11-1-2012

Graphical Modeling for High Dimensional Data

Munni Begum
Ball State University, Muncie, IN

Jay Bagga
Ball State University, Muncie, IN

C. Ann Blakey

Graphical Modeling for High Dimensional Data

Munni Begum   Jay Bagga   C. Ann Blakey
Ball State University,
Muncie, IN

With advances in science and information technologies, many scientific fields are able to meet the challenges of managing and analyzing high-dimensional data. A so-called large p small n problem arises when the number of experimental units, n, is equal to or smaller than the number of features, p. A methodology based on probability and graph theory, termed graphical models, is applied to study the structure and inference of such high-dimensional data.

Key words: High dimensional data, graphical Markov models, conditional independence, Markov properties, chain graphs.

Introduction

Graphical models are the result of a marriage between probability distribution theory and graph theory; these models have been used to study the associations among stochastic variables for decades in many disciplines. Graphical model methodologies evolved through a blend of statistical techniques: log-linear and covariance selection models with constructs of path analysis and the concept of conditional independence (Whitaker, 1990; Edwards, 2000).

Munni Begum is an Associate Professor of Statistics in the Department of Mathematical Sciences. Her research interests include biostatistics methods and applications, computational statistics, graphical models, Bayesian inference and longitudinal and survival data analysis. Email her at: mbegum@bsu.edu. Jay Bagga is a Professor of Computer Science at Ball State University. His research interests include bioinformatics, bioinformatics algorithms, graphical models, graph theory and graph algorithms. Email him at: jbagga@bsu.edu. C. Ann Blakey is an Associate Professor of Genetics in the Department of Biology. Her research interests include molecular genetics, genomics and bioinformatics, biocomputational population genetics modeling and applications. Email her at: ablakey@bsu.edu.

Classical examples of graphical model applications include: fitting complex patterns of associations among the factors cross-classifying multidimensional contingency tables, and studying relationships among variables using their covariance structure. The current state of the science of this area includes general methodology on the structural properties of graphical models suggested by the conditional independence and Markov properties.

Conditional independence and Markov properties of graphical models are keys to developing methodologies for high dimensional data analysis in a growing number of computational science fields. Structural learning and computational techniques/algorithms with running time and space complexity are of significant interest. This article outlines a method to address the challenge of making efficient statistical inferences with high dimensional data using the elegant features of graphical models.

Dimension Reduction Using Regression and Classification

Advances in science and information technology have allowed many scientific fields, such as bioinformatics, computational biology, medicine, pharmacology and toxicology to produce high-dimensional data at an astounding rate. The common scenario of a small number of features p from a large number of experimental units n has resulted in what is termed a large p

small *n* problem, as the number of experimental units *n* is equal to or even smaller than the number of features (or parameters) *p*. To address this problem there has been a surge in research activities offering data reduction methods and statistical inference.

Modern methods for subset selection include the least absolute shrinkage and selection operator, or LASSO (Tibshirani, 1996) method, which provides a sparse solution by picking influential regression coefficients and shrinking others to zero. The LASSO method reduces the dimension from *p* to *k* < *p, k* being a subset of the features. A major concern with this method – as well as other related methods such as Garotte (Breiman, 1993) and ridge regression – is whether they successfully identify the correct subset of non-zero regression coefficients. Thus, the question remains as to whether it is possible to achieve a proper projection of the coefficient matrix onto a computationally feasible lower dimension. Even with a substantial dimension reduction, questions still remain as to whether sufficient reduction was achieved.

The idea of sufficient dimension reduction in regression is addressed along the similar line of Fisher's sufficient statistics. Inverse regression (Cook, 2007) based on the principle component regression model has been applied to achieve a sufficiently reduced subspace of predictor variables. If *X* is a predictor vector in $\mathfrak{R}^p$ and *Y* is the response variable, then a sufficient reduction R: $\mathfrak{R}^p$ $\mathfrak{R}^q$, $q \leq p$ implies that at least one of the statements (1) *X*| (*Y, R(X)*) ~ *X*|*R(X)*, (2) *Y*|*X*~ *Y*|*R(X)* or (3) $Y \perp\!\!\!\perp X|R(X)$ holds (Cook, 2007). Whereas statements (i) and (ii) correspond directly to inverse and forward regression respectively, statement (iii) connotes the conditional independence of *Y* and *X* given the reduced predictor subspace, which is the basis for graphical models. This also implies no loss of relevant information under the sufficient reduction of predictor space. Sparse additive models (Ravikumar, et al., 2007) based on the so-called generalized additive models (Hastie & Tibshirani, 1990) and Bayesian additive regression trees (Chipman, et al., 2009) are other modern data reduction methods in linear and generalized linear regression problem settings.

Dimension Reduction using Graphical Models

Graphical models can be used as efficient tools to investigate the dependence structure of a large number of attributes. Probabilistic expert systems and Bayesian networks (Neapolitan, 1990, 2004, 2009; Cowell et al., 1999) based on directed acyclic graphs are commonly used graphical models in medical diagnosis, disease spread modeling and gene interaction and protein interaction networks. Advancements in the mathematical theory of general graphical models over the past decade through the pioneering work of Lauritzen (1996), Wermuth and Lauritzen (1983), Frydenberg (1990), Andersson (1993, 1995, 1997), Madigan (1995, 1997) and Perlman (1993, 1995, 1997) among others, facilitates the development of a general methodology for practical problems in diverse scientific fields.

