Exact Evaluation of Marginal Likelihood Integrals

Shaowei Lin (Joint work with B. Sturmfels, Z. Xu)

1 May 2008

shaowei@math.berkeley.edu

University of California, Berkeley

Menu

Appetizer

The Occasionally Dishonest Coin-Tosser

Main Course

Marginal Likelihood Integrals

Mixtures of Independence Model

Exact Formula for the Integral

Approximations of the Integral

Dessert

Two Different Examples

Occasionally Dishonest Coin-Tosser

- *The Deal*:
  Four coin tosses. If all equal, you lose.

Occasionally Dishonest Coin-Tosser

- *The Deal*:
  Four coin tosses. If all equal, you lose.

- *The Real Deal*:
  Two coins are involved, one fair and one biased.
  Most of the time, he uses the fair coin,
  but occasionally, he uses the biased coin.

Occasionally Dishonest Coin-Tosser

- *The Deal*:
  Four coin tosses. If all equal, you lose.

- *The Real Deal*:
  Two coins are involved, one fair and one biased.
  Most of the time, he uses the fair coin,
  but occasionally, he uses the biased coin.

- *The Data,* U:

#Heads	0	1	2	3	4
#Occurrences	51	18	73	25	75

Occasionally Dishonest Coin-Tosser

- Model 1:

Coin: $0 \leq \theta_h, \theta_t \leq 1, \quad \theta_h + \theta_t = 1.$

Probability of i heads, p_i = $\binom{4}{i}\theta_h^i \theta_h^{4-i}$.

Likelihood of data, $L_U(\theta)$ = $Z p_0^{51} p_1^{18} p_2^{73} p_3^{25} p_4^{75}$ = $Z 4^{43} 6^{73} \theta_h^{539} \theta_t^{429}$,

where Z = $242!/(51! \cdot 18! \cdot 73! \cdot 25! \cdot 75!)$.

Occasionally Dishonest Coin-Tosser

- Model 1:

Coin: $0 \leq \theta_h, \theta_t \leq 1, \quad \theta_h + \theta_t = 1.$

Probability of i heads, p_i = $\binom{4}{i}\theta_h^i \theta_h^{4-i}$.

Likelihood of data, $L_U(\theta)$ = $Zp_0^{51}p_1^{18}p_2^{73}p_3^{25}p_4^{75}$ = $Z4^{43}6^{73}\theta_h^{539}\theta_t^{429}$,

where Z = $242!/(51! \cdot 18! \cdot 73! \cdot 25! \cdot 75!)$.

- Model 2:

Coin 0: $0 \leq \theta_h, \theta_t, \leq 1, \quad \theta_h + \theta_t = 1.$

Coin 1: $0 \leq \rho_h, \rho_t \leq 1, \quad \rho_h + \rho_t = 1.$

Choice of coin: $0 \leq \sigma_0, \sigma_1 \leq 1, \quad \sigma_0 + \sigma_1 = 1.$

Probability of i heads, p_i = $\binom{4}{i}(\sigma_0\theta_h^i\theta_h^{4-i} + \sigma_1\rho_h^i\rho_t^{4-i})$.

Likelihood of data, $L_U(\theta)$ = $Zp_0^{51}p_1^{18}p_2^{73}p_3^{25}p_4^{75}$.

Occasionally Dishonest Coin-Tosser

- **Question:** How do we do model selection?

Occasionally Dishonest Coin-Tosser

- **Question:** How do we do model selection?
- **Method 1:** Maximum Likelihood
  Compare the maximum values of the likelihood functions.

Occasionally Dishonest Coin-Tosser

- **Question:** How do we do model selection?
- **Method 1:** Maximum Likelihood
  Compare the maximum values of the likelihood functions.
- **Method 2:** Marginal Likelihood
  Integrate the likelihood functions over the parameter space.

$$\int_{\Theta} L_U(\theta) d\theta$$

  - Can be viewed as “average” probability of model.
  - Probability measures on the parameter space represent prior beliefs.
  - Max. likelihood is integrating with unit measure on the set of optimal parameters.

