Published in final edited form as:

J Chromatogr A. 2017 November 10; 1523: 148–161. doi:10.1016/j.chroma.2017.06.036.

Orthogonality measurements for multidimensional chromatography in three and higher dimensional separations

Mark R. Schure and

Theoretical Separation Science Laboratory, Kroungold Analytical, Inc., 1299 Butler Pike, Blue Bell, Pennsylvania 19422

Joe M. Davis¹

Department of Chemistry and Biochemistry, Southern Illinois University at Carbondale, Carbondale, IL 62901-4409 USA

Abstract

Orthogonality metrics (OMs) for three and higher dimensional separations are proposed as extensions of previously developed OMs, which were used to evaluate the zone utilization of two-dimensional (2D) separations. These OMs include correlation coefficients, dimensionality, information theory metrics and convex-hull metrics. In a number of these cases, lower dimensional subspace metrics exist and can be readily calculated. The metrics are used to interpret previously generated experimental data. The experimental datasets are derived from Gilar's peptide data, now modified to be three dimensional (3D), and a comprehensive 3D chromatogram from Moore and Jorgenson.

The Moore and Jorgenson chromatogram, which has 25 identifiable 3D volume elements or peaks, displayed good orthogonality values over all dimensions. However, OMs based on discretization of the 3D space changed substantially with changes in binning parameters. This example highlights the importance in higher dimensions of having an abundant number of retention times as data points, especially for methods that use discretization. The Gilar data, which in a previous study produced 21 2D datasets by the pairing of 7 one-dimensional separations, was reinterpreted to produce 35 3D datasets. These datasets show a number of interesting properties, one of which is that geometric and harmonic means of lower dimensional subspace (i.e., 2D) OMs correlate well with the higher dimensional (i.e., 3D) OMs. The space utilization of the Gilar 3D datasets was ranked using OMs, with the retention times of the datasets having the largest and smallest OMs presented as graphs. A discussion concerning the orthogonality of higher dimensional techniques is given with emphasis on molecular diversity in chromatographic separations.

In the information theory work, an inconsistency is found in previous studies of orthogonality using the 2D metric often identified as %O. A new choice of metric is proposed, extended to higher dimensions, characterized by mixes of ordered and random retention times, and applied to

Correspondence to: Mark R. Schure; Joe M. Davis.

¹Current address: 733 Schloss Street, Wrightsville Beach, NC 28480 USA

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the experimental datasets. In 2D, the new metric always equals or exceeds the original one. However, results from both the original and new methods are given.

1. Introduction

The techniques known collectively as multidimensional liquid chromatography are practiced typically with two columns, whereby the effluent from the first column is provided to the second column for further separation based on molecular structure. These are referred to as two-dimensional (2D) separations [1]. While 2D gas chromatography has been widely available for a number of years, instrumentation for 2D liquid chromatography (2DLC) is now becoming widely available, and the ability to resolve compounds with similar molecular structure has been invaluable in studies of complex materials, most notably biomolecules.

Three-dimensional (3D) and higher dimensional separations have been discussed for years, but only one report exists on comprehensive 3DLC instrumentation able to use three different columns [2]. However, with increasing needs of speed and resolution for demanding biomolecule characterization, more attention is being given to using n chromatographic dimensions and a number of mass spectrometry (MS) dimensions, which might include ion mobility mass spectrometry, MS/MS or a combination of other techniques.

Orthogonality metrics (OMs) have been used to measure space occupancy and the uniformity of the spreading of components within these spaces [3–9]. This has become important in one-dimensional (1D) separations in determining and optimizing the resolution of specific components [4]. In 2D chromatography, OMs [3–5,7–9] measure the utilization of the separation space with occupancy metrics, peak spacing metrics and uniformity metrics, among others. These metrics can be utilized to measure the effectiveness of the separation and to optimize the separation.

Higher dimensional separations are indeed rare. One of these, published by Moore and Jorgenson [2], is the separation of peptides using 3D chromatography. In this experiment, the first dimension column is a size exclusion chromatography (SEC) column, followed by reversed-phase liquid chromatography (RPLC) column, followed by a very fast optically-gated capillary zone electrophoresis (CZE) experiment. As is a rather common practice, the columns and techniques were chosen to put the slower separation in the first dimension followed by faster and faster steps in succeeding dimensions. This is due to the sampling requirements [1,10,11], which dictate that the slower separations are typically put in the first dimension and the fastest steps are performed in the last dimension.

In this paper, we explore some of the aspects of 3D and higher dimensional OMs. The OMs described here are extensions of known 2D OMs and include correlation coefficients (CCs), dimensionality, information theory metrics and convex-hull metrics. For the dimensionality and information theory metrics, which employ discretization of retention times into bins, recommendations are given to the number of bins and the implications of binning in higher dimensions.

Two applications of 3D OM to experimental systems are discussed. First, a 3D dataset is reconstructed from the published 3D comprehensive separation of Moore and Jorgenson [2]. Second, we utilize the Gilar peptide data [7], which in a previous study [9] produced 21 unique 2D datasets by the pairing of 7 1D separations. This data now is reinterpreted to produce 35 unique 3D datasets. In all cases, datasets are $m \times 3$ matrices, where the i th column vector contains m retention times for the i th dimension. The evaluation of these datasets using 3D OM produces many interesting insights that are discussed in detail and point out many aspects and difficulties in analyzing and optimizing 3D separations.

With the exception of CCs, a global OM in 3D is available, along with lower dimensional subspace OM. For 3D systems, there are 3 2D subspace metrics based on paired dimensions. They will be shown to be quite useful. This scheme can be extended to n D (n -dimensional) systems, such that for $n \geq 3$ all of the lower dimensional subspace OM exist.

This work uncovers a problem with the previously derived 2D OM from information theory known as %O [8]. We will show that this OM is not symmetrical with respect to the input data, and we will derive a more universal equation able to be employed in any number of dimensions. The analysis of the datasets shows the 2D OM from the original and new methods are nearly equal, but we prove the new method always gives the same or a greater OM than the original one.

2. Mathematical development

2.1. Notation and terminology

The letter “D” in 1D, 2D, 3D, n D, etc., henceforth means both “dimension” and “dimensional”, depending on context. Its dual interpretation provides simplicity. The symbol D is the traditional symbol for the OM, dimensionality; it differs from these abbreviations in the absence of a number preceding it.

Various notations for column vectors of retention times are used, depending on circumstances. Vector groupings are referred to as pairs in 2D and triplets in 3D. In 2D, the individual vectors are labeled X (first dimension) and Y (second dimension); in 3D, they are usually X , Y , and Z (third dimension). In n D, they are $X_1, X_2, \dots, X_n$. Depending on circumstances, pairs and triplets are represented by letter or number combinations, e.g., XZ , XYZ , and 246 (the numbers are explained later). Individual coordinates in any vector are represented by lower case letters, e.g., x , y , and z , and represent retention times of peaks. The notation style used here for the most part is found in information theory texts, for example, in the well-known book of Cover and Thomas [12].

Generic symbols for OM have no subscripts, e.g., dimensionality D . Subscripts are used to identify specific types, e.g., D_{2D} for 2D dimensionality.

The information theory OM has the traditional name, orthogonality. The word “orthogonality” also is used in its general sense to indicate a good spreading of zones over a separation. We have tried to avoid confusion between the two meanings through context.

Zones in the SEC and CZE dimensions of the Moore and Jorgenson separation have elution times, not retention times, because retentive chromatography was not used. However, for consistency we describe all such times as retention times and describe this 3D separation as a 3D chromatogram. We also use the older abbreviation, CZE, to conform to the Moore and Jorgenson chromatogram reproduced here.

2.2. General properties of OMs

To be useful, OMs must have some essential properties. The first is the scale. The scale should be defined between 0 and 1, or some other values that convey characteristics of zone ordering or zone coverage. The CC analyses scale from 0 to 1 after a suitable normalization. The information theory results are typically scaled between 0 and 100%, although the 0 to 1 scale is more natural, and we will use that scaling here. The fractal dimension scale is between 0 and D_C , where D_C is the number of chromatographic instrumental dimensions, i.e., the number of columns that are used for successive fractionation. The convex hull relative area, volume, or hypervolume in any number of dimensions varies between 0 and 1. These scalings, with the exception of the fractal dimension, can be applied so that an OM can be suitably normalized between 0 and 1 no matter how many dimensions are utilized as the separation space.

Another important property is that OMs should give the same answer upon swapping the column data vectors. In other words, the OM should preserve the data symmetry and give the same answer for any order of the vectors. This is not to say that the order in which the different dimensions of a multidimensional separation are developed is not important; obviously it is. However, the vector order does not affect the intervals between these times, and therefore should not affect the OM. For example, in 2D, $OM(X, Y)$ should equal $OM(Y, X)$, where OM is the specific orthogonality metric calculation. In 3D, it should be true that $OM(X, Y, Z) = OM(X, Z, Y) = OM(Y, X, Z) = OM(Y, Z, X) = OM(Z, X, Y) = OM(Z, Y, X)$.