Graphical models are a flexible class of models based on both graph and probability theory and can capture complex dependence structure among a large number of stochastic variables efficiently. Although the general methodology is well developed, only a handful of these methods are implemented in practice. There is a need for addressing the computational aspects of these models under a general framework.

This research is based on the challenges of analyzing a large volume of high dimensional molecular interaction data. For example, the interaction between genes, proteins and metabolites are the focus of such emerging fields as transcriptomics, proteomics and metabolomics. A complete biological network consists of all of these interacting components. To understand this complex system it is necessary to apply the divide and conquer rule: break the system into small parts and map out the interactions. Each of these smaller components may be regarded as a unique complex network; thus gene, protein and metabolic interaction networks can be studied under a single framework. Such an endeavor will help address the challenges of high dimensional data analysis and statistical inference.

Preliminaries on Graphical Models

The fields of graph theory and probability theory are well developed. Graph theory is generally studied as a branch of discrete and combinatorial mathematics, however, graph theory and graph algorithms provide an applicable framework in many fields including computer science, mathematical and statistical sciences, biological and chemical sciences and several branches of engineering. The notion of graph has been tied with conditional independence among stochastic variables and their Markov properties. In graphical modeling and localized computations for probabilistic inference, Markov properties play a fundamental role. General chain graphs and their specializations, directed and undirected graphs, each have different types of Markov properties and conditional independence is a common theoretical tool to investigate these fundamental properties of a class of graphs.

In its simplest form, a graph $G = (V, E)$ constitutes a finite set of vertices $V = \{1, 2, \ldots, v\}$ and a set of edges $E \subseteq V \times V$. Each edge is thus a pair of vertices $(u, v) \in E$ that incorporates a relationship between two vertices. In a graphical model, the vertices may represent discrete or continuous variables and the edges, which may be undirected or directed, represent conditional dependence. A directed graph $G_D = (V, E)$ contains only directed edges drawn as arrows, where V is a set of vertices and E is a set of ordered pair vertices. Directed graphs, with no directed cycles, are known as directed acyclic

graphs (*DAGs*) and play a significant role in causal inference. In an undirected graph $G_U = (V, E)$, the edges are undirected and are used mainly to study the association among attributes. Chain graphs (Lauritzen, 1996) have both directed and undirected edges. For a chain graph $G_{ch} = (V, E)$, the vertex set V can be partitioned into numbered subsets that form a dependence chain $V = V_1 \cap V_2 \cap \ldots \cap V_T$, such that all the edges between vertices in the same subset are undirected and all edges between different subsets are directed, pointing from a set with lower number to the one with higher number. Figure 1 illustrates directed, undirected and chain graphs (Lauritzen, 1996).

The structural properties of general chain graph models and specialized undirected and directed acyclic graph models are of great interest. Theoretical tools to study the structural properties of a class of graphs are their corresponding Markov properties and characterization of their Markov equivalent classes. An important fact of conditional independence properties in localized computations is that these enable factorization of the joint probability distribution of the random variables associated with the nodes of a graph. Following the notations of Cowell, et al. (1999), let X_v, $v \in V$ be a collection of random variables, taking values in probability spaces X_v, $v \in V$, and let **B** be a collection of subsets of V. For $B \in$ **B**, if $a_B(x)$ denotes a non-negative function that depends only on $x_B = (x_v)_{v \in B}$, then a joint

Figure1: Graphical Model Illustrations

| (a) Directed Graph | (b) Undirected Graph | (c) Chain Graph |
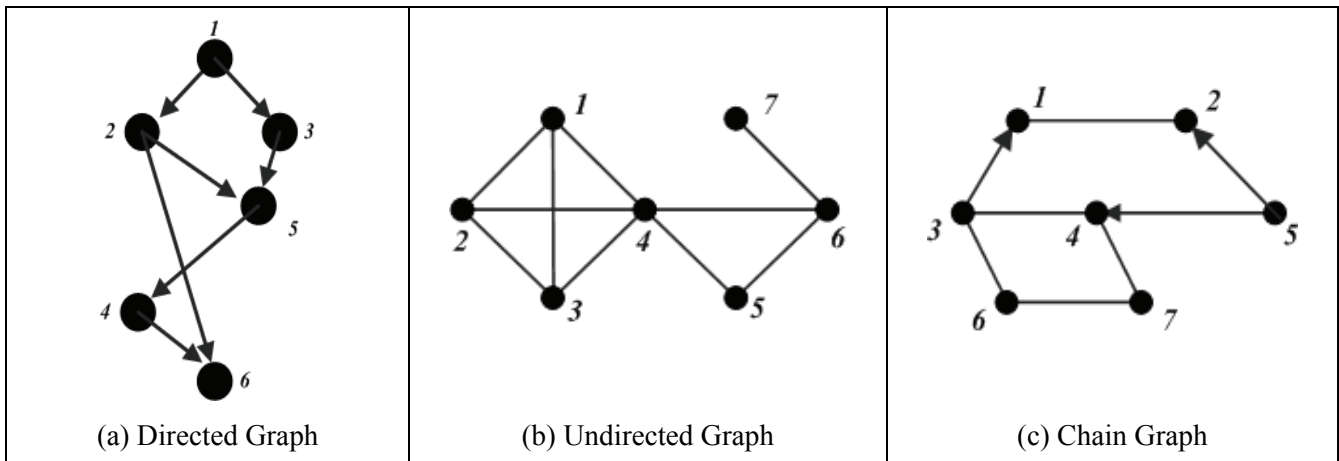
distribution P for X is **B-** hierarchical if its probability density p factorizes as

