Moran's I: a statistic in search of a model

Renato Assunção

Departmento de Estatística

Universidade Federal de Minas Gerais

assuncao@est.ufmg.br

1 Introduction

Ver Moran's spatial autocorrelation in www.scholar.google.com

Moran's I is the most popular statistic to test for the presence of spatial autocorrelation and to evaluate its strength in maps partitioned in geographical areas (REFS ????). Consider a region divided in n areas and let y_i be a random variable measured in area i, with $i = 1, \ldots, n$. Moran's I is given by

$$I = \frac{n}{\sum_{ij} w_{ij}} \frac{\sum_{ij} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_i (y_i - \overline{y})^2} \tag{1}$$

where the value w_{ij} is the weight assigned to areas i and j, and $\overline{y} = \sum_i y_i/m$. Usually, w_{ij} will reflect the geographical distance between areas i and j, being defined, for example, as $w_{ij} = 1$ if the areas are adjacent and $i \neq j$, and by $w_{ij} = 0$, otherwise. However, weights can be more general depending, for example, on functions of distances between the areas (e.g., ???). Moran's I usually ranges between -1 and 1 (be more precise???), with large positive values indicating neighborhood similarity of the rates and values close to zero indicating absence of spatial autocorrelation.

Its distribution under the null hypothesis (which ONE???) is well studied....

However, despite its popularity, Moran's I lacks a more ??? well defined model ??? there is not a model for ???? there is no parameter AND LIKELIHOOD model associated with this statistic. Why to have one is good?

IN this paper, ...

2 Moran's I as an estimator of a parameter

2.1 The SAR model

A well known spatial model is the simultaneous autoregression (SAR) model proposed by (REFS???). Consider a finite grid with n sites or areas indexed by $i = 1, \ldots, n$ and with associated random variables y_i. Then, the SAR model is a set of n simultaneous equations for the y_i random variables:

$$y_i = \mu_i + \rho \sum_{j \neq i} w_{ij}(y_j - \mu_j) + \epsilon_i \,, \tag{2}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent normal random variables with mean zero and variance λ_i.

The weights w_{ij} reflect the spatial structure of the sites such that a pair of sites close to each other has more weight than a pair of sites farther apart. One of the most common choices for w_{ij}, and the one we follow here, is to let $\boldsymbol{W}$ be a $n \times n$ binary matrix with elements $w_{ij} = 1$ if i and j share boundaries in a map and $w_{ij} = 0$, otherwise. We define also $w_{ii} = 0$ for all i. If i and j are neighbors, we denote it by $i \sim j$.

If $\boldsymbol{I} - \rho\, \boldsymbol{W}$ is invertible then

$$\boldsymbol{y} \sim N_n \left(\boldsymbol{\mu}, (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \, \boldsymbol{\Lambda} \, (\boldsymbol{I} - \boldsymbol{W}^t)^{-1} \right) \tag{3}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with element ii equal to λ_i and $\boldsymbol{\mu}$ is the n-dim vector with the means μ_i.

2.2 The pairwise interaction model

The set of n simultaneous stochastic equations (2) relate each area i at the same time with all its neighbors. Rather than that, consider a set of $\sum_{ij} w_i j = \mathbf{1}^t \boldsymbol{W} \mathbf{1}$ simultaneous equations, each one connecting a observation y_i with one of its neighboring

areas:

$$y_i = \mu_i + \rho(y_j - \mu_j) + \epsilon_{ij} \tag{4}$$

if $w_{ij} = 1$.

The set of random variables ϵ_{ij} can not be independent. To see this, imagine that $\rho \approx 1$ and $\mu_i = 0$ for all i. If $i \sim j$, $j \sim k$, and $k \sim i$, and if we know the value of

$$\epsilon_{ij} = y_i - \rho y_j \approx y_i - y_j$$

and also of

$$\epsilon_{jk} = y_j - \rho y_k \approx y_j - y_k$$

then we can make a pretty good guess about the value of

$$\epsilon_{ik} = y_i - \rho y_k \approx y_i - y_j - (y_j - y_k) = \epsilon_{ij} - \epsilon_{jk}$$

However, the set of pairwise equations (4) suggests the proposal of a probability distribution for the vector $\boldsymbol{y}$ of the form

$$\begin{aligned} f(\boldsymbol{y}) &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i\sim j} \epsilon_{ij}^2 \right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i\sim j} (y_i - \mu_i - \rho(y_j - \mu_j))^2 \right) \qquad (5) \\ & \qquad (6) \end{aligned}$$

We assume for now that $\mu_i = 0$ for all i. To find what the probability distribution $f(\boldsymbol{y})$ is, we write

$$\begin{aligned} \sum_{i\sim j} (y_i - \rho y_j)^2 &= (1+\rho^2) \sum_{i\sim j} y_i^2 - 2\rho \sum_{i\sim j} y_i y_j \\ &= (1+\rho^2) \left[\sum_i n_i y_i^2 - \frac{2\rho}{1+\rho^2} \sum_{i\sim j} y_i y_j \right] \\ &= (1+\rho^2) \left[\sum_{i,j} A_{ij} y_i y_j \right] \\ &= (1+\rho^2)\, \boldsymbol{y}^t \boldsymbol{Q} \boldsymbol{y} \end{aligned}$$

where

$$\boldsymbol{Q} = \boldsymbol{In} - \frac{2\rho}{1+\rho^2}\, \boldsymbol{W}\,.$$

The vector $\boldsymbol{n}$ is a n-dim vector with i-th element equal to the number n_i of neighbors of area i.

Therefore, if $\boldsymbol{Q}$ is invertible, (5) must be a normal distribution density with precision matrix $(1+\rho^2)\,\boldsymbol{Q}$. For all $\rho \in (-1, 1)$, this is true since $\boldsymbol{Q}$ is diagonally dominant in this case. In fact, we have

$$\boldsymbol{Q}_{ii} = n_i > n_i \, \frac{2\,\rho}{1+\rho^2} = \sum_j w_{ij} \, \frac{2\,\rho}{1+\rho^2} = \sum_{j \neq i} \boldsymbol{Q}_{ij}$$

2.3 Conditional independence and ρ

Since $\boldsymbol{Q}$ is a sparse matrix, properties of conditional independence follows immediately. That is, off-diagonal zeros of the matrix W identifies the pairs of areas that are conditionally independent given the values of all the other areas. This is in sharp contrast with the SAR model which has a more complicated conditional distribution structure. To compare the two models, let $\lambda_i = \sigma^2$ for all i. Then, precision matrix is the inverse of the covariance matrix of (3) and it is given by:

2.4 Properties of the distribution

MRF, find MLE, find moments, find MQ, find score. Find the determinant.

2.5 Cor(y_i, y_j) and ρ

Depends on the graph structure? Is there a maximum over the pairs of neighbors? A minimum? An average?

2.6 Properties of I as an estimator

Is it EMV? Is it unbiased? Consistent? Asymptotically normal?

References

Fahrmeir, L., and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York.

Thomas, L.C., Edelman, D.B., Crook, J.N. (2002) *Credit Scoring and Its Applications*. SIAM Monographs on Mathematical Modeling and Computation. Philadelphia.