Learning Neighborhoods: Method description part I – The models

Forest Gregg

February 20, 2013

Say we would like to think about how patterns of racial segregation in a city could arise from the local interactions of neighbors. We think there may be some process at the household level, that all else being equal, makes it more likely that neighbors will be of the same race than different races. We'd like to reason about what kinds of global patterns of segregation are likely to arise from these local processes.

To get going, let's make a couple of simplifying assumptions. First, we'll assume there are only two races: black and white. Second, we'll assume that only direct neighbors affect whether a black or white family lives in a house. More distant neighbors can influence the house, but only by influencing a directly neighboring house.

From these assumptions, we can build a probability distribution over every global patterns of racial segregation in a city, from complete segregation to complete integration.

Let's look at four houses, that we will call A, B, C, and D, Houses A and C are not neighbors and neither are B and D (Figure 1).

Now, let's say there is a particular pattern of black and white houses, for example House A–black, House B–black, House C–white, House D–white. We'll

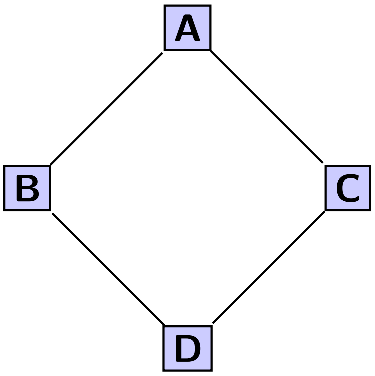

Figure 1: Markovian Dependent Variables

call this particular pattern x, and we'll say that $\Pr(X = x)$ is the probability of that pattern, while $\Pr(X)$ is the probability distribution over all possible patterns

We'd like to express this probability distribution only in terms of local interactions between neighboring houses. We'll do this in two steps. First we'll show that any probability distribution over these four houses can be expressed as a product of functions that only take in the value of neighboring houses, 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.

Independence and Factorization

Because we think that houses 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 $\phi_4(C) = \sum_B \phi_2(B, C)$.

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

Similarly

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

From which we can see that

$$\Pr(A, B \mid C) = \Pr(A \mid C) \Pr(B \mid C) \qquad (13)$$

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

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

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 houses as edges. We drew an edge between two houses when the houses 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* $\mathrm{sep}_H(X; Y \mid Z)$ *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* $I(H) = \{(X \perp\!\!\!\perp Y \mid Z) : \mathrm{sep}_H(X; Y \mid Z)\}$*.*

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 house example, we saw that we can express the probability of the pattern of an assignment to all the houses as the product of functions that only look at the interactions between neighboring houses. 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 house has some influence on neighboring houses to make them black. Let's call that influence w. And we'll assume that white houses have the same influence, but in the opposite direction, i.e. $-w$.[1]

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

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

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

$$\Pr(\text{house}_i = \text{white}) = \frac{e^{h_i}}{e^{h_i} + e^{-h_i}} \tag{22}$$
$$\Pr(\text{house}_i = \text{black}) = \frac{e^{-h_i}}{e^{h_i} + e^{-h_i}} \tag{23}$$
$$\tag{24}$$

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 house is white and a value of -1 if the house is black.

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

$$\Pr(X) = \prod_i \frac{e^{R_i h_i}}{e^{h_i} + e^{-h_i}} \tag{26}$$
$$= \prod_i \frac{\prod_{j \in <ij>} e^{R_i R_j w}}{e^{h_i} + e^{-h_i}} \tag{27}$$
$$\tag{28}$$

[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.

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

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

$$= \frac{1}{Z} \exp(\sum_{<ij>'} R_i R_j w') \tag{30}$$

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

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

3 Extensions

We have shown how we can develop an relatively simple formula for the probability distribution over global patterns that arise from local interactions. In the model we worked through, the smallest unit is only influenced by it's immediate neighbors and can only take on two values–either the house is occupied by a black family or a white one.

Perhaps we think that there are some other differences among houses that make it more likely for one race to occupy it than another. We can naturally extend the model to include a house specific term u_i.

$$\Pr(X) = \frac{1}{Z} \exp(\sum_{<ij>'} R_i R_j w' + \sum_i u_i R_i) \tag{31}$$

Indeed, we are not limited to multiplication, but can use choose any functions $\epsilon_{i,j}$ and ϵ_i, and we will still have a valid probability distribution over global patterns.

$$\Pr(X) = \frac{1}{Z} \exp(\sum_{<ij>'} \epsilon_{i,j}(R_i, R_j) + \sum_i \epsilon_i(R_i)) \tag{32}$$

A simple choice of $\epsilon_{i,j}$ that will allow us to deal with more than two races is

$$\epsilon_{i,j}(R_i, R_j) = \begin{cases} 0 & R_i = R_j \\ \lambda_{i,j} & R_i \neq R_j \end{cases} \tag{33}$$

Or the model can be extended so that $\epsilon_{i,j}$ is a site specific distance function between x_i and x_j so that likelihood of neighbors being the same race can be function of how similar they are in other ways.

