match no. 31 pp. 7-21 1994

Cluster–of–variables decomposition, a new mathematical tool for analysing structure–property relations

Werner Hässelbarth

Bundesanstalt für Materialforschung und -prüfung (BAM)
12200 Berlin
(received: February 1994)

Abstract

A method of data analysis is presented, designed for the interpretation of molecular property data in terms of additive increments relating to molecular fragments. It is based on a mathematical theory according to which any function of several discrete variables can be decomposed into a sum of unique orthogonal contributions, one from each cluster of variables. These mathematical results provide a consistent scheme for defining and investigating the contributions from molecular fragments to a given molecular property.

1 Introduction

This paper presents a survey of a mathematical toolkit for use in the field of quantitative structure–property relations. More specifically, its scope is the analysis of experimental data on a molecular property, in terms of (additive) increments relating to molecular fragments.

The typical application that we have in mind would be the development of an increment system for the study of multiple substituent effects. Consider a class of chemical compounds derived from a common parent compound by multiple substitution. The basic idea of the method is to model a (quantitative) property of the derivatives by a sum of contributions due to the various substituents, followed by sums of contributions due to interactions between pairs of substituents, triples of substituents, and so on.

For this purpose, a mathematical method has been developed in [2], and partly published in [3], by which it is possible to decompose any real-valued function of several discrete variables into a sum of unique orthogonal contributions, one from each cluster of variables. Truncated expansions obtained by restriction of cluster

size yield best approximations in the least squares sense. Certain linear identities provide means to test whether such truncation is numerically sufficient.
In the picture of multiple substituent effects used above, these mathematical results provide

(i) an empirical definition of the contribution of a molecular fragment, as specified by a set of substitution positions and the interactions among the substituents sitting there

(ii) means to estimate the degree of complexity of those fragments that have to be included when using a truncated cluster expansion as an approximation.

A summary of the mathematical derivations and characterisations developed in [2,3] is given in an annex.

2 The basic procedure

As an introductory example, consider the class of benzene derivatives with substituents -H, -CH_3, -OH, -NO_2, or -Cl at position 1, 2, and 4, and hydrogen everywhere else. Any such derivative is uniquely represented by a triple (X, Y, Z), where X, Y, and Z specify the type of substituent at position 1, 2, and 4, respectively. Given a quantitative property F, its value for the derivative (X, Y, Z) is denoted by $F(X, Y, Z)$.
Suppose now to be given the complete collection of data $F(X, Y, Z)$ – measured values, as a rule – of some property F under investigation, for all these compounds, and to be asked to analyse the relationship between the variation of property values and the variation of molecular structure.
Adopting the familiar approach of additivity schemes, we try to model the property values $F(X, Y, Z)$ by a sum $f_1(X) + f_2(Y) + f_4(Z)$ of contributions due to the individual substituents. Note that in this "Ansatz" a substituent may contribute differently, depending on its position, 1, 2, or 4. This feature is of immediate importance in the case of a non-symmetrical parent compound, but it is also needed in the benzene case, e.g. when analysing substituent effects on ^{13}C-NMR spectra. If we decide to use the familiar least squares error measure, the following problem has to be solved: Determine three parameters $f_1(X), f_2(X), f_4(X)$ for each of the five substituent types, X = H, CH_3, OH, NO_2, Cl, (that is, altogether 15 parameters), such that

$$\sum_X \sum_Y \sum_Z [F(X, Y, Z) - f_1(X) - f_2(Y) - f_4(Z)]^2 = \text{minimum.}$$

Let us assume that this approximation problem has been solved. If we were lucky to pick an almost additive quantity, the residuals will be negligible, and we are ready, ending up with a simple scheme of additive substituent effects.
Otherwise there will be a more or less considerable residual term,

$$G(X, Y, Z) = F(X, Y, Z) - f_1(X) - f_2(Y) - f_4(Z),$$

that needs to be analysed further. Continuing the approach of additivity schemes by interactions terms, we try to model this residue by a sum of contributions due to interactions between pairs of substituents, $f_{12}(X,Y) + f_{14}(X,Z) + f_{24}(Y,Z)$. Using again the least squares error measure, the problem is to determine three parameters $f_{12}(X,Y)$, $f_{14}(X,Y)$, and $f_{24}(X,Y)$ for each pair (X,Y) of substituent types (altogether $3 \times 5 \times 5 = 75$ parameters) such that

$$\sum_X \sum_Y \sum_Z [G(X,Y,Z) - f_{12}(X,Y) - f_{14}(X,Z) - f_{24}(Y,Z)]^2 = \text{minimum.}$$

If after that there should still be a considerable residue, this would be interpreted as an "indecomposable" 3-way interaction term, that is, a contribution due to interactions between all three substituents at a time.

