Exact Evaluation of Marginal Likelihood Integrals

Shaowei Lin [*]

Abstract

In Bayesian statistics, marginal likelihood integrals are important for model selection. Unfortunately, they are generally difficult to compute. We present algebraic algorithms for evaluating such integrals exactly for small sample sizes, and compare them with some existing approximations. Our methods will be applied to mixtures of discrete independence models, and are available as a library in `Maple`.

1 Introduction

Model selection is an important area of study in statistics because it occurs frequently in applications such as machine learning and computational biology. Typically, given some data and some parametric models, we wish to determine which model best describes the data. There are two general approaches to accomplishing this. The first is to solve for the parameters in each model that maximizes the likelihood function for the given data. The model which generates the largest likelihood value is deemed the best model. The second method involves integrating the likelihood function over the parameter space of each model. This integral is known as the marginal likelihood, and may be thought of as the "average" probability of the model over its parameter space. The model with the largest integral is chosen. This method is particularly useful when there are prior beliefs about the parameters defining the model. In this case, the integral is in some sense performed over a weighted parameter space.

[*] Joint work with Bernd Sturmfels and Zhiqiang Xu.

Unfortunately, marginal likelihood integrals are generally difficult to compute. Rather than evaluating them exactly, the common practice is to approximate them via Markov Chain Monte Carlo (MCMC) or importance sampling methods. There are also some approximation formulas which only work for certain classes of models. Furthermore, the accuracy of the estimates achieved by these methods and formulas is usually questionable. In this paper, we present methods for evaluating the integrals exactly for discrete data with small sample sizes. The marginal likelihoods of the models we study are rational numbers, and *exact* evaluation means finding that rational number rather than its floating point approximation. Our interest in developing these methods was inspired by recent developments which have found a link between the approximations and resolution of singularities in algebraic geometry [3, 5, 8, 9].

To demonstrate the scope of our methods, consider the following example of the *occasionally dishonest coin tosser*, a variant of the occasionally dishonest casino mentioned in [2] and [7, Ex 1.3]. Suppose a friend, an avid coin tosser, approaches you with a deal. He will take a quarter and toss it four times. If all four outcomes are heads or all four of them are tails, he wins; otherwise, you win. Thinking that you have got yourself a good deal, you agree. However, after playing the game several times with him, you suspect that things are not really what they seem. It appears as though he has two coins up his sleeves which he switches between occasionally. Being the cunning statistician that you are, you decide to observe several games and do some analysis. Below is a table summarizing your findings [6, Ex 9].

Number of heads	0	1	2	3	4
Number of occurrences	51	18	73	25	75

In particular, there are two models under consideration. This first model assumes that there is one coin, with head and tail probabilities θ_h and θ_t, (θ_h, θ_t) in the 1-simplex Δ_1. The second model supposes that there are two coins, with probabilities $(\theta_h, \theta_t) \in \Delta_1$ and $(\rho_h, \rho_t) \in \Delta_1$ respectively. The coin tosser chooses between the first and second coins with probabilities σ_0 and σ_1, $(\sigma_0, \sigma_1) \in \Delta_1$. We denote the observed data by the vector $\tilde{U} = (\tilde{U}_0, \ldots, \tilde{U}_4) = (51, 18, 73, 25, 75)$ and the sample size by $N = \sum_i \tilde{U}_i = 242$. Then, the marginal likehood of the first model is given by

$$\int_{\Delta_1} C \cdot \prod_{i=0}^{4} (\theta_h^i \theta_t^{4-i})^{\tilde{U}_i} d\theta \qquad (1)$$

while that of the second model is given by the integral

$$\int_{\Delta_1} C \cdot \prod_{i=0}^{4} (\sigma_0 \theta_h^i \theta_t^{4-i} + \sigma_1 \rho_h^i \rho^{4-i})^{\tilde{U}_i} d\sigma d\theta d\rho \tag{2}$$

Here, C is the normalizing constant

$$\frac{N!}{\tilde{U}_0! \cdot \tilde{U}_1! \cdot \tilde{U}_2! \cdot \tilde{U}_3! \cdot \tilde{U}_4!} \cdot 1^{\tilde{U}_0} 4^{\tilde{U}_1} 6^{\tilde{U}_2} 4^{\tilde{U}_3} 1^{\tilde{U}_4}$$

which ensures that the marginal likelihoods over all possible observed data $\tilde{U}$ with the same sample size sum to 1. The exponents in the integrands of (1) and (2) may be summarized with the matrix

$$\tilde{A} \quad = \quad \begin{array}{c} \\ \theta_t \\ \theta_h \end{array} \begin{array}{c} \begin{array}{ccccc} \tilde{U}_0 & \tilde{U}_1 & \tilde{U}_2 & \tilde{U}_3 & \tilde{U}_4 \end{array} \\ \begin{pmatrix} 4 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \end{array}. \tag{3}$$

