A Semi-Algebraic Description of Discrete Naive Bayes Models with Two Hidden Classes

Vincent Auvray[*]
EE and CS Dept.
University of Liège
Vincent.Auvray@ulg.ac.be

Pierre Geurts[†]
EE and CS Dept.
University of Liège
geurts@montefiore.ulg.ac.be

Louis Wehenkel
EE and CS Dept.
University of Liège
L.Wehenkel@ulg.ac.be

Abstract

Discrete Bayesian network models with hidden variables define an important class of statistical models. These models are usually defined parametrically, but can also be described semi-algebraically as the solutions in the probability simplex of a finite set of polynomial equations and inequations. In this paper we present a semi-algebraic description of discrete Naive Bayes models with two hidden classes and a finite number of observable variables. The identifiability of the parameters is also studied. Our derivations are based on an alternative parametrization of the Naive Bayes models with an arbitrary number of hidden classes.

1 Introduction

A Bayesian network model (see [Cowell *et al.*, 1999]) is a set of probability densities over a set of random variables. It can be specified parametrically as a product of conditional probability densities or implicitely by a set of (conditional) independence constraints between the random variables.

On the other hand, a Bayesian network model with hidden variables (see [Geiger *et al.*, 2001]) is usually only specified parametrically. The density over the non hidden variables is obtained by marginalisation of the hidden variables, which implies non-independence constraints that are difficult to formulate.

Let us focus on Bayesian network models over a finite set of discrete random variables. In [Geiger *et al.*, 2001], the authors show non-constructively that, with or without hidden variables, these models are semi-algebraic sets, i.e. they are implicitely described by a finite set of polynomial equalities and inequalities. Finding semi-algebraic descriptions of discrete bayesian networks with hidden variables is a difficult problem that has been the focus of recent research (see [Geiger, 1998] and [Settimi and Smith, 2000]). In particular, important progress has been reported in [Garcia *et al.*, 2005] and [Garcia, 2004] where discrete Bayesian networks are placed in the realm of computational algebraic statistics.

An (implicit) semi-algebraic representation allows to easily check whether a given distribution belongs to a model. Statistical tests can then be possibly devised for model selection.

Despite their simplicity, discrete Naive Bayes models can be successfully used for density estimation and allow for fast inference (see [Lowd and Domingos, 2005]). In this paper, we investigate those models in the special case where the hidden variable has two possible classes. Section 2 gives the definition of the discrete Naives Bayes models and introduces an alternative parametrization. Section 3 focuses on the inversion of the mapping from the parameter space to the probability distribution space. In section 4, we derive a semi-algebraic description by eliminating the parameters. Finally, we conclude in section 5. The proofs of two lemmas are given in the appendix.

2 Discrete Naive Bayes models

A probability distribution p on a finite sample space Ω with n configurations is naturally represented as a vector $(p(o))_{o\in\Omega} \in \mathbb{R}^n$ if we fix an order for the elements of Ω. Such a vector satisfies the following constraints.

$$\sum_{o\in\Omega} p(o) = 1, \tag{1}$$

$$(\forall o \in \Omega)(p(o) \geq 0). \tag{2}$$

[*]Vincent Auvray is a Research Fellow of the F.N.R.S.

[†]Pierre Geurts is a Postdoctoral Researcher of the F.N.R.S.

In fact, equations (1) and (2) give a semi-algebraic representation of the set of all probability distributions on Ω, the $(n-1)$-dimensional probability simplex.

Let X be an ordered finite set of discrete random variables and let $v(i)$ be the set of possible values of $i \in X$. The sample space considered is thus $\Omega = \times_{i \in X} v(i)$. A vector $\Theta = (\theta(i))_{i \in X}$ where $\theta(i)$ is a probability distribution on i can be mapped to the distribution on X given by

$$p(X) = \prod_{i \in X} \theta(i). \tag{3}$$

Let D_0 be the set of distributions parametrized by this mapping.

