2008 International Conference on Applied Probability and Statistics
with emphasis on Business and Industrial Statistics
Hanoi, Vietnam 1 - 3 December 2008

Invited Session 5C: DOE3, chair M. P. Rogantin

Polynomial algebra in kriging over a regular design

Giovanni Pistone and Grazia Vicario

Tuesday December 2, 2009

A Statistical Toolkit

- Most practical statistical model are algebraic in nature and where developed in the pre-computer to provide computable solutions in closed form.
- Today many applied model are still algebraic in nature, but cannot be solved in closed form because of the high dimensionality: large linear models, Bayes nets, ...
- With the advent of modern Symbolic Computation Software the "closed form" solution approach can be pushed on a little further, so that small Bayes nets can be solved.
- The algebraic approach provides further insight concerning such issues as the existence of likelihood estimators in contingency table models.
- Recently, a special issue of the journal Statistical Sinica has been devoted to this new area we called Algebraic Statistics in a book published on 2001. In particular, see Stephen E. Fienberg. Expanding the statistical toolkit with algebraic statistics. *Statistica Sinica*, 17:1261–1272, 2007 at `http://www3.stat.sinica.edu.tw/statistica/J17N4/editorial.pdf`.
- There is a nice tutorial by Seth Sullivan at `http://www3.stat.sinica.edu.tw/statistica/`.

Summary

1. A short review of algebraic statistics of toric models on a **finite state space**.
2. A toy example of application in **continuous models**: gaussian models as used in Kriging and Computer Experiments.
3. Conclusion.

Ideals

- $R = \mathbb{Q}[x_j : j = 1, \ldots, n]$ is the **ring** of polynomials with rational coefficients and n indeterminate $x_1, \ldots, x_n$.
- $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_x}$ is a **monomial**; $x^\alpha - x^\beta$ is a **binomial**.
- $I \subset R$ is an **ideal** if it is closed under the (internal) sum operation and under multiplication by any element of the ring; typically, an ideal is the set of all polynomials that are zero on some subset of $\mathbb{Q}^n$; all ideals are finitely generated, i.e. there exist a finite set of polynomials $g_1, \ldots, g_m$ such that all element $f \in I$ can be written as

$$f = \sum_{i=1}^{m} h_i g_i, \quad h_i \in R, i = 1, \ldots, m$$

- Given a system of polynomial equations $\begin{cases} g_1(x) = 0 \\ \vdots \\ g_m(x) = 0 \end{cases}$, the ideal generated by $g_1, \cdots, g_m$ consists of all the polynomials $f(x) = 0$ such that the equation $f = 0$ is obtained by manipulating the given system.

Toric ideals in Statistics

- It has been first shown by Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1): 363–397, 1998. ISSN 0090-5364 that a special case of ideal, called toric ideal, is of primary importance in Statistics.
- The previous paper is devoted to an application MCMC to exact testing in contingency tables. The same basic idea has been applied to statical models arising in Bayes networks in Giovanni Pistone, Eva Riccomagno, and Henry P. Wynn. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman&Hall, 2001, where such models are shown to be a type of exponential model. A general theory is in Dan Geiger, Christopher Meek, and Bernd Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34:1463–1492, 2006.
- The setup consists of a ring $S = Q[p_1, \ldots, p_n, t_1, \ldots, t_d]$, where $\Omega = \{1, \ldots, n\}$ is a finite sample space, p_i is the probability of i, and t_j is a parameter. The statistical model is given as $\begin{cases} p_1 = t^{\alpha_1} \\ \vdots \\ p_n = t^{\alpha_n} \end{cases}, \sum_i p_i = 1.$

Elimination and binomial invariants

- Given the system of polynomial equations

$$\begin{cases} p_1 - t^{\alpha_1} & = & 0 \\ & \vdots & \\ p_n - t^{\alpha_n} & = & 0 \\ p_1 + \cdots + p_n - 1 & = & 0 \end{cases}$$

in the ring $S = Q[p_1, \ldots, x_n, t_1, \ldots, t_d]$,

the algebraic elimination of the indeterminate t_j produces a system of equations in the indeterminate p_i only, which are called **algebraic invariants** of the toric model.

