Directed acylic graphical models and parameterizations of graphical models

Tianfang Zhang

Mathematical Statistics, Department of Mathematics, KTH

RaySearch Laboratories

April 16, 2021

Outline

- Recap of conditional independence and undirected graphical models
- Parameterizations of undirected graphical models
- Directed acyclic graphical models
- Parameterizations of directed acyclic graphical models

Conditional independence and undirected graphical models

Conditional independence and undirected graphical models

Conditional independence

- In short, we say that x and y are *conditionally independent* given z if

$$p(x, y \mid z) = p(x \mid z)p(y \mid z).$$

  This may be understood as x and y not providing any further information about each other when already knowing z.

- As a concrete example, suppose that the sample $\{x_i\}_i$ is drawn from a normal distribution $\mathrm{N}(\theta, 1)$. We usually (in a *Bayesian* setting) say that the x_i are conditionally independent given the mean θ—that is, the x_i "communicate" through θ. Indeed, we have the factorization

$$p(\{x_i\}_i \mid \theta) = \prod_i p(x_i \mid \theta) = \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \theta)^2}{2}\right).$$

Conditional independence and undirected graphical models

Undirected graphical models

- In certain settings, it is useful to represent conditional independence relations for a set $\{x_i\}_i$ of random variables by an undirected graph. This is known as an *undirected graphical model* or *Markov random field*. Each variable is represented as a vertex.
- The *pairwise Markov property* is the assertion that each pair of non-adjacent variables are conditionally independent given all other variables. The *local Markov property* is the assertion that given its neighbors, a variable is conditionally independent of all other variables. The *global Markov property* is the assertion that $\{x_i\}_{i\in A} \perp \{x_i\}_{i\in B} \mid \{x_i\}_{i\in C}$ if and only if C separates A and B in the graph.
- The properties are in general ordered from weakest to strongest, but equivalent for positive distributions.

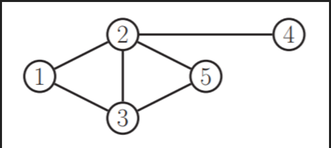

Figure 1: Example of an undirected graphical model.

Conditional independence and undirected graphical models

Undirected graphical models

- As an example, consider measuring some quantity in the ground (e.g. pH value) at different locations on a site. The measurement values should then be dependent on each other, with correlations higher the closer the locations are. Suppose, for simplicity, that the locations are distributed on a grid. Similar setups are common in *spatial statistics*.
- One way of constructing a conditional independence model is through an undirected graphical model. In particular, we may construct it according to below:

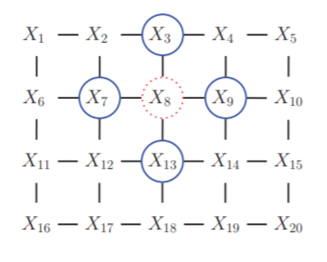

Figure 2: Undirected graphical model according to a rectangular grid.

Parameterizations of undirected graphical models

Parameterizations of undirected graphical models

Preliminaries

- Let $G = ([m], E)$ be an undirected graph with vertices $[m] = \{1, \ldots, m\}$ and edges E. We consider an undirected graphical model of $\{x_i\}_{i=1}^m$ associated with G.
- A *clique* $C \subseteq [m]$ is a collection of fully connected vertices, i.e. $(i, j) \in E$ for all $i, j \in C$. The set of maximal cliques is denoted $\mathcal{C}(G)$.
- For each $C \in \mathcal{C}(G)$, we introduce a *potential function* $\psi_C(\cdot \mid \theta)$ of $\{x_i\}_{i \in C}$ required to be continuous and such that $\psi_C(\{x_i\}_{i \in C} \mid \theta) \geq 0$ everywhere.
- The *parameterized undirected graphical model* consists of all joint likelihoods on the form

  $$p(\{x_i\}_{i=1}^m \mid \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}(G)} \psi_C(\{x_i\}_{i \in C} \mid \theta),$$

  where

  $$Z(\theta) = \int \prod_{C \in \mathcal{C}(G)} \psi_C(\{x_i\}_{i \in C} \mid \theta) \prod_{i=1}^m dx_i$$

  We say that the likelihood *factorizes* according to G if it can be written in the above form.
- The *Hammersley–Clifford theorem* states that a positive density satisfies the Markov properties on G if and only if it factorizes according to G. This is fundamental for working with parameterizations of undirected graphical models.

