Stat Comput (2015) 25:797–808
DOI 10.1007/s11222-015-9562-9

de Finetti Priors using Markov chain Monte Carlo computations

Sergio Bacallado[1] · **Persi Diaconis**[1] · **Susan Holmes**[1]

Accepted: 8 March 2015 / Published online: 11 June 2015

Abstract Recent advances in Monte Carlo methods allow us to revisit work by de Finetti who suggested the use of approximate exchangeability in the analyses of contingency tables. This paper gives examples of computational implementations using Metropolis Hastings, Langevin, and Hamiltonian Monte Carlo to compute posterior distributions for test statistics relevant for testing independence, reversible or three-way models for discrete exponential families using polynomial priors and Gröbner bases.

Keywords Priors · MCMC · Contingency tables · Bayesian inference · Independence

Mathematics Subject Classification 62C10 · 62C25

1 Introduction

In a little-known article de Finetti (1938), de Finetti suggests a useful general procedure for quantifying the idea that collections of random variables are *approximately* exchangeable. He also developed methods that extend the idea to partial exchangeability. Surprisingly, de Finetti worked in terms of parameters—usually he eschewed parametric versions in favor of observables.

This paper suggests using de Finetti's ideas to quantify things like multinomial contingency tables being approximately independent or Markov chains being close to reversible. It presents de Finetti's examples in modern language, gives a dozen further examples, and a development for log-linear models using the tools of algebraic statistics. A suite of python programs, some practical suggestions and examples complete the picture.

To begin, here is de Finetti's original example.

Example 1 (Men and women) Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be binary random variables with $(X_i)_{1\le i\le m}$ representing, say, the results of a test on men, and $(Y_i)_{1\le i\le n}$ representing the result of the same test on women. Suppose that, with the $(Y_i)_{1\le i\le n}$ fixed, the $(X_i)_{1\le i\le m}$ are judged exchangeable with each other, and similarly, the $(X_i)_{1\le i\le m}$ are exchangeable with the $(Y_i)_{1\le i\le n}$ fixed. If the two sequences are judged extendable, de Finetti's basic representation theorem, suitably extended, shows that there is a probability measure $\mu(dp_1, dp_2)$ on the Borel sets of the unit square $[0,1]^2$ such that

$$P(X_1 = x_1, \ldots, X_m = x_m, Y_1 = y_1, \ldots, Y_m = y_m) = \int_0^1 \int_0^1 p_1^a (1-p_1)^{m-a} p_2^b (1-p_2)^{n-b} \mu(dp_1, dp_2), \tag{1}$$

for any binary sequences $(x_i)_{1\le i\le m}, (y_i)_{1\le i\le n}$ with $x_1+\cdots+x_m = a$ and $y_1+\cdots+y_n = b$.

Note that this kind of partial exchangeability is very different than marginal exchangeability. If X_i are independent and identically distributed from p drawn from $\mu(dp)$ and $Y_i =$

Electronic supplementary material The online version of this article (doi:10.1007/s11222-015-9562-9) contains supplementary material, which is available to authorized users.

✉ Susan Holmes
susan@stat.stanford.edu

Sergio Bacallado
sergiob@stanford.edu

Persi Diaconis
diaconis@math.stanford.edu

[1] Sequoia Hall, Stanford University, Stanford, USA

X_i, then $\{X_i\}$ are 'marginally' exchangeable, as are the $\{Y_j\}$ but $\{X_i, Y_j\}$ are not partially exchangeable.

de Finetti considered the situation where we suspect that the covariate men/women hardly matters so that, in a rough sense, all of $(X_i)_{1\le i\le m}$, $(Y_i)_{1\le i\le n}$ are approximately exchangeable. While he did not offer a sharp definition, he observed that if $(X_i)_{1\le i\le m}$ and $(Y_i)_{1\le i\le n}$ are exactly exchangeable, the mixing measure $\mu(dp_1, dp_2)$ is concentrated on the diagonal: $p_1 = p_2$. One way to build approximate exchangeability is to work with measures concentrated near the diagonal. de Finetti suggested

$$Z^{-1}e^{-A(p_1-p_2)^2} \quad \text{on } [0,1]^2,$$

with scale factor A and a normalizing constant Z. The use of the quadratic in the exponent is not arbitrary, it is motivated by the central limit theorem. We will find it beneficial to reparametrize this as

$$Z^{-1}e^{-\frac{B}{\lambda}(p_1-p_2)^2}, \tag{2}$$

where B will play the role of a concentration parameter and λ is the average of $(p_1 - p_2)^2$ over the unit square.

For example, in a classical discussion of testing in 2×2 tables, Egon Pearson (1947) gave the data in Table 1.

In what Pearson (1947) names his second scenario (II), the data are supposedly collected on two different sets of shots tested against metal plates. The two samples are considered interchangeable and Pearson proposed a method for testing whether $p_1 = p_2$. He constructed a table of the confidence values (see Supplementary material, Figure 14) for his statistics for a whole table of possible values ($a, b, m = 18, n = 12$). An instructive paper of Howard (1998) uses this data to compare many different Bayesian and non-Bayesian procedures. Figure 1 shows the posterior distribution for $p_1 - p_2$ for differing values of the concentration parameter B. If $B = 0$ the prior is uniform on $[0, 1]^2$, as B increases, the prior becomes concentrated on $p_1 = p_2$.

In his original paper, Pearson presents a graphical display of the confidence values. We offer a Bayesian version of this in the supplementary material section.

This example is explained in Sect. 2. Nowadays, it is easy to work with such priors, sampling from the posterior using standard Markov chain Monte Carlo techniques. In de Finetti's time, this was a more difficult numerical analysis task; de Finetti and his students carefully worked out several numerical examples. These were presented in an English translation de Finetti (1980). The examples involve more covariates (e.g., men/women, smoker/non smoker gives four possible covariates). Exponentiated quadratics are used to force the various $\{p_i\}$ toward each other, or toward a null model. Different weights allowed in front of various quadratic terms permit quantification of more or less prior belief in various sub-models. Section 2 presents some of de Finetti's other examples, an actuarial example and a proportional hazards example. Section 3 presents further new examples: independence, no three-way interaction in contingency tables, Markovian and reversible Markov chains, log-linear models and models from algebraic statistics. Section 4 discusses the computational methods used to compute the posterior distributions used for inference in this framework. Finally Sect. 5 presents some more detailed scientific applications.

