analytical chemistry

Article

pubs.acs.org/ac

Comparison of Orthogonality Estimation Methods for the Two-Dimensional Separations of Peptides

Martin Gilar,*[,†] Jessica Fridrich,[‡] Mark R. Schure,[§] and Aleksander Jaworski[⊥]

[†]Waters Corporation, 34 Maple St., Milford, Massachusetts 01757, United States

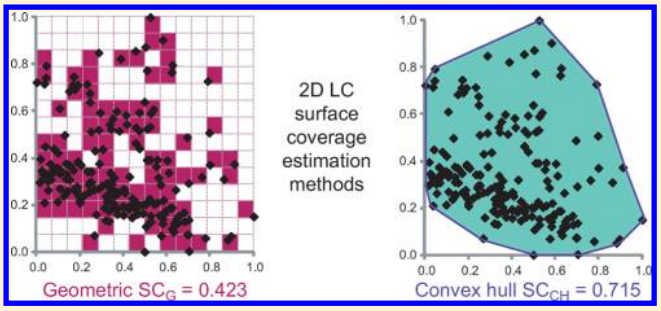
[‡]Department of Electrical and Computer Engineering, SUNY Binghamton, Binghamton, New York 13902-6000, United States

[§]Theoretical Separation Science Laboratory, Kroungold Analytical, Inc., 1299 Butler Pike, Blue Bell, Pennsylvania 19422, United States

[⊥]51 Palomino Drive, Franklin, Massachusetts 02038, United States

Supporting Information

ABSTRACT: In two-dimensional chromatography, the orthogonality of separation is important for achieving high peak capacity. In this paper, a number of different metrics are compared as measures of orthogonality. Six peptide elution data sets acquired on different stationary phases are plotted against reversed phase retention data and examined as two-dimensional chromatographic pairs. The data, including six in silico prepared data pairs, are utilized to challenge and compare selected orthogonality metrics. The metrics include correlation coefficients, mutual information, box-counting dimensionality, and surface fractional coverage with different hulls. Although correlation coefficients were found to be less suited for the intended purpose, other methods can provide a suitable measure of orthogonality. The presented results are discussed in terms of method utility, simplicity, and applicability for statistically small sets of chromatographic data. Two of the methods, box counting dimensionality and fractional coverage, were found to be mathematically related.



In two-dimensional chromatography, the concept of orthogonality is often discussed as a quantitative measure of the efficacy of separation space utilization. However, a mathematically rigorous definition of orthogonal separations has yet to be offered. The term orthogonality is used differently from statistics to mathematics, and this has been articulated previously in the context of two-dimensional separations.[1] For example, in mathematics two lines are orthogonal if they form right angles at their point of intersection.[2] In vector analysis, two vectors are orthogonal if the inner product is zero.[3] In statistics, independent variables are said to be orthogonal if they are uncorrelated.[4] In computer science, orthogonal range trees are data structures used in database searching,[5] and orthogonal refers to rectangular ranges that are aligned with coordinate axes.

Statisticians study the correlation of two variables. From a multidimensional chromatography perspective, a good two-dimensional chromatogram has little correlation in the two retention time pairs and adequately fills the two-dimensional space with a minimal amount of clumping of retention times in specific regions. It is in this context that one of the earliest studies of orthogonality used data from a number of single-dimension techniques, formed two-dimensional data sets, and calculated a number of metrics based on information theory.[6] These metrics included the information entropy, the mutual information, the synentropy or cross information, and the information similarity.

A wide variety of orthogonality metrics have been proposed. Many of these have been recently described by Carr and co-workers in a thorough review of two-dimensional liquid chromatography[7] up to 2012. However, active research continues in this area. For example, the fractional coverage metric continues to be developed with different hull conditions.[8] Although the fractional surface coverage in a stated area can be utilized to estimate peak capacity, it is also useful as an orthogonality metric based on the idea that orthogonality is related to the effective utilization of the separation space (the greater is the fractional area coverage, the higher should be the orthogonality). In addition, metrics with an information theoretic basis,[9] different correlation coefficients,[10] and the box counting (fractal) dimension[11] continue to be developed for orthogonality measurement in two-dimensional chromatography.

In a perfectly ordered two-dimensional chromatogram, as embodied in the two-dimensional peak capacity concept exemplified by Giddings,[12] all peaks are equispaced, and it is

Received: July 17, 2012
Accepted: September 12, 2012
Published: September 12, 2012

   dx.doi.org/10.1021/ac3020214 | *Anal. Chem.* 2012, 84, 8722–8732

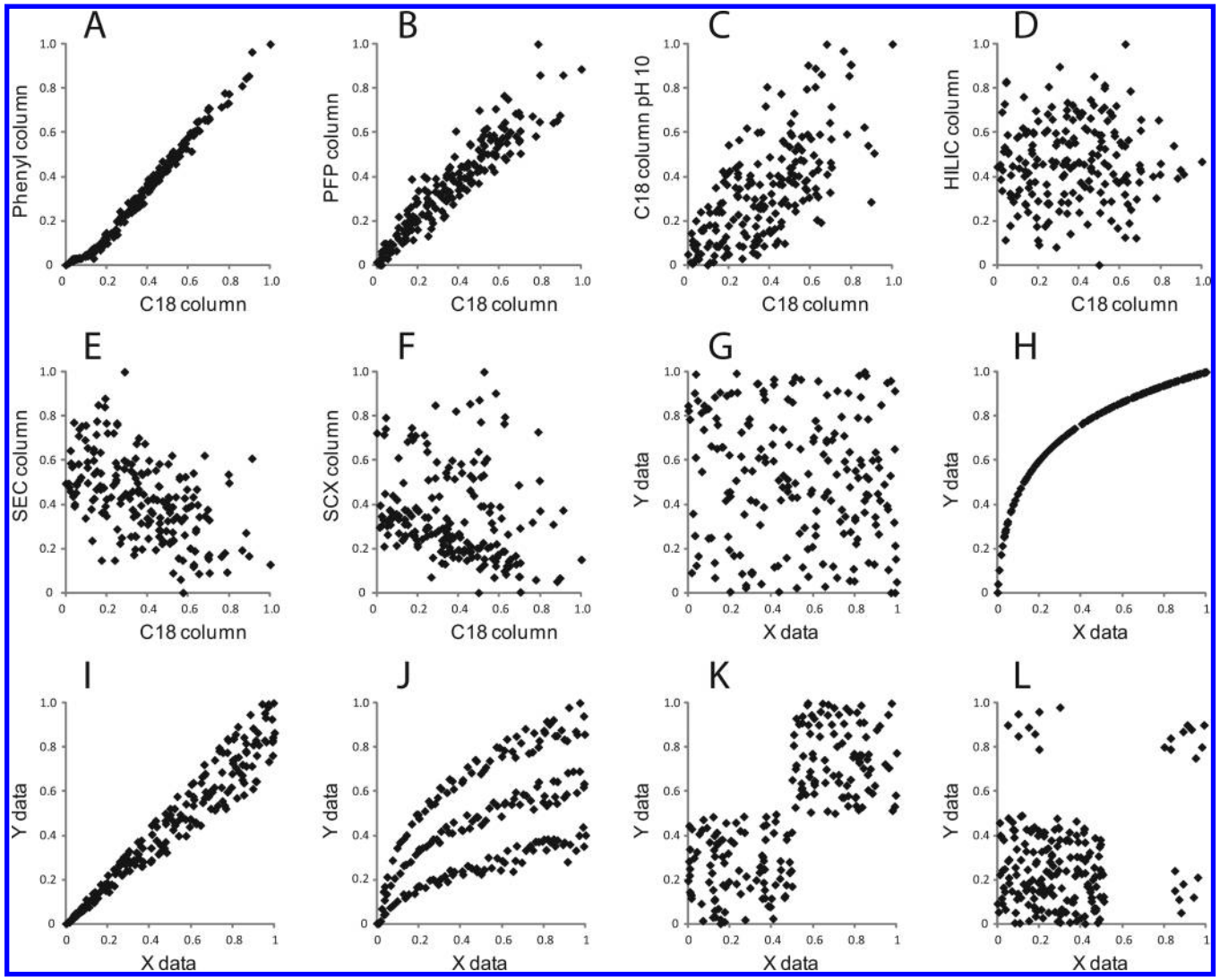


Figure 1. Normalized sets of 2D data, each containing 196 data points. Panels A–F represent peptide samples; panels G–L were generated in silico with the goal of challenging the orthogonality estimation methods.

expected that the maximum number of single component peaks (SCPs) can be put into that separation space. Obviously, few chromatograms if any utilize this rectangular space as efficiently as this idealization. Hence, a quantitative measure of how effective the separation space is utilized is highly desirable. This measure should be useful in choosing column pairs and in selecting optimal gradient conditions.[7,11,13] However, the metric used to measure the effective use of the separation space in 1D, 2D, and nD separations is elusive because different metrics measure different attributes of the efficacy problem. In addition, a global metric, such as surface coverage,[14−16] information theory metrics,[6] correlation coefficients,[17,18] and the fractal dimension,[11] does not consider short length scale effects, such as nearest neighbor statistics, but rather is averaged over the whole separation space. Clearly, no one metric may therefore be useful as the "end all" measurement of orthogonality.