Given that also the second approximation problem has been solved, we end up with an expansion of $F(X,Y,Z)$ into a sum of contributions from clusters of substituents of increasing size: contributions from single substituents, followed by contributions from pairs of substituents, and finally from triples. Amazingly, these highdimensional nested approximation problems admit a very simple closed form solution, involving only averaging operations.

For an optimum expression of this result, we introduce a constant term f_0 as the level-0 approximation to $F(X,Y,Z)$. That is, the function $F(X,Y,Z)$ is expanded as follows:

$$\begin{aligned} F(X,Y,Z) = \; & f_0 && (\text{level}-0) \\ + \; & f_1(X) + f_2(Y) + f_4(Z) && (\text{level}-1) \\ + \; & f_{12}(X,Y) + f_{14}(X,Z) + f_{24}(Y,Z) && (\text{level}-2) \\ + \; & f_{124}(X,Y,Z) && (\text{level}-3) \end{aligned}$$

In this expansion the different terms are successively defined as best approximations in the least squares sense as follows:

Level-0
f_0 = best approximation to $F(X,Y,Z)$ by a constant.

Level-1
$f_1(X) + f_2(Y) + f_4(Z)$ = best approximation to the level-0 residue by a sum of single substituent terms.

Level-2
$f_{12}(X,Y) + f_{14}(X,Z) + f_{24}(Y,Z)$ = best approximation to the level-1 residue by a sum of pairwise interaction terms.

Level-3
$f_{124}(X,Y,Z)$ = level-2 residue, that is, the ultimate residue.

The solution parameters $f_0, f_i(X)$, and $f_{ij}(X,Y)$ are obtained as averages of the original property values over subsets of derivatives with partly fixed and partly

varying substituents, as follows:

$$f_0 = \langle F(X,Y,Z)\rangle_{XYZ}$$

$$\begin{aligned} f_1(X) &= \langle F(X,Y,Z)\rangle_{YZ} - f_0 \\ f_2(Y) &= \langle F(X,Y,Z)\rangle_{XZ} - f_0 \\ f_4(Z) &= \langle F(X,Y,Z)\rangle_{XY} - f_0 \end{aligned}$$

$$\begin{aligned} f_{12}(X,Y) &= \langle F(X,Y,Z)\rangle_{Z} - f_1(X) - f_2(Y) - f_0 \\ f_{14}(X,Z) &= \langle F(X,Y,Z)\rangle_{Y} - f_1(X) - f_4(Z) - f_0 \\ f_{24}(Y,Z) &= \langle F(X,Y,Z)\rangle_{X} - f_2(Y) - f_4(Z) - f_0 \end{aligned}$$

$$\begin{aligned} f_{124}(X,Y,Z) = {} & F(X,Y,Z) - f_{12}(X,Y) - f_{14}(X,Z) - f_{24}(Y,Z) \\ & - f_1(X) - f_2(Y) - f_4(Z) - f_0 \end{aligned}$$

In these expressions brackets denote averaging over the subscript variables, so that e.g.

$$\begin{aligned} \langle F(X,Y,Z)\rangle_{XYZ} &= 5^{-3} \cdot \textstyle\sum_X \sum_Y \sum_Z F(X,Y,Z), \\ \langle F(X,Y,Z)\rangle_{YZ} &= 5^{-2} \cdot \textstyle\sum_X \sum_Y F(X,Y,Z), \\ \langle F(X,Y,Z)\rangle_{Z} &= 5^{-1} \cdot \textstyle\sum_Z F(X,Y,Z), \end{aligned}$$

where each sum runs over the five substituent types H, CH_3, OH, NO_2, Cl.
Using these formulas, the various cluster components can be most easily determined directly from the collection of property data.
Which now is the benefit that can be expected from such exercise? By this procedure the original data file is recast into another format, designed for a consistent interpretation of data in terms of substituent effects. By comparing the cluster components, information can be extracted about the effects due to the types of substituents, their position, and their interactions.
Up to here we have been using the term *cluster expansion*, while the title contains the term *cluster decomposition*. This wording takes account of the fact that, while starting with the idea of an expansion according to a scheme of successive approximations, the final result is, in fact, a decomposition into a sum of mutually independent terms, one for each cluster of substituents.