Using the methods described in this paper, the first marginal likelihood works out to be the rational number

$$\frac{242! \cdot 429! \cdot 539! \cdot 4^{43} \cdot 6^{73}}{51! \cdot 18! \cdot 73! \cdot 25! \cdot 75! \cdot 969!} \tag{4}$$

while the second marginal likelihood is the rational number with numerator

28057480352222313067135398014075361975978864622235225616054475981
67473678179944347671964920094262857814142954778919484575794494 63
45970873531023042489712762833760845774052573250231055298084652 70
32258197855156758075892511025767529711754486138526055065915281 25
47614120802176732047030181879109493690844304745407842533226543 56
70406065197838062752909347743870834021204638972697649334519554 41
34714220439905754357896320656893049737172976960604156324007410 50
56347734223863639964738475530800977857245483838909692596887698 04
869503436965543936

and denominator

$$\begin{array}{l}
3602324071338125874577562671962054628339147256791746496077298664\\
5794994368368890494866895070514638792643281538451620022851782244\\
5366346027908075890415694594639097772451285931203609676574631396\\
9020541775346907766998180397769609299339804266010207548603870980\\
8611293581738396072604546834020830055089592489029033403476636706\\
0574717661999313960788983299986760335032007048283774068706760885\\
2004726493742428623588390160566874549440724360484442163404900024\\
3965166858513718054240138217757464446986147063001051399626377515\\
37933349768190601412833540994898650618 75.
\end{array}$$

The organization of the paper is as follows. In Section 2, we will define independence models, their mixtures and their marginal likelihoods. In Section 3, we will give a formula for the marginal likelihood integral of a mixture model, and discuss techniques for evaluating the formula efficiently. In Section 4, we give an overview of three existing approximation formulas for the integral: the BIC, the Laplace approximation and Watanabe's approximation. In Section 5, we compare these approximations with the actual values computed by our methods, and end off by evaluating the marginal likelihood of a real data set.

2 Independence Models and their Mixtures

The coin toss example in the previous section is an example of an independence model and its mixture. In general, suppose we have random variables

$$\begin{array}{c}
X_1^{(1)}, X_2^{(1)}, \ldots, X_{s_1}^{(1)}\\
X_1^{(2)}, X_2^{(2)}, \ldots, X_{s_2}^{(1)}\\
\cdots\\
X_1^{(k)}, X_2^{(k)}, \ldots, X_{s_k}^{(1)}
\end{array}$$

where $X_1^{(i)}, \ldots, X_{s_i}^{(i)}$ are identically distributed with values in $\{0, 1, \ldots, t_i\}$. We may store the state of all the above variables with a string

$$v \quad \in \quad V := \{0,1,\ldots,t_1\}^{s_1} \times \{0,1,\ldots,t_2\}^{s_2} \times \cdots \times \{0,1,\ldots,t_k\}^{s_k}.$$

of length $n = \prod_{i=1}^{k}(1+t_i)^{s_i}$. Then, the independence model $\mathcal{M}$ for these variables is the subset of the simplex Δ_{n-1} given parametrically by

$$p_v(\theta) \quad = \quad \text{Prob}(X_j^{(i)} = v_j^{(i)} \text{ for all } i, j) \quad = \quad \prod_{i=1}^{k} \prod_{j=0}^{t_i} \theta^{(i)}_{v_j^{(i)}}$$

for some model parameters $\theta_j^{(i)}$ satisfying

$$\theta^{(i)} \quad = \quad (\theta_0^{(i)}, \theta_1^{(i)}, \ldots, \theta_{t_i}^{(i)}) \quad \in \quad \Delta_{t_i}.$$

The model $\mathcal{M}$ is thus described by $d = \sum_{i=1}^{k}(1+t_i)$ parameters and may be summarized with a $d \times n$ matrix A where each column a_v corresponds to the exponents of the parameters θ in the formula for p_v. Note that A has a constant column sum $a = \sum_{i=1}^{k} s_i$. We may represent the data for our model by a non-negative integer vector $U \in \mathbb{N}^n$ where each entry U_v counts the number of times the state v is observed in a sample of size $N = \sum_v U_v$. The likelihood for a given data U is

$$L_U(\theta) \quad = \quad C \cdot \prod_v p_v(\theta)^{U_v} \quad = \quad C \cdot \theta^b. \tag{5}$$

where $b = AU$ is the sufficient statistic for this model and C is the normalizing constant $N!/\prod_v U_v!$. Here, θ^b is the monomial

$$\prod_{i=1}^{k} \prod_{j=0}^{t_i} (\theta_j^{(i)})^{b_j^{(i)}}$$

where the $b_j^{(i)}$ denote the entries of b. The marginal likelihood is defined to be the integral $\int_P L_U(\theta)\, d\theta$, where P is the parameter space $\Delta_{t_1} \times \Delta_{t_2} \times \cdots \times \Delta_{t_k}$. In fact, we can give an explicit formula for this integral. It is interesting that for the independence model, evaluating the marginal likelihood is easy, just as solving the maximum likelihood equations is easy [7, Prop 1.13].

