Comparing Structure Learning Methods for RKHS Embeddings of Protein Structures

Anonymous Author(s)
Affiliation
Address
email

Abstract

Non-parametric graphical models, embedded in reproducing kernel Hilbert spaces, provide a framework to model multi-modal and arbitrary multi-variate distributions, which are essential when modeling complex protein structures. Non-parametric belief propagation requires the structure of the graphical model to be known a priori. Currently there are nonparametric structure learning algorithms available for tree structures, but a tree structure is not reasonable when modeling protein molecular networks. In this paper, we compare parametric neighborhood selection structure learning method, which is capable of recovering true general graph structures, to the non-parametric tree learning method, for the particular task of modeling protein structures represented as sequences of torsion angles. Our experiments, performed on molecular dynamics simulation data of Engrailed Homeodomain protein, show that neighborhood selection method outperforms nonparametric tree structure learning method. We also find that non-parametric models outperform the semi-parametric non-paranormal model as well as the parametric sparse Gaussian graphical model, when an appropriate kernel is used.

1 Introduction

The three dimensional structures of proteins and other molecules vary in time according to the laws of thermodynamics. The ability to model the resulting probability distribution over structures — and its response to environmental perturbations, has a number of practical applications, including computer-aided drug design. Probabilistic graphical models are a natural choice for encodings these distributions.

A protein's structure can be defined as a set of torsion angles, corresponding to the rotatable bonds within the molecule. Hence, graphical models over continuous-variables can be used to encode distributions over these angles, and thus over protein structures. Parametric models, such as Gaussian graphical models[5],[1], or von-Mises graphical models[11], facilitate inference and learning owing to their compact parametric forms. Gaussian graphical models are especially common because they have closed form analytical solutions to inference queries. Unfortunately, the distributions of angles seen in protein data are generally non-Gaussian. Thus, there is a need for graphical models over continuous variables that capture the statistical features of real proteins.

The literature on non-Gaussian, continuous-variable graphical models includes mixture models[17], semi-parametric models[8][7], and non-parametric models[14][13]. Mixture models, such as mixtures of Guassians, are powerful but introduce serious computational challenges, unless simplifying assumptions are made. Semi parametric models, such as non-paranormals[8], define a parametric form over the $transformed$ data, where the transformations are smooth functions around the data points. The non-paranormal model, specifically, restricts the transformed data to have a multivariate Gaussian distribution. Nonparametric graphical models, such as reproducing kernel Hilbert space (RKHS) embedded models[14][13], do not enforce any parametric form on the transformed data, and are based on kernel density estimation of any query that is sent to the graphical model. Since these models use the data samples

themselves to represent the data, they can reflect more model complexity as the amount of available data increases. This property allows them to scale appropriately with the data, and model arbitrary and multi-modal distributions.

Previous work on nonparametric methods have assumed the structure of the graph is known, or the graph is fully connected. In large protein structures, residues tend to interact based on their proximity in 3D structure, leading to a relatively sparse set of interactions. Fully connected networks are not reasonable for inference, and we must find other methods to estimate these networks before we can perform any nonparametric inference. In this paper, we compare two existing structure learning methods for RKHS embedded models, and then perform experiments using protein molecular dynamics simulation data.

2 Belief Propagation in Reproducing Kernel Hilbert Space

A Hilbert space, H, is a complete vector space, endowed with a dot product operation. When elements of H are vectors, each with elements from some space, F, a Hilbert space requires that the result of the dot product be in F as well. For example, the space of vectors in $\Re^n$ is a Hilbert space, since the dot product of any two elements is in $\Re$. [3]. Reproducing kernel Hilbert space is a Hilbert space defined over a $reproducing$ kernel function. Reproducing kernels are the family of kernels that define a dot product function space, which allows any new function, $f(x)$, to be evaluated as a dot product of the feature vector of x, $\phi(x)$, and the f function. Thus:

$$f(x) = \langle K(x,.), f(.) \rangle$$

$$Consequently, k(x, x_0) = \langle K(x,.), K(x_0,.) \rangle$$

This reproducing property is essential to define operations required for calculating expected values of functions and belief propagation messages in kernel space.

Given an *iid* dataset $X = \{x_1, ..., x_m\}$, Smola *et al.*[12] define two main mappings, $\mu[P_x] = E_x[k(x,,)]$ and $\mu[x] = 1/m \sum_{i=1}^{m} k(x_i,.)$, which allows us to estimate the expected value of any function $f(x)$, using the reproducing property, as: $E_x[f(x)] = \langle \mu[P_x], f \rangle$, and $\langle \mu[X], f \rangle = 1/m \sum_{i=1}^{m} f(x_i)$. Song *et al* [15] then define $Covariance\ operator$, $C_{X,Y} = E_{X,Y}[\phi(X) \bigotimes \phi(Y)] - \mu_X \bigotimes \mu_Y$, which allows us to calculate the expected value of product of any two functions in the $f(x)$ and $g(y)$ non-parametrically, using the reproducing property. Covariance operator is then used to represent conditional mean mappings: $\mu_{Y|x} = C_{Y,X} C_{X,X}^{-1} \phi(x)$, which can be estimated from the samples for any value x, using the reproducing property. Using these two main ideas, Song *et al* [14] provide a unified framework to embed tree graphical models as a set of conditional probabilities, and perform nonparametric inference on trees [14] and loopy[13] graphical models, by representing messages and beliefs in kernel space.

