Research Statement

Sandipan Roy

November 19, 2014

My research focuses on modeling complex data having high-dimensional network structure and provide statistical methodology for estimating the corresponding network structure using tools from graphical models and high dimensional inference. I have been working with my advisors Dr. George Michailidis and Dr. Yves Atchadé on my dissertation on developing computational tools and providing theoretical justifications to some intriguing problems in the area of high-dimensional networks. The emphasis is placed on developing new theoretical techniques and computational tools for network problems and applying the corresponding methodology in many fields, including biomedical and social science research, where network modeling and analysis plays an exceedingly important role. I also have an active interest in graphical models and their applications to real world networks, theoretical issues involving high dimensional inference, nonparametric regression and Bayesian inference. In the next section I briefly describe my research and demonstrate how I have dealt with some of the challenges that comes with modeling and estimating complex high dimensional network structure.

• Phd Research: Modeling data with complex high dimensional network structure

 Broadly speaking two key aspects of my research are the following -(a) Modeling network structures that vary over time and (b) estimating the network structures with possible presence of groups or clusters. The first problem is important since the examples of time-varying network structures are ubiquitous in the nature and the increasing availability of data sets that evolve over time has accentuated the need for developing models for time varying networks. Examples of such data sets include time course gene expression data, voting records of legislative bodies, etc.

Estimation of time invariant networks from independent and identically distributed data based on the MRF model has been a very active research area (see e.g. Höfling and Tibshirani (2009); Ravikumar et al. (2010)). On the other hand, there is significant less work on time varying networks (see Kolar et al. (2010), Kolar and Xing (2012)).The second aspect is a popular problem in the field of Physics, Computer science and Statistics. While physicists and computer scientists used algorithms based on modularity or some other graph properties (see for example Clauset, Newman and Moore (2004), Newman and Girvan (2004)), statistics literature (see for example Airoldi, Blei, Fienberg and Xing (2008),Chen, Amini, Bickel and Levina (2012)) considered the problem as a model fitting problem for independent Bernoulli data using node labels as latent variables.

Specifically, for modeling a time-varying network we model the network which undergoes a big shift/jump in the structure over time and aim to estimate the change-point of the structural alteration. Such examples of networks would be political networks, a gene regulatory network that exhibits a significant change at a particular dose of a drug treatment, or in finance
where major economic announcements may disrupt financial networks. We adapt the framework of Markov random field to model the network structures pre and post change-point and allow the number of nodes to grow with time to introduce the high dimensional nature of the data.

Dimension reduction in the high-dimensional regime is typically achieved under the assumption of some relevant, lower dimensional structure on the underlying network like sparsity of edges (Ravikumar et al. (2010)) or presence of a few underlying latent variables. The framework of regularized regressions has gained a lot of attention in the last decade as a rigorous way to incorporate such structural assumptions in the models.

Change-point estimation problem has a long history in statistics (see Bai (2010), Hinkley (1970), Lan, Banerjee and Michailidis (2009)) but its use in a high-dimensional network problem is novel and motivated by the US Senate voting record application. The key contribution comes in establishing the finite sample estimation error bound involving the key parameters (length of the observation time interval, dimensionality of the networks and the corresponding level of sparsity of the network before and after the change-point) in the model.

The derivation of the result requires a careful handling of model misspecification in Markov random fields, a novel aspect not present when estimating a single Markov random field from independent and identically distributed observations.

The second aspect of my research is more about exploring the structure of the underlying network. Many of the examples of real world networks have a underlying group/community structure. A prime example would be online social networks such as Facebook/Twitter where one can clearly observe distinct groups/clusters of friends. Especially, in this age of massive data or “Big Data” the goal would be how to unearth the underlying community structure reasonabley fast. Many algorithms have been proposed for fitting network models with communities, but most of them do not scale well to large networks and often fail on sparse networks. We focus on the computational aspect of the problem and develop an approach based on subsampling a chunk of the “Big Data” and use a EM type algorithm on that small portion of the original data. We repeat it multiple times independently in parallel cores and aggregate the estimate result from different cores by simple averaging.

The key idea again goes back to the fact that using a few latent variables one can achieve the dimension reduction in high-dimensional regime. Although we have $n$ individuals, the number of clusters $K$ is quite small compared to $n$ and we can work with a small portion of the initial “Big Data” to retrieve the groups/ clusters in the network as well as estimate the model parameters with reasonable accuracy.

- **Future Research Plan**

  The possible directions that I want to work on as a researcher primarily comprise high dimensional network structure estimation via graphical modeling and investigations of theoretical properties of the underlying high dimensional methodology used for the estimation. Further, Up until now, the Statistics community has focused mainly on structure identification and provided consistency results for point estimates. On the theoretical front, I plan to develop meaningful inference framework (confidence intervals, hypothesis testing) for quantifying uncertainty in network models. Even in the low dimensional setting, it is challenging to develop efficient inference framework that can associate p-values to the presence or absence of particular subnetworks. There are interesting and non-trivial challenges in the domain of uncertainty quantification. For example, to find a suitable confidence interval for the change-point estimate that I discussed in my current
research is certainly a difficult problem and is quite different from the way classical change-point estimate interval is constructed. The challenge of constructing such confidence interval for the change-point estimator comes from the fact that the estimate itself is a function of the lasso estimates of the other model parameters (viz. the ones that encode the network structure before and after the change-point) and those estimators do not have a tractable limiting distribution. The idea of subsampling the “Big data” could be seen as a useful idea applicable in many other instances where one faces the challenge of dealing with massive amount of data. For example, consider a recommender system where you have a number of customers and there are ratings for different items given by the individual customers. We can use our subsampling technique to cluster the individuals who have similar kinds of preferences for the corresponding items. In a recent paper by Kleiner et al. (2014) the idea of scalable bootstrap is discussed for massive data. This work can also be seen as a way of constructing computationally efficient estimators in the “Big Data” setting. In the modern age where we have access to huge amount of data, extending the idea of bootstrap and subsampling to this settings along with the use of parallel and distributed computing could be seen as an area with great scope of exploring efficient computational algorithms. Further, there will be interesting theoretical challenges that comes in constructing those efficient algorithms as well.

In summary I am interested in developing theory and methods to address different theoretical and computational issues in the context of network learning. Motivated by many real world applications in diverse scientific fields including biology, economics and the social sciences I see this as a great opportunity to work in the field of modeling and learning network structure using tools from machine learning, optimization and statistics.

References


