Mathematical Geology, Vol. 31, No. 6, 1999

Scale Effect on Principal Component Analysis for Vector Random Functions[1]

J. A. Vargas-Guzmán,[2] A. W. Warrick,[2] and D. E. Myers[3]

Principal component analysis (PCA) is commonly applied without looking at the "spatial support" (size and shape, of the samples and the field), and the cross-covariance structure of the explored attributes. This paper shows that PCA can depend on such spatial features. If the spatial random functions for attributes correspond to largely dissimilar variograms and cross-variograms, the scale effect will increase as well. On the other hand, under conditions of proportional shape of the variograms and cross-variograms (i.e., intrinsic coregionalization), no scale effect may occur. The theoretical analysis leads to eigenvalue and eigenvector functions of the size of the domain and sample supports. We termed this analysis "growing scale PCA," where spatial (or time) scale refers to the size and shape of the domain and samples. An example of silt, sand, and clay attributes for a second-order stationary vector random function shows the correlation matrix asymptotically approaches constants at two or three times the largest range of the spherical variogram used in the nested model. This is contrary to the common belief that the correlation structure between attributes become constant at the range value. Results of growing scale PCA illustrate the rotation of the orthogonal space of the eigenvectors as the size of the domain grows. PCA results are strongly controlled by the multivariate matrix variogram model. This approach is useful for exploratory data analysis of spatially autocorrelated vector random functions.

KEY WORDS: dispersion covariances, spatial support, Pearson correlation, spatial scales of variability, PCA, matrix variogram.

INTRODUCTION

Principal component analysis (PCA) is used widely to compute orthogonal components that are linear combinations of the correlated original variables (attributes). PCA is related to *R*-mode factor analysis when performed for the attributes (e.g., soil features). Principal component analysis may be computed from the covariance matrix or from the correlation matrix be-

[1]Received 18, May 1998; accepted 16 November 1998.

[2]Department of Soil, Water and Environmental Science, University of Arizona, 429 Shantz 38, Tucson, Arizona 85721.

[3]Department of Mathematics, University of Arizona, Tucson, Arizona 85721.

tween attributes. Results are in general different. Although PCA does not require any assumptions about autocorrelation or independence between samples, classic maximum likelihood estimators of covariance and variances used to compute Pearson correlations assume samples are independent of each other. It is accepted that in many earth science problems, samples are spatially autocorrelated (e.g., Journel and Huilbregts, 1978). However, PCA is used commonly as a data analysis tool without further considerations. The PCA eigenvectors are assumed valid without domain and sample support considerations. Also, exploration of data some times need to compare the eigenvector structure from one place to another without ''spatial support'' restrictions. As we will show later, such a use of PCA and *R*-mode factor analysis can be inadequate whether the coregionalization structure of the vector random function shows non-intrinsic coregionalization.

In this paper, we analyze the effect of multivariate spatial auto and cross-dependency on classical PCA computed for attributes that are spatial realizations of a vector random function. We provide a theoretical analysis of the effect of autocovariance, cross-covariance, sample supports, and size of the domain for the PCA results. We called this approach *growing scale PCA*. An example is included for completeness to illustrate the scale effect on PCA performed for real field data.

From Classic PCA Method to Geostatistics

The literature about PCA and factor analysis is extensive. Mardia, Kent, and Bibby (1979) and Basilevsky (1994) give many references. Preisendorfer (1988) provides an extensive applied explanation. In recent years, multivariate geostatistics has used PCA to simplify co-kriging. The basic idea is to apply PCA to get independent principal components, then rotate the data to obtain scores that could be kriged separately avoiding the necessity of modeling the cross-variograms required by co-kriging. The kriged scores can then be back rotated to the space of the original variables (Davis and Greenes, 1983). However, the statistically independent PCA scores are rarely spatially orthogonal. In other words, the cross-variograms for the PCA scores are not zero for all lag distances, and the idea of kriging the PCA scores as an alternative to co-kriging becomes limited.

From early works in geostatistics, it is known that the variogram of a random function can be decomposed by nested structures. Wackernagel (1985) explains the factorial kriging method which applies PCA to each multivariate nested structure separately. See also Sandjivy (1984). Thus,

PCA techniques have been incorporated in geostatistics or more directly in variogram modeling and factorial kriging. In the above, there is no mention of the effect of autocorrelation, size and geometry of the domain, and sample supports into the classic PCA for a finite physical domain.

The Classic PCA Method

Principal component analysis transforms a correlated set of attributes into an orthogonal set. The starting hypothesis is that the geometry of the problem allows the existence of orthogonal directions of variability $\mathbf{E}$ in the space of attributes. A data set is defined as a $n \times p$ matrix $\mathbf{Z}$ of n samples and p attributes. Also, a $n \times p$ matrix $\mathbf{Y}$ of independent scores is defined. For simplicity, sample data $\mathbf{Z}$ are assumed to be centered (i.e., mean zero). Then,

$$\mathbf{Y} = \mathbf{Z}\,\mathbf{E} \tag{1}$$

and conversely

$$\mathbf{Z} = \mathbf{Y}\,\mathbf{E}^{-1} \tag{2}$$

The matrix $\mathbf{E}$ is a $p \times p$ orthonormal, so that $\mathbf{E}^{-1}$ can be replaced by $\mathbf{E}^T$. Thus,

$$\mathbf{E}^T\mathbf{E} = \mathbf{E}\,\mathbf{E}^T = \mathbf{I} \tag{3}$$

For each possible new basis there exists a diagonal variance matrix $\mathbf{L}^2$ of $\mathbf{Y}$ given by

$$\mathbf{L}^2 = \frac{1}{n}(\mathbf{Y}^T\mathbf{Y}) = \mathrm{diag}[l_1, \ldots, l_p];\ l_1 \geq \ldots l_p \geq 0 \tag{4}$$

where n must be a large number of samples to obtain an unbiased estimator. Equivalently, $\mathbf{U}^2$ is the estimated covariance matrix for $\mathbf{Z}$:

$$\mathbf{U}^2 = \frac{1}{n}(\mathbf{Z}^T\mathbf{Z}) \tag{5}$$

Total invariant variance is given by

$$\mathrm{trace}\,\mathbf{U}^2 = \sum_{j=1}^{p} u_{jj} = \sum_{j=1}^{p} l_j \tag{6}$$

where u_{jj} are diagonal terms of $\mathbf{U}^2$.