3 Structure Learning Methods for RKHS inference

Recently [16], Song *et al* proposed a method to perform structure learning for tree graphical models in RKHS. Their method is based on the structure learning method proposed by Choi *et al* [2], where a tree metric is first used to estimate a distance measure between node pairs. A minimum spanning tree is then calculated from the distances [6]. Choi *et al* define the distance metric based on correlation coefficient, and Song *et al* define the nonparametric version of this metric, based on kernel space covariance operator. This method is suitable for nonparametric structure learning, but it is limited to tree structures, which is not usually the case in protein structures.

For general loopy graphs, neighborhood selection is also a popular method for structure learning in graphical models. This method, proposed by Meinshausen *et al* [10], breaks the overall optimization problem into a set of smaller optimization problems, by maximizing $Pseudo - likelihood$, instead of the full likelihood. Each optimization term in the pseudo likelihood becomes equivalent to a regression problem, and can be solved efficiently with the Lasso regression. In this paper, we use this method to recover our structures, and compared it with the nonparametric tree structure learning algorithm. We currently limit our experiments to sparse linear regression, which does not take the nonlinear relations between the variables into account, and leave the use of sparse nonparametric regression for neighborhood selection for future work.

Figure 1: Engrailed Homeodomain

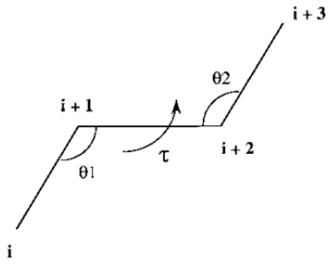

Figure 2: Theta and tau angles encoding the structure of the protein's alpha carbon atoms.

Dataset Dataset	RKHS: Gaussian Kernel	RKHS: Triangular Kernel	Non-paranormal Graphical Model	Gaussian Graphical Model
First1000	8.42	7.30	8.43	8.46
Uniform1000	54.76	51.34	63	59.4

Table 1: RMSE for RKHS with neighborhood selection(using two kernels), compared with non-paranormal and Gaussian graphical models

4 Experiments

We performed our experiments over Engrailed Homeodomain (Protein ID: 1ENH) MD simulations data, which is a 54-residue DNA binding domain (Figure 1). The DNA-binding domains of the homeotic proteins, called homeodomains (HD), play an important role in the development of all multicellular animals, and certain mutations to HDs are known to cause disease in humans [4]. This protein is an ultra-fast folding protein that is expected to exhibit substantial conformational fluctuations at equilibrium. [9]

Dataset We performed 50-microsecond simulations of the protein at 350 degrees Kelvin. This simulation was performed on ANTON, a special-purpose supercomputer designed to perform long-timescale simulations. We sampled more than 500,000 frames through the simulation, and used angular sequence of (θ, τ) to represent each frame. Figure 2 shows what these angles represent on a section of an imaginary protein sequence. Currently our nonparametric inference method can not scale to the whole dataset, so we created two sub-sampled versions of the data. First1000 includes the first 1000 samples, while Uniform1000 data contains the uniformly sampled 1000 samples. Both these datasets exhibit multi-modal and non-Gaussian marginal distributions.

Evaluations In our experiments, we performed leave-one-out cross-validation, and calculated the Root Mean Squared Error (RMSE) of the test frame, given that the model is learned from the training set. For each test frame, we have also assumed randomly selected 50% of the variables of the frame are observed, and have predicted the rest of the variables, given these observations and the training data. For each frame we have repeated this 50% subset selection 10 times.

Results Table 1 shows the results of running the full cross validation experiment, on RKHS method with neighborhood selection as the structure learning method with two different kernels, versus the non-paranormal and sparse Gaussian graphical model. In all cases, we show the RMSE (measured in degrees) of the predicted hidden variables conditioned on the observed variables, and the RMSE is calculated from the difference of predicted and actual values of the hidden variables.

We note that in this particular case, where our data is angular, we try two different kernels: Gaussian kernel, $K_1 = e^{-\lambda||x-y||^2}$, and triangular kernel, $K_2 = e^{-\lambda(sin(||x-y||))^2}$.

Table 2 shows the RMSE results of tree structure learning, versus the neighborhood selection method, on the first 1000 sample dataset. We used a triangular kernel in both cases, since it outperformed the Gaussian kernel on our angular data.

Neighborhood selection with Triangular kernel	Tree structure learning with Triangular kernel
7.30	7.41

Table 2: RMSE result for RKHS using neighborhood selection, versus tree structure learning on First1000 dataset. Both methods used triangular kernel.

