Markov Random Fields

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Probabilistic Models, Spring, 2010

Outline

Outline

From Markov Chains to Markov Random Fields

Markov chains are suitable models for time-series/sequence data. For spatial data, variables can no longer be placed on a line. What would be an analogous model and an analogous Markov property?

A natural generalization stems from the following conditional independence property of a Markov chain $X_1, \ldots, X_n$:

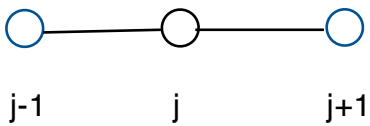
$$P(X_j = x_j \,|\, X_{-j} = x_{-j}) = P(X_j = x_j \,|\, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}), \qquad (1)$$

where $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$. (We proved this in slide 22, Lec. 2.)

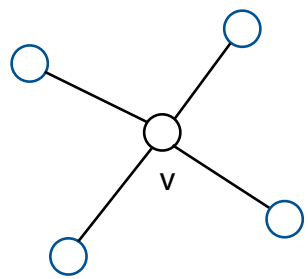
Consider a collection of random variables $\{Y_v, v \in V\}$, where V consists of "sites" in some space.

To generalize property (1) for $Y = \{Y_v\}$, we introduce the notion of *neighbors* to site v:

neighbors of j in a Markov chain:



neighbors of v, in space (illustration):



A Markov Property Analogous to Eq. (1)

Notation:

- $\mathcal{N}_v$: the set of neighbors of v
- $Y_{\mathcal{N}_j} = (Y_v)_{v \in \mathcal{N}_j}$, $\forall j \in V$
- $Y_A = (Y_v)_{v \in A}$, $\forall A \subseteq V$

Eq. (1) for a Markov chain:

$$P(X_j = x_j \,|\, X_{-j} = x_{-j}) = P(X_j = x_j \,|\, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}).$$

Extension of the above property for $\{Y_v, v \in V\}$:

$$P(Y_j = y_j \,|\, Y_{-j} = y_{-j}) = P(Y_j = y_j \,|\, Y_{\mathcal{N}_j} = y_{\mathcal{N}_j}). \tag{2}$$

- The generalized property is identical to

  $$Y_j \perp Y_{V \setminus \mathcal{N}_j} \,|\, Y_{\mathcal{N}_j}.$$

  In other words, Y_j depends on the other variables only through the neighboring variables.
- Full conditional distributions vs. local characteristics

For Markov chains, transition probabilities determine P.
A natural question here is whether local characteristics determine P.

Neighborhood System and Undirected Graph

Call $\mathcal{N} = \{\mathcal{N}_j, j \in V\}$ a *neighborhood system*, if

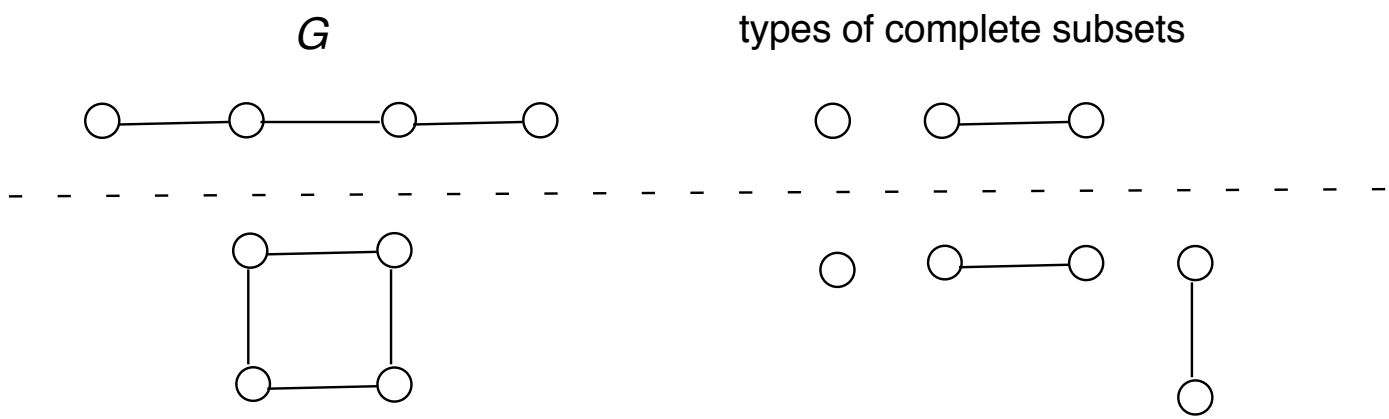
(i) $j \notin \mathcal{N}_j$, (ii) $i \in \mathcal{N}_j \Leftrightarrow j \in \mathcal{N}_i$.

