Exact Evaluation of Marginal Likelihood Integrals

Shaowei Lin[1], Bernd Sturmfels[1], and Zhiqiang Xu[2]

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

1. Department of Mathematics, University of California, Berkeley
2. Academy of Mathematics and System Sciences, Chinese Academy of Sciences

Abstract

Inference in Bayesian statistics involves the evaluation of marginal likelihood integrals. We present algebraic algorithms for computing such integrals exactly for discrete data of small sample size. The underlying statistical models are mixtures of independent distributions, or, in geometric language, secant varieties of Segre-Veronese varieties.

Main Problem

How do we evaluate an integral like

$$\int_0^1 \int_0^1 \int_0^1 \prod_{i=0}^{4} [\sigma\theta^i(1-\theta)^{4-i} + (1-\sigma)\rho^i(1-\rho)^{4-i}]^{U_i} d\sigma d\theta d\rho$$

quickly and *exactly* for large U_i?

Incidentally, this integral is the rational number

$$\frac{\begin{array}{l} 280574803522231306713539801407536197597886462223522561605447598167473678\ 179944347671964920094262857814142954778919 48 \\ 457579449463459708735310230424897127628337608457740525732502310552980846\ 527032258197855156758075892511025767529711 75 \\ 448613852605506591528125476141208021767320470301818791094936908443047454\ 078425332265435670406065197838062752909347 74 \\ 387083402120463897269764933451955441347142204399057543578963206568930497\ 371729769606041563240074105056347734223863 63 \\ 996473847553080097785724548383890969259688769804869503436965543936 \end{array}}{\begin{array}{l} 360232407133812587457756267196205462833914725679174649607729866457949943\ 683688904948668950705146387926432815384516 20 \\ 022851782244536634602790807589041569459463909777245128593120360967657463\ 139690205417753469077669981803977696092993 39 \\ 804266010207548603870980861129358173839607260454683402083005508959248902\ 903340347663670605747176619993139607889832 99 \\ 986760335032007048283774068706760885200472649374242862358839016056687454\ 944072436048444216340490002439651668585137 18 \\ 054240138217757464446986147063001051399626377515379333497681906014128335\ 4099489865061875 \end{array}}.$$

Statistical Motivation

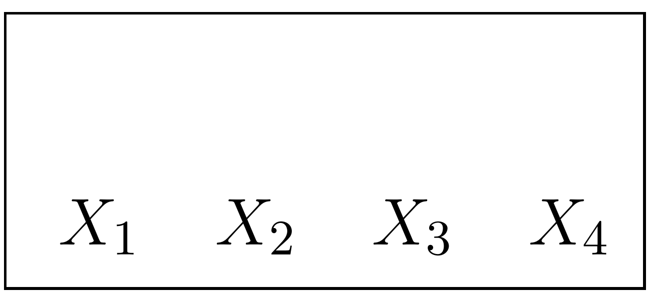
Example: The Cheating Coin Flipper

The Deal: Each game consists of four coin tosses.
The Data: Out of 242 games,

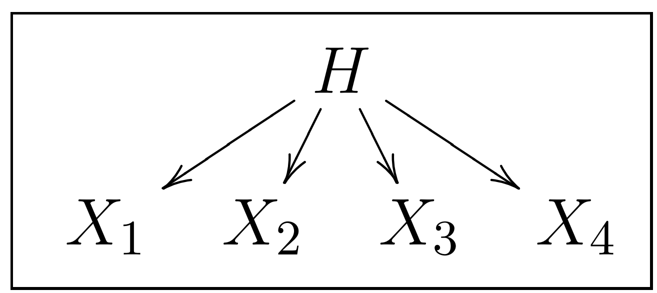
#Heads	0	1	2	3	4
Frequency	51	18	73	25	75

The Dilemma:
Was the coin swapped between the games?

Model *One*



Model *Two*



- Model *One*:

Parameters	Coin:	$0 \leq \theta_h, \theta_t \leq 1,\ \theta_h + \theta_t = 1$
Prob(i heads)	p_i =	$\binom{4}{i}\theta_h^i\theta_t^{4-i}$
Likelihood of data U	$L_U(\theta)$ =	$p_0^{51}p_1^{18}p_2^{73}p_3^{25}p_4^{75}$ = $4^{43}6^{73}\theta_h^{539}\theta_t^{429}$

- Model *Two*:

Parameters	Coin 0:	$0 \leq \theta_h, \theta_t, \leq 1,\ \ \theta_h + \theta_t = 1$
	Coin 1:	$0 \leq \rho_h, \rho_t \leq 1,\ \ \rho_h + \rho_t = 1$
	Choice of coin:	$0 \leq \sigma_0, \sigma_1 \leq 1,\ \ \sigma_0 + \sigma_1 = 1$
Prob(i heads)	p_i =	$\binom{4}{i}(\sigma_0\theta_h^i\theta_t^{4-i} + \sigma_1\rho_h^i\rho_t^{4-i})$
Likelihood of data U	$L_U(\theta)$ =	$p_0^{51}p_1^{18}p_2^{73}p_3^{25}p_4^{75}$

Question: How do we do model selection?

