On the geometry of DAG models with hidden variables

Dan Geiger[1,2], **David Heckerman**[1], **Henry King**[3], **Christopher Meek**[1]
[1]Microsoft Research, WA, USA
[2]Computer Science Department, Technion, Israel
[3]Math Department, University of Maryland, USA
{dgeiger,heckerma,meek}@microsoft.com, hck@math.umd.edu

Abstract

We prove that many graphical models with hidden variables are not curved exponential families. This result, together with the fact that some graphical models are curved and not linear, implies that the hierarchy of graphical models, as linear, curved, and stratified, is non-collapsing; each level in the hierarchy is strictly contained in the larger levels. This result is discussed in the context of model selection of graphical models.

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 statistician, provide flexible and modular definitions of families of probability distributions, and are amenable to scalable 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 (e.g., Spirtes et al., 1993, Pearl, 1998).

Graphical models can be viewed as a hierarchy according to their representation as exponential families. Undirected graphical models with no hidden variables are linear exponential families (LEFs) (Lauritzen, 1996), directed acyclic graphical models with no hidden variables are curved exponential families (CEFs) (Geiger and Meek, 1998), and graphical models with hidden variables are stratified exponential families (SEFs) (Geiger and Meek, 1998).

Herein we prove that many graphical models with hidden variables are not curved exponential families. This result, together with the fact that some graphical models are curved and not linear, implies that the hierarchy of graphical models, as linear, curved, and stratified, is non-collapsing; each level in the hierarchy is strictly contained in the larger levels. We also show how to compute the dimension of a SEF by proving a connection between the dimension of the highest stratum and the regular rank of a Jacobian matrix.

Our work 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 a linear 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). 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 graphical models without hidden variables.

However, 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. Herein, we show that graphical models with hidden variables are not CEFs. This 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.

2 Background

In this background section we recall the definitions of smooth manifolds, topological manifolds, and stratified sets, based on (Spivak 1965, Akbulut and King, 1992, Benedetti and Risler, 1990). We then provide the definitions for linear, curved, and stratified exponential families based on (Lauritzen, 1996, Kass and Vos, 1997, Geiger and Meek, 1998).

2.1 Manifolds and stratified sets

A *diffeomorphism* $f : U \subset R^n \rightarrow R^m$ 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 $f : U \cap M \rightarrow R^k$. When f is only assumed to be continuous and to have a continuous 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 $g : A \subset R^n \rightarrow B \subset R^n$ 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 $h : A \rightarrow R^{m-n}$ 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.*

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 *stratified set* is a set that has a stratification. The dimension of a stratified set is d_1— the largest dimension of a stratum. We note that if E is a stratified set and f is a diffeomorphism, then $f(E)$ is also a stratified set.

An example of a smooth manifold, a topological manifold and a stratified set are shown in Figure 1.

2.2 Exponential families

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

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

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.

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.

A subfamily of a 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 $k - p$.

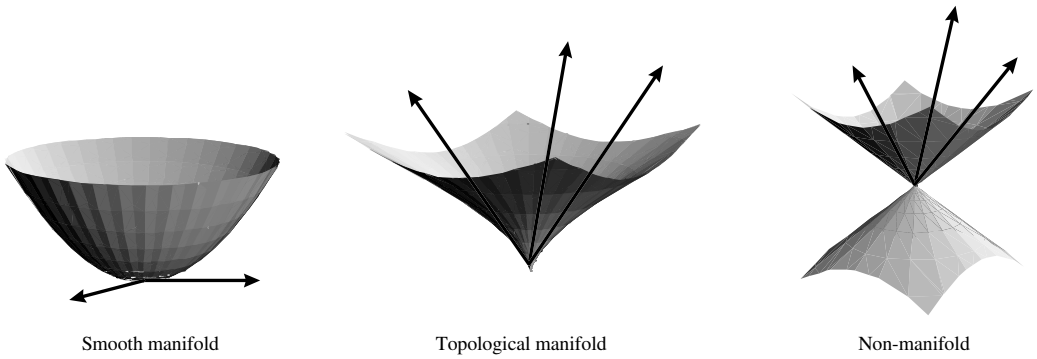

Figure 1: Three types of surfaces

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.

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 other words an SEF is a finite union of CEFs of various dimensions satisfying some regularity condition.

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 former class contains graphical models which are not contained in the smaller class. In this section we show that many graphical models with a hidden variable (which are SEFs) are not CEFs.

We first study in detail a class of graphical models which are often called *naive Bayes models*. 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 multinomial distributions for the sample space $dom(F_1) \times \ldots \times dom(F_n)$ such that

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

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. 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, also called the network parameters, 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), also called the joint-space parameters. 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:

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

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 2 *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. (See, for example, Figure 1 where there are three independent tangent vector at the origin while the surface has a dimension only of 2).