$$p(x) = \prod_{B \in \mathbf{B}} a_B(x). \qquad (1)$$

(Cowell, et al., 1999).

This factorization holds only when **B** is a complete subset of the underlying graph. For an undirected graph $G_U = (V, E)$, and a collection of random variables X_v, $v \in V$ taking values in probability spaces X_v, $v \in V$, the joint probability density $p(x)$ is $\mathbf{C}$ - hierarchical where $\mathbf{C}$ is the set of cliques of G_U. In this case $p(x)$ factorizes as,

$$p(x) = \prod_{C \in \mathbf{C}} {}_C(x_C), \qquad (2)$$

where the function ψ_C is referred to as factor potential of the probability measure P on $\mathbf{C}_v$.

A probability distribution P is said to admit a recursive factorization according to a directed acyclic graph $G_D = (V, E)$ if the joint density $p(x)$ factorizes as,

$$p(x) = \prod_{v \in V} p(x_v \mid x_{pa(v)}), \qquad (3)$$

where $pa(v)$ is the set of parents of the vertex v. If $(u, v) \in E$, but $(v, u) \notin E$, then u is a parent of v and the set of all parents of v is denoted by $pa(v)$. Recursive factorization according to G_D implies $\mathbf{C}$ hierarchical factorization according to the corresponding undirected moral graph of G_D, denoted as $G_D{}^m$. The moralization process of a directed acyclic graph involves adding undirected edges between all pairs of parents of each vertex which are not already joined and then making all edges undirected (Lauritzen, 1996). For a chain graph $G_{ch} = (V, E)$, with dependence chain $V = V_1 \cap V_2 \cap ... \cap V_T$, the joint density $p(x)$ factorizes as

$$p(x) = \prod_{t=1}^{T} p(x_{V_t} \mid x_{C'_{t-1}}), \qquad (4)$$

where C'_t are the concurrent variables defined as $C'_t = V_1 \cap V_2 \cap ... \cap V_t$. If $B'_t = pa(V_t) = bd(V_t)$, then the above factorization reduces to

$$p(x) = \prod_{t=1}^{T} p(x_{V_t} \mid x_{B'_t}). \qquad (5)$$

For an undirected graph the parent set of a vertex v becomes the neighbor set $nb(v)$. For a chain graph, $bd(v)$ is the set of parents and neighbors of the vertex v. This factorization takes an identical form to that of a directed acyclic graph due to the fact that a chain graph forms a directed acyclic graph of its chain components. One drawback of this representation is that the factorization does not reveal all conditional relationships. To investigate the relationships that are not revealed, if an undirected graph G^*_{ch} with vertex set $V_t \cap B'_t$ is considered, then, for a chain graph, the joint density of a collection of discrete random variables X_v factorizes as,

$$p(x) = \prod_{t=1}^{T} \frac{p(x_{V_t \cup B'_t})}{p(x_{B'_t})}, \qquad (6)$$

and each of the numerators factorizes on the graph G^*_{ch} (Lauritzen, 1996). In addition, if a density $p(x)$ factorizes as in (6), it also factorizes according to the moral graph G^m_{ch} (Lauritzen, 1996).

Associated with a graph G, there are primarily three Markov properties: pairwise, local and global. A probability measure P on X is said to follow the pairwise Markov property relative to G if, for any pair (u, v) of non-adjacent vertices, $u \perp\!\!\!\perp v \mid V/\{u, v\}$. It follows the local Markov property relative to G, if for any vertex $v \in V$, $v \perp\!\!\!\perp V/cl(v) \mid bd(v)$. Here the closure $cl(E)$ of a subset $E \subset V$ is the set of vertices such that $cl(E) = E \cap bd(E)$. Finally, a probability measure follows the global Markov property, relative to G, if for any triple (P, Q, S) of disjoint subsets of V such that S separates P from Q in G so that $P \perp\!\!\!\perp Q \mid S$. Because the global Markov property implies the local, which in turn implies the pairwise Markov property, it is the strongest of the three. A probability distribution

P on a discrete sample space with strictly positive density satisfies the pairwise Markov property if only if it factorizes (Lauritzen, 1996). Thus an undirected graph automatically satisfies the pairwise Markov property.

