Open Science Framework DOI:10.31219/osf.io/***

Combining conditional exchangeability and Bayesian networks

P.G.L. Porta Mana
Kavli Institute, Trondheim <pgl@portamana.org>

14 May 2020; updated 5 June 2020

[draft] A formula is given for conditionally, infinitely exchangeable probability distributions.

0 Introduction and notation

De Finetti's theorem for exchangeable distributions yields some of the formulae, derivable from the probability calculus, with the richest practical and philosophical consequences.It mathematically connects, under specific assumptions,

- what can be interpreted as the frequencies in a superpopulation,
- our uncertainty about those frequencies,
- the empirical frequencies in collected data,
- the probability for individual instances.

In this study I present three results. Each result builds on the preceding one(s), and each may interest a different audience.

The first result, derived in § 1, is a reformulation of partial exchangeability and of its representation theorem in a slightly unfamiliar form. Though not remarkable, this reformulation gives some insights into the connection between partial and conditional exchangeability.

The second result, derived is § 2, is an integral representation for joint predictive distributions with particular symmetries: some of the conditional distributions obtained from them satisfy partial exchangeability. This integral representation has the usual de Finetti form but its density must factorize in a specific way.

The third result, § 3, brings together exchangeability and Bayesian networks. As this is the main result, the rest of this introduction focuses on it.

A Bayesian network is a graphical representation of judgements about the informational relevance or irrelevance of some statements to other statements. Such relevance or irrelevance is expressed by implicit equalities between conditional probabilities involving those statements. The graph of the network is accompanied by tables of conditional distributions.

Unfortunately the use of Bayesian networks is a little ambiguous in the literature. Sometimes their application seems to refer to an *individual* instance, and therefore to involve degrees of belief. For example our degree of belief about the effect of a treatment upon a specific person, given the person's age, gender, health condition, and other factors. Sometimes their application seems instead to refer to a whole *population* or even superpopulation, and therefore to involve 'long-run' frequencies. And sometimes their application seems to unsystematically shift between both points of view.[1]

The individual and population applications are not equivalent. In particular, the distributions they involve are numerically different. In general we cannot apply to a population the conditional-probability tables of a Bayesian network constructed for an individual, nor vice versa. De finetti's theorem shows why: empirical frequencies, superpopulation frequencies, and probabilities for individual instances are numerically different. They may become approximately equal after updating on very large amounts of data. But in practical applications, especially in medicine and social sciences, we do *not* have large amounts of data. The distinction must be kept.

Several questions thus arise in applying a Bayesian network. In the application to individuals, how are our degrees of belief about a specific instance related to, or updated by, knowledge about similar instances? In the application to populations, what is our uncertainty about the (usually unknowable) superpopulation frequencies? and what is our uncertainty about the empirical frequencies of future observations of small groups of individuals? These questions are all facets of the same question.

And these are precisely the questions addressed by exchangeability and its theorems. Exchangeability expresses judgements about 'similarity'; more precisely, about the relevance (or irrelevance) of statements concerning individual events (observations, experiments, and so on) to

[1] See e.g. Pearl 2009; Wiegerinck et al. 2013; the discussion by Lindley & Novick 1981 is more precise in this respect: they assume, for simplicity, a limit in which the two problems become numerically similar.

sets of statements concerning other individual events. Such relevance is expressed by symmetries of conditional and unconditional probabilities involving those statements. The result is a mathematical formula that, as already mentioned, quantitatively relates empirical frequencies, superpopulation frequencies and their uncertainty, and degrees of beliefs about individual instances.

It seems therefore necessary to mathematically combine Bayesian networks and exchangeability. Such combination is explored in § 3. A formula is given that connects the individual and population applications of a Bayesian network. The formula has the typical structure of the de Finetti integral representation: informally speaking, the conditional distributions of the Bayesian network for an individual instance appear as mixtures of the distributions for the superpopulation. The structure of the network is preserved, but its preservation depends on specific assumptions about conditional exchangeability.

The remainder of this section introduces some notation and summarizes de Finetti's theorem for full exchangeability. It can be skimmed through by readers familiar with exchangeability theorems, only to grasp the notation I use.

0.1 Notation and summary of representation for full exchangeability

For the details about exchangeable distributions I refer to Bernardo & Smith[2], Diaconis & Freedman[3], and Dawid's[4] review.