In that vein, several methods are investigated to evaluate peptide separations including a variety of correlation coefficients, mutual information, the box counting (fractal) dimension, and surface fractional coverage obtained from the rectangular and the convex hull. The concept of geometric surface coverage (SC) proposed by Gilar et al.[14] received some criticism in the literature.[8,15,19] In this paper, we clarify the concept of SC, compare this method to other metrics, and investigate the robustness of orthogonality measurement for small sets of data, mimicking the practical scenario when the orthogonality is estimated with samples containing a limited number of components. The implications of chromatographic undersampling on band broadening and achievable peak capacity are left aside; these aspects of multidimensional separations were discussed in the recent review.[7] Furthermore, the effect of zone broadening on this data set, which may limit the resolution of overlapping peaks when mass spectrometry is not used as a detector, will be discussed in a later publication.

■ EXPERIMENTAL SECTION

The retention data of 196 common peptides were obtained as described earlier.[14] Chromatographic columns and LC-UV-MS conditions were described in refs 14 and 19 with the exception of hydrophilic interaction chromatography (HILIC) and modified strong cation exchange (SCX) conditions, which are described here.

Materials and Reagents. Formic acid (FA) and trifluoroacetic acid (TFA) were purchased from Sigma-Aldrich (St.

dx.doi.org/10.1021/ac3020214 | Anal. Chem. 2012, 84, 8722−8732

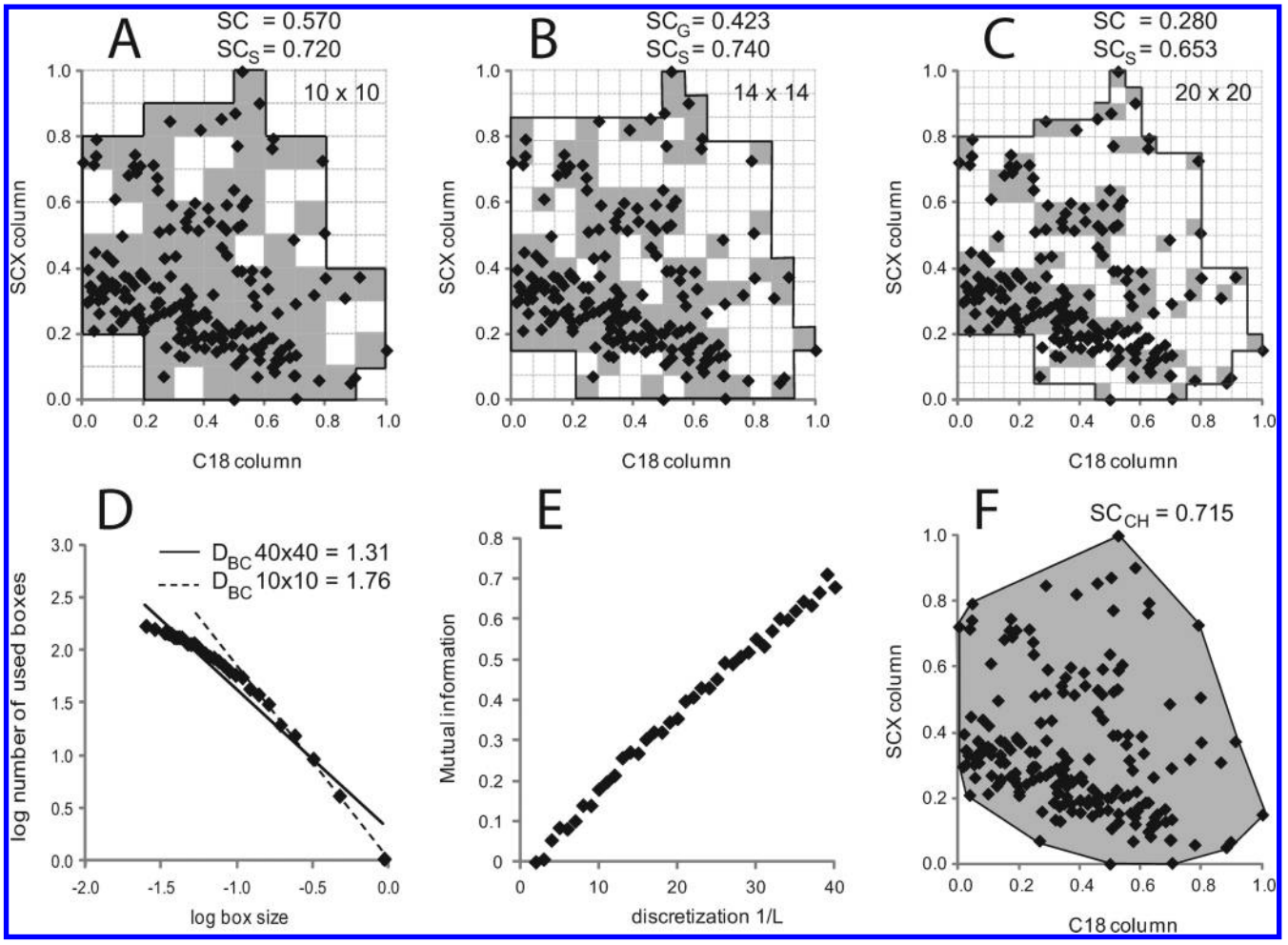


Figure 2. Graphical representation of methods used for description of 2D orthogonality. (A–C) Gilar and Stoll surface coverage; (D) box-counting dimensionality; (E) mutual information; (F) convex hull surface coverage methods. The value of D_{BC} measured as a slope of linear regression (panel D) depends on the data range used for regression (entire range or a linear subset).

Louis, MO, USA). HPLC grade acetonitrile (MeCN) and ammonium hydroxide were purchased from J.T. Baker (Phillipsburg, NJ, USA). A Milli-Q system (Millipore, Bedford, MA, USA) was used to prepare deionized water (18 MΩ cm) for HPLC mobile phases. MassPREP protein tryptic digestion standards of enolase, alcohol dehydrogenase (ADH), phosphorylase b, hemoglobin, bovine serum albumin (BSA), and MassPREP peptide standards were obtained from Waters (Milford, MA, USA).

LC-UV-MS Data. Experiments were conducted with an Atlantis silica HILIC (Waters, Milford, MA, USA), 150 × 2.1 mm, 3 μm column with gradient conditions of 90–48% MeCN in 50 min, constant 10 mM ammonium formate buffer. In addition, a SCX PolySULFOETHYL A 150 × 2.1 mm, 5 μm, 300 Å column (PolyLC Inc., Columbia, MD, USA) was utilized at pH 4.5 and eluted with a gradient of 40–300 mM ammonium formate, pH 3.25 in 40 min. The mobile phase contained 25% MeCN. All 196 peptides were extracted from total ion current chromatograms using their expected m/z value, and the retention times were recorded. The MS signal of the peptides was weak in the SCX mode but sufficient to obtain retention times from the SCX MS chromatograms.

Figure 1A–F summarizes the data sets considered in this study for 196 tryptic peptides. A C_{18} column (pH 2.6) is used as the x-axis in all cases, representing typical second dimension separation experiments. The data were plotted against various LC separation modes/columns (y-axis), including a phenyl reversed-phase column (Figure 1A), a pentafluoro phenyl (PFP) reversed-phase column (Figure 1B), a C_{18} column operated at pH 10 (Figure 1C), a HILIC column (Figure 1D), a size exclusion chromatography (SEC) column (Figure 1E), and a SCX column (Figure 1F).

