Abstract

We here develop a general framework for the residue fluctuations that simultaneously incorporates both anharmonicity and mode-coupling in a unified formalism. We show that both deviations from the pure Gaussian model are important for modeling the multidimensional energy landscape in the vicinity of the native state, even at physiological conditions where fluctuations are relatively small. (at what temperature is the MD?)

Anharmonicity and mode-coupling in a fluctuating protein

April 20, 2010

1 Introduction

Residue fluctuations of a protein around its native state reveal information that bridges the molecule's structural and functional properties. At the lowest order, these fluctuations can be treated as a collection of independent harmonic modes, yielding "Elastic Network Models" [1]. However, it was recently shown that the slowest oscillatory modes of a protein are strongly anharmonic [2], in contrast with the assumption underlying ENMs. The coupling between different modes is another aspect of protein dynamics that is believed to be relevant for the information/energy transfer between different parts of the molecule [3] and not captured by the harmonicity assumption...

2 Equilibrium fluctuations

2.1 Beyond Gaussian: the Hermite expansion

Sampling the time evolution of a protein by using molecular dynamics reveals a multivariate probability distribution function $f(\mathbf{\Delta R})$ for the deviations of atoms (assume there are N of them) from equilibrium coordinates, i.e. $\Delta R_i = R_i - R_i^{eq}$, $i = 1, \ldots, 3N$. We here adopt a coarse-grained representation of this p.d.f., where only C_α atoms are considered; i.e., N is the number of residues. Since the deviations from the free energy minimum should be harmonic for sufficiently small amplitudes, Hermite polynomials, which have a Gaussian kernel, constitute a natural basis for representing $f(\mathbf{\Delta R})$. First, following Ref. [2], we perform the transformation

$$\mathbf{\Delta r} \quad = \quad \langle \mathbf{\Delta R \Delta R}^T \rangle^{-1/2} \mathbf{\Delta R} \tag{1}$$

which diagonalizes the covariance matrix $Gamma \equiv \langle \mathbf{\Delta R \Delta R}^T \rangle$ and would give the normal modes of the protein if fluctuations were harmonic. Otherwise, the distribution function for $\{\Delta r\}$ in its most general form, can be expressed as [4]

$$f(\mathbf{\Delta r}) \quad = \quad \frac{1}{\sqrt{(2\pi)^{3N}}} e^{-\sum_{i=1}^{3N} \Delta r_i^2/2} \left[1 + \sum_{\nu=3}^{\infty} \mathbf{C}_\nu \cdot \mathbf{H}_\nu(\mathbf{\Delta r}) \right] \tag{2}$$

where $\mathbf{C}_\nu$ (constant) and $\mathbf{H}_\nu$ (derived below) are tensors of rank ν, and the dot product refers to $\sum_{ij..k} C_\nu^{ij..k} H_\nu^{ij..k}$. The fluctuations in this "normal" basis are meanless, i.e., $\langle \Delta r_i \rangle = 0$, and decoupled at the lowest (second) order, i.e.,

$$\langle \Delta r_i^T \Delta r_j \rangle = \delta_{ij} \ . \tag{3}$$

A purely harmonic model is given by $\mathbf{C}_\nu = 0, \ \forall \nu$.

As a reminder to the reader, tensor Hermite polynomials can be obtained by successive differentiation using Rodrigues' formula:

$$H_\nu^{ij..k}(\mathbf{\Delta r}) \ = \ \frac{(-1)^\nu}{g(\mathbf{\Delta r})} \nabla^{ij..k} g(\mathbf{\Delta r}) \ . \tag{4}$$

Above, $g(\mathbf{x}) = (2\pi)^{3N/2} \exp(-\mathbf{x}^2/2)$ is the multi-dimensional Gaussian distribution and $\nabla^{ij..k} = \nabla^i \nabla^j .. \nabla^k$ is the gradient tensor with $\nabla^i \equiv \partial/\partial x_i$. Explicitly,

$$\begin{aligned} \mathbf{H}_2^{ij}(\mathbf{\Delta r}) &= \Delta r_i \Delta r_j - \delta_{ij} \\ \mathbf{H}_3^{ijk}(\mathbf{\Delta r}) &= \Delta r_i \Delta r_j \Delta r_k - \delta_{ij} \Delta r_k - \delta_{jk} \Delta r_i - \delta_{ki} \Delta r_j \\ &\cdots \end{aligned} \tag{5}$$