Parameterizations of undirected graphical models

Discrete models

- Suppose that each x_i is one-dimensional and takes values in $[r_i]$, so that the joint state space is $\mathcal{R} = \prod_{i=1}^{m}[r_i]$. The graphical model associated with G is a subset of the simplex $\Delta_{\mathcal{R}-1}$.
- The Hammersley–Clifford parameterization is on the following monomial form:

$$p(x_1 = i_1, \dots, x_m = i_m \mid \theta) = \phi_{i_1 \dots i_m}(\theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}(G)} \theta^{(C)}_{i_C},$$

  with $\theta = (\theta^{(C)})_{C \in \mathcal{C}(G)}$ nonnegative.
- The parameterized discrete undirected graphical model associated with G consists of all probability distributions in $\Delta_{\mathcal{R}-1}$ of the form $p(x_1 = i_1, \dots, x_m = i_m \mid \theta) = \phi_{i_1 \dots i_m}(\theta)$. In particular, the positive part is precisely the *hierarchical log-linear* model on the complex $\mathcal{C}(G)$ of cliques.

Parameterizations of undirected graphical models

Discrete models

- Denote by I_G the *toric ideal* of the graphical model at hand. From before, we know that I_G is the ideal generated by the binomials $p^u - p^v$ corresponding to the Markov basis. Let also $V_\Delta(I_G)$ be the variety of I_G in the closed simplex $\Delta_{\mathcal{R}-1}$. We want to compare $V_\Delta(I_G)$ with conditional independence models $V_\Delta(I_\mathcal{C})$, where $\mathcal{C}$ ranges over conditional independence constraints implied by G.
- Let
$$\text{pairs}(G) = \{i \perp j \mid [m]\backslash\{i,j\} : (i,j) \notin E\}$$
and
$$\text{global}(G) = \{A \perp B \mid C : C \text{ separates } A \text{ from } B\}.$$
- It turns out that the following conditions are equivalent:
  1. $I_G = I_{\text{global}(G)}$.
  2. I_G is generated by quadrics.
  3. The ML degree of $V_\Delta(I_G)$ is one.
  4. G is a *decomposable* graph.
- This connects the Hammersley–Clifford parameterization to the global (and also pairwise) Markov property.

Parameterizations of undirected graphical models

Gaussian models

- Again, suppose that each x_i is one-dimensional and that $x = (x_i)_{i=1}^m \sim \mathrm{N}(\mu, K^{-1})$. The likelihood is written as

$$p(x \mid \theta) \propto \exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}} K (x-\mu)\right)$$
$$= \prod_{i=1}^{m} \exp\left(-\frac{K_{ii}}{2}(x_i-\mu_i)^2\right) \prod_{1 \le i < j \le m} \exp\big(-K_{ij}(x_i-\mu_i)(x_j-\mu_j)\big),$$

  with $\theta = (\mu, K)$. In particular, this factorizes into pairwise potentials and according to $G = ([m], E)$ if and only if $K_{ij} = 0$ for all $(i,j) \notin E$.
- In other words, the parameterized Gaussian undirected graphical model consists of the set of pairs $(\mu, K) \in \mathbb{R}^m \times \mathrm{PD}_m$ with $K_{ij} = 0$ for all $(i,j) \notin E$.
- In terms of covariance matrices $\Sigma = K^{-1}$, by the adjoint formula for the inverse, we may obtain a rational parameterization of the covariance matrices satisfying the Markov properties of the graphical model.

Parameterizations of undirected graphical models

Other models

- k-nearest neighbor classification may be extended to a probabilistic setting by modeling distance relations graphically.
- Suppose that we have a dataset $\{(x_i, y_i)\}_{i=1}^m$, where the x_i are covariates and y_i are binary labels. Let $\mathrm{nb}_k(i)$ be the set of the k nearest neighbors to x_i. We may define the joint likelihood as

$$p(\{y_i\}_{i=1}^m \mid \{x_i\}_{i=1}^m, \theta) \propto \exp\left(\frac{\beta}{k}\sum_{i=1}^{m}\sum_{j\in\mathrm{nb}_k(i)} 1_{y_i = y_j}\right),$$

with $\theta = (\beta, k)$. The *full conditionals* are

$$p(y_i \mid \{y_j\}_{j\neq i}, \{x_j\}_{j=1}^m, \theta) \propto \exp\left(\frac{\beta}{k}\left(\sum_{j\in\mathrm{nb}_k(i)} + \sum_{j\,:\,i\in\mathrm{nb}_k(j)}\right) 1_{y_i = y_j}\right).$$

Directed acyclic graphical models

Directed acyclic graphical models

Preliminaries

- Undirected graphical models were useful for variables which could not be arranged into any particular hierarchical order. For many models, however, there is a natural hierarchical order, and this may then be articulated through directed edges.
- In a *directed acyclic graph* (DAG) $G = (V, E)$, the edges are directed and there exists no sequence of vertices $\{v_i\}_{i=1}^n$ such that $(v_1, v_2), \dots (v_{n-1}, v_n), (v_n, v_1)$ are all in E. The set $\mathrm{pa}(v)$ of *parents* of a node $v \in V$ are the nodes w such that $(w, v) \in E$. The set $\mathrm{de}(v)$ of *descendants* comprise the nodes w such that there is a directed path from v to w. The *non-descendants* $\mathrm{nd}(v)$ are all nodes that are not v or a descendant to v.