For $p \in D_0$, the random variables are mutually independent. This may be too restrictive for density estimation and we may want to consider mixtures of such distributions. Suppose that the number of mixture components s is fixed. As a parameter space, consider the set Π_0 of vectors $((\omega_t, \Theta_t))_{t=1}^s$ where $\Theta_t = (\theta_t(i))_{i \in X}$ and such that

$$\omega_t \geq 0, \tag{4}$$

$$\theta_t(i) \geq 0, \tag{5}$$

$$\sum_{t=1}^{s} \omega_t = 1, \tag{6}$$

$$\sum_{i_0 \in v(i)} \theta_t(i_0) = 1. \tag{7}$$

A vector $((\omega_t, \Theta_t))_{t=1}^s \in \Pi_0$ can be mapped by the function f_0 to the distribution p over X given by

$$p(X) = \sum_{t=1}^{s} \omega_t \prod_{i \in X} \theta_t(i), \tag{8}$$

The set of distributions $NB_s = f_0(\Pi_0)$ parametrized by this mapping is the Naive Bayes model with s hidden classes. Of course, $D_0 = NB_1$.

The parameters of a distribution $p \in NB_s$ are identifiable if f_0 is injective at p, i.e. if the set $f_0^{-1}(p)$ contains only one element. The commutativity of the addition in equation (8) implies that any permutation of the s components of a vector $((\omega_t, \Theta_t))_{t=1}^s$ produces the same mixture. Therefore, we can only hope for identifiability up to such a permutation.

Equations (4) and (6) allow for some weights to be 0. As trivial consequences, $NB_s \subseteq NB_{s+1}$ and any mixture of s components parametrized as a mixture of more than s components will have non identifiable parameters.

2.1 Illustration

To illustrate, let us briefly consider NB_1. Given $p \in NB_1$, we have $\omega_1 = 1$ and by marginalisation

$$p(i) = \theta_1(i). \tag{9}$$

Hence, we have

$$p(X) = \prod_{i \in X} p(i) \tag{10}$$

Conversely, any distribution on X satisfying equation (10) is uniquely parametrized by the vector $(\omega_1, \theta_1) = (1, (p(i))_{i \in X})$. Hence, together with equations similar to (1) and (2), equation (10) is a semi-algebraic description of NB_1 and the parameters are identifiable.

2.2 About the notations

As the reader may have noticed, we adopted rather informal notations in this paper.

First, expressions like $p(i = i_0)$ and $\theta(i = i_0)$ for some random variable $i \in X$ and some value $i_0 \in v(i)$ are abbreviated as $p(i_0)$ and $\theta(i_0)$.

Moreover, if a variable or an index x is not constrained in an expression $q(x)$ then this expression should be read as

$$(\forall x \text{ such that } q(x) \text{ is defined})(q(x)). \tag{11}$$

For example, supposing that $X = \{i, j\}$, equation (3) stands for

$$(\forall i_0 \in v(i))(\forall j_0 \in v(j)) \\ (p(i = i_0, j = j_0) = \theta(i = i_0)\theta(j = j_0)), \tag{12}$$

and equation (5) stands for

$$(\forall t \in \{1, \ldots, s\})(\forall i \in X)(\forall i_0 \in v(i))(\theta_t(i = i_0) \geq 0). \tag{13}$$

2.3 An alternative parametrization

To simplify the developments in the remaining sections, we introduce an alternative parametrization of NB_s. As a new parameter space, consider the set Π of vectors $(\Lambda, ((\omega_t, \Delta_t)_{t=1}^s))$ where

$$\Lambda = (\lambda(i))_{i \in X}, \tag{14}$$

$$\Delta_t = (\delta_t(i))_{i \in X}, \tag{15}$$

and such that

$$\omega_t \geq 0, \tag{16}$$
$$\delta_t(i) \leq \lambda(i), \tag{17}$$
$$\sum_{t=1}^{s} \omega_t = 1, \tag{18}$$
$$\sum_{i_0 \in v(i)} \lambda(i_0) = 1, \tag{19}$$
$$\sum_{i_0 \in v(i)} \delta_t(i_0) = 0, \tag{20}$$
$$\sum_{t=1}^{s} \omega_t \delta_t(i) = 0. \tag{21}$$

Let us show that the function h given by

$$(\Lambda, ((\omega_t, \Delta_t)_{t=1}^s)) \mapsto ((w'_t, \Theta_t)_{t=1}^s) \tag{22}$$

with

$$\omega'_t = \omega_t, \tag{23}$$
$$\theta_t(i) = \lambda(i) - \delta_t(i), \tag{24}$$

is a bijection between Π and the original parameter space Π_0 whose inverse h^{-1} is given by

$$((\omega_t, \Theta_t)_{t=1}^s) \mapsto (\Lambda, ((w'_t, \Delta_t)_{t=1}^s)) \tag{25}$$

with

$$\omega'_t = \omega_t, \tag{26}$$
$$\lambda(i) = \sum_{t=1}^{s} \omega_t \theta_t(i), \tag{27}$$
$$\delta_t(i) = \lambda(i) - \theta_t(i). \tag{28}$$

