# MODEL INTRINSIC DIMENSIONALITY IN PATTERN RECOGNITION ANALYSIS OF STRUCTURAL DATA OF COMPLEX HYDRIDES

Oldřich ŠTROUF and Jiří FUSEK

*Institute of Inorganic Chemistry,*
*Czechoslovak Academy of Sciences, 250 68 Řež*

Received July 21st, 1978

The intrinsic dimensionality of the MODEL-1 representing the system of the classes of stable and unstable complex hydrides of the type $ABH_nD_{4-n}$ (A = alkali metal, B = IIIB atom, D = ligand and $n$ = number of hydride atoms) has been estimated to be eleven. The proposed intrinsic dimensionality determination is based on Karhunen–Loève expansion followed by a relatively simple procedure for the formation of near optimum feature set.

Chemical structure represents a highly concise information source for data analysis methods including those dealing with classification – pattern recognition methods. However, the crucial problem is the encoding the structure into an appropriate input for pattern recognition analysis with minimum loss of discriminatory information. Formal structural data (descriptors) have been frequently used for these purposes[1-3]. Recently, the encouraging results of pattern recognition classification of a set of complex hydrides into stable and unstable ones have been described[4] using MODEL-1 (ref.[5]). This model was constructed mainly from physically significant structural data combined with some descriptors for more complicated formula fragments, *cf.* the ligands. Such a type of model could be generally advantageous from the points of view of availability as well as accuracy of fundamental physical data and the possible use of the analysis results in a theoretical interpretation of the system under consideration.

In this paper the compression of discriminatory information of the MODEL-1 by reduction of dimensions up to the intrinsic one is described. The reduction is checked by the performance of the novel classification method described in our preceding paper[6].

## *Methodical Approach*

Initially[5], 49 available variables (Table I) have been included in the original hypothesis, 8 of which for alkali metal, 8 for the central IIIB atom and 11 for each ligand. The prevailing portion of them (82%) bear a simple physical character. Only remaining

TABLE I

Reduction of the MODEL-1 Dimensionality by the Deletion of Linearly Dependent Features

The deleted variables or features are marked by — and the remaining features by R. The first ligand atom is abbreviated by FLA.

| 49[a] | 28[b] | 23 | 11 | 10 |
|---|---|---|---|---|
| (variables) | (features) | | ($D^{\mathrm{m}}$) | |
| Alkali metal A | | | | |
| Melting point | — | — | — | — |
| Boiling point | — | — | — | — |
| Density | — | — | — | — |
| Atomic radius | R | R | — | — |
| Covalent radius | R | — | — | — |
| Ionic radius | R | R | — | — |
| 1st ionization energy | R | R | R | R |
| Electronegativity | R | — | — | — |
| Central atom B | | | | |
| Melting point | R | R | R | R |
| Boiling point | R | — | — | — |
| Density | R | R | R | R |
| Atomic radius | R | R | — | — |
| Covalent radius | R | — | — | — |
| Ionic radius | R | — | — | — |
| 1st ionization energy | R | — | — | — |
| Electronegativity | R | R | — | — |
| The first ligand D | | | | |
| Molecular weight | R | R | R | R |
| No of chain atoms | R | R | R | — |
| B.p. of DH | — | — | — | — |
| Density of DH | — | — | — | — |
| π-Donor or acceptor | R | R | — | — |
| No of substituents on FLA | R | R | R | R |
| Average electronegativity of these | — | — | — | — |
| Covalent radius of FLA | — | — | — | — |
| Ionic radius of FLA | — | — | — | — |
| 1st ionization energy of FLA | — | — | — | — |
| Electronegativity of FLA | R | R | R | R |
| The second ligand D | | | | |
| Molecular weight | R | R | — | — |
| No of chain atoms | R | R | — | — |
| B.p. of DH | — | — | — | — |

TABLE I
(*continued*)

| 49[a]<br>(variables) | 28[b]<br>(features) | 23 | 11<br>($D^m$) | 10 |
|---|---|---|---|---|
| Density of DH | — | — | — | — |
| π-Donor or acceptor | R | R | R | R |
| No of substituents on FLA | R | R | R | R |
| Average electronegativity of these | — | — | — | — |
| Covalent radius of FLA | — | — | — | — |
| Ionic radius of FLA | — | — | — | — |
| 1st ionization energy of FLA | — | — | — | — |
| Electronegativity of FLA | R | R | — | — |
| The third ligand atom D | | | | |
| Molecular weight | R | R | R | R |
| No of chain atoms | R | R | — | — |
| B.p. of DH | — | — | — | — |
| Density of DH | — | — | — | — |
| π-Donor or acceptor | R | R | R | R |
| No of substituents on FLA | R | R | — | — |
| Average electronegativity of these | — | — | — | — |
| Covalent radius of FLA | — | — | — | — |
| Ionic radius of FLA | — | — | — | — |
| 1st ionization energy of FLA | — | — | — | — |
| Electronegativity of FLA | R | R | — | — |
| Measure of reliability | 72 | 85 | 111 | 56 |
| Measure of correctness, % | 89 | 90 | 83 | 82 |

[a] Ref.[5]; [b] ref.[4].

