Global estimation based on indicators factorization

Nicolas Bez

Global estimation based on indicators factorization

nicolas.bez@ird.fr

june 2021

1 Abstract

In the context of highly 0-inflated skewed data, a non linear approach is recommended. However a full disjunctive model is probably too demanding. An intermediate approach consists in using indicators to code the data through a set of disjunctive indicator variables (one by class of density including the class of null densities). However, this gets the following two drawbacks. First, the dimension of the model is increased (moving from mono-variate to multivariate which notably wight up the kriging system). Second, the mathematical coherence of the model is not guaranteed for indicators variables. These two drawbacks were solved by using MAFs. This makes the variables orthogonal both locally and for small distances. Assuming full orthogonality, the new variables are factorized and handleable one by one with their corresponding appropriate variograms. By the way, punctual uncorrelation as produced by standard PCA is not sufficient to handle variables separately. While MAFs are linear combination of the input indicator variables, one can back-transform all the outputs and produce relevant estimations with their estimation variance.

2 Method

2.1 Discretization of $Z(x)$

The random function is decomposed into N disjunctive classes of densities.

$$Y(x) = \phi\big(Z(x)\big) = \sum_{1}^{N} m_i 1_i(x)$$

$$\sum_{1}^{N} 1_i(x) = 1$$

$$m_i = E\big(Z(x)1_i(x)\big)$$

$$p_i = E\big(1_i(x)\big)$$

$$\bar{m} = \sum_{1}^{N} m_i p_i$$

2.2 Closed form

Being a closed formalization,

$$1_N(x) = 1 - \sum_{i=1}^{N-1} 1_i(x)$$

the decomposition of $Z(x)$ into N indicator variables should be reduced to $N-1$:

$$Y(x) = \sum_{1}^{N-1} m_i 1_i(x) + m_N\big(1 - \sum_{1}^{N-1} 1_i(x)\big) = \sum_{1}^{N-1} \big(m_i - m_N\big)1_i(x) + m_N$$

that is finally wrote as:

$$Y(x) = m_N + \sum_{1}^{N-1} \check{m}_i 1_i(x)$$

Given their disjunctive characteristic, the indicator variables are locally correlated:

$$cov\big(1_j(x), 1_k(x)\big) = -p_j p_k < 0$$

They are also not spatially uncorrelated:

$$cov\big(1_j(x), 1_k(x+h)\big) \neq 0$$

Given the above considerations, the modelization of the indicator variables should be done jointly (multivariate geostatistics). However, rigorous multivariate models for indicator variables do not have explicit mathematical characterization. Even if such a model would exist, the dimension of the coKriging system would drastically slow down computations, especially when global kriging is concerned. A factorization of the system permits, without lose of information and generality, to reduce the N-dimensional question into N separate monovariate systems much easily handleable.

2.3 MAFactorization

The factorization (see below) works on centered variables. Denoting:

$$O_i(x) = 1_i(x) - p_i$$

We have:

$$Y(x) = \bar{m} + \sum_{1}^{N-1} \breve{m}_i O_i(x)$$

There is $N-1$ spatially correlated variables that we would like to transform into $N-1$ spatially uncorrelated variables. A sequence of two PCAs (also called Min-max Autocorrelation Factors - MAFs [3]) allow computing factors that are uncorrelated locally ($h = 0$) and at a short distance to be chosen (h_{ref}). As all PCAs, the factors $F_j(x), j = 1, ..., N-1$, are nothing but a linear combination of the input variables. In turns, the input variables are also linear combination of the factors:

$$O_i(x) = \sum_{j=1}^{N-1} a_{j,i} F_j(x)$$

Cautious with the order of the subscripts: $a_{j,i}$ corresponds to the j^{th} line of the i^{th} column of matrix A. In matrix notation this amounts to:

$$O = F \cdot A$$

$$\begin{bmatrix} \bullet & \cdot & \cdot \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix} = \begin{bmatrix} \rightarrow & \rightarrow & \rightarrow \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \downarrow & \cdot & \cdot \\ \downarrow & \cdot & \cdot \\ \downarrow & \cdot & \cdot \end{bmatrix}$$