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

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

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

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:

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

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 3 *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 2 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 2, 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 2 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. □

The proof technique of Theorems 2 and 3 can, with minor modifications, be used to prove that many DAG models with a hidden variable do not correspond to a smooth manifold.

Theorem 4 *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 4, 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. We derive a formula that provides the two possible source points for every non singular point in the image of a binary naive Bayes model with n binary features.

4 Computation of the dimension

We now show how to compute the dimension of a SEF when specified as an image of a polynomial mapping composed with a diffeomorphism, by proving a connection between the dimension of the highest stratum and the regular rank of some Jacobian matrix. 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 Jacobian matrix of a polynomial mapping. The algorithm and Lemma 5 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 5 *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. □

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 5 guarantees that, with probability 1, the resulting rank is the maximal rank of $J(x)$.

The next theorem shows that this algorithm computes the dimension of the image of a polynomial mapping. Such an image is a stratified set and its dimension is defined to be the dimension of the highest stratum (Benedetti and Risler, 1990).

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. When only equalities are used the set is *algebraic*.

Theorem 6 *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:

Theorem 7 *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 5.) 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 $\dim(g(C \cap V)) = d$. □

In the context of graphical models g is the mapping from the network parameters to the joint-space parameters. 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 $N((m_x, m_y), I)$. Suppose also that we have, a priori, 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 first model has one singularity at (0,0). Although this singularity has measure zero with respect to the first model, we cannot exclude it from the model. In particular, the MAP value for the first model will converge to (0,0) whenever the second model contains the true distribution, an event that will happen with probability 1/2 according to our prior. A more careful asymptotic analysis of the behavior at singular 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 and King, 1992] Akbulut, S. and King, H. (1992). *Topology of real algebraic sets.* Springer-Verlag, New York.

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

[Bröcker and Jänich, 1982] Bröcker, TH. and Jänich, K, (1982). An introduction to differential topology, Cambridge university press.

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

[Geiger et al., 1996] 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 and Meek, 1998] 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)*, pages 156–165, San Francisco, CA. Morgan Kaufmann Publishers.

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

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

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

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

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

[Pearl, 1998] Pearl, J. (1998). Graphs, Causality, and Structural Equation Models, *Sociological Methods and Research*, 27(2).

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

[Settimi and Smith, 1998] Settimi, R. and Smith, J. (1998). On the geometry of Bayesian graphical models with hidden variables. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, pages 472–479, San Francisco, CA. Morgan Kaufmann Publishers.

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

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

[Whittaker, 1990] 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:

$$\begin{aligned}
z_i &= x_i \\
z_{ij} &= x_i x_j + (1-s^2) u_i u_j \\
z_{ijk} &= x_i x_j x_k + (1-s^2)(x_i u_j u_k + u_i x_j u_k \\
&\quad + u_i u_j x_k) + u_i x_j u_k + u_i u_j x_k) \\
&\quad - 2s(1-s^2) u_i u_j u_k \\
z_{12\ldots r} &= x_1 x_2 \cdots x_r + \\
&\quad \sum_{i=2}^{r} p_i(s) \sum (\text{products of i u's and r-i x's})
\end{aligned}$$

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:

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

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

$$\begin{aligned}
0 < x_i < 1, \quad -1 < s < 1 \\
-x_i < (1-s) u_i < 1 - x_i \\
x_i - 1 < (1+s) u_i < x_i.
\end{aligned}$$

We denote the coordinates of F^n with $F^n_i(x, u, s) = x_i$, $F^n_{ij}(x, u, s) = p_2(s) u_i u_j$, $F^n_{ij\ldots r}(x, u, s) = p_r(s) u_i u_j \cdots u_r$, 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 8 *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 given in Figure 5 where $p(s) = 1 - s^2$. Using two row operations, we get a diagonal matrix with a maximal rank of $n + 1$ as claimed. □

Theorem 9 *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.

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

Figure 2: A submatrix of J_n

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. □

Theorem 10 *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:

$$s^\delta = 2z'_1 - 1$$
$$u_1^\delta = 1/2 - \delta$$
$$u_2^\delta = z'_{12}/((1/2 - \delta)p_2(s^\delta))$$
$$u_3^\delta = \epsilon/b$$
$$u_i^\delta = 0 \text{ for } i > 3$$

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:

$$-x_1 < 0 < (1 - s^\delta)u_1^\delta = (1 - 2\delta)(1 - x_1) < 1 - x_1$$
$$x_1 - 1 < 0 < (1 + s^\delta)u_1^\delta = (1 - 2\delta)x_1 < x_1$$

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. □

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 $R^{n+1} \times c(R^2 \times S^{n-3})$.