Our domain of discourse consists of a countably infinite set of atomic statements (in the logical sense)

$$\{X_i = x_i \mid i \in \mathbf{N},\ \forall i\ x_i \in \mathfrak{X}\} \tag{1}$$

where $\mathfrak{X}$ is a finite set. For each i the statements $\{X_i = x \mid x \in \mathfrak{X}\}$ are assumed mutually exclusive on information I. (The theorem holds for any set of statements with these properties, even if the statements are not of the form '$X = x$'.)

[2] Bernardo & Smith 2000 §§ 4.3, 4.6. [3] Diaconis & Freedman 1980a,b. [4] Dawid 2013.

A probability distribution over these atomic statements is called fully (infinitely) exchangeable if

$$\left.\begin{aligned}&\text{for every } N\text{, every subset } \{i_1,\dots,i_N\}\subset\mathbf{N}\text{, every permutation } \pi\\ &\text{thereof, and every set of values } x_{i_1},\dots,x_{i_N}\in\mathfrak{X}\text{:}\\ &\mathrm{P}(X_{i_1}=x_{i_1}, X_{i_2}=x_{i_2},\dots,X_{i_N}=x_{i_N}\mid I) =\\ &\qquad\qquad\mathrm{P}(X_{i_1}=x_{\pi(i_1)}, X_{i_2}=x_{\pi(i_2)},\dots,X_{i_N}=x_{\pi(i_N)}\mid I)\\ &\text{and all such probabilities are consistently related by marginalization.}\end{aligned}\right\}\tag{2}$$

This property is equivalent to declaring the empirical frequencies of the values x to be sufficient statistics.

In the following, a comma and '$\wedge$' will both denote logical conjunction. The notation $\bigwedge_i$ will express conjunction with the index i running over a subset of N elements from $\mathbf{N}$.

Denote by $\boldsymbol{f}_{\boldsymbol{x}} \coloneqq (f_x)$ a normalized distribution over the values $x \in \mathfrak{X}$. The set of all such distributions is a simplex of dimension $|\mathfrak{X}| - 1$.

For each $x \in \mathfrak{X}$, denote by F_x the empirical relative frequency of x in the set $\{x_i\}$:

$$NF_x \coloneqq \textstyle\sum_i \delta(x, x_i), \quad x \in \mathfrak{X}\,. \tag{3}$$

De Finetti's theorem states that a fully exchangeable distribution can be written as follows:

$$\begin{aligned}\mathrm{P}\big(\bigwedge_i X_i = x_i \,\big|\, I\big) &= \int \prod_i f_{x_i}\; \mathrm{p}(\boldsymbol{f}_{\boldsymbol{x}} \mid I)\,\mathrm{d}\boldsymbol{f}_{\boldsymbol{x}}\\ &\equiv \int \prod_x {f_x}^{NF_x}\; \mathrm{p}(\boldsymbol{f}_{\boldsymbol{x}} \mid I)\,\mathrm{d}\boldsymbol{f}_{\boldsymbol{x}}\,,\end{aligned}\tag{4}$$

where the integral is over the simplex of distributions $\{\boldsymbol{f}_{\boldsymbol{x}}\}$. In the first integral form, the product is over the set of instances $\{i\}$. In the second, equivalent integral form, the product is over the set of values x. This form shows that the empirical frequency distribution (F_x) is a sufficient statistic. It also hints at the important role played in the theorem by the relative entropy of (F_x) with respect to (f_x).

The theorem establishes a one-one correspondence between the set of (nested) joint predictive distributions and the set of densities over the $(|\mathfrak{X}| - 1)$-dimensional simplex. Thus any special properties of a predictive distribution must be reflected in special properties of its corresponding density, and vice versa.

For enough large N, the probability of observing an empirical frequency distribution $\boldsymbol{F_x}$ within a small volume v centred around the distribution $\boldsymbol{f_x}$ is approximately given by the density $\mathrm{p}(\boldsymbol{f_x} \mid I)\,\mathrm{d}\boldsymbol{f_x}$:

$$\mathrm{P}(\boldsymbol{F_x} \in v \mid N \text{ large}, I) \approx \mathrm{p}(\boldsymbol{f_x} \mid I)\, v\ . \tag{5}$$

For this reason the parameter $\boldsymbol{f_x}$ can be interpreted as a long-run frequency distribution[5], or the frequency in a 'superpopulation'. I will therefore call it so sometimes, but without the intention of forcing such interpretation on you.