Equivalence: a neighborhood system $\mathcal{N}$ $\Leftrightarrow$ an undirected graph $G = (V, E)$ with $E = \{(i, j) \,|\, i \in \mathcal{N}_j\}$.

Recall the definitions of *complete* subsets and *cliques* of G:

- Complete subset: a subset of vertices that are all pairwise connected in G
- Clique: a maximal complete subset of G

Illustrations of cliques and their non-empty subsets:



Markov Random Fields: Definition

P is called a *Markov random field* with respect to a neighborhood system $\mathcal{N}$ (equivalently, G), if P satisfies the Markov property [Eq. (2)]:

$$P(Y_j = y_j \,|\, Y_{-j} = y_{-j}) = P(Y_j = y_j \,|\, Y_{\mathcal{N}_j} = y_{\mathcal{N}_j}), \quad \forall j \in V,\ \forall y.$$

This property will later be referred to as the *local Markov property*.

Question: P determines the conditional distributions, but

- is P also uniquely determined by its conditional distributions

$$P(Y_j \,|\, Y_{\mathcal{N}_j}) = P(Y_j \,|\, Y_{-j}),\ j \in V\,?$$

Answer: generally, no; under a positivity condition, yes.

Positivity condition: for any $y_1, \ldots, y_n$,

$$P(Y_j = y_j) > 0,\ \ \forall j = 1, \ldots, n \quad \Rightarrow \quad P(Y_1 = y_1, \ldots, Y_n = y_n) > 0.$$

In words, "if y_j's can occur singly they can occur together."

Besag's Theorem

- Besag (1974): Under the positivity condition, P is uniquely determined by its full conditional distributions $P(Y_j \mid Y_{-j})$, $j \in V$.

  Indeed, for any possible values y, y' of Y,

$$\frac{P(Y=y)}{P(Y=y')} = \prod_{j=1}^{n} \frac{p(y_j \mid y_1, \ldots, y_{j-1}, y'_{j+1}, \ldots, y'_n)}{p(y'_j \mid y_1, \ldots, y_{j-1}, y'_{j+1}, \ldots, y'_n)}.$$

  (Proving it is an exercise.) So fixing y', we can determine P from the relation

$$P(Y=y) \propto \prod_{j=1}^{n} \frac{p(y_j \mid y_1, \ldots, y_{j-1}, y'_{j+1}, \ldots, y'_n)}{p(y'_j \mid y_1, \ldots, y_{j-1}, y'_{j+1}, \ldots, y'_n)}.$$

Note: P here is general, not necessarily a Markov random field.

Representation of P and Hammersley-Clifford Theorem

- Hammersley-Clifford (1971): (i) and (ii) are equivalent:

  (i) P satisfies the positivity condition and the Markov property w.r.t. $\mathcal{N}$.

  (ii) P has the form

$$P(Y = y) \propto \exp\{-\psi(y)\} \quad \text{where} \quad \psi(y) = \sum_{C \in \mathcal{C}} \phi_C(y_C), \qquad (3)$$

  $\mathcal{C}$ denotes the set of cliques of G and their subsets, and ϕ_C is a real-valued function of y_C.

This is a powerful representation because it enables systems with very complex global behavior to be built from simple local components.

Terminology: P of the form (3) is called the Gibbs distribution; $\{\phi_C, C \in \mathcal{C}\}$ is called a *potential*.

