On parametric and implicit algebraic descriptions of maximum entropy models

ON PARAMETRIC AND IMPLICIT ALGEBRAIC DESCRIPTIONS OF MAXIMUM ENTROPY MODELS

AMBEDKAR DUKKIPATI

Abstract. Main aim of this paper is to present some notions on how results from commutative algebra and algebraic geometry could be used in representation and computation of maximum and minimum entropy (ME) models. We show that various formulations of estimation of ME models can be transformed to solving systems of polynomial equations in cases where an integer valued sufficient statistic exists. We give an implicit description of ME-models by embedding them in algebraic varieties for which we give a Gröbner bases method to compute it.

1. Introduction

Algebra has always played an important role in statistics, a classical example being linear algebra. There are also many other instances of applying algebraic tools in statistics (*e.g.*, Diaconis, 1988; Viana and Richards, 2001). But, treating statistical models as algebraic objects, and thereby using tools of computational commutative algebra and algebraic geometry in the analysis of statistical models is very recent and has led to the still evolving field of *algebraic statistics*.

The use of computational algebra and algebraic geometry in statistics was initiated in the work of Diaconis and Sturmfels (1998) on exact hypothesis tests of conditional independence in contingency tables, and in the work of Pistone et al. (2001) in experimental design. The term 'Algebraic Statistics' was first coined in the monograph by Pistone et al. (2001) and appeared recently in the title of the book by Pachter and Sturmfels (2005).

To extract the underlying algebraic structures in discrete statistical models, algebraic statistics treats statistical models as algebraic varieties. (An algebraic variety is the set of all solutions to a system of polynomial equations.) Parametric statistical models are described in terms of a polynomial (or rational) mapping from a set of parameters to distributions. One can show that many statistical models, for example independence models, Bernoulli random variables etc. (see Pachter and Sturmfels, 2005) can be given this algebraic formulation, and such models are referred to as algebraic statistical models.

Information theory has a well established role (*cf.* Kullback, 1959; Csiszár and Shields, 2004), and within this line of research this paper attempts to treat maximum entropy models with the above mentioned algebraic formalisms.

We organize our paper as follows. In § 2 we present various formulations of estimation of ME-models in terms of solving a system of polynomial equations. We present

Key words and phrases. Gröbner Bases, Shannon Entropy, Toric Ideals.

This work is down when the author was at EURANDOM, Eindhoven and supported by the same.

basics of Gröbner bases theory in § 3 and in § 4 we present ME-models in an algebraic statistical framework. Finally, we present the implicit descriptions of ME-models in § 5 and we make concluding remarks in § 6.

2. Estimation of ME-models as Polynomial System Solving

2.1. **Maximum Entropy models.**

Let X be a discrete random variable taking finitely many values from the set $[m] = \{1, 2, \ldots m\}$. A probability distribution p of X is naturally represented as a vector $p = (p_1, \ldots, p_m) \in \mathbb{R}^m$ if we fix the order on $[m]$. The set of all probability density functions (pdfs) of X w.r.t counting measure on $[m]$ (we refer to such a pdf as probability mass functions or pmf) is called the probability simplex

$$\Delta_m = \{p = (p_1, \ldots, p_m) \in \mathbb{R}^m_{\geq 0} : \sum_{i=1}^{m} p_i = 1\} \ . \tag{2.1}$$

Suppose that the only information (observations) available about the pmf $p = (p_1, \ldots, p_m)$, of X is in the form of expected values of the functions $t_i : [m] \to \mathbb{R}$, $i = 1, \ldots, d$ (sufficient statistic). We therefore have

$$\sum_{j=1}^{m} t_i(j) p_j = T_i \ \ , i = 1, \ldots d \ , \tag{2.2}$$

where T_i, $i = 1, \ldots, d$, are assumed to be known. In an information theoretic approach to statistics, known as Jayens maximum entropy principle (Jaynes, 1957), one would choose the pmf $p \in \Delta_m$ that maximizes the Shannon entropy functional

$$S(p) = -\sum_{j=1}^{m} p_j \ln p_j \tag{2.3}$$

with respect to the constraints (2.2). The set

$$\left\{ p \in \Lambda_m : \sum_{j=1}^{m} t_i(j) p_j = T_i \ \ , i = 1, \ldots, d \right\} \ ,$$

if non-empty, is called a *linear family* of probability distributions. The corresponding Lagrangian can be written as

$$\Xi(p, \xi) \equiv S(p) - \xi_0 \left(\sum_{j=1}^{m} p_j - 1 \right) - \sum_{i=1}^{d} \xi_i \left(\sum_{j=1}^{m} t_i(j) p_j - T_i \right) \ . \tag{2.4}$$

Holding $\xi = (\xi_1, \ldots, \xi_d)$ fixed, the unconstrained maximum of Lagrangian $\Xi(p, \xi)$ over all $p \in \Delta_m$ is given by an exponential family

$$p_j(\xi) = Z^{-1}(\xi) \exp \left(-\sum_{i=1}^{d} \xi_i t_i(j) \right) \ , j = 1, \ldots, m, \tag{2.5}$$

where $Z(\xi)$ is a normalizing constant (or partition function) given by

$$Z(\xi) = \sum_{j=1}^{m} \exp\left(-\sum_{i=1}^{d} \xi_i t_i(j)\right) \ . \tag{2.6}$$

This model is an exponential family and is known as maximum entropy (ME) model.

One can show that the Lagrange parameters in ME-model (2.5) can be estimated, when the values of T_i, $i = 1, \ldots, d$ are available, by solving the set of partial differential equations (Jaynes, 1968)

$$\frac{\partial}{\partial \xi_i} \ln Z(\xi) = T_i \ \ , i = 1, \ldots, d, \tag{2.7}$$

which has no explicit analytical solution. In this case one could employ Newton-Raphson procedures. One of the important methods is Darroch and Ratcliff's generalized iterative scaling algorithm (Darroch and Ratcliff, 1972), which has a geometric interpretation in information theoretic statistics (Csiszár, 1989).

Here one can show that, by a simple transformation of coordinates and imposing certain constraints on feature functions one could transform this problem into solving polynomial equations.