1 Partial exchangeability: alternative form

ref to *unrestricted exchangeability* of Bernardo & Smith[6]

In de Finetti's theorem for *partially* exchangeable distributions, the set $\{X_i = x_i\}$ of § 0.1 is divided into two or more categories represented by subsets $\{Y_j = y_j\}$, $\{Z_k = z_k\}$, ... Partial exchangeability of the distribution for such statements means that permutations are allowed within each subset but not necessarily across subsets. The integral representation in this case is usually given in the form

$$\mathrm{P}\Bigl(\bigwedge_{i'} Y_{i'} = y_{i'},\ \bigwedge_{i''} Z_{i''} = z_{i''} \Bigm| J\Bigr) = \iint \prod_j g_{y_j} \prod_k h_{z_k}\ \mathrm{p}(\boldsymbol{g}, \boldsymbol{h} \mid J)\,\mathrm{d}\boldsymbol{g}\,\mathrm{d}\boldsymbol{h}\ , \tag{6}$$

with distinct normalized distributions $\boldsymbol{g}$, $\boldsymbol{h}$ for each category. If the density $\mathrm{p}(\boldsymbol{g}, \boldsymbol{h} \mid J)\,\mathrm{d}\boldsymbol{g}\,\mathrm{d}\boldsymbol{h}$ is diagonal, that is, if it contains a term $\delta(\boldsymbol{g} - \boldsymbol{h})$, the fully exchangeable form (4) is recovered.

A little reflection shows that if we know the X_i to belong to category A in instances i', and to category B in instances i'', then (i) there are some other quantities C_i that allows us to distinguish the two categories, and (ii) the values of these quantities *are known* for all instances.

Let us say, for example, that the statements $\{X_i = x_i\}$ refer to the results of behavioural experiments with insects, with x-values 'S'uccess

[5] 'But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead' (Keynes 2013 § 3.I p. 65). [6] Bernardo & Smith 2000.

and 'F'ailure. A refers to the results for experiments with 'A'nts, and B with 'B'eetles. If we write

$$\mathrm{P}(A_3 = \mathsf{S},\ B_5 = \mathsf{F} \mid J) = 0.2\ ,$$

then we must already know that insect number 3 is an ant: $C_3 = \mathsf{A}$, and insect number 5 is a beetle: $C_5 = \mathsf{B}$. This is clear from our very notation, otherwise we would not have known whether to use the symbol A or B for those instances. This information is evidently implicit in our background information J.

We now make this categorical information more explicit. A slightly different definition of partial exchangeability is thus obtained, with a slightly different form of its representation theorem.

Besides the statements $\{X_i = x_i\}$, we introduce an additional set of atomic statements

$$\left\{C_i = c_i \,\middle|\, i \in \mathbf{N},\ \forall i\ c_i \in \mathfrak{C}\right\}\ . \tag{7}$$

For each i the statements $\{C_i = c \mid c \in \mathfrak{C}\}$ are mutually exclusive on information I. These statements let us identify each instance i as belonging to one or another category out of the finite set $\mathfrak{C}$.

A probability distribution over the statements $\{X_i = x_i\}$ given the statements $\{C_i = c_i\}$ is called partially exchangeable if

$$\left.\begin{array}{l}\text{for every } N\text{, every subset of } N \text{ indices } \{i\} \subset \mathbf{N},\\ \text{every permutation } \pi \text{ thereof, every permutation } \sigma \text{ thereof such that}\\ \sigma(j) = k \implies c_j = c_k,\ \text{and every sets of values } x_i \in \mathfrak{X},\, c_i \in \mathfrak{C}\text{:}\\ \qquad \mathrm{P}\Big(\bigwedge_i X_i = x_i \,\Big|\, \bigwedge_i C_i = c_i\,,\ I\Big) = \mathrm{P}\Big(\bigwedge_i X_i = x_{\pi(i)} \,\Big|\, \bigwedge_i C_i = c_{\pi(i)}\,,\ I\Big)\\ \qquad \mathrm{P}\Big(\bigwedge_i X_i = x_i \,\Big|\, \bigwedge_i C_i = c_i\,,\ I\Big) = \mathrm{P}\Big(\bigwedge_i X_i = x_{\sigma(i)} \,\Big|\, \bigwedge_i C_i = c_i\,,\ I\Big)\\ \text{and all such probabilities are consistently related by marginalization.}\end{array}\right\} \tag{8}$$

Not the particular permutation restriction: the only allowed permutations are those *which exchange indices having the same c value*, that is, indices belonging to the same category.