Marginal Likelihood Integrals

Current State of Affairs

- Very difficult to compute exactly.
- Tackled using MCMC, importance sampling methods.
- Approximation formulas limited to special cases.
- Accuracy of above methods and formulas questionable.

Our Goal

- Show that they can be computed *exactly* in *many* cases previously thought impractical.

Mixtures of Independence Models

Coin Toss Example

Random Variables

$X_1, X_2, \ldots, X_4 \in \{0,1\}$ identically distributed.

Model Parameters

$\theta_0, \theta_1, \quad \theta \in \Delta_1.$

Independence Model

$p_v = \theta^{a_v}$, where a_v are the columns of a 2×16 matrix

$$A \quad = \quad \begin{array}{c c} & \begin{array}{ccccccc} p_{0000} & p_{0001} & p_{0010} & \cdots & p_{1101} & p_{1110} & p_{1111} \end{array} \\ \begin{array}{c} \theta_0 \\ \theta_1 \end{array} & \begin{pmatrix} 4 & 3 & 3 & \ldots & 1 & 1 & 0 \\ 0 & 1 & 1 & \ldots & 3 & 3 & 4 \end{pmatrix} \end{array}$$

Two-Mixture

$$p_v = \sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v}, \quad \sigma \in \Delta_1.$$

Three-Mixture

$$p_v = \sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v} + \sigma_2 \tau^{a_v}, \quad \sigma \in \Delta_2.$$

Mixtures of Independence Models

- Random Variables
$$X_1^{(1)}, X_2^{(1)}, \ldots, X_{s_1}^{(1)} \in \{0, \ldots, t_1\} \text{ identically distributed,}$$
$$\ldots$$
$$X_1^{(k)}, X_2^{(k)}, \ldots, X_{s_k}^{(k)} \in \{0, \ldots, t_k\} \text{ identically distributed.}$$

Mixtures of Independence Models

- Random Variables

$$X_1^{(1)}, X_2^{(1)}, \ldots, X_{s_1}^{(1)} \in \{0, \ldots, t_1\} \text{ identically distributed,}$$
$$\ldots$$
$$X_1^{(k)}, X_2^{(k)}, \ldots, X_{s_k}^{(k)} \in \{0, \ldots, t_k\} \text{ identically distributed.}$$

- Model Parameters

$$\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_{t_1}^{(1)}, \quad \theta^{(1)} \in \Delta_{t_1}.$$
$$\ldots$$
$$\theta_0^{(k)}, \theta_1^{(k)}, \ldots, \theta_{t_k}^{(k)}, \quad \theta^{(k)} \in \Delta_{t_k}.$$

Mixtures of Independence Models

- Random Variables

$$X_1^{(1)}, X_2^{(1)}, \ldots, X_{s_1}^{(1)} \in \{0, \ldots, t_1\} \text{ identically distributed,}$$
$$\cdots$$
$$X_1^{(k)}, X_2^{(k)}, \ldots, X_{s_k}^{(k)} \in \{0, \ldots, t_k\} \text{ identically distributed.}$$

- Model Parameters

$$\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_{t_1}^{(1)}, \quad \theta^{(1)} \in \Delta_{t_1}.$$
$$\cdots$$
$$\theta_0^{(k)}, \theta_1^{(k)}, \ldots, \theta_{t_k}^{(k)}, \quad \theta^{(k)} \in \Delta_{t_k}.$$

- Independence Model

Can be represented by a $d \times n$ matrix A, where

d = #parameters = $(t_1 + 1) + (t_2 + 1) + \cdots + (t_k + 1)$,

n = #outcomes = $(t_1 + 1)^{s_1} (t_2 + 1)^{s_2} \cdots (t_k + 1)^{s_k}$.

