

Article

# Partial Exchangeability for Contingency Tables

Persi Diaconis

Department of Mathematics and Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305, USA; diaconis@math.stanford.edu

**Abstract:** A parameter free version of classical models for contingency tables is developed along the lines of de Finetti’s notions of partial exchangeability.

**Keywords:** algebraic statistics; contingency tables; de Finetti representation theorem; Markov basis; partial exchangeability

## 1. Introduction

Consider cross-classified data:  $X_1, X_2, \dots, X_n$ , where  $X_a = (i_a, j_a)$ ,  $i_a \in [I]$ ,  $j_a \in [J]$  (for  $[I] = \{1, 2, \dots, I\}$ ). Such data are often presented as an  $I \times J$  contingency table  $T = (t_{ij})$  where  $t_{ij}$  is the number of times  $(i, j)$  happens. Suppose that  $X_1, \dots, X_n$  are exchangeable and extendible. Then, de Finetti’s theorem says:

**Theorem 1.** For exchangeable  $\{X_i\}_{i=1}^\infty$  taking values in  $[I] \times [J]$

$$P[X_1 = (i_1, j_1), \dots, X_n = (i_n, j_n)] = \int_{\Delta_{I \times J}} \prod_{i,j} p_{ij}^{t_{ij}} \mu(dp),$$

where  $\Delta_{I \times J} = \{p_{ij} \geq 0, \sum_{i,j} p_{ij} = 1\}$ . The representing measure  $\mu$  is unique.

A popular model for cross classified data is

$$p_{ij} = \theta_i \eta_j.$$

Here is a Bayesian, parameter free, description.

**Theorem 2.** For exchangeable  $\{X_i\}_{i=1}^\infty$  taking values in  $[I] \times [J]$ , a necessary and sufficient condition for the mixing measure  $\mu$  in Theorem 1 to be supported on  $\Delta_I \times \Delta_J$  (with  $\Delta_I = \{p_1, \dots, p_I : p_i \geq 0, \sum_i p_i = 1\}$ ), so

$$P[X_1 = (i_1, j_1), \dots, X_n = (i_n, j_n)] = \int_{\Delta_I \times \Delta_J} \prod \theta_i^{t_i^*} \eta_j^{t_j^*} \mu(d\theta, d\eta),$$

is that

$$P[X_1 = (i_1, j_1), X_2 = (i_2, j_2), X_3 = (i_3, j_3), \dots, X_n = (i_n, j_n)] = P[X_1 = (i_1, j_2), X_2 = (i_2, j_1), X_3 = (i_3, j_3), \dots, X_n = (i_n, j_n)]. \quad (1)$$

Condition (1) is to hold for any  $n \geq 2$  and any  $(i_a, j_a)$   $1 \leq a \leq n$ .

**Proof.** Condition (1) implies for all  $n$  and  $h \geq 1$  (surpressing  $P$  a.s. throughout)

$$P[X_1 = (i_1, j_1), X_2 = (i_2, j_2) | X_n = (i_n, j_n), \dots, X_{n+h} = (i_{n+h}, j_{n+h})] = P[X_1 = (i_1, j_2), X_2 = (i_2, j_1) | X_n = (i_n, j_n), \dots, X_{n+h} = (i_{n+h}, j_{n+h})]. \quad (2)$$



**Citation:** Diaconis, P. Partial Exchangeability for Contingency Tables. *Mathematics* **2022**, *10*, 442. <https://doi.org/10.3390/math10030442>

Academic Editors: Emanuele Dolera and Federico Bassetti

Received: 30 December 2021

Accepted: 20 January 2022

Published: 29 January 2022

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Let  $h \uparrow \infty$  and then  $n \uparrow \infty$ . Let  $\mathcal{T}$  be the tail field of  $\{X_i\}_{i=1}^\infty$ . Then, Doob’s increasing and decreasing martingale theorems show

$$P[X_1 = (i_1, j_1), X_2 = (i_2, j_2) | \mathcal{T}] = P[X_1 = (i_1, j_2), X_2 = (i_2, j_1) | \mathcal{T}].$$

However, a standard form of de Finetti’s theorem says that, given  $\mathcal{T}$ , the  $\{X_i\}_{i=1}^\infty$  are i.i.d. with  $P[X_1 = (i, j)] = p_{ij}$ . Thus

$$p_{ij}p_{i'j'} = p_{ij'}p_{i'j} \quad \text{for all } i, i', j, j'. \tag{3}$$

Finally, observe that (3) implies (writing  $p_{i*} := \sum_j p_{ij}$ ,  $p_{*j} := \sum_i p_{ij}$ )

$$p_{i*}p_{*j} = \sum_{h,l} p_{ih}p_{lj} = \sum_{hl} p_{ij}p_{hl} = p_{ij}.$$

□

We remark the following points.

1. If  $X_i = (Y_i, Z_i)$  condition (2) is equivalent to

$$\mathcal{L}((Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)) = \mathcal{L}((Y_1, Z_{\sigma(1)}), \dots, (Y_n, Z_{\sigma(n)}))$$

for all  $n$  and  $\sigma \in S_n$  ( $S_n$  is the symmetric group over  $1, 2, \dots, n$ ). Since  $\{(Y_i, Z_i)\}_{i=1}^n$  are exchangeable this is equivalent to saying the law is invariant under  $S_n \times S_n$ .

