Do genes keep us awake?

Research notes

Cüneyt
<cuneyt.guzey@ntnu.no>

Daniela
<daniela.bragantini@ntnu.no>

Luca
<piero.mana@ntnu.no>

Yasser
<yasser.roudi@ntnu.no>

Draft of 24 September 2018 (first drafted 22 August 2018)

Research notes

1 Introductory notes

1.1 Preliminary remarks about Bayesian theory

Bayesian theory is not just a set of new, better recipes meant to replace old ones. It also requires a different – and simpler – mindset about problems of inference. Three points are especially important:

1. The only purpose of Bayesian theory is to give the degree of belief in some statements – more exactly, 'propositions' (Copi et al. 2014; Barwise et al. 2003) – given other statements that may concern data, facts, hypotheses. For example, Bayesian theory can tell us that we have a degree of belief x in hypothesis A, given some data C and initial information I, and a degree of belief y in hypothesis B given the same conditions:

$$\mathrm{P}(A|\,C, I) = x, \qquad \mathrm{P}(B|\,C, I) = y.$$

That's all there is to it. We can then use these degrees of belief as we like; in particular, we can use them within decision theory to choose courses of action (Raiffa et al. 2000; Pratt et al. 1996; Sox et al. 2013). But notions like 'statistical significance', 'acceptance level', 'confidence', and similar are foreign to Bayesian theory; or at best they're just secondary notions.

2. Bayesian theory is an extension of formal logic, the truth calculus. In fact we'll call it *plausibility calculus* from now on.

In formal logic, to prove a theorem we need some axioms to start from. These may partly include experimental facts or data, but they always also include assumptions that are purely conjectural. It's impossible to avoid

this conjectural element (see for example Harding 1976). [1] Likewise, in the plausibility calculus we need to specify initial degrees of belief. These may originate in data, but they always also include additional assumptions. The motto 'let the data speak for themselves' is simply impossible.

The difference between Bayesian methods and traditional methods is *not* that the former need additional assumptions while the latter don't. Rather, Bayesian methods make these assumptions explicit, while traditional methods hide them. This is the reason why many traditional results can be obtained as special cases of Bayesian ones.

3. Conditional degrees of belief like $P(A|B)$ do not express a causal connection between A and B, but an *informational* connection. In that conditional degree of belief, A could be the cause of B, or B of A, or neither could be the cause of the other. The classical example of this is

$$\mathrm{P}(\text{clouds in the sky}|\,\text{rain on the pavement}, I) > 0.5, \tag{1}$$

not because the rain is the cause of the clouds, but because its presence gives us *relevant information* about the cloudiness of the sky.

The previous remarks may appear pedantic, but they're important lest we misuse Bayesian methods.

1.2 What is the question?

✠ Luca: the following thoughts may be naive; I must still read (Stingo et al. 2015) and (Bush et al. 2012)

[1] This impossibility is well known in modern science; we can quote Poincaré (1992): 'But upon more reflection we realize the position held by hypothesis; we see that the mathematician wouldn't know how to do without it, and the experimenter can't do without it at all' (Introduction); 'Every generalization is a hypothesis' (ch. IX, p. 176). Duhem (1991): 'In sum, the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed' (§ VI.2, p. 187); 'Unlike the reduction to absurdity employed by geometers, experimental contradiction does not have the power to transform a physical hypothesis into an indisputable truth; in order to confer this power on it, it would be necessary to enumerate completely the various hypotheses which may cover a determinate group of phenomena; but the physicist is never sure he has exhausted all the imaginable assumptions' (§ VI.3, p. 190); 'the realization and interpretation of no matter what experiment in physics imply adherence to a whole set of theoretical propositions' (§ VI.5, p. 200). Medawar (1963): 'the starting point of induction, naive observation, innocent observation, is a mere philosophic fiction. There is no such thing as unprejudiced observation' Jeffreys [quote][ref].

We want to assess the informational relevance between some genetic variations $\{G\}$ and (combinations of) insomnia symptoms $\{S\}$ in the Norwegian or European population. To assess this relevance we use data D from a population sample. Some assumptions or background information I are also always present in our assessment.

The traditional approach to this kind of problems is to set two hypotheses against each other: 'there is a correlation' vs 'there isn't a correlation', and to assess which is more 'significant' in view of the data. Mathematically this corresponds to a dichotomy between an exactly zero value and non-zero values of a correlation-like quantity.

Here we approach this problem differently. Rather than contrasting zero against non-zero values, we simply calculate how *relevant* one quantity is to make inferences about the other quantity. ✠ check (Stephens et al. 2009): they seem to have a reference with a similar philosophy One quantity, for example G, is *inferentially* or *informationally* relevant when the knowledge of its value leads us to have a different degree of belief regarding the value of the other, for example S, compared to when we don't know its value. In other words, the degrees of belief $\mathrm{P}(S|\,G\,D I)$ and $\mathrm{P}(S|\,D I)$ are numerically different. If these two degrees of belief are approximately equal then the particular genetic variation G are *ir*relevant for our prediction of the insomnia symptom S. The same conclusion holds with G and S exchanged: the plausibility calculus says that

$$\mathrm{P}(S|\,G\,D I) = \mathrm{P}(S|\,D I) \quad \Longleftrightarrow \quad \mathrm{P}(G|\,S\,D I) = \mathrm{P}(G|\,D I) \tag{2}$$

if $\mathrm{P}(S|\,D I)$, $\mathrm{P}(G|\,D I)$ aren't zero.