Proposition 2.1. *The marginal likelihood for the model $\mathcal{M}$ and data U is*

$$\int_P L_U(\theta)\, d\theta \quad = \quad C \cdot \prod_{i=1}^{k} \frac{t_i!\, b_0^{(i)}!\, b_1^{(i)}! \, \cdots \, b_{t_i}^{(i)}!}{(s_i N + t_i)!}. \tag{6}$$

Proof. The main ingredients for this formula are the facts that

$$\int_{\Delta_t} \theta_0^{b_0}\theta_1^{b_1}\cdots\theta_t^{b_t}d\theta \quad = \quad \frac{t!\,b_0!\,b_1!\,\cdots\,b_t!}{(b_0+b_1+\cdots+b_t+t)!}$$

and that for $b = AU$, $b_0^{(i)} + b_1^{(i)} + \cdots + b_{t_i}^{(i)} = s_iN$. □

In general, A has many repeated columns, and in our applications, it is useful to consider the reduced matrix $\tilde{A}$ where all the repeated columns are removed. This reduced matrix has

$$\tilde{n} \quad = \quad \prod_{i=1}^{k}\binom{s_i+t_i}{s_i}$$

columns. Analogously, we have the reduced data vector $\tilde{U}$ which is derived from the original data vector U by summing entries corresponding to the same repeated column. Care must be taken when talking about the likelihood of the data, because the likelihood of the reduced data $\tilde{U}$ is different from that of the original data U. Indeed, let $\tilde{v}$ be an element of the reduced state space, and let $a_{\tilde{v}}$ be the corresponding column of $\tilde{A}$. Suppose $\alpha_{\tilde{v}}$ is the number of columns of A equal to $a_{\tilde{v}}$. Then, the probability of observing the state $\tilde{v}$ is

$$p_{\tilde{v}}(\theta) \quad = \quad \alpha_{\tilde{v}}\cdot\theta^{a_{\tilde{v}}}$$

Therefore, the likelihood of the reduced data $\tilde{U}$ is $L_{\tilde{U}}(\theta) = \tilde{C}\cdot\theta^b$, where $b = AU = \tilde{A}\tilde{U}$ and the normalizing constant is

$$\tilde{C} \quad = \quad N!\prod_{\tilde{v}}\alpha_{\tilde{v}}^{\tilde{U}_{\tilde{v}}}/\tilde{U}_{\tilde{v}}!$$

It follows that the marginal likelihood of the reduced data is also given by the formula (6) except with the constant C replaced with $\tilde{C}$.

We now turn our attention to mixtures of independence models. In statistics, a mixture model is a probability distribution that is a convex combination of given probability distributions. In this paper, we will study the two-mixture $\mathcal{M}^{(2)}$, although the same analysis extends to three or larger mixtures. The mixture model $\mathcal{M}^{(2)}$ has the same state space as $\mathcal{M}$, and its parameter space is the polytope $\Theta = \Delta_1\times P\times P$. Given some $(\sigma,\theta,\rho)\in\Delta_1\times P\times P$, the probability of observing the state $v\in V$ is

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

and the likelihood of data U is

$$L_U(\sigma,\theta,\rho) \quad = \quad C\cdot\prod_v(\sigma_0\theta^{a_v}+\sigma_1\rho^{a_v})^{U_v} \tag{7}$$

where C is the same normalizing constant as that in (5). Our objective is to evaluate exactly the marginal likelihood

$$\int_\Theta L_U(\sigma,\theta,\rho)\,d\sigma d\theta d\rho \quad = \quad C\int_\Theta\prod_v(\sigma_0\theta^{a_v}+\sigma_1\rho^{a_v})^{U_v} \tag{8}$$

It is meaningful to talk about evaluating this *exactly* because it is a rational number, which follows from expanding the integrand into a sum of monomials and noting that the integral of each monomial is rational. The formulas and techniques for doing so will be discussed in the next section. Meanwhile, we will point out that the likelihood and marginal likelihood for the reduced data $\tilde{U}$ is simply (7) and (8) except with the constant C replaced with $\tilde{C}$.

3 Exact Evaluation of the Integral

In this section, we will focus on computing the integral

$$\int_\Theta\prod_v(\sigma_0\theta^{a_v}+\sigma_1\rho^{a_v})^{U_v}. \tag{9}$$

We will assume that the matrix $A=(a_v)$ and data vector $U=(U_v)$ are both reduced because reducing them does not change the value of the integral. For simplicity, we will also denote the state space of all v by $[n]:=\{1,2,\ldots,n\}$.

