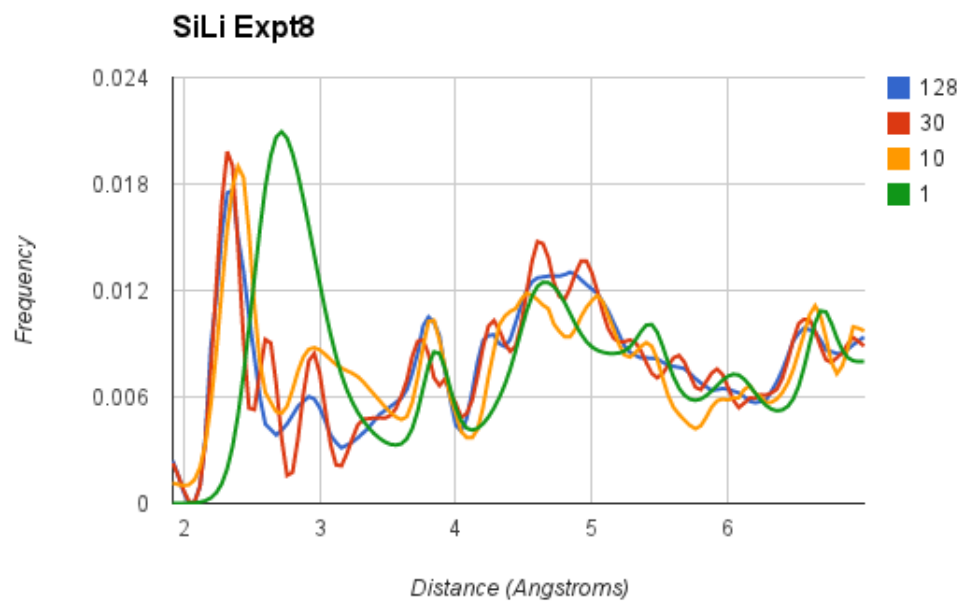Figure 9: SiLi Expt1 Projections

2.4.3 SiLi Expt8

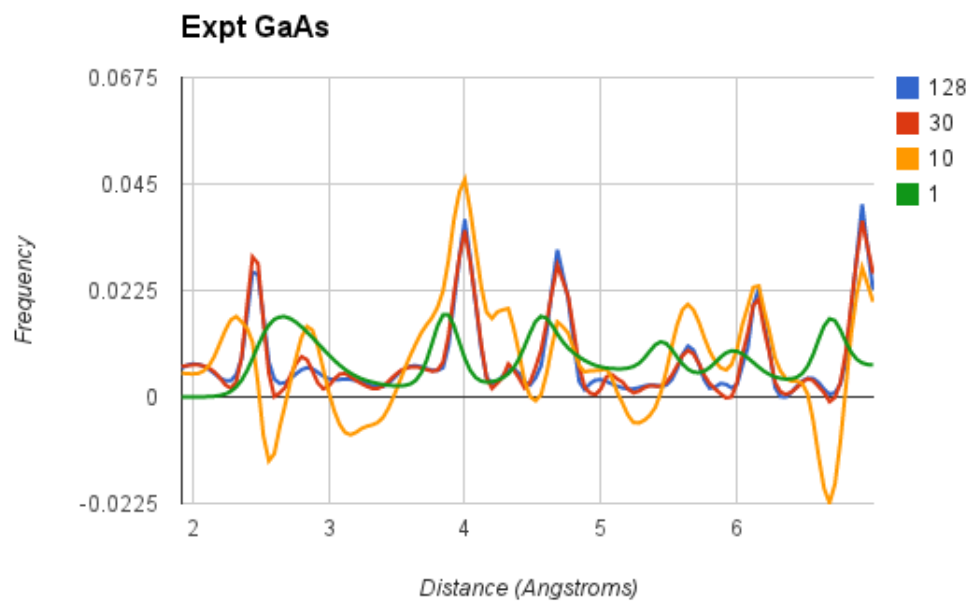Figure 10: SiLi Expt8 Projections

2.4.4 GaAs Expt

Figure 11: GaAs Expt Projections

2.5 Observations

From the proportion of variance explained and the selected projections, clearly thirty principal components are sufficient to capture the essence of the image. Also, it shows that one principal component is not sufficient in most cases. It is strange however that the first principal component captures so clearly