Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Sample-effective calibration design for multiple components

Dmitry Kirsanov[a,b], Vitaly Panchuk[a], Marina Agafonova-Moroz[c], Maria Khaydukova[a,b], Alexander Lumpov[c], Valentin Semenov[a], Andrey Legin[a,b]

a Institute of Chemistry, St. Petersburg State University, St. Petersburg, Russia

b Laboratory of Artificial Sensory Systems, University ITMO, St. Petersburg, Russia

c Khlopin Radium Institute, St. Petersburg, Russia

corresponding author: d.kirsanov@gmail.com

Abstract

The experimental design of mixtures for multivariate calibration is introduced. The idea of this design is based on uniform distribution of experimental points in a concentration hypercube. Unlike the already reported uniform designs this one is pretty simple and not computationally demanding. The suggested approach does not employ the concept of fixed "levels" and allows for designs with any number of experimental mixtures and any number of components depending on "time and money" considerations for each particular calibration experiment. The performance of the design is assessed with UV-Vis spectroscopic experiment for simultaneous quantification of four inorganic components in complex mixtures. The performance of the PLS regression models derived from design is compared with that of cyclic permutation and Kennard-Stone designs. The suggested approach allows for comparable or higher prediction accuracy with the lower number of experimental points.

Introduction

Multivariate calibration is a well-established and widely applied tool in modern analytical chemistry. Usually multivariate regression is employed when one is interested in substitution of some expensive, time-consuming and tedious method with a simpler and faster technique. Multivariate calibration is especially useful when dealing with multicomponent mixtures where ordinary least squares approach with a single variable fails due to a complex signal shape, an absence of distinct bands, etc. A classic example usually referred to is near-infrared spectroscopy. One will normally require a set of reference data from another method/instrument to establish regression model. For example, one has to analyze first all calibration samples with standard technique (e.g. burning in ash oven) making a calibration of a NIR spectrometer for prediction of ash content in grain [1]. This is the most straight-forward way of multivariate calibration and it allows for taking into account the influence of all the components in real complex multicomponent mixtures dealing with real samples. In certain cases, however, this direct approach can hardly be implemented, since real samples might be very expensive or not readily available. In these cases one can work with model mixtures to establish regression model. The design of these mixtures obviously must follow the composition of the real future samples in terms of concentration ranges and ratios between the components. With a single component of interest the design of calibration samples is quite intuitive: one has to prepare the samples with evenly spaced concentrations of analyte covering the whole relevant concentration range. Some important issues regarding this are summarized in [2]. The example of calibration design for two components is described in [3], similar design was employed in [4]. Numerous other types of design are known in literature [5].

For example, central composite design [6] can deal with three-four factors, however it is intended rather for optimization experiments then for calibration. It is important to mention that all these designs are operating with fixed concentration levels of components which are running through their certain values thus providing different combinations of factors. As the number of components grows higher this concept leads to an elevated number of experiments. A study of all possible combinations of seven components with five particular concentration levels of each component will require 5^7=78125 different mixtures to be studied. Obviously this is far from being doable in common laboratory practice. These limitations were successfully circumvented in the works of Brereton [7, 8] using the concept of cyclic generator which allows to avoid studying all possible combinations, while the studied factors are strictly orthogonal. Cyclic generator approach can construct designs of a calibration where the number of experiments is equal to the squared number of levels. The maximum number of factors permitted is N-1, where N is a number of experiments, e.g. for five levels, 24 factors can be employed in design. Thus instead of a huge number of mixtures one can design very compact experiments. However the number of calibration samples is strictly fixed to provide for the orthogonality of components. In the situations when this number is constrained by considerations of mixture preparation price and labor (e.g. for very complex samples) every extra sample counts. It would be very convenient for chemists to have an instrument to design calibration sets with random, voluntary adjustable number of mixtures. Implementation of such design can be based on the idea of uniform distribution of experimental points in concentration hypercube. This idea is somehow obvious and there were already numerous works addressing this idea of space filling, see e.g. [9-12] however the algorithms reported there are quite computationally intensive, while the works [11, 12] are mainly concentrated on designs for computer simulation experiments, thus the number of considered experimental points is usually about several hundred and this is poorly connected with physical calibration in a laboratory. Another example of uniform design is famous Kennard-Stone design [13], however its procedure requires that corner points in hypercube must be filled first and this leads to elevated number of samples when studying multiple components at multiple levels.