- More formally, the elimination ideal is the ideal

$$\text{Ideal}\,(p_1 - t^{\alpha_1}, \ldots, p_n - t^{\alpha_n}, p_1 + \cdots + p_n - 1) \cap Q[p_1, \ldots, p_n]$$

- Algebraic invariants are related with the orthogonal space of the log-linear model

$$\log p_i = \sum_{j=1}^{d} \alpha_{ij} \log t_j$$

Examples of toric model and algebraic invariants

Independence On a 2×2 table independence is

$$p_{ij} = t_{i+} t_{+j},$$

which is a toric model. The algebraic invariant is

$$\boxed{p_{11}p_{22} - p_{12}p_{21} = 0}.$$

Binomial The binomial probability $p_k = \binom{n}{k}(1-\theta)^{n-k}\theta^k$ is not directly a toric model. However, $q_k = p_k / \binom{n}{k}$ is, because

$$q_k = (1-\theta)^{n-k}\theta^k = t_1^{n-k} t_2^k, \qquad k = 1, \ldots, n$$

The model matrix of the log-model is $\begin{bmatrix} n & 0 \\ n-1 & 1 \\ \vdots & \\ 0 & n \end{bmatrix}$.

In case $n = 2$ the equation

$$\boxed{4p_0p_2 - p_1^2 = 0},$$

which is called Hardy-Weinberg constraint in Genetics, is obtained,

Computer Algebra

Computation The examples we have shown have been known for a long time. Algebraic computation are quite difficult, so that the algebraic approach is not usually taken seriously. However, computer programs able to do the exact algebraic computations that are needed.

CoCoA is a program to compute with numbers and polynomials. It is free. It works on many operating systems. It is used by many researchers, but can be useful even for "simple" computations. See `http://cocoa.dima.unige.it/` for a full documentation and downloading.

4ti2 is not as general as CoCoA, but in specific cases can be faster and more friendly. See a moderate documentation and downloading in `http://www.4ti2.de/`.

More options are available, both academic free software such as Singular or Macaulay2 or commercial such as Maple and Matematica.

Polynomial model with fixed fraction

Let us consider the polynomial model

$$y = a + bx + cx^2 \qquad x \in D = \{0 \dots 9\}$$

The model is evaluated on the fraction $F = \{x_1, x_2\}$. We compute the confounding structure with the methods of polynomial algebra. F is the set of rational solutions of the equations

$$x(x-1)\cdots(x-9) = 0, \quad (x-x_1)(x-x_2) = 0,$$

In particular, the second equation, that we call defining equation of the fraction, gives the aliasing of the squared term which is introduced by the fraction:

$$x^2 = (x_1 + x_2)x - x_1 x_2$$

Therefore. the original model is aliased with the lower degree model

$$y = a+bx+cx^2 = y = a+bx+c\left[(x_1+x_2)x - x_1x_2\right] = (a - cx_1x_2)+[b + c(x_1+x_2)]\,x$$

and we have found two parametric functions which are estimable (by interpolation).

$$a - cx_1x_2, \quad b + c(x_1 + x_2)$$

Polynomial model with random fraction

Let us assume now that the fraction is a random set $\{X_1, X_2\} \subset D$, where X_1 is uniformly distributed on D and $X_2|X_1 = x_1$ is uniformly distributed on $D \setminus x_1$. Therefore, the fraction F is sampled uniformly among the 90 cases. The relevant expected values of are

$$\mathrm{E}(X_1 + X_2) = \frac{1}{90} \sum_{0 \le x_2 \le 9, x_1 \neq x_2} (x_1 + x_2) = 9$$

$$\mathrm{E}(X_1 X_2) = \frac{1}{90} \sum_{0 \le x_2 \le 9, x_1 \neq x_2} x_1 x_2 = \frac{9}{2}$$

In this case, the interpolation procedure provides unbiased estimators of the parametric functions

$$a - \frac{9}{2}c, \quad b + 9c, \quad 2a + b$$

the latter being a consequences of the formers.
It should be noticed that the previous computations involves non linear functions of the random variables representing the random fraction: this is a consequence of the randomization procedure.

Kriging model with random fraction

Let us assume now that the response Y_x at point $x \in D = \{0 \dots 9\}$ is a jointly Gaussian centered random variable such that $\mathrm{Cov}\,(Y_{x_1}, Y_{x_2}) = \gamma(x_1 - x_2)$. We want to estimate the overall total $Y_+ = \sum_{x \in D} Y_x$ with the conditional expectation $\mathrm{E}\,(Y_+ | Y_{x_i}, i = 1, 2, 3)$ for a given 3-fraction $F = \{x_1, x_2, x_3\} \subset D$. The simplest procedure to estimate the covariances $\gamma(d)$ is the estimator

$$G(d) = \sum_{\{u,v\} \subset F, |u-v|=d} Y_u Y_v, \quad d = 0, 1, \dots, 9$$

We have

$$\mathrm{E}\,(G(d)) = n_d \gamma(d), \quad \text{where } n_d = \sum_{\{u,v\} \subset F, |u-v|=d} 1$$

