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 11 November 2018 (first drafted 22 August 2018)

Research notes

1 Problem setup

Every single-nucleotide polymorphism (SNP), together with the huge variety of external factors, can in principle affect the appearance of a disease's symptoms. It is extremely complicated to identify and untangle these causal mechanisms and interactions and to ascertain their degrees, although their causal graph (Pearl 2009) is easy to draw (fig.**). The interacting mechanisms represented by the arrows are difficult to study, and the external factors X are innumerable and largely unknown.

A qualitative indication of the causal strength of groups of SNPs on some symptoms can be obtained by replacing the causal graph with a corresponding simplified Bayesian network (Pearl 2009) of *conditional probabilities* (fig.**). The external factors X disappear from the graph but their presence is implicit in the probabilistic relation between the nodes, which also accounts for the unknown causal mechanisms. Consider a particular SNP and a symptom S. In an arbitrarily large population, if the individuals having one allele and those having the other allele show markedly different *conditional frequencies* of incidence of the symptom, then we can conclude that the SNP must have some causal relevance, however indirect, for the symptom.

These conditional frequencies are far easier to estimate than causes, given the conditional frequencies in a population sample.

In this study we show how to quantify our degrees of belief about such conditional frequencies in an arbitrarily large population, given the conditional frequencies in a population sample and a representation

of our initial information about such frequencies. The calculation is intuitively simple: using Bayes's theorem we have

$$\mathrm{p}(\text{frequencies}|\,\text{sample data}, \text{initial info}) \propto$$
$$\mathrm{p}(\text{data}|\,\text{frequencies}, \text{initial info}) \times \mathrm{p}(\text{frequencies}|\,\text{initial info}). \quad (1)$$

The first degree of belief in the product above is given by a simple sampling formula; the second can be modelled in several ways, but if the sample data are enough it will lead to basically the same final degrees of belief about the conditional frequencies.

Once we have a quantified distribution of belief about the conditional frequencies in a arbitrarily large population, it is easy to see whether we expect the latter to be significantly different for different alleles. See for example the distributions in fig. 1: from the sample data we expect conditional frequencies 0.259 and 0.284 for the symptom conditional on the two alleles, with standard deviations 0.008 and 0.006. From the two belief distributions we can even calculate our belief that the two conditional frequencies are equal to within some range; in the case of the figure our belief that the two frequencies are the same within 0.01 is 3.2%. Many other quantifications are possible; for example our belief that a conditional frequency lies between two particular values, and so on.

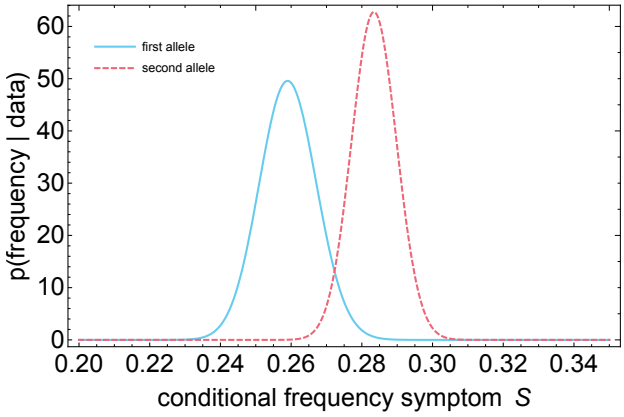

Figure 1 Example of distributions of belief

✠ add something about dependence of broadness on sample size, and 'smoothing' as discussed by MacKay & Bauman Peto (1995 § 2.6).

This approach is not dichotomous, like a 'significance' test. Rather, we will find a graduation of cases: from frequencies predicted to be clearly distinct, to frequencies predicted to be very similar, or with uncertainties too large for drawing definite conclusions. These cases can be sorted, obtaining a sequence of SNPs with a decreasing belief of causal association with the symptom. How many of these SNPs are to be selected for further study depends on one's experimental and computational resources.

1.1 Summary of the main formulae

We have a sample of size n. We check the subsample of individuals that have a particular allele, say Bx, for a particular gene, say rs697680_A. Suppose that in this subsample n_0 individuals *don't* show symptom A and n_1 *do* show symptom A. This also means that the size of our subsample (individuals with allele Bx) is $n \coloneqq n_0 + n_1$.

Our degree of belief about the frequency f_1 of symptom A among the individuals with allele Bx in an *infinite* population is a Beta distribution with parameters $n_0 + \theta_0$, $n_1 + \theta_1$, with $\theta \coloneqq \theta_0 + \theta_1$:

$$\mathrm{p}(f_1 | \, n_0, n_1, \theta_0, \theta_1) \, \mathrm{d}f_1 = \frac{\Gamma(n+\theta)}{\Gamma(n_0+\theta_0)\,\Gamma(n_1+\theta_1)} \, (1-f_1)^{n_0+\theta_0-1} \, {f_1}^{n_1+\theta_1-1} \, \mathrm{d}f_1 \quad (2)$$

This distribution has expected value and variance

$$\begin{aligned} \mathrm{E}(f_1 | \, n_0, n_1, \theta_0, \theta_1) &= \frac{n_1+\theta_1}{n+\theta}, \\ \mathrm{var}(f_1 | \, n_0, n_1, \theta_0, \theta_1) &= \frac{(n_0+\theta_0)\,(n_1+\theta_1)}{(n+\theta)^2\,(n+\theta+1)}. \end{aligned} \quad (3)$$

✠ Possible further developments: use of hyper-Dirichlet priors, use of graphical models to infer causal relationships (Pearl 2009)

Appendix

✠ 2018–10–21: This is the old version, before the approach via conditional degrees of belief

A Introductory notes

A.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,

[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

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

$$\mathrm{P}(\text{clouds in the sky}|\text{ rain on the pavement}, I) > 0.5, \tag{4}$$

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.