Proposition 2.1. *Estimation of the maximum entropy model (2.5), given the information in the form of (2.2) amounts to solving a set of polynomial equations provided that the sufficient statistic t_i, $i = 1, \ldots, d$, is nonnegative and integer valued.*

Proof. Set $\xi_i = -\ln \theta_i$, $i = 1, \ldots, d$. Now, (2.5) gives us

$$p_j = Z^{-1}(\theta) \exp\left(\sum_{i=1}^{d} t_i(j) \ln \theta_i\right) = Z^{-1}(\theta) \prod_{i=1}^{d} \theta_i^{t_i(j)} \ \ , \tag{2.8}$$

where

$$Z(\theta) = \sum_{j=1}^{m} \prod_{i=1}^{d} \theta_i^{t_i(j)} \ \ . \tag{2.9}$$

By substituting these maximum entropy distributions into (2.2) we get

$$\sum_{j=1}^{m} t_\ell(j) \prod_{i=1}^{d} \theta_i^{t_i(j)} = T_\ell Z(\theta) \ \ , \quad \ell = 1, \ldots, d. \tag{2.10}$$

This can be written as

$$\sum_{j=1}^{m} (t_\ell(j) - T_\ell) \prod_{i=1}^{d} \theta_i^{t_i(j)} = 0 \ \ , \quad \ell = 1, \ldots, d, \tag{2.11}$$

which is a system of polynomial equations in indeterminates $\theta_1, \ldots, \theta_d$ since t_i, $i = 1, \ldots, d$, are nonnegative integer valued functions. The solution of system of polynomial equations (2.11) gives the maximum entropy model specified by the available information (2.2). □

Later in this paper, we show that one could alternatively give an 'implicit' representation of ME-models (polynomial equations involving only $p_1, \ldots, p_m$). Before we study algebraic formalisms for representation of ME-models, we study applicability of two other cases of ME estimation, namely, the dual method and Kullback-Csiszár's iterative methods.

2.2. **Dual Method.** By using the Karush-Kuhn-Tucker theorem one can calculate the Lagrange parameters ξ_i, $i = 1, \ldots, d$, in (2.5) by optimizing the dual of $\Xi(p, \xi)$. That is, the task is to find ξ that maximizes

$$\Psi(\xi) \equiv \Xi(p, \xi) \ . \tag{2.12}$$

Note that $\Psi(\xi)$ is nothing but the entropy of ME-distribution (2.5), and we get

$$\Psi(\xi) = \ln Z + \sum_{i=1}^{d} \xi_i T_i \ . \tag{2.13}$$

This can be written as

$$\begin{aligned} \Psi(\xi) &= \ln \sum_{j=1}^{m} \exp\left(-\sum_{j=1}^{d} \xi_i t_i(j) \right) + \sum_{i=1}^{d} \xi_i T_i \\ &= \ln \sum_{j=1}^{m} \exp\left(\sum_{i=1}^{d} \xi_i (T_i - t_i(j)) \right) \ . \end{aligned} \tag{2.14}$$

Note that maximizing $\Psi(\xi)$ is equivalent to maximizing

$$\hat{\Psi}(\xi) = \sum_{j=1}^{m} \exp\left(\sum_{i=1}^{d} \xi_i (T_i - t_i(j)) \right) \ . \tag{2.15}$$

By introducing $\xi_i = -\ln \theta_i$, $i = 1, \ldots, d$, we maximize (for convenience we use same notation $\hat{\Psi}$)

$$\hat{\Psi}(\theta) = \sum_{j=1}^{m} \prod_{i=1}^{d} \theta_i^{t_i(j) - T_i} \ . \tag{2.16}$$

The solution is given by solving the following set of equations

$$\frac{\partial \hat{\Psi}}{\partial \theta_j} = 0 \ , j = 1, \ldots d. \tag{2.17}$$

Unfortunately, this does not give rise to solving polynomial equations. With this observation, we now consider the case where the expected values are available as sample means.

In most practical problems the information in the form of expected values is available via sample or empirical means. That is, given a sequence of observations $X_1, \ldots, X_N$ (i.i.d. random variables) the sample means $\widetilde{T}_i$, $i = 1, \ldots, d$, with respect to the functions t_i, $i = 1, \ldots, d$ are given by

$$\widetilde{T}_i = \frac{1}{N} \sum_{\ell=1}^{N} t_i(X_\ell), i = 1, \ldots, d, \tag{2.18}$$

and the underlying hypothesis is that $T_i \approx \widetilde{T}_i$. That is

$$\sum_{j=1}^{m} p_j t_i(j) \approx \frac{1}{N} \sum_{\ell=1}^{N} t_i(X_\ell) \ \ , i = 1, \ldots, d. \tag{2.19}$$

We show that, by choosing alternate Lagrangian in the place of (2.4) we can transform the parameter estimation of ME-model to a problem of solving a set of polynomial (Laurent) equations. (Laurent polynomial is a polynomial where exponents can be negative; see § 3).

Proposition 2.2. *Given the hypothesis (2.19), the problem of estimating the ME-model (2.5) with respect to (2.2) in the dual method amounts to solving a set of Laurent polynomial equations (assuming that the sufficient statistic is nonnegative and integer valued).*

Proof. To retain the integer valued exponents in our final solution, by the hypothesis (2.19) we consider the constrains in the form

$$N \sum_{j=1}^{m} t_i(j) p_j = \sigma_i \ \ , \quad i = 1, \ldots d \ \ , \tag{2.20}$$

where $\sigma_i = \sum_{l=1}^{N} t_i(O_l)$ denotes the integer valued sample sum. In this case Lagrangian may be written as

$$\widetilde{\Xi}(p, \widetilde{\xi}) \equiv S(p) - \widetilde{\xi}_0 \left(\sum_{j=1}^{m} p_j - 1 \right) - \sum_{i=1}^{d} \widetilde{\xi}_d \left(N \sum_{j=1}^{m} p_j t_i(j) - \sigma_i \right) \quad . \tag{2.21}$$

This results in the ME-distribution

$$p_j(\widetilde{\xi}) = \widetilde{Z}(\widetilde{\xi})^{-1} \exp \left(-N \sum_{i=1}^{d} \widetilde{\xi}_i t_i(j) \right) \ \ , \quad j = 1, \ldots, m, \tag{2.22}$$

where $\widetilde{Z}(\widetilde{\xi})$ is the normalizing constant given by

$$\widetilde{Z}(\widetilde{\xi}) = \sum_{j=1}^{m} \exp \left(-N \sum_{i=1}^{d} \widetilde{\xi}_i t_i(j) \right) \ \ . \tag{2.23}$$

To calculate the parameters we maximize the dual $\widetilde{\Psi}(\widetilde{\xi})$ of $\widetilde{\Xi}(p, \widetilde{\xi})$. That is, we maximize the functional

$$\widetilde{\Psi}(\widetilde{\xi}) = \ln \widetilde{Z} + \sum_{i=1}^{d} \widetilde{\xi}_i \sigma_i \ \ . \tag{2.24}$$

This is equivalent to optimizing the functional

$$\hat{\widetilde{\Psi}}(\widetilde{\xi}) = \sum_{j=1}^{m} \exp \left(\sum_{i=1}^{d} \widetilde{\xi}_i \sigma_i - N \sum_{i=1}^{d} \widetilde{\xi}_i t_i(j) \right) \ \ .$$

By setting $\ln \widetilde{\theta}_i = \widetilde{\xi}_i$ we have

$$\hat{\widetilde{\Psi}}(\widetilde{\theta}) = \sum_{j=1}^{m} \prod_{i=1}^{d} \widetilde{\theta}_i^{(\sigma_i - N t_i(j))} \quad . \tag{2.25}$$

The solution is given by solving the following set of equations

$$\frac{\partial \hat{\widetilde{\Psi}}}{\partial \widetilde{\theta}_i} = 0 \quad , i = 1, \ldots d. \tag{2.26}$$

Note that $\frac{\partial \hat{\widetilde{\Psi}}}{\partial \widetilde{\theta}_i}$ is a Laurent polynomial and hence the result is implied. □

We mention here that Gröbner bases methods have been extended to the case of Laurent polynomials (Pauer and Unterkircher, 1999); we will not address this issue in this paper.