We suggest the approach for design of multicomponent calibration mixtures, which does not use fixed concentration “levels” of the components. Relaxation of this constraint allows for a certain freedom of choice of the number of experiments in design. Moreover, this approach is a general one and allows for any number of components to be included.

Theory

The idea of this calibration design is based on the uniform distribution of points in a space of arbitrary dimensionality suggested in [14]. This task is known in mathematics. Consider we have to distribute uniformly *N* points in *n*-dimensional space. With initial task in mind here *n* will be the number of components in the mixture and *N* will be the number of calibration mixtures. The volume of this space is fixed by the upper limits of components concentrations. Let us divide the whole *n*-dimensional volume into the *m* identical equilateral sub-volumes in the way assuming that during the filling of the whole volume with points each sub-volume would contain at least one point, i.e. $m \le N$. At the same time the number of these sub-volumes must be maximal, e.g. for the two-dimensional space containing 9-15 points the *m* will be 9, if the number of points 16-25 then *m*=16, etc. In general case $m = k^n$, where *k* is the integer part of $\sqrt[n]{N}$.

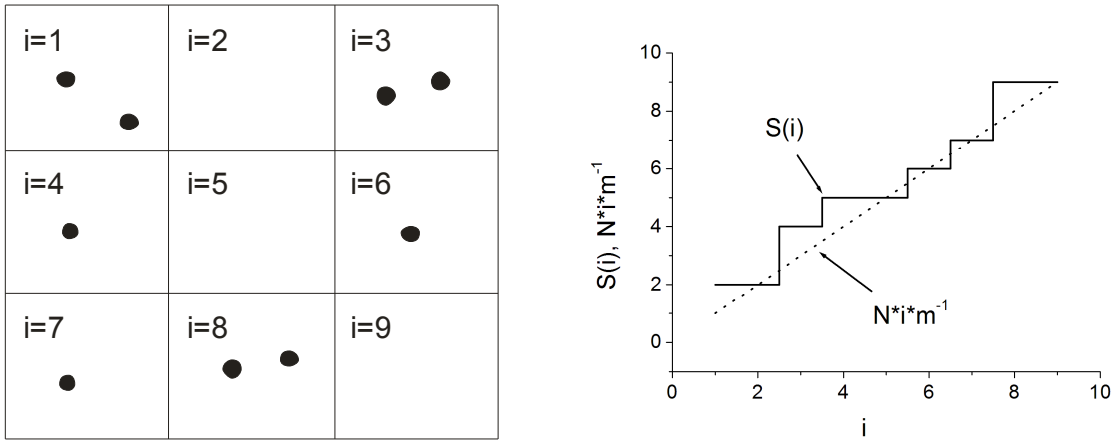
We denote by *S(i)* the number of points falling into all sub-volumes from 1st to i^{th}, where $i \in [1,m]$, and by *D(i)* – the deviation from uniformly filled volume:

$$D(i)=\{|S(i) - N\cdot i\cdot m^{-1}|\} \quad (1)$$

As the main criterion of the uniform filling we will consider the minimum of maximal deviation $D_{max} = max\{D(i)\}$ based on the approach reported in [12].

Let us illustrate this with an example of two-dimensional space with nine points.

a)



b)

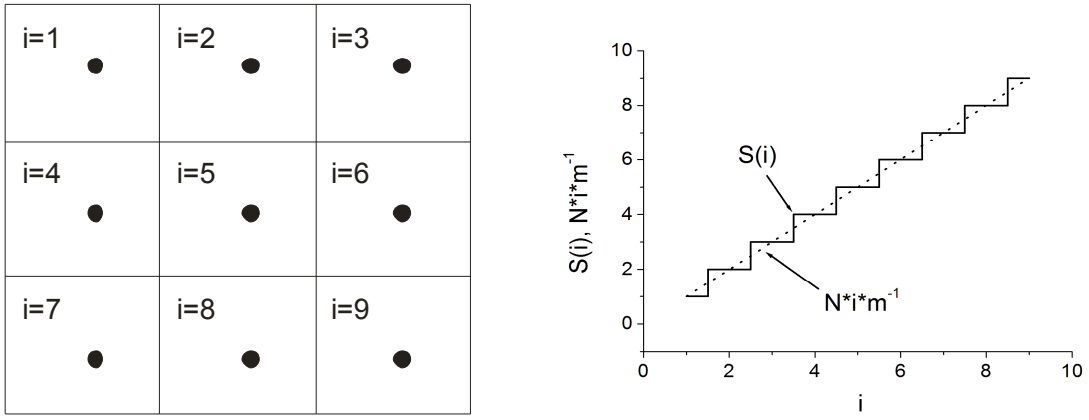


Figure 1. The examples of non-uniform (a) and uniform (b) filling of two-dimensional space.