2. The mixing measure  $\mu(d\theta, d\eta)$  allows general dependence between the row parameters  $\theta$  and column parameters  $\eta$ . Classical Bayesian analysis of contingency tables often chooses  $\mu$  so that  $\theta$  and  $\eta$  are independent. A parameter free version is that under  $P$ , the row sums  $t_{i*}$  and column sums  $t_{*j}$  are independent. It is natural to weaken this to “close to independent” along the lines of [1] or [2]. See also [3].
3. Theorems 1 and 2 have been stated for discrete state spaces. By a standard discretization argument, they hold for quite general spaces. For example:

**Theorem 3.** Let  $X_i = (Y_i, Z_i)$  be exchangeable with  $Y_i \in \mathcal{Y}$ ,  $Z_i \in \mathcal{Z}$ , complete separable metric spaces,  $1 \leq i < \infty$ . Suppose

$$P[X_1 \in (A_1, B_1), X_2 \in (A_2, B_2), \dots, X_n \in (A_n, B_n)] = P[X_1 \in (A_1, B_2), X_2 \in (A_2, B_1), \dots, X_n \in (A_n, B_n)]$$

for all measurable  $A_i, B_i$  and all  $n$ . Then,

$$P(X_1 \in (A_1, B_1), \dots, X_n \in (A_n, B_n)) = \int_{\mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Z})} \prod_1^n \theta(A_i)\eta(B_i)\mu(d\theta, d\eta),$$

with  $\mathcal{P}(\mathcal{Y}), \mathcal{P}(\mathcal{Z})$  the probabilities on the Borel sets of  $\mathcal{Y}, \mathcal{Z}$ . The mixing measure  $\mu$  is unique.

4. Theorem 2 is closely related to de Finetti’s work in [1,4].
5. De Finetti’s law of large numbers holds as well, in Theorem 3

$$\frac{1}{n} \sum \delta_{X_i}(A \times B) \rightarrow \mu(\theta(A), \eta(B)).$$

One object of this paper is to develop similar parameter free de Finetti theorems for widely used log-linear models for discrete data. Section 2 begins by relating this to an ongoing conversation with Eugenio Regazzini. Section 3 provides needed background on discrete exponential families and algebraic statistics. Sections 4 and 5 apply those tools to give de Finetti style partially exchangeable theorems for some widely used hierarchical and graphical models for contingency tables. Section 6 shows how these exponential

family tools can be used for other Bayesian tasks: building “de Finetti priors” for “almost exchangeability” and running the “exchange” algorithm for doubly intractable Bayesian computation. Some philosophy and open problems are in the final section.

### 2. Some History

I was lucky enough to be able to speak at Eugenio Regazzini’s 60<sup>TH</sup> birthday celebration, in Milan, in 2006. My talk began this way:

« Hello, my name is Persi and I have a problem. »

For those of you not aware of the many “10 step-programs” (alcoholics anonymous, gamblers anonymous, ...) they all begin this way, with the participants admitting to having a problem. In my case the problem was this:

- (a) After 50 years of thinking about it, I think that the subjectivist approach to probability, induction and statistics is the only thing that works;
- (b) At the same time, I have done a lot of work inventing and analyzing various schemes for generating random samples for things like contingency tables with given row and column sums; graphs with given degree sequences; ...; Markov Chain Monte Carlo. These are used for things like permutation tests and Fisher’s exact test.

There is a lot of nice mathematics and hard work in (b) but such tests violate the likelihood principle and lead to poor scientific practice. Hence my problem (I still have it): (a) and (b) are incompatible.

There has been some progress. I now see how some of the tools developed for (b) can be usefully employed for natural tasks suggested by (a). Not so many people care about such inferential questions in these ‘big data’ days. However, there are also lots of small datasets where the inferential details matter. There are still useful questions for people like Eugenio (and me).

### 3. Background on Exponential Families and Algebraic Statistics

The following development is closely based on [5], which should be considered for examples, proofs and more details.

Let  $\mathcal{X}$  be a finite set. Consider the exponential family:

$$p_\theta(x) = \frac{1}{Z(\theta)} e^{\theta \cdot T(x)} \quad \theta \in \mathbb{R}^d, x \in \mathcal{X}. \tag{4}$$

Here,  $Z(\theta)$  is a normalizing constant and  $T : \mathcal{X} \rightarrow \mathbb{N}^d - \{0\}$ . If  $X_1, X_2, \dots, X_n$  are independent and identically distributed from (4), the statistic  $t = T(X_1) + \dots + T(X_n)$  is sufficient for  $\theta$ . Let

$$\mathcal{Y}_t = \{(x_1, \dots, x_n) : T(x_1) + \dots + T(x_n) = t\}.$$

Under (4), the distribution of  $X_1, \dots, X_n$  given  $t$  is uniform on  $\mathcal{Y}_t$ . It is usual to write

$$t = \sum_{i=1}^n T(X_i) = \sum_{\mathcal{X}} \sigma(x) T(x) \quad \text{with } \sigma(x) = \#\{i : T(X_i) = T(x)\}.$$

Let

$$\mathcal{F}_t = \{f : \mathcal{X} \rightarrow \mathbb{N} : \sum f(x) T(x) = t\}.$$

**Example 1.** For contingency tables  $\mathcal{X} = \{(i, j) : 1 \leq i \leq I, 1 \leq j \leq J\}$ . The usual model for independence has  $T(i, j) \in \mathbb{N}^{I+J}$  a vector of length  $I + J$  with two non zero entries equal 1. The 1’s in  $T(i, j)$  are in the  $i^{\text{th}}$  place and position  $j$  of the last  $j$  places. The sufficient statistic  $t$  contains the row and column sums of the contingency table associated to the first  $n$  observations. The set  $\mathcal{F}_t$  is the set of an  $I \times J$  tables with these row and column sums.

A Markov chain on this  $\mathcal{F}_t$  can be based on the following moves: pick  $i \neq i', j \neq j'$  and change the entries in the current  $f$  by adding  $\pm 1$  in pattern

$$\begin{matrix} & j & j' \\ i & + & - \\ i' & - & + \end{matrix} \quad \text{or} \quad \begin{matrix} - & + \\ + & - \end{matrix}$$

This does not change the row sums and it does not change the column sums. If told to go negative, just pick new  $i, i', j, j'$ . This gives a connected, aperiodic Markov chain on  $\mathcal{F}_t$  with a uniform stationary distribution. See [6].