A few of these OMs do not have this property of data symmetry. For example, the Pearson and Kendall CCs have this property while the Spearman CC does not, and this suggests that the Spearman CC is not suitable for use as an OM. Furthermore, the %O OM [8] from information theory [12] does not have this property. However, on further inspection, a similar OM can be derived so that this symmetry property is preserved. We derive such a metric below.

OMs can be classified as non-discretized and discretized [9]. The non-discretized metrics do not require the binning or scaling of retention times, although scaling does not change the OM value. Discretized OMs require both binning and scaling.

The details of binning are introduced as needed. For consistency, all retention times are scaled here over the range as 0 to 1 using the maximum (max) and minimum (min) times of each dimension. For example, the scaled value x_i^s is calculated from the i th retention time x_i in dimension X as

$$x_i^s = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

This scaling assumes that each data vector (i.e., dimension) is to be considered equally important and that even though the range of the data is different, they are weighed equally.

2.3. Orthogonality metrics

2.3.1. Non-discretized OMs

2.3.1.1. Correlation coefficient: A simple formulation for the Pearson CC is to define two vectors, e.g., X and Y , which are of length m . In chromatographic terms, the m retention times in a 2D chromatography experiment are stored as separate vectors, X and Y . They could as easily be just two columns in a spreadsheet.

The sample Pearson CC r_{XY} is defined as

$$r_{XY} = \frac{\text{cov}(X, Y)}{s_X s_Y} \quad (2)$$

where cov is the sample covariance function which can be explicitly written such that

$$r_{XY} = \frac{1}{m-1} \sum_{i=1}^m \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right) \quad (3)$$

where $\bar{x}$ and $\bar{y}$ are the averages, and s_X and s_Y are the standard deviations, of vectors X and Y .

With simple matrix algebra, the vectors X , Y , and Z of a 3D chromatogram can be processed to obtain

$$R = \begin{bmatrix} r_{XX} & r_{XY} & r_{XZ} \\ r_{YX} & r_{YY} & r_{YZ} \\ r_{ZX} & r_{ZY} & r_{ZZ} \end{bmatrix} = \begin{bmatrix} 1 & r_{XY} & r_{XZ} \\ r_{YX} & 1 & r_{YZ} \\ r_{ZX} & r_{ZY} & 1 \end{bmatrix} \quad (4)$$

where R is the correlation matrix (CM). The CM is a symmetric matrix, i.e., $r_{ij} = r_{ji}$, where r_{ij} is the CC between the i th and j th vectors or dimensions. Only the upper or lower triangular matrix elements of R are unique and necessary. Note that the CM is composed of pair interactions – higher dimensions just contain more two-dimensional $r_{X_1 X_2}$ values, where X_1 and X_2 are generic column vectors. The diagonal elements in Eq. (4) are 1's because the CC of two identical vectors is 1.

The size of the matrix R is $D_C \times D_C$. The concept of D_C can also be applied to the sequential elution style of separation [13] or to D_C columns in series, but its usage here is specific to the (integral) number of columns in a multidimensional experiment. In the context of multidimensional chromatography, $D_C = s$, where s is Giddings' sample dimensionality [14]. The extension of CCs to higher dimensions where $D_C = 4, 5, 6, \dots$ is a simple extension of Eq. (4).

The number of unique off-diagonal matrix elements in Eq. (4) is equal to $(D_C^2 - D_C)/2$. For the usual and common case of $D_C = 2$ (e.g., 2DLC), the number of unique elements in R is 1; this is the common CC.

As noted [9], previous transformations of CCs for 2D chromatography include $1 - |r_{ij}|$ and $1 - r_{ij}^2$ [3,6,15,16]; these are helpful towards scaling the results for two reasons. The first is to make the number vary from 0 to 1; the non-diagonal elements of R can be less than 0 (due to negative slopes formed from vector pairs), and this operation forces positivity. Second, taking the complement of r (or r^2) allows a rescaling of low correlation coefficients (more uniformity in the 2D space or subspace) to give higher numbers indicative of better separations. The opposite also occurs: higher values of r indicate low orthogonality; the complement operation reorders this quantity to a more "separation scientist relevant" quantity. We use the transformation, $1 - r_{ij}^2$.

2.3.1.2. Convex hull relative volume: The 2D convex hull [17–19] is a mathematical device that wraps a graph of 2D retention times with an outer hull. It can be visualized by taking cellophane tape and wrapping the points as if they were pegs; the tape wraps the outermost points only. The internal area is a general measure of space utilization: the larger the area, the more space that is used. The convex hull relative area is an OM equaling the ratio of this area to the rectangular area mapped by the four extrema, i.e., the smallest and largest retention times in both dimensions [9]. It is symbolized by A_H and lies between 0 and 1. For scaled coordinates, the rectangular area is 1 and A_H equals the hull area.

The concept can be extended to three or more dimensions. For example, in 3D the convex hull relative volume V_H is the volume of a 3D hull, divided by the volume of the rectangular cuboid formed by the 8 points comprising the volume extrema. For n D, the number of extrema points is 2^n .

Algorithms for determining the convex hull in 2D, 3D and higher dimensional space are available directly in MATLAB (convhulln) and other math libraries. These routines are based on Qhull [20, 21], which can be easily ported to most computer platforms. The MATLAB routine convhulln returns the volume of the polyhedron in 3D and the hypervolume in higher dimensions, in addition to the point coordinates of the hull itself. In 2D, the returned volume is an area.

2.3.2. Discretized OMs—The two OMs discussed here are based on interpreting the numbers of scaled retention times falling in different bins superposed on a graph of these times. When we write about the number of bins, we implicitly mean the number of bins per

dimension, e.g., 5 bins in 2D means a 5×5 array of 25 bins, and 5 bins in 3D means a $5 \times 5 \times 5$ array of 125 bins. The bin specifications differ for each OM and are discussed later.

2.3.2.1. Fractal (box counting) dimension: This method determines the OM, dimensionality D [9,22], of chromatographic experiments and has been used in one and two dimensions [9,22]. The box-counting algorithm [22] can be used in any number of dimensions but is very inefficient for dimension numbers $n > 3$, because the number of boxes (or bins) becomes very large. The algorithm for 3D (and higher dimensions) is a simple extension of that used in the analysis of 2D separation data described in ref. 22.

The fractal or box-counting dimension measures how the spacing between peaks scales and indirectly how effectively the separation space is utilized. We refer to this quantity in this paper as the dimensionality. The dimensionality is defined as [22]: $D = \log N(\epsilon) / \log (1/\epsilon)$ in the limit of $\epsilon \rightarrow 0$, where $N(\epsilon)$ is the number of objects (such as boxes in two dimensions and cubes in three dimensions) with object size ϵ needed to cover some structure, such as points in a multidimensional plot that represent retention times in a multidimensional chromatogram. D describes how the mass or density of an object scales with the length scale that describes the object; it need not be an integral number. The so-called box-counting algorithm uses a least-squares fit of $\log N(\epsilon_i)$ versus $\log \epsilon_i$ and gives an estimate of D by multiplying the slope of this fit by -1 [22].

For 2D chromatography, $0 \leq D \leq 2$; for 1D chromatography, $0 \leq D \leq 1$, and in general $0 \leq D \leq D_c$. If retention time pairs are lined up on the diagonal in 2D chromatography, e.g., if the same stationary phase and solvent system are used for both columns, then $0 \leq D \leq 1$. If the pairs are perfectly ordered on the diagonal, i.e., if the peak spacing is constant, then $D = 1$. If the peak spacing on the diagonal is random, then $0.6 \leq D \leq 0.75$. This is also true for a 1D chromatogram with randomly spaced peaks. For 3D chromatography, if the retention times are uniformly spaced on a 3D lattice, then $D = 3$. If the 3D retention times cover a uniform 2D lattice, then $D = 2$. Hence, no matter how many chromatographic dimensions are utilized, D is reflective of how many dimensions are actually utilized in effecting the separation. In practical cases D will almost always be non-integral.

With fractal box-counting, there is always a question of the proper range of the number of bins in the discretization scheme. Typically the range in each dimension of the 2D algorithm is 1 to $\lfloor m^{1/2} \rfloor$, where the partial brackets signify the floor function, which for positive numbers discards any fractional component and returns an integer [23]. An alternative to this upper limit is Sturges's relationship [24], which uses the base-2 logarithm as the basis for the number of bins.

The largest number of bins can be generalized to n D separations using $\lfloor m^{1/n} \rfloor$ as an extension to the square-root bin number. For example, the largest number of bins for calculating D_{3D} from 105 3D retention times is $\lfloor 105^{1/3} \rfloor = 4$. The largest number for calculating the lower dimensional subspace D_{ij} from any 2 of these 3 vectors is $\lfloor 105^{1/2} \rfloor = 10$, just as if one had 105 2D retention times without the third dimension.