Our first step is to expand the integrand in (9). Write

$$\prod_v(\sigma_0\theta^{a_v}+\sigma_1\rho^{a_v})^{U_v} \quad = \quad \sum_{\substack{b\in Z^{\mathbb{L}}_A(U)\\ c=AU-b}}\phi_A(b,U)\cdot\sigma_0^{|b|/a}\sigma_1^{|c|/a}\cdot\theta^b\cdot\rho^c \tag{10}$$

We will now explain the notations in the above formula. First, in a broad sense, $Z^{\mathbb{L}}_A(U)$ represents the set of all exponents b which appears as the coefficient of θ in the expansion. To be precise, we first consider the zonotope

$$Z_A(U) \quad = \quad \sum_{v=1}^n U_v\cdot[0,a_v]$$

where $[0, a_v]$ is the line segment between the origin and the point $a_v \in \mathbb{R}^d$ and the sum is the Minkowski sum of line segments. This zonotope is isomorphic to the Newton polytope of the integrand, i.e. the convex hull of the points given by the exponents of the terms in the polynomial. Now, let $\mathbb{L}$ be the lattice formed by the image of the linear transformation $A : \mathbb{Z}^n \to \mathbb{Z}^d$. Then, define $Z_A^{\mathbb{L}}(U) = Z_A(U) \cap \mathbb{L}$. It is not difficult to see that the exponents b appearing in the expansion are contained in this set. Second, the $\phi_A(b, U)$ represent the coefficients of the expansion. A simple calculation shows that

$$\phi_A(b, U) \;\; = \sum_{\substack{Ax=b \\ x \in \mathbf{D}(U)}} \prod_{v=1}^{n} \binom{U_v}{x_v}. \tag{11}$$

where $\mathbf{D}(U) = \{(x_1, \ldots, x_n) \in \mathbb{Z}^n \,:\, 0 \le x_v \le U_v\}$. Lastly, $|b|$ represents the L^1-norm $\sum_{i=1}^{d} b_i$ of b. Now, we may integrate (10) and apply (6) to get

Proposition 3.1. *The integral (9) is given by*

$$\sum_{\substack{b \in Z_A^{\mathbb{L}}(U) \\ c = AU - b}} \phi_A(b, U) \cdot \frac{(|b|/a)!\,(|c|/a)!}{(|U|+1)!} \cdot \prod_{i=1}^{k} \left(\frac{t_i!\, b_0^{(i)}! \cdots b_{t_i}^{(i)}!}{(|b^{(i)}| + t_i)!} \; \frac{t_i!\, c_0^{(i)}! \cdots c_{t_i}^{(i)}!}{(|c^{(i)}| + t_i)!} \right). \tag{12}$$

We will now discuss several techniques for computing (12) rapidly. Let us consider the storage of $\phi_A(b, U)$. The first important observation is that the coefficients $\phi_A(b, U)$ depend only on the exponent $b \in \mathbb{R}^d$ of θ and not on the exponents of σ and ρ, so they can be stored in a d-dimensional array. Furthermore, b lies in the image of the matrix A which is of rank $d_0 = d - k + 1$, so we may store the coefficients in an array of even smaller dimension. Thirdly, by replacing x_v with $U_v - x_v$ in (11), we see that $\phi_A(b, U) = \phi_A(AU - b, U)$. The means that we only need to store half the values of $\phi_A(b, U)$.

As for the computation of $\phi_A(b, U)$, one may use (11) but in practice, this is too slow because of the many binomial coefficients which have to be evaluated. Instead, by substituting $\sigma = \mathbf{1}$ and $\rho = \mathbf{1}$ into (10), we get

$$\prod_{v} (\theta^{a_v} + 1)^{U_v} \quad = \sum_{\substack{b \in Z_A^{\mathbb{L}}(U) \\ c = AU - b}} \phi_A(b, U) \cdot \theta^b.$$

This means that the $\phi_A(b, U)$ can be extracted from the expansion of the RHS via a powerful computer algebra software such as `Maple`. This is the easiest and quickest way, though it is often not memory efficient because the software does not exploit the storage strategies discussed previously. If we want to avoid this software method, we may use the recurrence formula

$$\phi_A(b,U) \;=\; \sum_{x_n=0}^{U_n} \binom{U_n}{x_n} \phi_{A\backslash a_n}(b - x_n a_n, U\backslash U_n). \tag{13}$$

which follows from inspecting (11). Together with the storage strategies above, this allows us to write the following dynamic programming algorithm which runs in $O(N^{d_0})$ space and $O(N^{d_0+1})$ time. In the algorithm, ϕ is a d_0-dimensional array and $\phi[b]$ denotes the entry which stores the value of $\phi_A(b,U)$. We leave out the implementation details for storing the $\phi_A(b,U)$ to avoid complicating the algorithm.

```
Algorithm 3.2 (RECURSIVE($A$, $U$)).
Input: The matrix $A$ and the vector $U$.
Output: The coefficients $\phi_A(b,U)$.
Step 1: Create a $d_0$-dimensional array $\phi$ of zeros.
Step 2: For each $x \in \{0,1,\ldots,U_1\}$, set
            $\phi[a_1 x] \;:=\; \binom{U_1}{x}.$
Step 3: Create a new $d_0$-dimensional array $\phi'$.
Step 4: For each $2 \le j \le n$, do
          1. Set all the entries of $\phi'$ to 0.
          2. For each $x \in \{0,1,\ldots,U_j\}$, do
               For each non-zero entry $\phi[b]$ in $\phi$ do
                  Increment $\phi'[b + a_j x]$ by $\binom{U_j}{x}\phi[b]$.
          3. Replace $\phi$ with $\phi'$.
Step 5: Output the array $\phi$.
```

