Vitale, R.; Palací-López, DG.; Kerkenaar, H.; Postma, G.; Buydens, L.; Ferrer, A. (2018). Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments. Chemometrics and Intelligent Laboratory Systems. 175:37-46. https://doi.org/10.1016/j.chemolab.2018.02.002

Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments

Raffaele Vitale[a,b,c,*], Daniel Palací-López[a,*], Harmen H.M. Kerkenaar[d], Geert J. Postma[d], Lutgarde M.C. Buydens[d], Alberto Ferrer[a]

[a]*Grupo de Ingeniería Estadística Multivariante, Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain*
[b]*Molecular Imaging and Photonics Unit, Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001, Leuven, Belgium*
[c]*Laboratoire de Spectrochimie Infrarouge et Raman - UMR 8516, Université de Lille - Sciences et Technologies, Bâtiment C5, 59655, Villeneuve d'Ascq, France*
[d]*Radboud University Nijmegen, Institute for Molecules and Materials, Analytical Chemistry, P.O. Box 9010, 6500 GL, Nijmegen, The Netherlands*

Abstract

This article explores the potential of Kernel-Partial Least Squares (K-PLS) regression for the analysis of data proceeding from mixture designs of experiments. Gower's idea of pseudo-sample trajectories is exploited for interpretation purposes. The results show that, when the datasets under study are affected by severe non-linearities and comprise few observations, the proposed approach can represent a feasible alternative to classical methodologies (i.e. Scheffé polynomial fitting by means of Ordinary Least Squares - OLS - and Cox polynomial fitting by means of Partial Least Squares - PLS). Furthermore, a way of recovering the parameters of a Scheffé model (provided that it holds and has the same complexity as the K-PLS one) from the trend of the aforementioned pseudo-sample trajectories is illustrated via a simulated case-study.

Keywords: mixture designs of experiments, Kernel-Partial Least Squares (K-PLS), pseudo-sample trajectories, Scheffé and Cox polynomials, Partial Least Squares (PLS), Ordinary Least Squares (OLS)

1. Introduction

A wide range of products currently used in daily life result from processing blends of two or more ingredients. Hence, the physicochemical properties of these products mainly depend on the raw materials being mixed and on the proportions in which they are added. Alloys, as well as drugs and foodstuffs, are just some of the numerous examples where this applies, and their

*Corresponding authors:
Telephone number: +34963877007 ext. 74935
Email address: rvitale86@gmail.com (Raffaele Vitale); dapalpe@gmail.com (Daniel Palací-López)
These authors have equal contributions

manufacturing can be considered a so-called *mixture problem* [1]. Traditionally, mixture problems are defined as those in which i) the proportions x_i of the I different constituents are related to the aforementioned properties, ii) these proportions are of at least as much relevance as their absolute quantities, and iii) their sum must be a fixed value (usually 1 or 100%):

$$\sum_{i=1}^{I} x_i = 1 \tag{1}$$

where $0 \leq x_i \leq 1$. This perfect collinearity restriction makes it impossible to modify the composition of any one of the ingredients independently from the rest. This implies that classical polynomial fitting by traditional methods, like Ordinary or Generalised Least Squares (OLS/GLS), is unfeasible as they assume the regressors to be linearly independent. Alternative approaches e.g. the Scheffé models or their reparametrisation, the Cox models, can be exploited in these circumstances. However, both of them show several limitations: the interpretation of Scheffé models' coefficients may not be straightforward and Cox models' ones might not be directly estimated by OLS/GLS [2, 3]. In order to solve such issues, Partial Least Squares (PLS) regression-based techniques can be resorted to [4]. Nevertheless, if the mixture data under study are affected by strong non-linear relationships, which is rather common in e.g. industrial scenarios, even applying classical PLS taking into account additional interaction, inverse and/or higher-degree terms may not constitute an appropriate modelling strategy since it assumes their underlying structure is linear [5]. A good option in these situations may be represented by the so-called kernel-based techniques [6], which also encompass Kernel-Partial Least Squares (K-PLS) regression and have already been broadly used in different fields of interest [7–11]. Unfortunately, kernel-based methodologies suffer from a particular drawback: the information about the weights or the loadings of the original variables is lost. Many possibilities to recover this information exist, but authors commonly abstain from exploiting them essentially because they do not permit the graphical interpretation of the final models. Recently, the idea of the pseudo-sample trajectories, originally described by Gower and Hardings in [12], has been extended to overcome this limitation [13–18].
The main aim of this article is to evaluate the feasibility and the possible advantages of coupling K-PLS to the pseudo-sample trajectories for the analysis of mixture designs of experiments. The potential of such a combination will be assessed via simulated and real case-studies.

2. Methods

2.1. Scheffé and Cox models

Applying the constraint in Equation 1, the linear (first-order) and quadratic (second-order) Scheffé canonical polynomials can be expressed as:

$$\text{Linear model:} \qquad y = \sum_{i=1}^{I} \beta_i x_i + \epsilon \tag{2}$$

$$\text{Quadratic model:} \qquad y = \sum_{i=1}^{I} \beta_i x_i + \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \beta_{i,j} x_i x_j + \epsilon \tag{3}$$

being y the value of the response property to be predicted, β_i the first-order model coefficient related to the i-th constituent of the mixture, $\beta_{i,j}$ the model coefficient for the interaction between the i-th and the j-th ingredient and ϵ an error term. In other words, β_i corresponds to the expected value of y for the hypothetical *pure mixture* composed by the only i-th constituent, while $\beta_{i,j}$ measures the synergism or the antagonism between the i-th and the j-th ingredient.
Although Scheffé polynomials can be fitted by conventional OLS, the interpretation of their parameters is not straightforward. For this reason, they are commonly reformulated into their equivalent Cox models:

$$\text{Linear model:} \qquad y = \alpha_0 + \sum_{i=1}^{I} \alpha_i x_i + \epsilon \tag{4}$$

$$\text{s.t.} \sum_{i=1}^{I} \alpha_i s_i = 0 \tag{5}$$

$$\text{Quadratic model:} \quad y = \alpha_0 + \sum_{i=1}^{I} \alpha_i x_i + \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \alpha_{i,j} x_i x_j + \sum_{i=1}^{I} \alpha_{i,i} x_i^2 + \epsilon \tag{6}$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{I} \alpha_i s_i = 0 \\ \sum_{j=1}^{I} c_{i,j} \alpha_{i,j} s_j = 0 \quad \forall i \in [1,\ 2,\ \ldots,\ I] \end{cases} \tag{7}$$

where s_i is the proportion of the i-th ingredient in a specific mixture set as reference *a priori*; α_0 connotes the zero-order term of the polynomial; α_i and $\alpha_{i,i}$ denote the first-order and second-order model coefficients related to the i-th constituent of the mixture, respectively; $\alpha_{i,j}$ is the model coefficient for the interaction between the i-th and the j-th ingredient; and $c_{i,j} = \frac{1}{2}$ if $i \neq j$ or $c_{i,j} = 1$ if $i = j$. Here, α_0 represents the expected value of y for the reference mixture, α_i equals the difference between the expected value of y for the pure mixture composed by the only i-th constituent and the expected value of y for the reference mixture, and both $\alpha_{i,i}$ and $\alpha_{i,j}$ contribute to the response function curvature as for classical polynomials.
The relationship between the Scheffé model coefficients and the Cox model ones is derived in the Supplementary Material.
Cox canonical polynomials are probably the most intuitive approaches for mixture problem solving. However, the computation of their coefficients cannot be carried out by OLS/GLS. PLS regression constitutes a way to bypass this obstacle.