2.3.2.2. Multidimensional orthogonality based on information theory: Unlike CCs and hulls, which have been used in 3D and n D systems, and the fractal dimension, whose extension beyond 2D separations was discussed before this paper [22], OMs based on information theory have been limited to 2D systems. This limitation, when coupled with the paucity of 3D and n D studies of information in separation science, warrants a detailed discussion.

Information or entropy is a metric of uncertainty or lack of knowledge. Slonecker et al. [5] cited expressions reported by Massart and Kaufman [3] for fractional information content and informational similarity, and proposed the concept of %synentropy, or “the percentage of the 2D informational entropy contributed equally from both dimensions.” The concept was used in several papers. More recently, Pourhaghighi et al. [8] defined the percent orthogonality, % O , of a 2D separation as % $O = 100 * O^o$, where

$$O^o = \frac{H(Y|X)}{H(Y)} \quad (5)$$

In Eq. (5), $H(Y)$ is the entropy of Y and $H(Y|X)$ is the conditional entropy of Y , given X , i.e., the entropy needed to describe the part of Y that remains unknown when X is specified or, equivalently, the entropy of the unique part of Y . Different entropies are shown in the 2D information diagram, Fig. 1a, with circles representing the entropies of X , $H(X)$, and of Y ; their union represents the joint entropy $H(X, Y)$. A useful feature of such diagrams is additivity, e.g., $H(Y|X) + I(X; Y) = H(Y)$, where $I(X; Y)$ is the mutual information between X and Y . The diagram shows $0 \leq O^o \leq 1$.

The rationale for Eq. (5) is straightforward: the less uncertainty one has in Y , given knowledge of X , the more X and Y are correlated and the lower the orthogonality. We have shown this metric correlates well with experts' subjective evaluations of 2D separations [9]. Consequently, it is not our intent to criticize it. However, as observed above, we think OMs should give the same answer upon swapping data vectors. If Eq. (5) were independent of this swap, then it would equal $H(X|Y)/H(X)$. However, conditional entropies are not symmetric; $H(Y|X) \neq H(X|Y)$ [12]. In addition, our calculations for 2D separations discussed in ref. [9] show that $H(Y|X)/H(Y) \neq H(X|Y)/H(X)$, although the values are usually similar.

A simple modification of Eq. (5) provides the symmetry. Fig. 1a shows that the ratio $I(X; Y)/H(X, Y)$ is the fraction of entropy that is shared between X and Y . Consequently, the complement of this ratio, $1 - I(X; Y)/H(X, Y)$, is the fraction of independent entropy that is not shared. Using the relationship

$$H(X, Y) = H(Y|X) + H(X|Y) + I(X; Y) \quad (6)$$

expressed by Fig. 1a, we can rearrange this fraction and propose it as a new entropy-based definition of 2D orthogonality

$$O = \frac{H(Y|X) + H(X|Y)}{H(X, Y)} \quad (7)$$

where the lack of superscript “o” distinguishes this result from the original one, Eq. (5). The swap of X and Y has no effect on Eq. (7). We show in Section 1 of the Supplemental Material that $O = O^o$.

Although we were led to Eq. (7) by independent thought, it has been used before. It is the “fractional information content” cited by Slonecker et al. (Eq. (7) in ref. 5) and the “relative measure” cited by Massart and Kaufman (Eq. (25), Chapter 1 in ref. 3) for calculating similarity. We simply have defined it as O .

The earlier use of this fraction suggests it is a basis for extending orthogonality to higher dimensions. Fig. 1b is a 3D information diagram; equations for the terms therein are given in Section 2 of the Supplemental Material. The diagram shows the conditional entropy $H(X|Y, Z)$ is the entropy needed to describe the part of X that remains unknown when Y and Z are specified, i.e., the entropy of the unique part of X . Similar arguments can be made for the conditional entropies, $H(X|Y, Z)$ and $H(Z|X, Y)$. In analogy with Eq. (7), we propose that the 3D orthogonality, O_{3D} , is the sum of the unique conditional entropies of the three dimensions, divided by the joint entropy

$$O_{3D} = \frac{H(X|Y, Z) + H(Y|X, Z) + H(Z|X, Y)}{H(X, Y, Z)} \quad (8)$$

As before, swaps among X , Y , and Z do not change the result.

Similarly, the n D orthogonality, O_{nD} , is the sum of the n conditional entropies of each variable, given knowledge of the other variables, divided by the joint entropy. As shown in Section 3 of the Supplemental Material, the chain rule of entropy [12] allows one to express O_{nD} as

$$O_{nD} = \left[n - \frac{\sum_{j=1}^n H(\{X_i\}_{i \neq j})}{H(X_1, X_2, \dots, X_n)} \right] \quad (9a)$$

where $X_1, X_2, \dots, X_n$ replace X, Y, Z , etc., and $\{X_i\}_{i \neq j}$ is the subset of the $n - 1$ X_i variables that excludes X_j . For example

$$\sum_{j=1}^n H(\{X_i\}_{i \neq j}) = H(X_2, X_3, \dots, X_n) + H(X_1, X_3, \dots, X_n) + \dots + H(X_1, X_2, \dots, X_{n-1}) \quad (9b)$$

Consequently, O_{2D} is

$$O_{2D} = \left[2 - \frac{H(X_2) + H(X_1)}{H(X_1, X_2)} \right] \quad (10)$$

or equivalently

$$O_{2D} = \left[\frac{[H(X_1, X_2) - H(X_2)] + [H(X_1, X_2) - H(X_1)]}{H(X_1, X_2)} \right] \quad (11)$$

where the “2” is incorporated in the fraction and terms are grouped. Using an information diagram, one sees that $H(X_1, X_2) - H(X_2)$ equals $H(X_1|X_2)$ and $H(X_1, X_2) - H(X_1)$ equals $H(X_2|X_1)$, with the final result identical to Eq. (7).

An attractive feature of Eqs. (9a) and (9b) is that they depend only on joint entropies, which are easily calculated using Shannon’s formula [12]

$$H(X_1, X_2, \dots, X_n) = - \sum p(X_1, X_2, \dots, X_n) \log_2 p(X_1, X_2, \dots, X_n) \quad (12)$$

where $p(X_1, X_2, \dots, X_n)$ is the probability that an n -dimensional retention time falls into a given bin and the summation is evaluated over all bins.

A calculation of O requires a binning of scaled retention times. The number of bins used here is $\lfloor m^{1/n} \rfloor$. However, to follow the tradition of entropy calculations [12] and to preserve the additivity of information diagrams, subspace OMs are calculated using the same number of bins as the space of higher dimension. For example, the bin number for calculating O_{2D} from 105 2D retention times is $\lfloor 105^{1/2} \rfloor = 10$. The number for calculating O_{3D} from 105 3D retention times is $\lfloor 105^{1/3} \rfloor = 4$. However, the number for calculating the lower dimensional subspace O_{ij} from any 2 of these 3 vectors remains $\lfloor 105^{1/3} \rfloor = 4$; this calculation differs from that for dimensionality.

2.4. Binomial coefficient

The binomial coefficient for the number $N = \binom{k}{i}$ of unique ways that k items can be taken i at a time, without regard to order, is

$$N = \frac{k!}{i!(k-i)!} \quad (13)$$

Eq. (13) has two applications here. The first is the number of subspace OMs that can be formed in an i dimensional space and used to characterize a higher space of k dimensions. For example, in the 3D space XYZ , the number of subspace 2D OMs that can be formed is $3!/(2!1!) = 3$ ($k = 3$, $i = 2$), i.e., 3 metrics exist for the 2D spaces XY , XZ , and YZ . In this application, k is equal to n , the number of instrumental dimensions (which in turn equals D_C).

The second application is the number of unique groupings of i columns that can be formed from an available pool of k columns. For example, in the $k = 7$ 1D separations of Gilar, $7!/(2!5!) = 21$ unique column pairs exist to construct 2D datasets ($i = 2$), and $7!/(3!4!) = 35$ unique column triplets exist to construct 3D datasets ($i = 3$). In this latter application, i is equal to n .

3. Procedures

3.1. Data processing

All OMs were calculated using MATLAB (Mathworks, Natick, MA), and additional validation of the information theory results was obtained from FORTRAN programs. The MATLAB programs used input and output from Microsoft Excel (Redmond, WA) spreadsheets. Data input from Gilar's 7 column retention data [7,9] was read as a spreadsheet. The 35 3D datasets were constructed by program, and results were accumulated in a spreadsheet.

3.2. Moore and Jorgenson's 3D experiment

As mentioned previously, one of the first and only comprehensive 3D chromatograms was obtained by Moore and Jorgenson [2]. Since the data for this chromatogram was not supplied as a supplemental file or listed in the text, we decided to reconstruct it by manual means from the figures in the original paper (Figs. 3 and 7A in ref. 2). 3D retention time data was obtained as an $m \times 3$ matrix, using a series of MATLAB programs.