Artificial Chromatograms. Six additional 2D chromatographic data were generated in silico (Figure 1G–L). Although created in silico, the data represent useful scenarios. Figure 1G is a random distribution of 196 components in a 2D space representing an orthogonal separation system. Figure 1H shows a nonlinear, nonorthogonal 2D separation often encountered in 2DGC or the 2DLC separation of polymers. Another variant of this 2D scenario termed a "bananagram"[20] is shown in Figure 1J. Figure 1I is a poorly orthogonal scenario similar to Figure 1B. Figure 1K shows two dominant clusters in separation space, and Figure 1L contains one separation cluster with a small number of outliers. The rationale for creating artificial data is to challenge and contrast the methods for orthogonality estimation.

Data Processing. Both experimentally acquired and in silico generated data were normalized in both separation dimensions using eq 1. The t_i is retention time of i-th component; t_{min} and t_{max} represent retention times of the first

dx.doi.org/10.1021/ac3020214 | Anal. Chem. 2012, 84, 8722–8732

Table 1. 2D Orthogonality Calculated Using Selected Methods for Data Sets Shown in Figure 1

data label[a]	2D system	Pearson R^2	Spearman R^2	Kendall R^2	SC_G	SC_G trend[b]	SC_{CH}	D_{BC}[c]	MI
A	phenyl × C_{18}	0.986	0.993	0.916	0.143	0.148	0.108	1.144	0.776
B	PFP × C_{18}	0.874	0.894	0.642	0.260	0.262	0.311	1.301	0.511
C	C18 × C_{18}	0.461	0.470	0.243	0.408	0.427	0.613	1.537	0.291
D	HILIC × C_{18}	0.001	0.001	0.000	0.510	0.487	0.693	1.648	0.152
E	SEC × C_{18}	0.271	0.281	0.134	0.434	0.431	0.668	1.549	0.203
F	SCX × C_{18}	0.048	0.107	0.057	0.423	0.423	0.715	1.531	0.272
G	random	0.014	0.014	0.007	0.643	0.628	0.966	1.792	0.222
H	curve	0.856	1.000	1.000	0.138	0.137	0.259	1.102	0.742
I	angle	0.931	0.934	0.726	0.281	0.274	0.197	1.370	0.632
J	3 trends	0.387	0.373	0.209	0.418	0.415	0.548	1.543	0.383
K	2 clusters	0.535	0.556	0.247	0.439	0.440	0.661	1.654	0.386
L	outliers	0.037	0.000	0.000	0.347	0.356	0.906	1.499	0.146

[a]Consistent with Figure 1. [b]Calculated from trend as illustrated in Figure 6C using binning from 10 × 10 to 25 × 25. [c]Calculated for area binning from 2 × 2 to 14 × 14.

and last eluting component in each set. All normalized retention times are between 0 and 1.

$$t_{i(norm)} = \frac{t_i - t_{min}}{t_{max} - t_{min}} \quad (1)$$

Orthogonality Calculations. Geometric surface coverage described by Gilar et al.,[14] the Gilar-Stoll surface coverage concept proposed elsewhere[8,15] (we will refer to them as SC_G and SC_S, respectively), and the box-counting dimensionality (D_{BC}) were calculated using an Excel spreadsheet, available in the Supporting Information S1 (ac3020214_si_001.zip). The data files used in this study are provided as Supporting Information, S2 (ac3020214_si_002.xlsx). In a recent paper, the Pearson, Spearman, and Kendall correlation coefficients are summarized,[10] and we use the same basic formulas in this paper implemented in a Matlab (Mathworks, Natick, MA) routine. Similarly, the convex hull[8,16] surface coverage (SC_{CH}) and mutual information (MI) were calculated using programs written in Matlab. Equations for MI calculation are defined in Supporting Information S3 (ac3020214_si_003.pdf); an example of numerical MI solution is given in Supporting Information S1.

Surface Coverage Background. Unlike the other metrics, the fractional surface coverage has a characteristically direct relationship to the overall 2D peak capacity and could be substituted for the fractional coverage f_c in eq 2, used to compute $n^*_{c,2D}$, the achievable 2D peak capacity:[7]

$$n^*_{c,2D} = \frac{{}^1n_c \times {}^2n_c \times f_c}{<\beta>} \quad (2)$$

where ${}^1n_c$ and ${}^2n_c$ are the peak capacities in the first and second dimension and $<\beta>$ is a factor[7] used to describe the additional loss of efficiency due to first dimension sampling in a comprehensive 2D experiment. For an ideally ordered two-dimensional chromatogram, as postulated by Giddings,[12,21] the surface coverage f_c is one. The achievable peak capacity $n^*_{c,2D}$ characterizes the separation potential of a 2D system rather than giving the number of actually resolved peaks. It appears intuitively that orthogonality should increase with an increase in the portion of the utilized separation area.[14−16,20,21]

Previous work[14] in developing the surface coverage metric, SC_G, has been criticized[19] for giving biased results that are strongly tied to the sample component number. Other reports[7,15] suggest that SC_G depends on bin size, as shown in Figure 2A−C. Although the surface coverage does change with bin size, a clarification of this point is necessary. The original method[14] assumed that 2D retention data are normalized by eq 1, and the separation space is divided into square bins, as shown in Figure 2A. The binning discretization is chosen so that the number of bins B corresponds to the number of single components N. For example, the separation space with 196 SCPs is divided into 14 × 14 bins (=196). The SC_G value is then obtained by determining the cumulative area of all bins containing data points and dividing by the total available separation space. Hence, the binning is not related to the actual peak capacity in each LC dimension.

In principle, it is possible to estimate SC for $N \neq B$, but all 2D systems must be compared with the same sample and binning. Extremes in separation space saturation may result in poor assessment of SC, as discussed below. For example, when using a large bin number and a small number of components, all peaks are "resolved" into one bin and a further increase in binning is meaningless. In the opposite extreme, when the number of bins is low, SC approaches the value of one for all 2D systems (extreme separation space saturation). We use SC_G notation for situations when $N = B$.

SC_G assumes a random distribution of peaks across the 2D separation area. The probability of peaks residing in bins is expressed in eqs 3 and 4, where M is the number of bins containing peaks; eqs 3 and 4 are derived in Supporting Information S3.

$$M = B - B\left(1 - \frac{1}{B}\right)^N \quad (3)$$

$$\frac{M}{B} = SC_G = 1 - \left(1 - \frac{1}{B}\right)^N \cong 1 - \left(1 - \frac{1}{e}\right) = 0.63 \quad (4)$$

when $N = B$ and an orthogonal (random) data set is assumed, eq 4 is obtained from eq 3. Although multiple peaks may occupy the same bin, some bins remain empty. The SC_G for $N = B$ and orthogonal 2D separation is 0.63.

When the number of components in a typical real sample is unknown, it may be difficult choosing an appropriate binning[8] for this case. However, when using mass spectrometers for detection, components may be detected even when coeluting in certain cases. Their identity does not have to be known; they can be treated as mass-retention time pairs.[22,23]

dx.doi.org/10.1021/ac3020214 | Anal. Chem. 2012, 84, 8722−8732

The choice of a representative sample and its number of components is critical for assessment of separation orthogonality. We believe that the sample used in this study represents the general retention behavior of tryptic peptides. This assumption is supported by reports that found similar retention patterns as those shown in Figure 1A−F for unrelated tryptic peptide samples.[24−27] The impact of sample selection and the number of components on estimated orthogonality will be discussed further below.