18% of the variables are descriptors characterizing the bulkiness of the ligands. Recently[4], 28 discriminatory relevant variables (features) have been selected (Table I) by means of the combined criterion based on Wold's discrimination and modelling powers[7] irrespective of their possible dependence. This fact provided a possibility of an additional dimensionality reduction. We found that the sequential approach consisting of the determinations of 1) features in the first step and 2) linearly independent features in the second step can be efficient in the compression of discriminatory information.

In the initial step variables proposed by the hypothesis must be ordered according to their ability to represent correctly the clustering of the objects under consideration into the corresponding classes. The evaluation of the classification ability of individual variables might be performed by any suitable mathematical approach without considering whether the variables are linearly dependent or not. Recently one such evaluation method was used[4] and it is shortly outlined here for the purpose of information. It is based on the combined use of discrimination and modelling powers. The discrimination power (here $P_W^d$) (Eq. (2)) has been formulated by Wold[7] in connection with his SIMCA method based on the disjoint principal components analogy model[7,8] which may be in principle expressed by relation (*1*)

$$y_{ik} = \alpha_i + \sum_{a=1}^{A} \beta_{ia}\theta_{ak} + \varepsilon_{ik}\,, \tag{1}$$

where $y_{ik}$ are experimental data, $\alpha_i$, $\beta_{ia}$ and $\theta_{ak}$ are parameters, $\varepsilon_{ik}$ is residual, $i$, $k$ and $a$ are indexes for the variable, the object and the component and $A$ is the number of components $\beta_i\theta_k$. The discrimination power represents a relevance measure of a given variable $i$ by comparing the variance of the residuals in the cases when all trained patterns (prototypes) are in the "false" classes ($s^{2+}$) with that when the prototypes are in their "own" class ($s^2$).

$$s_i^{2+}/s_i^2 = \sum_{\substack{q=1\\ q\neq r}}^{Q} \sum_{r=1}^{Q} \sum_{k=1}^{n_r} (\varepsilon_{ikr}^{(q)})^2/(Q-1)\sum_{r=1}^{Q} \sum_{k=1}^{n_r} (\varepsilon_{ikr}^{(r)})^2\,, \tag{2}$$

$Q$ is the number of classes and $\varepsilon_{ikr}^{(q)}$ represents the residuals after fitting a $k$-th object of class $r$ to class $q$. Thus the discrimination power represents a measure of the importance of individual variables from the classification point of view. The second mesure – modelling power[4,7] – (Eq. (3*b*)) is based on ratio $U_i$ (Eq. (3*a*)) of the variance of the SIMCA-residuals $\varepsilon$ and the variance of the data $y$ of the training matrix for the variable $i$.

$$U_i = s_{\varepsilon,i}^2/s_{y,i}^2\,. \tag{3a}$$

However, the relevance is indirectly proportional to the value of $U_i$. Therefore, 1-$U_i$, modelling power $P_W^m$ is used having values close to one for highly relevant variables and close to zero for slightly relevant ones.

$$1 - U_i = P_W^m\,. \tag{3b}$$

The modelling power is a measure of the significance of the individual variable for the similarity within the classes. It may also be considered as a difference of the degree of organization (entropy) in a clustered and a non-clustered system. Generally speaking, the Wold's discrimination power $P^d_W$ emphasizes interclass information and the Wold's modelling power $P^m_W$ the intraclass one. For pattern recognition purposes the combined use of both powers as a relevance criterion for feature determination was successfully applied in a man-computer interactive manner[4].

### *The Determination of Linearly Independent Features*

The fundamental characteristic of each system is the minimum number of mutually independent variables which exactly define all objects in the system. The set of linearly independent variables can be represented, geometrically, as an orthogonal basis which defines the dimensionality of the system under study. If the system is composed of classes of similar objects, then the dimensionality estimation for pattern recognition purposes can be carried out more economically using only a set of features preliminarily selected by any appropriate method. Hencefore, for pattern recognition analysis, the intrinsic dimensionality of the system, $D^s$, can be defined as a minimum number of linearly independent features. Real systems are usually very complex and can only be approximated by a model. The quality of the model depends on the number of included features from the total hypothetical set of them as well as on their relative relevancies. Hence, the quality depends on the level of *a priori* knowledge available in the initial stage of the analysis – hypothesis formulation. The intrinsic dimensionality of the model, $D^m$, is thus not generally identical with $D^s$ and it only approximates $D^s$ to some extent.

