Adapted Nested Dirichlet Processes for Built Environment Data

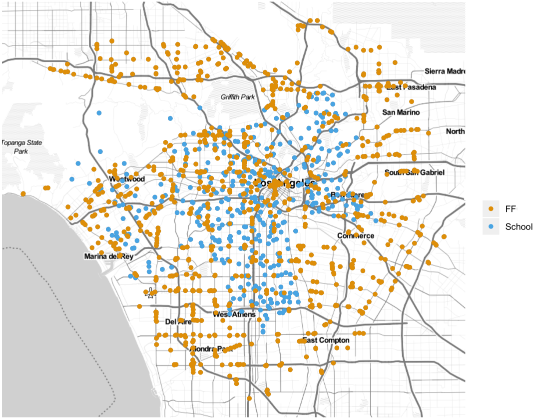
Adam Peterson Veronica Berrocal Brisa Sánchez

DataPhilly

June 17, 2020

Motivating Questions

- Does where we live with respect to stores, schools, parks, etc. matter?
  1. Are there patterns in accessibility to these amenities?
  2. Are these patterns relevant to our health?



Illustration

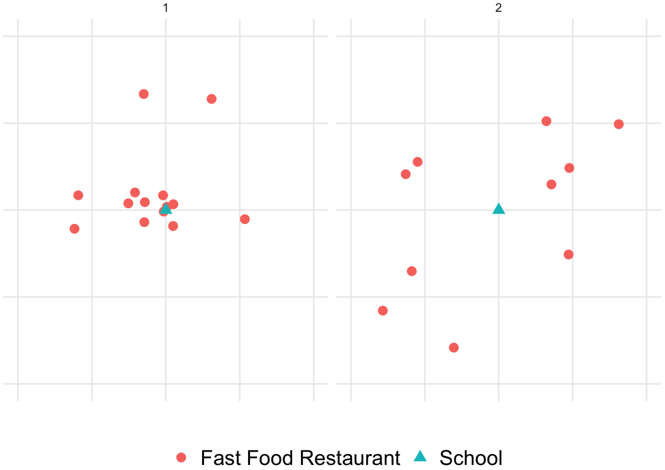
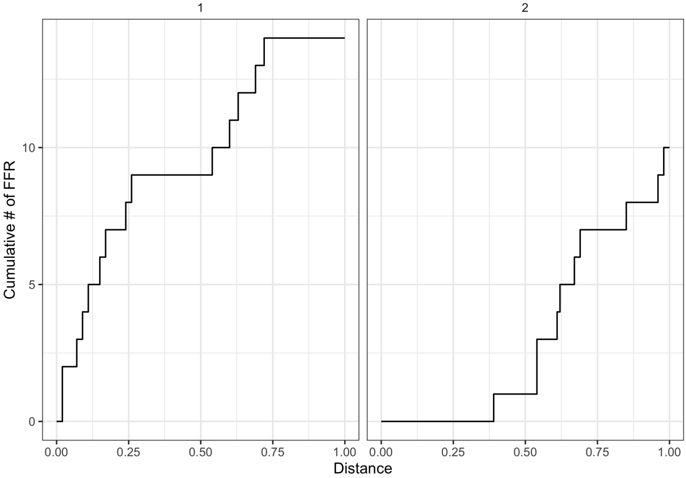
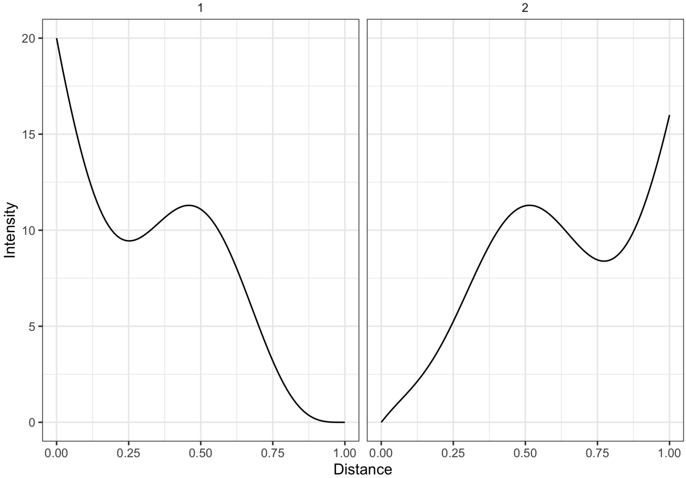


