Say we would like to think about patterns of racial segregation in a city, and we are willing to make a couple of simplifying assumptions. First, we'll assume that a city block is either all white or all black. Second, we'll assume that a white block that is surrounded by black blocks is likely to turn black and vice versa, and that only neighboring blocks have a direct influence. A block that is further away can influence a block, but only by influencing a directly neighboring block.

How can we use these beliefs to reason about what kinds of patterns of segregation and integration are more or less likely?

Independence and Factoring

Let's look at four blocks, that we will call A, B, C, and D, Blocks A and C are not neighbors and neither is B and D (figure X).

Now, let's say there is a particular pattern of black and white blocks, for example Block A–black, Block B–black, Block C–white, Block D–white. We'll call this particular pattern x, and we'll say that $\Pr(X = x)$ is the probability of that pattern, and $\Pr(X)$ is the probability distribution over all possible patterns

So that the model is clear enough to think with, we'd like to express this probability distribution only in terms of local interactions between neighboring blocks. We'll do this in two steps. First we'll show that any probability distribution over these four blocks can be expressed as a product of functions that only take in the value of neighboring blocks, i.e. :

$$\Pr(X) = \phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A) \tag{1}$$

Second, we'll choose functions that compactly express our assumptions about local interactions.

Factorization

Because we think that blocks can only influence each other through a direct neighbor, we can say that A and C are independent of each other given their immediate neighbors B and D, and that B and D are independent given A and C. This does not mean that A can not influence C just that the influence must operate through B and D. If we already know B and D, then we have fully taken into account the influence of A on C.

As I'll demonstrate these independencies imply that

$$\Pr(X) = \phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A) \tag{2}$$

Factorization

Theorem 1. *Let A, B, and C be three disjoint sets of variables such that $X = A \cup B \cup C$. $\Pr(X)$ satisfies $(A \perp\!\!\!\perp B) \mid C$ if and only if*

$$\Pr(X) = \phi_1(A,C)\phi_2(B,C) \tag{3}$$

for some functions ϕ_1 and ϕ_2.

Proof. Assume that $(A \perp\!\!\!\perp B) \mid C$

$$\begin{aligned}
\Pr(A,B,C) &= \Pr(A,B \mid C)\Pr(C) && (4)\\
&= \Pr(A \mid C)\Pr(B \mid C)\Pr(C) && (5)\\
&= \phi_1(A,C)\phi_2(B,C) && (6)
\end{aligned}$$

Where we set $\phi_1(A,C) = \Pr(A \mid C)$ and $\phi_2 = \Pr(B \mid C)\Pr(C)$.

Now assume that $\Pr(A,B,C) = \phi_2(A,C)\phi_2(B,C)$. Let $\phi_3(C) = \sum_A \phi_1(A,C)$ and $\phi_4(C) = \sum_B \phi_2(B,C)$.

$$\begin{aligned}
\Pr(A,B \mid C) &= \frac{\Pr(A,B,C)}{\sum_{A,B}\Pr(A,B,C)} && (7)\\
&= \frac{\phi_1(A,C)\phi_2(B,C)}{\sum_{A,B}\phi_2(A,C)\phi_2(B,C)} && (8)\\
&= \frac{\phi_1(A,C)\phi_2(B,C)}{\phi_3(C)\phi_4(C)} && (9)
\end{aligned}$$

Similarly

$$\begin{aligned}
\Pr(A \mid C) &= \frac{\sum_B \Pr(A,B,C)}{\sum_{A,B}\Pr(A,B,C)} && (10)\\
&= \frac{\phi_1(A,C)\phi_4(C)}{\phi_3(C)\phi_4(C)} && (11)\\
&= \frac{\phi_1(A,C)}{\phi_3(C)} && (12)
\end{aligned}$$

From which we can see that

$$\Pr(A,B \mid C) = \Pr(A \mid C)\Pr(B \mid C) \tag{13}$$

Which was to be proven. □

Now, if we let $\phi_5(A,B,D) = \phi_1(A,B)\phi_4(D,A)$ and $\phi_6(C,B,D) = \phi_2(B,C)\phi_3(C,D)$ we can see that

$$\begin{aligned}
\Pr(X) &= \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A) && (14)\\
&= \phi_5(A,B,D)\phi_6(C,B,D) && (15)\\
&= \phi_5(A,\{B,D\})\phi_6(C,\{B,D\}) && (16)
\end{aligned}$$

which implies $(A \perp\!\!\!\perp C) \mid (B,D)$, and if we combine the factors another way it also implies $(B \perp\!\!\!\perp D) \mid (A,C)$.

1 Factors to Distributions

We have see that there is an intimate connection between the independencies respected by a probability distribution and the factorization of that distribution into functions. In particular, we have shown that a probability distribution with certain independencies can be factored into functions that have only directly dependent random variables in their scope.

Remember that we represented the dependencies between the blocks as edges. We drew an edge between two blocks when the blocks had a direct, unmediated effect upon each other. Networks of this kind are called Markov networks, and what we saw for our example is true for all networks of this type.

