Combinatorial Analysis of Factorial Designs with Ordered Factors

Combinatorial Analysis of Factorial Designs with Ordered Factors

Analisi Combinatoria di Piani Fattoriali con Fattori Ordinali

Roberto Fontana and Fabio Rapallo

Abstract In recent literature a new combinatorial algorithm for the selection of robust fractional factorial designs has been introduced. In this work we analyze the application of this algorithm in the case of ordered factors.
Abstract *È stato sviluppato recentemente un nuovo algoritmo combinatorio per la selezione di piani fattoriali frazionari robusti. In questo lavoro analizziamo la sua applicazione nel caso di fattori ordinali.*

Key words: Algebraic statistics, Design of experiments, Optimality, Robust fractions

1 D-optimality, robustness, and combinatorial objects

The choice of a design from a set of candidate runs is one of the most relevant problems in Design of Experiments. When working in the framework of factorial designs, the candidate set is usally the full-factorial design containing all the possible level combinations of the factors. There are several criteria for choosing a design. Here we restrict our attention to model based techniques. Thus the linear model on the candidate set $\mathscr{D}$ is written in the form

$$\mathbf{y} = X_{\mathscr{D}}\beta + \varepsilon\,, \tag{1}$$

Roberto Fontana
Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, e-mail: roberto.fontana@polito.it

Fabio Rapallo
Dipartimento di Economia, Università di Genova, via Vivaldi 5, 16126 Genova, e-mail: fabio.rapallo@unige.it

where $X_{\mathscr{D}}$ is the full-design model matrix with dimensions $K \times p$, β is the p-dimensional vector of the parameters, ε is the error, and $\mathbb{E}(\mathbf{y}) = X_{\mathscr{D}}\beta$.

The classical theory leading to the class of alphabetical optimality criteria (D-optimality, A-optimality, etc.) is based on the maximization of some quantities computed using the model matrix $X_{\mathscr{F}}$ of the selected fraction $\mathscr{F} \subset \mathscr{D}$. In their basic form, the selection algorithms work with a pre-defined and fixed dimension $n = \#\mathscr{F}$ of the fraction. As general reference for optimal designs, refer to [6].

When the design may be incomplete, e.g. for time limitations, there are methods to choose the order of the runs in order to achieve first the most informative runs, so that a possibly incomplete design is as much effective as possible for parameter estimation. Fractional Factorial Designs with removed runs are studied in, e.g., [1], [7]. In such a case the set $\mathscr{D}$ is usually a candidate set different from the full-factorial design.

Both optimality with a fixed run size and with possibly incomplete designs has been recently analyzed under a geometric and combinatorial point of view using a special representation of the basis of the kernel $\ker(X_{\mathscr{D}}^t)$ of the model matrix for the candidate set, namely the circuit basis. In particular, the property that naturally reflects the geometry of the design points is the robustness, first introduced in [4].

Definition 1. The robustness of a fraction $\mathscr{F}$ with design matrix $X_{\mathscr{F}}$ is defined as

$$r(X_{\mathscr{F}}) = \frac{\#\left\{ \text{ saturated } \mathscr{F}_p \right\}}{\#\left\{ \mathscr{F}_p \right\}} = \frac{\#\left\{ \text{ saturated } \mathscr{F}_p \right\}}{\binom{n}{p}}$$

where $\mathscr{F}_p$ denotes a fraction with p runs and $\#\{\cdot\}$ denotes the cardinality of the set $\{\cdot\}$. $\mathscr{F}$ is a saturated fraction if $\#\mathscr{F} = p$ and the parameters β are estimable.

On the other hand, the circuit basis of the matrix $A_{\mathscr{D}} = X_{\mathscr{D}}^t$ is defined as follows.

Definition 2. 1. A vector $\mathbf{u} = (u(1), \ldots, u(K))$ in $\ker(A_{\mathscr{D}})$ is a circuit if it has relatively prime entries and minimal support. The support of a vector is the set of indices for which the entries are non-zero.
2. The (finite) set of all the circuits is the circuit basis of $\ker(A_{\mathscr{D}})$, and it is denoted by $\mathscr{C}(A_{\mathscr{D}})$.

