Letter to the Editor

Comment on "Indicator Principal Component Kriging" by V. Suro-Pérez and A. G. Journel

INTRODUCTION

This paper is interesting in different respects, and it is well worth commenting on it. The authors are proposing a technique to look for orthogonal random functions, which are linear transforms of the indicators at a given set of cutoffs, and which generates the same linear space that the considered indicators. Such functions are called factors. From this, indicator cokriging (or discrete disjunctive kriging, which is the same method under a different name),[1] can be performed as a sequence of separate krigings of each factor. In other words, the method proposed is an attempt to build discrete isofactorial models.

A possible basis for this construction is the indicator covariance matrix estimated at a particular distance h_0. Factorial data analysis of this matrix can be performed to compute factors orthogonal at h_0. Checks have then to be done to make sure that orthogonality also holds at other distances.

This approach, was used as well by C. Lantuéjoul and myself, and presented by us in Avignon geostatistic congress (Lajaunie and Lantuéjoul, 1989). More details on the factorial analysis aspects are given in Lajaunie (1986).

However, the two approaches diverge in several places, and here we mention three of them:

(i) Choice of cumulative indicators coding in Suro-Pérez and Journel's paper (subsequently referred to as SPJ), or of complete disjunctive coding of the information, in our approach (referred to as LL). These are defined from the cutoffs considered $z_1 < z_2 < \ldots < z_n$ as:

[1] I do not agree with the authors statement on top of p. 763, which mixes up an estimation method (projection on space spanned by sums $\Sigma\, f_i(Z_i)$) with a particular type of bivariate distributions. Incidently, the property given by the authors is true for some, but not every isofactorial distribution model (see for instance Matheron, 1984a, and Hu, 1988 for some case studies).

0882-8121/92/0700-0555$06.50/1 © 1992 International Association for Mathematical Geology

$$I(x;\, z_i) = \mathbf{1}_{z(x) \leq z_i}$$

$$J(x;\, z_i) = \mathbf{1}_{z(x) \in D_i} \qquad \text{with } D_i = [z_{i-1} - z_i]$$

(ii) Choice of principal component in SPJ, or correspondence analysis in LL, as a factorial method.

(iii) Use of empirical factors in the estimation (SPJ), or as a basis for further modeling (LL).

Suro-Pérez and Journel preferred to use cumulative indicator coding, and principal component analysis. This choice has some unfortunate consequences:

CAN THE FACTORS BE ORTHOGONAL?

In the cumulative indicator coding, the covariances are associated with cumulative bivariate distributions:

$$C_{i,j}(h_0) = F_{i,j}(h_0) - F_i F_j$$

Then the analysis of $C_{i,j}(h_0)$ produces a set of eigenvectors $a_n(i)$:

$$\forall_i \sum_j C_{ij}(h_0) a_n(j) = \lambda_n a_n(i)$$

orthogonal relatively to $C_{ij}(h_0)$:

$$\sum_{i,j} a_n(i) a_m(j) C_{ij}(h_0) = \delta_{nm} \lambda_n \tag{1}$$

orthonormal relatively to the Euclidean distance:

$$\sum_{i,j} a_n(i) a_m(j) = \delta_{nm} \tag{2}$$

and complete:

$$\sum_n a_n(i) a_n(j) = \delta_{ij}$$

From this, the following decomposition can be derived:

$$C_{i,j}(h_0) = \sum_n \lambda_n a_n(i) a_n(j)$$

or in matrix form, using the author's notations:

$$C(h_0) = A \Lambda A^t \tag{3}$$

$$AA^t = A^t A = I \tag{4}$$

From the matrix A it is possible to calculate empirical factors:

$$Y_n(x) = \sum_i a_n(i) \mathbf{1}_{z(x) \leq z_i}$$

Due to Eq. (1) these factors are effectively orthogonal for the particular distance h_0 considered:

$$\mathrm{Cov}\,[Y_n(x + h_0), Y_m(x)] = \sum_{ij} a_n(i)a_m(j)C_{ij}(h_0) = \delta_{nm}\lambda_n$$

At this point, the authors make the assumption that orthogonality holds for every possible distance h as well.

Unfortunately, this does not seem to be so even for $h = 0$. For this would require: Cov $[Y_n(x), Y_m(x)] = \delta_{nm}$. Instead of this, we have:

$$\begin{aligned}\mathrm{Cov}\,[Y_n(x), Y_m(x)] &= \sum_{i,j} a_n(i)a_m(j)\,\mathrm{Cov}\,[I(x; z_i), I(x; z_j)]\\ &= \sum_{i,j} a_n(i)a_m(j)\,[F_{i\wedge j} - F_iF_j]\end{aligned}$$

(where $i \wedge j$ stands for the minimum of i and j), a condition which is by no means implied by Eqs. (2) and (4). (In this regard, I think that on Figure 7b, p. 780, the line segment added by SPJ, joining the origin to the first experimental point on the curve, is misleading.)

