Exact Bayesian evidence for mixture models: from two components to infinity

Matthew D. Schwartz
Department of Physics, Harvard University

Subhabrata Sen
Department of Statistics, Harvard University

October 1, 2026

PRELIMINARY

Abstract

Whether data that look heterogeneous contain one population or several is one of the oldest questions in statistics. The Bayesian answer is a ratio of marginal likelihoods, and for mixture models the defining integrals have been treated as intractable for decades. Exact values existed only for isolated small datasets, the asymptotic theory of singular models supplies growth rates without the constants, and practice falls back on Monte Carlo estimators or criteria built for regular models. For mixtures of a discrete variable the evidence is far more explicit than this record suggests. The two-component evidence is a closed form in harmonic numbers, exact at every sample size, for any counts, any number of states, and any integer Dirichlet priors. At any number of components g, under uniform component priors, the evidence collapses to a sum over component sizes weighted by contingency-table counts, computable exactly up to a thousand states and a hundred thousand observations. The limit $g \to \infty$ at fixed concentration is taken exactly: the finite-g evidence is a polynomial in $1/g$ with the Dirichlet-process evidence as its constant term. For data with every count at its expected value, and for binomial mixtures, two-class naive Bayes networks, and 3×3 contingency-table mixtures, we prove the constant-order terms that singular learning theory leaves open. The constants convert rates into sample sizes: in a trial recording four binary outcomes per participant, strong evidence against a spurious second component takes 3.3 million participants at a fifty-fifty response rate; BIC reaches the same threshold at 22. Against the exact values we grade the marginal-likelihood estimators in common use; we compute the evidence of a Dirichlet process mixture exactly on a real benchmark, a computation previously done only by Monte Carlo. Applications to capture–recapture, record linkage, environmental DNA, and mutational signatures appear in the supplement.

Keywords: marginal likelihood; Bayes factor; mixture model; singular learning theory; Dirichlet process; model selection.

1 Introduction

Finite mixture models describe data drawn from several latent subpopulations, and the first question asked of such data, since Pearson (1894) fit two normal components to Weldon's crab measurements, is how many subpopulations there are. In the Bayesian formulation each candidate number of components g defines a model, and the models are compared through their marginal likelihoods,

or evidences (Kass and Raftery, 1995): the probability of the observed data under each model, with the model parameters integrated out against their prior. For mixtures of a single categorical variable we obtain the evidence in closed form at two components and, under uniform component priors, compute it exactly at any finite number of components and in the Dirichlet process limit. For three singular families we also determine the large-sample expansion of the evidence, including its constant term, on expected-count data, in which every cell count equals its expectation under the smaller model.

The comparison of mixture models by their evidences arises in several of the applications treated below. In a capture–recapture study in which 76 rabbits were caught over eighteen trapping occasions, the posterior mean of the population size is about 98 if the animals form one catchability class and about 150 if they form two (Supplement S5). For a trial that records four binary outcomes per participant, we determine the number of participants needed before the Bayes factor can favor a single homogeneous population over one with a spurious subpopulation of responders (Section 6.1). A targeted cancer gene panel reports only a handful of mutations per tumor, and the Bayes factor for the presence of a mutational signature must be computed from those few counts (Supplement S10). In all three problems the conclusion depends on a ratio of evidences of mixture models with different numbers of components, at a sample size small enough, or at target odds large enough, that the approximations reviewed below are in doubt. Mixture models present several well-known difficulties for inference, and most of them bear directly on this comparison of g with $g + 1$ components.

1.1 Known difficulties

A model with g components contains every model with fewer. Setting a mixing weight to zero, or making two components identical, reproduces a $(g-1)$-component distribution on a set of parameter values of positive dimension, the singular set, on which the parameters are not identifiable and the Fisher information matrix is singular. This degeneracy is distinct from the permutation symmetry of the component labels, and unlike that symmetry it invalidates the regular large-sample theory of estimation and testing. For a single categorical variable the degeneracy is complete: a mixture of k-state distributions is again a k-state distribution, so the mixture parameters cannot be recovered from the data at any sample size (Teicher, 1963).

The classical theory of testing $g-1$ against g components fails on the singular set: the likelihood-ratio statistic has no χ^2 limit there. For normal components with an unrestricted mean the statistic diverges (Hartigan, 1985), and on compact parameter sets its limit law is the supremum of a squared Gaussian process indexed by the directions of approach to the smaller model (Dacunha-Castelle and Gassiat, 1997). In an overfitted Bayesian mixture the posterior either empties the superfluous components or splits a true component among them, according to the Dirichlet exponent of the prior on the weights (Rousseau and Mengersen, 2011). Mixtures of normal distributions add difficulties of their own: the likelihood is unbounded, the likelihood surface has many local modes and the largest finite one is difficult to locate, and an improper prior on the component parameters gives an improper posterior at every sample size. Wasserman (2012) listed these facts and concluded that mixtures, "like tequila, are inherently evil and should be avoided at all costs."

The same singular set controls the large-sample behavior of the evidence, and hence the accuracy of BIC, the usual criterion for choosing the number of components in model-based clustering and latent-class analysis (Fraley and Raftery, 2002; Linzer and Lewis, 2011; Nylund et al., 2007). Schwarz's approximation $\log Z_N = \log P(D \mid \hat{\theta}) - \frac{d}{2} \log N + \mathcal{O}(1)$, the basis of BIC, is a Laplace expansion about a nondegenerate maximum (Schwarz, 1978). When the data are con-

PRELIMINARY

sistent with fewer components than the model allows, the maximum is attained on the whole singular set and the expansion does not apply. The coefficient of $\log N$ then changes: $-\log Z_N = NS + \lambda \log N - (m-1)\log\log N + \mathcal{O}(1)$ on expected-count data, with S the log loss of the true distribution, a rational learning coefficient $\lambda \leq d/2$, and a multiplicity m (Watanabe, 2009). The pairs (λ, m) have been derived for several mixture and related singular families by resolution of singularities (Yamazaki and Watanabe, 2004; Rusakov and Geiger, 2005; Aoyagi and Watanabe, 2005; Lin, 2011) and are the inputs of the singular BIC of Drton and Plummer (2017). However, singular learning theory, which gives the expansion and the learning coefficients, does not determine the $\mathcal{O}(1)$ term at the degenerate configurations that matter for model choice. Lin's leading-coefficient calculus (Lin, 2011, Ch. 5) supplies it where the phase is Newton-nondegenerate, and to our knowledge it has not been computed for any of the families above (Section 2.2). Any sample-size calculation based on the singular expansion needs the constant, and the resulting sample size is sensitive to its value. When the competing models have learning coefficients that differ by $1/4$, as they do for the binomial mixtures of Section 6.1, the Bayes factor in favor of the smaller model grows only like $c\,N^{1/4}$. The sample size at which it reaches a threshold B is then $(B/c)^4$, proportional to the inverse fourth power of the unknown constant.

Direct computation of the evidence, without asymptotic approximation, is also difficult, because the likelihood of N observations is a product of N sums over components and the integrand of the evidence therefore expands into g^N allocation terms. Under conjugate priors each term integrates in closed form, but the sum has too many terms to enumerate. Lin et al. (2009) showed that for mixtures of independence models on discrete data the sum can be reorganized over the values of a sufficient statistic and evaluated exactly for datasets of moderate size, and that the resulting Bayes factors are rational numbers. For normal components the reorganization gives no reduction, because with continuous data distinct allocations almost surely have distinct component sums and sums of squares. In applied work the evidence is instead estimated from Markov chain Monte Carlo output, by the harmonic mean, Chib's method, bridge sampling, stepping-stone sampling, or nested sampling (Newton and Raftery, 1994; Chib, 1995; Meng and Wong, 1996; Xie et al., 2011; Skilling, 2006). Mixtures are difficult for these estimators as well: Chib's estimator is biased by $\log g!$ unless the sampler visits every labeling of the components (Neal, 1999), and the harmonic mean estimator, which remains in common use, often has infinite variance (Newton and Raftery, 1994). Each of these methods also gives an error bar computed from the same Monte Carlo samples, and the accuracy of that error bar on a given mixture problem is unknown unless an exact value is available for comparison.

1.2 Contributions and techniques

We work with the family on which the exact computations of Lin et al. (2009) were built, mixtures of independence models for discrete data, under proper uniform or Dirichlet priors (Section 2.1). In these models the likelihood is bounded and the evidence is a finite sum that can be evaluated in exact arithmetic. Because the evidence is an integral over the whole parameter space, no likelihood mode has to be located and no identifiability constraint is imposed. The label symmetry and the singular set are integrated over, and the latter affects only the large-sample expansion of the evidence, which for these models can be compared with exact finite-sample values. The main results are the following.

Closed forms at two components (Section 3). For one k-state variable the summand depends on the allocation only through the number m of observations assigned to the first component. The

PRELIMINARY

allocation sum therefore collapses to a single sum over $m = 0, \ldots, N$ of a lattice-point count $W(m)$, piecewise polynomial in m, against a ratio of factorials with simple poles in m. Evaluating this sum by partial fractions shows that the evidence is an explicit $\mathbb{Q}$-linear combination of 1 and harmonic numbers, for any counts, any k, and any integer Dirichlet hyperparameters (Theorems 1–3; the general-k reduction is proved in Appendix A). A first consequence concerns Teicher's non-identifiable case: for one balanced binary variable the Bayes factor between one and two components converges to $1/(2\log 2)$, so that increasing the sample size never produces strong evidence for either model (Corollary 1). More generally, for one categorical variable a single component already ranges over the whole simplex of cell-probability vectors, and the models with different numbers of components therefore contain the same set of distributions. Their evidences differ only through the priors that the component structure induces on that simplex.

Any number of components, and the Dirichlet process limit (Section 4). The cancellation that reduces the two-component evidence to a single sum persists at every g when the component priors are uniform: the summand depends on the $k \times g$ allocation table only through its column sums. The evidence then collapses to a sum over component sizes $m_1 + \cdots + m_g = N$ weighted by the number $\mathrm{CT}_U(m)$ of contingency tables with margins (U, m), that is, with the observed cell counts as row sums and the component sizes as column sums. The generating function of these table counts is a product of complete homogeneous symmetric polynomials (Theorem 4). At two components the exact rational evaluation takes about five minutes on one core for $k = 1{,}000$ states and $N = 10^5$ observations (Table 1). With symmetric $\mathrm{Dir}(\alpha/g, \ldots, \alpha/g)$ weights the g-component evidence is a polynomial in $1/g$ of degree at most $N - 1$. Its constant term is the evidence of the Dirichlet process mixture with concentration α at sample size N, an explicit finite sum over cluster sizes (Theorem 5). The limit $g \to \infty$ is therefore taken by the substitution $1/g = 0$, with no interchange of limits to justify. Non-uniform component priors at general g remain open.

The constant term for three singular families (Section 5). On deterministic expected-count rays $U = Nq$ with q in the smaller model, and under uniform priors, we compute the $\mathcal{O}(1)$ term of the singular expansion. Laplace's method does not apply as it stands, because the exponent is maximal on a set of positive dimension with degenerate transverse directions. The dominant part of this set is either an interior ridge or a stratum on the boundary of the mixing simplex. In each family we integrate exactly along the flat directions and expand to the first non-vanishing transverse order. In the binomial mixture with four trials at a one-component truth, the allocation sum on the expected-count ray is a lattice sum whose summand has an explicit large-deviation exponent. The exponent is already in normal-crossing form in the allocation coordinates, so that no resolution of singularities is needed, and the dominant set is an interior ridge with quartic transverse decay. For data $U = n\,(1, 4, 6, 4, 1)$, that is $N = 16n$ observations at $\theta_0 = \frac{1}{2}$, the expansion about the ridge gives $Z = C\, n^{-3/4}\, 2^{-64n}(1 + o(1))$ with $C = \sqrt{2}\,\pi^{3/2}\,\Gamma(5/4)/(32 \cdot 3^{1/4})$ (Theorem 6), and at every one-component truth the constant is a universal number times the per-trial variance $\theta_0(1 - \theta_0)$ (Theorem 7).

The ridge computation extends to continuous components (Section 5.3). We compare one unit-variance normal component with a mixture of two, under a uniform prior on the mixing weight and an $N(0, s_0^2)$ prior on each mean, with the log likelihood replaced by its expectation under the truth $N(\mu_0, 1)$. In this setting we prove that the Bayes factor for one component grows with the sample size n as $c\, n^{1/4}$, where $c = 2s_0\, e^{\mu_0^2/(2s_0^2)}/(\sqrt{\pi}\,\Gamma(1/4))$ (Proposition 1; Supplement S1.3). For standard normal priors centered at the truth its leading term reaches 3 near $n = 8{,}600$ and 20 near $n = 1.7 \times 10^7$. Reduced-rank regression, treated in the same way, gives a second continuous

PRELIMINARY

example in which the evidence and the constants of its large-sample expansion have closed forms.

In multivariate latent-class models at a one-class truth the evidence localizes at the boundary of the mixing simplex, along which the exponent is exactly flat (Lemma 3) and across which the Hessian factorizes into per-factor χ^2 divergences (Theorem 8). The Laplace step across the boundary is controlled by dominated convergence with an N-free integrable envelope. The boundary computation yields the constants for the two-class naive Bayes model on three binary features and the two-class mixture on 3×3 tables (Theorems 9 and 10), whose exponents are due to Rusakov and Geiger (2005) and Lin (2017, 2011). Section 5 states which localization steps are proved and which rest on numerically determined decay orders.

Consequences of the exact values (Section 6). Once the constant term is known, the singular expansion can be solved for the sample size at which the Bayes factor reaches a given threshold. In a design recording four binary outcomes per participant, the Bayes factor against a spurious second component grows on expected-count data as $0.2311\, N^{1/4}/\sqrt{\theta_0(1-\theta_0)}$ (Corollary 2). At a fifty-fifty response rate the exact Bayes factor reaches 3 at $N = 1{,}281$ and 20 near $N = 3.3 \times 10^6$. On the median dataset generated from the one-component model, the BIC approximation to this Bayes factor gives twenty-to-one odds against the second component already at $N = 22$ (Section 6.1). The corollary extends to a sample-size formula for any target odds and any number of outcomes per participant. We then use the exact values to assess five standard marginal-likelihood estimators and the coarsened posterior of Miller and Dunson (2019), and compare a Gibbs sampler for record linkage with exact partition posteriors (Section 6.2). For the eyetracking counts of Escobar and West (1998) we compute exactly the Dirichlet-process evidence and the posterior distribution of the number of clusters (Section 6.3). Six further applications are summarized in Section 6.4 and given in full in the supplement: capture–recapture, the die count of a Roman coin issue, latent classes in an environmental-DNA survey, record linkage, a comparison of partition priors by their evidence, and mutational signatures in small gene panels.

1.3 Organization

Section 2 fixes notation and reviews the models, the asymptotic and computational work on the evidence, and the partition-model formulation of mixtures with infinitely many components. Sections 3–6 contain the results, in the order in which Section 1.2 describes them, and Table 6 in Section 6 gives an overview of all the applications. Section 7 discusses a limitation of the method and open problems. Appendix A proves the general-k reduction, and the supplement contains the remaining proofs, the numerical methods, an analysis of the regime in which the number of states grows in proportion to the sample size, and the six further applications.

2 Mixture models and their evidence

The models of Sections 3–5 are, apart from the two continuous examples of Section 5.3, mixtures of independence models on discrete data in the sense of Lin et al. (2009), with two components, any finite number g of components, or the Dirichlet-process limit $g \to \infty$. Throughout, the quantity computed for each model is its evidence under a proper prior. Section 2.1 fixes notation and the measure conventions, Section 2.2 reviews the singular asymptotic theory and the approximate and exact methods of computing the evidence, and Section 2.3 recalls the partition-model form of the infinite-component limit used in Sections 4 and 6.

PRELIMINARY

2.1 Models, notation, and conventions

The evidence, or marginal likelihood, of a model $\mathcal{M}$ with parameters θ, prior density $\pi(\theta)$, and data D is

$$Z \;=\; P(D \mid \mathcal{M}) \;=\; \int d\theta\, \pi(\theta)\, P(D \mid \theta, \mathcal{M})\,. \tag{2.1}$$

The ratio of two evidences, the Bayes factor, converts the prior odds of two models into their posterior odds; a factor of 20 is conventionally read as strong evidence (Kass and Raftery, 1995). The data type throughout is categorical: each observation falls in one of k cells, and a dataset is summarized by counts $U = (U_1, \ldots, U_k)$ with $\sum_v U_v = N$. We write Z for the probability of the observed sequence. The multinomial coefficient by which it differs from the probability of the counts is common to every model of the same data and cancels in every Bayes factor. For the one-component model, independent draws from $\theta \in \Delta_{k-1}$, the evidence under a uniform prior is the Dirichlet integral

$$Z_1(U) \;=\; (k-1)!\,\frac{U_1!\,U_2!\cdots U_k!}{(N+k-1)!}\,, \tag{2.2}$$

and under a general Dirichlet prior a ratio of Gamma functions. Conjugacy gives the one-component evidence in closed form wherever it appears below.

The two-component mixture replaces the cell probabilities by $p_v = \sigma\,\theta_v + (1-\sigma)\,\rho_v$, and its evidence, with uniform priors on all three parameter blocks, is

$$Z_2(U) \;=\; \int_0^1 d\sigma \int_{\Delta_{k-1}} d\mu(\theta) \int_{\Delta_{k-1}} d\mu(\rho) \prod_{v=1}^{k} \left(\sigma\,\theta_v + (1-\sigma)\,\rho_v\right)^{U_v}\,, \tag{2.3}$$

with $d\mu$ the normalized simplex measure. The integrand is a product of N sums, and no change of variables makes it a monomial. The parameterization is also degenerate, and most severely so at the one-component distributions against which the mixture is compared: swapping (θ, σ) with $(\rho, 1-\sigma)$ leaves every p_v unchanged; on the surface $\theta = \rho$ every value of σ gives the same distribution; and at $\sigma \in \{0, 1\}$ one component decouples with its parameters unconstrained. Evans et al. (1989) wrote down finite-sum expansions of Eq. (2.3) for the two-component latent-class model and judged them too complicated for practical use. The problem of closed-form evaluation, meaning an explicit formula in the counts rather than a finite sum over allocations or over their sufficient statistics, appears to have been open since.

