Journal of Machine Learning Research 10 (2009) 1611-1631 Submitted 5/08; Revised 2/09; Published 7/09

Marginal Likelihood Integrals for Mixtures of Independence Models

Shaowei Lin SHAOWEI@MATH.BERKELEY.EDU
Bernd Sturmfels BERND@MATH.BERKELEY.EDU
Department of Mathematics
University of California
Berkeley, CA 94720, USA

Zhiqiang Xu XUZQ@LSEC.CC.AC.CN
LSEC, Academy of Mathematics and System Sciences
Chinese Academy of Sciences
Beijing, 100080, China

Editor: Tommi Jaakkola

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. Our methods apply to both uniform priors and Dirichlet priors. The underlying statistical models are mixtures of independent distributions, or, in geometric language, secant varieties of Segre-Veronese varieties.

Keywords: marginal likelihood, exact integration, mixture of independence model, computational algebra

1. Introduction

Evaluation of marginal likelihood integrals is central to Bayesian statistics. It is generally assumed that these integrals cannot be evaluated exactly, except in trivial cases, and a wide range of numerical techniques (e.g., MCMC) have been developed to obtain asymptotics and numerical approximations (Chickering and Heckerman, 1997). The aim of this paper is to show that exact integration is more feasible than is surmised in the literature. We examine marginal likelihood integrals for a class of mixture models for discrete data. Bayesian inference for these models arises in many contexts, including machine learning and computational biology. Recent work in these fields has made a connection to singularities in algebraic geometry (Drton, 2009; Geiger and Rusakov, 2005; Watanabe, 2001; Watanabe and Yamazaki, 2003, 2004). Our study augments these developments by providing tools for symbolic integration when the sample size is small.

The numerical value of the integral we have in mind is a rational number, and exact evaluation means computing that rational number rather than a floating point approximation. For a first example consider the integral

$$\int_{\Theta} \prod_{i,j \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \left(\pi \lambda_i^{(1)} \lambda_j^{(2)} + \tau \rho_i^{(1)} \rho_j^{(2)}\right)^{U_{ij}} d\pi\, d\tau\, d\lambda\, d\rho, \tag{1}$$

where Θ is the 13-dimensional polytope $\Delta_1 \times \Delta_3 \times \Delta_3 \times \Delta_3 \times \Delta_3$. The factors are probability simplices,

$$\begin{aligned}
\Delta_1 &= \{(\pi,\tau) \in \mathbb{R}^2_{\geq 0} : \pi+\tau = 1\}, \\
\Delta_3 &= \{(\lambda_{\mathtt{A}}^{(k)},\lambda_{\mathtt{C}}^{(k)},\lambda_{\mathtt{G}}^{(k)},\lambda_{\mathtt{T}}^{(k)}) \in \mathbb{R}^4_{\geq 0} \ : \textstyle\sum_i \lambda_i^{(k)} = 1\}, \quad k = 1,2, \\
\Delta_3 &= \{(\rho_{\mathtt{A}}^{(k)},\rho_{\mathtt{C}}^{(k)},\rho_{\mathtt{G}}^{(k)},\rho_{\mathtt{T}}^{(k)}) \in \mathbb{R}^4_{\geq 0} \ : \textstyle\sum_i \rho_i^{(k)} = 1\}, \quad k = 1,2.
\end{aligned}$$

and we integrate with respect to Lebesgue probability measure on Θ. If we take the exponents U_{ij} to be the entries of the particular contingency table

$$U = \begin{pmatrix} 4 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{pmatrix}, \tag{2}$$

then the exact value of the integral (1) is the rational number

$$\frac{571 \cdot 773426813 \cdot 17682039596993 \cdot 625015426432626533}{2^{31} \cdot 3^{20} \cdot 5^{12} \cdot 7^{11} \cdot 11^8 \cdot 13^7 \cdot 17^5 \cdot 19^5 \cdot 23^5 \cdot 29^3 \cdot 31^3 \cdot 37^3 \cdot 41^3 \cdot 43^2}. \tag{3}$$

The table (2) is taken from Example 1.3 of Pachter and Sturmfels (2005), where the integrand

$$\prod_{i,j \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \bigl(\pi \lambda_i^{(1)} \lambda_j^{(2)} + \tau \rho_i^{(1)} \rho_j^{(2)}\bigr)^{U_{ij}} \tag{4}$$

was studied using the EM algorithm, and the problem of validating its global maximum over Θ was raised. See Feinberg et al. (2007, §4.2) and Sturmfels (2008, §3) for further discussions. That optimization problem, which was widely known as the 100 *Swiss Francs problem*, has in the meantime been solved by Gao et al. (2008).

The main difficulty in performing computations such as (1) = (3) lies in the fact that the expansion of the integrand has many terms. A first naive upper bound on the number of monomials in the expansion of (4) would be

$$\prod_{i,j \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} (U_{ij}+1) \quad = \quad 3^{12} \cdot 5^4 \ = \ 332,150,625.$$

However, the true number of monomials is only $3,892,097$, and we obtain the rational number (3) by summing the values of the corresponding integrals

$$\int_\Theta \pi^{a_1} \tau^{a_2} (\lambda^{(1)})^u (\lambda^{(2)})^v (\rho^{(1)})^w (\rho^{(2)})^x d\pi\, d\tau\, d\lambda\, d\rho \quad =$$
$$\frac{a_1!\,a_2!}{(a_1+a_2+1)!} \cdot \frac{3! \prod_i u_i!}{(\sum_i u_i+3)!} \cdot \frac{3! \prod_i v_i!}{(\sum_i v_i+3)!} \cdot \frac{3! \prod_i w_i!}{(\sum_i w_i+3)!} \cdot \frac{3! \prod_i x_i!}{(\sum_i x_i+3)!}.$$

The geometric idea behind our approach is that the Newton polytope of (4) is a *zonotope* and we are summing over its lattice points. Definitions for these geometric objects are given in Section 3.

This paper is organized as follows. In Section 2 we describe the class of algebraic statistical models to which our method applies, and we specify the problem. In Section 3 we examine the Newton zonotopes of mixture models, and we derive formulas for marginal likelihood evaluation using tools from geometric combinatorics. Our algorithms and their implementations are described

in detail in Section 4. Section 5 is concerned with applications in Bayesian statistics. We show how *Dirichlet priors* can be incorporated into our approach, we discuss the evaluation of *Bayes factors*, we compare our setup with that of Chickering and Heckerman (1997), and we illustrate the scope of our methods by computing an integral arising from a data set of Evans et al. (1989).

A preliminary draft version of the present article was published as Section 5.2 of the Oberwolfach lecture notes (Drton et al., 2009). We refer to that volume for further information on the use of computational algebra in Bayesian statistics.

2. Independence Models and their Mixtures

We consider a collection of discrete random variables

$$\begin{array}{cccc} X_1^{(1)}, & X_2^{(1)}, & \ldots, & X_{s_1}^{(1)}, \\ X_1^{(2)}, & X_2^{(2)}, & \ldots, & X_{s_2}^{(2)}, \\ \vdots & \vdots & \ddots & \vdots \\ X_1^{(k)}, & X_2^{(k)}, & \ldots, & X_{s_k}^{(k)}, \end{array}$$

where $X_1^{(i)},\ldots,X_{s_i}^{(i)}$ are identically distributed with values in $\{0,1,\ldots,t_i\}$. The independence model $\mathcal{M}$ for these variables is a toric model (Pachter and Sturmfels, 2005, §1.2) represented by an integer $d \times n$-matrix A with

$$d = t_1 + t_2 + \cdots + t_k + k \quad \text{and} \quad n = \prod_{i=1}^{k} (t_i+1)^{s_i}. \tag{5}$$

The columns of the matrix A are indexed by elements v of the state space

$$\{0,1,\ldots,t_1\}^{s_1} \times \{0,1,\ldots,t_2\}^{s_2} \times \cdots \times \{0,1,\ldots,t_k\}^{s_k}. \tag{6}$$

The rows of the matrix A are indexed by the model parameters, which are the d coordinates of the points $\theta = (\theta^{(1)},\theta^{(2)},\ldots,\theta^{(k)})$ in the polytope

$$P \;=\; \Delta_{t_1} \times \Delta_{t_2} \times \cdots \times \Delta_{t_k}, \tag{7}$$

and the model $\mathcal{M}$ is the subset of the simplex Δ_{n-1} given parametrically by

$$p_v \quad = \quad \text{Prob}\big(X_j^{(i)} = v_j^{(i)} \text{ for all } i,j\big) \quad = \quad \prod_{i=1}^{k}\prod_{j=1}^{s_i} \theta_{v_j^{(i)}}^{(i)}. \tag{8}$$

This is a monomial in d unknowns. The matrix A is defined by taking its column a_v to be the exponent vector of this monomial.

