Notes on Urban Information Theory

v0.0

Phil Chodrow

March 18th, 2016

1 Purpose

The purpose of this document is to record Phil's notes on information-theoretic methods in summarizing spatially-structured arrays of empirical distributions of data. An example of such a data set would be income distributions, summarized on a block level for an entire city. At each (recorded) point in the data set, there is an empirical probability distribution of income.

1.1 Motivation

A spatially-organized summary of such a data set would serve at least three useful purposes, corresponding to answers to 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. How complex is a city's distribution of income, as compared to other cities?

The answers an information-theoretic summary would provide are:

1. Suppose that we run our model and find that the spatial organization of income distributions in Dorchester is best modeled as a mixture of three different 'core' distributions. This would be reason to think that there are (at least) three phenomena deserving of study and intervention in the area.

2. A mixture model would 'compress' the spatial organization of distributions, recording precisely those features that contain the most information about the system. A reasonable approach would be to insist that a generative theory recover approximately these features, at some chosen level of detail.

3. A model of the spatial organization of distributions can be considered as a form of data compression. To see this, suppose that we have a collection of $\mathcal{X}$ of spatial observations, where $|\mathcal{X}| = N$. At each point $x \in \mathcal{X}$, we have a distribution p^x over an alphabet $\mathcal{Y}$ of possible categories, where $|\mathcal{Y}| = M$. As an example, $\mathcal{X}$ might consist of the centroids of Census tracts, and $\mathcal{Y}$ the income categories used in the survey. To fully describe the uncompressed data would require $N(M-1)$ numbers – since each p^x must be a valid probability distribution, we have $\sum_{y\in\mathcal{Y}} p^x(y) = 1$ for all $x \in \mathcal{X}$, which eliminates a degree of freedom at each location. Suppose that, for a given number P, we were able to provide an optimal description of the organization using just P numbers, where $P < N(M-1)$ and where optimality is measured in terms of minimal information loss. We could then plot information loss against P, obtaining a complexity profile curve. Such curves could be compared between cities, or between spatially-organized observables in the same city, giving a quantitative way to compare complexity across locations and phenomena.

2 Methods

2.1 Main Problem

Suppose that at each location $x \in \mathcal{X}$, we have an observed distribution p^x over $\mathcal{Y}$. Thus, our data consist of the matrix $p^x(y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We would like to approximate $p^x(y) \approx q^x(y)$, where

$$q^x(y) = \sum_{i=1}^{K} \lambda_i(x) q_i(y) \tag{1}$$

Intuitively: our model consists of K 'representative' distributions q_i, $i = 1, \dots, K$, that mix spatially acording to coefficients $\lambda_i(x)$ that measure how influential the distribution q_i is at point x. Our computational task is to specify λ_i and q_i to make this approximation 'good' in some relevant sense.

Note that, while $q_i \in \mathcal{P}^{\mathcal{Y}}$ and is therefore finite-dimensional, λ_i is as yet unconstrained. We need to impose an assumption about the functional form of λ_i in order to make this problem finite-dimensional. Suppose that we require that each q_i be a member of a family of functions specified by J parameters. Then, we can specify the model in $K(M-1+J)$ parameters, which will be an improvement if $K << N$.

A natural objective function is the divergence of the approximation q^x from the data data p^x. We are therefore moved to consider the following problem:

$$\min_{\lambda_i, q_i} \sum_{x\in\mathcal{X}} D[p^x \| q^x] = \min_{\lambda_i, q_i} \sum_{x\in\mathcal{X}} D\left[p^x \middle\| \sum_{i=1}^{K} \lambda_i(x) q_i \right] . \tag{2}$$

A solution to this problem would enjoy a brand of information-theoretic optimality, in the sense that such a model would require the least additional information to fully recover the original data p^x.

2.2 Specifying λ_i

For the sake of concreteness, we will primarily focus on the case in which each λ_i is proportional to an element of a parameterized family of functions:

$$\lambda_i(x) \propto \pi(x|\theta_i) \ , \tag{3}$$

where each θ_i is a vector of parameters. Since we must have $\sum_i \lambda_i(x) = 1$ for all $x \in \mathcal{X}$, this gives

$$\lambda_i(x) = \frac{c_i \pi(x|\theta_i)}{\sum_i c_i \pi(x|\theta_i)} \tag{4}$$

where c_i are constants. As above, our problem then has $K(M+J)$ dimensions, since we must also estimate K multiplying constants. This should be on the order of 10^2 for the kinds of problems Phil has in mind.

2.3 Approximation

One way to approach this problem is to use the convexity of the divergence D to generate an upper bound on (2) and attempt to minimize that. Such an approach might proceed as follows.

