SPATIAL STATISTICS AND RECOMMENDATION SYSTEMS

Overview

Broadly speaking, statistics is the study of *variance*. Spatial statistics, therefore, is the study of variance across space (which could be interpreted as physical space or some virtual space). The key question in spatial statistics is whether a set of observations of a random variable (or sets of variables) are spatially *autocorrelated*, that is whether adjacent values tend to be similar, and different from distal values. Autocorrelation violates the typical IID assumption in statistics, whereby samples collected are drawn independently and are identically distributed. Autocorrelated random variables tend to generate samples that are spatially similar.

For recommendation systems, spatial statistics provides a useful framework since it enables understanding the extent to which user biases are geographically influenced. In this project, we are trying to understand the extent to which movie recommendation systems can exploit knowledge of local variability among users.

In this short document, I explain some key terms from spatial statistics, and show results from the MovieLens dataset that Stefan has been analyzing in the past few weeks. The results are preliminary, but seem to suggest interesting spatial patterns that may play a useful role in our subsequent work.

Similarity

To define metrics for spatial autocorrelation, we need to define some notion of similarity. In the experiments I carried out, I am going to measure similarity using a simple *heat kernel*, which is widely used in machine learning. Given two observations x_i and x_j, the distance between them is defined as

$$w_{ij} = e^{\frac{-\|x_i - x_j\|^2}{\sigma}}$$

The distance between two users is simply measured in the numerator as the L2 norm between their locations as measured by latitude and longitude.

Spatial Autocorrelation

There are a number of ways of defining spatial autocorrelation. This includes the *Geary C index*, the *Moran I index*, and other measures, such as the *variogram* and the *semi-variogram*.

A. Moran's I

Moran's I (Moran 1950) tests for global spatial autocorrelation for continuous data. It is based on cross-products of the deviations from the mean and is calculated for n observations on a variable x at locations i, j as:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2},$$

where $\bar{x}$ is the mean of x, w_{ij} are the elements of the weight matrix, and S_0 is the sum of the elements of the weight matrix: $S_0 = \sum_i \sum_j w_{ij}$.

Moran's I has properties similar to correlation. It varies from -1 to +1. In the absence of autocorrelation and independent of the weight matrix, the expectation of Moran's I statistic is $-1/(n-1)$, which converges to zero for large sample sizes. For a row-standardized spatial weight matrix, the normalizing factor S_0 equals n (since each row sums to 1), and the statistic simplifies to a ratio of a spatial cross product to a variance. A Moran's I coefficient larger than $-1/(n-1)$ indicates positive spatial autocorrelation, and a Moran's I less than $-1/(n-1)$ indicates negative spatial autocorrelation. The variance is:

$$\begin{aligned} \mathrm{Var}(I) &= \frac{n\{n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2\} - k\{n(n-1)S_1 - 2nS_2 + 6S_0^2\}}{(n-1)(n-2)(n-3)S_0^2} \\ &\quad - \frac{1}{(n-1)^2} \end{aligned}$$

where

$S_1 = \frac{1}{2}\sum\sum_{i \neq j}(W_{ij} + W_{ij})^2 = 2S_0$ for symmetric W containing 0's and 1's.

$S_2 = \sum_i (W_{i0} + W_{0i})^2$ where $W_{i0} = \sum_j W_{ij}$ and $W_{0i} = \sum_j W_{ji}$

B. Geary's C

Geary's C statistic (Geary 1954) is based on the deviations in responses of each observation with one another:

$$C = \frac{n-1}{2S_0} \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2}.$$

Geary's C ranges from 0 (maximal positive autocorrelation) to a positive value for high negative autocorrelation. Its expectation is 1 in the absence of autocorrelation and regardless of the specified weight matrix (Sokal & Oden 1978). If the value of Geary's C is less than 1, it indicates positive spatial autocorrelation. The variance is:

$$\begin{aligned}\mathrm{Var}(c) &= \frac{1}{n(n-2)(n-3)S_0^2} \\ &\{S_0^2[(n^2-3)-k(n-1)^2]+S_1(n-1)[n^2-3n+3-k(n-1)] \\ &+\frac{1}{4}S_2(n-1)[k(n^2-n+2)-(n^2+3n-6)]\}\end{aligned}$$

where S_0, S_1, and S_2 are the same as in Moran's I.

