

Markov bases algorithm for ML in contingency tables

*Algoritmo con basi di Markov per la stima di massima verosimiglianza
nelle tabelle di contingenza*

Giovanni Pistone
Dipartimento di Matematica
Politecnico di Torino
pistone@calvino.polito.it

Fabio Rapallo¹
Dipartimento di Matematica
Università di Genova
rapallo@dima.unige.it

Riassunto: In questo lavoro presentiamo un metodo MCMC per calcolare una approssimazione Monte Carlo della stima di massima verosimiglianza di modelli log-lineari. Questo metodo, alternativo ai classici metodi numerici, è basato sul fatto che la stima di massima verosimiglianza è la media di una appropriata distribuzione ipergeometrica. La catena di Markov è definita utilizzando l'algoritmo algebrico descritto in Diaconis e Sturmfels (1998).

Keywords: log-linear models; maximum likelihood; MCMC methods; Markov bases.

1. Introduction

The problem of the maximum likelihood (ML) estimation of log-linear models is an important topic in the analysis of discrete data and, especially, in the analysis of contingency tables. A classical reference for the ML estimation theory of log-linear models is Haberman (1974), where a review on the theorems of existence and uniqueness of ML estimates of log-linear models can be found.

Consider a statistical model on a finite sample space $\mathcal{X}$. Although in contingency tables the sample space is usually a Cartesian product of the form $\{(i, j) \mid i = 1, \dots, I, j = 1, \dots, J\}$, we assume here, without loss of generality, $\mathcal{X} = \{1, \dots, k\}$. The model is characterized by the parameters $p_i = \mathbb{P}[i]$ with $i = 1, \dots, k$. In this paper we use the vector notation, that is $\mathbf{p}$ is the vector $(p_1, \dots, p_k)$. The parameter space is given by $p_i \geq 0, i = 1, \dots, k$ and $\sum_{i=1}^k p_i = 1$.

In the log-linear models for contingency tables, the likelihood equations can be classified in two types:

- (a) The system of equations has a closed form solution $\hat{\mathbf{p}}$ which is actually computable, so that no approximation is needed. One example of this situation is the independence model in two-way tables.
- (b) A computable solution of the system of equations is not known, so that a numerical method is needed in order to get an approximate solution. This is the case, for example, of the quasi-independence model in two-way tables.

In the case (b) the most widely used methods for approximating the ML estimate are the Newton–Raphson one and the Iterative Proportional Fitting one. For details about these methods, see (Haberman, 1974, Chapter 3). The Newton–Raphson method is used

¹Il lavoro è stato svolto con i contributi Ricerca di Base 2001 dei rispettivi Dipartimenti

in PROC CATMOD of SAS software, see SAS/STAT User's Guide (2000) for a concise discussion of the computations and for further references on the relevant numerics.

Recently, some alternative methods have been proposed for particular models. For example in Rudas (1998) a method for the computation of the ML estimate of graphical models for multi-way contingency tables can be found. His algorithm is based on the successive application of the independence and conditional independence constraints, but it does not apply to general log-linear models.

We will show in the next Section that the ML estimate $\hat{\mathbf{p}}$ of the parameters p_i 's is the mean of a probability distribution derived from the data. In the literature, it is usual to use Monte Carlo methods, and in particular MCMC methods, in order to approximate some parameters of interest of a probability distribution. These methods, based on simulation techniques, are widely used in statistical applications. In particular we recall here the Metropolis–Hastings algorithm, see Chib and Greenberg (1995) for a review of the algorithm and some practical examples.

In the past few years, the application of new algebraic non-linear techniques to Statistics have been presented. The first work on this topic is Diaconis and Sturmfels (1998). For a more general reference see Pistone *et al.* (2001a). In particular the work of Diaconis and Sturmfels (1998) leads to the possibility of sampling from conditional distributions using the key notion of Markov basis, finding a general method for defining a connected, aperiodic and reversible Markov chain on the appropriate sample space.