We now turn our attention to evaluating the sum (12) after the coefficients $\phi_A(b,U)$ have been computed. Firstly, because of the many factorials appearing in the sum, a great way to speed things up is to precompute and store the values of the factorials. A second way to reduce the computation time is to find the common denominator of all the terms in the sum so that effectively, the summands are integers and not fractions. Thirdly, note that the summand corresponding to a given b is equal to the summand corresponding to $AU - b$. Thus, only half the terms need to be summed. Lastly, we discuss what to do if there is insufficient memory to store all the values of $\phi_A(b, U)$. We overcome this problem by writing the integrand as a product of smaller factors which can be expanded separately. In particular, we partition the columns of A into submatrices $A^{[1]}, \ldots, A^{[m]}$ and let $U^{[1]}, \ldots, U^{[m]}$ be the corresponding partition of U. Thus the integrand becomes

$$\prod_{j=1}^{m} \prod_{v} (\sigma_0 \theta^{a_v^{[j]}} + \sigma_1 \rho^{a_v^{[j]}})^{U_v^{[j]}},$$

where $a_v^{[j]}$ is the v-th column in the matrix $A^{[j]}$. The resulting algorithm for evaluating the integral is as follows:

Algorithm 3.3 (Fast Integral)**.**

```
Input: The matrices $A^{[1]}, \ldots, A^{[m]}$, vectors $U^{[1]}, \ldots, U^{[m]}$ and the vector $t$.
Output: The value of the integral (9) in exact rational arithmetic.
Step 1: For $1 \leq j \leq m$, compute $\phi^{[j]} := \mathrm{RECURSIVE}(A^{[j]}, U^{[j]})$.
Step 2: Set $I := 0$.
Step 3: For each non-zero entry $\phi^{[1]}[b^{[1]}]$ in $\phi^{[1]}$, do
            ⋮
          For each non-zero entry $\phi^{[m]}[b^{[m]}]$ in $\phi^{[m]}$, do
              Set $b := b^{[1]} + \cdots + b^{[m]}$, $c := AU - b$, $\phi := \prod_{j=1}^{m} \phi^{[j]}[b^{[j]}]$.
              Increment $I$ by
                  $\phi \cdot \frac{(|b|/a)!(|c|/a)!}{(|U|+1)!} \cdot \prod_{i=1}^{k} \frac{t_i!\, b_0^{(i)}! \cdots b_{t_i}^{(i)}!}{(|b^{(i)}| + t_i)!} \frac{t_i!\, c_0^{(i)}! \cdots c_{t_i}^{(i)}!}{(|c^{(i)}| + t_i)!}$.
Step 4: Output the sum $I$.
```

The space and time complexity of this algorithm is $O(N^S)$ and $O(N^T)$ respectively, where $S = \max_i \operatorname{rank} A^{[i]}$ and $T = 1 + \sum_i \operatorname{rank} A^{[i]}$. From this, we see that the splitting of the integrand should be chosen wisely to achieve a good pay-off between the two complexities.

The above computational techniques have been implemented in `Maple`. The library and documentation for its use are made available at

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

4 Approximations of the Integral

In this section, we will discuss some well-known approximations formulas of the marginal likelihood integral. We will not, however, be touching on Monte

Carlo or importance sampling methods. Our overview of this broad topic is far from being indepth or complete, so the reader is invited to refer to [1, 5] for further information.

The motivating question behind these approximations is this: given the data vector $U = NY$ where Y is a fixed vector satisfying $|Y| = \sum_v Y_v = 1$, what can we say about the asymptotic behavior of the marginal likelihood integral $\int_\Theta L_U(\theta)\, d\theta$ as $N \to \infty$? Typically, one studies the log-marginal likelihood $\log \int_\Theta L_U(\theta)\, d\theta \in [-\infty, 0]$ instead because of its sensitivity to small changes. Also, the normalizing constant C in the likelihood function $L_U(\theta)$ is often ignored because it is equal across all models for the same data vector and we are only interested in the relative values of the integral between the different models for the purpose of model selection. The approximations we present below are hold under certain conditions for exponential models, i.e. models where the log-likelihood function (ignoring the normalizing constant) for NY is equal to N times the log-likelihood function for the averaged statistics Y. Note that independence models and their mixtures are all examples of exponential models.

The first approximation is the *Bayesian Information Criterion* (BIC) which states that

