CS/CNS/EE/IDS 165: Foundations in Machine Learning and Statistical Inference

Markov Random Fields/ Graphical Models

Anima Anandkumar
Computing and Mathematical Sciences
California Institute of Technology, Pasadena, CA
anima@caltech.edu

Outline

Concepts

- Graph-Theory Basics
- Markov Random Field.
- Hammersley-Clifford Theorem.
- Gauss-Markov Random Field.
- Acyclic Dependency Graph.
- Gibbs Field.

References

1. P. Brémaud, Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues, Springer, 1998.
2. S. Lauritzen, Graphical Models, Oxford, 1996.
3. H. Rue, L. Held, Gaussian Markov random fields. Theory and applications, Chapman & Hall, 2005.

Preliminary Definitions

Random Field

Let V be a finite set, with elements denoted by v and called nodes, and let Λ be a finite set called the phase space. A random field on V with phases in Λ is a collection $\mathbf{Y} = \{Y_v\}_{v \in V}$ of random variables Y_v with values in Λ.

Graph-Theory Basics

Undirected Graph $\mathcal{G} = (\mathcal{V}, E)$, where $\mathcal{V}$ is the vertex set and $E = \{(i, j)\}$ is the edge set.

Neighborhood function $\mathcal{N}(i; \mathcal{G})$ of a node i is the set of all other nodes having an edge with it in $\mathcal{G}$.

A Complete graph has edges between any two nodes.

A subgraph of a graph $\mathcal{G}$ is a graph whose vertex and edge sets are subsets of those of $\mathcal{G}$.

A clique in a graph is a set of pairwise adjacent vertices. In other words, it is a complete subgraph.

A maximal clique is a clique that is not the subset of any other clique.

Markov Random Field (MRF)

Markov random fields were introduced by Besag in 1974.

Definition

Let $\mathbf{Y}_{\mathcal{V}} = [Y_i, i \in \mathcal{V}]^T$ denote the random vector of measurements in set $\mathcal{V}$. $\mathbf{Y}_{\mathcal{V}}$ is a Markov random field with an (undirected) dependency graph $\mathcal{G} = (\mathcal{V}, E)$, if $\forall\, i \in \mathcal{V}$,

$$Y_i \perp \mathbf{Y}_{\mathcal{V}\setminus\{i,\mathcal{N}(i)\}} | \mathbf{Y}_{\mathcal{N}(i)},$$

where $\perp$ denotes conditional independence. In words, the above definition states that the value at any node, given the values at its neighbors, is conditionally independent of the rest of the network.

Examples

When $\mathbf{Y}_{\mathcal{V}}$ are independent, then $\mathcal{N}(i) = \emptyset$, $\forall i \in V$. In other words, the dependency graph has no edges.

Any general random field without special properties can be represented as a MRF with a complete dependency graph.

An Example for Markov Random Field

Autoregressive process of order 1 (AR-1)

$$Y_t = A_{t-1} Y_{t-1} + \epsilon_{t-1}, \qquad Y_{t-1} \perp \epsilon_{t-1}, \quad \forall t \in V = \{1, \ldots, n\}.$$

The above equation implies that

$$Y_t \perp \mathbf{Y}_{V \setminus \{t-1, t+1\}} | \{Y_{t-1}, Y_{t+1}\}, \quad t \neq 1, n.$$

Hence, $\mathcal{N}(t) = \{t-1, t+1\}$ for $t \neq 1, n$, $\mathcal{N}(1) = 2, \mathcal{N}(n) = n - 1$ or the dependency graph is a linear chain.

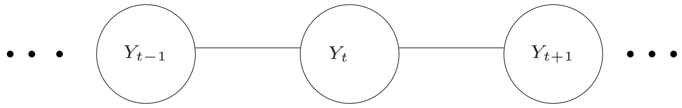


Figure 1: Linear dependency graph for autoregressive process of order 1.

The dependency graph does not capture all the information in the above example: the fact that the AR-1 process evolves in time. In other words, it satisfies causality.

The above example shows that the Markov random field can capture more general dependencies than temporal processes. Hence, the Markov random field model is routinely used to model spatial dependencies. e.g., to represent relationship between nearby pixels in images, correlated sensor measurements.

