

Limits of exponential models on a lattice¹

Limiti di modelli esponenziali su reticoli

Giovanni Pistone²

Dipartimento di Matematica, Politecnico di Torino,

giovanni.pistone@polito.it

Riassunto: vengono elencati e illustrati con un esempio alcuni risultati nuovi che riguardano i modelli esponenziali di tipo reticolo. In particolare si discute l'uso dell'Algebra Commutativa Computazionale per il calcolo dei limiti che non soddisfano la condizione di positività stretta delle probabilità.

Keywords: Exponential models, graphical models, algebraic statistics.

1. Introduction

The term “Algebraic Statistics” (AS) was introduced in Pistone *et al.* (2001) to denote the use in Statistics of mathematical technologies taken from Computational Commutative Algebra (CCA). The subject, started as early as 1993 in a preprint of the seminal paper Diaconis and Sturmfels (1998), is currently under intensive development; the best description of the current state of the art is obtained by looking at the papers and discussion presented in the specialized workshops: GrostatVI³, Menton Fr 2003, Computational Algebraic Statistics⁴, AIM Palo Alto Ca 2003, Computational Algebra for Algebraic Geometry and Statistics⁵, Torino It 2004. We cannot include in this short note an extensive bibliography.

A very active area of development consists of the exploration of multiple application of algebraic computations to graphical models. The present paper illustrates the matter through a specific example, the 4-cycle as described in Lauritzen (1996). Two of the results below are published here, in short form, for the first time. Preliminary versions were published in Rapallo (2003) and Pistone and Wynn (2003). Related results with full proofs are in Rapallo (2004). The present research is joint work with Fabio Rapallo and Henry P. Wynn. At the Conference time, all relevant computer listings and results will be posted on the Author's web-site⁶.

2. Algebraic statistical models

Loosely speaking, a statistical model is *algebraic* if it is described, either parametrically or by constraints, by algebraic (polynomial or rational) relations. This simple idea can lead to difficult problems as we will show below.

¹Il lavoro è parte delle attività del Cofin2003 (G. Consonni) e le ricerche sono cofinanziate dal progetto intergruppi *Algebraic Statistics* Indam 2003

²Indirizzo per corrispondenza: DIMAT Politecnico, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

³<http://cocoa.dima.unige.it/conference/grostatvi/>

⁴<http://www.aimath.org/>

⁵<http://calvino.polito.it/~geometri/>

⁶<http://staff.polito.it/giovanni.pistone/sis2004>

Let $\mathcal{X}$ be a finite sample space, with number $n = \#\mathcal{X}$. We will describe our statistical model by assigning integer valued random variables $T_j : \mathcal{X} \rightarrow \mathbb{Z}_+ = \{0, 1, 2, \dots\}$, $j = 1, \dots, l$, which are assumed to be *sufficient* for that model. If we assume the probabilities to be *strictly* positive $p(x) > 0$, $x \in \mathcal{X}$, then the relevant model is the exponential model (or log-linear model)

$$\ln p(x) = \sum_{j=0}^l \psi_j T_j(x) \quad T_0(x) = 1 \quad x \in \mathcal{X} \quad (1)$$

As in this case the sufficient statistics are integers, the model can be given, by the parameters change $e^{\psi_j} = \zeta_j$, the following form, that we call *Generalized MultiNomial* (GMN) or *toric*.

$$p(x) = \prod_{j=0}^l \zeta_j^{T_j(x)} \quad x \in \mathcal{X} \quad (2)$$

It is easy to prove that

T1 *The exponential model (1) is the strictly positive part of the GMN model (2).*

The object of the present note is to discuss the less obvious questions:

1. Which other probabilities are included in the GMN model?
2. Is the closure of the exponential model a GMN model?
3. Is it possible to describe the GMN model as a union of exponential models with variable support?

Let us first introduce the contribution of CCA to the matter of our interest. The following results are contained (in a much more complex and general form) in Sturmfels (1996).

T2 *The probabilities $p(x)$, $x \in \mathcal{X}$, of the GMN model are characterized by a system of polynomial equations of a specific form, namely binomial and homogeneous.*

T3 *The previous equations can be solved by a finite algorithm with is implemented in Computer Algebra Software Systems⁷.*

One important part of (T2) can be derived as follows. In fact, the log-linear model can be equivalently described by constraints, and such constraints can be realized by integer vectors. Let us introduce the model matrix $Z = [T_j(x)]_{x \in \mathcal{X}; j=0 \dots l}$ and let us orthogonalize the columns to produce

$$Z_2 = [R_1 \cdots R_{n-l-1}] \quad Z_2^T Z = 0 \quad Z_2 \text{ is full rank} \quad (3)$$

Let $R_j = U_j - V_j$, $U_j = R_j^+$ i.e. U_j is the positive part of R_j and V_j is its negative part. Let us consider the following equations that we term *residual equations* (REQ):

$$\prod_{x \in \mathcal{X}} p(x)^{U_j(x)} = \prod_{x \in \mathcal{X}} p(x)^{V_j(x)} \quad (4)$$

Then we can prove the following.