RESULTS

I used Stefan's data, which contains 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. This dataset contains 6040 users who rated 3952 movies. After filtering out users from outside the US and some with spurious zip codes, Stefan ended up with 5899 users. He implemented a gradient-based algorithm for constructing 10 latent factors. I used these, as well as MATLAB's non-negative matrix factorization (NNMF) algorithm to construct 5 latent factors.

In the results I show below, I am plotting Geary's C Index on user groups of various sizes. Generally, the strategy I used was to select randomly 50 among the 5899 users, find k of the nearest neighbors, and then compute the spatial autocorrelation of one of the latent factors. What is interesting is to look at both spatial autocorrelation of individual groups of users, as well as averages across a large sample of users for a fixed population size.

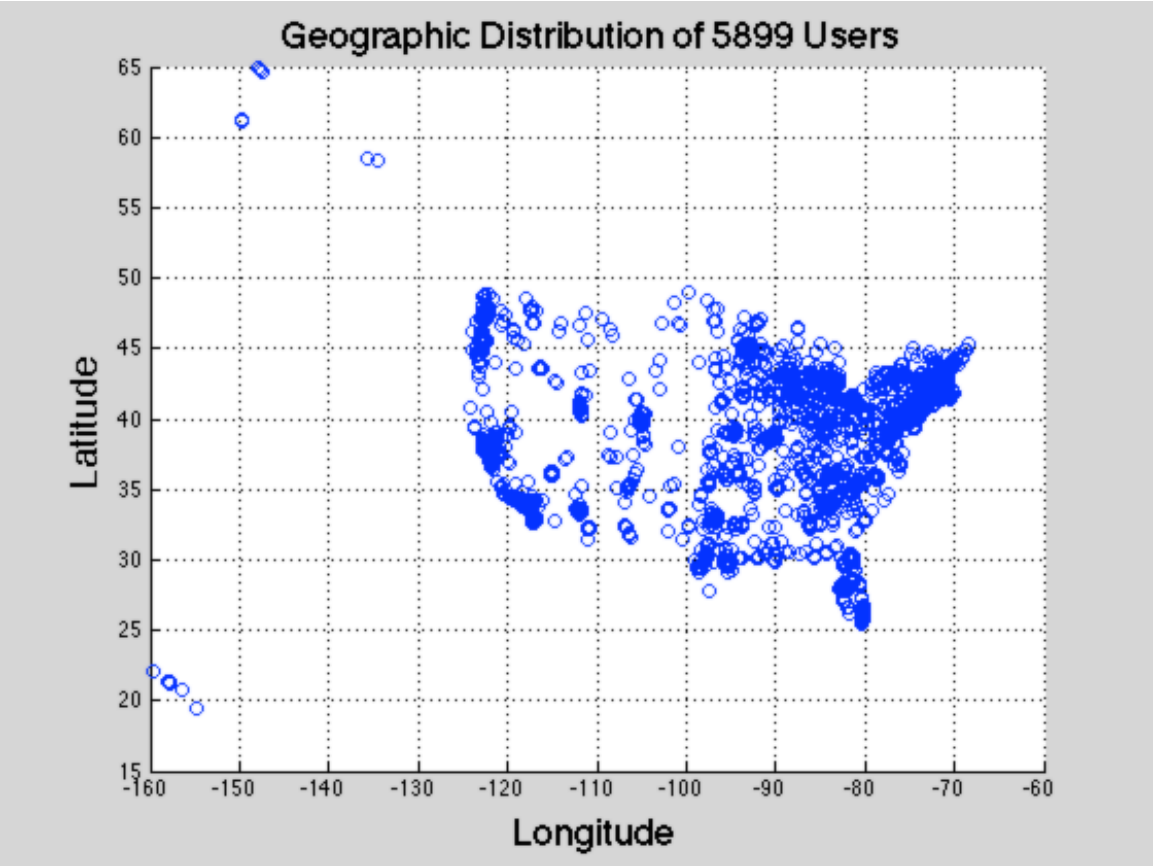
To interpret the plots, it is necessary to understand the range of values that the spatial autocorrelation metrics can take. Recall that if the value of Geary's C is less than 1, it indicates positive spatial autocorrelation. A value of 1 indicates no spatial autocorrelation.