2.3. **Estimation by the Minimum I-Divergence principle.** The Minimum I-divergence principle is a generalization of the maximum entropy principle that considers the cases where a prior estimate of the distribution p is available. Given a prior estimate $r \in \Delta_m$ and information in the form of (2.2), one would choose the pdf $p \in \Delta_m$ that minimizes the Kullback-Leibler divergence

$$I(p\|r) = \sum_{j=1}^{m} p_j \ln \frac{p_j}{r_j} \tag{2.27}$$

with respect to the constraints (2.2). The corresponding minimum entropy distributions are in the form of

$$p_j(\xi) = Z(\xi)^{-1} r_j \exp\left(- \sum_{i=1}^{d} \xi_i t_i(j) \right) \quad , j = 1, \ldots, m, \tag{2.28}$$

where $Z(\xi)$ is normalizing constant given by

$$Z(\xi) = \sum_{j=1}^{m} r_j \exp\left(- \sum_{i=1}^{d} \xi_i t_i(j) \right) \quad . \tag{2.29}$$

It is easy to see that estimating minimum entropy distributions can be translated to solving polynomial equations, when the feature functions are nonnegative and integer valued. Polynomial system one would solve in this case is

$$\sum_{j=1}^{m} r_j (t_\ell(j) - T_\ell) \prod_{i=1}^{d} \theta_i^{t_i(j)} = 0 \quad , \quad \ell = 1, \ldots, d. \tag{2.30}$$

Hence we have the following proposition.

Proposition 2.3. *The estimation of minimum entropy model (2.28) given the information in the form of (2.2) amounts to solving a set of polynomial equations provided that the sufficient statistic t_i, $i = 1, \ldots, d$ are nonnegative and integer valued.*

Estimating ME-distributions involves solving a system of nonlinear equations, simultaneously, which becomes computationally inefficient. Here, one would employ an iterative method where one would estimate the distribution considering only one constraint at a time. We describe this procedure as follows.

At the N^{th} iteration, the algorithm computes the distribution $p^{(N)}$ that minimizes $I(p^{(N)} \| p^{(N-1)})$ with respect the i^{th} constraint, if $N = ad+i$ $(1 \leq i \leq d)$ for any positive integer a. This iteration will ultimately converge to the maximum entropy distribution, which is demonstrated for discrete distributions by Ireland and Kullback (1968), and for continuous distributions by Kullback (1968). A general and rigorous treatment of convergence, existence, and uniqueness analysis is given by Csiszár (1975). We refer to this iteration as Kullback-Csiszár iteration.

In this iterative procedure, we have $p^{(0)} = r$ and $p^{(1)}$ is given by

$$p_j^{(1)} = r_j \big(Z^{(1)}\big)^{-1} \zeta_1^{t_1(j)},$$

where $\big(Z^{(1)}\big)^{-1} = \sum_{j=1}^{m} r_j \zeta_1^{t_1(j)}$. The first constraint in (2.2) can be estimated by solving polynomial equation

$$\sum_{j=1}^{m} r_j (t_1(j) - T_1) \zeta_1^{t_1(j)} = 0 \ , \tag{2.31}$$

with indeterminate ζ_1. Similarly we have

$$p_j^{(2)} = r_j \big(Z^{(1)}\big)^{-1} \big(Z^{(2)}\big)^{-1} \zeta_1^{t_1(j)} \zeta_2^{t_2(j)} \ ,$$

where $\big(Z^{(2)}\big)^{-1} = \sum_{j=1}^{m} \zeta_2^{t_1(j)}$. Considering the first two constrains in (2.2) the ME distribution can be estimated by solving

$$\sum_{j=1}^{m} r_j (t_2(j) - T_2) \zeta_1^{t_1(j)} \zeta_2^{t_2(j)} = 0 \ , \tag{2.32}$$

along with (2.31).

In general, when $N = ad + i$ for some positive integer a, $p_j^{(N)}$, for $N = 1, 2 \ldots$ is given by

$$p_j^{(N)} = r_j \big(Z^{(1)}\big)^{-1} \ldots \big(Z^{(N)}\big)^{-1} \zeta_1^{t_1(j)} \ldots \zeta_N^{t_i(j)}$$

and is determined by the following system of polynomial equations

$$\left.\begin{aligned} \sum_{j=1}^{m} r_j (t_1(j) - T_1) \zeta_1^{t_1(j)} &= 0 \ , \\ \sum_{j=1}^{m} r_j (t_2(j) - T_2) \zeta_1^{t_1(j)} \zeta_2^{t_2(j)} &= 0 \ , \\ &\vdots \\ \sum_{j=1}^{m} r_j (t_i(j) - T_i) \zeta_1^{t_1(j)} \zeta_2^{t_2(j)} \ldots \zeta_N^{t_i(j)} &= 0 \ . \end{aligned}\right\} \tag{2.33}$$

Note that this system has a triangular structure which makes it easy to solve. Also one could observe that the technique of Kullback-Csiszár iteration reflects in the system of polynomial equations which determines the ME distribution.

3. Gröbner bases Fundamentals

Gröbner bases were introduced by Buchberger (1965) as a computational tool for testing solvability of polynomial equations. It serves as a general algorithmic solution of some fundamental problems in commutative algebra (polynomial ideal theory and algebraic geometry). Since then the theory of Gröbner bases has held an important place in mathematical research. In this section we give a brief introduction to Gröbner bases.

Basic problem of *algebraic geometry* is to understand the set of points $(a_1, \ldots, a_n) \in k^n$ satisfying a system of polynomial equations $f_1(x_1, \ldots, x_n) = 0, \ldots, f_s(x_1, \ldots, x_n) = 0$ where $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$. Here, and throughout this paper, k represents a field (*e.g.*, $\mathbb{R}$, $\mathbb{C}$). The algebraic closure of the field k is represented by $\bar{k}$ (the algebraic closure of $\mathbb{R}$ is $\mathbb{C}$). $k[x_1, \ldots, x_n]$ denotes the set of all polynomials in indeterminates $x_1, \ldots, x_n$. From now on k represents the field $\mathbb{R}$ and $\bar{k}$ represents $\mathbb{C}$. Though we work with $\mathbb{R}$, we use k for a field.