This measure of relevance can be extended to sets of (combinations of) symptoms $\{S\}$ and of genetic variations $\{G\}$ by using the conditional entropy (Shannon 1948; Kelly 1956; Press et al. 2007 § 14.7; Cover et al. 2006 ch. 2)

$$H(\{S\}|\,\{G\}, D I) \coloneqq -\sum_{G} \mathrm{P}(G|\,D I) \sum_{S} \mathrm{P}(S|\,G\,D I) \ln \mathrm{P}(S|\,G\,D I), \tag{3}$$

which is zero only if G gives us certainty about S, and is equal to the entropy

$$H(\{S\}|\,D I) \coloneqq -\sum_{S} \mathrm{P}(S|\,D I) \ln \mathrm{P}(S|\,D I) \tag{4}$$

if G is irrelevant for predicting S (Press et al. 2007 § 14.7; Cover et al. 2006 ch. 2). Another, symmetric measure of relevance is the mutual information, discussed in § 2.

If we find that there is mutual informational relevance between genetic variations and insomnia symptoms, we can conclude from biologic reasons that those variations must have a direct or indirect influence on the symptoms, for example they may give susceptibility to insomnia.

There are three main ways to calculate the conditional degrees of belief: By calculating first $\mathrm{P}(SG|\,DI)$, or $\mathrm{P}(S|\,G\,DI)$, or $\mathrm{P}(G|\,S\,DI)$. Also, we can consider all possible combinations of genetic variations from the outset, or consider combination of few variations, gradually increasing the numbers. We shall try all these approaches and see whether their results are mutually consistent.

1.3 Why exchangeability

The sentence 'the degree of belief in an insomnia symptom given a genetic variation' is vague. What we mean is our guess that a given individual with that variation presents that symptom. Our guess about a particular individuals in the full population is updated from our knowledge about individuals we have sampled. Our guess about a new individual is affected by the sample data only insofar we believe those data to be representative for that individual.

The notion of exchangeability expresses this representativeness in terms of degrees of belief, as explained at length by de Finetti (1931; 1937; 1938). Denote by $S_s^{(i)}$ the statement that individual i has symptom S_s, and likewise with $G_g^{(i)}$ for the combination of genetic variations g. The fact that we believe, for inferential purposes, that the individuals from a population having the same genetic variation g are representative of one another is expressed by

$$\mathrm{P}(S_{s_1}^{(1)}\,S_{s_2}^{(2)}\,S_{s_3}^{(3)}\,\ldots|\,G_g^{(1)}\,G_g^{(2)}\,G_g^{(3)}\,\ldots\,I) = \mathrm{P}(S_{s_{\pi(1)}}^{(1)}\,S_{s_{\pi(2)}}^{(2)}\,S_{s_{\pi(3)}}^{(3)}\,\ldots|\,G_g^{(1)}\,G_g^{(2)}\,G_g^{(3)}\,\ldots\,I) \quad (5)$$

where π is an arbitrary permutation of the individuals' labels i. For example,

$$\mathrm{P}(S_{\mathrm{c}}^{(1)}\,S_{\mathrm{a}}^{(2)}\,S_{\mathrm{b}}^{(3)}\,\ldots|\,G_g^{(1)}\,G_g^{(2)}\,G_g^{(3)}\,\ldots\,I) = \mathrm{P}(S_{\mathrm{b}}^{(1)}\,S_{\mathrm{c}}^{(2)}\,S_{\mathrm{a}}^{(3)}\,\ldots|\,G_g^{(1)}\,G_g^{(2)}\,G_g^{(3)}\,\ldots\,I). \quad (6)$$

This mathematical property is called *exchangeability*. If the number of individuals is finite it is called *finite* exchangeability; letting this number increase indefinitely we reach *infinite* exchangeability as a limit (Heath et al. 1976).

Assuming exchangeability for each group of individuals sharing the same genetic variation, the mathematical relations above generalize as follows. Our degree of belief is the same if we exchange symptom labels among individuals *having the same genetic variation*; but it may be different if we exchange symptom labels among individuals with different genetic variations. The exact mathematical expression may look complicated; a concrete example is

$$\begin{aligned}\mathrm{P}(S_{\mathrm{a}}^{(1)}\, S_{\mathrm{c}}^{(2)}\, S_{\mathrm{d}}^{(3)}\, S_{\mathrm{a}}^{(4)}\, S_{\mathrm{b}}^{(5)}\, S_{\mathrm{d}}^{(6)}\, \dots \mid G_{\alpha}^{(1)}\, G_{\beta}^{(2)}\, G_{\beta}^{(3)}\, G_{\gamma}^{(4)}\, G_{\alpha}^{(5)}\, G_{\gamma}^{(6)}\, \dots\, I) = \\ \mathrm{P}(S_{\mathrm{b}}^{(1)}\, S_{\mathrm{d}}^{(2)}\, S_{\mathrm{c}}^{(3)}\, S_{\mathrm{d}}^{(4)}\, S_{\mathrm{a}}^{(5)}\, S_{\mathrm{a}}^{(6)}\, \dots \mid G_{\alpha}^{(1)}\, G_{\beta}^{(2)}\, G_{\beta}^{(3)}\, G_{\gamma}^{(4)}\, G_{\alpha}^{(5)}\, G_{\gamma}^{(6)}\, \dots\, I) \end{aligned} \tag{7}$$

where our degree of belief remains the same as we exchange symptoms a and b between individuals 1 and 5, both having genetic variation α; symptoms c and d between individuals 2 and 3, both having genetic variation β; symptoms a and d between individuals 4 and 6, both having genetic variation γ. These exchanges can involve an arbitrary number of individuals, symptoms, genetic variations. This general property is called *partial* exchangeability (de Finetti 1938; Diaconis et al. 1980; Diaconis 1988; for a connection with sampling theory see Sugden 1982; 1993).