Some important features of this definition will be discussed shortly. For the moment let us rewrite the representation formula (6) according to the new definition.

For each category $c \in \mathfrak{C}$, introduce a normalized distribution $\{f_{x|c} \mid x \in \mathfrak{X}\}$ over the values x. As the notation suggests, it can be considered as a *conditional* distribution over x given c. Denote (with some abuse of symbols) by $\boldsymbol{f_{x|c}} \coloneqq (f_{x|c})$ the set of all such conditional distributions. This set is the Cartesian product of $|\mathfrak{C}|$ simplices, each of dimension $|\mathfrak{X}| - 1$.

Denote by $F_{x,c}$ the empirical joint relative frequency of the pair of values (x, c) occurring in the set of pairs $\{(x_i, c_i)\}$:

$$NF_{x,c} \coloneqq \textstyle\sum_i \delta(x, x_i)\, \delta(c, c_i), \quad x \in \mathfrak{X},\ c \in \mathfrak{C}\,. \tag{9}$$

Thus $NF_{x,c}$ is the total number of times value x appears among the pairs with $c_i = c$.

De Finetti's theorem states that the partially exchangeable distribution (8) can be written as follows:

$$\mathrm{P}\Bigl(\bigwedge_i \mathrm{X}_i = x_i \Bigm| \bigwedge_i \mathrm{C}_i = c_i\,,\, I\Bigr) = \int \prod_{c,x} {f_{x|c}}^{NF_{x,c}}\ \mathrm{p}\bigl(\boldsymbol{f_{x|c}} \bigm| I\bigr)\, \mathrm{d}\boldsymbol{f_{x|c}}\,. \tag{10}$$

Scrutiny of this formula shows that this form is equivalent to the more familiar representation (6). The integral contains one product of $f_{\dots|c}$ terms for every category c. In each such product, $f_{x_i|c}$ terms are multiplied together for those i such that $c_i = c$. There are exactly $NF_{x,c}$ such terms.

This alternative formulation of partial exchangeability shows that this symmetry could also be called *conditional* exchangeability instead. The role of conditional distributions is clear in the representation (10). In the following I will use the term 'conditional exchangeability' in place of 'partial exchangeability' to emphasize this.

2 Representation for joint distributions with conditional-exchangeability symmetries

Suppose that we would assign a conditionally (i.e., 'partially': see previous section) exchangeable distribution of probability to the statements $\{X_i = x_i\}$, if we knew the true $\{C_i = c_i\}$. But we do not know the latter. What kind of properties does the joint probability distribution of these statements have? And the marginal distribution for $\{X_i = x_i\}$?

The joint probability distribution can be rewritten

$$\mathrm{P}\Bigl(\bigwedge_i (X_i = x_i\,,\ C_i = c_i) \Bigm| I\Bigr) = \mathrm{P}\Bigl(\bigwedge_i X_i = x_i \Bigm| \bigwedge_i C_i = c_i\,,\ I\Bigr) \times \mathrm{P}\Bigl(\bigwedge_i C_i = c_i \Bigm| I\Bigr)\,, \tag{11}$$

where the first factor, conditionally exchangeable, can be represented by the integral of eq. (10).

Let us suppose that our degrees of belief about the statements $\{C_i = c_i\}$ are expressed by a fully exchangeable marginal probability distribution. An integral representation analogous to (4) then holds:

$$\mathrm{P}\Bigl(\bigwedge_i C_i = c_i \Bigm| I\Bigr) = \int \prod_c f_{,c}{}^{NF_{,c}}\ \mathrm{p}(\boldsymbol{f_c} \,|\, I)\,\mathrm{d}\boldsymbol{f_c}\,, \tag{12}$$

where $\boldsymbol{f_c} \coloneqq \{f_{,c}\}$ (the reason of the subscript comma will be soon clear), and $F_{,c} \coloneqq \sum_x F_{x,c}$ is the marginal empirical distribution for the c values.

