Analytica Chimica Acta 951 (2017) 46–57

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Diagonal designs for a multi-component calibration experiment

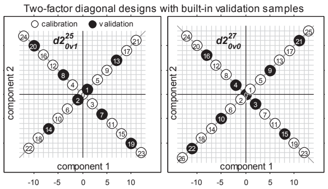
Andrey Bogomolov [a], [b], *

[a] Samara State Technical University, Molodogvardeyskaya Street 244, 443100 Samara, Russia
[b] Global Modelling, Rembrandtstraße 1, 73433 Aalen, Germany

HIGHLIGHTS

- Requirements of an optimal multi-component calibration experiment are considered.
- A family of diagonal designs for two-component calibration experiment is proposed.
- Design generalization to three or more components and its extensions are outlined.
- Diagonal design scheme contains a built-in validation subset.
- Diagonal DoE was compared to other multi-level designs using a new simulation-based approach.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 9 June 2016
Received in revised form
17 October 2016
Accepted 16 November 2016
Available online 26 November 2016

Keywords:
Design of experiment
Mixture analysis
Multivariate regression analysis
Multi-component calibration
Training set
Validation set

ABSTRACT

Modern spectroscopic and sensor technologies combined with multivariate modelling are increasingly used for the quantitative analysis of complex mixtures. Their performance depends directly on the data design chosen for model training and validation. A well-balanced calibration experiment with the fewest samples possible presents additional challenges when several mixture components (factors) need to be calibrated on the same dataset and subsequently quantified from the same multivariate measurement. This practically important problem stays poorly addressed by the theory of experimental design. This theoretical work systematically formulates the requirements to an optimal calibration/validation dataset and introduces a new family of calibration designs, where the samples are placed along the diagonals of an experimental space that is a hypercube. Such placement is appropriate due to reasonable assumptions about the linear nature of analytical response. Suggested filling schemes allow economical diagonal designs with intrinsic validation to be built for multiple factors presented in as many levels as the number of samples. The most important practical cases of two and three factors are considered in detail, and generalization to higher dimensions is outlined. Diagonal designs of any complexity can be generated using a simple geometrical scheme or with a supplied script.

1. Introduction

The importance of multivariate design of experiment (DoE) for enhancing analysis efficiency is well recognized [1]. DoE is traditionally focused on a number of problems [2] including the study of factor significance and variability effects as well as discovery of optimal experimental conditions that maximize the experimental outcomes, such as reaction yield, chromatographic separation, product quality attribute, taste perception, etc. Planning efficient calibration experiments for quantitative mixture analysis with an

* Rembrandtstr. 1, 73433 Aalen, Germany.
E-mail address: ab@globalmodelling.com.

http://dx.doi.org/10.1016/j.aca.2016.11.038
0003-2670/© 2016 Elsevier B.V. All rights reserved.

indirect (typically, spectroscopic) technique offers additional challenges that are poorly addressed by DoE theory [2]. These challenges are particularly relevant to quality control of pharmaceutical formulations [3–12], medical diagnostics and metabolomics [13,14], food quality analysis [15–18], environmental [19–26] or process monitoring [27–29], and more. In all those cases the analysis demands for the simultaneous determination of several constituents from the same multivariate measurement. The real-life mixtures do not usually comply with the closure constraint required by classical mixture designs [30].

The effects of experimental design on the prediction accuracy of multivariate calibrations have been illustrated in several studies [8,31–35]. The traditional DoE framework that is building an economical dataset providing an adequate coverage of the experimental space and keeping the factors mutually uncorrelated equally applies to multivariate calibrations. But in the latter case it needs further elaboration due to distinctive purposes, samples, and modelling requirements.

One of the most important distinctions of the calibration DoE is related to the number of necessary levels. A well-balanced calibration set consists of a suitably large number of samples uniformly distributed across each factor's (e.g. analyte concentration) range and therefore involves many levels. In multivariate regression the levels should significantly outnumber the factors. This reality makes the classical DoE's minimal use of just a few levels badly suited to the case of quantitative mixture analysis.

