A note on mixture experiments including process variables

Hugo Maruri-Aguilar[2] and Eva Riccomagno[1,2]

[1] Department of Mathematics, Polytechnic of Turin, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy Eva.Riccomagno@polito.it
[2] Department of Statistics, The University of Warwick, Gibbet Hill Road, CV4 7AL, UK H.Maruri-Aguilar@warwick.ac.uk

1 Models for mixture experiments with process variables

When a mixture experiment cannot be performed under uniform conditions or when the responses depend on factors other than the mixture components, like the total mixture amount or some process variables, then the cross product of a standard mixture design and a full factorial design in the non mixture factors is often recommended (see e.g. [3, Ch. 7] and [13]). But product designs might have a larger number of runs than desired and only a subset of the design is implemented.

Let $x = (x_1, \ldots, x_q) \in \mathbb{R}^q$ be the mixture components and $z = (z_1, \ldots, z_k) \in \mathbb{R}^k$ the process variables. The x_i, $i = 1, \ldots, q$ are to be interpreted as proportions, maybe scaled, of a total amount which might be one of the z_j, which often indicated by m. We assume that a design $\mathcal{D}$ is a finite set of points in $\mathbb{R}^{q+k}$, usually the mixture components listed first, and that there are no replicated runs. The projection of $\mathcal{D}$ over the x-space is $\mathcal{D}_x$ and D_z is the projection over the process variable space. Both D_x and D_z admit replicates. For small values of q and k the full product design, $D_x \times D_z$, is recommended in the literatures, where D_x is a simplex lattice design and D_z is a full factorial design [3, Ch. 7].

In the literature various models to study the combined effect of the x and z factors are proposed. Often they have a fairly regular structure derived from a standard cubic or quadratic model for factorial designs (g below) and Scheffé quadratic or cubic polynomial model, in a relevant parametrization, in the mixture components (f). In [5] for a pure mixture designs Kronecker type models are recommended of degree two or three. Thus typically proposed models are additive regression type models like $y(x, z) = f(x) + g(z)$, complete cross product models of the type $y(x, z) = f(x)g(z)$, or in between models as $y(x, z) = f(x) + g(z) + \sum_{i=1}^{q} \sum_{j=1}^{k} f_{ij}(x_i, z_j)$ (see e.g.[3, §7.10]), where the f_{ij} comprises products of terms in f and g. For a mixture amount experiment in

[3, §7.5 and page 405] a mixture amount model of the form $f_0(x) + mf_1(x) + \ldots + m^p f_p(x)$ is suggested where

$$f_p(x) = \sum_i \gamma_i^{(p)} x_i + \sum_{i<j} \gamma_{ij}^{(p)} x_i x_j + \ldots + \sum_{i_1<\ldots<i_l} \gamma_{1,\ldots,l}^{(p)} x_{i_1} \ldots x_{i_l},$$

$l \leq q$, p is a positive integer and the $\gamma^{(p)}$ are regression parameters. A component amount model has a smaller number of terms and takes the form $f(a_1, \ldots, a_q)$ for $a_i = x_i m$, $i = 1, \ldots, q$ and a suitable polynomial function f.

2 Homogeneous representation of a mixture experiment

In algebraic statistics [12] an indeterminate x_i is associated to the i-th factor in the experiment and the design, $\mathcal{D}$, is described and defined by the set of polynomials in the x_i's vanishing on all the design points. This infinite set of polynomials is called the design ideal, Ideal($\mathcal{D}$). The ideal generated by the polynomials $f_1, \ldots, f_v$ is defined as $\langle f_1, \ldots, f_v \rangle = \{\sum_{i=1}^v s_i f_i : s_i \text{ are polynomials}\}$.

Example 1. A $\{q, m\}$ *simplex lattice design* [16] is the intersection of the simplex in $\mathbb{R}^q$ and the full factorial design in q factors and with the uniformly spaced levels $\{0, 1/m, \ldots, 1\}$. Thus the ideal of the $\{q, m\}$ simplex lattice design is

$$\langle \prod_{j=0}^{m} (x_1 - j/m), \ldots, \prod_{j=0}^{m} (x_q - j/m), \sum_{i=1}^{q} x_i - 1 \rangle$$

that is the first q polynomials give the full factorial design and the last one is the simplex condition which selects the points of the full factorial whose components sum to one.