where i, j, k run over the components of the vector $\mathbf{\Delta r}$. Higher order Hermite polynomials can be obtained by summing all possible $0, 1, .., \lfloor \frac{\nu}{2} \rfloor$ pairwise contractions (δ_{ij}s) of $\Delta r_i \Delta r_j .. \Delta r_k$, a minus sign accompanying the terms obtained by an odd number of contractions. Diagonal elements of $\mathbf{H}_\nu$ (corresponding to a scalar Δr) are the usual Hermite polynomials

$$H_\nu(\Delta r) \ = \ \sum_{m=0}^{\nu/2} (-1)^m \binom{\nu}{2m} m!! \, \Delta r^{\nu-2m} \tag{6}$$

where $m!! \equiv \frac{(2m)!}{2^m m!} = (2m-1)(2m-3) \cdots 3 \cdot 1$, and the combinatorial expression counts the number of m pairwise contractions of ν variables.

The orthogonality condition for the Hermite tensor polynomials,

$$\int_{-\infty}^{\infty} \mathbf{dx} \left[\mathbf{A}_\nu \cdot \mathbf{H}_\nu(\mathbf{x}) \right] \mathbf{H}_\mu(\mathbf{x}) e^{-\mathbf{x}^2/2} \ = \ \begin{cases} \mathbf{A}_\nu \nu! \ , & \nu = \mu \\ 0 \ , & \nu \neq \mu \end{cases} \tag{7}$$

is true for any constant tensor $\mathbf{A}_\nu$ of rank ν. In particular, the tensor coefficients that appear in $f(\mathbf{\Delta r})$ follow from the orthogonality relation as

$$\mathbf{C}_\nu \ = \ \frac{1}{\nu!} \int_{-\infty}^{\infty} \mathbf{H}_\nu(\mathbf{x}) f(\mathbf{\Delta r}) \, \mathbf{d\Delta r} = \langle \mathbf{H}_\nu(\mathbf{\Delta r}) \rangle / \nu! \tag{8}$$

Therefore, the problem reduces to obtaining the expectation values of the polynomial tensor elements for the system. At the lowest nonvanishing order they read

$$\begin{aligned} \mathbf{H}_3^{111}(\mathbf{x}) &= x_1^3 - 3x_1 \\ \mathbf{H}_3^{112}(\mathbf{x}) &= x_1^2 x_2 - x_2 = \mathbf{H}_3^{121}(\mathbf{x}) = \mathbf{H}_3^{211}(\mathbf{x}) \\ \mathbf{H}_3^{123}(\mathbf{x}) &= x_1 x_2 x_3 \end{aligned} \tag{9}$$

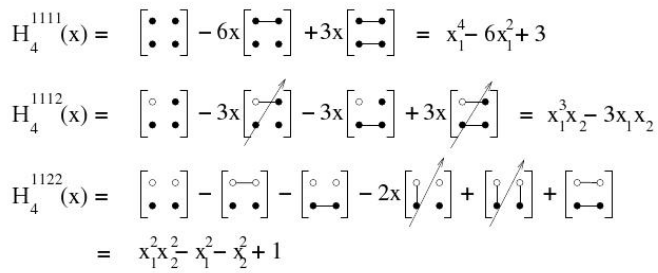

Figure 1: Graphical representation of $H_4(x)$ tensor elements in two dimensions. Terms that vanish by virtue of Eq. (3) are crossed with an arrow.

and some of the next order tensor elements are:

$$\begin{aligned}
\mathbf{H}_4^{1111}(\mathbf{x}) &= x_1^4 - 6x_1^2 + 3 \\
\mathbf{H}_4^{1112}(\mathbf{x}) &= x_1^3 x_2 - 3x_1 x_2 \\
&= \mathbf{H}_3^{1121}(\mathbf{x}) = \mathbf{H}_3^{1211}(\mathbf{x}) = \mathbf{H}_3^{2111}(\mathbf{x}) \\
\mathbf{H}_4^{1122}(\mathbf{x}) &= x_1^2 x_2^2 - x_1^2 - x_2^2 + 1 \\
&= \mathbf{H}_4^{1212}(\mathbf{x}) = \mathbf{H}_4^{1221}(\mathbf{x}) \\
\mathbf{H}_4^{1123}(\mathbf{x}) &= x_1^2 x_2 x_3 - x_2 x_3 = \mathbf{H}_4^{1231}(\mathbf{x}) = \mathbf{H}_4^{1213}(\mathbf{x}) = \cdots
\end{aligned} \tag{10}$$

A graphical representation of H_4 in one and two dimensions is given in Fig.1.