In algebraic geometry, the model $\mathcal{M}$ is known as *Segre-Veronese variety*

$$\mathbb{P}^{t_1} \times \mathbb{P}^{t_2} \times \cdots \times \mathbb{P}^{t_k} \quad \hookrightarrow \quad \mathbb{P}^{n-1}, \tag{9}$$

where the embedding is given by the line bundle $O(s_1,s_2,\ldots,s_k)$. The manifold $\mathcal{M}$ is the toric variety of the polytope P. Both objects have dimension $d-k$, and they are identified with each other via the moment map (Fulton, 1993, §4).

Example 1 *Consider three binary random variables where the last two random variables are identically distributed. In our notation, this corresponds to* $k = 2$, $s_1 = 1$, $s_2 = 2$ *and* $t_1 = t_2 = 1$. *We find that* $d = 4, n = 8$, *and*

$$A \quad = \quad \begin{array}{c} \\ \theta_0^{(1)} \\ \theta_1^{(1)} \\ \theta_0^{(2)} \\ \theta_1^{(2)} \end{array} \begin{array}{c} \begin{array}{cccccccc} p_{000} & p_{001} & p_{010} & p_{011} & p_{100} & p_{101} & p_{110} & p_{111} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 0 & 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 & 0 & 1 & 1 & 2 \end{pmatrix} \end{array}.$$

The columns of this matrix represent the monomials in the parametrization (8). The model $\mathcal{M}$ *lies in the* 5*-dimensional subsimplex of* Δ_7 *given by* $p_{001} = p_{010}$ *and* $p_{101} = p_{110}$, *and it consists of all rank one matrices*

$$\begin{pmatrix} p_{000} & p_{001} & p_{100} & p_{101} \\ p_{010} & p_{011} & p_{110} & p_{111} \end{pmatrix}.$$

In algebraic geometry, the surface $\mathcal{M}$ *is called a* rational normal scroll.

The matrix A has repeated columns whenever $s_i \geq 2$ for some i. It is sometimes convenient to represent the model $\mathcal{M}$ by the matrix $\tilde{A}$ which is obtained from A by removing repeated columns. We label the columns of $\tilde{A}$ by elements $v = (v^{(1)}, \ldots, v^{(k)})$ of (6) whose components $v^{(i)} \in \{0, 1, \ldots, t_i\}^{s_i}$ are weakly increasing. Hence $\tilde{A}$ is a $d \times \tilde{n}$-matrix with

$$\tilde{n} \quad = \quad \prod_{i=1}^{k} \binom{s_i + t_i}{s_i}. \tag{10}$$

The model $\mathcal{M}$ and its mixtures are subsets of a subsimplex $\Delta_{\tilde{n}-1}$ of Δ_{n-1}.

We now introduce *marginal likelihood integrals*. All our domains of integration in this paper are polytopes that are products of standard probability simplices. On each such polytope we fix the standard Lebesgue probability measure. In other words, our discussion of Bayesian inference refers to the uniform prior on each parameter space. Naturally, other prior distributions, such as Dirichlet priors, are of interest, and our methods are extended to these in Section 5. In what follows, we simply work with uniform priors.

We identify the state space (6) with the set $\{1, \ldots, n\}$. A *data vector* $U = (U_1, \ldots, U_n)$ is thus an element of $\mathbb{N}^n$. The *sample size* of these data is $U_1 + U_2 + \cdots + U_n = N$. If the sample size N is fixed then the probability of observing these data is

$$\mathbf{L}_U(\theta) \quad = \quad \frac{N!}{U_1! U_2! \cdots U_n!} \cdot p_1(\theta)^{U_1} \cdot p_2(\theta)^{U_2} \cdot \cdots \cdot p_n(\theta)^{U_n}.$$

This expression is a function on the polytope P which is known as the *likelihood function* of the data U with respect to the independence model $\mathcal{M}$. The *marginal likelihood* of the data U with respect to the model $\mathcal{M}$ equals

$$\int_P \mathbf{L}_U(\theta)\, d\theta.$$

The value of this integral is a rational number which we now compute explicitly. The data U will enter this calculation by way of the *sufficient statistic* $b = A \cdot U$, which is a vector in $\mathbb{N}^d$. The

coordinates of this vector are denoted $b_j^{(i)}$ for $i = 1,\ldots,k$ and $j = 0,\ldots,t_k$. Thus $b_j^{(i)}$ is the total number of times the value j is attained by one of the random variables $X_1^{(i)},\ldots,X_{s_i}^{(i)}$ in the i-th group. Clearly, the sufficient statistics satisfy

$$b_0^{(i)} + b_1^{(i)} + \cdots + b_{t_i}^{(i)} \;=\; s_i \cdot N \qquad \text{for all } i = 1,2,\ldots,k. \tag{11}$$

The likelihood function $\mathbf{L}_U(\theta)$ is the constant $\frac{N!}{U_1!\cdots U_n!}$ times the monomial

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

The logarithm of this function is concave on the polytope P, and its maximum value is attained at the point $\hat{\theta}$ with coordinates $\hat{\theta}_j^{(i)} = b_j^{(i)}/(s_i \cdot N)$.

Lemma 1 *The integral of the monomial θ^b over the polytope P equals*

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

The product of this number with the multinomial coefficient $N!/(U_1!\cdots U_n!)$ equals the marginal likelihood of the data U for the independence model $\mathcal{M}$.

Proof Since P is the product of simplices (7), this follows from the formula

$$\int_{\Delta_t} \theta_0^{b_0}\theta_1^{b_1}\cdots\theta_t^{b_t} d\theta \quad = \quad \frac{t!\cdot b_0!\cdot b_1!\cdots b_t!}{(b_0+b_1+\cdots+b_t+t)!} \tag{12}$$

for the integral of a monomial over the standard probability simplex Δ_t. ■

Our objective is to compute marginal likelihood integrals for the mixture model $\mathcal{M}^{(2)}$. The natural parameter space of this model is the polytope

$$\Theta \;=\; \Delta_1 \times P \times P.$$

Let $a_v \in \mathbb{N}^d$ be the column vector of A indexed by the state v, which is either in (6) or in $\{1,2,\ldots,n\}$. The parametrization (8) can be written simply as $p_v = \theta^{a_v}$. The mixture model $\mathcal{M}^{(2)}$ is defined to be the subset of Δ_{n-1} with the parametric representation

$$p_v \;=\; \sigma_0 \cdot \theta^{a_v} + \sigma_1 \cdot \rho^{a_v} \qquad \text{for } (\sigma,\theta,\rho) \in \Theta. \tag{13}$$

The likelihood function of a data vector $U \in \mathbb{N}^n$ for the model $\mathcal{M}^{(2)}$ equals

$$\mathbf{L}_U(\sigma,\theta,\rho) \quad = \quad \frac{N!}{U_1!U_2!\cdots U_n!}\, p_1(\sigma,\theta,\rho)^{U_1}\cdots p_n(\sigma,\theta,\rho)^{U_n}. \tag{14}$$

The *marginal likelihood* of the data U with respect to the model $\mathcal{M}^{(2)}$ equals

$$\int_\Theta \mathbf{L}_U(\sigma,\theta,\rho)\, d\sigma d\theta d\rho \;=\; \frac{N!}{U_1!\cdots U_n!}\int_\Theta \prod_v (\sigma_0\theta^{a_v} + \sigma_1\rho^{a_v})^{U_v} d\sigma\, d\theta\, d\rho. \tag{15}$$

The following proposition shows that we can evaluate this integral *exactly*.

