Received: 15 May 2018 | Revised: 18 August 2018 | Accepted: 5 November 2018

DOI: 10.1002/cem.3103

RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Multiresponse and multiobjective latent variable optimization of modern analytical instrumentation for the quantification of chemically related families of compounds: Case study—Solid-phase microextraction (SPME) applied to the quantification of analytes with impact on wine aroma

Marco S. Reis[1] | Ana C. Pereira[1,2] | João M. Leça[2,3] | Pedro M. Rodrigues[2] | José C. Marques[2,3]

[1] CIEPQPF, Department of Chemical Engineering, University of Coimbra, Coimbra, Portugal

[2] Faculty of Exact Sciences and Engineering, University of Madeira, Funchal, Portugal

[3] Institute of Nanostructures, Nanomodelling and Nanofabrication (I3N), University of Aveiro, Aveiro, Portugal

Correspondence
Marco S. Reis, CIEPQPF, Department of Chemical Engineering, University of Coimbra, Pólo II-Rua Sílvio Lima, 3030-790 Coimbra, Portugal.
Email: marco@eq.uc.pt

Funding information
Agência Regional para o Desenvolvimento da Investigação Tecnologia e Inovação (ARDITI), Grant/Award Number: M1420-09-5369-FSE-000001

Abstract

The optimized operation of modern analytical instrumentation is a critical but complex task. It involves the simultaneous consideration of a large number of factors, both qualitative and quantitative, where multiple responses should be quantified and several goals need to be adequately pondered, such as global quantification performance, selectivity, and cost. Furthermore, the problem is highly case specific, depending on the type of instrument, target analytes, and media where they are dispersed. Therefore, an optimization procedure should be conducted frequently, which implies that it should be efficient (requiring a low number of experiments), as simple as possible (from experimental design to data analysis) and informative (interpretable and conclusive). The success of this task is fundamental for achieving the scientific goals and to justify, in the long run, the high economic investments made and significant costs of operation. In this article, we present a systematic optimization procedure for the prevalent class of situations where multiple responses are available regarding a family of chemical compounds (instead of a single analyte). This class of problems conducts to responses exhibiting mutual correlations, for which, furthermore, several goals need to be simultaneously considered. Our approach explores the latent variable structure of the responses created by the chemical affinities of the compounds under analysis and the orthogonality of the interpretable extracted components to conduct their simultaneous optimization with respect to different analysis goals. The proposed methodology was applied to a real case study involving the quantification of a family of analytes with impact on wine aroma.

KEYWORDS

definitive screening designs, design of experiments, HS-SPME, latent variable modeling, optimizing analytical instrumentation, principal component analysis, wine production

1 | INTRODUCTION

Modern analytical instrumentation comprises a wide variety of equipment implementing distinct measurement principles in order to extract fundamental information for the qualitative or quantitative chemical characterization of samples. Common measurement principles include spectroscopy Fourier-transform infrared spectroscopy (FTIR), Ultaviolet-visible spectroscopy (UV-VIS), fluorescence, and Nuclear Magnetic Resonance spectroscopy (NMR), RAMAN), chromatography High-performance liquid chromatography (HPLC) and Gas chromatography (GC), electrochemistry, and mass spectrometry, as well as their hyphenation, ie, combination for the characterization of samples, leading to highly informative responses structured as high-order tensors or profiles.[1,2] Analytical equipment is also expensive (in terms of investment and operation costs), justifying the existence of strong economic and scientific pressures to take the most out of their capabilities when collecting information from the samples under analysis. However, this is a far from trivial task. There are always a variety of equipment settings, operation modes, and procedures to handle and prepare the samples that are essentially problem specific (ie, dependent on the device, nature of the analytes of interest, and the matrix where they are present) and that may have a critical impact on the final performance of the device and on the quality of the measurements. Therefore, the adoption of efficient and systematic approaches to carry out this optimization is of paramount importance in modern analytical laboratories, in order to take full advantage of the equipment potential and convert the high capital investments into high added-value scientific and economic outcomes.

Besides the number of factors and case-dependent nature, other important difficulties are the existence of multiple responses (usually several analytes need to be quantified) and the coexistence of multiple perspectives and concerns to accommodate during the optimization process. In fact, there is usually more than one response of interest to be considered and several optimization goals, which makes the entire tuning process rather involved. Understandably, this may represent a daunting task for highly skilled analysts, strongly focused on the science behind the chemical phenomena they are studying, but that lack, for a variety of reasons, the necessary contact, mindset, and training on statistical design of experiments (DOE) and data analysis (including multivariate data analysis and model building) that are instrumental for meeting all these challenges.