3 Extensions and supplements

The procedure outlined above can be generalized to arbitrary numbers of substitution positions and substituent types. Moreover, it is not restricted to the study of substituent effects. It can, in fact, be used to analyse the effect of variations of molecular structure on property data in every case where the variations of structure are of multivariate type, i.e. if the structure variations can be parametrized - not necessarily in a one-to-one fashion – by a set of independent variables with finite range.
The mathematical basis of the method has been elaborated, in a linear algebraic framework, in ref. [3]. The mathematical object studied there is the Euclidean vector space formed by the real-valued functions of a finite number of independent

discrete variables. It turns out that this space decomposes into a series of orthogonal subspaces, one for each cluster of variables. The functions in such a cluster subspace have two characteristic properties: their values do not depend on the variables outside of the cluster, and they average to zero in any of the variables inside the cluster. Owing to this structure, any function F of the given variables, say $x_1, x_2, ..., x_n$, admits a unique cluster decomposition as follows:

$$F(x_1, x_2, ..., x_n) = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) +$$

$$+ \sum_{i<j<k} f_{ijk}(x_i, x_j, x_k) + ... + f_{1...n}(x_1, x_2, ..., x_n).$$

The cluster components are obtained as orthogonal projections of the given function F onto the various cluster subspaces. The corresponding projection maps can be explicitely constructed. Like in the preceding example, they involve nothing more than averaging of function values over subsets of variables. The relevant mathematical constructions are given in section A.1 of the annex.
After this brief excursion into mathematical technicalities, let us return to the example of benzene derivatives in order to illustrate other features and uses of the mathematical results developed in refs. [2] and [3].

3.1 Problems relating to incomplete data files

Unfortunately the method requires complete data files. For the property to be analysed the data must be available for all the derivatives that can be built by placing substituents of a given collection of types to a given choice of substitution positions. In the introductory example these are 125 compounds. In practice it is an extremely difficult task to acquire complete property data files like this.
If the data file acquired is incomplete, but with only a moderate number of blanks, the method of cluster decomposition can still be used on a heuristical basis. In such a case some cluster terms cannot be evaluated rigorously, because of missing terms in the averages. They can, however, be estimated by extending the averages over those data that are available. Using this approach, it is possible to analyse the available data, and, at the same time, estimate the missing data. However, we should like to emphasize that the scope of this method primarily is data analysis and data interpretation rather than data estimation.

3.2 Significance tests for maximum cluster size

For the purpose of analysing incomplete data files it would be useful to have prior information about the possibility to truncate the cluster expansion by restriction of cluster size. In the case of benzene derivatives, for example, one should like to know whether the 3-way interaction term is numerically significant, or whether even the pair interaction terms can be omitted without any serious loss of accuracy.
This can be tested by using linear identities, as follows. Let us call a property k-representable, if it can be rigorously expressed as a sum of terms due to clusters of

maximum size k. For example a property $F(X,Y,Z)$ of triply substituted benzene derivatives would be 1-representable, if it could be expressed as a sum of single substituent terms,

$$F(X,Y,Z) = a(X) + b(Y) + c(Z),$$

and it would be 2-representable if it could be expressed as a sum of terms relating to pairs of substituents,

$$F(X,Y,Z) = d(X,Y) + e(X,Z) + f(Y,Z).$$

In these expressions, we have only used maximum size cluster terms because they incorporate all the terms relating to clusters of smaller size.
It can be shown by projection methods (comp. [3]) that a function $F(X,Y,Z)$ is 1-representable if and only if the following identity holds for all values of X,Y,Z and an arbitrary value of V.

$$F(X,Y,Z) = -2F(V,V,V) + F(X,V,V) + F(V,Y,V) + F(V,V,Z).$$

Similarly, $F(X,Y,Z)$ is 2-representable if and only if

$$\begin{aligned} F(X,Y,Z) &= F(V,V,V) - F(X,V,V) - F(V,Y,V) - F(V,V,Z) \\ &+ F(X,Y,V) + F(X,V,Z) + F(V,Y,Z). \end{aligned}$$

Spot checking of these identities – i.e. how badly they are broken – will provide a preliminary impression of the magnitude of errors to be expected when using a truncated cluster expansion as an approximation.
Section A.2 of the annex describes a general procedure for constructing such testing identities for any given complexity k. In fact, these identities can largely be tailored to meet the restrictions encountered in the highly incomplete data sets that are typically available in practice.
Identities of this type, relating to additivity schemes, have been discussed in the physico-chemical literature quite a number of times, e.g. by Bernstein in his investigation on relations between molecular properties in homologous series [1], by Kauzmann, Clough and Tobias in the presentation of their "Principle of pairwise interactions" [4] in molecular chirality, and in the theory of "Chirality functions" due to Ruch and Schönhofer [5].