In [9] it is shown that for a pure mixture design, $\mathcal{D} = \mathcal{D}_x$, an alternative polynomial representation is meaningful and useful. The design is identified with the set of lines through the origin of the x-space and a point in $\mathcal{D}$, to indicate that in a pure mixture experiment the relative proportions of the component are of interest irrelevant of the total mixture amount. The set of all such lines is called the *design cone* and indicated as $\mathcal{C}_\mathcal{D}$. The set of all polynomials vanishing on all points of $\mathcal{C}_\mathcal{D}$ is Ideal($\mathcal{C}_\mathcal{D}$).

Example 2. The simplex centroid design, $\mathcal{D}$ in $\mathbb{R}^q$ [17] is the projection, on the simplex in $\mathbb{R}^q$ with respect to the origin, of the full factorial design with levels 0 and 1. This shows that it has $2^q - 1$ points, the coordinates of each point are either zero or equal to each other and moreover it holds that

$$\text{Ideal}(\mathcal{C}_\mathcal{D}) = \langle x_i^2 x_j - x_i x_j^2 : i, j = 1, \ldots, q; i \neq j \rangle$$

In particular the design point with projective coordinates $(1 : \ldots : 1) \in \mathbb{R}^q$ is the barycenter point of the simplex in $\mathbb{R}^q$. See [9, §4.2].

In mathematical terms, a mixture experiment is thus to be interpreted as a projective variety. For the consequence of this interpretation we refer to [9]. Note that Ideal($\mathcal{C}_\mathcal{D}$) is a homogeneous ideal, that is an ideal generated by homogeneous polynomials. A polynomial is homogeneous of degree $s \in \mathbb{Z}_{\geq 0}$ if each one of its terms has sum of exponents equal to s. By convention $a \in \mathbb{R}$ has degree zero. Indeed if f is a homogeneous polynomials of degree s and $f(d) = 0$ for all $d \in \mathcal{D}_x$, then for $\lambda \in \mathbb{R}$ $f(\lambda d) = \lambda^s f(d)$. Viceversa see Theorem 1 in [9].

2.1 Homogeneous models for pure mixture experiments

Cone design ideals lead naturally to consider homogeneous polynomial regression models. We need to recall the basics of algebraic statistics for design of experiments [12]. Let $\mathbb{R}[x_1, \ldots, x_d]$ be the set of all polynomials in $x_1, \ldots, x_d$ with real coefficients. The set of real valued functions over $\mathcal{D}$ is isomorphic to the quotient space $\mathbb{R}[x_1, \ldots, x_d]/\operatorname{Ideal}(\mathcal{D})$ defined by the equivalence relationship stating that two polynomials f and g are equivalent if they take the same values over all the points of $\mathcal{D}$. The quotient space is a $\mathbb{R}$-vector space, of dimension equal to the number of points in $\mathcal{D}$ and it admits vector space bases formed by monomials.

In Lemma 3 and Theorem 4 of [9] it is proved that if $\mathcal{D}$ is a mixture design then there are bases formed by monomials of the same total degree larger than a suitable integer and an algorithm to compute them is provided. Any such basis can be used to construct homogeneous polynomial regression models of the Kronecker type [5] and submodels.

Example 3. For the design in Example 2 the largest set of degree three, linearly independent monomials in the quotient space is

$$x_i^3, \text{ for all } i = 1, \ldots, q \qquad x_i^2 x_j, \text{ for all } i < j, i, j = 1, \ldots, q$$
$$x_i x_j x_k, \text{ for all } i < j < k, i, j, k = 1, \ldots, q$$

A full basis of a given degree, equivalently a saturated homogeneous model identified by $\mathcal{D}$, can be retrieved only for a degree larger than three.