The column a_v corresponds to the probability $p_v = \theta^{a_v}$.

Mixtures of Independence Models

- Random Variables

$$X_1^{(1)}, X_2^{(1)}, \ldots, X_{s_1}^{(1)} \in \{0, \ldots, t_1\} \text{ identically distributed,}$$
$$\ldots$$
$$X_1^{(k)}, X_2^{(k)}, \ldots, X_{s_k}^{(k)} \in \{0, \ldots, t_k\} \text{ identically distributed.}$$

- Model Parameters

$$\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_{t_1}^{(1)}, \quad \theta^{(1)} \in \Delta_{t_1}.$$
$$\ldots$$
$$\theta_0^{(k)}, \theta_1^{(k)}, \ldots, \theta_{t_k}^{(k)}, \quad \theta^{(k)} \in \Delta_{t_k}.$$

- Independence Model

Can be represented by a $d \times n$ matrix A, where

d = #parameters = $(t_1 + 1) + (t_2 + 1) + \cdots + (t_k + 1)$,

n = #outcomes = $(t_1 + 1)^{s_1} (t_2 + 1)^{s_2} \cdots (t_k + 1)^{s_k}$.

The column a_v corresponds to the probability $p_v = \theta^{a_v}$.

- Mixtures

$$p_v = \sigma_0 \theta^{a_v} + \ldots + \sigma_l \rho^{a_v}, \quad \sigma \in \Delta_l.$$

Mixtures of Independence Models

- Random Variables
  $$X_1^{(1)}, X_2^{(1)}, \ldots, X_{s_1}^{(1)} \in \{0, \ldots, t_1\} \text{ identically distributed,}$$
  $$\ldots$$
  $$X_1^{(k)}, X_2^{(k)}, \ldots, X_{s_k}^{(k)} \in \{0, \ldots, t_k\} \text{ identically distributed.}$$
- Model Parameters
  $$\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_{t_1}^{(1)}, \quad \theta^{(1)} \in \Delta_{t_1}.$$
  $$\ldots$$
  $$\theta_0^{(k)}, \theta_1^{(k)}, \ldots, \theta_{t_k}^{(k)}, \quad \theta^{(k)} \in \Delta_{t_k}.$$
- Independence Model

  Can be represented by a $d \times n$ matrix A, where

  d = #parameters = $(t_1 + 1) + (t_2 + 1) + \cdots + (t_k + 1)$,

  n = #outcomes = $(t_1 + 1)^{s_1} (t_2 + 1)^{s_2} \cdots (t_k + 1)^{s_k}$.

  The column a_v corresponds to the probability $p_v = \theta^{a_v}$.
- Mixtures
  $$p_v = \sigma_0 \theta^{a_v} + \ldots + \sigma_l \rho^{a_v}, \quad \sigma \in \Delta_l.$$
- Data
  $$U = (U_v), \quad N = \sum_v U_v.$$

Exact Formula for the Integral

Main Formula:

$$\int_{\Delta_m} \theta_0^{b_0}\theta_1^{b_1}\cdots\theta_m^{b_m}d\theta = \frac{m!\cdot b_0!\cdot b_1!\cdot\cdots\cdot b_m!}{(b_0+b_1+\cdots+b_m+m)!}$$

Exact Formula for the Integral

Main Formula:

$$\int_{\Delta_m} \theta_0^{b_0} \theta_1^{b_1} \cdots \theta_m^{b_m} d\theta = \frac{m! \cdot b_0! \cdot b_1! \cdot \cdots \cdot b_m!}{(b_0 + b_1 + \cdots + b_m + m)!}$$

Independence Model:

$$\begin{aligned} L_U(\theta) &= Z \cdot \theta^b \\ \int_\Theta L_U(\theta)\ d\theta &= Z \cdot \prod_{i=1}^{k} \frac{t_i!\, b_0^{(i)}!\, b_1^{(i)}! \cdots b_{t_i}^{(i)}!}{(s_i N + t_i)!} \end{aligned}$$

where $Z = N!/\prod_v U_v!$ and $b = AU$.