In this article, we present an efficient, simple, and informative solution to this class of problems and illustrate its application to a case study presenting all the sources of complexity identified above and that are representative of a wide class of modern instrumentation optimization scenarios. More specifically, the case study regards the optimization of a state-of-the-art extraction methodology for the analysis of volatile compounds in liquid samples, namely, headspace solid-phase microextraction (HS-SPME). This technique is known to present good sensitivity and accuracy in the quantification of volatiles, having also the advantage of not requiring the use of solvents, as well as implying a simple procedure and limited manipulation of the samples. Furthermore, the possibility for scaling up the analysis pace through full automation makes it one of the main technological solutions in the analysis of volatile compounds. In the present work, HS-SPME is applied to the quantification of several critical wine compounds for establishing its flavor, in particular, the family of volatile fatty acids (VFA), namely, isobutyric acid, butyric acid, isovaleric acid, valeric acid, hexanoic acid, octanoic acid, nonanoic acid, decanoic acid, and dodecanoic acid. These analytes are extracted from wine samples through HS-SPME and quantified in a hyphenated instrument that comprises a gas chromatograph (for separating the analytes in distinct peaks of the moving media—the carrier gas) combined with a mass spectrometer detector (for identifying the analytes in each peak); this equipment is known as a Gas chromatography coupled with Mass spectroscopy (GC-MS). The responses of interest are the peaks area, or chromatographic responses, for each one of the analytes. The larger this area is, the better a given compound is detected and quantified. The optimization procedure consists in finding the optimal settings for all the equipment, operation, and materials options (levels) that lead to the maximum peak area for these compounds, while meeting some additional, user-defined goal. After reviewing the relevant technical literature, the following seven HS-SPME factors were identified as potentially interfering with the final chromatographic outcomes: preincubation time, headspace sample volume, type of fiber, time and extraction temperature, agitation during extraction, and ethanol content.

Currently, the process for optimizing the operation of analytical instrumentation in the analytical laboratories is often carried out through sequential tuning or one-factor-at-a-time (OFAT) approaches,[3-11] that are not only inefficient with respect to the use of resources but also prevent the identification of any relevant interactions among factors, which may in turn lead to suboptimal performances.[12] The specificity of this application scenario recommends the adoption of a strategy that is efficient not only in terms of planning, execution, and analysis[6,13-19] but also able to cope with multiple correlated responses arising from the need to quantify families of chemically related compounds and effective in finding the best levels for all the extraction factors under analysis. Therefore, in this work, a two-stage systematic procedure was implemented that combines the estimation efficiency of definitive screening designs (DSDs) with the ability of latent

variable modeling methods, namely, principal components analysis (PCA), to infer the main orthogonal directions of variability in the correlated responses and to use such orthogonality to conduct multicriterion optimization in order to quickly get to the optimal configuration and access to additional degrees of freedom to tune other interesting aspects of the operation of the equipment, maximizing the added value of its outcomes.

The rest of this article is organized as follows. In Section 2, we briefly describe the experimental setup considered in this work. The proposed methodology is presented in detail in Section 3. Afterwards, the results obtained are presented in Section 4, as well as data analysis, which include the latent variable modeling of the responses, where an interpretation of the components' meaning is provided and the latent variables' scores are regressed onto the factors. The analysis of these regression models leads not only to the optimal settings for the factor levels but also to some interesting additional capabilities for tuning extra HS-SPME extraction factors, according to the user preferences. Section 5 concludes the article, with a summary of the central aspects of the case study, the general proposed workflow, and the main findings of the study.

2 | EXPERIMENTAL BACKGROUND

The extraction study was carried out in a TriPlus autosampler, in SPME mode. The seven extraction factors (manipulated variables) considered are presented in Table 1, together with their operational range (minimum and maximum) and data type (categorical or quantitative). The factor levels were defined after a preliminary screening of the relevant information available in the technical literature. The fiber coatings where those most often considered to analyse VFA in food matrixes. The minimum extraction temperature was fixed according to the equipment capabilities and the maximum value fixed to cover the highest values often reported in literature. The preincubation and extraction time (ETI) levels were set to reasonable values in order to secure operational reasonability, for which an excessive time-consuming methodology must be avoided. The ethanol influence in extraction procedure was tested by comparing no dilution (18% of ethanol) with 1:2 and 1:4 dilution levels. The operation ranges resulted from a careful analysis of the literature and accumulated experience on using HS-SPME in other experimental scenarios.

After extraction of the volatile fractions from the samples through HS-SPME, these fractions are fed to a chromatographic column equipped with a mass spectroscopy detector. The first device (GC) separates the analytes of interest (the nine VFA) while the second Mass spectroscopy (MS) collects molecular information for their identification. This analysis was carried out in a hyphenated GC-MS system (TRACE GC Ultra gas chromatograph coupled with an ISQ single quadrupole from Thermo Scientific, Hudson, New Hampshire). The carrier gas was helium at a constant flow rate of 1 mL min^{-1}. The total GC run time was set to 45 minutes. The mass spectrometer was operated in electron impact (EI) mode, at 70 eV. A chromatogram of a standard solution and also an example of a wine sample, are presented in Figure 1.

3 | EXPERIMENTAL DESIGN AND DATA ANALYSIS WORKFLOW

The present application scenario is characterized by the existence of the following aspects: (a) multiple responses related to the quantification of several chemically related compounds, (b) multiple goals with different degrees of relevance to be contemplated, and (c) multiple factors to be simultaneously manipulated during the optimization stage. Besides the

TABLE 1 Experimental factors considered in the optimization of the headspace solid-phase microextraction (HS-SPME) extraction performance and the associated levels (for categorical variables) and ranges (for continuous variables)

Factor	Categorical/Quantitative	Factor Levels
Type of fiber (-)	Categorical	{L1-PA, L2-DVB/Car/PDMS}
Sample volume (ml)	Categorical[a]	{5,10}
Preincubation time (min)	Quantitative	[0,10]
Extraction time (min)	Quantitative	[15, 20]
Extraction temperature (°C)	Quantitative	[40, 55]
Agitation (-)	Categorical	{L1-yes, L2-no}
Ethanol content (%)	Quantitative	[4.5, 18]

[a]Sample volume was considered as a two-level categorical factor, for practical reasons related to limitations on sample preparation.

