Stratified Exponential Families: Graphical Models and Model Selection

Dan Geiger
Microsoft Research
dgeiger@microsoft.com

David Heckerman
Microsoft Research
heckerma@microsoft.com

Henry King
University of Maryland
College Park, Maryland 20742
hck@math.umd.edu

Christopher Meek
Microsoft Research
meek@microsoft.com

July, 1998

Technical Report
MSR-TR-98-31

Microsoft Research
Advanced Technology Division
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Abstract

We provide a classification of graphical models according to their representation as exponential families. Undirected graphical models with no hidden variables are linear exponential families (LEFs), directed acyclic graphical (DAG) models and chain graphs with no hidden variables, including DAG models with several families of local distributions, are curved exponential families (CEFs) and graphical models with hidden variables are stratified exponential families (SEFs). A SEF is a finite union of CEFs of various dimensions satisfying some regularity conditions. The main results of this paper are that graphical models are SEFs and that many graphical models are not CEFs. That is, roughly speaking, graphical models when viewed as exponential families correspond to a set of smooth manifolds of various dimensions and usually not to a single smooth manifold. These results are discussed in the context of model selection.

Keywords : Bayesian networks, graphical models, hidden variables, curved exponential families, stratified exponential families, semi-algebraic sets, model selection.

1 Introduction

A graphical model is a family of probability distributions specified via a set of conditional independence constraints that a graph represents or via a parametric definition dictated by a graph. The wide applicability of graphical models to many problems in Statistics is due to several features. Graphical models provide a language to facilitate communication between a domain expert and a statistian, provide flexible and modular definitions of families of probability distributions, and are amenable to scaleable computational techniques (e.g., Pearl, 1988; Whittaker, 1990; Lauritzen, 1996). Furthermore, graphical models based on directed acyclic graphs (DAGs), which are called DAG models or Bayesian networks, are useful for modeling causal relationships. For this reason, the problem of model selection has been examined for the purpose of identifying cause and effect from observational data (e.g., Spirtes et al., 1993). (e.g., Spirtes et al., 1993, Pearl, 1997).

We provide a classification of graphical models according to their representation as exponential families. Undirected graphical models with no hidden variables are known to be linear exponential families (LEFs) (see Lauritzen, 1996), directed acyclic graphical models and chain graphs with no hidden variables, including DAG models with several families of local distributions, are shown to be curved exponential families (CEFs), and graphical models with hidden variables are, what we term stratified exponential families (SEFs). An SEF is a finite union of CEFs of various dimensions satisfying some regularity condition. The main results of this paper are that graphical models are SEFs and that many graphical models are not CEFs. That is, roughly speaking, graphical models when viewed as exponential families correspond to a set of smooth manifolds of various dimensions and usually not to a single smooth manifold.

This classification is motivated by results on model selection within linear and curved exponential families. A Bayesian approach to model selection is to compute the probability that the data is generated by a given model via integration over all possible parameter values with which the model is compatible and to select a model that maximizes this probability. We call this probability the marginal likelihood. Although, in principle, this Bayesian approach is appealing, in practice, it is often impossible to evaluate the integral (even by sampling techniques) when the number of parameters is large. When the dataset consists of many cases, asymptotic results for approximating the marginal likelihood are useful.

Schwarz (1978) considered the problem of evaluating the marginal likelihood when a model is an affine subspace of the natural parameter space of an exponential family. He derived an asymptotic formula for the marginal likelihood, $P(Data|Model) = L(\hat{\theta})N - d/2 \log N + O_p(1)$, where L is the likelihood, $\hat{\theta}$ is the maximum likelihood estimator, d is the dimension of the affine subspace, and N is the sample size. This formula has become known as the Bayesian Information Criteria (BIC). We note that Schwarz's original result applies to the undirected graphical models discussed in Section 2, because these models define a linear subspace of the natural parameter space.

Haughton (1988) established, among other results, that BIC, under some regularity assumptions, is an $O_p(1)$ asymptotic approximation of the marginal likelihood for curved exponential families. The main regularity assumption of her work, and of Schwarz's work, is that the prior distribution expressed in a local coordinate system near the maximum likelihood solution is bounded and bounded away from zero. Other regularity assumptions are used to insure that with sufficient data, a unique model is selected with high probability. When these assumptions are acceptable, Haughton's results on model selection apply to all graphical models discussed in Section 3, since these graphical models are shown to be curved exponential families. In particular these results on model selection apply to DAG models with several families of local distributions including decision trees and leaky noisy-or distributions. Several of these families do not have known closed-form formula for the marginal likelihood.

We note that although researchers have been using BIC for selecting models among graphical models with hidden variables, this methodology has not yet been established as an asymptotic approximation of a Bayesian procedure as it has for CEFs. In Section 4, we show that graphical models with hidden variables are SEFs and usually not CEFs. This characterization implies that the justifications given by Schwartz and Haughton for BIC do not apply to graphical models with hidden variables and that a generalization of their arguments is needed. We offer stratified exponential families as a natural class for which the validity of BIC might be proven.

2 Linear Exponential Families

In this background section we give a definition of linear exponential families (LEFs) and discuss the well-known representation of undirected graphical models as LEFs (e.g., Barndorff-Nielsen, 1978, Lauritzen, 1996, respectively).

2.1 Definition of Linear Exponential Families

A *family* (or model) is a set of probability density functions. A probability density in an exponential family is given by

$$p(x|\eta) = e^{<\eta, t(x)> - \psi(\eta)} \tag{1}$$

where x is an element of a sample space $\mathcal{X}$ with a dominating measure μ and $t(x)$ is a sufficient statistics defined on $\mathcal{X}$ taking values in R^k with an inner product $< .,. >$. The sample space $\mathcal{X}$ is typically either a discrete set, R^n, or a product of these. We use the notion of a variable to describe the product sample space. A variable has a domain which is either finite or R and the product sample space is the cartesian product of the domains for the variables of interest. The quantity $\psi(\eta)$ is the normalization constant.

Every probability distribution for a finite sample space $\mathcal{X}$ belongs to an exponential family. For example, a sample space that consists of four outcomes can be written in the form of Eq. (1) by choosing $t(x)$ and η as follows: $t(x) = (t_1(x), t_2(x), t_3(x))$ where $t_i(x) = 1$ if x is outcome i, $1 \leq i \leq 3$, and zero otherwise, and $\eta_i = \log(w_i/w_0)$ where w_i is the probability of outcome i, $1 \leq i \leq 3$, and $w_0 = 1 - \sum_{i=1}^{3} w_i$ is the probability of the forth outcome.

When the vector η has k coordinates and when $p(x|\eta)$ cannot be represented with a parameter vector smaller than k, then the representation is *minimal* and the *order* (or *dimension*) of this family is k, and the parameters are called *natural parameters*. It is known that this order is unique for each family. The natural parameter space is given by

$$N = \{\eta \in R^k | \int e^{t(x)\eta - \psi(\eta)} d\mu(x) < \infty\}$$

The set of probability distributions having the form (1) are denoted by $\mathcal{S}$. If for each η in N there exists P_η in $\mathcal{S}$, then $\mathcal{S}$ is said to be a *full* exponential family; if, in addition, N is an open subset of R^k, then $\mathcal{S}$ is said to be a *linear* exponential family. The name linear exponential family comes from the fact that the log densities form a vector space over R where the coordinates of $t(x)$, called the canonical statistics, are the basis of the vector space and its dimension is the order of the family. Linear exponential families include many common distribution functions, such as multivariate Normal and multinomial distributions. (A linear exponential family in a minimal representation is often called a *regular* exponential family).

A subfamily of linear exponential family is a subset $\mathcal{S}_0$ of $\mathcal{S}$. A subfamily can be described by a mapping $f : \Theta \rightarrow N$ which defines $\mathcal{S}_0$ via $N_0 = \{f(\theta)|\theta \in \Theta\}$. When f is a linear mapping of rank p, and Θ is an open set, a new linear exponential family is formed of order $k - p$. In other words, a linear transformation f imposes p independent linear constraints on the parameters and these constraints can be used to reparameterize the family with $k - p$ natural parameters. In Sections 3 and 4, we discuss exponential families that are formed by non-linear transformations f.

2.2 Undirected graphical models

In this section, we discuss the representation of undirected graphical models as linear exponential families.

Let G be an undirected graph such that each vertex i in the vertex set corresponds to a variable x_i. We consider three cases: (1) all x_i are discrete; (2) all are continuous and their joint density is a multivariate non-singular Gaussian; (3) some are continuous and some are discrete with a joint Conditional Gaussian (CG) distribution. An *undirected graphical model w.r.t. G* is the set of probability distribution functions such that all of the saturated independence facts implied by the graph hold; that is x_i and x_j are conditionally independent given the remaining variables whenever nodes

i and j are not adjacent in G. Since multinomial, multivariate Gaussian, and CG distributions over a fixed set of variables belong to a linear exponential family and since saturated independence constraints are linear restrictions when expressed in terms of the natural parameters, undirected graphical models define linear exponential families. We now discuss the three cases.