Returning to the general case, an analog of  $\begin{matrix} + & - \\ - & + \end{matrix}$  moves is given by the following:

**Definition 1** (Markov basis). A Markov basis is a set of functions  $f_1, f_2, \dots, f_L$  from  $\mathcal{X}$  to  $\mathbb{Z}$  such that

$$\sum_{\mathcal{X}} f_i(x)T(x) = 0 \quad 1 \leq i \leq L \tag{5}$$

and that for any  $t$  and  $f, f' \in \mathcal{F}_t$  there are  $(t_1, f_{i_1}), \dots, (t_A, f_{i_A})$  with  $t_i = \pm 1$ , such that

$$f' = f + \sum_{j=1}^A t_j f_{i_j} \quad \text{and} \quad f + \sum_{j=1}^a t_j f_{i_j} \geq 0, \text{ for } 1 \leq a \leq A. \tag{6}$$

This allows the construction of a Markov chain on  $\mathcal{F}_t$ : from  $f$ , pick  $I \in \{1, 2, \dots, L\}$  and  $t = \pm 1$  at random and consider  $f + t f_I$ . If this is positive, move there. If not, stay at  $f$ . Assumptions (5) and (6) ensure that this Markov chain is symmetric and ergodic with a uniform stationary distribution. Below, I will use a Markov basis to formulate a de Finetti theorem to characterize mixtures of the model (4).

One of the main contributions of [5] is a method of effectively constructing Markov bases using polynomial algebra. For each  $x \in \mathcal{X}$ , introduce an indeterminate, also called  $x$ . Consider the ring of polynomials  $k[\mathcal{X}]$  in these indeterminates where  $k$  is a field, e.g., the complex numbers. A function  $g : \mathcal{X} \rightarrow \mathbb{N}$  is represented as a monomial  $\mathcal{X}^g = \prod_{\mathcal{X}} x^{g(x)}$ . The function  $T : \mathcal{X} \rightarrow \mathbb{N}^d$  gives a homomorphism

$$\begin{aligned} \varphi_T : k[\mathcal{X}] &\longrightarrow k[t_1, \dots, t_d] \\ x &\longmapsto t_1^{T_1(x)} t_2^{T_2(x)} \dots t_d^{T_d(x)}, \end{aligned}$$

extended linearly and multiplicatively ( $\varphi_T(x + y) = \varphi_T(x) + \varphi_T(y)$  and  $\varphi_T(x^2) = \varphi_T(x)^2$  and so on). The basic object of interest is the kernel of  $\varphi_T$ :

$$I_T = \{p \in k[\mathcal{X}] : \varphi_T(p) = 0\}.$$

This is an ideal in  $k[\mathcal{X}]$ . A key result of [5] is that a generating set for  $I_T$  is equivalent to a Markov basis. To state this, observe that any  $f : \mathcal{X} \rightarrow \mathbb{Z}$  can be written  $f = f_+ - f_-$  with  $f_+(x) = \max(f(x), 0)$  and  $f_-(x) = \max(-f(x), 0)$ . Observe  $\sum f(x)T(x) = 0$  iff  $\mathcal{X}^{f_+} - \mathcal{X}^{f_-} \in I_T$ . The key result is

**Theorem 4.** A collection of functions  $f_1, f_2, \dots, f_L$  is a Markov basis if and only if the set

$$\mathcal{X}^{f_{i+}} - \mathcal{X}^{f_{i-}} \quad 1 \leq i \leq L$$

generates the ideal  $I_T$ .

Now, the Hilbert Basis Theorem shows that ideals in  $k[\mathcal{X}]$  have finite bases and modern computer algebra packages give an effective way of finding bases.

I do not want (or need) to develop this further. See [5] or the book by Sullivant [7] or Aoki et al. [8]. There is even a Journal of Algebraic Statistics.

I hope that the above gives a flavor for what I mean by “working in (b) is hard honest work”. Most of the applications are for standard frequentist tasks. In the following sections, I will give Bayesian applications.

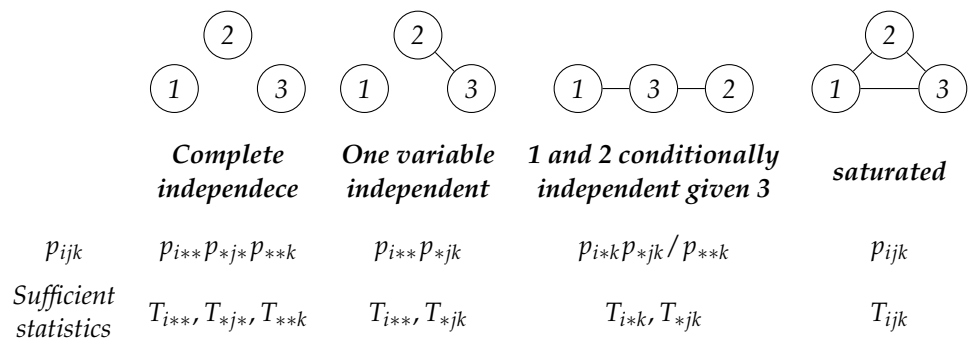
#### 4. Log Linear Model for Contingency Tables

Log linear models for multiway contingency tables are a healthy part of the modern statistics. The index set is  $\mathcal{X} = \prod_{\gamma \in \Gamma} I_\gamma$  with  $\Gamma$  indexing categories and  $I_\gamma$  the levels of  $\gamma$ . Let  $p(x)$  be the probability of falling into cell  $x \in \mathcal{X}$ . A log linear model can be specified by writing:

$$\log p(x) = \sum_{a \subseteq \Gamma} \varphi_a(x).$$

The sum ranges over subsets  $a$  of  $\Gamma$  and  $\varphi_a(x)$  means a function that only depends on  $x$  through the coordinates in  $a$ . Thus,  $\varphi_\emptyset(x)$  is a constant and  $\varphi_\Gamma(x)$  is allowed to depend on all coordinates. Specifying  $\varphi_a = 0$  for some class of sets  $a$  determines a model. Background and extensive references are in [9]. If the  $a$  with  $\varphi_a \neq 0$  permitted form a simplicial complex  $\mathcal{C}$  (so  $a \in \mathcal{C}$  and  $\emptyset \neq a' \subseteq a \Rightarrow a' \in \mathcal{C}$ ) the model is called *hierarchical*. If  $\mathcal{C}$  consists of the cliques in a graph, the model is called *graphical*. If the graph is chordal (every cycle of length  $\geq 4$  contains a chord) the graphical model is called *decomposable*.