We follow the conventions of Lin et al. (2009), whose published rational values serve as test cases. In their setting an observation consists of F discrete variables, or factors; factor f takes $t_f + 1$ states and is observed r_f times exchangeably within the observation, and a cell v is specified by the number of copies of each factor in each state. The independence model gives factor f a probability vector $\theta^{(f)} \in \Delta_{t_f}$, independently across factors and copies, so that every ordered outcome whose state counts are v has probability the monomial

$$p_v(\theta) \;=\; \theta^{a_v} \;=\; \prod_{f=1}^{F} \prod_{j=0}^{t_f} \left(\theta_j^{(f)}\right)^{a_{v,fj}}\,, \qquad \theta = (\theta^{(1)}, \ldots, \theta^{(F)}) \in P = \Delta_{t_1} \times \cdots \times \Delta_{t_F}\,, \tag{2.4}$$

where $a_{v,fj}$ is the number of copies of factor f in state j in cell v, and A denotes the matrix with columns a_v. The cell v itself has probability $c_v\, p_v(\theta)$ with $c_v = \prod_f \left(r_f! / \prod_j a_{v,fj}!\right)$, and following Lin et al. (2009) we work with the unscaled monomials, the factor $\prod_v c_v^{U_v}$ being common to all

PRELIMINARY

models of the same data. One k-state variable is the case $F = 1$, $t_1 = k - 1$, $r_1 = 1$, with $p_v = \theta_v$; F binary features observed once each give $p_v = \prod_f \theta_f^{v_f}(1-\theta_f)^{1-v_f}$ for $v \in \{0,1\}^F$; a $k \times k$ table is $F = 2$ with $p_{ij} = \theta_i^{(1)}\theta_j^{(2)}$. The two-component mixture draws each observation from one of two such parameter vectors, all copies of all factors in the observation sharing the component:

$$p_v \;=\; \sigma_0\,\theta^{a_v} + \sigma_1\,\rho^{a_v}\,, \qquad (\sigma_0,\sigma_1) \in \Delta_1\,, \quad \theta,\rho \in P\,. \tag{2.5}$$

For one variable observed once, Eq. (2.5) reduces to the cell probabilities $p_v = \sigma\theta_v + (1-\sigma)\rho_v$ of Eq. (2.3), with $\sigma_0 = \sigma$. The evidence is the integral of $\prod_v p_v^{U_v}$ over $\Theta = \Delta_1 \times P \times P$ with the Lebesgue probability measure on each simplex (total mass one per simplex). The one exception is the contingency-table computations of Section 5.4, where Z carries the coordinate Lebesgue measure $da_1\,da_2$ on each Δ_2 (mass $\frac{1}{2}$ per simplex); the mass-one convention multiplies those Z by $2^4 = 16$ and their constants accordingly. Two conventions matter and are easy to get wrong: the measure normalization contributes a factor $t_f!$ per simplex Δ_{t_f} through

$$\int_{\Delta_t} \theta_0^{b_0}\cdots\theta_t^{b_t}\,d\theta \;=\; \frac{t!\;b_0!\cdots b_t!}{(b_0+\cdots+b_t+t)!}\,, \tag{2.6}$$

and for models with repeated variables the reduced-state parameterization is unscaled: no binomial coefficients multiply the monomials. Throughout, g denotes the number of mixture components ($g = 2$ until Section 4) and k the number of states of the observed variable. Both log and ln denote the natural logarithm, and base-10 logarithms are written $\log_{10}$; harmonic numbers carry a subscript ($H_j = \sum_{i\le j} 1/i$); $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy; an unsubscripted H is a transverse Hessian in Section 5.4.

Three models recur:

1) the *binary model*: one variable with two states, cells $p_0 = \sigma_0\theta + \sigma_1\rho$, $p_1 = \sigma_0(1-\theta) + \sigma_1(1-\rho)$, counts $U = (u_0, u_1)$;

2) the *k-state model*: one variable with k states;

3) the *binomial mixture*: four exchangeable binary variables, reduced states $v \in \{0,\ldots,4\}$ with unscaled monomials $f_v(\theta) = \theta^{4-v}(1-\theta)^v$, cells $p_v = \sigma_0 f_v(\theta) + \sigma_1 f_v(\rho)$, counts $U = (U_0,\ldots,U_4)$; the secant of the quartic rational normal curve and the standard singular-learning example (Yamazaki and Watanabe, 2004; Lin et al., 2009).

Before using these conventions we checked our implementation against the three published values of Lin et al. (2009): the coin-toss rationals of their §4.1 (data from Hoşten et al., 2005), the $N = 132$ schizophrenic-patients table of their Example 5.5, and the 100 Swiss Francs table of Pachter and Sturmfels (2005) (Supplement S3).

2.2 Asymptotic theory and approximate computation

The standard large-sample treatment of the integral Eq. (2.3) assumes a regular model. Laplace approximation at an interior maximum of the likelihood then gives Schwarz's Bayesian information criterion, $\log Z = \log P(D \mid \hat\theta) - \frac{d}{2}\log N + \mathcal{O}(1)$ with d the parameter count (Schwarz, 1978), the maximized likelihood computed in practice by expectation–maximization (Dempster et al., 1977). Keribin (2000) showed that BIC, among other penalized-likelihood criteria, selects the true number

PRELIMINARY

of mixture components almost surely as $N \to \infty$, under regularity conditions on the component family. However, every hypothesis in Schwarz's derivation fails for Eq. (2.3) at data consistent with a single component: the likelihood is maximized on the whole degenerate set, the Hessian is singular along it, and the Gaussian integral does not exist. Models with this structure are called singular, and mixtures are the canonical example. Because the likelihood is flat along the degenerate directions, the effective number of parameters is smaller than d, and a $d/2$ penalty is likely to overstate the case against the larger model.

The correct asymptotics for singular models are Watanabe's:

$$-\log Z_N \;=\; NS \;+\; \lambda \log N \;-\; (m-1)\log\log N \;+\; \mathcal{O}(1)\,, \tag{2.7}$$

with S the log loss of the true distribution, λ the real log canonical threshold (learning coefficient; $d/2$ for a regular model), and m its multiplicity (Watanabe, 2009). For sampled data the first term is the empirical loss NS_N with an $O_p(1)$ remainder. On the deterministic count rays $U = Nq$ used throughout this paper the expansion holds as written with an $\mathcal{O}(1)$ remainder. The pair (λ, m) is computable through resolution of singularities and has been derived family by family (Yamazaki and Watanabe, 2003, 2004; Rusakov and Geiger, 2005; Aoyagi and Watanabe, 2005). The criterion built on the theory, the singular BIC of Drton and Plummer (2017), replaces $d/2$ by λ and takes the λ as inputs. However, singular learning theory does not in general determine the $\mathcal{O}(1)$ term of Eq. (2.7). Resolution methods yield the constant only where the phase is Newton-nondegenerate; there the leading-coefficient calculus of Lin (2011, Ch. 5) computes it, including an exact data-dependent constant for a two-parameter coin-mixture model (his Example 1.18). At the degenerate configurations of interest here the theory fixes evidence ratios only up to an unknown constant factor. Drton and Plummer note this limitation: "We are not aware of any successful work on higher-order approximations in statistically relevant singular settings" (their §2.2). The 2026 extension of the criterion to probabilistic principal component analysis is described by its authors in the same terms, as "equal to the true marginal likelihood up to an error term of order $O_p(1)$" (Drton et al., 2026). The constants derived below are this $\mathcal{O}(1)$ term, for three families where the λ are known.

Monte Carlo methods do not rely on asymptotic expansions: reversible-jump Markov chain Monte Carlo (Richardson and Green, 1997) and samplers for mixtures with a prior on the number of components (Miller and Harrison, 2018) estimate posteriors well, but a Markov chain samples the posterior without computing its normalizing constant, which is the evidence. Extracting it requires estimators of its own (harmonic mean, bridge sampling, Chib's method, stepping-stone, nested sampling; Newton and Raftery, 1994; Meng and Wong, 1996; Chib, 1995; Xie et al., 2011; Skilling, 2006). Each gives a point estimate and an error bar computed from the same Monte Carlo samples, so the calibration of the error bar can be checked only against an exact value; we carry out this check in Section 6.2. On the exact side, Lin et al. (2009) observed that for mixtures of independence models with uniform or Dirichlet priors $Z(U)$ is a rational number, gave algorithms that compute it for datasets of moderate size, and used the values to test the singular-learning asymptotics numerically. Borinsky et al. (2023) later replaced exact evaluation with tropical Monte Carlo sampling and noted that the error constants of their estimator grow exponentially with the sample size. They concluded that "exact symbolic algorithms are still surprisingly competitive." The algorithms of Lin et al. evaluate the evidence exactly, one dataset at a time, and their scaling study (their Example 4.3) measured the exponent λ but left the $\mathcal{O}(1)$ term open.

PRELIMINARY

2.3 Infinitely many components and partition models

Bayesian nonparametric mixtures place no upper bound on the number of components: give the mixing weights the symmetric Dirichlet prior $\mathrm{Dir}(\alpha/g, \ldots, \alpha/g)$ and send $g \to \infty$ at fixed concentration $\alpha > 0$. The limit of the predictive distribution is the Chinese restaurant process, and the probability of a partition π of the N observations into t blocks is

$$P(\pi) \;=\; \frac{\alpha^t \prod_{B\in\pi} (|B|-1)!}{\alpha(\alpha+1)\cdots(\alpha+N-1)}\,. \tag{2.8}$$

The random measure underlying the Chinese restaurant process is Ferguson's Dirichlet process (Ferguson, 1973), and the resulting model is the Dirichlet process mixture (Antoniak, 1974), sampled in practice following Escobar and West (1998). Its evidence follows from Eq. (2.8) by conditioning on the partition: given π the blocks are independent one-component problems, so

$$Z_{\mathrm{DPM}}(U;\alpha) \;=\; \sum_{\pi} P(\pi) \prod_{B\in\pi} Z_1(U_B)\,, \tag{2.9}$$

with U_B the cell counts of block B: a finite sum with roughly 10^{115} terms at $N = 100$. Equation (2.9) is more general than the Dirichlet process: any prior over partitions together with any per-block marginal likelihood defines a partition model, the Pitman–Yor two-parameter family (Pitman and Yor, 1997) being a standard alternative. In record linkage, for example, each $B \in \pi$ is the set of records attributed to one entity, and $Z_1(U_B)$ is the probability of generating those records as distorted copies of that entity. Sections 6.2 and 6.4 evaluate Eq. (2.9) exactly for linkage models under several partition priors, which Section 6.4 ranks by evidence. In each comparison the likelihood and hyperparameters are fixed and links are allowed only within prespecified groups of records ("blocking"). The scaling α/g is essential: under $\mathrm{Dir}(c, \ldots, c)$ with c fixed, two observations share a component with prior probability $(c+1)/(gc+1) \to 0$, against $1/(1+\alpha)$ under α/g, so as $g \to \infty$ every observation has its own component. The data are then iid from the prior predictive of one component, uniform on the k states under uniform component priors, and the evidence is $(1/k)^N$ whatever the data (Section 4.3).

3 Closed forms at two components

The evidence of the two-component mixture of one k-state variable is a closed form: an explicit $\mathbb{Q}$-linear combination of 1 and harmonic numbers, valid for arbitrary counts, arbitrary k, and arbitrary integer Dirichlet priors. We state and prove the formulas and then show, as their first statistical consequence, that for binary data the Bayes factor between the one-component model and the two-component mixture converges to an explicit constant as the sample size grows.

3.1 The allocation sum and the binary model

Expanding each cell of Eq. (2.5) binomially and integrating term by term with Eq. (2.6) gives

$$Z(U) \;=\; \sum_{0\le x\le U} \prod_{v} \binom{U_v}{x_v}\, D\big(|x|, N-|x|\big)\; E(Ax)\; E\big(A(U-x)\big)\,, \tag{3.1}$$

PRELIMINARY

where x_v counts the observations of cell v allocated to the first component, $D(a,b) = a!\,b!/(a+b+1)!$ is the mixing-weight integral, and $E(b)$ is the product of simplex integrals (2.6) of the sufficient statistic $b = Ax$, with A the matrix of exponent vectors. This is the expansion of Lin et al. (2009), whose algorithms organize the sum over the attainable values of the sufficient statistic b, which are the lattice points of a polytope. However, for the one-variable models the sum collapses to closed form.

Theorem 1 (Binary model, uniform prior)**.** *For the binary model with counts $u_0 \le u_1$ and $N = u_0 + u_1$,*

$$Z(u_0,u_1) \;=\; \frac{u_0!\,u_1!}{(N+1)!}\cdot\frac{2}{N+2}\Big[(u_0{+}1) + (u_0{+}1)(H_{u_1+1}{-}H_{u_0+1}) + (N{+}2)(H_{N+1}{-}H_{u_1+1}) - (N{-}u_1)\Big]\,, \tag{3.2}$$

where H_j is the j-th harmonic number. In particular, for symmetric counts,

$$Z(n,n) \;=\; \frac{(n!)^2}{(2n+1)!}\left[\frac{1}{n+1} + 2\big(H_{2n+1} - H_{n+1}\big)\right] \;=\; B(n{+}1,n{+}1)\left[\frac{1}{n+1} + 2\big(H_{2n+1} - H_{n+1}\big)\right]. \tag{3.3}$$

Proof. For a single k-state variable the exponent matrix is the identity, so $b = x$ in Eq. (3.1) and the binomials combine with the factorials of E: $\binom{u_v}{x_v} x_v!(u_v - x_v)! = u_v!$. The sum therefore depends on x only through $m = |x|$,

$$Z(U) = \frac{\prod_v u_v!\,\big((k-1)!\big)^2}{(N+1)!}\sum_{m=0}^{N} W(m)\,\frac{m!\,(N-m)!}{(m+k-1)!\,(N-m+k-1)!}\,, \qquad W(m) = \#\{x \le U : |x| = m\}\,. \tag{3.4}$$

For $k = 2$ the weight is $1/\big[(m{+}1)(N{-}m{+}1)\big]$, which partial-fractions as $\frac{1}{N+2}\big[\frac{1}{m+1} + \frac{1}{N-m+1}\big]$; since $W(N-m) = W(m)$, both halves contribute equally. The count $W(m)$ for a box $[0,u_0]\times[0,u_1]$ is the trapezoid $W(m) = m+1$ for $m \le u_0$, u_0+1 for $u_0 \le m \le u_1$, and $N - m + 1$ above, and the sum $\sum_m W(m)/(m+1)$ evaluates in harmonic numbers piece by piece. Collecting terms gives Eq. (3.2). □

The harmonic numbers arise from the mixing-weight integral D: after the partial-fraction step of the proof, each allocation total m carries a factor $1/(m+1)$, and $H_{N+1} - H_{u_1+1}$ is the sum of these factors over the totals m that exceed u_1. Expanding the harmonic numbers gives the complete asymptotic series of the evidence; the leading terms already show the singular structure treated in general in Section 5, e.g. from Eq. (3.3),

$$Z(n,n) \;=\; \sqrt{\pi}\,\log 2\;\cdot 4^{-n}\,n^{-1/2}\,\big(1 + \mathcal{O}(1/n)\big)\,, \tag{3.5}$$

a learning coefficient $\lambda = 1/2$ with constant $\sqrt{\pi}\log 2$.

3.2 The k-state family and Dirichlet priors

The reduction of the allocation sum to the single sum Eq. (3.4) over $m = |x|$ holds for every k and gives the evidence in closed form for the whole one-variable family.

PRELIMINARY

Theorem 2 (k-state model, uniform prior)**.** *For the k-state model with arbitrary counts U, the evidence $Z(U)$ is an explicit $\mathbb{Q}$-linear combination of* 1 *and harmonic numbers with arguments determined by the partial sums of the counts. The reduction is constructive: partial-fraction the weight in Eq.* (3.4) *over its simple poles, apply inclusion–exclusion to $W(m)$ (a piecewise polynomial), and divide each polynomial piece by each linear factor; every resulting sum is either polynomial (rational contribution) or a harmonic-number difference.*

Theorem 3 (Integer Dirichlet priors)**.** *Theorems 1 and 2 extend to arbitrary integer Dirichlet priors* $\mathrm{Dir}(\beta_0, \beta_1)$ *on the mixing weight and* $\mathrm{Dir}(\alpha)$, $\mathrm{Dir}(\alpha')$ *on the two component parameter vectors: the evidence remains an explicit combination of* 1 *and harmonic numbers. The Gamma-ratio kernel keeps simple poles for integer hyperparameters, so no generalized harmonic numbers arise.*

Both proofs are in Appendix A, together with the general formula written out as an explicit procedure; as an example, for $k = 3$,

$$Z(2,2,2) \;=\; \frac{2}{315}\Big[\tfrac{1}{40} + \tfrac{1}{4}H_3 - \tfrac{2}{5}H_4 - \tfrac{5}{4}H_6 + \tfrac{12}{5}H_7 - H_8\Big] \;=\; \frac{37}{63000}. \tag{3.6}$$

To illustrate the use of the closed forms in model comparison, take counts $(u_0, u_1) = (34, 66)$ and ask whether the population is homogeneous or a two-component mixture. The one-component evidence is the conjugate integral $B(u_0{+}1, u_1{+}1)$, the mixture evidence is Eq. (3.2), and their ratio is the exact Bayes factor

$$\mathrm{BF}_{\text{ind:mix}} \;=\; \frac{B(35,67)}{Z(34,66)} \;=\; 0.78343\ldots, \tag{3.7}$$

so the data slightly favor the mixture. That the ratio is a rational number follows from Corollary 5.4 of Lin et al. (2009); here both evidences are closed forms in the counts, so the value of the ratio is obtained by direct substitution.

To our knowledge the formulas of Theorems 1–3 have not appeared before; the nearest antecedents are the unreduced finite sums of Evans et al. (1989), the allocation-sum representations of the exact-mixture literature, collapsed at most to per-component sufficient-statistic tables (Nobile, 1994; Fearnhead, 2005; Robert and Mengersen, 2011) (Section 4 discusses these), and the per-dataset rational values of Lin et al. (2009). However, the quantity these formulas express is not new: the evidences are moments of Dirichlet-weighted random means, computed in transform form in the literature on such means (Van Assche, 1987; Johnson and Kotz, 1990). The formula in Theorem 2 can be long: as an inclusion–exclusion over subsets of the k cells, its worst-case term count is $O(\min(k\,2^k, N{+}k))$ and individual instances exceed 2^k terms, so it may grow exponentially in k. Its length is nevertheless no measure of the cost of computing the value, which is obtained from the single sum Eq. (3.4) without expanding the formula (Section 4.2).

3.3 Convergence of the Bayes factor to a constant

The mild preference for the mixture at counts $(34, 66)$ persists at large sample sizes. In the balanced case, comparing Eq. (3.5) with the one-component evidence $B(n{+}1, n{+}1) = \sqrt{\pi/n}\;4^{-n}/2\,(1 + \mathcal{O}(1/n))$ gives the limit of the Bayes factor.