A.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)

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.

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].

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 $\mathrm{P}(S|\,G\,DI)$ and $\mathrm{P}(S|\,DI)$ 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

$$\mathrm{P}(S|\,G\,DI) = \mathrm{P}(S|\,DI) \quad \Longleftrightarrow \quad \mathrm{P}(G|\,S\,DI) = \mathrm{P}(G|\,DI) \tag{5}$$

if $\mathrm{P}(S|\,DI)$, $\mathrm{P}(G|\,DI)$ 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)

$$H(\{S\}|\,\{G\},DI) \coloneqq -\sum_{G} \mathrm{P}(G|\,DI) \sum_{S} \mathrm{P}(S|\,G\,DI) \ln \mathrm{P}(S|\,G\,DI), \tag{6}$$

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

$$H(\{S\}|\,DI) \coloneqq -\sum_{S} \mathrm{P}(S|\,DI) \ln \mathrm{P}(S|\,DI) \tag{7}$$

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 § B.

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.

A.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

$$\mathrm{P}(S_{s_1}^{(1)}\, S_{s_2}^{(2)}\, S_{s_3}^{(3)}\, \ldots \,|\, G_g^{(1)}\, G_g^{(2)}\, G_g^{(3)}\, \ldots\, I) = \mathrm{P}(S_{s_{\pi(1)}}^{(1)}\, S_{s_{\pi(2)}}^{(2)}\, S_{s_{\pi(3)}}^{(3)}\, \ldots \,|\, G_g^{(1)}\, G_g^{(2)}\, G_g^{(3)}\, \ldots\, I) \quad (8)$$

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

$$\mathrm{P}(S_{\mathrm{c}}^{(1)}\, S_{\mathrm{a}}^{(2)}\, S_{\mathrm{b}}^{(3)}\, \ldots \,|\, G_g^{(1)}\, G_g^{(2)}\, G_g^{(3)}\, \ldots\, I) = \mathrm{P}(S_{\mathrm{b}}^{(1)}\, S_{\mathrm{c}}^{(2)}\, S_{\mathrm{a}}^{(3)}\, \ldots \,|\, G_g^{(1)}\, G_g^{(2)}\, G_g^{(3)}\, \ldots\, I). \quad (9)$$

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

$$\begin{aligned}\mathrm{P}(S_{\mathrm{a}}^{(1)}\, S_{\mathrm{c}}^{(2)}\, S_{\mathrm{d}}^{(3)}\, S_{\mathrm{a}}^{(4)}\, S_{\mathrm{b}}^{(5)}\, S_{\mathrm{d}}^{(6)}\, \ldots \mid G_{\alpha}^{(1)}\, G_{\beta}^{(2)}\, G_{\beta}^{(3)}\, G_{\gamma}^{(4)}\, G_{\alpha}^{(5)}\, G_{\gamma}^{(6)}\, \ldots\, I) = \\ \mathrm{P}(S_{\mathrm{b}}^{(1)}\, S_{\mathrm{d}}^{(2)}\, S_{\mathrm{c}}^{(3)}\, S_{\mathrm{d}}^{(4)}\, S_{\mathrm{a}}^{(5)}\, S_{\mathrm{a}}^{(6)}\, \ldots \mid G_{\alpha}^{(1)}\, G_{\beta}^{(2)}\, G_{\beta}^{(3)}\, G_{\gamma}^{(4)}\, G_{\alpha}^{(5)}\, G_{\gamma}^{(6)}\, \ldots\, I) \end{aligned} \tag{10}$$

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. (10) is more general than just separately stating exchangeability for the distributions of degrees of belief in the individuals' sharing the same genetic variations. Property (10) 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

A.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

$$\mathrm{P}(S \mid G_1\, DI), \tag{11}$$

or of the combination of any number of variations, say

$$\mathrm{P}(S \mid G_1 \neg G_2 \neg G_3 G_4\, DI). \tag{12}$$

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. '

B First approach: joint degree of belief and mutual information

B.1 Notation

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

$$\begin{aligned} \exp \boldsymbol{x} &:= (\exp x_1, \dots, \exp x_C) & \boldsymbol{x}\,\boldsymbol{y} &:= (x_1\, y_1, \dots, x_C\, y_C) \\ a\boldsymbol{x} &:= (a x_1, \dots, a x_C) & \boldsymbol{x}^a &:= ({x_1}^a, \dots, {x_C}^a) \qquad \text{and so on.} \end{aligned} \tag{13a}$$

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

$$\sum \boldsymbol{x} := x_1 + \dots + x_C \qquad \prod \boldsymbol{x} := x_1 \cdots x_C \tag{13b}$$

so that, for example,

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

Note also the conventions

$$\binom{a}{\boldsymbol{x}} := \binom{a}{x_1, \dots, x_C} := \frac{a!}{x_1! \cdots x_C!} \qquad \prod \binom{\boldsymbol{y}}{\boldsymbol{x}} := \binom{y_1}{x_1} \cdots \binom{y_C}{x_C}, \tag{13c}$$

where the first expression is the multinomial coefficient.

