Spatial Compositional Complexity

v0.1

Phil Chodrow

March 29th, 2016

1 Purpose

The purpose of this document is to record Phil's notes on information-theoretic methods for summarizing spatially-structured arrays of empirically-observed compositional data.

1.1 Motivation

A spatially-organized summary of such a data set would provide candidate answers to each of the following questions:

1. What is the "scale" of spatial heterogeneity in a city? For example, if Dorchester is substantially poorer than the rest of Boston, is that a 'Dorchester-level' phenomenon? Or is it an amalgamation of smaller spatial phenomena that may need a separate or coordinated approach?

2. Suppose we'd like to construct a generative theory of a spatially-distributed phenomenon based on some data. In the study of complex systems, we may not insist on a theory that perfectly reconstructs the data, but we should expect it to get the 'high-level' features right. What are those features that such a theory should replicate?

3. To what extent can a set of compositional data be explained by spatial effects? What phenomena in the data are decidedly "nonspatial"?

4. How complex is a city's distribution of income, as compared to other cities?

2 Methods

2.1 Data

Suppose that, for each $i = 1, \dots, I$, we are given:

1. A vector $x^i \in \mathbb{R}^n$.

2. A vector $\mathbf{p}^i \in \mathcal{S}^J$, where $\mathcal{S}^J = \{\mathbf{p} | \mathbf{p} \geq 0,\ \sum_{j=1}^J p_j = 1\}$. Such data is sometimes referred to as "compositional data" because it reflects the composition of a whole, rather than individual magnitudes.

A natural class of such data are demographics organized by spatial tracts, for which $n = 2$. For example, the Census collects data on income by block group. For example, in the 2012 American Community Survey data for the city of Boston, the number of block groups $I = 646$ and the number of categories $J = 16$. Each $x^i \in \mathbb{R}^2$ would then be the centroid of the corresponding tract.

2.2 Objective Function

We would like to generate, at each location x^i, a model prediction $\mathbf{q}^i$ of the data $\mathbf{p}^i$ that is in some sense "spatially structured." Since $\mathbf{p}^i$ is compositional, a natural loss function is the Kullback-Leibler divergence of the estimate $\mathbf{q}^i$ from the observed data $\mathbf{p}^i$:

$$D[\mathbf{p}^i \| \mathbf{q}^i] \triangleq \sum_{j=1}^{J} p_j^i \log\left(\frac{p_j^i}{q_j^i}\right) \tag{1}$$

If

$$\mathbf{P} = \begin{bmatrix} | & \cdots & | \\ \mathbf{p^1} & \cdots & \mathbf{p}^I \\ | & \cdots & | \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} | & \cdots & | \\ \mathbf{q^1} & \cdots & \mathbf{q}^I \\ | & \cdots & | \end{bmatrix}, \tag{2}$$

then we can write the total objective function for the problem as

$$f_{\mathbf{P}}(\mathbf{Q}) = \sum_{i=1}^{I} D[\mathbf{p}^i \| \mathbf{q}^i]\ . \tag{3}$$

The objective function is interpretable as the number of additional bits needed to fully specify the true data $\mathbf{P}$ after approximating it with $\mathbf{Q}$. Of course, setting $\mathbf{Q} = \mathbf{P}$ achieves the lower bound $f_{\mathbf{P}}(\mathbf{P}) = 0$, but this achieves no modeling value.

2.3 Spatial Structure

To impose spatial structure on our model, we require that our estimates $\mathbf{Q}$ have the structure

$$\mathbf{q}^i = \mathbf{R}\mathbf{\Lambda}(x^i) \tag{4}$$

where $K \leq I$,

$$\mathbf{R} = \begin{bmatrix} | & \cdots & | \\ \mathbf{r}^1 & \cdots & \mathbf{r}^K \\ | & \cdots & | \end{bmatrix} \in \mathbb{R}^{J \times K} \tag{5}$$

is a matrix of "representative" distributions, and $\mathbf{\Lambda} : \mathbb{R}^n \to \mathbb{R}^K$ describes how the representative distributions $\mathbf{R}$ mix spatially. In order for $\mathbf{q^i}$ to be a valid probability distribution, we require that $\mathbf{\Lambda}(x^i) \geq 0$ and $\sum_{k=1}^K \lambda_k(x^i) = 1$ for all i. Thus, (4) states

that each $\mathbf{q}^i$ is a convex combination of the representative distributions $\{\mathbf{r}^k\}$, where the convex coefficients are determined by the spatial location x^i.

