From Markov moves in contingency tables to linear model estimability
(joint work with Roberto Fontana and Maria Piera Rogantin)

Fabio Rapallo

Dipartimento di Scienze e Innovazione Tecnologica
Università del Piemonte Orientale, Alessandria (Italy)

mODa 10, Łagów Lubuski, Poland, 10-14 June 2013
The main goal of this work is to characterize saturated fractions of a factorial design in terms of algebraic and combinatorial objects derived from the design matrix.

This goal is accomplished by merging Design of Experiments and Contingency tables analysis, and then applying some tools from Algebraic Statistics.
Notation

- **D**: a full factorial design with \(d\) factors, \(A_1, \ldots, A_d\) with \(s_1, \ldots, s_d\) levels respectively:

\[
D = \{0, \ldots, s_1 - 1\} \times \cdots \times \{0, \ldots, s_d - 1\}
\]

- a linear model on **D**:

\[
Y = X\beta + \varepsilon
\]

- **p**: the number of estimable parameters.

Under a suitable (full-rank) parametrization, the design matrix **X** has dimensions \(s_1 \cdots s_d \times p\).
A subset $\mathcal{F}$, or fraction, of a full design $\mathcal{D}$, is a saturated fraction or saturated design if:

- it has minimal cardinality $\# \mathcal{F} = p$;
- it allows us to estimate the model parameters.

By definition, the design matrix $X_{\mathcal{F}}$ of a saturated design (under a full-rank parametrization) is a non-singular matrix with dimensions $p \times p$. 
Example

$2^4$ design and the model with simple effects and 2-way interactions.

```
  1  a_0  b_0  c_0  d_0  a_0b_0  a_0c_0  a_0d_0  b_0c_0  b_0d_0  c_0d_0
(0,0,0,0)  1  1  1  1  1  1  1  1  1  1  1
(0,0,0,1)  1  1  1  1  0  1  1  0  1  0  0
(0,0,1,0)  1  1  1  0  1  1  0  1  0  1  0
(0,0,1,1)  1  1  1  0  0  1  0  0  0  0  0
(0,1,0,0)  1  1  0  1  1  0  1  1  0  0  0
(0,1,0,1)  1  1  0  1  0  0  1  0  0  0  0
(0,1,1,0)  1  1  0  0  1  0  0  1  0  0  0
(0,1,1,1)  1  1  0  0  0  0  0  0  0  0  0
(1,0,0,0)  1  0  1  1  1  0  0  0  1  1  1
(1,0,0,1)  1  0  1  1  0  0  0  0  1  0  0
(1,0,1,0)  1  0  1  0  1  0  0  0  0  1  0
(1,0,1,1)  1  0  1  0  0  0  0  0  0  0  0
(1,1,0,0)  1  0  0  1  1  0  0  0  0  0  1
(1,1,0,1)  1  0  0  1  0  0  0  0  0  0  0
(1,1,1,0)  1  0  0  0  1  0  0  0  0  0  0
(1,1,1,1)  1  0  0  0  0  0  0  0  0  0  0
```
Example

As the matrix $X$ has rank 11, we search for fractions with 11 points.

1. The fraction

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

$$\quad (1, 0, 1, 0), (1, 1, 0, 0), (1, 1, 0, 1), (1, 1, 1, 0), (1, 1, 1, 1)\}$$

**is saturated.**

2. The fraction

$$\mathcal{F}_2 = \{(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 1, 1), (0, 1, 0, 0), (0, 1, 1, 0), (1, 0, 0, 1),$$

$$\quad (1, 0, 1, 0), (1, 1, 0, 0), (1, 1, 0, 1), (1, 1, 1, 0), (1, 1, 1, 1)\}$$

**is not** saturated.
Example

A direct computation shows that in the $2^4$ design with simple effects and 2-way interactions there are $\binom{16}{11} = 4,368$ fractions with 11 points:

- 3,008 are saturated
- 1,360 are non saturated

Our goal

To check whether a fraction is saturated or not with a purely combinatorial criterion.
Fractions and contingency tables

We identify:

- a **fraction** $\mathcal{F}$ of a factorial design

$$\mathcal{F} \subset \mathcal{D} = \{0, \ldots, s_1 - 1\} \times \cdots \times \{0, \ldots, s_d - 1\}$$

- a $s_1 \times \cdots \times s_d$ **contingency table** $N(\mathcal{F})$ whose entries are the indicator functions of the fraction (i.e., equal to 1 if the point belongs to the fraction and zero 0 otherwise).
**Example**