B.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 size of the full population. The Norwegian or European populations amount roughly to 5.3×10^6 and 740×10^6, but we'll see later that it makes sense to consider the limit $N \to \infty$. Our initial information I says that each individual is characterized by two groups of quantities or variates:

1. An insomnia variate $\sigma := (\sigma_1, \sigma_2, \sigma_3) \in \{0,1\}^3$ with $C_\sigma := 8$ 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 $\gamma := (\gamma_1, \dots, \gamma_l) \in \{0,1\}^l$ with $C_\gamma := 2^l$ 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 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 $\xi := (\sigma, \gamma)$ can thus assume $C := C_\sigma \times C_\gamma$ possible values, depending on how many gene alleles we consider.

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

Denote the relative frequency of the insomnia-variate value σ in our data by s_σ, and the frequency distribution of all values by $\boldsymbol{s} := (s_0, \dots, s_7)$. 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 $\boldsymbol{g} := (g_0, \dots, g_{2^l})$. The relative frequency for the joint variate value (σ, γ) is $x_{\sigma,\gamma}$, and the frequency distribution $\boldsymbol{x} := (x_{0,0}, \dots, x_{7,2^l})$. 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

$$\mathrm{p}(\boldsymbol{X}|\, D, I), \qquad \mathrm{p}(\boldsymbol{S}|\, D, I), \qquad \mathrm{p}(\boldsymbol{G}|\, D, I). \tag{14}$$

With N individuals, there are $\binom{N+C-1}{C-1}$ possible distributions $\boldsymbol{X}$; also $\binom{N+C_\sigma-1}{C_\sigma-1}$ possible distributions $\boldsymbol{S}$ and $\binom{N+C_\gamma-1}{C_\gamma-1}$ possible distributions $\boldsymbol{G}$ (Csiszár et al. 2004 § 2.1).

(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:

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, D, I\bigr), \qquad \mathrm{p}\bigl(\sigma^{(0)}|\, D, I\bigr), \qquad \mathrm{p}\bigl(\gamma^{(0)}|\, D, I\bigr). \tag{15}$$

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,

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, I\bigr) = \mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, D, I\bigr) = X_{\sigma^{(0)},\gamma^{(0)}}, \tag{16}$$

that is, the frequency by which the value $\bigl(\sigma^{(0)}, \gamma^{(0)}\bigr)$ appears in the population. 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

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, D, I\bigr) = \sum_{\boldsymbol{X}} \mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, D, I\bigr)\, \mathrm{p}(\boldsymbol{X}|\, D, I) \equiv \sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}}\, \mathrm{p}(\boldsymbol{X}|\, D, I), \tag{17}$$

where the sum comprises $\binom{N+C-1}{C-1}$ terms. This formula says that our degree of belief in the individual's variates equals our expectation of the

frequency of those variates. Formula (17) thus connects the degrees of belief (14) and (15). Analogous equations hold for the marginal frequency distributions $\boldsymbol{S}$, $\boldsymbol{G}$ and the variates $\sigma^{(0)}$, $\gamma^{(0)}$.

If the number of alleles and considered $l > \log_2 n$, 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 § A.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):

$$\begin{aligned} I(\sigma^{(0)} : \gamma^{(0)}|\,D, I) &\coloneqq H(\sigma^{(0)}|\,D, I) - H(\sigma^{(0)}|\,\gamma^{(0)}; D, I) \\ &\equiv \sum_{\sigma^{(0)},\gamma^{(0)}} \mathrm{p}(\sigma^{(0)}, \gamma^{(0)}|\,D, I) \ln \frac{\mathrm{p}(\sigma^{(0)}, \gamma^{(0)}|\,D, I)}{\mathrm{p}(\sigma^{(0)}|\,D, I)\,\mathrm{p}(\gamma^{(0)}|\,D, I)} \\ &\equiv \sum_{\sigma^{(0)},\gamma^{(0)}} \mathrm{p}(\sigma^{(0)}, \gamma^{(0)}|\,D, I) \ln \frac{\mathrm{p}(\sigma^{(0)}|\,\gamma^{(0)}, D, I)}{\mathrm{p}(\sigma^{(0)}|\,D, I)}. \end{aligned} \tag{18}$$