Therefore $G(d)/n_d$ is un unbiased estimator of $\gamma(d)$, whenever $n_d \neq 0$. Assume now that F is uniformly sampled among the $\binom{10}{3} = 120$ 3-subsets of D. The probability of the estimator is defined, $p(d) = \mathrm{P}\,(N_d \neq 0)$, is given in the table.

d:	1	2	3	4	5	6	7	8	9
$p(d)$:	$\frac{64}{120}$	$\frac{58}{120}$	$\frac{52}{120}$	$\frac{46}{120}$	$\frac{40}{120}$	$\frac{32}{120}$	$\frac{24}{120}$	$\frac{16}{120}$	$\frac{8}{120}$

Computer experiments

Computer experiments are usually contrasted to physical experiments:

- A physical experiments force constrains on the choice of experimental treatments, while a numerical experiment does not.
- A computer experiment does not show experimental error, while a physical experiment does.

However, the very use of the word *experiment* in both cases, hints for the use of the same basic methodology. We refer to Jerome Sacks, Susannah B. Schiller, and William J. Welch. Designs for computer experiments. *Technometrics*, 31(1): 41–47, 1989. ISSN 0040-1706 and Thomas J. Santner, Brian J. Williams, and William I. Notz. *The design and analysis of computer experiments*. Springer Series in Statistics. Springer-Verlag, New York, 2003. ISBN 0-387-95420-1.

- Is that possible to compute in the general case the aliasing introduced by a given trial set?
- Are randomization and statistical inference meaningful in a computer experiment?
- How to number, list, generate trial sets (fractions) of a given class, e.g. LH's orthogonal arrays?

The model $\Gamma_{xy} = \exp\left(-\theta\|x-y\|_1\right)$, $t = \mathrm{e}^{-\theta}$

- The covariance matrix is

$$\Gamma = \begin{bmatrix} t^0 & t^1 & t^2 & t^1 & t^2 & t^3 & t^2 & t^3 & t^4 \\ t^1 & t^0 & t^1 & t^2 & t^1 & t^2 & t^3 & t^2 & t^3 \\ t^2 & t^1 & t^0 & t^3 & t^2 & t^1 & t^4 & t^3 & t^2 \\ t^1 & t^2 & t^3 & t^0 & t^1 & t^2 & t^1 & t^2 & t^3 \\ t^2 & t^1 & t^2 & t^1 & t^0 & t^1 & t^2 & t^1 & t^2 \\ t^3 & t^2 & t^1 & t^2 & t^1 & t^0 & t^3 & t^2 & t^1 \\ t^2 & t^3 & t^4 & t^1 & t^2 & t^3 & t^0 & t^1 & t^2 \\ t^3 & t^2 & t^3 & t^2 & t^1 & t^2 & t^1 & t^0 & t^1 \\ t^4 & t^3 & t^2 & t^3 & t^2 & t^1 & t^2 & t^1 & t^0 \end{bmatrix} \begin{matrix} 11 \\ 21 \\ 31 \\ 12 \\ 22 \\ 32 \\ 13 \\ 23 \\ 33 \end{matrix}$$
$$\begin{matrix} 11 & 21 & 31 & 12 & 22 & 32 & 13 & 23 & 33 \end{matrix}$$

- Using CoCoA, we obtain the value of the determinant as $\Gamma = (t^2-1)^{12}$

CoCoA program 1st part

```
Use R::= Q[x[1..9,1..9],t]; -- specify the ring!
                            -- x's to be used later
Gamma := Mat([              -- a matrix is a list of lists
                            -- scanned by rows
 [t^0,t^1,t^2,t^1,t^2,t^3,t^2,t^3,t^4],
 [t^1,t^0,t^1,t^2,t^1,t^2,t^3,t^2,t^3],
 [t^2,t^1,t^0,t^3,t^2,t^1,t^4,t^3,t^2],
 [t^1,t^2,t^3,t^0,t^1,t^2,t^1,t^2,t^3],
 [t^2,t^1,t^2,t^1,t^0,t^1,t^2,t^1,t^2],
 [t^3,t^2,t^1,t^2,t^1,t^0,t^3,t^2,t^1],
 [t^2,t^3,t^4,t^1,t^2,t^3,t^0,t^1,t^2],
 [t^3,t^2,t^3,t^2,t^1,t^2,t^1,t^0,t^1],
 [t^4,t^3,t^2,t^3,t^2,t^1,t^2,t^1,t^0]
]);
DetGamma := Det(Gamma);     -- is a polynomial
Factor(DetGamma);           -- factorization of the polynomial
```