Illustration (2)



Underlying Model



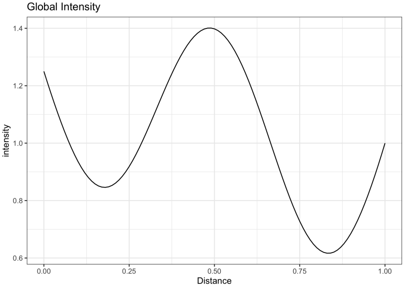
Complicating Questions

1. How do we identify these intensity functions?
   - We don't know what shape they are - need to estimate them flexibly!
2. How many intensity functions?
   - Could be as many as there are schools! (Probably not)

Intensity Estimation

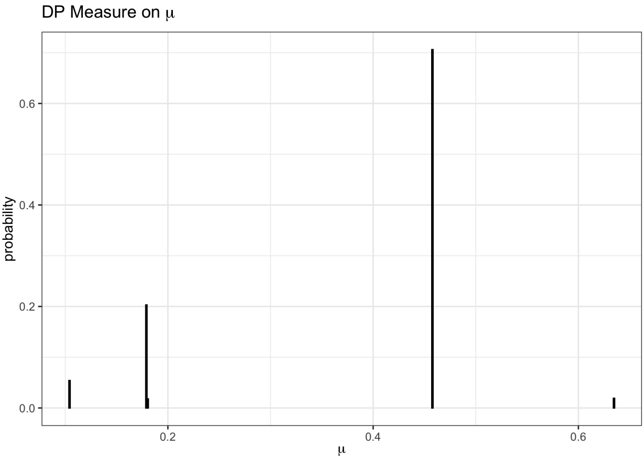
Mixture model

- Express the observed density as a mixture of simpler, more easily parameterized densities
- Obstacle: How many simpler densities should we use?
- Solution: Dirichlet Process



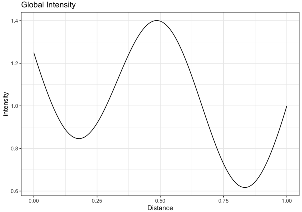
Dirichlet Process

Dirichlet Process(DP): A distribution on distributions

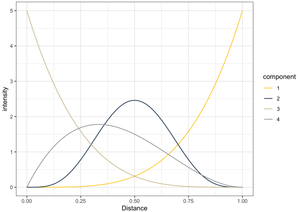


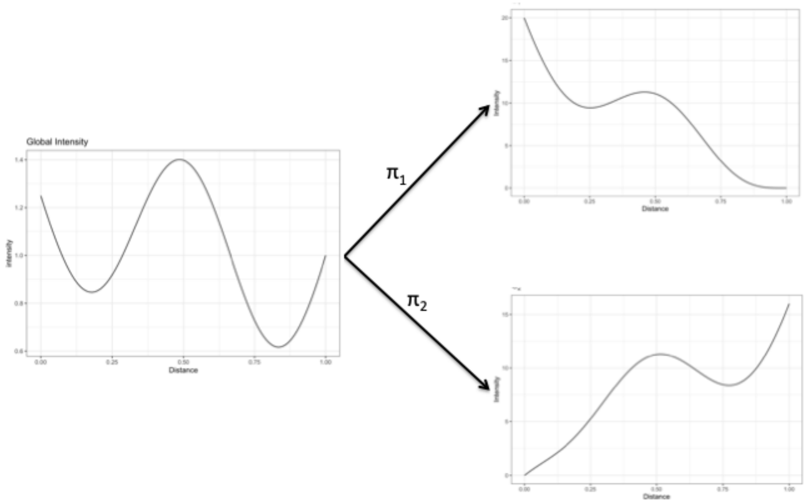
Intensity Estimation - Dirichlet Process

This will allow us to estimate the *global* intensity ...