- Method 1: *Maximum Likelihood*

  Compare the maximum values of the likelihood functions.

$$\max_{\theta \in \Theta} L_U(\theta)$$

- Method 2: *Marginal Likelihood*

  Integrate the likelihood functions over the parameter space.

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

Marginal Likelihood Integrals

- Very difficult to compute exactly.
- Approximated using:
  1. MCMC, importance sampling methods
  2. Asymptotic formulas like BIC, Laplace, etc.
- Accuracy of above methods and formulas questionable.

Our Goals

- Show that they can be computed *exactly* in many cases previously thought impractical.
- Provide a standard for comparison in research on approximation methods.
- Develop new algebraic, combinatorial and geometric methods for solving such problems.

Computation

We compute marginal likelihood integrals exactly for the following class of statistical models:

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

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

- Independence Model

d = #parameters = $(t_1 + 1) + (t_2 + 1) + \cdots + (t_k + 1)$,
n = #outcomes = $(t_1 + 1)^{s_1} (t_2 + 1)^{s_2} \cdots (t_k + 1)^{s_k}$.
Can be represented by a $d \times n$ matrix A, where the column a_v corresponds to the probability $p_v = \theta^{a_v}$.

- Two-mixtures

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

- Data

$$U = (U_v), \quad N = \textstyle\sum_v U_v.$$

Key Formula:

Integrating a monomial over a simplex

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

Formula for Independence Model:

Let $b = AU$, $P = \Delta_{t_1}\times\cdots\times\Delta_{t_k}$. Since $L_U(\theta)=\theta^b$,

$$\int_P L_U(\theta)\ d\theta = \prod_{i=1}^{k}\int_{\Delta_{t_i}}\theta^{b^{(i)}}d\theta^{(i)} = \prod_{i=1}^{k}\frac{t_i!\,b_0^{(i)}!\,b_1^{(i)}!\ \cdots\ b_{t_i}^{(i)}!}{(s_iN+t_i)!}$$

Formula for Mixture Model:

Let $\Theta = \Delta_1 \times P \times P$. Expanding $\prod_v (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v}$ gives

$$\begin{aligned} L_U(\sigma, \theta, \rho) &= \sum_b \phi_A(b, U) \sigma^{(b,c)/a} \theta^b \rho^c \\ \int_\Theta L_U(\sigma, \theta, \rho)\ d\sigma d\theta d\rho &= \sum_b \phi_A(b, U) \int_{\Delta_1} \sigma^{(b,c)/a} d\sigma \int_P \theta^b d\theta \int_P \rho^c d\rho \end{aligned}$$

where $\phi_A(b, U)$ is the coefficient of θ^b in the expansion of $\prod_v (\theta^{a_v} + 1)^{U_v}$, $c = AU - b$, and a the column sum of A.

Computational Considerations:

- In the expansion of $L_U(\sigma, \theta, \rho) = \prod_v (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v}$, naive estimate of number of monomials is $\prod_v (U_v + 1)$.
  Actual number of monomials is *a lot less*.
  e.g. Coin Flipper model: 144,469,312 vs 48,646.
  Idea: exploit this reduction in the computation.

- Bottleneck is in computing $\phi_A(\cdot, U)$. A naive method is to use the formula $\phi_A(b, U) \;=\; \sum_{Ax=b} \prod_{v=1}^{n} \binom{U_v}{x_v}$.
  Idea: use recurrence formula

$$\phi_A(b, U) \;=\; \sum_{x_n=0}^{U_n} \binom{U_n}{x_n} \phi_{A \setminus a_n}(b - x_n a_n, U \setminus U_n)$$

- Monomials correspond to certain lattice points in a zonotope Z of dimension rank(A). In fact, these points are the image of the lattice points of the hypercuboid $\prod_v [0, U_v]$ under the linear transformation A.
  Idea: exploit low rank of A to store $\phi_A(\cdot, U)$ efficiently.

- Some other tricks:
  1. Only need to sum half the terms because of symmetry.
  2. Precompute and look-up values of factorials.
  3. Computation is highly parallelizable.

Computational results

	Time(seconds)	Memory(bytes)
Ignorant Integration$_1$	16.331	155,947,120
Naive Expansion$_2$	0.007	458,668

	Time(minutes)	Memory(bytes)
Naive Expansion$_2$	43.67	9,173,360
Our method	1.76	13,497,944

1. Using Maple's standard integration command `int`.
2. Using naive expansion of integrand with $\prod_v (U_v + 1)$ terms.

Comparison with Approximations

Given a statistical model, let d be its dimension and $L(\hat{\theta})$ the maximum likelihood.