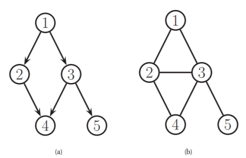

Figure 3: A directed and an undirected graph.

Directed acyclic graphical models

Definition and examples

- For a set $\{x_i\}_{i=1}^m$ of random variables and a DAG $G = ([m], E)$, the *directed local Markov property* associates the conditional independence constraints

$$x_i \perp x_{\mathrm{nd}(i) \setminus \mathrm{pa}(i)} \mid x_{\mathrm{pa}(i)}.$$

  A *directed graphical model* (or *Bayesian network*) of $\{x_i\}_{i=1}^m$ with respect to a DAG G is a graphical model with the local Markov property on G. In many practical examples, the directedness of the edges represent *causal* relationships.
- As an example, we have the (Bayesian) parametric model of $\{x_i\}_{i=1}^m$ being conditionally independent given θ (see figure below). Here, θ may be seen to "cause" the x_i.

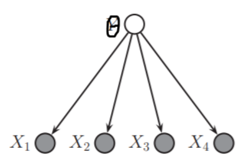

Figure 4: A parametric model as an directed graphical model.

Directed acyclic graphical models

Definition and examples

- Another example is the modeling of variables measured in an intensive care unit. The so called ALARM network (see below) models causal relationships of these variables, 37 in total.

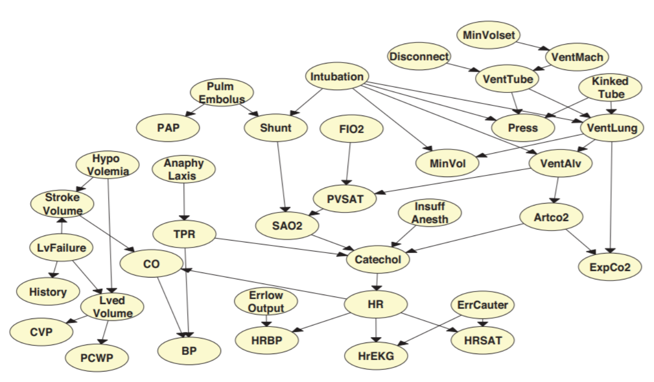

Figure 5: The ALARM network.

Directed acyclic graphical models

Definition and examples

- Yet another example are (feedforward) *Bayesian neural networks*, in which the network weights are regarded as model parameters with priors and posteriors. These are useful for articulating uncertainties in predictions.

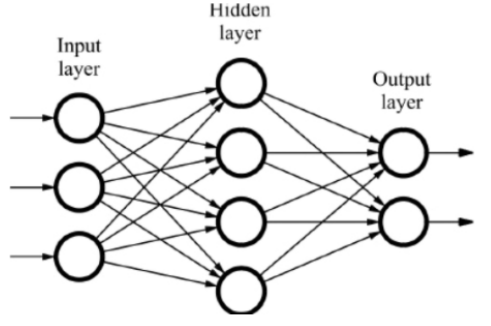

Figure 6: A feedforward neural network.

Directed acyclic graphical models

d-separation

- To understand the conditional independence constraints that the local Markov property implies, we need a more refined notion of separation.
- Let $\mathrm{an}(C) = \{w \in V : \text{there exists a } v \in V \text{ such that } v \in \mathrm{de}(w)\}$ be the set of *ancestors* of a subset $C \subseteq V$. On an undirected path $\pi = (v_0, \ldots, v_n)$, the vertex v_i is a *collider* if the indicent edges are on the form

$$v_{i-1} \to v_i \leftarrow v_{i+1}.$$

- We say that $v, w \in V$ are *d-connected* given a conditioning set $C \subseteq V \backslash \{v, w\}$ if there is a path π from v to w such that
  1. all colliders on π are in $\mathrm{an}(C)$, and
  2. no non-collider on π is in C.

  If A, B, C are pairwise disjoint with A and B nonempty, then C *d-separates* A and B provided that no two nodes $v \in A$ and $w \in B$ are d-connected given C.