RESULTS AND DISCUSSION

Comparison of Orthogonality Metrics. Figure 2 illustrates the unique aspects of the selected methods of

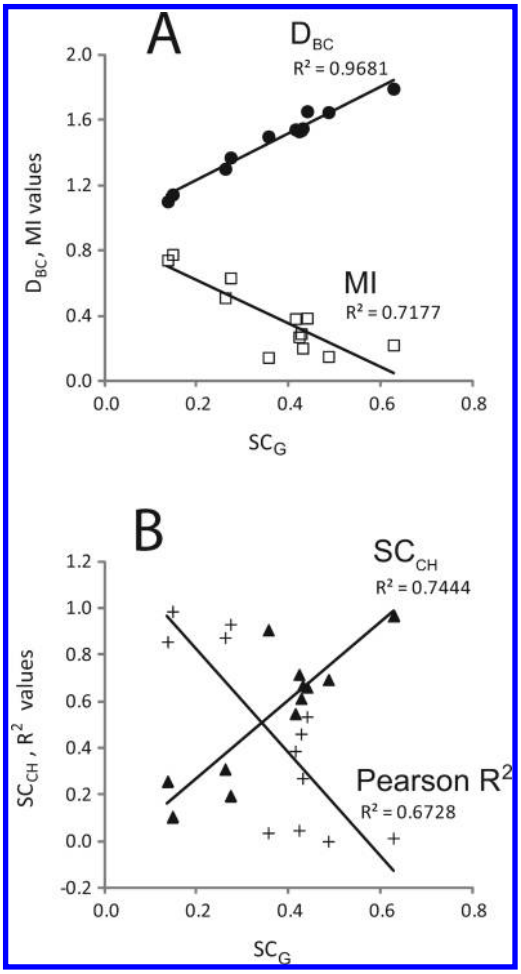


Figure 3. Correlation between orthogonality values calculated for 2D data sets shown in Figure 1. (A) SC_G versus D_{BC} and MI; (B) SC_G versus SC_{CH} and Pearson R^2.

orthogonality estimation. The principles of SC_G and SC_S methods are demonstrated in Figure 2A−C. The bin dimensions for a selected 2D system (Figure 1F, 196 components) are shown for 10 × 10, 14 × 14, and 20 × 20 bins. The bins highlighted by gray contain the eluting peptides; they were summed and the resulting surface coverage was 0.570, 0.423, and 0.280, respectively (geometric SC method). Figure 2B shows the correct SC_G estimate, adhering to the rule where N, the number of components, is equal to B, the number of bins. Stoll proposed a variant of the surface coverage method,[15] SC_S, that includes all of the bins within the area framed by the black line in Figure 2A−C. The SC_S value is less dependent on the discretization of separation space but includes area not utilized for separation. This may lead to overestimation of surface coverage, especially for 2D systems with clustered components (e.g., Figure 1F) and for scenarios with chromatographic outliers (Figure 1L). The Stoll method[15] uses a rectangular hull and is in principle similar to the convex hull method,[8,16] SC_{CH}, mentioned previously. Therefore, we utilize the SC_{CH} for method comparison in this manuscript.

The D_{BC} method is illustrated in Figure 2D. Similarly to SC_G and SC_S methods, the separation area is divided into square bins of different sizes, and the log of the number of boxes occupied by peaks (log M; y-axis) is plotted against the log of the box size (log L; axis x). If the distribution of peaks in 2D space or time (or any number of chromatographic dimensions for that matter) has fractal properties, the resulting relationship is a straight line with a slope corresponding to the fractal dimension D multiplied by −1. Even if the spacing of the peaks is not fractal, a scaling relationship which fits this data is operationally useful as it describes the length and time scales of the peak spacing.[11] Figure 2D suggests that for chromatographic systems the relationship is linear only in a certain range, approximately up to box size 0.1, i.e., log box size of −1.0.

At finer discretization levels, the boxes are smaller than the distance between components and will not find any more number of components. Therefore, these plots plateau when the number of utilized boxes approaches the number of components as shown in Figure 2D. This nonlinear behavior was observed for all data sets shown in Figure 1 and is known to be a characteristic of the box counting method in other applications. Hence, a proper choice of the linear portion of the plot has to be made for D_{BC} estimation. For example, the solid line in Figure 2D shows a linear trend across the entire data range, indicating dimensionality D_{BC} = 1.31, although the slope of the dashed line calculated for a more linear subset of data (log L range from −1.0 to 0) of the relationship has fractal dimension (slope) D_{BC} = 1.76. We empirically observed that a good linear fit ($R^2 > 0.99$) is obtained up to discretization 14 × 14; 1 × 1 discretization (one bin, always occupied) leads in log M = 0 in all cases. The data point log L = 0/log M = 0 represents an outlier, decreasing the linearity of relationship. In further discussion, we report D_{BC} values calculated for discretization range from 2 × 2 to 14 × 14.

The mutual information (MI) method[6,10] is a measure of the statistical dependency between two data sets. MI returns a lower value for independent data sets than for statistically dependent sets. For small data sets (all LC data sets in Figure 1), its value strongly depends on discretization as illustrated by Figure 2E. This suggests that MI values can be used for measurement and comparison of orthogonality of 2D separations only when using a defined test sample and discretization.

The results of using the convex hull method[8,16] are illustrated in Figure 2F. This method may be visualized as a hypothetical rubber band snapped around the data points in 2D space. The area within the resulting shape (gray in Figure 2F) is used for calculation of surface coverage. This method does not use discretization, but similar to the SC_S method, it includes the sparsely populated area into a usable separation space. One can envision that the method will be sensitive to outlying components; few outliers or even one data point may significantly alter the results.[16]

dx.doi.org/10.1021/ac3020214 | Anal. Chem. 2012, 84, 8722−8732

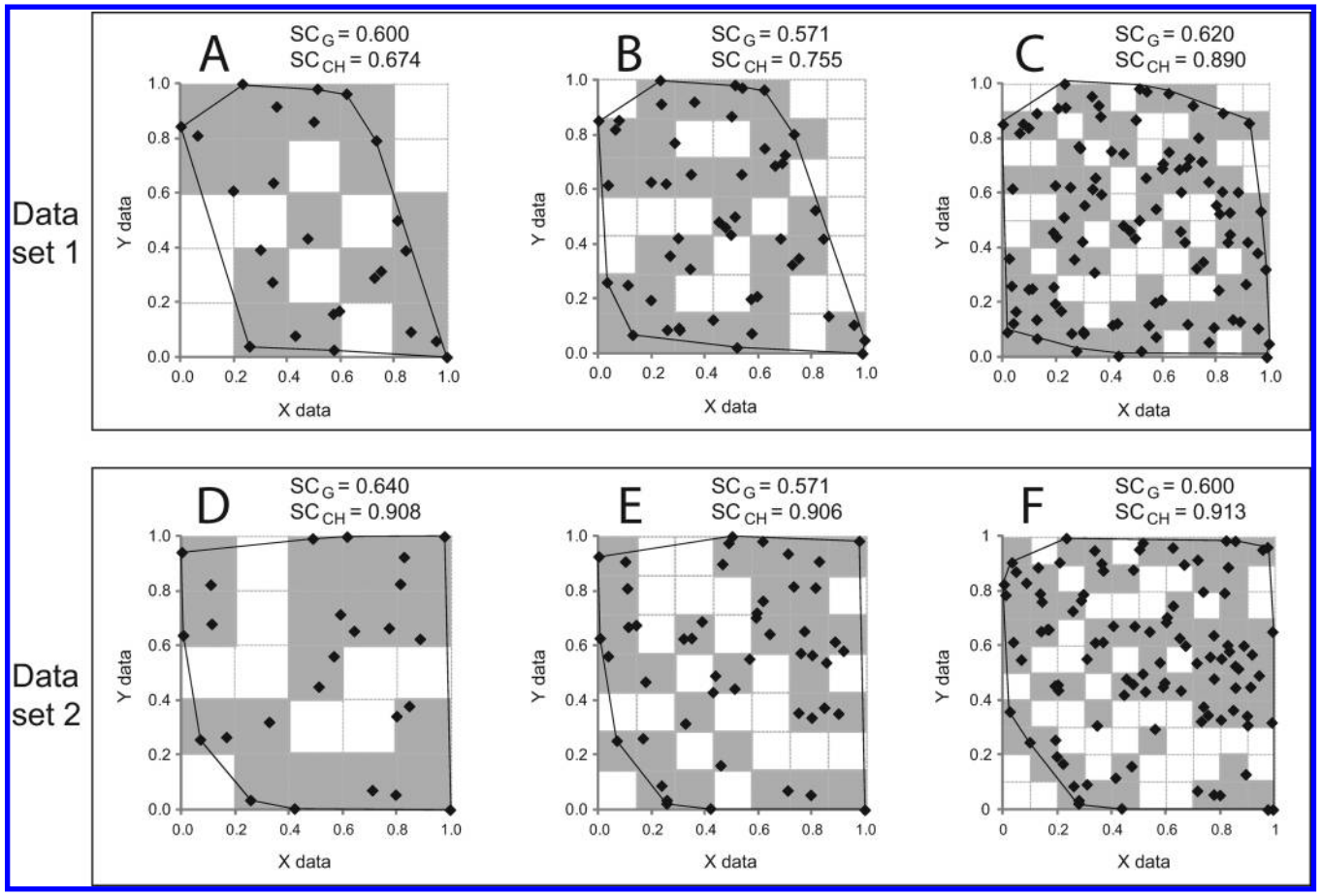


Figure 4. Illustration of the impact of sample size on orthogonality estimation. Data set 1 (A–C) and 2 (D–F). Binning was adjusted to maintain the number of bins equal to the number of components. A and D panels show 25 data points with 5 × 5 binning; B and E show 49 data points with 7 × 7 binning, and C and F show 100 data points with 10 × 10 binning. Gray bins were summed for SC_G calculation. The area framed by the black line was used for computation of the SC_{CH}. The orthogonality results averaged for eight data sets are shown in Table 2.