3 A model selection algorithm

Consider a product design $\mathcal{D} = \mathcal{D}_x \times \mathcal{D}_z$ with no replicated runs. Let $\mathrm{E}_x = \{x^\alpha : \alpha \in L_x\} \in \mathbb{R}[x_1, \ldots, x_q]$ be a set of linearly independent monomials in $\mathbb{R}[x_1, \ldots, x_q]/\operatorname{Ideal}(\mathcal{D}_x)$ and $\mathrm{E}_z = \{z^\alpha : \alpha \in L_z\} \in \mathbb{R}[z_1, \ldots, z_k]$ a set of linearly independent monomials in $\mathbb{R}[z_1, \ldots, z_k]/\operatorname{Ideal}(\mathcal{D}_z)$. Let $\mathrm{E}_x \otimes \mathrm{E}_z$ be the Kronecker product of E_x and E_z. Then by the basic property of Kronecker and tensor products, $\mathrm{E}_x \otimes \mathrm{E}_z$ is a set of linearly independent monomials in $\mathbb{R}[x, z]/\operatorname{Ideal}(\mathcal{D})$. Moreover if also $\mathcal{D}_z$ and $\mathcal{D}_x$ have no replicated points, then it is a $\mathbb{R}$-vector space basis and it has dimension $n_x n_z$ where n_i is the number

of points in $\mathcal{D}_i$, $i = z, x$. Then typically E_x is a set of monomials of the same degree, two or three, and E_z is an order ideal. In [12, §3.5] and [9, §3] algorithms are provided to compute E_x and E_z for generic $\mathcal{D}_x$ and $\mathcal{D}_z$.

In practice and when q and k are not small values, it has to be expected that $\mathrm{E}_x \otimes \mathrm{E}_z$ is large and a subset has to be considered to construct response surface models for the problem at hand, and also that not all runs in $\mathcal{D}_x \times \mathcal{D}_z$ can be implemented. We suggest an algorithm to select a subset L of $\mathrm{E}_x \otimes \mathrm{E}_z$ to be used as support for a model identifiable by a given fraction $\mathcal{F} \subset \mathcal{D}$. The subset L is selected according to "statistical" criterion.

The design/model matrix for $\mathcal{D}$ and a model supported over a subset of $\mathrm{E}_x \otimes \mathrm{E}_z$ is full rank, independently of the selected representatives of the homogeneous points in $\mathcal{C}_{\mathcal{D}_x}$. (It is an immediate corollary of Lemma 3 in [9]). That is, identifiability does not depend on the homogeneous coordinates. However, other properties of the design/model matrix are strongly effected by the representatives used, for example the eigenvalues of the corresponding information matrix.

We choose to minimise the condition number λ of the information matrix. It is defined as $\lambda = \lambda_{max}/\lambda_{min}$ where λ_{max} and $\lambda_{min} \geq 0$ are the maximum and minimum eigenvalues of the information matrix, $X_L^t X_L$ where ^{t} indicates transpose. If $X_L^t X_L$ is close to singular then its columns are almost linearly dependent and this is signaled by a minimum eigenvalue close to zero. Thus small condition number indicates more stability in the least square estimates and smaller variance inflation factor than for larger condition numbers.

In [5] Kronecker type models are studied for pure mixture experiments and in [14] quadratic Kronecker models are conjectured to be the most robust to miss-specification of the information matrix among second order models for experiments with mixtures. In particular in Corollary 1 the authors in [14] show that any model in a K-chain has higher maximum eigenvalue of the information matrix than the Kronecker type model. A K-chain is a chain of mixture models all of the same size, a model in the chain differs from the next one by one term and the final model is of Kronecker type. Thus E_x is of Kronecker type.

Other statistical criteria can be considered. A referee suggested to couple the condition number criterion with a criterion related to the goodness of fit of the model. The major change in the algorithm below is in the definition of λ_i which could become a vector or remain a scalar number and should now correspond to the new criterion or combinations of criteria. In Example 6 of Section 3.2 we simply checked that the R^2 values expressing a goodness of fit of the selected sub model were not worse than that presented in the literature.