The inclusion of mode-coupling necessitates consideration of mixed indices (nondiagonal tensor elements). Here, we focus on the coupling between mode pairs and ignore threesome and higher order mixing, i.e., we consider only the bi-polynomials $\mathbf{H}_\nu^{i_1 i_2 \ldots i_\nu}(\Delta r_k, \Delta r_l)$ with $i_m \in \{k, l\}$, $k, l = 1, 2, \ldots, 3N$. At first sight, estimating the contribution of mode-coupling even at this lowest level appears to be a formidable task, because for the optimal cut-off rank $\nu = 16$ (see below), the number of distinct expectation values to be extracted from the data grows combinatorially. We show below that, the factorization property of the off-diagonal tensor elements and the orthogonality of the modes at the second order bring a significant reduction in complexity, which we exploit to investigate the impact of anharmonicity and mode-coupling separately on the protein dynamics.

2.2 Mode-coupling corrections

In order to reach the desired formulation, we first note that the value of a tensor element $\mathbf{H}_\nu^{i_1 i_2 \ldots i_\nu}(\mathbf{\Delta r})$ does not depend on the order of the indices due to the commutativity of the gradient operator, $\nabla_k \nabla_l - \nabla_l \nabla_k = 0$. Therefore,

$$\mathbf{H}_\nu^{i_1 i_2 \ldots i_\nu}(\mathbf{\Delta r}) \quad = \quad \mathbf{H}_\nu^{p}(\Delta r_k, \Delta r_l)$$

where p is the number of indices equal to k (and the remaining $\nu - p$ indices are equal to l). The fact that the covariance matrix in $\{\mathbf{\Delta r}\}$ basis is diagonal further implies that

$$\mathbf{H}_\nu^{p}(\mathbf{\Delta r}) \quad = \quad H_p(\Delta r_1) \times H_{\nu-p}(\Delta r_2) \tag{11}$$

as is also evident from the Rodrigues's formula in Eq. (4).

Combining Eqs.(8)&(11), the Hermite expansion in Eq. (2) can be cast into the following form:

$$\begin{aligned} f(\mathbf{\Delta r}) \quad = \quad & \frac{1}{\sqrt{(2\pi)^N}} e^{-\sum_i \Delta r_i^2/2} \Bigg[1 + \sum_i \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \Big\langle H_\nu(\Delta r_i) \Big\rangle H_\nu(\Delta r_i) \\ + \quad & \sum_{i \neq j} \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \sum_{p=1}^{\nu-1} \binom{\nu}{p} \Big\langle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \Big\rangle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \\ & + \sum_{i \neq j \neq k} \cdots \Bigg] \end{aligned} \tag{12}$$

The first term in $[\cdots]$ corresponds to a purely harmonic given by the Gaussian probability distribution

$$f_0(\mathbf{\Delta r}) \quad = \quad \frac{1}{\sqrt{(2\pi)^N}} e^{-\sum_i \Delta r_i^2/2} \ . \tag{13}$$

This is the starting point for most of the past studies on protein fluctuations[]. The next in Eq. (16) term is appreciable when the modes are anharmonic, but gives no information about mode-coupling. In fact, the most general mode-amplitude distribution of an anharmonic model composed of decoupled modes is

$$f_1(\mathbf{\Delta r}) \quad = \quad \frac{1}{\sqrt{(2\pi)^N}} \ e^{-\sum_i \Delta r_i^2/2} \times \prod_i \left[1 + \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \Big\langle H_\nu(\Delta r_i) \Big\rangle H_\nu(\Delta r_i) \right] \tag{14}$$