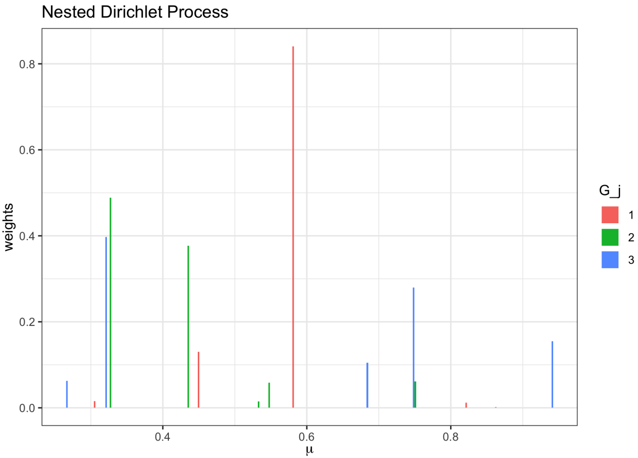
=




Global Intensity
1.4
1.2
1.0
0.8
0.6
Intensity
0.00
0.25
0.50
0.75
1.00
Distance
π_1
π_2
Intensity
Distance
Intensity
Distance

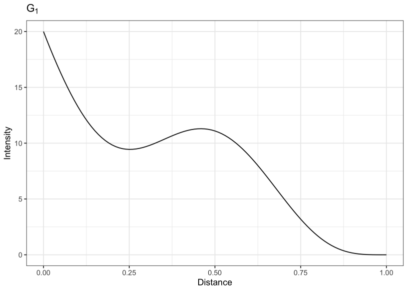
Sub Density Estimation - Nested Dirichlet Process

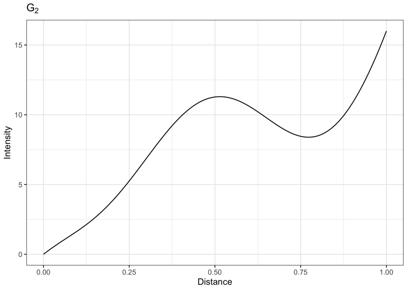
"Just as the DP is a distribution on distributions, the NDP can be characterized as a distribution on the space of distributions on distributions." (Rodriguez et al. 2008)



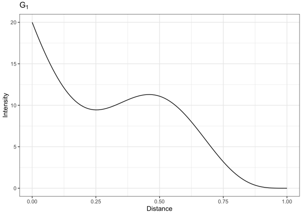
Heirarchy Layer 1

School	Distances			
1	0.05	0.09	0.15	0.23
2	0.03	0.06	0.18	
…	…	….	…	…
…	…	…	…	…
J-1	0.55	0.67		
J	0.75	0.84	0.93	

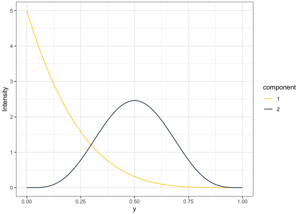


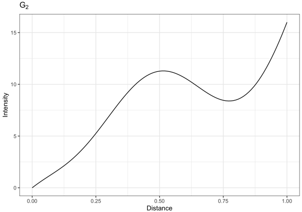


Heirarchy Layer 2

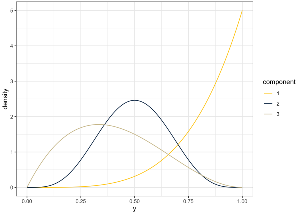


=





=



Adapting the NDP: Connecting to Health Outcomes

- The NDP only helps us to *identify* the differing patterns in spatial exposure.
- We need a different strategy to *link* these patterns to a health outcome of interest.
- Health Outcomes Models:
  - "Conservative" GLM (CGLM)
  - Bayesian Kernel Machine Regression (BKMR)

Second Stage Analysis: Health Outcomes Models

BKMR

$$\text{logit}(\pi_j) = \alpha + \boldsymbol{Z}_i^T \boldsymbol{\delta} + h_j(\boldsymbol{P})$$
$$h_j(\boldsymbol{P}) \sim \mathcal{GP}(\boldsymbol{0}, \kappa(\boldsymbol{p}_j, \boldsymbol{p}_{j'} | \sigma, \phi))$$

- $\boldsymbol{P}$ is the pairwise probability matrix of co-cluster membership derived from the cluster assignment labels
- $\kappa(\cdot, \cdot | \sigma, \phi)$ a valid covariance function