The first step was to digitize the 12 slice images [2] contained in Fig. 3 of that publication. This was accomplished using a program that digitized each slice using a mouse-driven cross-hair cursor with subsequent identification of the middle of each zone, giving retention times in the RPLC (Y) and CZE (Z) dimensions. The program required the selection of the minimum and maximum y and z times in each slice as linear scaling constraints. Each slice, which was stored as a separate file, is associated with a different SEC (x) time, with a difference of 6.35 minutes between slices.

In addition, the data in Fig. 7A of ref. 2, which consists of the sum of all SEC slice data projected into the RPLC-CZE plane, was also digitized using the same MATLAB program. Having the summed data allowed isolation of the RPLC (y) and CZE (z) retention times of

each zone, with subsequent determination of each SEC retention time x by the following procedure.

The distance of a point in the summed data RPLC-CZE plane to each point in all of the SEC slice data (also in the RPLC-CZE plane) was calculated using scaled (0 to 1) coordinates. The mean of the closest three points in the SEC slice data was then used as the SEC retention time, x . The standard deviation was calculated from these three closest points. Three points were chosen because most zones had sufficient width in the SEC dimension to occupy more than three slices, so this number was a compromise between noise sensitivity and sampling accuracy. The relatively low standard deviations suggest that this approach was consistent across slices.

3.3. Gilar's dataset

Gilar and coworkers [7] examined 7 different LC columns with different stationary phases or solvent pH values to ascertain how peptides of known sequence would be separated. The columns and conditions used to obtain these datasets are summarized in Table 2. We previously formed 21 unique retention time pairs from these 7 columns and interpreted them with 2D OM's [9,25]. This action was extended to form 35 unique triplets (i.e., 35 groupings of 3 LC columns), similar to the pairs in 2D. The numbers 21 and 35 were calculated from the binomial coefficient N , Eq. (13), as discussed in Section 2.4.

The retention times for the Gilar dataset are available as supplementary material [9].

3.4. Characterization of O by diagnostic coordinate patterns

The new O metrics were characterized by calculating values for mixes of ordered and uniformly random coordinates in scaled squares (2D) or cubes (3D). These coordinates represent the scaled retention times of chromatographic peaks. In one case, coordinates were purely random. In another case, some coordinates were assigned to diagonal bins, while others were distributed randomly elsewhere. The percentage of diagonal coordinates was varied on placing multiple coordinates in diagonal bins, with the same number in each bin. In a third case, some coordinates were placed in bins like checkers on a checkerboard, with one coordinate per bin, while others were distributed randomly elsewhere. The percentage of checkerboard-like coordinates was varied by systematically filling contiguous rows and columns of bins. Details on the generation of these coordinate patterns and example generations are given in Section 4 of the Supplementary Material.

For each characterization, the average of 1000 simulations was calculated. For the purely random case the coordinate density, or the number of coordinates per number of bins, was varied; in the other cases, the numbers of coordinates and bins were equal.

4. Results

4.1. 3D chromatogram of Moore and Jorgenson

Fig. 2A shows the original 3D chromatogram of Moore and Jorgenson (their Fig. 4) [2]. The retention times and standard deviations determined here are shown in Fig. 2B and summarized in Table 1. Fig. 2C shows Fig. 7A from Moore and Jorgenson [2], which is the

superposition of the 12 SEC slices. A view of our Fig. 2B from above looks very much like Fig. 2C, signifying the essential features of this chromatogram have been reconstructed successfully.

These $m = 25$ retention times form a triplet, which was processed for the 3D OMs, dimensionality (D_{3D}), convex hull relative volume ($V_{H,3D}$), and orthogonality (O_{3D}). It also was processed for the 3 subspace 2D OMs based on CCs ($1-r_{ij}^2$), dimensionalities (D_{ij}), convex hull relative areas ($A_{H,ij}$), and orthogonalities (original O_{ij}^o and new O_{ij}), with ij representing the vector pairs, XY , XZ , and YZ . Table 3A gives CCs and relative convex hull metrics, which are non-discretized and independent of binning. Table 3B gives the dimensionality and orthogonality metrics, which depend on binning.

A number of features are observed here. First, all three CCs of the 2D subspace in Table 3A show relatively high independence; in other words, this separation appears to be quite orthogonal. Clearly, the SEC (X) and RPLC (Y) dimensions are not as orthogonal as the other dimension pairs. The convex hull relative volume, $V_{H,3D}$, is somewhat low (0.310), but this may be partly due to the small number of retention times in the chromatogram. The convex hull relative areas for the 2D pairs, $A_{H,ij}$, show adequate spatial coverage. Interestingly, the product of the 3 $A_{H,ij}$ values, 0.282, almost equals $V_{H,3D} = 0.310$. These OMs show that one can have a good independence of retention mechanisms and medium utilization of the available subspace. However, one must view these results with caution as the number of data points, 25, is relatively small.

This point is particularly emphasized in Table 3B, where the discretized 3D OMs dimensionality D_{3D} and orthogonality O_{3D} are given for two different bin numbers, 2 and 3, with 2 being the number determined by the formula $\lfloor m^{1/3} \rfloor$. This bin number is inadequately small, and this is emphasized by the large change in D_{3D} (from 3.000 to 2.406) and O_{3D} (from 0.998 to 0.668), when 3 bins are used. One would expect equi-spaced data when $D_{3D} = 3$ but this bin number is just too small for reliable estimation. Even 3 bins are most likely too few for D_{3D} estimation; this is verified by the poor least-squares fit of a line to the log of the number of occupied boxes (bins) as a function of the log of the box size (not shown).

The 3 subspace 2D OMs of D are given in Table 3B. These D_{ij} 's are relatively high, which suggests good zone spacing in the subspaces. This is not surprising because of the small number of components in the separation.

Also shown in Table 3B are the 3 subspace 2D OMs derived using the original method for orthogonality, signified by the notation O_{ij}^o , and the new method signified by O_{ij} . The original method gives smaller values than the new method (this is expected, as shown in the Supplementary Material) but these values are typically quite close. Both the original and new methods of O estimation are affected by bin size. While the estimates are clearly dependent on the number of bins for both the original and new methods, the numbers are near unity suggesting a very good separation.

4.2. OMs for the Gilar dataset and correlations among them

The OMs for the 35 triplets of the Gilar 3D dataset are given in Table 4. This data is also given in Supplemental Material as an Excel spreadsheet. The numbers assigned to X , Y , and Z in the first column of the table identify the chromatographic columns in Table 2. The order of these numbers is immaterial, since the triplet is a grouping of 1D vectors and the OMs are the same on vector exchange. Some of the 2D subspace OMs are repeated in Table 4, e.g., the 2D subspace OMs having subscript XY are the same for the triplets 123, 124, 125, 126, and 127, because they are calculated using retention times from columns 1 and 2 in Table 2.

Also shown in Table 4 is the 3D OM product $O_{3D} \cdot V_{H,3D}$ that is reminiscent of the 2D product of orthogonality O and the convex hull relative area A_H suggested in previous 2D work [9, 26]. The concept underlying this product is that both the ordering of retention times in the available space (i.e., large O) and the size of that space (i.e., large $V_{H,3D}$) are important. These numbers show a general correlation across the metrics whereby “good” 3D chromatograms show higher products than less desirable chromatograms.

An evaluation of trends and correlations among these 35 triplets may be more useful (and simpler) than an interpretation of each. The correlation between the O_{3D} and D_{3D} vectors in Table 4, over all of the triplets, gives a Pearson CC of 0.938. This implies that these OMs give similar results for judging space utilization in 3D, at least for this peptide dataset. This is not surprising because both OMs use the discretization of data in bins as the basis for the method. The correlation also is encouraging; if it were absent either one or both OMs might be wrong. However the dimensionality gives an estimate of peak spacing. The total correlation matrix will be presented in a subsequent paper.

In the 2D subspaces, the original orthogonality (O_{ij}^o) appears to track the new one (O_{ij}) well, with the new OM always being larger than the original one. Inspection of these values shows that the original metric gets closer to the new one as both values asymptotically approach 1.

The near equivalence of $V_{H,3D}$ and the product of the 3 2D subspace $A_{H,ij}$ values, first observed in the Moore and Jorgenson data, also exists in the Gilar dataset as observed in the data given in Table 4. For $V_{H,3D}$ (ordinate) vs. the $A_{H,ij}$ product (abscissa), the slope, intercept, and Pearson CC are 1.10, -1.10×10^{-4} , and 0.986, respectively. Further study is needed to understand the basis of this equivalence and to ascertain if it is general.

Although the 3D OMs for dimensionality, orthogonality, and hull relative volume derive from data in all 3 dimensions simultaneously, the 2D subspace OMs have been suspected of other types of correlation to the 3D results. We have anticipated correlations by taking the harmonic and geometric means of various 2D subspace OMs and comparing them to the 3D OMs. These correlations are established in Table 5.