The joint densities associated to a directed acyclic graph and a chain graph factorize according to their moral graph respectively. In this case, the probability distributions follow the strongest global Markov property, which in turn implies local and pairwise Markov properties. Thus, for a directed acyclic graph and a chain graph, it is important to obtain their corresponding moral graphs as factorization of the joint densities according to these moral graphs to directly imply Markov properties.

There has been extensive research activity in developing and extending the links between graphical structures and conditional independence properties (Andersson, et al., 1995, 1997, 2001, 1993, 2006). A logical research question to explore is whether a probability distribution exists displaying the underlying properties and only the conditional properties displayed by a given graphical representation (Geiger & Pearl, 1990, 1993; Studeny & Bouckaert, 1998). It is important to note that Markov properties and conditional independence lay out one of several possible structures of an underlying graph because there may be more than one graph representing the same conditional independence relations. Over the last few decades, there has been a focus on characterizing the Markov equivalence class of graphs and nominating a natural representative of an equivalence class. After a Markov equivalence class is established and Markov properties are fulfilled, factorization of a joint probability of the attributes under study takes place uniquely, facilitating simplified computation for statistical inference.

Computational Issues in Graphical Models: Graph Structure and Statistical Learning

Graphical models have been studied extensively in order to investigate associations among discrete, continuous and mixed variables. Lauritzen and Wermuth (1989) examined properties of conditional Gaussian (CG) distributions and their applications to conditional Gaussian regression, with emphasis on Markovian properties to attain tractable form in the subsequent analysis of these models. Due to their flexible structures and abilities to represent both structural and associative dependences, chain graph models have been studied extensively in the literature of graphical modeling (Andersson, 1997, 2006; Frydenberg, 1990). The characterization of Markov properties for chain graphs, undirected and directed acyclic graphs, has important implications to the context of factorization of underlying probability models. The Markov properties of a graph directly impact computational issues based on joint likelihood function, however, Markov properties and conditional independence may only provide one of several possible underlying graph structures. Thus, it is important to characterize Markov equivalence class of graphs and nominate a natural representative of an equivalence class.

The characterization of Markov equivalence classes has significant implications to the context of the structure of graphical models. Two graphs are Markov equivalent if they have the same Markov properties. Using results from Verma and Pearl (1991) for directed acyclic graphs, Frydenberg (1990) showed that two chain graphs are Markov equivalent if and only if they have the same skeletons, or the undirected versions, and the same complexes. A complex is a subgraph induced by a set of nodes $\{v_1, v_2, ..., v_k\}$ with $k \geq 3$, whose edge set consists of $v_1 \rightarrow v_2$, $v_{k-1} \leftarrow v_k$, and $v_i \sim v_{i+1}$ for $2 \leq i \leq k-2$.

For a class of Markov equivalent chain graphs, a unique largest chain graph having the maximum number of undirected edges exists. The arrows of this largest graph are present in every other member of the class and thus may be considered as the representative graph of the class. There is no natural representative of an equivalence class within the class of directed acyclic graphs although it can be characterized by what is referred to as its essential graph (Andersson, et al., 1997). The natural representative of an equivalent class of chain graphs is the one with same skeleton in which an edge has an arrow, if and only if at least one member of the equivalence class has that arrow

and none has the reverse arrow (Andersson, et al., 1997).

Alternative graphical representations of conditional independence and Markov properties have been considered in the graph theory literature. Markov equivalence classes for chain graphs, undirected and directed acyclic graphs examined by Lauritzen and Wermuth (1989) and Frydenberg (1990) are referred to as LWF by Andersson, Madigan and Perlman (2001) who considered an alternative Markov property with a new semantics AMP to facilitate a direct mode of data generation (Cox, 1993; Cox & Wermuth, 1993). Andersson and Perlman (1993) showed that for AMP chain graphs, each Markov equivalence class can be uniquely represented by a single distinguished chain graph, the AMP essential graph, which plays a fundamental role in inference and model search. However, the AMP approach does not correspond to factorization of joint density in a straightforward manner; a crucial aspect for computational efficiency (Cowell, et al., 1999). Koster (1996) considered a generalization of chain graphs to reciprocal graphs and Drton (2009) showed that the block recursive Markov property of discrete chain graph models is equivalent to the global Markov property. The practical use of these models lies in developing algorithms for efficient computation characterizing running time and space complexities.

Exact and approximate inference algorithms for graphical Markov models based on independence graphs are proposed to address computational issues. Computational advancement for graphical models, particularly for the probabilistic expert systems evolved through construction of fundamental graph algorithms namely, moralization, triangulation and junction tree. The joint distribution of a graphical model can be represented and manipulated efficiently using a junction tree derived from the original graph. The junction tree algorithm starts with a moralized graph. A directed graphical model can be converted to an equivalent undirected model by the moralization process. The algorithm first selects an elimination order for all nodes and applies a triangulation operator to the moralized graph yielding a triangulated graph, then the triangulated graph creates a data structure

known as a junction tree on which a generalized message-passing algorithm can be defined (Xing, 2004). Figures 2 and 3 show an example of this process (Xing, 2004).