A *Multinomial undirected graphical model* is a family of probability distributions over a finite set U of variables each having a finite domain such that for some set of pairs of indices $\{(i,j)\}$, x_i and x_j are conditionally independent given $U \setminus \{x_i, x_j\}$. Consider, for example, the graph given by a cycle of size 4 with variables $x_1, \ldots, x_4$ arranged clockwise. Then the independence constraints imposed by this graphical model are that x_1 and x_3 are conditionally independent given $\{x_2, x_4\}$, and that x_2 and x_4 are conditionally independent given $\{x_1, x_3\}$. Suppose, for simplicity, that the four random variables are binary (having exactly two states) and denote by w_i the probability of the joint ith state of the four binary variables ($1 \leq i \leq 15$) where $w_0 = 1 - \sum w_i$. Each independence constraint translates to 4 equations of the form $w_i w_j = w_k w_l$. Dividing each equation by $(w_0)^2$ and taking the log, yields 8 linear equations in terms of the natural parameters $\eta_i = \log w_i/w_0$. In general, multinomial graphical models are log-affine models which are LEFs (Lauritzen, 1996, pp 76).

A *Gaussian undirected graphical model* is a family of multivariate non-singular Gaussian distributions in which some of the off-diagonal elements t_{ij} of the precision matrix (the inverse of the covariance matrix) are set to zero. Note that setting t_{ij} to zero is equivalent to requiring that variable x_i and x_j are conditionally independent given the remaining variables. Recalling that a multivariate non-singular Gaussian distribution belongs to a linear exponential family and the fact that setting the off-diagonal elements of the precision matrix to zero is equivalent to placing linear restrictions on the natural parameter space yields the conclusion that Gaussian undirected graphical models are linear exponential families. For details see (Lauritzen, 1996, pp. 124–132).

A *Conditional Gaussian undirected graphical model* is a family of Conditional Gaussian (CG) distributions over a set of discrete and continuous variables defined by a set of saturated independence constraints stating that variables i and j are conditionally independent given the remaining variables. That CG undirected graphical models can be represented as linear exponential families is shown in Lauritzen and Wermuth (1989). See also, Lauritzen (1996, pp. 171–175).

3 Curved Exponential Families

A *curved exponential family* of dimension n is defined to be a subfamily of an exponential family of order k such that N_0 is a n-dimensional smooth manifold in R^k. A subfamily of an exponential family $\mathcal{S}_0 \subseteq \mathcal{S}$ is often described by a mapping $f : \Theta \to N$ which defines $\mathcal{S}_0$ via $N_0 = \{f(\theta)|\theta \in \Theta\}$ and where Θ is an open set. Alternatively, a subfamily can be described by a set of constraints on S_0 given by $N_0 = \{\eta \in R^n | h(\eta) = 0\}$ where $h : R^k \to R^{k-n}$. The relationship of these alternatives and a method, called implicitization, for finding constraints from a mapping f is discussed in (Geiger and Meek, 1998) .

In this section we recall the definitions of smooth manifolds and show that DAG models correspond to smooth manifolds and are therefore curved exponential families (and not linear exponential families). Conditional-Gaussian DAG models and Conditional-Gaussian chain graphs are also curved exponential models.

Curved exponential families were studied by Efron who explored geometrical interpretation of various statistical measures using these families (e.g., Efron, 1978). A treatment of this topic is given by Kass and Vos (1997). We study curved exponential models because the standard asymptotic theory is valid for these models. In particular Haughton's (1988) results on model selection applies to all graphical models discussed in this and the previous section.

3.1 Manifolds

A *diffeomorphism* $f : U \subset R^n \to R^m$ is a smooth (C^∞) 1-1 function having a smooth inverse. A subset M of R^n is called a k-dimensional *smooth manifold* in R^n if for every point $x \in M$ there exists an open set U in R^n containing x and a diffeomorphism $f : U \cap M \to R^k$. When f is only assumed to be continous and to have a continous inverse (namely, a *homeomorphism*), then the set M is called a *topological manifold*. Since composition of diffeomorphisms is a diffeomorphism, we get the following proposition.

Proposition 1 *If $g : A \subset R^n \to B \subset R^n$ is a diffeomorphism, then $M \subseteq A$ is a smooth manifold if and only if $g(M)$ is a smooth manifold and $N \subseteq B$ is a smooth manifold if and only if $g^{-1}(N)$ is a smooth manifold.*

Another way to verify whether a subset of R^n is a smooth manifold is given by the following Theorem (e.g., Spivak, 1965).

Theorem 1 *Let $A \subset R^m$ be open and let $h : A \to R^{m-n}$ be a smooth function such that $h'(x)$ has rank $m-n$ whenever $h(x) = 0$. Then $h^{-1}(0)$ is a n-dimensional smooth manifold in R^m.*

Note that the rank of the Jacobian matrix h' in Theorem 1 is $m-n$ if h has the form $h_i(x_1, \ldots, x_m) = x_{n+i} - f_i(x_1, \ldots, x_n)$ for $i = 1, \ldots, m-n$ where f_i are smooth functions because in this case the $(m-n) \times m$ matrix h' factors as $[Q_{(m-n)\times n}|I_{m-n}]$ where I_{m-n} is the identity matrix of size $m-n$.

3.2 Discrete DAG models

A *Discrete DAG model* $B(\Theta, n, m)$ is a mapping $B_{n,m} : \Theta \subset R^n \to R^m$ where Θ, n, m and $B_{n,m}$ are given as follows (Pearl, 1988). Let $(x_1, \ldots, x_k)$ be an ordered sequence of discrete variables each having a finite set of values. Let p_i be a subset of $\{x_1, \ldots, x_{i-1}\}$, called the *parents set* of x_i, and let $u_i = \{x_1, \ldots, x_{i-1}\} \setminus p_i$. Let x_i^j, p_i^j and u_i^j be the jth value of x_i, p_i and u_i with $j \geq 0$. Let $|x_i|$, $|p_i|$ and $|u_i|$ be the domain sizes respectively. The components of $B_{n,m} : \Theta \subseteq R^n \to R^m$ are defined by $\theta_{x_i^a|p_i^b,u_i^c} = \theta_{x_i^a|p_i^b}$, for all $a > 0$, $b \geq 0$, and $c \geq 0$. Note that there are $n = \sum_i (|x_i| - 1)|p_i|$ source coordinates denoted by $\theta_{x_i^a|p_i^b}$ and $m = \sum_i (|x_i| - 1)|p_i||u_i| = (\prod_i |x_i|) - 1$ target coordinates denoted by $\theta_{x_i^a|p_i^b,u_i^c}$. The set Θ is the cartesian product of $\Theta_{i,j}$ over i and j where $\Theta_{i,j} = \{(\theta_{x_i^1|p_i^j}, \ldots, \theta_{x_i^{|x_i|-1}|p_i^j})|0 < \theta_{x_i^k|p_i^j} < 1, \sum_{k>0} \theta_{x_i^k|p_i^j} < 1\}$. The target coordinates of $B_{n,m}$ are called the *conditional-space parameters.*

Theorem 2 *For every Discrete DAG model $B(\Theta, n, m)$ the set $B_{n,m}(\Theta)$ is an n-dimensional smooth manifold in R^m.*

Proof: Define the components of a function h by $h_{i,a,b,c}(\theta) = \theta_{x_i^a|p_i^b,u_i^c} - \theta_{x_i^a|p_i^b,u_i^0}$ where $a > 0$, $b \geq 0$ and $c > 0$. Thus, h has $\sum_i (|x_i|-1)|p_i|(|u_i|-1) = m-n$ components. In other words, h imposes $m-n$ constraints on the target coordinates $\theta_{x_i^a|p_i^b,u_i^c}$. Note that in light of the definition of h and $B_{n,m}$, we have $h^{-1}(0) = B_{n,m}(\Theta)$. Also note that h' has the form $[Q_{(m-n)\times n}|I_{m-n}]$ where I_{m-n} is the identity matrix and so h' has full rank. Thus, according to Theorem 1, $B_{n,m}(\Theta)$ is a n-dimensional smooth manifold in R^m. $\Box$

A second definition of a discrete DAG model $\hat{B}$ is obtained by defining $\hat{B}_{n,m}$ with the equations: $w_{x_1^{i_1},\ldots,x_k^{i_k}} = \prod_{i=1}^k \theta_{x_i^b|p_i^c}$ where x_i^j is the j-th value of x_i and p_i^c is the c-th value of p_i obtained by the projection of $(x_1^{i_1}, \ldots, x_k^{i_k})$ to the coordinates that correspond to the variables in p_i. The mapping $B_{n,m}(\Theta) \rightarrow \hat{B}_{n,m}(\Theta)$ is a diffeomorphism for positive θ values and so the conclusion of Theorem 2 remains valid under this definition. The components of the image of Θ under $\hat{B}_{n,m}$ are called the *joint-space parameters.*

The practical significance of DAG models stems, among other reasons, from the small number of network parameters compared to the number of joint-space parameters. When the number of network parameters is still too large because $|p_i|$ is too large for some i's, additional factorizations are usually introduced. These include decision tree and decision graph models (Friedman and Goldszmidt 1996; Chickering, Meek, and Heckerman, 1997), noisy-or gates, leaky noisy-or gates, max-gates and causal independence models (Pearl, 1988; Henrion, 1987; Heckerman and Breese, 1996; Meek and Heckerman, 1997). These models share the following characteristic.