Recall that the maximum likelihood of an independence model is easy to compute.

Here, we see that its marginal likelihood is also easy to compute.

Exact Formula for the Integral

Main Formula:

$$\int_{\Delta_m} \theta_0^{b_0}\theta_1^{b_1}\cdots\theta_m^{b_m} d\theta = \frac{m!\cdot b_0!\cdot b_1!\cdot \cdots \cdot b_m!}{(b_0+b_1+\cdots+b_m+m)!}$$

Mixture of Independence Model:

$$\begin{aligned} L_U(\sigma,\theta,\rho) &= Z\cdot \prod_v (\sigma_0\theta^{a_v}+\sigma_1\rho^{a_v})^{U_v} \\ &= Z\cdot \sum_b \phi(b)\cdot \sigma^{\alpha(b)}\cdot \theta^b \cdot \rho^{\beta(b)} \\ \int_{\Delta_1\times\Theta\times\Theta} L_U(\sigma,\theta,\rho)\ d\sigma d\theta d\rho &= Z\cdot \sum_b \phi(b) \int_{\Delta_1}\sigma^{\alpha(b)}d\sigma \int_\Theta \theta^b d\theta \int_\Theta \rho^{\beta(b)}d\rho \end{aligned}$$

where $\phi(b)$ is the coefficient of θ^b in the expansion of $\prod_v(\theta^{a_v}+1)^{U_v}$,
$\beta(b) = AU - b$, and $\alpha(b) = \frac{1}{\text{column sum of A}}(b,\beta(b))$.

Exact Formula for the Integral

Formula:

$$\int_{\Delta_1 \times \Theta \times \Theta} L_U(\sigma, \theta, \rho)\ d\sigma d\theta d\rho = Z \cdot \sum_b \phi(b) \int_{\Delta_1} \sigma^{\alpha(b)} d\sigma \int_{\Theta} \theta^b d\theta \int_{\Theta} \rho^{\beta(b)} d\rho$$

Computational Considerations:

- Bottleneck is in computing $\phi(\cdot)$.

  Use the sum-product algorithm (dynamic programming).

  Exploit low rank of matrix A to store, compute $\phi(\cdot)$ efficiently.
- Only need to sum half the terms because of symmetry.
- Precompute and look-up values of factorials.
- Computation is highly parallelizable.
- Maple library:

  `http://math.berkeley.edu/~shaowei/integrals.html`

Approximations of the Integral

Question:

Suppose $U = NY$ where Y is a fixed vector with $\sum_v Y_v = 1$.

As $N \to \infty$, how does the log marginal likelihood behave?

$$\log \int_\Theta L_U(\theta) d\theta$$

Approximations of the Integral

Question:

Suppose $U = NY$ where Y is a fixed vector with $\sum_v Y_v = 1$.

As $N \to \infty$, how does the log marginal likelihood behave?

Answer 1:

$$\log \int_\Theta L_U(\theta) d\theta \to -\infty$$

.

Approximations of the Integral

Question:

Suppose $U = NY$ where Y is a fixed vector with $\sum_v Y_v = 1$.
As $N \to \infty$, how does the log marginal likelihood behave?