2.2. Partial Least Squares (PLS) and Kernel-Partial Least Squares (K-PLS) regression

Partial Least Squares (PLS) regression is a latent variable-based method for modelling the inner relationships between a matrix of predictors, $\mathbf{X}$, and a set of response variables, $\mathbf{Y}$. The basic idea behind it is predicting $\mathbf{Y}$ from the A-dimensional subspace of $\mathbf{X}$, which maximises its covariance with $\mathbf{Y}$ by resorting to the principles of the Nonlinear Iterative PArtial Least Squares algorithm originally developed by Herman Wold in [19]. In the mixture scenario, $\mathbf{X}$ ($N \times I$) would contain the proportions of the I ingredients for the N sampled blends, while $\mathbf{Y}$ ($N \times K$) the values of their K properties of interest. In this sense, the PLS structure model can be written as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F} \tag{8}$$

being $\mathbf{B}$ ($I \times K$) the array of PLS regression coefficients and $\mathbf{F}$ ($N \times K$) the so-called PLS $\mathbf{Y}$-residuals matrix, respectively.
By PLS one does not need to assume linearly independent regressors. Hence, it can be utilised to calculate the parameters of Cox models of various complexity by possibly augmenting $\mathbf{X}$ with interaction and/or higher-than-first-order terms to take into account their corresponding effect on $\mathbf{Y}$. However, as detailed before, PLS assumes the underlying structure of the data under study is linear. Therefore, when strong non-linearities affect them, which is rather common when mixture problems are dealt with, the application of this technique may not result in satisfactory outcomes, even if the aforementioned augmentation procedure is exploited. An alternative to classical PLS is represented by Kernel-Partial Least Squares (K-PLS) regression. K-PLS is based on the so-called *kernel transformation* of the original data matrix, $\mathbf{X}$. Its mathematical formulation is given by:

$$K(\mathbf{x}_n, \mathbf{x}_{n'}) = < \phi(\mathbf{x}_n), \phi(\mathbf{x}_{n'}) > \tag{9}$$

where $\mathbf{x}_n^{\mathrm{T}}$ and $\mathbf{x}_{n'}^{\mathrm{T}}$ are two generic row vectors of $\mathbf{X}$ to which a specific mapping function ϕ is applied, while $<$ and $>$ denote the inner product. By performing such a transformation for all the possible couples of vectors in $\mathbf{X}$, this array is converted into a squared symmetric *kernel* matrix, $\mathbf{K}$ ($N \times N$), whose elements measure the dissimilarity or distance between two different observations. This implies projecting the analysed data onto a new space, also known as *feature space*, allowing possible non-linear relationships to be modelled in a linear way. A comprehensive survey on the properties of the feature space and on the basic principles of the kernel transformation can be found in [5]. Finally, once the kernel matrix has been constructed, a standard PLS model is calibrated between $\mathbf{K}$ (double-centred) and $\mathbf{Y}$ as:

$$\mathbf{Y} = \mathbf{K}\mathbf{B}_{\mathrm{K-PLS}} + \mathbf{F}_{\mathrm{K-PLS}} \tag{10}$$

being $\mathbf{B}_{\mathrm{K-PLS}}$ ($N \times K$) the array of K-PLS regression coefficients and $\mathbf{F}_{\mathrm{K-PLS}}$ ($N \times K$) the so-called K-PLS $\mathbf{Y}$-residuals matrix, respectively. When K-PLS is resorted to, various mathematical functions can be applied to address the kernel transformation of $\mathbf{X}$. Table 1 lists those considered

Table 1 – Kernel functions referred to in this article and list of their adjustable parameters

Kernel type	Kernel function	Adjustable parameters
Second-order polynomial	$(\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_{n'})^2$	-
Third-order polynomial	$(\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_{n'})^3$	-
Fourth-order polynomial	$(\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_{n'})^4$	-
Radial-Basis-Function (RBF)	$\exp(-\frac{\|\mathbf{x}_n - \mathbf{x}_{n'}\|^2}{2\sigma})$	σ

in this article together with their adjustable parameters.
However, although K-PLS allows severe data non-linearities to be easily handled and, if such non-linearities exist, better fit and prediction quality to be achieved, its main disadvantage is associated to the fact that, due to the conversion of $\mathbf{X}$ into $\mathbf{K}$, the information about the importance of the original variables is not carried by $\mathbf{B}_{\mathrm{K-PLS}}$ and, thus, cannot be recovered by simply plotting the weights of the final models. To enable their interpretation, one can take advantage of Gower's idea of pseudo-sample trajectories.

2.3. Pseudo-samples and pseudo-sample trajectories

The term *pseudo-sample* connotes a particular observation carrying all the weight in one single regressor. For example, the vector [0, 0, ..., 1, 0, ..., 0] ($1 \times I$) can be looked at as one of the possible pseudo-samples related to e.g. the i-th ingredient of a generic mixture. If [0, 0, ..., 1, 0, ..., 0] is multiplied by the estimated regression coefficients, $\mathbf{b}$ ($I \times 1$), of a 1-response variable PLS model as in Equation 11:

$$\hat{y}_{\text{new}} = [0,\ 0,\ \ldots,\ 1,\ 0,\ \ldots,\ 0]\mathbf{b} = b_i \tag{11}$$

its predicted y-value, $\hat{y}_{\text{new}}$, equals the i-th element of $\mathbf{b}$, and thus provides insights into the contribution of the i-th constituent of the mixture to such a response variable.
Suppose now a pseudo-sample matrix, $\mathbf{Z}_i$ ($Z \times I$), is built so that its i-th column contains values ranging from 0 (minimum proportion of the i-th constituent) to 1 (maximum proportion of the i-th constituent) and 0 is set for all the other entries. Notice that the minimum and maximum proportion may differ depending on the specific mixture design of experiments taken into account. It follows:

$$\mathbf{Z}_i\mathbf{b} = \begin{bmatrix} 0, & 0, & \ldots, & 0, & 0, & \ldots, & 0 \\ & & & \vdots & & & \\ 0, & 0, & \ldots, & 1, & 0, & \ldots, & 0 \end{bmatrix} \mathbf{b} = \begin{bmatrix} 0 \\ \vdots \\ b_i \end{bmatrix} = \mathbf{b}_i \tag{12}$$

By graphing for all the I ingredients the respective $\mathbf{b}_i$ vectors as trajectories, an outline of the PLS regression coefficient plot would be obtained. For standard PLS, these trajectories would not render any additional information, but, as demonstrated in [14, 17], they could be extremely useful to interpret K-PLS models. Here, it would be only needed to i) execute on every $\mathbf{Z}_i$ the same kernel transformation as for $\mathbf{X}$ and ii) double-centre the resulting pseudo-sample kernel matrices as for $\mathbf{K}$ before carrying out the operation in Equation 12.