The approximation to the true distribution given in Eq. (14) is named f_1 in order to remind the reader that it qualitatively improves on the Gaussian approximation f_0 of Eq. (13). The difference between the full pdf given in Eq. (16) and the approximation f_1 is the mode-coupling corrections such as

$$\langle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \rangle - \langle H_p(\Delta r_i) \rangle \langle H_{\nu-p}(\Delta r_j) \rangle \quad \neq \quad 0$$

and higher order cumulants. Note that, these corrections are transparent to the marginal distributions

$$f(\Delta r_i) \equiv \int_0^\infty \prod_{j\neq i} d\Delta r_j \; f(\mathbf{\Delta r}) \tag{15}$$

as a merit of the orthogonality relation in Eq. (7). Therefore, even if the marginal distributions are reproduced to good accuracy, the multi-dimensional free-energy landscape of the protein may still be very different from that implied by a model based on Eq. (14). We demonstrate below that this is the case for the protein Crambin 2CI2. To this end, we improve the approximation in Eq. (14) one step further and set $f(\mathbf{\Delta r})$ to

$$\begin{aligned} f_2(\mathbf{\Delta r}) &= \frac{1}{\sqrt{(2\pi)^N}} e^{-\sum_i \Delta r_i^2/2} \Bigg[1 + \sum_i \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \Big\langle H_\nu(\Delta r_i) \Big\rangle H_\nu(\Delta r_i) \\ &+ \sum_{i\neq j} \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \sum_{p=1}^{\nu-1} \binom{\nu}{p} \Big\langle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \Big\rangle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \Bigg] \end{aligned} \tag{16}$$

which takes into account the mode-coupling corrections at the lowest order they appear and ignores higher-order terms cubic in H_ν. In the next section, we present our results for Crambin.

3 Experiment

3.1 The Simulation

Crambin (Protein Data bank code 1EJG.pdb) was selected as an example since it is a relatively small protein and its dynamics is widely studied [**?**, **?**, **?**]. **Need to get the references from "MD EXPERIMENTS.DOC"**

THIS PART COULD GO TO THE SUPPLEMENTS MAYBE?
All Molecular dynamics simulations were performed for an N,P,T ensemble in explicit solvent (water) at 310 K using NAMD 2.5 package with CHARMM27 force field. The protein was solvated in a waterbox of 15 A cushion and periodic boundary conditions were applied. Ions were added in order to represent a more typical biological environment. Langevin dynamics was used to control the system's temperature and pressure. All atoms were coupled to the heat bath of temperature of 310 K. A time step of 1fs was used. Nonbonded and electrostatic forces were evaluated each time step. In order to keep all degrees of freedom no rigid bonds were used.

*We used two minimization-equilibration cycles: The first one was applied to relax the water in the first place and the second one was applied to find a local minimum of the whole systems energy [**?**]. The energy of the initial system was first minimized for 20000 steps. The system was then equilibrated by*

*first keeping the Protein fixed for the first 0.1 ns. It took 0.02ns for the volume to converge. During the remaining 0.08ns volume fluctuated around 159000 Angstrom^3. Then, the protein was released stepwise by applying harmonic constraining forces to every backbone atom of 1, 0.5 and 0.25 kcal/(mol*Angstrom^2) in magnitude each for 0.05 ns. Finally the simulation was performed for an additional 0.05ns without applying any force. Having finished the first cycle, the second minimization-equilibration cycle was performed, this time the protein was free to move. Again 20000 steps of minimization were applied and system was equilibrated for 6ns. At every 100th time step, the instantaneous atomic coordinates $\hat{R}$ of all atoms, the velocities, the pressures and the energies were recorded. A 0.9 ns long part of the trajectory consisting of 9000 frames was used. In order to eliminate all the rotational and translational motions, all structures were aligned with respect to the initial structure using the transformation matrix which shows the best mass weighted fit with the initial structure. All transformation matrices were constructed via TCL commands in VMD.*

The 46 residue Crambin consists of 657 atoms. Taking only the alpha carbons into account a set of 138 modes were obtained. Then the overall rotational and translational motions were eliminated since they are irrelevant for the internal motion [**?**]. All structures were translated so that their center of mass is positioned at the origin and rotated to obtain the best mass weighted RMSD fit with the initial structure.

The fluctuation $\mathbf{\Delta R}$ of atoms are defined by $\mathbf{\Delta R} = \mathbf{R} - \bar{\mathbf{R}}$, where $\bar{\mathbf{R}}$ are the mean atomic coordinates and hence is a time independent quantity, which defines an average configuration obtained by the protein during the part of the trajectory that we use for calculations. The potential energy of this mean configuration is V_0. In the remaining sections of the paper we characterize the deviations of the potential energy of the residues from that of the reference conformation.