Pearson, Spearman, and Kendall correlation coefficients were recently utilized for measuring chromatographic orthogonality.[10] Although correlation coefficients do not require data normalization and discretization, they are designed to measure a degree of data correlation, rather than lack of it. Of the three, analytical chemists most commonly use the Pearson correlation coefficient R or the quantity R^2. Although R^2 is intuitive (equal to 1 for linear correlation and 0 for lack of linear correlation), a classical paper from Anscombe[28] illustrated that the same R^2 value is obtained for poorly correlated, clustered, and mutually related data sets with few outliers. This could be seen for example for the clearly correlated nonlinear data set (Figure 1H) which has lower R^2 (=0.856) than the more orthogonal system in Figure 1B (R^2 = 0.874).

Although the Spearman and Kendall correlation coefficients correctly characterize the system in Figure 1H with value 1, they still have some limitations as measures of chromatographic orthogonality, as discussed below. In general, correlation coefficients are better suited for characterization of correlated systems shown in Figure 1A–C than for noncorrelated 2D data with clustering shown in Figure 1F,J,K,L.

The orthogonality values calculated using selected metrics are summarized in Table 1. For methods requiring discretization, we used binning 14 × 14 (MI, SC_G) or the range from 2 × 2 to 14 × 14 (D_{BC}). When comparing the results in Figure 3, several trends emerge. Figure 3A shows that the D_{BC} and SC_G method values are closely correlated; higher chromatographic dimensionality corresponds to larger SC_G values. The mathematical relationship of D_{BC} and SC_G methods will be illustrated in the next section.

Mutual information does loosely correlate with the SC_G method (Figure 3A); the two most apparent outliers are the random data set and the cluster with outliers (data sets in Figure 1G,L, respectively). The SC_{CH} and SC_G methods correlate relatively well (Figure 3B); as expected, the most significant outlier is the scenario in Figure 1L which leads to a large value of SC_{CH}. The Pearson R^2 values correlate poorly with the SC_G method (Figure 3B).

Some other correlations could be plotted from Table 1 (figures are provided as Supporting Information S4, in file ac3020214_si_002.xlsx). For example, SC_G versus Spearman R^2 or Kendall R^2 correlations are 0.697 and 0.802, respectively (measured using Pearson R^2). The Kendall correlation coefficient appears to be the most acceptable for the characterization of chromatographic orthogonality. When plotting MI data against SC_{CH}, the correlation coefficient is R^2 = 0.852 and the plot of MI versus Kendall coefficient has R^2 = 0.951.

Similarly to SC_{CH} and SC_G methods, Liu's et al. geometric approach[29] can be used to estimate SC. The model assumes a situation, when the analytes "fan out" from the origin and form a geometric shape for which the spreading angle could be calculated. This assumption is not valid in all 2D LC scenarios. For example, the spreading angle cannot be meaningfully

dx.doi.org/10.1021/ac3020214 | Anal. Chem. 2012, 84, 8722–8732

Table 2. Orthogonality Values Calculated for Smaller Sample Sets Randomly Selected from the Original 196 Data Points[a]

		analytes/binning					
method	2D system	25/5 × 5	36/6 × 6	49/7 × 7	64/8 × 8	100/10 × 10	196/14 × 14
Pearson R^2	random	0.071 (89.5%)	0.039 (123.9%)	0.023 (89.9%)	0.026 (77.1%)	0.02 (107.5%)	0.014
Pearson R^2	C_{18} × C_{18}	0.451 (26.5%)	0.476 (25.4%)	0.473 (21.1%)	0.48 (15.6%)	0.46 (15.7%)	0.461
Pearson R^2	SCX × C_{18}	0.102 (117.5%)	0.069 (109.1%)	0.066 (101.5%)	0.047 (59.2%)	0.055 (58.5%)	0.048
Pearson R^2	HILIC × C_{18}	0.016 (121.4%)	0.015 (106.4%)	0.013 (148.6%)	0.009 (140.7%)	0.004 (114.4%)	0.001
Spearman R^2	random	0.067 (85.1%)	0.035 (106.1%)	0.022 (95.3%)	0.027 (88.2%)	0.021 (116.7%)	0.015
Spearman R^2	C_{18} × C_{18}	0.442 (25.4%)	0.478 (30.0%)	0.476 (24.4%)	0.492 (15.0%)	0.466 (16.4%)	0.470
Spearman R^2	SCX × C_{18}	0.174 (75.9%)	0.128 (75.6%)	0.118 (63.1%)	0.096 (33.7%)	0.104 (38.7%)	0.107
Spearman R^2	HILIC × C_{18}	0.019 (96.9%)	0.016 (70.8%)	0.014 (104.7%)	0.009 (102.2%)	0.006 (141.9%)	0.001
Kendall R^2	random	0.034 (78.4%)	0.018 (99.4%)	0.011 (104.2%)	0.014 (94.2%)	0.01 (117.5%)	0.007
Kendall R^2	C_{18} × C_{18}	0.246 (26.7%)	0.273 (32.9%)	0.266 (27.7%)	0.267 (17.7%)	0.245 (20.9%)	0.243
Kendall R^2	SCX × C_{18}	0.095 (73.7%)	0.068 (72.9%)	0.063 (61.0%)	0.051 (34.5%)	0.056 (35.8%)	0.057
Kendall R^2	HILIC × C_{18}	0.01 (86.1%)	0.008 (60.5%)	0.007 (105.4%)	0.005 (109.3%)	0.003 (138.8%)	0.000
SC_G	random	0.63 (9.4%)	0.649 (5.6%)	0.612 (4.7%)	0.635 (5.4%)	0.641 (6.3%)	0.643
SC_G	C_{18} × C_{18}	0.495 (14.9%)	0.458 (9.7%)	0.477 (12.1%)	0.471 (5.5%)	0.439 (3.7%)	0.408
SC_G	SCX × C_{18}	0.52 (7.1%)	0.514 (11.9%)	0.508 (8.4%)	0.486 (6.7%)	0.44 (6.2%)	0.423
SC_G	HILIC × C_{18}	0.59 (10.7%)	0.604 (8.4%)	0.561 (6.1%)	0.543 (5.7%)	0.539 (8.8%)	0.51
SC_G trend	random	0.624 (8.7%)	0.637 (5.3%)	0.623 (4.8%)	0.645 (3.6%)	0.623 (2.8%)	0.628
SC_G trend	C_{18} × C_{18}	0.476 (15%)	0.477 (7.0%)	0.475 (6.3%)	0.458 (5.0%)	0.438 (4.0%)	0.427
SC_G trend	SCX × C_{18}	0.501 (7.3%)	0.512 (7.5%)	0.513 (6.4%)	0.485 (8.5%)	0.455 (3.3%)	0.423
SC_G trend	HILIC × C_{18}	0.587 (7.6%)	0.599 (5.5%)	0.558 (4.2%)	0.541 (7.1%)	0.518 (5.8%)	0.487
SC_{CH}	random	0.818 (10.7%)	0.856 (8.6%)	0.888 (6.3%)	0.901 (5.4%)	0.925 (3.5%)	0.966
SC_{CH}	C_{18} × C_{18}	0.501 (17.2%)	0.518 (14.6%)	0.529 (12.3%)	0.542 (11.0%)	0.569 (7.5%)	0.613
SC_{CH}	SCX × C_{18}	0.636 (13.9%)	0.663 (9.3%)	0.703 (10.4%)	0.723 (11.0%)	0.718 (6.8%)	0.715
SC_{CH}	HILIC × C_{18}	0.648 (10.0%)	0.673 (11.1%)	0.665 (6.4%)	0.682 (5.8%)	0.686 (5.4%)	0.693
D_{BC}	random	1.507 (6.1%)	1.615 (3.0%)	1.63 (2.7%)	1.708 (1.7%)	1.745 (1.9%)	1.792
D_{BC}	C_{18} × C_{18}	1.234 (12.6%)	1.303 (6.5%)	1.394 (5.3%)	1.422 (2.0%)	1.47 (1.6%)	1.537
D_{BC}	SCX × C_{18}	1.325 (8.6%)	1.337 (13.0%)	1.454 (4.2%)	1.48 (4.9%)	1.482 (1.9%)	1.531
D_{BC}	HILIC × C_{18}	1.458 (8.6%)	1.553 (4.5%)	1.544 (3.5%)	1.561 (2.8%)	1.613 (2.8%)	1.648
MI	random	0.179 (22.1%)	0.196 (15.7%)	0.222 (12.7%)	0.204 (15.6%)	0.21 (12.7%)	0.222
MI	C_{18} × C_{18}	0.22 (9.1%)	0.262 (21.6%)	0.266 (18.1%)	0.254 (12.7%)	0.26 (12.0%)	0.29
MI	SCX × C_{18}	0.208 (26.1%)	0.228 (12.1%)	0.231 (12.6%)	0.242 (10.5%)	0.247 (7.2%)	0.272
MI	HILIC × C_{18}	0.136 (31.6%)	0.155 (23.3%)	0.148 (24.8%)	0.168 (10.5%)	0.148 (13.8%)	0.146