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

<table>
<thead>
<tr>
<th>( A_3 = 0 )</th>
<th>( A_1 = 0 )</th>
<th>( A_1 = 1 )</th>
<th>( A_2 = 0 )</th>
<th>( A_2 = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_4 = 0 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( A_4 = 1 )</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( A_3 = 1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( A_4 = 0 )</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( A_4 = 1 )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Example

\[ \mathcal{F}_2 = \{(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 1, 1), (0, 1, 0, 0), (0, 1, 1, 0), (1, 0, 0, 1), (1, 0, 1, 0), (1, 1, 0, 0), (1, 1, 0, 1), (1, 1, 1, 0), (1, 1, 1, 1)\} \]

<table>
<thead>
<tr>
<th>( N(\mathcal{F}_2) )</th>
<th>( A_1 = 0 )</th>
<th>( A_1 = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( A_2 = 0 )</td>
<td>( A_2 = 1 )</td>
</tr>
<tr>
<td>( A_3 = 0 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( A_4 = 0 )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( A_4 = 1 )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( A_3 = 1 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( A_4 = 0 )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( A_4 = 1 )</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Given a contingency table with $K$ cells and a $p \times K$ integer matrix $A$ ("the design matrix"), we define:

- the **polynomial ring** $\mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_K]$ of all polynomials with indeterminates $x_1, \ldots, x_K$ and real coefficients, i.e., we define an indeterminate for each cell of the table or, equivalently, for each point of the full-factorial design.

- the **toric ideal** defined by $A$ is the binomial ideal

$$\mathcal{I}_A = \langle x^a - x^b \mid Aa = Ab \rangle$$

where the monomials $x^a$ are written in vector notation

$$x^a = x_1^{a_1} \cdots x_K^{a_K}.$$
The computation of a system of generators of an ideal is a non trivial task in Computer Algebra, but special algorithms exist for toric ideals. In our computations we have used 4ti2.

An actual way to do that is to compute the reduced Gröbner basis of the ideal.

**Remark**

The computation of a Gröbner basis depends on the term-order chosen in the polynomial ring $\mathbb{R}[x]$, but for a given term-order the reduced Gröbner basis is unique and can be computed through symbolic software.

In Algebraic Statistics for contingency tables, Gröbner bases are used to find Markov bases.
Algebraic objects for contingency tables

Among all term-orders, the elimination term-order for a given indeterminate, say $x_K$, leads to the Gröbner basis of the projection $\text{Elim}(x_K; I) := I \cap \mathbb{R}[x_1, \ldots, x_{K-1}]$.

Remark

The statistical counterpart of elimination of indeterminates is the definition of a statistical model for incomplete tables.

Here we consider three different bases of a toric ideal

1. the circuits
2. the Universal Gröbner basis
3. the Graver basis
Universal Gröbner bases

Definition

Let $I_A$ be a toric ideal. The union of all reduced Gröbner bases of $I_A$ is called the *Universal Gröbner basis* $\mathcal{U}_A$ of $I_A$.

The computation of the Universal Gröbner basis is unfeasible for most ideals, but fortunately there are special algorithms for doing that in the case of toric ideals.
Graver bases

Definition

A binomial \( f = x^a - x^b \in \mathcal{I}_A \) is **primitive** if there is no binomial \( g = x^c - x^d \in \mathcal{I}_A \), with \( g \neq f \), such that \( c \leq a \) and \( d \leq b \). The **Graver basis** \( Gr_A \) of \( A \) is the set of all primitive binomials in \( \mathcal{I}_A \).

Definition

The **support** of a binomial \( f = x^a - x^b \) is the set of indices \( i \) \((i = 1, \ldots, K)\) such that \( a(i) \neq 0 \) or \( b(i) \neq 0 \). We denote the support of \( f \) with \( \text{supp}(f) \).
Circuits

Definition

An irreducible binomial \( f = x^a - x^b \in I_A \) is a circuit if there is no other binomial \( g \in I_A \) such that \( \text{supp}(g) \subset \text{supp}(f) \) and \( \text{supp}(g) \neq \text{supp}(f) \). We denote the set of all circuits of \( I_A \) with \( C_A \).

In general the following inclusions hold:

\[
C_A \subseteq U_A \subseteq Gr_A.
\]
The main result

In order to match the algebraic notation with the statistical notation, we use here the transpose of the design matrix $X$ in place of the design matrix $X$. We denote by $A = X^t$ the transpose of the design matrix, and we call it again *design matrix*.