Given $\pi = (\Lambda, ((\omega_t, \Delta_t)_{t=1}^s)) \in \Pi$, it is immediate to see that $h(\pi) \in \Pi_0$. Suppose that $h(\Lambda, ((\omega_t, \Delta_t)_{t=1}^s)) = h(\Lambda^*, ((\omega_t^*, \Delta_t^*)_{t=1}^s))$. This implies that

$$\omega_t = \omega_t^*, \tag{29}$$
$$\lambda(i) - \delta_t(i) = \lambda^*(i) - \delta_t^*(i). \tag{30}$$

Hence, we have

$$\sum_{t=1}^{s} \omega_t(\lambda(i) - \delta_t(i)) = \sum_{t=1}^{s} \omega_t^*(\lambda^*(i) - \delta_t^*(i)). \tag{31}$$

By equations (18) and (21), this implies that

$$\lambda(i) = \lambda^*(i), \tag{32}$$

and, in turn,

$$\delta_t(i) = \delta_t^*(i). \tag{33}$$

Hence, h is injective. On the other hand, given $\pi_0 = ((\omega_t, \Theta_t)_{t=1}^s) \in \Pi_0$, it is immediate to show that $h^{-1}(\pi_0) \in \Pi$ and that $h(h^{-1}(\pi_0)) = \pi_0$. Therefore, h is surjective and h^{-1} is its inverse.

This shows that NB_s is also parametrized by Π and the function f such that $f(\Lambda, ((\omega_t, \Delta_t)_{t=1}^s)) = p$ with

$$p(X) = \sum_{t=1}^{s} \omega_t \prod_{i \in X} (\lambda(i) - \delta_t(i)). \tag{34}$$

Similarly, we describe a distribution p over X in a non standard way by means of a function g defined on the subsets S of X as follows

$$g(S) = p(S) - \sum_{P \text{ s.t. } P \subset S} g(P) \prod_{i \in S \setminus P} p(i), \tag{35}$$

with the conventions that $p(\emptyset) = 1$ and $P \subset S$ stands for P is a proper subset of S. For example,

$$g(\emptyset) = 1, \tag{36}$$
$$g(i) = 0, \tag{37}$$
$$g(i,j) = p(i,j) - p(i)p(j), \tag{38}$$
$$g(i,j,k) = p(i,j,k) - p(i)p(j,k) - p(j)p(i,k) \tag{39}$$
$$- p(k)p(i,j) + 2p(i)p(j)p(k). \tag{40}$$

Note that g has the following properties

$$i \perp j \Leftrightarrow g(i,j) = 0, \tag{41}$$
$$i \perp S \Rightarrow g(S \cup \{i\}) = 0, \tag{42}$$
$$\sum_{i_0 \in v(i)} g(i_0, j) = 0. \tag{43}$$

and $g(S)$ is a polynomial of degree $|S|$.

3 Inversion of $f : \Pi \to NB_2$

In this section, we discuss the inversion of the parametrization mapping f. Suppose that a probability distribution $p \in NB_2$ is given. Let us attempt to identify the set of parameters $f^{-1}(p)$.

Equation (18) is equivalent to

$$\omega_2 = 1 - \omega_1. \tag{44}$$

Also, by marginalisation in equation (34) of the variables in $X \setminus \{i\}$, we obtain

$$\lambda(i) = p(i). \tag{45}$$

This implies that equation (19) is satisfied.

The following lemma is proved in the appendix.