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 third expression shows that it is an average value of the log-ratio between our degrees of belief about the insomnia symptoms conditioned and unconditioned on the genetic data. A mutual information of around 0.1 nats means that the two degrees of belief roughly differ in their second significant digit; generally 10^{-d} nats means a difference in the $(d+1)$th significant digit. The mutual information vanishes if the symptoms and genetic variations are completely irrelevant to one another, and is equal

to the entropy $H\bigl(\sigma^{(0)}|\, D, I\bigr)$ if knowledge of the gene variate gives us complete certainty about the insomnia symptoms, since $H\bigl(\sigma^{(0)}|\, \gamma^{(0)}; D, I\bigr)$ 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

$$I\bigl(\sigma^{(0)} : \gamma^{(0)}|\, I\bigr) = 0, \tag{19}$$

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 (18) given the data we need the joint degrees of belief (15). To calculate the latter we use eq. (17), which needs our degrees of belief about the joint frequency distribution (14). These can be calculated from our initial degrees of belief $\mathrm{p}(\boldsymbol{X}|\, I)$ via Bayes's theorem:

$$\mathrm{p}(\boldsymbol{X}|\, \boldsymbol{x}, n, I) = \frac{\mathrm{p}(\boldsymbol{x}|\, n, \boldsymbol{X}, I)\; \mathrm{p}(\boldsymbol{X}|\, I)}{\sum_{\boldsymbol{X}} \mathrm{p}(\boldsymbol{x}|\, n, \boldsymbol{X}, I)\; \mathrm{p}(\boldsymbol{X}|\, I)}. \tag{20}$$

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:

$$\mathrm{p}(\boldsymbol{x}|\, n, \boldsymbol{X}, I) = \binom{N}{n}^{-1} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}} \equiv \binom{N}{N\boldsymbol{X}}^{-1} \binom{n}{n\boldsymbol{x}} \binom{N-n}{N\boldsymbol{X}-n\boldsymbol{x}} \tag{21}$$

(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). For N very large this distribution simplifies to a multinomial one:

$$\mathrm{p}(\boldsymbol{x}|\, n, \boldsymbol{X}, I) = \binom{n}{n\boldsymbol{x}} \prod \boldsymbol{X}^{n\boldsymbol{x}}. \tag{22}$$

Combining eqs (17), (20), (21), and simplifying we obtain

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{x}, n, I\bigr) = \frac{\sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}}\; \mathrm{p}(\boldsymbol{X}|\, I)\; \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}}{\sum_{\boldsymbol{X}} \mathrm{p}(\boldsymbol{X}|\, I)\; \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}}. \tag{23}$$

With an analogous reasoning we find analogous formulae for $\mathrm{p}(\sigma^{(0)}|\,\boldsymbol{x}, n, I)$ by replacing $\boldsymbol{X}$, $\boldsymbol{x}$ with $\boldsymbol{S}$, $\boldsymbol{s}$; and for $\mathrm{p}(\gamma^{(0)}|\,\boldsymbol{x}, n, I)$ by replacing $\boldsymbol{X}$, $\boldsymbol{x}$ with $\boldsymbol{G}$, $\boldsymbol{g}$. From these we can calculate the mutual information (18). If N is very large the formula above becomes

$$\mathrm{p}(\sigma^{(0)}, \gamma^{(0)}|\,\boldsymbol{x}, n, I) = \frac{\int X_{\sigma^{(0)},\gamma^{(0)}} \left(\prod \boldsymbol{X}^{nx}\right) \mathrm{p}(\boldsymbol{X}|\,I)\,\mathrm{d}\boldsymbol{X}}{\int \left(\prod \boldsymbol{X}^{nx}\right) \mathrm{p}(\boldsymbol{X}|\,I)\,\mathrm{d}\boldsymbol{X}}. \tag{24}$$

B.3 Motivation for the infinite-population limit

In the previous sections we have considered the cases of finite and infinite N. There are two reasons for focusing on the infinite-N case.

First: we are interested in a general connection between insomnia symptoms and gene alleles, independent of the size of the full population. Thus we can consider our sample as coming from a hypothetical population of statistically similar biological characteristics.

Second: if the number of alleles $l \gg \log_2 N$, then every individual will have a unique combination of alleles, leading to degrees of belief (23) about the frequencies $\boldsymbol{G}$ and $\boldsymbol{X}$ that are either zero or unity, since each combination appears only once or not at all. This in turn leads to a mutual information with maximal value. But the reason of this maximal value is in this case not biological, but purely statistical. The way to bypass this is to consider an infinite population size. We will still have to take care of this statistical phenomenon in what follows.

B.4 Calculation: Dirichlet distributions

✠ The results of this section will be improved using the analysis by Good (1965 chs 4–5; 1980)

We need to assess our initial degree of belief for $\boldsymbol{X}$ according to some background knowledge. We'd also like our initial degree of belief to have a convenient mathematical expression.

We choose to express initial state of knowledge I_0 with a Dirichlet distribution ✠ refs, which depends on a positive parameter Θ and a non-negative C-tuple $\boldsymbol{\theta}$ such that $\sum \boldsymbol{\theta} = 1$, which we take equal to $(1/C, \dots, 1/C)$:

$$\mathrm{p}(\boldsymbol{X}|\,\boldsymbol{\theta}, \Theta, I_0)\,\mathrm{d}\boldsymbol{X} = \frac{\Gamma(\Theta)}{\prod \Gamma(\Theta\boldsymbol{\theta})} \prod \boldsymbol{X}^{\Theta\boldsymbol{\theta}-1}\,\mathrm{d}\boldsymbol{X}, \qquad \boldsymbol{\theta} \coloneqq (\underbrace{1/C, \dots, 1/C}_{C \text{ elements}}). \tag{25}$$

Our choice is motivated by the following reasons.