2. ML estimation for contingency tables

Let us write the probability model in the form $\mathbb{P}[x] = \varphi(T(x))$, $x \in \mathcal{X}$, where $T : \mathcal{X} \longrightarrow \mathbb{N}^s$ is the sufficient statistic. Let $\{A_1, \dots, A_s\}$ be a covering of $\mathcal{X}$, possibly overlapping. The probability models of interest in the analysis of contingency tables have the sufficient statistic of the form

$$T : \mathcal{X} \longrightarrow \mathbb{N}^s$$

$$x \longmapsto (\mathbb{I}_{A_1}(x), \dots, \mathbb{I}_{A_s}(x))$$

where $\mathbb{I}_A$ is the indicator function of the set A . Let $\mathbf{x} = (x_1, \dots, x_N)$ be a sample of size N drawn from a vector $\mathbf{X} = (X_1, \dots, X_N)$ of independent and identically $\mathbb{P}$ -distributed random variables with values in $\mathcal{X}$ and let us denote by $\mathbb{P}^{*N} = \psi(T_N)$ the distribution of the sample of size N , where T_N is the corresponding sufficient statistic. We assume that the sufficient statistic for the sample of size N verifies

$$T_N(\mathbf{X}) = \sum_{j=1}^N T(X_j), \quad (1)$$

that is, the sufficient statistic of $\mathbf{X}$ is the sum of the sufficient statistics of the unidimensional random variables X_j , $j = 1, \dots, N$. Necessary and sufficient conditions for such additivity property will be discussed in a joint paper currently in progress.

We denote by $\mathcal{Y}_{\mathbf{t}}$ the set of all samples with fixed value $\mathbf{t}$ of the sufficient statistic T_N and by $\mathcal{F}_{\mathbf{t}}$ the set of all frequency tables obtained from samples with value $\mathbf{t}$ of the sufficient statistic T_N . Formally

$$\mathcal{Y}_{\mathbf{t}} = \left\{ (x_1, \dots, x_N) \left| \sum_{j=1}^N T(x_j) = \mathbf{t} \right. \right\} \quad (2)$$

and

$$\mathcal{F}_{\mathbf{t}} = \left\{ f : \mathcal{X} \longrightarrow \mathbb{N} \mid \sum_{x \in \mathcal{X}} f(x)T(x) = \mathbf{t} \right\}. \quad (3)$$

In the following proposition we state some basic properties of these models.

Proposition 1 *The distribution of $\mathbf{X}$ given $\{T_N = \mathbf{t}\}$ is uniform on $\mathcal{Y}_{\mathbf{t}}$; moreover, the image probability of $\mathbb{P}^{*N}[\cdot \mid T_N = \mathbf{t}]$ on $\mathcal{F}_{\mathbf{t}}$ is*

$$\mathcal{H}_{\mathbf{t}}(f) = \mathbb{P}^{*N}[F^{-1}(f) \mid T_N = \mathbf{t}] = \frac{\#\{\mathbf{x} \mid F(\mathbf{x}) = f\}}{\#\mathcal{Y}_{\mathbf{t}}},$$

where $F : \mathcal{Y}_{\mathbf{t}} \longrightarrow \mathcal{F}_{\mathbf{t}}$ is the map which associates to every sample the corresponding table.

For further details and for the proof of these properties, see Diaconis and Sturmfels (1998). The distribution $\mathcal{H}_{\mathbf{t}}$ in the above proposition is the multivariate hypergeometric distribution on $\mathcal{F}_{\mathbf{t}}$ with the parameters depending on the value of the sufficient statistic T_N . From this we derive our main results as follows.

Theorem 2 *The ML estimate of the vector of parameters $N\mathbf{p}$ is the mean of the appropriate hypergeometric distribution.*

The detailed proof is not included here. It follows by direct computation on the hypergeometric distribution. It is actually a generalization of a result on the expectation parameters of the exponential models.

3. Algebraic method

As the set $\mathcal{F}_{\mathbf{t}}$ is usually very large, we apply an improvement of the algorithm proposed by Diaconis and Sturmfels (1998) in order to approximate the ML estimate of the parameters.