We can now replace the integral representations (10) and (12) into the product (11). The products within their integrals can be combined considering that

$$f_{,c}{}^{NF_{,c}} = f_{,c}{}^{N\sum_x F_{x,c}} = \prod_x f_{,c}{}^{NF_{x,c}}\ . \tag{13}$$

We obtain

$$\mathrm{P}\Bigl(\bigwedge_i (X_i = x_i, C_i = c_i) \Bigm| I\Bigr) = \int \prod_{c,x} f_{x,c}{}^{NF_{x,c}}\ \mathrm{p}(\boldsymbol{f_{xc}} \,|\, I)\,\mathrm{d}\boldsymbol{f_{xc}} \tag{14a}$$

$$\boxed{\text{with}\qquad \mathrm{p}(\boldsymbol{f_{xc}} \,|\, I)\ \mathrm{d}\boldsymbol{f_{xc}} = \mathrm{p}(\boldsymbol{f_{x|c}} \,|\, I)\ \mathrm{p}(\boldsymbol{f_c} \,|\, I)\ \mathrm{d}\boldsymbol{f_{x|c}}\,\mathrm{d}\boldsymbol{f_c}} \tag{14b}$$

where we have defined $f_{x,c} = f_{x|c}\, f_{,c}$ and $\boldsymbol{f_{xc}} \coloneqq (f_{x,c})$. Note that $(f_{x,c})$ has indeed the properties of a joint distribution.

The boxed equality above comes from the one-one correspondence between the variables $(\boldsymbol{f_{x|c}}\,, \boldsymbol{f_c})$ and $\boldsymbol{f_{xc}}$, so that the product of density functions for $\boldsymbol{f_{x|c}}$ and $\boldsymbol{f_c}$ is just a specific case of a density function for $\boldsymbol{f_{xc}}$ (which includes a Jacobian determinant).

The integral expression (14) is the representation of a fully exchangeable predictive distribution. Thus the joint distribution for the set of *pairs* of statements $\{(X_i = x_i, C_i = c_i)\}$ is fully exchangeable.

The noteworthy feature of the integral expression (14) is that *the density for the joint distribution $\boldsymbol{f_{xc}}$ is factorizable into the product of a density* *for the conditional distribution $f_{x|c}$ and a density for the marginal distribution f_c.*

This factorization is, on the one hand, trivial: it must be so because the integral itself must factorize, to yield the factorization (11) of the predictive distributions.

But from the point of view of the joint predictive distribution (14a), on the other hand, this factorization is generally not valid and a priori not necessary. The following general identity would hold instead:

$$\mathrm{p}(f_{xc} \,|\, I)\, \mathrm{d}f_{xc} \equiv \mathrm{p}[(f_{x|c}, f_c) \,|\, I]\, \mathrm{d}f_{x|c}\, \mathrm{d}f_c \equiv \\ \mathrm{p}(f_{x|c} \,|\, f_c, I)\, \mathrm{p}(f_c \,|\, I)\, \mathrm{d}f_{x|c}\, \mathrm{d}f_c\,, \qquad (15)$$

which does not lead to a factorization of the integral. The factorization is thus a special property of the density, which reflects a special property of the joint predictive distribution: the conditional exchangeability for the statements $\{X_i = x_i\}$ given the $\{C_i = c_i\}$.

3 Bayesian networks and exchangeability

3.1 Graphical representation of conditional exchangeability

The conditional (i.e., 'partial') exchangeability (10) of the probabilities of the X-statements given the C-statements, the full exchangeability (12) of the probabilities of the latter, and the final integral representation (14) for their join probability distribution, can be all together expressed in the guise of a Bayesian network:

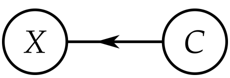

Figure 1

The two nodes represent the two sets of statements. The arrow from the C-node to the X-node represents the conditional exchangeability of the probabilities for X-statements conditional on the C-statements. The absence of incoming arrows to the C-node represents the full exchangeability of the distribution for the C-statements. The final integral representation for the joint probability of the full network has a factorizable density, with one factor per node.

Using the reasoning and integral representations of §§ 1–2 it is possible to generalize these rules to more complex networks of statements. It can be calculated that the following network, for example:

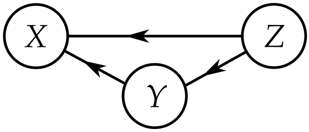

Figure 2

represents (i) the conditional exchangeability of the probabilities for the X-statements given *pairs* of Y- and Z-statements:

for every N, every subset of N indices $\{i\} \subset \mathbf{N}$, and every permutation π thereof such that $\quad \pi(j) = k \;\Rightarrow\; y_j = y_k \land z_j = z_k$,

$$\mathrm{P}\Bigl(\bigwedge_i X_i = x_i \Bigm| \bigwedge_i (Y_i = y_i\ ,\ Z_i = z_i)\ ,\ I\Bigr) = \mathrm{P}\Bigl(\bigwedge_i X_i = x_{\pi(i)} \Bigm| \bigwedge_i (Y_i = y_i\ ,\ Z_i = z_i)\ ,\ I\Bigr)\ , \tag{16}$$

(ii) the conditional exchangeability of the probabilities for the Y-statements given the Z-statements, and (iii) the full exchangeability of the probabilities of the Z-statements.