Note that the property exemplified by eq. (7) is more general than just separately stating exchangeability for the distributions of degrees of belief in the individuals' sharing the same genetic variations. Property (7) allows data about individuals with a genetic variation to be *relevant* for prediction of data about individuals with *another* genetic variation, as we'll see shortly.

✠ add: de Finetti's representation theorem for distributions of degrees of belief with the property above

1.4 Selection of variables and robustness

Denote the presence of the genetic variation labelled i by G_i and its absence by $\neg G_i$. We can consider the relevance of each variation individually, say

$$\mathrm{P}(S \mid G_1\, DI), \tag{8}$$

or of the combination of any number of variations, say

$$\mathrm{P}(S|\, G_1\neg G_2\neg G_3 G_4\, DI). \tag{9}$$

The plausibility calculus allows us to assign all these degrees of belief for any amount of data D – since they represent beliefs. If the number of combinations is high compared with the number of data, however, our degrees of belief will usually change noticeably when updated with new data; we can say that they are less 'robust' to the acquisition of new data. This robustness can be quantified in various ways to be discussed later.

From this point of view it makes sense to first consider each genetic variation individually and then larger and larger combinations of variations, as long as we see that our degrees of belief conditional on data D are robust.

✠ Jeffreys (1983) § 3.2 *very* relevant to our problem! Also Broad (1918)

✠ See Jeffreys (1983 § 3.1, p. 124) on the degree of belief to be given to the ratio values $\{0, 1\}$: 'In genetics the suggested values are usually intermediate, such as $1/2$, $1/4$, and $3/8$'. Also, 'we cannot give a universal rule for them beyond the common-sense one, that if anybody does not know what his suggested value is, or whether there is one, he does not know what question he is asking and consequently does not know what his answer means. '

2 First approach: joint degree of belief and mutual information

2.1 Notation

The following notation produces compact but readily understandable formulae. Functions and operations on tuples $\boldsymbol{x} \coloneqq (x_1, \dots, x_C)$, $\boldsymbol{y} \coloneqq (y_1, \dots, y_C)$, and numbers a operate component-wise. For example:

$$\begin{aligned} \exp \boldsymbol{x} &\coloneqq \left(\exp x_1, \dots, \exp x_C\right) & \boldsymbol{x}\boldsymbol{y} &\coloneqq (x_1\, y_1, \dots, x_C\, y_C) \\ a\boldsymbol{x} &\coloneqq (a x_1, \dots, a x_C) & \boldsymbol{x}^a &\coloneqq \left({x_1}^a, \dots, {x_C}^a\right) \qquad \text{and so on.} \end{aligned} \tag{10a}$$

The exception are the sum and multiplication operators $\sum$, $\prod$:

$$\sum \boldsymbol{x} \coloneqq x_1 + \cdots + x_C \qquad \prod \boldsymbol{x} \coloneqq x_1 \cdots x_C \tag{10b}$$

so that, for example,

$$\sum \ln(\boldsymbol{x}/\boldsymbol{y}) \coloneqq \sum_{i=1}^{C} \ln(x_i/y_i).$$

Note also the conventions

$$\binom{a}{\boldsymbol{x}} \coloneqq \binom{a}{x_1, \dots, x_C} \coloneqq \frac{a!}{x_1! \cdots x_C!} \qquad \prod \binom{\boldsymbol{y}}{\boldsymbol{x}} \coloneqq \binom{y_1}{x_1} \cdots \binom{y_C}{x_C}, \tag{10c}$$

where the first expression is the multinomial coefficient.

2.2 Scheme of this approach

Here is the way of thinking and general form of the calculations for this approach. We'll see later if these calculations are practically feasible.

Denote by N the amount of Norwegian or European population, roughly equal to 5.3×10^6 or 740×10^6. Our initial information I says that each individual is characterized by two groups of quantities or variates:

1. An insomnia variate $\sigma \coloneqq (\sigma_1, \sigma_2, \sigma_3) \in \{0,1\}^3$ with $C_\sigma \coloneqq 8$ possible values. This variate consists of three binary variates representing the presence or absence of three insomnia symptoms. An individual with no insomnia symptoms ('control') has therefore $\sigma = (0,0,0)$. When convenient we shall use a binary-digit notation like $\sigma = 2 \equiv (0,1,0)$ or $\sigma = 5 \equiv (1,0,1)$.
2. A genetic variate $\gamma \coloneqq (\gamma_1, \dots, \gamma_l) \in \{0,1\}^l$ with $C_\gamma \coloneqq 2^l$ possible values. This variate consists of l binary variates, each representing the presence of either of two variants of a particular gene allele. We have data for $l = 94$ allele pairs, but we'll often consider a smaller subset of pairs. When convenient we shall use a binary-digit notation also for this variate.

The combined variate $\xi \coloneqq (\sigma, \gamma)$ can thus assume $C \coloneqq C_\sigma \times C_\gamma \approx 1.6 \times 10^{29}$ possible values at most, depending on how many gene alleles we consider.