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

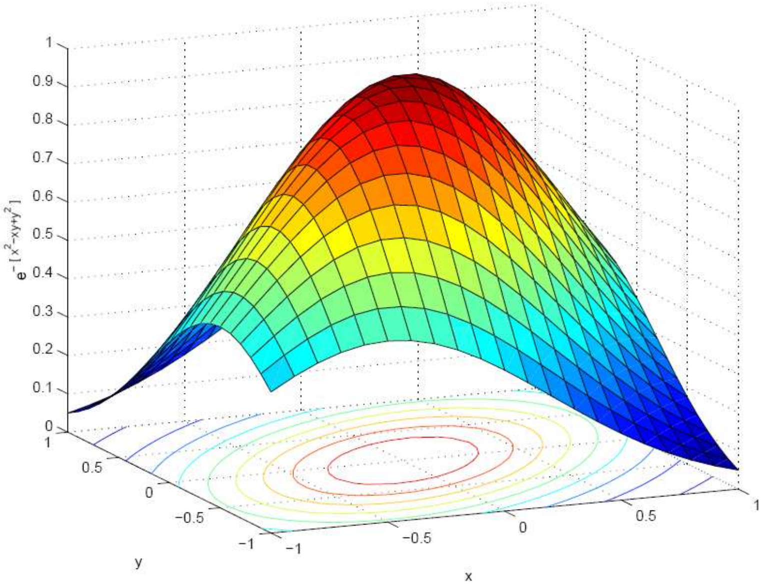
where $\hat{\theta}$ is the maximum likelihood estimate (MLE) and D is the dimension of the model. For the independence model $\mathcal{M}$ described in Section 2, the dimension is $D = d - k$ while for the mixture $\mathcal{M}^{(2)}$, it is $D = 2(d-k)+1$. It is assuring to see the maximum likelihood appearing in a formula for the marginal likelihood because it shows that these two different methods of model selection are actually closely related. Also, the correction term $\frac{D}{2} \log N$ may be thought of as a penalty for models with large dimensions, because given enough parameters, one can always design a model that produces the observed data vector with large probability.

The BIC is a relaxation of the *Laplace approximation*, which states that

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

where $H(\theta)$ is the Hessian matrix of the log-likelihood function $\log L_U(\theta)$. The relaxation to the BIC comes from approximating $\log |\det H(\hat{\theta})|$ with

$D \log N$ and ignoring the constant $\frac{D}{2}\log(2\pi)$. The conditions under which the Laplace approximation holds are as follows: fix the data vector U and consider the log-likelihood function $\log L_U(\theta)$ as a function of $\theta \in \Theta$. We require that $\log L_U$ be twice differentiable and convex, and that $\log L_U$ attains its maximum on a single internal point $\hat{\theta} \in \Theta$. The main idea behind this approximation is that as $N \to \infty$, a large bulk of the integrand $L_U(\theta)$ is concentrated near the unique maxima, as illustrated in the figure below [5].



For mixtures of independence models, problems occur when the model is non-identifiable, i.e. the mapping from the the parameter space Θ to the distribution space Δ_{n-1} is not one-to-one. An example of a non-identifiable model is the *100 Swiss Francs* example in [7, Ex 1.3]. In such cases, the likelihood function will have more than one MLE, so the BIC and Laplace approximations are not guaranteed to work. Furthermore, the set $\mathcal{V}$ of MLEs could be a variety of dimension larger than zero, rather than a finite set of points. The situation becomes worse when the variety $\mathcal{V}$ has singularities. Such problems occur frequently in areas such as machine learning. To overcome this problem, Watanabe developed algebraic geometric methods that relates the approximation of the marginal likelihood integral to the geometry of the singularities [8]. His approximation may be stated roughly as

$$\log \int_{\Theta} L_U(\theta)\, d\theta \quad = \quad \log L_U(\hat{\theta}) + \lambda \log N + (m-1)\log\log N + O(1),$$

where λ and m are the largest pole and its multiplicity of the meromorphic function that is analytically continued from

$$J(\lambda) = \int_{\Theta} (\log L_U(\theta) - \log L_U(\hat{\theta}))^{\lambda}\, d\theta, \quad \mathrm{Re}(\lambda) > 0.$$

Finding the largest pole and its multiplicity of $J(\lambda)$ is not easy, but some progress has been made using the technique of resolution of singularities from algebraic geometry [3, 5, 9]. This is a intense area of research.

5 Examples

In this section, we present three examples which illustrate the themes described in this paper. Our first example revisits the occasionally dishonest coin tosser story in Section 1. We compare some of the various approximations against the actual value of the marginal likelihood integral computed by our `Maple` code. Our second example is from [5] and it shows how model selection based on the BIC approximation chooses the wrong model, while Watanabe's approximation picks the right one. We confirm their result by producing the actual value of the marginal likelihood. Our last example demonstrates the scope of our methods by evaluating an integral from a data set in [4] previously thought to be infeasible.

5.1 Occasionally Dishonest Coin Tosser

In the story of the occasionally dishonest coin tosser, we may represent the outcomes of the four coin tosses when there is only one coin by binary random variables X_1, X_2, X_3 and X_4. This is the independence model $\mathcal{M}$ with $k = 1$, $s_1 = 4$, $t_1 = 1$. The state space $V = \{0, 1\}^4$ consists of binary strings of length four, with 1 and 0 representing heads and tails respectively. The 2×16 matrix A corresponding to the model is

$$A \;=\; \begin{array}{c} \\ \theta_0 \\ \theta_1 \end{array} \begin{array}{c} \begin{array}{ccccccccc} p_{0000} & p_{0001} & p_{0010} & p_{0100} & p_{1000} & p_{0011} & \cdots & p_{1110} & p_{1111} \end{array} \\ \left(\begin{array}{ccccccccc} 4 & 3 & 3 & 3 & 3 & 2 & \cdots & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 2 & \cdots & 3 & 4 \end{array} \right). \end{array}$$

