Duality between faithfulness assumptions in Graphical models

Dhafer Malouche, Bala Rajaratnam

The Perfect Markovianity and faithful assumption in Graphical Models : The case of Multivariate Gaussian distributions.

Dhafer Malouche & Bala Rajaratnam

University of 7th-November at Carthage, Tunisia
&
Stanford University, USA

Graphical Models and Bayesian Networks have found widespread use in statistics - especially in high dimensional settings. One important application of these models is in the area of analyzing gene expression level data (see Friedman *et al.* (2000), Magwene and Kim (2004), Castelo and Roverato (2006), Wille and Bühlman (2006), Malouche and Sevestre (2008)...). One of the main objectives in this active area of research is to reconstruct or recover the graph, $G = (V, E)$, representing the interactions between genes from observed data. This graph is often termed as a *Gene Network Interaction* (see Toh and Horimoto (2002)). Under the assumption of *Gaussianity*, this gene network coincides with the *concentration* graph (see Lauritzen (1996)) associated with an unknown probability distribution generating the data. These types of datasets typically have a low number of observations compared to the number of variables, i.e., $n \leq p = |V|$, and classical estimation procedures for this concentration graph model are no longer applicable. Classical procedures include those based on the maximum likelihood estimation(mle) of the covariance matrix or its inverse (see for example Lauritzen (1996) or Edwards (2000)). Existence of the mle is not even guaranteed in high dimensions, let alone obtaining a stable estimator with good properties. Buhl (1993) showed that the maximum likelihood estimator exists with probability one if the number of observations, n, is greater than the number of variables, p, however, this probability can be smaller than one in the case when $n \leq p = |V|$.

As a result many authors have proposed estimation procedures for concentration graphs by checking for low order conditioning (for example Magwene and Kim (2004), Castelo and Roverato (2006), Wille and Bhlman (2006), Malouche and Sevestre (2008)). These approaches aim to discover conditional independences given a certain fixed number of variables. Generally this fied number is very low. Wille and Bühlman (2006) consider one variable, Friedman *et al.* (2000) consider two variables... The assumption that allows them in such procedures to estimate the true, but unknown graph, is to impose the perfect or faithfulness assumption. Indeed, the faithfulness assumption and perfect Markovianity are two hypothesis which are commonly used in estimation procedures for graphical models.

First let us briefly state these two hypothesis. Assume that the graph associated with a given multivariate probability distribution has a set of vertices that corresponds to the

random variables in the random vector. Each vertex corresponds to one such random variable. A distribution is called *faithful* to a graph if any conditional independence statement in the distribution can be represented by a *separation* statement in a graph. The perfect Markovianity assumption means that no other conditional independences exist in the distribution than the ones given by the separation statement appearing on the graph. So for example, for the faithful hypothesis, we can find separations that do not correspond to any conditional independences in the graph. This is not exactly the case for the *perfect* Markovianity hypothesis. Because with perfect Markovianity there is a one-to-one association between separations in the graph and conditional independences in the probability distribution.

In the case of multivariate Gaussian distributions these two properties are rather similar. The precision or inverse covariance matrix reflects conditional independences in multivariate Gaussian distributions. The concentration graph model associated with a given multivariate Gaussian distribution is constructed by looking through the pattern of zeros in the precision matrix. The separation statement in this graph always leads to a conditional independence statement in the generating probability distribution. So if the probability distribution here is *faithful* to that graph, it is then systematically *perfectly* Markov to its concentration graph.

In this note we study the theoretical aspects of the *perfect* Markovianity assumption in general and in the special case of multivariate Gaussian distributions. Though these assumptions are widely used in current statistical estimation procedures, their validity has been questioned in very recent work and is widely perceived as too restrictive. Obtaining a better understanding of this assumption and checking to see if it is valid forms the basis of our current work. We prove that there is a simple and elegant way to know from the pattern of zeros in the covariance matrix if the Gaussian distribution is not *perfectly* Markov to its concentration graph model. We prove that if all the rows, and by symmetry, the columns contain at least two zeros coefficient in the covariance matrix, the corresponding probability distribution cannot be *perfectly* Markov to its concentration graph.

Bibliography

[1] Buhl, S. L., 1993. On the existence of maximum likelihood estimators for graphical gaussian models. Scan. Journal of Statistic. 20, 263–270.

[2] Castelo, R., Roverato, A., 2006. A robust procedure for gaussian graphical models search for microarray data with p larger than n. Journal of Machine Learning Research 57, 2621–2650.

[3] Edwards, D., 2000. Introduction to graphical modelling. Springer texts in statistics.

[4] Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using bayesian networks to analyse expression data. J. Comput. Biol. 7(3-4), 601–620.

[5] Lauritzen, S. L., 1996. Graphical Models. New York : Oxford University Press.

[6] Magwene, P., Kim, J., 2004. Estimating genomic coexpression networks using first-order conditional independence. Genom Biol. 5(12).
[7] Malouche, D., Sevestre-Ghalila, S., 2008. Estimating high dimensional faithful gaussian graphical models by low-order conditioning. Proceeding, of 26th IASTED International Multi-Conference on Applied Informatics, Artificial Intelligence and Applications 595-025, 1–6.
[8] Toh, H., Horimoto, K., 2002. Inference of genetic network by combined approach of cluster analysis and graphical gaussian modelling. Bioinformatics 18(2), 287–297.
[9] Wille, A., Bühlman, P., 2006. Low-order conditional independence graphs for inferring genetic network. Statistical Applications in Genetics and Molecular Biology 5, 1–32.