Constructing Hypothesis Testing

- Our primary goal is to construct confidence intervals or hypothesis testing for the codispersion coefficient.

- This implies to find the distribution of the sample version of the codispersion coefficient.

- The conditions for which the consistency and the asymptotic normality hold are not trivial and restrictive.

- The main drawback is that the asymptotic variance is not easy to compute in practice, specially when the model has a large number of parameters.

9. Properties

- Three broad categories of spatial models that have been studied are simultaneous autoregressive (SAR) models (Whittle, 1954), conditional autoregressive (CAR) models (Besag, 1974), and moving average (MA) models (Haining, 1978).
- The simultaneous AR model is defined as

$$\Phi(B_1, B_2)X(i,j) = \epsilon(i,j),$$

where

$$\Phi(B_1, B_2) = \sum_k \sum_l \phi(k,l) B_1^k B_2^l,$$

with $B_1 X(i,j) = X(i-1,j)$, $B_2 X(i,j) = X(i,j-1)$ and $\epsilon(i,j)$ are independent random variables with $\mathbb{E}[\epsilon(i,j)] = 0$ and var$[\epsilon(i,j)] = \sigma^2$.

Properties

- Tjostheim (1978) examined a unilateral AR model that the value at the site (i, j) is a finite autoregression on the values at the sites that are in the lower quadrant of (i, j).
- $$X(i,j) = \sum_{k=0}^{p_1} \sum_{l=0}^{p_2} \phi(k,l) X(i-k, j-l) + \epsilon(i,j),$$

  with $\phi(0,0) = 0$.
- This model leads to the Wold-type representation

$$X(i,j) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \psi(k,l) \epsilon(i-k, j-l).$$

Properties

- Basu and Reinsel (1993) investigated the correlation structure of a general first-order autoregressive process of the form

$$X(i,j) = \phi(1,0)X(i-1,j) + \phi(0,1)X(i,j-1) + \phi(1,1)X(i-1,j-1) + \epsilon(i,j).$$

- They showed that the conditions

$$|\phi(k,l)| < 1,$$
$$1 - \phi(0,1)^2 > |\phi(1,0) + \phi(0,1)\phi(1,1)|,$$
$$(1 + \phi(1,0)^2 - \phi(0,1)^2 - \phi(1,1))^2 > 4(\phi(1,0) + \phi(0,1)\phi(1,1))^2,$$

guarantee the stationarity of the process.

- There exist a representation of the form

$$X(i,j) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{r=0}^{\infty} \frac{(k+l+r)!}{k!l!r!}\phi(1,0)^k\phi(0,1)^l\phi(1,1)^r\epsilon(i-k-r, j-l-r).$$

Properties

To establish the asymptotic results, we consider general models for which increments in the two-dimensional process $\boldsymbol{Z}(\boldsymbol{s}) = (X(\boldsymbol{s}), Y(\boldsymbol{s}))^\top$ admit the following structure,

$$\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s}) = \sum_{\boldsymbol{t}\in\mathbb{T}} \boldsymbol{A}(\boldsymbol{t})\boldsymbol{\epsilon}(\boldsymbol{s}-\boldsymbol{t}), \tag{4}$$

where $\boldsymbol{A}(\boldsymbol{t}) = \boldsymbol{A}_{\boldsymbol{h}}(\boldsymbol{t})$ are 2×2 matrices defined for $\boldsymbol{t} \in \{\boldsymbol{t}\in\mathbb{Z}^2 : A(\boldsymbol{t}) \neq \boldsymbol{0}\}$ such that

$$\sum_{\boldsymbol{t}\in\mathbb{T}} ||\boldsymbol{A}(\boldsymbol{t})||^2 < \infty.$$

Here, $||\cdot||$ denotes any matrix norm, and two-dimensional random vectors $\boldsymbol{\epsilon}(\boldsymbol{t})$ are independent with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

Properties

For two processes satisfying (4), we derive the formula for (3).

Proposition

Let

$$\boldsymbol{K} = \mathbb{E}[\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s})][\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s})]^\top$$

be the covariance matrix of the vector process for which (4) holds. Then,

$$\boldsymbol{K} = \sum_{\boldsymbol{t}\in\mathbb{T}} \boldsymbol{A}(\boldsymbol{t})\boldsymbol{\Sigma}\boldsymbol{A}(\boldsymbol{t})^\top, \qquad (5)$$

and

$$\rho_{XY}(\boldsymbol{h}) = \frac{\kappa_{12}}{\sqrt{\kappa_{11}\kappa_{22}}},$$

where for $i, j = 1, 2$, κ_{ij} denote elements of $\boldsymbol{K}$.