Table 1 Pearson (1947) used this example to visit three different possible interpretations of testing when faced with a 2×2 contingency table

	Successes	Failures	Totals
Sample 1	3	15	18
Sample 2	7	5	12
Totals	10	20	30

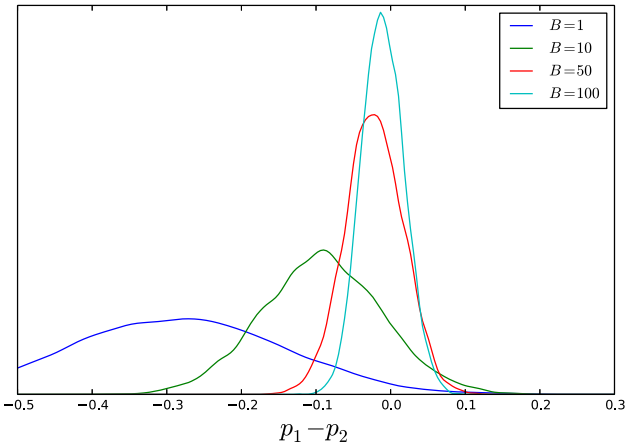


Fig. 1 Posterior distribution of the difference $p_1 - p_2$ for different values of the concentration parameter B as defined in (2)

2 Background from de Finetti's paper

Consider the sequence $(X_1, C_1), (X_2, C_2), \ldots, (X_n, C_n)$ with $X_i \in \{0, 1\}$ random variables and $C_i \in \mathcal{C}$ observable covariates such as male/female, hair color, or weight. Suppose for now that $\mathcal{C}$ is a finite set of cardinality c, this means that C is a factor variable with a finite set of possible levels. If all of the X_i with the same value of the covariate are considered exchangeable with each other (for fixed values of the other variables) and if the data are extendable to infinite sequences then a standard variant of de Finetti's theorem is in force: there is a joint probability measure $\mu(dp_1, \ldots, dp_c)$ on $[0, 1]^c$ such that

$$P(r_i \text{ successes out of } n_i \text{ trials of type } i;\ 1 \le i \le c) = \int \prod_{i=1}^{c} p_i^{r_i}(1-p_i)^{n_i-r_i}\mu(dp_1,\ldots,dp_c). \quad (3)$$

Observing r_i successes out of n_i trials for each type i, the posterior is of the form

$$Z^{-1}\prod_{i=1}^{c} p_i^{r_i}(1-p_i)^{n_i-r_i}\mu(dp_1,\ldots,dp_c). \quad (4)$$

In large samples, with $n = n_1 + \cdots + n_c$, the Central Limit Theorem applies under mild conditions on μ [see Ghosh et al. (1982)]. This says the posterior is well approximated by the normal density

$$Z^{-1}e^{-\frac{n}{2}R(p_1,\ldots,p_c)}, \quad (5)$$

with

$$R(p_1,\ldots,p_m) = \left(\frac{p_1-p_1^*}{\sigma_1}\right)^2 + \cdots + \left(\frac{p_c-p_c^*}{\sigma_c}\right)^2, \quad (6)$$

$$p_i^* = \frac{r_i}{n_i} \quad ; \quad 1 \le i \le c, \quad (7)$$

$$\sigma_i^2 = \frac{p_i^*(1-p_i^*)}{n_i/n} \quad ; \quad 1 \le i \le c. \quad (8)$$

For large n, provided n_i/n are bounded away from zero and one, the posterior converges weakly to a point mass $\delta_{(p_1^*,\ldots,p_c^*)}$ at the observed proportions, see Diaconis and Freedman (1990) for more precise versions of this. Under the approximate point mass posterior, the future observations for each covariate are judged approximately independent with probabilities $p_1^*,\ldots,p_c^*$, respectively.

de Finetti considers priors μ which are informative. Based on the above central limit behavior, he takes priors of Gaussian form restricted to $[0,1]^c$ and renormalized.

Example 2 (de Finetti's almost exchangeability) To reflect "almost exchangeability" among the classes, de Finetti suggests taking μ of the form

$$Z^{-1}e^{-\frac{1}{2}Q(p_1,\ldots,p_c)}, \quad \text{with} \quad Q(p_1,\ldots,p_c) = \sum_{1\le i<j\le c} a_{ij}(p_i-p_j)^2. \quad (9)$$

The posterior is approximately of the form

$$Z^{-1}e^{-\frac{1}{2}\left(Q(p_1,\ldots,p_c)+nR(p_1,\ldots,p_c)\right)}, \quad (10)$$

with R from Eq. 6. This allows a rough guide for how to choose $\{a_{ij};\ 1 \le i < j \le c\}$. Thinking about them in terms of a prior sample size. As a_{ij} tend to infinity, the prior becomes the uniform distribution concentrated on the main diagonal. As all a_{ij} tend to zero, the prior becomes uniform on $[0,1]^c$. A modern thought is to put a prior on $\{a_{ij};\ 1 \le i < j \le c\}$.

In the examples from de Finetti (1972) Chapters 9 and 10 the $\{a_{ij};\ 1 \le i < j \le c\}$ are quite large. This means a very large sample size is required to change the initial opinion that the covariates do not matter (complete exchangeability).

Example 3 (de Finetti's almost constant probability) As a first variation, de Finetti considers the situation where all the probabilities are judged approximately equal to a constant (fixed) value p_*. Think of different subjects flipping the same symmetrical coin. He suggests taking the quadratic form

$$Q(p_1,\ldots,p_c) = A\sum_{1\le i<j\le c}(p_i-p_j)^2 + B(p_1+\cdots+p_c-cp_*)^2. \quad (11)$$

Adjusting A and B allows trading off between "all p_i equal" and "all p_i equal to p_*." If just the latter is considered, the form $A\sum_i(p_i-p_*)^2$ is appropriate.

Example 4 (de Finetti's actuarial example) Consider (X_i, C_i) with X_i one if the ith family has an accident in the coming year and zero otherwise; C_i is the number of family members in the ith family. Suppose that all of the X_i with the same value of C_i are judged exchangeable with each other. By de Finetti's theorem, the problem can be represented in terms of parameters p_j—the chance that a family with j members has an accident in the coming years.

It is also natural to let $q_j = 1-p_j$ and consider priors concentrated about the curve $q_2 = q_1^2, q_3 = q_1^3,\ldots,q_c = q_1^c$. One way to do this is to use a polynomial which vanishes at the null model such as

$$Q(p_1,\ldots,p_c) = A\sum_j\left((1-p_j)-(1-p_1)^j\right)^2. \quad (12)$$

Other choices are possible; e.g., using $\left((1-p_j)^{1/j}-(1-p_1)\right)^2$ or inserting weights A_j into the sum.

Example 5 (de Finetti's proportional rates) Consider a medical experiment involving four categories:

- Treated lab animals—p_1,
- Untreated lab animals—p_2,
- Treated humans—p_3, and
- Untreated humans—p_4.

If the units within each category are judged exchangeable, de Finetti's theorem shows that the problem can be described in

terms of the form of the four parameters shown. de Finetti was interested in cases where the proportion of improvement due to treatment is roughly constant; $p_1/(p_1+p_2) = p_3/(p_3+p_4)$ or equivalently $p_1/p_2 = p_3/p_4$ or $p_1 p_4 = p_2 p_3$. The polynomial $A(p_1 p_4 - p_2 p_3)^2$ does the job.

de Finetti (1938) develops this example extensively, leading to insights into how we come to believe laws of the constant proportionality type in the first place.

3 Further examples

This section uses de Finetti's idea to suggest natural priors for a variety of standard statistical models: contingency tables (independence, no three-way interaction), directed graphical models and more general log-linear models. Here the task is to translate the statistically natural version of the model into an equivalent form involving the vanishing of systems of polynomial equations. Then, a quadratic expression can be used to build priors which live close to the model.

Before proceeding, note that one use of these priors is for testing if the model holds versus the more general alternative. The de Finetti priors can be used on the alternative; more standard priors (e.g., Dirichlet) may be used for the null model. Then, relevant Bayes factors may be computed. The de Finetti priors are centered around the null model. This is standard practice; for example, if $X_1, X_2, \ldots, X_n$ are Normal$(\mu, 1)$, a test for $\mu = 0$ would be constructed using a point mass prior at $\mu = 0$ and a normal prior on μ otherwise. Usually, this normal prior is centered at $\mu = 0$. In testing if data are, say, Normal(0, 1) a point mass is put on the normal and, for a non-parametric test, a non-parametric prior is put on the alternative. Usually, this prior is centered on the Normal(0, 1). Thus, if a Dirichlet prior is considered, the parameter measure is taken as Normal(0, 1). For further discussion, especially for cases where an antagonistic prior is put on the alternative, see the tests for independence in Diaconis and Efron (1985).