4 Neighborhoods

The time has come to leave discussion of racial segregation, and take up the phenomena we are really interested in here: neighborhoods.

If we make the assumption that a household can belong to one and only one neighborhood, we can use all the machinery of Markov Random Fields we just developed to try to learn what are the features, block by block that cohere into a neighborhood. This is a very troublesome assumption, and we we'll return to it, but let's see what we can do if we make it.

We'll use this machinery twice. First, we'll use it to help create a set of target "neighborhoods" and then we'll use it to learn what local features might have produced those neighborhoods.

4.1 MRF's as priors

I have a nightly updated database of geocoded Craigslist apartment rentals, sublet, and roommate listings. For most of these listings, the poster entered some text in the "Specific Location" field. With some minimal pre-processing, we can use these data as observations of claims that geographical points are in some neighborhood.

Using kernel density estimation, we can use this point data to estimate a continuous probability distributions that any point in the city will be claimed to be in any of the neighborhoods (Figure 2.

However, KDE and other smoothers are... smooth. The estimates may converge to true probability estimates as the amount of data increases, but in the meantime these methods will produce overly smooth decision boundaries between neighborhoods.

We can correct for these artifacts if we have valid prior beliefs about what kind of geographical features are likely to divide neighborhoods. For example, we might believe that major roads, the river, and railroad embankments are more likely to divide neighborhoods than residential streets.

We can encode these beliefs in the following model

$$\Pr(Y) = \frac{1}{Z}\exp\left(\sum_{<ij>} \epsilon_{i,j}(\mathrm{Name}_i, \mathrm{Name}_j) + \sum_i \epsilon_i(\mathrm{Name}_i)\right) \tag{34}$$

Where Y is a configuration of neighborhood names over census blocks, i indexes census blocks and the function $\epsilon_{i,j}$ has the following form (blocks count as neighbors if they share any edge).

$$\epsilon_{i,j}(\mathrm{Name}_i, \mathrm{Name}_j) = \begin{cases} 0 & \mathrm{Name}_i = \mathrm{Name}_j \\ 0 & \text{if } i,\, j \text{ are separated by a feature} \\ \lambda & R_i \neq R_j \text{and not separated} \end{cases} \tag{35}$$

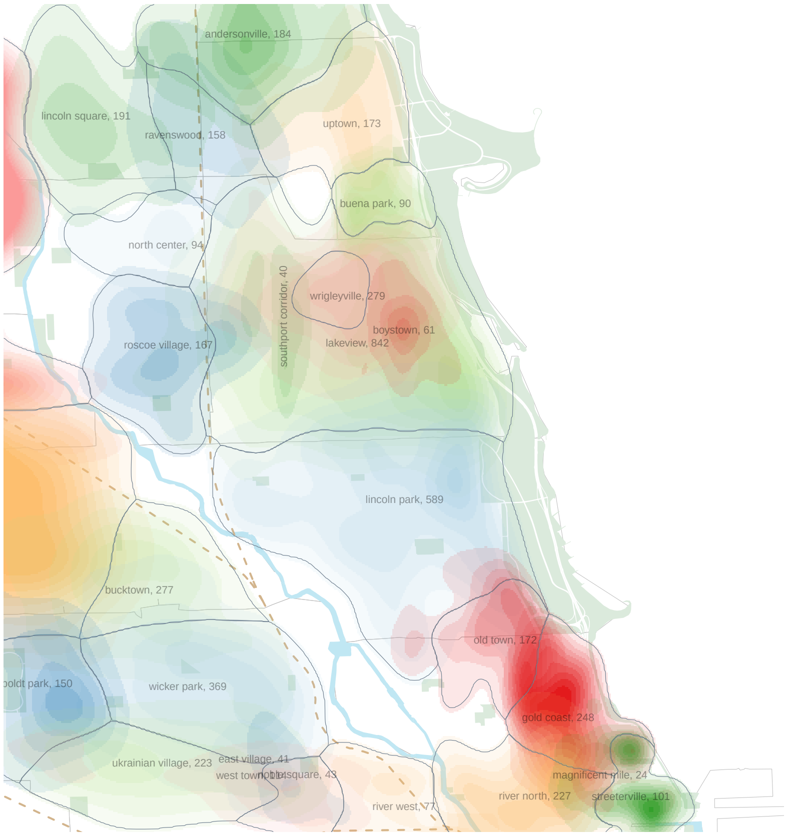

Figure 2: KDE probability estimates of neighborhood claims on North Side

ϵ_i is just the estimated probabilities that a block i will be claimed to belong to a neighborhood for every neighborhood in the city. This comes from calculating the densities kde estimates at the centroid of a census block.