Also,

$$n\mathbf{U}^2 = \mathbf{Z}^T\mathbf{Z} = (\mathbf{Y}\mathbf{E}^T)^T\,\mathbf{Y}\mathbf{E}^T = \mathbf{E}\,\mathbf{Y}^T\mathbf{Y}\mathbf{E}^T \tag{7}$$

and

$$\mathbf{L}_2 = \mathbf{E}^T \mathbf{U}^2 \mathbf{E} \tag{8}$$

$\mathbf{L}^2$ is maximized when $\mathbf{E}$ is the matrix of eigenvectors $\mathbf{Q}$, and $\mathbf{L}^2$ the diagonal matrix of eigenvalues $\bar{\lambda}$ of the positive definite matrix $\mathbf{U}^2$.

Special care should be taken when applying PCA rotation. Note that when data are standardized (i.e., mean zero and variance 1), the covariance matrix $\mathbf{U}^2$ is the correlation matrix between attributes, otherwise $\mathbf{U}^2$ is the covariance matrix. Matrices $\mathbf{Q}$ obtained from the two cases are not the same. If standardized data $\mathbf{z}$ are used, eigenvectors obtained from the correlation matrices may be scaled by the square roots of the eigenvalues $[\bar{\lambda}]^{1/2}$ to get the factors of *R*-mode factor analysis

$$\mathbf{A} = \mathbf{Q}\,[\bar{\lambda}]^{1/2} \tag{9}$$

Factors can be used for computation of nonstandardized scores. Then,

$$\mathbf{Y} = \mathbf{z}\,\mathbf{A} = \mathbf{Z}\,\mathbf{Q}$$

In general, use the eigenvectors to rotate the covariance matrix.

$$\mathbf{U}^2 = \mathbf{Q}\,\bar{\lambda}\,\mathbf{Q}^T = \mathbf{A}\,\mathbf{A}^T \tag{10}$$

and

$$\mathbf{Z} = \mathbf{Y}\,\mathbf{Q}^T \tag{11}$$

The order of the rows of data in $\mathbf{Z}$ does not affect the results. PCA as just described does not take account of the physical location of the samples. Neither does it account for the size of the domain or terrain where the samples came from. Also, it does not consider whether samples are autocorrelated and have different supports (i.e., size and shape of the samples). Note the difference with *Q*-mode factor analysis where the matrix of similarity is made between samples. In practice, *R*-mode factor analysis is applied without restrictions to autocorrelated data when *Q*-mode is also applied to the same data. We show that the results of PCA and therefore *R*-mode factor analysis can be scale dependent if samples are autocorrelated and cross-correlated.

Multivariate Coregionalization

From the classic early works in geostatistics (e.g., Journel & Huijbregts, 1978), the multivariate matrix variogram for p attributes can be considered as a nested structure of q independent random functions

$$\overline{\Gamma}_Z(h) = \sum_{u=1}^{q} \overline{\Gamma}^{u}_{Z^u}(h) \tag{12}$$

where $\Gamma_Z(h)$ is the multivariate matrix variogram for the original random function and $\overline{\Gamma}^u_{Z^u}(h)$ are the q nested structures (i.e., Journel and Huijbregts, 1978). The linear model of coregionalization is

$$\overline{\Gamma}_Z(h) = \sum_{u=1}^{q} \mathbf{B}^u g^u(h) \tag{13}$$

$$\overline{\Gamma}_z(h) = \begin{bmatrix} b^1_{11} & & b^1_{1p} \\ b^1_{21} & \cdots & b^1_{2p} \\ & \vdots & \\ b^1_{p1} & \cdots & b^1_{pp} \end{bmatrix} g^1(h) + \cdots + \begin{bmatrix} b^q_{11} & & b^q_{1p} \\ b^q_{21} & \cdots & b^q_{2p} \\ & \vdots & \\ b^q_{pp} & \cdots & b^q_{pl} \end{bmatrix} g^q(h) \tag{14}$$

The coregionalization matrices $\mathbf{B}^u$ can be used to compute regionalized correlations for each spatial scale of variability (Wackernagel 1985, 1995).

$$r^u_{ij} = \frac{b^u_{ij}}{\sqrt{b^u_{ii} b^u_{jj}}} \tag{15}$$

These coefficients have the disadvantage that they depend on the structures utilized for modeling the multivariate variogram. Different modelers may derive different nested structures and therefore different regionalized correlations. Goulard and Voltz (1992), Myers (1994), and Xie and Myers (1995) provide tools for modeling multivariate variograms. Wackernagel (1985) explains that coregionalization matrices $\mathbf{B}^u$ can be diagonalized to give spatially orthogonal coregionalized factors $\mathbf{A}^u = \mathbf{E}^u \sqrt{\bar{\lambda}^u}$ where $\mathbf{E}^u$ is a matrix of eigenvectors and $\bar{\lambda}^u$ a matrix of eigenvalues. Then,

$$\mathbf{E}^{\mu^T} \mathbf{B}^u \mathbf{E}^u = \bar{\lambda}^u \tag{16}$$

A particular case of the linear model of coregionalization is when the coregionalization matrices are proportional. Then, the intrinsic coregionalization is

$$\overline{\Gamma}_Z(h) = \mathbf{B} \sum_{u=1}^{q} b^u g^u(h) \tag{17}$$

In this situation, matrix $\mathbf{B}$ provides a global set of spatially orthogonal eigenvectors.

In the space of the increments, intrinsic coregionalization has long been recognized as a property of certain multivariate matrix variogram or autocovariance (Journel and Huijbregts, 1978). These authors have explained the advantages of the intrinsic coregionalization in the space of

increments. Under the intrinsic coregionalization model of the variogram, factors or principal components at each lag distance in the space of the increments are independent of spatial structure (i.e., unique principal directions, see also Sandjivy, 1984; Wackernagel, 1985). Because of the lack of spatial orthogonality of the global scores, Wackernagel (1985) has applied the linear model of coregionalization to construct orthogonal coregionalized factors for each nested structure in the multivariate matrix variogram. From Goovaerts (1993) the cross-variograms for the rotated data show a lack of spatial orthogonality depending on the nested structures. Then, the computed global PCA principal components are not orthogonal at each lag distance due to the nonintrinsic coregionalization.

Note that all of the above analyses (already classic) were made in the space of the increments (i.e., the eigenvectors are computed from the matrix variogram at each lag distance). In statistics, we wish an analysis in the average physical space or a region. Therefore, a spatial average analysis is introduced that could provide information about correlation and PCA analysis in the average physical space.