A junction tree possesses a key property of a running intersection, which implies that, when a node appears in any two cliques in the tree it appears in all cliques lying on the path between the two cliques. The running intersection property of the junction tree enables the joint probability distribution to be factorized as,

$$p(x) = \frac{\prod_{C_i \in \mathbf{C}_T} \psi_i(x_{C_i})}{\prod_{S_j \in \mathbf{S}_T} \phi_j(x_{S_j})}, \qquad (7)$$

where $\mathbf{C}_T$ is the set of all cliques in the triangulated graph and S_T is the set of separators spanned by the junction tree (Xing, 2004). A message passing scheme on the junction tree updates the clique potentials $\psi(.)$ and the separator potentials $\phi(.)$ according to the rule,

$$\phi_j^*(x_{S_j}) = \sum_{x_{C_i \setminus S_j}} \Psi_i(x_{C_i}), \quad \Psi_k^*(x_{C_k})$$

$$= \frac{\phi_j^*(x_{S_j})}{\phi_j(x_{S_j})} \Psi_k(x_{C_k}),$$

$$\qquad (8)$$

where x_{Sj} denotes the set of variables separating cliques x_{Ci} and x_{Ck}, and the message being passed from clique i to k via separator j. Running time and space complexity of the junction tree algorithm is determined by the size of the maximal clique in the triangulated graph, which is affected by the choice of elimination order that induces the triangulated graph. Tree width of a graph is the minimum of the maximal clique size among all possible triangulations. Selecting an elimination order that minimizes the maximal clique size is an NP-hard problem for arbitrary graphs. The implementation of this exact inference algorithm based on the junction tree is not efficient – or possible – for graphical models under high dimensional data. Although exact inference algorithms are simple to interpret, their implementation in high

dimensional problems becomes prohibitive due to running time and space complexities.

Computational Issues in Graphical Models: Efficient Learning/Inference Engines, Algorithms and Complexities

Approximate efficient inference algorithms, such as variational approach under a complex scenario, are considered. The approach involves converting the original optimization problem into an approximated optimization problem that is solved for an approximate

solution to the exact inference problem. Given a probability distribution $p(x|\theta)$ that factors according to a graph, the variational methods yields approximations to marginal probabilities by solving an optimization problem exploiting the underlying graphical structure (Xing, 2004). Many graphical models can be naturally viewed as an exponential family of distributions, a broad class of distributions for both discrete and continuous random variables, through the principle of maximum entropy (Wainwright & Jordan, 2008). This principle depends on a
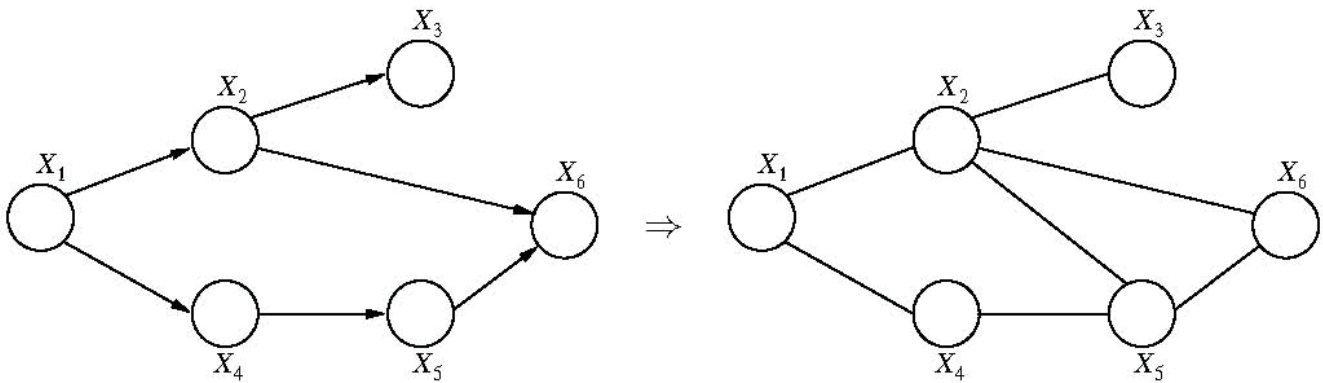
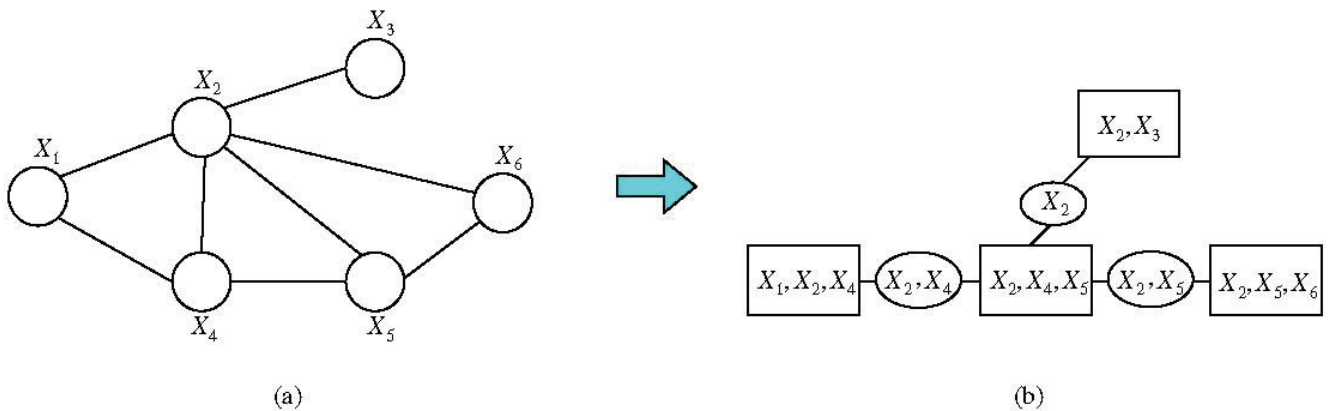Figure2: Moralization of a Directed Graph (Xing, 2004)