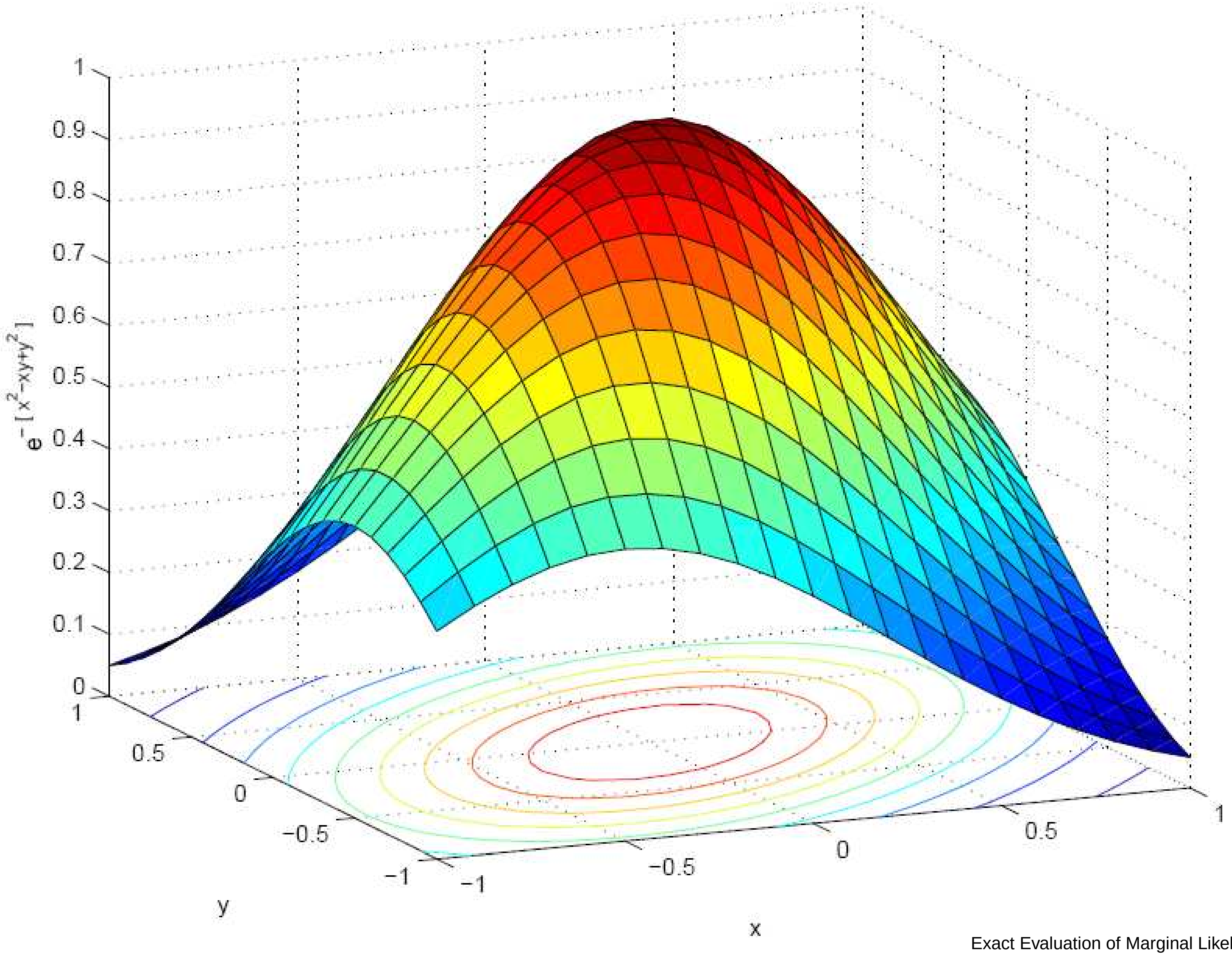
Answer 2: BIC Score

$$\log \int_\Theta L_U(\theta) d\theta \quad \approx \quad \log L(\hat{\theta}) - \frac{d}{2} \log N$$

where d is the dimension of the model and $L(\hat{\theta})$ is the *maximum* likelihood.
BIC stands for Bayesian Information Criterion.

Assumes that the model is in the exponential family. In particular, the model has one local maxima. As $N \to \infty$, the “main bulk” of the integral accumulates near the maximum likelihood.

Approximations of Integral



Approximations of the Integral

Question:

Suppose $U = NY$ where Y is a fixed vector with $\sum_v Y_v = 1$.