Lemma 3.1 *Given a distribution p over X and $(\Lambda, ((\omega_t, \Delta_t)_{t=1}^2)) \in \Pi$ such that $\lambda(i) = p(i)$, we have*

$$p = f(\Lambda, ((\omega_t, \Delta_t)_{t=1}^2)) \tag{46}$$

if and only if

$$g(S) = -(\delta_1(i)+\delta_2(i))g(S\backslash\{i\})-\delta_1(j)\delta_2(k)g(S\backslash\{j,k\}). \tag{47}$$

In accordance with our notational conventions, equation (47) should be read as

$$\begin{aligned}(\forall S \subseteq X)(|S| \geq 2)(\forall s_0 \in \times_{u\in S}\, v(u))(\forall i \in S)\\ (\forall \{j,k\} \subseteq S)(g(S = s_0) = -(\delta_1(i = i_0) + \delta_2(i = i_0))\\ g(S \setminus \{i\} = s_0 \setminus \{i_0\}) - \delta_1(j = j_0)\delta_2(k = k_0)\\ g(S \setminus \{j,k\} = s_0 \setminus \{j_0, k_0\})),\end{aligned} \tag{48}$$

where i_0, j_0 and k_0 are, respectively, the values of i, j and k in the instantiation s_0.

Lemma 3.1 has an immediate corollary that can be used for identification purposes.

Corollary 3.2 *Given $p = f(\Lambda, ((\omega_t, \Delta_t)_{t=1}^2))$, we have*

$$g(i,j) = -\delta_1(i)\delta_2(j), \tag{49}$$
$$\delta_t(i_l)g(i_m, j) = \delta_t(i_m)g(i_l, j), \tag{50}$$
$$g(i,j)g(i,k) = -\delta_1(i)\delta_2(i)g(j,k), \tag{51}$$
$$g(i,j,k) = -(\delta_1(i) + \delta_2(i))g(j,k). \tag{52}$$

Let us partition the set of probability distributions on X into three subsets D_0, D_1 and D_2 with $D_0 = NB_1$ defined previously. Let D_1 be the set of distributions that have a single pair of non-independent random variables, i.e.

$$(\exists!\{a,b\} \subseteq X)(a \not\perp b). \tag{53}$$

Let D_2 be the set of distributions with at least two distinct pairs of non-independent random variables, i.e.

$$\begin{aligned}(\exists\{a,b\},\{c,d\} \subseteq X)\\ (\{a,b\} \neq \{c,d\}) \wedge (a \not\perp b) \wedge (c \not\perp d)).\end{aligned} \tag{54}$$

Note that D_1 is defined only if $|X| \geq 2$ and D_2 is defined only if $|X| \geq 3$. Also, $\omega_1 = 0$ or $\omega_1 = 1$ implies that $p \in NB_1$.

Let us now analyse these three cases separately.

3.1 $p \in NB_1$

As mentioned earlier, the parameters are not identifiable up to a permutation. Let us describe $f^{-1}(p)$ in more details. If $|X| = 1$, then $f^{-1}(p)$ is the set of parameters belonging to Π and such that equations (44) and (45) are satisfied.

Suppose that $|X| \geq 2$. All the variables are mutually independent and hence, by equation (42), we have

$$g(S) = 0 \tag{55}$$

for $S \geq 1$. Hence, equation (47) becomes

$$\delta_1(i)\delta_2(j) = 0. \tag{56}$$

$f^{-1}(p)$ is the set of parameters that belongs to Π, satisfy this equation, equation (44) and equation (45).

3.2 $p \in D_1 \cap NB_2$

For some $a_0 \in v(a)$ and $b_0 \in v(b)$, we have

$$g(a_0, b_0) \neq 0. \tag{57}$$

Hence, equation (49) implies that

$$\delta_1(a_0)\delta_2(b_0) = \delta_1(b_0)\delta_2(a_0) \neq 0. \tag{58}$$

By equation (21), we have

$$\omega_1(\delta_2(a_0) - \delta_1(a0)) = \delta_2(a_0). \tag{59}$$

Recall that $p \notin NB_1$ implies that $\omega_1 \neq 0$. Hence, we have $\delta_2(a_0) \neq \delta_1(a0)$ and

$$\omega_1 = \frac{\delta_2(a_0)}{\delta_2(a_0) - \delta_1(a0)}. \tag{60}$$

Equation (49) yields

$$\delta_1(e) = -\frac{g(a_0, e)}{\delta_2(a_0)}, \tag{61}$$
$$\delta_2(e) = -\frac{g(a_0, e)}{\delta_1(a_0)}, \tag{62}$$

for $e \neq a$. Moreover, by equation (50), we have

$$\delta_t(a) = \delta_t(a_0)\frac{g(a, b_0)}{g(a_0, b_0)}. \tag{63}$$

$p \in D_1$ implies that $e \perp a$, for $e \neq a$ and $e \neq b$. Hence, for all such e we have

$$\delta_1(e) = \delta_2(e) = 0, \tag{64}$$

and p can be represented by a Naive Bayes network where only a and b are connected to the hidden variable.