THEORY

Multivariate Dispersion Covariance Matrix

We introduced the extension of classic univariate dispersion variance to the multivariate case (Vargas-Guzmán, Warrick, and Myers, 1999). As a consequence of considering cross-variograms and dispersion (cross) covariances, a covariance matrix for p attributes measured in elements of size v that exactly make up the domain of size V is given by

$$\mathbf{D}^2(v|V) = \frac{1}{V^2}\int_V dx \int_{V'} \overline{\Gamma}_Z(x - x')\, dx' - \frac{1}{v^2}\int_v dx \int_{v'} \overline{\Gamma}_Z(x - x')dx' \qquad (18)$$

where $\overline{\Gamma}_Z(h)$ is the multivariate matrix variogram for a vector spatial random function $Z(x)$

$$\overline{\Gamma}_Z(h) = \begin{bmatrix} \gamma_{11}(h) & & \gamma_{1p}(h) \\ \gamma_{21}(h) & \cdots & \gamma_{2p}(h) \\ & \vdots & \\ \gamma_{p1}(h) & \cdots & \gamma_{pp}(h) \end{bmatrix} \qquad (19)$$

and

$$\mathbf{D}^2(v|V) = \begin{bmatrix} D^2_{11}(v|V) & \cdots & D^2_{1p}(v|V) \\ D^2_{21}(v|V) & \cdots & D^2_{2p}(v|V) \\ & \vdots & \\ D^2_{p1}(v|V) & \cdots & D^2_{pp}(v|V) \end{bmatrix} \tag{20}$$

Diagonal entries $D^2_{ii}(v|V)$ are the univariate dispersion variances and the off-diagonal entries $D^2_{ij}(v|V)$ are the dispersion (cross) covariances between pairs of attributes (e.g., soil or geological features). The meaning of the dispersion covariance between two regionalized attributes is the same as the classic statistical covariance, but it is conditioned to the geometry of v and V and the governing matrix variogram. Although the term dispersion cross-covariance could be used, the term dispersion covariances is maintained because for an infinite second-order stationary domain and point sample support $\mathbf{D}^2(v|V)$ converges to the classic statistical covariance between attributes. A single dispersion covariance between two attributes i and j is given by the difference between the average cross-variogram function within the domain V and within the elements v that exactly make up V

$$D^2_{ij}(v|V) = \frac{1}{V^2}\int_V dx \int_{V'} \gamma_{ij}(x - x')\, dx' - \frac{1}{v^2}\int_v dx \int_{v'} \gamma_{ij}(x - x')\, dx' \tag{21}$$

where $\gamma_{ij}(h)$ for $i \neq j$ is the cross-variogram between i and j.

We also introduced a geostatistically scaled multivariate correlation matrix,

$$\mathbf{R}(v|V) = ((\mathbf{D}^2(v|V))(\mathbf{S}^2)^{-1/2})^{\mathrm{T}}(\mathbf{S}^2)^{-1/2} \tag{22}$$

where $(\mathbf{S}^2)^{1/2}$ is a diagonal matrix of dispersion standard deviations constructed with the root square of the diagonal terms of the dispersion covariance matrix $\mathbf{D}^2(v|V)$. Thus, the scaled correlation matrix is

$$\mathbf{R}(v|V) = \begin{bmatrix} 1 & \cdots & r_{1p}(v|V) \\ r_{21}(v|V) & \cdots & r_{2p}(v|V) \\ & \vdots & \\ r_{p1}(v|V) & \cdots & 1 \end{bmatrix} \tag{23}$$

These correlations depend on the sample supports, size and shape of the domain, and the multivariate matrix variogram. Such a correlation matrix

for a second-order stationary infinite domain converges to the classic multivariate correlation matrix of Pearson coefficients for independent samples. We propose that dispersion covariance matrices and scaled correlation matrices computed for regionalized variables can be used for PCA. We term this to be a *growing scale PCA*. Several advantages can be exploited by conditioning the results of PCA to the shape and size of the domain and samples taken. This technique is useful for exploratory data analysis of spatially auto and cross-correlated data. It includes analysis with samples at different support and domains of different size. Also, this technique is useful for multivariate time series where autocorrelation can produce an effect on the PCA results.

Growing Scale PCA

For a vector random function $\mathbf{Z}(x)$, dispersion covariance matrices such as Eq. (20) can be expressed as continuous variance functions of V with v constant. For a given size v of the elements or blocks, a matrix of functions can be obtained by a variable w to describe V:

$$\mathbf{D}^2(w) = \begin{bmatrix} D^2_{11}(w) & \cdots & D^2_{1p}(w) \\ D^2_{21}(w) & \cdots & D^2_{2p}(w) \\ & \vdots & \\ D^2_{p1}(w) & \cdots & D^2_{pp}(w) \end{bmatrix} \tag{24}$$

In practice w may change quasi-continuously by steps depending on the support v. For point support, $\mathbf{D}^2(w)$ is truly continuous. In this case, $\mathbf{D}^2(w)$ is a spatial average from the model multivariate matrix variogram in the region of size w. Therefore, it is a variance estimated from the ensemble and not only for a particular realization. Now define a new spatial vector random function within the neighborhood $w = W$

$$\mathbf{Y}(x|W) = \{Y_1(x|W), \ldots, Y_2(x|W)\} \tag{25}$$

where x represents the physical location; this leads to

$$\frac{1}{W}\int_0^w Y_j(x|W)\, Y_k(x|W)\, dx = \lambda_k \delta_{jk} \tag{26}$$

where $\delta_{jk} = 0$ for $j \neq k$ and 1 otherwise is a Dirac delta function and λ_k the corresponding eigenvalue. Note that the $Y_j(x|W)$ functions are uncorrelated on the average, but may or may not be spatially orthogonal. The

representation of $\mathbf{Y}(x|W)$ is a vector of score spatial random functions rotated from the original random functions with the eigenvectors $\mathbf{Q}$ and eigenvalues $\bar{\lambda}$ obtained for a neighborhood of size W. Then,

$$\mathbf{Y}(x|W) = \mathbf{Z}(x)\,\mathbf{Q} \tag{27}$$

where $\mathbf{Z}(x)$ is centered by the mean vector.

Now, instead of taking a single size W, we will allow the size of the domain to change a differential of size dw. So the size of the domain will be growing continuously with its shape conserved. The variable or set of variables that will measure such a growth is w. As will be explained, $\mathbf{Q}(w)$ will be a matrix of functions changing as the size w changes. Then $\mathbf{Y}(x, w)$ from

$$\mathbf{Y}(x, w) = \mathbf{Z}(x)\,\mathbf{Q}(w) \tag{28}$$

is a family of random functions that change as the size of the domain w changes.

$$[Y_1(x, w) \cdots Y_p(x, w)] = [Z_1(x) \cdots Z_p(x)] \begin{bmatrix} Q_{11}(w) & \cdots & Q_{1p}(w) \\ & \cdots & \\ Q_{p1}(w) & \cdots & Q_{pp}(w) \end{bmatrix} \tag{29}$$

where all elements are functions. $\mathbf{Z}(x)$ may be standardized by the standard deviation.