Corollary 1. *For balanced counts (n, n), the Bayes factor between the one-component model and its two-component mixture converges to*

$$\mathrm{BF}_{\text{ind:mix}} \;\longrightarrow\; \frac{1}{2\log 2} \;=\; 0.72134\ldots \tag{3.8}$$

PRELIMINARY

Proof. At counts (n, n) the one-component evidence is $B(n{+}1, n{+}1)$ and the mixture evidence is Eq. (3.3) of Theorem 1, so that

$$\mathrm{BF}_{\text{ind:mix}} \;=\; \frac{B(n{+}1, n{+}1)}{Z(n, n)} \;=\; \Big[\frac{1}{n+1} + 2\big(H_{2n+1} - H_{n+1}\big)\Big]^{-1} \tag{3.9}$$

at every n, and $H_{2n+1} - H_{n+1} \to \log 2$. □

Both models have $\lambda = 1/2$ here, so the Bayes factor Z_1/Z_2 stays bounded as the sample size grows and converges to a finite nonzero constant that mildly favors the larger model. The mixture is non-identifiable: one binary trial per subject is below Teicher's identifiability threshold (Teicher, 1963), so the mixture parameters cannot be recovered from the data law. Doob-style consistency therefore does not apply, and convergence of the Bayes factor to a constant is the expected behavior; Chib and Kuffner (2016) review the conditions under which the Bayes factor is instead consistent. The constant is a ratio of induced prior densities at the truth. The mixture prior induces on the success probability $w = \sigma\theta + (1-\sigma)\rho$ a density whose value at $w = \frac{1}{2}$ is $2\log 2$ (the elementary integral $\int_0^1 \max(\sigma, 1-\sigma)^{-1} d\sigma$), the one-component prior is uniform, and $1/(2\log 2)$ is the density ratio $\pi_1(\frac{1}{2})/\pi_2(\frac{1}{2})$. The convergence is visible at small samples: $\mathrm{BF} = 0.7265$ already at $(50, 50)$. Away from balance the Bayes factor also converges to a constant: comparing Eq. (3.2) with $B(u_0{+}1, u_1{+}1)$ on the ray $(qN, (1{-}q)N)$ gives $\mathrm{BF} \to 1/f(q)$ with $f(p) = -2[p\ln p + (1{-}p)\ln(1{-}p)]$, twice the binary entropy. The limiting Bayes factor thus favors the one-component model where $f(q) < 1$, i.e. for q outside approximately $(0.20, 0.80)$. Although convergence of a Bayes factor between overlapping models to a ratio of induced prior densities is expected in general, to our knowledge neither the constant $1/(2\log 2)$ nor the family $1/f(q)$ has appeared in print. Section 6.1 treats the case of four trials per subject, in which the Bayes factor grows without bound and the constant of Theorem 7 determines the sample size needed to reject a spurious component.

4 Any number of components, and the Dirichlet process limit

The closed forms of Section 3 are for two components, but for the one-variable family the evidence of the g-component mixture collapses, for every g, to a sum over the g component sizes weighted by contingency-table counts. The collapsed sum is cheap to evaluate, and the limit $g \to \infty$ at fixed concentration can be taken exactly, giving an explicit finite sum for the evidence of the Dirichlet process mixture at every finite sample size. For one unrestricted k-state variable the one-component model already covers the whole simplex Δ_{k-1} of cell-probability vectors, so the set of distributions in the g-component mixture model is that same simplex at every $g \geq 2$. On this family the evidences at different g, including the $g \to \infty$ limit, therefore differ only through the priors that the component structure induces on that simplex. A summation of the same kind, taken over set partitions within blocks, gives the partition-model posteriors used for record linkage (Section 4.4). The evidence integrals are also holonomic in the counts, and the two-component closed forms can be used in the proportional regime $N = \nu k$; the last subsection summarizes both properties, with the details in the supplement. In this section t denotes a number of occupied clusters, $m = (m_1, \ldots, m_g)$ the vector of component sizes, and $(x)_n = x(x+1)\cdots(x+n-1)$.

PRELIMINARY

4.1 The contingency-table collapse

Let $p_v = \sum_{a=1}^{g} \sigma_a \, \theta_v^{(a)}$, with mixing weights $\sigma \in \Delta_{g-1}$ and component parameters $\theta^{(a)} \in \Delta_{k-1}$, all carrying the uniform probability measures of Section 2.1.

Theorem 4 (General g)**.** *For the g-component mixture of a single k-state variable with uniform priors,*

$$Z_g(U) \;=\; \frac{\prod_v u_v! \; (g-1)! \; \big((k-1)!\big)^g}{(N+g-1)!} \sum_{\substack{m_1+\cdots+m_g=N \\ m_a \geq 0}} \mathrm{CT}_U(m) \prod_{a=1}^{g} \frac{m_a!}{(m_a+k-1)!} \,, \tag{4.1}$$

where $\mathrm{CT}_U(m)$ is the number of $k \times g$ nonnegative-integer tables with row sums U and column sums m, with generating function

$$\sum_{m} \mathrm{CT}_U(m) \; z_1^{m_1} \cdots z_g^{m_g} \;=\; \prod_{v=1}^{k} h_{u_v}(z_1, \ldots, z_g) \,, \tag{4.2}$$

h_j the complete homogeneous symmetric polynomial of degree j.

Proof. Expanding $\prod_v p_v^{u_v}$ by the multinomial theorem gives a sum over allocations of the u_v observations of each cell v among the components, that is, over $k \times g$ tables $x = (x_{va})$ with row sums U. By Eq. (2.6) the integral over component a contributes $(k-1)! \prod_v x_{va}! \,/ (m_a + k - 1)!$ with $m_a = \sum_v x_{va}$, the mixing integral contributes $(g-1)! \prod_a m_a! / (N+g-1)!$, and the multinomial coefficients contribute $\prod_v u_v! / \prod_a x_{va}!$, canceling the table factorials. The summand therefore depends on x only through its column sums, and the number of tables with margins (U, m) is $\mathrm{CT}_U(m)$. Equation (4.2) is the standard fact that products of complete homogeneous polynomials enumerate nonnegative-integer matrices with fixed margins (Stanley, 1999, §7.5). □

At $g = 2$, $\mathrm{CT}_U(m, N-m) = W(m)$ and Eq. (4.1) reduces to Eq. (3.4), from which Theorems 1–3 were derived. The two priors enter Eq. (4.1) asymmetrically. The mixing prior need not be uniform: any symmetric $\mathrm{Dir}(\beta_0, \ldots, \beta_0)$ on the weights replaces $\prod_a m_a!$ by $\prod_a (\beta_0)_{m_a}$ and the factor $(g-1)!/(N+g-1)!$ by $1/(g\beta_0)_N$. However, the uniform component priors are essential: for Dirichlet concentration $\beta \neq 1$ on the states, the component integral depends on the full table column instead of its sum alone, and no collapse to CT_U occurs. The restriction is likely to matter in practice, because a sensitivity analysis over the component prior is a standard part of a mixture-based report. Every general-g and Dirichlet-process value in this paper is computed with the uniform component prior, $\beta = 1$, and non-uniform component priors at general g would need a different method, possibly the tilting of Rousseau and Mengersen (2011).

The collapse has partial precedents in the literature on exact evidence computation by allocation sums. Nobile (1994) wrote the conjugate allocation sum for general mixtures and enumerated it exactly for a Gaussian mixture at $n = 8$. Fearnhead (2005) computed exact posteriors, and as a byproduct evidences, for discrete mixtures by recursions over allocation sufficient statistics. Robert and Mengersen (2011) collapsed the g^N allocations to the per-component sufficient-statistic tables, writing that, outside the cases treated there, the allocation sum "does not simplify into a smaller number of terms because there exists no summary statistics" (their Section 2.2). For the present family the summand depends on such a table only through its g column sums, so the sum collapses further. The same cancellation is implicit in Proposition 3.3(1) of Lin et al. (2009), which collapses

PRELIMINARY

allocations to the lattice points of a polytope; for the one-variable model their summand depends on the lattice point b only through $|b|$, the number of observations allocated to a component. In a contemporaneous note, Kim (2026) obtains the same margin collapse for a finite-support contamination model, applied there to a likelihood rather than to the evidence and without the table-count identity. To our knowledge, however, the evidence identity Eq. (4.1), which expresses the evidence of a g-component multinomial mixture as a margin sum over contingency-table counts for every g, has not been stated before.

4.2 Cost of evaluation

At $g = 2$ the margin counts $R_U(m) = \mathrm{CT}_U(m, N-m)$ are the coefficients of $\prod_v (1 + z + \cdots + z^{u_v})$, computable by prefix sums in at most $2k(N+1)$ big-integer additions, with operand bit lengths bounded by $k \log_2(1 + N/k)$; assembling the $N+1$ terms costs N big-rational additions. Evaluating the evidence therefore takes time polynomial in k and N, whereas the defining allocation sum Eq. (3.1) has $\prod_v (u_v+1)$ terms, about 10^{200} at $k = 100$, $N = 10^4$. However, the exact output is itself large: the reduced fraction at $k = 1000$, $N = 10^4$ has 129,674 bits. The same generating-function arithmetic has been used in linear-time algorithms for a different model-selection score, the normalized maximum likelihood of multinomial and naive-Bayes latent-class models (Kontkanen and Myllymäki, 2007; Mononen and Myllymäki, 2007), which replaces the evidence by the maximized likelihood divided by a non-Bayesian complexity term.

For general g we use one of two algorithms, according to whether g or k is small. A dynamic program over the k states costs $O(kN^{g-1})$ big-integer operations at fixed small g; alternatively, exchanging the roles of states and components gives

$$\sum_m \mathrm{CT}_U(m) \prod_{a=1}^{g} w(m_a) \;=\; [\xi^U]\, P(\xi)^g\,, \qquad P(\xi) \;=\; \sum_{0 \le c \le U} w(|c|)\, \xi^c\,, \tag{4.3}$$

for any weight function w, where P is a polynomial in k formal variables $\xi_1, \ldots, \xi_k$, one for each state, truncated componentwise at degree U, so that the cost is set by the number of retained coefficients, $\prod_v (u_v + 1) \le (1 + N/k)^k$. Taking the better of the two algorithms gives a cost governed by $\min(N^{g-1}, N^k)$: polynomial in N at fixed g, and nearly independent of g at small k. For $g \ge N$ there is a third option: components with no observations contribute trivially, and the sum can be rewritten over the integer partitions of N, with at most $p(N)$ terms independent of g.

Table 1 reports evaluation times on count vectors drawn as random compositions of N, balanced in the rows that use Eq. (4.3); the $g = 10^6$ rows were computed with Eq. (4.3) and the $g = \infty$ rows with the Dirichlet-process generating function Eq. (4.7) below. For comparison with a model outside the one-variable family of Theorem 4, Lin et al. (2009) report that their exact algorithms evaluated the five-cell binomial mixture at $N = 242$ in 45 s, and they estimated a running time of more than 13 days for $N = 5{,}000$ (their §4.4). For the one-variable family, the largest cases in Table 1 ($k = 1{,}000$ states with $N = 10^5$ observations, and $g = 10^6$ components or the Dirichlet process limit with $N = 5 \times 10^4$) run in minutes to a few hours.

4.3 The exact Dirichlet process limit

With the symmetric mixing prior $\mathrm{Dir}(\alpha/g, \ldots, \alpha/g)$ at fixed $\alpha > 0$, the g-component mixture tends as $g \to \infty$ to the Dirichlet process mixture with concentration α (Ferguson, 1973; Antoniak, 1974).

PRELIMINARY

g	k	N	time	$\ln Z$	arithmetic
2	100	10,000	0.42 s	-41336.614	exact rational
2	1,000	10,000	18 s	-66101.556	exact rational
2	1,000	100,000	5 min	-652834.747	exact rational
3	10	1,000	1.35 s	-2027.445	exact rational
3	3	3,000	11.2 s	-2718.593	exact rational
3	30	10,000	19.4 min		exact rational
4	5	40	0.84 s	-57.569	exact rational
10^6	2	5,000	31 s	-3465.738	floating point
10^6	2	50,000	8217 s	-34657.375	floating point
∞ ($\alpha = 1$)	2	5,000	79 s	-3469.221	floating point
∞ ($\alpha = 2$)	2	5,000	64.7 s	-3468.969	floating point
∞ ($\alpha = 1$)	2	50,000	9114 s	-34661.995	floating point

Table 1: Cost of evaluating the evidence for the g-component mixture of one k-state variable under uniform priors. Times are single-core wall times from single measurements. Exact-rational rows return Z as a fraction, whose logarithm is printed; one timing-only run has no $\ln Z$ entry. Floating-point rows use exponentially tilted arithmetic, checked against exact evaluation at smaller sizes (Supplement S3).

This finite-Dirichlet construction of the limit was studied by Neal (1992) (see also Neal, 2000); Green and Richardson (2001) state the convergence including the joint law with the data, crediting Neal, and Ishwaran and Zarepour (2000, 2002) developed the finite-Dirichlet approximation in detail. Given the finite-g partition law in these works, the new content of Theorem 5 is the polynomiality in $1/g$ and the collapse of the evidence sum to cluster sizes, both of which follow from that law by a short regrouping of terms. Substituting the moments of the $\mathrm{Dir}(\alpha/g, \ldots, \alpha/g)$ mixing prior into Eq. (4.1) gives

$$Z_g(U;\alpha) \;=\; \frac{\prod_v u_v!}{(\alpha)_N} \sum_{\substack{m_1+\cdots+m_g=N \\ m_a \geq 0}} \mathrm{CT}_U(m) \prod_{a=1}^{g} (\alpha/g)_{m_a} \, \frac{(k-1)!}{(m_a+k-1)!}\,, \tag{4.4}$$

which coincides with Eq. (4.1) at $\alpha = g$.

Theorem 5 (Exact Dirichlet process limit)**.** *$Z_g(U;\alpha)$ is a polynomial in $1/g$ of degree at most $N-1$. Its constant term is the finite-N evidence of the Dirichlet process mixture,*

$$Z_{\mathrm{DPM}}(U;\alpha) \;=\; \frac{\prod_v u_v!}{(\alpha)_N} \sum_{t=1}^{N} \frac{\alpha^t}{t!} \sum_{\substack{m_1+\cdots+m_t=N \\ m_i \geq 1}} \mathrm{CT}_U(m) \prod_{i=1}^{t} \frac{(m_i-1)!\,(k-1)!}{(m_i+k-1)!}\,, \tag{4.5}$$

the set-partition sum Eq. (2.9) *of the Chinese restaurant process with per-cluster Dirichlet-multinomial factors, collapsed to cluster sizes by the same table-count cancellation.*

The proof is a regrouping of the sum Eq. (4.4), in which an empty component contributes a factor of exactly one, identically in g. Grouping the weak compositions by their t-element support

PRELIMINARY

then makes each term an explicit polynomial in $1/g$ of total degree $N-1$, and the limit $g \to \infty$ is the substitution $1/g = 0$ in a finite, g-independent expression, so that no interchange of the limit with a sum or an integral has to be justified. The coefficient of $1/g$ has a closed form, again in terms of harmonic numbers:

$$\frac{Z_g}{Z_{\mathrm{DPM}}} \;=\; 1 \;+\; \frac{1}{g}\,E\Big[\,\alpha \sum_{B} H_{|B|-1} \;-\; \tfrac{1}{2}T(T-1)\Big] \;+\; \mathcal{O}(1/g^2)\,, \tag{4.6}$$

where the expectation is over the posterior on partitions of the data under the Dirichlet process mixture, B ranges over the blocks and T is their number. The sign of the $1/g$ term depends on the data (at $U = (2,1)$, $\alpha = 1$ it is $-\frac{1}{48g}$; at $U = (3,0)$ it is $+\frac{1}{16g}$), so finite-g evidences do not bound the limit from one side. For $g \geq N$ they are sandwiched, $Z_g/Z_{\mathrm{DPM}} \in [1 - N(N-1)/2g,\ e^{\alpha N/g}]$. The $1/g$ scaling of the concentration is essential: with a fixed-concentration prior $\mathrm{Dir}(c,\ldots,c)$ the evidence converges to the iid uniform value $(1/k)^N$ rather than to Eq. (4.5) (at $k=2$, $U=(3,1)$: $Z_g \to 0.0625$ against $Z_{\mathrm{DPM}}(\alpha{=}1) = \frac{337}{5760} = 0.0585069\ldots$).

Evaluating Eq. (4.5) does not require enumerating partitions, because the double sum over the number of clusters and their sizes is a single coefficient of an exponential generating function,

$$Z_{\mathrm{DPM}}(U;\alpha) \;=\; \frac{\prod_v u_v!}{(\alpha)_N}\,[\xi^U]\,\exp\left(\alpha \sum_{m\geq 1} \frac{(k-1)!\,(m-1)!}{(m+k-1)!}\,h_m(\xi_1,\ldots,\xi_k)\right), \tag{4.7}$$

with every polynomial truncated at U and the truncation exact, so the arithmetic of Section 4.2 applies verbatim. Equation (4.7) is the Cifarelli–Regazzini (Markov–Krein) transform of the mean of a Dirichlet process whose base measure is uniform on the simplex Δ_{k-1} (Cifarelli and Regazzini, 1990; Diaconis and Kemperman, 1996; Kerov and Tsilevich, 2001; Regazzini et al., 2002; James, 2005): $Z_{\mathrm{DPM}}(U;\alpha)$ is the U-th moment of that random mean, and at $k=2$, $\alpha=1$ a one-dimensional integral against the Diaconis–Kemperman density. The finite-N evidence of a Dirichlet process mixture over discrete data is thus computable by the same contingency-table dynamic programming as the finite mixtures; in Section 6.3 we evaluate it on a real benchmark at $N = 101$. Mena et al. (2025) give exact finite-mixture representations for the wider class of species-sampling processes, constructed for posterior sampling rather than for evidence computation.

4.4 Partition models under blocking

The record-linkage computations of Sections 6.2 and 6.4 use the same kind of summation, taken over set partitions instead of over allocations. There the latent object is a partition of records into entities, and the prior is an exchangeable partition probability function factorizing into a per-cluster term and a global term $g_K(K)$ in the number of clusters. The records are grouped into prespecified blocks, with no links across blocks. Under these hypotheses the likelihood-weighted sum over the partitions of each block is a polynomial in a formal variable whose exponent counts the clusters in the block; we compute it by enumerating those partitions. The posterior of the total number of entities K and the evidence then follow by multiplying the block polynomials (a convolution of their coefficient sequences) and weighting the coefficient of degree K by $g_K(K)$. Replacing the convolution by a per-block product is wrong for priors whose g_K is not a pure exponential. On the RLdata500 linkage benchmark (Section 6.2; Supplement S8) the product misstates the log evidence by $+103.39$ nats for the Pitman–Yor prior and -100.13 nats for the uniform-labelings prior, while it reproduces the uniform and Dirichlet-process values. Enumeration within blocks is

PRELIMINARY

cheaper than the partition counts suggest: the largest block of that benchmark has 14 records and Bell(14) = 190,899,322 partitions, enumerated in under a minute.