We have data D consisting of the values of these variates for a sample of $n \coloneqq 6029$ individuals. The values for individual i are denoted $\xi^{(i)} \coloneqq \left(\sigma^{(i)}, \gamma^{(i)}\right)$, $i \in \{1, \dots, n\}$.

Denote the relative frequency of the insomnia-variate value σ in our data by s_σ, and the frequency distribution of all values by $\boldsymbol{s} \coloneqq (s_0, \dots, s_7)$. The frequency distribution is normalized, $\sum \boldsymbol{s} = 1$. The relative frequency for the gene-variate value γ is denoted g_γ, and the frequency distribution $\boldsymbol{g} \coloneqq (g_0, \dots, g_{2^l})$. The relative frequency for the joint variate value (σ, γ) is $x_{\sigma,\gamma}$, and the frequency distribution $\boldsymbol{x} \coloneqq (x_{0,0}, \dots, x_{7,2^l})$. By marginalization we must have $\sum_\gamma x_{\sigma,\gamma} = s_\sigma$ and $\sum_\sigma x_{\sigma,\gamma} = g_\gamma$. We can

also consider the relative frequency of a particular gene allele, say the jth one; we'll denote it by g_{γ_j}.

From our data D and from some initial knowledge I we want to make inferences about two connected unknowns:

(a) the frequency distributions of insomnia symptoms, of gene variations, and of both jointly in the full population of N individuals. In other words, we must guess what these frequency distributions are, and therefore quantify our degrees of belief in their possible values. Let's denote these frequency distributions with the same symbols as for our data, but with capital letters: $\boldsymbol{S}$ is the frequency distribution of insomnia symptoms, $\boldsymbol{G}$ of gene variations, and $\boldsymbol{X}$ the joint distribution of both. In formulae we want to assign values to

$$\mathrm{p}(\boldsymbol{X}|\, D, I), \qquad \mathrm{p}(\boldsymbol{S}|\, D, I), \qquad \mathrm{p}(\boldsymbol{G}|\, D, I). \tag{11}$$

(b) The symptoms and gene variations of an individual '0' chosen at random from the full population. That is, we want to quantify our degrees of belief in joint and separate possible values of these variates for this individual:

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, D, I\bigr), \qquad \mathrm{p}\bigl(\sigma^{(0)}|\, D, I\bigr), \qquad \mathrm{p}\bigl(\gamma^{(0)}|\, D, I\bigr). \tag{12}$$

These two kinds of degree of belief are mathematically connected: If we knew the joint frequency distribution $\boldsymbol{X}$ in the full population, then our degree of belief that individual 0 have variates $\bigl(\sigma^{(0)}, \gamma^{(0)}\bigr)$ would be, by symmetry,

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, I\bigr) = \mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, D, I\bigr) = X_{\sigma^{(0)},\gamma^{(0)}}. \tag{13}$$

The conditionals in these degrees of belief indicate that we know $\boldsymbol{X}$ besides our initial knowledge I, and the first equality says that the data D would be irrelevant if we knew $\boldsymbol{X}$. If $\boldsymbol{X}$ is unknown, then by the theorem of total degree of belief we have

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, D, I\bigr) = \sum_{\boldsymbol{X}} \mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, D, I\bigr)\, \mathrm{p}(\boldsymbol{X}|\, D, I) \equiv \sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}}\, \mathrm{p}(\boldsymbol{X}|\, D, I), \tag{14}$$

which says that our degree of belief in the individual's variates equals our expectation of the frequency of those variates. Formula (14) thus

connects the degrees of belief (11) and (12). Analogous equations hold for the marginal frequency distributions $\boldsymbol{S}$, $\boldsymbol{G}$ and the variates $\sigma^{(0)}$, $\gamma^{(0)}$.

If the number of alleles considered $l > 20$, the C possible joint variate values are much more numerous than the data n; not all of them can therefore appear in the data. Hence many of these gene variate values have frequencies $g_\gamma = 0$ and consequently $x_{\sigma,\gamma} = 0$. Denote by C^+_γ the number of distinct gene variations present in the data, and by C^-_γ the number of those absent. Let C^+ and C^- have the same meaning regarding the joint variate $(\sigma^{(0)}, \gamma^{(0)})$. Obviously $C_\gamma = C^+_\gamma + C^-_\gamma$, $C = C^+ + C^-$, and $C^+ \leqslant C^+_\gamma$.

As explained in § 1.2, the relevance of our knowledge of an individual's genetic data for our degree of belief about his or her insomnia symptoms resides in the difference between the distributions $\mathrm{p}(\sigma^{(0)}|\,\gamma^{(0)}, D, I)$ and $\mathrm{p}(\sigma^{(0)}|\,D, I)$. This can be quantified as the difference in the corresponding entropy $H(\sigma^{(0)}|\,D, I)$ and conditional entropy $H(\sigma^{(0)}|\,\gamma^{(0)}; D, I)$, which is the mutual information (Shannon 1948; Kelly 1956 in these called 'rate of transmission'; Press et al. 2007 § 14.7; Cover et al. 2006 ch. 2):

$$\begin{aligned} I(\sigma^{(0)} : \gamma^{(0)}|\,D, I) &\coloneqq H(\sigma^{(0)}|\,D, I) - H(\sigma^{(0)}|\,\gamma^{(0)}; D, I) \\ &\equiv \sum_{\sigma^{(0)},\gamma^{(0)}} \mathrm{p}(\sigma^{(0)}, \gamma^{(0)}|\,D, I)\, \ln \frac{\mathrm{p}(\sigma^{(0)}, \gamma^{(0)}|\,D, I)}{\mathrm{p}(\sigma^{(0)}|\,D, I)\, \mathrm{p}(\gamma^{(0)}|\,D, I)}. \end{aligned} \tag{15}$$