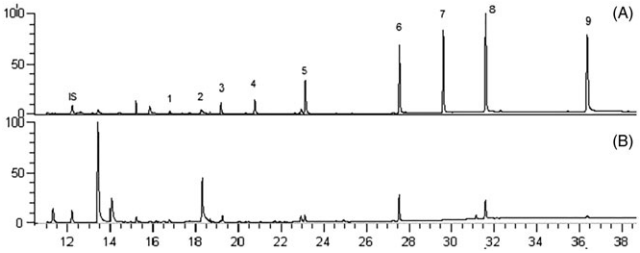


FIGURE 1 Chromatograms of the following: A, a 1-mg L^{-1} standard solution and B, a fortified wine. IS, internal standard; 1, isobutyric acid; 2, butyric acid; 3, isovaleric acid; 4, valeric acid; 5, hexanoic acid; 6, octanoic acid; 7, nonanoic acid; 8, decanoic acid; 9, dodecanoic acid

capability to handle multiple responses/goals/factors, the analysis workflow should also be efficient, as the purpose is to implement it in a routine way for other equipment and analyte families as well, which will require the repetition of the entire plan/execute/analyze/validate cycle. In this context, the several building blocks were carefully considered and the options assessed regarding their fitness to these practical requirements, in order to facilitate the sustainable acceptance and adoption by the laboratory technicians and researchers. The resulting methodological workflow proposed for this application scenario, and for others with a similar typology, is summarized in Figure 2.

In brief terms, the workflow consists of two stages. The first stage is dedicated to the planning and execution of experiments, while the second one addresses the analysis of data collected, with special focus on the latent variable modeling of responses and their interpretation.

3.1 | Stage 1—Experimental design and execution

The first stage starts with the definition of the factors and responses of interest. The former are identified through a risk-based approach informed by the accumulated experience of analysts and an in-depth coverage of the relevant literature regarding applications of the experimental set up. The responses are usually clear from the context of the problem under analysis. They are typically directly related to the analytes to be detected and quantified. The next step is the DOE

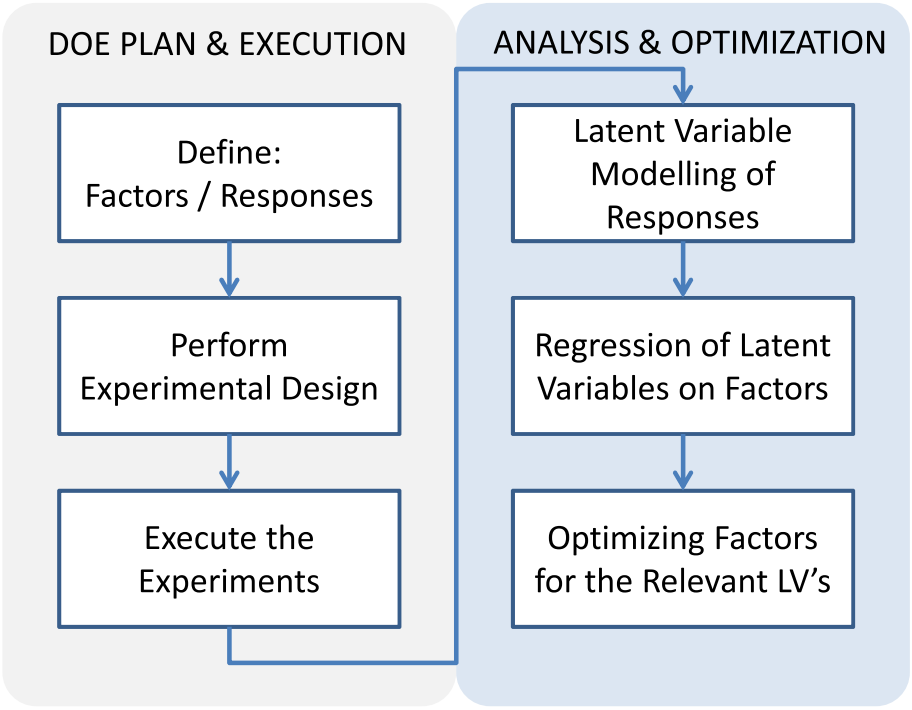


FIGURE 2 Workflow followed for the optimization of multiple responses from chemically related compounds in analytical instrumentation scenarios

following an appropriate approach. In this work, we opt to implement DSD for setting the treatments to be tested. Other approaches could have been also used, but we opt for DSDs due to their efficiency (one of the criteria to be considered in the workflow) and interesting estimation features. In fact, DSDs have been attracting a considerable interest, as they only require one more experiment than twice the number of factors under analysis and still allow for the estimation of the main effects without any aliasing with each other or with two-factor interactions.[21-23] Furthermore, contrary to classical screening designs, pure quadratic effects can also be estimated, if the number of factors is greater than six and their strength is large enough to be detected.[24] The original version of the DSD was restricted to quantitative factors[25] but was recently extended to incorporate two-level categorical factors,[21] allowing its application for addressing the factors described in Table 1. These constitute interesting features that fit very well in the requirements established for building the analysis workflow presented in Figure 2. Furthermore, the technical literature still lacks case studies on the application of DSD to real world applications, which was another motivation to include them in this work. The final step of the first stage consists of performing all experiments by the random order established in the design and collecting the outcomes from the analytical protocol and instrumentation. This step is rather labor intensive, requiring high-skilled personnel, and lengthy operation periods. Therefore, the efficiency requirement is critical for the success of a systematic workflow and its wide acceptance.