**Remark**

Each column of $A$ identifies a design point, and therefore the definition of a set of column-indices is equivalent to the definition of the fraction with the corresponding design points.

Given $\mathcal{F} = \{i_1, \ldots, i_p\}$, $A_{\mathcal{F}}$ is the submatrix of $A$ obtained by selecting the columns of $A$ according to $\mathcal{F}$. 
The main result is:

**Theorem**

Let $A$ be a full-rank design matrix with dimensions $p \times K$ and let $C_A = \{f_1, \ldots, f_L\}$ be the set of its circuits. Given a set $F$ of $p$ column-indices of $A$, the submatrix $A_F$ is non-singular if and only if $F$ does not contain any of the supports $\text{supp}(f_1), \ldots, \text{supp}(f_L)$.

**Remark**

This theorem replaces a linear algebra condition with a combinatorial property for checking whether a fraction is saturated or not.
“⇐”: Suppose that there is a circuit $f \in C_A$ with support in $\mathcal{F}$, i.e., $x^{m^+} - x^{m^-}$ is in $\mathcal{I}_A$. Hence, $Am = 0$ and projecting onto the subspace $\mathbb{R}[x_i : i \in \mathcal{F}]$ we obtain $A_{\mathcal{F}}m = 0$, i.e., $A_{\mathcal{F}}$ is singular.
Proof ("if" direction)

“⇒”: If $A_F$ is singular, then there is a null linear combination of its columns with coefficients in $\mathbb{Z}$.

1. There exists a non-zero vector $m \in \mathbb{Z}^p$ with $A_F m = 0$.
2. Writing $m = m^+ - m^-$, the binomial $x^{m^+} - x^{m^-}$ belongs to $I_{A_F}$, the toric ideal associated to $A_F$.
3. As $I_{A_F} = \text{Elim}(x_i : i = 1, \ldots, K, i \notin F ; I_A)$, $I_{A_F}$ is non-empty and there is a binomial in the Gröbner basis $G_{A, \tau}$ of $I_A$, where $\tau$ is the elimination term order for the indeterminates $(x_i : i = 1, \ldots, K, i \notin F)$.
4. By definition of Universal Gröbner basis, this implies that there is a binomial $g$ in $U_A$ and its support in $F$.
5. If the binomial $g$ is a circuit, the proof if complete. If not, there do exist a circuit $h \in C_A$ with $\text{supp}(h) \subset \text{supp}(g)$ and it is enough to choose such binomial $h$. 
Example

In our examples we label the design points lexicographically. For instance in the $2^4$ case we define the indeterminates as follows:

<table>
<thead>
<tr>
<th>(0, 0, 0, 0)</th>
<th>(0, 0, 0, 1)</th>
<th>(0, 0, 1, 0)</th>
<th>(0, 0, 1, 1)</th>
<th>(0, 1, 0, 0)</th>
<th>⋮</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$x_2$</td>
<td>$x_3$</td>
<td>$x_4$</td>
<td>$x_5$</td>
<td>⋮</td>
</tr>
</tbody>
</table>

Moreover, we use the log notation for binomials:

$$x^a - x^b \mapsto a - b$$

For instance the binomial $x_1 x_3^2 - x_2 x_7 x_8$ in $\mathbb{R}[x_1, \ldots, x_8]$ is written as $(1, -1, 2, 0, 0, 0, -1, -1)$. 
We consider again the $2^4$ design and the model with simple effects and 2-way interactions. The design matrix has 140 circuits. They can be divided into three classes, up to permutations of factors or levels:

- 20 circuits of the form
  \[ f_1 = (0, 0, 0, 0, 1, -1, -1, 1, -1, 1, 1, -1, 0, 0, 0, 0) \]

- 40 circuits of the form
  \[ f_2 = (1, -2, 0, 1, 0, 1, -1, 0, 0, 1, -1, 0, -1, 0, 2, -1) \]

- 80 circuits of the form
  \[ f_3 = (1, 0, -2, 1, 0, -1, 1, 0, -2, 1, 3, -2, 1, 0, -2, 1) \]
### Example

<table>
<thead>
<tr>
<th>$A_3$</th>
<th>$A_4$</th>
<th>$A_1 = 0$</th>
<th>$A_1 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_3 = 0$</td>
<td>$A_4 = 0$</td>
<td>$A_2 = 0$</td>
<td>$A_2 = 1$</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>$*$</td>
<td>$*$</td>
</tr>
<tr>
<td>$A_3 = 1$</td>
<td>$A_4 = 0$</td>
<td>$*$</td>
<td>$*$</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>$*$</td>
<td>$*$</td>
</tr>
<tr>
<td>$A_3 = 0$</td>
<td>$A_4 = 0$</td>
<td>$*$</td>
<td>$*$</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>$*$</td>
<td>$*$</td>
</tr>
<tr>
<td>$A_3 = 1$</td>
<td>$A_4 = 0$</td>
<td>$*$</td>
<td>$*$</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>$*$</td>
<td>$*$</td>
</tr>
</tbody>
</table>
To see that the fraction $N(F_2)$ in the previous example is not saturated, we compare it with the support of the circuit $f_2$.