The second expression shows that the mutual information also measures the discrepancy between our degree of belief about the variates jointly and the product of our degrees of belief about them separately. The mutual information vanishes if the symptoms and genetic variations are completely irrelevant to one another, and is equal to the entropy $H(\sigma^{(0)}|\,D, I)$ if knowledge of the gene variate gives us complete certainty about the insomnia symptoms, since $H(\sigma^{(0)}|\,\gamma^{(0)}; D, I)$ vanishes in this case.

The mutual information depends on the knowledge on which our degree of belief is based, in this case the data and initial knowledge DI. We'll obviously consider initial states of knowledge I such that

$$I(\sigma^{(0)} : \gamma^{(0)}|\,I) = 0, \tag{16}$$

that is, we assume no a priori relevance of one variate upon the other. This kind of initial information I can still strongly emphasize or de-emphasize

the effect of the data on our knowledge when the latter are few compared to the range of the variates.

To calculate the mutual information (15) given the data we need the joint degrees of belief (12). To calculate the latter we use eq. (14), which needs our degrees of belief about the joint frequency distribution (11). These can be calculated from our initial degrees of belief $\mathrm{p}(\boldsymbol{X}|\, I)$ via Bayes's theorem:

$$\mathrm{p}(\boldsymbol{X}|\, \boldsymbol{x}, n, I) = \frac{\mathrm{p}(\boldsymbol{x}|\, n, \boldsymbol{X}, I)\; \mathrm{p}(\boldsymbol{X}|\, I)}{\sum_{\boldsymbol{X}} \mathrm{p}(\boldsymbol{x}|\, n, \boldsymbol{X}, I)\; \mathrm{p}(\boldsymbol{X}|\, I)}. \tag{17}$$

Bayes's theorem requires our degrees of belief about the joint frequency distribution $\boldsymbol{x}$ of the variates in the sampled population, given the distribution $\boldsymbol{X}$ in the full population. This problem is similar to 'drawing from an urn without replacement', for which our degrees of belief are represented by the multivariate hypergeometric distribution:

$$\mathrm{p}(\boldsymbol{x}|\, n, \boldsymbol{X}, I) = \binom{N}{n}^{-1} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}} \tag{18}$$

(Ghosh et al. 1997; Freedman et al. 2007 parts I, VI; summaries in Gelman et al. 2014 ch. 8; Jaynes 2003 ch. 3; properties of this distribution are discussed in Ross 2010 § 4.8.3; Feller 1968 § II.6).

Combining eqs (14), (17), (18), and simplifying we obtain

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{x}, n, I\bigr) = \frac{\sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}\; \mathrm{p}(\boldsymbol{X}|\, I)}{\sum_{\boldsymbol{X}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}\; \mathrm{p}(\boldsymbol{X}|\, I)}. \tag{19}$$

With an analogous reasoning we find analogous formulae for $\mathrm{p}\bigl(\sigma^{(0)}|\, \boldsymbol{x}, n, I\bigr)$ by replacing $\boldsymbol{X}$, $\boldsymbol{x}$ with $\boldsymbol{S}$, $\boldsymbol{s}$; and for $\mathrm{p}\bigl(\gamma^{(0)}|\, \boldsymbol{x}, n, I\bigr)$ by replacing $\boldsymbol{X}$, $\boldsymbol{x}$ with $\boldsymbol{G}$, $\boldsymbol{g}$. From these we can calculate the mutual information (15).

2.3 First calculation: constant initial degree of belief

We need to assess our initial degree of belief for $\boldsymbol{X}$ according to some background knowledge. As a first tentative let's consider background knowledge I_l such that the degree of belief is constant for all possible

distributions $\boldsymbol{X}$. There are $\binom{N+C-1}{C-1}$ possible distributions $\boldsymbol{X}$ (Csiszár et al. 2004 § 2.1); hence

$$\mathrm{p}(\boldsymbol{X}|\, I_l) = \binom{N+C-1}{C-1}^{-1}. \tag{20}$$

This yields a constant initial distribution of degree of belief for $(\sigma^{(0)}, \gamma^{(0)})$:

$$\mathrm{p}\big(\sigma^{(0)}, \gamma^{(0)}|\, I_l\big) = \sum_{\boldsymbol{X}} \mathrm{p}\big(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, I_l\big)\, \mathrm{p}(\boldsymbol{X}|\, I_l) = \sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \binom{N+C-1}{C-1}^{-1} = 1/C, \tag{21}$$

because the last sum is a convex combination, with equal weights, of all points in a simplex of dimension $C-1$, giving its centre of mass $(1/C, 1/C, \dots)$.

