MULTIVARIATE SPATIAL STATISTICS AND RECOMMENDATION SYSTEMS

Overview

Last time we explored univariate spatial autocorrelation, whereby individual latent factors were spatially analyzed using the Geary C index. In this meeting, I am going to show how to generalize spatial analysis to multivariate variables, and instead using a modified Moran's I index proposed by Wartenberg.

A. Moran's I

Recall from last week's notes that Moran's I (Moran 1950) tests for global spatial autocorrelation for continuous data 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:

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

where $\overline{x}$ 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

$$\begin{aligned} S_1 &= \frac{1}{2}\sum_{i \neq j}\sum (W_{ij} + W_{ij})^2 = 2S_0 \text{ for symmetric } W \text{ containing 0's and 1's.} \\ S_2 &= \sum_i (W_{i0} + W_{0i})^2 \text{ where } W_{i0} = \sum_j W_{ij} \text{ and } W_{0i} = \sum_j W_{ji} \end{aligned}$$

Defining the normalized variables $z_i = x_i - \mu_I$, and also normalizing the weights such that

$$S_0 = \sum_i \sum_j w_{ij} = 1$$

we can write Moran's I index simply as

$$I = \sum_{(2)} w_{ij}^* z_i'^* z_j',$$

where

$$z_i' = \frac{(x_i - \bar{x})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}}$$

We can now easily generalize Moran's I index in the multivariate case to simply be

$$\mathbf{M} = \mathbf{Z}^t\mathbf{W}\mathbf{Z},$$

where

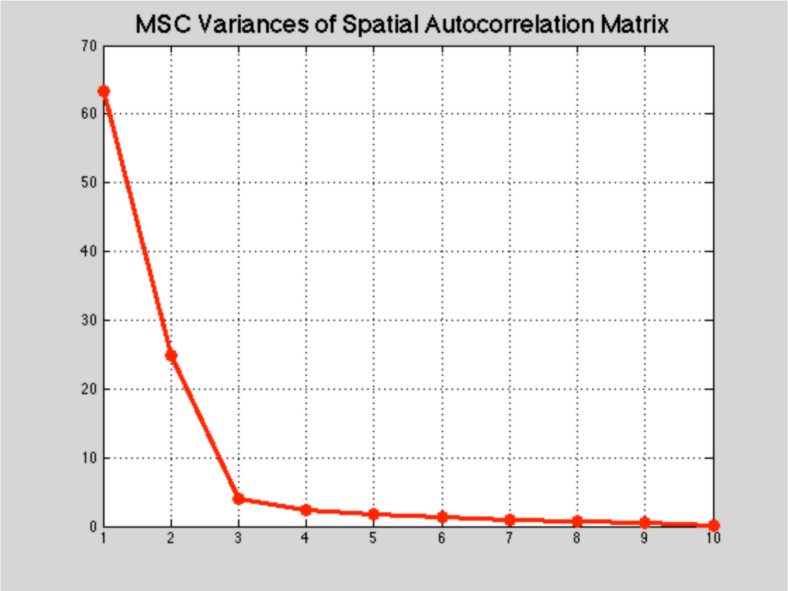
M is an m by m, variable by variable, spatial correlation matrix
Z is an n by m, location by variable, standardized and centered (by variable) data matrix
$\mathbf{Z}^t$ is an m by n, variable by location, standardized and centered (by variable) data matrix, the transpose of **Z**
W is an n by n, locality by locality, weight matrix.

Note that the spatial correlation matrix M is symmetric quadratic form, and hence by the spectral theorem in linear algebra, we can decompose it into