Example 6 (Testing for independence in contingency tables) Consider $X_1, X_2, \ldots, X_n$ taking values (i, j) with $1 \le i \le I, 1 \le j \le J$. If N_{ij} is the number of X_k equalling (i, j), under exchangeability and extendability, the $N_{i,j}$ are conditionally multinomial with parameters $\{p_{ij}; 1 \le i \le I, 1 \le j \le J\}$ and mixing measure $\mu(dp_{11}, \ldots, dp_{IJ})$. A test for independence can be based on the prior

$$Z^{-1} e^{-\frac{A}{2}\sum_{i,j}(p_{ij}-p_{i\cdot}p_{\cdot j})^2} \quad \text{on } \Delta_{IJ-1}, \tag{13}$$

with $p_{i\cdot} = \sum_j p_{ij}$ and $p_{\cdot j} = \sum_i p_{ij}$.

Of course, this is just one of many priors that can be used. An alternative parametrization uses the form

$$\sum_{i_1,i_2,j_1,j_2} (p_{i_1 j_1} p_{i_2 j_2} - p_{i_1 j_2} p_{i_2 j_1})^2, \tag{14}$$

with $1 \le i_1, i_2 \le I$ and $1 \le j_1, j_2 \le J$. No attempt will be made here to review the extensive Bayesian literature on testing for independence in contingency tables. The book by Good (1965), and paper by Diaconis and Efron (1985) survey much of this. The de Finetti priors suggested above allow a simple way to concentrate the prior around the null. We have not seen these believable possibilities (13) and (14) treated seriously previously.

3.1 Birthday–deathday

The question we pose is *Can older people hold on to die just past their birthdays?* This question was popularized by Phillips (1978) and is still of current research interest Ajdacic-Gross et al. (2012). The effect is claimed to be potent on important people. A well-known dataset (Andrews and Herzberg 1985, p. 429) records the month of birth and death of 82 descendants of Queen Victoria. The usual χ^2 test for independence provides a statistic whose value is 115.6 on 121 degrees of freedom. An exact test Diaconis and Sturmfels (1998) shows the permutation probability of $\chi^2 \le 115.6$ is 0.321 suggesting no association.

For our Bayesian analysis, the parameters $p_{ij}, 1 \le i \le j \le 12$ representing the chance that a subject is born in month i and dies in month j. The null hypothesis is that $p_{ij} = \eta_i \gamma_j$. We used a product uniform distribution on the 12-simplex for the null hypothesis. For the alternative, we used the density (13) where $A = 0$ corresponds to the uniform distribution and large A forces the prior close to the surface of independence. As described in Sect. 4 we take $A = B/\lambda$ where λ is the expectation of Q with respect to the uniform distribution on $\{p_{ij}\}$. Thus large values of B make the prior concentrate at the surface of independence and small values of B make the prior uniform.

In this example λ, the expected value of Q is approximately 1.62×10^{-6}.

Figure 2 shows the logarithmic Bayes factor comparing the alternative model with $B = 0$, to models with increasing parameter B in which the prior pulls the model toward the null hypothesis of independence. The plot suggests weak evidence against independence. Figure 3 shows the posterior distribution of the form $Q(p)$ for different values of B, illustrating the effect of the parameter.

Example 7 (No three-way interaction) If n objects are classified into three categories with I, J, and K levels, respectively, under exchangeability and extendability, conditionally the objects have a multinomial distribution with probability p_{ijk} of falling into category (i, j, k). There is also a prior

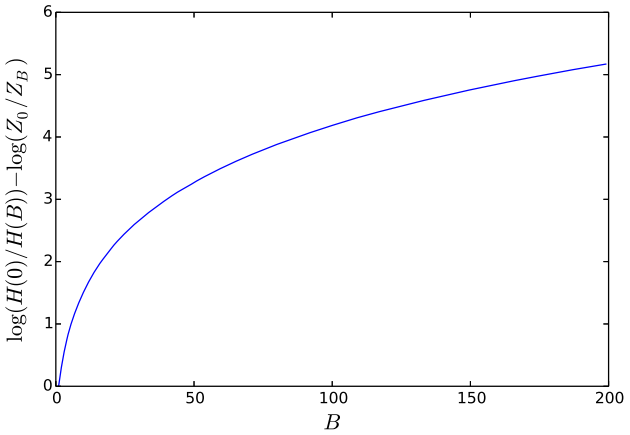


Fig. 2 Independence model for the birth and death month of the descendants of queen Victoria. The plot shows the logarithmic Bayes factor comparing the model with $B = 0$ to a range of models with increasing parameter B, i.e., increasing concentration around the independence hypothesis

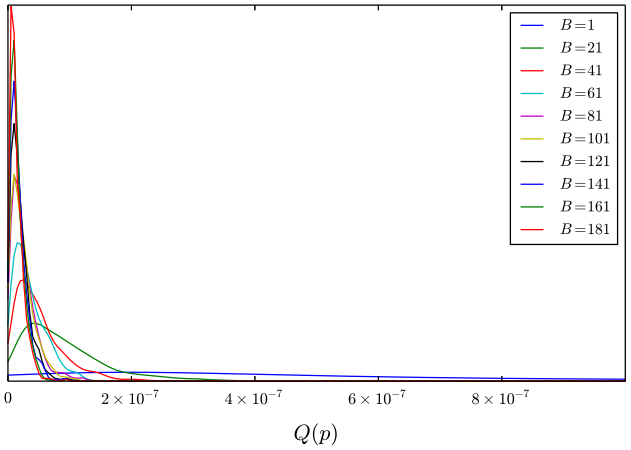


Fig. 3 Independence model for the birth and death months of the descendants of queen Victoria. Histogram of posterior samples of the polynomial $Q(p)$ for different values of the prior parameter B. The least peaked Q is for the value $B = 1$, the very peaked Q is for the value $B = 181$

$\mu(dp_{111}, \ldots, dp_{IJK})$ on Δ_{IJK-1}. The "no three factor interaction" model may be specified by

$$\frac{p_{111}\, p_{ij1}}{p_{i11}\, p_{1j1}} = \frac{p_{11k}\, p_{ijk}}{p_{i1k}\, p_{1jk}} \quad 2 \le i \le I,\ 2 \le j \le J,\ 2 \le k \le K. \tag{15}$$

Using the form

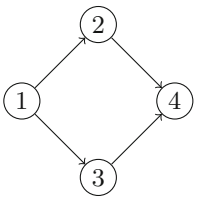
$$\sum_{i,j,k} A(p_{111}\, p_{ij1}\, p_{i1k}\, p_{1jk} - p_{i11}\, p_{1j1}\, p_{11k}\, p_{ijk})^2 \tag{16}$$

summed over the same indices gives a simple way to put a prior near this model.

The no three-way interaction model is an example of a hierarchical model Goodman (1970), Haberman (1978), Darroch et al. (1980). The widely studied subclass of graphical models Lauritzen (1996) having simple interactions in terms of conditional independence. All of these models are amenable to present analysis. For example, with three factors on I, J, and K levels, the three graphical models with an appropriate polynomial form for the prior are shown below (a dot in the index denotes summing over the index).