2.4. Pseudo-sample trajectories for mixture data and pseudo-sample-based response surfaces

Unfortunately, the described way of defining the different $\mathbf{Z}_i$ is not adequate when mixture problems are concerned, because it violates the constraint in Equation 1, i.e. it is impossible to modify the composition of any one of the blend constituents independently from the rest. In order to account for such a constraint, the pseudo-samples matrices $\mathbf{Z}_i$ should be adapted and structured so that the values in their i-th column range from the minimum to the maximum proportion of the i-th ingredient and all the elements of each one of their rows sum up to 1, provided that the fraction of the other constituents is equal. E.g. if a ternary mixture problem is faced, a hypothetical $\mathbf{Z}_1$ may have the following aspect:

$$\mathbf{Z}_1 = \begin{bmatrix} 0, & 0.5, & 0.5 \\ 0.2, & 0.4, & 0.4 \\ 0.4, & 0.3, & 0.3 \\ 0.6, & 0.2, & 0.2 \\ 0.8, & 0.1, & 0.1 \\ 1, & 0, & 0 \end{bmatrix} \tag{13}$$

As shown in Figure 1, this would mean spanning in the design space (a triangle) the direction connecting the vertex associated to the pure blend composed by only the i-th constituent ([1, 0, 0],

if $i = 1$) and the midpoint of its opposite side ([0, 0.5, 0.5], if $i = 1$).
More generically speaking:

$$\mathbf{Z}_i = \begin{bmatrix} \frac{1-z_{1,i}}{I-1}, & \frac{1-z_{1,i}}{I-1}, & \dots, & \min(\mathbf{x}_i), & \frac{1-z_{1,i}}{I-1}, & \dots, & \frac{1-z_{1,i}}{I-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1-z_{z,i}}{I-1}, & \frac{1-z_{z,i}}{I-1}, & \dots, & \vdots & \frac{1-z_{z,i}}{I-1}, & \dots, & \frac{1-z_{z,i}}{I-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1-z_{Z,i}}{I-1}, & \frac{1-z_{Z,i}}{I-1}, & \dots, & \max(\mathbf{x}_i), & \frac{1-z_{Z,i}}{I-1}, & \dots, & \frac{1-z_{Z,i}}{I-1} \end{bmatrix} \tag{14}$$

where $\mathbf{x}_i$ is the i-th column vector of $\mathbf{X}$ and $z_{z,i}$ refers to the $z \times i$ entry of $\mathbf{Z}_i$. Mind that this is not valid if the design space is not a simplex or if it is a simplex but the ingredient proportions do not vary from 0 to 1. Anyway, it is straightforward to extend the described approach to handle such situations [1].
As will be highlighted in Sections 4.3 and 4.4, the representation of the pseudo-sample trajectories derived in this way yields the so-called trace plot, traditionally used in Cox model analysis to get an idea of the linear and non-linear effects generated on the property of interest by the change in the proportion of every i-th ingredient. However, as most of these effects are confounded with those due to the simultaneous variation of the proportion of the other constituents, a precise identification of the individual coefficients of the corresponding Scheffé polynomial (if the Scheffé model holds) cannot be achieved by directly investigating it.

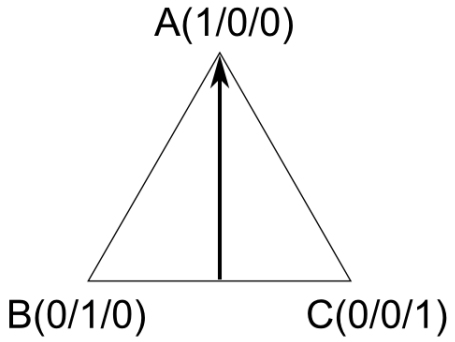

Figure 1 – Graphical representation of the direction spanned by the pseudo-sample trajectory associated to the constituent A in a generic ternary mixture design space. Notice that the three vertices correspond to the three pure mixtures composed by the only ingredient A, B or C, respectively

Alternatively, by using a combination of multiple pseudo-sample trajectories and graphing them in a contour plot, the response surface for the full mixture design space can be retrieved. To this end, every pseudo-sample matrix has to be constructed by i) fixing the proportions of all but two constituents, ii) increasing the proportion of one of these two constituents, and iii) decreasing the proportion of the other accordingly, for keeping the sum in Equation 1 constant and equal to 1. Such a procedure is iterated for different values of the fixed proportions of the rest of the ingredients. Graphically, this implies moving over the design space in the directions displayed in Figure 2, if, recalling the previous example, the proportion of A is fixed and the proportions of B

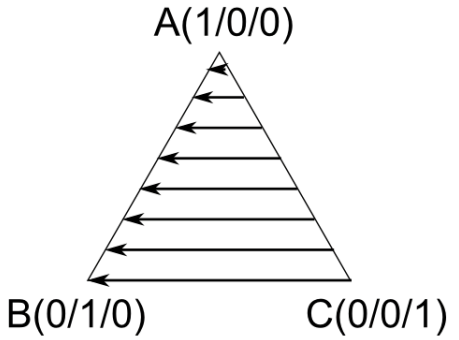

Figure 2 – Graphical representation of the direction spanned by the pseudo-sample trajectories exploited for retrieving the response surface for a generic ternary mixture design space. Notice that the higher the number of such trajectories, the higher the resolution of the final plot. In this specific case, in every single pseudo-sample matrix, the proportion of A is fixed, while that of both B and C varies

and C vary. In this case, a measure of the Scheffé model coefficients for the first-order effects of B and C and for their interaction can be recovered from the trajectory covering the BC side of the triangle, as will be illustrated in Section 4.1. Clearly, this is also valid for the trajectories covering the AB and the AC side of the triangle, not represented in Figure 2.

3. Datasets

Two simulated and four real datasets from mixture designs of experiments will be object of this study.

3.1. Data simulated according to a second-order polynomial model

66 artificial samples (with no replicates) of a ternary mixture homogeneously distributed inside a simplex and a single response variable were simulated according to the following second-order Scheffé model:

$$\begin{gathered} y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{1,2} x_1 x_2 + \beta_{1,3} x_1 x_3 + \beta_{2,3} x_2 x_3 \\ \beta_1 = 1.89,\ \beta_2 = -1.33,\ \beta_3 = 0.67,\ \beta_{1,2} = -2.89,\ \beta_{1,3} = 0.54,\ \beta_{2,3} = -1.33 \\ x_i \in [0,\ 1] \qquad \text{s.t} \sum_{i=1}^{3} x_i = 1 \end{gathered} \tag{15}$$

whose reformulation as a Cox model for a reference mixture where $s_1 = s_2 = s_3 = \frac{1}{3}$ can be written as (see Section 2.1 and Section I of the Supporting Material):

$$\begin{gathered} y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_{1,2} x_1 x_2 + \alpha_{1,3} x_1 x_3 + \alpha_{2,3} x_2 x_3 + \alpha_{1,1} x_1^2 + \alpha_{2,2} x_2^2 + \alpha_{3,3} x_3^2 \\ \alpha_0 = 0,\ \alpha_1 = 2,\ \alpha_2 = -2.67,\ \alpha_3 = 0.67,\ \alpha_{1,2} = -1.67,\ \alpha_{1,3} = 0.44, \\ \alpha_{2,3} = 0,\ \alpha_{1,1} = -0.11,\ \alpha_{2,2} = 1.33,\ \alpha_{3,3} = 0 \\ x_i \in [0,\ 1] \qquad \text{s.t} \begin{cases} \sum_{i=1}^{3} x_i = 1 \\ \sum_{i=1}^{3} \alpha_i s_i = 0 \\ \sum_{j=1}^{3} c_{i,j} \alpha_{i,j} s_j = 0 \\ s_1 = s_2 = s_3 = \frac{1}{3} \end{cases} \end{gathered} \tag{16}$$

According to Equation 16, the first constituent is characterised by a positive first-order and a small negative second-order term. Conversely, the second one features a negative first-order and a positive second-order term. The third ingredient exhibits a small positive first-order and no second-order term. Positive interaction terms were generated for both x_1x_2 and x_1x_3, while no interaction was assumed to involve x_2 and x_3. No noise was added after the data simulation.

3.2. Tablet data

This dataset was first described in [3]. 10 pharmaceutical tablets resulting from distinct blends of cellulose, lactose and phosphate were prepared to assess the influence of these substances on the release time of the active ingredient of the final manufactured drug. No replicates were performed.

3.3. Bubbles data

The bubbles data relate to an experiment also reported in [3]. Different proportions of two dish-washing liquids (DWL1 and DWL2), water and glycerol were combined to produce 24 soap mixtures (21 unique samples and 3 replicates) and determine which composition would have yielded the longest bubble lifetime.