The harmonic mean $\overline{OM}_H$ is defined as

$$\overline{OM}_H = \frac{N}{\sum_{i=1}^N \frac{1}{OM_i}} \quad (14)$$

where N is the number of OM's to be averaged. For 2D subspace OM's in a 3D system, $N=3$, as calculated from Eq. (13). The geometric mean $\overline{OM}_G$ is defined as:

$$\overline{OM}_G = \left[\prod_{i=1}^N OM_i \right]^{\frac{1}{N}} \quad (15)$$

The Pearson CCs for a select group of correlations are shown in Table 5. As can be seen, all but 2 of 14 of them exceed 0.9, indicating good correlation. Specifically, a good correlation exists between $V_{H,3D}$ and the corresponding geometric and harmonic means of its 2D subspace counterparts, $A_{H,ij}$. Furthermore, there is a good correlation between D_{3D} or O_{3D} and the harmonic and geometric means of not only their specific 2D subspace counterparts but other 2D subspace OM's as well (e.g., D_{3D} correlates well not only with the means of D_{ij} but with those of $1-r_{ij}^2$ and O_{ij}). The correlations in Table 5, derived from the geometric means, appear to be slightly larger in most cases than those derived from the harmonic means. The (fractal) dimensionality D_{3D} values appear to correlate better with the subspace means than do the orthogonality O_{3D} values. This suggests that if one has 3D data, the use of the geometric mean of the three subspace CCs (specifically as the complement of the CC squared) may suffice as an easily obtainable OM that can be calculated with a simple calculator. However, we suspect that both D_{3D} and O_{3D} are the OM's of choice for these types of analyses. These correlations do not imply that 3D OM's can be predicted from 2D subspace OM's, at least at this time. We do not know if the regression coefficients will describe other data sets. However, as previously observed, the convex hull relative volume $V_{H,3D}$ appears to be predictable by simply multiplying the 2D dimensionless areas $A_{H,XY}$, $A_{H,XZ}$, and $A_{H,YZ}$. For example, the areas for the good 3D separation of triplet 256 are 0.758, 0.772, and 0.652; their product, 0.382, is similar to $V_{H,3D} = 0.438$. Also, the areas for the poor 3D separation of triplet 124 are 0.108, 0.694, and 0.695; their product, 0.052, is similar to $V_{H,3D} = 0.055$. For this poor 3D separation, $A_{H,XY}$ is small, flattening $V_{H,3D}$ into a small volume even though $A_{H,XZ}$ and $A_{H,YZ}$ are large. If this relationship is general, then all 3 2D areas must be large, if $V_{H,3D}$ is to be large.

4.3. Predictions of O for diagnostic coordinate patterns

Fig. 3a is a graph of O (circles, 2D; squares, 3D) and coordinate density (solid line, 2D; dashed line, 3D) vs. the number of random coordinates in simulations of scaled retention time sequences. (Strictly, the randomness is distorted on scaling but the distortion is small.) The OM equals about 0.9 for O_{2D} and varies between 0.8 and 0.9 for O_{3D} . Both OM's oscillate in phase with the coordinate density, i.e., the number of coordinates per number of bins. The oscillation is greater for O_{3D} than for O_{2D} because the relative density change is

larger. The oscillation is an artifact of binning. Another binning artifact is that O does not equal unity, as expected for a uniformly random distribution. The artifacts are explored in Sections 5 and 6 of the Supplemental Material.

Fig. 3b is a graph of O vs. the percentage of coordinates in diagonal bins. The coordinate density is one. Both O_{2D} and O_{3D} decrease to zero as the percentage approaches 100 – the expected result for full coordinate alignment along the diagonal – although they decrease at different rates. Interestingly, at constant coordinate density the resultant curves are almost independent of bin number, except as the coordinates approach total randomness (not shown).

Another coordinate ordering is shown in Fig. 3c, in which some coordinates are placed in bins like checkers on a checkerboard. Fig. 3d is a graph of O vs. the percentage of such ordered coordinates. As the percentage increases from 0 to 100, both O_{2D} and O_{3D} systematically increase in a nearly linear manner from the random limit shown in Fig. 3a to unity. Unlike in Fig. 3b, no universal curve exists.

These characterizations show that O varies largely as expected, although it has limitations imposed by binning.

5. Discussion

In an attempt to determine the “best” and “worst” chromatograms of the Gilar dataset, we sorted the 3D OM_s given in Table 4. The triplets with the 10 largest OM_s (the best) and the 10 triplets with the lowest OM_s (the worst) are listed in Table 6. The top 10 are listed in descending order; the bottom 10, in ascending order. As can be seen, the 3D OM_s and one OM product have similar rank for the best and worst triplets, with strong correlation among these OM_s. For example, the triplet 345 (i.e., the grouping of columns 3, 4, and 5 in Table 2) has the second largest O_{3D} , the second largest D_{3D} , the ninth largest $V_{H,3D}$, and the fifth largest OM product.

For each ranked triplet, we have included the corresponding OM in parentheses. We see that the best triplets have OM_s that roughly range from 0.8 to 0.9 for O_{3D} , 2.5 to 2.6 for D_{3D} , 0.35 to 0.40 for $V_{H,3D}$, and 0.9 to 1.1 for the product, $D_{3D} V_{H,3D}$ values are similar to those for purely random peaks (see Section 4.3). In These O_{3D} contrast, the worst triplets have OM_s that roughly range from 0.5 to 0.7 for O_{3D} , 1.9 to 2.3 for D_{3D} , 0.05 to 0.15 for $V_{H,3D}$, and 0.05 to 0.35 for the product, $D_{3D} V_{H,3D}$.

This result is similar to that found previously [9], where we observed that for subjective evaluation by experienced researchers, very good and very bad chromatograms are easily distinguishable but intermediate ones are not obvious as to their ordering. In the evaluation here, which involves a purely numerical ranking, similar trends are found. This is not surprising, even though the hull relative volume is certainly different than orthogonality, dimensionality, and the product of dimensionality and hull relative volume. Nonetheless, the extreme ranks correspond to strikingly different chromatograms. Fig. 4 show the top ranked or “best” 3D chromatogram in the Gilar dataset, triplet 356, and the bottom ranked or “worst” one, triplet 127. The best 3D chromatogram is spread out across all 3D axes. In

contrast, the worst chromatogram has little spread and mostly lies within the diagonal region of the 3D axes.

The best column triplet, 356, is based on the C₁₈ silica-polymer hybrid, SEC and SCX (strong cation exchange) columns. However, the C₁₈ silica-polymer hybrid, bare silica and SEC columns (345) also do quite well. In addition, the C₁₈ silica-polymer hybrid, bare silica and SCX column columns (346) also do well across all 4 OMs. The successful combination of these surface chemistries is not surprising because peptides are known to separate well on hydrophobic and SCX phases [27]. These were the column phase chemistries used for the MudPIT proteomics system [1, 28], which is well-known for peptide characterization.

From the results at the bottom of Table 6 it appears that a number of chromatograms based on C₁₈ and phenyl stationary phases (columns 1 and 2) do not form good orthogonal separation systems. These phases are too similar. All 3D chromatograms built around these two phases appear to have poor orthogonality regardless of the third column, i.e., columns with C₁₈ silica polymer hybrid, bare silica, SEC, SCX and PFP (pentafluorophenyl) phases (columns 3 – 7) do not help. Poor orthogonality in any two dimensions is not corrected by the addition of a third column, whatever it is.

The PFP phase of column 7 is typically used for diastereomers, polar, basic, hydrophobic and electron-rich compounds, and it has many interesting properties [29]. However, in Table 6, PFP shows up in triplets having both the highest and lowest OMs, i.e., “7” appears in both high and low rankings. On further inspection of the subspace pairs given in Table 4, it appears that PFP has low orthogonality with C₁₈ and phenyl phases (pairs 17, 27, and 37) but has excellent orthogonality with SEC and SCX phases (pairs 57 and 67).

Even for higher dimensional separations where $n > 3$, it is highly desirable to analyze the OMs in the lower subspaces. Eq. (13) was used to show that 3 2D subspace OMs exist for a 3D system. For the hypothetical combination of $n = 4$ columns labeled a, b, c, and d, Eq. (13) shows there are 4 3D subspace OMs ($k = 4, i = 3$) for the triplets abc, abd, acd and bcd. Also, there are 6 2D subspace OMs ($k = 4, i = 2$) for the pairs ab, ac, ad, bc, bd, and cd. One can build this scheme for any number n of dimensions. These subspace relationships are useful as one can possibly eliminate a dimension when OMs indicate that it does not provide adequate orthogonality.

