Identifying Nonlinear Correlations in High Dimensional Data with Application to Protein Molecular Dynamics Simulations

UMBC REU Site: Interdisciplinary Program in High Performance Computing

Team members: William J. Bailey[1], Claire A. Chambless[2], Brandynne M. Cho[3], and Jesse D. Smith[4]
Graduate assistant: Andrew M. Raim[4], Faculty mentor: Kofi P. Adragni[4], Client: Ian F. Thorpe[5]
[1]Kenyon College, [2]Elizabethtown College, [3]Saint Mary's College of California, [4]Math. and Stat., UMBC, [5]Chemistry, UMBC

Biological Motivation

Molecular Dynamics (MD) Simulations simulate the movement of atoms in a biomolecule and generate data sets with high dimensionality.

Allostery is the process by which an event at one location affects the properties of another location in a biomolecule. Allostery allows information to be communicated over long distances.

How can simulations of allostery lead to an in-depth understanding of this biological process?

Statistical Background

Covariance maps, often used to asses protein motional correlations, can fail to detect nonlinear statistical dependency.

Many methods have been developed to find nonlinear correlations (e.g., kernel PCA, Isomap, etc.) but also have drawbacks.

Principal Fitted Components (PFC) is a novel technique that overcomes many of the issues of high dimensional data.

Principal Fitted Components

PFC is an inverse regression methodology with p predictors $\mathbf{X} = [x_1, \ldots, x_p]$ and a response Y. The general model of PFC is

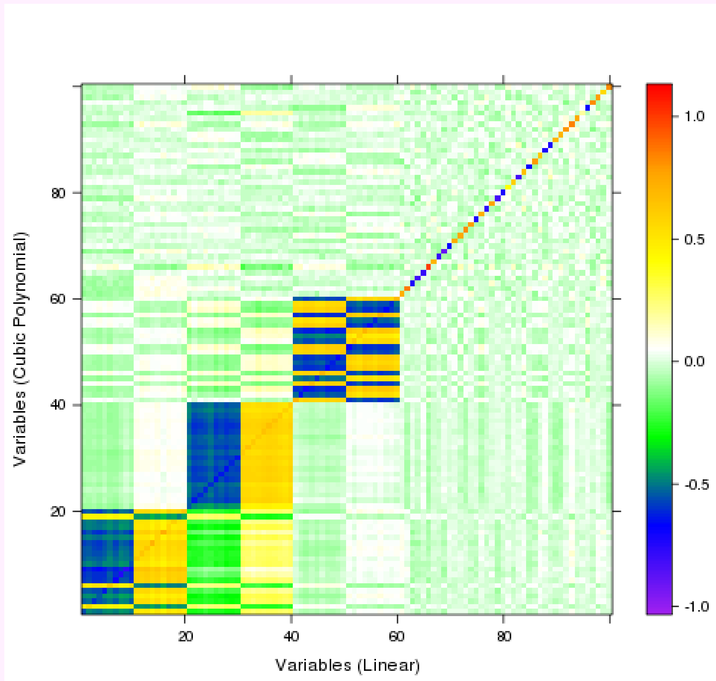
$$E(X|Y) - E(X) = \Gamma\, \nu_Y + \Delta^{-\frac{1}{2}}\, \epsilon.$$

ν_Y is an unknown function of Y, Δ is the conditional variance function independent of Y, and $\epsilon \sim N(0, I)$.
It is important to chose an appropriate **basis function** when approximating ν_Y as $\nu_Y \approx \beta\, \mathbf{f}_Y$, where $\mathbf{f}_Y$ is a collection of r functions used to capture the underlying structure of the data and β is an unconstrained $1 \times r$ vector.
There are three possible error structures: isotropic, anisotropic, and unstructured. For MD data, the isotropic $\Delta = \sigma^2 I_p$ is appropriate, because all predictors are on the same measurement scale.

Implementation of PFC

Parallelization

- Parallelization offers the opportunity to analyze much larger data sets.
- We wrote a parallel wrapper using the pbdMPI library to implement the 'pfc' function from the R library 'ldr' [2].



Association map of simulated data with $\mathbf{f_Y} = (\mathbf{Y}, \mathbf{Y^2}, \mathbf{Y^3})$

A correlation matrix $\Theta_{p\times p}$ describes the linear $(r = 1)$ correlation between each of the variables; it has the form

$$\Theta = [\hat{\Gamma}_1\beta, \ldots, \hat{\Gamma}_p\beta]$$

where $\hat{\Gamma}_k$ is the estimate of Γ using the k^{th} column of $\mathbf{X}$ as the response. Here, $\beta \in \mathbb{R}$.

For a basis with $r \geq 1$, $\hat{\Gamma}_k$ is $p \times 1$, and β is $1 \times r$. To obtain a unified illustration of the associations, the columns must have a different form:

$$\Theta_j = \hat{\Gamma}_j ||\beta|| \hat{\Gamma}_j^T \hat{\Gamma}_j.$$

Simulated Data
To characterize the effectiveness of PFC a set of data was simulated containing known linear and nonlinear correlations. The data matrix $\mathbf{X_{300\times 100}}$ has the form

$$\mathbf{X} = [Y, Y^2, Y\cos(6\pi Y)]^T \Lambda + \epsilon,$$

where $\epsilon_{n\times p} \sim N(0, I)$.

Application to MD Simulations

The Data Set

- The columns of data matrix $\mathbf{X}$ are the atom number $1, 2, \ldots, 531$ and each atom's coordinates, x, y, and z
- The rows are the times the coordinates were recorded with between 100 and 10,000 time steps

Time shifting
In natural physical motion, as in MD simulated motion, instantaneous reactions do not occur. Thus, we created a method of time shifting, which compares responses to predictors within 'tmax' time steps:

- Consider the j^{th} column of data matrix $\mathbf{X}$ to be the response
- Shift each entry in the column up by one row, so that the i^{th} element is now in the $(i-1)^{\text{th}}$ location.
- Remove 'tmax' rows
- Repeat for all p columns of $\mathbf{X}$.
- Repeat process for all time steps from 1 to maximum (e.g., 10)

Example of a time shift with $t = 1$ and column $j = 2$ is the response:

$$\begin{bmatrix} x_{1,1} & x_{2,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{3,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-1,1} & x_{n,2} & \cdots & x_{n-1,p} \\ \cancel{x_{n,1}} & \cancel{x_{n,2}} & \cancel{\cdots} & \cancel{x_{n,p}} \end{bmatrix}$$

Application Performance
An efficiency study with $p = 531$ demonstrated the effectiveness and scalability of parallelized PFC, by reducing the serial time of over 2 hours to about 1 minute on 256 processes.

References and Acknowledgments

1. Full technical report HPCF–2013–12 at `www.umbc.edu/hpcf` > Publications
2. R documentation for 'ldr' library

– REU Site: `www.umbc.edu/hpcreu`
– NSF, NSA, UMBC, HPCF, CIRC