The plot of S(i) for random space filling is shown in the Fig.1a. In this particular example

$$S(i)=\{2, 2, 4, 5, 5, 6, 7, 9, 9\};$$

$$N\cdot i\cdot m^{-1}=\{1, 2, 3, 4, 5, 6, 7, 8, 9\};$$

$$D(i)=\{1, 0, 1, 1, 0, 0, 0, 1, 0\},$$

thus $D_{max} = 1$. Fig.1b shows an example of uniform space filling:

$$S(i) = \{1, 2, 3, 4, 5, 6, 7, 8, 9\};$$

$N \cdot i \cdot m^{-1} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\};$

$D(i) = \{0, 0, 0, 0, 0, 0, 0, 0, 0\},$

and $D_{max} = 0$. The condition of the minimum of D_{max} is necessary, but not sufficient since it still allows for a quite close position of two points in two neighboring sub-volumes. Moreover, in a general case of $m<N$ with condition of $min(D_{max})$ some sub-volumes will contain several points which positions are not included in the criterion of D_{max}. Let us introduce an additional condition of uniform filling and denote the distance between two points of n-dimensional space as $r_{l,k}$ $(l,k \in [1,N]; l \neq k)$:

$$r_{l,k} = \sqrt{\sum_{j=1}^{n} (x_{l,j} - x_{k,j})^2} \qquad (2),$$

where $x_{l,j}$ is a coordinate of point l on the j axis, $x_{k,j}$ is a coordinate of point k on the j axis, n is the space dimensionality and denote $r_{min} = min\{r_{l,k}\}$. Then an additional criterion of uniformity will be maximal value of r_{min}. According to these two criteria ($min\{D_{max}\}$ and $max\{r_{min}\}$) the following algorithm for uniform filling of n-dimensional space with N points is suggested (Fig.2).

At the first stage the whole n-dimensional space is divided into m sub-volumes as described above and matrix $\boldsymbol{X}$ with coordinates of all N points is initialized:

$$X = \begin{vmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots \\ x_{N,1} & x_{N,2} & \dots & x_{N,n} \end{vmatrix}.$$

The coordinates are being chosen in a random way: $x_{l,j} = random[x_{min}; x_{max}]$, where x_{min} and x_{max} are the lower and the upper limits of x; $l \in [1, N]; j \in [1, n]$. Then the maximum deviation of these points from uniform distribution D_{max}^{X} is computed.

At the second stage the coordinates of the points are iteratively shifted to find $min\{D_{max}\}$. For this purpose the shifting matrix ΔX is defined at every iteration step, which elements contain random shifts in coordinates of points in X:

$\Delta x_{l,j} = random[-\Delta x_{max}/2; \Delta x_{max}/2]$,

where Δx_{max} – is the maximum shift value for coordinate. The matrix X' with shifted coordinates is defined as:

$$X' = X + \alpha * \Delta X \qquad (3),$$

where α is regularization coefficient and $\alpha \in (0,1]$. If some of the coordinates in X' appear being outside of the prescribed concentration ranges these coordinates are assigned with the nearest boundary value.

Next step is computation of $D_{max}^{X'}$ (maximum deviation of new shifted points from uniform distribution) and its comparison with D_{max}^{X}. If $D_{max}^{X'} \leq D_{max}^{X}$ then new shifted coordinates are considered as initial,

otherwise the regularization coefficient α gets smaller value and X' is being redetermined. This procedure continues until $D_{max}^{X'} \leq D_{max}^{X}$. The coordinate shifting repeats until specified number of iterations (NumItMax) after the last change in D_{max} is reached.