With the initial knowledge I_l, eq. (19) simplifies to

$$\mathrm{p}\big(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{x}, n, I_l\big) = \frac{\sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}}{\sum_{\boldsymbol{X}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}}. \tag{22}$$

The sum in the denominator can be calculated with an identity for multinomial coefficients ✠ add refs for summation formula:

$$\sum_{\boldsymbol{X}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}} = \binom{N+C-1}{n+C-1}, \tag{23}$$

which, substituted in eq. (17), also leads to

$$\mathrm{p}(\boldsymbol{X}|\, \boldsymbol{x}, n, I_l) = \binom{N+C-1}{n+C-1}^{-1} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}. \tag{24}$$

The sum in the numerator of eq. (22) can be rewritten ✠ explain the steps obtaining

$$\sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}} = \frac{n x_{\sigma^{(0)},\gamma^{(0)}} + 1}{N} \binom{N+C}{n+C} - \frac{1}{N} \binom{N+C-1}{n+C-1}. \tag{25}$$

Simplifying we finally find

$$\mathrm{p}\big(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{x}, n, I_l\big) = \frac{(N+C) n x_{\sigma^{(0)},\gamma^{(0)}} + N - n}{N\,(n+C)} \tag{26}$$

This expression can be interpreted as a weighted sum of the observed frequency $x_{\sigma^{(0)},\gamma^{(0)}}$ in the data and the initial degree of belief $1/C$, eq. (21):

$$\mathrm{p}\big(\sigma^{(0)},\gamma^{(0)}\big|\,\boldsymbol{x},n,I_l\big) \propto x_{\sigma^{(0)},\gamma^{(0)}} + \frac{N-n}{n}\frac{C}{N+C}\frac{1}{C}, \tag{27}$$

the ratio of the second to the first weight being $(N-n)/n \times C/(N+C)$ ✠ Luca: I don't find this intuitively satisfying, though I don't know why. It'd be good to try another initial state of knowledge. When $n = N$, that is, when we've sampled the full population, the second weight is zero and we're left with $x_{\sigma^{(0)},\gamma^{(0)}}$, which is also equal to $X_{\sigma^{(0)},\gamma^{(0)}}$, consistent with eq. (13).

For the marginal distributions we obtain, by summation,

$$\mathrm{p}\big(\sigma^{(0)}\big|\,\boldsymbol{x},n,I_l\big) = \frac{(N+C)ns_{\sigma^{(0)}} + (N-n)C_\gamma}{N\,(n+C)}, \tag{28}$$

$$\mathrm{p}\big(\gamma^{(0)}\big|\,\boldsymbol{x},n,I_l\big) = \frac{(N+C)ng_{\gamma^{(0)}} + (N-n)C_\sigma}{N\,(n+C)}. \tag{29}$$

Note that if the initial distribution of degree of belief about the joint frequencies $\boldsymbol{X}$ is uniform, then the degrees of belief for the marginal frequencies $\boldsymbol{S}$ and $\boldsymbol{G}$ are *not* uniform.

Using the distributions (26) and (28) in the formula for the mutual information (15) and simplifying we find

$$I\big(\sigma^{(0)}:\gamma^{(0)}\big|\,D,I_l\big) = \ln[N\,(n+C)] + \sum_{\sigma^{(0)},\gamma^{(0)}} \frac{(N+C)nx_{\sigma^{(0)},\gamma^{(0)}} + N - n}{N\,(n+C)} \times \ln \frac{(N+C)nx_{\sigma^{(0)},\gamma^{(0)}} + N - n}{[(N+C)ns_{\sigma^{(0)}} + (N-n)C_\gamma]\,[(N+C)ng_{\gamma^{(0)}} + (N-n)C_\sigma]} \tag{30}$$

We can divide this sum into two parts: one sum over the values of $\sigma^{(0)}$ and $\gamma^{(0)}$ which appear in our data, for which $g_{\gamma^{(0)}} > 0$; and one sum

over the C_γ^- remaining $\gamma^{(0)}$ values, for which $g_{\gamma^{(0)}} = x_{\sigma^{(0)},\gamma^{(0)}} = 0$:

$$
\begin{aligned}
I\big(\sigma^{(0)} : \gamma^{(0)} \big| D, I_l\big) &= \ln[N\,(n+C)] \\
&+ \sum_{\sigma^{(0)}} \sum_{\gamma^{(0)} \in D} \frac{(N+C)n x_{\sigma^{(0)},\gamma^{(0)}} + N - n}{N\,(n+C)} \times \\
&\qquad \ln \frac{(N+C)n x_{\sigma^{(0)},\gamma^{(0)}} + N - n}{[(N+C)n s_{\sigma^{(0)}} + (N-n)C_\gamma]\,[(N+C)n g_{\gamma^{(0)}} + (N-n)C_\sigma]} \\
&- C_\gamma^- \frac{N-n}{N\,(n+C)} \sum_{\sigma^{(0)}} \ln\{C_\sigma\,[(N+C)n s_{\sigma^{(0)}} + (N-n)C_\gamma]\}
\end{aligned}
\tag{31}
$$

2.4 Second calculation: uniform marginals

✠ IMPORTANT: check (Zhang et al. 2012) and its refs in § 2

The initial state of knowledge I_l used in the previous section was denoted with 'l' because it depends on the number of gene variations we consider. Two different values l' and l'' correspond to two different states of knowledge, $I_{l'} \neq I_{l''}$, which yield different initial degrees of belief for the marginal frequency distribution of a single gene. For example, for very large N, if we consider a single gene, $= 1$, the density of degree of belief for its marginal frequency f is constant in $\mathrm{d}f$. If we consider an additional gene, $l = 2$, the density of degree of belief for the marginal frequency of the previous gene becomes $6f\,(1-f)\,\mathrm{d}f$. If we consider an additional third gene, this density will be again different. As a consequence, if we decide to consider additional genes we cannot compare our inference with the ones previously made, because they come from *incompatible* initial states of knowledge.