- Bayesian Information Criterion (BIC)

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

- 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 $\log L$.

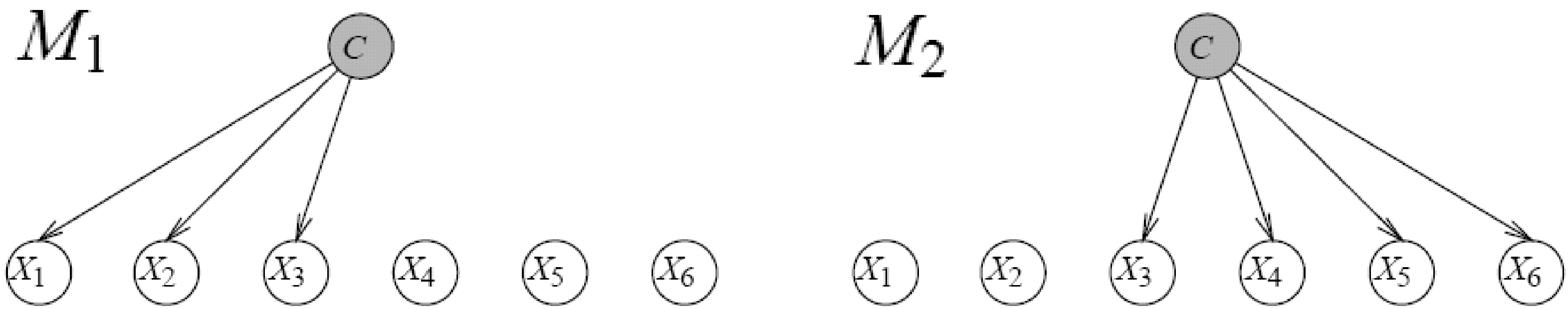
Computational Results

We compute the marginal likelihoods exactly for the Cheating Coin Flipper example using our method, and compare them with the BIC and Laplace approximations.

Model One:	$0.5773010423 \times 10^{-56}$	(Actual)
	$0.9279595380 \times 10^{-56}$	(BIC)
	$0.5777455911 \times 10^{-56}$	(Laplace)
Model Two:	$0.7788716339 \times 10^{-22}$	(Actual)
	$0.3706788423 \times 10^{-22}$	(BIC)
	$0.4011780794 \times 10^{-22}$	(Laplace)

Approximation via Resolution of Singularities

Consider the two hidden binary tree models below (cf. Geiger and Rusakov, 2002)



$$
\begin{aligned}
M_1: \quad p_v &= (\sigma_0 \theta_{v_1}^{(1)} \theta_{v_2}^{(2)} \theta_{v_3}^{(3)} + \sigma_1 \rho_{v_1}^{(1)} \rho_{v_2}^{(2)} \rho_{v_3}^{(3)}) \theta_{v_4}^{(4)} \theta_{v_5}^{(5)} \theta_{v_6}^{(6)} \\
M_2: \quad p_v &= \theta_{v_1}^{(1)} \theta_{v_2}^{(2)} (\sigma_0 \theta_{v_3}^{(3)} \theta_{v_4}^{(4)} \theta_{v_5}^{(5)} \theta_{v_6}^{(6)} + \sigma_1 \rho_{v_3}^{(3)} \rho_{v_4}^{(4)} \rho_{v_5}^{(5)} \rho_{v_6}^{(6)})
\end{aligned}
$$

Using Watanabe's method of approximating the integral using resolution of singularities, it was shown that despite having a lower BIC score than M_1, asymptotically model M_2 has a higher marginal likelihood than M_1.

We generated data of sample size $N = 36$ using the prescribed true distribution and computed the marginal likelihoods exactly with our methods.

M1
$$\frac{267362025735827910080192483006357146129828 6189}{59538979132667209233616524443109056635813657694291780556 0000000}$$

$$\approx 0.449 \times 10^{-17}$$

M2
$$\frac{4829340197554788427936519709643060370350820175724880921163731 5169}{87324840297149981832828656317845952488159658986431128744344415229529448320 00000000}$$

$$\approx 0.553 \times 10^{-17}$$

This calculation agrees with the earlier prediction.

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. M. Evans, Z. Gilula and I. Guttman: Latent class analysis of two-way contingency tables by Bayesian methods, *Biometrika* **76** (1989) 557–563.
3. D. Geiger and D. Rusakov: Asymptotic model selection for naive Bayesian networks, *Journal of Machine Learning Reseach* **6** (2005) 1–35.
4. S. Hoşten, A. Khetan and B. Sturmfels: Solving the likelihood equations, *Foundations of Computational Mathematics* **5** (2005) 389–407.
5. L. Pachter and B. Sturmfels: *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.
6. K. Yamazaki and S. Watanabe: Newton diagram and stochastic complexity in mixture of binomial distributions, in *Algorithmic Learning Theorem*, Springer Lecture Notes in Computer Science **3244** (2004) 350–364.