3.2 | Stage 2—Data analysis and optimization

After concluding the experimental design and execution stage, the workflow proceeds to the analysis of results and the multiobjective optimization of the factor levels. This stage begins with the usual exploratory data analysis activities, strongly focused on the application of graphical tools to preliminarily assess the existence of potential problems in the data and infer the quality of the results and main trends. Afterwards, the focus changes to the analysis of the underlying structure of the responses (the first step of stage 2) and their relationship with the experimental factors (second step of stage 2). This analysis deviates from the classical workflow for treating DOE data, which is centered in a linear regression model structure as the basis for analysis and inference. For the case of DSD designs, this corresponds to the quadratic model,

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{i,j} + \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \beta_{jk} x_{i,j} x_{i,k} + \sum_{j=1}^{m} \beta_0 x_{i,j}^2 + \varepsilon_i, \tag{1}$$

where y_i is the response for ith observation; $x_{i,j}$ the level of factor j in the ith observation; β_0, ...,$\beta_{m,m}$ are the model parameters to be estimated (most of them will likely be zero, according to the sparsity principle); and ε_i is the residual term, which is assumed to be a random variable following an *i. i. d.* $N(0, \sigma^2)$ distribution. However, in the present scenario, one has to deal with multiple responses that, by their chemical characteristics and the consequent impact on the measurement system, present strong correlations. This correlation structure is a manifestation of an underlying latent structure linking all the responses. Remembering that the responses are the chromatographic peak areas of the analytes, they can be expected to be related with each other according to mutual chemical structure affinities. Thus, the right modeling formalism to describe the phenomena under observation should contemplate this structure as one of its pillars, which can be achieved by adopting a latent variable description for the responses.

This characteristic is not found in classical DOE situations, or even in the usual application of latent variable approaches in regression. In fact, the overwhelming majority of applications of latent variable regression approaches are for the case where both the **X**-block and the **Y**-block of variables are related with the same underlying latent variables (**T**), according to the following model structure:

$$\begin{aligned} \mathbf{X} &= \mathbf{T} \cdot \mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{T} \cdot \mathbf{Q}^T + \mathbf{G}, \end{aligned} \tag{2}$$

where, **X** is a $(n \times m)$ matrix of input variables (n stands for the number of observations and m for the number of input variables); **Y** is a $(n \times p)$ matrix of output variables (p stands for the number of output variables); **T** is the $(n \times a)$ matrix of latent variables or scores (a stands for the dimension of the latent variable space, also called the pseudorank); **P** and **Q** are two matrices of coefficients (usually called loadings) with dimensions $(m \times a)$ and $(p \times a)$, respectively; and **E** and **G** are $(n \times m)$ and $(n \times p)$ residual matrices of the **X** and **Y** spaces, respectively. The model structure presented in equation (2) is usually estimated through principal components regression (PCR)[26-29] or partial least squares (PLS).[30-33]

This model structure finds wide applicability and success in chemometric problems,[34-36] soft sensor development,[37,38] large-scale process monitoring and predictive analysis,[20,39,40] and biosystems.[41] However, there is a mismatch between model structure in equation (2) and the structure of data collected in the present situation, as detailed next.

Data generated by the current DOE plan presents a strong latent variable structure in the responses, the **Y**-block, but the **X**-block is full-rank and completely free of any latent structure. More, the latent variable model presented in equation (2) is based on the assumption that the variability drivers, **T**, are unobservable quantities; only the **X**- and **Y**-blocks can be observed, which are dependent on such unmeasured sources of variability. However, in the current situation the variability drivers can in fact be measured and appear in the **X**-block. This implies that neither the model structure in equation (1) nor that in equation (2) conform to the data structure found in instrumental applications such as the one under analysis and an alternative model should be set as the basis to conduct the analysis. In this context, the following model structure provides a suitable match with the main variability components of experimental data in the present class of problems[1]:

$$\mathbf{Y} = \mathbf{T}\cdot\mathbf{Q}^T + \boldsymbol{\varepsilon}_2,\ \boldsymbol{\varepsilon}_2 \sim i.i.d. N_m(\mathbf{0}_m, \boldsymbol{\Sigma}_{\varepsilon_2}), \tag{3}$$

$$\mathbf{T} = \widetilde{\mathbf{X}}\cdot\mathbf{B} + \boldsymbol{\varepsilon}_1,\ \boldsymbol{\varepsilon}_1 \sim i.i.d. N_a(0, \boldsymbol{\Lambda}_{\varepsilon_1}). \tag{4}$$

Equation 3 models the correlation pattern of the responses through a latent variable model, whereas Equation 4 establishes the dependence between the latent variables of the responses and experimental factors. Thus, the underlying causal dependence chain is $\mathbf{X}\xrightarrow{B}\mathbf{T}\xrightarrow{Q}\mathbf{Y}$. In this pair of equations, **T** corresponds to the unobservable latent variables structuring the variability of the responses (related to the chemical structure of the analytes), and **Q** is the loading matrix for the responses (**Y**-block). The latent variables for the responses depend on the factor settings, **X**, through the model presented in Equation 4, where $\widetilde{\mathbf{X}}$ is the **X**-matrix extended to the model space. $\boldsymbol{\varepsilon}_1$ stands for the regression residuals with covariance $\boldsymbol{\Lambda}_{\varepsilon_1}$ (a diagonal matrix), and $\boldsymbol{\varepsilon}_2$ represents the reconstruction error of the responses, modeled as an *i.i.d.* N_m process (m being again the number of output variables considered), with zero mean and covariance $\boldsymbol{\Sigma}_{\varepsilon_2}$.