4.5 Holonomic structure and the proportional regime

A function of integer arguments is called holonomic if its generating function is D-finite, that is, satisfies in each variable a nontrivial linear differential equation with polynomial coefficients. Such a function satisfies, in each argument, a linear recurrence with polynomial coefficients, and for one argument the two properties are equivalent. Integrals of a product of powers $\prod_v p_v(\theta)^{U_v}$ over a semialgebraic parameter set, such as the evidence Eq. (2.3), are holonomic in the exponents (Borinsky et al., 2023), so that $Z(U)$ obeys count-shift relations of hypergeometric type, the contiguity relations. Where such relations can be written down, they give a second route to the exact evidence by iteration from small counts, as in the holonomic gradient method for contingency tables (Tachibana et al., 2020). A scalar recurrence in N along a data ray would also deliver the form of the large-sample expansion of Section 5 (rate, power of N and correction structure) through the analysis of its singular points. The constant would then follow to any precision from exact initial values.

For the binary model we derive the contiguity system explicitly (Supplement S3): the sum rule $Z(U) = Z(U + e_0) + Z(U + e_1)$ and a third-order recurrence in one count with cubic coefficients, whose iteration from small counts reproduces the values of Theorem 1. For the binomial mixture on its singular ray $U = n\,(1, 4, 6, 4, 1)$ we searched for a scalar recurrence in n and found none. No operator of order r and coefficient degree s in the tested region, which lies inside $(r+1)(s+1) \le 670$ and includes all degrees $s \le 140$ at orders $r \le 3$, annihilates the exact series through $n = 800$ (Supplement S3). Diagonal restrictions of multivariate holonomic systems can have very high scalar complexity, and this family appears to be an instance. The asymptotic analysis of Section 5 therefore starts from the summand of the allocation sum and does not use recurrence operators.

Separately from the holonomic structure, the closed forms hold at every k and N jointly and can therefore be evaluated in the proportional regime $N = \nu k$, in which the number of states grows with the sample size (Supplement S4). There the evidence density $f(\nu) = \lim_k [\ln Z + N \ln k]/k$, measured on two count shapes by extrapolation in k from exact rational values, falls below the plug-in optimum by $c \ln \nu$, with $c = 0.7498 \pm 0.0027$ on a boundary shape and $c = 0.506 \pm 0.018$ on an interior shape. The boundary value matches a boundary-corrected singular-learning density of $3/4$ derived there and lies far from the uncorrected value $1/2$; the interior value is consistent with $1/2$. On rays $k \sim N^\alpha$ the coefficient of $k \ln N$ falls linearly as $c\,(1 - \alpha)$, an empirical relation measured at three exponents that may not hold beyond them. At the fixed-k end the same measurement gives $\lambda(k) = (k - 1)/2$ for the interior shape, whereas the theorem of Watanabe and Watanabe (2022) for multinomial mixtures with at least two trials per observation would give $k/2$. The difference appears to be their component-splitting cost, a contribution to the learning coefficient from parameter directions along which the one-trial-per-observation likelihood considered here is constant, so that this contribution vanishes for the present model.

5 Asymptotic expansions with exact constants

We compute the $\mathcal{O}(1)$ term of Eq. (2.7) for three families whose learning coefficients are known: the binomial mixture, the naive Bayes network on three binary features, and the two-class mixture

PRELIMINARY

of independence models on 3×3 tables. All results are for deterministic expected-count data rays $U = Nq$ under the uniform priors of Section 2.1, except in Section 5.3, where two continuous models are treated under Gaussian priors with the log likelihood replaced by its expectation under the truth (the population ray of Eq. (5.10)). We work on these rays because there the $\mathcal{O}(1)$ term of Eq. (2.7) is a number; for sampled counts it is a random variable of order one, and its deterministic part is what we compute. In each family the evidence is dominated by a neighborhood of the set where the exponent is maximal. The dominant part of this set has one of two local geometries: an interior ridge with quartic transverse decay for the binomial mixture, and a boundary stratum with a quadratic edge for the latent-class models. Which of the two dominates depends on the power of N that each part of the maximal set contributes: a flat direction contributes no decay, each quadratic transverse direction a factor $N^{-1/2}$, and each quartic one a factor $N^{-1/4}$. With one parameter per component, as in the binomial mixture, the interior ridge is flat in the mixing weight and has one quadratic and one quartic transverse direction, giving $N^{-3/4}$. On the boundary stratum of that model only the parameter of the empty component is flat, while the weight and the other parameter are quadratic, giving N^{-1}. For the latent-class models of Section 5.4, with $d_1 \geq 3$ parameters per component, the order reverses: the boundary stratum leaves the whole parameter block of the empty component flat and gives $N^{-(d_1+1)/2}$, which decays more slowly than the contribution of any interior stratum. Section 5.4 proves this for the tables and supports it numerically for the naive Bayes model.

5.1 The interior ridge: the binomial mixture

This section computes the constant term of Eq. (2.7) for the binomial mixture of Section 2.1 (four trials, two components, uniform priors) when the data lie on the one-component submodel, $q = \mathrm{Bin}(4, \theta_0)$. Over such data the fiber of the mixture map is positive-dimensional: it contains every parameter with both components at θ_0, whatever the weight, and every parameter with one component at θ_0 carrying all the weight. This fiber is the singular locus responsible for $\lambda = 3/4$ in place of the regular value $d/2 = 3/2$. We take the symmetric point $\theta_0 = \frac{1}{2}$, that is, counts $U = n\,(1, 4, 6, 4, 1)$ with $N = 16n$, where the formulas are shortest, and Section 5.2 extends the result to every $\theta_0 \in (0, 1)$. No closed form is available on this ray. Theorem 6 below gives the evidence on this ray as an explicit constant times $n^{-3/4}\, 2^{-64n}$; the exponent $3/4$ was known (Yamazaki and Watanabe, 2004). We carry the binary model along as a check, since there Theorem 1 gives the answer in closed form.

For the binomial mixture the sufficient statistic of an allocation x is the pair (m, ℓ) with $m = |x|$ and $\ell = \sum_v v\, x_v$. Writing $\phi_n(m, \ell)$ for the count of allocations (weighted by the binomial coefficients of Eq. (3.1)) at fixed (m, ℓ) for data $U = n\,(1, 4, 6, 4, 1)$, Eq. (3.1) becomes

$$Z(n) \;=\; \sum_{m=0}^{16n} \sum_{\ell} \phi_n(m, \ell)\; w(m, \ell, n)\,, \qquad \phi_n(m, \ell) \;=\; [\zeta^m \eta^\ell]\; G(\zeta, \eta)^n\,, \tag{5.1}$$

with generating polynomial $G(\zeta, \eta) = (1 + \zeta)(1 + \zeta\eta)^4(1 + \zeta\eta^2)^6(1 + \zeta\eta^3)^4(1 + \zeta\eta^4)$ and weight

$$w(m, \ell, n) \;=\; \frac{m!\,(16n - m)!}{(16n + 1)!} \cdot \frac{(4m - \ell)!\,\ell!}{(4m + 1)!} \cdot \frac{(4(16n - m) - (32n - \ell))!\,(32n - \ell)!}{(4(16n - m) + 1)!}\,. \tag{5.2}$$

By the Legendre transform (large deviations of the coefficient array of G^n) and Stirling's formula,

$$\frac{1}{n} \log\left[\phi_n(nx, ny)\, w(nx, ny, n)\right] \;\longrightarrow\; E(x, y) \;=\; s(x, y) + \gamma(x, y)\,, \tag{5.3}$$

PRELIMINARY

with $s(x,y) = \inf_{\zeta,\eta>0} [\log G(\zeta,\eta) - x\log\zeta - y\log\eta]$, γ the explicit Stirling rate of Eq. (5.2), and (x,y) interior to the allocation domain. The evidence decays at the rate $\max E$, and the power of n multiplying the exponential is set by the geometry of E near its maximum. For a regular model the maximum would be a nondegenerate critical point; for the mixture it is degenerate, and the degenerate locus is the image of the singular fiber of the mixture map. In the allocation coordinates (x,y) the exponent E is already in normal-crossing form near its degenerate maximum, so no blowup is needed.

Lemma 1 (Ridge identity)**.** *Along the line $y = 2x$, $0 < x < 16$, the exponent is constant:*

$$E(x,2x) \;=\; -64\log 2\,. \tag{5.4}$$

The Legendre saddle sits at $(\zeta^*,\eta^*) = (\frac{x}{16-x},\,1)$, and the identity is exact: $s(x,2x) = 16\,h(x/16)$ while the Stirling brackets of γ telescope to $-64\log 2 - 16h(x/16)$.

Proof. At $\eta = 1$ the generating polynomial is $G(\zeta,1) = (1+\zeta)^{16}$, and the saddle equations $\zeta\,\partial_\zeta \log G = x$, $\eta\,\partial_\eta \log G = 2x$ hold at (ζ^*,η^*) identically in x. Then $s(x,2x) = 16\log(1+\zeta^*) - x\log\zeta^* = 16h(x/16)$. Along $y = 2x$ the sufficient statistics balance, $4m - \ell = \ell = 2nx$ and likewise in the second component, so each bracket of γ contributes a pure binomial rate, and these sum with the mixing bracket to $-64\log 2 - 16h(x/16)$. □

The ridge is the allocation-sum image of the singular fiber: under the condition $\ell = 2m$ the subsample allocated to component one has the same mean state as the full data, and this equality defines the locus on which the mixture degenerates to the independence model.

Lemma 2 (Quartic transverse decay)**.** *Transverse to the ridge, at fixed x,*

$$E(x,2x+\epsilon) \;=\; -64\log 2 \;-\; q_4(x)\,\epsilon^4 \;+\; \mathcal{O}(\epsilon^6)\,, \qquad q_4(x) \;=\; \frac{3}{x^2\,(16-x)^2}\,. \tag{5.5}$$

The quadratic and cubic terms vanish identically in x: the decay is quartic uniformly along the ridge.

The vanishing of the quadratic and cubic terms and the value of the quartic coefficient are polynomial identities in x, established by a symbolic computation (Supplement S1.2). An envelope simplification ($\partial_\epsilon s = -\log\eta(\epsilon)$ along the constraint, $\eta(\epsilon)$ being the saddle value of η) reduces that computation to three orders of the saddle expansion. The transverse Hessian vanishes because at the singular point the Fisher information of the mixing degrees of freedom degenerates, so the leading transverse behavior of the likelihood is quartic. The same degeneracy produces $\lambda = 3/4$ in the resolution-of-singularities computation of Yamazaki and Watanabe (2004); in Lemma 2 the degeneracy is visible directly, as the vanishing of the second and third transverse derivatives of the explicit function E.

The asymptotics of Eq. (5.1) come from a tube around the ridge, and the constant is the product of three factors, which we first describe informally. Summing along the ridge, a one-dimensional lattice of about $16n$ points, gives a Riemann sum and hence a factor $n\int dx$. At fixed $x = m/n$, the transverse sum of $e^{-q_4(x)j^4/n^3}$ over integer offsets $j = \ell - 2m$ contributes a factor $n^{3/4}\cdot 2\Gamma(5/4)\,q_4(x)^{-1/4}$. The pointwise prefactor $\hat{p}(x) = \sqrt{2\pi}\,n^{-5/2}/[64\,x(16-x)]$ has two sources: the Gaussian normalization of ϕ_n, whose values are point probabilities of an exponentially tilted random walk on $\mathbb{Z}^2$ (local limit theorem and step covariance in Supplement S1.2), and the Stirling prefactors of the three Beta brackets of w. Multiplying and using $\int_0^{16} dx/\sqrt{x(16-x)} = \pi$ gives the constant of the following theorem.

PRELIMINARY

Theorem 6 (Binomial mixture at the singular point)**.** *For data $U = n\,(1,4,6,4,1)$, the evidence of the binomial mixture with uniform priors satisfies*

$$Z \;=\; C\, n^{-3/4}\, 2^{-64n}\,(1+o(1))\,, \qquad C \;=\; \frac{\sqrt{2}\,\pi^{3/2}\,\Gamma(5/4)}{32\cdot 3^{1/4}} \;=\; 0.1694847389\ldots \tag{5.6}$$

The proof is in Supplement S1.2. The learning coefficient $\lambda = 3/4$ agrees with the resolution-of-singularities value (Yamazaki and Watanabe, 2004) and with the numerical observation of Lin et al. (2009); the constant is, to our knowledge, new. Because the transverse decay is quartic uniformly along the ridge, the multiplicity is $m = 1$ in Eq. (2.7) and no $\log\log$ corrections arise. The proof bounds the corrections by $\mathcal{O}(n^{-1/4})$ on compact ridge segments, and Table 2 indicates that $n^{-1/4}$ is their actual order. Beyond the two lemmas, the assembly rests on four localization estimates, all proved in Supplement S1.2 (classical background in Bhattacharya and Ranga Rao, 1976; Dembo and Zeitouni, 1998). The first is off-ridge domination, the non-asymptotic bound $\phi_n w\, 2^{64n} \le \exp\{-(\ell-2m)^4/(131072\,n^3)\}$, valid at every lattice point (m,ℓ) for every n. The others are a local limit theorem for the tilted lattice walk that is uniform on compact ridge segments, convergence of the transverse sum on such segments with relative error $\mathcal{O}(n^{-1/4})$, and direct bounds on the lattice rows at fixed m near the two ends of the ridge. For every $\kappa \in (0,8)$, the rows with $m < \kappa n$ or $m > (16-\kappa)n$ contribute at most a constant times $\sqrt{\kappa}\,n^{-3/4} + n^{-1}$ to $2^{64n}Z(n)$. Here $\sqrt{\kappa}$ is the order of the arcsine tail $\int_0^\kappa dx/\sqrt{x(16-x)}$ of the ridge integral, so that the ends enter the leading term only through the integrable density $dx/\sqrt{x(16-x)}$. The same assembly applied to the binary model, where the exponent is constant on the whole two-dimensional allocation domain, reproduces the constant $\sqrt{\pi}\log 2$ of Eq. (3.5) obtained from the closed form. The power of n follows from a simple count: in this family,

$$\text{power of } n \;=\; -\tfrac{5}{2} \;+\; \dim(\text{flat locus}) \;+\; \sum_{\text{transverse}} \left(1 - \tfrac{1}{\text{order}}\right), \tag{5.7}$$

giving $-\frac{1}{2}$ for the binary model (flat plane, no transverse directions) and $-\frac{3}{4}$ for the binomial mixture (flat line, one quartic direction). We state Eq. (5.7) for the models computed here; it may hold for a wider class of singular models, a question we leave open (Section 7).

5.2 The constant on the singular stratum

Although Theorem 6 is stated at the symmetric point, nothing in the ridge computation is special to it. The singular stratum of the binomial mixture is the whole independence locus: for any $\theta_0 \in (0,1)$, data at $\mathrm{Bin}(4,\theta_0)$ lie over a degenerate fiber of the mixture map, and every ingredient generalizes with the same proofs. For rational $\theta_0 = a/b$ (lowest terms, $c = b-a$) and integer data rays $U = \frac{N}{b^4}\left(\binom{4}{v}a^{4-v}c^v\right)_v$, the exponent is flat along the ridge $y = (4c/b)x$, the quadratic and cubic transverse terms vanish identically in (a,b,x), and the quartic coefficient is $q_4(x;\theta_0) = 3b^8/[256\,a^2c^2\,x^2(b^4-x)^2]$, all symbolic identities. We state the result in the sample size N; the subscript on C_N records the normalization, and the constant does not depend on N.

Theorem 7 (The constant as a function on the singular stratum)**.** *For rational $\theta_0 \in (0,1)$ and data $U = \frac{N}{b^4}\left(\binom{4}{v}a^{4-v}c^v\right)_v$,*

$$Z \;=\; C_N(\theta_0)\; N^{-3/4}\; e^{\,4N[\theta_0\log\theta_0 + (1-\theta_0)\log(1-\theta_0)]}\,(1+o(1))\,, \tag{5.8}$$

$$\boxed{C_N(\theta_0) \;=\; \frac{\sqrt{2}\;3^{3/4}\,\pi^{3/2}\,\Gamma(1/4)}{12}\;\theta_0\,(1-\theta_0) \;=\; 5.4235116\ldots \times \theta_0(1-\theta_0)\,.} \tag{5.9}$$

PRELIMINARY

The learning coefficient is 3/4 uniformly on the interior of the stratum, and the constant factorizes into a universal number times the variance of the underlying binomial. All representation dependence (the denominator b) cancels once the expansion is written in N, and Theorem 6 is the case $\theta_0 = \frac{1}{2}$. Numerically, exact rational values at $\theta_0 = 1/3$ and log-stabilized evaluations up to $N = 1944$ approach the predicted constant with the expected $N^{-1/4}$ and $N^{-1/2}$ corrections. Lin obtained a data-dependent constant of this kind for a coin-mixture model on its data strata (Lin, 2011, Example 1.18) within the Newton-nondegenerate calculus, which does not cover the present model.

Theorem 6 also determines the additive constant in the scaling benchmark of Lin et al. (2009), who tested Watanabe's prediction by computing $F_N = N \sum_i q_i \log_{10} q_i - \log_{10} I_N$ on this data direction and checking that consecutive differences track $\frac{3}{4}$; the differences cancel the additive constant, so their test measured the exponent alone. In their normalization convention (bare integral times column multiplicities), Theorem 6 is equivalent to $F_N = \frac{3}{4} \log_{10} N + c_0 + o(1)$ with $c_0 = -\frac{3}{4} \log_{10} 16 - \log_{10} C = -0.132220\ldots$. Table 2 extends their table by a factor of sixteen in sample size; over that range the residuals decay approximately like $N^{-1/4}$ and the convergence to c_0 appears unambiguous.

N	F_N	$\frac{3}{4}\log_{10} N + c_0$	$F_N - (\frac{3}{4}\log_{10} N + c_0)$
16	0.8593832	0.770869	0.0885
32	1.0696536	0.996642	0.0730
64	1.2849713	1.222414	0.0626
96	1.4122027	1.354483	0.0577
384	1.8506364	1.806028	0.0446
1024	2.1625779	2.125504	0.0371
2048	2.3836327	2.351277	0.0324

Table 2: The scaling benchmark F_N of Lin et al. (2009) against $\frac{3}{4}\log_{10} N + c_0$ with the constant c_0 implied by Theorem 6. The published differences of Lin et al. (2009) end at $N = 112$. Rows above the rule use exact rational evidence values; rows below use log-stabilized floating-point evaluation (Supplement S3). The last column decays roughly as $N^{-1/4}$, the order of the first correction.