For each variable x_i in the DAG model, a subset of k_i states of p_i are designated as *reference states.* The components of $B_{n,m} : \Theta \subset R^n \rightarrow R^m$ are defined by $\theta_{x_i^a|p_i^b,u_i^c} = f_i(\theta_{x_i^a|p_i^0}, \ldots, \theta_{x_i^a|p_i^{k_i-1}})$ for all $a > 0$, $b \geq k_i$, and $c \geq 0$ where f_i are smooth functions. We call DAG models defined in this way DAG models with *explicit local constraints.* The number of network parameters is given by $n = \sum_i (|x_i| - 1)k_i$ where k_i is often much smaller than p_i.

When the number of reference states is zero, namely each f_i is the constant function, we get a discrete DAG model. In the case of a noisy-or model, the reference states are the states where exactly one parent is on and the other parents are off (see Pearl 1988). For leaky noisy-or model the reference states also include the state when all the parents of x_i are off. For decision tree models, the reference states are those that correspond to a path from the root to a leaf in the decision tree; all parents on the path are at a specified state and all those not on the path are at state zero. Note that for decision trees, noisy-or and leaky noisy-or models the functions f_i are all polynomial functions.

Theorem 3 *For every discrete DAG model $B(\Theta, n, m)$ having explicit local constraints the set $B_{n,m}(\Theta)$ is an n-dimensional smooth manifold in R^m.*

Proof: Suppose the local constraints are given by f_i. Define the components of a function h by

$$h_{a_i,b_i,c_i}(\theta) = \\ \theta_{x_i^a|p_i^b,u_i^c} - f_i(\theta_{x_i^a|p_i^0 u_i^0}, \ldots, \theta_{x_i^a|p_i^{k_i-1} u_i^0})$$

where ($a > 0,\ b \geq 0,\ c > 0$) or ($a > 0,\ b \geq k_i,\ c = 0$). Note that h has $\sum_i (|x_i| - 1)\,[|p_i|(|u_i| - 1) + (|p_i| - k_i)] = m - n$ components. The conclusion now follows from Theorem 1 and the comment that follows. □

Recall that for a multinomial distribution with u states each associated with a positive parameter w_i such that $\sum_i w_i = 1$, the map $\eta_i = \log w_i / w_0$, $i = 1, \ldots, u-1$ defines a diffeomorphism between the natural parameter space η and the multinomial parameters $\{w_i\}_0^{u-1}$. Consequently, due to Theorem 2, we have established the following claim.

Theorem 4 *Every discrete DAG model $B(\Theta, n, m)$ with explicit local constraints is a curved exponential family of dimension n.*

3.3 Gaussian graphical models

The parameters of a multivariate non-singular Gaussian distribution can be described in various ways. The most common representation is by the elements of a covariance matrix Σ and a vector of means μ. A second representation is by a precision matrix Σ^{-1} and μ. These two representations are related by the diffeomorphism $f : \Sigma \rightarrow \Sigma^{-1}$. A third representation is constructed as follows. Assign a total order to the k variables. Specify the regression coefficients $b_{i,j}$ of x_i given $x_1, \ldots, x_{i-1}$, and the conditional variance and conditional means of x_i given $x_1, \ldots, x_{i-1}$. The third representation is called the *regression parameterization* and is related to the second representation by a well-known diffeomorphism (e.g., Shachter and Kenley, 1989).

A *Gaussian DAG model* is a family of multivariate non-singular Gaussian distributions in which some b_{ij} are set to zero (Shachter and Kenley, 1989). A *Gaussian undirected graphical model* was defined in Section 2.2 to be a family of multivariate non-singular Gaussian distributions in which some of the off-diagonal elements of the precision matrix are set to zero. Both models define a map $B_{n,m} : \Theta \subset R^n \rightarrow R^m$. It follows from Theorem 1 that $B_{n,m}(\Theta)$ is a n-dimensional smooth manifold in R^m since the components of h can be defined as projections and so h' has the form $[Q_{(m-n)\times n} | I_{m-n}]$ where I_{m-n} is the identity matrix and Q is a matrix of zeros.

The difference between the two models is that the restrictions formed by setting elements of the precision matrix to zero define linear constraints in the natural parameter space and therefore Gaussian undirected graphical models are also LEFs while the restrictions set by a Gaussian DAG model are not linear in the natural parameter space. To demonstrate the latter fact we note that the restriction $b_{31} = 0$ imposed by the Gaussian DAG model $x_1 \to x_2 \leftarrow x_3$ can, in terms of the precision parameters, be written as $t_{1,2}t_{3,3} = t_{1,3}t_{2,3}$ and thus is not linear in the natural parameter space. See Geiger and Heckerman (1994) for the relationships between $t_{i,j}$ and $b_{i,j}$ for this three-node model.

We note that Spirtes, Richardson, and Meek (1997) show that Gaussian MAGs define smooth manifolds. Since Gaussian MAGs are a generalization of Gaussian DAG model, their results also imply that Gaussian DAG models define smooth manifolds.

4 Stratified Exponential Families

This section is divided into four parts. First, we provide some mathematical background, then we define stratified exponential families (SEFs), and show that graphical models representing discrete, Gaussian, and Conditional-Gaussian with or without hidden variables are SEFs. In Section 4.3, we show that graphical models with hidden variables are usually not CEFs and in the final section we discuss a method to compute the dimension of a parametrically-defined SEF.

4.1 Mathematical Prerequisites

The set of all polynomials in $x_1, \ldots, x_n$ with real coefficients is denoted by $R[x_1, \ldots, x_n]$. Let $q_1, \ldots, q_t$ be polynomials in $R[x_1, \ldots, x_n]$. A *variety* $\mathbf{V}(q_1, \ldots, q_t)$ is the set $\{(x_1, \ldots, x_n) \in R^n | q_i(x_1, \ldots, x_n) = 0 \text{ for all } 1 \leq i \leq t\}$. A variety is also called an *algebraic set.*

A subset V of R^n is called a *semi-algebraic set* if $V = \cup_{i=1}^{s} \cap_{j=1}^{r_i} \{x \in R^n | P_{i,j}(x) \Leftrightarrow_{ij} 0\}$ were P_{ij} are polynomials in $R[x_1, \ldots, x_n]$ and $\Leftrightarrow_{ij}$ is one of the three comparison operators $\{<, =, >\}$. Loosely speaking, a semi-algebraic set is simply a set that can be described with a finite number of polynomial equalities and inequalities. A variety is clearly a semi-algebraic set.

A map $f : X \to Y$ where $X \subseteq R^n$ and $Y \subseteq R^m$ are semi-algebraic sets, is called *semi-algebraic* if the graph of f is a semi-algebraic set of R^{n+m}. Note that if f is a polynomial map then f is a semi-algebraic map because its graph can be described by m polynomial equalities: $y_j - f_j(x) = 0$, where $1 \leq j \leq m$. A key result about semi-algebraic sets is given by the Tarski-Seidenberg theorem (see, e.g., Benedetti and Risler, 1990).

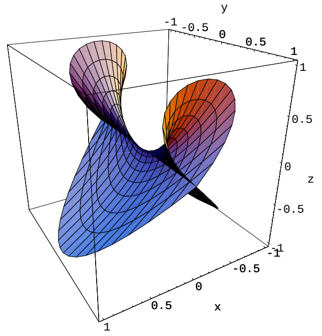

Figure 1: A plot of part of the variety $\mathbf{V}(x^2 - y^2z^2 + z^3)$.

Theorem 5 (Tarski-Seidenberg) *Let $f : X \to Y$ be a semi-algebraic map. Then the image $f(X) \subseteq Y$ is a semi-algebraic set.*

We note that some smooth manifolds are semi-algebraic sets and some are not. Similarly, some semi-algebraic sets are smooth manifolds and some are not. Consider, for example, the variety $\mathbf{V}(x^2 - y^2z^2 + z^3)$ which can be described parametrically as a (two dimensional) surface in R^3 by $x = t(u^2 - t^2)$, $y = u$, and $z = u^2 - t^2$ (see plot in Figure 1). This variety is not a smooth manifold because, locally, at each point of the y-axis other than the origin the surface looks like the intersection of two smooth manifolds, as evident from the figure. To prove that the variety $\mathbf{V}(x^2 - y^2z^2 + z^3)$ is not a smooth manifold it suffices to observe that as we approach any point on the y-axis other than the origin we have two (two dimensional) tangent planes where each plane contains a tangent vector that is not spanned by the other tangent plane.

Another important result about semi-algebraic sets is that they admit a *stratification*. We will first illustrate this concept with the variety $\mathbf{V}(x^2 - y^2z^2 + z^3)$. This variety can be described as a union of several 2-dimensional smooth manifolds along with a 1-dimensional smooth manifold— the y-axis. These smooth manifolds define a stratification of the variety.