After removing the repeated columns, the *reduced* matrix is the one shown in (3). The two coin scenario is the two-mixture model $\mathcal{M}^{(2)}$. We compute the marginal likelihood (4) of the independence model using our `Maple` library with the command

```
ML([4],[1],[51,18,73,25,75],unmixed=true);
```

while the marginal likelihood of the mixture model is computed with

```
ML([4],[1],[51,18,73,25,75]);
```

The numerical values of these two rational numbers are listed below.

$$\begin{array}{lrl} \text{a.} & \text{Independence} & 0.5773010420 \times 10^{-56} \\ \text{b.} & \text{Mixture} & 0.7788716339 \times 10^{-22} \end{array} \qquad (14)$$

We also compute the BIC and Laplace approximations for the mixture model. According to [6, Ex 9], the likelihood function $p_0^{51} p_1^{18} p_2^{73} p_3^{25} p_4^{75}$ has three local maxima $(\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ in the model $\mathcal{M}^{(2)}$, and these translate into six local maxima $(\hat{\sigma}, \hat{\theta}, \hat{\rho})$ in the parameter space P, which is the 3-cube. The two global maxima $(\hat{\sigma}, \hat{\theta}, \hat{\rho})$ in P are

$$(0.3367691969, 0.02877132371, 0.6536073424)$$
$$(0.6632308031, 0.02877132371, 0.6536073424).$$

Both of these points in P give the same point $(\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ in the model:

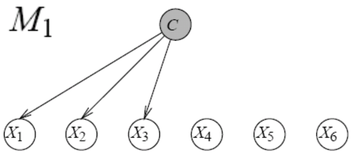
$$(0.12104, 0.25662, 0.20556, 0.10758, 0.30920).$$

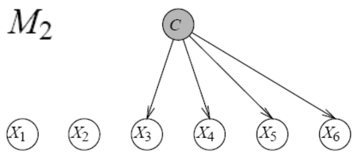
Thus, the maximum likelihood value is $0.1395471101 \times 10^{-18}$. The following table compares the various approximations. Here, "Actual" refers to the base-10 logarithm of the marginal likelihood (14b).

BIC	-22.43100220
Laplace	-22.39666281
Actual	-22.10853411

5.2 Incorrect Model Selection

We consider an experiment with six binary random variables $X_1, X_2, \ldots, X_6$, and represent the states in the experiment by strings $v \in \{0,1\}^6$. In [5], two graphical models M_1 and M_2 for these six variables were studied and they are shown in the diagrams below.





In each diagram, the node C is a hidden binary variable, and observed variables X_i linked to C by an edge participate in a two-mixture. The other observed variables are independent of the mixture and of each other. For instance, the probability of observing the state v in M_1 is given by

$$p_v = (\sigma_0 \theta^{(1)}_{v_1} \theta^{(2)}_{v_2} \theta^{(3)}_{v_3} + \sigma_1 \rho^{(1)}_{v_1} \rho^{(2)}_{v_2} \rho^{(3)}_{v_3}) \theta^{(4)}_{v_4} \theta^{(5)}_{v_5} \theta^{(6)}_{v_6} \tag{15}$$

for some parameters $\sigma, \theta^{(i)}, \rho^{(i)} \in \Delta_1$, while in M_2, it is given by

$$p_v = \theta^{(1)}_{v_1} \theta^{(2)}_{v_2} (\sigma_0 \theta^{(3)}_{v_3} \theta^{(4)}_{v_4} \theta^{(5)}_{v_5} \theta^{(6)}_{v_6} + \sigma_1 \rho^{(3)}_{v_3} \rho^{(4)}_{v_4} \rho^{(5)}_{v_5} \rho^{(6)}_{v_6}). \tag{16}$$

Now, suppose that we have the following data for a sample size of $N = 36$.

$$\begin{array}{ccc}
 & & X_4X_5X_6 \\
 & & \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{array} \\
X_1X_2X_3 & \begin{array}{c} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} & \begin{pmatrix}
2 & 3 & 0 & 1 & 3 & 5 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 4 & 1 & 1 & 2 & 3 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\end{array}$$

For example, the entry 4 in the row 100 and column 001 implies that the state $v = 100001$ was observed four times. This data was derived from averaged statistics Y designed so that the maximum likelihood values under M_1 and M_2 for Y are exactly the same.