**Example 2** (3 way contingency tables). *The graphical models for three way tables are:*



The simplest hierarchical model that is not graphical is *No Three Way Interaction Model*. This can be specified by saying ‘the odds rate of any pair of variables does not depend on the third’. Thus,

$$\frac{p_{ijk}p_{i'j'k}}{p_{ij'k}p_{i'jk}} \text{ is constant in } k \text{ for fixed } i, i', j, j'. \tag{7}$$

As one motivation, recall that for two variables, the independence model is specified by

$$p_{ij} = \theta_i \eta_j.$$

For three variables, suppose there are parameters  $\theta_{ij}, \eta_{jk}, \psi_{ik}$  satisfying:

$$p_{ijk} = \theta_{ij} \eta_{jk} \psi_{ik} \quad \text{for all } i, j, k. \tag{8}$$

It is easy to see that (8) entails (7) hence ‘no three way interaction’. Cross multiplying (7) entails

$$p_{ijk}p_{i'j'k}p_{ij'k'}p_{i'jk'} = p_{ijk'}p_{i'j'k}p_{ij'k}p_{i'jk}. \tag{9}$$

This is the form we will work with for the de Finetti theorems below.

For background, history and examples (and some nice theorems) see ([10], Section 8.2), [11,12], Simpsons ‘paradox’ [13] is based on understanding the no three way interaction model. Further discussion is in Section 5 below.

**5. From Markov Bases to de Finetti Theorems**

Suppose  $\mathcal{X}$  is a finite set,  $T : \mathcal{X} \rightarrow \mathbb{N}^d - \{0\}$  is a statistic and  $\{f_i\}_{i=1}^L$  is a Markov basis as in Section 3. The following development shows how to translate this into de Finetti theorems for the contingency table examples of Section 4. The first argument abstracts the argument used for Theorem 2 above.

**Lemma 1 (Key Lemma).** *Let  $\mathcal{X}$  be a finite set and  $\{X_i\}_{i=1}^\infty$  an exchangeable sequence of  $\mathcal{X}$ -valued random variables. Suppose for all  $n > m$*

$$P[X_1 = x_1, \dots, X_m = x_m, X_{m+1} = x_{m+1}, \dots, X_n = x_n] = P[X_1 = y_1, \dots, X_m = y_m, X_{m+1} = x_{m+1}, \dots, X_n = x_n]. \tag{10}$$

In (10),  $x_1, \dots, x_m, y_1, \dots, y_m$  are fixed and  $x_{m+1}, \dots, x_n$  are arbitrary. Then, if  $\mathcal{T}$  is the tail field of  $\{X_i\}_{i=1}^\infty$  and  $p(x) = P[X_1 = x | \mathcal{T}]$ ,

$$\prod_{i=1}^m p(x_i) = \prod_{i=1}^m p(y_i). \tag{11}$$

**Proof.** From (10) and exchangeability

$$P[X_1 = x_1, \dots, X_m = x_m, X_{n+1} = x_{n+1}, \dots, X_{n+h} = x_{n+h}] = P[X_1 = y_1, \dots, X_m = y_m, X_{n+1} = x_{n+1}, \dots, X_{n+h} = x_{n+h}]$$

so

$$P[X_1 = x_1, \dots, X_m = x_m | X_{n+1} = x_{n+1}, \dots, X_{n+h} = x_{n+h}] = P[X_1 = y_1, \dots, X_m = y_m | X_{n+1} = x_{n+1}, \dots, X_{n+h} = x_{n+h}].$$

Let  $h \uparrow \infty$  and then  $n \uparrow \infty$ , use Doob’s upward and then downward martingale convergence theorems to see:

$$P[X_1 = x_1, \dots, X_m = x_m | \mathcal{T}] = P[X_1 = y_1, \dots, X_m = y_m | \mathcal{T}].$$

Now, de Finetti’s theorem implies (11).  $\square$

**Remark 1.** *The Key Lemma shows that the  $p(x)$  satisfy certain relations. Using choices of  $\{x_i\}, \{y_i\}$  derived from a Markov basis will show that  $p(x)$  satisfy the required independence properties. Suppose that  $\sum_{\mathcal{X}} f(x)T(x) = 0, \sum_{\mathcal{X}} f(x) = 0$  and  $f \in \{0, \pm 1\}$ . Let  $S_+ = \{x : f(x) = 1\}, S_- = \{y : f(y) = -1\}$ . Say  $|S_+| = |S_-| = m$ . Enumerate  $S_+ = \{x_1, \dots, x_m\}, S_- = \{y_1, \dots, y_m\}$ . Assumptions (10) and conclusion (11) will give our theorems.*

**Example 3 (Independence in a two way table).** *Let  $\mathcal{X} = [I] \times [J]$ . A minimal basis for the independence model is given by  $f_{i,j,i',j'}$ :*

$$\begin{array}{c|cc} & j & j' \\ \hline i & + & - \\ i' & - & + \end{array} \quad (\text{all other entries} = 0).$$