3.4. Colorant data

This dataset was referenced in [20]. 49 blends (46 unique samples and 3 replicates) of different proportions of white (C_w), black (C_b), violet (C_v) and magenta (C_b) paints were manufactured to optimise the values of three specific colour responses: lightness (L^*), red-green tone (a^*) and yellow-blue tone (b^*).

3.5. Gasoline data

Different proportions of three gasoline constituents, catalytically cracked, C_5-isomer and reformate were mixed to produce 10 distinct blend samples according to an augmented simplex-centroid design of experiments [21]. Here, the idea is to evaluate the effect of these constituents on the octane rating of the final product, and possibly maximise it.

3.6. Data simulated according to a highly non-linear model

An additional small dataset, constituted by 12 artificial samples of a ternary mixture simulated according to an augmented simplex centroid design of experiments (with 2 replicates for the design centroid), was also generated based on the following model:

$$\begin{gathered} y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \log(x_1 + 0.01) + \beta_5 x_3^4 + \beta_6 \sin\left[(1.01 - x_2) * \pi\right] \\ \beta_1 = 4.87,\ \beta_2 = -1.35,\ \beta_3 = 5.67,\ \beta_4 = -1.52,\ \beta_5 = -0.35,\ \beta_6 = 8.00 \\ x_i \in [0,\ 1] \qquad \text{s.t} \sum_{i=1}^{3} x_i = 1 \end{gathered} \tag{17}$$

Normally distributed random noise was added to the response variable estimated by Equation 17.

4. Results

Both the simulated and real data were used for addressing an exploratory comparison among Scheffé polynomial fitting by means of OLS, Cox polynomial fitting by means of PLS, and K-PLS in terms of goodness-of-fit in calibration (R^2), goodness-of-fit in leave-one-out cross-validation (Q^2)[i], and Root Mean Square Error in leave-one-out Cross-Validation (RMSECV)[i] [22], and for illustrating that under certain conditions K-PLS can guarantee improved predictions and interpretation. Moreover, a way of retrieving the coefficients of a Scheffé polynomial (when they hold) from the pseudo-sample trajectories yielded by a K-PLS model with the same complexity was derived.
The whole set of routines resorted to for data processing and analysis was self-coded in MATLAB R2012b (Version 8.0.0.783) and is available on request.

4.1. Data simulated according to a second-order polynomial model

This section will be focused on demonstrating how pseudo-sample trajectories can be resorted to for recovering the coefficients of the Scheffé model in Equation 15, which the generation scheme outlined in Section 3.1 is based on. Figure 3 shows the shape of the trajectories spanning the three sides of the ternary mixture space of the first simulated dataset. The three lines reproduce the evolution of the values of the response variable, predicted by means of a 3-latent variable second-order polynomial K-PLS model, while moving from a vertex (pure blend) to another vertex of a triangle like that in Figure 2. Recall that every β_i ($\forall i \in [1,\ 3]$) measures the expected y for the pure mixture composed by the only i-th constituent. Therefore, each one of such parameters should match the predicted response at one of the two extremes of the respective pseudo-sample trajectories. As indicated in Figure 3, since the data at hand are noiseless, an exact match was here observed for β_1, β_2 and β_3. Analogously, the coefficients for the interaction terms x_1x_2, x_1x_3 and x_2x_3 can be computed as:

$$\begin{aligned}
\beta_{1,2} &= \frac{\hat{y}_{0.5,0.5,0} - 0.5\beta_1 - 0.5\beta_2}{0.25} = \frac{-0.44 - 0.5(1.89) - 0.5(-1.33)}{0.25} = -2.89 \\
\beta_{1,3} &= \frac{\hat{y}_{0.5,0,0.5} - 0.5\beta_1 - 0.5\beta_3}{0.25} = \frac{1.42 - 0.5(1.89) - 0.5(0.67)}{0.25} = 0.54 \\
\beta_{2,3} &= \frac{\hat{y}_{0,0.5,0.5} - 0.5\beta_2 - 0.5\beta_3}{0.25} = \frac{-0.44 - 0.5(-1.33) - 0.5(0.67)}{0.25} = -1.33
\end{aligned} \tag{18}$$

where $\hat{y}_{0.5,0.5,0}$, $\hat{y}_{0.5,0,0.5}$ and $\hat{y}_{0,0.5,0.5}$ denote the estimated y value for the binary blends with composition $x_1 = x_2 = 0.5$, $x_1 = x_3 = 0.5$, $x_2 = x_3 = 0.5$, respectively (the mid-points of the three trajectories in Figures 3a, 3b and 3c).

[i]Notice that when extreme observations are left out of the original data, responses for mixtures which are outside the calibration experimental domain are predicted (extrapolation). However, as this is the case for all the approaches under study, a fair comparison of the RMSECV values is still guaranteed.
Furthermore, for K-PLS, the objects/samples to be iteratively left out are removed from the datasets before the kernel transformation.

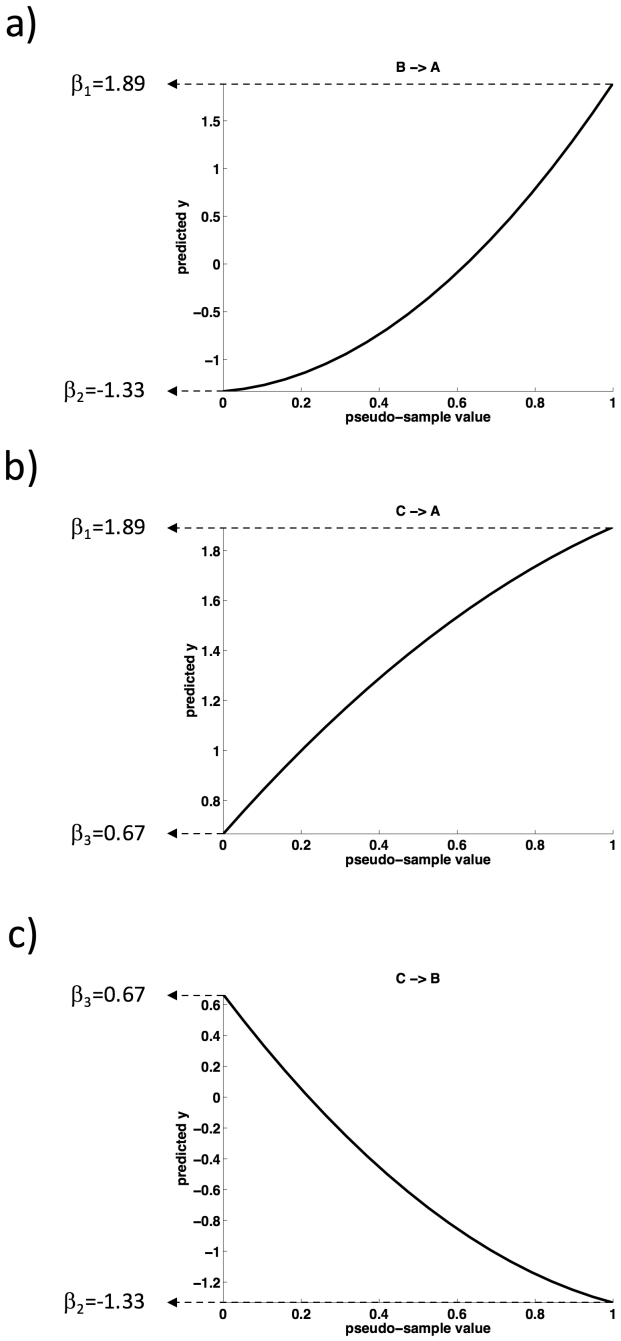

Figure 3 – Data simulated according to a second-order polynomial model (generation scheme in Equations 15 and 16): pseudo-sample trajectories representing the evolution of the predicted response (estimated by means of a cross-validated 3-latent variable second-order polynomial K-PLS model) while moving from a) the pure mixture composed by the only B constituent to the pure mixture composed by the only A constituent, b) the pure mixture composed by the only C constituent to the pure mixture composed by the only A constituent, and c) the pure mixture composed by the only C constituent to the pure mixture composed by the only B constituent