The joint predictive distribution has the expected representation

$$\mathrm{P}\Bigl(\bigwedge_i X_i = x_i \ \bigwedge_i Y_i = y_i \ \bigwedge_i Z_i = z_i \Bigm| I\Bigr) =$$
$$\int \prod_{x,y,z} \underbrace{f_{x,y,z}}_{=f_{x|y,z}\, f_{,y|z}\, f_{,,z}}{}^{NF_{x,y,z}} \quad \underbrace{\mathrm{p}(\boldsymbol{f}_{x|yz} \mid I)\, \mathrm{p}(\boldsymbol{f}_{y|z} \mid I)\, \mathrm{p}(\boldsymbol{f}_{z} \mid I)\, \mathrm{d}\boldsymbol{f}_{x|yz}\, \mathrm{d}\boldsymbol{f}_{y|z}\, \mathrm{d}\boldsymbol{f}_{z}}_{=\mathrm{p}(\boldsymbol{f}_{xyz} | I)\, \mathrm{d}\boldsymbol{f}_{xyz}} \tag{17}$$

with $f_{,,z} \coloneqq \sum_{x,y} f_{x,y,z}$ and so on.

3.2 Representing independence assumptions

In the last example the joint predictive distribution for all the statements is decomposed in full accord with the product rule:

$$\begin{aligned}\mathrm{P}\Bigl(\bigwedge_i X_i=x_i \ \bigwedge_i Y_i=y_i \ \bigwedge_i Z_i=z_i \Bigm| I\Bigr) \equiv \\ \mathrm{P}\Bigl(\bigwedge_i X_i=x_i \Bigm| \bigwedge_i Y_i=y_i \ \bigwedge_i Z_i=z_i \ , I\Bigr) \times \\ \mathrm{P}\Bigl(\bigwedge_i Y_i=y_i \Bigm| \bigwedge_i Z_i=z_i \ , I\Bigr) \times \mathrm{P}\Bigl(\bigwedge_i Z_i=z_i \Bigm| I\Bigr) ,\end{aligned} \tag{18}$$

which is an identity in the probability-calculus. In other words, no other special properties hold, besides the conditional exchangeability of the separate predictive distributions.

In this case the integral representation (17) involves an integration over a set of conditional or marginal long-run frequencies which is equivalent to the set of joint frequencies. That is, the two sets

$$\{f_{x|y,z} \, , \, f_{,y|z} \, , \, f_{,,z} \mid x \in \mathfrak{X}, y \in \mathfrak{Y}, z \in \mathfrak{Z}\} \leftrightarrow \{f_{x,y,z} \mid x \in \mathfrak{X}, y \in \mathfrak{Y}, z \in \mathfrak{Z}\} \tag{19}$$

are in one-one correspondence.

The factorizability of the density $\mathrm{p}(\boldsymbol{f}_{xyz} \mid I)\,\mathrm{d}\boldsymbol{f}_{xyz}$ therefore implies, and is implied by, the exchangeability symmetries of conditional probability distributions. But it does not imply additional independences.

Now we want to consider independence assumptions characteristics of non-trivial Bayesian networks, and see how they can be combined with exchangeability assumptions.

Take for example the non-trivial Bayesian network

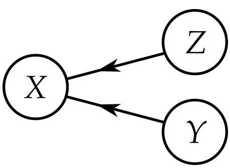

Figure 3

Interpreted as an individual instance, this network expresses the factorization

$$\begin{aligned}\mathrm{P}(X_1=x \, , \, Y_1=y \, , \, Z_1=z \mid I) \equiv \\ \mathrm{P}(X_1=x \mid Y_1=y \, , \, Z_1=z \, , \, I) \times \mathrm{P}(Y_1=y \mid I) \times \mathrm{P}(Z_1=z \mid I)\end{aligned} \tag{20}$$

for all x, y, z, which implies the conditional independence

$$\mathrm{P}(Y_1 = y \mid Z_1 = z \ , I) = \mathrm{P}(Y_1 = y \mid I) \tag{21}$$

or an equivalent one with Y and Z exchanged.