CGLM

$$\text{logit}(\pi_{j*}) = \alpha_{j^*,k} + \boldsymbol{Z}_{j*}^T \boldsymbol{\delta}$$

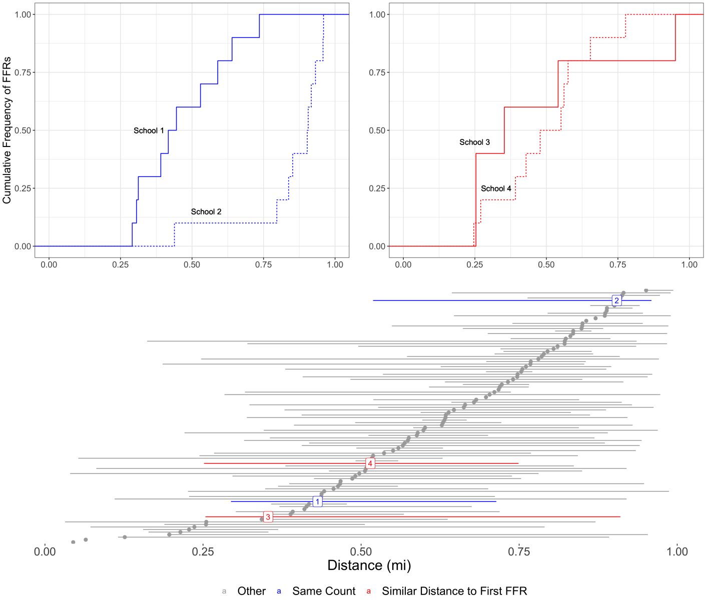
j^* selected by intersection of posterior credible ball bounds

Application: FFR Exposure around CA highschools

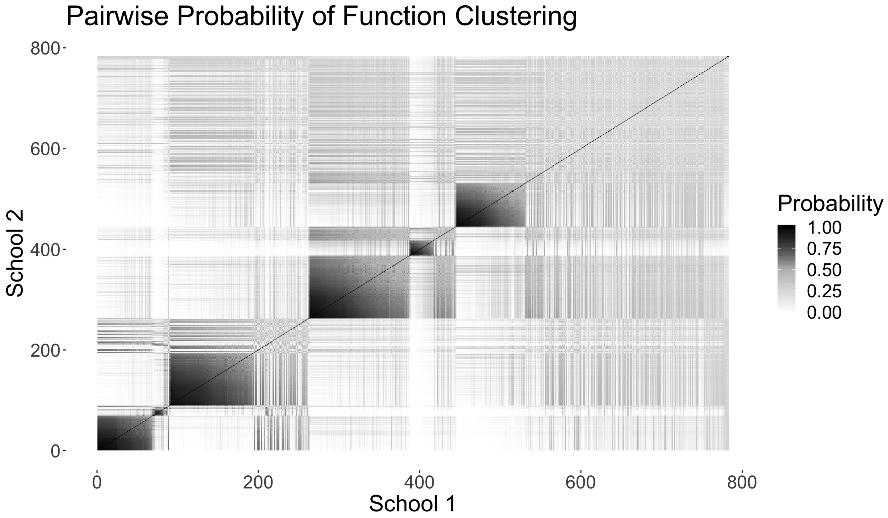
- 782 high schools in CA during academic year 2010
  - $\approx$ 4000 Fast Food Restaurants within 1 mile of the school.
- Proportion of obese 9th graders estimated as a function of exposure profile, adjusting for relevant covariates.

