

Markov Random Fields

1. **Markov property** The Markov property of a stochastic sequence $\{X_n\}_{n \geq 0}$ implies that for all $n \geq 1$, X_n is independent of $(X_k : k \notin \{n-1, n, n+1\})$, given (X_{n-1}, X_{n+1}) . Another way to write this is:

$$X_n \perp (X_k : k \notin \partial\{n\}) \mid (X_k : k \in \partial\{n\})$$

where $\partial\{n\}$ is the set of *neighbors* of *site* n . We would like to now generalize this Markov property from one-dimensional index sets to more arbitrary domains.

2. **Random field** Let S be a finite set, with elements denoted by s and called *sites*, and let Λ be a finite set called the *phase space*. A *random field* on S with phases in Λ is a collection $X = \{X(s) : s \in S\}$ of random variables $X(s)$ with values in Λ .

In fact, a random field can be regarded as a random variable taking its values in the *configuration space* Λ^S . Thus, a configuration $x \in \Lambda^S$ is of the form $x = \{x(s) : s \in S\}$, where $x(s) \in \Lambda$ for all $s \in S$. We also define, for a given configuration x and any given subset $A \subset S$, $x(A) = \{x(s) : s \in A\}$. By this notation, we can write: $x = \{x(A), x(S-A)\}$.

3. **Dependence structure** How do we impose *dependence* structures on a random field? There are two clear paths to take. One may either define a *joint* distribution on X , treating a configuration x as a realization from that field, and directly model the correlations using the variance-covariance matrix. Alternatively, one may try to build Markovian dependence structures on the set S . It is this latter approach that we address here.
4. **A discrete topology on S** We introduce a *neighborhood system* (also called a *topology*) on S by defining a symmetric relation \sim on S and defining neighborhoods as the relational sets:

$$N = \{(i, j) \in S \times S : i \sim j\}.$$

This generates the graph $\mathcal{G} = (S, N)$ as a topology on S , with neighborhoods, say $\partial s = N_s$ associated with each element s as:

- (a) $N_s = \{t \in S : t \sim s\}$
- (b) $s \notin N_s$
- (c) $t \in N_s \implies s \in N_t$.

5. **Markov Random Field** A random field $X = \{X(s) : s \in S\}$ is called a *Markov Random Field* with respect to $\mathcal{G} = (S, N)$ if for all $s \in S$,

$$X(s) \perp X(S - \{s \cup N_s\}) \mid X(N_s).$$

That is, the random variable $X(s)$ is conditionally independent of all other sites in S , given its values in N_s .

6. **Local characteristic** The *local characteristic* of an MRF at site s is given by the function:

$$\pi_s(x) = P(X(s) = x(s) \mid X(N_s) = x(N_s)).$$

The family $\{\pi_s\}_{s \in S}$ is called the *local specification* of the MRF. The central question underlying the validity of a MRF is *when* this local specification of an MRF leads to a joint distribution.

To highlight this problem, consider two random variables Y_1 and Y_2 such that $Y_1|Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, 1)$ and $Y_2|Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, 1)$. Then, clearly

$$\begin{aligned} E[Y_1] &= \alpha_0 + \alpha_1 E[Y_2] \\ E[Y_2] &= \beta_0 + \beta_1 E[Y_1^3]. \end{aligned}$$

To see that Y_1 and Y_2 are *incompatible*, note that by the first equation $E[Y_1]$ and $E[Y_2]$ have a linear relationship. So, the second equation will be compatible with the first if $E[Y_1^3]$ also has a linear relationship for $E[Y_1]$. But this is not true except for only trivial situations. As another example, note that even if we define $Y_2|Y_1 \sim N(Y_1, 1)$ and $Y_1|Y_2 \sim N(Y_2, 1)$, we end up with an *improper* joint distribution $p(y_1, y_2) \propto \exp(-(y_1 - y_2)^2/2)$.

Thus, it is of interest to investigate the conditions when local specifications lead to unambiguous joint distributions.

7. **Positivity condition** The probability distribution π on the finite configuration space Λ^S , where $S = \{1, \dots, K\}$ is said to satisfy the *positivity condition* if for all $j \in S$, $x_j \in \Lambda$,

$$(\pi_j(x_j) = 0) \implies (\pi(y_1, \dots, y_{j-1}, x_j, y_{j+1}, \dots, y_K) = 0)$$

for all $y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_K \in \Lambda$, where π_j is the marginal probability distribution on site j .