$$M = V \Lambda V^T$$

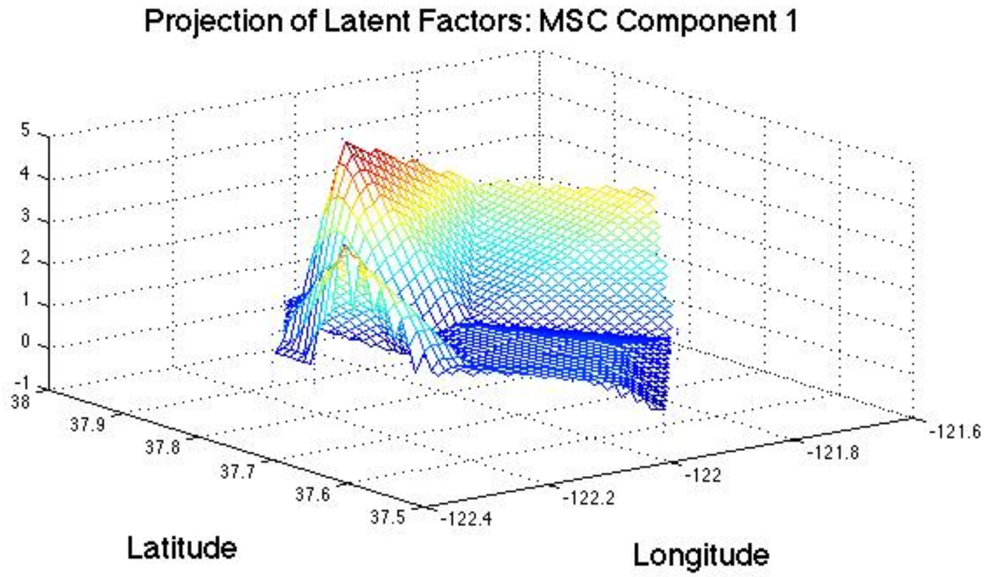
Note that unlike eigendecomposition of a covariance matrix, there is no guarantee that M is positive definite. Consequently, M may have positive or negative real eigenvalues. Nonetheless, we can still try to attach meaning to these in terms of explaining the variance in the data, analogous to PCA, by comparing the magnitude of the largest eigenvalues to the remaining ones.

In what follows, I show examples of the spectral decomposition of spatial autocorrelation matrix M, and refer to this as MSC (multivariate spatial correlation) analysis. I am also going to analyze the user recommendation matrix after subtracting out the most popular 500 movies. The procedure is as follows. In our case, Z is a n x 10 matrix, where n is around 6000 users. Z is a normalized (mean 0) user latent factor matrix. W remains the same exponentially weighted similarity matrix, except that W is normalized such that it sums to 1. Thus, M will be a 10x10 matrix, which yields 10 eigenvalues and eigenvectors. The eigenvector associated with the largest eigenvalue corresponds to the linear combination of the latent factors that "explains" the maximum spatial variance in the data. We can now project the 10 user latent factors onto the first eigenvector to get a 1-dimensional signal across all users. We can do this for the remaining eigenvectors as well. I am going to show this spatial variance plot for just 2 eigenvectors for simplicity. Here then are the example plots following this line of analysis.

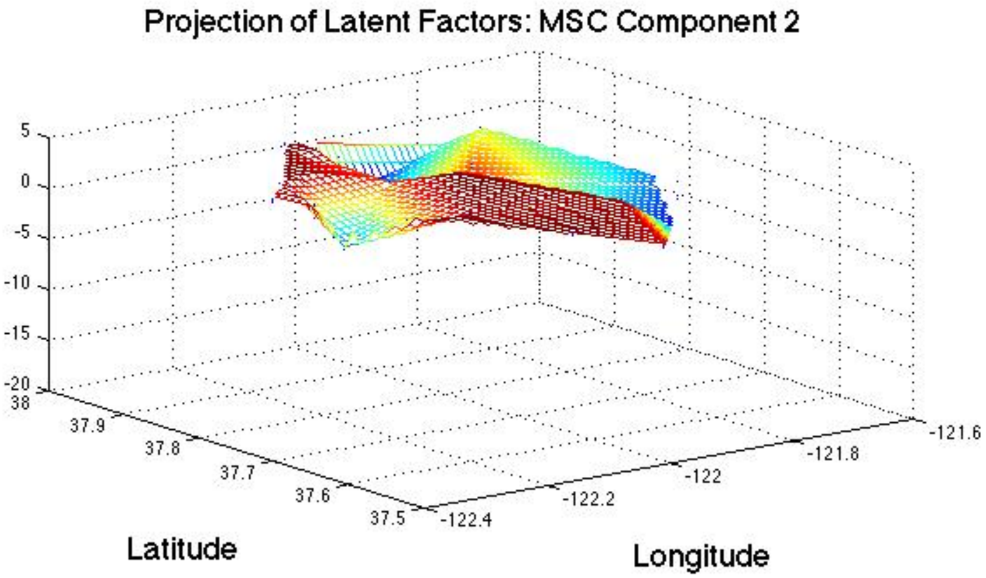


Shown above is an example of the spectral decomposition of the M matrix. Notice that the largest eigenvalue corresponds to about 65 of the spatial variance. The next eigenvalue corresponds to about 25 percent. Consequently, the first two eigenvalues represent about 85 percent of the variance in the data. We can take the eigenvectors corresponding to these two eigenvalues, and project the user latent factors onto them. This gives us projections of all the latent factors and reveals low-dimensional spatial smoothness in user latent factors (i.e., similarities across users). On the next page is an example of what these look like for a particular sample of users.