The matrix of eigenvectors $\mathbf{Q}(w)$ is orthonormal for each w. Then,

$$\mathbf{Q}(w)\,\mathbf{Q}^T(w) = \mathbf{I} \tag{30}$$

The eigenvector functions or factor functions can be computed from the dispersion covariances

$$\mathbf{Q}^T(w)\,\mathbf{D}^2(w)\,\mathbf{Q}(w) = \bar{\lambda}(w) \tag{31}$$

where $\bar{\lambda}(w)$ is the diagonal matrix of eigenvalue functions. If $\mathbf{Q}(w)$ becomes constant for a certain size of neighborhood, then $\mathbf{Q}(w) = \mathbf{Q}$ and we can call $\bar{\lambda}(w)$ the dispersion covariances of the principal component random functions for the size of the neighborhood.

In the same way, PCA can be performed from the matrix of correlation functions for the attributes such as Eq. (23). Such a correlation function is not the spatial correlation, but it is an average correlation between attributes or feature variables within a domain of size w when samples are taken at the size v. Thus, by knowing the correlation functions between attributes as a function of the size of the domain, a matrix of functions $\mathbf{R}(w)$ can be constructed. Consider again uniform support. Then, we have

$$\mathbf{e}^T(w)\,\mathbf{R}(w)\,\mathbf{e}(w) = \overline{\lambda}_R(w) \tag{32}$$

where $\mathbf{e}(w)$ and $\overline{\lambda}_R(w)$ are the eigenvectors and eigenvalue functions for the size w. $\mathbf{R}(w)$ is similar to the dispersion covariance for standardized data. Note, standardization is not constant because the diagonal matrix $\mathbf{d}(w)$ of standard deviations $\sqrt{D_{ii}^2(w)}$ is a function of the size of the domain. Therefore, assuming the mean is independent of the size of the domain,

$$z_{\text{standardized}}(x, w) = (\mathbf{Z}(x) - \mu)(\mathbf{d}(w))^{-1} \tag{33}$$

If data are available, a different set of standardized z can be computed for each size of neighborhood w.

The Linear Model of Coregionalization and Growing Scale PCA

Intrinsic correlation in the average space of the regionalized variables is a particular case where the classic correlation matrix **R** between attributes is independent of the size of the domain and sample supports. Therefore, a unique numerical matrix of eigenvectors is obtained for any size w. Under intrinsic correlation, if the covariance matrix is used for PCA, the eigenvalues grow proportionally to the elementary dispersion variances. The eigenvectors remain independent of size and shape of the domain and sample supports. This allows, in some ways, legitimate application of some techniques designed for independent samples (i.e., classic PCA) to a spatially correlated vector random functions. However, in most of the cases correlation is nonintrinsic. We have shown that it is wrong, although commonly done, simply to ignore the effect of the size and shape of the domain and sample supports when analyzing results of PCA or *R*-mode factor analysis.

For nonintrinsic correlation, dispersion covariances can be calculated separately for each nested structure u of the multivariate variogram.

$$\mathbf{D}^{u2}(v|V) = \frac{1}{V^2}\int_V dx \int_{V'} \overline{\Gamma}^u_{Z^u}(x - x')\,dx' - \frac{1}{v^2}\int_v dx \int_{v'} \overline{\Gamma}^u_{Z^u}(x - x')\,dx' \tag{34}$$

Also, we have

$$\mathbf{D}^{u2}(v|V) = \mathbf{B}^u d^u(v|V) \tag{35}$$

where $d^u(v|V)$ is the scalar elementary dispersion variance computed from the elementary average variogram $\overline{g}^u(h)$.

$$d^{u2}(v|V) = \frac{1}{V^2}\int_V dx \int_{V'} g^u(x - x')\,dx' - \frac{1}{v^2}\int_v dx \int_{v'} g^u(x - x')\,dx' \tag{36}$$

Holding v constant and considering $\mathbf{D}^{u2}(v|V)$ as a function of the size w of the growing domain, the eigenvector property is

$$\mathbf{Q}^{u^T}(w)\,\mathbf{D}^{u2}(w)\,\mathbf{Q}^{u}(w) = d^{u2}(w)[\mathbf{Q}^{u^T}\mathbf{B}^{u}\mathbf{Q}^{u}] = d^{u2}(w)\overline{\lambda}^{u} \tag{37}$$

In this way, the coregionalized factors (constant) are related to the eigenvector functions for growing domains.

The total dispersion variance for the elements of support v in the growing domain is

$$\mathbf{D}^2(w) = \sum_{u=1}^{q} \mathbf{D}^{u2}(w) = \sum_{u=1}^{q} \mathbf{B}^{u}\, d^{u2}(w) \tag{38}$$

Then, $\mathbf{D}^2(w)$ is

$$\mathbf{D}^2(w) = \sum_{u=1}^{q} d^{u2}(w)[\mathbf{Q}^{u}\overline{\lambda}^{u}\mathbf{Q}^{u^T}] \tag{39}$$

and the total eigenvalues matrix is

$$\mathbf{L}^2(w) = \mathbf{Q}(w)\left(\sum_{u=1}^{q} d^{u2}(w)[\mathbf{Q}^{u}\overline{\lambda}^{u}\mathbf{Q}^{u^T}]\right)(\mathbf{Q}(w))^T \tag{40}$$

As can be seen from the equations above, no direct relationship exists between the eigenvalues and eigenvectors of each nested structure (i.e., coregionalized factors) and the global eigenvalues and eigenvectors for a given w. Rotation is required to make the q spaces parallel (on the original variables space) and then a sum and a global rotation on the total eigenvectors will produce the $\mathbf{L}^2(w)$ matrix.

Under the intrinsic correlation

$$\mathbf{D}^2(w) = \sum_{u=1}^{q} \mathbf{D}^{u2}(w) = \mathbf{B}\sum_{u=1}^{q} b^{u} d^{u2}(w) \tag{41}$$

$$\mathbf{L}^2(w) = \mathbf{Q}(w)\left(\sum_{u=1}^{q} b^{u} d^{u2}(w)\right)\mathbf{B}(\mathbf{Q}(w))^T \tag{42}$$

there is a single constant eigenvectors structure computed from matrix $\mathbf{B}$. This case is not common in practice and the eigenvectors will change whether the multivariate matrix variogram model is not an intrinsic coregionalization.