Figure3: Triangulated Graph of Figure 2 Directed Graph and Junction Tree

(a)

(b)

(a) Triangulated Graph of Figure 2 Directed Graph   (b) The Junction Tree

functional of the probability density p, absolutely continuous with respect to some measure ν. $H(p)$ is known as Shannon entropy and is defined as,

$$H(p) := -\int_x (\log p(x)) p(x)(dx) \qquad (9)$$

Consider variational inference approaches for the exponential family representations of the graphical models. Approximate inference methods, such as sum-product algorithms, generalized belief propagating methods and generalized mean field inference algorithms are some of the most recent computational methodologies for graphical models in a high dimensional scenario. For variational inference, the exponential family of joint distributions determined by a collection of potential functions or sufficient statistics $\phi = \{\phi_\alpha \in C\}$ is expressed as,

$$p(\boldsymbol{x} | \theta) = exp\left\{\sum_{\alpha \in C} \theta_\alpha \phi_\alpha(\boldsymbol{x}_{C\alpha}) - A(\theta)\right\}, \qquad (10)$$

where C is the set of cliques, C_α is the clique corresponding to the node α, $A(\boldsymbol{\theta})$ is the log partition function or cumulant function defined as

$$A(\theta) = \log \int_{X^m} exp\left\{\sum_{\alpha \in C} \theta_\alpha \phi_\alpha(\boldsymbol{x}_{C_\alpha})\right\} \nu(dx) ,$$

where X^m is a product space for m random variables. The conjugate dual function to $A(\boldsymbol{\theta})$, central to the variational principle, is defined as $A^*(\mu) := sup_{\theta \in \Theta}\{< \mu, \theta > - A(\boldsymbol{\theta})\}$. Here θ and μ represent canonical and mean parameters respectively of the exponential family of distributions. The conjugate dual function A^* takes the form $A^*(\mu) = -H(p_{\theta(\mu)})$, where the functional $H(.)$ is defined as the Shannon entropy of the density $p_{\theta(\mu)}$ given that μ is in the interior of the set of realizable mean parameters M which is defined as,

$$M := \{\mu \in R^d | \exists p \ s.t. \ E_p[\phi(X)] = \mu\}$$

(Wainwright & Jordan, 2008). Here R^d indicates number of elements to be specified in the vector

of sufficient statistics and the variational representation of the log partition function in terms its dual A^* is $A(\theta) = sup_{\mu \in M} \{<\theta, \mu > - A^*(\mu)\}$. Thus, under the variational representation it is necessary to maximize or minimize over the set of M as opposed to the entire parameter space Θ. The optimization problem for the variational representation of specialized graphs such as trees is computationally feasible, however, for a general structure graphical model with a large number of nodes, exact optimization becomes infeasible due to the complexity in characterizing the constraint set M and dual function $A^*(\mu)$. Approximate methods seek approximations to M and $A^*(\mu)$.

Mean-Field Methods as an Approximation to the Exact Variational Principle

In order to implement a variational inferential approach, the nature of the constraint set M and an explicit form for the dual function A^* must be known (Wainwright & Jordan, 2008); this, however, may not be easy to obtain for most practical problems. Mean field approaches permit limiting of the optimization to a subset of distributions, referred to as tractable distributions, for which both M and A^* are relatively simple to characterize (Wainwright & Jordan, 2008). For a graphical model based on a graph $G(V, E)$, the tractability can be obtained in terms of a tractable sub-graph. A sub-graph F is tractable if it is possible to carry out exact calculations on F.

A straightforward example of a tractable sub-graph is the fully disconnected sub-graph $F(V, \varnothing)$ containing all the vertices of $G(V, E)$ but none of the edges. This tractable sub-graph F leads to a product distribution for which computations are easy to carry out. However, completely disconnected sub-graphs do not capture dependencies among vertices, if any. Thus, as opposed to a fully disconnected sub-graph, an arbitrarily structured sub-graph from the given graph $G(V, E)$ is considered in generalized mean field methods. The question then becomes how to select a tractable sub-graph leading to an efficient factorization of the joint probability distribution so that feasible solution set M and the optimizing function A^* can be

characterized with less intensive computational and mathematical background. In addition, a generalized version of the mean field methods to the context of chain graph models is of great importance for practical problems in numerous scientific fields.