Consistency

- Here, we sketch the consistency and asymptotic normality of the sample codispersion coefficient following Rukhin and Vallejos (2008).

- Assume that both processes can be observed on the increasing part of the positive lattice $S = \{0 \leq s_1 < M, 0 \leq s_2 < M\}$.

- For convenience, we assume normality of the error vectors $\boldsymbol{\epsilon}(\boldsymbol{t}) = (\epsilon_1(\boldsymbol{t}), \epsilon_2(\boldsymbol{t}))^\top$ in (4), although the main results hold for other distributions with four finite moments.

- By the law of large numbers, the following convergence in probability holds

$$\frac{1}{M^2} \sum_{0 \leq s_i < M - h_i} (\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s}))(\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s}))^\top \to \boldsymbol{K},$$

where $\boldsymbol{K}$ is given in (5).

Asymptotic Normality

Theorem

If the observed data admit representation (4) with zero mean error $\boldsymbol{\epsilon}(\boldsymbol{t})$ possessing a fourth finite moment, the limiting distribution of $M[\widehat{\rho}_{XY}(\boldsymbol{h}) - \rho_{XY}(\boldsymbol{h})]$ is normal with mean 0 and variance

$$v_{\boldsymbol{h}}^2 = \frac{\varphi_{1122}}{\kappa_{11}\kappa_{22}} + \frac{\kappa_{12}^2\varphi_{1111}}{4\kappa_{11}^3\kappa_{22}} + \frac{\kappa_{12}^2\varphi_{2222}}{4\kappa_{11}\kappa_{22}^3} - \frac{\kappa_{12}\varphi_{1112}}{\kappa_{11}^2\kappa_{22}} - \frac{\kappa_{12}\varphi_{1222}}{\kappa_{11}\kappa_{22}^2} + \frac{\kappa_{12}^2\varphi_{1212}}{2\kappa_{11}^2\kappa_{22}^2}, \tag{6}$$

where $\kappa_{i,j}$ are the elements of matrix $\boldsymbol{K}$, as in Proposition 2, and φ_{ijkl} are the entries of 4×4 matrix $\boldsymbol{\Phi}$ defined as

$$\begin{aligned}\boldsymbol{\Phi} =& \mathbb{E}\left[(\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s}))(\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s}))^\top - \boldsymbol{K}\right] \\ &\otimes \left[(\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s}))(\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{Z}(\boldsymbol{s}))^\top - \boldsymbol{K}\right].\end{aligned}$$

Hypothesis Testing

- As a consequence of the asymptotic normality of the codispersion, we study the hypothesis testing problem

$$\mathrm{H}_0 : \rho_{XY}(\boldsymbol{h}) = 0 \ \text{ versus } \mathrm{H}_A : \rho_{XY}(\boldsymbol{h}) > 0$$

for a fixed $\boldsymbol{h}$.

- We consider $Q = \frac{M(\widehat{\rho}_{XY}(\boldsymbol{h}) - \rho_{XY}(\boldsymbol{h}))}{v} \underset{approx.}{\sim} \mathcal{N}(0,1)$.
- Specifically, under H_0

$$Q_0 = \frac{M\widehat{\rho}_{XY}(\boldsymbol{h})}{v} \sim \mathcal{N}(0,1).$$

- Then, we reject H_0 if $Q_0 > z_\alpha$, where z_α is the $(1-\alpha)$-th percentile of the standard normal distribution.
- The power of the test for $\rho_{XY}(\boldsymbol{h}) = \rho_0 > 0$ is given by

$$\beta = P\left(Q_0 > z_\alpha | H_A\right).$$

An Application: Similarity Between Images

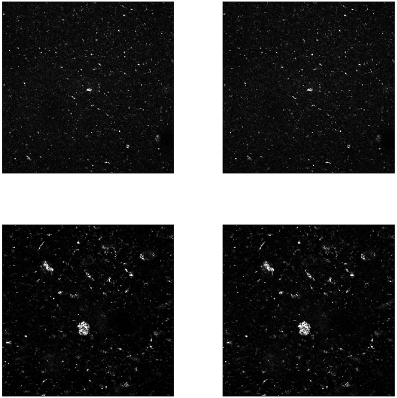


Figure: Dispersion of nanotubes. Images taken at NIST, USA. The images of the top were taken at the same distance. The images of the bottom were taken at the same distance but closer.