For a concise reference on the circuits, their combinatorial properties, and their applications to optimization problems, the reader can refer to [8]. The circuit basis of an integer matrix $A_{\mathscr{D}}$ can be computed through several packages for symbolic computation. The computations presented in the present paper are carried out with `4ti2`, see [9].

In [2] and [3] it is shown that robust fractions correspond to the fractions which minimize the intersections between the fraction and the support of the circuits. For space limits, we do not introduce here the full details of the theory, but we summarize the algorithm for finding nested robust fractions using the circuit basis.

1. Start with an arbitrary fraction $\mathscr{F}$ of a specified size n;
2. Repeat:

a. Consider the circuits of $\mathscr{C}(X_{\mathscr{D}})$ which are contained in $\mathscr{F}$;
b. For each run R in $\mathscr{F}$, compute the number of circuits in which R is contained. This is the loss function associated to R;
c. Remove from the fraction the run with the highest loss function. In case of ties, randomize.

2 Circuits in case of ordered factors

In the previous section we have not mentioned the problem of the choice of the coding of the factor levels. Indeed, the combinatorial analysis introduced above is usually applied in the framework of qualitative nominal factors. In such a case, the linear model in Eq. (1) can be written in the standard ANOVA form. For instance:

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{2}$$

with the constraints $\sum_i \alpha_i = 0$, $\sum_i \beta_i = 0$, $\sum_i (\alpha\beta)_{ij} = 0$,$\sum_j (\alpha\beta)_{ij} = 0$.

Using the model written in the form of Eq. (2) with qualitative factors, there is a large class of codings which are equivalent in terms of the kernel of the matrix $A_{\mathscr{D}} = X_{\mathscr{D}}^t$. Among these parametrizations, one can use a polynomial model in the form $\mathbb{E}(Y) = \sum_{\alpha \in L} c_\alpha X^\alpha$, where c_α are real coefficients, X^α are monomials, and L is a suitable list of exponents. The proof of the equivalence is based on the Identity Theorem for Polynomials, see [5] for a detailed analysis and examples.

To encode ordered factors we use here two codings:

1. For a linear ordered factor (e.g., the discretization of a quantitative factor) with s levels, we use the set $\{0, \ldots, s-1\}$ for a linear effect, and its powers for higher-order effects;
2. For a cyclic factor with s levels, we use the coding based on the roots of the unity:

$$\left\{ \omega_k = \frac{2\pi i k}{s} \ : \ k = 0, \ldots, s-1 \right\} \tag{3}$$

With this choice, the monomials $X, X^2, X^3, \ldots$ encode the cyclical nature of the factor. For computational reasons, the roots of the unity in Eq. (3) can be replaced with suitable Fourier-type functions, such as linear combinations of sin and cos functions.

Notice that there is a major difference between nominal and ordered factors. While for nominal factors all parametrizations are equivalent, when ordered factors are considered one can add to the model only a linear effect, or the linear effect plus some powers. This implies that when using the algorithm described in the previous section with ordered factors, special attention must be given on the circuit basis, taking the correct effects in the model matrix. In the next section we illustrate two examples involving ordered factors.

3 Examples

The first example considers two 2-level factors (X_1, X_2) and one 5-level factor (X_3) with two different models. The first model is linear in X_1, X_2, and X_3 and contains a constant term: $\mathbb{E}(Y_{(x_1,x_2,x_3)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. The number of degrees of freedom of this model is $p = 1+3 = 4$. For this model, the circuit basis is formed by 44 circuits with cardinality of the supports ranging from 2 to 5. The robustness of a D-optimal design with $n = 10$ runs is analyzed. The 10-run D-optimal design has been obtained using the full factorial design $\mathscr{D} = \{0,1\}^2 \times \{0,\ldots,4\}$ as candidate set. The exact distributions of the values of the robustness of the fractions which are obtained removing $k = 1,\ldots,n-p = 6$ points are computed and compared with the values of the robustness corresponding to the fractions found by the algorithm. Table 1 compares the values of the robustness of the fractions found by the algorithm (r_*) with the 75th, 90th and 95th percentile of the distributions of the robustness (p_{75}, p_{90}, p_{95} respectively) for different number of points (k) removed by the initial design. The value corresponding to the robustness of the initial design (r_0) is given at $k = 0$. It is worth noting that for each number k of points removed the algorithm provides values of robustness equal to the 95th percentile.