The condition of the Key Lemma becomes:

$$P[X_1 = (i, j), X_2 = (i', j'), X_3 = (i_3, j_3), \dots, X_n = (i_n, j_n)] =$$

$$P[X_1 = (i, j'), X_2 = (i', j), X_3 = (i_3, j_3), \dots, X_n = (i_n, j_n)].$$

Passing to the limit gives

$$p_{ij}p_{i'j'} = p_{ij'}p_{i'j}$$

and so

$$p_{i*}p_{*j} = \sum_{i'j'} p_{ij'}p_{i'j} = p_{ij}.$$

This is precisely Theorem 2 of the Introduction.  $\square$

**Example 4** (Complete independence in a three way table). The sufficient statistics are  $T_{i**}, T_{*j*}, T_{**k}$ . From [5], there are two kinds of moves in a minimal basis. Up to symmetries, these are:

Class I				Class II			
	$j$	$j'$			$j$	$j'$	
$i$	+	-		$i$	+	-	
$i'$	-	+		$i'$	-	+	

Passing to the limit, this entails:

$$p_{ijk}p_{i'j'k} = p_{ij'k}p_{i'jk} \text{ and } p_{ijk}p_{i'j'k'} = p_{ij'k}p_{i'jk'}.$$

These may be said as 'the product of any  $p_{ijk}, p_{i'jk}$  remains unchanged if the middle coordinates are exchanged'. By symmetry, this remains true if the two first or last coordinates are exchanged. As above, this entails

$$p_{i**}p_{*j*}p_{**k} = p_{ijk}.$$

These observations can be rephrased into a statement that looks more similar to the classical de Finetti theorem; using symmetry:

**Theorem 5.** Let  $\{X_i\}_{i=1}^\infty$  be exchangeable, taking values in  $[I] \times [J] \times [K]$ . Then

$$P[X_1 = (i_1, j_1, k_1), \dots, X_n = (i_n, j_n, k_n)] = P[X_1 = (\sigma(i_1), \zeta(j_1), \eta(k_1)), \dots, X_n = (\sigma(i_n), \zeta(j_n), \eta(k_n))]$$

for all  $n, \{(i_a, j_a, k_a)\}_{a=1}^n$  and  $(\sigma, \zeta, \eta) \in S_I \times S_J \times S_K$  is necessary and sufficient for there to exist a unique  $\mu$  on  $\Delta_I \times \Delta_J \times \Delta_K$  with

$$P[X_a = (i_a, j_a, k_a), 1 \leq a \leq n] = \int_{\Delta_I \times \Delta_J \times \Delta_K} \prod_{a=1}^n p_{i_a} q_{j_a} r_{k_a} \mu(dp, dq, dr).$$

$\square$

**Example 5** (One variable independent of the other two). Suppose, without loss, that the graph is



Identify the pairs  $(j, k)$  with  $\{1, 2, \dots, L\}$  with  $L = JK$ . The problem reduces to Example 4. A minimal basis consists of (again, up to relabeling)

	$l$	$l'$
$i$	+	-
$i'$	-	+

We may conclude

**Theorem 6.** Let  $\{X_i\}_{i=1}^\infty$  be exchangeable, taking values in  $[I] \times [J] \times [K]$ . Then

$$P[X_1 = (i_1, j_1, k_1), \dots, X_n = (i_n, j_n, k_n)] = P[X_1 = (\sigma(i_1), \zeta(j_1, k_1)), \dots, X_n = (\sigma(i_n), \zeta(j_n, k_n))]$$

for all  $n$ ,  $\{(i_a, j_a, k_a)\}_{a=1}^n$  and  $(\sigma, \zeta) \in S_I \times S_{J \times K}$  is necessary and sufficient for there to exist a unique  $\mu$  on  $\Delta_I \times \Delta_{JK}$  with

$$P[X_a = (i_a, j_a, k_a), 1 \leq a \leq n] = \int_{\Delta_I \times \Delta_{JK}} \prod_{a=1}^n p_a q_a \mu(dp, dq).$$

□

**Example 6** (Conditional independence). Suppose variable  $i$  and  $j$  are conditionally independent given  $k$ .



Rewrite the parameter condition of section four as

$$p_{***} p_{ijk} = p_{i*k} p_{*jk} \text{ for all } i, j, k$$

The sufficient statistics are  $\{T_{i*k}\}_{i,k}, \{T_{*jk}\}_{j,k}$ . From [5], a minimal generating set is

	$j_k$	$j'_k$
$i_k$	+	-
$i'_k$	-	+

$$K \times \frac{I(I-1)}{2} \times \frac{J(J-1)}{2} \text{ moves in all.}$$

From this, the Key Lemma shows (for all  $i, j, k$ )

$$p_{ijk} p_{i'j'k} = p_{ij'k} p_{i'jk}.$$

This entails:

$$p_{i*k} p_{*jk} = \sum_{i',j'} p_{ij'k} p_{i'jk} = \sum_{i',j'} p_{ij'k} p_{i'jk} = p_{ijk} p_{***}.$$

Again, phrasing the condition (10) in terms of symmetry.

**Theorem 7.** Let  $\{X_i\}_{i=1}^\infty$  be exchangeable, taking values in  $[I] \times [J] \times [K]$ . Then,

$$P[X_1 = (i_1, j_1, k_1), \dots, X_n = (i_n, j_n, k_n)] = P[X_1 = (\sigma^{k_1}(i_1), \zeta^{k_1}(j_1), k_1), \dots, X_n = (\sigma^{k_n}(i_n), \zeta^{k_n}(j_n), k_n)] \quad (12)$$

for all  $n$ ,  $\{(i_a, j_a, k_a)\}_{a=1}^n$  and  $\sigma^k, \zeta^k \in S_I \times S_J, 1 \leq k \leq K$ , is necessary and sufficient for there to exist a unique family  $\mu \times \prod_{b=1}^k \mu_{b,r}$  on  $\Delta_K \times (\Delta_I \times \Delta_J)^K$

$$P[X_a = (i_a, j_a, k_a), 1 \leq a \leq n] = \int_{\Delta_K \times (\Delta_I \times \Delta_J)^K} \prod_{a=1}^n r_{k_a} p_{i_a}^{k_a} q_{j_a}^{k_a} \prod_{b=1}^k \mu_{b,r}(p^{i_b} q^{j_b}) \mu(dr). \quad (13)$$

Both (12) and (13) have a simple interpretation. For (12),  $\{X_i\}_{i=1}^n$  are exchangeable 3-vectors. For any  $k$  and specified sequence of values  $\{(i_a, j_a, k)\}_{a=1}^n$  the chance of observing these values is unchanged under permuting the  $(i_a, j_a, k)$ , by permutations  $\sigma^k \in S_I, \zeta^k \in S_J$ . Here  $\sigma^k, \zeta^k$  are allowed to depend on  $k$ .