The modeling task is to find the matrix $\mathbf{R}$ and function $\mathbf{\Lambda}$. We need to restrict the set of candidate functions $\mathbf{\Lambda}$ such that:

1. Elements of this set set reflect reasonably intuitive spatial structure.
2. Optimizing over this set is computationally feasible, which we take to imply finite-dimensional.

A set of functions that fits the bill is:

$$\mathbf{L} = \left\{ \Lambda(x) = \frac{\mathbf{c} \bullet \Phi(x|\Theta)}{\mathbf{c} \cdot \Phi(x|\Theta)},\ \Phi(x|\Theta) = (\phi(x|\theta^1), \ldots, \phi(x|\theta^K)),\ \Theta = (\theta^1, \ldots, \theta^K),\ \mathbf{c} \in \mathbb{R}_+^K \right\}, \tag{6}$$

where $\bullet$ is the Hadamard product and $\phi(x|\theta)$ is the multivariate Gaussians distribution with parameters θ evaluated at x. The set $\mathbf{L}$ is an attractive setting for us because:

1. Elements of this set depend spatially on x through unimodal distributions, reflecting the idea that the influence of each representative distribution $\mathbf{q}^i$ is "centered" at the mean of the corresponding density and decays with distance. Some may be more absolutely "influential" than others, and will have high corresponding entries of $\mathbf{c}$.
2. This set is finite dimensional: we need to fit KJ entries of $\mathbf{R}$, $\frac{n(n+3)}{2}$ parameters for each of K Gaussians, and K parameters $\mathbf{c}$, giving a total problem dimension of $K\left(J + \frac{n(n+3)}{2} + 1\right)$.

2.4 Problem Statement

With this framework in place, we can define our optimization problem. First, since

$$D[\mathbf{p}^i \| \mathbf{q}^i] = \sum_{j=1}^{J} p_j^i \log \frac{p_j^i}{q_j^i} = \sum_{j=1}^{J} p_j^i \log p_j^i - \sum_{j=1}^{J} p_j^i \log q_j^i \ , \tag{7}$$

we can define $g_i(\mathbf{q}^i) = -\mathbf{p}^i \cdot \log \mathbf{q}^i$ (the logarithm is evaluated componentwise) and minimize

$$-\sum_{i=1}^{I} g_i(\mathbf{q^i}) \ . \tag{8}$$

Then, if we define

$$h_i(\mathbf{R}, \Theta, \mathbf{c}) \triangleq \mathbf{R} \frac{\mathbf{c} \bullet \Phi(x^i|\Theta)}{\mathbf{c} \cdot \Phi(x^i|\Theta)} \ , \tag{9}$$

we can write our main optimization problem as

$$\begin{aligned} \min_{\mathbf{R},\Theta,\mathbf{C}} \quad & \sum_i (g_i \circ h_i)(\mathbf{R}, \Theta, \mathbf{c}) \\ \text{subject to} \quad & \mathbf{c} \geq 0 \\ & \mathbf{r}^k \in \mathcal{S}^J \quad \forall k \ . \end{aligned} \tag{10}$$

The problem (10) has the following properties:

1. The feasible region is convex.
2. The objective function is likely nonconvex.
3. The objective function is smooth.

Thus, finding a local optimum would likely be easy with simple first- or second-order methods, but finding a global optimum could be nontrivial.

3 Open Questions

- Is (10) convex up to permutations of the indices $k = 1, \ldots, K$? This would imply that the only local optima are $K!$ global optima, which are identical up to relabeling the parameters.
- How much "signal" should we expect in a standard data set?
- How nearly-optimal are local minima? How bad would it be to "settle" for a local optimum?