⁶We mention, among others: CoCoA <http://cocoa.dima.unige.it/> and Macaulay2 <http://www.math.uiuc.edu/Macaulay2/>

T4 The exponential model (1) is the strictly positive part of the REQ model (4).

T5 Let $\mathcal{X}_j = \{x \in \mathcal{X} : T_j(x) = 0\}$. Then the restriction to $\mathcal{X}_j$ of the GMN model (2) is a GMN model with parametric equations $p(y) = \prod_{h \neq j} \zeta_h^{T_h(y)}$, $y \in \mathcal{X}_j$.

T6 Any GMN model is a union of exponential models.

We finally discuss the closure of an exponential model. Let us consider all one dimensional exponential models whose REQ's include (4). The set of all sufficient S is a semi-group of non-negative vectors. Let $S_1, \dots, S_k$ be its minimal set of generators.

T7 The GMN model with $S_1 \cdots S_k$ is the closure of the original exponential model.

Other results, not mentioned here, relates to the Markov Basis (MB) of the original model, see Diaconis and Sturmfels (1998) on MB and application to conditional inference.

3. The 4-cycle

Consider 4 binary variables X_1, X_2, X_3, X_4 which are assumed to be dependent only through the interactions 12, 23, 34, 41. Using the computations in Pistone *et al.* (2001) together with the previous results, we can see that the equations describing the parametrically GMN model are

$$\begin{aligned}
 0000 &\mapsto p_1 = \zeta_0 & 0110 &\mapsto p_9 = \zeta_0 \zeta_2 \zeta_3 \zeta_7 \\
 1000 &\mapsto p_2 = \zeta_0 \zeta_1 & 0101 &\mapsto p_{10} = \zeta_0 \zeta_2 \zeta_4 \\
 0100 &\mapsto p_3 = \zeta_0 \zeta_2 & 0011 &\mapsto p_{11} = \zeta_0 \zeta_3 \zeta_4 \zeta_8 \\
 0010 &\mapsto p_4 = \zeta_0 \zeta_3 & 1110 &\mapsto p_{12} = \zeta_0 \zeta_1 \zeta_2 \zeta_3 \zeta_5 \zeta_7 \\
 0001 &\mapsto p_5 = \zeta_0 \zeta_4 & 1101 &\mapsto p_{13} = \zeta_0 \zeta_1 \zeta_2 \zeta_4 \zeta_5 \zeta_6 \\
 1100 &\mapsto p_6 = \zeta_0 \zeta_1 \zeta_2 \zeta_5 & 1011 &\mapsto p_{14} = \zeta_0 \zeta_1 \zeta_3 \zeta_4 \zeta_6 \zeta_8 \\
 1010 &\mapsto p_7 = \zeta_0 \zeta_1 \zeta_3 & 0111 &\mapsto p_{15} = \zeta_0 \zeta_2 \zeta_3 \zeta_4 \zeta_7 \zeta_8 \\
 1001 &\mapsto p_8 = \zeta_0 \zeta_1 \zeta_4 \zeta_6 & 1111 &\mapsto p_{16} = \zeta_0 \zeta_1 \zeta_2 \zeta_3 \zeta_4 \zeta_5 \zeta_6 \zeta_7 \zeta_8
 \end{aligned} \tag{5}$$