The estimation of the model structure in equations (3)-(4) consists of a two-stage process:

- First, the latent variable structure of the responses is modeled, through an approach capable to infer its scores and loadings. This can be very efficiently achieved through PCA, using singular value decomposition, spectral analysis or the NIPALS algorithm.[26,27,42] Then, the scores are estimated and their meaning and importance assessed through the analysis of the loadings and eigenvalues, respectively.
- The second stage consists of building a regression model to predict the relevant scores found in stage 1, using the experimental factors as regressors and possibly some higher order bilinear or quadratic terms found to be statistically significant. The models for the different scores are constructed independently, as they are uncorrelated quantities.

The components should be selected based on considerations of the amount of original variability they are able to explain and their individual interpretation with respect to the optimization goals. With the models for the response scores developed and their meaning clearly established by the analysis of the associated loadings, the workflow proceeds to the optimization step. The meaning of the scores establishes the nature of the optimization task to be conducted. At this point, the orthogonality property of the scores is exploited to address complementary optimization goals. It is well-known that PCA models have signal indeterminacy, meaning that the loading for each variable may have one value or its symmetric, without changing the nature of the PCA solution. Thus, the loading coefficients' magnitudes and the corresponding signals should be considered case by case, in order to formulate the adequate optimization problems to solve for each latent variable (or principal component scores). This task is more effectively described with the support of real data, and therefore more information about it is deferred to the next section, where it will be fully covered.

4 | RESULTS AND DISCUSSION

In this section, we report the results, data analysis, and main findings obtained with the implementation of the procedure described in Section 3. Following the workflow presented in Figure 2, we have considered seven extraction parameters (fiber coating, preincubation time, ETI, extraction temperature, headspace sample volume, agitation during extraction, and ethanol content) and nine responses to be simultaneously optimized (maximized), which are the

analytical response (peak area) of the following VFAs: isobutyric acid, butyric acid, isovaleric acid, valeric acid, hexanoic acid, octanoic acid, nonanoic acid, decanoic acid, and dodecanoic acid.

4.1 | Experimental design and execution

A DSD design was implemented leading to 18 runs with different combinations of factor levels, as described in the design matrix presented in Table 2.

The generation of the DSD design matrix and the subsequent data analysis (model estimation and optimization) were conducted in JMP-PRO ver. 12.1.0 (64-bit) (SAS Institute Inc.). All extraction experiments were carried out using the same wine matrix, namely a 3-year-old Madeira wine sample.

4.2 | Latent variable modeling of the responses

This stage is part of the two-step estimation procedure for the model structure represented by equations (3)-(4) and the main interest here is the description of the conjoint behavior of the responses under analysis, ie, the chromatographic responses of the nine VFA compounds involved in establishing the wine flavor. The responses were autoscaled and then analyzed through PCA (or, equivalently, PCA was conducted over the correlation matrix of responses). PCA reveals a strong eigenstructure in the responses (Table 3), with the first two latent variables explaining approximately 96% of the original (scaled) data variability (henceforth designated as PC 1 and PC 2, which are also the ones having eigenvalues greater than 1, ie, complying with the Kaiser criterion for components selection[26,27]).

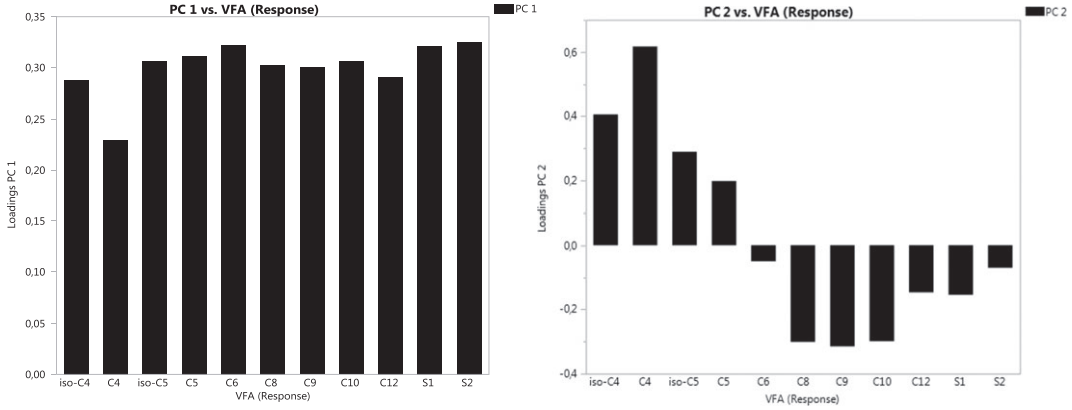
The loadings (or eigenvectors) for the first two components are presented in Figure 3. The analysis of the loadings profiles reveals a clear meaning for the nature of variability they are capturing. PC1 represents the mean performance of all chromatographic responses. It is very important to realize that they are highly and positively correlated, and therefore their overall performance can be, to a large extent, optimized simultaneously (remember that PC1 explains 85% of

TABLE 2 Design matrix obtained from the application of definitive screening design (DSD) to the factor settings presented in Table 1. (For the coding used in the levels L1 and L2, see also Table 1)