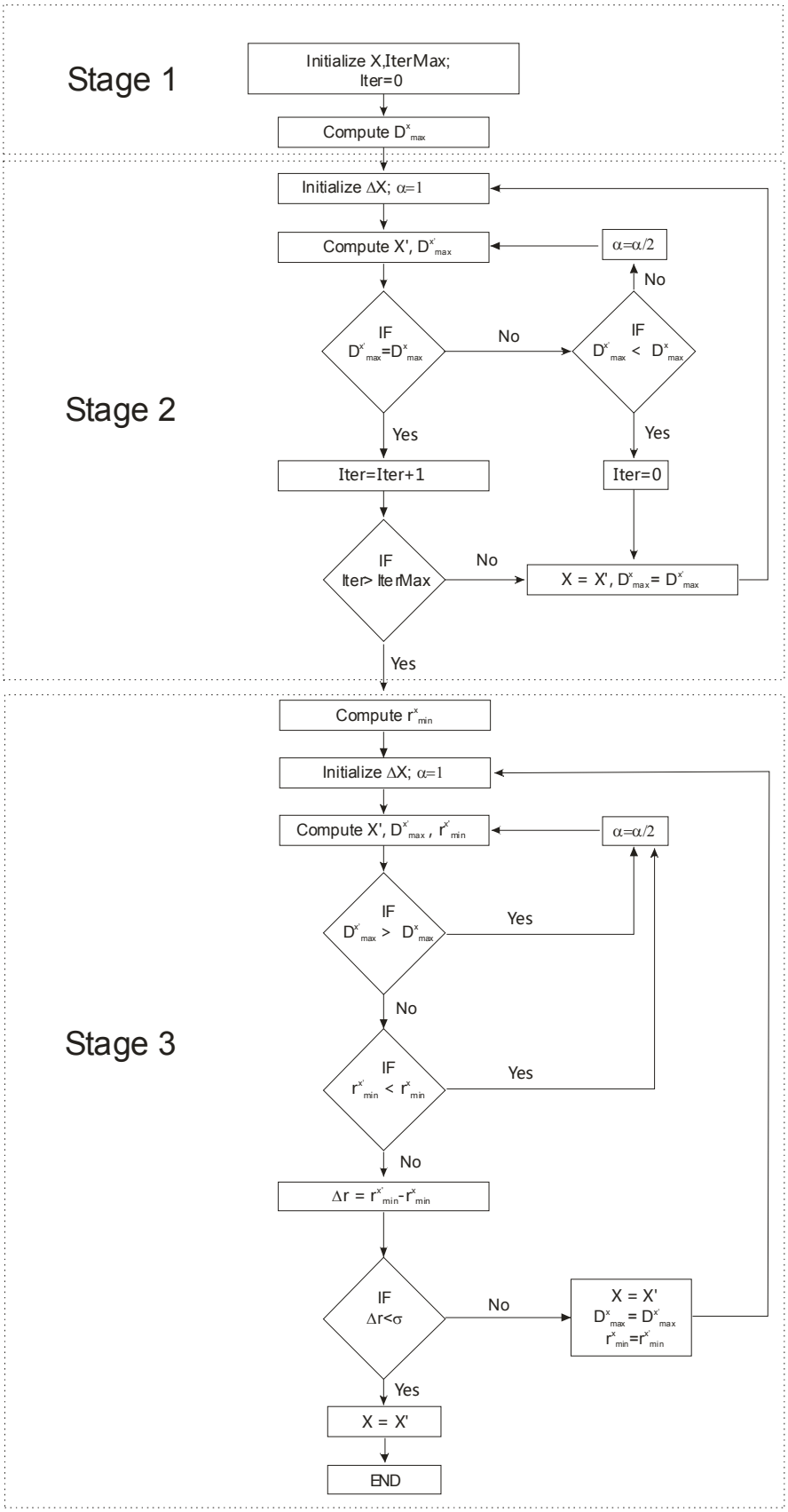


Figure 2. The algorithm of the uniform filling of *n*-dimensional space with *N* points.

The final stage is devoted to optimization of r_{min}. For this purpose we again shift the coordinates of points in X, however this time we add an extra condition: $r_{min}^{X'} > r_{min}^{X}$ (besides $D_{max}^{X'} \leq D_{max}^{X}$). Coordinate shifting continues until $r_{min}^{X'} - r_{min}^{X} \geq \sigma$, where σ has preliminary defined value.

Examples

The Fig.3 shows the results of algorithm performance for two-dimensional space with *N* growing from 2 to 10. Concentration levels of two components A and B are given in arbitrary units and vary from 1 to 5.