By equation (42), the independence relationships between the random variables imply that

$$g(S) = 0 \tag{65}$$

for $S \neq \{a, b\}$ and $S \neq \phi$. Introducing the parameters already identified, equation (47) reduces to

$$g(a, b) = \frac{g(a, b_0)g(a_0, b)}{g(a_0, b_0)}. \quad (66)$$

This equation does not imply any additional constraint on the parameters. In particular, it can not be used for the identification of $\delta_1(a_0)$ and $\delta_2(a_0)$.

To ensure that the parameters belong to Π, equations (16) to (21) must be satisfied. Equation (43) implies that equation (20) holds. Similarly, one can see that equation (21) holds. Equations (16) and (17) are equivalent to

$$\delta_1(a_0)\delta_2(a_0) \leq 0, \quad (67)$$

$$\delta_t(a_0)\frac{g(a, b_0)}{g(a_0, b_0)} \leq p(a), \quad (68)$$

$$-\frac{g(a_0, b)}{\delta_t(a_0)} \leq p(b). \quad (69)$$

Hence, $f^{-1}(p)$ is the set of parameters given by equations (44), (45) and (60) to (64) and with $\delta_1(a_0)$ and $\delta_2(a_0)$ satisfying equations (67) to (69) and such that

$$\delta_1(a_0) \neq \delta_2(a_0), \quad (70)$$

$$\delta_t(a_0) \neq 0. \quad (71)$$

We see that, in general, the parameters are not identifiable up to a permutation.

3.3 $p \in D_2 \cap NB_2$

For some $a_0 \in v(a)$, $b_0 \in v(b)$, $c_0 \in v(c)$ and $d_0 \in v(d)$, we have

$$g(a_0, b_0) \neq 0, \quad (72)$$

$$g(c_0, d_0) \neq 0. \quad (73)$$

Equation (49) implies that $g(a_0, c_0) \neq 0$ and $g(b_0, c_0) \neq 0$, i.e. $a \not\perp c$ and $b \not\perp c$.

Equations (60) to (63) are still applicable. To identify $\delta_1(a_0)$ and $\delta_2(a_0)$, we simply use equations (51) and (52) and obtain

$$\delta_1(a_0) + \delta_2(a_0) = -\frac{g(a_0, b_0, c_0)}{g(b_0, c_0)}, \quad (74)$$

$$\delta_1(a_0)\delta_2(a_0) = -\frac{g(a_0, b_0)g(a_0, c_0)}{g(b_0, c_0)}. \quad (75)$$

We conclude that $f^{-1}(p)$ contains two elements and that the parameters are identifiable up to a permutation.

4 A semi-algebraic description of NB_2

Using our identification results, let us derive the semi-algebraic description of NB_2 given by the following theorem.

Theorem 4.1 *Let p be a distribution over X. $p \in NB_2$ if and only if*

$$g(l, m)g(S) = g(S \setminus \{i\})g(i, l, m) + g(j, l)g(k, m)g(S \setminus \{j, k\}), \quad (76)$$

and

$$g(i, j)g(i, k)g(j, k) \geq 0, \quad (77)$$

$$g(j, k)(g(i, j, k) + 2p(i)g(j, k)) \geq 0, \quad (78)$$

$$g(j, k)(p(i)^2 g(j, k) + p(i)g(i, j, k) - g(i, j)g(i, k)) \geq 0. \quad (79)$$

Note that the number of equations and inequations represented by these expressions can be very large. This could cause algorithmic complexity issues for model selection.