Definition 1. *A distribution* $\Pr_\phi$ *is a Gibbs distribution defined by the factors* $\{\phi_1(D_1), ..., \phi_K(D_k)\}$ *if*

$$\Pr_\phi(X_1, ...X_n) = \frac{1}{Z}\tilde{P}_\phi(X_1, ..., X_n) \tag{17}$$

where

$$\tilde{P}_\phi(X_1, ..., X_n) = \phi_1(D_1)\phi_2(D_2)...\phi_{K-1}(D_{K-1})\phi_K(D_k) \tag{18}$$

and

$$Z = \sum_{X_1,...,X_n} \tilde{P}_\phi(X_1, ..., X_n) \tag{19}$$

A Gibbs distribution is the probability distribution with maximum entropy given some constraint. For example, the uniform distribution is the maximum entropy distribution that has support on a given interval $[a, b]$, the exponential distribution is the maximum entropy distribution given a positive mean, and the normal distribution is the maximum entropy distribution given a mean and a standard deviation. The ϕ's can be thought of as Lagrangian encoding of constraints.

Definition 2. *A distribution* $P(X) = \frac{1}{Z}\phi_1(D_1)\phi_2(D_2)...\phi_{K-1}(D_{K-1})\phi_K(D_k)$ *factorizes over a Markov network* H *if each* D_k *is a complete subgraph of* H*.*

As a reminder a complete subgraph is set of nodes in a network where there is a direct connection between all of the nodes, also called a clique.

Definition 3. *Let* H *be a Markov network structure, and let* $X_1 - -... - -X_k$ *be a path in* H*. Let* $Z \in X$ *be a set of observed variables. The path* $X_1 - ... - X_k$ *is active given* Z *if none of the* X_i*'s,* $i = 1, ...k$ *is in* Z*.*

Definition 4. *A set of nodes* Z *separates* X *and* Y *in* H*, which we denote* $\text{sep}_H(X; Y \mid Z)$ *if there is no active path between nodes* $X \in X$ *and* $Y \in Y$ *given* Z*. We define the global independencies associated with* H *to be* $I(H) = \{(X \perp\!\!\!\perp Y \mid Z) : \text{sep}_H(X; Y \mid Z)\}$*.*

From a proof similar to the one above and the Hammersly-Clifford theorem, it turns out that the following theorems hold

Theorem 2. *If P is a Gibbs distribution, then P factorizes over a Markov network H if and only if every independency in P is encoded in H*

2 Parameterization

Returning to our four block example, we saw that we can express the probability of the pattern of an assignment to all the blocks can be factored into functions that only look at the interactions between neighboring blocks. But we still need to choose the form of the functions ϕ.

$$\Pr(X) = \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A) \tag{20}$$

To return to our original assumptions, we said that we think that a black block has some influence on neighboring blocks to make them black. Let's call that influence w. And we'll assume that white blocks have the same influence, but in the opposite direction, i.e. $-w$.[1]

For a particular block i, the balance of influence h_I depends upon the number of white neighboring blocks, n_W and black neighboring blocks n_B.

$$h_i = w(n_W - n_B) \tag{21}$$

and we'll say that the probability the race of a block is

$$\Pr(\text{block}_i = \text{white}) = \frac{e^{h_i}}{e^{h_i} + e^{-h_i}} \tag{22}$$

$$\Pr(\text{block}_i = \text{black}) = \frac{e^{-h_i}}{e^{h_i} + e^{-h_i}} \tag{23}$$

$$\tag{24}$$

And that the probability for any particular pattern of race is

$$\Pr(X) = \prod \frac{e^{R_i h_i}}{e^{h_i} + e^{-h_i}} \tag{25}$$

Where R_i is an indicator variable that takes a value of 1 where the race of the block is white and a value of -1 if the block is black.

We can also express this probability in terms of the pairwise interactions of neighbors.

[1] I believe, but have not proven that this assumption of 'symmetry of influence' is required if we want represent the system as a Markov field.

$$\Pr(X) = \prod_i \frac{e^{R_i h_i}}{e^{h_i} + e^{-h_i}} \tag{26}$$

$$= \prod_i \frac{\prod_{j \in <ij>} e^{R_i R_j w}}{e^{h_i} + e^{-h_i}} \tag{27}$$

$$\tag{28}$$

Where $j \in < ij >$ are all the sites that are nearest neighbors of site i. Each pair will be appear in the numerator twice. So that

$$\Pr(X) = \frac{1}{Z} \prod_{<ij>'} e^{R_i R_j w'} \tag{29}$$

Where Z is normalizing constant that is the sum of all possible assignments, $< ij >'$ is every nearest neighbor only counted once, and $w' = w/2$.

This model is classically called the Ising model. We can see that it is a Gibbs distribution, which and we can also see that it encodes the same independencies of our racial preference block model, so that this distribution factorizes over network of influence.