3.3 Symmetry properties

The cluster decomposition is fully adapted to permutation symmetry. That is, if the function $F(X,Y,Z)$ has some symmetry with respect to permutations of variables, any sum of cluster terms of a given size automatically transforms alike.
As a consequence, there are characteristic relations between symmetry-equivalent cluster terms. Consider, e.g., (1,3,5)-tri-substituted benzene derivatives. Due to the symmetrical arrangement of the substitution positions, a scalar property $F(X,Y,Z)$ is invariant under all permutations of X,Y, and Z. This symmetry has a number of consequences for the cluster terms, as follows.

$$f_1(X) = f_3(X) = f_5(X),$$
$$f_{13}(X,Y) = f_{13}(Y,X) = f_{15}(X,Y) = f_{15}(Y,X) = f_{35}(X,Y) = f_{35}(Y,X).$$

That is, the contribution of a substituent does not depend on its position, and neither do the contributions due to pairwise interactions.
In any given case, the complete set of symmetry relations for cluster components can be constructed using group theory, similar to the construction of selection rules for quantum mechanical matrix elements. The relevant tools are summarized in section A.3 of the annex.
Symmetry based relations like this, with emphasis on pseudoscalar (i.e. chiral) properties, have been extensively studied by Ruch and Schönhofer, compare, e.g., ref. [5].

3.4 Interpretation relating to average substituents

The cluster expansion may be loosely interpreted as an expansion with respect to the deviations of the substituents from their average, as follows. Suppose that an "average substituent type" Θ could be found, such that any average of $F(X,Y,Z)$ over a subset of variables coincides with the corresponding single value for the derivative with Θ in these positions. That is, e. g.,

$$\begin{aligned} \langle F(X,Y,Z)\rangle_{XYZ} &= F(\Theta,\Theta,\Theta), \\ \langle F(X,Y,Z)\rangle_{XY} &= F(\Theta,\Theta,Z), \\ \langle F(X,Y,Z)\rangle_{X} &= F(\Theta,Y,Z). \end{aligned}$$

Then the cluster decomposition could be expressed as follows.

$$\begin{aligned} F(X,Y,Z) = {} & F(\Theta,\Theta,\Theta) - F(X,\Theta,\Theta) - F(\Theta,Y,\Theta) - F(\Theta,\Theta,Z) + \\ & F(X,Y,\Theta) + F(X,\Theta,Z) + F(\Theta,Y,Z) + \text{residue} \end{aligned}$$

With an arbitrary "real" substituent type V instead of the hypothetical Θ, the values of $F(X,Y,Z)$ can be expressed analogously. This might be viewed as an expansion with respect to the deviations of substituents from a given standard type, say hydrogen. Such procedure, though lacking the benefit of optimality with respect to the least squares error measure, may nevertheless provide a useful tool for rationalizing and interpreting structure–property relations.

4 Summary

The mathematical procedures for data analysis presented in this contribution are applicable to any collection of quantitative property data for any set of compounds which can be parametrized by a set of independent parameters, such as a complete set of derivatives of a given parent compound based on a specified set of substituent types. In the case of a complete set of compounds the result is a set of increments, one for each cluster of the molecular fragments considered, intended to serve as a basis for developing an understanding, in molecular science terms, on how the property under investigation depends on the nature of the molecular fragments considered, their spatial arrangement, and their interactions.

In the case of an incomplete data set, the methods can be used for the same purpose, with some loss of mathematical benefits. In addition, they can be used as an interpolation scheme for estimating the missing data.
The degree of cluster complexity that is necessary for an appropriate representation of the data under investigation can be assessed using specific identities, one for each cluster size.
If the parent molecular structure posesses some symmetry, and if the property under investigation transforms according to an irreducible representation of the corresponding symmetry group, the number of cluster components is greatly reduced by symmetry relations. These relations can be constructed systematically, using similar tools like in the construction of symmetry based selection rules for quantum mechanical matrix elements.

Annex: Mathematical framework and main results

A.1 Cluster decomposition of the property space

Consider a composite system Σ, built up from a finite number p of subsystems, each with a finite state space[1]. Let the subsystems be labelled $1, 2, ..., p$, and let $S_1, S_2, ..., S_p$ denote their state spaces. Let the composite system Σ be such that each of its states is completely characterized by specifying the states of all the subsystems. Then its state space is the Cartesian product $\Omega := S_1 \times S_2 \times ... \times S_p$ of the state spaces of its components.
The elements of Ω will be denoted by small Greek letters $\mu, \sigma, \ldots$ that is, $\mu := (\mu_1, \mu_2, ..., \mu_p)$ denotes a p-tuple of subsystem states $\mu_i \in S_i$ $(i = 1, 2, ..., p)$.
With $\Omega := S_1 \times S_2 \times \ldots \times S_p$ taking the part of the state space of a composite system Σ, real-valued functions $F : \Omega \to \mathbf{R}$ are readily interpreted as (real number valued) properties of the system in the sense that for $\mu \in \Omega$ the number $F(\mu)$ is the numerical value of the property F, measured on an appropriate scale, for the system in its state μ. We may, e. g., consider F to represent a measuring apparatus, and $F(\mu)$ to be the result of the corresponding measurement performed on the system Σ in its state μ.
The main result of this section will be a decomposition,