More complex models are treated below as special cases of log-linear models. Several specific examples of explicit parametrizations of graphical models are in sections 3.2 and 3.3 of Drton et al. (2009). For directed graphical models, an explicit representation of the model is available in terms of the vanishing of homogeneous quadratic polynomials; here is a special case.

Example 8 (Directed graphical model) Let X_1, X_2, X_3, X_4 be binary variables with distributions restricted by the graph



Thus, probabilities are to be assigned so that

$$X_2 \perp X_3 \mid X_1, \quad X_4 \perp X_1 \mid X_2, X_3.$$

Results in Example 3.3.11 of Drton et al. (2009) yield the following quadratic in the sixteen parameters $(p_{0000}, \ldots, p_{1111})$

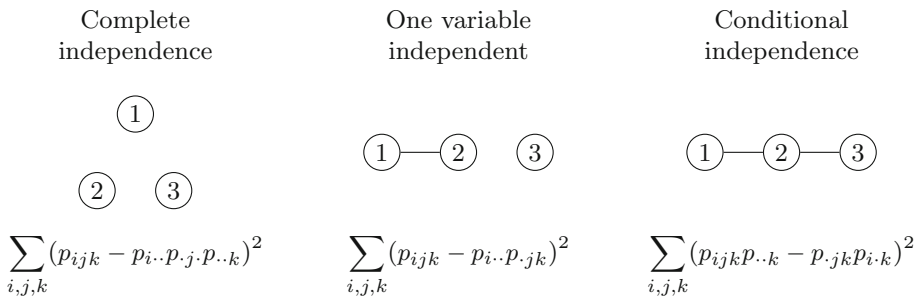


Table 2 Two by two table for Pearson's independence testing scenario

	Black hair	Blonde	Totals
Men	3	15	18
Women	7	5	12
Totals	10	20	30

to be used in the exponent of a prior on $[0,1]^{16}$ concentrating near this model:

$$(p_{000\cdot}p_{001\cdot} - p_{001\cdot}p_{010\cdot})^2 + (p_{100\cdot}p_{111\cdot} - p_{101\cdot}p_{110\cdot})^2 + (p_{0000}p_{1001} - p_{0001}p_{1000})^2 + (p_{0011}p_{1011} - p_{0011}p_{1010})^2 + (p_{0100}p_{1101} - p_{0101}p_{1100})^2 + (p_{0110}p_{1111} - p_{0111}p_{1110})^2. \quad (17)$$

Example 9 (Log-linear models) Some of the models above fall into the general class of log-linear models, equivalently, discrete exponential families. This classical subject is the start of algebraic statistics and we refer to Diaconis and Sturmfels (1998), Drton et al. (2009) for extensive development, examples, and reference.

Let $\mathcal{X}$ be a finite set, $T : \mathcal{X} \to \mathbb{N}^d$ a function. The exponential family through T is the family of probability measures on $\mathcal{X}$:

$$p_\theta(x) = Z^{-1}e^{\theta\cdot T(x)}, \quad x \in \mathcal{X}, \theta \in \mathbb{R}^d. \quad (18)$$

This includes some of the models above as well as many other models in wide-spread use in applied statistics (e.g., logistic regression). Here $\{p_\theta; \theta \in \mathbb{R}^d\}$ is a subfamily of $\mathcal{P}(\mathcal{X})$ – all probabilities on $\mathcal{X}$. The following development uses the tools of algebraic statistics Diaconis and Sturmfels (1998), Drton et al. (2009) to produce suitable quadratic exponents that vanish near $\{p_\theta; \theta \in \mathbb{R}^d\}$ in $\mathcal{P}(\mathcal{X})$.

For simplicity, assume throughout that $T(x)\cdot 1$ is equal to a constant for all $x \in \mathcal{X}$. Let $\{f_i(x) : \mathcal{X} \to \mathbb{Z}\}_{i\in\mathcal{I}}$ be a linear generating set for

$$\left\{ f : \mathcal{X} \to \mathbb{Z} ; \sum_{x\in\mathcal{X}} f(x)T(x) = 0 \right\}.$$

As described in Sect. 3 of Diaconis and Sturmfels (1998) and Chapter 1 of Drton et al. (2009), the Hilbert basis theorem shows that such generating sets exist with $|\mathcal{I}| < \infty$. They can be found using computational algebra (e.g., Gröbner basis) techniques. There is a large library of precomputed bases for standard examples. Observe that if $\sum_x f(x)T(x) = 0$, then $\prod_x p_\theta(x)^{f(x)} = 1$ (using the fact that $T(x)\cdot 1$ is constant). Write $f(x) = f_+(x) - f_-(x)$, with $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$. It follows that $\prod_x p_\theta(x)^{f_+(x)} - \prod_x p_\theta(x)^{f_-(x)} = 0$. These considerations imply the following.

Proposition 3.1 *Let $\mathcal{X}$ be a finite set, and let $T : \mathcal{X} \to \mathbb{N}^d$ have $T(x)\cdot 1$ constant. Let $\{f_i\}_{i\in\mathcal{I}}$ be a linear generating set for*

$$\left\{ f : \mathcal{X} \to \mathbb{Z} ; \sum_{x\in\mathcal{X}} f(x)T(x) = 0 \right\}.$$

Then, the polynomial Q in entries $p(x)$, $p \in \mathcal{P}(\mathcal{X})$,

$$Q(p(x); x \in \mathcal{X}) = \sum_{i\in\mathcal{I}} \left(\prod_{x\in\mathcal{X}} p(x)^{f_{+,i}(x)} - \prod_{x\in\mathcal{X}} p(x)^{f_{-,i}(x)} \right)^2 \quad (19)$$

vanishes at the model.

Remark 1 There are many generating sets available. As explained in Chapter 1.3 of Drton et al. (2009), these include lattice bases, Markov bases, Gröbner bases, universal Gröbner bases, and Graver bases. Any of these may be used. Parsimony suggests using the smallest $|\mathcal{I}|$. Even here, there is evident non-uniqueness; the pros and cons of the various choices have yet to be studied.

Remark 2 Literature Note: While there does not seem to be a convenient conjugate prior for the 2×2 table which allows for dependencies between p_1 and p_2, several non-conjugate priors have been suggested. (Howard 1998, section 7) develops the prior

$$e^{-\frac{1}{2}U^2}p_1^{\alpha-1}(1-p_1)^{\beta-1}p_2^{\gamma-1}(1-p_2)^{\delta-1} \quad \text{with} \quad U = \frac{1}{\sigma}\log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right).$$

Agresti and Min (Agresti and Min 2005, p. 520) compare this with a bivariate logit Normal prior. The present efforts can be viewed as a way of extending these from the 2×2 table to more general exponential families.