Note: the potential $\{\phi_C\}$ in the representation of P is not unique.

Representation of P in the Hammersley-Clifford Theorem

It is easy to show "(ii) $\Rightarrow$ (i)" in the Hammersley-Clifford Theorem.

Suppose $p(y) \propto exp\{-\sum_{C\in\mathcal{C}} \phi_C(y_C)\}$. Then for some constant α,

$$p(y_j, y_{-j}) = \alpha \exp\Big\{ - \sum_{C: C\in\mathcal{C}, j\in C} \phi_C(y_C)\Big\} \cdot \exp\Big\{ - \sum_{C: C\in\mathcal{C}, j\notin C} \phi_C(y_C)\Big\}$$

$$= \exp\Big\{ - \sum_{C: C\in\mathcal{C}, j\in C} \phi_C(y_C)\Big\} \cdot h(y_{-j})$$

for some function $h(y_{-j})$. Since $C \in \mathcal{C}$ is a complete subset of G,

$$j \in C \quad \Rightarrow \quad C \setminus \{j\} \subseteq \mathcal{N}_j.$$

Hence p has the form $p(y_j, y_{-j}) = g(y_j, y_{\mathcal{N}_j})\, h(y_{-j})$ for some functions h and g. This implies (as shown in the first exercise)

$$p(y_j \,|\, y_{-j}) = p(y_j \,|\, y_{\mathcal{N}_j}).$$

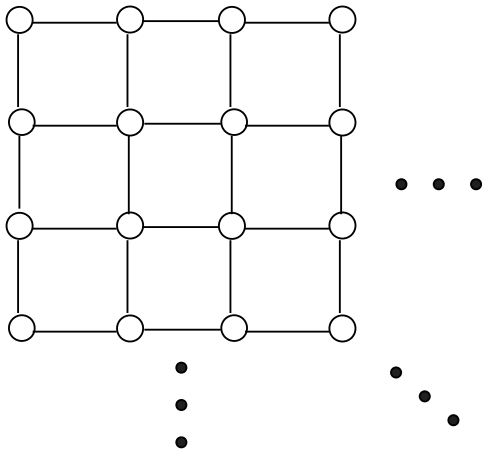
Furthermore, as a function of y_j for fixed $y_{\mathcal{N}_j}$,

$$p(y_j \,|\, y_{\mathcal{N}_j}) \propto \exp\Big\{ - \sum_{C: C\in\mathcal{C}, j\in C} \phi_C(y_C)\Big\}.$$

Outline

Ising Model

G:



Sites: V
$Y_j \in \{0, 1\},\ \forall j \in V$

Ising model and its variants are used in

- statistical physics (phase transition)
- image analysis

- A homogeneous Ising model:

$$\psi(y) = \sum_j b(y_j) + \sum_{i \sim j} c(y_i, y_j).$$

(Homogeneous in the sense that the potential functions b and c do not vary with sites.) The second summation is over all edges (cliques). We have

$$p(y) \propto \exp\Big\{ - \sum_j b(y_j) - \sum_{i \sim j} c(y_i, y_j)\Big\},$$

$$p(y_j \,|\, y_{\mathcal{N}_j}) \propto \exp\Big\{ - b(y_j) - \sum_{i \in \mathcal{N}_j} c(y_i, y_j)\Big\}.$$

Interpret 0 as white, 1 as black.
We calculate next

$$P\big(Y_j = \text{white} \,|\, Y_{-j} = y_{-j}\big) = p(0 \,|\, y_{\mathcal{N}_j}).$$

Ising Model

From $p(y_j \,|\, y_{\mathcal{N}_j}) \propto \exp\big\{ -b(y_j) - \sum_{i\in\mathcal{N}_j} c(y_i, y_j)\big\}$ we obtain

$$p(0\,|\,y_{-j}) = \frac{\exp\big\{ -b(0) - \sum_{i\in\mathcal{N}_j} c(y_i,0)\big\}}{\exp\big\{ -b(0) - \sum_{i\in\mathcal{N}_j} c(y_i,0)\big\} + \exp\big\{ -b(1) - \sum_{i\in\mathcal{N}_j} c(y_i,1)\big\}}$$
$$= \frac{1}{1+\exp\big\{b(0)-b(1)+\sum_{i\in\mathcal{N}_j}\big(c(y_i,0)-c(y_i,1)\big)\big\}}.$$