In chemistry, different techniques have been used for the determination of the number of linearly independent variables by the computation of the rank of data matrix (refs[9,10] and the references cited therein). We use here the technique of the estimation of the number of non-zero eigenvalues, $\lambda$, from the second order moment matrix based on Karhunen–Loève expansion (rotation)[11,12]. The Karhunen–Loève rotation carried out by the Jacobi procedure results in a set of $\lambda$'s which are arranged in an increasing order. However, eigenvalues (the variances in the transformed space) indicate only the number of linearly independent features and do not represent any of them explicitly. The "come-back" into measurement space is not trivial and can be achieved by *e.g.* the rotational part of factor analysis[13]. Here, a simple stepwise deletion of features is used for this purpose. The approach is based on the idea that the deletion of the feature corresponding to the largest component of the eigenvector related to the lowest eigenvalue, $\lambda_{\min}$, minimizes the loss of information. Geometrically, the idea can be demonstrated in three-dimensional space by Fig. 1.

Unfortunately, the Karhunen-Loève approach does not account for the discriminatory effect[14] because it works with the matrix of whole system without considering the clustering. However, in the case of minimum eigenvalues, $\lambda_{\min}$, approaching to zero this "insensivity" does not represent any serious problem because practically no information is lost. It is mostly preserved in the set of features after such a deletion. On the other hand, in the case with $\lambda_{\min} \neq 0$ some information (including discriminatory information) is lost. Hence, for pattern recognition purposes any additional unique deletion of feature is not possible. Nevertheless, a remarkable increase in the percentual loss of information $(100\lambda_{\min}^{(i)}/\sum_{r=1}^{R}\lambda_r^{(i)}$ where $R$ is the dimension after $i$-th deletion) may be considered as an indication of the absence of strongly linearly dependent features. Hencefore, such an increase can be accounted for as a criterion for the determination of the model intrinsic dimensionality $D^m$ as defined above. Naturally, a remarkable deteriorating of classification performance should be expected if any deletion below $D^m$ is carried out. In the present paper, this effect is studied by means of the classification method described elsewhere[6].

## EXPERIMENTAL

### Algorithm

Our approach towards formation of a near optimum set of linearly almost independent features can be summarized by the following algorithm: *1*) Data for the prototypes are autoscaled so that the mean be zero and the variance be one. *2*) The covariance matrix is formed from the autoscaled data. *3*) The rotation from feature space into eigenvalue space is performed by the Jacobi procedure. *4*) $\lambda$'s are arranged according to their values. *5*) $\lambda_{\min}$ is searched. *6*) The eigenvector corresponding to the $\lambda_{\min}$ is picked up. *7*) The largest component of this vector is found. *8*) The feature related to this component is delected. *9*) The cycle is repeated until a significant increase of percentual loss of information occurs.

### Data

In the present study, the input data of complex hydrides are used as summarized in the "Data Base I" (ref.[5]). Analogously to the recent work[4], 115 complex hydrides are selected for the training

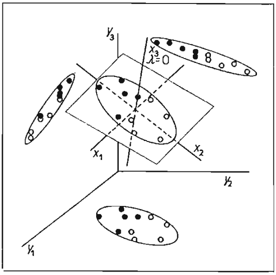

FIG. 1

Projection from Eigenvalue Space into Measurement Space

$y_i$ ($i = 1, 2, 3$) the original coordinates, $x_i$ the coordinates after the Karhunen–Loève rotation, ● objects in class 1, ○ objects in class 2.

procedure from the total set of 224 hydrides — 95 of them as stable prototypes and 20 as unstable ones.

### Computation

Programs were written in the GIER ALGOL-III version of ALGOL-60. All calculations were carried out on the GIER computer in the Computer Centre of the Institute of Nuclear Research, Řež. Some procedures were translated into the machine code.

## RESULTS AND DISCUSSION

Some representative results for dimensionality reduction from 28 features of the MODEL-1 up to its intrinsic dimensionality $D^m = 11$ are summarized in Table I. As expected, the deletion of five features corresponding to the largest components of eigenvectors related to $\lambda = 0$ did not cause any significant changes in the measure of correctness[6] nor in the measure of reliability[6].

Because of rather poor population of prototypes in the class of unstable complex hydrides, additional deletion of features for $\lambda_{\min} \approx 0$ was made. It was performed in a stepwise way according to the increasing order of $\lambda$'s. Such a dimensionality reduction up to $D^m = 11$ gave rise to a very mild decrease (about 7%) in the measure of correctness only, while the measure of reliability even increased to some extent. This counter-intuitive result for the latter measure may provide an additional example of the "peaking phenomenon" discussed recently[15] in connection with independence and dimensionality.