4.2. Tablet data

Second-order Scheffé, Cox and polynomial K-PLS models were fitted for the analysis of the tablet dataset[ii]. The number of extracted PLS and K-PLS latent variables was selected by leave-one-out cross-validation. As Tables 2 points out, the three modelling strategies as well as RBF K-PLS returned comparable and satisfactory performance indices. Figure 4 displays their corresponding response surface plots (almost identical - including the one for RBF K-PLS, not shown). They enabled a similar interpretation of the effects of the single constituents on the active ingredient release time. High contents of phosphate, moderate contents of cellulose and low contents of lactose clearly led to high values of such property of interest. More concretely, binary mixtures composed by roughly $\frac{2}{3}$ of phosphate and $\frac{1}{3}$ of cellulose are expected to exhibit the longest release time. Short release times are instead yielded by blends consisting of e.g. $\frac{2}{3}$ of lactose and $\frac{1}{3}$ of cellulose. Thus, it is quite reasonable to assume the presence of a positive contribution for the interaction phosphate/cellulose and a negative contribution for the interaction lactose/cellulose. As illustrated in Section 4.1, one can look at the pseudo-sample trajectories spanning the sides of the triangle in Figure 4c for an accurate determination of the Scheffé model first-order and interaction parameters (see Table SM.1).

Table 2 – Tablet data: R^2, Q^2 and RMSECV values resulting from second-order Scheffé model fitting by means of OLS, second-order Cox model fitting by means of PLS, second-order polynomial K-PLS, and RBF K-PLS

	# LV	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.98	0.84	38.86
Second-order Cox model (PLS)	5	0.99	0.83	39.69
Second-order polynomial K-PLS model	5	0.98	0.83	38.86
Radial-Basis-Function (RBF) K-PLS model (σ = 1.5)	5	0.99	0.85	36.26

4.3. Bubbles data

As for the previous example, the second-order Scheffé, Cox and polynomial K-PLS models and the RBF K-PLS model adjusted for the bubbles dataset rendered acceptable and equivalent R^2, Q^2 and RMSECV values (see Table 3)[ii]. Since this particular mixture problem embraces up to four constituents, the proportion of one of them has to be fixed to allow the response surfaces

Table 3 – Bubbles data: R^2, Q^2 and RMSECV values resulting from second-order Scheffé model fitting by means of OLS, second-order Cox model fitting by means of PLS, second-order polynomial K-PLS, and RBF K-PLS

	# LV	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.94	0.81	0.042
Second-order Cox model (PLS)	9	0.94	0.81	0.042
Second-order polynomial K-PLS model	9	0.94	0.81	0.042
Radial-Basis-Function (RBF) K-PLS model (σ = 1.5)	9	0.94	0.81	0.042

[ii]The use of second-order models was originally suggested in [3]

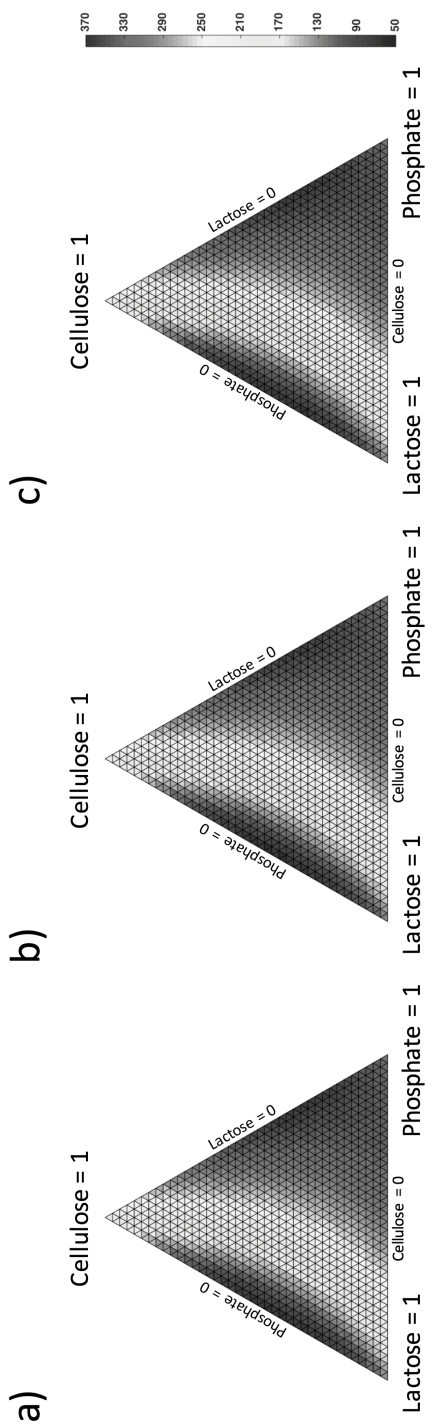

Figure 4 – Tablet data: response surface plots resulting from a) second-order Scheffé model fitting by means of OLS, b) second-order Cox model fitting by means of PLS, and c) the combination of second-order polynomial K-PLS and pseudo-sample trajectories

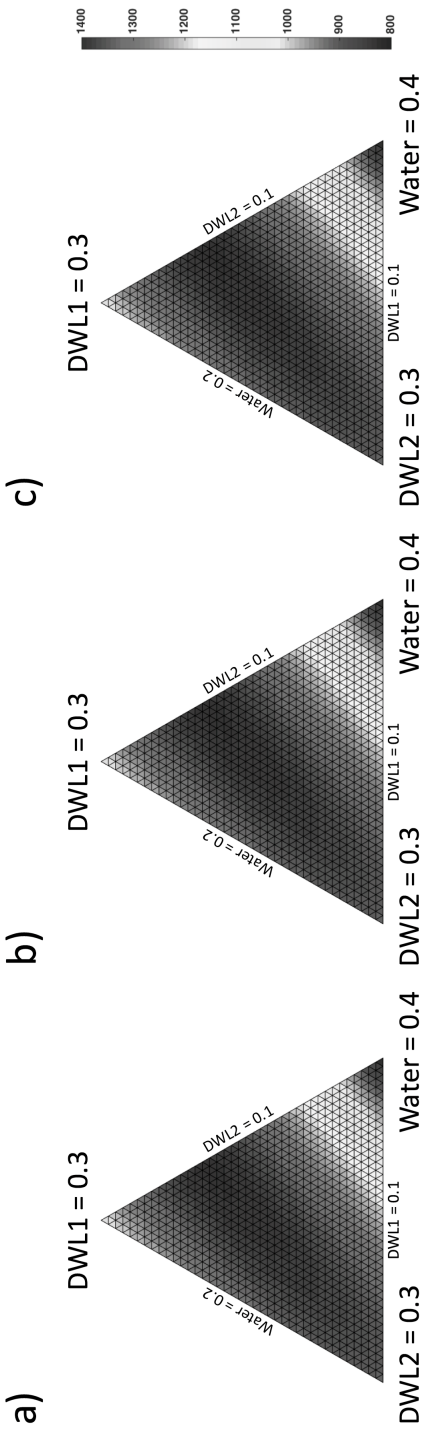

Figure 5 – Bubbles data: response surface plots resulting from a) second-order Scheffé model fitting by means of OLS, b) second-order Cox model fitting by means of PLS, and c) the combination of second-order polynomial K-PLS and pseudo-sample trajectories