k	p_{75}	p_{90}	p_{95}	r_*
0		r_0=0.457		
1	0.476	0.476	0.476	0.476
2	0.529	0.529	0.529	0.529
3	0.629	0.629	0.629	0.629
4	0.6	0.6	0.867	0.867
5	0.8	1	1	1
6	1	1	1	1

Table 1: Example 1, $\mathbb{E}(Y_{(x_1,x_2,x_3)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Unlike the first model, the second one contains also the powers of X_3 of order 2,3,4: $\mathbb{E}(Y_{(x_1,x_2,x_3)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sum_{k=1}^{4} \beta_{3k} x_3^k$. The number of degrees of freedom of this model is $p = 1+2+4 = 7$. From the combinatorial point of view this model is simpler than the linear one: there are only 4 circuits in the circuit basis. Also in this case the robustness of a D-optimal design with $n = 10$ runs and obtained using the full factorial $\mathscr{D}$ as candidate set is analyzed. The exact distributions of the values of the robustness of the fractions which are obtained removing $k = 1,\ldots,n-p = 3$ points are computed and compared with the values of the robustness corresponding to the fractions found by the algorithm. In this case we do not display all the results, but the performance of the algorithm is similar to the previous example.

The second example is taken from [11] and is based on [10]. An animal scientist wants to compare wildlife densities in four different habitats over a year. However, due to the cost of experimentation, only $n = 16$ observations can be made (in the original example the requested size was $n = 12$, but $n = 16$ allows us to describe the

method better than $n = 12$). The following model is postulated for the density $Y_j(t)$ in habitat j during the month m:

$$\mathbb{E}(Y_j(t)) = \mu_j + \gamma m + \sum_{k=1}^{4} \alpha_k \cos\left(k\frac{\pi m}{4}\right) + \sum_{k=1}^{3} \beta_k \sin\left(k\frac{\pi m}{4}\right) \tag{4}$$

The model includes the habitat as a classification variable ($\mu_j, j = 1,\ldots,4$), the effect of time with an overall linear drift term γm, $m = 1,\ldots,12$ ($m = 1$ corresponds to January, $\ldots$, $m = 12$ corresponds to December), and cyclic behavior in the form of a Fourier series. There is no intercept term in the model and the number of parameters is $p = 4+1+4+3 = 12$.

The Optex procedure [11] is used to generate a D-optimal design $\mathscr{F}$ with $n = 16$ runs using the full factorial arrangement of four habitats by 12 months (48 runs) as candidate set. The model matrix $X_{\mathscr{F}}$ corresponding to the 16-run D-optimal design that has been generated by the Optex procedure is reported in Table 2. The month $m \in \{1,\ldots,12\}$ is expressed as a number $t \in [-1,+1]$ using the linear transformation $t = -1 + (2/11)(m-1)$.

μ_1	μ_2	μ_3	μ_4	γ	α_1	α_2	α_3	α_4	β_1	β_2	β_3
0	0	0	1	-0.636	-0.707	0	0.707	-1	0.707	-1	0.707
0	0	0	1	-0.091	0	-1	0	1	-1	0	1
0	0	0	1	0.273	1	1	1	1	0	0	0
0	0	0	1	0.636	0	-1	0	1	1	0	-1
0	0	1	0	-0.818	0	-1	0	1	1	0	-1
0	0	1	0	-0.273	-0.707	0	0.707	-1	-0.707	1	-0.707
0	0	1	0	0.091	0.707	0	-0.707	-1	-0.707	-1	-0.707
0	0	1	0	0.818	-0.707	0	0.707	-1	0.707	-1	0.707
0	1	0	0	-0.455	-1	1	-1	1	0	0	0
0	1	0	0	-0.273	-0.707	0	0.707	-1	-0.707	1	-0.707
0	1	0	0	0.273	1	1	1	1	0	0	0
0	1	0	0	0.455	0.707	0	-0.707	-1	0.707	1	0.707
1	0	0	0	-1	0.707	0	-0.707	-1	0.707	1	0.707
1	0	0	0	-0.091	0	-1	0	1	-1	0	1
1	0	0	0	0.091	0.707	0	-0.707	-1	-0.707	-1	-0.707
1	0	0	0	1	-1	1	-1	1	0	0	0