Similarity Between Images

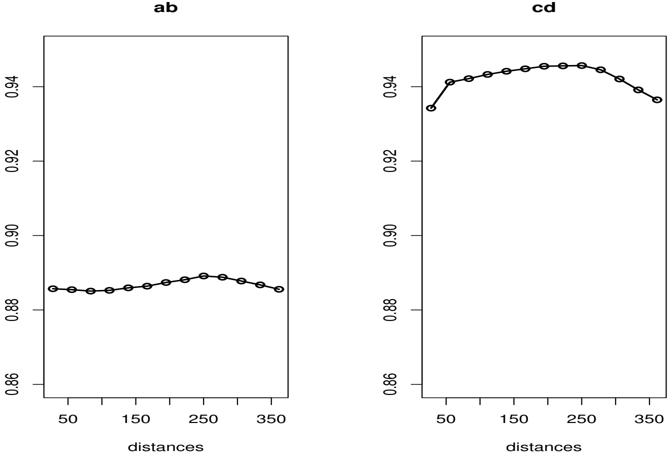


Figure: Codispersion coefficient between the pairs of images (ab) (left) and (cd) (right).

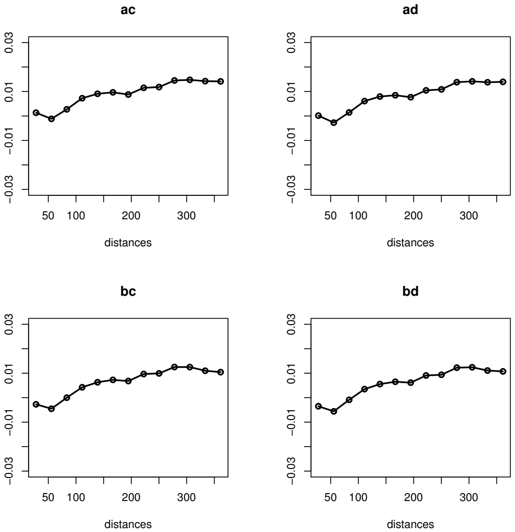


Figure: Codispersion coefficient between pairs of images (ac) (top left), (ad) (top right), (bc) (bottom left), and (bd) (bottom right).

R Code for the Codispersion

```
data(murray)
# defining the arsenic (As) and lead (Pb) variables from the
#murray dataset
x <- murray$As
y <- murray$Pb
# extracting the coordinates from Murray dataset
coords <- murray[c("xpos","ypos")]
# computing the codispersion coefficient
z <- codisp(x, y, coords) z
# plotting the codispersion coefficient vs. the lag distance
plot(z)
```

Complexity of the computations

- To explore the computational time associated with the functions that compute the codispersion coefficient and the modified t-test, a Monte Carlo simulation experiment was conducted.

Complexity of the computations

- To explore the computational time associated with the functions that compute the codispersion coefficient and the modified t-test, a Monte Carlo simulation experiment was conducted.
- Enlarged images sized 8×8, 16×16, 32×32, and 64×64 were generated to study the behavior of the procedures `modified.ttest` and `codisp` as a function of the image size.

Complexity of the computations

- To explore the computational time associated with the functions that compute the codispersion coefficient and the modified t-test, a Monte Carlo simulation experiment was conducted.
- Enlarged images sized 8×8, 16×16, 32×32, and 64×64 were generated to study the behavior of the procedures `modified.ttest` and `codisp` as a function of the image size.
- The generated images were the realizations of two correlated spatial normal variablesusing a well defined bivariate covariance structure (Gneiting et. al., 2010).

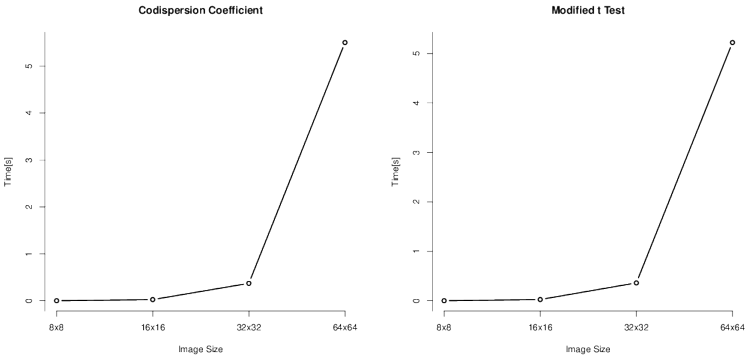
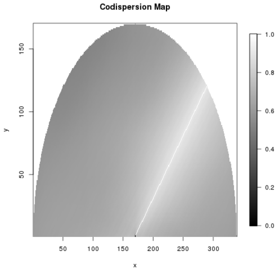


Figure: computational time to evaluate the codispersion coefficient (left) and the modified t test (right) as a function of the image size (from 8×8 to 64×64).