8. **Brook's Lemma** An important use of the positivity condition to identify joint distributions from local specifications is provided by this lemma due to Brook: For any $x, y \in \Lambda^S$, with strictly positive probability:

$$\frac{\pi(x)}{\pi(y)} = \prod_{i=1}^K \frac{\pi(x_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_K)}{\pi(y_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_K)}.$$

Proof: The simplest way to prove this is to observe that:

$$\pi(x) = \frac{\pi(x_K | x_1, \dots, x_{K-1})}{\pi(y_K | x_1, \dots, x_{K-1})} \pi(x_1, \dots, x_{K-1}, y_K),$$

and then to proceed recursively by considering:

$$\pi(x_1, \dots, x_{K-1}, y_K) = \frac{\pi(x_{K-1} | x_1, \dots, x_{K-2}, y_K)}{\pi(y_{K-1} | x_1, \dots, x_{K-2}, y_K)} \pi(x_1, \dots, x_{K-2}, y_{K-1}, y_K)$$

and so on. Note that the above calculation makes sense because the positivity condition and the strict positivity of $\pi(x)$ and $\pi(y)$ imply that for all $j \in \{1, \dots, K\}$, $\pi(x_1, \dots, x_j, y_{j+1}, \dots, y_K) > 0$.

9. **Uniqueness under a local specification** Brook's lemma indeed reveals a situation when the conditional specifications completely determine a *joint* distribution. Algorithmically speaking, fixing any $y \in \Lambda^S$ and with the local specifications at our disposal, we can identify $\pi(x)$ up to a proportionality constant.

10. **Cliques, Potentials and Gibbs distributions** A family of distributions that play a central role in constructing MRF's arose in the Physics literature where Gibbs introduced them (hence called Gibbs distributions):

$$\pi_T(x) \propto \exp\left(-\frac{1}{T}\mathcal{E}(x)\right)$$

on Λ^S , where T is the *temperature*, $\mathcal{E}(x)$ is the *energy* of configuration x , and the normalizing constant (actually a function of T) is called the *partition function*. Such distributions are interesting for physicists when the energy is expressed in terms of a potential function describing the local interactions. It is here that *cliques* play a central role.

Cliques are actually *complete subgraphs*, or subgraphs where any sites are neighbors. By convention, any singleton $\{s\}$ is also a clique. A clique is called *maximal* if inclusion of any other additional site prevents it from remaining a clique.

A *Gibbs potential* on Λ^S relative to the neighborhood system N is a collection $\{V_C\}_{C \subset S}$ of functions $V_C : \Lambda^S \rightarrow \mathfrak{R} \cup \{+\infty\}$ such that:

- (a) $V_C \equiv 0$ if C is not a clique
(b) for all $x, x' \in \Lambda^S$ and all $C \subset S$,

$$\{x(C) = x'(C)\} \implies \{V_C(x) = V_C(x')\}.$$

The energy function $\mathcal{E}(x)$ is said to *derive from the potential* $\{V_C\}_{C \subset S}$ if:

$$\mathcal{E}(x) = \sum_C V_C(x).$$

Note that the function $V_C(x)$ depends only upon the phases at the sites inside subset C . It is in this context that the distribution mentioned above is called a *Gibbs distribution*.

11. **Gibbs fields are MRF's** If X is a random field with a Gibbs distribution (without loss of generality assume $T = 1$) over $\mathcal{G} = (S, N)$:

$$\pi(x) = \frac{1}{Z} \exp\left(-\sum_{C \subset S} V_C(x)\right),$$

then X is a Markov Random Field over (S, N) with local specification given by:

$$\pi_s(x) = P(X(s) = x(s) \mid X(N_s) = x(N_s)) = \frac{\exp(-\sum_{C \ni s} V_C(x))}{\sum_{\lambda \in \Lambda} \exp(-\sum_{C \ni s} V_C(\lambda))}.$$

Proof: It is enough to prove that

$$P(X(s) = x(s) \mid X(S-s) = x(S-s)) = \pi_s(x),$$

since $\pi_s(x)$ depends only upon $x(s)$ and $x(N_s)$. Also note that, by virtue of our definition of the potential function,

$$P(X(s) = x(s) \mid X(S-s) = x(S-s)) = \frac{\pi(x)}{\sum_{\lambda \in \Lambda} \pi(\lambda, x(S-s))}.$$

Let $C_1(s) = \{C : C \ni s\}$ and $C_2(s) = C_1(s)^c$. Then,

$$\pi(x) = \frac{1}{Z} \exp\left(-\sum_{C_1(s)} V_C(x) - \sum_{C_2(s)} V_C(x)\right),$$

and similarly,

$$\pi(\lambda, x(S-s)) = \frac{1}{Z} \exp\left(-\sum_{C_1(s)} V_C(\lambda, x(S-s)) - \sum_{C_2(s)} V_C(\lambda, x(S-s))\right).$$

Since C is a clique, note that $s \notin C$ implies that $V_C(\lambda, x(S-s)) = V_C(x)$ does not depend upon λ .