5.3 The Gaussian location mixture and reduced-rank regression

The ridge assembly of Section 5.1 needs only a flat locus, the transverse decay orders, and a density on the fiber, and all three exist for continuous models. The continuous analog of the deterministic count ray is the population ray, obtained by replacing the log likelihood with its expectation under the truth:

$$Z(n) \;=\; \int d\pi(\theta)\; e^{\,nL(\theta)}\,, \qquad L(\theta) \;=\; \int q(y)\, \log p_\theta(y)\, dy\,, \tag{5.10}$$

with q the true density and n the sample size. This deterministic integral was introduced by Watanabe (1999); it bounds the expected free energy by Jensen's inequality, a bound going back to Barron (1987), and analyzing it before the sampled problem is standard in the singular-learning literature (Bhattacharya et al., 2026). The leading-coefficient calculus of Lin (2011) gives its constant-level asymptotics in general form under the Newton-nondegeneracy hypothesis. No constant-order term

PRELIMINARY

of the evidence at a fixed prior seems to have been computed for either of the two models treated below, the two-component Gaussian mixture and reduced-rank regression. Existing results in this direction are an existence statement (Chen et al., 2025), stochastic limit laws under $n^{-1/2}$-shrinking priors (Kariya and Watanabe, 2022), a variational analysis with a stochastic constant (Kariya and Watanabe, 2020), and the exact evidence of a related restricted model, deep linear networks with one output and zero noise (Hanin and Zlokapa, 2023).

Consider the two-component Gaussian mixture $p_\theta(y) = \alpha\, N(y;\mu_1,1)+(1-\alpha)\, N(y;\mu_2,1)$ against the one-component model $N(y;\mu,1)$, with α uniform, means iid $N(0,s_0^2)$, and truth $N(\mu_0,1)$. In fiber-adapted coordinates $w = \alpha t_1 + (1-\alpha)t_2$, $s = t_1 - t_2$, $t_h = \mu_h - \mu_0$, the population Kullback–Leibler divergence has the fourth-order normal form

$$K \;=\; \frac{w^2}{2} \;-\; \frac{\alpha(1-\alpha)}{2}\, w^2 s^2 \;+\; \frac{\alpha^2(1-\alpha)^2}{4}\, s^4 \;+\; \mathcal{O}\big((|w|+|s|)^5\big)\,. \tag{5.11}$$

The divergence K has the structure used in the ridge computation of Section 5.1: flat along the fiber, quadratic in one transverse direction, and quartic in the other, with the quadratic and cubic terms vanishing identically in α. The counting rule Eq. (5.7), in continuous form, gives $\lambda = 3/4$ with $m = 1$ on this ray, the value of Aoyagi (2010) for this family (see also Drton and Plummer, 2017, Example 2.1). The Bayes factor for one component against two then has an explicit constant.

Proposition 1 (Gaussian location mixture on the population ray)**.** *Let $Z_1(n)$ and $Z_2(n)$ be the population-ray evidences Eq.* (5.10) *of the one- and two-component models above, with truth $q = N(\mu_0,1)$, α uniform on $[0,1]$, the means μ, μ_1, μ_2 iid $N(0,s_0^2)$ and independent of α, and $L_0 = \int q \log q$. For fixed $s_0 > 0$ and $\mu_0 \in \mathbb{R}$, as $n \to \infty$, $n^{3/4}\, e^{-nL_0} Z_2(n) = \pi^{3/2}\,\Gamma(\frac{1}{4})\,\varphi(\mu_0)^2\,(1 + \mathcal{O}(n^{-1/4}))$, where φ is the $N(0,s_0^2)$ density, and*

$$\mathrm{BF}(n) \;=\; \frac{Z_1(n)}{Z_2(n)} \;=\; c(s_0,\mu_0)\; n^{1/4}\,\big(1 + \mathcal{O}(n^{-1/4})\big)\,, \qquad c(s_0,\mu_0) \;=\; \frac{2\, s_0\, e^{\mu_0^2/(2s_0^2)}}{\sqrt{\pi}\,\Gamma(1/4)}\,. \tag{5.12}$$

The proof is in Supplement S1.3. This Bayes factor parallels Corollary 2 below: the rate $n^{1/4}$ and the factor $\Gamma(1/4)$ from the quartic direction are the same, and the truth enters only through the prior density at μ_0, which takes the place of the binomial variance $\theta_0(1-\theta_0)$ in Eq. (5.9). A direct quadrature of Eq. (5.10) on a doubling sequence in n, with exponent and constant fitted by Richardson extrapolation, agrees with Eq. (5.12) in four prior/truth configurations (Table 3). The bound $\mathcal{O}(n^{-1/4})$ on the relative correction in Eq. (5.12) comes from the ends of the fiber, where the quartic width $(\alpha(1-\alpha))^{-1/2} n^{-1/4}$ is no longer small, and from the neighborhoods of the boundary strata $\alpha \in \{0,1\}$. The proof bounds the correction without determining its coefficient. Direct quadrature indicates that the coefficient is nonzero (Supplement S1.3), so that the corrections appear to begin at relative order $n^{-1/4}$, as on the lattice. Section 6.1 gives the sample sizes at which this Bayes factor reaches specified odds.

Reduced-rank regression gives a second instance, with nontrivial multiplicity. Take $y = BAx + \varepsilon$ with two-dimensional input and output, a rank-one factorization, standard normal noise, design, and priors, and truth $BA = 0$. The population-ray evidence then has the closed form

$$\bar{Z}(n) \;=\; \frac{1}{2n}\, e^{1/(2n)}\, E_1\Big(\frac{1}{2n}\Big) \;=\; \frac{\frac{1}{2}\ln n + \frac{1}{2}\big(\ln 2 - \gamma_E\big)}{n} \;+\; \mathcal{O}\Big(\frac{\ln n}{n^2}\Big)\,, \tag{5.13}$$

with γ_E Euler's constant. The exponents are the pair $(\lambda, m) = (1,2)$ of Aoyagi and Watanabe (2005) for these dimensions, and both constants are explicit: the coefficient $c_1 = \frac{1}{2}$ of $\ln n$ and

PRELIMINARY

s_0	μ_0	c derived	c fitted	pull
1	0	0.311225	0.311224 ± 0.000051	$+0.005\,\sigma$
4	0	1.244899	1.244889 ± 0.000264	$+0.04\,\sigma$
1	$\frac{1}{2}$	0.352664	0.352663 ± 0.000053	$+0.01\,\sigma$
1	1	0.513123	0.513119 ± 0.000052	$+0.07\,\sigma$

Table 3: The constant $c(s_0, \mu_0)$ of Eq. (5.12) against a numerical fit of the population-ray Bayes factor that does not use the formula, for four configurations of prior width s_0 and truth location μ_0. The pull is the difference between the two in units of the fit uncertainty. The fitted exponent agrees with 1/4 within its uncertainty in every configuration.

$c_0 = \frac{1}{2}(\ln 2 - \gamma_E) = 0.0579657\ldots$. A fit of the same kind, run without the closed form, recovers $\lambda = 1$, $m = 2$, and both constants.

The normal form Eq. (5.11) is the degree-five Taylor polynomial of K at the fiber, derived in Supplement S1.3; the E_1 reduction in Eq. (5.13) is an identity obtained by symbolic computation, and the expansion in Eq. (5.13) follows from the closed form by expanding E_1 at small argument. Supplement S1.3 proves the Laplace localization behind Proposition 1. In $e^{-nL_0} Z_2(n)$, the region away from the zero set of K contributes an exponentially small amount and the neighborhoods of the boundary strata $\alpha \in \{0, 1\}$ contribute $\mathcal{O}(n^{-1})$, against order $n^{-3/4}$ from the fiber. Near the fiber, a lower bound $K \geq c_0(w^2/2 + \alpha^2(1-\alpha)^2 s^4/4)$ with a constant $c_0 > 0$, uniform in $\alpha \in [0, 1]$, gives an n-free integrable envelope after rescaling w by $n^{-1/2}$ and s by $n^{-1/4}/\sqrt{\alpha(1-\alpha)}$. The ends of the fiber then enter the leading term only through the integrable density $d\alpha/\sqrt{\alpha(1-\alpha)}$. Proposition 1 and Eq. (5.13) are population-ray statements under the stated priors, and the corresponding sampled-data constants remain open for both models.

5.4 Boundary strata: latent-class models

This section treats two multivariate mixtures of independence models of the form Eq. (2.5) and computes the constant term of Eq. (2.7) for each when the data lie on the one-component submodel, the analog of the rays $\mathrm{Bin}(4, \theta_0)$ of Section 5.2. The first is the two-class naive Bayes model on F binary features: the cells are $v \in \{0, 1\}^F$, one class has cell probabilities $p_v(\omega) = \prod_{f=1}^{F} \theta_f^{v_f} (1 - \theta_f)^{1-v_f}$ with $\omega = (\theta_1, \ldots, \theta_F) \in [0, 1]^F$, and the two-class model has cell probabilities $M_v = \sigma\, p_v(\omega_1) + (1 - \sigma)\, p_v(\omega_2)$ with σ, ω_1 and ω_2 independent and uniform. The second is the two-class mixture of independence models on a $k \times k$ table: the cells are the pairs (i, j), one class has $p_{ij}(\omega) = a_i b_j$ with row and column probability vectors $\omega = (a, b) \in \Delta_{k-1} \times \Delta_{k-1}$, and M_{ij} mixes two such classes with weight σ, under the measure convention of Section 2.1. In both models the data ray is $U_v = N q_v$ with $q = p(\omega^*)$ for an interior one-class parameter ω^*, and Z_N denotes the parameter integral of $\prod_v M_v^{U_v}$, with no multinomial coefficient. The learning coefficients at such data are known: $(F + 1)/2$ for the naive Bayes model with $F \geq 3$ (Rusakov and Geiger, 2005) and 5/2 for 3×3 tables (Lin, 2017, 2011). Theorems 9 and 10 give the constant terms at the uniform points, $q_v = 2^{-F}$ with $F = 3$ and $q_{ij} = 1/9$. At such data the dominant contribution to the evidence comes from the boundary of the mixing simplex, where one class is empty and the transverse decay is quadratic, and the constants are obtained in closed or near-closed form.

Lemma 3 below, which gives the local form of the exponent at the boundary of the mixing simplex, holds for an arbitrary parametric component family $\{p_v(\omega)\}$ on a finite cell set. Let

PRELIMINARY

$M_v = \sigma p_v(\omega_1) + (1-\sigma)p_v(\omega_2)$ be its two-component mixture, and suppose the data direction is realizable by a single component, $q_v = p_v(\omega^*)$ with ω^* interior. The evidence exponent is $\Phi = \sum_v q_v \ln M_v$, whose maximum value is $-E = \sum_v q_v \ln q_v$.

Lemma 3 (Exact flatness of the mixing edge)**.** *On the stratum $S_0 = \{\sigma = 0,\ \omega_2 = \omega^*\}$, for every value of the free block ω_1: (i) $\Phi = -E$; (ii) $\nabla\Phi = 0$, including $\partial_\sigma\Phi = \sum_v q_v\,[p_v(\omega_1) - q_v]/q_v = 1 - 1 = 0$; (iii) the full Hessian of $-\Phi$ is block diagonal, vanishing identically on the ω_1 block and the mixed blocks, positive semidefinite on the transverse block (σ, ω_2).*

Proof. (i) is immediate. For (ii), the ω_2-gradient is the score of the true model at the truth, and the σ-derivative is the sum-to-one identity displayed. For (iii), Φ depends on ω_1 only through $\sigma\, p_v(\omega_1)$, so every ω_1-derivative carries a factor σ; the mixed $\partial_{\omega_1}\partial_\sigma$ entry is $\sum_v \partial_{\omega_1} p_v = \partial_{\omega_1}(1) = 0$. □

The first consequence of Lemma 3 is that the edge at $\sigma = 0$ is exactly quadratic, with no linear term in σ for any ω_1 and hence no boundary-layer cubic correction. The second is that the flat directions are the entire parameter block ω_1 of the vanished component, the largest flat set on the critical variety. Laplace localization then gives, summing the strata $\sigma \in \{0, 1\}$,

$$Z_N \;=\; e^{-NE}\Big(\frac{2\pi}{N}\Big)^{d_t/2} \int_{\mathcal{F}} \frac{d\omega_1}{\sqrt{\det H(\omega_1)}}\,(1 + o(1))\,, \tag{5.14}$$

with $\mathcal{F}$ the flat block, $d_t = 1 + \dim\omega_2$ the transverse dimension, and H the transverse Hessian of $-\Phi$ on S_0. Lemma 3 is, for the evidence integral, the counterpart of the classical boundary-χ^2 behavior of mixture likelihoods at vanishing weight (Hartigan, 1985; Dacunha-Castelle and Gassiat, 1997). The determinant in Equation (5.14) can be evaluated in closed form by the following theorem, proved in Supplement S2: for any two-component mixture of product distributions over a product cell set, the transverse Hessian factorizes into χ^2-divergences.

Theorem 8 (χ^2-factorization for bipartite latent-class models)**.** *Let each of F factors carry a saturated multinomial family, let $p_v(\omega) = \prod_f p_{vf}^{(f)}(\theta_f)$ be the product family, M_v its two-component mixture, and let the data direction be $q = p(\omega^*)$ with ω^* interior. On the stratum S_0, the transverse Hessian of the q-normalized exponent satisfies*

$$\det H \;=\; \Big[\sum_{j=2}^{F} e_j(\chi_1^2, \ldots, \chi_F^2)\Big] \Big/ \prod_f \prod_{s \in A_f} q_s^{(f)}\,, \tag{5.15}$$

where χ_f^2 is the χ^2-divergence of the f-th marginal of the flat component against the true f-th marginal and e_j the elementary symmetric polynomials.

Per-factor saturation is essential: the proof uses the fact that the scores of a full multinomial family span all mean-zero functions on its alphabet, which fails for a curved factor family. The alphabets of the factors may differ, however, so Eq. (5.15) gives the Hessian integrand of Eq. (5.14) in closed form for every bipartite latent-class model, while the flat integral and the dominance of the boundary stratum must be established model by model. The boundary dominates here because of the powers of N in Eq. (5.14): at these models the interior critical strata appear numerically to have quadratic transverse decay, which would make them subdominant to the boundary, whose flat block is larger.

PRELIMINARY

The first of the two families we work out is the two-class naive Bayes network on F binary features, the model of Rusakov and Geiger (2005), whose Theorem 4(c) gives $\lambda = (F+1)/2$ at independent data for $F \geq 3$; their method does not produce the constant. Theorem 8 specializes to $\det H = 2^{F(F+3)} \sum_{j=2}^{F} e_j(u_1, \ldots, u_F)$ with $u_f = (2\theta_f - 1)^2$ at uniform data, and three substitutions reduce the flat integral to a one-dimensional integral:

Theorem 9 ($2 \times 2 \times 2$ naive Bayes at uniform data)**.** *For $F = 3$ at the uniform ray $U_v = N/8$,*

$$Z_N \;=\; \frac{\pi^2 J}{2}\, N^{-2}\, 8^{-N}(1 + o(1))\,, \qquad J = \frac{1}{8\sqrt{\pi}} \int_0^\infty s^{-1/2} e^{-s} E_1(s)^3\, ds = 5.092204308\ldots\,, \quad (5.16)$$

where $E_1(s) = \int_s^\infty t^{-1} e^{-t}\, dt$ is the exponential integral.

The number J does not appear to reduce to standard constants, and we take the one-dimensional E_1^3 integral as its definition. As in Theorem 7, the constant extends to an explicit function on the full singular stratum: at any interior product data $q_v = \prod_f (\theta_f^*)^{v_f} (1 - \theta_f^*)^{1 - v_f}$,

$$C(\theta^*) = (2\pi)^2 \prod_f \theta_f^*(1 - \theta_f^*) \cdot \frac{1}{8\sqrt{\pi}} \int_0^\infty \frac{e^{-s}}{\sqrt{s}} \prod_{f=1}^{3} \Big[E_1(s\, r_f) + E_1(s/r_f) \Big] ds\,, \qquad r_f = \frac{1 - \theta_f^*}{\theta_f^*} \quad (5.17)$$

(proof in Supplement S2), with the prefactor $\prod_f \theta_f^*(1-\theta_f^*)$ the analog of the binomial variance $\theta_0(1-\theta_0)$ in Eq. (5.9). For general $F \geq 3$ the same computation gives $Z_N = C_F\, N^{-(F+1)/2}\, 2^{-FN}(1 + o(1))$ with $C_F = (2\pi)^{(F+1)/2}\, 2^{-F} J_F$ and J_F an explicit F-dimensional integral ($J_3 = J$, and $J_4 \approx 2.118$ by quadrature of the integral). At $F = 2$ the integral J_F diverges logarithmically on both axes. With the integral cut off where the transverse Gaussian approximation fails, each axis contributes $\frac{1}{2} \ln N$, which reproduces the anomalous $\lambda = 3/2$, $m = 3$ behavior that Rusakov and Geiger (2005) prove for the degenerate two-feature case by a separate argument. This agreement suggests reading the multiplicity as the divergence order of the integral J_F, although the cutoff argument is heuristic.

The second family is the two-class mixture of independence models on $k \times k$ tables, studied in the table examples of Lin et al. (2009). For $k = 3$, Lin (2017, 2011) proved by fiber-ideal blowups that at positive rank-one data $(\lambda, m) = (5/2, 1)$, noting that computing the constant would sharpen the result. Lemma 3 applies to the rank-one tables $q = \rho \otimes \kappa$, with flat block $(a, b) \in \Delta_{k-1}^2$ and transverse dimension $2k - 1$, and Theorem 8 gives $\det H = \chi^2(a\|\rho)\, \chi^2(b\|\kappa) / [\prod_i \rho_i \prod_j \kappa_j]$. In this family the priors entering Z are the coordinate Lebesgue measures on each simplex; normalizing them to total mass one multiplies every constant by 16 at $k = 3$.

Theorem 10 (3×3 tables at uniform data)**.** *At the uniform ray $U_{ij} = N/9$,*

$$\boxed{Z_N \;=\; C_u\, N^{-5/2}\, 9^{-N}(1 + o(1))\,, \qquad C_u = \frac{(2\pi)^{5/2} \cdot 2 \ln^2(2 + \sqrt{3})}{81} = 4.23778022\ldots} \quad (5.18)$$

The one ingredient beyond the Hessian factorization is the flat integral $\int_{\Delta_2} da/|a-u| = \sqrt{2}\, \ln(2+\sqrt{3})$, an inverse-distance integral over the triangle about its barycenter, evaluated in closed form in Supplement S2. At any positive rank-one data the constant is again an explicit function on Lin's stratum,

$$C(\rho, \kappa) = (2\pi)^{5/2} \sqrt{\prod_i \rho_i \prod_j \kappa_j}\; I(\rho)\, I(\kappa)\,, \qquad I(\rho) = \int_{\Delta_2} \frac{da}{\sqrt{\chi^2(a\|\rho)}}\,, \quad (5.19)$$

PRELIMINARY

with $I(\rho)$ elementary for every ρ (Supplement S2). For general $k \geq 3$ the boundary strata contribute $C_k N^{-(2k-1)/2} k^{-2N}$ with C_k explicit, and they dominate: every non-boundary stratum enters at $O(N^{-3(k-1)/2} \ln N)$, a gap of $(k-2)/2$ powers, and an elementary Pinsker–volume argument gives the log-free bound $Z_N e^{NE} \leq B_k N^{-(2k-1)/2}$ at all N. Hence $\lambda = (2k-1)/2$ and $m = 1$ for every $k \geq 3$, recovering Lin's theorem at $k = 3$. The learning coefficient λ is also known exactly for several related models: Zwiernik (2011) treats binary latent trees, Watanabe and Watanabe (2022) two-component multinomial mixtures of a single feature, and Hayashi (2021) stochastic matrix factorization under a column-stochastic normalization. Their learning coefficients do not transfer to the joint table model, and none of these works computes a constant.