3.1 Selection based on term orderings

As mentioned, the number of columns in the design model matrix, X, for $\mathrm{E}_x \otimes \mathrm{E}_z$ and $\mathcal{D}$ or $\mathcal{F}$ might be prohibitive to a full search. Nevertheless term orderings can be used to guide this search. Thus the search might start with

a Kronecker type model E_x (likely to have a low condition number). Terms are substituted one at a time according to a term ordering which favours the x-indeterminates. We suggest and sketch a variation of the algorithms in [1, 9] to scan the class of models obtaines with term orderings. This class of models is typically smaller than the class of submatrices of X which are full rank and of size n. The search we suggest is based on the algorithms for exchange of basis introduced in [6], developed in [1], and described for designs in [12, §3.5] and [9, §3]. In [2] an algorithm is given to list all saturated models which are order ideal and have the same support size.

The idea is to order the finite set of monomials $\mathrm{E}_x \otimes \mathrm{E}_z$, equivalently the columns of X they label, in all possible ways that can be extended to a full term ordering. We do so by using vectors of weights, i.e. $w \in \mathbb{Z}_{\geq 0}^{q+k}$. It turns out that a finite set of weighing vectors is sufficient to describe all such possible ways. The set of all weighing vectors, W, depends only on the exponents of monomials in the candidate set $\mathrm{E}_x \otimes \mathrm{E}_z$. Thus W can be computed once for each set $\mathrm{E}_x \otimes \mathrm{E}_z$ independently of the design and become part of a library. The computation of W is straightforward for models in two dimensions, but for models in higher dimensions there is still need for efficient algorithms. This is largely investigated in the first author's Ph.D. thesis. See also Example 6 below.

Example 4. There are only two ways of ordering the three monomials $x_1^2, x_1x_2, x_2^2 \in \mathbb{R}[x_1, x_2]$. They are $x_1^2 < x_1x_2 < x_2^2$ and $x_2^2 < x_1x_2 < x_1^2$. The first one corresponds to the weighing vector $w = (1, 2)$, indeed $((1, 2) \cdot (2, 0)) = 2 < ((1, 2) \cdot (1, 1)) = 3 < ((1, 2) \cdot (0, 2)) = 4$. Many other weighing vectors can be equivalently considered. See [8].

Call X_w the matrix X whose columns are reordered according to w. Then the first n linearly independent columns of X_w can be used as support for regression models. We select the model with n terms and smallest condition number by varying $w \in W$. The algorithm can be outlined as follows.

Input: $\mathcal{D}_x$ and $\mathcal{D}_z$, a fraction $\mathcal{F} \subseteq \mathcal{D}_x \times \mathcal{D}_z$ and the number of terms in the final submodel n and the sets of monomials E_x and E_z, which are determined following the guidelines at the beginning of Section 3. The final submodel size n cannot be greater than the number of points in $\mathcal{F}$ to ensure identifiability.

Output: a submodel L_0 with a minimal condition number λ_0. The final submodel is formed by the smallest terms of $\mathrm{E}_x \otimes \mathrm{E}_z$ with respect to a term ordering.

Technique: the search space of candidate submodels is generated by ordering $\mathrm{E}_x \otimes \mathrm{E}_z$ with different weight vectors, and within this search space we look for the submodel with the smallest condition number.

Step 1: Compute the design-model matrix X using the points in $\mathcal{F}$ and the terms in $\mathrm{E}_x \otimes \mathrm{E}_z$. Compute, see also Example 6, the set of weight vectors, $W := \{w_1, \ldots, w_s\}$. Initialize $i := 1$, $\lambda_0 := \infty$ and $L_0 := []$.

Step 2: Order $\mathrm{E}_x \otimes \mathrm{E}_z$ and the corresponding columns of X using the weight vector w_i. Let L be the first n monomials of $\mathrm{E}_x \otimes \mathrm{E}_z$ such that the rank of X_L is n. Let λ_i be the condition number of $X_L^t X_L$.

Step 3: If $\lambda_i < \lambda_0$ then $\lambda_0 \leftarrow \lambda_i$ and $L_0 \leftarrow L$.

Step 4: Update $i \leftarrow i+1$. If $i \leq s$ then repeat from Step 3.1, otherwise L_0 is the set of terms of the wanted model.