The deficiency in theory, methodology, and software for multi-component calibration design forces researchers to construct their own custom designs [3,6–8,12,19,24,27,28] or to adopt known DoEs even though they may be non-optimal or poorly suited. Full or fractional factorial and composite designs in three to five levels are most ubiquitous in the reviewed publications of the last two decades [3,5,11,13,19,21,23,35–37]. Non-uniform sample distribution over the experimental space (and worse, over their individual concentration scales) is a serious disadvantage of widely used central composite designs and their derivatives [5,19,31,32,35]; it may make resulting calibration model biased and the prediction error-prone. Besides, the sample grouping hinders the selection of an independent validation subset, thus enhancing the risk of overfitting. Other examples of adaptations of classical DoE approaches to the calibration problem are given by Doehlert design [20,38] and by different orthogonal designs [4,9,10,18,29]. Some authors use random selection of samples to construct the calibration [14,20,25,38–40], but it is generally detrimental: it requires many experimental runs to ensure adequate coverage and to fill the design space uniformly.

Publications devoted to the development of multi-level DoE for simultaneous calibration of two or more factors are rare. Brereton et al. extended the classical Plackett-Burman two-level screening design to five [41] and later to seven levels [42]; this approach has been applied in some later works [22,26,34]. While the number of samples (N) is relatively high ($N = l^2$), the number of levels (l) in this DoE remains limited. Latin hypercube sampling (LHS) approach [43], where each point on the grid of levels contains one and only one sample, is most efficient in terms of the number of experiments. Properties of LHS designs depend on the filling scheme and can be adjusted. Uniform designs by Fang and Lin [44,45] exploit the concept of maximum uniformity determined from the discrepancy function. The use of uniform designs in multi-component calibration experiments in chemistry is straightforward, especially when the underlying response function is unknown [44]. Belonging to the LHS type the uniform designs are computationally very intensive, which makes them mainly usable in table form. A simpler uniformity-based approach was recently suggested [46]. Reviewed calibration designs neither considered a validation strategy nor included a built-in validation set, which should be a necessary attribute of practical modelling.

This paper presents a new multi-component calibration design based on an assumption that under certain conditions, which hold for the majority of analytical objects in chemistry and industry, the samples can be placed along the diagonals of the concentration space, inherently minimizing the components' mutual correlations. This diagonal design basically belongs to the Latin hypercube family with $N = l$, independently of the number of factors, but can be extended to more saturated schemes. The suggested simple and intuitive population algorithm is aimed at reaching the maximum uniformity of both calibration and predefined validation samples along individual factor scales.

2. Theoretical basis

2.1. Requirements of a multi-component calibration dataset

While classical optimization DoE is aimed at revealing an external objective function with as few experiments as possible, calibration design aims instead at minimizing the prediction error being a function of the design itself. Before suggesting calibration design approaches, let us summarize key criteria responsible for the quality of multi-component calibration datasets.

2.1.1. Uncorrelated factors

Pairwise correlation coefficients (r) between the factors should be possibly close to zero. This requirement has a paramount importance to avoid confounding effects, where the regression model for a factor is based on a spurious correlation with another factor.

2.1.2. Uniformity

A well-balanced calibration dataset must be sufficiently large and have uniformly distributed samples along the factor ranges; hence, the necessity of having many design levels. The levels can be predefined as in LSH scheme or generated by the design algorithm, as in the random DoE or in Kirsanov's design [46]. In practice, calibration sets include tens or hundreds of measurements depending on the analytical problem, model complexity, and data availability. From the multivariate modelling point of view, the number of design levels should be larger than the model complexity, e.g. the number of latent variables in partial least-squares (PLS) regression. The letter statement may require some further explanation. Even in the case of univariate (one factor) regression under the linearity assumption a two-point calibration is statistically wrong. Populating these two levels with experimental points is also non-optimal because: first, the error is generally non-even over the calibration region and its correct modelling requires additional points between the two edge levels; and second, the linearity is typically a hypothesis that can only be proven experimentally using multiple levels. Similar reasons for having more levels than LVs are valid in the multivariate calibration case. Besides, if the multivariate regression model is based on only a few levels containing sample groups there is a higher risk of overfitting, when LVs start to describe irrelevant spectral differences between the groups giving an overoptimistic calibration/validation statistics.

2.1.3. Design space coverage

A multivariate regression model reliably operates only on the space defined by the factor variation intervals. The samples, therefore, should adequately cover the experimental space. But absolute uniformity of their distribution in the whole k-factor space—as in uniform design [44]—is not generally required. The samples may follow any pattern provided that it is consistent with