Logratio Approach

v0.0

Phil Chodrow

March 20th, 2016

1 Purpose

The purpose of this document is to fix notation and establish preliminary results for an approach to spatial compositional complexity measurement using logratio transforms.

2 Problem Setup and Notation

We consider the following problem. Suppose that we observe a vector of $J+1$-dimensional data $p^i = [p^i_0, \dots, p^i_J]$ at each of I points $x^i \in \mathbb{R}^M$. We assume that our observations are *compositional* data with $p^i > 0$ and $\sum_{j=0}^J p^i_j = 1$ at each x^i. We can then express our data as a matrix $P_{ij} = p^i_j$, where the ith row represents a point in $\mathbb{R}^M$ and the jth column represents a compositional category. An example of such data would be a Census data set of income categories for multiple tracts, in which case the ijth entry of P would be the proportion of respondents in the ith Census tract who reported that they fall in the jth category of income.

We would like to construct a model matrix Q of

1. Provides a good representation of P, where "good" is defined in information-theoretic terms and
2. Reflects the spatial structure of the data.

Like P, the rows of Q should satisfy $q^i > 0$ and $\sum_{j=0}^J q^i_j = 1$. The information-theoretic sense of "goodness" we will use is the Kullback-Leibler divergence. Define $f : \mathbb{R}^{I\times(J+1)} \times \mathbb{R}^{I\times(J+1)} \to \mathbb{R}$ by

$$f(P, Q) = \sum_{i=1}^I D[p^i \| q^i] = \sum_{i=1}^I \sum_{j=0}^J p^i_j \log \frac{p^i_j}{q^i_j} \,. \tag{1}$$

The (summed) divergence has an intuitive characterization: it is the amount of information (in bits) I would need to give you in addition to the model Q in order for you to reconstruct the true data P. For each i, Gibbs' inequality implies that $D[p^i \| q^i]$ attains its minimum of 0 if and only if $p^i = q^i$; thus, $f(P, Q) = 0$ iff $P = Q$.

To represent spatial structure, we stipulate that

$$q^i = \sum_{k=1}^{K} \lambda^k(x^i) Q^k, \tag{2}$$

where Q^k is one of $K < M$ "representative" compositions. The function $\lambda^k : \mathbb{R}^M \to \mathbb{R}$ reflects the "influence" of the representative composition $\bar{q}^k$ at x^i. Thus, Q can be fully determined by specifying K representative compositions Q^k and influence functions λ^k.

The compositional constraints $q^i > 0$ and $\sum_{j=0}^{J} q^i_j = 1$ may seem to require some rather stringent assumptions on Q^k and λ^k. We can circumvent this technical issue using the standard toolbox of compositional data analysis.

3 Compositional Tools

The logratio transformation is a standard map that transforms compositional data into general data. The map $L : \mathbb{R}^{J+1} \to \mathbb{R}^J$ is defined by

$$\bar{p} = L(p) = \left[\log \frac{p_1}{p_0}, \ldots, \log \frac{p_J}{p_0} \right] . \tag{3}$$

Its inverse is

$$p = L^{-1}(\bar{p}) = \mathcal{C} \exp \{[0, \bar{p}]\} , \tag{4}$$

where the exponential operator is applied componentwise and $\mathcal{C}$ is the closure operator

$$\mathcal{C}y = \frac{y}{\mathbf{1} \cdot y} . \tag{5}$$

We would like to use this transformation to simplify the objective function (1). To do so, we first note that

$$D[p^i \| q^i] = \sum_{j=0}^{J} p^i_j \log \frac{p^i_j}{q^i_j} \tag{6}$$

$$= \sum_{j=0}^{J} p^i_j \log p^i_j - \sum_{j=0}^{J} p^i_j \log q^i_j \tag{7}$$

$$= H[p^i] - \sum_{j=0}^{J} p^i_j \log q^i_j . \tag{8}$$

The first term is the entropy of the ith row of P and depends only on the data. We can therefore focus on the second term when minimizing. Using the logratio transformation,

we can write this term as

$$-\sum_{j=0}^{J} p_j^i \log q_j^i = -\sum_{j=0}^{J} p_j^i \log L^{-1}(\bar{q}^i)_j \tag{9}$$

$$= -\sum_{j=0}^{J} p_j^i \log \mathcal{C} \exp\{[0, \bar{q}^i]\}_j \tag{10}$$

$$= -\sum_{j=0}^{J} p_j^i \log \exp\{[0, \bar{q}^i]\}_j + \sum_{j=0}^{J} p_j^i \log \left(1 + \sum_{j=1}^{J} e^{\bar{q}_j^i}\right) \tag{11}$$

$$= -\sum_{j=0}^{J} p_j^i \log \exp\{[0, \bar{q}^i]\}_j + \log \left(1 + \sum_{j=1}^{J} e^{\bar{q}_j^i}\right) \tag{12}$$

$$= -\sum_{j=1}^{J} p_j^i \bar{q}_j^i + \log \left(1 + \sum_{j=1}^{J} e^{\bar{q}_j^i}\right) \ . \tag{13}$$