Proposition 2 *The marginal likelihood (15) is a rational number.*

Proof The likelihood function $\mathbf{L}_U$ is a $\mathbb{Q}_{\geq 0}$-linear combination of monomials $\sigma^a\theta^b\rho^c$. The integral (15) is the same $\mathbb{Q}_{\geq 0}$-linear combination of the numbers

$$\int_\Theta \sigma^a\theta^b\rho^c d\sigma\, d\theta\, d\rho \quad = \quad \Big(\int_{\Delta_1}\sigma^a d\sigma\Big)\cdot\Big(\int_P \theta^b d\theta\Big)\cdot\Big(\int_P \rho^c d\rho\Big).$$

Each of the three factors is an easy-to-evaluate rational number, by (12). ■

Example 2 *The integral (1) expresses the marginal likelihood of a* 4×4*-table of counts* $U=(U_{ij})$ *with respect to the mixture model* $\mathcal{M}^{(2)}$*. Specifically, the marginal likelihood of the data (2) equals the normalizing constant* $40!\cdot(2!)^{-12}\cdot(4!)^{-4}$ *times the number (3). The model* $\mathcal{M}^{(2)}$ *consists of all non-negative* 4×4*-matrices of rank* ≤ 2 *whose entries sum to one. Here the parametrization (13) is not identifiable because* $\dim(\mathcal{M}^{(2)})=11$ *but* $\dim(\Theta)=13$*. In this example,* $k=2$*,* s_1=s_2=1*,* t_1=t_2=3*,* $d=8$*,* $n=16$*.*

In algebraic geometry, the model $\mathcal{M}^{(2)}$ is known as the first secant variety of the Segre-Veronese variety (9). We could also consider the higher secant varieties $\mathcal{M}^{(l)}$, which correspond to mixtures of l independent distributions, and much of our analysis can be extended to that case, but for simplicity we restrict ourselves to $l=2$. The variety $\mathcal{M}^{(2)}$ is embedded in the projective space $\mathbb{P}^{\tilde{n}-1}$ with $\tilde{n}$ as in (10). Note that $\tilde{n}$ can be much smaller than n. If this is the case, it is convenient to aggregate states whose probabilities are identical and represent the data by a vector $\tilde{U}\in\mathbb{N}^{\tilde{n}}$. Here is an example.

Example 3 *Let* k=1*,* s_1=4 *and* t_1=1*, so* $\mathcal{M}$ *is the independence model for four identically distributed binary random variables. Then* $d=2$ *and* $n=16$*. The corresponding integer matrix and its row and column labels are*

$$A \;=\; \begin{array}{c} \\ \theta_0 \\ \theta_1 \end{array}\!\!\begin{array}{c} \begin{array}{cccccccccc} p_{0000} & p_{0001} & p_{0010} & p_{0100} & p_{1000} & p_{0011} & \cdots & p_{1110} & p_{1111} \end{array} \\ \begin{pmatrix} 4 & 3 & 3 & 3 & 3 & 2 & \cdots & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 2 & \cdots & 3 & 4 \end{pmatrix} \end{array}.$$

However, this matrix has only $\tilde{n}=5$ *distinct columns, and we instead use*

$$\tilde{A} \;=\; \begin{array}{c} \\ \theta_0 \\ \theta_1 \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} p_0 & p_1 & p_2 & p_3 & p_4 \end{array} \\ \begin{pmatrix} 4 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \end{array}.$$

The mixture model $\mathcal{M}^{(2)}$ *is the subset of* Δ_4 *given by the parametrization*

$$p_i \quad = \quad \binom{4}{i}\cdot\big(\sigma_0\cdot\theta_0^{4-i}\cdot\theta_1^i \,+\, \sigma_1\cdot\rho_0^{4-i}\cdot\rho_1^i\big) \qquad \text{for } i=0,1,2,3,4.$$

In algebraic geometry, this threefold is the secant variety of the rational normal curve in $\mathbb{P}^4$*. This is the cubic hypersurface with the implicit equation*

$$\det\begin{bmatrix} 12p_0 & 3p_1 & 2p_2 \\ 3p_1 & 2p_2 & 3p_3 \\ 2p_2 & 3p_3 & 12p_4 \end{bmatrix} \quad = \quad 0.$$

In Hoşten et al. (2005, Example 9), the likelihood function (14) was studied for the data

$$\tilde{U} \quad = \quad (\tilde{U}_0, \tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{U}_4) \quad = \quad (51, 18, 73, 25, 75).$$

It has three local maxima (modulo swapping θ *and* ρ*) whose coordinates are algebraic numbers of degree* 12*. Using the methods to be described in the next two sections, we computed the exact value of the marginal likelihood for the data* $\tilde{U}$ *with respect to* $\mathcal{M}^{(2)}$*. The rational number (15) is found to be the ratio of two relatively prime integers having* 530 *digits and* 552 *digits, and its numerical value is approximately* $0.77887163388386786113357 42 \cdot 10^{-22}$.

3. Summation over a Zonotope

Our starting point is the observation that the Newton polytope of the likelihood function (14) is a zonotope. Recall that the *Newton polytope* of a polynomial is the convex hull of all exponent vectors appearing in the expansion of that polynomial, and a polytope is a *zonotope* if it is the image of a standard cube under a linear map. See Cox et al. (2005, §7) and Ziegler (1995, §7) for further discussions. We are here considering the zonotope

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

where $[0, a_v]$ represents the line segment between the origin and the point $a_v \in \mathbb{R}^d$, and the sum is a Minkowski sum of line segments. We write $Z_A = Z_A(1, 1, \ldots, 1)$ for the basic zonotope spanned by the vectors a_v. Hence $Z_A(U)$ is obtained by stretching Z_A along those vectors by factors U_v respectively. Assuming that the counts U_v are all positive, we have

$$\dim(Z_A(U)) \;=\; \dim(Z_A) \;=\; \operatorname{rank}(A) \;=\; d-k+1. \tag{16}$$

The zonotope Z_A is related to the polytope $P = \operatorname{conv}(A)$ in (7) as follows. The dimension $d-k = t_1 + \cdots + t_k$ of P is one less than $\dim(Z_A)$, and P appears as the *vertex figure* of the zonotope Z_A at the distinguished vertex 0.

Remark 3 *For higher mixtures* $\mathcal{M}^{(l)}$*, the Newton polytope of the likelihood function is isomorphic to the Minkowski sum of* $(l-1)$*-dimensional simplices in* $\mathbb{R}^{(l-1)d}$*. Only when* $l = 2$*, this Minkowski sum is a zonotope.*

The marginal likelihood (15) we wish to compute is the integral

$$\int_{\Theta} \prod_{v=1}^{n} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} d\sigma d\theta d\rho \tag{17}$$

times the constant $N!/(U_1! \cdots U_n!)$. Our approach to this computation is to sum over the lattice points in the zonotope $Z_A(U)$. If the matrix A has repeated columns, we may replace A with the reduced matrix $\tilde{A}$ and U with the corresponding reduced data vector $\tilde{U}$. If one desires the marginal likelihood for the reduced data vector $\tilde{U}$ instead of the original data vector U, the integral remains the same while the normalizing constant becomes

$$\frac{N!}{\tilde{U}_1! \cdots \tilde{U}_{\tilde{n}}!} \cdot \alpha_1^{\tilde{U}_1} \cdots \alpha_{\tilde{n}}^{\tilde{U}_{\tilde{n}}},$$

where α_i is the number of columns in A equal to the i-th column of $\tilde{A}$. In what follows we ignore the normalizing constant and focus on computing the integral (17) with respect to the original matrix A.