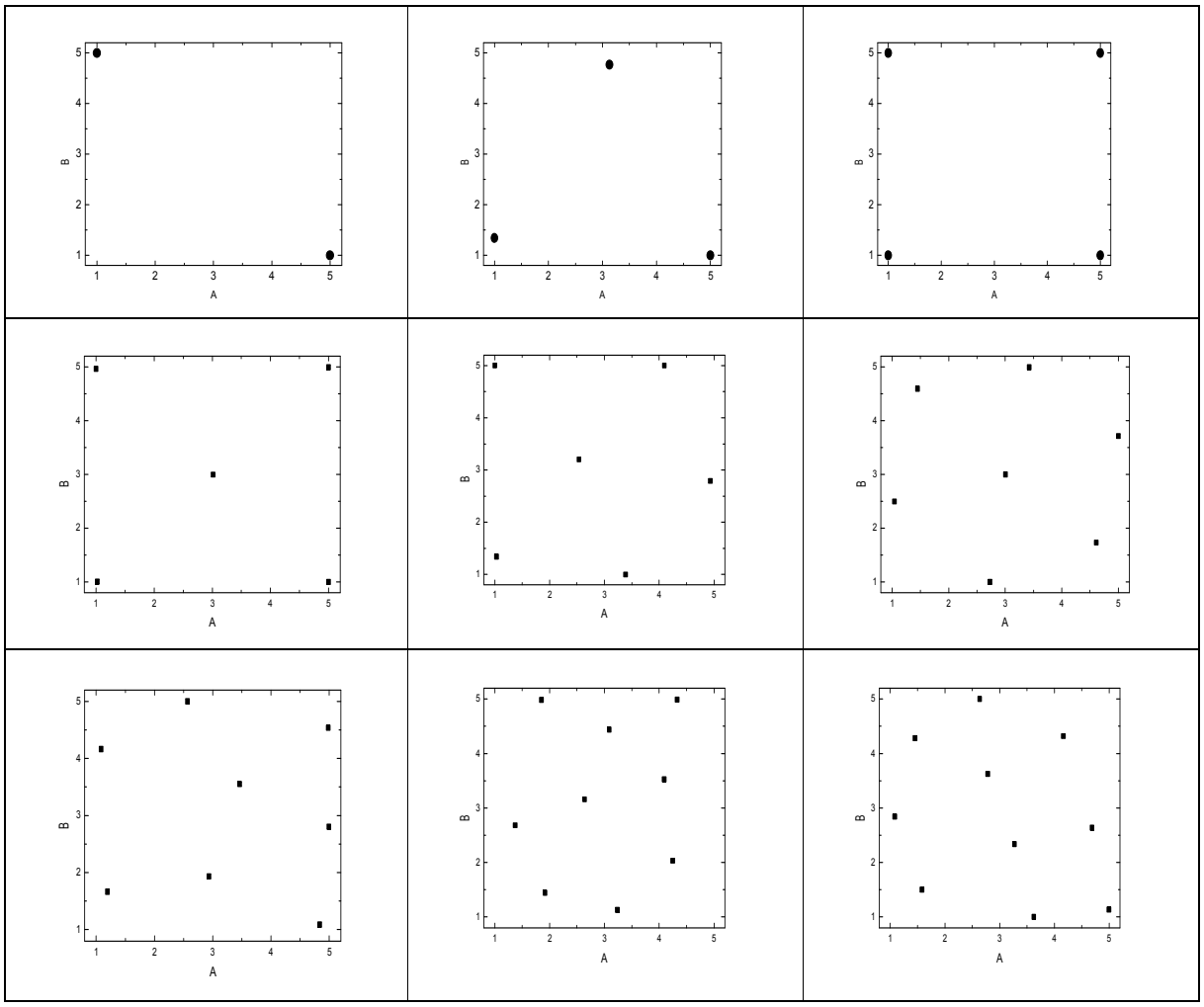


Figure 3. Filling the two-dimensional space with the growing number of points.

As can be seen the distribution of points is visually quite uniform. It is important to notice that the results in Fig.3 are not unique. The algorithm is based on the iteration procedure and includes random coordinate shifts, thus there could be several solutions obeying the conditions of $min\{D_{max}\}$ and $max\{r_{min}\}$.

Another example we would like to demonstrate is the design of seven component mixtures with 30 experiments. The results from the suggested algorithm are shown in Fig. 4 as two-dimensional slices of the seven-dimensional concentration hypercube. For successful design it is important that the concentrations of components (coordinates of points) would be not correlated to each other.

Correlation coefficients for each pair of mixture components were calculated and provided under the corresponding plots to evaluate this issue. The correlation is computed as the covariance between the two variables divided by the square root of the product of their variances. It varies from -1 to +1. For the sake of brevity only selected projections are shown. Apparent slight non-uniformity in 2D projections can be explained by the fact that the algorithm is aimed at uniform distribution of points in 7D hypercube which can hardly be visualized in a correct way.