4 Computational methods

Pearson (1947) visited three different possible interpretations of testing when faced with a 2×2 contingency table such as the data in Table 1. In the introduction we presented his second scenario (II) in which data are two interchangeable samples for which we have a two-dimensional model. Now we turn to Pearson's scenario (III) where the data are considered as a sample of size 30 from a larger population classified

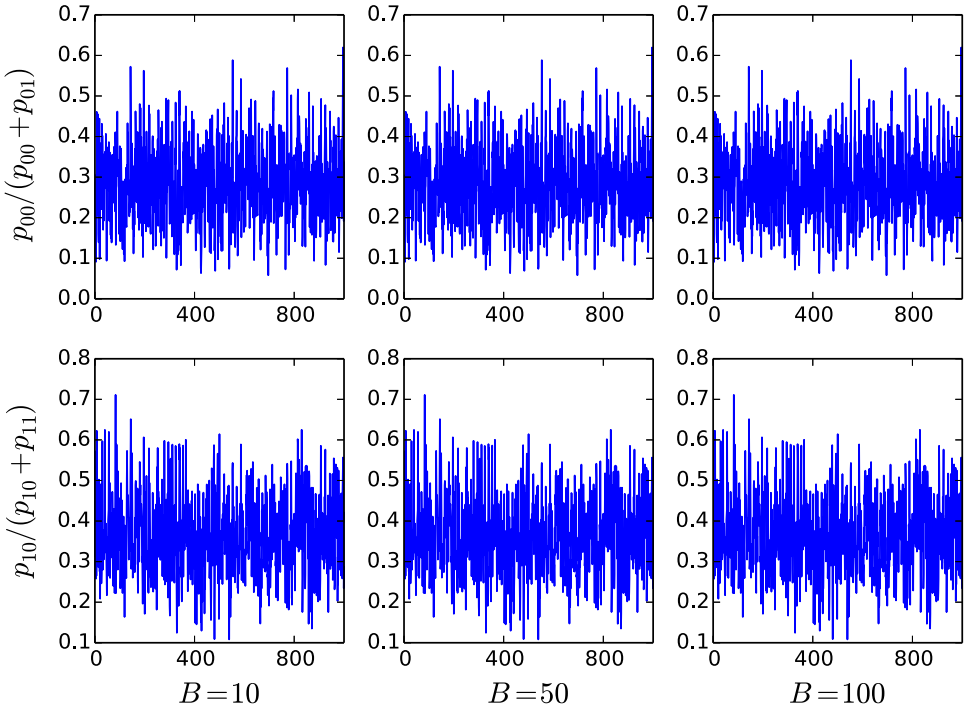


Fig. 4 The *top row* shows the probability of success for Sample 1 as a function of the MCMC iteration; the *bottom row* shows the same for Sample 2. The columns indicate the parameter B of the de Finetti prior. The unconstrained parameters $\tilde{p}$ were sampled by the Metropolis–Hastings algorithm with a log-normal proposal kernel

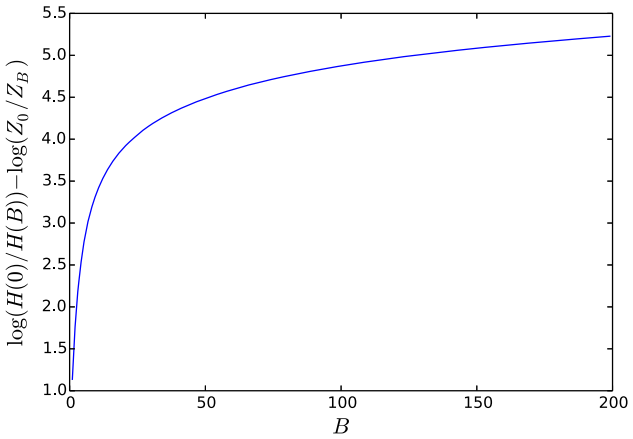


Fig. 5 The logarithmic Bayes factor comparing the model with $B = 0$ to a range of models with increasing parameter B for the example from Pearson (1947)

according to two attributes (for instance gender and hair color as in Table 2).

Now, a relevant model involves a multinomial distribution with four parameters labeled

$$(p_{00}, p_{01}, p_{10}, p_{11}).$$

Our goal in this section is to illustrate what is possible using standard Markov chain Monte Carlo techniques, which have been implemented in the Python package `de finetti`. The package abstracts discrete models like the ones in Sects. 2 and 3 using a symbolic representation of their characterizing polynomial $Q(p)$.

Consider the prior with density proportional to

$$e^{-AQ(p)} = e^{-A(p_{01}p_{10}-p_{11}p_{00})^2},$$

with respect to a measure $\mu(dp)$, which we assume in the Dirichlet family. The quadratic term vanishes on the surface of independence.

The scale of A is determined by the typical value of the polynomial Q, so it will be convenient to define

$$\lambda = \int Q(p)\mu(dp)$$

and set $A = B/\lambda$ to define a parameter B whose scale will be comparable from model to model.

A posterior distribution with density proportional to

$$e^{-\frac{B}{\lambda}Q(p)} \prod_{x\in\mathcal{X}} p_x^{n_x}$$

will be sampled via two variants of Markov chain Monte Carlo. In each case, it will be convenient to sample an unconstrained vector $\tilde{p}$, related to the original parameter via $p_x = \tilde{p}_x / \sum_{x\in\mathcal{X}} \tilde{p}_x$, which has posterior distribution

$$Z^{-1} e^{-\frac{B}{\lambda}Q(\tilde{p}/\sum_{x\in\mathcal{X}}\tilde{p}_x)} \prod_{x\in\mathcal{X}} \frac{\tilde{p}_x^{n_x+\ell_x}}{(\sum_{x'\in\mathcal{X}}\tilde{p}_{x'})^{n_x}} e^{-\tilde{p}_x} \prod_{x\in\mathcal{X}} d\tilde{p}_x,$$

for some normalization constant Z. Here ℓ_x are parameters in the Dirichlet prior.

The first algorithm is a Metropolis–Hastings Markov chain with a log-normal proposal kernel. The second algorithm is a simplified version of a Riemannian manifold

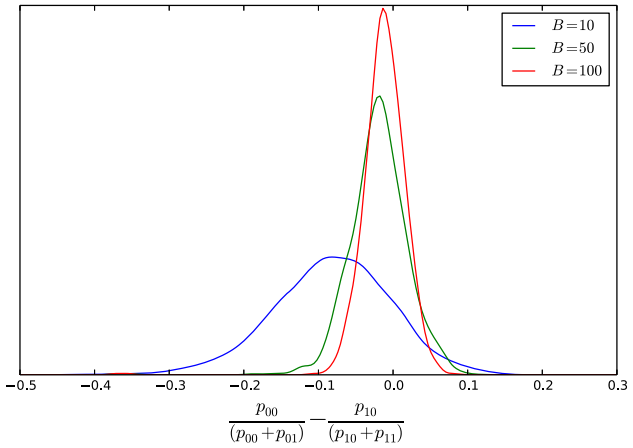


Fig. 6 Histogram of the difference between the probability of success in Sample 1 and Sample 2, for different values of the prior parameter B

Langevin dynamics Girolami and Calderhead (2011) with metric tensor $G(\tilde{p})$. As in Girolami and Calderhead, we define a Metropolis–Hastings Markov chain with a normal proposal kernel from $\tilde{p}$ with mean $\tilde{p}+\frac{\epsilon^2}{2}G(\tilde{p})^{-1}\nabla_{\tilde{p}}\log f(\tilde{p})$ and covariance $\epsilon^2 G(\tilde{p})^{-1}$, where f is the target posterior and the constant ϵ tunes the step size (Figs. 4 and 10).

The motivation for Riemannian manifold samplers lies in their ability to move quickly in spaces with strong correlations between variables, by virtue of making bigger proposals in directions of low curvature. Thus, a common choice for the metric tensor is the Hessian of the negative logarithm of the target posterior. Crucially, for a de Finetti posterior, the entries of this matrix are polynomials in $\tilde{p}$ which can be computed symbolically. Since this matrix is not guaranteed to be positive-definite, it will be regularized through the SoftAbs mapping introduced by Betancourt (2013).