A monomial in indeterminates $x_1, \ldots, x_n$ is a power product of the form $x_1^{\alpha_1} \ldots x_n^{\alpha_n}$, where all the exponents are nonnegative integers, i.e. $\alpha_i \in \mathbb{Z}_{\geq 0}$, $i = 1, \ldots n$. One can simplify the notation for monomial as follows: denote $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}^n_{\geq 0}$ and by using multi-index notation we set

$$x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$$

with the understanding that $x = (x_1, \ldots, x_n)$. Note that $x^\alpha = 1$ whenever $\alpha = (0, \ldots, 0)$. Once the order of the indeterminates is fixed, the monomial $x_1^{\alpha_1} \ldots x_n^{\alpha_n} = x^\alpha$ is identified by $(\alpha_1, \ldots, \alpha_n)$. Hence, the set of all monomials in indeterminates $x_1, \ldots, x_n$ can be represented by $\mathbb{Z}^n_{\geq 0}$. A polynomial f in inderterminates $x_1, \ldots, x_n$ with coefficients in k is a finite linear combination of monomials and can be written in the form of

$$f = \sum_{\alpha \in \Lambda_f} a_\alpha x^\alpha \ ,$$

where $\Lambda_f \subset \mathbb{Z}^n_{\geq 0}$ is a finite set and $a_\alpha \in k$. The collection of all such polynomials $k[x_1, \ldots, x_n]$ has the structure not only of a vector space but also of a ring. Indeed the ring structure of $k[x_1, \ldots, x_n]$ plays a main role in computational algebra and algebraic geometry. The ring $k[x_1, \ldots, x_n]$ is called the ring of polynomials in n indeterminates. If we allow negative exponents, *i.e.*, the polynomial of the form $f = \sum_{\alpha \in \Lambda_f} a_\alpha x^\alpha$ where $\alpha \in \mathbb{Z}^n$, it is known as Laurent polynomial ($\Lambda_f \subset \mathbb{Z}^n_{\geq 0}$ is finite). The set of all Laurent polynomials in the indeterminates $x_1, \ldots, x_n$ is denoted by $k[x_1, \ldots, x_n, x_1^{-1}, \ldots, x_n^{-1}]$ or $k[x_1^{\pm}, \ldots, x_n^{\pm}]$ and it has the structure of a ring.

A subset $\mathfrak{a} \subset k[x_1, \ldots, x_n]$ is said to be an ideal if it satisfies: (i) $0 \in \mathfrak{a}$ (ii) $f, g \in \mathfrak{a}$, then $f + g \in \mathfrak{a}$ (iii) $f \in \mathfrak{a}$ and $h \in k[x_1, \ldots, x_n]$ then $hf \in \mathfrak{a}$. A set $V \subset k^n$ is said to be an algebraic variety if there exist $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$ such that

$$V = \{(c_1, \ldots c_n) \in k^n : f_i(c_1, \ldots c_n) = 0, 1 \leq i \leq s\} \ .$$

We use the notation $\mathcal{V}(f_1, \ldots, f_s) = V$ to represent the varieties. In this case V is uniquely determined by the ideal generated by $f_1, \ldots, f_s$. This ideal is denoted by $\langle f_1, \ldots, f_s \rangle$ and hence we have

$$\mathcal{V}(f_1, \ldots, f_s) = \mathcal{V}(\langle f_1, \ldots, f_s \rangle) \ .$$

Further,

$$\langle f_1, \ldots, f_s \rangle = \langle g_1, \ldots, g_t \rangle \Longrightarrow \mathcal{V}(f_1, \ldots, f_s) = \mathcal{V}(g_1, \ldots, g_t) \ .$$

Though one can recover varieties from the ideals, the converse is not true. To see this we need the following notion of vanishing ideal of a variety. The vanishing ideal of a variety $E \subset k^n$ is defined as

$$\mathcal{I}(E) = \{f \in k[x_1, \ldots, x_n] : f(a) = 0, \forall a \in E\} \ ,$$

which is, indeed an ideal. Further, this definition can be extended to any arbitrary subset of k^n. Now, varieties can be uniquely identified by their corresponding vanishing ideals. That is

$$V = W \Longleftrightarrow \mathcal{I}(V) = \mathcal{I}(W) \ .$$

For any algebraic variety $V \subset k^n$ we have

$$V \subseteq \mathcal{V}(\mathcal{I}(V)) \ .$$

But when it comes to recovering ideals from varieties we have

$$\langle f_1, \ldots, f_s \rangle \subseteq \mathcal{I}(\mathcal{V}(f_1, \ldots, f_s)) \ ,$$

and by *Hilbert's Nullstellensatz* (Adams and Loustaunau, 1994, p. 62) given a variety, we can recover the ideal up to its radical only in the case of algebraically closed fields.

The basic idea behind Gröbner bases is the generalization of the division algorithm in a single variable case ($k[x]$) to the multivariate case ($k[x_1, \ldots, x_n]$). For this we need the notion of monomial order.

Definition 3.1. *A monomial order or term order on $k[x_1, \ldots, x_n]$ is a relation $\prec$ (we use $\succ$ for the corresponding 'greater than') on $\mathbb{Z}^n_{\geq 0}$ that satisfies following conditions (i) $\prec$ is a total (or linear) ordering on $\mathbb{Z}^n_{\geq 0}$, (ii) if $\alpha \prec \beta$, for $\alpha, \beta \in \mathbb{Z}^n_{\geq 0}$ then for any $\gamma \in \mathbb{Z}^n_{\geq 0}$ it holds $\alpha + \gamma \prec \beta + \gamma$, and (iii) $\prec$ is a well-ordering on $\mathbb{Z}^n_{\geq 0}$.*

Given such an ordering $\prec$, one can define the leading term of non-zero polynomial $f \in k[x_1, \ldots, x_n]$ as the term of f (the coefficient times its monomial) whose monomial is maximal for $\prec$. We denote this leading term by $\mathrm{LT}_\prec(f)$ and the corresponding monomial is denoted by $\mathrm{LM}_\prec(f)$.

Definition 3.2. *An ideal $\mathfrak{a} \subset k[x_1, \ldots, x_n]$ is said to be a monomial ideal if there is a set $A \subset \mathbb{Z}^n_{\geq 0}$, possibly infinite, such that $\mathfrak{a} = \langle x^\alpha : \alpha \in A \rangle$.*

Given any ideal $\mathfrak{a} \subset k[x_1, \ldots, x_n]$, the ideal defined as $\langle \mathrm{LM}_\prec f : f \in \mathfrak{a} \rangle$ is a monomial ideal and is denoted by $\mathrm{LM}_\prec(\mathfrak{a})$, which is known as leading monomial ideal of $\mathfrak{a}$. By Dickson's lemma (Cox et al., 1991, p. 69), the ideal $\mathrm{LT}_\prec(\mathfrak{a})$ is generated by a finite set of monomials. Dickson's lemma and the multivariate division algorithm leads to a proof of Hilbert bases theorem which states that every polynomial ideal can be finitely generated, which further lead to a definition of Gröbner basis (see Cox et al., 1991, § 2.5).