Equivalent Markov Properties

Local Markov Property $Y_i \perp \mathbf{Y}_{V\setminus(i\cup\mathcal{N}(i))}|\mathbf{Y}_{\mathcal{N}(i)}$, $\forall i \in V$

Global Markov Property $\mathbf{Y}_A \perp \mathbf{Y}_B|\mathbf{Y}_C$, where A, B, C are disjoint sets. A, B are non-empty and C separates A, B.

Pairwise Markov Property $Y_i \perp Y_j|\mathbf{Y}_{\mathcal{V}\setminus\{i,j\}} \iff (i,j) \notin E$

1. Global Markov implies Local Markov

$$A = \{i\}, B = \mathcal{V}\setminus\{i, \mathcal{N}(i)\}, C = \mathcal{N}(i)$$

2. Global Markov implies Pairwise Markov

$$A = \{i\}, B = \{j\}, C = \mathcal{V}\setminus\{i,j\}, \quad \forall (i,j) \notin E.$$

3. The three properties are equivalent under positivity condition, defined later.

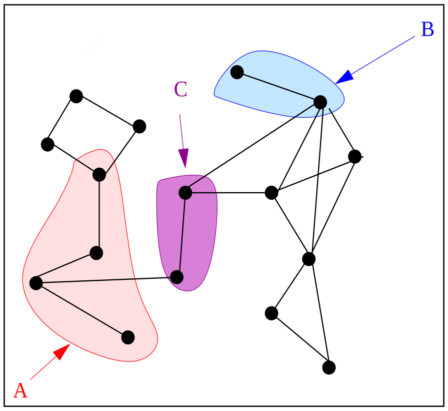


Figure 2: Illustration of Global Markov Property.

Hammersley-Clifford Theorem

Let π be the distribution of a Markov random field with respect to a graph $\mathcal{G} = (V, E)$ satisfying positivity condition. Then

$$\pi(\mathbf{Y}) = \frac{1}{Z} \exp[-\sum_{c \in \mathcal{C}} \Phi_c(\mathbf{Y}_c)]$$

where $\mathcal{C}$ is a collection of maximal cliques in $\mathcal{G}$, $Z > 0$ is the normalizing constant and $\{\Phi_c\}_{c \subset V}$ are non-negative functions called the Gibbs potentials associated with graph $\mathcal{G}$.

Proof: See Brémaud. □

The joint probability distribution is vastly simplified for sparse dependency graphs. Here, the conditional independence relations results in the factorization of the joint distribution into a product of components, each of which depends on a small set of variables. This is exploited by distributed algorithms such as belief propagation.

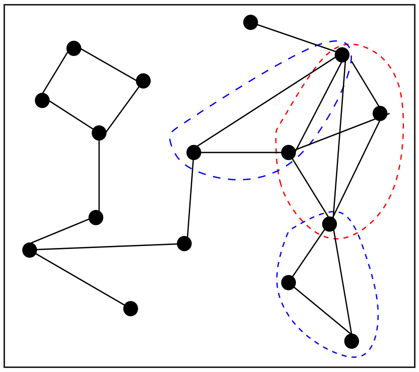


Figure 3: A dependency graph shown with some of the maximal cliques.

Gauss-Markov Random Field

For a Gaussian vector $\mathbf{Y} = [Y_1, \cdots, Y_n]^T$, for simplicity, assume the mean vector $\boldsymbol{\mu} = \mathbf{0}$. Let the inverse of covariance matrix $\boldsymbol{\Sigma}^{-1} = \mathbf{A}$. The PDF is

$$f_{\mathbf{Y}}(\mathbf{y}) = \sqrt{\frac{\det(\mathbf{A})}{(2\pi)^n}} exp\{-\frac{1}{2}[\sum_i A(i,i)Y_i^2 + \sum_{i,j} A(i,j)Y_iY_j]\}$$

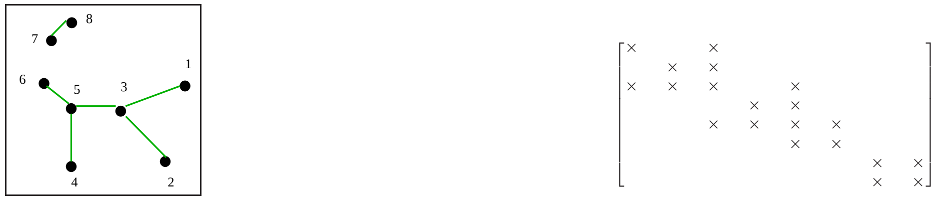
If $\mathbf{Y}$ is a Gauss-Markov random field with graph $\mathcal{G} = (V, E)$, then it can have only pairwise dependencies

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{Z}\exp[-\sum_{(i,j)\in E} \Phi_{i,j}(Y_i, Y_j)].$$

Comparing the two equations,

$$A(i,j) = 0 \iff (i,j) \notin E.$$

Since $\mathbf{A}$ is associated with the potentials, it is called the potential matrix. Hence, for Gaussian distribution, we only need the edges of the dependency graph and not the higher-order cliques.