This suggests that for 3D separations, it is important that all three columns are orthogonal at the pair level. If one pair is not orthogonal it is certainly easier to just implement a 2D separation. Furthermore, it may be difficult to find more than two phase chemistries for certain solute families that are orthogonal enough to justify the difficulties of running a 3D separation. However, in Table 6, the four highest orthogonal separations utilize SEC; SEC is “universal” as a separation system and is unique in selectivity, because it is less chemistry-dependent as the mechanism is based on molecular size.

This promotes the idea that using mass spectrometry (MS) as a separating dimension and detector will offer an additional element of orthogonality. MS is both different and universal when compared to a particular surface phase, although it is still dependent on chemistry (mass and charge). This notion of a universal orthogonal element is quite important, because

as the number of columns increases it will be more difficult to find orthogonal phase systems.

There are many options here as to how to compartmentalize MS data. One can take an extrema value in mass to charge ratio or intensity and easily incorporate that as a dimension for both MS and MS/MS (with each MS as a separate dimension). However, MS data has much more to offer and one may envision specific algorithms for MS that determine a local value, e.g., taking the dimensionality of a mass spectral slice normalized suitably as if it were some type of pseudo-retention time for incorporation in the context of the analysis given here. Other methods to express the inherently multichannel data of MS into the retention time order of these analyses are indeed desirable and possible.

The consideration of 3D separations promotes the concept of running 1D separations on different columns as a process of methods development, as in Gilar's original study [7], and then ranking them as in Table 6. This development requires MS as a detector, because it is necessary to identify as many components as possible in an initial screening process. Solute tracking is essential in this scheme. The use of MS detection is becoming more routine as MS, in most cases, is the detector of choice for many analytical tasks.

Many practical obstacles remain in implementing 3D and higher dimensional separations. One of the least obvious is picking solvent systems that are compatible across all of the columns. This may seem trivial compared to the sampling problem, where all succeeding columns must be run at faster and faster speeds. However, solvent system compatibility can be difficult in 2D separations, and the addition of another column makes the problem even more difficult in 3D separations. One other difficulty that must be addressed is the dilution effect that additional columns place on the limit of detection [30]. This is a multiplicative effect that can limit the suitability of using more than two dimensions in a separation. However, this can be reduced with new higher performance columns.

As the number of separation dimensions increases, it becomes more difficult to keep the convex hull relative volume (and hypervolume for $n \geq 4$) anywhere near unity. As the modes of unique solute interaction with different phases become limited on the addition of more columns, it becomes progressively more difficult to utilize the space completely in an area (2D), volume (3D) or hypervolume (n D, with $n \geq 4$). This difficulty leads to the question: can one have high dimensionality within a small volume or hypervolume of the available separation space? The answer is, of course, extremely dependent on the sample's complexity in terms of functional groups, which can be selectively and independently retained by the third and higher stationary phases. This aspect may be the limitation on using higher dimensional separations, with regard to the change in retention with respect to the solute structure.

In regards to the reduction of relative volume and hypervolume with increasing numbers of columns, one also must question whether the product of dimensionality and relative hull volume will be a useful metric for higher dimensional spaces. We suspect the answer is no. While this product worked well in 2D [9,26], the use of metric products needs to be

reexamined in higher dimensions, as the relative occupied volume intuitively appears to be less.

Measurements of practical systems need to be made so that these OMs are further examined and relationships understood with other solute systems. This includes single OMs and products of OMs. The study of chromatographic selectivity and optimization places special value on these higher order OMs as well as the subspace OMs.

Despite the insights obtained in this study, we would be negligent if we did not acknowledge that limitations exist on OMs based solely on retention times. In all but the simplest separations, some retention times belong not to pure peaks but to multiplets of overlapping peaks. This causes a bias, regardless of metric. In addition, methods development based on sequences of 1D retention times is an idealistic but somewhat ironic consideration; if one can determine the 1D retention times of mixture components, one may not need a multidimensional separation. It is evident that any OM, however sophisticated, must consider at least in part the location of peaks (or at least signals), as measured by some metric such as retention time.

Nonetheless, we believe having the tools to explore orthogonal phase chemistries, such as the development of higher dimensional OMs given here, are first steps necessary to extend separation science into higher resolving technologies able to approach the goal of complete component resolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Part of this research was supported by the National Institutes of Health under grant R44-GM108122-02. We appreciate the review by Martin Gilar of Table 2 and adding some of the column data. Happy birthday Jim and we both look forward to any joint research you might like to do with us in higher dimensional space.

References

1. Cohen, SA., Schure, MR., editors. Multidimensional Liquid Chromatography: Theory and Applications in Industrial Chemistry and the Life Sciences. John Wiley; New York: 2008.
2. Moore AW, Jorgenson JW. Comprehensive three-dimensional separation of peptides using size exclusion chromatography/reversed phase liquid chromatography/optically gated capillary zone electrophoresis. *Anal Chem.* 1995; 67:3456–3463. [PubMed: 8686894]
3. Massart, DL., Kaufman, L. The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis. John Wiley & Sons, Inc; New York: 1983.
4. Liu Z, Patterson DG Jr, Lee ML. Geometric approach to factor analysis for the estimation of orthogonality and practical peak capacity in comprehensive two-dimensional separations. *Anal Chem.* 1995; 67:3840–3845.
5. Slonecker PJ, Li X, Ridgway TH, Dorsey JG. Informational orthogonality of two-dimensional chromatographic separations. *Anal Chem.* 1996; 68:682–689. [PubMed: 8999742]
6. Van Gysegheem E, Crosiers I, Gourvénec S, Massart DL, Vander Heyden Y. Determining orthogonal and similar chromatographic systems from the injection of mixtures in liquid chromatography-diode array detection and the interpretation of correlation coefficients color maps. *J Chromatogr A.* 2004; 1026:117–128. [PubMed: 14763739]

7. Gilar M, Olivova P, Daly AE, Gebler JC. Orthogonality of separation in two-dimensional liquid chromatography. *Anal Chem.* 2005; 77:6426–6434. [PubMed: 16194109]
8. Pourhaghighi MR, Karzand M, Girault HH. Orthogonality of two-dimensional separations based on conditional entropy. *Anal Chem.* 2011; 83:7676–7681. [PubMed: 21913671]
9. Schure MR, Davis JM. Orthogonal separations: comparison of orthogonality metrics by statistical analysis. *J Chromatogr A.* 2015; 1414:60–76. [PubMed: 26338213]
10. Murphy RE, Schure MR, Foley JP. Effect of sampling rate on resolution in comprehensive two-dimensional liquid chromatography. *Anal Chem.* 1998; 70:1585–1594.
11. Davis JM, Stoll DR, Carr PW. Effect of first-dimension undersampling on effective peak capacity in comprehensive two-dimensional separations. *Anal Chem.* 2008; 80:461–473. [PubMed: 18076145]
12. Cover, TM., Thomas, JA. *Elements of Information Theory.* John Wiley & Sons, Inc; New York: 1991.
13. Little EL, Jeansonne MS, Foley JP. Sequential multimodal elution for pseudomultidimensional liquid chromatography on a single column. *Anal Chem.* 1991; 63:33–44.
14. Giddings JC. Sample dimensionality: a predictor of order-disorder in component peak distribution in multidimensional separation. *J Chromatogr A.* 1995; 703:3–15. [PubMed: 7599743]
15. Van Gysegheem E, Jimidar M, Sneyers R, Redlich D, Verhoeven E, Massart DL, Vander Heyden Y. Orthogonality and similarity within silica-based reversed-phase chromatographic systems. *J Chromatogr A.* 2005; 1074:117–131. [PubMed: 15941047]
16. Neue UD, O’Gara JE, Méndez A. Selectivity in reversed-phase separations: influence of the stationary phase. *J Chromatogr A.* 2006; 1127:161–174. [PubMed: 16806238]
17. de Berg, M., Cheong, O., van Kreveld, M., Overmars, M. *Computational Geometry: Algorithms and Applications.* 3. Springer; Berlin: 2008.
18. Semard G, Peulon-Agasse V, Bruchet A, Bouillon JP, Cardinaël P. Convex hull: a new method to determine the separation space used and to optimize operating conditions for comprehensive two-dimensional gas chromatography. *J Chromatogr A.* 2010; 1217:5449–5454. [PubMed: 20633886]
19. Rutan SC, Davis JM, Carr PW. Fractional coverage metrics based on ecological home range for calculation of the effective peak capacity in comprehensive two-dimensional separations. *J Chromatogr A.* 2012; 1255:267–276. [PubMed: 22226455]
20. Barber CB, Dobkin DP, Huhdanpaa H. The Quickhull Algorithm for Convex Hulls. *ACM Transactions on Mathematical Software.* 1996; 22:469–483.
21. <http://www.qhull.org/>.
22. Schure MR. The dimensionality of chromatographic separations. *J Chromatogr A.* 2011; 1218:293–302. [PubMed: 21159346]
23. Floor and ceiling functions. https://en.wikipedia.org/wiki/Floor_and_ceiling_functions
24. Sturges HA. The choice of a class interval. *J Am Stat Assoc.* 1926; 21:65–66.
25. Gilar M, Fridrich J, Schure MR, Jaworski A. Comparison of orthogonality estimation methods for the two-dimensional separations of peptides. *Anal Chem.* 2012; 84:8722–8732. [PubMed: 22970653]
26. Donato P, Rigano F, Cacciola F, Schure M, Farnetti S, Russo M, Dugo P, Mondello L. Comprehensive two-dimensional liquid chromatography-tandem mass spectrometry for the simultaneous determination of wine polyphenols and target contaminants. *J Chromatogr A.* 2016; 1458:54–62. [PubMed: 27372414]
27. Mant, CT., Hodges, RS., editors. *High-Performance Liquid Chromatography of Peptides and Proteins: Separation, Analysis, and Conformation.* CRC Press; Boca Raton, FL: 1991.
28. Washburn MP, Wolters D, Yates JR III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol.* 2001; 19:242–247. [PubMed: 11231557]
29. Bell DS, Jones AD. Solute attributes and molecular interactions contributing to “U-shape” retention on a fluorinated high-performance liquid chromatography stationary phase. *J Chromatogr A.* 2005; 1073:99–109. [PubMed: 15909511]