Definition 3.3. *Fix a monomial order $\prec$ on $k[x_1, \dots, x_n]$. A finite subset $G = \{g_1, \dots, g_t\}$ of an ideal $\mathfrak{a} \subset k[x_1, \dots, x_n]$ is a Gröbner basis if and only if*

$$\mathrm{LT}_{\prec}(\mathfrak{a}) = \langle \mathrm{LT}_{\prec}(g_1), \dots, \mathrm{LT}_{\prec}(g_t) \rangle.$$

Given a set of generators of an ideal, the Buchberger algorithm (Buchberger and Gröbner, 1985) can be used to compute a Gröbner basis of the ideal with respect to various term orders. The algorithm and its variants are implemented in most symbolic computation programs. Note that a Gröbner basis is not unique, but one can transform it to a reduced Gröbner basis which is unique for every ideal in $k[x_1, \dots, x_n]$. The Buchberger algorithm provides a common generalization of the Euclidean algorithm and the Gaussian elimination algorithm to multivariate polynomial rings.

One of the important results in Gröbner bases theory is the elimination theorem, which we will use in the next section, and here we present a brief discussion on it.

Definition 3.4. *Consider $k[x_1, \dots, x_n, y_1, \dots, y_m]$ a polynomial ring in indeterminates $x_1, \dots, x_n$, $y_1, \dots, y_m$. We refer to $\{x_1, \dots, x_n\}$ as x-variables and $\{y_1, \dots, y_m\}$ as y-variables. Let $\prec_x$ and $\prec_y$ be monomial orderings on x and y variables respectively. Define an ordering relation $\prec_{[\{x\} \succ \{y\}]}$ on $\mathbb{Z}_{\geq 0}^{n+m}$ (i.e set of all monomials in indeterminates $x_1, \dots, x_n, y_1, \dots, y_m$) as follows:*

$$x^{\alpha^{(1)}} y^{\beta^{(1)}} \prec_{[\{x\} \succ \{y\}]} x^{\alpha^{(2)}} y^{\beta^{(2)}} \Longleftrightarrow \begin{cases} \alpha^{(1)} \prec_x \alpha^{(2)} \\ \text{or} \\ \alpha^{(1)} = \alpha^{(2)} \text{ and } \beta^{(1)} \prec_y \beta^{(2)} \end{cases},$$

where $\alpha^{(1)}, \alpha^{(2)} \in \mathbb{Z}_{\geq 0}^n$ and $\beta^{(1)}, \beta^{(2)} \in \mathbb{Z}_{\geq 0}^m$. The term order $\prec_{[\{x\} \succ \{y\}]}$ is called elimination order with the x variables larger than the y variables (which is indeed a term order).

Similarly one can define elimination order for more than two subsets of indeterminates. For example, one can consider a polynomial ring $k[x, y, z]$ in x, y and z variables and one can define an elimination term order $\prec_{[\{x\} \succ \{y\} \succ \{z\}]}$. An example for this is term order $\prec_{[\{x_1\} \succ \dots \succ \{x_n\}]}$ (or simply we denote this as $\prec_{[x_1 \succ \dots \succ x_n]}$) on polynomial ring $k[x_1, \dots, x_n]$.

Now we are ready to define important algebraic objects called *elimination ideals*.

Definition 3.5. *Given $\mathfrak{a} \subset k[x_1, \dots, x_n]$, the l^{th} elimination ideal $\mathfrak{a}_l$ in the polynomial ring $k[x_{l+1}, \dots, x_n]$ is defined as $\mathfrak{a}_l = \mathfrak{a} \cap k[x_{l+1}, \dots, x_n]$.*

With these concepts we now state the elimination theorem.

Theorem 3.6. *(Adams and Loustaunau, 1994, p. 69) Let $\mathfrak{a} \subset k[x_1, \dots, x_n]$ be an ideal and let $G \subset k[x_1, \dots, x_n]$ be a Gröbner basis of $\mathfrak{a}$ with respect a term order $\{x_1, \dots, x_l\} \succ \{x_{l+1}, \dots, x_n\}$ (for example consider the term order $x_1 \succ \dots \succ x_n$) for $0 \leq l \leq n$. Then the set $G_l = G \cap k[x_{l+1}, \dots, x_n]$ is a Gröbner basis of l^{th} elimination ideal $\mathfrak{a}_l$.*

4. ME in the framework of Algebraic Statistics

There has been some attempts to give a common terminology for the rapidly developing field of algebraic statistics (*e.g.*, Drton and Sullivant, 2006). Here, we adopt the appropriate definition of statistical model from (Pachter and Sturmfels, 2005).

A statistical model $\mathcal{M}$ is a subset of Δ_m and is said to be algebraic if $\exists f_1, \ldots, f_s \in k[p_1, \ldots, p_m]$ such that

$$\mathcal{M} = \mathcal{V}(f_1, \ldots, f_s) \cap \Delta_m \ .$$

Let $\Theta \subseteq \mathbb{R}^d$ be a parameter space and $\kappa : \Theta \to \Delta_m$ be a map. The image $\kappa(\Theta)$ is called a parametric statistical model. Given a statistical model $\mathcal{M} \subseteq \Delta_m$, by parametrization of $\mathcal{M}$ we mean, identifying a set $\Theta \subseteq \mathbb{R}^d$ and a function $\kappa : \Theta \to \Delta_m$ such that $\mathcal{M} = \kappa(\Theta)$. To define an algebraic counter part of parametric statistical models one needs some restrictions on Θ and κ which are of algebraic nature. First, we need the following definition.

Definition 4.1. *A set $\Theta \subseteq \mathbb{R}^d$ is called a semi-algebraic set, if there are two finite collections of polynomials $F \subset k[x_1, \ldots, x_d]$ and $G \subset k[x_1, \ldots, x_d]$ such that*

$$\Theta = \{\theta \in \mathbb{R}^d : f(\theta) = 0, \forall f \in F \textit{ and}\, g(\theta) \geq 0, g \in G\} \ .$$

The simplest example of a semialgebraic set is the probability simplex Δ_m itself. In general, any convex polyhedron or polytope is semialgebraic (the requisite G and G will consist of linear polynomials).

Now we have the following definition of parametric algebraic statistical model.

Definition 4.2. *Let Δ_m be a probability simplex and $\Theta \subset \mathbb{R}^d$ be a semi-algebraic set. Let $\kappa : \mathbb{R}^d \to \mathbb{R}^m$ be a rational function (a rational function is a quotient of two polynomials) such that $\kappa(\Theta) \subseteq \Delta_m$. Then the image $\mathcal{M} = \kappa(\Theta)$ is a parametric algebraic statistical model.*

Conversely, a parametric statistical model $\mathcal{M} = \kappa(\Theta) \subseteq \Delta_m$ is said to be algebraic if Θ is a semi-algebraic set and κ is a rational function. From now on we refer to 'parametric algebraic statistical models' as 'algebraic statistical models'.