[a]The average value and RSD were calculated from n = 8 samples in each experiment.

defined in scenarios shown in Figure 1D−F. Therefore, the Liu et al. method[29] was not included in this study.

Impact of Sample Size on Orthogonality Estimation. The choice of sample is crucial for meaningful estimation of 2D orthogonality. Some scientists recommend performing method development with the same sample for which the 2D chromatographic analysis is intended.[30] However, this may not be practical, and model mixtures may be often used instead. The difficulty lies in selecting the model mixture. Many components of the target sample may not be known or only occasionally exceed the limits of detection.[23,31]

When characterizing the 2D orthogonality with a test mixture, compounds of representative chromatographic behavior must be chosen. When a limited set of compounds is utilized (often less than 100),[6,13,16,32−35] some variability of orthogonality estimation may arise. Whether the results can be generalized for other samples is an open question.

In this work, we used a moderate set of 196 tryptic peptides. In order to evaluate the impact of sample size on calculated orthogonality values, we randomly selected sets of 25, 36, 47, 64, and 100 peptide retention times from the original pool of 196. The selection process was repeated eight times at each sample complexity level. The orthogonality for the resulting 40 sets (eight repetitions on five complexity levels) was estimated with selected methods. The study was executed for data in Figure 1G,C,D,F.

Figure 4 outlines the study execution for SC_{CH} and SC_G methods using random data sets. Only the selected examples are shown for simplicity (two repetitions were labeled set 1 and 2) with separation space divided into 25, 49, and 100 bins containing 25, 49, and 100 peptides, respectively. The area of bins containing data points (in gray) was summed and reported as SC_G.

The upper graphs (Figure 4A−C) represent one data set and can be compared against the second data set shown in Figure 4D−F. Notice that the distribution of components in the space is different; however, the SC_G values are similar for both sets. The data for all eight sets were averaged, and these values were reported in Table 2 along with the relative standard deviation expressed as a percentage (% RSD). For example, for a random set and 100 data points, the average for SC_G is 0.641 with a RSD of 6.3% (n = 8).

Figure 4 also depicts the orthogonality estimation for the SC_{CH} method; the area in these plots is framed by the black lines. Two extreme cases, Figure 4A−C versus Figure 4D−F, were selected out of eight experiments to illustrate the impact of the choice of the data set on the SC_{CH} value. The overall robustness of the SC_{CH} metric is shown in Table 2. Similarly to other methods, the variability increases for experiments with a

dx.doi.org/10.1021/ac3020214 | Anal. Chem. 2012, 84, 8722−8732

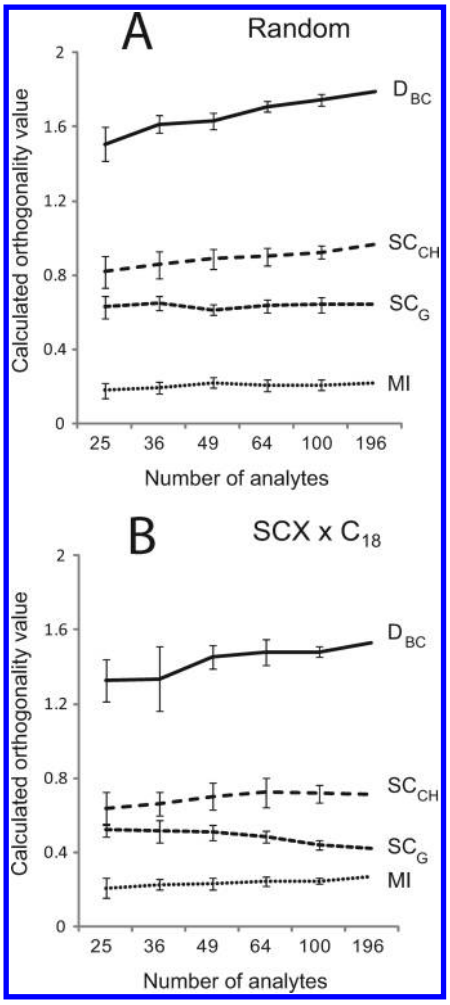


Figure 5. Trends in calculated orthogonality value when changing the sample size. Error bars indicate the standard deviation from eight experiments. (A) random data set; (B) SCX × C_{18} 2D data set.

smaller number of components. The average SC_{CH} for 100 data points is 0.925 with an RSD of 3.5% (n = 8).

Table 2 summaries the average orthogonality values and the RSDs for selected methods. MI and correlation coefficients exhibit an increased scatter. MI appears to be a less useful measure of 2D orthogonality, because it is less robust for statistically smaller numbers of data points. As mentioned earlier, correlation coefficients are more suited for the description of correlated rather than uncorrelated data sets. This is why the highest variability is seen for random and HILIC × C_{18} data sets.

Figure 5 illustrates greater variability of calculated D_{BC}, SC_{CH}, SC_G, and MI values. The error bars show the standard deviation for 8 data sets at each data size level. Despite the random data selection from the 196 components pool, these calculated orthogonality estimators vary as more outlying data points are included or excluded. This effect becomes more dominant for data sets containing smaller number of components.

Figure 5 also shows the trends in the orthogonality estimator values. Higher values of D_{BC}, SC_G, and SC_{CH} in Figure 5A indicate that random data are more orthogonal than the SCX × C_{18} 2D system (Figure 5B). The values of SC_{CH} are higher than SC_G because the SC_{CH} method includes the potentially underutilized 2D separation space (e.g., Figure 2F). Interestingly, SC_{CH} values increase with number of components, although SC_G values show the opposite trend (see also Table 2). The D_{BC} and MI values also increase for larger data sets. The plots for HILIC × C_{18} and C_{18} × C_{18} experiments are provided as Supporting Information S5, in file ac3020214_si_002.xlsx.

The data in Figure 5 suggest that the apparent orthogonality depends moderately on sample size and the variability of estimation is higher for small data sets. Therefore, the preferred practice is to compare the 2D orthogonality using large and identical sample types for all 2D systems and report the number of components when estimating orthogonality by any method.

Relationship between SC_G and D_{BC}. A close correlation between SC_G and D_{BC} observed in Figure 3A is intriguing. The question arises whether these two methods are mathematically related.

Both methods have formal similarities, including the discretization of a normalized separation space, and counting of bins used for separation. When surface coverage is plotted against the number of bins (Figure 6A) or the log of the number of bins (Figure 6B,C), the resulting graphs are formally similar to the D_{BC} representation shown in Figure 2D. In the central range of the x-axis (Figure 6C), the relationship between surface coverage and log B is linear. SC_G is defined for the condition when the number of bins is equal to the number of components, $B = N$; this position is marked by a gray dashed line in Figure 6B,C.