Another bothering feature of the state of knowledge I_l is that the marginal degree of belief in the insomnia symptoms alone given the full set of data, eq. (28), depends on the number of gene variations considered in the data. Likewise, the degree of belief in the gene variations depend on the number of insomnia symptoms.

It seems reasonable to consider an initial state of knowledge that doesn't lead to different marginal distributions of degrees of belief in the frequencies when we want to consider an additional gene variation, and that doesn't have the counter-intuitive features above.

One such a state of knowledge I_0 exists (perhaps it isn't unique) and is characterized by a Dirichlet distribution for $\boldsymbol{X}$ in the $N \to \infty$ limit:

$$\mathrm{p}(\boldsymbol{X}|\, I_0)\,\mathrm{d}\boldsymbol{X} = \frac{1}{\prod \Gamma\left(\frac{2}{C}\right)} \prod \boldsymbol{X}^{\frac{2}{C}-1}\, \delta(\textstyle\sum \boldsymbol{X} - 1)\,\mathrm{d}\boldsymbol{X} \tag{32}$$

This distribution has the property of leaving the distribution of degree of belief for the marginal frequencies of any number of gene variations invariant if we consider one more gene, thanks to the marginalization properties of the Dirichlet distribution (Ferguson 1973 § 2 p. 211). It moreover assigns a uniform distribution to the frequency of each gene variation. It doesn't assign equal degrees of belief to the possible frequency distributions for the $C_\sigma \equiv 8$ symptom combinations, but rather

$$\mathrm{p}(\boldsymbol{S}|\, I_0)\,\mathrm{d}\boldsymbol{S} = \frac{1}{\prod \Gamma\left(\frac{2}{C_\sigma}\right)} \prod \boldsymbol{S}^{\frac{2}{C_\sigma}-1}\, \delta(\textstyle\sum \boldsymbol{S} - 1)\,\mathrm{d}\boldsymbol{S} = \frac{1}{8\Gamma\left(\frac{1}{4}\right)} \prod \boldsymbol{S}^{-3/4}\, \delta(\textstyle\sum \boldsymbol{S} - 1)\,\mathrm{d}\boldsymbol{S}. \tag{33}$$

Note, however, that the density of degree of belief for the frequency of controls vs cases is uniform.

The approximation $N \to \infty$ needs to be justified. If we consider a limited amount of gene variations, so that N/C is large (say, less than 12 gene variations), this approximation should be valid. If we consider a large amount of gene variations, so that C/N is large (say, more than 25 gene variations), the possible frequencies lie on the faces of the $(C-1)$-dimensional simplex; but the density (32) is concentrated in regions at the boundary of the simplex, and thus the approximation may still be reasonable.

Result:

$$\mathrm{p}(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{x}, nI_0) = \frac{n x_{\sigma^{(0)},\gamma^{(0)}} + 2/C}{n+2} \tag{34}$$

[Luca's memoranda:]

- Use of partial exchangeability *has to* distinguish also between men and women: see Gehrman et al. (2013 p. 327).
- This study could also be used to detect most relevant genes, by eliminating them in turn (and in pairs etc) and checking the ensuing predictions.

- Is it computationally possible to use a 'nonparametric model'? It would avoid unwarranted assumptions and phenomena like overtranining.

Bibliography

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

Barwise, J., Etchemendy, J. (2003): *Language, Proof and Logic*. (CSLI, Stanford). Written in collaboration with Gerard Allwein, Dave Barker-Plummer, Albert Liu. First publ. 1999.

Bernardo, J.-M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M., eds. (1988): *Bayesian Statistics 3*. (Oxford University Press, Oxford).

Broad, C. D. (1918): *On the relation between induction and probability. – (Part I.)* Mind **27**[108], 389–404. See also Broad (1920).

— (1920): *On the relation between induction and probability. – (Part II.)* Mind **29**[113], 11–45. See also Broad (1918).

Bush, W. S., Moore, J. H. (2012): *Genome-wide association studies*. PLoS Comput. Biol. **8**[12], e1002822.

Copi, I. M., Cohen, C., McMahon, K. (2014): *Introduction to Logic*, 14th ed. (Pearson, Harlow, UK). First publ. 1953.

Cover, T. M., Thomas, J. A. (2006): *Elements of Information Theory*, 2nd ed. (Wiley, Hoboken, USA). First publ. 1991.

Csiszár, I., Shields, P. C. (2004): *Information theory and statistics: a tutorial*. Foundations and Trends in Communications and Information Theory **1**[4], 417–528. http://www.renyi.hu/~csiszar/.

de Finetti, B. (1931): *Probabilismo*. Logos **14**, 163–219. Transl. as de Finetti (1989). See also Jeffrey (1989).

— (1937): *La prévision : ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**[1], 1–68. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

— (1938): *Sur la condition d'équivalence partielle*. In: *Colloque consacré à la théorie des probabilités. VI : Conceptions diverses*. Ed. by B. de Finetti, V. Glivenko, G. Neymann (Hermann, Paris), 5–18. Transl. in Jeffrey (1980), pp. 193–205, by P. Benacerraf and R. Jeffrey.