First, this distribution is uniquely determined by the assumption, called 'sufficientness' (Zabell 1982; Diniz et al. 2016) ✠ more refs, that only the data about a specific variate value, say $(\sigma^{(0)}, \gamma^{(0)})$ are relevant for our belief about the frequency $X_{\sigma^{(0)},\gamma^{(0)}}$ of that value:

$$\mathrm{p}(X_{\sigma^{(0)},\gamma^{(0)}} | \boldsymbol{x}, n, \boldsymbol{\theta}, \Theta, I_0)\,\mathrm{d}\boldsymbol{X} \equiv \mathrm{p}(X_{\sigma^{(0)},\gamma^{(0)}} | x_{\sigma^{(0)},\gamma^{(0)}}, n, \boldsymbol{\theta}, \Theta, I_0)\,\mathrm{d}\boldsymbol{X}. \tag{26}$$

This in particular implies that our degrees of belief about the insomnia symptoms or any particular gene allele are not affected by the number of alleles we're considering. That is, if we decide to consider additional genes, our inferences are consistent with the ones previously made when those genes were not considered; likewise if we decide to neglect some genes.

Second, the Dirichlet distribution is the conjugate of the multinomial distribution (22), which means that our degree of belief updated from the data will still be expressed by a Dirichlet distribution but with new parameters Θ', $\boldsymbol{\theta}'$:

$$\mathrm{p}(\boldsymbol{X} | \boldsymbol{x}, n, \boldsymbol{\theta}, \Theta, I_0)\,\mathrm{d}\boldsymbol{X} = \frac{\Gamma(\Theta')}{\prod \Gamma(\Theta'\boldsymbol{\theta}')} \prod \boldsymbol{X}^{\Theta'\boldsymbol{\theta}'-1}\,\mathrm{d}\boldsymbol{X},$$
$$\text{with} \qquad \Theta' := n + \Theta, \quad \boldsymbol{\theta}' := \frac{n\boldsymbol{x} + \Theta\boldsymbol{\theta}}{n + \Theta}. \tag{27}$$

This distribution has mean and covariance matrix

$$\mathrm{E}(\boldsymbol{X} | \Theta, \boldsymbol{\theta}, I_0) = \boldsymbol{\theta}, \qquad \mathrm{cov}(\boldsymbol{X} | \Theta, \boldsymbol{\theta}, I_0) = \frac{1}{\Theta + 1}(\mathrm{diag}\,\boldsymbol{\theta} - \boldsymbol{\theta} \otimes \boldsymbol{\theta}). \tag{28}$$

Equations (27) and (28) show the meanings of the parameters $\boldsymbol{\theta}$, Θ:

✠ Rest of the section has to be changed from here on Let's preliminarily 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 $\boldsymbol{X}$; hence

$$\mathrm{p}(\boldsymbol{X} | I_l) = \binom{N + C - 1}{C - 1}^{-1}. \tag{29}$$

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

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)} \mid I_l\bigr) = \sum_{\boldsymbol{X}} \mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)} \mid \boldsymbol{X}, I_l\bigr)\, \mathrm{p}(\boldsymbol{X} \mid I_l) = \sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \binom{N+C-1}{C-1}^{-1} = 1/C, \quad (30)$$

This state of knowledge is 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, $I_{l'} \neq I_{l''}$. It's important to note that these states of knowledge yield different initial degrees of belief for the marginal frequency distribution of a single gene. As a consequence, if we suddenly decide to consider additional genes, or to neglect some, we cannot compare our inference with the ones previously made.

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. (56), 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 state of knowledge I_0 exists (perhaps it isn't unique) and is characterized by a Dirichlet-multinomial distribution for $\boldsymbol{X}$, which depends on a positive parameter Θ and a non-negative C-tuple $\boldsymbol{\theta}$ such that $\sum \boldsymbol{\theta} = 1$, which we take equal to $(1/C, \dots, 1/C)$:

$$\mathrm{p}(\boldsymbol{X} \mid \Theta, \boldsymbol{\theta}, I_0) = \binom{N+\Theta-1}{N}^{-1} \prod \binom{N\boldsymbol{X} + \Theta\boldsymbol{\theta} - 1}{N\boldsymbol{X}}, \qquad \boldsymbol{\theta} \coloneqq \underbrace{(1/C, \dots, 1/C)}_{C \text{ elements}} \quad (31)$$

(Johnson et al. 1996 § 13.1; Minka 2012 § 3; and especially Basu et al. 1982 §§ 3–4).

