ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Maximum Likelihood Estimation and Polynomial System Solving

Kim Batselier Philippe Dreesen Bart De Moor *

Department of Electrical Engineering (ESAT), SCD, Katholieke Universiteit Leuven
/IBBT-K.U.Leuven Future Health Department
3001 Leuven, Belgium, Email: {kim.batselier,bart.demoor}@esat.kuleuven.be

Abstract. This article presents an alternative method to find the global maximum likelihood estimates of the mixing probabilities of a mixture of multinomial distributions. For these mixture models it is shown that the maximum likelihood estimates of the mixing probabilities correspond with the roots of a multivariate polynomial system. A new algorithm, set in a linear algebra framework, is presented which allows to find all these roots by solving a generalized eigenvalue problem.

1 Introduction

The term maximum likelihood was first coined by Fisher in 1922 [1, 2]. Since then, the use of maximum likelihood estimation has become extremely popular in a vast number of fields. The two most common methods for finding maximum likelihood estimates are Expectation Maximization (EM) [3] and Markov Chain Monte Carlo (MCMC) [4, 5]. EM is an iterative hill climbing algorithm. Starting from some initial guess, model parameters are updated consecutively such that the likelihood increases until convergence has occurred. This dependence of the solution on the initial guess means that for the case of many solutions only a local maximum is obtained. MCMC methods are typically used in a Bayesian Learning setting where one is usually more interested in posterior distributions than in point estimates. These methods allow to generate samples from unknown distributions which then can be used to calculate point estimates (such as the mode or mean). Although this method is commonly used to sample the posterior distribution it can be also utilized to obtain maximum likelihood estimates [6]. A more recent method has come from the field of algebraic statistics which seeks to mix algebraic geometry and commutative algebra with statistics [7]. This method relies on Buchberger's algorithm which is a symbolic method

*Kim Batselier is a research assistant at the Katholieke Universiteit Leuven, Belgium. Philippe Dreesen is supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare Belgian Federal Science Policy Office: IUAP.

however and therefore has inherent difficulties when dealing with real numbers. This article presents a numerical method for finding maximum likelihood estimates which is guaranteed to find the global maximum. This is achieved by first showing that for a mixture of multinomial distributions finding the maximum likelihood estimates of the mixing probabilities corresponds with solving a multivariate polynomial system. Then an algorithm is presented which allows to find all solutions of polynomial systems by solving a generalized eigenvalue problem.

2 Maximum Likelihood and Multivariate Polynomial Systems

The models considered in this paper are mixtures of multinomial distributions. n will denote the number of distributions in the mixture and K the total number of possible outcomes in an experiment. Each nth multinomial distribution is characterized by K probabilities $p(k|i)$ with $i = 1 \dots n$ and $k = 1 \dots K$. These are assumed to be known. The probability of an observed outcome y_k is then given by

$$p_{y_k}(x) \;=\; x_1\, p(k|1) + \ldots + x_n\, p(k|n) \;=\; \sum_{i=1}^{n} x_i\, p(k|i) \tag{1}$$

where $x = (x_1, \ldots, x_n)$ are the unknown mixing probabilities. Data are typically given as a sequence of observations. The integer N denotes the sample size. When all observations are independent and identically distributed, the data can then be summarized in a data vector $u = (u_1, \ldots, u_K)$. Each possible outcome y_k is observed u_k times and therefore $u_1 + u_2 + \ldots + u_K = N$. We can now define the likelihood function.

Definition 2.1 *Given a mixture of n multinomial distributions and a sequence of N independent and identical distributed samples then the likelihood function $L(x)$ is given by*

$$L(x) = p_{y_1}(x)^{u_1} p_{y_2}(x)^{u_2} \ldots p_{y_K}(x)^{u_K} = \prod_{i=1}^{K} p_{y_i}(x)^{u_i}. \tag{2}$$

This function depends on the parameter vector x and data vector u and is hence called the likelihood function. Note that it is the assumption of independent and identical distributed observations that allows us to factorize the likelihood. Any reordering of the observations leads to the same data vector u and has therefore no effect. Multiplying probabilities leads to very small numbers which could lead on a computer to numerical underflow. By taking the logarithm of (2) the expression is reduced to

$$l(x) = \log L(x) = \sum_{i=1}^{K} u_i \log p_{y_i}(x)$$

which effectively transforms the product of probabilities into a sum. This takes care of the numerical underflow problem. The maximum log-likelihood estimate of x is the solution of the following optimization problem

$$\hat{x} = \operatorname*{argmax}_{x} l(x) \tag{3}$$

which is equivalent with maximizing $L(x)$ since the logarithm is a monotonic function. The optimization problem (3) is solved by taking the partial derivatives of $l(x)$ with respect to each x_i and equating these to zero. This results in the following system of n rational equations in n unknowns

$$\left\{\begin{array}{lclcl} \frac{\partial l(x)}{\partial x_1} & = & \sum_i \frac{u_i}{p_{y_i}} \frac{\partial p_{y_i}}{\partial x_1} & = & 0 \\ \vdots & & & & \\ \frac{\partial l(x)}{\partial x_n} & = & \sum_i \frac{u_i}{p_{y_i}} \frac{\partial p_{y_i}}{\partial x_n} & = & 0. \end{array}\right. \tag{4}$$