In this paper we consider the following special case of algebraic statistical models (*cf.* Pachter and Sturmfels, 2005, p. 7). Consider a map

$$\begin{aligned} \kappa : \Theta(\subseteq \mathbb{R}^d) &\to \mathbb{R}^m \\ \theta = (\theta_1, \ldots, \theta_d) &\mapsto (\kappa_1(\theta), \ldots, \kappa_m(\theta)) \ , \end{aligned} \tag{4.1}$$

where $\kappa_i \in k[\theta_1, \ldots, \theta_d]$. We assume that Θ satisfies $\kappa_i(\theta) \geq 0$, $i = 1, \ldots, m$, and $\sum_{i=1}^{m} \kappa_i(\theta) = 1$ for any $\theta \in \Theta$. Under these conditions $\kappa(\Theta)$ is indeed an algebraic statistical model (Definition 4.2) since $\kappa(\Theta) \subset \Delta_m$, κ is a polynomial function and Θ is a semi-algebraic set ($F = \{\sum_{i=1}^{m} \kappa_i - 1\}$ and $G = \{\kappa_i : i = 1, \ldots, m\}$ in the Definition 4.1).

Some statistical models are naturally given by a polynomial map κ (4.1) for which the condition $\sum_{i=1}^{m} \kappa_i(\theta) = 1$ does not hold. If this is the case, one may consider the following algebraic statistical model:

$$\kappa : \theta = (\theta_1, \ldots, \theta_d) \mapsto \frac{1}{\sum_{i=1}^{m} \kappa_i(\theta)} \left(\kappa_1(\theta), \ldots, \kappa_m(\theta)\right) \quad , \tag{4.2}$$

assuming that the remaining conditions that have been specified for the model (4.1) are valid here too. The only difference is that instead of κ being a polynomial map, we have it as a rational map. One can show that many statistical models are algebraic statistical models, for reference see (Pachter and Sturmfels, 2005).

Now by setting $\xi_i = -\ln \theta_i$, $i = 1, \ldots, d$ in (2.5) one can verify that ME-models are algebraic statistical models.

5. Embedding ME-models in Algebraic Varieties

Given a positive integer valued sufficient statistic $t_i, i = 1, \ldots, d$, we have maximum entropy model as image of

$$\begin{aligned} f : k^d &\to k^m - W \\ (\theta_1, \ldots, \theta_d) &\mapsto \left(\frac{\prod_{i=1}^{d} \theta_i^{t_i(1)}}{\sum_{j=1}^{m} \prod_{i=1}^{d} \theta_i^{t_i(j)}}, \ldots, \frac{\prod_{i=1}^{d} \theta_i^{t_i(m)}}{\sum_{j=1}^{m} \prod_{i=1}^{d} \theta_i^{t_i(m)}} \right) \quad . \end{aligned} \tag{5.1}$$

where $W = \mathcal{V}(\sum_{j=1}^{m} \prod_{i=1}^{d} \theta_i^{t_i(j)})$. Note that this is bigger parametrization as we allow negative probabilities. Also note that the image of a polynomial map (or rational map) need not be a variety. In this case, the usual technique employed is to take the Zariski closure (Zariski closure of a set $A \subset k^n$ is the smallest variety that contains A (Adams and Loustaunau, 1994, §2.5)). By the following result, one could give an algebraic description for the ME-model (*i.e.*, $\mathrm{im}(f)$) in (5.1).

Theorem 5.1. *Let f be a polynomial function that parametrizes a maximum entropy model with respect to sufficient statistics $t_i : \mathbb{R} \to \mathbb{Z}_{\geq 0}$ according to (5.1). Then*

$$\mathrm{im}(f) \subseteq \mathcal{V}(\ker(\tilde{f}^*)) \cap \mathcal{V}(\sum_{j=1}^{m} p_j - 1) \tag{5.2}$$

where $\tilde{f}^*$ is a k-algebra homomorphism[1]

$$\begin{aligned} \tilde{f}^* : k[p_1, \ldots, p_m] &\to k[\theta_0, \ldots, \theta_d] \\ p_j &\mapsto \theta_0 \prod_{i=1}^{d} \theta_i^{t_i(j)} \quad . \end{aligned} \tag{5.3}$$

Further $\mathcal{V}(\ker(\tilde{f}^*)) \cap \mathcal{V}(\sum_{j=1}^{m} p_j - 1)$ is the smallest variety that contains the maximum entropy model.

[1] k-algebra homomorphism is a ring homomorphism

$$\psi : k[y_1, \ldots, y_m] \longrightarrow k[x_1, \ldots, x_n]$$

which is also a k-vector space linear transformation. Such a map is uniquely determined by

$$\psi : y_i \mapsto h_i$$

where $h_i \in k[x_1, \ldots, x_n]$, $i = 1, \ldots, m$.

Proof. Define an 'unnormalized' ME model given sufficient statistic $t_i, i = 1, \ldots, d$, as image of k-algebra homomorphism

$$\begin{aligned} \tilde{f} : k^{d+1} &\to k^m \\ (\theta_0, \theta_1, \ldots, \theta_d) &\mapsto \left(\theta_0 \prod_{i=1}^d \theta_i^{t_i(1)}, \ldots, \theta_0 \prod_{i=1}^d \theta_i^{t_i(m)}\right) \quad . \end{aligned} \tag{5.4}$$

Now we claim that

$$\mathrm{im}(f) = \mathrm{im}(\tilde{f}) \cap \Delta_m \quad . \tag{5.5}$$

Clearly we have $\mathrm{im}(f) \subseteq \mathrm{im}(\tilde{f}) \cap \Delta_m$. Let $a = (a_1, \ldots, a_m) \in \mathrm{im}(\tilde{f}) \cap \Delta_m$. Then $\exists (b_0, b_1, \ldots, b_d) \in k^{d+1}$ such that

$$a_j = b_0 \prod_{i=1}^d b_i^{t_i(j)} \quad , j = 1, \ldots, m.$$

Since $a \in \Delta_m$, $\sum_{j=1}^m a_j = 1$ implies

$${b_0}^{-1} = \sum_{j=1}^m \prod_{i=1}^d b_i^{t_i(j)} \quad .$$

Hence $a \in \mathrm{im}(f)$ and $\mathrm{im}(f) \supseteq \mathrm{im}(\tilde{f}) \cap \Delta_m$.

Now we only have to show that $\mathrm{im}(\tilde{f}) \subseteq \mathcal{V}(\ker(\tilde{f}^*))$.

By the theorem of polynomial implicitization (Cox et al., 1991, p. 126) we have

$$\mathrm{im}(\tilde{f}) \subseteq \mathcal{V}(\mathfrak{a} \cap k[p_1, \ldots, p_m]) \quad , \tag{5.6}$$

where

$$\mathfrak{a} = \left\langle p_j - \theta_0 \prod_{i=1}^d \theta_i^{t_i(j)} : j = 1, \ldots, m \right\rangle \quad .$$

$\mathfrak{a} \cap k[p_1, \ldots, p_m]$ is nothing but the kernel of k-algebra homomorphism defined by (5.3) (Adams and Loustaunau, 1994, Theorem 2.42, p. 81). Hence the result. □

The following corollary is part of the above proof.