For a vector $b \in \mathbb{R}^d_{\geq 0}$ we let $|b|$ denote its L^1-norm $\sum_{t=1}^d b_t$. Recall from (8) that all columns of the $d \times n$-matrix A have the same coordinate sum

$$a \ := \ |a_v| \ = \ s_1 + s_2 + \cdots + s_k, \ \text{ for all } v = 1,2,\ldots,n,$$

and from (11) that we may denote the entries of a vector $b \in \mathbb{R}^d$ by $b_j^{(i)}$ for $i = 1,\ldots,k$ and $j = 0,\ldots,t_k$. Also, let $\mathbb{L}$ denote the image of the linear map $A : \mathbb{Z}^n \to \mathbb{Z}^d$. Thus $\mathbb{L}$ is a sublattice of rank $d-k+1$ in $\mathbb{Z}^d$. We abbreviate $Z_A^{\mathbb{L}}(U) := Z_A(U) \cap \mathbb{L}$. Now, using the binomial theorem, we have

$$(\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \ = \ \sum_{x_v=0}^{U_v} \binom{U_v}{x_v} \sigma_0^{x_v} \sigma_1^{U_v - x_v} \theta^{x_v \cdot a_v} \rho^{(U_v - x_v)\cdot a_v}.$$

Therefore, in the expansion of the integrand in (17), the exponents of θ are of the form of $b = \sum_v x_v a_v \in Z_A^{\mathbb{L}}(U)$, $0 \leq x_v \leq U_v$. The other exponents may be expressed in terms of b. This gives us

$$\prod_{v=1}^n (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \ = \ \sum_{\substack{b \in Z_A^{\mathbb{L}}(U) \\ c = AU - b}} \phi_A(b,U) \cdot \sigma_0^{|b|/a} \cdot \sigma_1^{|c|/a} \cdot \theta^b \cdot \rho^c. \tag{18}$$

Writing $\mathbf{D}(U) = \{(x_1,\ldots,x_n) \in \mathbb{Z}^n : 0 \leq x_v \leq U_v, v = 1,\ldots,n\}$, the coefficient in (18) equals

$$\phi_A(b,U) \ = \ \sum_{\substack{Ax = b \\ x \in \mathbf{D}(U)}} \prod_{v=1}^n \binom{U_v}{x_v}. \tag{19}$$

Thus, by formulas (12) and (18), the integral (17) evaluates to

$$\sum_{\substack{b \in Z_A^{\mathbb{L}}(U) \\ c = AU - b}} \phi_A(b,U) \cdot \frac{(|b|/a)!\,(|c|/a)!}{(|U|+1)!} \cdot \prod_{i=1}^k \left(\frac{t_i!\, b_0^{(i)}! \cdots b_{t_i}^{(i)}!}{(|b^{(i)}|+t_i)!} \ \frac{t_i!\, c_0^{(i)}! \cdots c_{t_i}^{(i)}!}{(|c^{(i)}|+t_i)!} \right). \tag{20}$$

We summarize the result of this derivation in the following theorem.

Theorem 4 *The marginal likelihood of the data U in the mixture model $\mathcal{M}^{(2)}$ is equal to the sum (20) times the normalizing constant $N!/(U_1!\cdots U_n!)$.*

Each individual summand in the formula (20) is a ratio of factorials and hence can be evaluated symbolically. The challenge in turning Theorem 4 into a practical algorithm lies in the fact that both of the sums (19) and (20) are over very large sets. We shall discuss these challenges and present techniques from both computer science and mathematics for addressing them.

We first turn our attention to the coefficients $\phi_A(b,U)$ of the expansion (18). These quantities are written as an explicit sum in (19). The first useful observation is that these coefficients are also the coefficients of the expansion

$$\prod_v (\theta^{a_v} + 1)^{U_v} \quad = \ \sum_{b \in Z_A^{\mathbb{L}}(U)} \phi_A(b,U) \cdot \theta^b, \tag{21}$$

which comes from substituting $\sigma_i = 1$ and $\rho_j = 1$ in (18). When the cardinality of $Z_A^{\mathbb{L}}(U)$ is sufficiently small, the quantity $\phi_A(b,U)$ can be computed quickly by expanding (21) using a computer algebra system. We used MAPLE for this and all other symbolic computations in this project.

If the expansion (21) is not feasible, then it is tempting to compute the individual $\phi_A(b,U)$ via the sum-product formula (19). This method requires summation over the set $\{x \in \mathbf{D}(U) : Ax = b\}$, which is the set of lattice points in an $(n-d+k-1)$-dimensional polytope. Even if this loop can be implemented, performing the sum in (19) symbolically requires the evaluation of many large binomials, causing the process to be rather inefficient.

An alternative is offered by the following recurrence formula:

$$\phi_A(b,U) \;=\; \sum_{x_n=0}^{U_n} \binom{U_n}{x_n} \phi_{A\setminus a_n}(b - x_n a_n, U \setminus U_n). \tag{22}$$

This is equivalent to writing the integrand in (17) as

$$\left(\prod_{v=1}^{n-1} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \right) (\sigma_0 \theta^{a_n} + \sigma_1 \rho^{a_n})^{U_n}.$$

More generally, for each $0 < i < n$, we have the recurrence

$$\phi_A(b,U) = \sum_{b' \in Z_{A'}^{\mathbb{L}}(U')} \phi_{A'}(b',U') \cdot \phi_{A\setminus A'}(b-b', U \setminus U'),$$

where A' and U' consist of the first i columns and entries of A and U respectively. This corresponds to the factorization

$$\left(\prod_{v=1}^{i} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \right) \left(\prod_{v=i+1}^{n} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \right).$$

This formula gives flexibility in designing algorithms with different payoffs in time and space complexity, to be discussed in Section 4.

The next result records useful facts about the quantities $\phi_A(b,U)$.

Proposition 5 *Suppose $b \in \mathbb{Z}_A^{\mathbb{L}}(U)$ and $c = AU - b$. Then, the following quantities are all equal to $\phi_A(b,U)$:*
(1) $\#\{z \in \{0,1\}^N : A^U z = b\}$, where A^U is the extended matrix

$$A^U \;:=\; (\underbrace{a_1,\ldots,a_1}_{U_1}, \underbrace{a_2,\ldots,a_2}_{U_2}, \ldots, \underbrace{a_n,\ldots,a_n}_{U_n}),$$

(2) $\phi_A(c,U)$,
(3)

$$\sum_{\substack{Ax=b \\ l_j \le x_j \le u_j}} \prod_{v=1}^{n} \binom{U_v}{x_v},$$

where $u_j = \min\{U_j\} \cup \{b_m/a_{jm}\}_{m=1}^n$ and $l_j = U_j - \min\{U_j\} \cup \{c_m/a_{jm}\}_{m=1}^n$.

Proof (1) This follows directly from (21).
(2) For each $z \in \{0,1\}^N$ satisfying $A^U z = b$, note that $\overline{z} = (1,1,\ldots,1) - z$ satisfies $A^U \overline{z} = c$, and vice versa. The conclusion thus follows from (1).
(3) We require $Ax = b$ and $x \in \mathbf{D}(U)$. If $x_j > u_j = b_m/a_{jm}$ then $a_{jm}x_j > b_m$, which implies $Ax \neq b$. The lower bound is derived by a similar argument. ■

One aspect of our approach is the decision, for any given model A and data set U, whether or not to attempt the expansion (21) using computer algebra. This decision depends on the cardinality of the set $Z_A^{\mathbb{L}}(U)$. In what follows, we compute the number exactly when A is unimodular. When A is not unimodular, we obtain useful lower and upper bounds.

Let S be any subset of the columns of A. We call S *independent* if its elements are linearly independent in $\mathbb{R}^d$. With S we associate the integer

$$\text{index}(S) \;:=\; [\mathbb{R}S \cap \mathbb{L} : \mathbb{Z}S].$$

This is the index of the abelian group generated by S inside the possibly larger abelian group of all lattice points in $\mathbb{L} = \mathbb{Z}A$ that lie in the span of S. The following formula is due to R. Stanley and appears in Stanley (1991, Theorem 2.2):

Proposition 6 *The number of lattice points in the zonotope $Z_A(U)$ equals*

$$\#Z_A^{\mathbb{L}}(U) \;=\; \sum_{S \subseteq A \text{ indep.}} \text{index}(S) \cdot \prod_{a_v \in S} U_v. \tag{23}$$

In fact, the number of monomials in (18) equals $\#M_A(U)$, where $M_A(U)$ is the set $\{b \in Z_A^{\mathbb{L}}(U) : \phi_A(b,U) \neq 0\}$, and this set can be different from $Z_A^{\mathbb{L}}(U)$. For that number we have the following bounds. The proof, which uses the methods in Stanley (1991, §2), will be omitted here.