The ingredients above that hold as identities are Lemma 3, the Hessian factorizations (proved for all sizes), and the closed-form flat integrals. The Laplace localization at the boundary stratum is proved by a dominated-convergence argument with an N-free integrable envelope, valid up to the degenerate corners precisely because $F \geq 3$, resp. $k \geq 3$ (Supplement S2). Subdominance of the remaining critical strata is proved for the tables by the global bound above; for naive Bayes it depends on their transverse decay orders, known so far only from numerical computation (see the discussion after Theorem 8). We also evaluated the evidence numerically at five data rays, from exact rational values at small N and log-space dynamic programming up to $N = 384$ for naive Bayes and $N = 216$ for the tables (Supplement S2). The naive Bayes sequences approach their constant at the few-percent level with $N^{-1/2}$ corrections. However, the table sequences have not reached the asymptotic regime by $N = 216$, where the residual in $\ln Z$ is still rising, so for the tables the constant rests on the localization proof. The computed table values are consistent with the correction class $(c_1 \ln N + c_2) N^{-1/2}$ expected from the non-boundary strata described above.

6 Applications

The applications range widely in problem size: two components and millions of observations in the design question below; ten components, ninety-six channels and five observations for mutational signatures; partitions of tens of thousands of latent entities in record linkage. Three of the applications are developed in this section: the sample size at which Bayesian model selection can reject a spurious mixture component (Section 6.1), a set of reference values for the approximate marginal-likelihood methods in common use (Section 6.2), and the evidence of a Dirichlet process mixture on a benchmark dataset (Section 6.3). Section 6.4 summarizes six further applications whose full analyses are in the supplement (Table 6).

6.1 Sample size to reject a spurious component

Consider a trial that records a yes-or-no outcome four times for each participant. Model selection must decide whether the participants form two cohorts, responders and non-responders with different response rates, or a single population whose apparent heterogeneity is sampling noise. The constant of Theorem 7 determines how much data that decision needs when the truth is a single population. Take data at $\mathrm{Bin}(4, \theta_0)$ and compare the independence model against its two-component mixture. The evidence of the smaller model is a conjugate Beta integral and the evidence of the mixture is given by Theorem 7. The exponential rates cancel, the polynomial factors are $N^{-1/2}$ for the smaller model and $N^{-3/4}$ for the mixture, and the constants combine into a closed form.

PRELIMINARY

Corollary 2. *For data at* $\mathrm{Bin}(4, \theta_0)$ *with sample size* N*,*

$$\mathrm{BF}_{\text{ind:mix}} \;=\; \frac{2 \cdot 3^{1/4}}{\pi\,\Gamma(1/4)} \;\cdot\; \frac{N^{1/4}}{\sqrt{\theta_0(1-\theta_0)}}\,\bigl(1 + o(1)\bigr)\,, \qquad \frac{2 \cdot 3^{1/4}}{\pi\,\Gamma(1/4)} \;=\; 0.231089\ldots \tag{6.1}$$

The Bayes factor grows, so the true model is eventually favored, unlike the binary case of Corollary 1. However, it grows only as $N^{1/4}$, with a known coefficient. Equation (6.1) holds for expected-count data, in which every cell count equals its expectation under the one-component model. For multinomial samples the $N^{1/4}$ rate is unchanged, but the $\mathcal{O}(1)$ term of $\log Z$, and with it the multiplicative constant, fluctuates by order one from sample to sample (Watanabe, 2009). For a particular multinomial sample the Bayes factor may therefore cross a given evidence threshold at an N that differs by a factor of that order from the sample sizes below, which should be read as design values for expected-count data.

On the scale of Kass and Raftery (1995), inverting the asymptotic law at $\theta_0 = \frac{1}{2}$ gives, as leading-order targets, $N \approx 1{,}775$ for positive evidence (BF = 3), $N \approx 3.5 \times 10^6$ for strong evidence (BF = 20), and $N \approx 2.2 \times 10^9$ for BF = 100. The leading $\mathcal{O}(N^{-1/4})$ correction to the law is positive, so the exact Bayes factor lies slightly above it, by 8.5% at $N \approx 1{,}300$ and 1.4% near the strong-evidence threshold. The exact Bayes factor therefore crosses each threshold somewhat earlier than the law does: BF = 3 at $N = 1{,}281$ and BF = 20 at $N \approx 3.3 \times 10^6$ (exact values by direct quadrature; Supplement S3). Whether one uses the asymptotic law or the exact values, strong evidence against a spurious second cohort takes millions of participants (Figure 1). If the trial instead records a single yes or no per participant, Corollary 1 applies and no sample size settles the question from that endpoint alone.

Earlier work has shown that Bayes factors reject a spurious mixture component only at a rate polynomial in the sample size. Fuquéne et al. (2019) state the general in-probability rate without making the exponent or the constant explicit for particular models, Chambaz and Rousseau (2008) bound the error probabilities of posterior-mode order estimators, and Kariya and Watanabe (2022) derive the distribution of the Bayesian likelihood ratio for homogeneity against a normal mixture with sampled data. Equation (6.1) gives the explicit constant for expected-count data. Its derivation also suggests why the Bayes factor grows so slowly with N: the exponent $\lambda_2 - \lambda_1 = 3/4 - 1/2 = 1/4$ is a difference of learning coefficients fixed by the singular geometry of the mixture model, and the constant is determined by the ridge volume. The design is the "cheating coin flipper" of Drton et al. (2009), studied further by Lin (2011); Lin's exact constant (Section 2.2) was derived for a two-parameter coin-mixture model, and his leading-coefficient calculus does not cover the three-parameter mixture treated here (Section 5.2). The singular BIC of Drton and Plummer (2017), run with the correct learning coefficients, predicts the $N^{1/4}$ rate but leaves the constant undetermined, so that the predicted law cannot be inverted for a sample size; the published discussion of that paper (pp. 372 and 375) notes that such constants are generally difficult to compute.

The corollary generalizes in both θ_0 and the number of readings m, which makes the constant, and with it the required sample size, an explicit function of the design. For data at $\mathrm{Bin}(m, \theta_0)$ the derivation of Theorem 7 carries through unchanged, giving

$$c_m(\theta_0) \;=\; \frac{\sqrt{2}\,\bigl(m(m-1)\bigr)^{1/4}}{\pi\,\Gamma(\frac{1}{4})\,\sqrt{\theta_0(1-\theta_0)}}\,, \qquad N^*(B) \;=\; \frac{B^4\,\pi^4\,\Gamma(\frac{1}{4})^4\,\bigl(\theta_0(1-\theta_0)\bigr)^2}{4\,m(m-1)}\,, \tag{6.2}$$

with $\lambda = 3/4$ for every $m \geq 2$, where $N^*(B)$ is the sample size at which the asymptotic Bayes factor reaches B. To our knowledge neither formula has appeared before; Chen et al. (2016) give a related

PRELIMINARY

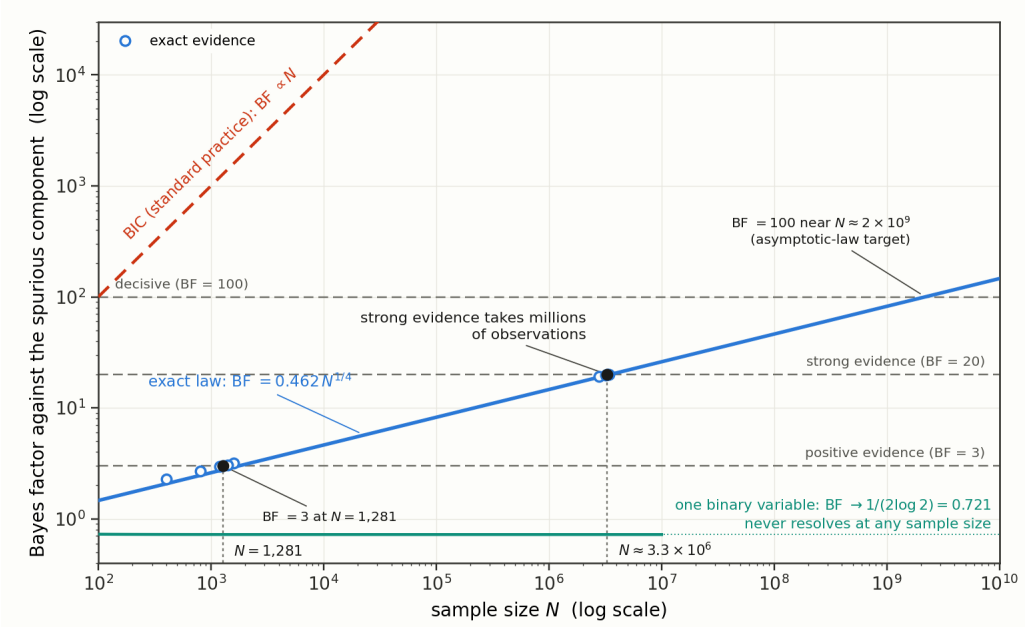

Figure 1: Bayes factor for one component against a two-component mixture on $\mathrm{Bin}(4, \frac{1}{2})$ expected-count data. Solid line: the asymptotic law (6.1); circles: exact values; filled dots: the crossings of $\mathrm{BF} = 3$ and $\mathrm{BF} = 20$. Red dashed: the $\mathrm{BF} \propto N$ rate implied by a parameter-counting BIC. Lower curve: the single-binary-variable case, which converges to $1/(2\log 2)$. Both axes are logarithmic.

fourth-power sample-size–separation law, which fixes error rates rather than an odds target. The factor $m(m-1)$ arises because the information separating the mixture from the one-component model is carried by the second factorial moment, with $\mathrm{KL}/t^4 = 4m(m-1)$ in the ridge coordinate, a factor that vanishes at $m = 1$, the classical identifiability threshold of the binary case. The coefficients in Eq. (6.1) and Eq. (6.2) assume the uniform priors of Section 2.1. The rate does not depend on the prior, whereas the constant does; for the one-variable model Theorem 3 gives the corresponding closed form under integer Dirichlet weights. The constants for the binomial design under non-uniform priors are left to future work.

The Gaussian location mixture of Section 5.3 obeys the same fourth-power law on the population ray (Proposition 1). With φ the $N(0, s_0^2)$ prior density of the means, Eq. (5.12) gives $n^*(B) = B^4\pi^4\,\Gamma(\frac{1}{4})^4\,\varphi(\mu_0)^4/4$, which is Eq. (6.2) with $\varphi(\mu_0)^4$ in place of $\big(\theta_0(1-\theta_0)\big)^2/(m(m-1))$. For standard normal priors centered at the truth, n^* is about 8,600 at $B = 3$ and 1.7×10^7 at $B = 20$. A prior with less density at the true mean favors one component sooner.

On the binomial trial design we also compare the evidence with BIC, the criterion most used in practice to choose the number of components (Section 1.1). The Laplace expansion behind it fails at singular models (Section 2.2), and the comparison measures the resulting error. In Bayes-factor form BIC gives $\mathrm{BF}_{\mathrm{BIC}} = Ne^{-\Delta}$, where $\Delta \geq 0$ is the gap in maximized log likelihoods (parameter difference 2, the Schwarz convention, no $\mathcal{O}(1)$ prior terms). For data generated from the one-component model $\mathrm{Bin}(4, \frac{1}{2})$, enumeration over all datasets gives $P(\mathrm{BF}_{\mathrm{BIC}} \geq 20) = 0.37$, 0.49 and 0.54 at $N = 20$, 21 and 22. On the median such dataset BIC therefore gives twenty-to-one odds against a second component already at $N = 22$. At that sample size the true Bayes factor on

PREPRINT

expected-count data is approximately 1.2, near even odds, and it reaches twenty-to-one only for N in the millions, five orders of magnitude larger. The label-permutation multimodality of the mixture likelihood does not appear to explain the gap. Correcting BIC by $g! = 2$ moves its crossing only from 22 to about 44. The BIC crossing is also independent of m: it stays at $N \approx 20$ whether each participant is measured once, four times, or ten times, including the case $m = 1$ in which, by Corollary 1, the question is never resolved. Many Bayesian statisticians assess models by out-of-sample predictive accuracy, estimated by cross-validation or the Watanabe–Akaike information criterion, and set BIC aside because it targets the marginal probability of the data (e.g. Gelman et al., 2014). The Bayes factor is a ratio of two such probabilities, exact here under the stated priors and approximated by BIC, and it converts prior odds on a second cohort into posterior odds, a different quantity from predictive accuracy.

6.2 Exact reference values for approximate methods

When the evidence is known exactly, the error of an approximate method can be measured directly instead of being inferred from the method's own output. We make three such comparisons: marginal-likelihood estimators against the exact evidences of the binary and binomial-mixture families, the coarsened posterior of Miller and Dunson (2019) on the binomial family, and a partition sampler for record linkage against the posteriors of Supplement S8. The reference values, run counts and numerical checks for the three comparisons are collected in Supplement S3.

We first compare five marginal-likelihood estimators with the true evidence: the harmonic mean (Newton and Raftery, 1994), bridge sampling (Meng and Wong, 1996), stepping-stone sampling (Xie et al., 2011), Chib's method (Chib, 1995), and nested sampling (Skilling, 2006), each run at commonly used default settings. The test problems are the one- and two-component evidences of the binary and binomial-mixture families of Section 2.1, at nine sample sizes from $N = 100$ to 10^6 with thirty simulated datasets per size. Table 4 summarizes the behavior of each estimator. The binary family may be the most demanding case, because there the true log Bayes factor is bounded. By Corollary 1 it tends to the constant $\log(2 \log 2) \approx 0.33$ (mixture over one component), so any bias that grows with N eventually dominates the quantity being estimated. The harmonic mean estimator is biased upward at every sample size, increasingly so as N grows. On the binary family, its estimate of the Bayes factor falls on the wrong side of one at some sample size for each of the thirty simulated datasets.

Chib's method diverges on the binomial family, with errors of order 10^6 nats at $N = 10^6$. Its known difficulty on mixtures is label switching (Neal, 1999), and the permutation-averaged correction of Berkhof et al. (2003) recovers the $\log 2$ missing from Chib's estimate on the binary family. However, the binomial divergence remains of the same order after the correction (Table 4), so it does not appear to be a label-switching artifact. Bridge sampling agrees with the true values up to $N \approx 3{,}000$ and drifts below them at larger N. The random-permutation bridge of Frühwirth-Schnatter (2004) extends the agreement to $N \approx 10^5$. The design question of Section 6.1 is decided only at sample sizes in the millions, beyond the range in which either bridge estimator agrees with the true values, so a pilot study that tested bridge sampling at moderate N would probably not detect the problem. Stepping-stone and nested sampling agree with the true values across the whole range, stepping-stone with a small negative drift at the largest N.

Second, we examine the coarsened posterior of Miller and Dunson (2019) on the two-component $\mathrm{Bin}(4, \theta)$ family, at coarsening parameters $\alpha \in \{10^2, 10^3, 10^4\}$ alongside the standard posterior, with twenty datasets per sample size. When the data are generated outside the family, both the

PRELIMINARY

estimator	behavior	worst bias (nats)	flips (binary, binomial)
harmonic mean	positive bias at every N, growing with N	$+7.6$ at $N = 10^6$	184/270, 24/270
Chib	diverges on the binomial family	$+1.4 \times 10^6$ at $N = 10^6$	213/270, 160/270
Chib, permutation-averaged	recovers $\log 2$ on the binary family; diverges on the binomial family	$+9.2 \times 10^5$ at $N = 10^6$	190/270, 155/270
bridge	calibrated to $N \approx 3 \times 10^3$, negative bias beyond	-6.4 ± 1.3 at $N = 10^6$	84/270, n/a
bridge, random-permutation	calibrated to $N \approx 10^5$, negative bias beyond	-5.8 ± 2.9 at $N = 10^6$	n/a
stepping-stone	calibrated	-0.49 ± 0.26 drift at $N = 10^6$	0/270, 0/270
nested sampling	calibrated	none detected	0/270, 0/250

Table 4: Marginal-likelihood estimators at default settings compared with the exact evidence on the binary and binomial-mixture families, $N = 10^2$ to 10^6, thirty datasets per size. Worst bias: mean error in nats at the least favorable N. Flips: datasets out of 270 per family (250 completed nested-sampling runs, binomial) with the estimated Bayes factor on the wrong side of one.

one-component and the two-component model are misspecified, and under the standard posterior the difference in log evidence between them grows in favor of the wrong model at 0.193 nats per observation (least-squares slope through the medians). Coarsening truncates that growth: this evidence gap rises to a plateau whose half-height sample size increases with α, from $N \approx 145$ at $\alpha = 10^2$ to $N \approx 1{,}031$ at 10^3 and $N \approx 9{,}785$ at 10^4.

Third, a partition sampler, which is usually validated only through diagnostics computed from its own output (Geweke, 2004; Talts et al., 2018), can instead be compared with an exactly known posterior. We compared a standard single-site Gibbs sampler for the record-linkage model of Supplement S8 with the entity-count posteriors computed there for the RLdata500 benchmark, in 24 configurations: four partition priors, three initializations, and two replicate chains, with 50,000 sweeps each. The largest total-variation distance from the exact posterior was 0.05, at the level of the Monte Carlo noise, and the largest shift in posterior mean was 0.04 posterior standard deviations. No configuration passed the standard convergence diagnostics (split-$\hat{R}$, effective sample size) while disagreeing with the true posterior. The comparison covers our own sampler under the plug-in distortion model of Supplement S8, in which the per-field distortion probability is fixed rather than given a prior. Released entity-resolution samplers were not included, because they use global distortion parameters, which under blocking define a different model. Partition samplers have been validated against fully enumerated posteriors in clustering at small n (Bouchard-Côté et al., 2017); with the convolution of Section 4.4 the comparison above was carried out at the full size of the benchmark.

6.3 The Dirichlet process mixture on a real dataset

By Theorem 5 the Dirichlet-process evidence at fixed concentration is obtained by a substitution in a finite sum, so it can be computed on real data in the same way as the finite-mixture evidences. We use the eyetracking data of Escobar and West (1998): counts of eyetracking anomalies in $N = 101$ schizophrenic patients, a standard test case in the Dirichlet-process literature (Frühwirth-Schnatter and Malsiner-Walli, 2019). The counts are binned in two ways, into two cells (zero versus nonzero anomalies, $k = 2$, cell totals $U = (46, 55)$) and into three cells (zero, one to four, five or more, $k = 3$, $U = (46, 29, 26)$). The unbinned counts ($k = 35$) would require about 1.6×10^{11} retained coefficients in the truncated generating function of Section 4.2, beyond the reach of this computation. For each binning we compute the Dirichlet-process evidence and the posterior $P(T = t \mid \text{data})$ of the number of occupied clusters at seven fixed rational concentrations from 1/3 to 355/113. We also compute the evidences of the finite mixtures with $g \leq 3$ components and the Bayes factors of the Dirichlet

PRELIMINARY

process mixture against each. The finite mixtures carry $\mathrm{Dir}(\alpha/g, \ldots, \alpha/g)$ weights and tend to the Dirichlet process mixture as g grows, so each Bayes factor compares the limit with its own truncation at the same concentration and kernel. Since for a single categorical variable the one-component model already covers the whole simplex Δ_{k-1} of cell-probability vectors (Section 4), these Bayes factors measure only the difference between the priors that the models induce on that simplex.