<table>
<thead>
<tr>
<th>$N(F_2)$</th>
<th>$A_1 = 0$</th>
<th>$A_1 = 1$</th>
<th>$A_2 = 0$</th>
<th>$A_2 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_3 = 0$</td>
<td>$A_4 = 0$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$A_3 = 1$</td>
<td>$A_4 = 0$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\text{supp}(f_2)$</th>
<th>$A_1 = 0$</th>
<th>$A_1 = 1$</th>
<th>$A_2 = 0$</th>
<th>$A_2 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_3 = 0$</td>
<td>$A_4 = 0$</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>$A_3 = 1$</td>
<td>$A_4 = 0$</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>$A_4 = 1$</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>
The elements of the bases (Graver or circuits) are the coefficients of the linear combination of fraction points to have zero as result, and therefore a singular matrix.

**Remark**

The elements of the bases with more than $p$ values different from zero are not of interest of our aims.

In our running examples, the circuits of the third kind like the circuit

$$f_3 = (1, 0, -2, 1, 0, -1, 1, 0, -2, 1, 3, -2, 1, 0, -2, 1)$$

can be excluded.
A special case happens when the design matrix is unimodular.

**Definition**

A nonnegative integer matrix with rank \( p \) is *unimodular* if all its non-zero \( p \times p \) minors are equal to \( \pm 1 \). A nonnegative integer matrix is *totally unimodular* if all its non-zero minors are equal to \( \pm 1 \).

Some properties:

- A totally unimodular matrix is unimodular.
- The entries of a totally unimodular matrix are 0 and 1.
- Each submatrix of a totally unimodular matrix is again totally unimodular.
- The transpose of a totally unimodular matrix is again totally unimodular.
Proposition

If $A$ is a unimodular matrix, then

$$ C_A = U_A = Gr_A. $$

Using some properties of the Lawrence lifting of a matrix, one can prove the following

Theorem

*The design matrix of all models for a $s_1 \times s_2 \times 2 \times \cdots \times 2$ factorial design is totally unimodular.*
Examples with binary factors

- Design $2^5$; model with simple factors and 2-way and 3-way interactions. The saturated model has 26 points. There are 3,254 circuits that can be divided into 12 classes.

- Design $2^5$; model with simple factors. The saturated model has 6 points. The circuits are 353,616 elements that can be divided into 38 classes, up to permutations of factors or levels.
Examples with multilevel factors

- Design $2 \times 3 \times 4$; model with simple factors and 2-way interactions.
  The saturated model has 18 points. There are 42 circuits that can be divided into two classes.

- Design $3 \times 3 \times 4$; model with simple factors and 2-way interactions.
  The saturated model has 24 points. There are 19,722 circuits that can be divided into 20 classes.
The circuits of a design matrix can be computed with the free software 4ti2.

- The circuits of the previous examples are computed in less than 2 seconds
- As usual in Computer Algebra and Combinatorics, computational problems may arise for large design matrices

**Important remark**

The circuits do not depend on the fraction, but only on the design matrix of the full factorial design $\mathcal{D}$. 
The algorithm based on circuits is useful for generating random saturated fractions without computing the determinant of the design matrix of each fraction. Moreover, we allow restrictions on the projections.

1. Starting from a given fraction, with a MCMC algorithm for contingency tables we are able to find a Markov chain on the set of all fraction with given projections.

2. Each table is compared with the set of circuits to check whether it is saturated or not.

Curiously, the basic moves to run the MCMC algorithm are the circuits, for a large class of design matrices.
Future directions

This work suggests several future directions:

- To explore how the results can be extended for the characterization of saturated fractions to more general designs.
- To study the connections between fractions and graphs.
- To investigate the classification of the saturated fractions with respect to some statistical criteria, such as minimum aberration and optimality.
- To make available the implementation of the algorithms in statistical softwares, such as SAS or R.
Thanks for your attention!

fabio.rapallo@unipmn.it

Preprint available at arXiv:1304.7914v1