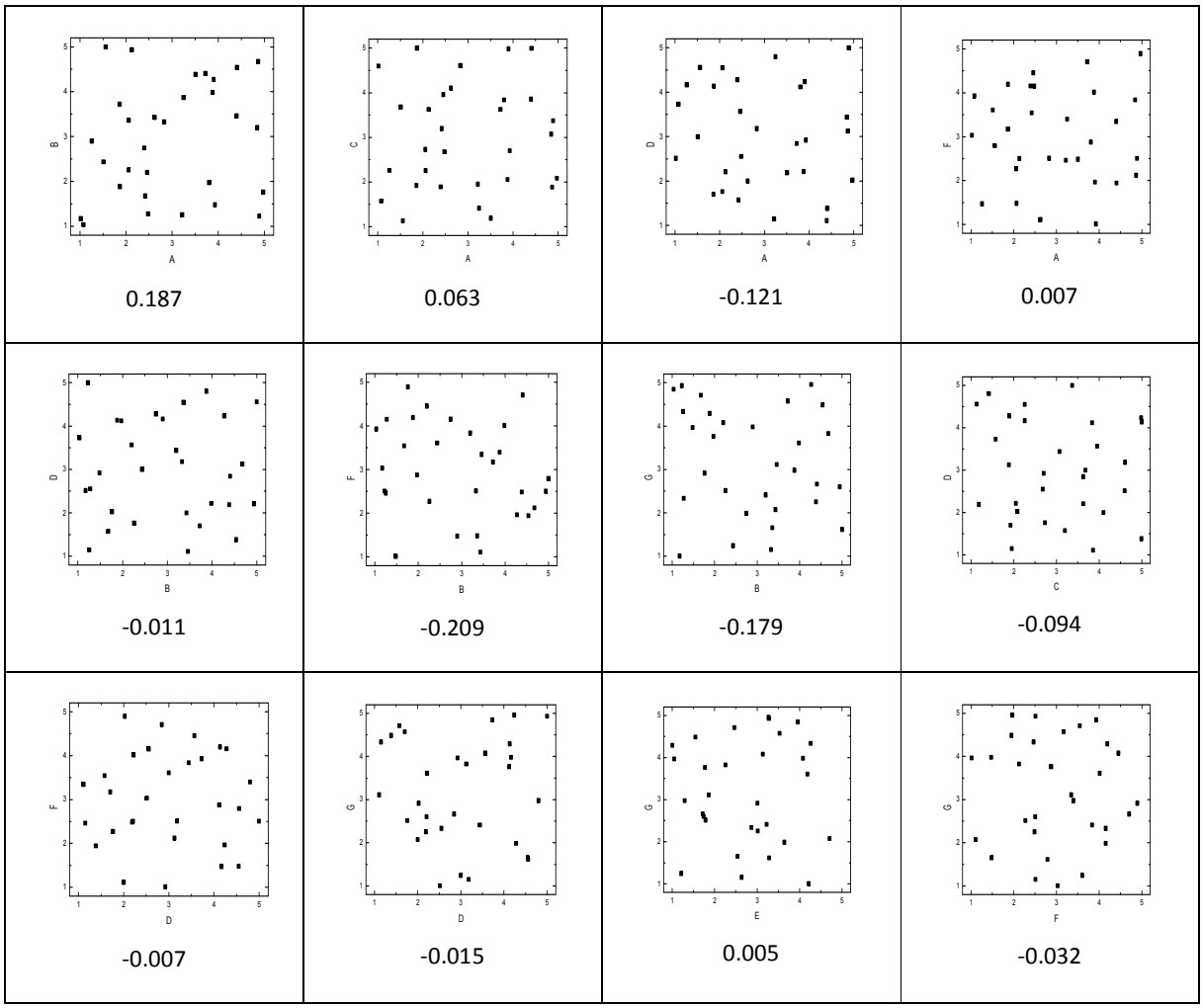


Figure 4. Distribution of 30 points in a seven-dimensional space with the suggested algorithm. Selected two-dimensional slices are shown.