Let

$$n_0 = \#\text{ white neighbors}, \qquad n_1 = \#\text{ black neighbors} = |\mathcal{N}_j| - n_0.$$

$$\sum_{i\in\mathcal{N}_j}\big(c(y_i,0)-c(y_i,1)\big) = n_0 c(0,0) + n_1 c(1,0) - n_0 c(0,1) - n_1 c(1,1)$$
$$= n_0\big(c(0,0)+c(1,1)-c(0,1)-c(1,0)\big) + |\mathcal{N}_j|\big(c(1,0)-c(1,1)\big).$$

So we can write

$$p(0\,|\,y_{-j}) = p(0\,|\,y_{\mathcal{N}_j}) = \frac{1}{1+\exp\big\{\beta+\gamma|\mathcal{N}_j|+\delta n_0\big\}}.$$

Interpretation of parameters: $\beta+\gamma|\mathcal{N}_j|$ controls the overall size of the probability; δ controls the degree of dependence of Y_j on its white neighbors: when $\delta = 0$, the color of site j is independent of the colors around it. When $\gamma \to -\infty$, $p(0\,|\,y_{-j})$ increases to 1.

Other Application Examples of MRF

Mulecular structure simulation: e.g.,

- protein structural simulation
- protein folding – finding "minimal-energy" configuration

Solving optimization problems: $\min_x \psi(x)$ where $\psi(x) = \sum_C \phi_C(x_c)$

- We embed the objective $\psi(x)$ in $e^{-\frac{1}{T}\psi(x)}$ and sample from the density

$$\pi(x; T) \propto e^{-\frac{1}{T}\psi(x)}.$$

As the "temperature" T decreases, the probability mass concentrates around the near-optimal points x:

$$\forall\ x, x' \text{with } \psi(x') - \psi(x) > 0,$$
$$\frac{\pi(x; T)}{\pi(x'; T)} = e^{\frac{1}{T}\left(\psi(x') - \psi(x)\right)} \to \infty, \quad \text{as } T \to 0.$$

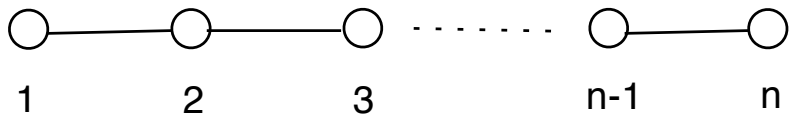
MRF with sampling methods provides a powerful approach to address large-scale problems of this kind in practice.

Outline

Markov Chains Revisited

Let $X_1, \dots X_n$ be a Markov chain with joint distribution P.

- Recall that by Eq. (1), P is a Markov random field w.r.t. the graph



- The joint PMF p factorizes as

$$p(x) = p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_2) \cdots p(x_n \,|\, x_{n-1}),$$

so, *if p is strictly positive*, p may be written as for some functions ϕ_i^a, ϕ_i^b,

$$p(x) \propto \exp\Big\{ - \sum_{i=1}^{n} \phi_i^a(x_i) - \sum_{i=1}^{n-1} \phi_i^b(x_i, x_{i+1}) \Big\}. \tag{4}$$

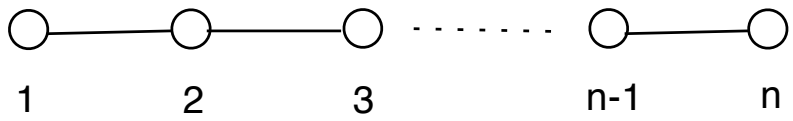
- Such expressions are not unique. For example, an alternative expression of p in terms of the marginals is

$$p(x) = \frac{p(x_1, x_2)p(x_2, x_3) \cdots p(x_{n-1}, x_n)}{p(x_2)p(x_3) \cdots p(x_{n-1})}.$$

The potential functions in (4) may be chosen as $\phi_i^a = -\ln p(x_i)$, and $\phi_i^b = -\ln p(x_i, x_{i+1})$. *Generally, ϕ_i^a, ϕ_i^b do not have to correspond to probability distributions.*

Markov Chains Revisited

Conversely, if P is a Markov random field w.r.t. the graph



is $\{X_n\}$ a Markov chain?