Figure 5 shows a simulation of the posterior distribution for the 2 by 2 contingency table in example 2, using the simple Metropolis–Hastings algorithm with a log-normal proposal.

A key quantity for comparing de Finetti priors with different levels of concentration around the null model are Bayes factors of the form

$$\frac{\int Z_{B_1}^{-1} e^{-\frac{B_1}{\lambda}Q(p)} \prod_{x\in\mathcal{X}} p_x^{n_x}\mu(dp)}{\int Z_{B_2}^{-1} e^{-\frac{B_2}{\lambda}Q(p)} \prod_{x\in\mathcal{X}} p_x^{n_x}\mu(dp)}, \tag{20}$$

where Z_B is the normalization constant of the de Finetti prior with parameter B. This statistic can be computed through a standard thermodynamic integration algorithm implemented in the `de finetti` package.

Define

$$H(b) = \int e^{-\frac{b}{\lambda}Q(p)} \prod_{x\in\mathcal{X}} p_x^{n_x}\mu(dp).$$

Table 3 Number of errors observed in a DNA amplification experiment after applying the DADA amplicon denoising algorithm

	A	C	G	T
A	1, 810, 943	142	4836	325
C	119	1, 482, 244	43	1212
G	1568	69	1, 768, 575	171
T	235	3651	565	1, 394, 303

The rows indicate the true DNA base, and the columns indicate the DNA base read after amplification

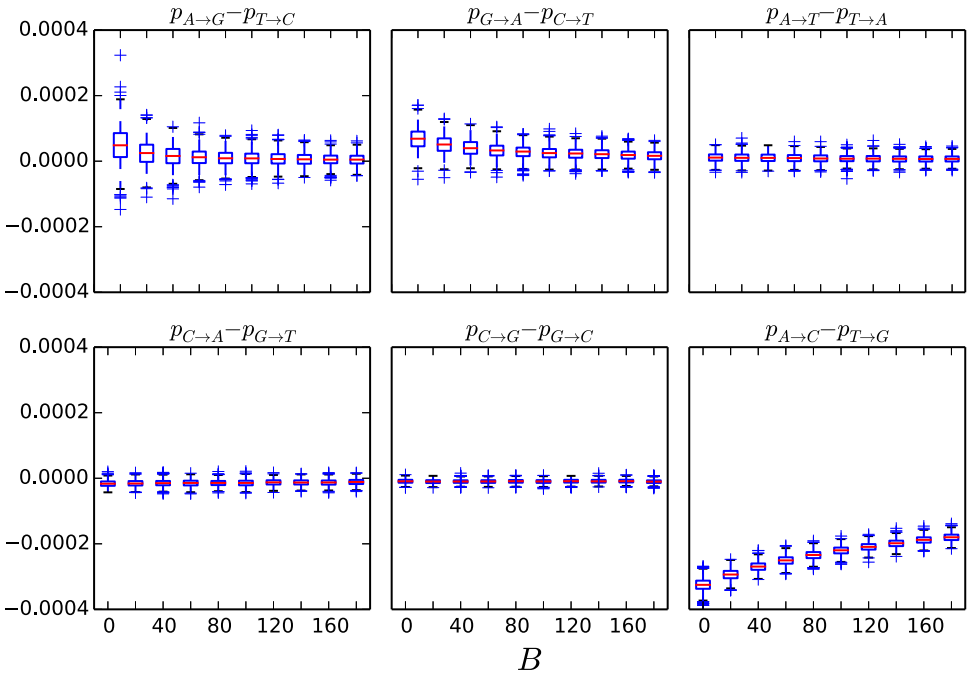


Fig. 7 DNA amplification example. The null model equates substitution probabilities for reverse-complementary pairs of DNA bases. The plot shows the logarithmic Bayes factor comparing the model with $B = 0$ to a range of models with increasing parameter B

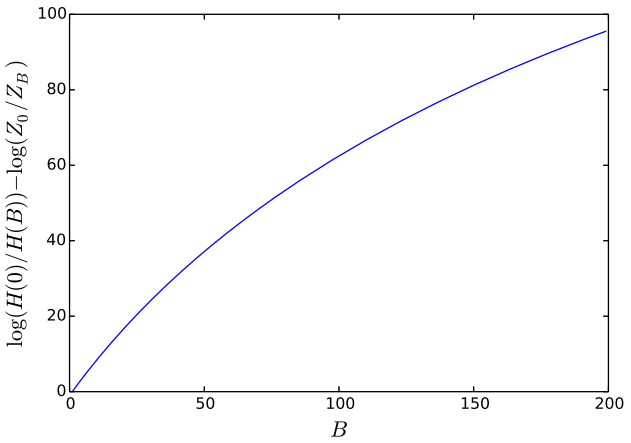


Fig. 8 DNA amplification example. The null model constrains the matrix of substitution probabilities to be reversible. The plot shows the logarithmic Bayes factor comparing the model with $B = 0$ to a range of models with increasing parameter B

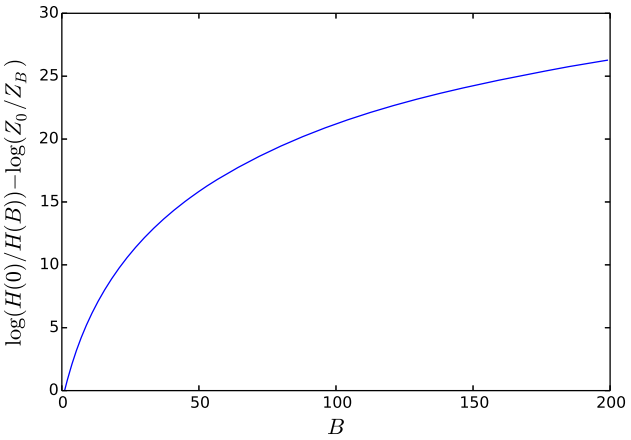


Fig. 9 Each panel corresponds to a reverse-complementary pair of errors in DNA amplification. The boxplots represent the posterior of the difference between the probability of one error and the other for increasing values of the prior parameter B

The basic identity

$$\frac{\mathrm{d}\log H(b)}{db} = \mathbb{E}^{(b)}_{\text{post}}\left[-\frac{Q(p)}{\lambda}\right],$$

where $\mathbb{E}^{(b)}_{\text{post}}$ denotes the expectation with respect to the posterior of p when the de Finetti prior has parameter b, allows us to write

$$\log\frac{H(B_1)}{H(B_2)} = \int_{B_1}^{B_2} \mathbb{E}^{(b)}_{\text{post}}\left[-\frac{Q(p)}{\lambda}\right] db.$$

The one-dimensional integral on the right hand side can be approximated numerically using Monte Carlo estimates of the integrand on a grid of values b ranging from B_1 to B_2. In the simplest variant of thermodynamic integration, this involves an application of the trapezoidal rule.

The logarithm of the Bayes factor in Eq. 20 can be written

$$\log\frac{H(B_1)}{H(B_2)} - \log\frac{Z_{B_1}}{Z_{B_2}},$$

where the second term can be approximated by thermodynamic integration as well, since $H(b)$ can be made equal to Z_b by setting $n_x = 0$ for all $x \in \mathcal{X}$.