All evidences and posteriors in this comparison are computed under the Dirichlet-categorical kernel with uniform cell priors on the binned counts. The published analyses of these data used Poisson–Gamma kernels on the raw counts, so their numbers are not directly comparable with ours. The inner sums of Eq. (4.5) are rational numbers, one per cluster count, and do not involve α. Once they are computed, the evidence at every concentration and the whole posterior on T follow by substitution (Supplement S3).

k	α	$\log_{10} Z_{\mathrm{DPM}}$	$\log_{10} \mathrm{BF}_{\infty:1}$	$\log_{10} \mathrm{BF}_{\infty:2}$	$\log_{10} \mathrm{BF}_{\infty:3}$	T mode	$E[T]$
2	1/3	−31.0322	0.107	0.055	0.038	2	2.77
2	1/2	−30.9948	0.144	0.076	0.052	3	3.52
2	1	−30.9133	0.225	0.124	0.086	5	5.50
2	3/2	−30.8573	0.281	0.159	0.113	7	7.21
2	2	−30.8151	0.324	0.187	0.134	8	8.76
2	22/7	−30.7468	0.392	0.234	0.172	12	11.86
2	355/113	−30.7468	0.392	0.234	0.171	12	11.86
3	1/3	−48.2082	0.194	0.100	0.067	3	2.91
3	1/2	−48.1456	0.256	0.133	0.090	3	3.67
3	1	−48.0176	0.384	0.207	0.142	5	5.65
3	3/2	−47.9363	0.466	0.256	0.177	7	7.35
3	2	−47.8792	0.523	0.292	0.203	9	8.86
3	22/7	−47.7963	0.606	0.344	0.242	12	11.90
3	355/113	−47.7963	0.606	0.344	0.242	12	11.90

Table 5: Dirichlet process mixture on the eyetracking data ($N = 101$), two-cell ($k = 2$) and three-cell ($k = 3$) binnings, seven concentrations α: $\log_{10}$ evidence, $\log_{10}$ Bayes factors against the matched finite mixtures with $g = 1, 2, 3$ components, and mode and mean of the posterior on the number of occupied clusters T. Dirichlet-categorical kernel, uniform cell priors.

Table 5 gives the results. At every concentration and under both binnings the evidence favors the Dirichlet process mixture over each of $g = 1, 2, 3$, though never strongly: the $\log_{10}$ Bayes factors range from 0.04 to 0.61, all below one order of magnitude. However, the posterior on the number of occupied clusters depends heavily on the concentration. Its mode moves from 2 to 12 as α increases across the grid (Figure 2), and it is broad at every concentration; under the two-cell binning the mode carries 0.31 of the mass at $\alpha = 1/3$ and 0.14 at $\alpha = 22/7$. Conclusions about the number of clusters in these data thus rest largely on the choice of α. A hyperprior on α could be handled by quadrature over a grid of these values; we have not done so here.

A comparison with published estimator output requires the same kernel and data. We are aware of only one published instance in which an analytic Dirichlet-process evidence is reported together with estimator output, and it concerns a different dataset. Basu and Chib (2003) printed an analytic value beside three of their estimators for a ten-observation subset of the rat-growth data, and the estimators differ from it by at most about 0.02. For the eyetracking data no published Dirichlet-

PRELIMINARY

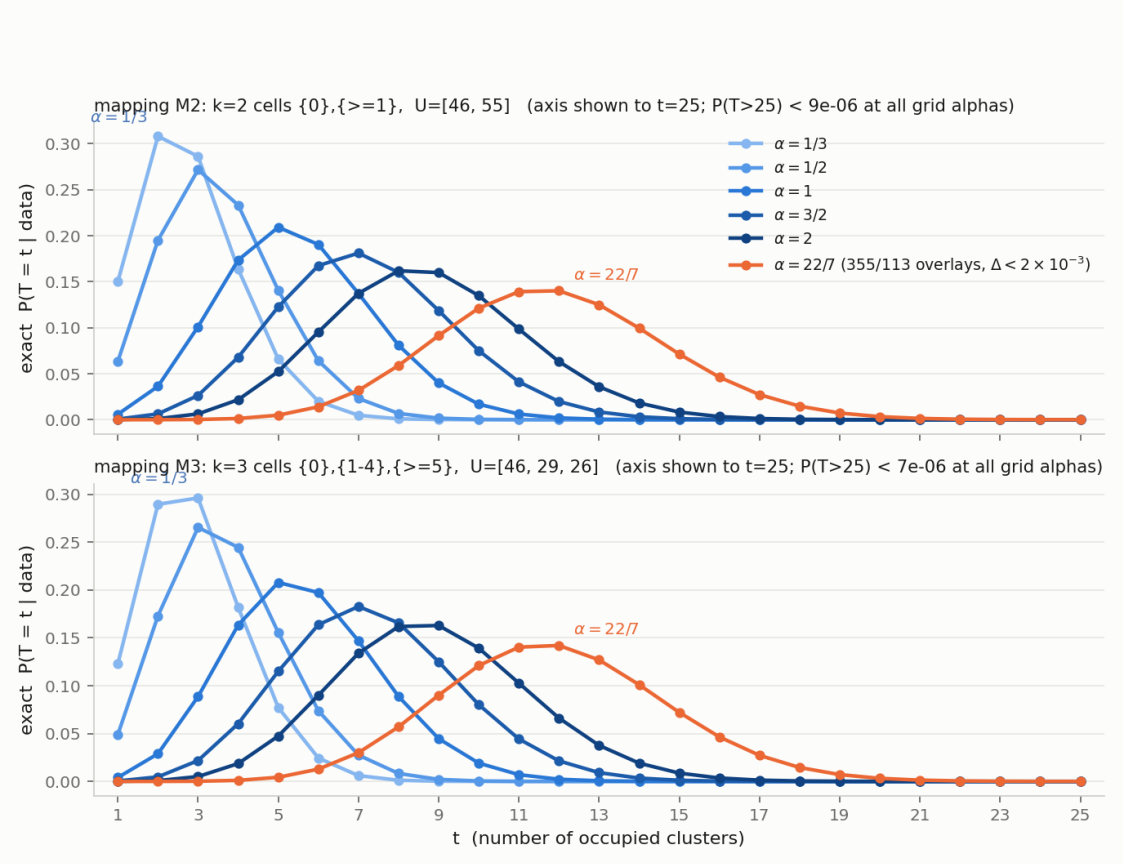

Figure 2: Posterior distribution of the number of occupied clusters T for the Dirichlet process mixture on the eyetracking data at each of the seven concentrations, under the two-cell binning (top) and the three-cell binning (bottom). The curves for $\alpha = 22/7$ and $\alpha = 355/113$ coincide at this scale. The horizontal axis is truncated at $t = 25$.

process evidence value appears to exist. The finite-g bridge-sampling values of Frühwirth-Schnatter and Malsiner-Walli (2019) differ from ours by hundreds of nats because of the different kernel, binning and weight prior. Their MCMC cluster-count table was computed under a hyperprior on α with mean $1/2$. Binned to that table's columns, our posterior at fixed $\alpha = 1/2$ differs from it by total variation 0.45 (two-cell) and 0.42 (three-cell), which again presumably reflects the difference in models.

Exact Dirichlet-process computations have been reported at small n: Ferguson (1983) at $n = 5$, Basu and Chib (2003) by enumerating the Bell(10) = 115,975 partitions at $n = 10$, and Kohonen and Corander (2016) by 3^n subset convolution at $n \leq 20$. A recent study of evidence estimation for Dirichlet process mixtures (Hairault et al., 2022) notes that only Basu and Chib had addressed the problem, and uses a Monte Carlo estimator as its reference value at $n = 6$. To our knowledge, the values in Table 5 are the first exact Dirichlet-process evidences and cluster-count posteriors for a real benchmark dataset with N in the hundreds. At $N = 101$ partition enumeration and subset convolution would take about 10^{117} and 10^{48} operations respectively, and the generating function of Section 4.3 avoids both.

PRELIM

6.4 Further applications

Table 6 lists every application in the paper by its number of components g, number of states or channels k, and number of observations N, together with its main result. Six of them are developed in full in the supplement and summarized below; the model and the range of priors for each are specified in the corresponding supplement section, and the results hold within those ranges.

application	g	k	N	result
trial design (§6.1)	1 vs 2	5	to 3.3×10^6	strong evidence (BF = 20) near 3.3×10^6 (four readings, $\theta_0 = \frac{1}{2}$, expected counts); BIC on the median dataset: $N = 22$
estimator comparison (§6.2)	1 vs 2	2, 5	10^2–10^6	stepping-stone and nested sampling calibrated throughout; bridge sampling to $N \approx 3 \times 10^3$; harmonic mean and Chib biased
capture–recapture (S5)	1–3	T=6, 18	68, 76 seen	known size 135 excluded by g=1, covered by g=2, 3
Roman dies (S6)	1, 2	6 cells (15 lists)	413 seen	die count 519 at the upper edge of the one-class 95% set, inside the two-class set; excluded by every issue-level form of Esty's estimator
eDNA survey (S7)	1–3, cont.	7	356	one class rejected ($\log_{10}$ BF $\geq$ 292.5); at least three classes among finite mixtures; beta-binomial preferred to three classes under every prior
record linkage (S8)	partition, 450	5 fields	500	posterior mean entity count moves from 327.1 (Dirichlet process) to 452.7 (Pitman–Yor) with the partition prior, same data and blocking
prior comparison (S9)	partitions to 3×10^4	5 fields	to 39,743 (34,275 analyzed)	Dirichlet process ranks first on all 3 corpora at α=1 (uniform prior ahead at larger α on the RL-data corpora); pairwise BF up to $10^{83,702}$
DPM eyetracking (§6.3)	∞, $g \leq 3$	2, 3	101	DPM favored over $g \leq 3$ at every α, by less than one order of magnitude
cancer panels (S10)	10 vs 9	96	5–30	largest BF at N=5 among 6,980 readouts: 8; maximum over all possible five-mutation readouts: 50.27
sampler comparison (§6.2)	partition	5 fields	500	Gibbs sampler agrees with the exact posterior within Monte Carlo error in all 24 configurations (largest total variation 0.05)
scale benchmark (Table 1)	2–10^6, ∞	to 1,000	to 10^5	g=2, k=1000, N=10^5 in exact rational arithmetic in 5 min; g=10^6 and g=∞ at N=5,000 in 31–79 s (floating point)

Table 6: Applications by number of components g, states or channels k, and observations N, with the main result of each; S-numbers are supplement sections. In the capture–recapture and die rows the population total is the estimand and N counts observed units; in the partition rows g is the scale of the latent partition.

Capture–recapture (Supplement S5). For Pledger's heterogeneity model with $g = 1, 2, 3$ latent classes we compute the joint posterior $P(N, g \mid \text{data})$, with the unknown population size N and the unobserved zero-count cell inside the integral, on two classic datasets, over sixteen prior settings (four hyperpriors, two priors on the population size, and two truncation caps). On the Edwards–Eberhardt rabbit data, from a penned population of known size 135, the one-class posterior places the true size above its 99.98th percentile under every prior, whereas the two- and three-class posteriors cover it. The evidence also favors heterogeneity under every prior setting, by Bayes factors of roughly 3.5 to 97. On the snowshoe hare data the preference between two and three classes changes sign with the hyperprior. The posterior mean of the population size also rises with the assumed number of classes, the finite-sample counterpart of Link's asymptotic result that the estimated size depends on that assumption.

PRELIMINARY

Counting the dies of a Roman mint (Supplement S6). The denarii of P. Crepusius carry mint-engraved die numbers up to 519, so the number of dies is known and the same latent-class models can be tested against it. From the capture-frequency histogram of fifteen modern holdings (413 distinct dies observed), the one-class 95% credible set includes 519 only at its upper edge, while the two-class 95% set contains it under every prior considered. Every issue-level form of Esty's estimator, the standard tool in numismatics, excludes it. However, the evidence computed from the histogram favors one class, at $\log_{10}$ Bayes factors of -0.99 to -2.21. Because the fifteen holdings differ in size and the histogram pools them, this preference may reflect unequal list effort rather than homogeneity of the mint's output.

Latent classes in an environmental-DNA survey (Supplement S7). The tidewater-goby survey we analyze comprises 356 water samples with six PCR replicates each. We treat the samples as exchangeable and model the histogram of detection counts. The one-class model is rejected decisively ($\log_{10}$ Bayes factors of 292.50 to 293.11 across the priors considered) and two classes are also clearly insufficient ($\log_{10}\mathrm{BF}(3{:}2)$ between 15.5 and 16.1), so at least three latent classes are needed among finite mixtures. A continuous beta-binomial alternative has still higher evidence than the three-class model under every prior. Standard eDNA occupancy analysis fixes two latent states a priori; those site-structured models are not among the exchangeable-sample models compared here. The survey's grouping into 39 sites is not modeled and could account for part of the heterogeneity.

Exact posteriors for record linkage (Supplement S8). On the RLdata500 benchmark (500 records, 450 true entities) we compute the exact posterior of the number of entities under a fixed blocking scheme; the published posteriors for this benchmark were obtained by MCMC. Because the blocking splits some true pairs across blocks, even error-free linkage within blocks would yield 458 entities rather than 450, and posterior mass below 458 can arise only from false merges. With the data and blocking held fixed, changing the partition prior from a Dirichlet process to a Pitman–Yor process moves the posterior mean from 327.1 to 452.7 under the plug-in distortion model. At this sample size the partition prior therefore appears to influence the posterior entity count more than the data do.

An evidence comparison of partition priors (Supplement S9). For eight families of partition prior, including the microclustering ESC priors, we compute the evidence on three linkage corpora of up to 39,743 records and rank the priors by it. Under the primary likelihood specification with concentration $\alpha = 1$ the Dirichlet process ranks first on all three corpora, with pairwise $\log_{10}$ Bayes factors as large as 83,702. In the sensitivity analysis the uniform prior overtakes it at $\alpha = 2$ and 4 on RLdata10000 and at $\alpha = 4$ on RLdata500. The position of the ESC priors in the ranking depends on the likelihood specification, and under the plug-in distortion model on RLdata500 they have higher evidence than the Dirichlet process, so no general statement about them follows. On the RLdata corpora the prior ranked last by the evidence is also the one whose posterior entity count comes closest to the value that error-free linkage within the blocks would give. Such a disagreement is possible because an evidence computed under a misspecified likelihood ranks priors by their fit under that likelihood, which need not coincide with the accuracy of a particular posterior summary such as the entity count.

PRELIMINARY

Mutational-signature evidence at panel scale (Supplement S10). We fix the signature profiles at their published COSMIC values and place a Dirichlet prior on the weights alone. We then compute per-sample Bayes factors for the presence of signature SBS3 (homologous-recombination deficiency) over $k = 96$ mutation channels, on panel-sized readouts drawn from 698 publicly available breast-cancer genomes. At five mutations, the readout size relevant to clinical panels, the largest Bayes factor among 6,980 readouts is 8, whereas the maximum over every possible five-mutation readout is 50.27. Strong evidence for the signature from five mutations is thus possible in principle, but no readout produced by a real tumor comes close to it. At thirty mutations, the formal significance test used in the signature literature flags about a quarter of the readouts whose Bayes factor gives positive evidence for the signature; this difference likely reflects the two methods' operating points, and we have not compared their discrimination.

7 Discussion

Bayesian evidence integrals for mixtures of independence models appear to be considerably more explicit than the literature suggests. For the one-variable family the evidence has a closed form at two components and collapses to a margin sum over contingency-table counts at every g, up to and including the Dirichlet process limit; the most recent work on mixtures of independence models estimated the evidence for individual datasets by Monte Carlo (Borinsky et al., 2023). For data on a singular stratum of the models treated here, we obtain the large-sample expansion of the log evidence to constant order, where singular learning theory had supplied the exponent. However, both results are conditional: the collapse at general g depends on the uniform component prior, and the constants are proved only for counts fixed at their expected values. By how much the sample-size laws of Section 6.1 shift for randomly sampled counts is therefore not settled here. The sensitivity of the applied conclusions of Section 6 to the component prior has likewise been examined only over the finite grids of priors reported with each application. Several ingredients are classical or due to others. The allocation-sum representation is the starting point of the algorithms of Lin et al. (2009), which already compute exact values of the evidence and of the Bayes factor for a given dataset. The learning coefficient $\lambda = 3/4$ for binomial mixtures is a theorem of Yamazaki and Watanabe (2004). The margin-collapse mechanism appears contemporaneously, for a contamination likelihood, in Kim (2026), and the table-count generating function is standard symmetric-function combinatorics (Stanley, 1999).

7.1 A latent-class model beyond reach

One attempted application, a latent-class analysis of diagnostic tests for latent tuberculosis infection, could not be completed. Such analyses rest on published test-agreement tables and, in the guideline-relevant cases, assume two latent classes (Stout et al., 2018); our aim was to test the two-class assumption by computing the evidence of the latent-class model with one, two, and three classes. We considered six tables: the five strata of the Stout et al. table (three tests, $k = 8$ sign patterns) and a four-assay serology table ($k = 16$; Saeed et al., 2021). The one-class evidence of the conditionally independent latent-class model was computed on all six. The two-class evidence was computed on the two small strata ($N = 464$ and $N = 140$), where two classes are preferred over one by ln Bayes factors of 33.95 to 36.92 and 55.81 to 59.73 across four prior settings. A variant of the dynamic program used for the small strata gave the two-class evidence for the two US-born strata,

PRELIMINARY

where the ln Bayes factors are 75.38 to 76.72 and 279.82 to 282.36. On the remaining and largest stratum, with $N = 7{,}931$ subjects, the two-class dynamic program requires a lattice of 1.32×10^{14} states, roughly a petabyte of memory, and was not run; no three-class evidence with an error bound was obtained on any table.

Although the collapse of Theorem 4 is insensitive to sample size, it holds for one categorical variable only. The conditionally independent latent-class model has no such collapse: its lattice grows with the sample size as well as with the multivariate table structure, and for the largest of the five strata the lattice no longer fits in memory. Treating the same stratum as one unstructured $k = 8$ variable, we computed the evidence up to $g = 3$. For a single categorical variable, however, the g-component mixtures cover the same simplex of cell probabilities at every g, so the resulting Bayes factors compare induced priors on one simplex and cannot distinguish two latent classes from three in the three-test latent-class model. A different reduction would be needed to compute the evidence of multivariate latent-class models at clinical sample sizes.