$$F = \sum_{Q \leq P} f_Q \qquad (A.1)$$

of any property F into a sum of (mutually orthogonal) components f_Q, one for each cluster Q of subsystems, that is, for each subset Q of the set $P := \{1, 2, ..., p\}$ of subsystems of Σ. By virtue of this decomposition, any property of the compound system is split up into a sum of contributions due to its clusters of subsystems.

[1]The terms finite space and finite set are used synonymously.

Rewritten in the form

$$F = f_0 + \sum_{i}^{1,p} f_i + \sum_{i<j}^{1,p} f_{ij} + \sum_{i<j<k}^{1,p} f_{ijk} + ... \qquad (A.2)$$

the cluster decomposition may be interpreted as an expansion into a sum of individual contributions f_i of the subsystems, followed by corrections f_{ij}, f_{ijk}, ... due to interactions of increasing complexity: interactions between pairs of subsystems, triples, etc. In this sum f_0 denotes a constant term (the grand average) which is associated with the empty subset of P.
The object that we are now going to investigate is the property space of Ω, that is, the set $X := \mathbf{R}^{\Omega}$ of all real-valued functions on Ω. Elements of X will be denoted by Latin letters $F, G, ...$ or $f, g, ...$ With addition of functions and multiplication by real numbers defined pointwise in the usual manner, X becomes a vector space over $\mathbf{R}$ of dimension $|\Omega| = s_1 \cdot s_2 \cdot ... \cdot s_p$, where $s_i = |S_i|$. Moreover, X may be endowed with the customary scalar product,

$$< F, G >:= \sum_{\mu \in \Omega} F(\mu) G(\mu) \qquad (A.3)$$

turning it into an Euclidean vector space.
Let us now consider, for each subset[2] $Q \leq P$, the subset $X_Q \leq X$ of those properties which only depend on the state of the cluster Q, that is, which are independent of the state of the complementary cluster $P \backslash Q$.

$$X_Q := \{ f \in X : \mu_Q = \sigma_Q \Rightarrow f(\mu) = f(\sigma) \text{ for all } \mu, \sigma \in \Omega \} \qquad (A.4)$$

In this expression the following notation is used: Let $\mu \in \Omega$, and let Q be a subset of $P = \{1, 2, ..., p\}$ with q elements ($0 \leq q \leq p$). Then μ_Q denotes the q-tuple derived from μ by restriction to the subsystems $i \in Q$, that is, by restricting the domain of the mapping represented by μ to $Q \leq P$.
Evidently, all the X_Q are linear subspaces of X. In particular, $X_P = X$, and $X_\emptyset$ ($\emptyset$ = the empty subset) is the 1-dimensional subspace of constant functions. There are simple relations between the algebra of these subspaces of X and the algebra of subsets of P, like the following:

$$X_Q \cap X_R = X_{Q \cap R} \qquad (A.5)$$

For the projection mappings onto these subspaces, the natural candidates are averaging operators. To this end we consider the following objects: For any $i \in P$, let $\mathcal{A}_i : X \to X$ be the mapping given by

$$[\mathcal{A}_i F](\mu_1, \mu_2, ..., \mu_p) := s_i^{-1} \cdot \sum_{x \in S_i} F(\mu_1, ..., \mu_{i-1}, x, \mu_{i+1}, ..., \mu_p) \qquad (A.6)$$

Thus $\mathcal{A}_i$ takes the average over the states of the i-th subsystem.

[2]For typographical reasons the symbols $<$ and $\leq$ are (mis)used to denote the proper and general subset relationship.