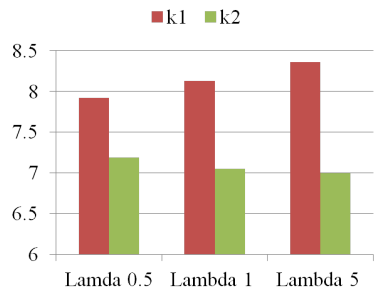

Figure 3: Effect of structure sparsity in neighborhood selecton on RMSE in RKHS inference

As can be seen from these results, without an appropriate kernel, RKHS models do not outperform the non-paranormal and Gaussian graphical models. However, if the kernel is well suited for the problem (i.e. triangular kernel function for our data), we see significant improvement in RMSE score of neighborhood selection for structure learning in RKHS, over non-paranormal and Gaussian graphical models. Additionally, we observe that neighborhood selection outperforms the tree-based nonparametric structure learning. However, we note that neighborhood selection with Gaussian kernel over angular data does worse than tree structured method with triangular kernel. This demonstrates the importance of kernel selection and learning.

We also investigated the effect of the density of the estimated structure learned by neighborhood selection on the RMSE of the predictions. Using different regularization penalties in Lasso regression, results in different levels of sparsity. Figure 3 shows the RMSE for different values of the regularization penalty, measured with both Gaussian and triangular kernels, when modeling the first1000 dataset. As the graph becomes denser (higher λ corresponds to denser graphs, as in our implementation λ corresponds to the upper-bound of sum of regression coefficients), the triangular kernel performs better. The Gaussian kernel, on the other hand, does not benefit from denser graphs.

5 Conclusions and Future Work

In this paper we compared non-parametric tree structure learning to parametric neighborhood selection method based on *l1*-regularized linear regression. Our experiments on modeling protein structures represented as sequences of torsion angles, showed that neighborhood selection outperforms non-parametric tree structure learning method. Furthermore, both these models outperform a semi-parametric non-paranormal model and parametric sparse Gaussian graphical model when a a triangular kernel is used.

As part of our future work, we are focusing on non-parametric sparse regression and variable selection methods to improve the neighborhood selection so that non-linear relationships can be utilized when estimating the sparse graph structure. Another direction for future work includes kernel optimization, as we have already observed the importance of the correct parameters and form for the kernel. Finally, scalability of the non-parametric belief propagation is an issue which we will address in our future work.

References

[1] BANERJEE, O., EL GHAOUI, L., AND D'ASPREMONT, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res. 9* (June 2008), 485–516.

[2] CHOI, M. J., TAN, V. Y. F., ANANDKUMAR, A., AND WILLSKY, A. S. Learning latent tree graphical models. *J. Mach. Learn. Res. 12* (July 2011), 1771–1812.

[3] DAUM'E III, H. From zero to reproducing kernel hilbert spaces in twelve pages or less. February 2004.

[4] DELIA, A. V., TELL, G., PARON, I., PELLIZZARI, L., LONIGRO, R., AND DAMANTE, G. Missense mutations of human homeoboxes: A review. *Human Mutation 18*, 5 (2001), 361–374.

[5] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*, 3 (2008), 432–441.

[6] KRUSKAL, JOSEPH B., J. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society 7*, 1 (1956), pp. 48–50.

[7] LAFFERTY, J., LIU, H., AND WASSERMAN, L. Sparse Nonparametric Graphical Models. *ArXiv e-prints* (Jan. 2012).

[8] LIU, H., LAFFERTY, J., AND WASSERMAN, L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res. 10* (Dec. 2009), 2295–2328.

[9] MAYOR, U., JOHNSON, C. M., DAGGETT, V., AND FERSHT, A. R. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proceedings of the National Academy of Sciences 97*, 25 (2000), 13518–13522.

[10] MEINSHAUSEN, N., AND BHLMANN, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 3 (2006), pp. 1436–1462.

[11] RAZAVIAN, N. S., KAMISETTY, H., AND LANGMEAD, C. J. The von mises graphical model:structure learning, 2011.

[12] SMOLA, A., GRETTON, A., SONG, L., AND SCHÖLKOPF, B. A hilbert space embedding for distributions. In *Algorithmic Learning Theory* (2007), Springer. Invited paper.

[13] SONG, L., GRETTON, A., BICKSON, D., LOW, Y., AND GUESTRIN, C. Kernel belief propagation. In *International Conference on Artifical Intelligence and Statistics (AISTATS)* (2011).

[14] SONG, L., GRETTON, A., AND GUESTRIN, C. Nonparametric tree graphical models. In *Artificial Intelligence and Statistics (AISTATS)* (2010).

[15] SONG, L., HUANG, J., SMOLA, A., AND FUKUMIZU, K. Hilbert space embeddings of conditional distributions. In *International Conference on Machine Learning* (2009).

[16] SONG, L., PARIKH, A., AND XING, E. Kernel embeddings of latent tree graphical models. In *Neural Information Processing Systems (NIPS)* (2011).

[17] SUDDERTH, E. B., IHLER, A. T., ISARD, M., FREEMAN, W. T., AND WILLSKY, A. S. Nonparametric belief propagation. *Commun. ACM 53*, 10 (Oct. 2010), 95–103.