The algorithm clearly ends as W is a finite set [1]. Moreover, any weight vector identifies a model of size n and thus the algorithm gives an answer. The algorithm is of order $O((n_x n_z)^{2(qk-1)} n^2)$ and, as the dimensions q and k are fixed, the algorithm is of polynomial order in $(n_x n_z)$ (see [1]. This argument is detailed in [8, Chapter 2]). The search space is certainly more restricted than the full combinatorial search of exponential order $\binom{n_x n_z}{n} = O((n_x n_z)^{n_x n_z})$. However the final model respects a hierarchical structure, unlike many of the models in the combinatorial search and the search is clearly much faster.

3.2 Examples

Example 5. A mixture-amount design $\mathcal{D}$ is given in the left-hand side of Table 1 in affine coordinates. Here x_1 and x_2 are proportions of a total amount m. The ideals of interest are $\mathrm{Ideal}(\mathcal{D}) = \langle x_1 + x_2 - m, (m-1)(m-2), (x_2 - 1)(m-2), (x_2-1)(x_2-2)\rangle$, $\mathrm{Ideal}(\mathcal{C}_{\mathcal{D}_x}) = \langle x_1 x_2 (x_1 - x_2)\rangle$ and $\mathrm{Ideal}(\mathcal{D}_m) = \langle (m-1)(m-2)\rangle$, from which we have that $\mathrm{E}_x = \{x_1^2, x_1x_2, x_2^2\}$ and $\mathrm{E}_z = \{1, m\}$. The corresponding X matrix is shown in the right side of Table 1. The algorithm of Section 3 returns $L_0 = \{x_1^2, x_2^2, x_1x_2, mx_2^2\}$ for the weighing vector $w = (1,2,3)$. In this simple case we can additionally perform a full combinatorial search, which returns the same result.

x_1	x_2	m	x_1^2	x_1x_2	x_2^2	mx_1^2	mx_1x_2	mx_2^2
0	1	1	0	0	1	0	0	1
0	2	2	0	0	4	0	0	8
1	1	2	1	1	1	2	2	2
2	0	2	4	0	0	8	0	0

Table 1. Mixture-amount design and matrix X for Example 5.

Example 6. We consider the well-known bread experiment introduced in [10], for which $\mathcal{D}_x$ is a simplex lattice with three factors and 10 runs and $\mathcal{D}_z$ is a factorial 3^2 design. The analysis in [13] returns a final model of $n = 15$ terms and with the condition number 86.83. See [13, Equation (11)].

For the natural sets $E_x = \{x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3\}$ and $E_z = \{1, z_1, z_2, z_1^2, z_1z_2, z_2^2\}$, the set $\mathrm{E}_x \otimes \mathrm{E}_z$ has 54 monomials, and the number of submodels with fifteen terms is $\binom{54}{15} \approx 8.6 \times 10^{12}$. A full search on this

space is impossible and the algorithm of Section 3 can be applied to select a candidate model with small condition number.

Instead of computing the full set of weighing vectors W, which as mentioned can be hard, we mimic it as follows. The $(q+k-1)$-dimensional simplex intersects all "cones of equivalence classes" of the weighing vectors. In this sense every point on the simplex is equivalent to an element of W. We apply our algorithm with a sample of random vectors uniformly distributed over the simplex. If the sample is large enough, there is a high chance to pick at least one w for each equivalence class. This alternative is properly quantified in [8, Chapter 2].

The variables are listed as $(x_1, x_2, x_3, z_1, z_2)$. The algorithm returns the submodel

$$L_0 = (\{x_1, x_2, x_3\} \otimes \{1, z_1, z_2\} \cup \{x_2, x_3\} \otimes \{z_1^2, z_1 z_2, z_2^2\})$$

for $w = (17, 12, 10, 3, 2)$. The model L_0 traded the monomials $x_1 z_1^2, x_1 z_2^2$ in [13, Equation(11)] for $x_2 z_1 z_2$ and $x_3 z_1 z_2$, and this slight asymmetry allows for the reduction of the condition number to 47.47. With respect to the model in [13], there is a slight increase in the root-mean-squared error, while the R^2 is practically the same.