Formally, a *stratification* of a subset E of R^m is a finite partition $\{A_i\}$ of E such that (1) each A_i (called a *stratum* of E) is a d_i-dimensional smooth manifold in R^m and (2) if $A_j \cap \overline{A_i} \neq \emptyset$, then $A_j \subseteq \overline{A_i}$ and $d_j < d_i$ (frontier condition) where $\overline{A_i}$ is the closure of A_i in R^m. See Akbulut and King (1992) for a more general definition.

A stratification is called *semi-algebraic* if every stratum is semi-algebraic. A *stratified set* is a set that has a stratification. The dimension of a stratified set is d_1— the largest dimension of a stratum. A key theorem about semi-algebraic sets is the Stratification theorem (see Benedetti and Risler, 1990).

Theorem 6 (Stratification) *Every semi-algebraic set has a semi-algebraic stratification.*

We note that if E is a stratified set and f is a diffeomorphism, then $f(E)$ is also a stratified set. This proposition, that stratification is preserved under a diffeomorphism f, is proven as follows. Let $\{A_i\}$ be a stratification of A. We show that $\{f(A_i)\}$ is a stratification of $f(A)$. Clearly, $\{f(A_i)\}$ is a partition of $f(A)$. Due to Proposition 1, the image of a smooth manifold A_i under a diffeomorphism f is a smooth manifold $f(A_i)$ and so condition (1) of the definition of stratified sets is satisfied. The frontier condition is satisfied because $A_i \subseteq \overline{A_j}$ implies $f(A_i) \subseteq f(\overline{A_j})$ which, due to continuity of f, implies $f(A_i) \subseteq \overline{f(A_j)}$ as needed for satisfying the frontier condition.

4.2 SEFs and Graphical models

We define a *stratified exponential family (*SEF*)* of dimension n as a subfamily of an exponential family having a natural parameter space N of order k if its parameter space $N_0 \subset N$ is a n-dimensional stratified set in R^k. In this section we show that N_0 defined by some graphical models with or without hidden variables is a stratified set because it is a semi-algebraic set or diffeomorphic to one. Consequently, these models are SEFs.

All graphical models considered in the previous sections are SEFs because LEFs and CEFs are subsets of SEFs. Every one of these models is a set of distributions that satisfy all the independence constraints represented by a graph g. For multinomial and Gaussian graphical models an independence fact is expressible as a finite set of polynomial equalities. Combined with the inequalities which state that multinomial parameters are positive, and that variances are positive, respectively, the resulting graphical model corresponds to a semi-algebraic set.

There are several classes of graphical models defined by a set of conditional independence constraints that can accommodate a combination of discrete and continuous variables using Conditional-Gaussian distributions. Among these models, in addition to the models discussed in the previous sections, are AMP chain graphs (Andersson, Madigan, and Perlman, 1996), and reciprocal graphs (Koster, 1997). These graphical models all correspond to semi-algebraic sets because independence facts in CG-distributions are expressible as polynomial equalities.

We now discuss graphical models with hidden variables. In particular we show that multinomial DAG models with hidden variables correspond to semi-algebraic sets. We note that a similar claim holds for any graphical model representing CG-distributions of which we are aware as long as the distribution over the observable variables is in the exponential family.

A *discrete DAG model* $B(\Theta, n, m)$ with hidden variables is a DAG model where Θ, n, m and $B_{n,m}$ are given as follows. Let $(x_1, \ldots, x_k)$ be an ordered sequence of discrete variables each having a finite set of values. Partition this set of variables into two disjoint non-empty sets H and X. The variables in H are *hidden*. Those in X are *observable*. For each x_i define two disjoint subsets of $\{x_1, \ldots, x_{i-1}\}$, the *observable parents* $p_i \subseteq X$ and the *hidden parents* $h_i \subseteq H$.

The components of $B_{n,m} : \Theta \subseteq R^n \to R^m$ are defined by $w_a = \sum_b \prod_{i=1}^k \theta_{x_i^a | p_i^a, h_i^b}$ where a are (vector) values of the observed variables X not all zero and b are (vector) values of the hidden variables H. The values x_i^a and p_i^a are obtained by obtained by the projection of a to the coordinates that correspond to x_i, p_i. Similarly, the value h_i^b is obtained from b. As before, the domain Θ of $B_{n,m}$ is the cartesian product of sets of the form $\{(t_1, \ldots, t_{|x_i|-1}) | 0 < t_a < 1, \sum_a t_a < 1\}$. Note that $n = \sum_{i=1}^k (|x_i| - 1)|p_i||h_i|$ and $m = \prod_{i=1}^k |x_i| - 1$.

The Tarski-Seidenberg theorem guarantees that for a discrete DAG model with hidden variables, $B_{n,m}(\Theta)$ is a semi-algebraic set because it is the image of a semi-algebraic set under a polynomial mapping. Similarly, we note that Gaussian DAG model with hidden variables also correspond to semi-algebraic sets due to their parametric definition via a polynomial mapping called the trek-rule (see, e.g., Spirtes et al. 1993). Consequently, the image of these graphical models can be described with a set of polynomial equalities and polynomial inequalities.

We have thus shown that N_0 defined by each of the models considered in this paper is a stratified set because it is a semi-algebraic set or diffeomorphic to one.

4.3 Graphical models with hidden variables are not CEFs

It is clear that SEFs is a class of models that is strictly larger than CEFs, however, it remains to show that the new class contains models that are used in practice which are not contained in the smaller class. In this section we show that many graphical models with a hidden variable are not CEFs.

We first study in detail a class of graphical models which are often called *naive Bayes models (NBM)*. We show that naive Bayes models are stratified exponential families but are usually not curved exponential families. Then we extend the proof to wider classes of graphical models.

Let $H, F_1, \ldots, F_n$ be a set of variables each having a finite set of possible values denoted by $dom(H), dom(F_i)$, respectively. Let $|dom(H)| = k$ and $|dom(F_i)| = k_i$ and let $p(h)$ stand for $p(H = h)$ where $h \in dom(H)$. A naive Bayes model is a set of distributions for the sample space $dom(F_1) \times \ldots \times dom(F_n)$ such that

$$p(f_1, \ldots, f_n) = \sum_{h \in dom(H)} p(h) \prod_{i=1}^{n} p(f_i|h), \tag{2}$$

where $f_i \in dom(F_i)$. The variable H is called the *class variable* and each F_i is called a *feature*. When $k = 2$ we get a *Binary naive Bayes model* and when $k_i = 2$ the feature F_i is binary and its domain is $\{f_i, \overline{f}_i\}$. In applications, H denotes a mutually exclusive and exhaustive set of classes and each F_i is a measurement that has a finite set of possible outcomes denoted by $dom(F_i)$. By observing outcomes of F_i, a common task is to infer how many classes should H have, or when the number of classes is known, to find the most likely class given the measurements. We focus on inferring the number of classes, and more generally on model selection.

We note that Eq. 2 defines a mapping $g^{n,k,k_1,\ldots,k_n} : A \subseteq R^{\hat{n}} \to R^m$ where $\hat{n} = k - 1 + \sum_{i=1}^{n}(k_i - 1)k$ is the number of coordinates on the right hand side and $m = (\prod_{i=1}^{n} k_i) - 1$ is the number of coordinates on the left hand side minus one (since these coordinates sum to 1). The set A is an open set of $R^{\hat{n}}$ defined by the following inequalities. For each $h \in dom(H)$ and $f_i \in dom(F_i)$, $1 \leq i \leq n$, we have $0 < p(h) < 1$, $0 < p(f_i|h) < 1$, and $\sum_{f_i \in dom(F_i)} p(f_i|h) < 1$. These are the usual restrictions regarding strict probabilities. Note that the set A depends on n, k, and k_i but this dependence is suppressed in our notation.

In order not to clutter our notation, we first present the results for naive Bayes models with binary features and then extend to naive Bayes models with features for which $k_i \geq 2$, and to other graphical models. When all k_i equal 2, the mapping defined by Eq. 2 is denoted by $g^{n,k} : A \subseteq R^{\hat{n}} \to R^m$ where $\hat{n} = nk + k - 1$ and $m = 2^n - 1$. For Binary naive Bayes models with n binary features, the mapping defined by Eq. 2 is denoted by $g^n : A \subseteq R^{\hat{n}} \to$

R^m where $\hat{n} = 2n+1$ and $m = 2^n - 1$. The set $g^{n,k,k_1,\ldots,k_n}(A)$ is called the *image* of a naive Bayes model.

We now show that the image of a naive Bayes model with k classes and n binary features is not a smooth manifold when $n \geq 2k$. Assume $\{h_1, \ldots, h_k\}$ are the k values of $dom(H)$ and $\{f_i, \overline{f}_i\}$ are the two values of $dom(F_i)$. Let the source coordinates of $g^{n,k}$ be $t_1, \ldots, t_{k-1}$, a_{ic}, $1 \leq i \leq n$, $1 \leq c \leq k$, where $t_c = p(h_c)$ and $a_{ic} = p(f_i|h_c)$. Note that $t_k = 1 - \sum_{c=1}^{k-1} t_c$ is not a source coordinate. The target coordinates of $g^{n,k}$ can be indexed as follows:

$$w_{i_1 i_2 \ldots i_r} = \sum_{c=1}^{k} t_c \prod_{i \in I} (1 - a_{ic}) \prod_{i \in \overline{I}} a_{ic} \tag{3}$$

where each index i has 2 possible values, I is the set of r indices $\{i_1, \ldots, i_r\}$ which are assigned with their second (or last) value and $\overline{I}$ is the set of the remaining $n-r$ indices. The first coordinate, when $I = \emptyset$, is denoted by $w_\emptyset$.