Applications of the Codispersion Coefficient



Geometry of Codispersion Coefficient

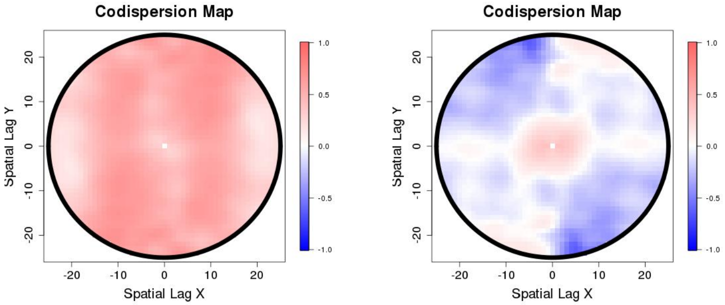
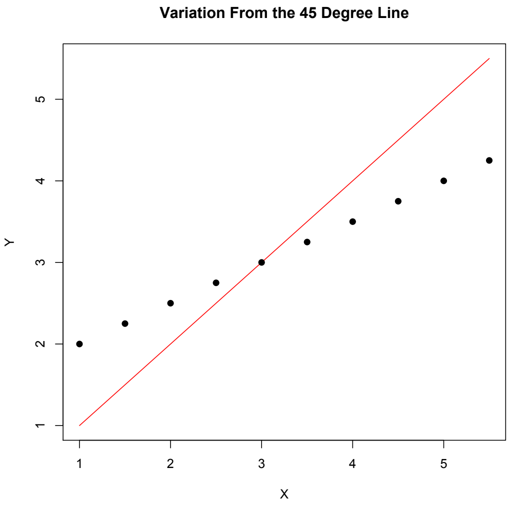


Figure: Soil chemistry variables. (Left) Al and P. (Right) Ca and P

Other Interesting Topics

- Codispersion maps for large datasets
- Spatial Concordance Coefficients (*)
- Effective Sample Size for Codispersion Maps
- Image Similarity Coefficients (SSIM)
- Coefficients of nonlinear association.

Concordance



The Concordance Correlation Coefficient

Definition

Assume that the join distribution of X and Y has finite second moment with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and covariance σ_{12}.

- The mean squared deviation of $D = Y - X$ is

$$\begin{aligned} \mathsf{MSD} = \epsilon^2 = \mathbb{E}[D^2] &= \mathbb{E}[(Y - X)^2] \\ &= (\mu_1 - \mu_2)^2 + \sigma_2^2 + \sigma_1^2 - 2\sigma_{21}. \end{aligned}$$

- The Concordance Correlation Coefficient Lin (1989) is

$$\rho_c = 1 - \frac{\epsilon^2}{\epsilon^2|\rho = 0} = \frac{2\sigma_{21}}{\sigma_2^2 + \sigma_1^2 + (\mu_2^2 - \mu_1^2)^2}.$$

The Concordance Correlation Coefficient

Properties

- $\rho_c = \alpha \cdot \rho$, where $\alpha = \dfrac{2}{w + 1/w + v^2}$, $w = \frac{\sigma_2}{\sigma_1}$, $v = \dfrac{\mu_2 - \mu_1}{\sqrt{\sigma_2 \sigma_1}}$, and $\rho = \mathrm{corr}(X, Y)$.
- $|\rho_c| \leq 1$.
- $\rho_c = 0$ if and only if $\rho = 0$.
- $\rho_c = \rho$ if and only if $\sigma_2 = \sigma_1$ and $\mu_2 = \mu_1$.

Sample Concordance

The sample counterpart of ρ_c is given as

$$\widehat{\rho}_c = \frac{2 s_{21}}{s_2^2 + s_1^2 + (\overline{y} - \overline{x})^2}.$$

Estimation of ρ_c

- Lin (1989) proved that

$$Z = \frac{1}{2}\log\left(\frac{1+\widehat{\rho}_c}{1-\widehat{\rho}_c}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(\psi, \sigma_Z^2), \text{ as } n \to \infty,$$

  where

$$\psi = \tanh^{-1}(\rho_c) = \tfrac{1}{2}\log\left(\frac{1+\rho_c}{1-\rho_c}\right),$$

$$\sigma_Z^2 = \tfrac{1}{n-2}\left[\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{2v^2(1-\rho_c)\rho_c^3}{(1-\rho_c^2)^2\rho} + \frac{v^4\rho_c^4}{2(1-\rho_c^2)^2\rho^2}\right].$$

- The asymptotic result leads to an approximate hypothesis test and a confidence interval for ρ_c.