7.2 Open problems

A general dictionary between flat-locus geometry and the resolution-based computations of the Watanabe school would make the constant terms computable for the other models of the singular-learning literature, and might give the singular BIC of Drton and Plummer (2017) its inputs mechanically. The counting rule Eq. (5.7) is stated only for the structures treated here, and the two models of Section 5.3, the Gaussian mixture and reduced-rank regression, are its only continuous cases. For the Gaussian mixture, the extension to unknown or unequal variances and the case of three or more components remain open, and the coefficient of the $n^{-1/4}$ correction in Proposition 1 is known only numerically. Exact values of the evidence for this mixture are out of reach beyond small samples (Nobile, 1994), since with continuous data the g^N allocation terms are almost surely distinct apart from the label symmetry. Asymptotic constants under non-uniform priors are open beyond the integer-Dirichlet family of Theorem 3, as are constants for sampled data in every family treated here, the Gaussian mixture included. The tilting of Rousseau and Mengersen (2011) might give the constants under arbitrary Dirichlet priors, and it is also a possible approach to non-uniform component priors at general g, for which the collapse of Theorem 4 does not hold (Section 4.1). Within the boundary regime the concrete gaps are the general-k family law at generic rank-one data (dominance is proved at uniform expected counts only), the general-F naive Bayes analog, and more than two components. On the applied side, each condition attached to an application could be removed by a further computation: an α-hyperprior treatment of the Dirichlet-process evidence, a comparison of the released linkage samplers under their own kernels, cross-block linkage, capture designs with more occasions, and the three-class model for the die data.

The numerical estimation of learning coefficients, pursued for singular models generally and lately for deep networks, has so far been validated only against the known values of the exponents. The expansions of this paper give such estimators a second quantity to recover: a fitted $\mathcal{O}(1)$ term can be compared with the constant $C_N(\theta_0)$ of Theorem 7 on the binomial-mixture stratum, with the boundary-stratum constants of Theorems 9 and 10, and, for estimators applied to the population integral Eq. (5.10), with the Gaussian-mixture and reduced-rank constants of Section 5.3 under their stated priors. Because finite-N values of the evidence are also available for the discrete families (Section 3, Table 2, Supplement S2), such a comparison could be made along an estimator's whole convergence path as well as at its limit.

PRELIMINARY

Supplementary material

The supplement contains the details of the ridge computation, with the proof of Theorem 6, and the localization proof for the Gaussian location mixture (S1); the boundary-stratum computations, with the proof of Theorem 8 and numerical checks (S2); the numerical methods, the numerical checks for the main-text computations, and the holonomic-structure results (S3); and the proportional regime $N = \nu k$ (S4). Sections S5 to S10 contain the six further applications: capture–recapture, the Roman die study, the environmental-DNA panel, record linkage, the partition-prior comparison, and mutational signatures.

Data and code availability

Every dataset analyzed in this paper is published and is cited where it is used. The evidence routines, the closed forms, the component-count posteriors, and the large-N evaluators are implemented in a software package, available from the authors on request.

Acknowledgments

Funding for this project was provided by Anthropic. This is not an official Anthropic project and results and views are not endorsed by Anthropic.

A The general-k reduction

This appendix proves Theorems 2 and 3. For the k-state model, Eq. (3.4) holds with weight $m!(N-m)!/[(m+k-1)!(N-m+k-1)!] = 1/\prod_{i=1}^{k-1}(m+i)\prod_{j=1}^{k-1}(N-m+j)$. Partial fractions give

$$\frac{1}{\prod_i (m+i)} = \sum_{i=1}^{k-1} \frac{A_i}{m+i}\,, \qquad A_i = \prod_{i' \neq i} \frac{1}{i'-i}\,, \tag{A.1}$$

and each cross term splits into single poles through $\frac{1}{(m+i)(N-m+j)} = \frac{1}{N+i+j}\big[\frac{1}{m+i} + \frac{1}{N-m+j}\big]$. Since $W(N-m) = W(m)$, the sum in Eq. (3.4) reduces to the single-pole sums $S_i = \sum_m W(m)/(m+i)$. By inclusion–exclusion, $W(m) = \sum_{T\subseteq[k]} (-1)^{|T|}\binom{m-a_T+k-1}{k-1}$ with $a_T = \sum_{v\in T}(u_v+1)$ (binomial coefficients vanishing when the upper index is negative), each piece a polynomial in m; dividing each polynomial by $(m+i)$ leaves a polynomial (summing to a rational) plus a residue times $\sum_{m=a}^{N} \frac{1}{m+i} = H_{N+i} - H_{a+i-1}$. Assembling the pieces proves Theorem 2. For integer Dirichlet priors, the binomial weights and rising factorials combine into a weighted count $W_\alpha(m)$ that is again piecewise polynomial (a convolution of polynomial-times-indicator sequences), the Gamma-ratio kernel still has only simple poles in m, now at integer points shifted by the hyperparameters, and the same reduction proves Theorem 3.

References

C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," Ann. Statist. **2** (1974) 1152–1174.

PRELIMINARY

M. Aoyagi, "A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities," Comm. Statist. Theory Methods **39** (2010) 2667–2687.

M. Aoyagi and S. Watanabe, "Stochastic complexities of reduced rank regression in Bayesian estimation," Neural Networks **18** (2005) 924–933.

A. R. Barron, "Are Bayes rules consistent in information?," in *Open Problems in Communication and Computation*, T. M. Cover and B. Gopinath, eds., Springer, New York (1987) 85–91.

S. Basu and S. Chib, "Marginal likelihood and Bayes factors for Dirichlet process mixture models," J. Amer. Statist. Assoc. **98** (2003) 224–235.

J. Berkhof, I. van Mechelen and A. Gelman, "A Bayesian approach to the selection and testing of mixture models," Statist. Sinica **13** (2003) 423–442.

A. Bhattacharya, D. Pati, S. Plummer and Y. Yang, "Evidence bounds in singular models: probabilistic and variational perspectives," Statist. Sci. **41** (2026) 281–302, [arXiv:2008.04537].

R. N. Bhattacharya and R. Ranga Rao, *Normal Approximation and Asymptotic Expansions*, Wiley, New York (1976); SIAM Classics edition, Philadelphia (2010).

M. Borinsky, A.-L. Sattelberger, B. Sturmfels and S. Telen, "Bayesian integrals on toric varieties," SIAM J. Appl. Algebra Geom. **7** (2023) 77–103, [arXiv:2204.06414].

A. Bouchard-Côté, A. Doucet and A. Roth, "Particle Gibbs split-merge sampling for Bayesian inference in mixture models," J. Mach. Learn. Res. **18**(28) (2017) 1–39, arXiv:1508.02663.

A. Chambaz and J. Rousseau, "Bounds for Bayesian order identification with application to mixtures," Ann. Statist. **36** (2008) 938–962.

J. Chen, P. Li, and Y. Liu, "Sample-size calculation for tests of homogeneity," Canad. J. Statist. **44** (2016) 82–101.

Y. Chen, C.-D. Fuh and C.-L. M. Kao, "Determine the number of states in hidden Markov models via marginal likelihood," J. Mach. Learn. Res. **26** (2025) 1–59, [arXiv:2405.12343].

S. Chib, "Marginal likelihood from the Gibbs output," J. Amer. Statist. Assoc. **90** (1995) 1313–1321.

S. Chib and T. A. Kuffner, "Bayes factor consistency," arXiv:1607.00292 (2016).

D. M. Cifarelli and E. Regazzini, "Distribution functions of means of a Dirichlet process," Ann. Statist. **18** (1990) 429–442; correction, **22** (1994) 1633–1634.

D. Dacunha-Castelle and É. Gassiat, "Testing in locally conic models, and application to mixture models," ESAIM Probab. Statist. **1** (1997) 285–317.

A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer, New York (1998).

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. B **39** (1977) 1–22.

PRELIMINARY

P. Diaconis and J. Kemperman, "Some new tools for Dirichlet priors," in *Bayesian Statistics 5*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds. (Oxford University Press, Oxford, 1996) 97–106.

M. Drton, A. McCormack and D. Windisch, "Asymptotics for model selection in probabilistic principal component analysis," arXiv:2608.23513 (2026).

M. Drton and M. Plummer, "A Bayesian information criterion for singular models," J. R. Stat. Soc. Ser. B **79** (2017) 323–380 (with discussion), [arXiv:1309.0911].

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

M. D. Escobar and M. West, "Computing nonparametric hierarchical models," in *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. D. Dey, P. Müller and D. Sinha, eds.), Lecture Notes in Statistics **133**, Springer, New York, 1998, pp. 1–22.

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

P. Fearnhead, "Direct simulation for discrete mixture distributions," Stat. Comput. **15** (2005) 125–133.

T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," Ann. Statist. **1** (1973) 209–230.

T. S. Ferguson, "Bayesian density estimation by mixtures of normal distributions," in *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.), Academic Press, New York, 1983, pp. 287–302.

C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," J. Amer. Statist. Assoc. **97** (2002) 611–631.

S. Frühwirth-Schnatter, "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques," Econometrics J. **7** (2004) 143–167.

S. Frühwirth-Schnatter and G. Malsiner-Walli, "From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering," Adv. Data Anal. Classif. **13** (2019) 33–64, arXiv:1706.07194.

J. Fuquéne, M. Steel, and D. Rossell, "On choosing mixture components via non-local priors," J. R. Stat. Soc. Ser. B **81** (2019) 809–837.

A. Gelman, J. Hwang and A. Vehtari, "Understanding predictive information criteria for Bayesian models," Stat. Comput. **24** (2014) 997–1016.

J. Geweke, "Getting it right: joint distribution tests of posterior simulators," J. Amer. Statist. Assoc. **99** (2004) 799–804.

P. J. Green and S. Richardson, "Modelling heterogeneity with and without the Dirichlet process," Scand. J. Statist. **28** (2001) 355–375.

PRELIMINARY

A. Hairault, C. P. Robert and J. Rousseau, "Evidence estimation in finite and infinite mixture models and applications," arXiv:2205.05416 (2022).

B. Hanin and A. Zlokapa, "Bayesian interpolation with deep linear networks," Proc. Natl. Acad. Sci. **120** (2023) e2301345120, [arXiv:2212.14457].

J. A. Hartigan, "A failure of likelihood asymptotics for normal mixtures," in Proc. Berkeley Conference in Honor of J. Neyman and J. Kiefer, vol. II, Wadsworth (1985) 807–810.

N. Hayashi, "The exact asymptotic form of Bayesian generalization error in latent Dirichlet allocation," Neural Networks **137** (2021) 127–137.

S. Hoşten, A. Khetan and B. Sturmfels, "Solving the likelihood equations," Found. Comput. Math. **5** (2005) 389–407.

H. Ishwaran and M. Zarepour, "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models," Biometrika **87** (2000) 371–390.

H. Ishwaran and M. Zarepour, "Dirichlet prior sieves in finite normal mixtures," Statist. Sinica **12** (2002) 941–963.

L. F. James, "Functionals of Dirichlet processes, the Cifarelli–Regazzini identity and Beta-Gamma processes," Ann. Statist. **33** (2005) 647–660.

N. L. Johnson and S. Kotz, "Randomly weighted averages: some aspects and extensions," Amer. Statist. **44** (1990) 245–249.

N. Kariya and S. Watanabe, "Testing homogeneity for normal mixture models: variational Bayes approach," IEICE Trans. Fundamentals **E103-A** (2020) 1274–1282, [arXiv:1912.07041].

N. Kariya and S. Watanabe, "Asymptotic analysis of singular likelihood ratio of normal mixture by Bayesian learning theory for testing homogeneity," Comm. Statist. Theory Methods **51** (2022) 5873–5888, [arXiv:1812.03510].

R. E. Kass and A. E. Raftery, "Bayes factors," J. Amer. Statist. Assoc. **90** (1995) 773–795.

C. Keribin, "Consistent estimation of the order of mixture models," Sankhyā Ser. A **62** (2000) 49–66.

S. V. Kerov and N. V. Tsilevich, "The Markov–Krein correspondence in several dimensions," Zap. Nauchn. Sem. POMI **283** (2001) 98–122; English transl. J. Math. Sci. (N.Y.) **121** (2004) 2345–2359.

J. Kim, "Marginal likelihoods for finite-support Huber contamination," arXiv:2605.26723 (2026).

J. Kohonen and J. Corander, "Computing exact clustering posteriors with subset convolution," Comm. Statist. Theory Methods **45** (2016) 3048–3058, arXiv:1310.1034.

P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," Inf. Process. Lett. **103** (2007) 227–233.

PRELIMINARY

S. Lin, *Algebraic Methods for Evaluating Integrals in Bayesian Statistics*, Ph.D. thesis, University of California, Berkeley, 2011.

S. Lin, "Ideal-theoretic strategies for asymptotic approximation of marginal likelihood integrals," arXiv:1003.5338 (2010); J. Algebraic Statistics **8** (2017).

S. Lin, B. Sturmfels and Z. Xu, "Marginal likelihood integrals for mixtures of independence models," J. Mach. Learn. Res. **10** (2009) 1611–1631, [arXiv:0805.3602].

D. A. Linzer and J. B. Lewis, "poLCA: An R package for polytomous variable latent class analysis," J. Stat. Softw. **42**(10) (2011) 1–29.

R. H. Mena, C. Merkatas, T. Nicoleris and C. E. Rodríguez, "Exact two-stage finite-mixture representations for species sampling processes," arXiv:2512.24414 (2025).

X.-L. Meng and W. H. Wong, "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration," Statist. Sinica **6** (1996) 831–860.

J. W. Miller and D. B. Dunson, "Robust Bayesian inference via coarsening," J. Amer. Statist. Assoc. **114** (2019) 1113–1125.

J. W. Miller and M. T. Harrison, "Mixture models with a prior on the number of components," J. Amer. Statist. Assoc. **113** (2018) 340–356, arXiv:1502.06241.

T. Mononen and P. Myllymäki, "Fast NML computation for naive Bayes models," in *Discovery Science 2007*, Lecture Notes in Computer Science **4755** (2007) 151–160.

R. M. Neal, "Bayesian mixture modeling," in *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop (Seattle, 1991)*, C. R. Smith, G. J. Erickson and P. O. Neudorfer, eds., Kluwer Academic, Dordrecht (1992) 197–211.

R. M. Neal, "Erroneous results in 'Marginal likelihood from the Gibbs output'," unpublished letter, University of Toronto (1999); submitted to J. Amer. Statist. Assoc. and declined, available at `https://glizen.com/radfordneal/chib-letter.html`.

R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," J. Comput. Graph. Statist. **9** (2000) 249–265.

M. A. Newton and A. E. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap," J. R. Stat. Soc. Ser. B **56** (1994) 3–26.

A. Nobile, *Bayesian Analysis of Finite Mixture Distributions*, Ph.D. thesis, Carnegie Mellon University, 1994.

K. L. Nylund, T. Asparouhov and B. O. Muthén, "Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study," Struct. Equ. Model. **14** (2007) 535–569.

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

PRELIMINARY

K. Pearson, "III. Contributions to the mathematical theory of evolution," Phil. Trans. R. Soc. Lond. A **185** (1894) 71–110.

J. Pitman and M. Yor, "The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator," Ann. Probab. **25** (1997) 855–900.

E. Regazzini, A. Guglielmi, and G. Di Nunno, "Theory and numerical analysis for exact distributions of functionals of a Dirichlet process," Ann. Statist. **30** (2002) 1376–1411.

S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," J. R. Stat. Soc. B **59** (1997) 731–792.

C. P. Robert and K. Mengersen, "Exact Bayesian analysis of mixtures," in *Mixtures: Estimation and Applications*, Wiley (2011), [arXiv:1010.6113].

J. Rousseau and K. Mengersen, "Asymptotic behaviour of the posterior distribution in overfitted mixture models," J. R. Stat. Soc. Ser. B **73** (2011) 689–710.

D. Rusakov and D. Geiger, "Asymptotic model selection for naive Bayesian networks," J. Mach. Learn. Res. **6** (2005) 1–35.

S. Saeed, S. F. O'Brien, K. Abe, Q.-L. Yi, B. Rathod, J. Wang, M. Fazel-Zarandi, A. Tuite, D. Fisman, H. Wood, K. Colwill, A.-C. Gingras and S. J. Drews, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) seroprevalence: Navigating the absence of a gold standard," PLoS One **16** (2021) e0257743.

G. Schwarz, "Estimating the dimension of a model," Ann. Statist. **6** (1978) 461–464.

J. Skilling, "Nested sampling for general Bayesian computation," Bayesian Anal. **1** (2006) 833–859.

R. P. Stanley, *Enumerative Combinatorics, Volume 2*, Cambridge Studies in Advanced Mathematics **62**, Cambridge University Press, 1999.

J. E. Stout, Y. Wu, C. S. Ho, A. C. Pettit, P.-J. Feng, D. J. Katz, S. Ghosh, T. Venkatappa and R. Luo, "Evaluating latent tuberculosis infection diagnostics using latent class analysis," Thorax **73** (2018) 1062–1070.

Y. Tachibana, Y. Goto, T. Koyama and N. Takayama, "Holonomic gradient method for two-way contingency tables," Algebraic Statistics **11** (2020) 125–153.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari and A. Gelman, "Validating Bayesian inference algorithms with simulation-based calibration," arXiv:1804.06788 (2018).

H. Teicher, "Identifiability of finite mixtures," Ann. Math. Statist. **34** (1963) 1265–1269.

W. Van Assche, "A random variable uniformly distributed between two independent random variables," Sankhyā Ser. A **49** (1987) 207–211.

L. Wasserman, "Mixture models: the twilight zone of statistics," Normal Deviate (blog), August 4, 2012, `https://normaldeviate.wordpress.com/2012/08/04/mixture-models-the-twilight-zone-of-statistics/`.

PRELIMINARY

S. Watanabe, "Algebraic analysis for non-regular learning machines," in *Advances in Neural Information Processing Systems* **12** (1999) 356–362.

S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, 2009.

T. Watanabe and S. Watanabe, "Asymptotic behavior of Bayesian generalization error in multinomial mixtures," arXiv:2203.06884 (2022).

W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen, "Improving marginal likelihood estimation for Bayesian phylogenetic model selection," Syst. Biol. **60** (2011) 150–160.

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

K. Yamazaki and S. Watanabe, "Newton diagram and stochastic complexity in mixture of binomial distributions," in *Algorithmic Learning Theory*, Lecture Notes in Computer Science **3244** (2004) 350–364.

P. Zwiernik, "An asymptotic behaviour of the marginal likelihood for general Markov models," J. Mach. Learn. Res. **12** (2011) 3283–3310.

PRELIMINARY