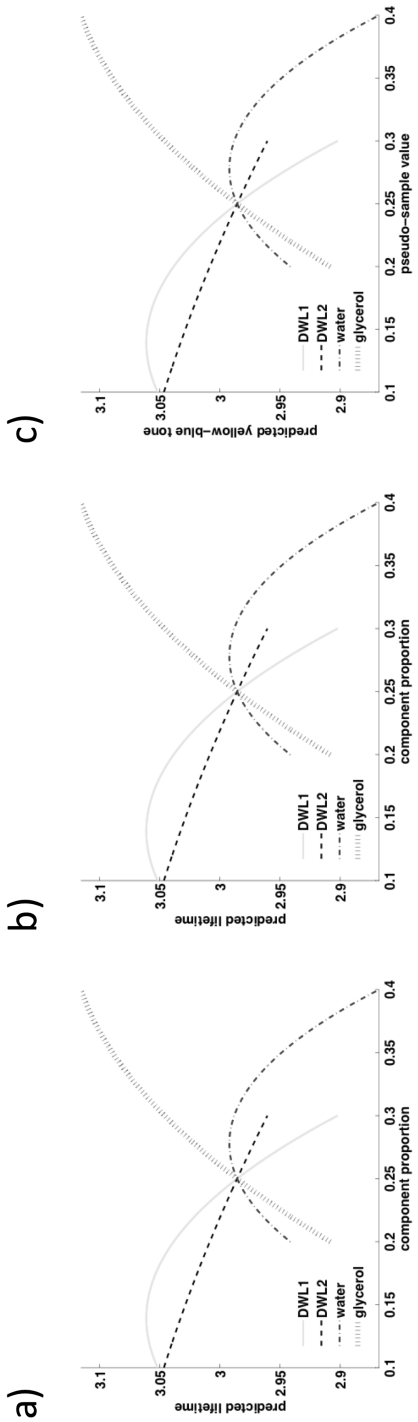

Figure 6 – Bubbles data: trace plots representing the evolution of the predicted lifetime while varying the proportion of the 4 ingredients of the blend (DWL1, DWL2, water and glycerol) and resulting from a) second-order Scheffé model fitting by means of OLS, b) second-order Cox model fitting by means of PLS, and c) the combination of second-order polynomial K-PLS and pseudo-sample trajectories. Here, $s_1 = s_2 = s_3 = s_4 = \frac{1}{4}$

to be graphed as in Section 4.2. Given that glycerol presented a much more positive effect on the bubble lifetime and a much higher cost than any other ingredient, as also suggested in [3], its relative amount was set at 0.4. The results (virtually indistinguishable - including the one for RBF K-PLS, not displayed) are represented in Figure 5. Figure 6 shows instead the corresponding trace plots (not displayed for RBF K-PLS). As one can easily see, although the effect of DWL2 on the response of interest seems to be more positive than that of DWL1 and water (see Table SM.2), the interaction of these latter is crucial for guaranteeing high bubble lifetimes, i.e. more equilibrated blends of DWL1, DWL2 and water would feature more durable bubbles.
The pseudo-sample trajectories spanning the sides of the triangle in Figure 5c cannot be directly resorted to for the estimation of the related Scheffé model coefficients in this situation owing to the fact that the design space of the bubbles data is just a portion of a whole tetrahedron, and then they do not reflect the evolution of the predicted response while moving from a pure mixture to another. On the other hand, if these trajectories are constructed so that they exactly overlap the entire edges of this hypothetical tetrahedron, the methodology proposed in Section 4.1 for the retrieval of the first-order and binary interaction parameters is still valid assuming that any effect involving the two constituents of the concerned binary mixture do not vary outside the actual data space [1].

4.4. Colorant data

When the colorant dataset was dealt with, second-, third- and fourth-order Scheffé, Cox and polynomial K-PLS models and RBF K-PLS models were fitted (separately for every response variable) in order to additionally assess the effect of their complexity on the final outcomes. Table 4 lists their main performance indices. It can be said that different approaches usually required a different complexity to achieve the minimum RMSECV, but, overall, their performance was found to be rather similar also in this case.
For the sake of interpretation, as an illustration, the trace plots resulting from the best Scheffé, Cox and K-PLS models built for the prediction of the yellow-blue tone (b^*) are displayed in Figure 7 (including the one for RBF K-PLS, not shown). They are almost in perfect agreement and only negligible variations with respect to the outcomes obtained by Alman and Pfeifer in [20] were observed (the same goes for those derived for both L^* and a^* - not shown). Concretely, all the constituents exhibited a positive effect on b^*.

4.5. Gasoline data

Second-order cross-validated Scheffé and Cox models were tentatively adjusted for the gasoline dataset. Due to the low number of mixture samples concerned, not enough degrees of freedom were available for the estimation of the coefficients of more complex polynomials. As highlighted by Table 5, both the approaches returned negative Q^2 and poor RMSECV values. On the other hand, their performance was clearly outmatched by that of RBF K-PLS. In addition, RBF K-PLS was found to yield figures of merit just slightly worse than those obtained by coding the constituent proportions in terms of so-called *pseudo-components* [21] and fitting a first-order Scheffé or Cox model including inverse terms, a more standard and common procedure for handling non-linear data coming from a mixture design of experiments [1]. Still, the main advantage of K-PLS over it is that there is no need of performing such a domain transformation and defining these inverse terms prior to the analysis: the optimisation of the kernel transformation function and, in this case, of

Table 4 – Colorant data: R^2, Q^2 and RMSECV values for the three concerned properties of interest (L^*, a^* and b^*) resulting from second/third/fourth-order Scheffé polynomial fitting by means of OLS, second/third/fourth-order Cox polynomial fitting by means of PLS, second/third/fourth-order polynomial K-PLS, and RBF K-PLS. For each response variable the best models in terms of RMSECV are highlighted in bold

	L^*				a^*				b^*			
	# LVs	R^2	Q^2	RMSECV	LVs	R^2	Q^2	RMSECV	LVs	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.97	0.92	6.29	**-**	**0.96**	**0.93**	**3.74**	**-**	**0.92**	**0.87**	**4.27**
Second-order Cox model (PLS)	5	0.97	0.92	6.27	5	0.96	0.93	3.71	5	0.92	0.87	4.24
Second-order polynomial K-PLS model	3	0.97	0.95	4.94	6	0.96	0.93	3.72	**6**	**0.92**	**0.86**	**4.27**
Third-order Scheffé model (OLS)	**-**	**0.99**	**0.99**	**1.35**	-	0.98	0.85	5.62	-	0.95	0.50	8.39
Third-order Cox model (PLS)	**12**	**0.99**	**0.99**	**1.33**	8	0.98	0.88	4.93	3	0.84	0.74	6.12
Third-order polynomial K-PLS model	**14**	**0.99**	**0.99**	**1.25**	**9**	**0.98**	**0.94**	**3.23**	10	0.95	0.86	4.41
Fourth-order Scheffé model (OLS)	-	0.99	0.90	7.15	-	0.99	0.76	7.01	-	0.99	0.21	10.56
Fourth-order Cox model (PLS)	12	0.99	0.96	4.61	**12**	**0.99**	**0.95**	**3.35**	**13**	**0.99**	**0.92**	**3.45**
Fourth-order polynomial K-PLS model	6	0.98	0.97	3.54	9	0.98	0.93	3.77	10	0.95	0.82	5.01
Radial-Basis-Function (RBF) K-PLS model ($\sigma_{L^*} = 25$, $\sigma_{a^*} = 50$, $\sigma_{b^*} = 100$)	**14**	**0.99**	**0.99**	**1.30**	**10**	**0.98**	**0.94**	**3.38**	**11**	**0.95**	**0.89**	**3.97**