Order	Type of Fiber	Preincubation Time, min	Sample Volume, ml	Extraction Time, min	Extraction Temperature, °C	Agitation	Ethanol Content, %
1	L2	0	L1	40	40	L1	4.50%
2	L1	5	L1	15	40	L1	18%
3	L2	0	L2	15	55	L1	18%
4	L2	10	L2	40	40	L2	18%
5	L1	0	L1	15	55	L1	4.50%
6	L1	0	L2	40	47.5	L1	18%
7	L2	5	L2	40	55	L2	4.50%
8	L1	0	L1	40	55	L2	9%
9	L2	10	L1	40	55	L1	18%
10	L1	5	L1	27.5	47.5	L1	9%
11	L2	5	L2	27.5	47.5	L2	9%
12	L1	10	L2	15	55	L2	18%
13	L2	0	L1	27.5	40	L2	18%
14	L2	10	L2	15	40	L1	9%
15	L1	0	L2	15	40	L2	4.50%
16	L1	10	L2	27.5	55	L1	4.50%
17	L2	10	L1	15	47.5	L2	4.50%
18	L1	10	L1	40	40	L2	4.50%

TABLE 3 Principal components analysis (PCA) analysis of the responses (volatile fatty acids [VFA]): eigenvalues and percent of variance explained of the scaled **Y**-block data.

Number	Eigenvalue	Percent	Cumulative Percentage
1	9.3562	85.056	85.056
2	1.2029	10.935	95.992
3	0.2621	2.383	98.257
4	0.1457	1.325	99.699
5	0.0257	0.233	99.933
6	0.0049	0.045	99.977
7	0.0017	0.016	99.993
8	0.0006	0.005	99.999
9	0.0002	0.0001	100.000
10	0.0000	0.0000	100.000
11	0.0000	0.0000	100.000



(A) (B)

FIGURE 3 Principal components analysis (PCA) analysis of the responses (volatile fatty acids [VFA]): Loading (or eigenvectors) of the following: A, PC1 and B, PC2

the responses variability). This can be accomplished by maximizing PC1, because all loadings have positive signs, and the goal is to maximize the VFAs responses.

On the other hand, PC2 presents a curious structure. It is a contrast between analytes with short chain lengths (iso-C4, C4, C5) with those with larger chain lengths (C8 and above). This leads to novel and very interesting possibilities to tune the performance of the analytical method, as will be discussed in the Section 4.4, dedicated to the optimization task. The distributions of the scores and of the PCA residuals (distance to the PCA subspace) were also analyzed and confirmed that no deviating observations are present either in the model space or around it. Therefore, the subsequent analysis proceeds with the use of the first two components, PC 1 and PC 2, which have a rather clear meaning and efficiently comprise 96% of the original variability.

4.3 | Regression of the response latent variables on the design factors

In this second step of the estimation procedure for the model in equations (3)-(4) each latent variable (or principal component) is regressed onto the design factors, or **X**-blocks. Given the orthogonality between PC1 and PC2, this model-building task was conducted individually, for each one of the components. A summary of the models developed is presented Table 4, which includes the variables selected to incorporate each model, their signal, and associated significance. The model-fitting quality is quite good, taking into consideration that by predicting the scores of PC1

TABLE 4 Summary of results for the models developed for predicting the scores of the first two principal components. The factors that were found significant for the prediction of each principal component scores' are indicated through the signal of the respective coefficient and the associated *P* value (for assessing the trend and the associated statistical significance of the coefficient). Also presented for each model is the coefficient of determination (R^2) and the adjusted coefficient of determination (R^2_{adj})

Design Factors	PC 1	PC 2
Fiber (F): L1-PA		[+] ($P = 0.0002$)
Preincubation time (PIT)		
Sample volume (SV)	[−] ($P = 0.0417$)	
Extraction time (ETI)	[+] ($P = 0.0029$)	[−] ($P = 0.0030$)
Extraction temperature (ETE)		[−] ($P = 0.0001$)
Agitation (A)		
Ethanol content (EC)	[−] ($P < 0.0001$)	[−] ($P = 0.0001$)
*EC*ETE*		[+] ($P = 0.0003$)
*EC*EC*	[+] ($P = 0.0699$)	
*VOL*ETI*		
*F*VOL*		
R^2	0.83	0.92
R^2_{adj}	0.77	0.89

and PC2 ($\widehat{\mathbf{T}}$) is actually predicting all the VFA's through the reconstruction PCA formula ($\widehat{\mathbf{T}}\cdot\mathbf{Q}^T$). Analyzing the models' composition, it is possible to verify that most of the variables affecting PC 1 are not active for PC 2 or have the same signal in both. The only exception is the ETI, which presents opposite signals in both models, even though its influence is higher for PC 1 (this is rather clear when analyzing the profile plots of the responses versus this factor). This scenario will be explored during the optimization stage, described next.

4.4 | Optimization of the design factors

With the relevant models derived for describing the variability of the responses, taking into account their latent variable nature, they were then used to conduct the optimization of the HS-SPME procedure. In this regard, the main goal in this case study was to maximize the chromatographic response that the instrument provides for the nine analytes of interest. As PC 1 represents the overall performance of the equipment (his loadings are of similar magnitudes and all positive), the simultaneous optimization of all responses—more precisely, of 85% of their variation—was suitably done by maximizing PC1. This is the primary goal and has priority over any additional concern to be considered. Its formulation and solution ($\mathbf{S}_1$) appear in the first line of Table 5.

However, the interpretation of PC 2 and the low overlap (or consistency) of its model terms with those of the model for PC 1, open interesting perspectives for contemplating additional dimensions in the definition of the optimal levels. As this component is still responsible for 11% of the original data variation, its influence remains important and justifies a closer analysis in order to find the levels for the factors not contemplated in the optimization of PC 1. This will improve even further the analytical performance of the method, allowing for a better use of the equipment and increasing the quality of results.