On the right of (13), the mixing measure may be understood as follows. There is a probability  $\mu$  on  $\Delta_K$ . Pick  $r = (r_1, \dots, r_k) \in \Delta_K$ . Given this  $r$ , pick  $(p^k, q^k)$  from  $\mu_{k,r}$  on the  $k^{\text{th}}$  copy of  $\Delta_I \times \Delta_J$ . These choices are allowed to depend on  $r$  but are independent, conditional on  $r, 1 \leq k \leq K$ .

All of this simply says that, conditional on the tail field,

$$P[X_a = (i, j, k) | \mathcal{T}] = P[X_a = (i, *, k) | \mathcal{T}]P[X_a = (*, j, k) | \mathcal{T}].$$

The first two coordinates are conditionally independent given the third.

**Example 7** (No three way interaction). *The model is described in Section 4. The sufficient statistics are  $\{T_{ij*}\}, \{T_{i*k}\}, \{T_{*jk}\}$ . Minimal Markov bases have proved intractable. See [5] or [8]. For any fixed  $I, J, K$ , the computer can produce a Markov basis but these can have a huge number of terms. See [7,8] and their references for a surprisingly rich development.*

*There is a pleasant surprise. Markov bases are required to connect the associated Markov chain. There is a natural subset, the first moves anyone considers, and these are enough for a satisfactory de Finetti theorem (!).*

*Described informally, for an  $I \times J \times K$  array, pick a pair of parallel planes, say the  $k, k'$  planes in the three dimensional array, and consider moves depicted as*

$$\begin{array}{c|cc} & j & j' \\ \hline i & + & - \\ i' & - & + \\ \hline & k & k' \end{array}$$

*These moves preserve all line sums (the sufficient statistics). They are **not** sufficient to connect any two datasets with the same sufficient statistics. Using the prescription in the Key Lemma, suppose:*

$$\begin{aligned} P[X_1 = (i, j, k), X_2 = (i', j', k), X_3 = (i, j', k'), X_4 = (i', j, k'), \\ X_a = (i_a, j_a, k_a) \ 5 \leq a \leq n] = \\ P[X_1 = (i, j', k), X_2 = (i', j, k), X_3 = (i, j, k'), X_4 = (i', j', k'), \\ X_a = (i_a, j_a, k_a) \ 5 \leq a \leq n]. \end{aligned} \tag{14}$$

Passing to the limit gives

$$p_{ijk}p_{i'j'k}p_{ij'k'}p_{i'jk'} = p_{ij'k}p_{i'jk}p_{ijk'}p_{i'j'k'}. \tag{15}$$

This is exactly the no three way interaction condition. Or, equivalently:

$$\frac{p_{ijk}p_{i'j'k}}{p_{ij'k}p_{i'jk}} = \frac{p_{ij'k'}p_{i'jk'}}{p_{ij'k'}p_{i'jk'}}.$$

The odds ratios are constant on the  $k^{\text{th}}$  and  $k'^{\text{th}}$  planes (of course, they depend on  $i, j, i', j'$ ). These considerations imply:

**Theorem 8.** *Let  $\{X_i\}_{i=1}^\infty$  be exchangeable, taking values in  $[I] \times [J] \times [K]$ . Then, condition (14) is necessary and sufficient for the existence of a unique probability  $\mu$  on  $\Delta_{IJK}$ , supported on the no three way interaction variety (15) satisfying*

$$P[X_a = (i_a, j_a, k_a), 1 \leq a \leq n] = \int_{\Delta_{IJK}} \prod p_{ijk}^{\eta_{ijk}} \mu(dp_{ijk}).$$

We remark on the following points.

1. It follows from theorems in [12] and [11] that, if all  $p_{ijk} > 0$ , condition (15) is equivalent to the unique representation,

$$p_{ijk} = r\alpha_{jk}\beta_{ki}\gamma_{ij}, \tag{16}$$

where  $r, \alpha, \beta, \gamma$  have positive entries and satisfy

$$\sum_k \alpha_{jk} = \sum_i \beta_{ki} = \sum_j \gamma_{ij} = 1 \text{ for all } i, j, k$$

and

$$r \sum_{i,j,k} \alpha_{jk}\beta_{ki}\gamma_{ij} = 1.$$

The integral representation in the theorem can be stated in this parametrization. The condition  $p_{ijk} > 0$  is equivalent to  $P(X_1 = (i, j, k)) > 0$  on observables.

2. Condition (14) does not have an obvious symmetry interpretation.
3. Conditions (14) and (15) are stated via varying the third variable when  $i, j, i', j'$  are fixed. Because of (16), if they hold in this form, they hold for any two variables fixed as the third varies.
4. It is possible to go on, but, as John Darroch put it, ‘the extensions to higher order interactions... are not likely to be of practical interest’. The most natural development—the generalization to decomposable models—is being developed by Paula Gablenz.
5. There are many extensions of the Key Lemma above. These allow a similar development for more general log linear models and exponential families.

### 6. Discussion and Conclusions

The tools of algebraic statistics have been harnessed above to develop partial exchangeability for standard contingency table models. I have used them for two further Bayesian tasks: approximate exchangeability and the problem of ‘doubly intractable priors’. As both are developed in papers, I will be brief.