SiLi Calc10136. Further investigation is needed to determine whether this is due the large number of images similar to Calc10136 in the data set.
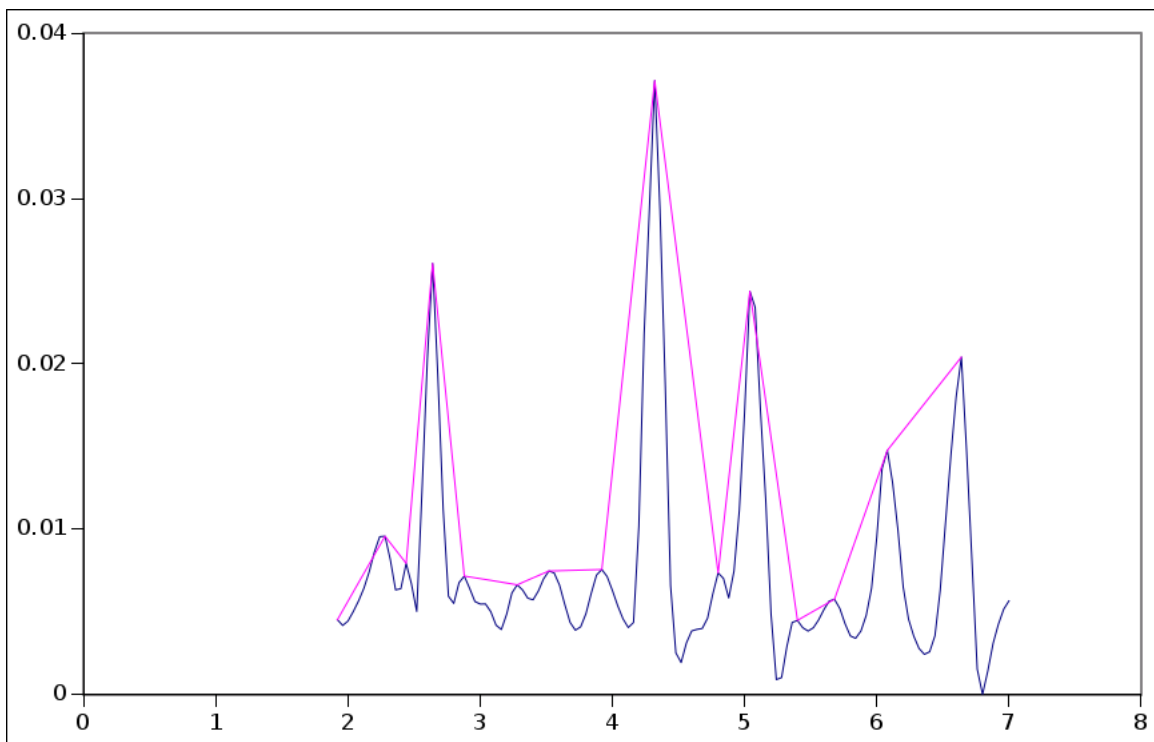
3 Noise Analysis

3.1 Peak Counts

Figure 12: InAs Expt, Max 7 Angstroms
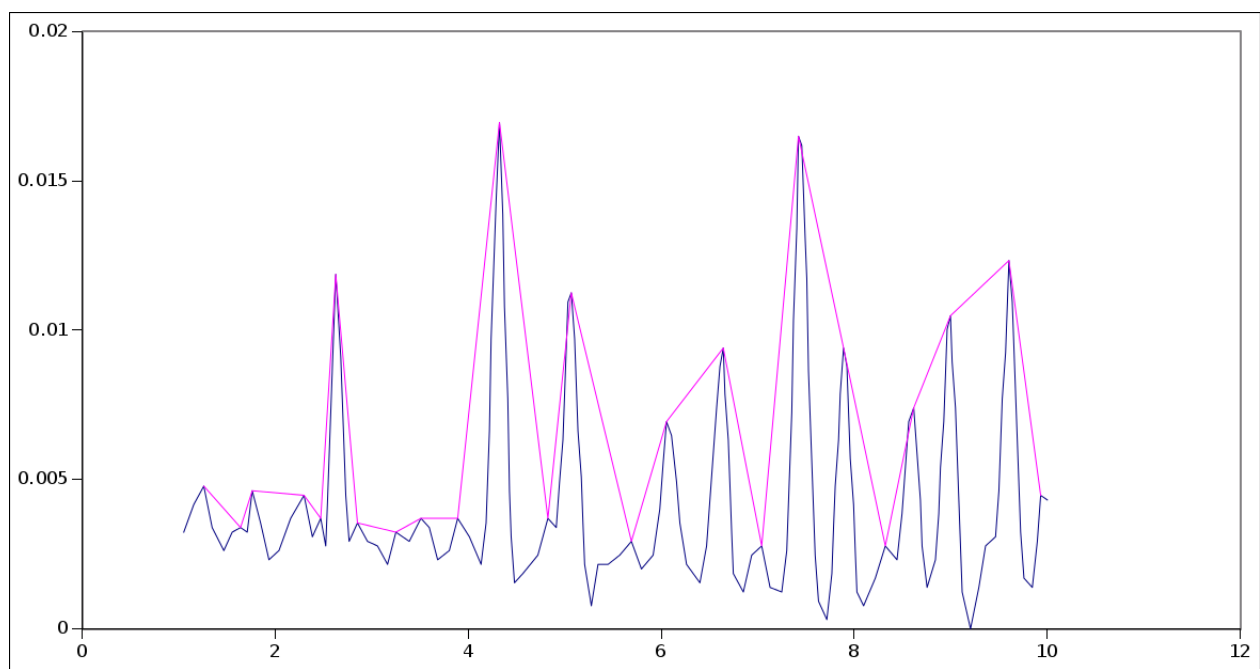
Figure 13: InAs Expt, Max 10 Angstroms

4 Sources

http://en.wikipedia.org/wiki/Atom_vibrations

http://en.wikipedia.org/wiki/Radial_distribution_function

http://en.wikipedia.org/wiki/Weierstrass_transform

http://matplotlib.org/api/mlab_api.html