This distribution has mean and covariance matrix It also 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 more genes (Basu et al. 1982 §§ 3–4), provided Θ is kept fixed, with no dependence on C for example. This parameter determines our degree of belief about the frequency of single gene variations, for example for the first gene allele, G_{γ_0}: it is a beta-binomial distribution

$$\mathrm{p}(G_{\gamma_0}|\,\Theta,\boldsymbol{\theta},I_0)=\binom{N+\Theta-1}{N}^{-1}\prod\binom{NG_{\gamma_0}+\Theta/2-1}{NG_{\gamma_0}}. \tag{32}$$

The parameter Θ represents the strength of our belief that $\boldsymbol{X}$ is a uniform distribution. It can be considered as a prior number of observations in which we found that the C values of the joint variate came all in equal proportions. Its value – in particularly in comparison with the sample size n – can therefore greatly influence our inferences. We consider two particular values:

- If $\Theta = 2$ then our degree of belief about the frequency of each gene allele is uniform; it doesn't assign equal degrees of belief to the possible frequency distributions for the $C_\sigma \equiv 8$ symptom combinations; but the density of degree of belief for the frequency of controls vs cases is uniform.
- If Θ is large, of the order of millions or more, then our degree of belief about the joint frequency distribution $\boldsymbol{X}$ of all variates is uniform – but no longer uniform for the marginal frequencies.

As shown by Basu & de Bragança Pereira (1982 § 4, Theorem 2), the initial degree of belief I_0 leads to a degree of belief for $N\boldsymbol{X}-n\boldsymbol{x}$ conditional on our data D that has the same form, but with new parameters $n+\Theta$ and $(n\boldsymbol{x}+\Theta\boldsymbol{\theta})/(n+\Theta)$:

$$\mathrm{p}(N\boldsymbol{X}-n\boldsymbol{x}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)=\binom{N+\Theta-1}{N-n}^{-1}\prod\binom{N\boldsymbol{X}+\Theta\boldsymbol{\theta}-1}{N\boldsymbol{X}-n\boldsymbol{x}}, \tag{33}$$

from which we can deduce the degree of belief (20), using (21). Moreover, the expectation of $N\boldsymbol{X}-n\boldsymbol{x}$ is (1982 § 3)

$$\mathrm{E}(N\boldsymbol{X}-n\boldsymbol{x}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)=(N-n)\frac{n\boldsymbol{x}+\Theta\boldsymbol{\theta}}{n+\Theta}, \tag{34}$$

from which we find

$$\mathrm{E}(\boldsymbol{X}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)=\frac{n}{N}x_{\sigma^{(0)},\gamma^{(0)}}+\frac{N-n}{N}\frac{n x_{\sigma^{(0)},\gamma^{(0)}}+\Theta/C}{n+\Theta}. \tag{35}$$

This expression can be combined with eq. (17) to finally find

$$\mathrm{p}(\sigma^{(0)}, \gamma^{(0)} | \boldsymbol{x}, n, \Theta, \boldsymbol{\theta}, I_0) = \frac{n}{N} x_{\sigma^{(0)},\gamma^{(0)}} + \frac{N-n}{N} \frac{n x_{\sigma^{(0)},\gamma^{(0)}} + \Theta/C}{n+\Theta}. \quad (36)$$

From equation (28) we can calculate the variance of our degree of belief in the frequency of a particular variate value:

$$\mathrm{var}(X_{\sigma,\gamma} | \boldsymbol{x}, n, \Theta, \boldsymbol{\theta}, I_0) = \frac{N+n+\Theta}{N(n+\Theta+1)} \frac{n x_{\sigma^{(0)},\gamma^{(0)}} + \Theta/C}{n+\Theta} \left(1 - \frac{n x_{\sigma^{(0)},\gamma^{(0)}} + \Theta/C}{n+\Theta}\right). \quad (37)$$

Calculation: Dirichlet distributions with finite N

✠ Old section, kept for reference

We need to assess our initial degree of belief for $\boldsymbol{X}$ according to some background knowledge. Let's preliminarily 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 $\boldsymbol{X}$; hence

$$\mathrm{p}(\boldsymbol{X} | I_l) = \binom{N+C-1}{C-1}^{-1}. \quad (38)$$

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

$$\mathrm{p}(\sigma^{(0)}, \gamma^{(0)} | I_l) = \sum_{\boldsymbol{X}} \mathrm{p}(\sigma^{(0)}, \gamma^{(0)} | \boldsymbol{X}, I_l)\, \mathrm{p}(\boldsymbol{X} | I_l) = \sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \binom{N+C-1}{C-1}^{-1} = 1/C, \quad (39)$$

This state of knowledge is 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, $I_{l'} \neq I_{l''}$. It's important to note that these states of knowledge yield different initial degrees of belief for the marginal frequency distribution of a single gene. As a consequence, if we suddenly decide to consider additional genes, or to neglect some, we cannot compare our inference with the ones previously made.

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. (**??**), 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 state of knowledge I_0 exists (perhaps it isn't unique) and is characterized by a Dirichlet-multinomial distribution for $\boldsymbol{X}$, which depends on a positive parameter Θ and a non-negative C-tuple $\boldsymbol{\theta}$ such that $\sum \boldsymbol{\theta} = 1$, which we take equal to $(1/C, \dots, 1/C)$:

$$\mathrm{p}(\boldsymbol{X}|\,\Theta, \boldsymbol{\theta}, I_0) = \binom{N+\Theta-1}{N}^{-1} \prod \binom{N\boldsymbol{X}+\Theta\boldsymbol{\theta}-1}{N\boldsymbol{X}}, \qquad \boldsymbol{\theta} \coloneqq (\underbrace{1/C, \dots, 1/C}_{C \text{ elements}}) \tag{40}$$

(Johnson et al. 1996 § 13.1; Minka 2012 § 3; and especially Basu et al. 1982 §§ 3–4).