The SC increases with the decreasing number of bins; for very low number of bins (1, 4, 9, 16, ...), the surface coverage plateaus at value of one. This is most clearly visible in Figure 6B. Contrary to the D_{BC} relationship in Figure 2D, which is linear in the small number of bins region, the SC function appears to be linear at the middle to low number of bins range (Figure 6C). The slope and intercept of this linear relationship can be used to calculate the SC_G value. We label this quantity "SC_G trend", and list them in Table 1 for comparison. As illustrated by dotted lines in Figure 6C, the intercept and slope are used to calculate SC_G value (196 bins, log $B \approx 2.3$). The "SC_G trend" values are more robust, because the trend eliminates a local SC_G scatter. Trend calculated SC_G values were used to construct the plots in Figure 3.

In the D_{BC} method, we count the number of occupied bins M as a function of the bin size L. The dimension $D \geq 0$ satisfies the following scaling:

$$M = AL^{-D} \tag{5}$$

where A is a constant. Taking the logarithms of both sides of eq 5 gives:

$$\log M = \log A - D \log L \tag{6}$$

Equation 6 shows that the slope of a line fit to the log M vs log L data yields $-D$ and the intercept of this line is equal to log A.

The surface coverage in a unit square divided into square bins can be expressed as eq 7:

$$SC = \frac{M}{B} = \frac{M}{L^{-2}} = ML^2 \tag{7}$$

By substituting the expression for M from eq 5 into eq 7, we obtain for SC:

$$SC = AL^{2-D} \tag{8}$$

After converting eq 8 to a logarithmic form, it becomes apparent that the box counting dimensionality D could be

dx.doi.org/10.1021/ac3020214 | Anal. Chem. 2012, 84, 8722−8732

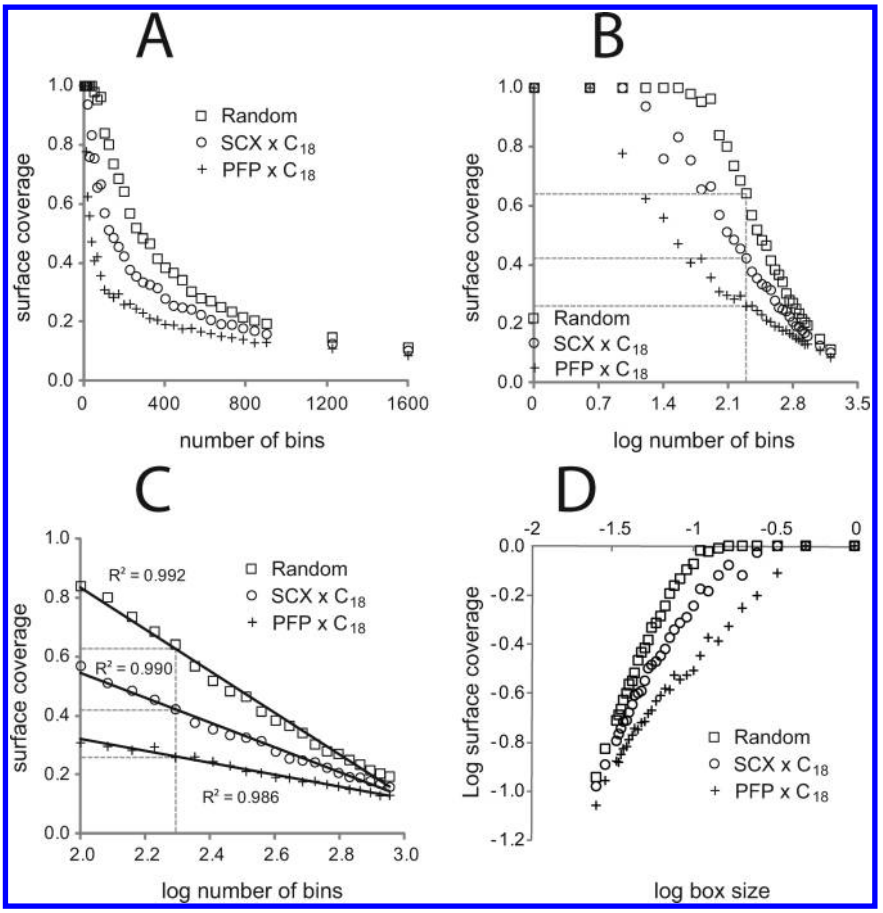


Figure 6. Various representation of SC trends for 196 components when varying the discretization of separation space. (A) Plot of SC versus *B*; (B) SC versus log *B*; (C) linear subset of the plot shown in panel B; (D) plot of log SC versus log *L*. Gray lines indicate the values of SC_G when the number of bins corresponds to the number of components. The "SC_G trend" values were calculated using a linear trend between 10 × 10 and 25 × 25 binning range (log *B* range 2–2.796). E.g., for Random data (panel C), the "SC_G trend" = −0.749 × log 196 + 2.345.

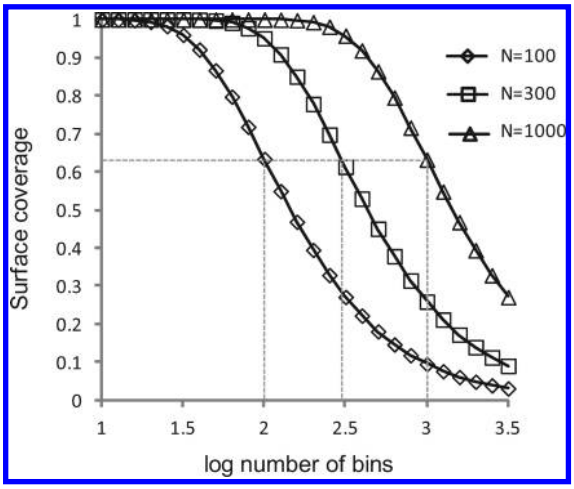


Figure 7. Surface coverage calculated from eq 4, for 100, 300, and 1000 uniformly randomly distributed components. Gray lines show that for condition *B* = *N* surface coverage is constant (0.63) regardless of the number of data points. SC at *B* = *N* condition is denoted as SC_G.

ascertained from the equation where SC is a function of both *N* and *L*:

$$\log SC = \log A + (2 - D)\log L \quad (9)$$

The slope of the logarithmic plot given by eq 9 is 2 − *D*, which leads to exactly the same value of D_{BC}. Indeed, for binning from 2 × 2 to 14 × 14 (the same as for the D_{BC} method), the D_{BC} values estimated from *D* in eq 9 were identical to those in Table 1.

The importance of binning range selection is apparent from Figure 6D which shows that relationship log SC versus log *L* is linear only in certain range of *L*. For large *L* (high saturation of separation space), the log SC value plateaus, approaching the limit value of 0 (full utilization of space, SC = 1). The proper range of log *L* has to be chosen to obtain a meaningful estimate of *D* value. The situation is analogical to box counting method as in Figure 2D. Inappropriate selection will lead to inaccurate estimate of dimensionality.

Using the condition *N* = *B*, eq 7 is equal to eq 4 (for random data set). The choice of *N* = *B* is rationalized in Figure 7, which shows SC as a function of log *B* (eq s3.8 in Supporting Information S3) for 100, 300, and 1000 data points. Regardless of the number of components *N*, SC_G is 0.63 where *N* = *B*. The SC relationship in Figure 7 is not a linear function of log *B*; the quasi-linear portion is centered at the inflection point with SC_G = 0.63. It could be argued that this point is the preferred and most robust point to measure SC_G. Further testing of SC_G with other types of data and samples will help clarify the widespread use of this metric and of D_{BC} as orthogonality metrics.