Theorem 7 *The image of a naive Bayes model with k classes and $n \geq 2k$ binary features is not a smooth manifold.*

Proof: The crucial fact we use is that if the image of $g^{n,k}$ were a smooth manifold, then the image would have a tangent hyperplane at each point and the dimension of that tangent hyperplane could not exceed the dimension of A which is $kn + k - 1$. Furthermore, if the image of $g^{n,k}$ were a smooth manifold, then $\partial g^{n,k} / \partial a_{ic}$ evaluated at a point x in the domain of $g^{n,k}$ would be a tangent vector to M at the point $g^{n,k}(x)$ in the image. This is because these partial derivatives are columns of the Jacobian matrix for $g^{n,k}$ and the Jacobian matrix gives the mapping between the tangent space of A and the tangent space of M. The proof provides a point in the image at which there are more than $kn+k-1$ linearly independent tangent vectors. Hence, the dimension of the tangent hyperplane is too large for the image to be a smooth manifold.

Suppose now that the image of $g^{n,k}$ is a smooth manifold M in R^{2^n-1}. Pick some $j \leq n$ and some point $x_j \in A$ with $t_c = 1/k$ and $a_{ic} = 1/2$ for all c and $i \neq j$. Furthermore, for x_j, let $a_{j1} \neq a_{j2}$, $a_{jc} = 1/2$ for $c > 2$, and $1/2 = \sum_{c=1}^{k} t_c a_{jc}$ (i.e., $a_{j1} + a_{j2} = 1$). Note that $y = g^{n,k}(x_j)$ is independent of which j we choose because $w_{i_1 i_2 \ldots i_r} = (1/2)^n$.

Consider the partial derivatives $\partial g^{n,k} / \partial a_{ic}$, $c = 1, 2$, evaluated at $x_1, \ldots, x_n$. Each partial derivative, as well as any linear combination of partial derivatives, is a tangent vector at y. We show that there are

$n + n(n-1)/2$ linearly independent tangent vectors at y. Consequently, since $kn + k - 1 < n + n(n-1)/2$ for $n \geq 2k$ we reach a contradiction: the number of independent tangent vectors is greater than the dimension of A. Consequently, M is not a smooth manifold at y.

We select the following $n + n(n-1)/2$ tangent vectors: $\partial g^{n,k}/\partial a_{i1} + \partial g^{n,k}/\partial a_{i2}$ evaluated at x_i, $1 \leq i \leq n$, and $\partial g^{n,k}/\partial a_{j1} - \partial g^{n,k}/\partial a_{j2}$ evaluated at x_i, $1 \leq i < j \leq n$. We consider these vectors as columns of a matrix and examine the submatrix formed by the first $1 + n + n(n-1)/2$ coordinates, denoted $w_\emptyset$, w_i, w_{ij}, $i < j$. By subtracting line $w_\emptyset$ from each of the other lines w_i and w_{ij}, removing $w_\emptyset$ from the matrix, and pulling the common constant from each column, we get a convenient square matrix of size $n + n(n-1)/2$. This matrix, which consists only of zeros and ones, has the form:

$$\begin{bmatrix} I & B' \\ B & C \end{bmatrix}$$

where I is the identity matrix of size $n \times n$, B' is the transpose of B and every line w_{ij} when restricted to B has two ones, in column i and j, and zeros otherwise (in B), and the square matrix C has zeros on the two main diagonals and ones otherwise. By subtracting lines w_i and w_j from line w_{ij}, $1 \leq i < j \leq n$, we get a diagonal matrix as needed. These calculations are facilitated by the equation

$$\partial w_{i_1 i_2 \ldots i_r}/\partial a_{jc}(x_l) = (1/k)(1/2)^{n-2} \cdot \begin{cases} -(1-a_{lc}) & j \in I,\ l \in I,\ j \neq l \\ -a_{lc} & j \in I,\ l \in \overline{I},\ j \neq l \\ 1 - a_{lc} & j \in \overline{I},\ l \in I,\ j \neq l \\ a_{lc} & j \in \overline{I},\ l \in \overline{I},\ j \neq l \\ -1/2 & j, l \in I,\ j = l \\ 1/2 & j, l \in \overline{I},\ j = l, \end{cases}$$

and by the fact that $a_{l1} + a_{l2} = 1$ for $1 \leq l \leq n$. $\Box$

Suppose now that the features are not all binary. Let f_{ij_i} be the jth element in $dom(F_i)$. Let a_{icj_i} stand for $p(f_{ij_i}|h_c)$, and let $t_c = p(h_c)$. Then the target coordinates of $g^{n,k,k_1,\ldots,k_n}$ can be indexed as follows:

$$w_{i_1 i_2 \ldots i_r} = \sum_{c=1}^{k} t_c \prod_{i \in I} (1 - \sum_{j_i=1}^{k_i - 1} a_{icj_i}) \prod_{i \in \overline{I}} a_{icj_i} \tag{4}$$

where each index i has k_i possible values, I is the set of r indices $\{i_1, \ldots, i_r\}$ which are assigned with their last value and $\overline{I}$ is the set of the remaining $n - r$ indices.

Theorem 8 *The image of a naive Bayes model with k classes and n features is not a smooth manifold, whenever $n \geq 2(k'-1)k$, where $k' = \max_i k_i$, $k_i = |\mathrm{dom}(F_i)|$.*

Proof: We use the same idea as in the proof of Theorem 7 and so we only describe the relevant changes. The image of a naive Bayes model is discussed in the notation of Eq 4. The point y for which we count the number of linearly independent tangent vectors is given as follows. Let $t_c = 1/k$ and $a_{icj_i} = 1/k_i$, for all $i \neq j$, $1 \leq j_i \leq k_i$, and $1 \leq c \leq k$. Let $a_{j11} \neq a_{j21}$, and $a_{jcj_i} = 1/k_j$ otherwise. Finally, let $1/k_j = \sum_{c=1}^{k} t_c a_{jcj_i}$ (i.e., $a_{j11} + a_{j21} = 2/k_j$). Note that $y = g^{n,k,k_1,\ldots,k_n}(x_j)$ is independent of which j we choose because $w_{i_1,\ldots,i_n} = \prod_i (1/k_i)$. We now compute the same derivatives as in Theorem 7, namely, with respect to a_{i11} and a_{i21} (which are denoted in the previous proof by a_{i1} and a_{i2}). The $1+n+n(n-1)/2$ lines are also selected as before; In line $w_\emptyset$ every index is assigned its first value. In line w_i, $1 \leq i \leq n$, index i is assigned its last value and all other indices are assigned their first value. In the next $n(n-1)/2$ lines, w_{ij}, $j > i$, the indices i and j are assigned their last value and all other $n-2$ indices are assigned their first value. The resulting matrix, after pulling constants from each column, is identical to the one given in the proof of Theorem 7 and so its rank is $n + n(n-1)/2$. Now, since the dimension of the image is at most $k - 1 + \sum_{i=1}^{n}(k_i - 1)k < k - 1 + n(k'-1)k$ and since $k - 1 + n(k'-1)k < n + n(n-1)/2$ when $n \geq 2(k'-1)k$, the image is not a smooth manifold at y. $\Box$

The proof technique of Theorems 7 and 8 can, with minor modifications, be used to prove that many DAG models with a hidden variable do not correspond to a smooth manifold. We outline the needed extensions.

First, we note that it suffices to examine the Markov blanket X of H (see Pearl, 1988) . The reason being that we can make a target coordinate change from $p(x,y)$ to $p(x)$ and $p(y|x)$ where y is a value of Y and Y are all nodes not in the Markov blanket of H. This is a diffeomorphism. Furthermore, the network coordinates that correspond to the Markov blanket determine $p(x)$ and the rest of the network coordinates determine $p(y|x)$. Hence we can analyze separately how the Markov network coordinates are mapped to $p(x)$. Thus we will restrict our discussion to the case where $Y = \emptyset$.

Suppose we have a DAG model with one hidden node having k classes and suppose it has n binary children and no parents. The target coordinates of this model can be indexed as follows:

$$w_{i_1 i_2 \ldots i_r} = \sum_{c=1}^{k} t_c \prod_{i \in I} (1 - a_{ic\pi_i}) \prod_{i \in \overline{I}} a_{ic\pi_i}$$

where each index i has 2 possible values, I is the set of r indices $\{i_1, \ldots, i_r\}$ which are assigned with their second (or last) value and $\overline{I}$ is the set of the remaining $n - r$ indices. The symbol π_i denotes the parents of i other than H and their values must be consistent with those in I.