In contrast to this, if we use the class indicators in a complete disjunctive coding (use of $J(x; z_i)$ instead of $I(x; z_i)$), the empirical bivariate joint probability is used:

$$W_{ij}(h_0) = \mathrm{Prob}\,[Z(x + h_0) \in D_i;\, Z(x) \in D_j]$$

Then the use of correspondence analysis[2] yields the following eigenvalue problem[3]:

$$\forall i \sum_j W_{ij}(h_0)\chi_n(j) = \mu_n \sum_j W_{ij}(0)\chi_n(j)$$

(The matrix $W_{ij}(0)$ is a diagonal matrix formed from the marginal distribution $W_i = \mathrm{Prob}\,[Z(x) \in D_i]$). The vectors χ_n are orthogonal relative to $W_{ij}(h_0)$:

$$\sum_{ij} \chi_n(i)\chi_m(j)W_{ij}(h_0) = \delta_{nm}\mu_n \tag{5}$$

orthonormal relative to $W_{ij}(0)$:

$$\sum_{ij} \chi_n(i)\chi_m(j)W_{ij}(0) = \delta_{nm} \tag{6}$$

and we have the following isofactorial decomposition of $W_{ij}(h_0)$:

$$W_{ij}(h_0) = w_i w_j \sum_n \mu_n\chi_n(i)\chi_n(j)$$

[2] Correspondence analysis (Benzécri, 1973) originated as a method to study discrete empirical bivariate distributions. This makes it more appropriate in this context.

[3] With a simplification on account of the symmetry of W.

The associated factors:

$$Y'_n(x) = \sum_i \chi_n(i)\mathbf{1}_{z(x)\in D_i}$$

are effectively orthogonal for distance h_0:

$$\mathrm{Cov}\ [Y'_n(x + h_0),\ Y'_m(x)] = \delta_{nm}\mu_n$$

as well as for distance 0:

$$\mathrm{Cov}\ [Y'_n(x),\ Y'_m(x)] = \delta_{nm}$$

So that we are in a more comfortable situation to make the further assumption that orthogonality also holds for other distances. This is due to the more appropriate choice of the metric, where the marginal distribution is taken into account.

DOES THE MODEL PRODUCE VALID INDICATORS COVARIANCE?

In the method proposed by SPJ, the covariances of $Y_n(x)$ are calculated and modeled independently one from the other as $\lambda_n(h)$. The form implicitly used in the indicator cokriging, for the indicator covariance is:

$$C(h) = A \cdot \Lambda(h) \cdot A^t \tag{7}$$

For $\lambda_n(h) \geq 0$ this is effectively a valid covariance matrix. But is it automatically a valid cumulative indicator covariance? For it to be so, one would require the quantities:

$$W_{i,j}(h) = F_{i,j}(h) - F_{i-1,j}(h) - F_{i,j-1}(h) + F_{i-1,j-1}(h)$$

to be positive. I believe that it is rarely the case if cumulative indicators covariance matrix is used. For instance the following matrix:

$$F_{ij} = \begin{pmatrix} 0.20 & 0.29 & 0.34 \\ 0.29 & 0.58 & 0.70 \\ 0.34 & 0.70 & 1. \end{pmatrix}$$

is a valid *cdf* matrix. The marginal *cdf* associated with it is:

$$F_i = (0.34 \quad 0.70 \quad 1.)^T$$

Then the eigenvalues of the covariance of cumulative indicators variables associated:

$$C_{ij} = F_{ij} - F_i \cdot F_j \tag{8}$$

are: $\lambda_0 = 0.0$, $\lambda_1 = 0.03512$, and $\lambda_2 = 0.1393$. If this matrix is interpreted as valid at distance h_0, and if exponential models are fitted to the factor covariances, then for distance $h = h_0/2$, the eigenvalues are changed to:

$$\lambda_k(h_0/2) = \sqrt{\lambda_k(h_0)}$$

The *cdf* for $h_0/2$, calculated by means of Eqs. (7) and (8) is:

$$F_{ij}(h_0/2) = \begin{pmatrix} 0.39 & 0.33 & 0.34 \\ 0.33 & 0.78 & 0.70 \\ 0.34 & 0.70 & 1. \end{pmatrix}$$

with the corresponding unacceptable probabilities:

$$W_{ij}(h_0/2) = \begin{pmatrix} 0.39 & -0.06 & 0.01 \\ -0.06 & 0.50 & -0.08 \\ 0.01 & -0.08 & 0.38 \end{pmatrix}$$

This example is thought to be typical of the behavior at short distances.

