Algebraic Geometry Applications in Model Selection

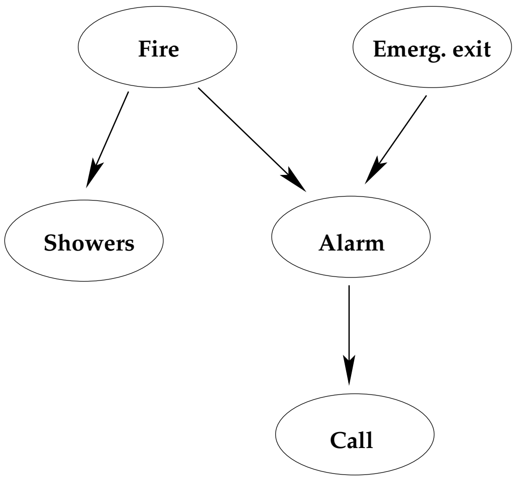
Luis David Garcia

lgarcia@math.vt.edu

Virginia Polytechnic Institute and State University

References

- Asymptotic Model Selection for Naive Bayesian Networks by D. Rusakov and D. Geiger, Uncertainty in Artificial Intelligence (UAI-02).
- Automated Analytic Asymptotic Evaluation of the Marginal Likelihood for Latent Models by Rusakov and Geiger, (UAI-03).
- Algebraic Geometry of Bayesian Networks by Garcia, M. Stillman, B. Sturmfels, JSC.
- Algebraic Statistics in Model Selection by L. D. Garcia, (submitted UAI-04).



- Binary Random Variables:
  $X = \{X_1, X_2, X_3, X_4, X_5\} = \{F, E, S, A, C\}$.
- Joint Probability Distribution: $p(X = u) = \prod_{i=1}^{n=5} p(X_i = u_i | \, \mathrm{pa}_i)$.

$$p(F, E, S, A, C) = p(F)p(E)p(S|F)p(A|FE)p(C|A)$$

- Number of joint space parameters $D = 2^5 = 32$.
- Number of model parameters $E = 1 + 1 + 2 + 4 + 2 = 10$.
- The image of $\phi : \mathbb{R}^E \longrightarrow \mathbb{R}^D$ contains the set of all joint distributions that factor according to G.

Homomorphisms and Recursive Factorization

- $p(F = u_1, E = u_2, S = u_3, A = u_4, C = u_5) = p(u_1)p(u_2)p(u_3|u_1)p(u_4|u_1, u_2)p(u_5|u_4)$.
- Let p_u be an indeterminate representing $p(u_1, u_2, u_3, u_4, u_5)$.
- Let $\mathbb{R}[D] = \mathbb{R}[p_u \mid u \in \{0,1\}^5]$.
- Let q_{ijk} be an indeterminate representing $p(X_i = j | \operatorname{pa}_i = k)$.
- Let $\mathbb{R}[E] = \mathbb{R}[q_{10}, q_{20}, q_{300}, q_{301}, \ldots, q_{501}]$.
- $\phi : \mathbb{R}^E \to \mathbb{R}^D$ is specified by $\Phi : \mathbb{R}[D] \to \mathbb{R}[E]$

$$p_{00000} \longrightarrow q_{10}q_{20}q_{300}q_{4000}q_{500}$$

$$\vdots$$

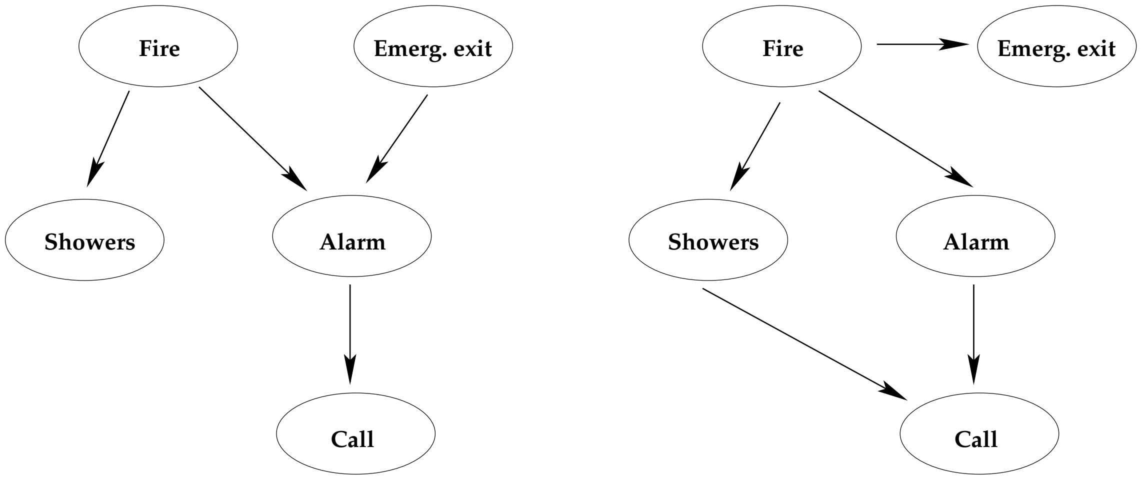
$$p_{11111} \longrightarrow (1 - q_{10})(1 - q_{20})(1 - q_{301})(1 - q_{4011})(1 - q_{501})$$

- The variety $V(\ker(\Phi))$ contains the set of all joint probability distributions that factor according to G.