The point we select is the one in which all edges are missing except one edge which goes from H to some node j. This is the same point as in the proof of Theorem 7. Also we take the same derivatives and obtain the same matrix with a dimension of $n + n(n-1)/2$. To get a contradiction we must have $n(n+1)/2$ be greater than the number of network parameters when n is large enough. This happens whenever the state space created by the parents of each node is sufficiently small compared to n. So, a contradiction is reached when

$$2^{\pi} kn + k - 1 < n(n+1)/2$$

where π is the maximal number of parents of a node (aside of H). This inequality is satisfied, for n large enough, whenever π is a constant (not depending on n). Obvious modifications are needed when variables in X are not binary. When H has parents and its children have parents (i.e., a general Markov blanket), we pick a point that makes H independent of its parents, and equally likely on each state. This is the same point as before—just more equalities need to be set. To summarize, we have justified the following claim:

Theorem 9 *The image of a discrete DAG model with a hidden variable H with n children is not a CEF whenever $n(n+1)/2$ is larger than the cardinality of the state space over the observable variables.*

We note that the proof of Theorem 9, as well as all other proofs in this section, exhibits one singular point y at which the image of a graphical model is not a smooth manifold. It does not describe the set of all singular points at which the image is not a smooth manifold. It also does not determine whether the point y is singular because the image is not a topological manifold at y or because it is not smooth at y. In the Appendix we give full answers to these questions for binary naive Bayes model with n binary features. In particular, we show that the image is not even a topological manifold at singular points, and that the singular points are precisely those for which $p(f_i|h) = p(f_i|\overline{h})$ for all values of i, except at most two values $\{i_1, i_2\}$ where inequality is possible. Additional results are provided in the appendix that shed light on the geometry of the image of binary naive Bayes models with binary features.

4.4 Computation of the Dimension

The dimension of a SEF is the dimension of the highest stratum. In this section we present an algorithm that computes the dimension of a SEF when specified as an image of a polynomial mapping composed with a diffeomorphism. For this discussion, it is sufficient to consider only the polynomial portion of the mapping because diffeomorphisms do not change the dimension.

The next lemma suggests a random algorithm for calculating the maximal rank of the Jacobiam matrix of a polynomial mapping. The algorithm and Lemma 10 were also studied more generally for analytical mappings in Bamber and van Santen (1985). A proof for polynomial mappings, which is all we need, is much simpler and thus included herein.

Lemma 10 *Let $g : R^m \to R^n$ be a polynomial mapping. Let $J(x) = \partial g/\partial x$ be the Jacobian matrix at x. Then the rank of $J(x)$ equals the maximal rank almost everywhere.*

Proof: Let d be the maximal rank of $J(x)$. Because the mapping g is polynomial, each entry in the matrix $J(x)$ is a polynomial in x. When diagonalizing $J(x)$, the leading elements of the first d lines remain polynomials in x, whereas all other lines, which are linearly dependent given every value of x, become identically zero. The rank of $J(x)$ falls below d only for values of x that are roots of some of the polynomials in the diagonalized matrix. The set of all such roots has measure zero. $\Box$

A random algorithm for computing the maximal rank of $J(x)$ is now evident. At the first step, the algorithm computes the Jacobian matrix $J(x)$ symbolically from $g(x)$. This computation is possible since g is a vector of polynomials in x. Then, it assigns a random value to x and diagonalizes the numeric matrix $J(x)$. Lemma 10 guarantees that, with probability 1, the resulting rank is the maximal rank of $J(x)$.

The next lemma shows that this algorithm computes the dimension of the image of a polynomial mapping. Recall that the dimension of the image is defined to be the dimension of the highest stratum of the image.

Theorem 11 *Let $g : A \subseteq R^m \to R^n$ be a polynomial mapping where A is a semialgebraic open set. Let $J(x) = \partial g/\partial x$ be the Jacobian matrix at x. Then the maximal rank of $J(x)$ is equal to the dimension of $g(A)$.*

This theorem is a special case (with $V = R^m$) of the following theorem (still in a draft form):

Theorem 12 *Let $g : R^m \to R^n$ be a polynomial mapping. Let A be an open semialgebraic subset of R^m and let V be an algebraic subset of R^m. Suppose that $A \cap V$ is contained in the nonsingular points of V. For $x \in A \cap V$, let $J(x) = \partial g/\partial x$ be the Jacobian matrix of g at x, and let $P_V(x)$ be the matrix of orthogonal projection to the tangent space of V at x. Let d be the maximum over $x \in A \cap V$ of the rank of the matrix $J(x)P_V(x)$. Then $g(A \cap V)$ is a semialgebraic set whose dimension is d.*

Proof: We recall a few facts about semialgebraic sets. Let A and B be semialgebraic sets. If $A \subset B$ then $\dim(A) \leq \dim(B)$. Also $\dim(A \cup B) = \max(\dim(A), \dim(B))$. The closure $\overline{A}$ is semialgebraic and $\dim(\overline{A}) = \dim(A)$. Finally, any semialgebraic set has only a finite number of connected components.

We prove this theorem by induction on d. By Proposition 2.4.3 of Akbulut and King (1992), we know the entries of $P_V(x)$ are rational functions, whose denominators do not vanish on the nonsingular points of V. Consequently, there is an algebraic subset $W \subset V$ so that $W \cap A$ is the set of points $x \in A \cap V$ at which $J(x)P_V(x)$ has rank less than d. (The subset W is given by the vanishing of all $d \times d$ minors of $J(x)P_V(x)$, or alternatively, see the proof of Lemma 10.) By induction, we know that $g(W \cap A)$ has dimension less than d. In particular, let $W_0 = W$ and let W_i be the singular points of W_{i-1} if $i \geq 1$. We apply this theorem with A replaced by $A - W_{i+1}$ and V replaced by W_i. Note that if $x \in W_i$ then the tangent space of W_i at x is contained in the tangent space of V at x and so the rank of $J(x)P_{W_i}(x)$ is less than or equal to the rank of $J(x)P_V(x)$ which is less than d. So by induction the dimension of $g(A \cap (W_i - W_{i+1}))$ is less than d. So if B is the closure of $g(A \cap W)$, then B is semialgebraic and $\dim(B) < d$.

Let $C = A - g^{-1}(B)$. Note that C is an open semialgebraic set and $J(x)P_V(x)$ has rank d at all points $x \in C \cap V$. We have reduced to showing that $\dim(g(C \cap V)) = d$. Take any point $y \in g(C \cap V)$ and any $x \in C \cap V \cap g^{-1}(y)$. Theorem 5.4 of Bröcker and Jänich (1982) gives a local description of g near x in V. In particular, there is a neighborhood U of x

in V so that $g(U)$ is a d dimensional submanifold of R^n and $g^{-1}(y) \cap U$ is a submanifold of V. So if $x' \in g^{-1}(y) \cap V$ is close enough to x, a neighborhood of x' in V will be mapped to the exact same d dimensional submanifold as a neighborhood of x. Consequently, if x' is any point in the same connected component of $C \cap V \cap g^{-1}(y)$ as x, a neighborhood of x' in V will be mapped to the exact same d dimensional submanifold as a neighborhood of x. Since $C \cap V \cap g^{-1}(y)$ is semialgebraic, it has only a finite number of connected components. Hence a neighborhood of y in $g(C \cap V)$ is a finite union of d dimensional submanifolds. So $\dim(g(C \cap V)) = d$. $\Box$

In the context of graphical models g is the mapping from the network parameters Θ to the joint-space parameters W. For example, for naive Bayes models g is replaced with $g^{n,k,k_1,\ldots,k_n}$. We have implemented the algorithm in Mathematica and used it to find the dimension of several graphical models with hidden variables. Here we summarize the results for $g^{n,k,k_1,\ldots,k_n}$. (Implementation details can be found in Geiger, Heckerman, and Meek, 1996).

For $k = 2$, the maximal rank of $g^{n,k}$ computed by the algorithm was full, namely, all results were consistent with the formula $\min(2n+1, 2^n - 1)$. In the appendix, among other results, we prove that the maximal rank is indeed full for every n. For $k > 2$, the maximal rank of $g^{n,k}$ found by the algorithm was $\min(nk+k-1, 2^n-1)$, except when $(n = 4, k = 3)$, where the maximal rank is 13 rather than 14. This drop in dimension has also been observed by Goodman (1974, pp. 221). When $n = 2$, the maximal rank of g^{n,k,k_1,k_2} can be far from full. Settimi and Smith (1998) show that for $k < \min(k_1, k_2)$ the dimension drops by $k(k-1)$. The algorithm confirms this dimension drop. Other examples are discussed in Geiger et al. (1996).

5 Discussion

An obvious challenge remains open: Is BIC a valid asymptotic expansion for the marginal likelihood $P(Data|model)$ when the model is a stratified exponential family ?

One solution to this problem may be as follows. Exclude from the stratified model all points aside of the highest stratum. As a result, only a measure zero set (with respect to the volume element of the highest stratum) of points is excluded. The remaining set is a smooth manifold and so BIC is a correct asymptotic expansion, under the appropriate regularity conditions, as long as the MAP point converges to a point that has not been excluded.

This requirement about convergence is not always satisfied. To be concrete, suppose points in R^2 are generated from a standard two dimensional normal distribution $N((m_x, m_y), I)$. We have two equally likely models. The first model consists of all standard two dimensional normal distributions for which $\{(m_x, m_y)|m_x^2 = m_y^3\}$ and the second model consists of all those distributions for which $\{(0, m_y)|m_y < -1\}$. The point $(0, 0)$ is not smooth in the first model. However, if the second model contains the true distribution then the MAP value for the first model will converge to a bad point. According to our prior probability, we expect this to happen with probability 1/2. Thus, given the relationship between the alternative plausible models, this point cannot be excluded in an asymptotic analysis. A more careful asymptotic analysis of the behavior at bad points is needed.