30. Schure MR. Limit of detection, dilution factors, and technique compatibility in multidimensional separations utilizing chromatography, capillary electrophoresis, and field-flow fractionation. *Anal Chem.* 1999; 71:1645–1657.

Highlights

- Orthogonality metrics (OMs) are developed for n -dimensional (nD) chromatograms ($n \geq 3$)
- A comprehensive 3D separation and 35 groupings of 3 1D LCs are interpreted by 3D OMs
- The best and worst groupings of 3 1D LCs agree with rankings by 3D OMs
- Previous 2D OM based on information theory is modified and extended to nD systems
- Geometric and harmonic means of subspace 2D OMs correlate well with 3D OMs

Fig. 1.

Information diagrams for a) vectors X and Y , and b) vectors X , Y , and Z . Circles represent entropies of vectors.

Fig. 2.

A. The original data from Moore and Jorgenson, Fig. 4 in [2], showing the tryptic digest of ovalbumin using a 3D comprehensive separation. B. The reconstruction of this data using slices provided in [2]. The points are the mean retention times and the red segments are one standard deviation in the SEC dimension. C. The original data from Moore and Jorgenson, Fig. 7A, showing the summation and projection of the 3D data slices into a 2D CZE-RPLC space. This projection can be compared to the projection in B above. Reprinted with permission from A. W. Moore, J. W. Jorgenson, Comprehensive three-dimensional separation of peptides using size exclusion chromatography/reversed phase liquid chromatography/optically gated capillary zone electrophoresis, *Anal. Chem.* 67 (1995) 3456–3463. Copyright 1995 American Chemical Society.

Fig. 3.

a) Graphs of orthogonality O for 2D (○) and 3D (□) uniformly random coordinates, and of 2D (—) and 3D (---) coordinate density vs. number of coordinates. b) Graph of O vs. percentage of coordinates in diagonal bins, with other coordinates uniformly random elsewhere. Coordinate density: one. Square root of numbers of coordinates and 2D bins: 30 (○), 25 (□), 20 (◇), 15 (△), 10 (▽), 5 (⋈); cube root of numbers of coordinates and 3D bins: 12 (○), 10 (□), 9 (◇), 8 (△), 6 (▽), 4 (⋈). c) Graphs of 2D and 3D bins containing ordered checkerboard-like coordinates (filled) and uniformly random coordinates (open). One column of ordered coordinates is added at a time to 2D systems; one column and multiple rows are added at a time to 3D systems. d) Graph of O vs. percentage of coordinates ordered as in panel c), with other coordinates uniformly random elsewhere. Coordinate density: one. Square root of numbers of coordinates and 2D bins: 30 (●), 25 (■), 20 (◆), 15 (▲), 10 (▼), 5 (▴); cube root of numbers of coordinates and 3D bins: as in panel b).

Fig. 4. Graphs of “best” and “worst” chromatograms in Gilar dataset, as assessed by largest and smallest 3D OM. a) Best chromatogram, triplet 356. b) Worst chromatogram, triplet 127.

Table 1

Retention times of the Moore and Jorgenson 3D chromatogram [2], reconstructed by manual digitization.

point number	z, CZE (sec)	y, RPLC (min)	x, SEC (min)	σ_x (min)
1	1.51	5.66	215.82	13.22
2	1.51	5.33	224.28	3.67
3	1.54	5.11	236.98	13.22
4	1.58	5.26	215.82	20.41
5	1.48	4.66	226.40	6.35
6	1.45	4.38	226.40	6.35
7	1.49	4.32	222.17	13.22
8	1.54	4.69	249.68	15.98
9	1.55	4.40	239.10	27.68
10	1.57	4.65	245.45	16.80
11	1.57	4.89	247.57	13.22
12	1.63	4.78	215.82	3.67
13	1.61	4.91	222.17	9.70
14	1.71	5.15	234.87	3.67
15	1.75	5.11	251.80	6.35
16	1.74	5.59	209.47	3.67
17	1.83	5.46	228.52	13.22
18	1.91	5.63	228.52	3.67
19	1.96	5.29	234.87	9.70
20	1.94	4.96	232.75	16.80
21	1.89	4.69	226.40	6.35
22	1.90	4.59	236.98	13.22
23	1.74	4.36	228.52	3.67
24	1.78	4.12	230.63	3.67
25	1.55	3.71	217.93	3.67

Table 2

The chromatographic columns used in Gilar's peptide dataset.

Column number	Phase	pH	Manufacturer name	Material	d_p (μm)	Column length x i.d. (mm x mm)
1	C ₁₈	2.6	Atlantis (Waters)	Silica	3	150x2.1
2	Phenyl	2.6	Xterra (Waters)	Silica-polymer hybrid	3.5	150x2.1
3	C ₁₈	2.6	Xterra (Waters)	Silica-polymer hybrid	3.5	150x2.1
4	HILIC (bare silica)	4.5	Atlantis (Waters)	Silica	3	150x2.1
5	SEC (diol phase)	4.5	YMC, 60A	Silica	5	750x4.6 (3 columns in series)
6	SCX	3.25	Polysulfoethyl (PolyLC) 300Å	Silica	5	150x2.1
7	PFP	3.25	(Waters)	Silica	5	150x2.1

Table 3

OMs for Moore and Jorgenson 3D chromatogram. Dimensions: X, SEC; Y, RPLC; Z, CZE. A. For non-discretized OMs.

$1-r_{XY}^2$	$1-r_{XZ}^2$	$1-r_{YZ}^2$	V_{H3D}	A_{HXY}	A_{HYZ}	A_{HYZ}
0.933	0.988	0.986	0.310	0.714	0.659	0.599

B. For discretized OMs.								
Bins, 3D	D_{3D}	D_{XY}	D_{XZ}	D_{YZ}	O_{3D}	O_{XY}^o	O_{XZ}^o	O_{YZ}^o
2	3.000	1.754	1.722	1.758	0.988	0.980	0.993	0.992
3	2.406	1.754	1.722	1.758	0.668	0.902	0.796	0.909
						0.990	0.996	0.996
						0.949	0.887	0.953

Table 4

OMs for the Gilar 3D dataset. Triplet numbers in first table column are numbers assigned to LC columns in Table 2.