- Choose the appropriate model M that best fits a given set of observations D.

F	E	S	A	C
0	1	0	1	1
0	0	0	1	1
1	0	1	0	0
0	1	0	0	0
0	0	0	0	0
0	1	0	1	1
0	1	0	1	0

- Choose the appropriate model M that best fits a given set of observations D.



Bayesian Approach to Model Selection

Choose a model M that maximizes the posterior model probability:

$$p(M|D) \propto p(M,D) = p(M)p(D|M)$$
$$= p(M)\int_{\Omega} p(D|M,\omega)p(\omega|M)d\omega$$

- $p(M)$ is the structure prior.
- $p(D|M)$ is called the marginal likelihood.
- Ω denotes the domain of the model parameters ω.
- $p(\omega|M)$ is the parameter prior.

BIC: Choose a model that maximizes $\ln p(D|M)$.

$$\ln p(D|M_1) = -23.26 \qquad \ln p(D|M_2) = -23.46$$

Bayesian Approach to Model Selection

Choose a model M that maximizes the posterior model probability:

$$p(M|D) \propto p(M,D) = p(M)p(D|M)$$
$$= p(M) \int_{\Omega} p(D|M,\omega)p(\omega|M)d\omega$$

- $p(M)$ is the structure prior.
- $p(D|M)$ is called the marginal likelihood.
- Ω denotes the domain of the model parameters ω.
- $p(\omega|M)$ is the parameter prior.

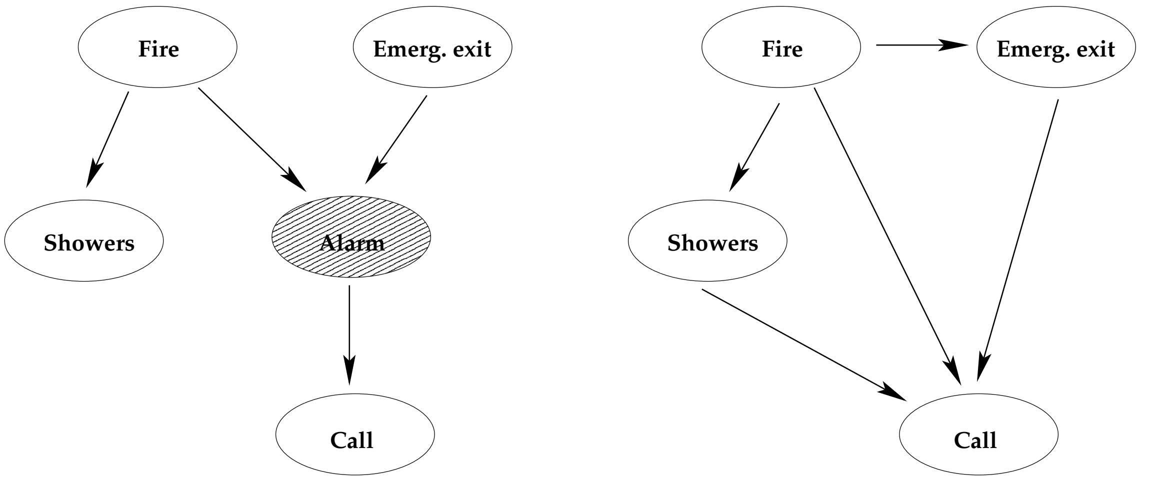
BIC: Choose a model that maximizes $\ln p(D|M)$.

BIC score: $\ln p(D|M) = N \ln p(D|\omega_{ML}) - \frac{d}{2} \ln N + O(1)$, [Haughton (1988)]