The instantaneous fluctuations are transformed into modal space using Eq. (1) where the average denoted by the angular brackets is taken over the trajectory. [**?**, **?**]. We let e represent the eigenvector matrix that diagonalizes $\Gamma^{-1/2} = \left\langle \mathbf{\Delta R \Delta R^T} \right\rangle^{-1/2}$, and λ represent the eigenvalues. Then, $\Gamma^{-1/2} = \text{diag}\ \lambda^{-1/2}\, e^T$ and the fluctuations $\mathbf{\Delta r}$ are the fluctuations in mode space spanned by the eigenvectors, e [**?**]. The components of the real trajectory that correspond to a given mode is obtained simply by keeping the eigenvalue of interest, equating all the others to zero, followed by a back transformation of Eq. (1).

3.2 Dataset

The full dataset consists of 8967 snapshots of 132 modes. To prevent overfitting, every 9-th snapshot (a total of 996) was reserved as the test set and the rest (7971 snapshots) were used as the training set. Because of normalization, each mode has zero mean and unit variance. Figure 2 gives the time plots of a few of the sample modes.

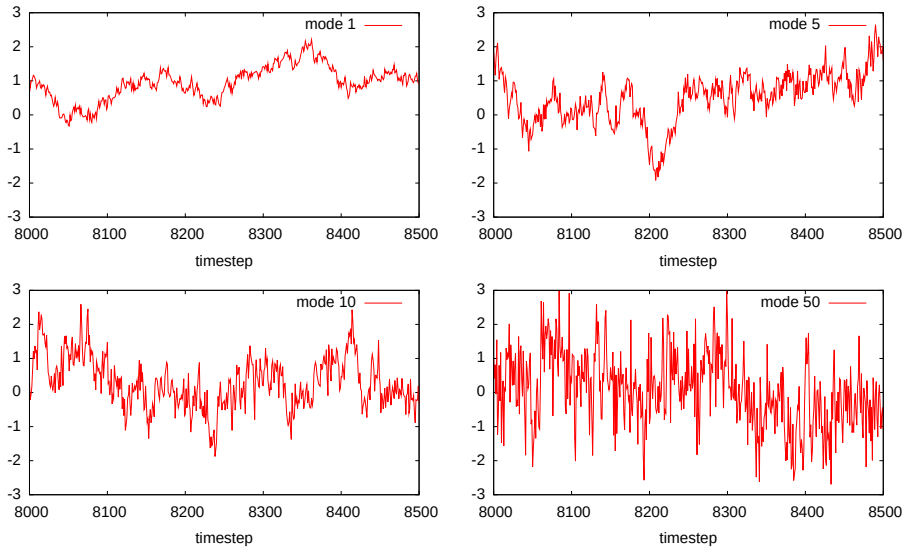

Figure 2: Time plots of sample modes.

3.3 Baselines

In this section we introduce two baseline density models that do not take into account mode coupling. The first baseline model, f_0, is the standard multivariate normal distribution with zero mean and unit variance in Eq. (13). The second baseline model, f_1, uses first-order hermite corrections based on statistics of the training data given by Eq. (14).

To compare the quality of the models, we use the average log likelihood of the snapshots in the test data given parameters optimized for the training data[1]. Eq. (17) defines the average log likelihood of the data. $\mathbf{\Delta r}^{(i)}$ denotes the i'th snapshot and N is the number of snapshots.

$$\langle \log f(\mathbf{\Delta r}) \rangle = \frac{1}{N} \sum_{i=1}^{N} \log f(\mathbf{\Delta r}^{(i)}) \tag{17}$$

The average log likelihood based on f_0 is -187.449 per snapshot, corresponding to 111.5 kcal/mol contribution to the free energy at room temperature ($= \langle \log f \rangle k_B \times 1.6 \cdot 10^{-19} \times 6.02 \cdot 10^{23}/4200$ - Alkan). The measurement of f_1 is a bit more problematic because the approximation gives negative probabilities on some test instances. This problem can be partially solved by applying the f_1 approximation only to the first few modes and assuming the rest are coming from the standard normal distribution. Figure 3 describes the outcome based on the number of modes where the f_1 approximation is applied. The first plot gives the average log likelihood based on only the positive probability instances.

[1] except for f_0, which has no parameters to optimize

The second plot gives the proportion of negative probability instances. If we ignore the negative probabilities the best log likelihood for f_1 is -186.43.