Theorem 7 *The number $\#M_A(U)$ of monomials in the expansion (18) of the likelihood function to be integrated satisfies the two inequalities*

$$\sum_{S \subseteq A \text{ indep.}} \prod_{v \in S} U_v \;\leq\; \#M_A(U) \;\leq\; \sum_{S \subseteq A \text{ indep.}} \text{index}(S) \cdot \prod_{v \in S} U_v. \tag{24}$$

By definition, the matrix A is *unimodular* if $\text{index}(S) = 1$ for all independent subsets S of the columns of A. In this case, the upper bound coincides with the lower bound, and so $M_A(U) = Z_A^{\mathbb{L}}(U)$. This happens in the classical case of two-dimensional contingency tables ($k = 2$ and $s_1 = s_2 = 1$). In general, $\#Z_A^{\mathbb{L}}(U)/\#M_A(U)$ tends to 1 when all coordinates of U tend to infinity. This is why we believe that for computational purposes, $\#Z_A^{\mathbb{L}}(U)$ is a good approximation of $\#M_A(U)$.

Remark 8 *There exist integer matrices A for which $\#M_A(U)$ does not agree with the upper bound in Theorem 7. However, we conjecture that $\#M_A(U) = \#Z_A^{\mathbb{L}}(U)$ holds for matrices A of Segre-Veronese type as in (8) and strictly positive data vectors U.*

Example 4 *Consider the* 100 Swiss Francs *example in Section 1. Here A is unimodular and it has* 16145 *independent subsets S. The corresponding sum of* 16145 *squarefree monomials in (23) gives the number of terms in the expansion of (4). For the data U in (2) this sum evaluates to* $3,892,097$.

Example 5 *We consider the matrix and data from Example 3.*

$$\begin{aligned} \tilde{A} &= \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix} \\ \tilde{U} &= (51, 18, 73, 25, 75) \end{aligned}$$

By Theorem 7, the lower bound is 22,273 and the upper bound is 48,646. Here the number $\#M_{\tilde{A}}(\tilde{U})$ *of monomials agrees with the latter.*

We next present a formula for $\text{index}(S)$ when S is any linearly independent subset of the columns of the matrix A. After relabeling we may assume that $S = \{a_1, \ldots, a_k\}$ consists of the first k columns of A. Let $H = VA$ denote the row Hermite normal form of A. Here $V \in SL_d(\mathbb{Z})$ and H satisfies

$$H_{ij} = 0 \text{ for } i > j \text{ and } 0 \leq H_{ij} < H_{jj} \text{ for } i < j.$$

Hermite normal form is a built-in function in computer algebra systems. For instance, in MAPLE the command is `ihermite`. Using the invertible matrix V, we may replace A with H, so that $\mathbb{R}S$ becomes $\mathbb{R}^k$ and $\mathbb{Z}S$ is the image over $\mathbb{Z}$ of the upper left $k \times k$-submatrix of H. We seek the index of that lattice in the possibly larger lattice $\mathbb{Z}A \cap \mathbb{Z}^k$. To this end we compute the column Hermite normal form $H' = HV'$. Here $V' \in SL_n(\mathbb{Z})$ and H' satisfies

$$H'_{ij} = 0 \text{ if } i > j \text{ or } j > d \text{ and } 0 \leq H_{ij} < H_{ii} \text{ for } i < j.$$

The lattice $\mathbb{Z}A \cap \mathbb{Z}^k$ is spanned by the first k columns of H', and this implies

$$\text{index}(S) \quad = \quad \frac{H_{11} H_{22} \cdots H_{kk}}{H'_{11} H'_{22} \cdots H'_{kk}}.$$

4. Algorithms

In this section we discuss algorithms for computing the integral (17) exactly, and we discuss their advantages and limitations. In particular, we examine four main techniques which represent the formulas (20), (21), (16) and (22) respectively. The practical performance of the various algorithms is compared by computing the integral in Example 3.

A MAPLE library which implements our algorithms is made available at

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

The input for our MAPLE code consists of parameter vectors $s = (s_1, \ldots, s_k)$ and $t = (t_1, \ldots, t_k)$ as well as a data vector $U \in \mathbb{N}^n$. This input uniquely specifies the $d \times n$-matrix A. Here d and n are as in (5). The output features the matrices A and $\tilde{A}$, the marginal likelihood integrals for $\mathcal{M}$ and $\mathcal{M}^{(2)}$, as well as the bounds in (24).

We tacitly assume that A has been replaced with the reduced matrix $\tilde{A}$. Thus from now on we assume that A has no repeated columns. This requires some care concerning the normalizing constants. All columns of the matrix A have the same coordinate sum a, and the convex hull of the columns is the polytope $P = \Delta_{t_1} \times \Delta_{t_2} \times \cdots \times \Delta_{t_k}$. Our domain of integration is the following polytope of dimension $2d - 2k + 1$:

$$\Theta \;=\; \Delta_1 \times P \times P.$$

We seek to compute the rational number

$$\int_{\Theta} \prod_{v=1}^{n} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} d\sigma d\theta d\rho, \tag{25}$$

where integration is with respect to Lebesgue probability measure. Our MAPLE code outputs this integral multiplied with the statistically correct normalizing constant. That constant will be ignored in what follows. In our complexity analysis, we fix A while allowing the data U to vary. The complexities will be given in terms of the sample size $N = U_1 + \cdots + U_n$.

4.1 Ignorance is Costly

Given an integration problem such as (25), a first attempt is to use the symbolic integration capabilities of a computer algebra package such as MAPLE. We refer to this method as *ignorant integration*:

```
U  := [51, 18, 73, 25, 75]:
f  := (s*t^4        +(1-s)*p^4        )^U[1] *
      (s*t^3*(1-t)  +(1-s)*p^3*(1-p)  )^U[2] *
      (s*t^2*(1-t)^2+(1-s)*p^2*(1-p)^2)^U[3] *
      (s*t  *(1-t)^3+(1-s)*p  *(1-p)^3)^U[4] *
      (s    *(1-t)^4+(1-s)    *(1-p)^4)^U[5]:
II := int(int(int(f,p=0..1),t=0..1),s=0..1);
```

In the case of mixture models, recognizing the integral as the sum of integrals of monomials over a polytope allows us to avoid the expensive integration step above by using (20). To demonstrate the power of using (20), we implemented a simple algorithm that computes each $\phi_A(b,U)$ using the naive expansion in (19). We computed the integral in Example 3 with a small data vector $U = (2,2,2,2,2)$, which is the rational number

$$\frac{66364720654753}{59057383987217015339940000},$$

and summarize the run-times and memory usages of the two algorithms in the table below. All experiments reported in this section are done in MAPLE.

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

For the remaining comparisons in this section, we no longer consider the ignorant integration algorithm because it is computationally too expensive.

4.2 Symbolic Expansion of the Integrand

While ignorant use of a computer algebra system is unsuitable for computing our integrals, we can still exploit its powerful polynomial expansion capabilities to find the coefficients of (21). A major advantage is that it is very easy to write code for this method. We compare the performance of this symbolic expansion algorithm against that of the naive expansion algorithm. The table below concerns computing the coefficients $\phi_A(b,U)$ for the original data $U = (51,18,73,25,75)$. The

column "Extract" refers to the time taken to extract the coefficients $\phi_A(b,U)$ from the expansion of the polynomial, while the column "Sum" shows the time taken to evaluate (20) after all the needed values of $\phi_A(b,U)$ had been computed and extracted.

	Time(seconds)				Memory
	$\phi_A(b,U)$	Extract	Sum	Total	(bytes)
Naive Expansion	2764.35	-	31.19	2795.54	10,287,268
Symbolic Expansion	28.73	962.86	29.44	1021.03	66,965,528

4.3 Storage and Evaluation of $\phi_A(b,U)$

Symbolic expansion is fast for computing $\phi_A(b,U)$, but it has two drawbacks: high memory usage and the long time it takes to extract the values of $\phi_A(b,U)$. One solution is to create specialized data structures and algorithms for expanding (21), rather using than those offered by MAPLE.

First, we tackle the problem of storing the coefficients $\phi_A(b,U)$ for $b \in Z_A^{\mathbb{L}}(U) \subset \mathbb{R}^d$ as they are being computed. One naive method is to use a d-dimensional array $\phi[\cdot]$. However, noting that A is not row rank full, we can use a d_0-dimensional array to store $\phi_A(b,U)$, where $d_0 = \text{rank}(A) = d-k+1$. Furthermore, by Proposition 5(2), the expanded integrand is a symmetric polynomial, so only half the coefficients need to be stored. We will leave out the implementation details so as not to complicate our discussions. In our algorithms, we will assume that the coefficients are stored in a d_0-dimensional array $\phi[\cdot]$, and the entry that represents $\phi_A(b,U)$ will be referred to as $\phi[b]$.