Table 2: Model matrix $X_{\mathscr{F}}$ corresponding to the 16-run D-Optimal design

The circuits of a matrix A can be computed for an integer matrix A. Then we have to build an approximate version $\tilde{X}_{\mathscr{F}}$ of $X_{\mathscr{F}}$. Let us denote with x_{ij} and $\tilde{x}_{ij}$ the elements of the matrices $X_{\mathscr{F}}$ and $\tilde{X}_{\mathscr{F}}$ respectively, $i = 1,\ldots,16$, $j = 1,\ldots,12$. We do not modify the coding of the habitats, $\tilde{x}_{ij} = x_{ij}, j = 1,\ldots,4$, we code the month using again the integer m as above ($\tilde{x}_{ij} = m$) and we round the remaining values defining $\tilde{x}_{ij} = 10r(x_{ij})$ where $r(x)$ returns the rounding of x to one decimal place. For example for the first row of $\tilde{X}_{\mathscr{F}}$ we have $\tilde{x}_{14} = 1$, $\tilde{x}_{15} = 3$ and $\tilde{x}_{16} = -7$ which correspond to the parameters μ_4, γ, and α_1 respectively. However the circuits

encodes the complexity of the model: there are 26 circuits in the circuit basis but the supports now range from 8 to 10 points.

From Table 3 it is worth noting that for each number k of points removed the algorithm provides values of robustness equal to the 95th percentile.

k	p_{75}	p_{90}	p_{95}	r_*
0		r_0=0.527		
1	0.527	0.527	0.527	0.527
2	0.571	0.571	0.571	0.571
3	0.615	0.769	0.769	0.769
4	1	1	1	1

Table 3: Example 2, $\mathbb{E}(Y_j(t)) = \mu_j + \gamma m + \sum_{k=1}^{4} \alpha_k \cos(k\frac{\pi m}{4}) + \sum_{k=1}^{3} \beta_k \sin(k\frac{\pi m}{4})$

Acknowledgements

Roberto Fontana gratefully acknowledges financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022. Roberto Fontana and Fabio Rapallo are members of the GNAMPA-INdAM group.

References

1. Butler, N.A., Ramos, V.M.: Optimal additions to and deletions from two-level orthogonal arrays. J. R. Stat. Soc. Ser. B **69**, 51–61 (2007)
2. Fontana, R., Rapallo, F., Wynn, H.P.: Circuits for robust designs. Stat. Pap. (Berl.), online first, 1–22 (2022)
3. Fontana, R., Rapallo, F.: Robustness of Fractional Factorial Designs through Circuits. In *SIS 2021 - Book of short papers*, C. Perna, N. Salvati, F. Schirripa Spagnolo Eds., Pearson (2021)
4. Ghosh, S.: On robustness of designs against incomplete data. Sankhya Ser. B **40**, 204–208 (1979)
5. Pistone, G., Riccomagno, E., Wynn, H.P.: Algebraic Statistics. Computational Commutative Algebra in Statitics Chapman & Hall/CRC, Boca Raton (2001)
6. Pukelsheim, F.: Optimal Design of Experiments, *Classics in Applies Mathematics*. SIAM, Philadelphia, PA (2006)
7. Street, D.J., Bird, E.M.: *D*-optimal orthogonal array minus *t* run designs. J. Stat. Theory Pract. **12**, 575–594 (2018)
8. Sturmfels, B.: Gröbner bases and convex polytopes, *University Lecture Series*, vol. 8. American Mathematical Society, Providence, RI (1996)
9. 4ti2 team: 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces (2018). URL https://4ti2.github.io
10. Mitchell, T. An algorithm for the construction of "*D*-optimal" experimental designs. Technometrics, **16**, 203–210 (1974)
11. SAS-Institute. SAS/QC 9.1: User's Guide. SAS Institute, Cary, NC (2004)