Journal of Statistical Planning and Inference 147 (2014) 204–211

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A characterization of saturated designs for factorial experiments

Roberto Fontana [a], Fabio Rapallo [b,*], Maria Piera Rogantin [c]

[a] Department of Mathematical Sciences, Politecnico di Torino, Italy
[b] Department of Sciences and Technological Innovation, Università del Piemonte Orientale, Italy
[c] Department of Mathematics, Università di Genova, Italy

ARTICLE INFO

Article history:
Received 2 October 2013
Received in revised form
30 October 2013
Accepted 30 October 2013
Available online 8 November 2013

Keywords:
Estimability
Linear models
Circuits
Graver basis
Universal Markov basis

ABSTRACT

In this paper we study saturated fractions of factorial designs under the perspective of Algebraic Statistics. Exploiting the identification of a fraction with a binary contingency table, we define a criterion to check whether a fraction is saturated or not with respect to a given model. The proposed criterion is based on combinatorial algebraic objects, namely the circuit basis of the toric ideal associated to the design matrix of the model.

1. Introduction

The search for minimal designs to estimate linear models is an active research area in design of experiments. Saturated fractions have a minimum number of points to estimate all the parameters of a given model. As a consequence, all information is used to estimate the parameters and there are no degrees of freedom to estimate the error term. Nevertheless, saturated fractions are of common use in sciences and engineering, and they become particularly useful for highly expensive experiments, or when time limitations impose the choice of the minimum possible number of design points. For general reference in design of experiments, the reader can refer to Raktoe et al. (1981) and Bailey (2008), where the issue of saturated fractions is discussed.

In this paper we characterize saturated fractions of a factorial design in terms of the circuits of its model design matrix and we define a criterion to check whether a given fraction is saturated or not without computing the determinant of the corresponding design matrix.

Our work falls within the framework of Algebraic Statistics. The application of polynomial algebra to the design of experiments was originally presented in Pistone et al. (2001), but with a different point of view. The techniques used here are mainly based on the combinatorial and algebraic objects associated to the design matrix of the model, such as the circuit, the Graver and the Universal Gröbner bases. All these bases have been used to solve enumeration problems, to make non-asymptotic inference, and to describe the geometric structure of the statistical models for discrete data, and in particular for the analysis of contingency tables. A recent account of this theory can be found in Drton et al. (2009).

* Corresponding author. Tel.: +39 0131 360 223; fax: +39 0131 360 250.
E-mail addresses: roberto.fontana@polito.it (R. Fontana), fabio.rapallo@unipmn.it (F. Rapallo), rogantin@dima.unige.it (M.P. Rogantin).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.10.011

In this paper, we benefit from the interplay between algebraic techniques for the analysis of contingency tables and some topics of the design of experiments. Indeed, we identify a factorial design with a binary contingency table whose entries are the indicator function of the fraction, i.e., they are equal to 1 for the fraction points and 0 for the other points. This implies that a fraction can also be considered as a subset of cells of the table. Some recent results in this direction can be found in Aoki and Takemura (2010). The connections between experimental designs and contingency tables have also been explored in Fontana et al. (2012), but were limited to the investigation of enumerative problems in the special cases of contingency tables arising from the Sudoku problems. We emphasize that our approach to design experiments differs from the standard approach based on Algebraic Statistics, and introduced in Pistone and Wynn (1996) and Caboara and Robbiano (2001). For instance, a difference lies in the number of indeterminates and their meaning: usually in the design approach an indeterminate corresponds to a factor, while in the contingency table approach an indeterminate corresponds to a point of the full factorial design.

The basic idea underlying the theory is as follows. A fraction is non-saturated with respect to a given model if there exists a null linear combination with non-zero coefficients of the rows of its corresponding design matrix. This vector of coefficients belongs to the kernel of the transpose of the model design matrix. In polynomial algebra, the counterpart of this kernel is the toric ideal associated to the transpose of the design matrix. Using a special basis of this ideal, namely the circuit basis, we define a combinatorial criterion to determine whether a fraction is saturated or not. Some preliminary results in the special case of two factor designs were presented in Kuhnt et al. (2013) and Fontana et al. (2013).