Our Proposal

- The goal is to construct a concordance coefficient that takes into account the spatial lag $\boldsymbol{h}$, similarly to the variogram and cross-variogram.

- $C_{11}(\boldsymbol{h}) = \mathrm{cov}[X(\boldsymbol{s}), X(\boldsymbol{s}+\boldsymbol{h})],$
  $C_{22}(\boldsymbol{h}) = \mathrm{cov}[Y(\boldsymbol{s}), Y(\boldsymbol{s}+\boldsymbol{h})].$
  $C_{12}(\boldsymbol{h}) = \mathrm{cov}[X(\boldsymbol{s}), Y(\boldsymbol{s}+\boldsymbol{h})].$

- The idea is to define a new coefficient of the form

$$\rho_c = 1 - \frac{\epsilon^2}{\epsilon^2|\rho = 0},$$

but using the above ingredients.

Our Definition

Definition

Let $(X(\boldsymbol{s}), Y(\boldsymbol{s}))^\top$ be a bivariate second order stationary random field with $\boldsymbol{s} \in \mathbb{R}^2$, mean $(\mu_1, \mu_2)^\top$ and covariance function

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} C_{11}(\boldsymbol{h}) & C_{12}(\boldsymbol{h}) \\ C_{21}(\boldsymbol{h}) & C_{22}(\boldsymbol{h}) \end{pmatrix}.$$

Then the spatial concordance coefficient is defined as

$$\rho_c(\boldsymbol{h}) = \frac{\mathbb{E}[(Y(\boldsymbol{s}+\boldsymbol{h}) - X(\boldsymbol{s}))^2]}{\mathbb{E}[(Y(\boldsymbol{s}+\boldsymbol{h}) - X(\boldsymbol{s}))^2 | C_{12}(\boldsymbol{0}) = 0]} = \frac{2C_{21}(\boldsymbol{h})}{C_{11}(\boldsymbol{0}) + C_{22}(\boldsymbol{0}) + (\mu_1 - \mu_2)^2}.$$

References

[1] Acosta, J., Vallejos, R. (2018). Effective sample size for spatial regression processes. Electronic Journal of Statistics 12, 3147-3180.

[2] Acosta, J., Osorio, F., Vallejos, R. (2016). Effective sample size for line transect models with an application to marine macroalgae. Journal of Agricultural, Biological and Environmental Statistics 21, 407-425.

[3] Bevilacqua, M., Vallejos, R., Velandia, D. (2015). Assessing the significance of the correlation between the components of a bivariate Gaussian random field. Environmetrics 26, 545–556.

[4] Clifford, P., Richardson, S., Hémon, D. (1989). Assessing the significance of the correlation between two spatial processes. Biometrics 45, 123-134.

[5] Cressie, N. (1993). Statistics for Spatial Data. Wiley, New York.

[6] Cuevas, F., Porcu, E., Vallejos, R. (2013). Study of spatial relationships between two sets of variables: A nonparametric approach. Journal of Nonparametric Statistics 25, 695-714.

[7] Dutilleul, P. (1993). Modifying the t test for assessing the correlation between two spatial processes. Biometrics 49, 305-314.

[8.] Duttilleul, P. Pelletier, B. Alpargu, G. (2008). Modified F tests for assessing the multiple correlation between one spatial process and several others. Journal of Statistical Planning and Inference 138, 1402-1415.

[9.] Osorio, F., Vallejos, R., Cuevas, F. (2012). `SpatialPack` - An R package for computing the association between two spatial or temporal processes.

[10.] Rukhin A., Vallejos, R. (2008). Codispersion coefficient for spatial and temporal series. Statistics & Probability Letters 78, 1290-1300.

[11.] Vallejos, R. (2012). Testing for the absence of correlation between two spatial or temporal sequences. Pattern Recognition Letters 33, 1741–1748.

[12] Vallejos, R., Osorio, F., (2014). Effective sample size for spatial process models. Spatial Statistics 9, 66-92.

[13.] Vallejos, R., Mancilla, D., Acosta, J. (2016). Image similarity assessment based on measures of spatial association. Journal of Mathematical Imaging and Vision 56, 77-98.

Postgrado en la UTFSM, Chile

- Magister en Ciencias Mención Matemática
  `http://www.magister.mat.utfsm.cl`

- Doctorado en Matemática
  `https://doctoradomat.uv.cl`

Thank you

ronny.vallejos@usm.cl