TABLE 5 Formulation for the different optimization goals considered in this case study

Optimization Goal	Formulation
Maximize overall performance	$\mathbf{S}_1 : \mathbf{X} = \arg\max\{\mathbf{PC1}(\mathbf{X})\}$
Minimize discrepancy in performances	$\mathbf{S}_2 : \mathbf{X} = \arg\{\mathbf{PC}2(\mathbf{X}) = 0\}$
Maximize performance of hard to extract components	$\mathbf{S}_3 : \mathbf{X} = \arg\max\{\mathbf{PC}2(\mathbf{X})\}$

TABLE 6 Optimal levels for the experimental factors optimizing headspace solid-phase microextraction (HS-SPME) performance

Factor	PC1 Overall Performance	PC2 Minimize discrepancy	PC2 Maximize Performance of Difficult Components	Optimal Compromise
Type of fiber	...	L1-PA	L1-PA	L1-PA
Sample volume	10 ml	...	...	10 ml
Preincubation time	...	...	...	0 min
Extraction time	40 min	Weak effect	Weak effect	40 min
Extraction temperature	...	55°C	40 °C	Case dependent
Agitation	...	...		L2-No
Ethanol content	18%	Weak effect	18%	18%

In this regard, we have considered two alternatives perspectives, between which the analyst can opt according to what may be more relevant for his/her final goal. In the first perspective, the concern is to have the same quantification performance for all components under analysis. In other words, no analyte should be privileged during the extraction and analysis procedure. As PC2 represent the contrast between compounds with shorter chain lengths (with positive loadings) and those with higher chain lengths (with negative loadings), an unbiased performance would be induced by finding the conditions for which PC 2 is zero. This is the formulation presented in the second row of Table 5.

The second perspective corresponds to attributing more importance to a certain type of components. For instance, we may want to favor the conditions that are more convenient to quantify analytes that are known to be harder to detect by the instrument. In the present situation, they correspond to the shorter chain length compounds. As these have positive loadings, the proper solution can be found by finding the factors that maximize PC 2 (see third row of Table 5).

Therefore, as a result of the eigenstructure structure of the responses and the orthogonality (and interpretability) of their latent variables, the final solution can be improved by considering the reunion of conditions $\mathbf{S}_1 \cup \mathbf{S}_2$ or $\mathbf{S}_1 \cup \mathbf{S}_3$, according to the particular interests of the analyst. The solutions found for the optimization problems formulated in Table 5 are presented in Table 6, where the optimal compromise is also proposed. In this compromise, there is one factor, extraction temperature, whose level is dependent on the user preference towards the secondary optimization goal, which involves PC 2. These secondary goals are referred on the second or third rows of Table 5 and appear in the columns of Table 6 regarding "PC2—Minimize Discrepancy" and "PC2—Maximize Performance of Difficult Components."

4.5 | Final comments

The present approach exploits the orthogonality of the latent variable (principal components), in order to address different optimization problems. Therefore, it is a multiple response, multiobjective (and of course multifactor) approach. The source of eigenstructure (or latent structure) is the regularity between the properties of chemically related compounds, which creates the correlation between their quantification when processed by the extraction and separation equipment. The proposed approach match all criteria initially established for adoption in practical scenarios, namely, it is simple, efficient, and informative. Furthermore, it can be readily extended to the optimization of other type of analytical instruments, whenever the goal is the analysis of families of analytes (ie, chemically related compounds), which greatly enlarges its domain of application.

5 | CONCLUSIONS

In this article, we propose an approach for optimizing modern analytical instrumentation for the quantification of multiple, chemically related analytes. These families of analytes originate a latent variable structure in the responses, which can be efficiently described by PCA. The orthogonality property of the extracted latent variables is used to incorporate multiple objectives in the optimization, increasing the value of data collected by the instrumentation.

This approach was applied to an analytical technique of great importance in modern laboratories: HS-SPME. The case study involves the quantification of an important family of compounds that influence the wine aroma. The methodology was successfully applied and validated.

The proposed two-stage method can be used for the large class of problems involving the optimization of analytical instrumentation for the quantification of families of chemically related analytes and has therefore a wide application scope.

ACKNOWLEDGEMENTS

Ana C. Pereira and João M. Leça are thankful to the Agência Regional para o Desenvolvimento da Investigação Tecnologia e Inovação (ARDITI) for the financial support through their Post Doc and PhD grants in the scope of the project M1420-09-5369-FSE-000001.