Latent Models

F	E	S	C
0	1	0	1
0	0	0	1
1	0	1	0
0	1	0	0
0	0	0	0
0	1	0	1
0	1	0	0

Latent Models

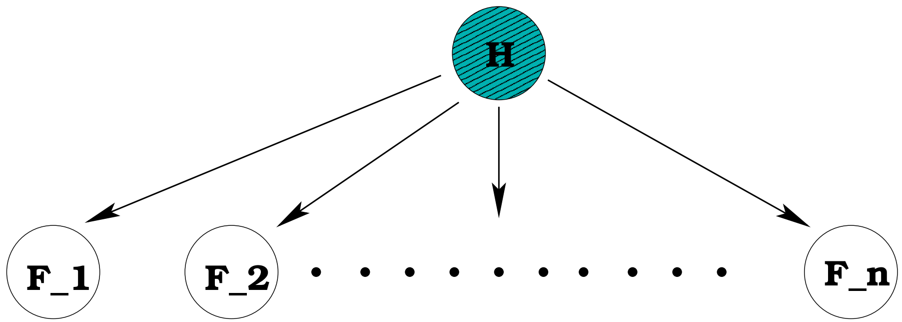


$$p(u_1, u_2, u_3, u_5) = \sum_{l=1}^{2} p(u_1, u_2, u_3, l, u_5)$$
$$= \sum_{l=1}^{2} p(u_1)p(u_2)p(u_3|u_1)p(l|u_1, u_2)p(u_5|l).$$

$$p(u_1, u_2, u_3, u_5) = \sum_{l=1}^{2} p(u_1, u_2, u_3, l, u_5)$$
$$= \sum_{l=1}^{2} p(u_1)p(u_2)p(u_3|u_1)p(l|u_1, u_2)p(u_5|l).$$

- Let $p_{u_1 u_2 u_3 + u_5} = \sum_{l=1}^{2} p_{u_1 u_2 u_3 l u_5}$ be a linear form representing the observable probabilities $p(u_1, u_2, u_3, u_5)$.
- Let $\mathbb{R}[D'] \subset \mathbb{R}[D]$ be the subring generated by these linear forms.
- The variety $V(\ker(\Phi) \cap \mathbb{R}[D'])$ contains the set of all observable joint probability distributions that factor according to G.

Naive Bayesian Network (1)



- H is the hidden variable, and its levels $1, 2, \ldots, r$ are called the classes.
- The observed random variables $F_1, \ldots, F_n$ are the features of the model.
- $\ker(\Phi)$ is the ideal of the join of r copies of the **Segre** variety

$$S_{r_1, r_2, \ldots, r_n} \quad := \quad \mathbb{P}^{r_1 - 1} \times \mathbb{P}^{r_2 - 1} \times \cdots \times \mathbb{P}^{r_n - 1} \quad \subset \quad \mathbb{P}^{r_1 r_2 \cdots r_n - 1}.$$

- The naive Bayesian network with r classes and n features corresponds to the r-th secant variety of a Segre product of n projective spaces:

$$V(\ker(\Phi) \cap \mathbb{R}[D']) \quad = \quad S^r_{r_1, r_2, \ldots, r_n}$$

- $S = S_{r_1,r_2,\ldots,r_n}$ is contained in a space of dimension $r_1 r_2 \cdots r_n - 1$ (number of joint distribution parameters).
- $\dim S_{r_1,r_2,\ldots,r_n}$ equals $d = r_1 + r_2 + \cdots + r_n - n$.
- The expected dimension of S^r equals

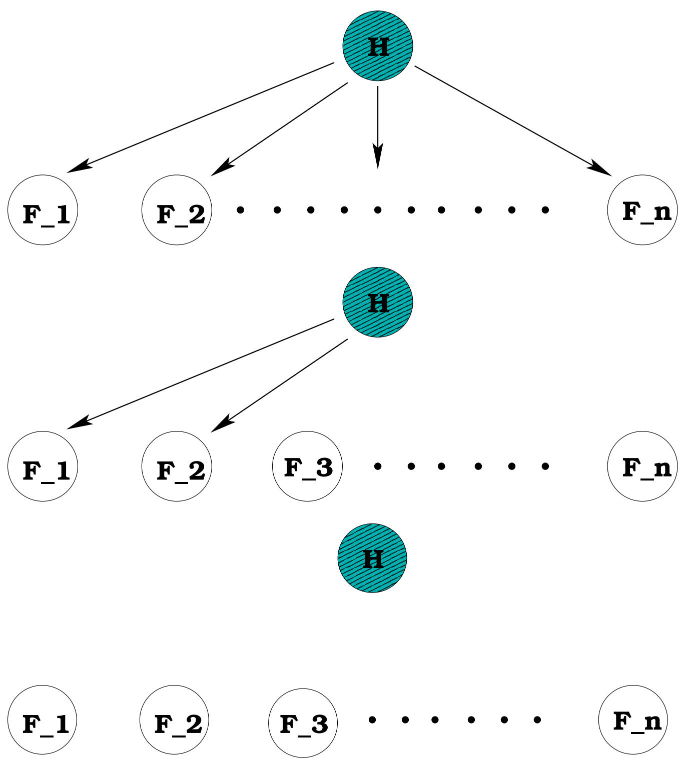
$$\min\{\prod_{i=1}^{n} r_i - 1, rd + r - 1\}.$$

- $rd + r - 1$ equals the number of model parameters of M.
- When S^r does not have the expected dimension, S is $(r-1)$–defective.