Eigenvalues and Dispersion Variances

Eigenvalues computed from dispersion covariance matrices such as Eq. (25) can be interpreted as dispersion variances when they are functions

of the size of the domain and sample supports. Then, the $\mathbf{L}^2(w)$ are dispersion variances of a vector random function. Unfortunately, since the principal components are not constant in most instances (i.e., they are nonintrinsic independent), $\mathbf{L}^2(w)$ corresponds to different orthogonal spaces for each w. This complicates its practical use for computing variograms. However, we could find cases in practice where the deviation of the eigenvectors from constant values is very small at least for a local neighborhood. Although the lack of orthogonality may arise at larger lag distances due to nonintrinsic correlation caused by different length of range or drift component in each attribute, still a local stationarity may be sufficient for kriging purposes. Also, we can consider locally intrinsic correlation as valid for a restricted neighborhood w depending on how much the experimental eigenvectors from growing scale PCA depart from constants within w. In such a case, we postulate computation of variograms from variances. In one dimension, the method starts by computing a numerically estimated covariance functions from experimental dispersion covariances for point support. Next is growing scale PCA in a sequence of regularly increasing size of the domain w, where w is less than or equal to the local neighborhood. The eigenvalues can be used to construct a numerical representation of a G^2 function

$$G^2(w) = w^2 L^2(w) \tag{43}$$

Then, the sample variogram for the probable intrinsic principal components in the neighborhood is

$$\hat{\gamma}_Y(h) = \frac{\partial^2 G^2(w)}{2\partial^2 w} \tag{44}$$

This approach is justified if sample eigenvectors deviate only slightly or not at all from constants. However, such vectors are taken from the sample, which in fact is obtained from a particular realization (i.e., regionalized random variable) of the random function. Therefore, they may deviate from the intrinsic correlation even if the ensemble or total vector random function is intrinsically correlated. In this sense, the question would be: How much tolerance should be given to the nonconstant sample eigenvectors? Of course, this question should be answered according to experimental observations depending on each particular case.

The principal components diagonal matrix variogram can be modeled. Then, the orthogonal variogram can be used to generate an intrinsic multivariate variogram of the original variables restricted to the neighborhood. Alternatively, the data can be rotated to work in the space of the eigenvectors. This approach may allow kriging of the scores.

FIELD EXAMPLE

To illustrate a numerical example of growing scale PCA, a typic torrifluvents soil classified as Trix soil series has been utilized. The fields are called MAC 28-31 located at the Maricopa Agricultural Center of the University of Arizona (Warrick and others, 1990). For this example, clay, sand, and silt were chosen, from 16 monitored attributes, because they guarantee nice PCA results. These attributes have significant linear correlations, close to normal probability distributions, second-order stationarity and a nice coregionalization model. These textural attributes are commonly used in soil classification and PCA analysis of soil multivariate systems including water content and chemical attributes. Note, one of the attributes can be expressed as a function of the other two; this additional piece of information is beneficial for our example because it assures the system of three attributes reduces to two orthogonal factors. Because the three attributes add to a constant everywhere, the variogram of their sum should be theoretically zero. Under intrinsic correlation these attributes should not show scale effects.

A matrix variogram computed for attributes clay, sand, and silt respectively was modeled with the linear model of coregionalization following Eq. (14)

$$\Gamma_z(h) = \begin{bmatrix} 7.098 & -7.820 & 0.721 \\ -7.820 & 8.616 & -0.794 \\ 0.721 & -0.794 & 0.073 \end{bmatrix} g^1(h) + \begin{bmatrix} 15.336 & -23.182 & 8.303 \\ -23.182 & 36.096 & -12.928 \\ 8.303 & -12.928 & 4.631 \end{bmatrix} g^2(h) + \begin{bmatrix} 1.918 & -4.086 & 3.281 \\ -4.086 & 20.918 & -16.797 \\ 3.281 & -16.797 & 13.489 \end{bmatrix} g^3(h)$$

where the elementary variograms $g^1(h) = (1 - \delta(h))$ is nugget with ($\delta(h) = 1$ if h = 0 and δ(h) = 0 otherwise), $g^2(h)$ is spherical of range 150 m with unit sill, and $g^3(h)$ = spherical range 365 m with unit sill. This model has been checked for closure by considering the variogram of the sum of the three attributes should be close to zero. This coregionalization model shows the studied attributes are nonintrinsically correlated. Thus, the scaled corre-

lation model expressing Eq. (22) as the linear combination of nested random functions is

$$\mathbf{R}(w) = \left[\begin{bmatrix} 1 & -0.999 & 0.9994 \\ -0.999 & 1 & -0.999 \\ 0.9994 & -0.999 & 1 \end{bmatrix}\mathbf{W}^0\right]^T\mathbf{W}^0 + \left[\begin{bmatrix} 1 & -0.9853 & 0.9853 \\ -0.9853 & 1 & -1 \\ 0.9853 & -1 & 1 \end{bmatrix}\mathbf{W}^1\right]^T\mathbf{W}^1 + \left[\begin{bmatrix} 1 & -0.645 & 0.6451 \\ -0.645 & 1 & -1 \\ 0.6451 & -1 & 1 \end{bmatrix}\mathbf{W}^2\right]^T\mathbf{W}^2$$

where the numerical matrices are the regionalized correlation coefficients. The diagonal matrices $\mathbf{W}^u$ for each domain V and support v are calculated from

$$\mathbf{W}^u = (d^{u2}(v_i|V_i)\mathbf{b}^u)^{1/2}\left[\sum_{u=1}^{q}(d^{u2}(v_i|V_i)\mathbf{b}^u)\right]^{-1/2} = \mathbf{S}^u\mathbf{S}^{-1} \tag{45}$$

where the elementary dispersion variances $d^{(1)2}(v|V)$ and $d^{(2)2}(v|V)$ for each spatial scale of variability are computed with Eq. (36) and $\mathbf{b}^u$ are diagonal matrices from the major diagonal of the coregionalization matrices. Note that the high correlations of the nugget component are attributed to the coregionalization matrix utilized. In some cases, the nugget effect is also associated to measurement errors. Also note, if the total nested multivariate matrix variogram is held constant, then changing the involved structures of spatial components should not change the global scaled correlations.