ORCID

Marco S. Reis https://orcid.org/0000-0002-4997-8865

REFERENCES

1. Reis MS, Saraiva PM. Prediction of profiles in the process industries. *Ind Eng Chem Res*. 2012;51:4524-4266.
2. Woodall WH, Spitzner DJ, Montgomery DC, Gupta S. Using control charts to monitor process and product quality profiles. *J Qual Tech*. 2004;36(3):309-320.
3. Hecht ES, Oberg AL, Muddiman DC. Optimizing mass spectrometry analyses: a tailored review on the utility of design of experiments. *J Am Soc Mass Spectrom*. 2016;27(5):767-785.
4. Araujo PW, Brereton RG. Experimental design I. Screening. *TrAC Trends Anal Chem*. 1996;15(1):26-31.
5. Leardi R. Experimental design in chemistry: a tutorial. *Anal Chim Acta*. 2009;652(1–2):161-172.
6. Bezerra MA, Santelli RE, Oliveira EP, Villar LS, Escaleira LA. Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta*. 2008;76(5):965-977.
7. Montgomery DC, Runger GC. Design of experiments with several factors. In: Statistics A, ed. *Probability for Engineers*. I (ed.) ed. United States of America: John Wiley & Sons; 2003:505-594.
8. Antony J. 3—understanding key interactions in processes. In: *Design of Experiments for Engineers and Scientists*. Oxford: Butterworth-Heinemann; 2003:17-28.
9. Lundstedt T, Seifert E, Abramo L, et al. Experimental design and optimization. *Chemom Intel Lab Syst*. 1998;42(1–2):3-40.
10. Callao MP. Multivariate experimental design in environmental analysis. *TrAC Trends Anal Chem*. 2014;62:86-92.
11. González AG. Two level factorial experimental designs based on multiple linear regression models: a tutorial digest illustrated by case studies. *Anal Chim Acta*. 1998;360(1–3):227-241.
12. Hibbert DB. Experimental design in chromatography: a tutorial review. *J Chromatogr B*. 2012;910:2-13.
13. Tarley CRT, Silveira G, dos Santos WNL, et al. Chemometric tools in electroanalytical chemistry: methods for optimization based on factorial design and response surface methodology. *Microchem J*. 2009;92(1):58-67.
14. Hanrahan G, Montes R, Gomez FA. Chemometric experimental design based optimization techniques in capillary electrophoresis: a critical review of modern applications. *Anal Bioanal Chem*. 2008;390(1):169-179.
15. Weston DJ. Ambient ionization mass spectrometry: current understanding of mechanistic theory; analytical performance and application areas. *Analyst*. 2010;135(4):661-668.
16. Ferreira SLC, Dos Santos WNL, Quintella CM, Neto BcB, Bosque-Sendra JM. Doehlert matrix: a chemometric tool for analytical chemistry—review. *Talanta*. 2004;63(4):1061-1067.
17. Dejaegher B, Vander Heyden Y. Experimental designs and their recent advances in set-up, data interpretation, and analytical applications. *J Pharmaceut Biomed*. 2011;56(2):141-158.
18. Oberg AL, Vitek O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res*. 2009;8(5):2144-2156.
19. Ferreira SLC, Bruns RE, Ferreira HS, et al. Box-Behnken design: an alternative for the optimization of analytical methods. *Anal Chim Acta*. 2007;597(2):179-186.
20. Kourti T. Application of latent variable methods to process control and multivariate statistical process control in industry. *Int J Adapt Contr Signal Process*. 2005;19(4):213-246.
21. Jones B, Nachtsheim CJ. Definitive screening designs with added two-level categorical factors. *J Qual Technol*. 2013;45(1):1-9.
22. Jones B. 21st century screening experiments: what, why, and how. *Qual Eng*. 2016;28(1):98-106.

23. Jones B, Nachtsheim CJ. Blocking schemes for definitive screening designs. *Dent Tech*. 2016;58(1):74-83.
24. Jones B, Nachtsheim CJ. A class of three-level designs for definitive screening in the presence of second-order effects. *J Qual Tech*. 2011;43(1):1-15.
25. Jones B, Nachtsheim CJ. A class of three-level designs for definitive screening in the presence of second-order effects. *J Qual Tech*. 2011;43(1):1-15.
26. Jackson JEA. *User's Guide to Principal Components*. New York: Wiley; 1991.
27. Jolliffe IT. *Principal Component Analysis*. New York (etc.): Springer; 2002.
28. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. *Chemometrics—a textbook*. Amsterdam: Elsevier; 1988.
29. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Dent Tech*. 1993;35(2):109-135.
30. Martens H, Naes T. *Multivariate Calibration*. Chichester: Wiley; 1989.
31. Höskuldsson A. PLS regression methods. *J Chemometr*. 1988;2(3):211-228.
32. Helland IS. On the structure of partial least squares regression. *Commun Statist-Simula*. 1988;17(2):581-607.
33. Helland IS. Rotational symmetry, model reduction and optimality of prediction from the PLS population model. In: *2nd International Symposium on PLS and Related Methods*; 2001.
34. Naes T, Isaksson T, Fearn T, Davies T. *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester (UK): NIR Publications; 2002.
35. Sahni NS, Aastveit AH, Naes T. In-line process and product control using spectroscopy and multivariate calibration. *J Qual Tech*. 2005;37(1):1-20.
36. Pereira AC, Reis MS, Saraiva PM, Marques JC. Development of a fast and reliable method for long- and short-term wine age prediction. *Talanta*. 2011;86:293-304.
37. Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Comput Chem Eng*. 2009;33(4):795-814.
38. Rato TJ, Reis MS. Multiresolution soft sensors (MR-SS): a new class of model structures for handling multiresolution data. *Ind Eng Chem Res*. 2017;56(13):3640-3654.
39. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Dent Tech*. 1995;37(1):41-59.
40. MacGregor JF, Jaeckle C, Kiparissides C, Koutoudi M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal*. 1994;40(5):826-838.
41. Burger J, Geladi P. Hyperspectral NIR imaging for calibration and prediction: a comparison between image and spectrometer data for studying organic and biological samples. *Analyst*. 2006;131(10):1152-1160.
42. Jobson JD. *Applied Multivariate Data Analysis*. New York: Springer-Verlag; 1992.

How to cite this article: Reis MS, Pereira AC, Leça JM, Rodrigues PM, Marques JC. Multiresponse and multiobjective latent variable optimization of modern analytical instrumentation for the quantification of chemically related families of compounds: Case study—Solid-phase microextraction (SPME) applied to the quantification of analytes with impact on wine aroma. *Journal of Chemometrics*. 2019;33:e3103. https://doi.org/10.1002/cem.3103