Let us show that $p \in NB_2$ implies equations (76) to (79). Equation (76) is obtained by multiplying both sides of equation (47) by $g(l, m)$ and noting that

$$-(\delta_1(i) + \delta_2(i))g(l, m) = g(i, l, m), \quad (80)$$

$$g(l, m)g(j, k) = \delta_1(l)\delta_1(m)\delta_1(k)\delta_1(j) \quad (81)$$

$$= g(j, l)g(k, m) \quad (82)$$

by corollary 3.2. Furthermore, equation (21) implies that $\delta_1(i)\delta_2(i) \leq 0$ and thus

$$g(i, j)g(i, k)g(j, k) = -\delta_1(i)\delta_2(i)\delta_1^2(j)\delta_2^2(k) \geq 0. \quad (83)$$

Finally, note that corollary 3.2 also implies that equations (78) and (79) are equivalent to

$$g(j, k)^2((p(i) - \delta_1(i)) + (p(i) - \delta_2(i))) \geq 0, \quad (84)$$

$$g(j, k)^2(p(i) - \delta_1(i))(p(i) - \delta_2(i)) \geq 0, \quad (85)$$

and these equations hold because the parameters are in Π.

Suppose now that a distribution p over X is given and that it satisfies equations (76) to (79). First note that the following equations hold

$$g(i, j)g(k, l) = g(i, k)g(j, l), \quad (86)$$

$$g(i, j)g(k, l, m) = g(i, j, k)g(l, m). \quad (87)$$

Let us now show that $p \in NB_2$. Once again, we partition the set of distributions over X into NB_1, D_1 and D_2.

4.1 $p \in NB_1$

Trivially, $p \in NB_1$ implies that $p \in NB_2$.

4.2 $p \in D_1$

Let us show that $p \in NB_2$ by finding parameters generating it.

Without loss of generality, suppose that

$$g(a_0, b_0) \neq 0. \tag{88}$$

Consider the set

$$A = \{-\frac{g(a_0, b_i)}{p(b_i)} | b_i \in v(b) \text{ and } p(b_i) \neq 0\}. \tag{89}$$

Let us choose

$$\delta_1(a_0) = \max A, \tag{90}$$
$$\delta_2(a_0) = \min A, \tag{91}$$

and the remaining parameters given by equations (60) to (64), equation (44) and equation (45).

Let us check that these parameters are well defined and belong to Π. By equation (43), we have

$$\sum_{b_i \in v(b)} g(a_0, b_i) = 0. \tag{92}$$

Together with equation (88), this implies that A contains strictly positive and strictly negative elements. Hence, we have $\delta_1(a_0) > 0$, $\delta_2(a_0) < 0$ and equations (70) to (67) hold. For some $b_i \in v(b)$ depending on t, we have

$$\delta_t(a_0) = -\frac{g(a_0, b_i)}{p(b_i)}. \tag{93}$$

By equation (86), equation (68) is thus equivalent to

$$-g(a, b_i) \leq p(a)p(b_i), \tag{94}$$

Hence, it is also equivalent to

$$p(a, b_i) \geq 0, \tag{95}$$

which holds because p is a probability distribution. Equation (69) trivially holds for our choice of parameters.

Finally, note that equation (76) implies equation (66). Therefore, by lemma 3.1, p is generated by the parameters chosen.

4.3 $p \in D_2$

Again, let us show that $p \in NB_2$ by finding parameters generating it.

Without loss of generality, suppose that

$$g(a_0, b_0) \neq 0, \tag{96}$$
$$g(a_0, c_0) \neq 0, \tag{97}$$
$$g(b_0, c_0) \neq 0. \tag{98}$$

Let us choose one of the two elements of $f^{-1}(p)$ that were given in section 3.3. We have

$$\delta_1(i) + \delta_2(i) = -\frac{g(a_0, i)g(a_0, b_0, c_0)}{g(a_0, b_0)g(a_0, c_0)}, \tag{99}$$
$$\delta_1(i)\delta_2(i) = -\frac{g(a_0, i)^2 g(b_0, c_0)}{g(a_0, b_0)g(a_0, c_0)}, \tag{100}$$
$$\delta_1(a)\delta_2(i) = -\frac{g(a, b_0)g(a_0, i)}{g(a_0, b_0)}, \tag{101}$$
$$\delta_1(i)\delta_2(j) = -\frac{g(a_0, i)g(a_0, j)g(b_0, c_0)}{g(a_0, b_0)g(a_0, c_0)}, \tag{102}$$
$$\delta_1(a) + \delta_2(a) = -\frac{g(a, b_0)g(a_0, b_0, c_0)}{g(a_0, b_0)g(b_0, c_0)}, \tag{103}$$
$$\delta_1(a)\delta_2(a) = -\frac{g(a, b_0)^2 g(a_0, c_0)}{g(a_0, b_0)g(b_0, c_0)}, \tag{104}$$