It has been pointed out earlier that SC can be used to compute the achievable 2D peak capacity $n^*_{c,2D}$ using eq 2. The

dx.doi.org/10.1021/ac3020214 | *Anal. Chem.* 2012, 84, 8722–8732

fractional coverage f_c can be in principle substituted by SC_G, SC_{CH}, or other SC values. However, as shown in Figures 5 and 3, the SC_G and SC_{CH} have limited correlation, and more importantly, they differ in absolute value. The scenario with random distribution of analytes in 2D space (Figure 1G) gives $SC_{CH} \approx 1$ and $SC_G \approx 0.63$. Therefore, the achievable 2D peak capacity will differ according to the method used, giving lower achievable peak capacity for SC_G method. While it is possible to normalize SC_G values to a common range of 0−1 by dividing the values by 0.63, we did not chose to do so for two reasons. First, for some selected examples, such as separation of polymers[36] where the analytes are rather evenly distributed in the separation plane, the normalized SC_G values may exceed the value of one. Second, we believe that the available 2D separation space is reduced by local crowding which results in under-utilization of space which is evident in all chromatograms including the simulated random scenario (Figure 1G). In other words, while SC_G method provides lower achievable peak capacity, it could be the more realistic one compared to SC_{CH} method. It remains to be seen which of the two achievable peak capacities (SC_{CH} or SC_G based) will correlate better with the empirical method of peak counting, which is also used for estimating the 2D LC separation quality.[37]

CONCLUSIONS

The results presented in this paper allow for several conclusions: (1) Several statistical methods are applicable for measuring 2D orthogonality. SC_G, SC_{CH}, and D_{BC} methods offer useful and intuitively understandable measures of orthogonality. The methods are sufficiently robust for small sample sets commonly used in chromatographic practice. (2) The MI method provides less intuitive orthogonality values; it shows greater variability for small data sets. Correlation coefficients are not well suited for measurement of chromatographic orthogonality, especially for systems with data clustering and data outliers (3) SC_G, D_{BC}, and MI methods depend on discretization (binning) of the separation space. Although the SC_{CH} method is independent of binning, it is sensitive to the degree of saturation of the separation space (number of components N; see Figure 4). Orthogonality between 2D systems should be compared and specified with the component number and binning dimensions. (4) The SC_G method is defined for conditions where the number of bins is equal to the number of components ($B = N$). In comparison to the SC_{CH} method, it returns lower values for the data used here. This is because the SC_{CH} includes sparsely used area in the accessible separation space. The SC_{CH} method may overestimate the orthogonality for some 2D scenarios found in the literature.[15,34,38−40] (5) The concept of SC as a measure of orthogonality is particularly attractive because the fractional surface coverage value is useful for calculating the achievable peak capacity from eq 2. SC_G can be used for this purpose providing that surface coverage is calculated for the condition $B = N$ and not varied with the actual peak capacities in first and second separation dimensions. (6) D_{BC} and SC_G methods show strong correlation and are mathematically related so that one can be obtained from the other using eq 9 (see also Supporting Information S3). (7) Although it is not clear whether the chromatographic data used here have fractal nature, for the range of binning, and data sets used, the calculated dimensionality, D_{BC}, follow the expected trends ($D_{BC} = 1$ for ordered nonorthogonal components and $D_{BC} = 2$ for fully ordered orthogonal 2D components[11]).

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: +1 508 482 2000. Fax: +1 508 482 3625. E-mail: Martin_Gilar@waters.com.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Schoenmakers, P.; Marriott, P.; Beens, J. *LCGC Europe* **2003**, *June 1*, 1−4.
(2) Tomas, G. B. *Calculus and analytical geometry*, 4th ed.; Addison-Wesley Publishing Company: Reading, MA, 1969.
(3) Strang, G. *Linear algebra and its application*, 2nd ed.; Academic Press: New York, 1980.
(4) Snedecor, G. W.; Cochran, W. G. *Statistical methods*, 4th ed.; The Iowa State University Press: Ames, IA, 1980.
(5) Brass, P. *Advanced data structures*; Cambridge University Press: New York, 2008.
(6) Slonecker, P. J.; Li, X.; Ridgway, T. H.; Dorsey, J. G. *Anal. Chem.* **1996**, *68*, 682−689.
(7) Carr, P. W.; Davis, J. M.; Rutan, S. C.; Stoll, D. R. In *Advances in chromatography*; Grushka, E., Grinberg, N., Eds.; CRC Press: Boca Raton, FL, 2012; Vol. *50*, p 139−235.
(8) Rutan, S. C.; Davis, J. M.; Carr, P. W. *J. Chromatogr., A* **2012**, *1255*, 267−276.
(9) Pourhaghighi, M. R.; Karzand, M.; Girault, H. H. *Anal. Chem.* **2011**, *83*, 7676−7681.
(10) Al Bakain, R.; Rivals, I.; Sassiat, P.; Thiebaut, D.; Hennion, M. C.; Euvrard, G.; Vial, J. *J. Chromatogr., A* **2011**, *1218*, 2963−2975.
(11) Schure, M. R. *J. Chromatogr., A* **2011**, *1218*, 293−302.
(12) Giddings, J. C. *Unified Separation Science*; Wiley Interscience: New York, 1991.
(13) Al Bakain, R.; Rivals, I.; Sassiat, P.; Thiebaut, D.; Hennion, M. C.; Euvrard, G.; Vial, J. *J. Chromatogr., A* **2012**, *1232*, 231−241.
(14) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. *Anal. Chem.* **2005**, *77*, 6426−6434.
(15) Davis, J. M.; Stoll, D. R.; Carr, P. W. *Anal. Chem.* **2008**, *80*, 8122−8134.
(16) Semard, G.; Peulon-Agasse, V.; Bruchet, A.; Bouillon, J. P.; Cardinael, P. *J. Chromatogr., A* **2010**, *1217*, 5449−5454.
(17) Gray, M.; Dennis, G. R.; Wormell, P.; Shalliker, R. A.; Slonecker, P. *J. Chromatogr., A* **2002**, *975*, 285−297.
(18) Neue, U. D.; Mendez, A. *J. Sep. Sci.* **2007**, *30*, 949−963.
(19) Watson, N. E.; Davis, J. M.; Synovec, R. E. *Anal. Chem.* **2007**, *79*, 7924−7927.
(20) van der Horst, A.; Schoenmakers, P. J. *J. Chromatogr., A* **2003**, *1000*, 693−709.
(21) Giddings, J. C. *J. Chromatogr., A* **1995**, *703*, 3−15.
(22) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. *Anal. Chem.* **2005**, *77*, 2187−2200.
(23) Huidobro, A. L.; Pruim, P.; Schoenmakers, P.; Barbas, C. *J. Chromatogr., A* **2008**, *1190*, 182−190.
(24) Gilar, M.; Olivova, P.; Chakraborty, A. B.; Jaworski, A.; Geromanos, S. J.; Gebler, J. C. *Electrophoresis* **2009**, *30*, 1157−1167.
(25) Dwivedi, R. C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2008**, *80*, 7036−7042.
(26) Delmotte, N.; Lasaosa, M.; Tholey, A.; Heinzle, E.; Huber, C. G. *J. Proteome Res.* **2007**, *6*, 4363−4373.
(27) Nakamura, T.; Kuromitsu, J.; Oda, Y. *J. Proteome Res.* **2008**, *7*, 1007−1011.

(28) Anscombe, F. J. *Am. Statistician* **1973**, *27*, 17−21.
(29) Liu, Z.; Patterson, D. G.; Lee, M. L. *Anal. Chem.* **1995**, *67*, 3840−3845.
(30) Stevenson, P. G.; Mnatsakanyan, M.; Francis, A. R.; Shalliker, R. A. *J. Sep. Sci.* **2010**, *33*, 1405−1413.
(31) Kalili, K. M.; de Villiers, A. *J. Chromatogr., A* **2009**, *1216*, 6274−6284.
(32) Venkatramani, C. J.; Zelechonok, Y. *Anal. Chem.* **2003**, *75*, 3484−3494.
(33) Du, C. M.; Valko, K.; Bevan, C.; Reynolds, D.; Abraham, M. H. *J. Chromatogr. Sci.* **2000**, *38*, 503−511.
(34) Tanaka, N.; Kimura, H.; Tokuda, D.; Hosoya, K.; Ikegami, T.; Ishizuka, N.; Minakuchi, H.; Nakanishi, K.; Shintani, Y.; Furuno, M.; Cabrera, K. *Anal. Chem.* **2004**, *76*, 1273−1281.
(35) Jandera, P.; Hajek, T.; Cesla, P. *J. Sep. Sci.* **2010**, *33*, 1382−1397.
(36) Murphy, R. E.; Schure, M. R.; Foley, J. P. *Anal. Chem.* **1998**, *70*, 1585−1594.
(37) Huang, Y.; Gu, H.; Filgueira, M.; Carr, P. W. *J. Chromatogr., A* **2011**, *1218*, 2984−2994.
(38) Gilar, M.; Jaworski, A. *J. Chromatogr., A* **2011**, *1218*, 8890−8896.
(39) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. *J. Sep. Sci.* **2005**, *28*, 1694−1703.
(40) Jandera, P.; Fischer, J.; Lahovska, H.; Novotna, K.; Cesla, P.; Kolarova, L. *J. Chromatogr., A* **2006**, *1119*, 3−10.