- Not necessarily, without the positivity condition.
- A counter-example:

$$X_1 = X_2, \quad X_3 \perp X_2, \quad X_4 = X_2 + X_3, \quad X_5 = X_4.$$

Then, X_j is independent of the other variables conditionally on the neighboring variables:

$$P(X_1 \,|\, X_{-1}) = P(X_1 \,|\, X_2); \quad P(X_2 \,|\, X_{-2}) = P(X_2 \,|\, X_1);$$

$$P(X_4 \,|\, X_{-4}) = P(X_4 \,|\, X_5); \quad P(X_5 \,|\, X_{-5}) = P(X_5 \,|\, X_4);$$

$$P(X_3 \,|\, X_{-3}) = P(X_3 \,|\, X_2, X_4).$$

But

$$P(X_4 \,|\, X_1, X_2, X_3) = P(X_4 \,|\, X_3, X_2).$$

So $X_1, X_2, \ldots, X_5$ is not a Markov chain.

Outline

Motivation

Observations from the preceding discussion on MRF and Markov chains:

- Different graphs can represent the same set of conditional independence relations for a distribution.
- Conditional independence seems to have similarity with separation between vertices in graphs.

Separation in an undirected graph $G = (V, E)$:

- For $A, B, S \subseteq V$, S is said to *separate* A from B, if every path from some $\alpha \in A$ to some $\beta \in B$ intersects S.

In what follows we introduce more Markov properties on undirected graphs, placing the Markov property in MRF among them.
We mention some relations between these properties without proofs.

Factorized Representation of Distribution and Global Markov Property

Notation and definitions:

- $G = (V, E)$: an undirected graph
- $\mathcal{C}$: the set of cliques of G
- For $A, B, S \subseteq V$, $A \perp B \,|\, S$: S separates A from B in G
- $X = \{X_v, v \in V\}$: random variables associated with V
- X_A for $A \in V$: $\{X_v, v \in A\}$
- $X_A \perp X_B \,|\, X_S$: X_A and X_B are independent conditional on X_S

We name some properties for $P(X)$: We say

- P *factorizes* according to G (F), if

$$p(x) = \prod_{C \in \mathcal{C}} \phi_C(x_C), \quad \text{for some nonnegative functions } \phi_C, C \in \mathcal{C}.$$

- P obeys the *global Markov property* (G) with respect to G, if for any disjoint subsets A, B, S of V,

$$A \perp B \,|\, S \quad \Rightarrow \quad X_A \perp X_B \,|\, X_S.$$

Note the direction of implication in (G).

Local Markov Property and Pairwise Markov Property

Notation: for $A \subseteq V$,

- *boundary* of A:

$$\mathrm{bd}(A) = \cup_{v \in A} \mathcal{N}_v \setminus A$$

  i.e., all neighbors of members of A that are not in A

- *closure* of A:

$$\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$$

  i.e., A and its boundary

(Example: for $v \in V$, what is $\mathrm{bd}(v), \mathrm{cl}(v)$?)

The Markov property that defines MRF [Eq. (2)] can be rewritten in this new notation. We say

- P obeys the *local Markov property* (L) with respect to G, if for all $v \in V$,

$$X_v \perp X_{V \setminus \mathrm{cl}(v)} \,|\, X_{\mathrm{bd}(v)}.$$

Clearly, the neighbors of v separate v from the rest of the vertices. So (L) is weaker than (G). A Markov property weaker than (L) is:

- P obeys the *pairwise Markov property* (P) with respect to G, if for all pairs of non-adjacent vertices (v, v') in G,

$$X_v \perp X_{v'} \,|\, V \setminus \{v, v'\}.$$

Relations between Markov Properties on Undirected Graphs

Fact: (F) $\Rightarrow$ (G) $\Rightarrow$ (L) $\Rightarrow$ (P).