Given the factors are Min-max Autocorrelation Factors (MAFs), we have by construction that:

$$cov\big(F_j(x), F_k(x)\big) = 0$$

$$cov\big(F_j(x), F_k(x + h_{ref})\big) = 0$$

The discrete version of $Z(x)$ finally writes down:

$$Y(x) = \bar{m} + \sum_{i=1}^{N-1} \breve{m}_i \sum_{j=1}^{N-1} a_{j,i} F_j(x)$$

which can be simplified into:

$$Y(x) = \bar{m} + \sum_{i=1}^{N-1} \alpha_i F_i(x)$$

where (be care full with the order of the subscripts):

$$\alpha_i = \sum_{i=1}^{N-1} \breve{m}_i \sum_{j=1}^{N-1} a_{i,j}$$

In matrix notations, this amounts to:

$$\alpha = A \cdot \breve{m}$$

$$\begin{bmatrix} \bullet \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \rightarrow & \rightarrow & \rightarrow \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \downarrow \\ \downarrow \\ \downarrow \end{bmatrix}$$

2.4 Hypothesis

We now assume that the factors are uncorrelated for all distances, i.e. not only for $h = 0$ and for $h = h_{ref}$:

$$cov\big(F_j(x), F_k(x+h)\big) = 0 \qquad \forall h$$

2.5 Punctual Kriging: mapping

Being spatially uncorrelated, the cokriging of the factors reduced to their monovariate krigings [1], so that:

$$\hat{Y}(x_0) = \bar{m} + \sum_{1}^{N-1} \alpha_i {F^K}_i(x_0)$$

One can also estimate the indicator variables, that is the probabilities to be in a class of density:

$$\hat{O}_i(x) = \sum_{j=1}^{N-1} a_{j,i} {F^K}_j(x)$$

so that

$$\hat{1}_i(x) = p_i + \sum_{j=1}^{N-1} a_{j,i} {F^K}_j(x), \text{ for i} = 1, ..., N-1$$

$$\hat{1}_N(x) = 1 - \sum_{i=1}^{N-1} \hat{1}_i(x)$$

2.6 Global Kriging: estimation of global sA

Thanks to the factorization, the global estimation can be readily deduced:

$$\hat{Y}(v) = \bar{m} + \sum_{1}^{N-1} \alpha_i F_i(v)$$

with

$$\hat{Y}(v) = \bar{m} + \sum_{1}^{N-1} \alpha_i {F^K}_i(v)$$

and

$$\sigma_E^2 = var(\hat{Y}(v) - Y(v)) = \sum_{1}^{N-1} \alpha_i^2 \sigma_{E,i}^2(v)$$

where

$$\sigma_{E,i}^2(v) = var(\hat{F}_i(v) - F_i(v))$$

As in standard estimation, the coefficient of variation is :

$$CV_E = \frac{\sigma_E}{\hat{Y}(v)}$$

3 Discussion

3.1 Assumptions

The factorization is not complete strictly speaking. The MAF are uncorrelated at 0 and, on average for distances falling in the interval used for their elaboration. While building recursive MAFs of larger and larger interval do not solve the problem, the most appropriate way of doing is to enlarge the interval. By doing so, we loose interpretability of the factors that are produced, but we gain in terms of absence of correlation. It is thus possible to generate MAFs whose spatial cross-correlation are on average null over a large distance interval.

As for any global approach, the model is considered relevant for all geographical distances at play. This means that order 2 stationarity is assumed.

We produced the estimation of $Y(v)$ i.e. the discretized version of $Z(v)$. This is <u>not</u> the estimation of $Z(v)$ which is not accessible with reasonable assumptions. This is a change of paradigm that must be acknowledged strongly. It brings the idea that we can estimate the of biomass

3.2 Practical considerations

Some practical questions can be addressed in real cases:

- Should we truncate to 0 the negative estimation ? How ? When in the process ?
- Should we use survey by survey thresholds or should we use fixed threshold e.g. powers of 10 with possible empty class in some surveys ?

References

[1] Chilès JP, Delfiner P, Geostatistics: Modeling spatial uncertainty. J. Wiley & Sons, New York, (1999)

[2] MINES ParisTech / ARMINES, RGeostats: The Geostatistical R Package. Version: 12.0.0, Free download from: http://cg.ensmp.fr/rgeostats (2020)

[3] Switzer P, Green AA, Min/Max Autocorrelation Factors for Multivariate Spatial Imagery. Stanford University, Stanford, Technical Report n6 (1984)