Chapter 8

Conclusion

The subject of this thesis has been the tree as a tool for statistical multivariate modeling. A tree, like any graphical model, has the ability to express the dependencies between variables (by means of the graph) separately from the detailed form of these dependencies (contained in the parameters), a property that makes it an excellent support for human intuition and allows the design of general inference and learning algorithms.

Trees are simple models: this is especially evident when one examines the algorithm that fits a tree to a given distribution. All the information about the target distribution that a tree can capture is contained in a small number (at most $n-1$) pairwise marginals. Simplicity leads to computational efficiency. Efficient inference is a direct consequence of the fact that trees are decomposable models with small clique size, but the existence of an efficient algorithm for learning the tree structure is a property that makes trees quasi-unique in the graphical models family. As this thesis shows, the limitation in the trees' modeling power can be overcome by combining them into mixtures. The number of trees in the mixture is a smoothing parameter that allows one to control the complexity of the resulting model.

From the point of view of their ability to graphically represent independencies, mixtures are not strictly speaking graphical models because the independence relationships between the visible variables cannot be determined from the graph topology alone. In other words, the dependency structure is not separated from the parametrization as it is for other classes of belief networks. But from a computational point of view, mixtures of trees are tributary to the belief network perspective and methods. They inherit the computational properties of trees, both in terms of inference and of learning from data, to such an extent that in fact all mixtures of trees algorithms in this thesis are combinations and modifications of the corresponding algorithms for trees.

The learning algorithm for mixtures of trees introduced here is a version of the well known EM algorithm. As such, it converges to a local optimum only. This property is a characteristic of mixtures in general, rather than of the present model. It is also an illustration of the fact that in the search for the best model structure there is no free lunch: just as learning optimal graphical model structure requires brute-force search, learning mixtures suffers from the local optima problem.

Therefore, it is sensible to ask: when should one class of models be chosen instead of the other? A rule suggested by the above discussion on structure search is: Use graphical models for those problems where there is significant knowledge about the dependency structure of the data, so that the structure search is minimal or not necessary; when there is little or no prior knowledge about the structure of the dependencies take mixtures of trees as the preferred candidate. And if the structure is unknown, but simple, like in the splice junction example, a single tree may suffice for the task. The reader should keep in mind however that the final answer will depend on the properties of the actual data.

The thesis has also examined the use of the mixture of trees learning algorithm as a heuristic method for hidden variable discovery. The research is motivated in part by the cases in science and everyday life when identifying a hidden variable enables one, by providing a computationally simple model, to change one's way of thinking (like diseases, hidden variables that are at the core of medical knowledge). A second motivation was seeing this problem as an instance of a more general divide and conquer approach to the problem of structure learning. Because structure search requires model selection and model validation, this thesis has explored these topics as well. It has devised and tested as a model selection criterion an empirical measure of a model's description length. On the model validation side it has introduced a new distribution free independence test.

Making the learning scale better with the dimensionality of the domain by taking advantage of certain properties of the data is the last, but not the least of the research topics that this work is concerned with. The property it was built onto is data sparsity, loosely defined as low entropy in the marginal distributions of the individual variables. The algorithms introduced reduced running time by up to 3 orders of magnitude, mainly by intelligently reorganizing the data in a manner that exploits sparsity. This underlines the importance of combining the graphical models perspective with other data mining techniques in enabling us to conquer the high dimensional domains that exist out there.

After devoting a whole thesis to trees, one may want to ask: are they unique in the properties and advantages demonstrated here? Are there other classes of graphical models that can be learned efficiently from data? There is one other such class. It is the class of Bayes nets for which the variables' topological ordering is fixed and the number of parents of each node is bounded by l. For this class the optimal model structure for a given target distribution can be found in $\mathcal{O}(n^{l+1})$ time by a greedy algorithm. These models and trees are instances of *matroids* [71]. The matroid is the unique algebraic structure for which the "Maximum Weight" problem, corresponding to the Maximum Weight Spanning Tree problem, is solved optimally by a greedy algorithm. Therefore, any class of graphical models that were also matroids would have a structure learning algorithm that is better than brute-force. Finding such classes is an open problem.

By introducing and studying the mixture of trees, this thesis has extended the scope of graphical modeling for unsupervised learning. At the same time, it has aimed to demonstrate that the potential of this exciting field of research is far from being exhausted.