The scaled correlation model has been plotted for square domains of side $0 \leq w \leq 2000$ m and square sample supports {point, 50, 100, and 150} in Figure 1. Point support has been assumed for samples that are much smaller than the sampled field. This figure provides abundant information about the correlation behavior. Note these curves give clear idea about the scale of spatial cross-correlation too. For point support, the correlation between two attributes of this soil approaches a constant at two and three times the largest range (i.e., largest spatial scale of variability) in the nested model matrix variogram. The asymptotic convergence for large domains follows from the second order stationarity of the random functions. For

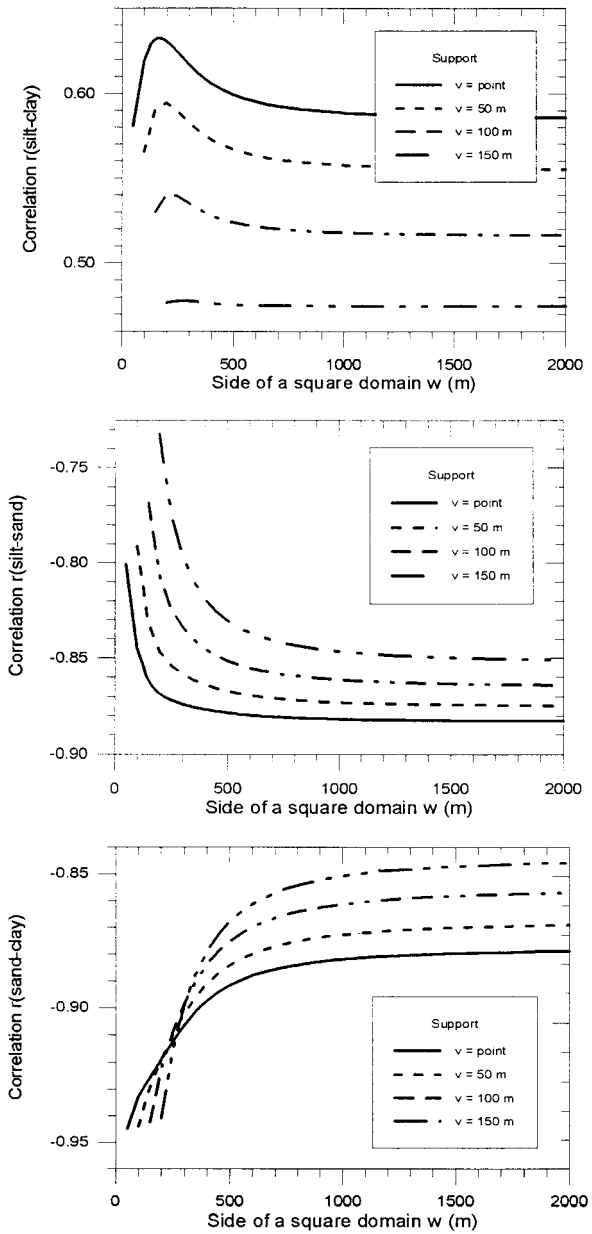

Figure 1. Correlation from square elements v in a domain of side w.

larger sample supports, correlation functions tend to be shifted to the right. The shape of those curves is controlled by the contributions of the spatial components to dispersion variances present in the $\mathbf{W}^u$ diagonal matrices multiplied by the regionalized correlation coefficients.

The correlation function for silt and clay shows a maximum for w close to 150 m. The shift to the right is because of the averaging within square domain and sample supports. Note that 150 m also corresponds to the range of one of the nested structures. The scaled correlation between silt and sand is a monotonically decreasing function. On the other hand, the correlation between sand and clay is a monotonically increasing function. An observation is that, in the three cases, the asymptotic correlation is decreasing in absolute value as the sample supports increases.

Growing scale PCA computed from the correlation functions of Figure 1 illustrates the scale dependency of the results. Figure 2 illustrates the eigenvalues as a function of the size of the domain and sample supports. The first eigenvalue has values higher than one meaning that it absorbs most of the variability. Recall that these eigenvalues are standardized variances. As expected, the third eigenvalue in Figure 2 takes very small values and could be dropped out from the system. For factor analysis, the smaller eigenvalue is interpreted as noise. However, in this example, it has some structure. Due to the strong deviation from the intrinsic correlation, the shapes exhibit a maximum for the first eigenvalue and a minimum for the second eigenvalue. Both extremes occur around 150 m. The magnitude of the scale effect depends on each particular vector random function. Such an effect is more dramatic for attributes having highly different variograms and cross-variograms.

Figures 3–5 illustrate the matrix of scale dependent eigenvectors. Note each figure is an eigenvector which has three components. Their locations in the matrix are given in parenthesis. Figure 3 is the plot for the first eigenvector or principal component. The three elements of the eigenvector correspond to the cosine of the angle (correlation) between the eigenvector and the original variables clay, sand, and silt respectively. For the intrinsic hypothesis we would get constant values (i.e., horizontal lines), so deviations from straight lines can be interpreted as rotations. Figures showing eigenvectors are strongly influenced by the correlations shown in Figure 1. The clay component exhibits a monotonically increasing behavior, the sand component shows a minimum, and the silt component is a monotonically decreasing function. The absolute values of the three components of the first eigenvector are similar. However, the first eigenvector has slightly higher correlation with sand.

Figure 4 shows a principal component that is strongly correlated in absolute value to clay and silt, but less correlated to sand. This eigenvector

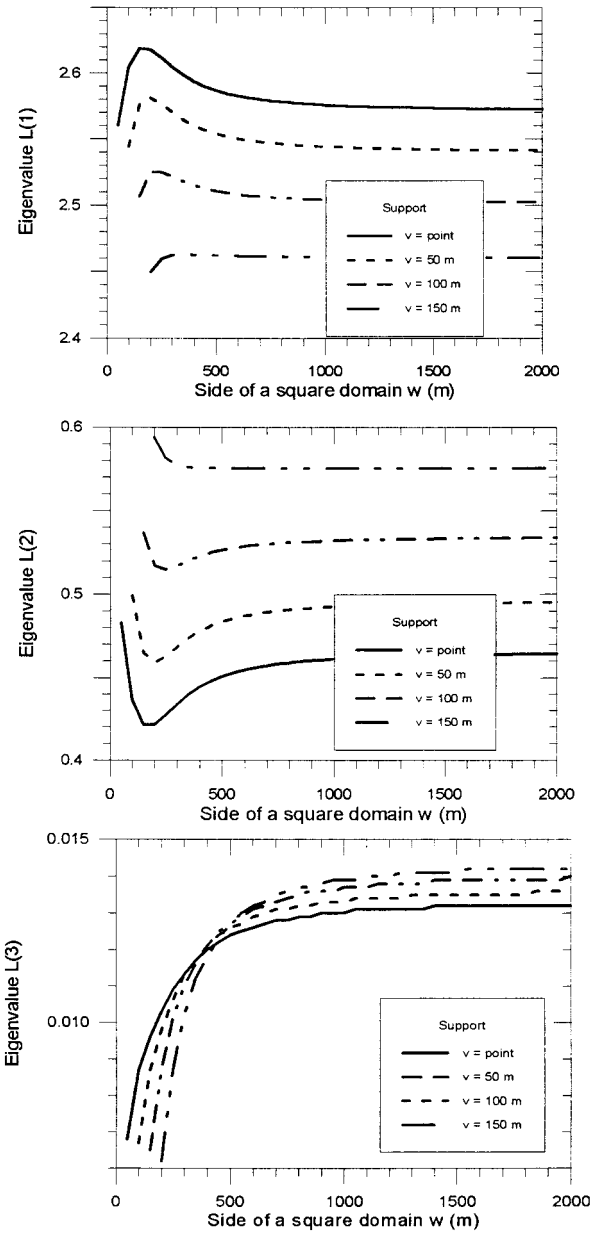

Figure 2. Eigenvalues computed from the correlation functions shown in Figure 1.