Given $\mathbf{t}$, we define a Markov chain with sample space $\mathcal{F}_{\mathbf{t}}$ and with stationary distribution $\mathcal{H}_{\mathbf{t}}$. The Markov chain is a random walk on $\mathcal{F}_{\mathbf{t}}$ with moves in a set $\mathcal{M} = \{m_1, \dots, m_L\}$: in order to obtain the connectedness of the Markov chain the set of moves must be a *Markov basis*. A Markov basis $\mathcal{M}$ is a set of tables, with possibly negative integer entries, such that the value of the sufficient statistic is $\mathbf{0}$ and with the following property: for every pair of tables f and f' in $\mathcal{F}_{\mathbf{t}}$ there is a path between f and f' with moves in $\mathcal{M}$ and through nonnegative tables. The Markov chain is simulated with the following algorithm:

- (a) at the time 0 the chain is in f ;
- (b) choose a move m uniformly in the Markov basis and $\epsilon = \pm 1$ with probability $1/2$ each independently of m ;
- (c) if $f + \epsilon m \geq \mathbf{0}$ then move the chain from f to $f + \epsilon m$ with probability

$$\min\{\mathcal{H}_{\mathbf{t}}(f + \epsilon m)/\mathcal{H}_{\mathbf{t}}(f), 1\}.$$

In all other cases, stay at f .

The proof of the validity of the basic algorithm can be found in Diaconis and Sturmfels (1998), while a detailed discussion and improvements are presented in Rapallo (2001). We emphasize that for the computation of the Markov basis the vector space theory is not sufficient, as pointed out in Rapallo (2001), where there is a simple counterexample. The computation of the Markov basis is performed through algebraic techniques based on the theory of toric ideals. A reference from the algebraic point of view is Kreuzer and Robbiano (2000), while the relevance of toric ideals in statistics is discussed in Pistone *et al.* (2001b).

The computation of the Markov basis is possible using symbolic software, such as CoCoA Capani *et al.* (2000) or Maple. For the explicit computation of the Markov basis for the most commonly used log-linear models, see Rapallo (2001). The algorithm for some log-linear models are implemented and freely available at the web page <http://www.mathprize.it>. Future work will concern numerical comparisons between this algebraic-based method and the classical ones.

Acknowledgments The present work is part of the PhD research program of F. Rapallo, Professor G. Pistone acting as supervisor. Joint work on related topics, and in particular about the algebraic techniques for ML estimation in discrete models, is in progress. While all the matter has been object of extensive joint discussion, the present redaction is fully due to F. Rapallo. Both authors gratefully acknowledge the friendly help by Professor L. Robbiano.

References

- Capani A., Niesi G. and Robbiano L. (2000) *CoCoA, a system for doing Computations in Commutative Algebra*, Available via anonymous ftp from cocoa.dima.unige.it, 4.0 edition.
- Chib S. and Greenberg E. (1995) Understanding the Metropolis–Hastings algorithm, *The American Statistician*, 49, 4, 327–335.
- Diaconis P. and Sturmfels B. (1998) Algebraic algorithms for sampling from conditional distributions, *Ann. Statist.*, 26, 1, 363–397.
- Haberman S.J. (1974) *The Analysis of Frequency Data*, The University of Chicago Press, Chicago and London.
- Kreuzer M. and Robbiano L. (2000) *Computational Commutative Algebra 1*, Springer, New York.
- Pistone G., Riccomagno E. and Wynn H.P. (2001a) *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman&Hall, New York.
- Pistone G., Riccomagno E. and Wynn H.P. (2001b) Computational commutative algebra in discrete statistics, in: *Algebraic Methods in Statistics and Probability*, Viana M.A.G. and Richards D.S.P., eds., American Mathematical Society, volume 287 of *Contemporary Mathematics*, 267–282.
- Rapallo F. (2001) Algebraic Markov bases and MCMC for two-way contingency tables, grostat V workshop, New Orleans, September 2001. Submitted.
- Rudas T. (1998) A new algorithm for the maximum likelihood estimation of graphical log-linear models, *Computational Statistics*, 13, 529–537.
- SAS/STAT User’s Guide (2000) *Version 8*, Cary, NC: SAS Institute, 1 edition.