If instead of principal components, correspondence analysis had been performed, then we would have ended up with:

$$W_{ij}(h) = W_i W_j \sum_n e^{-\tau_n |h|} \chi_n(i) \chi_n(j)$$

(with $\tau_n = -\log(\mu_n)/h_0$ if $\mu_n > 0$. The zero eigenvalues can safely be disregarded in the summation). This matrix is indeed a valid bivariate probability law.[4] But anyway, even using this more appropriate factorial decomposition, there is no guarantee of this, for a more general modeling of $\lambda_n(h)$. Some warning should be stressed on this point.

CHANGE OF SUPPORT MODELS

Very often, when indicator estimation is needed, a change of support is involved. This was the case in the problem we were dealing with in LL, which was for mining applications, but it is likely to be the case in environment applications as well. The volume at which the measurements are made is generally different from the volume at which estimation is needed.

A typical illustration of this is the case of nutrient deficiency in the soils studied in Rivoirard and Webster (1991). As the cattle move while feeding, they might suffer from nutrient deficiency if the average concentration over the field

[4] This is the bivariate distribution of a stationary Markovian process with generator $A_{ij} = -W_j \Sigma \tau_n \chi_n(i) \chi_n(j)$. More generally, expressions like $W_{ij}(h) = W_i W_j \Sigma e^{-\tau_n t(h)} \chi_n(i) \chi_n(j)$ where $t(h)$ is a valid variogram that could be used.

where they stay is less than a critical level. Therefore, the required distribution is for field support, not for sample support.

For this reason, we preferred to use the empirical factors as a basis for inference of models. The discrete isofactorial models, proposed in Matheron (1984b), are a very flexible framework for modeling the bivariate distributions of discrete random functions. In addition to that, a large family of change of support models is available in this methodology. Therefore, it was just the modeling tool needed to produce more consistent models from the empirical factors.

CONCLUSIONS

The indicators covariance is equivalent to the bivariate distribution of the class index. The factor decomposition of the indicator covariance proposed by SPJ is an isofactorial model of the corresponding bivariate distribution. However, the choices of cumulative indicators and of principal component analysis produce unacceptable inconsistencies. In LL, a correspondence analysis of the bivariate distribution was used to produce more satisfactory empirical factors. These were used in the procedure of identification of discrete isofactorial models, with improved consistency, and the benefit of change of support models.

Ch. Lajaunie
Centre de Géostatistique
École des Mines de Paris
35 Rue St. Honore
77305 Fontainebleau
France

REFERENCES

Benzécri, J. P., 1973, L'analyse des Données, Vol. 2: L'analyse des Correspondances: Dunod, Paris, 619 p.

Hu, L., 1988, Mise en Oeuvre du Modèle Gamma pour l'Estimation de Distributions Spatiales: Doctoral thesis, Ecole des Mines de Paris, 141 p.

Lajaunie, C. 1986, Estimation Directe des Paramètres de Diffusion: Centre de Géostatistique de l'École des Mines de Paris, Fontainebleau, Report N-25/86/G; 21 p.

Lajaunie, C., and Lantuéjoul, C., 1989, Setting up the General Methodology for Discrete Isofactorial Models, *in* M. Armstrong (Ed.), Geostatistics: Proc. Third Int. Geost. Congress, Avignon, Kluwer Academic Press, Dordrecht Holland, Vol. 1, p. 323–334.

Matheron, G., 1984a, Isofactorial Models and Change of Support: Proc. 2nd NATO A.S.I., Geostatistics for Natural Resources Characterisation, Part 1, Reidel Pub. Co., Dordrecht, Netherlands, p. 449–467.

Matheron, G., 1984b, Une Méthodologie Générale pour les Modèles Isofactoriels Discrets: *in* "Sciences de la terre," Annales de l'Ecole Supérieure de Géologie Appliquée et de Prospection Minière, Nancy, N 21, p. 1–64.

Matheron, G., 1989, Two Classes of Isofactorial Models, *in* M. Armstrong (Ed.), Geostatistics: Proc. Third Int. Geost. Congress, Avignon, Kluwer Academic Press, Dordrecht Holland, Vol. 1, p. 309–322.

Rivoirard, J. and Webster R., 1991, Copper and Cobalt Deficiency in Soil—A Study Using Disjunctive Kriging: Cahiers de Géostatistique, Fascicule 1, Compte Rendu des Journées de Geostatistique 6-7 Juin 1991, ENSMP Fontainebleau, p. 205–226.