In doing the plot on the following page, what I did is compute the projection of all 10 latent factors onto the two largest eigenvectors, and then plot the corresponding 1-dimensional projections onto the latitude/longitude locations of the users. I then used MATLAB's spatial interpolation routines (meshgrid and griddata) to generate a smooth plot across geographical space. We can then look at the "smoothness" of the user latent factors across space to judge how similar users are. Here are two groups of 50 users for whom I am plotting the similarities in user latent factors.

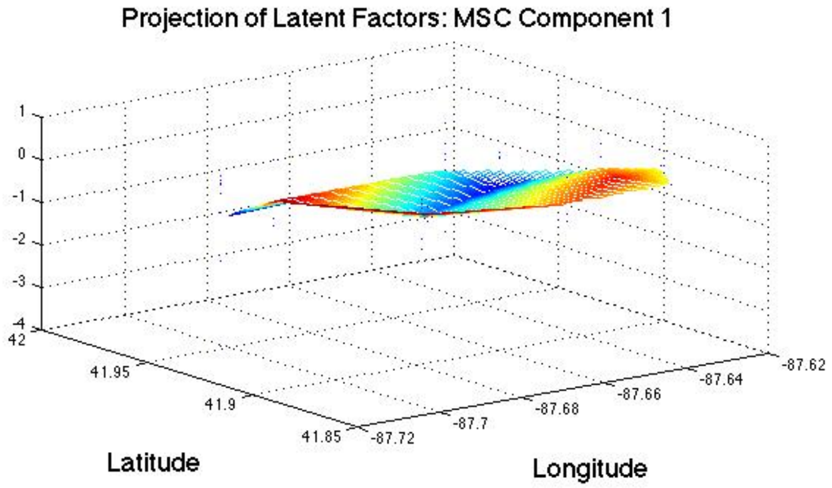


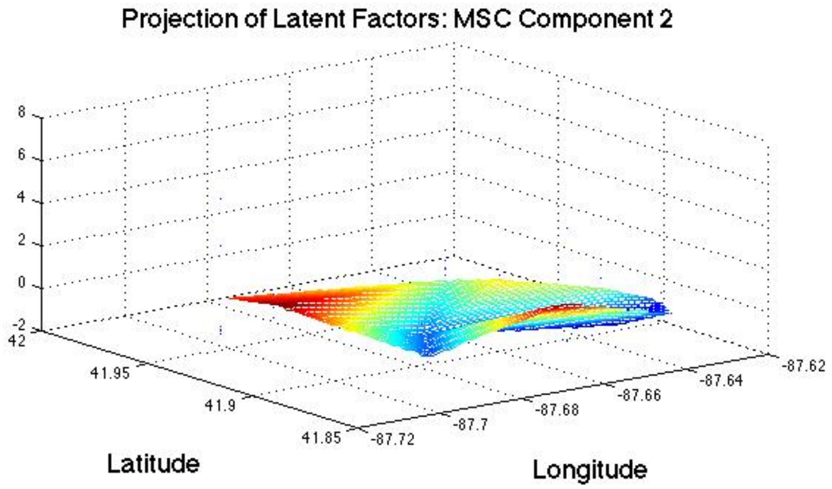
Projection of Latent Factors: MSC Component 1



Projection of Latent Factors: MSC Component 2

Users: 160 192 222 456 1157 1205 1220 1221 1257
1369 1400 1407 1454 1486 1594 1614 1618 1622
1693 1722 1839 1841 1850 1886 1897 1909 2515
2526 2610 2767 3098 3193 3272 3350 3389 3516
3693 3699 3713 4383 4493 4718 4859 4884 5254
5381 5405 5550 5779 5784


Projection of Latent Factors: MSC Component 1


Projection of Latent Factors: MSC Component 2

43 815 1298 1340 1459 1599 2041 2109 2110

2162 2165 2189 2193 2467 2585 2665 2679 2724

2745 2749 2772 2863 3136 3156 3761 3826 3880

3895 4233 4448 4471 4605 4677 4934 5094 5117

5135 5157 5158 5271 5310 5379 5553 5587 5607

5651 5702 5759 5798 5844