# COMP 543, Tools and Models for Data Science

Chengyin Liu, cl93

## *Research #3, EM algorithm*

This paper [1] was written by Nir Friedman, who described Bayesian Structural EM, an algorithm for learning the structure of a network from incomplete data. In his previous paper [2] in 1997, Friedman introduced the Structural EM algorithm for searching over model structures in the presence of missing values or hidden variables. However, that algorithm is limited to use score functions that approximate the Bayesian score instead of dealing directly with Bayesian model selection. Therefore, he extended structural EM algorithm to directly optimize the Bayesian score in this paper [1] in 1998.

As it follows the basic intuition of the EM algorithm, it attempts to complete the data using the best estimate of the distribution by far, and then performs a structural search using the same procedure for complete data. To maximize the expected Bayesian score at each iteration, this algorithm runs the EM algorithm to compute the maximum a posteriori (MAP) parameters for the current Bayesian Network, followed by searching over the Bayesian models. The search procedure in Bayesian Structural EM algorithm exploits the decomposition properties of Bayesian Networks and its convergence is proved by Friedman in his paper.

Based on this algorithm, J.M. Pena proposed an improvement [3] in 2000. His paper mainly focuses on learning Bayesian Networks for data clustering problem. The key idea is to use the BC + EM method instead of only using the EM algorithm. Here the Bound and Collapse (BC) method [4] is a deterministic method to estimate conditional probabilities from incomplete databases. This paper uses BC method to bound the set of possible estimates by computing the minimum and the maximum estimate that would be obtained from all possible completions of the data. And then it collapses the bounds into a unique value by a convex combination of the extreme points with weights depending on the assumed pattern of missing data. The author applied the BC method before the EM algorithm is executed to improve the MAP computation step in the original Structural EM algorithm. As expected, the resulted Bayesian Structural BC + EM algorithm exhibits a faster convergence rate and a more efficient and robust behavior than the Bayesian Structural EM algorithm.

**References:**

[1] Friedman, Nir. "The Bayesian structural EM algorithm." Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998.

[2] Friedman, Nir. "Learning belief networks in the presence of missing values and hidden variables." ICML. Vol. 97. No. July. 1997.

[3] Peña, José M., Jose Antonio Lozano, and Pedro Larrañaga. "An improved Bayesian structural EM algorithm for learning Bayesian networks for clustering." Pattern Recognition Letters 21.8 (2000): 779-786.

[4] Ramoni, Marco, and Paola Sebastiani. "Learning Bayesian networks from incomplete databases." Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1997.