The result now follows by simply factoring out $\exp(-\sum_{C_2(s)} V_C(x))$.

12. **Mobius inversion formula** Let E be a finite set and let $\mathcal{P}(E)$ denote the *power set* of E (i.e., the set of all subsets of E). Let ϕ and ψ be two set functions defined on $\mathcal{P}(E)$. Let $A \subset E$ be any subset of E . Then the following two statements are equivalent:

(a)

$$\phi(A) = \sum_{B \subset A} (-1)^{|A-B|} \psi(B)$$

(b)

$$\psi(A) = \sum_{B \subset A} \phi(B)$$

Proof: We will first prove (b) implies (a). Substitute (b) into the right hand expression for (a) to obtain:

$$\begin{aligned} \sum_{B \subset A} (-1)^{|A-B|} \psi(B) &= \sum_{B \subset A} (-1)^{|A-B|} \sum_{D \subset B} \phi(D) \\ &= \sum_{D \subset B \subset A} (-1)^{|A-B|} \phi(D) = \sum_{D \subset A} \sum_{F \subset A-D} (-1)^{|F|} \phi(D) \\ &= \sum_{D \subset A} \phi(D) \sum_{F \subset A-D} (-1)^{|F|}. \end{aligned}$$

But note that if $A - D$ is the null set then $\sum_{F \subset A-D} (-1)^{|F|} = (-1)^0 = 1$. On the other hand, when $A - D$ is non-empty then

$$\begin{aligned} \sum_{F \subset A-D} (-1)^{|F|} &= \sum_{k=0}^{|A-D|} (-1)^k \times |[F : |F| = k, F \subset A - D]| \\ &= \sum_{k=0}^{|A-D|} (-1)^k \binom{|A-D|}{k} = (1-1)^{|A-D|} = 0, \end{aligned}$$

which shows

$$\sum_{D \subset A} \phi(D) \sum_{F \subset A-D} (-1)^{|F|} = \phi(A),$$

yielding the desired result.

HW: Show that (a) implies (b).

Hammersley-Clifford theorem Let π be the distribution of an MRF with respect to $\mathcal{G} = (S, N)$ satisfying the positivity condition. Then, $\pi(x) \propto \exp(-\mathcal{E}(x))$ for some energy function deriving from a Gibbs potential $\{V_C\}_{C \subset S}$ associated with the topology (S, N) .

Proof: Let 0 be the element of Λ^S with all elements 0 . Let $A \subset S$ be a subset of S and denote x^A to be the vector derived from x such that $x_i^A = x_i$ if $i \in A$ and 0 otherwise. Define for $A \subset S$, $x \in \Lambda^S$,

$$V_A(x) = \sum_{B \subset A} (-1)^{|A-B|} \log \frac{\pi(0)}{\pi(x^B)}.$$

From the Mobius formula,

$$\log \frac{\pi(0)}{\pi(x^A)} = \sum_{B \subset A} V_B(x).$$

Taking $A = S$ gives

$$\log \frac{\pi(0)}{\pi(x)} = \sum_{B \subset S} V_B(x),$$

so that

$$\pi(x) = \pi(0) \exp\left(-\sum_{A \subset S} V_A(x)\right).$$

Furthermore, if $x, y \in \Lambda^S$ are such that $x(A) = y(A)$, then for any $B \subset A$, $x^B = y^B$ and so $V_A(x) = V_A(y)$. Finally, another combinatorial argument will prove that $V_A(x) = 0$ if A is not a clique. **Try it as a BONUS HW.**

13. **The Gibbs sampler** Here we consider how a given random field with a Gibbs probability distribution $\pi(x)$ can arise as a stationary distribution of an MRF (or a field-valued Markov Chain). The problem is interesting in the following context. Suppose there exists an irreducible aperiodic homogeneous Markov Chain with state-space $E = \Lambda^S$ and stationary distribution $\pi(x)$. If, then, one is able to generate a realization of the HMC, its distribution at a large time n will be close to π , and one will therefore have simulated π .