This distribution has mean and covariance matrix

$$\mathrm{E}(\boldsymbol{X}|\,\Theta, \boldsymbol{\theta}, I_0) = \boldsymbol{\theta}, \qquad \mathrm{cov}(\boldsymbol{X}|\,\Theta, \boldsymbol{\theta}, I_0) = \frac{\Theta+N}{N\,(\Theta+1)}(\mathrm{diag}\,\boldsymbol{\theta} - \boldsymbol{\theta}\otimes\boldsymbol{\theta}). \tag{41}$$

It also 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 more genes (Basu et al. 1982 §§ 3–4), provided Θ is kept fixed, with no dependence on C for example. This parameter determines our degree of belief about the frequency of single gene variations, for example for the first gene allele, G_{γ_0}: it is a beta-binomial distribution

$$\mathrm{p}(G_{\gamma_0}|\,\Theta, \boldsymbol{\theta}, I_0) = \binom{N+\Theta-1}{N}^{-1} \prod \binom{NG_{\gamma_0}+\Theta/2-1}{NG_{\gamma_0}}. \tag{42}$$

The parameter Θ represents the strength of our belief that $\boldsymbol{X}$ is a uniform distribution. It can be considered as a prior number of observations in which we found that the C values of the joint variate came all in equal proportions. Its value – in particularly in comparison with the sample size n – can therefore greatly influence our inferences. We consider two particular values:

– If $\Theta = 2$ then our degree of belief about the frequency of each gene allele is uniform; it doesn't assign equal degrees of belief to the possible frequency distributions for the $C_\sigma \equiv 8$ symptom combinations; but the density of degree of belief for the frequency of controls vs cases is uniform.

– If Θ is large, of the order of millions or more, then our degree of belief about the joint frequency distribution $\boldsymbol{X}$ of all variates is uniform – but no longer uniform for the marginal frequencies.

As shown by Basu & de Bragança Pereira (1982 § 4, Theorem 2), the initial degree of belief I_0 leads to a degree of belief for $N\boldsymbol{X}-n\boldsymbol{x}$ conditional on our data D that has the same form, but with new parameters $n+\Theta$ and $(n\boldsymbol{x}+\Theta\boldsymbol{\theta})/(n+\Theta)$:

$$\mathrm{p}(N\boldsymbol{X}-n\boldsymbol{x}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)=\binom{N+\Theta-1}{N-n}^{-1}\prod\binom{N\boldsymbol{X}+\Theta\boldsymbol{\theta}-1}{N\boldsymbol{X}-n\boldsymbol{x}}, \tag{43}$$

from which we can deduce the degree of belief (**??**), using (**??**). Moreover, the expectation of $N\boldsymbol{X}-n\boldsymbol{x}$ is (1982 § 3)

$$\mathrm{E}(N\boldsymbol{X}-n\boldsymbol{x}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)=(N-n)\frac{n\boldsymbol{x}+\Theta\boldsymbol{\theta}}{n+\Theta}, \tag{44}$$

from which we find

$$\mathrm{E}(\boldsymbol{X}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)=\frac{n}{N}x_{\sigma^{(0)},\gamma^{(0)}}+\frac{N-n}{N}\frac{nx_{\sigma^{(0)},\gamma^{(0)}}+\Theta/C}{n+\Theta}. \tag{45}$$

This expression can be combined with eq. (**??**) to finally find

$$\mathrm{p}(\sigma^{(0)},\gamma^{(0)}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)=\frac{n}{N}x_{\sigma^{(0)},\gamma^{(0)}}+\frac{N-n}{N}\frac{nx_{\sigma^{(0)},\gamma^{(0)}}+\Theta/C}{n+\Theta}. \tag{46}$$

From equation (41) we can calculate the variance of our degree of belief in the frequency of a particular variate value:

$$\mathrm{var}(X_{\sigma,\gamma}|\,\boldsymbol{x},n,\Theta,\boldsymbol{\theta},I_0)= \frac{N+n+\Theta}{N\,(n+\Theta+1)}\frac{nx_{\sigma^{(0)},\gamma^{(0)}}+\Theta/C}{n+\Theta}\left(1-\frac{nx_{\sigma^{(0)},\gamma^{(0)}}+\Theta/C}{n+\Theta}\right). \tag{47}$$

Constant initial degree of belief

✠ This section has become irrelevant. Left here just for reference

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 $\boldsymbol{X}$; hence

$$\mathrm{p}(\boldsymbol{X}|\, I_l) = \binom{N+C-1}{C-1}^{-1}. \tag{48}$$

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

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, I_l\bigr) = \sum_{\boldsymbol{X}} \mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{X}, I_l\bigr)\, \mathrm{p}(\boldsymbol{X}|\, I_l) = \sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \binom{N+C-1}{C-1}^{-1} = 1/C, \tag{49}$$

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. (23) simplifies to

$$\mathrm{p}\bigl(\sigma^{(0)}, \gamma^{(0)}|\, \boldsymbol{x}, n, I_l\bigr) = \frac{\sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}}{\sum_{\boldsymbol{X}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}}. \tag{50}$$