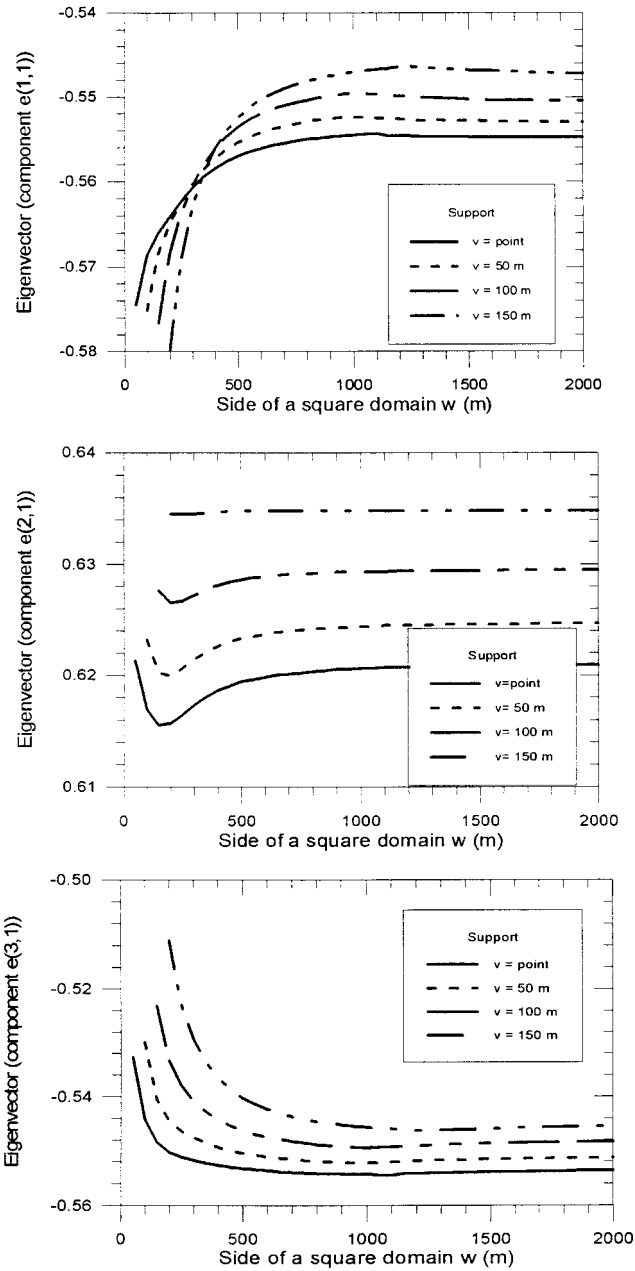

Figure 3. First eigenvector (components for clay, sand, and silt).

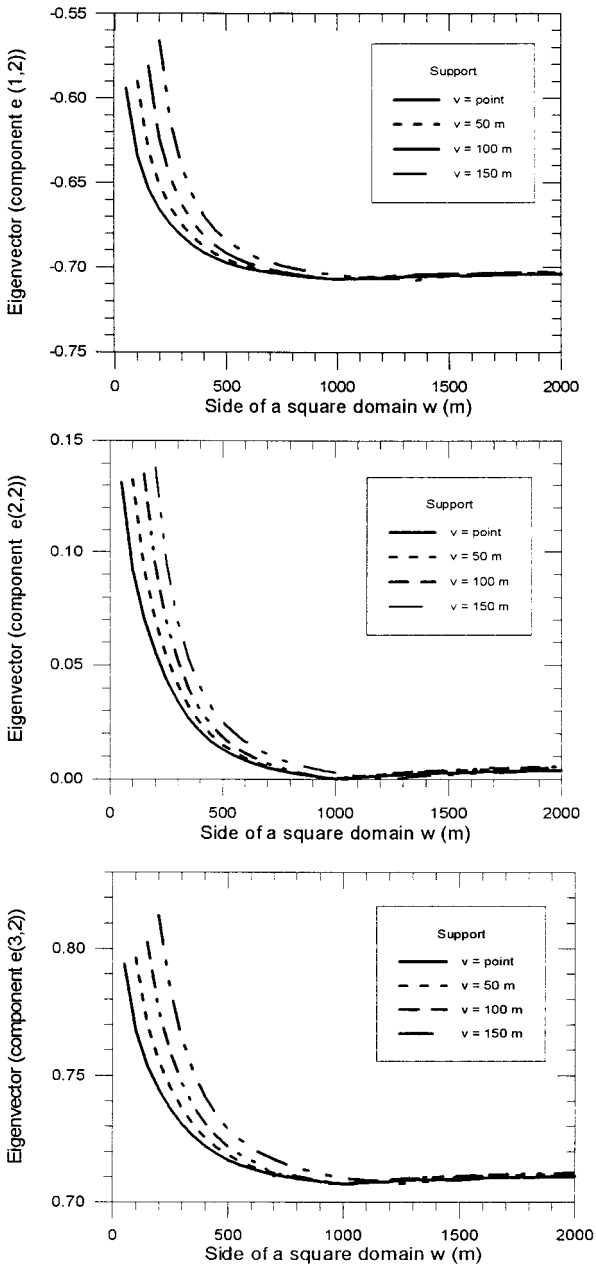

Figure 4. Second eigenvector (components for clay, sand, and silt).