(a) Dependency Graph

(b) × : Non-zero element of Potential Matrix

Figure 4: Relationship between potential matrix and dependency graph.

Besag's Auto Model

Definition

A Besag's auto model is a Markov random field with only only pairwise dependencies.

$$\pi(\mathbf{Y}) = \frac{1}{Z}\exp[-\sum_{(i,j)\in E}\Phi_{i,j}(Y_i, Y_j)].$$

Hence, Besag's model is a generalization of GMRF and can be generated from the exponential-family distributions.

Example: Ising Model

It was introduced in 1925 to study the phase transition in ferromagnetic materials. Here, $V = \mathbb{Z}^2$ (integer lattice) and $\Lambda = \{-1, +1\}$ and

$$\mathcal{N}[v_{i,j}] = \{v_{i,j+1}, v_{i,j-1}, v_{i+1,j}, v_{i-1,j}\},$$

where $v_{i,j}$ is a node placed at coordinates (i, j). The potential functions are given by

$$\begin{aligned}\Phi_v(Y_v) &= -\frac{H}{k}Y_v, \quad \forall v \in V\\ \Phi_{v,w}(Y_v, Y_w) &= -\frac{J}{k}Y_vY_w, \quad \forall w \in \mathcal{N}(v),\end{aligned}$$

where k is the Boltzmann constant, H is the external magnetic field and J is the internal energy of an elementary magnetic dipole.

Acyclic Dependency Graph

Joint distribution in terms of marginals

Since an acyclic graph does not have cliques of order more than 2, the corresponding Markov random field can have only pairwise dependencies. More interestingly, the joint distribution π can be expressed in terms of marginals at nodes π_i and pairwise distributions $\pi_{i,j}$

$$\pi(\mathbf{y}) = \prod_{i \in V} \pi_i(y_i) \prod_{(i,j)\in E} \frac{\pi_{i,j}(y_i, y_j)}{\pi_i(y_i)\pi_j(y_j)}.$$

Gauss-Markov random field with acyclic dependency

The coefficients of the potential matrix $\mathbf{A}{:=}\mathbf{\Sigma}^{-1}$, for a positive-definite covariance matrix $\mathbf{\Sigma}$ and acyclic dependency graph $\mathcal{G}(\mathcal{V}, \mathrm{E})$, are

$$\begin{aligned} A(i,i) &= \frac{1}{\Sigma(i,i)}\Big(1 + \sum_{j \in \mathcal{N}(i)} \frac{\Sigma(i,j)^2}{\Sigma(i,i)\Sigma(j,j) - \Sigma(i,j)^2}\Big), \\ A(i,j) &= \begin{cases} \dfrac{-\Sigma(i,j)}{\Sigma(i,i)\Sigma(j,j) - \Sigma(i,j)^2} & \text{if } i \sim j, \\ 0 & \text{o.w.} \end{cases} \end{aligned}$$

The determinant of the potential matrix of $\mathbf{A}$ is given by

$$|\mathbf{A}| = \frac{1}{|\mathbf{\Sigma}|} = \frac{\prod_{i\in\mathcal{V}} \Sigma(i,i)^{\text{Deg}(i)-1}}{\prod\limits_{\substack{i\sim j\\ i<j}} [\Sigma(i,i)\Sigma(j,j) - \Sigma(i,j)^2]}.$$

Positivity Condition

Definition

The probability distribution π on a finite space Λ^V, where $V = \{1, \ldots, n\}$, is said to satisfy positivity if,

$$\forall j \in V, y_j \in \Lambda, \;\; \pi_j(y_j) > 0 \Rightarrow \pi(y_1, y_2, \ldots, y_n) > 0,$$

where π_j is the marginal distribution at node j.

The positivity condition implies that if variables individually can take certain values, then they can also take those values jointly. This rules out cases such as $Y_1 = Y_2 = \ldots = Y_n$.