4 Final comments

In this note we considered a design $\mathcal{F} \subseteq \mathcal{D}_x \times \mathcal{D}_z$ and a set of linearly independent monomial functions over the vector space of real functions defined over $\mathcal{D}_x \times \mathcal{D}_z$.

An algorithm to select a model identified by $\mathcal{F}$, with a given number of terms and of minimal condition number is described. It has polynomial complexity in the number of design points and model size. Its search space is smaller than the one of a full search. In the authors' experience, see also [1], not only it is fast (especially when coupled with a search of the W vectors over a grid as in Example 6) but also it performs well returning good models. One possible drawback is that it might exclude models which are symmetric in the factors. This is inherent in the use of term orderings and thus w vectors. Indeed there is no term ordering for which $x^2 < x_1^2 < x_1 x_2$. Symmetric models might be added to the search space or one can use only partially weighing vectors w. Methods to work with monomial bases of the quotient space which are based on term-ordering free computations with multiplication tables are being studied in the algebraic community. See [15] for a first example. Other criteria can be substituted to the minimal condition number criterion and general designs, even with replicated runs, can be considered. We focused on mixture designs with process variables or mixture amount experiments. The final model we obtain is usually not one suggested in the literature, it is different from the model obtained by running the standard algorithm in

[12, §3.5] and when comparable it performs statistically not worse than other models suggested in the literature for the examples we tried.

References

1. Babson E, Onn S, Thomas R (2003). The Hilbert zonotope and a polynomial time algorithm for universal Gröbner bases. Advances in Applied Mathematics 30,3:529–544.
2. Bates, R.A., Giglio, B., Wynn, H.P. (2003). A global selection procedure for polynomial interpolators *Technometrics* **45**, 246-255.
3. Cornell JA (2002). Experiments with mixtures. John Wiley & Sons, New York, Third Edition.
4. Cox D, Little J, O'Shea D (1997). Ideal, Varieties, and Algorithm. Springer-Verlag, New York, Second Edition.
5. Draper NR, Pukelsheim F (1998). Mixture models based on homogeneous polynomials. Journal of Statistical Planning and Inference 71,1-2:303–311.
6. Faugère, JC, Gianni P, Lazard D, Mora, T (1993). Efficient computation of zero-dimensional Gröbner bases by change of ordering. Journal of Symbolic Computation 16,4:329–344.
7. Holliday T, Pistone G, Riccomagno E, Wynn HP (1999). The Application of Algebraic Geometry to Analysis of Designed Experiments: A Case Study. Computational Statistics 14,2:213–231.
8. Maruri-Aguilar H (2006), PhD thesis, Department of Statistics, The University of Warwick (work in progress).
9. Maruri-Aguilar H, Notari R, Riccomagno, E (2006). On the description and identifiability analysis of mixture designs. CRiSM, Department of Statistics, The University of Warwick.
10. Næs T, Færgestad E M, Cornell J (1998). A comparison of methods for analysing data from a three component mixture experiment in the presence of variation created by two process variables. Chemometrics and Intelligent Laboratory Systems 41, 221–235.
11. Pistone G, Riccomagno E, Wynn HP (2000). Gröbner basis methods for structuring and analysing complex industrial experiments. International Journal of Reliability, Quality and Safety Engineering 7,4:285-300.
12. Pistone G, Riccomagno E, Wynn HP (2001). Algebraic Statistics. Chapman & Hall/CRC, Boca Raton.
13. Prescott P (2004). Modelling in mixture experiments including interactions with Process variables. Quality Technology & Qualitative Management 1,1:87–103.
14. Prescott P, Dean AM, Draper N R, Lewis SM (2002). Mixture experiments: ILL-Conditioning and Quadratic Model Specification. Technometrics 44,3:260–268.
15. Rouillier, F (1999). Solving zero-dimensional systems through the rational univariate representation. Applicable Algebra in Engineering, Communication and Computing 9,5:433–461.
16. Scheffé H (1958). Experiments with mixtures. Journal of the Royal Statistical Society. Series B. 20:344–360.

17. Scheffé H (1963). The simplex-centroid design for experiments with mixtures. Journal of the Royal Statistical Society. Series B 25:235–263.