There are other obstacles in applying Haughton's results to graphical models with hidden variables. These consist of Haughton's (1988) technical assumptions, as well as the assumptions that the prior is bounded and bounded away from zero in a local coordinate system on the natural parameter space. Priors are usually defined on the network parameters and when the prior is transformed to the natural parameter space, it is not necessarily bounded. In particular, for a DAG model with a hidden variable, the prior on the natural parameter space is usually not bounded whenever the prior on the network parameters is bounded and bounded away from zero.

Acknowledgement

We thank Mike Freedman for fruitful discussions on the mathematics related to this paper and to Steffen Lauritzen for guiding us through the mysteries of exponential families. We have also benefited from conversations with and comments by many other people including Christian Borges, Jennifer Chayes, Dominique Haughton, Jim Kajiya, Rob Kass and Paul Vos.

References

Akbulut, S. and King, H. (1992). *Topology of real algebraic sets.* Springer-Verlag, New York.

Andersson, S., Madigan, D., and Perlman, M. (1996). An alternative Markov property for chain graphs. In *Proceedings of the Twelfth conference on Uncertainty in Artificial Intelligence*, pages 40–48. Morgan Kaufmann.

Bamber, D. and van Santen, J. (1985). How many parameters can a model have and still be testable? *Journal of mathematical pshychology*, 29:443–473.

Barndorff-Nielsen, O. (1978). *Information and exponential families*. Wiley, New York.

Benedetti, R. and Risler, J. (1990). *Real algebraic and semi-algebraic sets*. Hermann, Paris.

Chickering, D., Heckerman, D., and Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of Uncertainty and Artificial Intelligence*, San Francisco. Morgan Kaufmann.

Efron, B. (1978). The geometry of exponential families. *Annals of Statistics*, 6(2):362–376.

Friedman, N. and Goldszmidt, M. (1996). Learning Bayesian networks with local structure. In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence (UAI–96)*, pages 252–262, San Francisco, CA. Morgan Kaufmann Publishers.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research.

Geiger, D., Heckerman, D., and Meek, C. (1996). Asymptotic model selection for directed networks with hidden variables. In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence (UAI–96)*, pages 283–290, San Francisco, CA. Morgan Kaufmann Publishers.

Geiger, D. and Meek, C. (1998). Graphical models and exponential families. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, page to appear, San Francisco, CA. Morgan Kaufmann Publishers.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16:342–555.

Heckerman, D. and Breese, J. (1996). Causal independence for probability assessment and inference using Bayesian networks. *IEEE, Systems, Man, and Cybernetics*, 26:826–831.

Henrion, M. (1987). Some practical issues in constructing belief networks. In *Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 132–139. Association for Uncertainty in Artificial Intelligence, Mountain View, CA.

Kass, R. and Vos, P. (1997). *Geometrical foundations of asymptotic inference.* Wiley, New York.

Koster, J. (1997). Gibbs and Markov properties of graphs. *Annals of Mathematics and Artificial Intelligence*, 21(1):13–26.

Lauritzen, S. (1996). *Graphical models.* Clarendon Press, Oxford.

Lauritzen, S. and Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17:31–57.

Meek, C. and Heckerman, D. (1997). Structure and parameter learning for causal independence and causal interaction models. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–97)*, pages 366–375, San Francisco, CA. Morgan Kaufmann Publishers.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent systems.* Morgan-Kaufmann, San Mateo.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Settimi, R. and Smith, J. (1998). Geometry and identifiability in simple discrete bayesian models. Technical Report 324, University of Warwick.

Shachter, R. and Kenley, R. (1986). Gaussian influence diagrams. *Management Science*, 35(5):527–550.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search.* Springer-Verlag.

Spirtes, P., Richardson, T., and Meek, C. (1997). The dimensionality of mixed ancestral graphs. Technical Report CMU-PHIL-83, Philosophy Department, Carnegie Mellon University.

Spivak, M. (1965). *Calculus on manifolds.* Addison-Wesley, New York.

Whittaker, J. (1990). *Graphical Models in applied multivariate statistics.* Wiley.

Appendix

In this appendix we study the image M of a binary naive Bayes model with n binary features. In particular, we characterize the set of points S for which the image is not a topological manifold, show that $M \setminus S$ is a smooth manifold, show that every point in $M \setminus S$ has exactly two sources and provide an explicit formula that computes these source points. In addition we resolve a conjecture made in Geiger et al. (1996) by showing that the dimension of these models is full, namely, $2n+1$ when $n \geq 3$. For $n = 1, 2$, the dimension is $2^n - 1$.

These results are facilitated by a sequence of diffeomorphisms some of which are applied to the source coordinates and some to the target coordinates. Such transformations are valid because they preserve the properties we study herein. Our starting point is Eq. 3 with $k = 2$, $a_{i1} = a_i$, $a_{i2} = b_i$, $t_1 = t$, and $t_2 = 1 - t$.

Using a non-singular linear transformation on the target coordinates we obtain the following mapping:

$$z_{ij\ldots r} = t a_i a_j \cdots a_r + (1-t) b_i b_j \cdots b_r$$

where z_i stands for the probability of the i-th feature being true, z_{ij} stands for the probability that the i-th and j-th features are both true, etc.

We now apply a diffeomorphism on the source coordinates where s, x_1, x_2, ... x_n, and u_1, ..., u_n are the new coordinates as given by,

$$t = (s+1)/2, \quad a_i = x_i + (1-s)u_i, \quad b_i = x_i - (1+s)u_i.$$

The mapping in the new source coordinates becomes:

$$\begin{aligned} z_i &= x_i \\ z_{ij} &= x_i x_j + (1-s^2) u_i u_j \end{aligned}$$

$$z_{ijk} = x_i x_j x_k + (1-s^2)(x_i u_j u_k + u_i x_j u_k + u_i u_j x_k) - 2s(1-s^2) u_i u_j u_k$$

$$z_{12\ldots r} = x_1 x_2 \cdots x_r + \sum_{i=2}^{r} p_i(s) \cdot (\sum(\text{products of i u's and r-i x's}))$$

where $p_i(s) = 1/2(1-s^2)((1-s)^{i-1} - (-1)^{i-1}(1+s)^{i-1})$, and, in particular, $p_2(s) = 1 - s^2$ and $p_3(s) = -2s(1-s^2)$.

Now we subtract products of the first n coordinates to get rid of the leading terms. So, we do $z_{ij} \leftarrow z_{ij} - z_i z_j$. Then we subtract products of the first n coordinates with one of the next n choose 2 coordinates to get rid of the second terms, namely, $z_{ijr} \leftarrow z_{ijr} - z_{ij} z_r - z_{ir} z_j - z_{jr} z_i - z_i z_j z_r$. And so forth. We end up with the mapping:

$$z_i = x_i, \quad z_{ij} = p_2(s) u_i u_j, \quad \ldots, \quad z_{ij\cdots r} = p_r(s) u_i u_i \cdots u_r$$

Let us denote this mapping with $F^n : U \subset R^{2n+1} \to R^{2^n - 1}$, where U is the set of $(x, u, s) \in R^n \times R^n \times R$ such that:

$$0 < x_i < 1, \quad -1 < s < 1$$

$$-x_i < (1-s) u_i < 1 - x_i$$

$$x_i - 1 < (1+s) u_i < x_i.$$

We denote the coordinates of F^n with $F_i^n(x, u, s) = x_i$, $F_{ij}^n(x, u, s) = p_2(s) u_i u_j$, $F_{ij\cdots r}^n(x, u, s) = p_r(s) u_i u_j \cdots u_r$, etc.

We are now ready to analize the image of U under F^n. Let $M = F^n(U)$ be the image of U. Let S be the set of points in M for which at most one of the coordinates z_{ij} is nonzero. Let S' be the set of points in M for which all coordinates z_{ij} are 0. Note that $S' \subset S$.

Theorem 13 *The dimension of the image of a naive Bayes model with $n \geq 3$ binary features is $2n + 1$.*

Proof. The dimension of the image of a naive Bayes model is equal to the maximal rank of F^n because F^n is obtained from g^n by composition with diffeomorphisms. Thus one just needs to compute the maximal rank of the Jacobian matrix of F^n. Let J_n denote this Jacobian matrix. We show that the maximal rank of J_n is $2n + 1$ for $n \geq 3$.

