Department of Applied and Computational Mathematics and Statistics Colloquium

Hui Zou

School of Statistics
University of Minnesota

will give a lecture entitled:

Statistical Inference of Sparse Ising Models, with Applications to HIV Mutation Data

Abstract

The Ising model is a very useful mathematical model that has been successfully exploited to model complex interactions for network exploration in various research fields such as physics, social-economics, protein modeling and statistical genetics, and so on. Consider an Ising model with K "magnetic dipoles" denoted by X_j , $1 \le j \le K$. Each X_j equals +1 or −1 and the joint distribution of $\mathbf{X} = (X_1, \ldots, X_K)$ is $\Pr(X_1 = x_1, \ldots, X_K = x_K) = 1/Z(\beta) \exp(\Sigma_{(i,j)} (\beta_{ij} x_j x_i)/4)$, where $Z(\beta)$ is the partition function and $\beta_{ii} = 0$, $\beta_{ij} = \beta_{ji}$. The number of unknown parameters in this model is equal to $p = K(K\text{-}1)/2$. Thus, the problem of estimating the Ising model naturally falls into the category of high dimensional inference. In the current literature penalized likelihood estimation has become a standard technique for sparse inference with high dimensions. Unfortunately, this nice idea cannot be used for estimating the Ising model because the likelihood function involves the partition function $Z(\beta)$ which is computationally intractable even for moderate size K (such as $K = 30$).

To overcome such difficulties, we develop the methodology and theory of *high-dimensional penalized composite likelihood estimation.* We consider both $\ell 1$ and concave penalized composite likelihood estimators. Let n be the sample size and let A be the support of the true sparse Ising model. Under weak regularity conditions, if $n \gg \log(K)|A|^3$ the concave penalized composite likelihood estimator enjoys the strong oracle properties. Compared to the concave penalized estimator, the ℓ_1 penalized estimator requires an extra model assumption in order to achieve consistent sparse recovery.

In practice, the utility of a statistical method heavily depends on its computational feasibility and efficiency. In the penalized least squares setting, the ℓ_1 estimator is often preferred over the concave estimators, mainly because of the computational consideration. In the much more challenging Ising model case, we derive a *new unified algorithm* for efficiently computing the solution paths of *both* ℓ_1 *and concave* penalized composite likelihood estimators. Through extensive numerical experiments we show that, with the aid of our new algorithm, the concave estimator can enjoy computational efficiency comparable to that of the ℓ_1 estimator.

We provide a detailed demonstration of our methodology by studying the Human Immunodeficiency Virus type 1 (HIV-1) protease structure based on data from the Stanford HIV Drug Resistance Database. Our statistical learning results match the known biological findings very well.

Monday, January 30, 2012
4:00 p.m. to 5:00 p.m.
127 Hayes-Healy Center