Next, we discuss how $\phi_A(b,U)$ can be computed. One could use the naive expansion (19), but this involves evaluating many binomials coefficients and products, so the algorithm is inefficient for data vectors with large coordinates. A more efficient solution uses the recurrence formula (22):

Algorithm 1 (RECURRENCE(*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* ϕ *of zeros.*
Step 2*: For each* $x \in \{0,1,\ldots,U_1\}$ *set*

$$\phi[xa_1] := \binom{U_1}{x}.$$

Step 3*: Create a new* d_0*-dimensional array* ϕ'.
Step 4*: For each* $2 \le j \le n$ *do*
1. *Set all the entries of* ϕ' *to* 0.
2. *For each* $x \in \{0,1,\ldots,U_j\}$ *do*
   For each non-zero entry $\phi[b]$ *in* ϕ *do*
   Increment $\phi'[b+xa_j]$ *by* $\binom{U_j}{x}\phi[b]$.
3. *Replace* ϕ *with* ϕ'.

Step 5*: Output the array* ϕ.

The space complexity of this algorithm is $O(N^{d_0})$ and its time complexity is $O(N^{d_0+1})$. By comparison, the naive expansion algorithm has space complexity $O(N^d)$ and time complexity $O(N^{n+1})$.

We now turn our attention to computing the integral (25). One major issue is the lack of memory to store all the terms of the expansion of the integrand. 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 vth column in $A^{[j]}$. The resulting algorithm for evaluating the integral is as follows:

Algorithm 2 (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 (25) in exact rational arithmetic.*
Step 1*: For $1\le j\le m$, compute $\phi^{[j]}:=\mathrm{RECURRENCE}(A^{[j]},U^{[j]})$.*
Step 2*: Set $I:=0$.*
Step 3*: For each non-zero entry $\phi^{[1]}[b^{[1]}]$ in $\phi^{[1]}$ do*

$$\vdots$$

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 algorithm can be sped up by precomputing the factorials used in the product in Step 3. The space and time complexity of this algorithm is $O(N^S)$ and $O(N^T)$ respectively, where $S=\max_i \mathrm{rank}A^{[i]}$ and $T=\sum_i \mathrm{rank}A^{[i]}$. From this, we see that the splitting of the integrand should be chosen wisely to achieve a good pay-off between the two complexities.

In the table below, we compare the naive expansion algorithm and the fast integral algorithm for the data $U=(51,18,73,25,75)$. We also compare the effect of splitting the integrand into two factors, as denoted by $m=1$ and $m=2$. For $m=1$, the fast integral algorithm takes significantly less time than naive expansion, and requires only about 1.5 times more memory.

	Time(minutes)	Memory(bytes)
Naive Expansion	43.67	9,173,360
Fast Integral (m=1)	1.76	13,497,944
Fast Integral (m=2)	139.47	6,355,828

4.4 Limitations and Applications

While our algorithms are optimized for exact evaluation of integrals for mixtures of independence models, they may not be practical for applications involving large sample sizes. To demonstrate their limitations, we vary the sample sizes in Example 3 and compare the computation times. The data vectors U are generated by scaling $U=(51,18,73,25,75)$ according to the sample size N and rounding off the entries. Here, N is varied from 110 to 300 by increments of 10. Figure 1 shows a logarithmic plot of the results. The times taken for $N=110$ and $N=300$ are 3.3 and 98.2 seconds

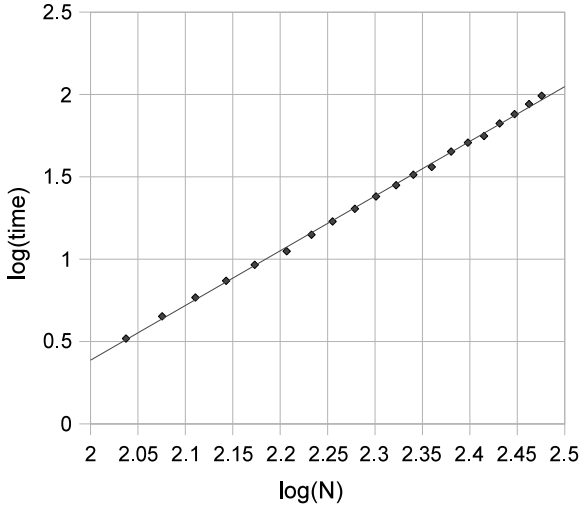

Figure 1: Comparison of computation time against sample size.

respectively. Computation times for larger samples may be extrapolated from the graph. Indeed, a sample size of 5000 could take more than 13 days.

For other models, such as the 100 *Swiss Francs* example in Section 1 and that of the schizophrenic patients in Example 9, the limitations are even more apparent. In the table below, for each example we list the sample size, computation time, rank of the corresponding A-matrix and the number of terms in the expansion of the integrand. Despite having smaller sample sizes, the computations for the latter two examples take a lot more time. This may be attributed to the higher ranks of the A-matrices and the larger number of terms that need to be summed up in our algorithm.

	Size	Time	Rank	#Terms
Coin Toss	242	45 sec	2	48,646
100 Swiss Francs	40	15 hrs	7	3,892,097
Schizophrenic Patients	132	16 days	5	34,177,836

Despite their high complexities, we believe our algorithms are important because they provide a gold standard with which approximation methods such as those studied in Chickering and Heckerman (1997) can be compared. Below, we use our exact methods to ascertain the accuracy of asymptotic formula derived in Watanabe (2001) and Watanabe and Yamazaki (2003, 2004) using desingularization methods from algebraic geometry.

Example 6 *Consider the model from Example 3. Choose data vectors* $U = (U_0, U_1, U_2, U_3, U_4)$ *with* $U_i = Nq_i$ *where* N *is a multiple of* 16 *and*

$$q_i = \frac{1}{16}\binom{4}{i}, \quad i = 0, 1, \ldots, 4.$$

Let $I_N(U)$ *be the integral (25). Define*

$$F_N(U) = N \sum_{i=0}^{4} q_i \log q_i - \log I_N(U).$$

According to Watanabe and Yamazaki (2004), for large N we have the asymptotics

$$\mathrm{E}_U[F_N(U)] \;=\; \frac{3}{4}\log N + O(1) \tag{26}$$

where the expectation E_U is taken over all U with sample size N under the distribution defined by $q = (q_0, q_1, q_2, q_3, q_4)$. Thus, we should expect

$$F_{16+N} - F_N \;\approx\; \frac{3}{4}\log(16+N) - \frac{3}{4}\log N =: g(N).$$

We compute $F_{16+N} - F_N$ using our exact methods and list the results below.

N	$F_{16+N} - F_N$	$g(N)$
16	*0.21027043*	*0.225772497*
32	*0.12553837*	*0.132068444*
48	*0.08977938*	*0.093704053*
64	*0.06993586*	*0.072682510*
80	*0.05729553*	*0.059385934*
96	*0.04853292*	*0.050210092*
112	*0.04209916*	*0.043493960*

Clearly, the table supports our conclusion. The coefficient $3/4$ of $\log N$ in the formula (26) is known as the real log-canonical threshold *of the statistical model. The example suggests that our method could be developed into a numerical technique for computing the real log-canonical threshold.*

5. Back to Bayesian Statistics

In this section we discuss how the exact integration approach presented here interfaces with issues in Bayesian statistics. The first concerns the rather restrictive assumption that our marginal likelihood integral be evaluated with respect to the uniform distribution (Lesbegue measure) on the parameter space Θ. It is standard practice to compute such integrals with respect to *Dirichlet priors*, and we shall now explain how our algorithms can be extended to Dirichlet priors. That extension is also available as a feature in our MAPLE implementation.