Although the suggested design can handle any number of components with any number of samples the latter one should be not too small as it will obviously lead to unbalanced and sparse filling of experimental space with points. For example, Table 1 shows correlation coefficients between two corresponding components in two-dimensional slices of seven-dimensional hypercube filled with only ten experimental points. These coefficients can be used as “cheap and dirty” measure of uniformity. Small number of points leads to significant increase of correlations (see e.g. A/C, C/E, E/F) compared to Fig. 4. It is hard to suggest some rule of thumb for the choice of particular number of mixtures but from general considerations [2] and our experience *n*6* mixtures (where *n* is a number of components in a calibration mixture) allow for quite uniform distribution of points.

Table 1. Pairwise correlation coefficients for the components of seven-component mixture design modelled with 10 mixtures only.

A/B -0.393	A/C -0.633	A/D 0.119	A/E -0.317	A/F 0.284	A/G -0.030
	B/C 0.298	B/D 0.058	B/E -0.043	B/F -0.272	B/G -0.011
		C/D 0.106	C/E -0.456	C/F 0.056	C/G -0.030
			D/E -0.167	D/F 0.124	D/G 0.151
				E/F -0.318	E/G -0.112
					F/G -0.205

It is important to mention that fulfillment of $min\{D_{max}\}$ and $max\{r_{min}\}$ criteria distributes mixtures (points) evenly in a concentration hypercube, however, corresponding coordinates of the mixtures in two-dimensional projections are not necessarily evenly spaced. The price to pay for selectable number of points in design is certain unbalance in some of two-dimensional projections, e.g. in Fig.4 for D-G slice there will be no mixtures studied with low concentrations of D and G simultaneously. Another drawback is that the solution is not unique in each particular case. Nevertheless, in general the selected criteria of distribution uniformity seem to provide reasonable tool for design of multicomponent mixtures. The correlations between individual factors are quite low in the vast majority of the corresponding two-dimensional slices.

In cases when heteroscedastic noise is expected in analytical signals the suggested design can lead to lower precision in calibration for particular regions with pronounced heteroscedasticity. These situations can be effectively handled with recently reported adaptive WSP design algorithm [12] which can condense/reduce experimental points in certain regions of hypercube; however this will obviously require more experimental points. Our approach can be extended to handle heteroscedastic situations by correcting the resulted uniform distribution of points on a concentration scale for each particular component e.g. by shifting some of them towards higher values.

Proper validation is of ultimate importance for multivariate modelling. In case of suggested approach a separate validation subset can be designed in the same way using appropriate number of samples.

Performance validation

To test the suggested design on a real-world application we addressed the problem of simultaneous UV-Vis spectroscopic determination of several lanthanides and nitric acid in mixtures simulating composition of certain stages of PUREX-process (Plutonium URanium EXtraction) [15]. The possibility of on-line monitoring of these components is of high importance in spent nuclear fuel reprocessing. Due to limitations associated with high radioactivity of a real reprocessing media and limited availability of reference data one has to deal with model solutions to calibrate spectroscopic instrument for determination of these components. For that purpose we designed four-component mixtures containing

cerium, praseodymium, neodymium and nitric acid. These lanthanides are typical fission products and their concentration in reprocessing solutions must be controlled to provide for smooth process run. Designs with 10 and 25 samples were produced. For comparison purposes we also employed cyclic permutation design by Richard Brereton and Kennard-Stone design for four components with 25 samples. Concentration ranges for the components were relevant to the compositions of real PUREX-process solutions and were as follows: Ce 10-2000 mg/L, Pr 10-1000 mg/L, Nd 10-3500 mg/L, HNO_3 0.4-4 mol/L [15].

UV-Vis measurements were performed in 187-1020 nm wavelength range with AvaSpec spectrometer (Avantes BV, Holland) in 1 cm cuvette. Measurements were averaged over 10 scans, each one performed in 0.5 sec. The resulted spectra were used for PLS (projection on latent structures) modelling [16] to produce regression models for prediction of concentration for each individual component in the sample; the models were validated with full cross-validation. The following spectral ranges were employed: Ce 330-890 nm, Pr 400-680 nm, Nd 350-880 nm, nitric acid 310-390 nm. The rest of the spectrum was ignored as irrelevant for particular component. The parameters of the validation plot in "measured vs. predicted" coordinates are given in Table 2. PLS modelling was performed in The Unscrambler 9.7 (CAMO, Norway) software.

Table 2. Parameters of the "measured vs. predicted" plot for PLS models according to different designs. Full cross-validation.