Additional independence properties, such as

$$\begin{aligned}
\mathrm{P}\Bigl(\bigwedge_i X_i = x_i \ \bigwedge_i Y_i = y_i \ \bigwedge_i Z_i = z_i \Bigm| I\Bigr) \equiv \\
\mathrm{P}\Bigl(\bigwedge_i X_i = x_i \Bigm| \bigwedge_i Y_i = y_i \ \bigwedge_i Z_i = z_i \ , I\Bigr) \times \\
\mathrm{P}\Bigl(\bigwedge_i Y_i = y_i \Bigm| I\Bigr) \times \mathrm{P}\Bigl(\bigwedge_i Z_i = z_i \Bigm| I\Bigr) ,
\end{aligned} \tag{22}$$

which is not an identity of the probability-calculus, are those typically expressed by Bayesian networks; in this case by the network

In the integral representation under discussion, these conditional independences are expressed by *a reduction in the number of conditional or marginal long-run frequencies that are integrated over* (similarly to what happens when partial exchangeability reduces to full exchangeability; see § 1).

For example, if the independence (22) hold, a little calculation shows that the representation (17) becomes

$$\begin{aligned}
\mathrm{P}\Bigl(\bigwedge_i X_i = x_i \ \bigwedge_i Y_i = y_i \ \bigwedge_i Z_i = z_i \Bigm| I\Bigr) = \\
\int \prod_{x,y,z} (f_{x|y,z}\, f_{,y}\, f_{,,z})^{NF_{x,y,z}}\ \mathrm{p}(\boldsymbol{f}_{x|yz} \mid I)\, \mathrm{p}(\boldsymbol{f}_{y} \mid I)\, \mathrm{p}(\boldsymbol{f}_{z} \mid I)\, \mathrm{d}\boldsymbol{f}_{x|yz}\, \mathrm{d}\boldsymbol{f}_{y}\, \mathrm{d}\boldsymbol{f}_{z} \ .
\end{aligned} \tag{23}$$

We see that the integration is now over the *reduced* set of distributions

$$\{f_{x|y,z} \ , \ f_{,y} \ , \ f_{,,z} \mid x \in \mathfrak{X}, y \in \mathfrak{Y}, z \in \mathfrak{Z}\} \ . \tag{24}$$

This is a reduced set in the sense that $f_{,y|z} = f_{,y}$ for all z, implying the presence of a delta term in the corresponding density.

3.3 Generalization and connection with Bayesian networks

Let us summarize what we have found thus far. We have

0. several sets of statements, $\{X_i = x_i\}$, $\{Y_i = y_i\}, \dots$. Each set is countably infinite. For each set we also have an associated set of values $x \in \mathfrak{X}$, $y \in \mathfrak{Y}, \dots$;
i. several assumptions of conditional independence representable in the guise of a Bayesian network, such as in eq. (22) and fig. 3;
ii. several assumptions of conditional exchangeability for the predictive probabilities of some groups of such statements conditional on some other groups, such as in eq. (10).

And we require that the two sets of assumptions i. and ii., taken together, be sufficient to obtain the joint predictive distribution for all statements through the product rule of the probability-calculus.

Then the joint predictive distribution for N statements from each group has an integral representation of the de Finetti form:

$$\mathrm{P}\Bigl(\bigwedge_i X_i = x_i \ \bigwedge_i Y_i = y_i \ \dots \Bigm| I\Bigr) = \int \prod_{x,y,\dots} (\xi_{x,y,\dots})^{N F_{x,y,\dots}} \ \mathrm{p}(\boldsymbol{\xi} \mid I)\, \mathrm{d}\boldsymbol{\xi} \tag{25}$$

where $N\, F_{x,y,\dots}$ are the joint empirical frequencies. This representation has these properties:

I. the set of integration variables $\{\xi_{x,y,\dots}\}$ consists of subsets of marginal and conditional distributions: the same that would formally be associated with the Bayesian network of point i. For example

$$\{\xi_{x,y,z}\} = \bigl\{\{f_{x|y,z}\},\ \{f_{,y}\},\ \{f_{,,z}\}\bigr\}\ .$$

II. the density $\mathrm{p}(\boldsymbol{\xi} \mid I)\,\mathrm{d}\boldsymbol{\xi}$ over the integration variables factorizes into a product of independent densities, one for each subset of the point above, for example

$$\mathrm{p}(\boldsymbol{\xi} \mid I)\,\mathrm{d}\boldsymbol{\xi} = \mathrm{p}(\boldsymbol{f}_{x|yz} \mid I)\,\mathrm{d}\boldsymbol{f}_{x|yz}\ \ \mathrm{p}(\boldsymbol{f}_{y} \mid I)\,\mathrm{d}\boldsymbol{f}_{y}\ \ \mathrm{p}(\boldsymbol{f}_{z} \mid I)\,\mathrm{d}\boldsymbol{f}_{z}\ .$$

3.4 Discussion

The representation just derived combines exchangeability and Bayesian networks. I believe such a combination is needed in concrete inference problems, especially when data are few with respect to the number of factors involved.