Figure 5 compares the marginal likelihood of the data in Table 2 using de Finetti models with increasing levels of concentration around the independence hypothesis. The figure suggests there is weak evidence against independence, with a logarithmic Bayes factor of roughly 4 in favor of the uniform prior on p versus the model with strongest concentration around the independence hypothesis. Figure 6 plots the posterior distribution of the difference between the probability of success in Sample 1 and Sample 2, for $B = 25, 50, 100$. In an extensive discussion on the 2 by 2 contingency table Howard (1998), Howard uses this as a test statistic for the independence hypothesis. Naturally, increasing B concentrates the posterior distribution around 0.

5 Examples

5.1 Error rates in the polymerase chain reaction

DNA sequencing experiments require the amplification of genetic material in the cells through the polymerase chain reaction (PCR). This process copies each strand of DNA in the sample, but it is not entirely faithful; the reaction makes substitutions between nucleic bases A, C, G, and T with certain probabilities. A number of denoising algorithms are used to correct these substitutions. The contingency Table 3 counts the number of mistakes of each type made in a sequencing experiment; this table contrasts the bases read (columns) with the true bases estimated by the denoising algorithm DADA (rows) Rosen et al. (2012).

The first null hypothesis considered equates the probabilities of pairs of mistakes related by reverse complementarity. The PCR reaction copies each sequence to its complement, then the complement to the original, and so forth. For example, a specific letter G in the sequence will follow a series of transitions G→C→G→C→ .... So, a mistake G→A in the amplification can be caused by introducing an A in place of a G, or a T in place of a C along this process. This suggests that the error probabilities for reverse complementary pairs $p_{G\to A}$ and $p_{C\to T}$ are equal.

A de Finetti prior was put on this hypothesis using a polynomial $Q(p)$ that is a sum of squared differences for the mistake probabilities in each reverse complementary pair. The base measure $\mu(dp)$ in this example is a product of

uniform measures on the 4-simplex. Figure 9 shows box plots of the posterior for each difference, as a function of the prior parameter B. While the differences are generally small, in some cases such as $p_{A\to C} - p_{T\to G}$ the credible intervals suggest that the difference is different from 0. As the parameter B is increased, the prior begins to swamp the data and the differences are shrunk to 0. Figure 7 shows Bayes factors comparing the model with $B = 0$ to models with increasing values of B, which provide strong evidence against the equality in probability of reverse complementary mistakes.

A second null hypothesis was considered for these data. The matrix of mistake probabilities is stochastic, and it is interesting to consider whether it defines a reversible Markov chain. Kolmogorov's criterion for reversibility states that the probability of every cycle is the same in the forward and backward direction. It is sufficient to verify this criterion for a finite basis of cycles, for example by taking the cycles obtained by adding every missing edge to a spanning tree on the set of states. In our DNA sequencing example, we choose the polynomial

$$\begin{aligned}&(p_{A\to C}\ p_{C\to G}\ p_{G\to A} - p_{A\to G}\ p_{G\to C}\ p_{C\to A})^2\\ &\quad + (p_{A\to C}\ p_{C\to T}\ p_{T\to A} + p_{A\to T}\ p_{T\to C}\ p_{C\to A})^2\\ &\quad + (p_{A\to G}\ p_{G\to T}\ p_{T\to A} + p_{A\to T}\ p_{T\to G}\ p_{G\to A})^2.\end{aligned}$$

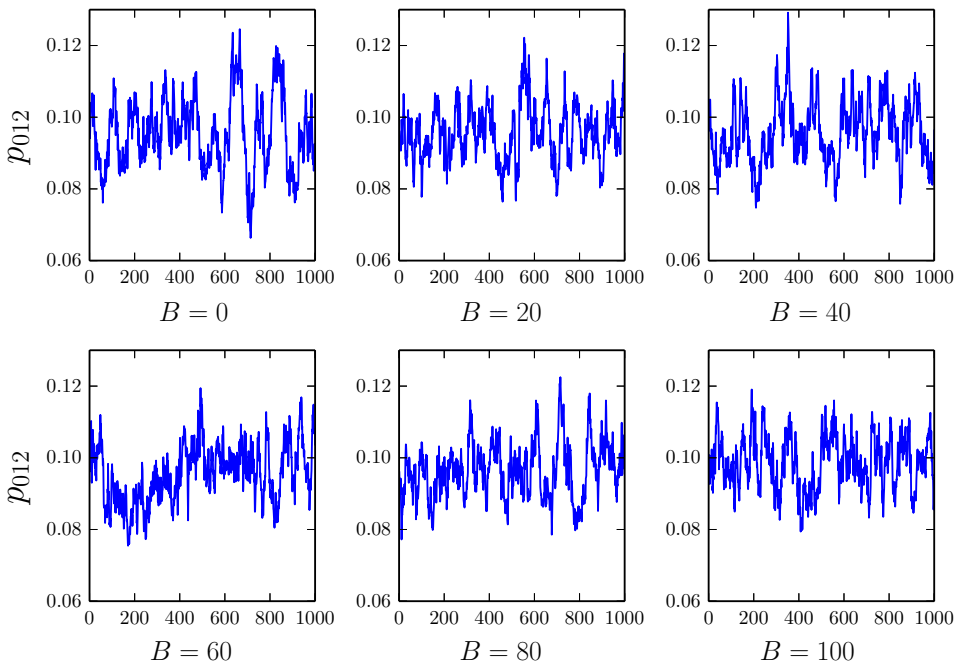


Fig. 10 The plots show the probability of an entry in the 3 by 3 by 3 contingency table as a function of the MCMC iteration, for 6 values of the prior parameter B. The parameter p_{012} is the probability of a Northern Protestant subject with a medium level of education and a negative attitude toward abortion. The prior is concentrated around the independence model, and the unconstrained parameters $\tilde{p} \in \mathbb{R}^{27}$ were sampled by a simplified Riemannian manifold Langevin algorithm

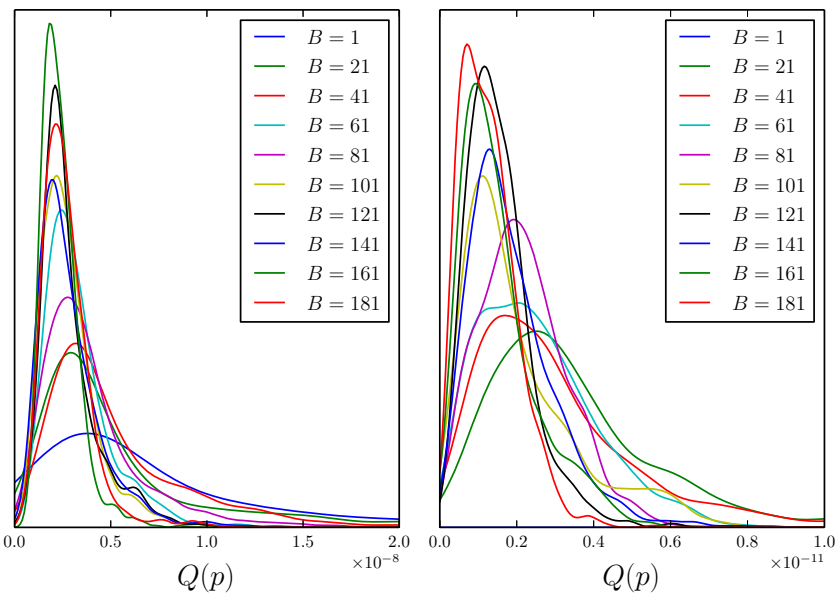


Fig. 11 Histogram of posterior samples of the polynomial $Q(p)$ characterizing the null model, for different values of the prior parameter B. *Left* independence model for the 3 by 3 by 3 contingency table on attitudes toward abortion. *Right* no three-way interaction model for the same data

Figure 8 shows Bayes factors computed under this prior, which provide strong evidence against reversibility.