Using convexity, we can write the objective function of (2) as

$$\begin{aligned} D[p^x \| q^x] &= D\left[p^x \middle\| \sum_{i=1}^{K} \lambda_i(x) q_i \right] && (5) \\ &\leq \sum_{i=1}^{K} \lambda_i(x) D\left[p^x \| q_i\right] \ . && (6) \end{aligned}$$

Equation (6) has nice interpretation: roughly, a large divergence of q_i from p^x should only concern us if we would expect q_i to be very influential at x; i.e. if $\lambda_i(x)$ is large. We can simplify (6) further as follows:

$$\begin{aligned}\sum_{i=1}^{K} \lambda_i(x) D\left[p^x \| q_i\right] &= \sum_{i=1}^{K} \lambda_i(x) \sum_{y\in\mathcal{Y}} p^x(y) \log \frac{p^x(y)}{q_i(y)} \\ &= \sum_{i=1}^{K} \lambda_i(x) \sum_{y\in\mathcal{Y}} \left[p^x(y) \log p^x(y) + p^x(y) \log \frac{1}{q_i(y)}\right] \\ &= \sum_{i=1}^{K} \lambda_i(x) \sum_{y\in\mathcal{Y}} p^x(y) \log p^x(y) + \sum_{i=1}^{K} \lambda_i(x) \sum_{y\in\mathcal{Y}} p^x(y) \log \frac{1}{q_i(y)} \\ &= \sum_{i=1}^{K} \lambda_i(x) H[p^x] + \sum_{i=1}^{K} \lambda_i(x) \sum_{y\in\mathcal{Y}} p^x(y) \log \frac{1}{q_i(y)} \\ &= H[p^x] + \sum_{i=1}^{K} \lambda_i(x) \sum_{y\in\mathcal{Y}} p^x(y) \log \frac{1}{q_i(y)} \ ,\end{aligned}$$

where the last equality follows from the fact that $\sum_{i=1}^{K} \lambda_i(x) = 1$ for all $x \in \mathcal{X}$. The first term depends only on the data, not on the model. We can therefore solve (6) by solving the equivalent problem

$$\min_{\lambda_i, q_i} - \sum_{x\in\mathcal{X}} \sum_{i=1}^{K} \sum_{y\in\mathcal{Y}} \lambda_i(x) p^x(y) \log q_i(y) \ . \tag{7}$$

2.4 EM Algorithm

Since this problem appears to be related to clustering tasks, we might reasonably expect that there should be an approach based on the EM algorithm.

2.5 KKT Conditions

We can express our defining problem a bit more explicitly as an optimization problem as follows. As above, we let

$$\lambda_i(x) = \frac{c_i \pi(x|\theta_i)}{\sum_i c_i \pi(x|\theta_i)} \ . \tag{8}$$

Furthermore, index the elements of $\mathcal{Y}$ by j, so that we can write $q_i(y_j) = q_{ij}$. Similarly, index the elements of $\mathcal{X}$ by k, so that we can write $p^{x_k}(y_j) = p_{jk}$. We can then write our main problem as

$$\min_{\lambda_i, q_i} \sum_{x\in\mathcal{X}} D[p^x \| q^x] = \sum_{x\in\mathcal{X}} H[p^x] + \sum_{x\in\mathcal{X}} p^x(y) \log \frac{1}{q^x(y)} \ , \tag{9}$$

so it's equivalent to solve

$$\min_{\lambda_i,q_i} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p^x(y) \log \frac{1}{q^x(y)} = \min_{\lambda_i,q_i} - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p^x(y) \log q^x(y) \tag{10}$$

$$= \min_{\lambda_i,q_i} - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p^x(y) \log \left(\sum_{i=1}^{K} \lambda_i(x) q_i(y) \right) \tag{11}$$

$$= \min_{c_i,\theta_i,q_i} - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p^x(y) \log \left(\sum_{i=1}^{K} \frac{c_i \pi(x|\theta_i)}{\sum_i c_i \pi(x|\theta_i)} q_i(y) \right) \tag{12}$$

$$= \min_{c_i,\theta_i,q_i} - \sum_{k} \sum_{j} p_{jk} \log \left(\sum_{i} \frac{c_i \pi(x_k|\theta_i)}{\sum_i c_i \pi(x_k|\theta_i)} q_{ij} \right) \tag{13}$$

$$= \min_{c_i,\theta_i,q_i} - \sum_{k} \sum_{j} p_{jk} \log \left(\sum_{i} c_i \pi(x_k|\theta_i) q_{ij} \right) - \log \left(\sum_{i} c_i \pi(x_k|\theta_i) \right) \tag{14}$$

subject to the constraints $q_{ij} \geq 0$, $\sum_j q_{ij} = 1$ for all i, and $c_i \geq 0$.