Using these (mutually commutative) operators we may associate with any cluster $Q \leq P$ a mapping $\mathcal{B}_Q : X \to X_Q$ as follows:

$$\mathcal{B}_Q := \prod_{i \in P \setminus Q} \mathcal{A}_i \tag{A.7}$$

That is, $\mathcal{B}_Q$ takes the average over the states of the complementary cluster $P \setminus Q$.
It is easily verified that $\mathcal{B}_Q$ is a linear mapping onto X_Q, which is symmetric with respect to the scalar product defined in the beginning, and idempotent. Hence $\mathcal{B}_Q$ is the unique orthogonal projector onto X_Q. In addition the following relation holds:

$$\mathcal{B}_Q \cdot \mathcal{B}_R = \mathcal{B}_{Q \cap R} \tag{A.8}$$

In view of the intention to attribute with any cluster a specific contribution to a given property, these subspaces and projectors are not quite what we are looking for. The reason is that a subspace $X_Q \leq X$ contains all the subspaces X_R for any subset $R \leq Q$. Therefore, given a property F and a cluster Q, the cluster contribution $\mathcal{B}_Q F$ incorporates contributions from every cluster contained in Q.
To meet these needs, we consider another series of subspaces, $Y_Q \leq X$, defined as follows:

$$Y_Q := \{F \in X_Q : \mathcal{B}_R F = 0 \text{ for all } R < Q\} \tag{A.9}$$

So Y_Q is the orthogonal complement of all the subspaces X_R which are contained in X_Q. From the definition given above, the following relations, valid for any two subsets S, T of X, are easily proved

(i) Y_S and Y_T are orthogonal unless $S = T$

(ii) Y_S and X_T are orthogonal unless $S < T$

From this it follows that, for any $Q \leq X$, the subspace X_Q is the direct sum of the mutually orthogonal subspaces $Y_R, R \leq Q$.

$$X_Q = \bigoplus_{R \leq Q} Y_R \tag{A.10}$$

The orthogonal projectors, $\mathcal{C}_Q$, onto the subspaces Y_Q are obtained from the $\mathcal{B}_Q$, using Moebius inversion, as follows:

$$\mathcal{C}_Q = \sum_{R \leq Q} (-1)^{q-r} \mathcal{B}_R \tag{A.11}$$

where $q = |Q|$ and $r = |R|$.
Alternatively, the $\mathcal{C}_Q$ may be expressed as follows:

$$\mathcal{C}_Q := \prod_{i \in P \setminus Q} \mathcal{A}_i \cdot \prod_{k \in Q} (1 - \mathcal{A}_k) \tag{A.12}$$

Summarizing these results, we have

Theorem 1 *The cluster spaces Y_Q ($\emptyset \leq Q \leq P$) defined in (A.9) constitute a decomposition of the property space X into a direct sum of mutually orthogonal subspaces,*

$$X = \bigoplus_{Q \leq P} Y_Q.$$

In the corresponding decomposition of an arbitrary function $F \in X$,

$$F = \sum_{Q \leq P} f_Q$$

the cluster components are given by $f_Q = \mathcal{C}_Q F$, where the $\mathcal{C}_Q$ are the projection mappings defined in (A.11) or (A.12).

A.2 Significance tests for maximum cluster size

For any integer $0 \leq k \leq p$, let us call a property $F \in X$ to be k-representable if it is a sum of contributions from clusters Q of size $q \leq k$. Formally this means that F is k-representable if and only if $F \in X^{(k)}$, where $X^{(k)}$ is the sum of all cluster subspaces X_Q for clusters Q of size $q \leq k$.

$$X^{(k)} = \sum_{Q \leq P,\ q \leq k} X_Q \tag{A.13}$$

Replacing the X_Q by Y_Q, this sum of subspaces is turned into a direct one:

$$X^{(k)} = \bigoplus_{Q \leq P,\ q \leq k} Y_Q \tag{A.14}$$

The decomposition above implies that the orthogonal projector onto $X^{(k)}$ is given by

$$\mathcal{C}^{(k)} = \sum_{Q \leq P,\ q \leq k} \mathcal{C}_Q \tag{A.14}$$

This result provides a test of k-representability as follows:

$$F \in X^{(k)} \Longleftrightarrow F = \mathcal{C}^{(k)} F \tag{A.15}$$

with $\mathcal{C}^{(k)}$ given above.

The right hand side of (A.15) constitutes a family of linear identities of the form

$$\sum_{\sigma \in \Omega} a(\mu, \sigma) F(\sigma) = 0 \quad \text{for all } \mu \in \Omega \tag{A.16}$$

where the coefficients $a(\mu, \sigma)$ will be nonzero for almost all pairs μ, σ in general. There are, however, much "shorter" identities for the same purpose, as indicated in section 3.2. The approach chosen in ref. [3] for constructing identities like that starts from the observation that (A.15) holds true for *any* projection map onto the subspace $X^{(k)}$ in place of the unique perpendicular projector $\mathcal{C}^{(k)}$.