and in polynomial form

$$\left\{ \begin{array}{ll}
 -p_3 p_5 + p_1 p_{10} = 0 & p_1 p_8 p_{10} p_{12} + p_2 p_5 p_9 p_{13} = 0 \\
 -p_2 p_4 + p_1 p_7 = 0 & -p_6 p_8 + p_2 p_{13} = 0 \\
 -p_9 p_{11} + p_4 p_{15} = 0 & p_3 p_7 p_{11} p_{13} - p_4 p_6 p_{10} p_{14} = 0 \\
 p_5 p_9 p_{14} - p_4 p_8 p_{15} = 0 & p_4 p_8 p_{10} p_{12} - p_5 p_7 p_9 p_{13} = 0 \\
 -p_1 p_9 p_{10} p_{14} + p_3 p_4 p_8 p_{15} = 0 & -p_5 p_6 p_{14} p_{15} + p_2 p_{10} p_{11} p_{16} = 0 \\
 -p_8 p_{11} + p_5 p_{14} = 0 & -p_4 p_6 p_{14} p_{15} + p_3 p_7 p_{11} p_{16} = 0 \\
 p_2 p_9 p_{10} p_{14} - p_3 p_7 p_8 p_{15} = 0 & -p_1 p_6 p_{14} p_{15} + p_2 p_3 p_{11} p_{16} = 0 \\
 p_3 p_{11} p_{12} - p_4 p_6 p_{15} = 0 & -p_{13} p_{15} + p_{10} p_{16} = 0 \\
 -p_1 p_{10} p_{11} p_{12} + p_4 p_5 p_6 p_{15} = 0 & p_4 p_8 p_{12} p_{15} - p_5 p_7 p_9 p_{16} = 0 \\
 -p_6 p_9 + p_3 p_{12} = 0 & p_2 p_9 p_{13} p_{14} - p_3 p_7 p_8 p_{16} = 0 \\
 p_2 p_{10} p_{11} p_{12} - p_5 p_6 p_7 p_{15} = 0 & p_1 p_8 p_{12} p_{15} - p_2 p_5 p_9 p_{16} = 0 \\
 p_2 p_{11} p_{13} - p_5 p_6 p_{14} = 0 & p_1 p_9 p_{13} p_{14} - p_3 p_4 p_8 p_{16} = 0 \\
 -p_1 p_7 p_{11} p_{13} + p_4 p_5 p_6 p_{14} = 0 & p_1 p_{11} p_{12} p_{13} - p_4 p_5 p_6 p_{16} = 0 \\
 -p_3 p_8 p_{12} + p_2 p_9 p_{13} = 0 & -p_{12} p_{14} + p_7 p_{16} = 0
 \end{array} \right. \tag{6}$$

white the REQ equations are restricted to the quadrics.

$$\left\{ \begin{array}{ll} -p_3p_5 + p_1p_{10} = 0 & -p_6p_8 + p_2p_{13} = 0 \\ -p_2p_4 + p_1p_7 = 0 & -p_{13}p_{15} + p_{10}p_{16} = 0 \\ -p_9p_{11} + p_4p_{15} = 0 & -p_{13}p_{15} + p_{10}p_{16} = 0 \\ -p_6p_9 + p_3p_{12} = 0 & -p_{12}p_{14} + p_7p_{16} = 0 \end{array} \right. \quad (7)$$

Actually these simpler equations encode the conditional independence properties of the model, as pointed out in Sturmfels (2002), and later systematically exploited by B. Sturmfels and co-workers.

All other quantities mentioned in the previous paragraph can be computed. In particular, the generator of the semi-group in (T7) can be computed with the software **2ti2**, <http://www.4ti2/de>.

4. Conclusions

The precise interplay of the exponential (log-linear) model with its “border” limit is a difficult problem, which is especially relevant when discussing such items as the existence of maximum likelihood estimation, identifiability, loss of degrees of freedom, model choice. We believe that a number of algebraic notions, together with the relevant symbolic software can be of some help. Especially on the notion of union of exponential models, see Geiger *et al.* (2001).

References

- Diaconis P. and Sturmfels B. (1998) Algebraic algorithms for sampling from conditional distributions, *Ann. Statist.*, 26, 1, 363–397.
- Geiger D., Heckerman D., King H. and Meek C. (2001) Stratified exponential families: graphical models and model selection, *The Annals of Statistics*, 29.
- Lauritzen S.L. (1996) *Graphical models*, The Clarendon Press Oxford University Press, New York, Oxford Science Publications.
- Pistone G., Riccomagno E. and Wynn H.P. (2001) *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman&Hall.
- Pistone G. and Wynn H.P. (2003) Statistical toric models, Lecture Notes. Grostat VI Menton.
- Rapallo F. (2003) *Log-linear models and toric ideals*, Ph.D. thesis, Dottorato in Matematica Università Genova, Università di Torino, Politecnico di Torino.
- Rapallo F. (2004) Toric statistical models: Parametric and binomial representations, Preprint 498, Dipartimento di Matematica, Università di Genova.
- Sturmfels B. (1996) *Gröbner bases and convex polytopes*, American Mathematical Society, Providence, RI.
- Sturmfels B. (2002) *Solving Systems of Polynomial Equations*, number 97 in CBMS Regional Conference Series in Mathematics, American Mathematical Society, Providence, RI.