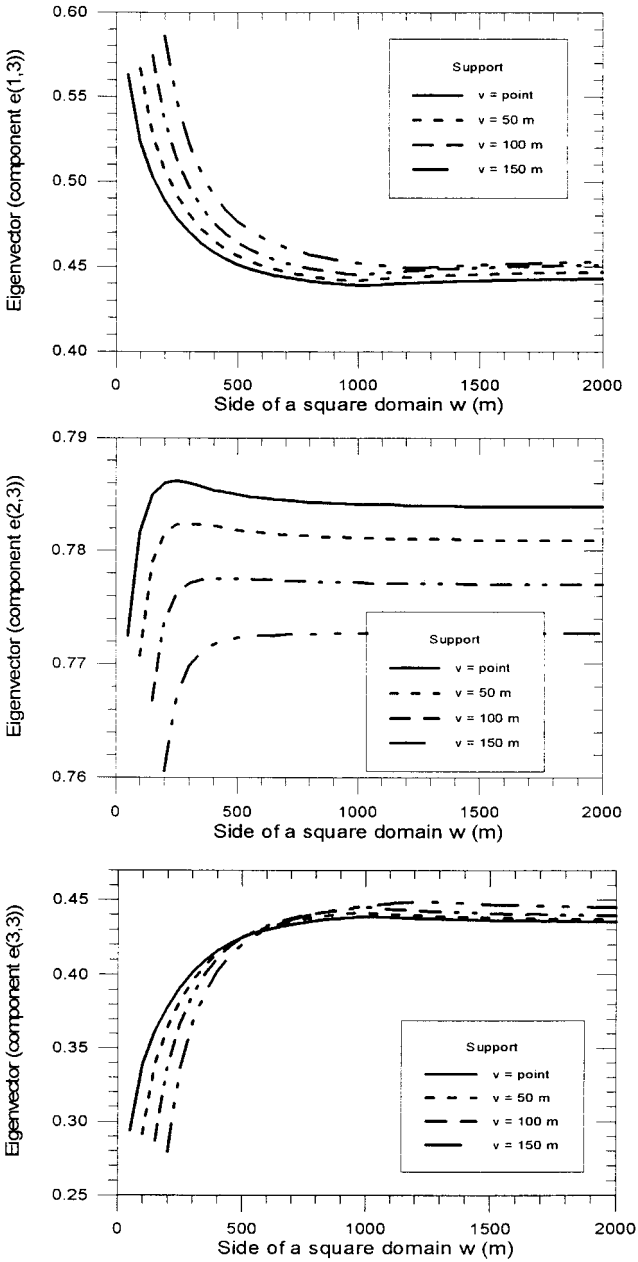

Figure 5. Third eigenvector (components for clay, sand, and silt).

becomes asymptotically scale independent for w larger that 1000 m. This is of course a consequence of the coregionalization. The three components are monotonically decreasing functions. The shift to the right due to larger sample supports is observed for smaller domains.

The third eigenvector is more correlated to sand. The sand component exhibits a maximum (Figure 5). The clay component exhibits a monotonically decreasing function. On the other hand, the correlation with silt is an increasing function.

The three eigenvectors and eigenvalues all exhibit the asymptotic behavior due to second order stationarity. In general, larger-scale effect may result from other attributes corresponding to highly nonproportional variograms and cross-variograms. Cases where samples are taken at different support may exhibit larger scale effects in correlation and PCA. Note that anisotropy is also important and case specific.

CONCLUSIONS

Scale-dependent PCA is a consequence of the nonintrinsic coregionalization present in the multivariate matrix variogram. Therefore, performing PCA for a vector random function should be done cautiously when the spatial correlation structure is unknown. Exploratory studies using PCA should consider the spatial autocorrelation existing between the data. Size, shape of domain, and sample supports can be taken into account by utilizing dispersion (cross) covariances. This leads to a PCA approach that accounts for the geometry and spatial correlation of the vector random function. This approach has been called a ''growing scale PCA.'' Results of this technique, applied to correlation matrices of textural soil data, demonstrate that asymptotic behavior of the PCA results exists under second-order stationarity. However, for smaller domains the results of PCA can be strongly scale dependent. In conclusion, PCA and *R*-mode factor analysis as traditionally used may be highly limited in earth science studies if are performed without considering the spatial auto and cross-correlation. Potential exists for using this type of exploratory tool for quantitatively detecting the spatial scales of variability and also for detecting intrinsic correlations for local neighborhoods.

ACKNOWLEDGMENTS

This research was supported by Western Regional Research Project W-188 and by the U.S. Nuclear Regulatory Commission (Contract No. NRC-04095-046).

REFERENCES

Basilevsky, A., 1994, Statistical factor analysis and related methods, theory and applications: John Wiley & Sons, New York, 737 p.

Davis, B. M., and Greenes, K. A., 1983, Estimation using spatially distributed multivariate data. An example with coal quality: Math. Geology, v. 15, n. 2, p. 287–300.

Goovaerts, P., 1993, Spatial orthogonality of the principal components computed from coregionalized variables: Math. Geology, v. 25, no. 3, p. 281–301.

Goulard, M., and Voltz, M., 1992, Linear coregionalization: Tools for estimation and choice of cross-variogram matrix: Math. Geology, v. 24, no. 3, p. 269–286.

Journel, A. G., and Huijbregts, Ch. J., 1978, Mining geostatistics: Academic Press, New York, 600 p.

Mardia, K. V., Kent, J. T., and Bibby, J. M., 1979, Multivariate analysis: Academic Press, London, 521 p.

Myers, D. E., 1994, The linear model and simultaneous diagonalization of the variogram matrix function, *in* Fabbri, A. G., and Royer, J. J., eds., 3rd CODATA Conference on Geomathematics and Geostatistics: Sci. de la terre, Sér. Inf., Nancy, v. 32, p. 125–139.

Preisendorfer, R., 1988, Principal component analysis in meteorology and oceanography: Elsevier, New York, 425 p.

Sandjivy, L., 1984, The factorial kriging analysis of regionalized data. Its application to geochemical prospecting, *in* Verly, G., and others, eds., Geostatistics for Natural Resources Characterization: NATO-ASI Series C., v. 122, Reidel Publ. Co., Dordrecht, p. 559–572.

Vargas-Guzmán, J. A., Warrick, A. W., and Myers, D. E., 1999, Multivariate correlation in the framework of support and spatial scales of variability: Math. Geology, v. 31, no. 1, p. 85–103.

Wackernagel, H., 1985, The inference of the linear model of coregionalization in the case of a geochemical data set: Ecole des Mines de Paris, Centre de Geostatisque et de Morphologie Mathematique, Fontainebleau, 14 p.

Wackernagel, H., 1995, Multivariate geostatistics: Springer, Berlin, 256 p.

Warrick, A. W., Musil, S. A., Artiola, J. F., Hendricks, D. E., and Myers, D. E., 1990, Sampling strategies for hydrological properties and chemical constituents in the upper vadose zone, final technical report: University of Arizona, Tucson, Arizona, p. 117.

Xie, T., and Myers, D. E., 1995, Fitting matrix valued variogram models by simultaneous diagonalization. I Theory: Math. Geology, v. 27, no. 7, p. 867–876.