In search of other projectors onto $X^{(k)}$, let us fix an arbitrary reference state $\delta \in \Omega$ and define a family of mappings $\mathcal{D}_Q : X \to X_Q$ as follows:

$$[\mathcal{D}_Q F](\mu) = F(\mu_Q + \delta_{P\setminus Q}) \tag{A.17}$$

where the p-tuple $\mu_Q + \delta_{P\setminus Q}$ is defined as follows

$$(\mu_Q + \delta_{P\setminus Q})_i := \begin{cases} \mu_i & \text{for } i \in Q \\ \delta_i & \text{for } i \in P\setminus Q \end{cases} \tag{A.18}$$

For any $Q \leq P$ the mapping $\mathcal{D}_Q$ from X to X_Q is linear, idempotent, and surjective, hence a projector onto X_Q. Moreover, these maps multiply according to

$$\mathcal{D}_Q \cdot \mathcal{D}_R = \mathcal{D}_{Q\cap R} \tag{A.19}$$

thus exhibiting close analogy with the mappings $\mathcal{B}_Q$ discussed in the previous section. Indeed, the theory developed there can be copied (cf. ref. [3]), giving essentially analogous results (that is, except for the loss of orthogonality of the subspaces involved). In particular, corresponding projection maps $\mathcal{E}^{(k)}$ onto the subspaces $X^{(k)}$ are obtained as follows:

$$\mathcal{E}^{(k)} = \sum_{Q\leq P,\ q\leq k} \mathcal{E}_Q \tag{A.20}$$

where the $\mathcal{E}_Q$ are obtained from the $\mathcal{D}_Q$ by Moebius inversion, like in the previous section, eq. (A.11).

$$\mathcal{E}_Q = \sum_{R\leq Q} (-1)^{q-r} \mathcal{D}_R \tag{A.21}$$

with $q = |Q|$ and $r = |R|$.
In consequence

$$F \in X^{(k)} \Longleftrightarrow F = \mathcal{E}^{(k)} F \tag{A.22}$$

with $\mathcal{E}^{(k)}$ given above.
After some arithmetic one arrives at another expression for the projection maps $\mathcal{E}^{(k)}$ which provides an efficient criterion for k-representability as follows

Theorem 2 *A function $F \in X$ is k-representable $(0 \leq k \leq p-1)$ if and only if*

$$F(\mu) = \sum_{q=0}^{k} (-1)^{q-k} \binom{p-q-1}{k-q} \sum_{Q\leq P, |Q|=q} F(\mu_Q + \delta_{P\setminus Q}) \tag{A.23}$$

for all states $\mu \in \Omega$ and an arbitrary reference state $\delta \in \Omega$.

The examples in section 3.2 are constructed using a uniform reference state $\delta = (V, V, V)$. It should be noted that the reference state δ can be chosen arbitrarily. This freedom of choice permits the construction of testing identities which are tailored to a given incomplete data file.

A.3 Symmetry properties

The cluster decomposition of a function $F \in X$ is fully adapted to permutation symmetry. That is, if the function F has some symmetry with respect to permutations of its variables, then for any integer $0 \leq q \leq p$ the sum of cluster terms $f_Q = \mathcal{C}_Q F$ for all clusters Q of size q automatically transforms alike. As a consequence, the individual cluster terms bear some inherited symmetry properties, in particular some cluster terms may vanish for symmetry reasons, and there are characteristic relations between different symmetry equivalent cluster terms.
For an example, let $F = F(x, y, z)$ be a function of three variables, and let its cluster decomposition be expressed as

$$\begin{aligned} F(x,y,z) &= a_0 + b_1(x) + b_2(y) + b_3(z) + \\ &\quad c_{12}(x,y) + c_{23}(y,z) + c_{31}(z,x) + d_{123}(x,y,z). \end{aligned}$$

Moreover, let the variables x, y, z range through the same finite set S.

1) Let F be totally symmetric, i.e. symmetric with respect to permutations of all its variables x, y, z. Then

   $b_1 = b_2 = b_3 = b$

   $c_{12} = c_{23} = c_{31} = c$ where $c(y, x) = c(x, y)$

   d_{123} is totally symmetric

2) Let F be totally antisymmetric, i.e. antisymmetric with respect to permutations of all its variables x, y, z. Then