bendr

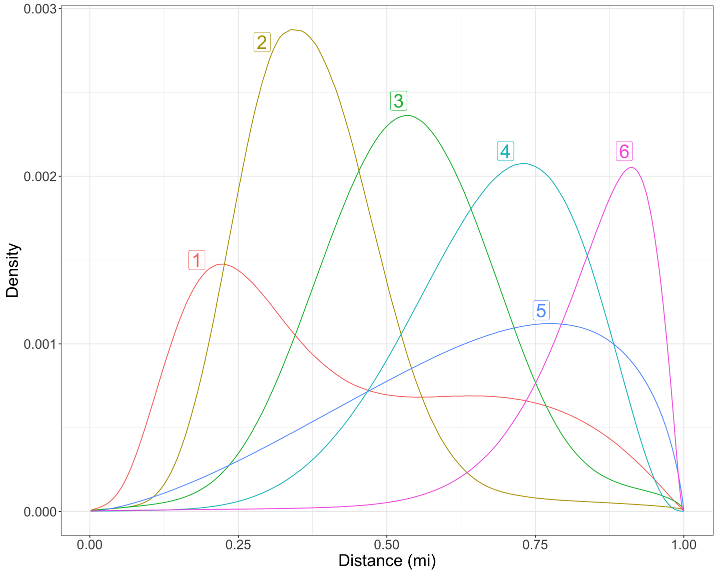
California FFR Exposure



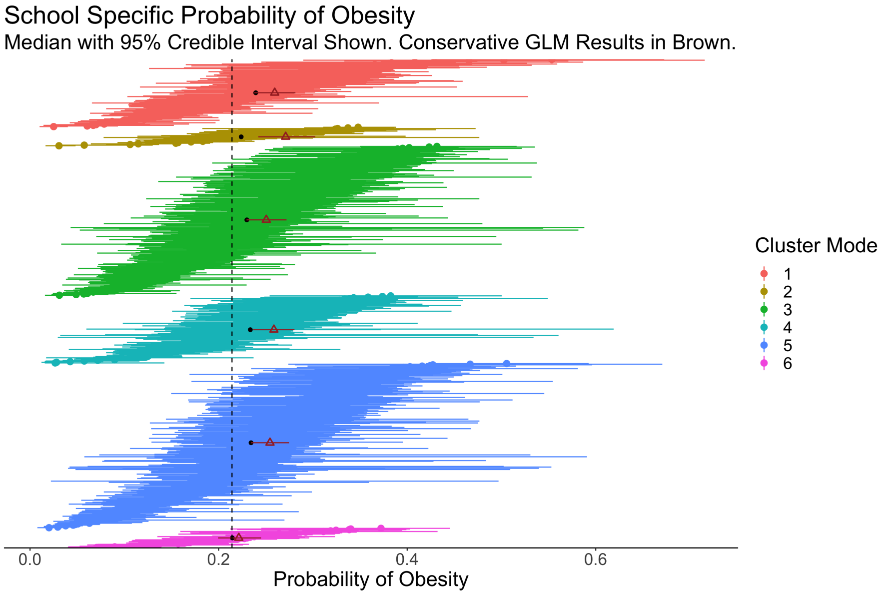
NDP Results: Co-Clustering Probabilities



NDP Results: Cluster Intensities



Health Outcome Models



Questions?

Supplementary Material

Adapted NDP: Model Assumptions

Model

$$
\begin{aligned}
p(\{r_{ij}\}_{(i,j)=(1,1)}^{(n_j,J)} | f_j(r), n_j) &\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} f_j(r_{ij}) \\
f_j(r) &= \int \mathcal{K}(r|\theta) G_j(\theta) \\
G_j &\overset{iid}{\sim} Q \\
Q &\sim DP(\alpha', DP(\rho, H_0))
\end{aligned}
$$

Assumptions

- Inhomogenous Poisson Process:
  - conditional on n_j the distances $r_{ij} \overset{iid}{\sim} f_j(\cdot)$
- Independence between schools

Model Specification

$$\lambda_j(r) = \gamma_j f_j(r) \quad \gamma_j \in \mathbb{R}^+$$

$$r'_{ij} = \text{probit}(r_{ij})$$

$$f_j(r') = \int \text{Normal}(r'|\mu, \tau) dG_j((\mu, \tau))$$

$$G_j \overset{iid}{\sim} Q$$

$$Q \equiv \sum_{k=1}^{\infty} \pi_k \delta_{G_k(\cdot)}(\cdot) \approx \sum_{k=1}^{K} \pi_k \delta_{G_k(\cdot)}(\cdot)$$

$$G_k \equiv \sum_{l=1}^{\infty} w_{lk} \delta_{(\mu,\tau)_{lk}}(\cdot) \approx \sum_{l=1}^{L} w_{lk} \delta_{(\mu,\tau)_{lk}}(\cdot)$$

$$Q \equiv DP(\alpha, DP(\beta, G_0))$$