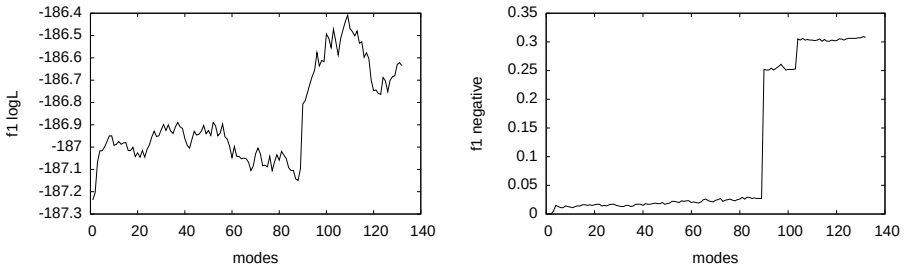

Figure 3: Log likelihood and proportion of negative probabilities given the number of modes where f_1 approximation is applied.

The slight improvement of f_1 over f_0 is due to the better fit f_1 provides for individual modes. Figure 4 compares the f_0 and f_1 distributions with the test data histogram for mode 1. Although f_1 fits the marginal mode distributions better, modes are still assumed independent of each other.

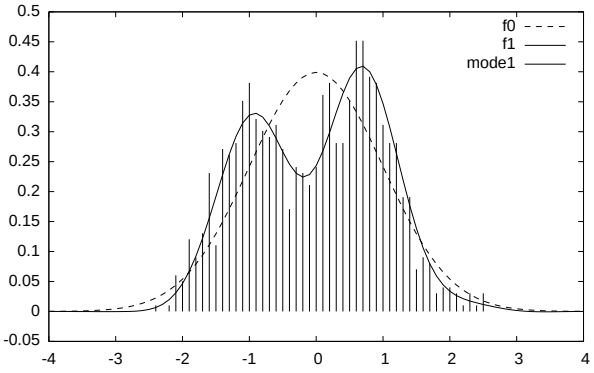

Figure 4: A comparison of f_0 and f_1 with the normalized histogram for mode 1

3.4 Mode Coupling Corrections

The f_2 approximation (Equation 16) can model pairwise interactions between modes. Figure 5 shows the log likelihood achieved and the proportion of negative probabilities based on the number of modes f_2 approximation is applied. The remaining modes are assumed to be generated independently from the standard normal distribution. If we ignore the negative probabilities the best log likelihood for f_2 is -180.05, *a correction of* ≈ 3.8 *kcal/mol to the free energy solely due to mode-coupling.*(Bu free energy farki daha anlamli herhalde. -Alkan)

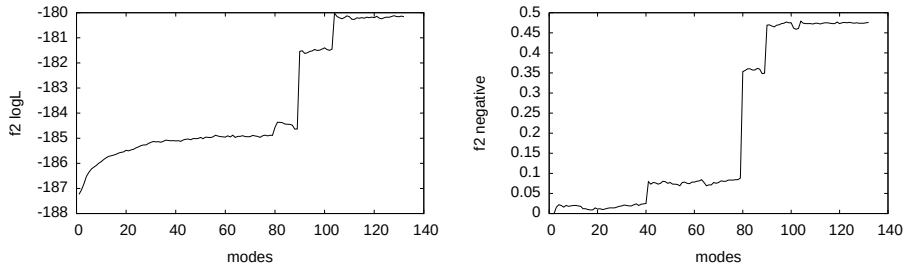

Figure 5: Log likelihood and proportion of negative probabilities given the number of modes where f2 approximation is applied.

f_2 achieves a significantly better fit to the data compared to f_1. The two models give identical marginal distributions to individual modes. The difference comes from f_2's better modeling of pairwise interactions. Figure 6 compares f_1 and f_2 distributions with the scatter plot of the first two modes.