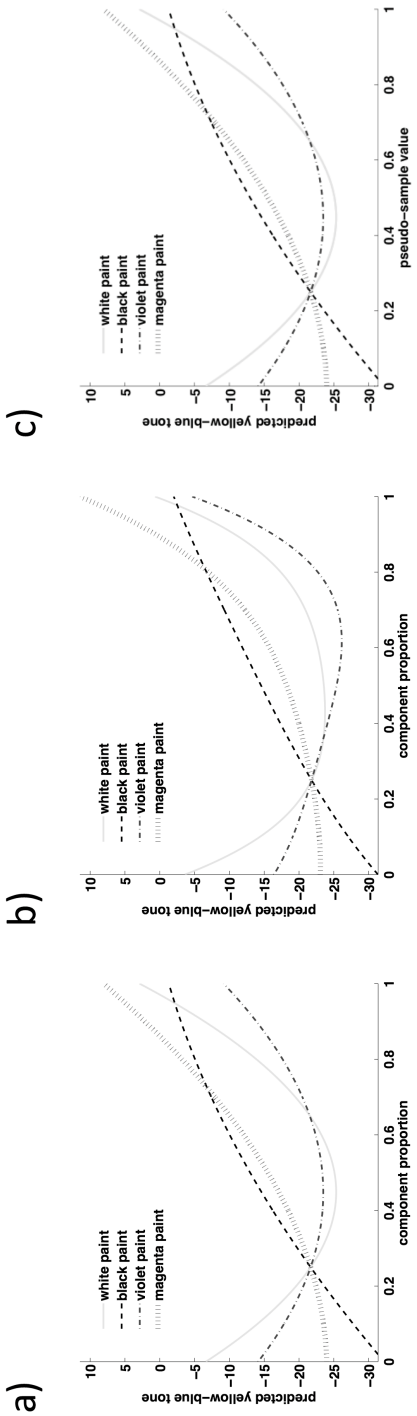

Figure 7 – Colorant data: trace plots representing the evolution of the predicted yellow-blue tone (b^*) while varying the proportion of the 4 ingredients of the blend (white, black, violet and magenta paints) and resulting from a) second-order Scheffé model fitting by means of OLS, b) fourth-order Cox model fitting by means of PLS, and c) the combination of second-order polynomial K-PLS and pseudo-sample trajectories. Here, $s_1 = s_2 = s_3 = s_4 = \frac{1}{4}$

the σ parameter guarantees a certain flexibility when modelling different types of non-linearities without requiring any further operation to be carried out. The first-order pseudo-component

Table 5 – Gasoline data: R^2, Q^2 and RMSECV values resulting from second-order Scheffé model fitting by means of OLS, second-order Cox model fitting by means of PLS, first-order (plus inverse terms) pseudo-component Scheffé model fitting by means of OLS, first-order (plus inverse terms) pseudo-component Cox model fitting by means of PLS, and RBF K-PLS. If required and feasible (i.e. a sufficient number of degrees of freedom was available), non-linearity degree (tuned through the value of the σ parameter in RBF K-PLS) and complexity (number of latent variables) were optimised within a leave-one-out cross-validation loop

	# LV	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.85	-1.42	12.40
Second-order Cox model (PLS)	2	0.99	-0.91	11.02
First-order (plus inverse terms) pseudo-component Scheffé model (OLS)	-	0.98	0.88	2.73
First-order (plus inverse terms) pseudo-component Cox model (PLS)	5	0.99	0.88	2.73
Radial-Basis-Function (RBF) K-PLS model K-PLS model (σ = 0.2)	8	0.99	0.82	3.46

Scheffé and Cox models encompassing inverse terms led to identical surface plots (see Figures 8a and 8b). Certain dissimilarities from the one rendered by RBF K-PLS (see Figure 8c) are instead observable, which was expected considering the intrinsic differences among the compared algorithmic methodologies and especially the fact that the first-order pseudo-component Scheffé and Cox models encompassing inverse terms are able to explain strong non-linearities mainly at the borders of the design space but not in its central area. Nevertheless, a common explanation of how the distinct ingredients affect the values of the response variable can be given: the ideal (maximum) octane rating can be achieved by blending a relatively high quantity of catalytically cracked and relatively low quantities of C_5-isomer and reformate.

4.6. Data simulated according to a highly non-linear model

This section will be focused on further emphasizing the added value of K-PLS with respect to the other methodologies concerned. The outcomes yielded by the application of Scheffé model fitting by means of OLS, Cox model fitting by means of PLS, and K-PLS to the second simulated dataset are reported in Table 6:
The R^2, Q^2 and RMSECV displayed values corroborate what stated before about K-PLS: when strong non-linear relationships (e.g. fourth-order, logarithmic, *etc.*) characterise the data under study, it may outperform in terms of fit and prediction quality both Scheffé model fitting by means of OLS and Cox model fitting by means of PLS. This applies even if first-order Scheffé or Cox models including inverse terms are fitted. Notice that here a Radial-Basis-Function (RBF) kernel transformation and not a polynomial one was found to guarantee the best Q^2 and RMSECV. RBF K-PLS requires the optimisation of an additional parameter, σ. The variation of such a parameter allows different types of complex trends to be modelled, thus its utilisation might be highly recommended when combinations of unknown non-linearities influence the nature of the interdependence between ingredient proportions and properties of interest.

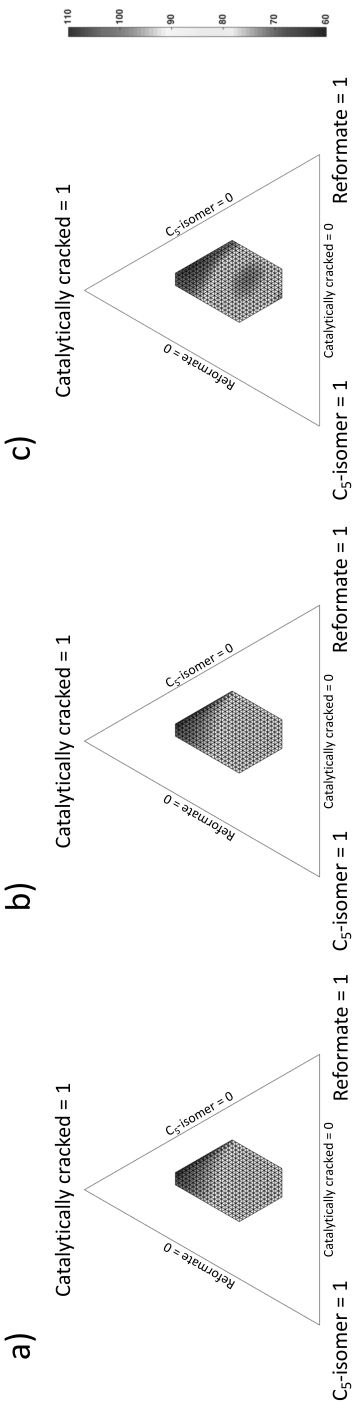

Figure 8 – Gasoline data: response surface plots resulting from a) first-order (plus inverse terms) pseudo-component Scheffé model fitting by means of OLS, b) first-order (plus inverse terms) pseudo-component Cox model fitting by means of PLS, and c) the combination of RBF K-PLS and pseudo-sample trajectories. As here the original mixture design space is irregular, the graphs are represented as parts of the corresponding simplex with vertices [1, 0, 0], [0, 1, 0] and [0, 0, 1] for the sake of an easier visualisation