Approximate exchangeability. Consider  $n$  men and  $m$  women along with a binary outcome. If the men are judged exchangeable (for fixed outcomes for the women) and vice versa, and, if both sequences are extendable, de Finetti [1] shows that there is a unique prior on the unit square  $[0, 1]^2$  such that, for any outcomes  $t_1, \dots, t_n, \sigma_1, \dots, \sigma_m$  in  $\{0, 1\}$

$$P[X_1 = t_1, \dots, X_n = t_n, Y_1 = \sigma_1, \dots, Y_m = \sigma_m] = \int_{[0,1]^2} p^S (1-p)^{n-S} \theta^T (1-\theta)^{m-T} \mu(dp, d\theta),$$

with  $S = \sum_{i=1}^n t_i, T = \sum_{j=1}^m \sigma_j$ .

If, for the outcome of interest,  $\{X_i, Y_j\}$  were almost fully exchangeable (so the men/women difference is judged practically irrelevant) the prior  $\mu$  would be concentrated near the diagonal of  $[0, 1]^2$ . De Finetti suggested implementing this by considering priors of the form

$$\mu(dp, d\theta) = Z^{-1} e^{-A(p-\theta)^2} dp d\theta$$

for  $A$  large.

In joint work with Sergio Bacallado and Susan Holmes [3], multivariate versions of such priors are developed. These are required to concentrate near sub-manifolds of cubes or products of simplices; think about ‘approximate no three way interaction’. We used the tools of algebraic statistics to suggest appropriate many variable polynomials which vanish on submanifold of interest. Many ad hoc choices were involved. Sampling from such priors or posteriors is a fresh research area. See [2,14,15].

Doubly intractable priors. Consider an exponential family as in Section 3:

$$p_\theta(x) = \frac{1}{Z(\theta)} e^{\theta \cdot T(x)}.$$

Here  $x \in \mathcal{X}$  a finite set,  $T : \mathcal{X} \rightarrow \mathbb{R}^d$  and  $\theta \in \mathbb{R}^d$ . In many real examples, the normalizing constant  $Z(\theta)$  will be unknown and unknowable. For a Bayesian treatment, let  $\Pi(d\theta)$  be a prior distribution on  $\mathbb{R}^d$ . For example, the conjugate prior.

If  $X_1, X_2, \dots, X_n$  is as i.i.d. sample from  $p_\theta$ ,  $T$  is a sufficient statistic and the posterior has the form

$$Z(Z^{-1}(\theta))^n e^{\theta F} \Pi(d\theta),$$

with  $F = \sum_{i=1}^n T(X_i)$  and  $Z$  another normalizing constant. The problem is that  $Z^{-1}(\theta)$  depends on  $\theta$  and is unknown!

The exchange algorithm and many variants offer a useful solution. See [16,17].

In practical implementations, there is an intermediary step requiring a sample from  $p_{\theta}^T$ , the measure induced by  $p_{\theta}^n$  under  $\sum_i^n T(x_i) : \mathcal{X}^n \rightarrow \mathbb{R}$ . This is a discrete sampling task and Markov basis techniques have been proved useful. See [16].

*A philosophical comment.* The task undertaken above, finding believable Bayesian interpretations for widely used log linear models, goes somewhat against the grain of standard statistical practice. I do not think anyone takes a reasonably complex, high dimensional hierarchical model seriously. They are mostly used as a part of exploratory data analysis; this is not to deny their usefulness. Making any sense of a high dimensional dataset is a difficult task. Practitioners search through huge collections of models in an automated way. Usually, any reflection suggests the underlying data is nothing like a sample from a well specified population. Nonetheless, models are compared using product likelihood criteria. It is a far far cry from being based on anyone’s reasoned opinion.

I have written elsewhere about finding Bayesian justification for important statistical tasks such as graphical methods or exploratory data analysis [18]. These seem like tasks similar to ‘how do you form a prior’. Different from the focus of even the most liberal Bayesian thinking.

*The sufficiency approach.* There is a different approach to extending de Finetti’s theorem. This uses ‘sufficiency’. Consider exchangeable  $\{X_i\}_{i=1}^\infty$ . For each  $n$ , suppose  $T_n : \mathcal{X}^n \rightarrow \mathcal{Y}$  is a function. The  $\{T_n\}$  have to fit together according to simple rules satisfied in all of the examples above. Call  $\{X_i\}$  *partially exchangeable with respect to  $T_n$*  if  $P[X_1 = x_1, \dots, X_n = x_n | T_n = t_n]$  is uniform. Then, Diaconis and Freedman [19] show that a version of de Finetti’s theorem holds. The law of  $\{X_i\}$  is a mixture of i.i.d. laws indexed by extremal laws. In dozens of examples, these extremal laws can be identified with standard exponential families. This last step remains to be carried out in the generality of Section 3 above. What is required is a version of the Koopman–Pitman–Darmois theorem for discrete random variables. This is developed in [19] when  $\mathcal{X} \subseteq \mathbb{N}$  and  $T_n(X_1, \dots, X_n) = X_1 + \dots + X_n$ . Passing to interpretation, this version of partial exchangeability has the following form:

$$\begin{aligned} \text{if } T_n(x_1, \dots, x_n) &= T_n(y_1, \dots, y_n), \\ \text{then } P[X_1 = x_1, \dots, X_n = x_n] &= P[X_1 = y_1, \dots, X_n = y_n]. \end{aligned}$$

This is neat mathematics (and allows a very general theoretical development). However, it does not seem as easy to think about in natural examples. Exchangeability via symmetry is much easier. The development above is a half-way house between symmetry and sufficiency. A close relative of the sufficiency approach is the topic of ‘extremal models’ as developed by Martin-Löf and Lauritzen. See [20] and its references. Moreover, Refs. [21,22] are recent extensions aimed at contingency tables.