CoCoA program 2nd part

We first compute Γ^{-1}. The computation is symbolic, so we do not care about the singular case $t = 1$.

```
InvGamma := Inverse(Gamma);
InvGamma                                  -- check the inverse
NumInvGamma := (t^2-1)^2InvGamma;         -- suggested by inspection
                                          -- the previous step
Latex(NumInvGamma);                       -- export to latex
                                          -- for the presentation
```

As the model is toric in the entries of Γ we compute the elimination ideal.

```
Eqs := [[Gamma[I,J]-x[I,J]|I In 1..9] | J In 1..9];
                                          -- list of polynomials
I := Ideal(Flatten(Eqs));                 -- Ideal is a data structure!
J := Elim(t,I);                           -- computes the elimination ideal
Latex(J);
```

Elimination ideal

- 9 equations for the diagonal: $-x_{99}+1=0, -x_{88}+1=0\ldots$
- 36 equations for the symmetry: $x_{69}-x_{89}=0, -x_{89}+x_{98}=0,\ldots$
- 6 *generating* nonlinear equations for the model

$$\begin{aligned}
-x_{98}^2+x_{79}&=0,\\
-x_{97}x_{98}+x_{49}&=0,\\
-x_{97}^2+x_{49}x_{98}&=0,\\
-x_{94}x_{98}+x_{19}&=0,\\
-x_{94}x_{97}+x_{19}x_{98}&=0,\\
-x_{94}^2+x_{19}x_{97}&=0
\end{aligned}$$

- The actual form of the equations is controlled by a CoCoA setting called *monomial order*.
- Algebraic invariants show the geometrical shape of the model.

CoCoA program 3rd part

```
-- 1 2 3  -- 3x3 grid
-- 4 5 6
-- 7 8 9

Locations := 1..9;
Design := [1,5,9];
Predicted := Diff(Locations,Design);

-- Gaussian conditional expectation

Covariance := Submat(Gamma,Predicted,Design);
GammaDesign := Submat(Gamma,Design,Design);
InvGammaDesign := Inverse(GammaDesign);
Variance := Covariance*InvGammaDesign*Transposed(Covariance);
DiagVariance := [Variance[1,1],Variance[2,2],Variance[3,3],
                 Variance[4,4],Variance[5,5],Variance[6,6]];
```

Gaussian conditional expectation

- The (symbolic) precision matrix can be computed notwithstanding the critical value $t = 1$.
- $\Gamma^{-1} =$

$$\begin{bmatrix}
1 & -t & 0 & -t & t^2 & 0 & 0 & 0 & 0 \\
-t & t^2+1 & -t & t^2 & -t^3-t & t^2 & 0 & 0 & 0 \\
0 & -t & 1 & 0 & t^2 & -t & 0 & 0 & 0 \\
-t & t^2 & 0 & t^2+1 & -t^3-t & 0 & -t & t^2 & 0 \\
t^2 & -t^3-t & t^2 & -t^3-t & t^4+2t^2+1 & -t^3-t & t^2 & -t^3-t & t^2 \\
0 & t^2 & -t & 0 & -t^3-t & t^2+1 & 0 & t^2 & -t \\
0 & 0 & 0 & -t & t^2 & 0 & 1 & -t & 0 \\
0 & 0 & 0 & t^2 & -t^3-t & t^2 & -t & t^2+1 & -t \\
0 & 0 & 0 & 0 & t^2 & -t & 0 & -t & 1
\end{bmatrix}$$

- The 0's in Γ^{-1} reveal a structure of conditional independence.

Prediction errors

- The distribution is stationary, therefore we look for the maximal variance of the prediction.
- The variance of the the prediction at the points is

Point:	2	3	4	6	7	8
Variance:	$\frac{2t^2}{t^2+1}$	$\frac{-t^6+3t^4}{t^2+1}$	$\frac{2t^2}{t^2+1}$	$\frac{2t^2}{t^2+1}$	$\frac{-t^6+3t^4}{t^2+1}$	$\frac{2t^2}{t^2+1}$

- We look for the maximum between

$$\frac{2t^2}{t^2+1} \quad \text{and} \quad \frac{-t^6+3t^4}{t^2+1}$$

at different values of t. his part of the computation is numerical and should be done outside CoCoA, e.g. in R.

Conclusion

- Symbolic algebraic computation is a useful tool in the analysis of statistical models used in Computer Experiment.
- The use of symbolic software requires special data structures, such as rings, ideal, ...
- Algebraic design of experiments is relevant outside standard DoE.
- Explicit randomization should be used in both physical and computer experiments.