2.6 KL Divergence Under Logratio Transformation

Logratio transformations are a helpful class of transformations for working with compositional data. In some sense, we are working with the data at point x under two aspects: first, as an empirically-observed distribution that has information-theoretic measures, and second as a data point that happens to have a normalization property. We need to keep these two aspects clearly delineated if we are going to avoid confusion.

The standard additive logratio transformation of a variable $p = (p_0, \ldots, p_n)$ is

$$P = L(p) = \left[\ln \frac{p_1}{p_0}, \ldots, \ln \frac{p_n}{p_0} \right] . \tag{15}$$

In term, we can invert L to write

$$p = L^{-1}(P) = \mathcal{C} \exp([0, P]) \ , \tag{16}$$

where $\mathcal{C}$ is the simplex closure operation given by

$$\mathcal{C}y = \frac{y}{\sum_i y_i} \ . \tag{17}$$

In this section, we'd like to understand how the divergence between two distributions changes under the logratio transform. The rationale for considering this question is that it may be easier to optimize the divergence in transformed coordinates. We know that

$$\operatorname*{argmin}_{\bar{p}} D[p\|q] = \operatorname*{argmin}_{q} \quad - \sum_{i=0}^{N} p_i \log q_i \ . \tag{18}$$

Writing this out with the decision variables in transformed coordinates, the objective function then becomes

$$\begin{aligned}
-\sum_{i=0}^{N} p_i \log q_i &= -\sum_{i=0}^{N} p_i \log \mathcal{C} \exp([0, Q])_i && (19)\\
&= -\sum_{i=0}^{N} p_i \log \exp([0, Q])_i + \sum_{i=0}^{N} p_i \log \left(1 + \sum_{j=1}^{N} e^{\bar{q}_j}\right) && (20)\\
&= \log \left(1 + \sum_{j=1}^{N} e^{Q_j}\right) - \sum_{j=1}^{N} p_j Q_j && (21)\\
&= \log(1 + \mathbf{1} \cdot \exp Q) - p \cdot Q \ . && (22)
\end{aligned}$$

So, we can minimize this last expression over the unconstrained decision variables Q instead of the compositionally-constrained q. We can interpret the first term as a normalizer that punishes us for choosing values of Q that are large in absolute value. Notably, the transformed expression is still convex in Q Furthermore, using Jensen's inequality, we have

$$\log \left(1 + \sum_{j=1}^{N} e^{Q_j}\right) \geq \sum_{j=1}^{N} Q_j \ , \tag{23}$$

so we can minimize an upper bound by solving

$$\underset{Q}{\operatorname{argmin}} \sum_{i=1}^{N} (1 - p_i) Q_i \ . \tag{24}$$

Our main problem now takes the form

$$\operatorname{argmin} \sum_{x \in \mathcal{X}} D[p_x \| q_x] = \operatorname{argmin} \sum_{x \in \mathcal{X}} \left[\log \left(1 + \sum_{j=1}^{N} e^{Q_j(x)}\right) - \sum_{j=1}^{N} p_j Q_j(x) \right] \ . \tag{25}$$

In this formulation, $\bar{q}(x)$ need not satisfy any normalization conditions, so we can express the space of admissible models $Q(x)$ as

$$Q(x) = \sum_{k=1}^{K} \lambda_k(x) Q^k = \sum_{k=1}^{K} \lambda(x|\theta_k) Q^k \tag{26}$$

where Q^k is the kth 'representative distribution' under the model and $\lambda_k(x)$ is the spatial influence function corresponding to Q^k. However, Q_k need not be normalized. So, we have the unconstrained problem

$$\underset{q}{\operatorname{argmin}} \sum_{x\in\mathcal{X}} D[p_x \| q_x] = \underset{Q,\theta}{\operatorname{argmin}} \sum_{x\in\mathcal{X}} \left[\log\left(1 + \sum_{j=1}^{N} e^{\sum_{k=1}^{K} \lambda(x|\theta_k) Q_j^k}\right) - \sum_{i=j}^{N} p_j \sum_{k=1}^{K} \lambda(x|\theta_k) Q_j^k \right] \tag{27}$$

$$= \underset{Q,\theta}{\operatorname{argmin}} \sum_{x\in\mathcal{X}} \left[\log\left(1 + \mathbf{1} \cdot \exp\left\{\sum_{k=1}^{K} \lambda(x|\theta_k) Q^k\right\}\right) - p \cdot \sum_{k=1}^{K} \lambda(x|\theta_k) Q^k \right] \tag{28}$$

for $Q^k \in \mathbb{R}^N$ and θ_k in the valid parameter space of the influence function λ used.

In principle, it may be possible to work with this form explicitly and do e.g. a gradient-based method, or maybe simulated annealing if this problem turns out to be nonconvex in the variables, which I would assume it wouldn't be.

3 Helpful Citations