As $N \to \infty$, how does the log marginal likelihood behave?

Answer 3: Laplace Approximation

$$\log \int_\Theta L_U(\theta) d\theta \quad \approx \quad \log L(\hat{\theta}) - \frac{1}{2} \log |\det H(\hat{\theta})| + \frac{d}{2} \log 2\pi$$

where H is the Hessian of the log-likelihood function $\log L$.

Only assumes that L is twice differentiable, convex and achieves maximum on internal point.

Back to the Coin Tosser

- **Maximum Likelihood**

  Independence: $0.1443566234 \times 10^{-54}$

  Mixture: $0.1395471101 \times 10^{-18}$

Back to the Coin Tosser

- **Maximum Likelihood**

  Independence: $0.1443566234 \times 10^{-54}$

  Mixture: $0.1395471101 \times 10^{-18}$

- **Marginal Likelihood**

  Independence: $0.5773010423 \times 10^{-56}$

  Mixture: $0.7788716339 \times 10^{-22}$ (Actual)

  $0.3706788423 \times 10^{-22}$ (BIC)

  $0.4011780794 \times 10^{-22}$ (Laplace)

Back to the Coin Tosser

- **Maximum Likelihood**

  Independence: $0.1443566234 \times 10^{-54}$

  Mixture: $0.1395471101 \times 10^{-18}$

- **Marginal Likelihood**

  Independence: $0.5773010423 \times 10^{-56}$

  Mixture: $0.7788716339 \times 10^{-22}$ (Actual)

  $0.3706788423 \times 10^{-22}$ (BIC)

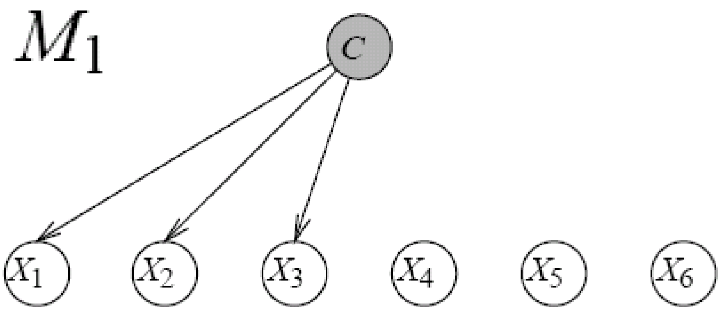
  $0.4011780794 \times 10^{-22}$ (Laplace)

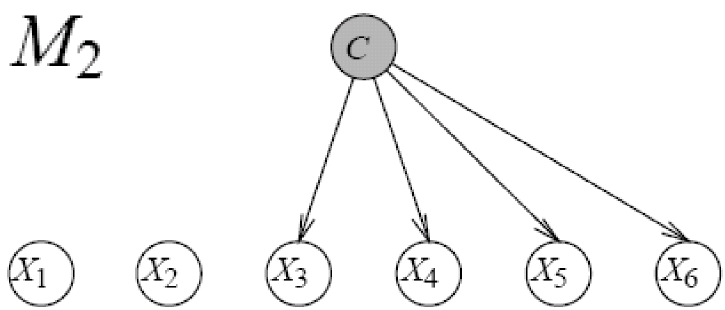
- In fact,
  the marginal likelihood for the mixture model is exactly...