We should note three points about the transformed objective (13). First, (13) is unconstrained in the decision variables $\bar{q}^i$; other than being strictly positive, no compositional constraint is enforced. This allows us to circumvent the possible computational difficulties associated with working with (1) and (2) directly. In particular, we can seek representative compositions $\bar{Q}^k$ and functions $\bar{\lambda}^k$ such that the assignment

$$\bar{q}^i = \sum_{k=1}^{K} \bar{\lambda}_k(x^i) \bar{Q}^k \tag{14}$$

minimizes (13). Second, the two terms of (13) are nicely interpretable: the first term rewards transformed decision variables $\bar{q}^i$ that are parallel to the last J coordinates of p^i, while the second term punishes us for choosing transformed decision variables that are too large in magnitude. Third and finally, Jensen's inequality allows us to bound (13) from above as

$$-\sum_{j=1}^{J} p_j^i \bar{q}_j^i + \log \left(1 + \sum_{j=1}^{J} e^{\bar{q}_j^i}\right) \geq -\sum_{j=1}^{J} p_j^i \bar{q}_j^i + \sum_{j=1}^{J} \bar{q}_j^i \tag{15}$$

$$= \sum_{j=1}^{J} (1 - p_j^i) \bar{q}_j^i \ , \tag{16}$$

which is linear in the decision variables $\bar{q}_j^i$. If we find that the full problem (13) is not tractable, it may be useful to consider (16) as a computational alternative. Our final simplification is to assume that the influence functions $\bar{\lambda}_k$ share a common functional form, expressed as $\bar{\lambda}_k(x^i) = \bar{\lambda}(x^i|\theta^k)$, where the parameters θ^k distinguish each influence

function. Then, (14) becomes

$$\bar{q}^i = \sum_{k=1}^{K} \bar{\lambda}(x^i|\theta^k)\bar{Q}^k \tag{17}$$

In practice, we'll be most interested in the case in which $\bar{\lambda}$ is a normal density, in which case the parameters θ^k will include the mean and covariance.

4 Computing Gradients

Since everything in sight is differentiable, we can consider elementary first-order methods. To do this requires the computation of derivatives. Let $g_i : \mathbb{R}^J \to \mathbb{R}$ be given by

$$g_i(\bar{q}^i) = -\sum_{j=1}^{J} p_j^i \bar{q}_j^i + \log\left(1 + \sum_{j=1}^{J} e^{\bar{q}_j^i}\right) , \tag{18}$$

and $h_i : \Theta \times \mathbb{R}^{K\times J} \to \mathbb{R}^J$ be given by

$$h_i(\theta, \bar{Q}) = \sum_{k=1}^{K} \lambda(x^i|\theta^k)\bar{Q}^k . \tag{19}$$

Then, the ith term of the objective function (13) can be written as $(g_i \circ h_i)(\theta, \bar{Q})$. The derivatives we need are now

$$Dg_i(\bar{q}) = -p^i + \frac{\exp \bar{q}^i}{1 + \mathbf{1} \cdot \exp \bar{q}^i} , \tag{20}$$

which is a $1 \times J$ row vector, and

$$Dh_i(\theta, \bar{Q}) = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \sum_{k=1}^{K} Q_j^k D\lambda(x^i|\theta^k) & \lambda(x^i|\theta^1) & \cdots & \lambda(x^i|\theta^K) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{21}$$

which is a $J \times (L + K)$ matrix. The gradient of the objective function can now be computed by the chain rule, which in this case takes the form

$$\nabla f(\theta, \bar{Q}) = \sum_{i=1}^{I} Dg_i(h(\theta, \bar{Q})) \cdot Dh_i(\theta, \bar{Q}) \tag{22}$$

The primary remaining task would be an implementation of a first-order method based on this gradient, such as classical gradient descent. Newton's method would be a possibility if we found the performance of classical gradient descent to be lacking. A globalization such as stochastic gradient descent or simulated annealing might be desirable.