Discussion and Future Direction
Statistical Learning on the Underlying Graph Structure from Empirical Data

In order to implement chain graph models to study relationships among stochastic variables in empirical data, consider the exponential family of probability distributions as the distribution of the random variables associated to the nodes of a graphical model. It is relatively straightforward to write the joint probability distribution of a set of discrete random variables utilizing the factorization under a given graphical structure. The factorization of joint distribution of continuous random variables representing the nodes of an underlying graph requires attention and a general framework for the factorization scheme of joint probability distributions of both discrete and continuous random variables using established graph theory properties is of interest for simplified computation.

Gaussian graphical models for continuous variables and the graphical counterpart of log-linear models for discrete attributes are proposed and implemented for empirical model building. For a large volume of attributes, as in biological network data, such as gene-gene interaction networks, gene-protein interaction networks and transcription regulatory networks, as well as network data in other scientific and social science fields, these methods can be computationally prohibitive.

The variational inference approach based on the mean parameterization of the exponential family of distributions and their mathematical properties, such as, convexity and conjugate duality is an efficient inference approach to graphical models. Implementation of these algorithms and complexity are of interest in the contexts of high dimensional gene-gene interaction networks, gene-protein interaction networks and transcription regulatory network data. Structural properties such as connectivity and existence of specific

substructures in the graphical models are of specific interest. It is necessary to identify Markov equivalence classes in order to narrow down possible representations of same conditional independence by many graphical structures.

In particular, this investigation considers when a chain graph G_{ch} is Markov equivalent to some unique undirected graph G_U, decomposable undirected graph and to some directed acyclic graph G_D. The directed acyclic graph models provide a convenient recursive factorization of the joint probability. The likelihood function factorizes and it is possible to implement maximum likelihood methods for estimation of model parameters. These tractable features are also available for decomposable undirected graphs, which are Markov equivalent to some directed acyclic graphs (Andersson, et al., 1997). For a directed acyclic Markov model G_D, the joint density factorizes as

$$p(x \mid \theta) = \prod_{v \in V} p(x_v \mid x_{pa(v)}, \theta^{v, x_{pa(v)}})$$

(11)

where $\theta^{v, x_{pa(v)}}$ is the minimal function of the overall parameter θ for the distribution determining the conditional distribution of $X_v \mid X_{pa(v)} = x_{pa(v)}$. For a complete case, each factor in the likelihood is maximized separately to attain a maximum likelihood estimate of $\theta^{v, x_{pa(v)}}$. For an incomplete case consider the expectation maximization (EM) algorithm. Let $f(x \mid \theta)$ denote the density function of a random variable X that is incomplete except one known function, $Y = g(X)$. Given an initial estimate θ, the E-step requires the current expected value of the log-likelihood function $Q(\theta' \mid \theta) = E_\theta \{log\ f(X \mid \theta') \mid g(X) = y\}$. The M-step maximizes Q over θ' yielding the next estimate. The algorithm alternates between these two steps until convergence is attained. The evidence propagation or message-passing on the junction tree can be exploited to perform the E-step of an EM algorithm for a discrete directed acyclic graph model with missing observations (Lauritzen, 1995). Gradient-descent search near the maximum can be considered to speed up the convergence of an expectation-maximum (EM)

algorithm in a graphical model with incomplete data.

A general factorization scheme for the joint probability distribution in the exponential family enabling tractability in subsequent statistical computation sets the foundation for efficient computation in graphical models. Searching within a collection of candidate models for one or more graphical structures consistent with the conditional independence relationships suggested by data follows; the point is to assess the adequacy of a single candidate graphical model as the so-called true objective process of generation of empirical data. In order to narrow down the possible high dimension of the space of the graphical structures, the Markov equivalence classes of graphical structures, identification of a unique graphical model as the probability model and checking identifiability of the model parameters are essential. Either the likelihood-based or the Bayesian methods can be implemented to address the estimation and model search problem. Complete case data are addressed through maximum likelihood estimation or a Bayesian updating scheme; incomplete case data are addressed through the EM algorithm coupled with gradient search methods for estimation using likelihood- and sampling- based methods using a Bayesian approach respectively.

Model Selection, Diagnostics and Checking Models against Data

A Markov equivalence class insures only proper graphical structure. The properties of the joint probability distribution of the variables must be inferred from the graphical structure and the conditional independence relationships suggested by the empirical data. According to the semantics of machine learning and data mining, unsupervised learning methods for model selection, diagnostics and model checking against data can be employed. In low dimensional problems with a number of variables $p\ q\ 3$, effective nonparametric methods are used for density estimation solely from the data (Silverman, 1986). However, these methods are not applicable in high-dimension problems due to the curse of dimensionality.