Back to the Coin Tosser

$$
\frac{\begin{array}{l}
2805748035222313067135398014075361975978864622235225616054475981674736781 \\
1799443476719649200942628578141429547789194845757944946345970873531023 04 \\
2489712762833760845774052573250231055298084652703225819785515675807589 25 \\
1102576752971175448613852605506591528125476141208021767320470301818791 09 \\
4936908443047454078425332265435670406065197838062752909347743870834021 20 \\
4638972697649334519554413471422043990575435789632065689304973717297696 06 \\
0415632400741050563477342238636399647384755308009778572454838389096925 96 \\
88769804869503436965543936
\end{array}}{\begin{array}{l}
3602324071338125874577562671962054628339147256791746496077298664579499 43 \\
6836889049486689507051463879264328153845162002285178224453663460279080 75 \\
8904156945946390977724512859312036096765746313969020541775346907766998 18 \\
0397769609299339804266010207548603870980861129358173839607260454683402 08 \\
3005508959248902903340347663670605747176619993139607889832999867603350 32 \\
0070482837740687067608852004726493742428623588390160566874549440724360 48 \\
4442163404900024396516685851371805424013821775746444698614706300105139 96 \\
263775153793334976819060141283354099489865061875
\end{array}}.
$$

BIC can be wrong!

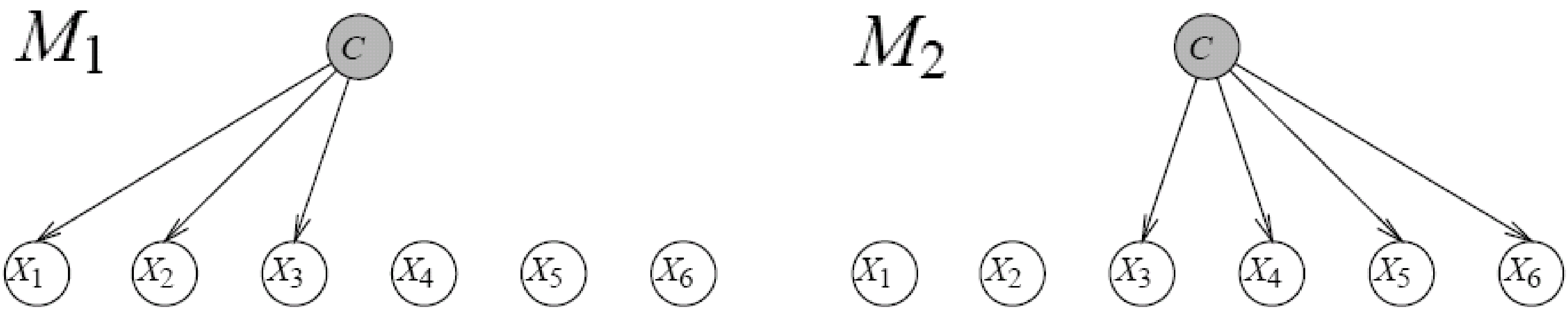




Consider the two models above and data below (N=36).

		$X_4X_5X_6$							
		000	001	010	011	100	101	110	111
$X_1X_2X_3$	000	2	3		1	3	5	1	1
	001								
	010						1		
	011								
	100	3	4	1	1	2	3	1	1
	101		1						
	110	1	1						
	111								

BIC can be wrong!



Model Selection:

- BIC Score: M1's score is better than M2's.
- Actual Marginal Likelihood:

M1 $$\frac{2673620257358279100801924830063571461298286189}{5953897913266720923361652444310905663581365769429178055600000000} \approx 0.449 \times 10^{-17}$$

M2 $$\frac{48293401975547884279365197096430603703508201757248809211637315169}{873248402971499818328286563178459524881596589864311287443444152295294483200000000} \approx 0.553 \times 10^{-17}$$

Thus, a true Bayesian should choose M2 over M1,
even though the BIC score tells him otherwise!

References

1. D.M. Chickering and D. Heckerman: Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables, *Machine Learning* **29** (1997) 181-212; Microsoft Research Report, MSR-TR-96-08.

2. D. Geiger and D. Rusakov: Asymptotic model selection for naive Bayesian networks, *Journal of Machine Learning Reseach* **6** (2005) 1–35.

3. S. Hoşten, A. Khetan and B. Sturmfels: Solving the likelihood equations, *Foundations of Computational Mathematics* **5** (2005) 389–407.

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