Table 6 – Data simulated according to a highly non-linear model (generation scheme in Equation 17): R^2, Q^2 and RMSECV values resulting from second-order Scheffé model fitting by means of OLS, second-order Cox model fitting by means of PLS, first-order (plus inverse terms) pseudo-component Scheffé model fitting by means of OLS, first-order (plus inverse terms) pseudo-component Cox model fitting by means of PLS, and RBF K-PLS. If required and feasible (i.e. a sufficient number of degrees of freedom was available), non-linearity degree (tuned through the value of the σ parameter in RBF K-PLS) and complexity (number of latent variables) were optimised within a leave-one-out cross-validation loop

	# LV	R^2	Q^2	RMSECV
Second-order Scheffé model (OLS)	-	0.95	0.56	2.14
Second-order Cox model (PLS)	5	0.99	0.56	2.14
First-order (plus inverse terms) Scheffé model (OLS)	-	0.69	-1.11	4.71
First-order (plus inverse terms) Cox model (PLS)	1	0.96	-0.06	3.35
Radial-Basis-Function (RBF) K-PLS model (σ = 0.8)	7	0.98	0.61	1.83

5. Conclusions

In this article, a novel approach for the analysis of data proceeding from mixture designs of experiments and based on the combination of K-PLS and pseudo-sample trajectories was proposed. Two interesting points arose from the discussed examples:

- if the considered mixture data were not affected by severe non-linearities and/or featured a sufficiently high number of observations, K-PLS and pseudo-sample trajectories yielded very similar results to classical Scheffé model fitting by means of OLS and Cox model fitting by means of PLS (see Sections 4.2, 4.3 and 4.4). Furthermore, a way of recovering the parameters of a Scheffé model (provided that it holds and has the same complexity as the K-PLS one) from the trend of the aforementioned pseudo-sample trajectories was derived and validated via a simulated case-study (see Section 4.1);

- on the contrary, when more non-linear and relatively small data structures had to be analysed, K-PLS proved to be a valid alternative for overcoming the main limitation of both Scheffé model fitting by means of OLS and Cox model fitting by means of PLS (see Sections 4.5 and 4.6): it resulted, in fact, in better fit and prediction quality when the nature of such non-linear data was not strictly polynomial. In addition, although the performance of these more classical methodologies can be improved by taking into account inverse terms, often not enough degrees of freedom are available for a stable estimation of the coefficients of these *augmented* models. K-PLS does not suffer from the same drawback. On top of that, RBF K-PLS through the optimisation of its parameter, σ, may allow different types of complex non-linear relationships to be modelled. Its use might then be highly recommended when combinations of unknown non-linearities influence the nature of the interdependence between constituent proportions and response variables.

Finally, it was also shown how graphs like the surface plots or the trace plots associated to the mixture design space can be retrieved by the pseudo-sample trajectories enabling a reliable interpretation of the influence of changing the proportion of the different ingredients of the blend on its properties of interest.

6. Supporting material

The supporting material associated to this paper includes also two tables containing the Scheffé model coefficients estimated by Scheffé polynomial fitting by means of OLS, Cox polynomial fitting by means of PLS and K-PLS, respectively, for the tablet and bubbles examples. As in the other two illustration cases the best outcomes were returned by Scheffé, Cox and K-PLS models with diverse complexity, such a comparison is skipped for them.

7. Acknowledgements

This research work was partially supported by the Spanish Ministry of Economy and Competitiveness under the project DPI2014-55276-C5-1R and Shell Global Solutions International B.V. (Amsterdam, The Netherlands).

8. References

[1] J. Cornell, Experiments with Mixtures - Designs, Models and the Analysis of Mixture Data, 3rd Edition, John Wiley & Sons, Inc., New York, USA, 2002.
[2] N. Kettaneh-Wold, Analysis of mixture data with Partial Least Squares, Chemometr. Intell. Lab. 14 (1992) 57–69.
[3] L. Eriksson, E. Johansson, C. Wikström, Mixture design - Design generation, PLS analysis, and model usage, Chemometr. Intell. Lab. 43 (1998) 1–24.
[4] H. Martens, T. Næs, Multivariate Calibration, 1st Edition, John Wiley & Sons Ltd., 1989.
[5] D. Cao, Y. Liang, Q. Xu, Q. Hu, L. Zhang, G. Fu, Exploring nonlinear relationships in chemical data using kernel-based methods, Chemometr. Intell. Lab. 107 (2011) 106–115.
[6] B. Schölkopf, A. Smola, Learning with Kernels, 1st Edition, MIT Press, Cambridge, USA, 2002.
[7] P. Williams, Influence of water on prediction of composition and quality factors: the aquaphotomics of low moisture agricultural materials, J. Near Infrared Spectroscop. 17 (2009) 315–328.
[8] M. Embrechts, S. Ekins, Classification of metabolites with Kernel-Partial Least Squares (K-PLS), Drug Metab. Dispos. 35 (2007) 325–327.
[9] V. Struc, N. Pavesic, Gabor-based Kernel Partial-Least-Squares discrimination features for face recognition, Informatica 20 (2009) 115–138.
[10] J. Lee, C. Yoo, S. Choi, P. Vanrolleghem, I. Lee, Nonlinear process monitoring using Kernel Principal Component Analysis, Chem Eng. Sci. 59 (2004) 223–234.
[11] R. Vitale, J. Prats-Montalbán, F. López-García, J. Blasco, A. Ferrer, Segmentation techniques in image analysis: a comparative study, J. Chemometr. 30 (2016) 749–758.
[12] J. Gower, S. Hardings, Nonlinear biplots, Biometrika 75 (1988) 445–455.
[13] P. Krooshof, B. Üstün, G. Postma, L. Buydens, Visualisation and recovery of the (bio)chemical interesting variables in data analysis with support vector machine classification, Anal. Chem. 82 (2010) 7000–7007.
[14] G. Postma, P. Krooshof, L. Buydens, Opening the kernel of Kernel Partial Least Squares and Support Vector Machines, Anal. Chim. Acta 705 (2011) 123–134.
[15] A. Smolinska, L. Blanchet, L. Coulier, K. Ampt, T. Luider, R. Hintzen, S. Wijmega, L. Buydens, Interpretation and visualization of nonlinear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis, PLoS One 7 (2012) e38163.
[16] J. Engel, G. Postma, I. van Peufflik, L. Blanchet, L. Buydens, Pseudo-sample trajectories for variable interaction detection in Dissimilarity Partial Least Squares, Chemometr. Intell. Lab. 146 (2015) 89–101.
[17] R. Vitale, O. de Noord, A. Ferrer, A kernel-based approach for fault diagnosis in batch processes, J. Chemometr. 28 (2014) 697–707.
[18] R. Vitale, O. de Noord, A. Ferrer, Pseudo-sample based contribution plots: innovative tools for fault diagnosis in kernel-based batch process monitoring, Chemometr. Intell. Lab. 149B (2015) 40–52.

[19] H. Wold, Estimation of principal components and related models by iterative least squares, in: P. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, USA, 1966, pp. 391–420.
[20] D. Alman, C. Pfeifer, Empirical colorant mixture models, Color Res. Appl. 12 (1987) 210–222.
[21] J. Cornell, How to Run Mixture Experiments for Product Quality, 1st Edition, American Society for Quality Control - Statistics Division, Milwaukee, USA, 1990.
[22] L. Eriksson, T. Byrne, E. Johansson, J. Trygg, C. Vikström, Multi- and Megavariate Data Analysis - Basic Principles and Applications, 3rd Edition, MKS Umetrics AB, Malmö, Sweden, 2013.