Our approach based on circuits avoids the computation of the determinant and has the advantage that the circuits do not depend on the fraction but they are computed once for all from the design matrix of the model. Moreover, this new criterion has an interesting connection with optimization.

The theory described in this paper suggests several extensions and applications. Firstly, it is interesting to explore how the results can be extended for the characterization of saturated fractions to more general designs, also using graph theory. It would also be interesting to study the classification of the saturated fractions with respect to some statistical criteria, for instance the minimum aberration in a classical sense, or the more recent state polytopes approach, see Berstein et al. (2010).

The paper is organized along these lines. In Section 2 we set some notations and we state the problem. In Section 3 we provide the basic algebraic framework to be used in the paper. In Section 4 we prove the main result, showing that the absence of circuits is a necessary and sufficient condition for obtaining a saturated fraction. In Section 5 we show how saturated designs can be seen as the solutions of properly defined integer linear programming problems, while in Section 6 we provide some examples to demonstrate the practical applicability of our theory, we discuss some computational issues, and we show how to generate a sample of saturated fractions by applying the theory of Markov bases when we add constraints to the projections.

2. Saturated designs

Let $\mathcal{D}$ be a full factorial design with d factors, $A_1, \dots, A_d$ with $s_1, \dots, s_d$ levels ($s_i \in \mathbb{N}, i = 1, \dots, d$). Here the levels are coded by natural numbers, so that $\mathcal{D} = \{0, \dots, s_1 - 1\} \times \cdots \times \{0, \dots, s_d - 1\}$. We consider a linear model on $\mathcal{D}$:

$$Y = X\beta + \varepsilon,$$

where Y is the response variable, X is the *design matrix*, β is the vector of model parameters, and ε is a vector of random variables that represent the error terms. We denote by p the number of estimable parameters.

For instance, in a two-factor design with the main effect model, we have $p = s_1 + s_2 - 1$ and an estimable model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{for } i = 0, \dots, s_1 - 2, \; j = 0, \dots, s_2 - 2,$$

where μ is the expected value of Y in the combination of reference levels (the last one here), $\mu = \mathbb{E}(Y_{s_1-1, s_2-1})$, the coefficient α_i is the difference between the expectation of Y in the level i and the expectation in the reference level, $\alpha_i = \mathbb{E}(Y_{i,\bullet}) - \mathbb{E}(Y_{s_1-1,\bullet})$, and analogously for β_j. The corresponding design matrix is

$$X = (m|a_0|\cdots|a_{s_1-2}|b_0|\cdots|b_{s_2-2}), \tag{1}$$

where m is a column vector of 1's, $a_0, \dots, a_{s_1-2}$ are the indicator vectors of the first $(s_1 - 1)$ levels of the factor A_1, and $b_0, \dots, b_{s_2-2}$ are the indicator vectors of the first $(s_2 - 1)$ levels of the factor A_2.

A subset $\mathcal{F}$, or fraction, of a full design $\mathcal{D}$, with minimal cardinality $\#\mathcal{F} = p$, that allows us to estimate the model parameters, is a saturated fraction or a *saturated design*. The design matrix $X_{\mathcal{F}}$ of a saturated design is a non-singular matrix with dimensions $p \times p$.

Running example 1. Let us consider the 2^4 design and the model with main factors and 2-way interactions. This example is at the same time not trivial and easy to handle, so that we use it as the running example in this paper. The design matrix X of the full design has 16 rows and 11 columns. The points are its row labels, while the model terms are its column labels. As the matrix X has rank 11, we search for fractions with 11 points. For instance the fraction

$$\mathcal{F}_1 = \{(0,0,0,0), (0,0,0,1), (0,0,1,1), (0,1,0,1), (0,1,1,0), (1,0,0,1), (1,0,1,0), (1,1,0,0), (1,1,0,1), (1,1,1,0), (1,1,1,1)\}$$