LATENT FACTORS USING NON-NEGATIVE MATRIX FACTORIZATION

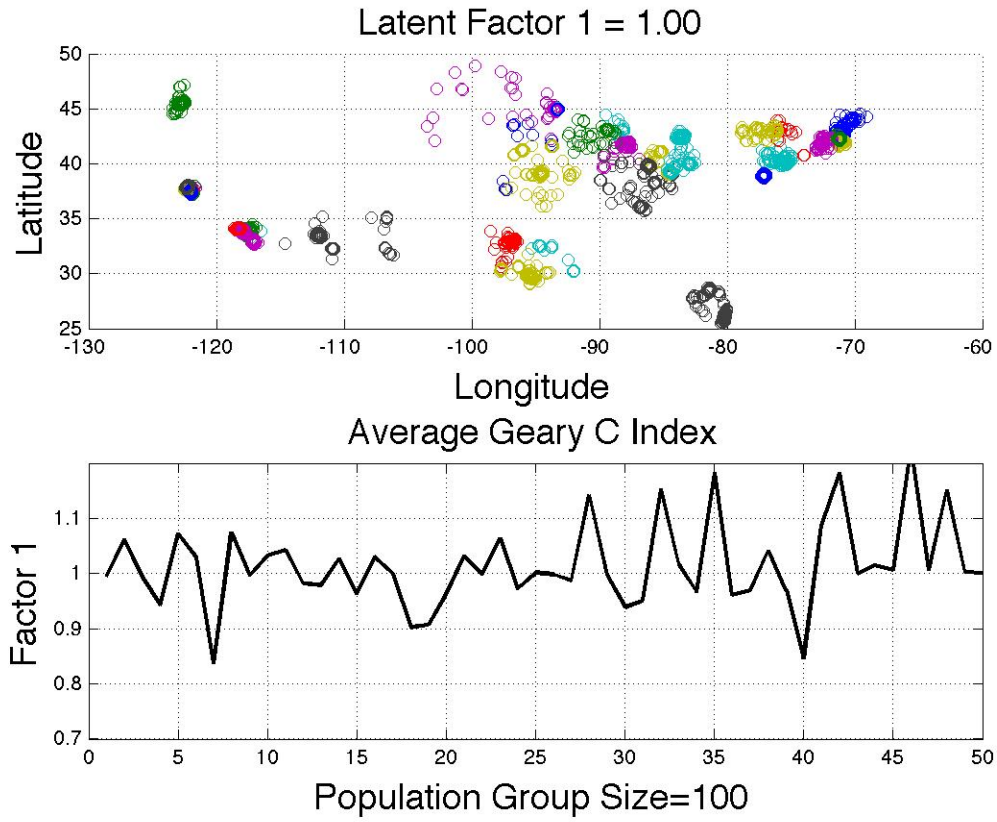
I generated 5 latent factors by factorizing the user ratings matrix R of size 5899x3952 into two matrices, W of size 5899 x 5, and H of size 5 x 3952. The columns of the W matrix can now be interpreted as latent factors indicating user biases in movie preferences. We can now compute the spatial autocorrelation of the NNMF-produced latent variables.