Recall that the *Dirichlet distribution* $\mathrm{Dir}(\alpha)$ is a continuous probability distribution parametrized by a vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m)$ of positive reals. It is the multivariate generalization of the beta distribution and is conjugate prior (in the Bayesian sense) to the multinomial distribution. This means that the probability distribution function of $\mathrm{Dir}(\alpha)$ specifies the belief that the probability of the ith among $m+1$ events equals θ_i given that it has been observed $\alpha_i - 1$ times. More precisely, the probability density function $f(\theta;\alpha)$ of $\mathrm{Dir}(\alpha)$ is supported on the m-dimensional simplex

$$\Delta_m \;=\; \bigl\{(\theta_0,\ldots,\theta_m) \in \mathbb{R}^m_{\geq 0} \,:\, \theta_0 + \cdots + \theta_m = 1\bigr\},$$

and it equals

$$f(\theta_0,\ldots,\theta_m;\alpha_0,\ldots,\alpha_m) \quad = \quad \frac{1}{\mathbb{B}(\alpha)} \cdot \theta_0^{\alpha_0-1}\theta_1^{\alpha_1-1}\cdots\theta_m^{\alpha_m-1} \;=:\; \frac{\theta^{\alpha-1}}{\mathbb{B}(\alpha)}.$$

Here the normalizing constant is the multinomial beta function

$$\mathbb{B}(\alpha) \quad = \quad \frac{m!\Gamma(\alpha_0)\Gamma(\alpha_1)\cdots\Gamma(\alpha_m)}{\Gamma(\alpha_0+\alpha_1+\cdots+\alpha_m)}.$$

Note that, if the α_i are all integers, then this is the rational number

$$\mathbb{B}(\alpha) \;=\; \frac{m!(\alpha_0-1)!(\alpha_1-1)!\cdots(\alpha_m-1)!}{(\alpha_0+\cdots+\alpha_m-1)!}.$$

Thus the identity (12) is the special case of the identity $\int_{\Delta_m} f(\theta;\alpha)d\theta = 1$ for the density of the Dirichlet distribution when all $\alpha_i = b_i+1$ are integers.

We now return to the marginal likelihood for mixtures of independence models. To compute this quantity with respect to Dirichlet priors means the following. We fix positive real numbers α_0,α_1, and $\beta_j^{(i)}$ and $\gamma_j^{(i)}$ for $i=1,\ldots,k$ and $j=0,\ldots,t_i$. These specify Dirichlet distributions on Δ_1, P and P. Namely, the Dirichlet distribution on P given by the $\beta_j^{(i)}$ is the product probability measure given by taking the Dirichlet distribution with parameters $(\beta_0^{(i)},\beta_1^{(i)},\ldots,\beta_{t_i}^{(i)})$ on the i-th factor Δ_{t_i} in the product (7) and similarly for the $\gamma_j^{(i)}$. The resulting product probability distribution on $\Theta = \Delta_1\times P\times P$ is called the *Dirichlet distribution* with parameters (α,β,γ). Its probability density function is the product of the respective densities:

$$f(\sigma,\theta,\rho;\alpha,\beta,\gamma) \;=\; \frac{\sigma^{\alpha-1}}{\mathbb{B}(\alpha)}\cdot\prod_{i=1}^{k}\frac{(\theta^{(i)})^{\beta^{(i)}-1}}{\mathbb{B}(\beta^{(i)})}\cdot\prod_{i=1}^{k}\frac{(\rho^{(i)})^{\gamma^{(i)}-1}}{\mathbb{B}(\gamma^{(i)})}. \tag{27}$$

By the marginal likelihood with Dirichlet priors we mean the integral

$$\int_\Theta \mathbf{L}_U(\sigma,\theta,\rho)\, f(\sigma,\theta,\rho;\alpha,\beta,\gamma)d\sigma d\theta d\rho. \tag{28}$$

This is a modification of (15) and it depends not just on the data U and the model $\mathcal{M}^{(2)}$ but also on the choice of Dirichlet parameters (α,β,γ). When the coordinates of these parameters are arbitrary positive reals but not integers, then the value of the integral (28) is no longer a rational number. Nonetheless, it can be computed exactly as follows. We abbreviate the product of gamma functions in the denominator of the density (27) as follows:

$$\mathbb{B}(\alpha,\beta,\gamma) \;:=\; \mathbb{B}(\alpha)\cdot\prod_{i=1}^{k}\mathbb{B}(\beta^{(i)})\cdot\prod_{i=1}^{k}\mathbb{B}(\gamma^{(i)}).$$

Instead of the integrand (18) we now need to integrate

$$\sum_{\substack{b\in Z_A^{\mathbb{L}}(U)\\ c=AU-b}} \frac{\phi_A(b,U)}{\mathbb{B}(\alpha,\beta,\gamma)}\cdot\sigma_0^{|b|/a+\alpha_0-1}\cdot\sigma_1^{|c|/a+\alpha_1-1}\cdot\theta^{b+\beta-1}\cdot\rho^{c+\gamma-1}$$

with respect to Lebesgue probability measure on Θ. Doing this term by term, as before, we obtain the following modification of Theorem 4.

Corollary 9 *The marginal likelihood of the data U in the mixture model $\mathcal{M}^{(2)}$ with respect to Dirichlet priors with parameters (α,β,γ) equals*

$$\frac{N!}{U_1!\cdots U_n!\cdot\mathbb{B}(\alpha,\beta,\gamma)}\cdot\sum_{\substack{b\in Z_A^{\mathbb{L}}(U)\\ c=AU-b}}\phi_A(b,U)\,\frac{\Gamma(|b|/a+\alpha_0)\Gamma(|c|/a+\alpha_1)}{\Gamma(|U|+|\alpha|)}$$
$$\cdot\prod_{i=1}^{k}\Big(\frac{t_i!\Gamma(b_0^{(i)}+\beta_0^{(i)})\cdots\Gamma(b_{t_i}^{(i)}+\beta_{t_i}^{(i)})}{\Gamma(|b^{(i)}|+|\beta^{(i)}|)}\;\frac{t_i!\Gamma(c_0^{(i)}+\gamma_0^{(i)})\cdots\Gamma(c_{t_i}^{(i)}+\gamma_{t_i}^{(i)})}{\Gamma(|c^{(i)}|+|\gamma^{(i)}|)}\Big).$$

A well-known experimental study (Chickering and Heckerman, 1997) compares different methods for computing numerical approximations of marginal likelihood integrals. The model considered in the study is the *naive-Bayes model*, which, in the language of algebraic geometry, corresponds to arbitrary secant varieties of Segre varieties. In this paper we considered the first secant variety of arbitrary Segre-Veronese varieties. In what follows we restrict our discussion to the intersection of both classes of models, namely, to the first secant variety of Segre varieties. For the remainder of this section we fix

$$s_1 = s_2 = \cdots = s_k = 1$$

but we allow $t_1, t_2, \ldots, t_k$ to be arbitrary positive integers. Thus in the model of Chickering and Heckerman (1997, Equation 1), we fix $r_C = 2$, and the n there corresponds to our k.

To keep things as simple as possible, we shall fix the uniform distribution as in Sections 1–4 above. Thus, in the notation of Chickering and Heckerman (1997, §2), all Dirichlet hyperparameters α_{ijk} are set to 1. This implies that, for any data $U \in \mathbb{N}^n$ and any of our models, the problem of finding the maximum a posteriori (MAP) configuration is equivalent to finding the maximum likelihood (ML) configuration. To be precise, the *MAP configuration* is the point $(\hat{\sigma}, \hat{\theta}, \hat{\rho})$ in Θ which maximizes the likelihood function $\mathbf{L}_U(\sigma, \theta, \rho)$ in (14). This maximum may not be unique, and there will typically be many local maxima. Chickering and Heckerman (1997, §3.2) used the expectation maximization (EM) algorithm (Pachter and Sturmfels, 2005, §1.3) to approximate the MAP configuration numerically

The Laplace approximation and the BIC score (Chickering and Heckerman, 1997, §3.1) are predicated on the idea that the MAP configuration can be found with high accuracy and that the data U were actually drawn from the corresponding distribution $p(\hat{\sigma}, \hat{\theta}, \hat{\rho})$. Let $\mathbf{H}(\sigma, \theta, \rho)$ denote the Hessian matrix of the log-likelihood function $\log \mathbf{L}(\sigma, \theta, \rho)$. Then the Laplace approximation (Chickering and Heckerman, 1997, Equation 15) states that the logarithm of the marginal likelihood can be approximated by

$$\log \mathbf{L}(\hat{\sigma}, \hat{\theta}, \hat{\rho}) - \frac{1}{2} \log |\det \mathbf{H}(\hat{\sigma}, \hat{\theta}, \hat{\rho})| + \frac{2d - 2k + 1}{2} \log(2\pi). \tag{29}$$

The Bayesian information criterion (BIC) suggests the coarser approximation

$$\log \mathbf{L}(\hat{\sigma}, \hat{\theta}, \hat{\rho}) \; - \; \frac{2d - 2k + 1}{2} \log(N), \tag{30}$$

where $N = U_1 + \cdots + U_n$ is the sample size.