	Slope	Offset	RMSECV	R^2
Cyclic permutation design, 25 points				
Ce	0.39	640	640	0.23
Pr	0.95	31	53	0.98
Nd	0.99	6	92	0.99
HNO_3	0.95	0.10	0.29	0.95
Kennard-Stone design, 25 points				
Ce	0.29	546	909	0.06
Pr	0.96	16	68	0.98
Nd	0.95	76	241	0.98
HNO_3	0.97	0.05	0.23	0.98
Suggested design, 10 points				
Ce	-0.38	1435	1027	0.23
Pr	0.88	73	144	0.85
Nd	0.99	9	112	0.99
HNO_3	0.95	0.09	0.23	0.97
Suggested design, 25 points				
Ce	0.12	930	563	0.22
Pr	0.97	13	30	0.99
Nd	0.99	5	50	0.99
HNO_3	0.92	0.14	0.22	0.94

It is important to point out that cerium has no distinct bands in the specified spectral region, thus its quantitative determination is hardly possible under selected experimental conditions however it was modelled to check whether chance correlation will be significant in the system.

Comparison of the results reveals that all designs can handle the problem in cases when relevant spectral signal is available. In case of cerium none of the designs was affected by chance correlations,

corresponding R^2 values are very low. Detailed inspection of RMSECV values leads to the conclusion that suggested design outperforms cyclic permutation design and Kennard-Stone design in terms of prediction performance. Even if using only ten points to fill the concentration hypercube still the precision of the resulted PLS models for Nd and nitric acid is comparable with that for two other designs. In case of equal number of experimental points RMSECV values are the lowest with suggested approach. An obvious problem with Kennard-Stone design in this application is that it starts filling the hypercube from the corners and with 25 points it achieves only three different concentration levels for each component. This number is five for cyclic permutation design and it is equal to the number of sample (10 or 25) with proposed algorithm.

Conclusion

The algorithm for design of multicomponent mixtures in calibration experiments is suggested. The proposed approach can produce designs with any required number of components and calibration samples due to rejection of the fixed concentration levels concept. The number of calibration samples can be adjusted according to the particular considerations of time and resources at hand. Uniform distribution of experimental points in multidimensional concentration space allows for accuracy of resulted multivariate regression models comparable with more sample-demanding results.

The software implementation of suggested calibration design algorithm is available by request from the authors as a stand-alone PC application.

Acknowledgment

Dr. Andrey Bogomolov (Samara State University, Russia and J&M Analytik, Germany) and Dr. Federico Marini (University of Rome “La Sapienza”) are gratefully acknowledged for a number of valuable suggestions. This work was partially financially supported by Government of Russian Federation, Grant 074-U01.

References

1 L. E. Agelet, C. R. Hurburgh, *Critical Reviews in Analytical Chemistry*, 2010, **40**, 246.
2 Analytical Methods Committee, *Analyst*, 1994, **119**, 2363.
3 A. Bogomolov, S. Dietrich, B. Boldrini, R.W. Kessler, *Food Chemistry*, 2012, **134** , 412.
4 D. Kirsanov, V. Babain, M. Agafonova-Moroz, A. Lumpov, A. Legin, *Radiochimica Acta*, 2012, **100**, 185.
5 R. Leardi, *Analytica Chimica Acta*, 2009, **652**,161.
6 P. W. Araujo, R. G. Brereton, *Trends in Analytical Chemistry*, 1996, **15**, 63.
7 R. Brereton, *Analyst*, 1997, **122**, 1521.
8 J. A. Muñoz, R. G. Brereton, *Chemometrics and Intelligent Laboratory Systems*, 1998, **43**, 89.
9 Y.-Z. Liang, K.-T. Fang, Q.-S. Xu, *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**, 43.
10 K.-T. Fang, D. K. J. Lin, P. Winker, Y. Zhang, *Technometrics*, 2000,**42**, 237.
11 J. Santiago, M. Claeys-Bruno, M. Sergent, *Chemometrics and Intelligent Laboratory Systems*, 2012, **113**, 26.
12 A. Beal, M. Claeys-Bruno, M. Sergent, *Chemometrics and Intelligent Laboratory Systems*, 2014, **133**, 84.
13 R. W. Kennard, L. A. Stone, *Technometrics*, 1969, **11**, 137.

14 I.M. Sobol, *USSR Computational Mathematics and Mathematical Physics*, 1976, **16**, 236.
15 K. L. Nash, G. J. Lumetta, Advanced Separation Techniques for Nuclear Fuel Reprocessing and Radioactive Waste Treatment Woodhead Publishing Ltd., Cambridge, UK, 2011.
16 S. Wold, M. Sjöström, L. Eriksson, *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**, 109.