The matrix J_n has two blocks along the main diagonal where the first block of size n is an identity matrix. It remains to argue that the second block has a maximal rank of $n + 1$. We establish this claim by selecting

$n+1$ rows and showing that this submatrix has full rank. The rows selected, among many other valid possibilities, are those that correspond to the target coordinates $z_{1,i}$, $2 \leq i \leq n$, z_{23} and z_{123}. Assuming the columns of the second block are organized according to the order, $u_2, \ldots, u_n, u_1, s$, then this submatrix of J_n is

$$\begin{pmatrix} p(s)u_1 & 0 & 0 & 0 & \ldots & p(s)u_2 & -2su_1u_2 \\ 0 & p(s)u_1 & 0 & 0 & \ldots & p(s)u_3 & -2su_1u_3 \\ 0 & 0 & p(s)u_1 & 0 & \ldots & p(s)u_4 & -2su_1u_4 \\ & \ldots & & & & & \ldots \\ 0 & 0 & 0 & 0 & p(s)u_1 & p(s)u_n & -2su_1u_n \\ p(s)u_3 & p(s)u_2 & 0 & 0 & \ldots & 0 & -2su_2u_3 \\ -2sp(s)u_1u_3 & -2sp(s)u_1u_2 & 0 & 0 & 0 & -2sp(s)u_2u_3 & -[2sp(s)]'u_1u_2u_3 \end{pmatrix}$$

where $p(s) = 1 - s^2$. Using two row operations, we get a diagonal matrix with a maximal rank of $n+1$ as claimed. $\square$

Theorem 14 *Let S be the set of points in M for which at most one of the coordinates z_{ij} is nonzero. The set $M - S$ is a smooth manifold and this set is double covered by F^n.*

Proof. Take any point $z \in M - S$. Then we have $z_{ij} \neq 0$ and $z_{k\ell} \neq 0$ with $ij \neq k\ell$. So if $F^n(x, u, s) = z$, we must have $u_a \neq 0$ for $a = i, j, k, \ell$. So u must have at least three nonzero coordinates. Without loss of generality, we may suppose that $u_i \neq 0$ for $i = 1, 2, 3$. Consequently, z_{12}, z_{13}, z_{23}, and z_{123} are all nonzero.

Then we can solve for $(x, u, s) = F^{n-1}(z)$ as follows:

$$x_i = z_i$$
$$u_1 = \pm\sqrt{z_{12}z_{13}z_{23} + (z_{123})^2/4}/z_{23}$$
$$s = -z_{123}/(2u_1z_{23})$$
$$u_i = z_{1i}/(p_2(s)u_1) \text{ for } i > 1$$

In particular, there are exactly two points in the inverse image, and if we choose one of these points (by choosing the $\pm$ sign) we have a smooth local inverse for F^n. Consequently, $M - S$ is a smooth manifold and it is double covered by F^n. $\square$

Theorem 15 *Let S be the set of points in M for which at most one of the coordinates z_{ij} is nonzero. The set M is not a topological manifold at points of S.*

Proof. A topological manifold is locally compact. (A space is locally compact if each point has a compact neighborhood. Since each point in a topological manifold has a neighborhood homeomorphic to closed disc, any topological manifold is locally compact.) We will show that M is not locally compact at points of $S \setminus S'$. Recall that S' is the set of points in M for which all coordinates z_{ij} are 0. Loosely stated, the reason M is not locally compact at points of $S \setminus S'$ is that points arbitrarily close to the edge of U are mapped arbitrarily close to any point of $S - S'$. Finally, we argue that M is also not locally compact at points of S'.

To be precise, pick any $z' \in S - S'$ and suppose it has a compact neighborhood N in M. Pick $\epsilon > 0$ small enough that N contains the intersection of M with the ball of radius ϵ around z. Pick a large constant b. We may as well suppose that $z'_{12} \neq 0$, but all other z_{ij} are 0. Consequently the only nonzero coordinates of z' are z'_i and z'_{12}. Pick any $(x', u', s') \in U$ so that $F^n(x', u', s') = z'$. after applaying σ, we may as well assume that $u'_1 > 0$. For small enough $\delta > 0$, consider the point (x', u^δ, s^δ) in U where:

$$s^\delta = 2z'_1 - 1$$
$$u^\delta_1 = 1/2 - \delta$$
$$u^\delta_2 = z'_{12}/((1/2 - \delta)p_2(s^\delta))$$
$$u^\delta_3 = \epsilon/b$$
$$u^\delta_i = 0 \text{ for } i > 3$$

We show here that $(x', u^\delta, s^\delta) \in U$ if δ is small enough. Since $x'_i \in (0,1)$ and $s^\delta \in (-1, 1)$, by the above description of U, we must only show that:

$$-x_i < (1-s)u_i < 1 - x_i$$
$$x_i - 1 < (1+s)u_i < x_i$$

These are trivially true if $i > 3$, and true for large enough b if $i = 3$. We also have:

$$-x_1 < 0 < (1 - s^\delta)u^\delta_1 = (1 - 2\delta)(1 - x_1) < 1 - x_1$$
$$x_1 - 1 < 0 < (1 + s^\delta)u^\delta_1 = (1 - 2\delta)x_1 < x_1$$

If $z'_{12} > 0$ then since $(x', u', s') \in U$ we have

$$x'_1 > (1 + s')u'_1 = z'_{12}/((1 - s')u'_2) > z'_{12}/(1 - x_2)$$

so $z'_{12}/x'_1 < 1 - x'_2$. Likewise $z'_{12}/(1 - x'_1) < x'_2$. So if δ is small enough, we have the remaining inequalities

$$-x_2 < 0 < (1 - s^\delta)u_2^\delta = z'_{12}/((1 - 2\delta)x'_1) < 1 - x'_2$$

$$x_2 - 1 < 0 < (1 + s^\delta)u_2^\delta = z'_{12}/((1 - 2\delta)(1 - x'_1)) < x'_2$$

Similarly, if $z'_{12} < 0$ then $u'_2 < 0$ and we have

$$x'_1 > (1 + s')u'_1 = z'_{12}/((1 - s')u'_2) > -z'_{12}/x'_2$$

$$1 - x'_1 > (1 - s')u'_1 = z'_{12}/((1 + s')u'_2) > z'_{12}/(x'_2 - 1)$$

and so for small enough δ,

$$-x_2 < z'_{12}/((1 - 2\delta)x'_1) = (1 - s^\delta)u_2^\delta < 0 < 1 - x_1$$

$$x_2 - 1 < z'_{12}/((1 - 2\delta)(1 - x'_1)) = (1 + s^\delta)u_2^\delta < 0 < x_1$$

Now we have

$$F_i^n(x', u^\delta, s^\delta) = z'_i$$
$$F_{12}^n(x', u^\delta, s^\delta) = z'_{12}$$
$$F_{13}^n(x', u^\delta, s^\delta) = p_2(s^\delta)\epsilon(1/2 - \delta)/b$$
$$F_{23}^n(x', u^\delta, s^\delta) = \epsilon z'_{12}/(b(1/2 - \delta))$$
$$F_{123}^n(x', u^\delta, s^\delta) = -2s^\delta z'_{12}\epsilon/b$$

and all other coordinates of $F^n(x', u^\delta, s^\delta)$ are 0. So if b is large enough (for example $b > 2 \geq 1/2 + 6|z'_{12}|$) we see that $F^n(x', u^\delta, s^\delta)$ is within ϵ of z', so it is in the compact N. Letting δ approach 0, compactness of N gives us a limit point $z'' \in N$. We see that $z''_i = z'_i$, $z''_{12} = z'_{12}$, $z''_{23} = 2\epsilon z'_{12}/b$, $z''_{13} = p_2(z'_1)\epsilon/(2b)$, $z''_{123} = -2s^\delta z'_{12}\epsilon/b$, and all other coordinates are 0.

Note that z'' is in $M - S$ so we have an explicit formula above for its inverse image. In particular, if $F^n(x'', u'', s'') = z''$ then $x'' = x'$, $s'' = s^\delta$, $u''_1 = 1/2$, $u''_2 = z'_{12}/p_2(s^\delta)$, $u''_3 = \epsilon/b$, and all other u''_i are 0. But this point is not in U which can be seen by converting back to the original coordinates: $a''_1 = x''_1 + (1 - s'')u''_1 = z'_1 + (2 - 2z'_1)(1/2) = 1$ which is outside the allowed range.

So we have a contradiction. Consequently, M is not locally compact at $S - S'$ and hence is not a manifold there. Note also that M cannot be locally compact at S' since any point of S' has arbitrarily close points in $S - S'$ so any compact neighborhood of a point in S' is also a compact neighborhood of a point in $S - S'$, which we have just shown cannot exist. $\square$

At this point one might argue that perhaps M is not a topological manifold for a mere technical reason. Suppose we considered $M' = F^n(\bar{U})$ where $\bar{U}$ is the closure of U. Since $\bar{U}$ is closed and bounded, it is compact, so its image M' is also compact, and hence locally compact. Hence, there is still the possibility that M' could be a topological manifold. Moreover, taking $\bar{U}$ is not unreasonable, we are just allowing our probabilities to be 0 or 1. Nevertheless M' is not a topological manifold. In fact, we can show that at points of $S - S'$, M' is locally homeomorphic to $R^{n+1} \times c(D^2 \times S^{n-3})$ where $c(D^2 \times S^{n-3})$ is the cone on a 2-disc D^2 cross the $n - 3$ sphere (A cone on a set A is the set of points lying on some straight line between a point in A and the origin). We can also show that at points of $S \setminus S'$, M is locally homeomorphic to $R^{n+1} \times c(R^2 \times S^{n-3})$.