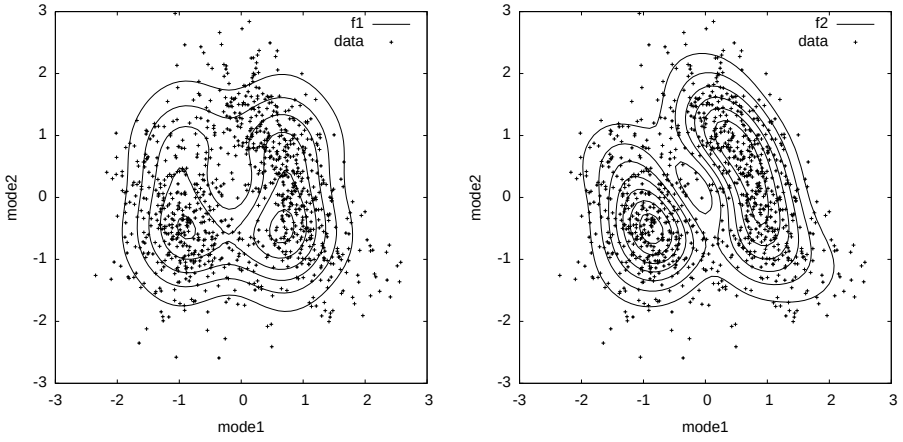

Figure 6: A comparison of f_1 and f_2 against a scatter plot of the first two modes.

3.5 Kernel Density Estimation

To quantify the effects of phenomena beyond pairwise interactions, we used non-parametric kernel density estimation (KDE) procedure to model the probability density of the training set. We fit second order Gaussian kernels with fixed bandwidths optimized using likelihood cross validation[2].

[2]We used the "npudensbw" and "npudens" procedures with default arguments from the "np" package for the "R" statistical computing environment by Tristen Hayfield and Jeffrey

Figure 7 shows the log likelihood achieved based on the number of modes KDE approximation is applied. The remaining modes are assumed to be generated independently from the standard normal distribution. There are no negative probability estimates in KDE.

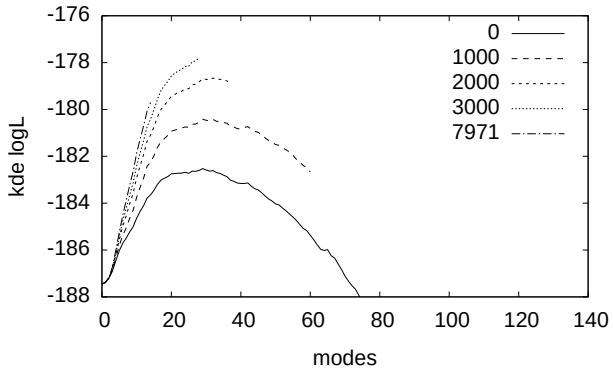

Figure 7: Log likelihood given the number of modes where KDE approximation is applied. Each curve represents a different training set size. The best result is -176.569 using 30 modes on the full training set of 7971 instances.

Figure 8 compares f_2 and KDE distributions with the scatter plot of the first two modes. KDE achieves a significantly better fit to the data compared to f_2. The peaks are higher and the contours are more detailed for the KDE figure. More importantly KDE is not restricted to pairwise interactions only, it can model higher order interactions as well.

References

[1] Reference GNM, ENM, etc.

[2] Yogurtcu *et al.* (2009).

[3] Reference on mode coupling, enegry transfer between modes, etc.

[4] Flory, 1976.

S. Racine.

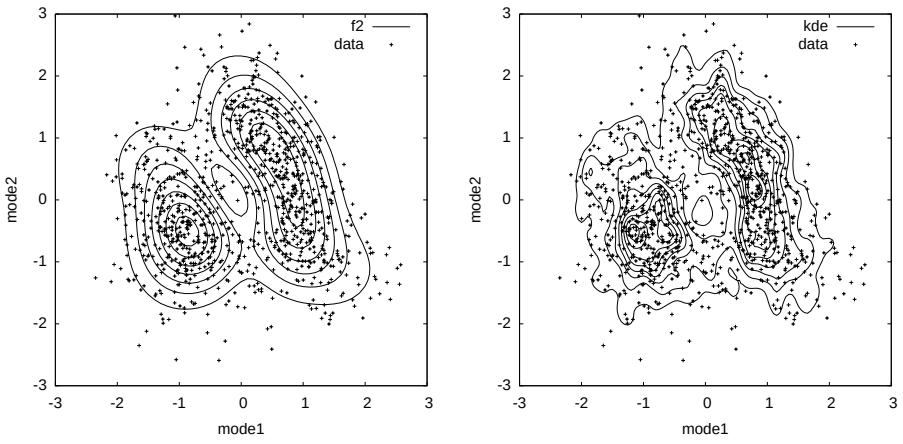

Figure 8: A comparison of f_2 and KDE against a scatter plot of the first two modes.