Triplet XYZ	$1-r_{xy}^2$	$1-r_{xz}^2$	$1-r_{yz}^2$	$V_{H,3D}$	$A_{H,XY}$	$A_{H,XZ}$	$A_{H,YZ}$	D_{3D}	D_{XY}	D_{XZ}	D_{YZ}	O_{3D}	O_{XY}^o	O_{XZ}^o	O_{YZ}^o	O_{XY}	O_{XZ}	O_{YZ}	$O_{3D} \cdot V_{H,3D}$
123	0.014	0.539	0.582	0.050	0.108	0.613	0.659	2.138	1.214	1.639	1.649	0.617	0.268	0.821	0.833	0.435	0.903	0.909	0.031
124	0.014	0.999	0.997	0.055	0.108	0.694	0.695	2.186	1.214	1.724	1.708	0.653	0.268	0.979	0.976	0.435	0.990	0.988	0.036
125	0.014	0.729	0.710	0.054	0.108	0.668	0.713	2.124	1.214	1.645	1.627	0.633	0.268	0.860	0.859	0.435	0.928	0.926	0.034
126	0.014	0.952	0.949	0.063	0.108	0.715	0.715	2.133	1.214	1.645	1.633	0.615	0.268	0.849	0.821	0.435	0.923	0.905	0.039
127	0.014	0.126	0.160	0.025	0.108	0.311	0.355	1.881	1.214	1.414	1.455	0.535	0.268	0.496	0.563	0.435	0.671	0.722	0.013
134	0.539	0.999	0.923	0.329	0.613	0.694	0.717	2.524	1.639	1.724	1.671	0.825	0.821	0.979	0.944	0.903	0.990	0.971	0.272
135	0.539	0.729	0.999	0.319	0.613	0.668	0.758	2.503	1.639	1.645	1.693	0.728	0.821	0.860	0.974	0.903	0.928	0.987	0.232
136	0.539	0.952	0.994	0.358	0.613	0.715	0.772	2.464	1.639	1.645	1.680	0.784	0.821	0.849	0.958	0.903	0.923	0.979	0.280
137	0.539	0.126	0.390	0.135	0.613	0.311	0.523	2.245	1.639	1.414	1.594	0.666	0.821	0.496	0.710	0.903	0.671	0.831	0.090
145	0.999	0.729	0.689	0.314	0.694	0.668	0.591	2.478	1.724	1.645	1.608	0.785	0.979	0.860	0.858	0.990	0.928	0.925	0.247
146	0.999	0.952	0.810	0.382	0.694	0.715	0.655	2.493	1.724	1.645	1.626	0.813	0.979	0.849	0.881	0.990	0.923	0.939	0.310
147	0.999	0.126	0.999	0.166	0.694	0.311	0.695	2.354	1.724	1.414	1.690	0.759	0.979	0.496	0.979	0.990	0.671	0.989	0.126
156	0.729	0.952	0.974	0.362	0.668	0.715	0.652	2.476	1.645	1.645	1.592	0.827	0.860	0.849	0.862	0.928	0.923	0.927	0.300
157	0.729	0.126	0.864	0.160	0.668	0.311	0.653	2.301	1.645	1.414	1.651	0.688	0.860	0.496	0.895	0.928	0.671	0.943	0.110
167	0.952	0.126	0.999	0.139	0.715	0.311	0.756	2.285	1.645	1.414	1.634	0.640	0.849	0.496	0.901	0.923	0.671	0.946	0.089
234	0.582	0.997	0.923	0.352	0.659	0.695	0.717	2.517	1.649	1.708	1.671	0.818	0.833	0.976	0.944	0.909	0.988	0.971	0.288
235	0.582	0.710	0.999	0.357	0.659	0.713	0.758	2.469	1.649	1.627	1.693	0.725	0.833	0.859	0.974	0.909	0.926	0.987	0.259
236	0.582	0.949	0.994	0.387	0.659	0.715	0.772	2.464	1.649	1.633	1.680	0.792	0.833	0.821	0.958	0.909	0.905	0.979	0.307
237	0.582	0.160	0.390	0.150	0.659	0.355	0.523	2.261	1.649	1.455	1.594	0.692	0.833	0.563	0.710	0.909	0.722	0.831	0.104
245	0.997	0.710	0.689	0.334	0.695	0.713	0.591	2.443	1.708	1.627	1.608	0.775	0.976	0.859	0.858	0.988	0.926	0.925	0.259
246	0.997	0.949	0.810	0.384	0.695	0.715	0.655	2.479	1.708	1.633	1.626	0.814	0.976	0.821	0.881	0.988	0.905	0.939	0.313
247	0.997	0.160	0.999	0.190	0.695	0.355	0.695	2.362	1.708	1.455	1.690	0.773	0.976	0.563	0.979	0.988	0.722	0.989	0.146
256	0.710	0.949	0.974	0.379	0.713	0.715	0.652	2.449	1.627	1.633	1.592	0.817	0.859	0.821	0.862	0.926	0.905	0.927	0.310
257	0.710	0.160	0.864	0.187	0.713	0.355	0.653	2.287	1.627	1.455	1.651	0.731	0.859	0.563	0.895	0.926	0.722	0.943	0.137
267	0.949	0.160	0.999	0.155	0.715	0.355	0.756	2.277	1.633	1.455	1.634	0.658	0.821	0.563	0.901	0.905	0.722	0.946	0.102
345	0.923	0.999	0.689	0.373	0.717	0.758	0.591	2.573	1.671	1.693	1.608	0.855	0.944	0.974	0.858	0.971	0.987	0.925	0.319

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Triplet XYZ	$1-r_{XY}^2$	$1-r_{XZ}^2$	$1-r_{YZ}^2$	V_{H3D}	A_{HXY}	A_{HYZ}	D_{3D}	D_{XY}	D_{XZ}	D_{YZ}	O_{3D}	O_{XY}^o	O_{XZ}^o	O_{YZ}^o	O_{XY}	O_{XZ}	O_{YZ}	$O_{3D} \cdot V_{H3D}$
346	0.923	0.994	0.810	0.404	0.717	0.772	2.530	1.671	1.680	1.626	0.826	0.944	0.958	0.881	0.971	0.979	0.939	0.334
347	0.923	0.390	0.999	0.262	0.717	0.523	2.435	1.671	1.594	1.690	0.773	0.944	0.710	0.979	0.971	0.831	0.989	0.202
356	0.999	0.994	0.974	0.438	0.758	0.772	2.574	1.693	1.680	1.592	0.873	0.974	0.958	0.862	0.987	0.979	0.927	0.382
357	0.999	0.390	0.864	0.282	0.758	0.523	2.386	1.693	1.594	1.651	0.730	0.974	0.710	0.895	0.987	0.831	0.943	0.206
367	0.994	0.390	0.999	0.316	0.772	0.523	2.483	1.680	1.594	1.634	0.793	0.958	0.710	0.901	0.979	0.831	0.946	0.251
456	0.689	0.810	0.974	0.270	0.591	0.655	2.395	1.608	1.626	1.592	0.753	0.858	0.881	0.862	0.925	0.939	0.927	0.203
457	0.689	0.999	0.864	0.319	0.591	0.695	2.478	1.608	1.690	1.651	0.816	0.858	0.979	0.895	0.925	0.989	0.943	0.261
467	0.810	0.999	0.999	0.391	0.655	0.695	2.505	1.626	1.690	1.634	0.826	0.881	0.979	0.901	0.939	0.989	0.946	0.323
567	0.974	0.864	0.999	0.385	0.652	0.653	2.486	1.592	1.651	1.634	0.840	0.862	0.895	0.901	0.927	0.943	0.946	0.323

Table 5

Correlation of 3D OMs in Gilar dataset with geometric mean $\overline{OM}_G$ and harmonic mean $\overline{OM}_H$ of 2D subspace OMs. ij represents XY , XZ , and YZ . Pearson r is reported.

3D OM	2D OM for columns i and j	r , 3D OM vs. $\overline{OM}_G$	r , 3D OM vs. $\overline{OM}_H$
D_{3D}	$1-r_{ij}^2$	0.943	0.916
O_{3D}	$1-r_{ij}^2$	0.926	0.907
D_{3D}	D_{ij}	0.972	0.973
O_{3D}	D_{ij}	0.885	0.887
D_{3D}	O_{ij}	0.979	0.968
O_{3D}	O_{ij}	0.913	0.904
$V_{H,3D}$	$A_{H,ij}$	0.956	0.925

Table 6

Ranking of ten triplets in Gilar dataset with highest 3D OM_s (in descending order) and lowest 3D OM_s (in ascending order). For easy comparison, a given triplet designation (e.g., 356) has the same color.

Rank	O_{3D}	D_{3D}	$V_{H,3D}$	$D_{3D} \cdot V_{H,3D}$
1	356 (0.873)	356 (2.574)	356 (0.438)	356 (1.127)
2	345 (0.855)	345 (2.573)	346 (0.404)	346 (1.022)
3	567 (0.840)	346 (2.530)	467 (0.391)	467 (0.979)
4	156 (0.827)	134 (2.524)	236 (0.387)	345 (0.960)
5	346 (0.826)	234 (2.517)	567 (0.385)	567 (0.957)
6	467 (0.826)	467 (2.505)	246 (0.384)	236 (0.954)
7	134 (0.825)	135 (2.503)	146 (0.382)	146 (0.952)
8	234 (0.818)	146 (2.493)	256 (0.379)	246 (0.952)
9	256 (0.817)	567 (2.486)	345 (0.373)	256 (0.928)
10	457 (0.816)	367 (2.483)	156 (0.362)	156 (0.896)
35	127 (0.535)	127 (1.881)	127 (0.025)	127 (0.047)
34	126 (0.615)	125 (2.124)	123 (0.050)	123 (0.107)
33	123 (0.617)	126 (2.133)	125 (0.054)	125 (0.115)
32	125 (0.633)	123 (2.138)	124 (0.055)	124 (0.120)
31	167 (0.640)	124 (2.186)	126 (0.063)	126 (0.134)
30	124 (0.653)	137 (2.245)	137 (0.135)	137 (0.303)
29	267 (0.658)	237 (2.261)	167 (0.139)	167 (0.318)
28	137 (0.666)	267 (2.277)	237 (0.150)	237 (0.339)
27	157 (0.688)	167 (2.285)	267 (0.155)	267 (0.353)
26	237 (0.692)	257 (2.287)	157 (0.160)	157 (0.368)