Because the maximum likelihood for M_1 and M_2 under this data is the same and BIC penalizes models with higher dimensions, the BIC approximation for M_1 will be higher than that for M_2. Hence, model selection via the BIC score will prefer M_1 to M_2. However, Watanabe's approximations for the two models are

$$\begin{array}{ll} M_1: & \log L_U(\hat{\theta}) - 5 \log N + O(1) \\ M_2: & \log L_U(\hat{\theta}) - \frac{9}{2} \log N + O(1) \end{array}$$

so model selection via this approximation chooses M_2 over M_1. We compute the actual values of the marginal likelihood integrals for the above data. It can be seen from (15) that the marginal likelihood for M_1 is

$$C \cdot I_{123} \cdot I_4 \cdot I_5 \cdot I_6$$

where C is the normalizing constant $36!/\prod_v U_v!$,

$$\begin{aligned}
I_{123} &= \int_{\Delta_1^7} \prod_{v_1v_2v_3} (\sigma_0\theta^{(1)}_{v_1}\theta^{(2)}_{v_2}\theta^{(3)}_{v_3} + \sigma_1\rho^{(1)}_{v_1}\rho^{(2)}_{v_2}\rho^{(3)}_{v_3})^{U_{v_1v_2v_3***}} d\sigma d\theta d\rho \\
I_4 &= \int_{\Delta_1} \prod_{v_4} (\theta^{(4)}_{v_4})^{U_{***v_4**}} d\theta \\
I_5 &= \int_{\Delta_1} \prod_{v_5} (\theta^{(5)}_{v_5})^{U_{****v_5*}} d\theta \\
I_6 &= \int_{\Delta_1} \prod_{v_6} (\theta^{(6)}_{v_6})^{U_{*****v_6}} d\theta
\end{aligned}$$

and $U_{v_1v_2v_3***} = \sum_{v_4,v_5,v_6} U_{v_1v_2v_3v_4v_5v_6}$. The other exponents are similarly defined. All of the four integrals above are the marginal likelihoods of some independence or mixture model, up to a constant, so we may use our `Maple` library to evaluate them. The marginal likelihood for M_2 may be decomposed in a similar way. The table below summarises the results.

M_1	$\frac{2673620257358279100801924830063571461298286189}{595389791326672092336165244431090566358136576942917805560000000}$
M_2	$\frac{48293401975547884279365197096430603703508201757248809211637315169}{873248402971499818328286563178459524881596589864311287443444152295294483200000000}$

Their numerical values are 0.449×10^{-17} and 0.553×10^{-17} respectively. Hence, model selection via the actual values of the marginal likelihoods agree with that using Watanabe's approximation.

5.3 Schizophrenic Patients

We conclude by applying our method to a data set taken from the Bayesian statistics literature. Evans, Gilula and Guttman [4, §3] analyzed the association between length (in years Y) of hospital stay of 132 schizophrenic patients and the frequency with which they are visited by their relatives. Their data set is the following contingency table of format 3×3:

			$2 \leq Y < 10$	$10 \leq Y < 20$	$20 \leq Y$	*Totals*
		Visited regularly	43	16	3	*62*
U	$=$	Visited rarely	6	11	10	*27*
		Visited never	9	18	16	*43*
		Totals	*58*	*45*	*29*	**132**

They present estimated posterior means and variances for these data, where *"each estimate requires a 9-dimensional integration"* [4, p. 561]. Computing their integrals is essentially equivalent to ours, for $k = 2, s_1 = s_2 = 1, t_1 = t_2 = 2$ and $N = 132$. The authors emphasize that *"the dimensionality of the integral does present a problem"* [4, p. 562], and they point out that *"all posterior moments can be calculated in closed form .... however, even for modest N these expressions are far to complicated to be useful"* [4, p. 559].

We disagree with that conclusion. In our view, the closed form expressions in Section 3 are quite useful for modest sample size N. Using Algorithm 3.3, we computed the integral (9). It is the rational number with numerator

$$\begin{array}{l}
278019488531063389120643600324989329103876140805 \\
285242839582092569357265886675322845874097528033 \\
99493069713103633199906939405711180837568853737
\end{array}$$

and denominator

$$\begin{array}{l}
122884028735919354006780947965998487454428331775722 04 \\
504488199792864569951855421959468150731124291699978 01 \\
335039001699219121673522392041537866450291539511764 22 \\
432983280461634722619620284616504320243563397065411 32 \\
3437531847188027481866765742374912000000000000000000.
\end{array}$$

To obtain the marginal likelihood for the data U above, that rational number (of moderate size) still needs to be multiplied with the normalizing constant

$$\frac{132!}{43! \cdot 16! \cdot 3! \cdot 6! \cdot 11! \cdot 10! \cdot 9! \cdot 18! \cdot 16!}.$$

6 Acknowledgements

I would like to thank Lior Pachter and Yun S. Song for giving me the opportunity to work on this subject as a class project, and for their valuable suggestions for future work. I would also like to thank Bernd Sturmfels and Zhiqiang Xu who are collaborators on this project. Many of the formulations, ideas and results in this paper are the fruits of their hard work.

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] R. Durbin, S. Eddy, A. Korgh and G. Mitchison: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[3] M. Drton: Likelihood ratio tests and singularities, to appear in *Annals of Statistics*, `arXiv:math/0703360`.

[4] M. Evans, Z. Gilula and I. Guttman: Latent class analysis of two-way contingency tables by Bayesian methods, *Biometrika* **76** (1989) 557–563.

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

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

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

[8] S.Watanabe: Algebraic analysis for nonidentifiable learning machines, in *Neural Computation* **13**(4) (2001) 899–933

[9] 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.