   $a_0 = 0$

   $b_1 = b_2 = b_3 = 0$

   $c_{12} = c_{23} = c_{31} = c$ where $c(y, x) = -c(x, y)$

   d_{123} is totally antisymmetric

3) Let F be symmetric with respect to permutations of x and y. Then

   $b_1 = b_2$

   $c_{12}(y, x) = c_{12}(x, y)$ and $c_{23}(x, y) = c_{31}(y, x)$

   d_{123} is symmetric in x and y

4) Let F be antisymmetric with respect to permutations of x and y. Then

   $a_0 = 0$

   $b_1 = -b_2$ and $b_3 = 0$

   $c_{12}(y, x) = -c_{12}(x, y)$ and $c_{23}(x, y) = -c_{31}(y, x)$

   d_{123} is antisymmetric in x and y

The systematic derivation of such symmetry relations proceeds as follows. Let all the variables μ_i have the same range S, that is, $\Omega = S \times S \times ... \times S$ (p copies). Let G be an arbitrary permutation group on $P = \{1, 2, ..., p\}$. Then this group acts on the state space Ω by permutations of the variables μ_i as follows: For any permutation $g \in G$ let gi denote the image of an index $i \in P$, and let $g\mu$ denote the image of a p-tuple $\mu \in \Omega$. Then

$$[g\mu]_{gi} := \mu_i \tag{A.24}$$

This action on Ω, in turn, induces an analogous action on the property space $X = \mathbf{R}^\Omega$

$$[gF](g\mu) := F(\mu) \tag{A.25}$$

Using this action, the following relationship results, where gQ denotes the image of a subset $Q \leq P$ under $g \in G$.

$$g\mathcal{B}_Q F = \mathcal{B}_{gQ} gF \tag{A.26}$$

The very same relationship holds for the perpendicular projections $\mathcal{C}_Q$ in place of the $\mathcal{B}_Q$,

$$g\mathcal{C}_Q F = \mathcal{C}_{gQ} gF \tag{A.27}$$

Now let O be a G-orbit of subsets of P, that is, a set of symmetry equivalent clusters, and let $\mathcal{C}^O$ denote the corresponding sum of projectors,

$$\mathcal{C}^O = \sum_{Q \in O} \mathcal{C}_Q \tag{A.28}$$

Then $g\mathcal{C}^O F = \mathcal{C}^O gF$ holds for any $g \in G$ and any $F \in X$. From this it follows that, for any $F \in X$, $\mathcal{C}^O F$ transforms like F under the group G. The same is true for any G-invariant set of clusters, that is, for a union of orbits instead of a single one. Thus it holds in particular for the set of clusters of a given size.
For the rest of this section let the function F transform according to a 1-dimensional real representation of the group G, i.e.,

$$gF = \tau(g)F \text{ with } \tau(g) = \pm 1. \tag{A.29}$$

Now let $f_Q = \mathcal{C}_Q F$ be an arbitrary term in the cluster decomposition of F, and let f_{gQ} be a symmetry equivalent term, where $g \in G$. Then f_{gQ} is related to f_Q by

$$f_{gQ} = \tau(g) g f_Q \tag{A.30}$$

or, more explicitely,

$$[f_{gQ}](\mu) = \tau(g)[f_Q](g^{-1}\mu) \tag{A.31}$$

Besides such relations between symmetry equivalent cluster terms, the cluster terms f_Q have intrinsic symmetry properties, inherited from those of the function F, as follows.

$$[f_Q](\mu) = \tau(g)[f_Q](g^{-1}\mu) \text{ for any } g \in G_Q \tag{A.32}$$

$$[f_Q](\mu) = \tau(g)[f_Q](\mu) \text{ for any } g \in G'_Q \tag{A.33}$$

Here G_Q and G'_Q denote the set-wise and the point-wise stabilizers of Q, respectively, i.e. $G_Q := \{g \in G : gi \in Q \text{ for all } i \in Q\}$ and $G'_Q := \{g \in G : gi = i \text{ for all } i \in Q\}$. These relations imply that f_Q transforms according to the representation of the set-wise stabilizer of Q subduced from that of G, and that $f_Q = 0$ unless the restriction of that representation to the point-wise stabilizer of Q is the identy.
Analogous relations can be derived for cases where F transforms according to other irreducible representations. Again these will be intrinsic symmetry properties inherited from those of the given F, and relations between terms from symmetry equivalent clusters. In the case of representations of dimension ≥ 2, these relations will incorporate, besides the given F, its partner(s) which together span a G-invariant subspace of X, that is, a representation space of the group G.

References

[1] H. J. Bernstein, J. Chem. Phys. **20**, 263 (1952)

[2] W. Hässelbarth, Habilitationsschrift, Berlin 1982

[3] W. Hässelbarth, J. Math. Phys. **25**, 828 (1984)

[4] W. Kauzmann, F.B. Clough, I. Tobias, Tetrahedron **13**, 57 (1961)

[5] E. Ruch, A. Schönhofer, Theoret. Chim. Acta **19**, 225 (1970)