Corollary 5.2.

$$\mathrm{im}(f) \subseteq \mathcal{V}(\mathfrak{a} \cap k[p_1, \ldots, p_m]) \cap \mathcal{V}(\sum_{j=1}^m p_j - 1) \quad , \tag{5.7}$$

where

$$\mathfrak{a} = \left\langle p_j - \theta_0 \prod_{i=1}^d \theta_i^{t_i(j)} : j = 1, \ldots, m \right\rangle$$

By the elimination theorem (Theorem 3.6) we can find the generators for ideal $\ker(\tilde{f}^*)$ easily by first computing the Gröbner bases for $\mathfrak{a}$ and removing all the polynomials involving indeterminates $\theta_1, \ldots, \theta_d$.

Corollary 5.3. *(By Elimination theorem)*

$$\mathrm{im}(f) \subseteq \mathcal{V}(G \cap k[p_1, \ldots, p_m]) \cap \mathcal{V}(\sum_{j=1}^{m} p_j - 1) \ , \tag{5.8}$$

where G is the Gröbner basis of

$$\mathfrak{a} = \left\langle p_j - \theta_0 \prod_{i=1}^{d} \theta_i^{t_i(j)} : j = 1, \ldots, m \right\rangle$$

with respect to a term order satisfying $\{p_1, \ldots, p_m\} \prec \{\theta_1, \ldots, \theta_d\}$.

The following comments are in place. Here, we compute an implicitly defined nonnormalized ME model and then intersect with the linear variety $\mathcal{V}(\sum_{k=1}^{1} p_k - 1)$ to give an implicit representation for the ME-model. This gives rise to an implicit representation of the ME-model, which contains only binomials. These ideals of the form $\ker(\tilde{f}^*)$ are known as toric ideals and they are very well studied in the literature. Also by using toric ideal theory (see Sturmfels, 1996) one could extend the above results to situations where our sufficient statistic could be negative.

On the other hand, the ME model (5.1) can be viewed as rational mapping and one can apply the implicitization theorem for rational functions (Cox et al., 1991, p. 130) to get another implicit representation. But in this case generators of the implicit representation need not be binomials.

Now we demonstrate the implicit ME models with an example. Consider the following sufficient statistic

$$[t_i(j)]_{((i=1,2)\times(j=1,\ldots,7))} = \begin{pmatrix} 2 & 1 & 3 & 1 & 5 & 2 & 1 \\ 1 & 2 & 1 & 4 & 3 & 3 & 1 \end{pmatrix} \ . \tag{5.9}$$

The corresponding ME model can be written as

$$(p_1, p_2, p_3, p_4, p_5, p_6, p_7) = \left(\frac{\theta_1^2\theta_2}{Z}, \frac{\theta_1\theta_2^2}{Z}, \frac{\theta_1^3\theta_2}{Z}, \frac{\theta_1\theta_2^4}{Z}, \frac{\theta_1^5\theta_2^3}{Z}, \frac{\theta_1^2\theta_2^3}{Z}, \frac{\theta_1\theta_2}{Z} \right) \ ,$$

where $Z = \theta_1^2\theta_2 + \theta_1\theta_2^2 + \theta_1^3\theta_2 + \theta_1\theta_2^4 + \theta_1^5\theta_2^3 + \theta_1^2\theta_2^3 + \theta_1\theta_2$.

The corresponding unnormalized ME model is

$$(p_1, p_2, p_3, p_4, p_5, p_6, p_7) = \left(\theta_0\theta_1^2\theta_2, \theta_0\theta_1\theta_2^2, \theta_0\theta_1^3\theta_2, \theta_0\theta_1\theta_2^4, \theta_0\theta_1^5\theta_2^3, \theta_0\theta_1^2\theta_2^3, \theta_0\theta_1\theta_2 \right) \ .$$

Now let $\mathfrak{a}$ be the ideal generated by the polynomials

$$\begin{array}{ll} p_1 - \theta_0\theta_1^2\theta_2, & p_2 - \theta_0\theta_1\theta_2^2, \\ p_3 - \theta_0\theta_1^3\theta_2, & p_4 - \theta_0\theta_1\theta_2^4, \\ p_5 - \theta_0\theta_1^5\theta_2^3, & p_6 - \theta_0\theta_1^2\theta_2^3, \\ p_7 - \theta_0\theta_1\theta_2 & \end{array}$$

in $k[p_1, p_2, p_3, p_4, p_5, p_6, p_7, \theta_0, \theta_1, \theta_2]$. By using *Mathematica* software, the Gröbner basis for the above ideal $\mathfrak{a}$ with the term order

$$p_1 \prec p_2 \prec p_3 \prec p_4 \prec p_5 \prec p_6 \prec p_7 \prec \theta_1 \prec \theta_2$$

can be found as

$$\begin{array}{rl}
-p_7+\theta_0\theta_1, & p_1\theta_1-p_3, \\
p_1\theta_2-p_2\theta_1, & p_2\theta_2-p_4\theta_1, \\
p_1p_7-p_3\theta_0, & p_1p_2-p_6p_7, \\
p_2p_6-p_4p_1, & p_2-p_4p_7, \\
p_2\theta_1-p_3\theta_2, & p_7\theta_2\theta_1-p_1, \\
p_2\theta_2\theta_1-p_6, & p_6p_7\theta_1-p_2p_3, \\
p_2p_3\theta_1-p_5\theta_0, & p_7\theta_2^2-p_2, \\
p_3\theta_2^2-p_6\theta_1, & p_2\theta_2^2-p_4, \\
p_5\theta_0\theta_2-p_6p_3, & p_3\theta_0\theta_2-p_2, \\
p_7^2\theta_2-p_1\theta_0, & p_5p_7\theta_2-p_6p_3\theta_1, \\
p_3p_7\theta_2-p_1^2, & p_2p_7\theta_2-p_6\theta_0, \\
p_3^2\theta_2-p_5p_7, & p_4p_3\theta_2-p_6^2, \\
p_2p_3\theta_2-p_6p_1, & p_6p_7^2-p_2p_3\theta_0, \\
p_2p_3p_7-p_2\theta_0^2, & p_1^3-p_5p_7\theta_0, \\
p_3p_1^2-p_5p_7^2, & p_6p_1^2-p_2p_5\theta_0, \\
p_4p_1^2-p_6^2p_7, & p_6p_3p_1-p_2p_5p_7, \\
p_6^2p_1-p_4p_5\theta_0, & p_4p_3^2-p_6p_5\theta_0, \\
p_2p_3^2-p_1p_5\theta_0, & p_6^2p_3-p_4p_5p_7, \\
p_6\theta_1^3-p_5, & p_4p_7\theta_1^2-p_6p_1, \\
p_6p_3\theta_1^2-p_1p_5, & p_4p_3\theta_1^2-p_2p_5, \\
p_6^2\theta_1^2-p_2p_5\theta_2, & p_5p_7^2\theta_1-p_1p_3^2, \\
p_4p_7^2\theta_1-p_6p_1\theta_0, & p_2p_7^2\theta_1-p_1^2\theta_0, \\
p_2p_5p_7\theta_1-p_6p_3^2, & p_4p_3p_7\theta_1-p_2p_5\theta_0, \\
p_2p_4p_7\theta_1p_6^2\theta_0, & p_6p_3^2\theta_1-p_1^2p_5, \\
p_4^2p_3\theta_1-p_6^3, & p_6^3\theta_1-p_2p_4p_5, \\
p_5p_7^3-p_3^2p_1\theta_0, & p_4p_7^3-p_6p_1\theta_0^2, \\
p_2p_7^3-p_1^2\theta_0^2, & p_2p_5p_7^2-p_6p_3^2\theta_0, \\
p_4p_3p_7^2-p_2p_5\theta_0^2, & p_2p_4p_7^2-p_6^2\theta_0^2, \\
p_2p_3^2p_7-p_1^2p_5\theta_0, & p_4^2p_3p_7-p_6^3\theta_0, \\
p_6^3p_7-p_2p_4p_5\theta_0, & p_6p_3^3-p_5^2p_7\theta_0, \\
p_4\theta_1^4-p_5\theta_2, & p_4^3p_5p_7\theta_1-p_6^5, \\
p_4^3p_5p_7^2-p_6^5\theta_0, & p_6^6-p_2p_4^3p_3p_5.
\end{array}$$