Directed acyclic graphical models

d-separation

- With this, we have the *directed global Markov property*, which is the assertion that

$$\{x_i\}_{i \in A} \perp \{x_i\}_{i \in B} \mid \{x_i\}_{i \in C}$$

  for all A, B, C such that C d-separates A and B.
- Analogously to undirected graphical models, it holds that a model for $\{x_i\}$ satisfies the directed local Markov property for a DAG G if and only if it satisfies the directed global Markov property for G.
- An issue with directed graphical models is that two DAGs may possess identical d-separation relations and thus encode the same conditional independence relations. The graphs are then called *Markov equivalent*. One can determine Markov equivalence by the fact that two DAGs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are Markov equivalent if and only if
  1. G_1 and G_2 have the same skeleton, and
  2. G_1 and G_2 have the same collider triplets.

Parameterizations of directed acyclic graphical models

Parameterizations of directed graphical models

Recursive factorization

- Every DAG $G = ([m], E)$ has a *topological ordering*, i.e. a permutation σ of $[m]$ such that the vertices $(\sigma(1), \ldots, \sigma(m))$ are ordered from starting earlier to starting later.
- Writing
$$p(\{x_i\}_{i=1}^m \mid \theta) = \prod_{i=1}^m p(x_{\sigma(i)} \mid \{x_{\sigma(j)}\}_{j=1}^{i-1}, \theta),$$
we may use the directed local Markov property to reduce this to
$$p(\{x_i\}_{i=1}^m \mid \theta) = \prod_{i=1}^m p(x_i \mid \{x_j\}_{j \in \mathrm{pa}(i)}, \theta).$$
This is called the *parametric directed graphical model*, and this factorization is equivalent to satisfying the local or global Markov properties.

Parameterizations of directed graphical models

Discrete models

- Suppose that each x_i is one-dimensional and takes values in $[r_i]$, so that the joint state space is $\mathcal{R} = \prod_{i=1}^{m} [r_i]$. The directed graphical model associated with a DAG G has the parametric form

$$p(x_1 = i_1, \ldots, x_m = i_m \mid \theta) = \phi_{i_1 \ldots i_m}(\theta) = \prod_{j=1}^{m} \theta^{(j)}(i_j \mid i_{\mathrm{pa}(j)})$$

  with the constraints

$$\sum_{k=1}^{r_j} \theta^{(j)}(k \mid i_{\mathrm{pa}(j)}) = 1$$

  for all j and tuples $i_{\mathrm{pa}(j)} \in \mathcal{R}_{\mathrm{pa}(j)}$.
- Denote by $\phi_{\geq 0}$ the restriction of ϕ to nonnegative parameters and $\mathrm{local}(G)$ the conditional independence constraints associated with the local Markov property. We have

$$\operatorname{im} \phi_{\geq 0} = V_{\Delta}(I_{\mathrm{local(G)}}).$$

Parameterizations of directed graphical models

Gaussian models

- To characterize Gaussian directed graphical models, we assume that $\{x_i\}_{i=1}^m$ are ordered from early to late. Let $\epsilon_i \sim \mathrm{N}(\nu_i, \omega_i^2)$ for all $i \in [m]$ independently and construct
$$x_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ji} x_j + \epsilon_i, \quad i \in [m].$$
This is sometimes known as an *autoregressive* model.
- The random vector $x = (x_i)_{i=1}^m$ is then multivariate normal. In particular, let $\nu = (\nu_i)_{i=1}^m$, $\Omega = \operatorname{diag} \omega^2 = \operatorname{diag} (\omega_i^2)_{i=1}^m$ and
$$\Lambda_{ij} = \begin{cases} 1 & \text{if } i = j, \\ -\lambda_{ij} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$
We then have by back-substitution that $\Lambda^\top x = \omega^2$ and thus
$$x \sim \mathrm{N}(\Lambda^{-\top} \nu, \Lambda^{-\top} \Omega \Lambda^{-1}).$$

Parameterizations of directed graphical models

Gaussian models

- The density $\mathrm{N}(x \mid \mu, \Sigma)$ of a multivariate normal distribution satisfies the recursive factorization property if and only if $\Sigma = \Lambda^{-\top}\Omega\Lambda^{-1}$.
- In other words, the parameterized Gaussian directed graphical model associated with G corresponds to all pairs $(\mu, \Sigma) \in \mathbb{R}^m \times \mathrm{PD}_m$ such that one can write $\Sigma = \Lambda^{-\top}\Omega\Lambda^{-1}$, where Λ is upper-triangular and Ω diagonal with positive diagonal entries.
- Let $I_{\mathrm{global}(G)}$ be the ideal generated by all constraints arising from the global Markov property of G, and let I_G be the vanishing ideal of all covariance matrices which factorize on the above form. We have that

$$V(I_{\mathrm{global}(G)}) \cap \mathrm{PD}_m = V(I_G) \cap \mathrm{PD}_m \,.$$