(F) P *factorizes* according to G.

(G) For any disjoint subsets A, B, S of V, $A \perp B \,|\, S \quad \Rightarrow \quad X_A \perp X_B \,|\, X_S$.

(L) For all $v \in V$, $X_v \perp X_{V \setminus \mathrm{cl}(v)} \,|\, X_{\mathrm{bd}(v)}$.

(P) For all pairs of non-adjacent vertices (v, v') in G, $X_v \perp X_{v'} \,|\, V \setminus \{v, v'\}$.

Implication:

$$\mathcal{M}_{\mathrm{F}}(G) \subseteq \mathcal{M}_{\mathrm{G}}(G) \subseteq \mathcal{M}_{\mathrm{L}}(G) \subseteq \mathcal{M}_{\mathrm{P}}(G)$$

where $\mathcal{M}_*(G)$ denotes the set of P satisfying the property indicated by * with respect to the graph G. (All inclusions are strict generally.)

- The above fact follows from the properties of conditional independence. (By a conclusion in the 1st exercise, (F) $\Rightarrow$ (L), (F) $\Rightarrow$ (P) are immediate. We verify the fact entirely at another time.)
- The version of Hammersley-Clifford theorem shown earlier establishes (L) $\Leftrightarrow$ (F) under the positivity condition. (You can find a proof in Davison's book.) There is another version for (P) $\Leftrightarrow$ (F) under the same condition (see *Graphical Models* by Lauritzen, 1996).

Positivity Condition and a Counter-Example

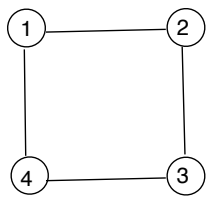
Without the positivity condition, (G) $\not\Rightarrow$ (F).

Counter-example due to Moussouris (1974) (see Lauritzen 1996):

Let X_1, X_2, X_3, X_4 be four binary random variables taking the follow values, each with equal probability of $1/8$:

$$\begin{array}{cccc} (0,0,0,0) & (1,0,0,0) & (1,1,0,0) & (1,1,1,0) \\ (0,0,0,1) & (0,0,1,1) & (0,1,1,1) & (1,1,1,1) \end{array}$$

As can be verified, P so defined satisfies the global Markov property w.r.t. the graph shown on the right. But P does not admit a factorized representation.



To see this, suppose it does. Then, from

$$\begin{aligned} 0 \neq 1/8 = p(0,0,0,0) &= \phi_{\{1,2\}}(0,0)\phi_{\{2,3\}}(0,0)\phi_{\{3,4\}}(0,0)\phi_{\{4,1\}}(0,0), \\ 0 = p(0,0,1,0) &= \phi_{\{1,2\}}(0,0)\phi_{\{2,3\}}(0,1)\phi_{\{3,4\}}(1,0)\phi_{\{4,1\}}(0,0), \\ 0 \neq 1/8 = p(0,0,1,1) &= \phi_{\{1,2\}}(0,0)\phi_{\{2,3\}}(0,1)\phi_{\{3,4\}}(1,1)\phi_{\{4,1\}}(1,0), \end{aligned}$$

we must have

$$\phi_{\{2,3\}}(0,1) \neq 0, \quad \phi_{\{3,4\}}(1,0) = 0.$$

But

$$0 \neq 1/8 = p(1,1,1,0) = \phi_{\{1,2\}}(1,1)\phi_{\{2,3\}}(1,1)\phi_{\{3,4\}}(1,0)\phi_{\{4,1\}}(0,1),$$

contradiction.

Further Readings

About MRF:

1. A. C. Davison. *Statistical Models*, Cambridge Univ. Press, 2003. Chap. 6.2.

About Markov properties on undirected graphs:

2. Robert G. Cowell et al. *Probabilistic Networks and Expert Systems*, Springer, 2007. Chap. 5.1, 5.2.