These are rational equations since each term contains a linear polynomial of the form (1) in the denominator. Therefore, in order to find the solutions of (4) all terms of each equation need to be put onto a common denominator. Then one needs to solve the polynomial system obtained from equating the nominators to zero. Note that the dependencies of p_k on x are dropped in the notation. A polynomial system like this typically has many solutions. Also note that it is in fact possible to eliminate 1 unknown. Using the relation $x_1 + \ldots + x_n = 1$ it is possible to reduce the number of equations and unknowns to $n - 1$. Now that it is established that finding the maximum likelihood estimates for the mixing probabilities corresponds with solving a multivariate polynomial system a new algorithm is introduced which guarantees to find all solutions (including the global optimum).

3 Solving Polynomial Systems as Eigenvalue Problems

An overview of the basic polynomial root-finding algorithm is given for the case that there are no roots with multiplicities and roots at infinity. More details for the case of multiplicities and roots at infinity can be found in [8, 9]. The main idea will be to generate a matrix $M(d)$ that contains all coefficients of the polynomial system and find its kernel (null space). This can be done using either the singular value decomposition (SVD) or rank-revealing QR decomposition. The computational complexity of this algorithm is therefore of the order $O(pq^2)$ where $M(d)$ is a $p \times q$ matrix.

Algorithm 3.1

```
Input: n n-variate polynomials $F = f_1, \dots, f_n$ of degrees $d_1, \dots, d_n$
Output: kernel $K$
$M \leftarrow$ coefficient matrix of $F$ up to degree $d = \sum_{i=1}^{n} d_i - n + 1$
$s \leftarrow$ nullity of $M$
$Z \leftarrow$ basis null space from $SVD(M)$ or $QR(M)$
$S_1 \leftarrow$ row selection matrix for $s$ linear independent rows of $Z$
$S_2 \leftarrow$ row selection matrix for shifted rows of $S_1\,Z$
$B \leftarrow S_1\,Z$
$A \leftarrow S_2\,Z$
$[V, D] \leftarrow$ solve eigenvalue problem $B\,V\,D = A\,V$
$K \leftarrow Z\,V$
```

As mentioned before, the first step in the algorithm is to construct the coefficient matrix of the polynomial system F. In order to explain how this coefficient matrix is made we first need to explain how multivariate polynomials are represented by their coefficient vectors. This is achieved by simply storing the coefficients of the polynomial into a row vector according to a certain monomial ordering. In principle any monomial ordering can be used. We refer to [10] for more details on monomial orderings. The following example illustrates this for a bivariate polynomial of degree 2.

Example 3.1 *The vector representation of* $2 + 3x_1 - 4x_2 + x_1x_2 - 7x_2^2$ *is*

$$\begin{array}{cccccc} 1 & x_1 & x_2 & x_1^2 & x_1x_2 & x_2^2 \\ \big(2 & 3 & -4 & 0 & 1 & -7\big). \end{array}$$

We can now define the coefficient matrix $M(d)$ of a multivariate polynomial system up to a degree d.

Definition 3.1 *Given a set of n-variate polynomials $f_1, \dots, f_n$, each of degree $d_i\,(i = 1, \dots, n)$ then the coefficient matrix of degree d, $M(d)$, is the matrix containing the coefficients of*

$$M(d) \;=\; \begin{pmatrix} f_1 \\ x_1 f_1 \\ \vdots \\ x_n^{d-d_1} f_1 \\ f_2 \\ x_1 f_2 \\ \vdots \\ x_n^{d-d_n} f_n \end{pmatrix} \tag{5}$$

where each polynomial f_i is multiplied with all monomials from degree 0 up to $d - d_i$ for all $i = 1, \dots, n$.