From Corollary 5.3 the maximum entropy model is contained in the variety defined by the polynomials

$$\begin{array}{rl}
p_1p_2-p_6p_7, & p_2p_6-p_4p_1, \\
p_2-p_4p_7, & p_3p_1^2-p_5p_7^2, \\
p_4p_1^2-p_6^2p_7, & p_6p_3p_1-p_2p_5p_7, \\
p_6^2p_3-p_4p_5p_7, & p_6^6-p_2p_4^3p_3p_5,
\end{array} \quad \sum_{i=1}^{7} p_i - 1.$$

We denote this collection of polynomials by $M \subset k[p_1,\ldots,p_7]$ and denote polynomials that captures the data in the form of constraints (2.2) by

$$D = \{\sum_{j=1}^{m} t_i(j)p_j - T_i : i = 1,2\} \subset k[p_1,\ldots,p_7] \; .$$

Now, we can say that the maximum entropy model given by the sufficient statistic (5.9) is contained in the algebraic variety $\mathcal{V}(M)$ and given the values of T_1 and T_2 the maximum entropy distribution is contained in the variety $\mathcal{V}(M) \cap \mathcal{V}(D)$.

Finally, one could conclude that in the case of maximum entropy both the model and the data can be embedded in algebraic varieties and the maximum entropy distribution itself is contained in the intersection of these varieties.

6. Closing Remarks

In this paper we attempted to give an algebraic descriptions of ME-models in the finitely discrete case. It might seem that this paper studied a particular case, but the reader should be aware of the fact that any distribution, be it one- or multidimensional, may be approximated by a discrete distribution, arbitrarily closely.

We showed that estimation of ME-models can be transformed to solving a system of polynomial equations and we discussed various cases, viz., Primal, Dual and Kullback-Csisźar iteration, in this respect. We also showed that the ME-model can be treated with Toric ideals and can be embedded in algebraic varieties by Gröbner bases methods. By this we demonstrated that in the case of ME, both model and data can be represented by algebraic varieties, *implicitly*, and we hope that this result paves a way to work on algebraic geometry of information theoretic statistics.

Acknowledgments

Author would like to thank Prof. C. A. J. Klaassen for patiently going through this work and providing valuable advise. Author thanks Dr. Peter Grünwald for discussions on this topic.

References

Adams, W. W. and P. Loustaunau (1994). *An Introduction to Gröbner Bases*, Volume 3 of *Graduate Studies in Mathematics*. American Mathematical Society.

Buchberger, B. (1965). *An Algorithm for Finding a Basis for the Residue Class Ring of a Zero-Dimensional Polynomial Ideal (in German)*. Ph. D. thesis, University of Innsbruck, Austria. (reprinted in Buchberger (2006)).

Buchberger, B. (2006). Bruno buchberger's phd thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. *Journal of Symbolic Computation 41*, 475–511.

Buchberger, B. and Gröbner (1985). An algorithmic method in polynomial ideal theory. In N. K. Bose (Ed.), *Multidimensional Systems Theory*, Chapter 6. D. Reidel.

Cox, D., J. Little, and D. O'Shea (1991). *Ideals, Varieties, and Algorithms* (2nd ed.). New York: Springer.

Csiszár, I. (1975). I-divergence of probability distributions and minimization problems. *Ann. Prob. 3*(1), 146–158.

Csiszár, I. (1989). A geometric interpreation of Darroch and Ratcliff's generalized iterative scaling. *The Annals of Statistics 17*(3), 1409–1413.

Csiszár, I. and P. Shields (2004). *Information Theory and Statistics: A Tutorial*, Volume 1 of *Foundations and Trends in Communications and Information Theory*. Now Publications.

Darroch, J. N. and D. Ratcliff (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics 43*(5), 1470–1480.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Volume 11 of *Institute of Mathematical Statistics–Monograph Series*. Hayward, CA: Institute of Mathematical Statistics.

Diaconis, P. and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics 26*, 363–397.

Drton, M. and S. Sullivant (2006). Algebraic statistical models. *(To be published in) Statistical Sinica*.

Ireland, C. and S. Kullback (1968). Contingency tables with given marginals. *Biometrika 55*, 179–188.

Jaynes, E. T. (1957). Information theory and statistical mechanics i. *Physical Review 106*(4), 620–630.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics sec-4*(3), 227–241.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Kullback, S. (1968). Probability densities with given marginals. *The Annals of Mathematical Statistics 39*(4), 1236–1243.

Pachter, L. and B. Sturmfels (2005). *Algebraic Statistics and Computational Biology*. Cambridge: Cambridge University Press.

Pauer, F. and A. Unterkircher (1999). Gröbner bases for ideals in Laurent polynomial rings and their applications to systems of differential equations. *Applicable Algebra in Engineering, Communication and Computing 9*, 271–291.

Pistone, G., E. Riccomagno, and H. Wynn (2001). *Algebraic Statistics. Computational Commutative Algebra in Statistics*. New York: Chapman and Hall.

Sturmfels, B. (1996). *Gröbner Bases and Convex Polytopes*, Volume 8 of *University Lecture Series*. Providence, RI: AMS Press.

Viana, M. A. G. and D. S. P. Richards (Eds.) (2001). *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics*. Providence, RI: American Mathematical Society.

EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

E-mail address: dukkipati@eurandom.tue.nl