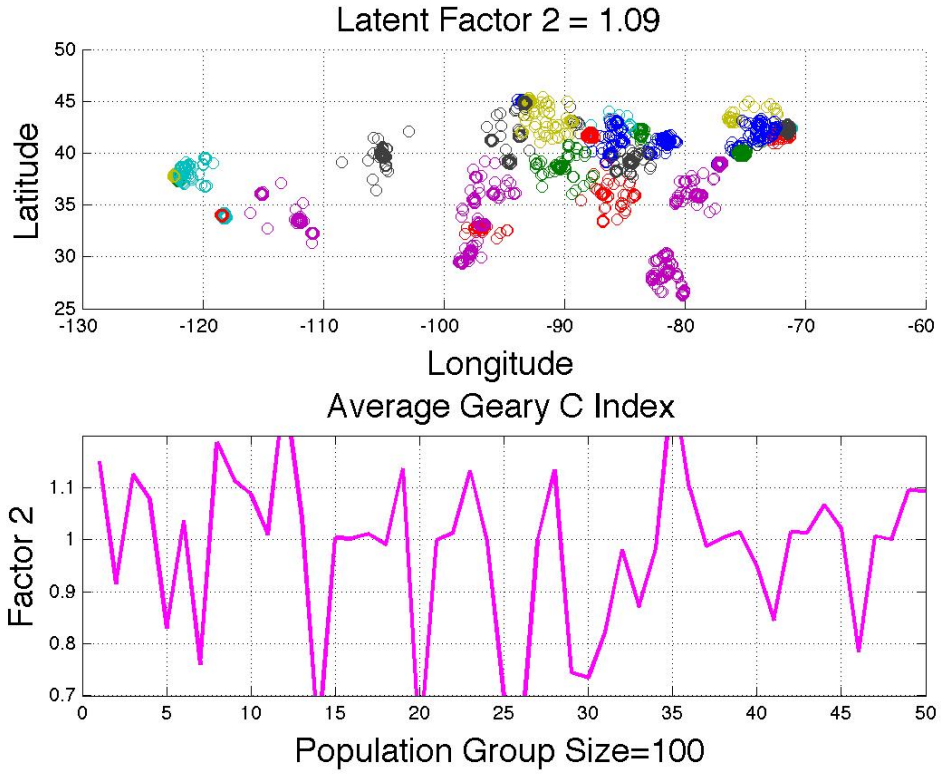
USER DISTRIBUTION

The plot below shows the geographical distribution of all the 5899 users in the MovieLens 1 million dataset. The map of the US is clearly visible.