*Classical Bayesian contingency table analysis.* There is a healthy development of parametric analysis for the examples of Section 5. This is based on natural conjugate priors. It includes nice theory and R packages to actually carry out calculations in real problems. Three papers that I like are [23–26]. The many wonderful contributions by I.J. Good are still very much worth consulting. See [27] for a survey. Section 5 provides ‘observable characterizations’ of the models. The problem of providing ‘observable characterizations’ of the associated conjugate priors (along the lines of [28]) remains open.

**Funding:** This research received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 817257 and funding from NSF grant No 1954042.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The author would like to thank Paula Gablenz, Sourav Chatterjee and Emanuele Dolera for help throughout.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

- de Finetti, B. On the condition of partial exchangeability. *Stud. Inductive Log. Probab.* **1980**, *2*, 193–205.
- Bruno, A. On the notion of partial exchangeability (Italian). In *Giornale dell’Istituto Italiano degli Attuari*; English Translation in: de Finetti, *Probability, Induction and Statistics*; International Statistical Institute: Leidschenveen, The Netherlands, 1964 ; Volume 27, Chapter 10; pp. 174–196.
- Bacalado, S.; Diaconis, P.; Holmes, S. De Finetti priors using Markov chain Monte Carlo computations. *J. Stat. Comput.* **2015**, *25*, 797–808. [[CrossRef](#)] [[PubMed](#)]
- de Finetti, B. *Probability, Induction and Statistics: The Art of Guessing*; Wiley: Hoboken, NJ, USA, 1972.
- Diaconis, P.; Sturmfels, B. Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.* **1998**, *26*, 363–397. [[CrossRef](#)]
- Diaconis, P.; Gangolli, A. Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms*; Springer: New York, NY, USA, 1995 ; Volume 72, pp. 15–41.
- Sullivant, S. *Algebraic Statistics*; AMS: Providence, RI, USA, 2018 .
- Aoki, S.; Hara, H.; Takemura, A. *Markov Bases in Algebraic Statistics*; Springer: Berlin/Heidelberg, Germany, 2012.
- Lauritzen, S.L. *Graphical Models*, 2nd ed.; Oxford University Press: Oxford, UK, 2004 .
- Agresti, A. *Categorical Data Analysis*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2002.
- Birch, M.W. Maximum likelihood in three-way contingency tables. *J. R. Stat. Soc. Ser. B* **1963**, *25*, 220–233. [[CrossRef](#)]
- Darroch, J.N. Interactions in multi-factor contingency tables. *J. R. Stat. Soc. Ser.* **1962**, *24*, 251–263. [[CrossRef](#)]
- Simpson, E.H. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Ser.* **1951**, *13*, 238–241. [[CrossRef](#)]
- Diaconis, P.; Holmes, S.; Shahshahani, M. Sampling From a Manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton* ; IMS Statistics Collections: Beachwood, OH, USA, 2013 ; pp. 102–125.
- Gerencsér, B.; Ottolini, A. Rates of convergence for Gibbs sampling in the analysis of almost exchangeable data. *arXiv* **2020**, arXiv:2010.15539v2.
- Diaconis, P.; Wang, G. Bayesian goodness of fit tests: A conversation for David Mumford. *Ann. Math. Sci. Appl.* **2018**, *3*, 287–308. [[CrossRef](#)]
- Wang, G. On the Theoretical Properties of the Exchange Algorithm. *arXiv* **2021**, arXiv:2005.09235v4.
- Diaconis, P. Theories of data analysis: From magical thinking through classical statistics. In *Exploring Data Tables, Trends, and Shapes*; Hoaglin, D.C., Mosteller, F., Tukey, J.W., Eds.; Wiley: Hoboken, NJ, USA, 1985; pp. 1–36.
- Diaconis, P.; Freedman, D. Partial Exchangeability and Sufficiency. In *Statistics: Applications and New Directions*; Ghosh, K., Roy, F. Eds.; Indian Statistical Institute: Calcutta, India, 1984; pp. 205–236
- Lauritzen, S.L. General Exponential Models for Discrete Observations. *Scand. J. Stat.* **1975**, *2*, 23–33.
- Lauritzen, S.L.; Rinaldo, A.; Sadeghi, K. Random Networks, Graphical Models, and Exchangeability. *arXiv* **2017**, arXiv:1701.08420v2.
- Lauritzen, S.L.; Rinaldo, A.; Sadeghi, K. On exchangeability in network models. *J. Algebr. Stat.* **2019**, *10*, 85–114. [[CrossRef](#)]
- Albert, J.H.; Gupta, A.K. Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Stat.* **1982**, *10*, 1261–1268. [[CrossRef](#)]
- Murray, I.; Ghahramani, Z.; MacKay, D.J.C. MCMC for doubly-intractable distributions. In Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence (UAI ’06), Cambridge, MA, USA, 13–16 July 2006.
- Letac, G.; Massam, H. Bayes factors and the geometry of discrete hierarchical loglinear models. *Ann. Stat.* **2012**, *40*, 861–890. [[CrossRef](#)]
- Tarantola, C.; Ntzoufras, I. Bayesian Analysis of Graphical Models of Marginal Independence for Three Way Contingency Tables. In *Quaderni di Dipartimento from University of Pavia*; No 172; Department of Economics and Quantitative Methods, University of Pavia: Pavia, Italy, 2012. Available online: <http://dem-web.unipv.it/web/docs/dipeco/quad/ps/RePEc/pav/wpaper/q172.pdf> (accessed on 30 December 2021 ).
- Diaconis, P.; Efron, B. Testing for independence in a two-way table: New interpretations of the Chi-Square statistic. *Ann. Stat.* **1985**, *13*, 845–874. [[CrossRef](#)]
- Diaconis, P.; Ylvisaker, D. Quantifying prior opinion. In *Bayesian Statistics, II*; Bernardo, J., DeGroot, M., Lindley, D., Smith, A.F.M., Eds.; North-Holland: Amsterdam, The Netherlands, 1985 ; pp. 133–156.