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

$$\sum_{\boldsymbol{X}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}} = \binom{N+C-1}{n+C-1}, \tag{51}$$

which, substituted in eq. (20), also leads to

$$\mathrm{p}(\boldsymbol{X}|\, \boldsymbol{x}, n, I_l) = \binom{N+C-1}{n+C-1}^{-1} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}}. \tag{52}$$

The sum in the numerator of eq. (50) can be rewritten ✠ explain the steps obtaining

$$\sum_{\boldsymbol{X}} X_{\sigma^{(0)},\gamma^{(0)}} \prod \binom{N\boldsymbol{X}}{n\boldsymbol{x}} = \frac{n x_{\sigma^{(0)},\gamma^{(0)}} + 1}{N} \binom{N+C}{n+C} - \frac{1}{N} \binom{N+C-1}{n+C-1}. \tag{53}$$

Simplifying we finally find

$$\mathrm{p}\big(\sigma^{(0)}, \gamma^{(0)} | \boldsymbol{x}, n, I_l\big) = \frac{(N+C) n x_{\sigma^{(0)},\gamma^{(0)}} + N - n}{N\,(n+C)} \tag{54}$$

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. (49):

$$\mathrm{p}\big(\sigma^{(0)}, \gamma^{(0)} | \boldsymbol{x}, n, I_l\big) \propto x_{\sigma^{(0)},\gamma^{(0)}} + \frac{N-n}{n} \frac{C}{N+C} \frac{1}{C}, \tag{55}$$

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. (16).

For the marginal distributions we obtain, by summation,

$$\mathrm{p}\big(\sigma^{(0)} | \boldsymbol{x}, n, I_l\big) = \frac{(N+C) n s_{\sigma^{(0)}} + (N-n) C_\gamma}{N\,(n+C)}, \tag{56}$$

$$\mathrm{p}\big(\gamma^{(0)} | \boldsymbol{x}, n, I_l\big) = \frac{(N+C) n g_{\gamma^{(0)}} + (N-n) C_\sigma}{N\,(n+C)}. \tag{57}$$

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 (54) and (56) in the formula for the mutual information (18) and simplifying we find

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

$$\ln \frac{(N+C) n x_{\sigma^{(0)},\gamma^{(0)}} + N - n}{[(N+C) n s_{\sigma^{(0)}} + (N-n) C_\gamma]\,[(N+C) n g_{\gamma^{(0)}} + (N-n) C_\sigma]} \tag{58}$$

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 $g_{\gamma^{(0)}} = x_{\sigma^{(0)},\gamma^{(0)}} = 0$:

$$
\begin{aligned}
I\big(\sigma^{(0)} : \gamma^{(0)} | \, D, I_l\big) &= \ln[N\,(n+C)] \\
&+ \sum_{\sigma^{(0)}} \sum_{\gamma^{(0)} \in D} \frac{(N+C)n x_{\sigma^{(0)},\gamma^{(0)}} + N - n}{N\,(n+C)} \times \\
&\qquad \ln \frac{(N+C)n x_{\sigma^{(0)},\gamma^{(0)}} + N - n}{[(N+C)n s_{\sigma^{(0)}} + (N-n)C_\gamma]\,[(N+C)n g_{\gamma^{(0)}} + (N-n)C_\sigma]} \\
&- C_\gamma^- \frac{N-n}{N\,(n+C)} \sum_{\sigma^{(0)}} \ln\{C_\sigma\,[(N+C)n s_{\sigma^{(0)}} + (N-n)C_\gamma]\}
\end{aligned}
\tag{59}
$$

[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.

Basu, D., de Bragança Pereira, C. A. (1982): *On the Bayesian analysis of categorical data: the problem of nonresponse*. J. Stat. Plann. Infer. **6**[4], 345–362.

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.

Diniz, M. A., De Bock, J., Van Camp, A. (2016): *Characterizing Dirichlet priors*. American Statistician **70**[1], 9–17. http://users.ugent.be/~jdbock/documents/MD-2015-TAS-paper.pdf.

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).

— (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.

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).

Good, I. J. (1965): *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. (MIT Press, Cambridge, USA).

— (1980): *Some history of the hierarchical Bayesian methodology*. Trabajos de Estadística y de Investigación Operativa **31**[1], 489–519. Repr. in Good (1983), ch. 9, pp. 95–105.

— (1983): *Good Thinking: The Foundations of Probability and Its Applications*. (University of Minnesota Press, Minneapolis, USA).

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.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1996): *Discrete Multivariate Distributions*. (Wiley, New York). First publ. 1969 in chapter form.

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.

MacKay, D. J. C., Bauman Peto, L. C. (1995): *A hierarchical Dirichlet language model*. Nat. Lang. Eng. **1**[3], 289–307.

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

Minka, T. P. (2012): *Estimating a Dirichlet distribution*. https://tminka.github.io/papers/. First publ. 2000.

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

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.

Zabell, S. L. (1982): *W. E. Johnson's "sufficientness" postulate*. Ann. Stat. **10**^{4}, 1090–1099. Repr. in Zabell (2005 pp. 84–95).

— (2005): *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. (Cambridge University Press, Cambridge).