In algebraic statistics, we do not content ourselves with the output of the EM algorithm but, to the extent possible, we seek to actually solve the likelihood equations (Hoşten et al., 2005) and compute all local maxima of the likelihood function. We consider it a difficult problem to reliably find $(\hat{\sigma}, \hat{\theta}, \hat{\rho})$, and we are concerned about the accuracy of any approximation like (29) or (30).

Example 7 *Consider the* 100 Swiss Francs *table (2) discussed in the Introduction. Here* $k = 2$, $s_1 = s_2 = 1$, $t_1 = t_2 = 3$, *the matrix A is unimodular, and (9) is the Segre embedding* $\mathbb{P}^3 \times \mathbb{P}^3 \hookrightarrow \mathbb{P}^{15}$. *The parameter space* Θ *is* 13*-dimensional, but the model* $\mathcal{M}^{(2)}$ *is* 11*-dimensional, so the given parametrization is not identifiable (Feinberg et al., 2007). This means that the Hessian matrix* $\mathbf{H}$ *is singular, and hence the Laplace approximation (29) is not defined.*

Example 8 *We compute (29) and (30) for the model and data in Example 3. According to Hoşten et al. (2005, Example 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* Θ*, which is the* 3*-cube. The two global maxima are*

$$(0.3367691969, 0.0287713237, 0.6536073424),$$
$$(0.6632308031, 0.6536073424, 0.0287713237).$$

Both of these points in Θ *give the same point in the model:*

$$(\hat{p}_0,\hat{p}_1,\hat{p}_2,\hat{p}_3,\hat{p}_4) \;=\; (0.12104, 0.25662, 0.20556, 0.10758, 0.30920).$$

The likelihood function evaluates to $0.1395471101 \times 10^{-18}$ *at this point. The following table compares the various approximations. Here, "Actual" refers to the base-10 logarithm of the marginal likelihood in Example 3.*

BIC	*-22.43100220*
Laplace	*-22.39666281*
Actual	*-22.10853411*

The method for computing the marginal likelihood which was found to be most accurate in the experimental study is the *candidate method* (Chickering and Heckerman, 1997, §3.4). This is a Monte-Carlo method which involves running a Gibbs sampler. The basic idea is that one wishes to compute a large sum, such as (20) by sampling among the terms rather than listing all terms. In the candidate method one uses not the sum (20) over the lattice points in the zonotope but the more naive sum over all 2^N hidden data that would result in the observed data represented by U. The value of the sum is the number of terms, 2^N, times the average of the summands, each of which is easy to compute. A comparison of the results of the candidate method with our exact computations, as well as a more accurate version of Gibbs sampling which is adapted for (20), will be the subject of a future study.

One of the applications of marginal likelihood integrals lies in model selection. An important concept in that field is that of *Bayes factors*. Given data and two competing models, the Bayes factor is the ratio of the marginal likelihood integral of the first model over the marginal likelihood integral of the second model. In our context it makes sense to form that ratio for the independence model $\mathcal{M}$ and its mixture $\mathcal{M}^{(2)}$. To be precise, given any independence model, specified by positive integers $s_1,\ldots,s_k,t_1,\ldots,t_k$ and a corresponding data vector $U \in \mathbb{N}^n$, the Bayes factor is the ratio of the marginal likelihood in Lemma 1 and the marginal likelihood in Theorem 4. Both quantities are rational numbers and hence so is their ratio.

Corollary 10 *The Bayes factor which discriminates between the independence model* $\mathcal{M}$ *and the mixture model* $\mathcal{M}^{(2)}$ *is a rational number. It can be computed exactly using Algorithm 2 (and our* MAPLE*-implementation).*

Example 9 *We conclude by applying our method to a data set taken from the Bayesian statistics literature. Evans, Gilula, and Guttman (1989, §3) analyzed the association between length of hospital stay (in years Y) of* 132 *schizophrenic patients and the frequency with which they are visited*

by relatives. Their data set is the following 3×3 *contingency table:*

$$U \;=\; \begin{array}{lccccc} & & 2\le Y<10 & 10\le Y<20 & 20\le Y & \textit{Totals} \\ & \textit{Visited regularly} & 43 & 16 & 3 & \textit{62} \\ & \textit{Visited rarely} & 6 & 11 & 10 & \textit{27} \\ & \textit{Visited never} & 9 & 18 & 16 & \textit{43} \\ & \textit{Totals} & \textit{58} & \textit{45} & \textit{29} & \mathbf{132} \end{array}$$

They present estimated posterior means and variances for these data, where "each estimate requires a 9-dimensional integration" *(Evans et al., 1989, 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" *(Evans et al., 1989, 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" *(Evans et al., 1989, p. 559).*

We differ on that conclusion. In our view, the closed form expressions in Section 3 are quite useful for modest sample size N*. Using Algorithm 2, we computed the integral (25). It is the rational number with numerator*

$$\begin{array}{l} 27801948853106338912064360032498932910387614080 5 \\ 28524283958209256935726588667532284587409752803 3 \\ 99493069713103633199906939405711180837568853737 \end{array}$$

and denominator

$$\begin{array}{l} 12288402873591935400678094796599848745442833177572204 \\ 50448819979286456995185542195946815073112429169997801 \\ 33503900169921912167352239204153786645029153951176422 \\ 43298328046163472261962028461650432024356339706541132 \\ 34375318471880274818667657423749120000000000000000. \end{array}$$

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*

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

Acknowledgments

Shaowei Lin was supported by graduate fellowship from A*STAR (Agency for Science, Technology and Research, Singapore). Bernd Sturmfels was supported by an Alexander von Humboldt research prize and the U.S. National Science Foundation (DMS-0456960). Zhiqiang Xu was supported by the NSFC grant 10871196 and a Sofia Kovalevskaya prize awarded to Olga Holtz.

References

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

D. Cox, J. Little, and D. O'Shea. *Using Algebraic Geometry*. Springer-Verlag, 2 edition, 2005.

M. Drton. Likelihood ratio tests and singularities. *Ann. Statist.*, 37(2):979–1012, 2009.

M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser, Basel, 2009.

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

S. Feinberg, P. Hersh, A. Rinaldo, and Y. Zhou. Maximum likelihood estimation in latent class models for contingency table data. arXiv:0709.3535, 2007.

W. Fulton. *Introduction to Toric Varieties*. Princeton Univ. Press, 1993.

S. Gao, G. Jiang, and M. Zhu. Solving the 100 swiss francs problem. arXiv:0809.4627, 2008.

D. Geiger and D. Rusakov. Asymptotic model selection for naive bayesian networks. *Journal of Machine Learning Research*, 6:1–35, 2005.

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

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

R. Stanley. A zonotope associated with graphical degree sequences. In P. Gritzmann and B. Sturmfels, editors, *Applied Geometry and Discrete Mathematics: The Victor Klee Festschrift*, volume 4 of *DIMACS Series in Discrete Mathematics*, pages 555–570. Amer. Math. Soc., 1991.

B. Sturmfels. Open problems in algebraic statistics. In M. Putinar and S. Sullivant, editors, *Emerging Applications of Algebraic Geometry*, volume 149 of *Volumes in Mathematics and its Applications*, pages 351–364. I.M.A., 2008.

S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13: 899–933, 2001.

S. Watanabe and K. Yamazaki. Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks*, 16:1029–1038, 2003.

S. Watanabe and K. Yamazaki. Newton diagram and stochastic complexity in mixture of binomial distributions. In *Algorithmic Learning Theorem*, volume 3244 of *Lecture Notes in Computer Science*, pages 350–364. Springer, 2004.

G. Ziegler. *Lectures on Polytopes*. Springer-Verlag, 1995.