We'd like to find the assignment Y that maximizes the $\Pr(Y)$. In other words, we'd like to find a labeling of every block in the city that is most likely given our data-driven estimates of labels and our beliefs about where boundaries should fall.

$$\begin{aligned}
&\arg\max_Y \frac{1}{Z} \exp\left(\sum_{<ij>} \epsilon_{i,j}(\text{Name}_i, \text{Name}_j) + \sum_i \epsilon_i(\text{Name}_i)\right) \\
&\arg\max_Y \exp\left(\sum_{<ij>} \epsilon_{i,j}(\text{Name}_i, \text{Name}_j) + \sum_i \epsilon_i(\text{Name}_i)\right) \\
&\arg\max_Y \sum_{<ij>} \epsilon_{i,j}(\text{Name}_i, \text{Name}_j) + \sum_i \epsilon_i(\text{Name}_i)
\end{aligned}$$

Unfortunately, our typical approaches to finding the maximum of a distribution cannot help us here. Configurations are discrete, so derivative based methods won't work. The number of possible, discrete configurations of neighborhood labels is $(\#\text{ of neighborhood names})^{\#\text{ of census blocks}}$ so we cannot search the space. Markov Chain Monte Carlo approaches are also intractable because they require estimating the normalizing constant Z which can is the sum of probability of the $(\#\text{ of neighborhood names})^{\#\text{ of census blocks}}$ configurations.

Surprisingly, when there are only two labels, we can exactly find the Y that maximizes get $\Pr(Y)$, and we can typically get very good approximations of Y in the general case. It turns out that, in the binary case, we can create a auxillary graph with a node corresponding to every block in our mode, with two additional special nodes called the source and the sink. This graph has a directed edge from the source node to every normal node and every normal node has a directed node to the target node. If we set the capacity of flow between nodes correctly, then finding the minimum number of cuts to separate the source and target nodes will correspond to finding the assignment labels to nodes that maximizes $\Pr(Y)$. Finding this minimum cut of graph can be done in polynomial time. There are extensions of this minimum cut method to the multi-label case that do not guarantee optimality, but perform very well in practice. The details of all this are fascinating, and perhaps takes us too far afield.

Using one of these graph-cut extensions, we find the segmentaion shown in Figure 3

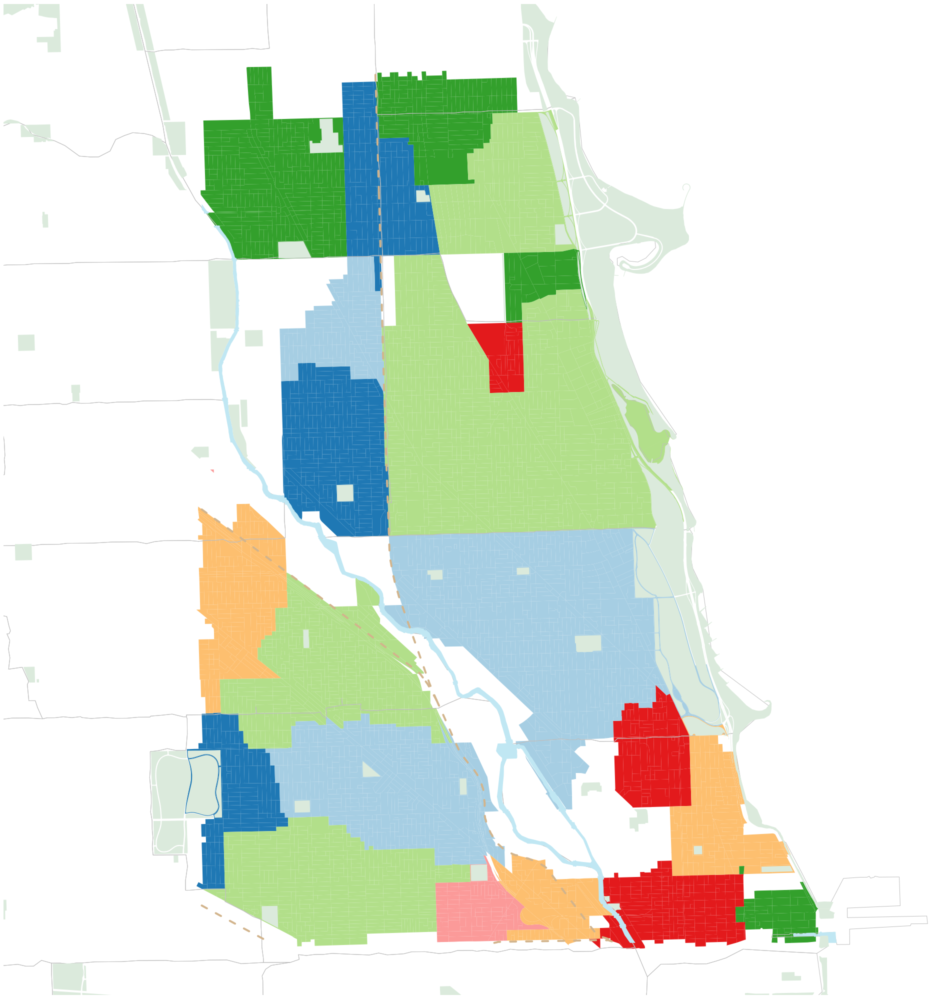

Figure 3: MAP estimate of neighborhood labels