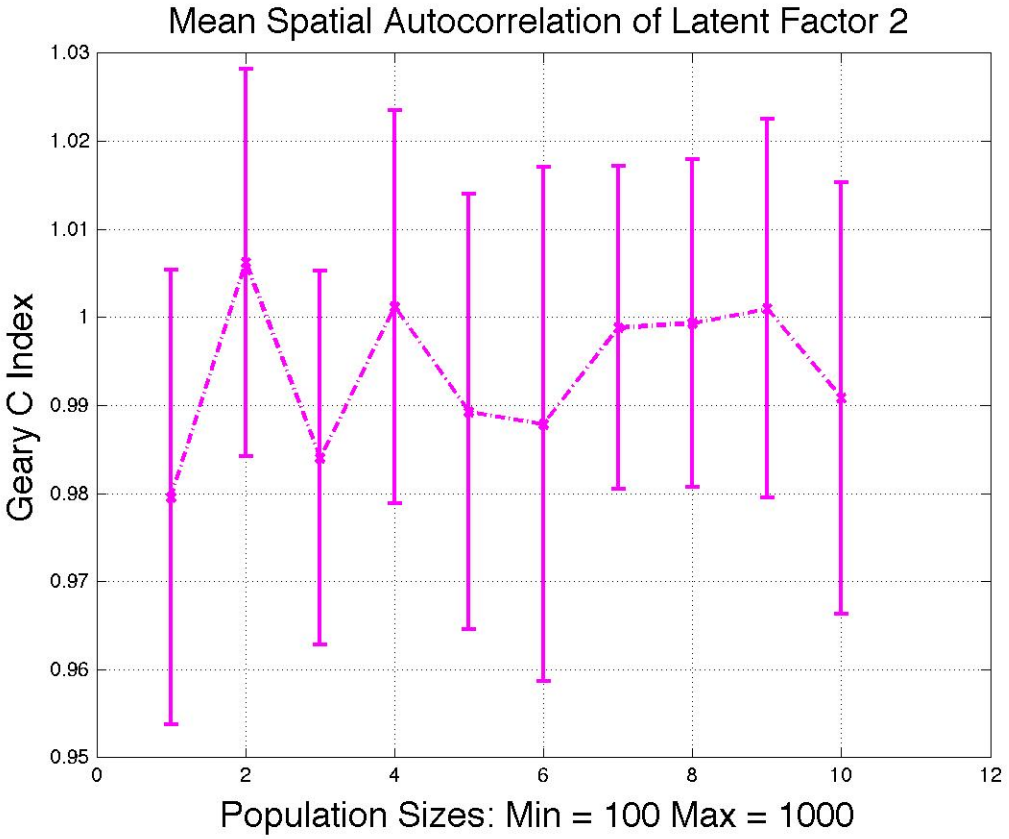


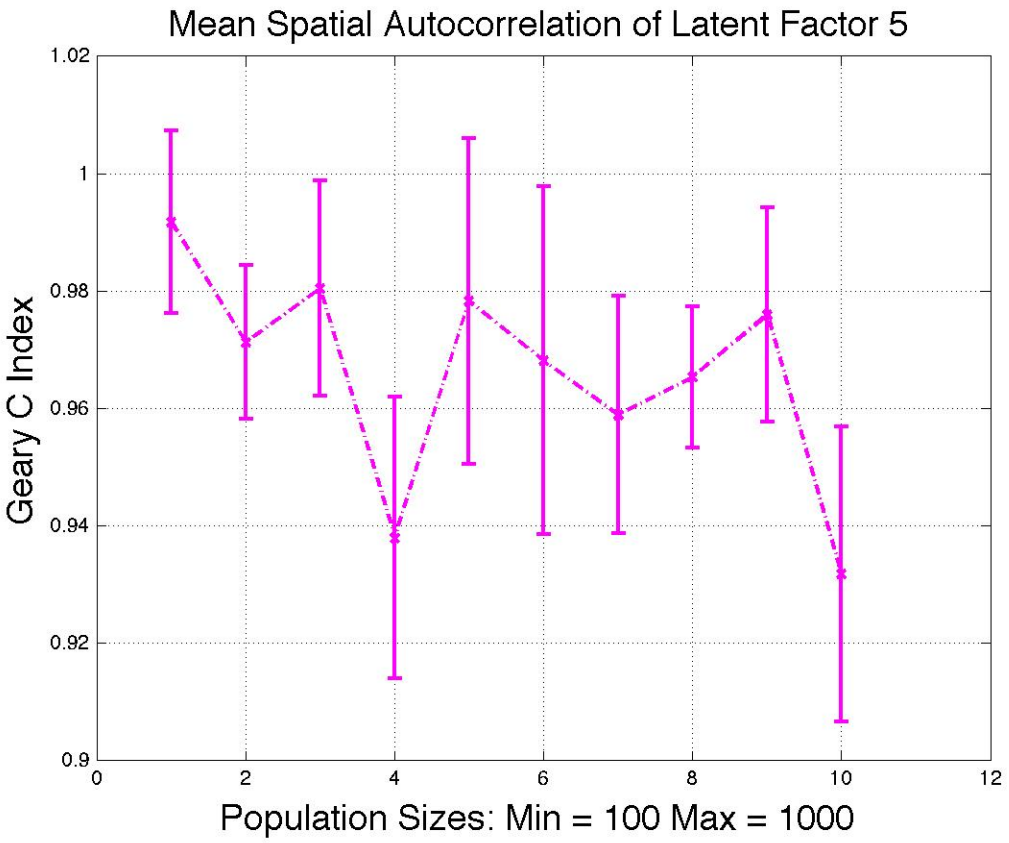
Let us contrast the spatial autocorrelation of two latent factors, factors 1 and 2 (computed using non-negative matrix factorization) on populations of size 100. It is clear from the plots below that factor 2 has higher spatial autocorrelation than factor 1. What this means is that for the samples selected, the users tend to have similar user preferences as measured by factor 2 than on factor 1.


Latent Factor 1 = 1.00
Latitude
Longitude
Average Geary C Index
Factor 1
Population Group Size=100


Latent Factor 2 = 1.09
Latitude
Longitude
Average Geary C Index
Factor 2
Population Group Size=100

We can also plot the average spatial autocorrelation across population of user groups of various sizes. These can indicate roughly the variation of spatial autocorrelation across spatial scales. A key question is whether users get uniformly more like each other as averaged at a fine scale or at a coarser scale. Once again, comparative results are more interesting than individual results. Comparing factors 2 and 5, shown in the plots below for population sizes of a 100 to a 1000, we see that the average spatial autocorrelation of latent factor 2 is less than that for latent factor 5. Also, latent factor 2 has higher variance than factor 5.





Hopefully, these results should give some indication of the usefulness of spatial statistics in designing recommendation systems that are geographically customized to groups of users. Much more remains to be done in exploring these connections, and we should discuss ideas for further study in our meetings.

REFERENCES

1. Geary, R. (1954). The contiguity ratio and statistical mapping. The Incorporated Statistician 5: pp115-45

2. Moran, P.A.P. (1950). Notes on continuous stochastic phenomena, Biometrika 37, pp17-23.

3. Ripley, B, Spatial Statistics, John Wiley, 2004.

4. Sokal, R.R. and Oden, N.L. (1978). Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, 10: 199-228.