The first problem is that of identifying a chain X_n with stationary distribution as $\pi(x)$. The Gibbs sampler uses a strictly positive probability distribution $(q_s, s \in S)$ on S , and the transition from $X_n = x$ to $X_{n+1} = y$ is made in the following manner. The new state y is obtained from the old state x by changing the value of the phase at *one site only*. The site s to be changed at time n is chosen independently of the past with probability q_s . When site s has been selected, the current configuration $x = (x(s), x(S-s))$ is changed into $y = (y(s), x(S-s))$ with probability $\pi(y(s)|x(S-s))$. Thus, the transition matrix is given by:

$$P(X_{n+1} = y | X_n = x) = q_s \pi(y(s)|x(S-s)) 1(y(S-s) = x(S-s)).$$

The corresponding chain is aperiodic and irreducible. To prove that π is the stationary distribution we prove the detailed balance condition:

$$\pi(x)P(X_{n+1} = y \mid X_n = x) = \pi(y)P(X_{n+1} = x \mid X_n = y).$$

Starting with the left hand side, we write:

$$\pi(x)P(X_{n+1} = y \mid X_n = x) = \pi(x)q_s\pi(y(s)|x(S-s))1(y(S-s) = x(S-s)) = \pi(x)q_s \frac{\pi(x)}{P(X(S-s) = x(S-s))},$$

and see immediately that:

$$\pi(x)q_s \frac{\pi(y(s), x(S-s))}{P(X(S-s) = x(S-s))} = \frac{\pi(x)}{P(X(S-s) = x(S-s))} q_s \pi(y(s), x(S-s)) = \pi(y)q_s \frac{\pi(x)}{P(X(S-s) = x(S-s))},$$

which is the detailed balance condition.

Clearly, Gibbs sampling applies to any multivariate probability distribution $\pi(x(1), \dots, x(N))$ on a set $E = \Lambda^N$. Thus, the basic step of the Gibbs sampler for the multivariate distribution π consists of selecting a coordinate number $i \in 1, \dots, N$ at random and choosing the new value $y(i)$ of the corresponding coordinate, given the present values of the other coordinates, with probability

$$\pi(y(i) \mid x(1), \dots, x(i-1), x(i+1), \dots, x(N)).$$

These distributions are known as the *full conditional distributions*.

Finally, note that we could easily design a *periodic Gibbs sampler* where the sites s are not updated at random, but rather are updated in a well-determined sequence $s(1), \dots, s(N)$ where $\{s(i)\}_{i=1}^N$ is a complete enumeration of all the sites in S . In fact, the state of the random field after the n^{th} sweep is $Z_n = X_{nN}$, where X_k denotes the state before the k^{th} update. At time k , site $s(k \bmod N)$ is updated to produce the new state X_{k+1} . If $X_k = x$ and $s(k \bmod N) = s$ then $X_{k+1} = (y(s), x(S-s))$ with probability $\pi(y(s)|x(S-s))$. The Gibbs distribution is stationary in the sense that if $X_k \sim \pi$, then $X_{k+1} \sim \pi$.

14. **Returning to the Metropolis-Hastings** We now consider an alternative to the earlier Metropolis-Hastings algorithm, which we call the *component-wise Metropolis-Hastings* algorithm. In particular consider the setting with a general state space like an MRF. Here, instead of using a straightforward M-H algorithm with a multivariate proposal distribution $q()$, we might wish to update the elements of the configuration x component-wise, using a *series* of transition kernels, say P_k for the k^{th} component

x_k , such that each of the P_k 's maintain the stationary distribution π . In that case, $P_1 \dots P_N$ will also maintain π . Now, each P_k will require its own proposal, say q_k , and we have greater flexibility in choosing the P_k 's so that the proposals and acceptance decisions are simple to implement.

Let us now openly acknowledge that $E = \Lambda^N$ and a configuration X has many components so that $X = (X_1, \dots, X_N)$. Then, the most common approach is to devise an algorithm in which a $q_k(x, x^*)$ is assigned to each individual component X_k . Thus, if x is the current state, then q_k proposes a replacement x_k^* for the k^{th} component of x , say x_k , but leaves the remainder of $x_{-k} = (x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_N)$ unchanged. Thus, in the component-wise algorithm the acceptance ratio $\alpha_k(x, x^*)$ can be expressed as:

$$\alpha_k(x, x^*) = \min \left\{ 1, \frac{\pi(x_k^* | x_{-k}) q_k(x^*, x)}{\pi(x_k | x_{-k}) q_k(x, x^*)} \right\}.$$

This identifies the crucial role played by the *full conditional* distributions in the component-wise Hastings algorithm. In fact, taking each proposal q_k as the full-conditional distribution

$$q_k(x, x^*) = \pi(x_k^* | x_{-k})$$

results in precisely the Gibbs sampler. Note that proposals are *always* accepted in this situation.