Further deletion of the dimensionality of ten caused a more remarkable loss of information (above 1%) as well as a decrease of the measure of reliability[6] (Fig. 2). We used these facts as an indication for $D^m = 11$ adjustment. It is worth of mention that the measure of correctness[6] remained practically unchanged in this situation

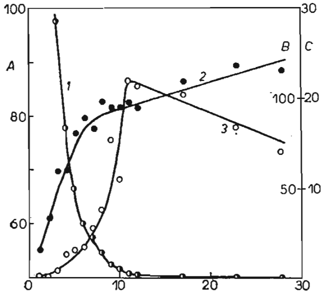

FIG. 2

Relative Loss of Information and the Classification Performance at Different Dimensionalities

*A* the measure of correctness (%) *2*, *B* the measure of reliability *3*, *C* the relative loss of information (%) *1*.

(Table I, Fig. 2). Generally, the preferential use of the measure of reliability will be probably more appropriate for the systems with closely joint classes, where the switching around the discrimination boundary may be caused even by very low errors in the data.

The further deletion up to single dimension gave values limiting to 50% for the measure of correctness and to one for the measure of reliability (Fig. 2). This clearly indicates the nontriviality of the problem under study.

For the structural fragments of the complex hydrides $ABH_nD_{4-n}$ the near optimum set of eleven linearly independent features is shown in Table I. The alkali metal A is represented in the 11-dimensional (reduced) MODEL-1 by the first ionization energy only. The central atom B is characterized by its melting point and density. This reduction is in a good agreement with recently discussed[4] correlation of variables for A and B due to their relation in the periodic system. In monosubstituted hydrides $ABH_3D$ four features for the ligand D are included in the reduced MODEL-1, namely molecular weight, electronegativity of the first atom of the ligand, the number of substituents on this atom and, finally, the number of atoms in the ligand chain. Different sets of features were found for the ligands in di- and trisubstituted derivatives $ABH_2D_2$ and $ABHD_3$. In the former case the ligands are characterized by the number of substituents on the first atom and by $\pi$-donor ability, in the latter case by $\pi$-donor ability and molecular weight. Thus, different sets of features have been found for a ligand in relation to the degree of substitution.

Any quantitative physico-chemical interpretation of the above results is not simple and is not a subject of this study. On the other hand, this introductory study demonstrates that the selection of features in the pattern recognition analysis could serve as a sound decision tool in the treatment of a complex problem.

*The authors would like to acknowledge to Dr K. Eckschlager, Institute of Inorganic Chemistry, Czechoslovak Academy of Sciences, Prague, and Dr K. Kuchynka, J. Heyrovský Institute of Physical Chemistry and Electrochemistry, Czechoslovak Academy of Sciences, Prague, for their interest and to Professor S. Wold, Umeå University, Sweden, Professor P. C. Jurs, Pennsylvania State University, U.S.A., and Professor C. F. Bender, University of California, U.S.A., for their critical comments to the manuscript of this paper.*

## REFERENCES

1. Brugger W. E., Stuper A. J., Jurs P. C.: J. Chem. Inf. Comput. Sci. *16*, 105 (1976).
2. Hodes L.: J. Chem. Inf. Comput. Sci. *16*, 88 (1976).
3. Woodward W. S., Isenhour T. L.: Anal. Chem. *46*, 422 (1974).
4. Štrouf O., Wold S.: Acta Chem. Scand. *A31*, 391 (1977).
5. Štrouf O., Wold S.: *Data Base I: Stability of Complex Hydrides. MODEL-1.* Umeå University, 1977.
6. Fusek J., Štrouf O.: This Journal *44*, 1362 (1979).
7. Wold S.: Pattern Recognition *8*, 127 (1976).

8. Wold S.: *Pattern Cognition and Recognition (Cluster Analysis) Based on Disjoint Principal Components Models.* Techn. Report No 357, March 1974, Univ. Wisconsin.
9. Katakis D.: Anal. Chem. *37*, 876 (1965).
10. Varga L. P., Veatch F. C.: Anal. Chem. *39*, 1101 (1967).
11. Fu K. S.: *Sequential Methods in Pattern Recognition and Machine Learning*, p. 29. Academic Press, New York 1965.
12. Andrews H. C.: *Mathematical Techniques in Pattern Recognition*, p. 24. Wiley — Interscience, New York 1972.
13. Weiner H., Malinowski E. R., Levinstone A. R.: J. Phys. Chem. *74*, 4537 (1970).
14. Kittler J.: IEEE Trans. Computers *26*, 604 (1977).
15. Chandrasekaran B., Jain A. K.: IEEE Trans. Systems, Man and Cybernetics *5*, 240 (1975).