A simple non-trivial example is this network: which says that the joint distribution factorizes as $f_{x,y,z} = f_{x|y,z} \times f_{,y} \times f_{,,z}$.

One way to combine a Bayesian network and exchangeability would be to mimic the way a hypothesis about a superpopulation frequency ('parametric or non-parametric model') is assimilated by exchangeability: into a mixture of such frequencies, which yields our degrees of belief for a finite number of individual instances. This is equivalent to assuming full exchangeability of our degrees of beliefs about an infinite number of individuals.

For example, if $\bigl\{\{f_{x|y,z}\}, \{f_{,y}\}, \{f_{,,z}\}\bigr\}$ are the conditional-probability tables of the Bayesian network of fig. 3, our beliefs about a single individual would be given by the mixture

$$\begin{aligned}&\mathrm{P}(X_1 = x_1\,,\, Y_1 = y_1\,,\, Z_1 = z_1 \,|\, I) = \\ &\qquad \int f_{x_1|y_1,z_1}\; f_{,y_1}\; f_{,,z_1}\; \mathrm{p}(\boldsymbol{f}_{x|yz}\,,\, \boldsymbol{f}_{y}\,,\, \boldsymbol{f}_{z} \,|\, I)\, \mathrm{d}\boldsymbol{f}_{x|yz}\, \mathrm{d}\boldsymbol{f}_{y}\, \mathrm{d}\boldsymbol{f}_{z}\;. \end{aligned} \tag{26}$$

This probability distribution, however, in general *does not factorize*. So our beliefs about the individual would *not* be represented by a Bayesian network.

If we surmise – as seems to be the case in the literature –that a Bayesian network's structure applies both to every single individual and also to a superpopulation's statistics, then some additional assumptions are required besides full exchangeability.

4 Discussion

*** first result can be useful for infinite limits, leading to regression

The result of the previous section is thus summarized: Given infinitely countable sets of statements $\{X_i = x_i\}$ and $\{C_i = c_i\}$, and assuming that

1. the marginal probability distribution for the C statements is fully exchangeable,

2. the probability distribution for the X statements is partially (or conditionally) exchangeable given the C,

Then the joint distribution for both sets is fully exchangeable, and the density within its integral representation *factorizes* into a density for a conditional long-run frequency distribution, and a density for a marginal long-run frequency distribution, eq. (14b).

Appendix: On the ambiguous meaning of '$X = x$'

Bibliography

('de X' is listed under D, 'van X' under V, and so on, regardless of national conventions.)

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, repr. (Wiley, New York). First publ. 1994.

Bianchini, M., Maggini, M., Jain, L. C., eds. (2013): *Handbook on Neural Information Processing*. (Springer, New York).

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford).

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013), ch. 2, 19–29.

Diaconis, P., Freedman, D. (1980a): *Finite exchangeable sequences*. Ann. Prob. **8**[4], 745–764.

— (1980b): *De Finetti's generalizations of exchangeability*. In: Jeffrey (1980), 233–249.

Jeffrey, R. C., ed. (1980): *Studies in inductive logic and probability. Vol. II*. (University of California Press, Berkeley).

Keynes, J. M. (2013): *A Tract on Monetary Reform*, repr. of 2nd ed. (Cambridge University Press, Cambridge). First publ. 1923.

Lindley, D. V., Novick, M. R. (1981): *The role of exchangeability in inference*. Ann. Stat. **9**[1], 45–58.

Pearl, J. (2009): *Causality: Models, Reasoning, and Inference*, 2nd ed. (Cambridge University Press, Cambridge). First publ. 2000.

Wiegerinck, W., Burgers, W., Kappen, B. (2013): *Bayesian networks, introduction and practical applications*. In: Bianchini, Maggini, Jain (2013), p. 12, 401–431.