— (1989): *Probabilism: A critical essay on the theory of probability and on the value of science*. Erkenntnis **31**[2–3], 169–223. Transl. of de Finetti (1931) by Maria Concetta Di Maio, Maria Carla Galavotti, and Richard C. Jeffrey.

Diaconis, P. (1988): *Recent progress on de Finetti's notions of exchangeability*. In: Bernardo, DeGroot, Lindley, Smith (1988), 111–125. With discussion by D. Blackwell, Simon French, and author's reply. http://statweb.stanford.edu/~cgates/PERSI/year.html, https://statistics.stanford.edu/research/recent-progress-de-finettis-notions-exchangeability.

Diaconis, P., Freedman, D. (1980): *De Finetti's generalizations of exchangeability*. In: Jeffrey (1980), 233–249.

Duhem, P. (1914): *La Théorie Physique : son objet – sa structure*, 2nd ed. (Marcel Rivière, Éditeur, Paris). http://virtualbooks.terra.com.br/freebook/fran/la_theorie_physique.htm. First publ. 1906. Transl. as Duhem (1991).

Duhem, P. (1991): *The Aim and Structure of Physical Theory*, Transl. of the 2nd ed. (Princeton University Press, Princeton). Transl. of Duhem (1914) by P. P. Wiener. First publ. in French 1906.

Feller, W. (1968): *An Introduction to Probability Theory and Its Applications. Vol. I*, 3rd ed. (Wiley, New York). First publ. 1950.

Ferguson, T. S. (1973): *A Bayesian analysis of some nonparametric problems*. Ann. Stat. **1**^{2}, 209–230.

Freedman, D. A., Pisani, R., Purves, R. (2007): *Statistics*, 4th ed. (Norton, London). First publ. 1978.

Gehrman, P. R., Pfeiffenberger, C., Byrne, E. M. (2013): *The role of genes in the insomnia phenotype*. Sleep Med. Clin. **8**^{3}, 323–331.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2014): *Bayesian Data Analysis*, 3rd ed. (Chapman & Hall/CRC, Boca Raton, USA). First publ. 1995.

Ghosh, M., Meeden, G. (1997): *Bayesian Methods for Finite Population Sampling*, reprint. (Springer, Dordrecht).

Harding, S. G., ed. (1976): *Can Theories Be Refuted?* (D. Reidel, Dordrecht).

Heath, D., Sudderth, W. (1976): *De Finetti's theorem on exchangeable variables*. American Statistician **30**^{4}, 188–189.

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffrey, R. (1989): *Reading* Probabilismo. Erkenntnis **31**$^{2-3}$, 225–237. See de Finetti (1931).

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

Jeffreys, H. (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.

Kelly Jr., J. L. (1956): *A new interpretation of information rate*. Bell Syst. Tech. J. **35**^{4}, 917–926. http://turtletrader.com/kelly.pdf, https://archive.org/details/bstj35-4-917.

Koch, G., Spizzichino, F., eds. (1982): *Exchangeability in Probability and Statistics*. (North-Holland, Amsterdam).

Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.

Medawar, P. B. (1963): *Is the scientific paper a fraud?* Listener **70**, 377–378.

Poincaré, H. (1905): *Science and Hypothesis*. (Walter Scott, London). Transl. of Poincaré (1992) by W. J. Greenstreet; with a Preface by J. Larmor. First publ. 1902. Partly repr. in Poincaré (1958).

— (1958): *The Value of Science*. (Dover, New York). Authorized transl. with an introduction by G. B. Halsted. First publ. 1913.

— (1992): *La science et l'hypothèse*. (Éditions de la Bohème, Rueil-Malmaison, France). http://gallica.bnf.fr/document?O=N026745. First publ. 1902; transl. as Poincaré (1905).

Pratt, J. W., Raiffa, H., Schlaifer, R. (1996): *Introduction to Statistical Decision Theory*, 2nd pr. (MIT Press, Cambridge, USA). First publ. 1995.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007): *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge). First publ. 1988.

Raiffa, H., Schlaifer, R. (2000): *Applied Statistical Decision Theory*, reprint. (Wiley, New York). First publ. 1961.

Ross, S. (2010): *A First Course in Probability*, 8th ed. (Pearson, Upper Saddle River, USA). First publ. 1976.

Shannon, C. E. (1948): *A mathematical theory of communication*. Bell Syst. Tech. J. **27**[3, 4], 379–423, 623–656. https://archive.org/details/bstj27-3-379, https://archive.org/details/bstj27-4-623, http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

Sox, H. C., Higgins, M. C., Owens, D. K. (2013): *Medical Decision Making*, 2nd ed. (Wiley, New York). First publ. 1988.

Stephens, M., Balding, D. J. (2009): *Bayesian statistical methods for genetic association studies*. Nat. Rev. Genet. **10**, 681–690.

Stingo, F. C., Swartz, M. D., Vannucci, M. (2015): *A bayesian approach to identify genes and gene-level SNP aggregates in a genetic analysis of cancer data*. Stat. Interface **8**[2], 137–151.

Sugden, R. A. (1982): *Exchangeability and survey sampling inference*. In: Koch, Spizzichino (1982), 321–330.

— (1993): *Partial exchangeability and survey sampling inference*. Biometrika **80**[2], 451–455.

Zhang, H., Ghosh, K., Ghosh, P. (2012): *Sampling designs via a multivariate hypergeometric-Dirichlet process model for a multi-species assemblage with unknown heterogeneity*. Comput. Stat. Data Anal. **56**[8], 2562–2573.