Singularities

- The r-th secant variety of any projective variety is singular along the $(r-1)$-st secant variety.
- If $r = r_i = 2$, the naive Bayesian model M with two features is singular along the Segre variety S.

[Geiger, Heckerman, King, Meek 2001]



$$S^2 = S^2_{2,2,\ldots,2} = Sec(\mathbb{P}^1 \times \cdots \times \mathbb{P}^1)$$

$$S' = S^2_{2,2} \times \mathbb{P}^{r_3 - 1} \times \cdots \times \mathbb{P}^{r_n - 1}$$

$$S'' = S = \mathbb{P}^1 \times \cdots \times \mathbb{P}^1$$

Asymptotic Approximation for the Marginal Likelihood

- Maximize $p(D|M) = \int_\Omega e^{\mathcal{L}(Y_D, N|\omega, M)} \mu(\omega|M) d\omega$.
- $N = |D|$, $\mu(\omega|M)$ is the prior parameter density for M, and $\mathcal{L}$ is the log-likelihood function of M.

Theorem (Watanabe 2001, Geiger and Rusakov 2002)

Let $I(N) = \int_{W_\epsilon} e^{-Nf(w)} \mu(w) dw$ where W_ϵ is some closed ϵ-box around w_0, which is a minimum point of f in W_ϵ, and $f(w_0) = 0$. Assume that f and μ are analytic functions, $\mu(w_0) \neq 0$. Then,

$$\ln I(N) = \lambda_1 \ln N + (m_1 - 1) \ln \ln N + O(1)$$

where the rational number $\lambda_1 < 0$ and m_1 are the largest pole and its multiplicity of the analytic continuation of

$$J(\lambda) = \int_{f(w)<\epsilon} f(w)^\lambda \mu(w) dw \qquad Re(\lambda) > 0$$

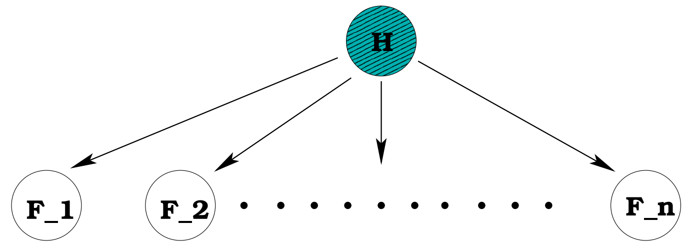
Resolution Theorem [Atiyah 1970]

Let $f(w)$ be a real analytic function defined in a neiborhood of $0 \in \mathbb{R}^d$. Then there exists an open set W that contains 0, a real analytic manifold U, and a proper analytic map $g : U \longrightarrow W$ such that:

1. $g : U \setminus U_0 \longrightarrow W \setminus W_0$ is an isomorphism, where $W_0 = f^{-1}(0)$ and $U_0 = g^{-1}(W_0)$.
2. For each point $p \in U$ there are local analytic coordinates $(u_1, \ldots, u_d)$ centered at p so that, locally near p,

$$f(g(u_1, \ldots, u_d)) = a(u_1, \ldots, u_d) u_1^{k_1} \cdots u_d^{k_d},$$

where $k_i \geq 0$ and $a(u)$ is an analytic function with analytic inverse $1/a(u)$.



- Let $a_i = p(F_i = 1|H = 1)$, $b_i = p(F_i = 1|H = 0)$, $t = p(H = 1)$, $\theta_x = p(F = x)$.

$$\theta_x = t \prod_{i=1}^{n} a_i^{x_i} (1 - a_i)^{1-x_i} + (1 - t) \prod_{i=1}^{n} b_i^{x_i} (1 - b_i)^{1-x_i}.$$

- Let $I[N, Y_D]$ be the marginal likelihood of data with averaged sufficient statistics Y_D

$$I[N, Y_D] = \int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega.$$

Theorem [Geiger and Rusakov 2002]

$$I[N, Y_D] = \int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega.$$

Assume the following conditions

1. The density $\mu(\omega)$ is bounded and bounded away from zero on Ω.
2. The statistics $Y_D = (Y_1, \ldots, Y_{2^n})$ satisfy $Y_i > 0$.
3. There exists N_0 such that Y_D equals the limiting statistics Y for all $N \geq N_0$.

Theorem [Geiger and Rusakov 2002]

$$I[N, Y_D] = \int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega.$$

Then for $n \geq 3$ as $N \longrightarrow \infty$:

- If $Y \in S^2 \setminus S'$ (regular point)

$$\ln I[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n+1}{2} \ln N + O(1),$$

- If $Y \in S' \setminus S''$ (type 1 singularity)

$$\ln I[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n-1}{2} \ln N + O(1),$$

- If $Y \in S''$ (type 2 singularity)

$$\ln I[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1),$$