for $a \neq i, j$. By equations (86) and (87), these equations simplify to

$$\delta_1(i)\delta_2(j) = -g(i, j), \tag{105}$$
$$\delta_1(i) + \delta_2(i) = -\frac{g(a_0, b_0, i)}{g(a_0, b_0)}, \tag{106}$$
$$\delta_1(b) + \delta_2(b) = -\frac{g(a_0, b, c_0)}{g(a_0, c_0)}, \tag{107}$$
$$\delta_1(a) + \delta_2(a) = -\frac{g(a, b_0, c_0)}{g(b_0, c_0)}, \tag{108}$$
$$\delta_1(i)\delta_2(i) = -\frac{g(a_0, i)g(b_0, i)}{g(a_0, b_0)}, \tag{109}$$
$$\delta_1(b)\delta_2(b) = -\frac{g(a_0, b)g(b, c_0)}{g(a_0, c_0)}, \tag{110}$$
$$\delta_1(a)\delta_2(a) = -\frac{g(a, b_0)g(a, c_0)}{g(b_0, c_0)}, \tag{111}$$

for $i \neq a, b$. Note that these sums and products are expressed as quotients of polynomials.

Let us now ensure that the parameters chosen belong to Π. As noted before, equations (18) to (21) are satisfied. Moreover, one can see that equations (16) and (17) are equivalent to

$$\delta_1(a_0)\delta_2(a_0) \leq 0, \tag{112}$$
$$\delta_1(i) + \delta_2(i) \leq 2p(i), \tag{113}$$
$$p(i)^2 - p(i)(\delta_1(i) + \delta_2(i)) + \delta_1(i)\delta_2(i) \geq 0. \tag{114}$$

Equations (77) to (79) imply equations (112) to (114).

Finally, we see that equation (76) implies equation (47). Hence, by lemma 3.1, p is generated by the parameters chosen.

5 Conclusion

In this paper, we first studied the inversion of the parametrization mapping of the discrete Naive Bayes models with two hidden classes. Then we applied these results to eliminate the parameters and obtain a complete semi-algebraic description. To our knowledge, these two contributions are original.

Several lines of future research seem interesting to us. First, this work should be generalized to Naive Bayes models with an arbitrary number of hidden classes. We tried to expose our ideas with this in mind. Second, our results should be compared with previous work and, if possible, stated in the algebraic framework of Garcia. Finally, they should be applied to model selection.

6 Appendix

Before tackling lemma 3.1, let us prove the following result.

Lemma 6.1 *Given a distribution p over X and $(\Lambda, ((\omega_t, \Delta_t)_{t=1}^s)) \in \Pi$ such that $\lambda(i) = p(i)$, we have*

$$p = f(\Lambda, ((\omega_t, \Delta_t)_{t=1}^s)) \tag{115}$$

if and only if

$$g(S) = (-1)^{|S|} \sum_{t=1}^{s} \omega_t \prod_{i \in S} \delta_t(i). \tag{116}$$

Proof of lemma 6.1. First, note that we have

$$\begin{aligned} &\sum_{t=1}^{s} \omega_t \prod_{i \in S} (p(i) - \delta_t(i)) \\ &= \sum_{t=1}^{s} \omega_t \sum_{P \text{ s.t. } P \subseteq S} \Big(\prod_{i \in S \setminus P} p(i) \Big) \Big(\prod_{i \in P} -\delta_t(i) \Big), \\ &= \sum_{P \text{ s.t. } P \subseteq S} \Big(\prod_{i \in S \setminus P} p(i) \Big) (-1)^{|P|} \Big(\sum_{t=1}^{s} \omega_t \prod_{i \in P} \delta_t(i) \Big). \end{aligned} \tag{117}$$

Suppose that (115) is satisfied. Let us prove by induction on $|S|$ that equation (116) holds. For $S = \phi$, we have

$$g(\phi) = 1 = \sum_{t=1}^{s} \omega_t. \tag{118}$$

For $|S| \geq 1$, by definition of g (see equation (35)) and the inductive hypothesis, we have

$$\begin{aligned} &g(S) = p(S) \\ &- \sum_{P \text{ s.t. } P \subset S} \Big(\prod_{i \in S \setminus P} p(i) \Big) (-1)^{|P|} \Big(\sum_{t=1}^{s} \omega_t \prod_{i \in P} \delta_t(i) \Big). \end{aligned} \tag{119}$$

Marginalising equation (115), we obtain

$$p(S) = \sum_{t=1}^{s} \omega_t \prod_{i \in S} (p(i) - \delta_t(i)) \tag{120}$$

Hence, inserting equation (117) into equation (119), we see that equation (116) holds.