Note that the coefficient matrix not only contains the original polynomials $f_1, \ldots, f_n$ but also 'shifted' versions where we define a shift as a multiplication with a monomial. The dependence of this matrix on the degree d is of crucial importance, hence the notation $M(d)$. It can be shown [11, 12] that the degree $d = \sum_{i=1}^{n} d_i - n + 1$ provides an upper bound for the degree for which all the solutions of the polynomial system appear in the kernel of $M(d)$. This brings us to step 2 of Algorithm 3.1. The number of solutions of F are counted by the dimension of the kernel (nullity) of $M(d)$. For the case that there are no multiplicities and no solutions at infinity, this is then simply given by the Bezout bound $m_B = \prod_{i=1}^{n} d_i$. As described in more detail in [8, 9], steps 3 up to 9 find all these solutions from a generalized eigenvalue problem which is constructed from exploiting the structure of the canonical kernel. The canonical kernel K of $M(d)$ is a n-variate Vandermonde matrix. It consists of columns of monomials, ordered according to the chosen monomial ordering and evaluated in the roots of the polynomial system. This monomial structure allows to use a shift property which is reminiscent of realization theory. This shift property tells us that the multiplication of rows of the canonical kernel K with any monomial corresponds with a mapping to other rows of K. This can be written as the following matrix equation

$$S_1 K D = S_2 K \tag{6}$$

where S_1 and S_2 are row selection matrices and D a diagonal matrix which contains the shift monomial on the diagonal. S_1 will select the first m_B linear independent rows of K. The canonical kernel K is unfortunately unknown but a numerical basis Z for the kernel can be computed from either the SVD or QR decomposition. This basis Z is then related to K by means of a linear transform V, $K = ZV$. Writing the shift property (6) in terms of the numerical basis Z results in the following generalized eigenvalue problem

$$BVD = AV \tag{7}$$

where $B = S_1 Z$ and $A = S_2 Z$ are square nonsingular matrices. The eigenvalues D are then the shift monomial evaluated in the different roots of the polynomial system. The canonical kernel K is easily reconstructed from $K = ZV$. The monomial ordering used in Algorithm 3.1 is such that the first row of the canonical kernel corresponds with the monomial of degree 0. Therefore, after normalizing K such that its first row contains ones, all solutions can be read off from the corresponding first degree rows.

4 Conclusion

It was shown that finding the maximum likelihood estimates for the mixing probabilities of a mixture of multinomial distributions is equivalent with solving a multivariate polynomial system. A new algorithm was introduced that is guaranteed to find all solutions of such multivariate polynomial systems. Since the algorithm uses basic linear algebra tools it can be easily implemented using any numerical linear algebra software package (e.g. Lapack[13]). This method is not limited in any way to mixture models. In fact, as soon as the maximum likelihood (or log-likelihood) estimation is equivalent to solving a polynomial system this method can be employed. Likewise, the assumption of having independent and identical distributed observations is not strictly necessary. It only allows to factorize the likelihood and hence reduces the complexity of writing down the polynomial system. It would be interesting to further investigate for which other discrete statistical models maximum likelihood estimation is equivalent with multivariate polynomial system solving.

References

[1] J Aldrich. R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, 12(3):162–176, Aug 1997.

[2] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

[3] Ap Dempster, Nm Laird, and Db Rubin. Maximum Likelihood from Incomplete Data via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1):1–38, 1977.

[4] Wk Hastings. Monte-Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57(1):97–&, 1970.

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[6] Cj Geyer. Markov-Chain Monte-Carlo Maximum-Likelihood. In Keramidas, Em, editor, *Computing Science and Statistics*, pages 156–163, 1991. 23rd Symp on the Interface between Computing Science and Statistics - Critical Applications of Scientific Computing : Biology, Engineering, Medicine, Speech, Seattle, Wa, Apr 21-24, 1991.

[7] L. Pachter and B. Sturmfels, editors. *Algebraic Statistics for Computational Biology.* Cambridge University Press, August 2005.

[8] P. Dreesen, K. Batselier, and B. De Moor. Back to the roots: Polynomial system solving, linear algebra, systems theory. Technical Report 10-191, ESAT/SCD/SISTA, Katholieke Universiteit Leuven, 2011.

[9] K. Batselier, P. Dreesen, and B. De Moor. Global optimization and prediction error methods. Technical Report 11-199, ESAT/SCD/SISTA, Katholieke Universiteit Leuven, 2011.

[10] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, Varieties and Algorithms.* Springer-Verlag, third edition, 2007.

[11] F. S. Macaulay. On some formulae in elimination. *Proc. London Math. Soc.*, 35:3–27, 1902.

[12] M Giusti. Combinatorial Dimension Theory of Algebraic-Varieties. *Journal of Symbolic Computation*, 6(2-3):249–265, Oct-Dec 1988.

[13] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide.* Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.