5.2 Attitudes of white Christian subjects toward abortion

Haberman (1978) reports a survey from 1972 by the national research center on the attitudes toward abortion among 1055 white Christian subjects, who are assumed a simple random sample from the US population. Three categorical variables characterize each subject: religion (Northern Protestant, Southern Protestant, or Catholic), education level (less than 9 years, 9 to 12 years, or higher), and attitude to nontherapeutic abortion (positive, negative, or mixed).

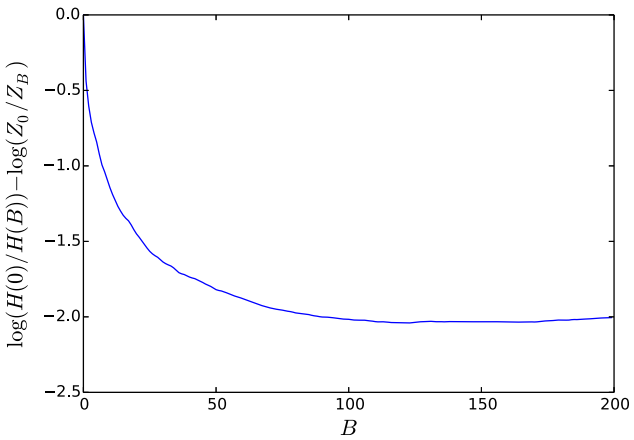


Fig. 12 Independence model for a 3 by 3 by 3 contingency table of education, religion, and attitudes toward abortion. The plot shows the logarithmic Bayes factor comparing the model with $B = 0$ to a range of models with increasing parameter B, i.e., increasing concentration around the independence hypothesis

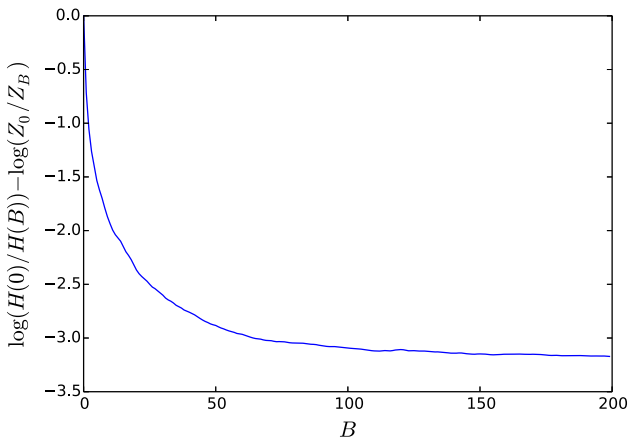


Fig. 13 No three-way interaction model for a 3 by 3 by 3 contingency table of education, religion, and attitudes toward abortion. The plot shows the logarithmic Bayes factor comparing the model with $B = 0$ to a range of models with increasing parameter B, i.e., increasing concentration around the null hypothesis

We test the hypotheses of independence and no three-way interaction on this $3 \times 3 \times 3$ contingency table. This model has 27 parameters, and we employed a simplified Riemannian manifold Langevin algorithm to sample the posterior distribution. Figure 10 shows some diagnostic plots for the sampler, in the de Finetti model around independence.

Figure 11 shows posterior histograms of the characterizing polynomial for each model, at different values of the prior parameter B. As expected, the posterior of this variable which measures the deviation from the hypothesis concentrates around 0 as B is increased. The Bayes factors in Figure 12 show weak evidence in favor of the independence hypothesis, and Figure 13 shows slightly stronger evidence in favor of the no three way interaction model, which is less restrictive.

Acknowledgments We thank Ben Callahan for discussions about the DNA denoising example. This work was partially funded by Grant NSF-DMS-1162538 to SH, Grant NSF-DMS-1208775 to PD and a CIMI fellowship that funded the travel of all three authors to Toulouse in 2014.

References

Agresti, A., Min, Y.: Frequentist performance of Bayesian confidence intervals for comparing proportions in 2× 2 contingency tables. Biometrics **61**(2), 515–523 (2005)

Ajdacic-Gross, V., Knöpfli, D., Landolt, K., Gostynski, M., Engelter, S.T., Lyrer, P.A., Gutzwiller, F., Rössler, W.: Death has a preference for birthdays—an analysis of death time series. Ann. Epidemiol. **22**(8), 603–606 (2012)

Andrews, D.F., Herzberg, A.M.: Data. Springer, New York (1985)

Betancourt, M.: A general metric for Riemannian manifold Hamiltonian Monte Carlo. In: Geometric Science of Information, pp. 327–334. Springer, New York (2013)

Darroch, J.N., Lauritzen, S.L., Speed, T.P.: Markov fields and log-linear interaction models for contingency tables. Ann. Stat **8**, 522–539 (1980)

Diaconis, P., Efron, B.: Testing for independence in a two-way table: New interpretations of the chi-square statistic. Ann. Stat **13**(3), 845–874 (1985)

Diaconis, P., Freedman, D.: On the uniform consistency of Bayes estimates for multinomial probabilities. Ann. Stat **18**(3), 1317–1327 (1990)

Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann Stat **26**(1), 363–397 (1998)

Drton, M., Sturmfels, B., Sullivant, S.: Lectures on algebraic statistics. Springer, Basel (2009)

de Finetti, B.: Sur la condition d'équivalence partielle (1938)

de Finetti, B.: Probability, induction and statistics: The art of guessing. Wiley, New York (1972)

de Finetti, B.: On the condition of partial exchangeability. Stud. Inductive Log. Probab. **2**, 193–205 (1980)

Ghosh, J., Sinha, B., Joshi, S.: Expansions for posterior probability and integrated Bayes risk. Stat. Decis. Theory Relat. Top. III **1**, 403–456 (1982)

Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. R Stat. Soc. **73**(2), 123–214 (2011)

Good, I.J.: The Estimation of Probabilities: An Essay on Modern Bayesian Methods, vol. 258. MIT press Cambridge, Cambridge (1965)

Goodman, L.A.: The multivariate analysis of qualitative data: Interactions among multiple classifications. J. Am. Stat. Assoc. **65**(329), 226–256 (1970)

Haberman, S.J.: Analysis of Qualitative Data. vol. 1: IntroductoryTtopics. Academic Press, New York (1978)

Howard, J.: The 2× 2 table: A discussion from a Bayesian viewpoint. Stat. Sci. **13**, 351–367 (1998)

Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)

Pearson, E.S.: The choice of statistical tests illustrated on the interpretation of data classed in a 2× 2 table. Biometrika **34**, 139–167 (1947)

Phillips, D.P.: Deathday and birthday - unexpected connection. In: Statistics: a guide to the unknown, pp. 71–85. Holden-Day Series in Probability and Statistics (1978)

Rosen, M.J., Callahan, B.J., Fisher, D.S., Holmes, S.P.: Denoising pcr-amplified metagenome data. BMC Bioinform. **13**(1), 283 (2012)