Suppose now that (116) holds. By definition of g, we have

$$p(X) = g(X) + \sum_{P \text{ s.t. } P \subset X} \Big(\prod_{i \in X \setminus P} p(i) \Big) g(P). \tag{121}$$

Hence, we have

$$\begin{aligned} &p(X) = \\ &\sum_{P \text{ s.t. } P \subseteq X} \Big(\prod_{i \in X \setminus P} p(i) \Big) (-1)^{|P|} \Big(\sum_{t=1}^{s} \omega_t \prod_{i \in P} \delta_t(i) \Big). \end{aligned} \tag{122}$$

Using equation (117), we see that equation (115) holds. ■

Proof of lemma 3.1. Suppose that equation (46) is satisfied. By equation (21) we have

$$\omega_1 \delta_1(i) + \omega_2 \delta_2(i) = 0. \tag{123}$$

Hence, by lemma 6.1, we have

$$\begin{aligned} g(i,j) &= \omega_1 \delta_1(i) \delta_1(j) + \omega_2 \delta_2(i) \delta_2(j) && (124) \\ &= -(\omega_1 + \omega_2) \delta_1(i) \delta_2(j) && (125) \\ &= -\delta_1(i) \delta_2(j). && (126) \end{aligned}$$

We have

$$\begin{aligned} &- g(S \setminus \{i\})(\delta_1(i) + \delta_2(i)) && (127) \\ &= g(S) - (-1)^{|S|-1} \Big(\delta_1(i) \omega_2 \prod_{v \in S \setminus \{i\}} \delta_2(v) && (128) \\ &\quad + \delta_2(i) \omega_1 \prod_{v \in S \setminus \{i\}} \delta_1(v) \Big) && (129) \\ &= g(S) + \delta_1(j) \delta_2(k) g(S \setminus \{j,k\}). && (130) \end{aligned}$$

Hence, equation (47) holds.

Suppose that equation (47) holds. First note that for $|S| = 2$, this implies that

$$\delta_1(i)\delta_2(j) = -g(i,j) = \delta_2(i)\delta_1(j). \tag{131}$$

Let us now prove by induction on $|S|$ that equation (116) holds for $S \subseteq X$. For $|S| = 0$ and $|S| = 1$, equation (116) simply holds because the parameters belong to Π, in particular by equations (18) and (21). For $|S| \geq 2$, equation (47) yields

$$g(S) = -g(S\backslash\{i\})(\delta_1(i)+\delta_2(i))-\delta_1(j)\delta_2(k)g(S\backslash\{j,k\}) \tag{132}$$

Using the inductive hypothesis and equation (131), a simple computation shows that

$$g(S) = (-1)^{|S|}\left(\omega_1 \prod_{i\in S}\delta_1(i) + \omega_2 \prod_{i\in S}\delta_2(i)\right). \tag{133}$$

Hence, by lemma 6.1, equation (46) holds. ■

References

[Cowell *et al.*, 1999] Robert Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer, New York, 1999.

[Garcia *et al.*, 2005] Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of bayesian networks. *Journal of Symbolic Computation*, 39(3-4):331–355, 2005.

[Garcia, 2004] Luis David Garcia. Algebraic statistics in model selection. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 177–184, Arlington, Virginia, 2004. AUAI Press.

[Geiger *et al.*, 2001] Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29:505–529, 2001.

[Geiger, 1998] Dan Geiger. Graphical models and exponential families. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 156–165, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

[Lowd and Domingos, 2005] Daniel Lowd and Pedro Domingos. Naive bayes models for probability estimation. In Luc de Raedt and Stefan Wrobel, editors, *Proceedings of 22nd International Conference on Machine Learning*, pages 529–536, August 2005.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, 1988.

[Settimi and Smith, 2000] Raffaella Settimi and Jim Q. Smith. Geometry, moments and conditional independence trees with hidden variables. *The Annals of Statistics*, 28:1179–1205, 2000.