Testable hypotheses, based on prior knowledge and expert opinion in the scientific field along with corresponding testing principles should be developed to address the graphical model selection problem. Efficient computational algorithms along with running time and space complexities must be formulated. Diagnostics in statistical modeling address outlier detection problems and development of robust methods against outliers. Outlier identification in high dimensional problems is an active research area where robust principal component analysis, k- nearest neighbor, local outlier factor and other distance and density based methods are commonly used. Future research interests should center on addressing these important statistical problems for high-dimensional data.

Conclusion

Graphical models originated as the marriage between graph and probability theories and are appealing methods for studying conditional (in)dependences among a large number of attributes in many scientific fields. Markov properties of various graphical models, directed, undirected and more general chain graph models, lead to efficient factorization of joint probability distributions of multivariate random variables. An explicit form of a joint distribution may not be known for many random variables, except some arbitrary dependence structure.

Graphical modeling is an efficient tool for studying dependence structure among an arbitrary number of random variables without specifying their joint distribution. This article described essential properties of graphical models that lead to factorization of a joint distribution. An exponential family representation of graphical models was demonstrated for a broad class of distributions of discrete and continuous random variables. Exponential family representation is essential for formulating approximate inference algorithms such as mean field algorithms. It was also indicated that studies regarding unique graph structure through a Markov equivalence class of graphs for specialized undirected, directed and general chain graphs is an area for future research. Finally, a graphical model derived from a unique graph structure illuminated the relationship among the attributes under study.

References

Andersson, S. A., Madigan, D., & Perlman, M. D. (1995). *A characterization of Markov equivalence classes for acyclic digraphs: Technical report 287*. University of Washington, Seattle: Department of Statistics.

Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). On the Markov equivalence of chain graphs, undirected graphs and acyclic digraphs. *Scandinavian Journal of Statistics*, *24*, 81-102.

Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, *25*, 505-541.

Andersson, S. A., Madigan, D., & Perlman, M. D. (2001). Alternative Markov Properties for chain graphs. *The Scandinavian Journal of Statistics*, *28*, 33-85.

Andersson, S. A., & Perlman, M. D. (1993). Lattice models for conditional independence in a multivariate normal distribution. *The Annals of Statistics*, *21*, 1318-58.

Andersson, S. A., & Perlman, M. D. (2006). Characterizing Markov equivalence classes for AMP chain graph models. *The Annals of Statistics*, *34*, 939-972.

Breiman, L. (1993). Better subset selection using the non-negative garotte. *Technical Report*. University of California, Berkeley.

Chipman, H. A., George, E. I., & McCulloch, R. (2009). *Bayesian additive regression trees*. Preprint, NI09002-SCH. Cambridge, MA: Isaac Newton Institute.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, *22*, 1-43.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. New York: Springer.

Cox, D. R. (1993). Causality and graphical models. *Bulletin of the International Statistical Institute, Proceedings of the 49th Session*, *1*, 363-372.

Cox, D. R., & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, *8*, 204-218; 247-277.

Drton, M. (2009). Discrete chain graph models. *Bernoulli*, *15(3)*, 736-753.

Edwards, D. (2000) *Introduction to graphical modeling*. New York, NY: Springer-Verlag, Inc.

Frydenberg, M. (1990).The Chain Graph Markov property. *Scandinavian Journal of Statistics*, *17*, 333-353.

Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *The Annals of Statistics*, *18*, 790-805.

Geiger, D., & Pearl, J. (1990). On the logic of causal models. In *Uncertainty in Artificial Intelligence*, R. D. Shachter, T. S. Levit, L. N. Kanal, & J.F. Lemmer, Eds., 4, 3-14. Amsterdam, The Netherlands: North-Holland.

Geiger, D., & Pearl, J. (1993). Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, *21*, 2001-2021.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London, UK: Chapman & Hall/CRC.

Koster, J. T. A. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, *24*, 2148-2177.

Lauritzen, S. L. (1996). *Graphical models*. Oxford, England: Oxford Science Publications.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, *19*, 191-201.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, *17*, 31-57.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems*. New York, NY: John Wiley and Sons.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Upper Saddle River, NJ: Prentice Hall.

Neapolitan, R. E. (2009). *Probabilistic methods for bioinformatics*. Burlington, MA: Morgan Kaufmann.

Ravikumar, P., Liu, H., Lafferty, J., & Wasserman, L. (2007). SpAM: sparse additive models. In *Advances in neural information processing systems*, J. C. Platt, D. Koller, Y. Singer & S. Roweis, Eds., 1201-1208. Cambridge, MA: MIT Press.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London, England: Chapman and Hall.

Studeny, M., & Bouckaert, R. R. (1998). On chain graph models for description of conditional independence structures. *The Annals of Statistics*, *26*, 1434-1495.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Sociecty*, *Series B*, *58*, 267-288.

Verma, T., & Pearl, J. (1991). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence*, R. D. Shachter, T. S. Levit, L. N. Kanal, & J.F. Lemmer, Eds., 255-268. Amsterdam, The Netherlands: North-Holland.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*, 1-305.

Wermuth, N., & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, *70*, 537-552.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York, NY: John Wiley and Sons.

Xing, P. (2004). *Probabilistic graphical models and algorithms for genomic analysis*. PhD. Dissertation, Department of Computer Science, University of California, Berkeley.