Does our DNA keep us awake?

Cüneyt
<cuneyt.guzey@ntnu.no>

Daniela
<daniela.bragantini@ntnu.no>

Luca
<piero.mana@ntnu.no>

Yasser
<yasser.roudi@ntnu.no>

Draft of 13 November 2018 (first drafted 22 August 2018)

abstract

1 Problem setup

✠ Some intro about insomnia and its symptoms here

Every single-nucleotide polymorphism (SNP), together with the huge variety of external factors, can in principle affect the appearance of insomnia 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.

An indication of the causal strength of one or more SNPs on one or more 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; the latter also accounts for the complexity of the causal mechanisms and our uncertainty about them.

Let's focus for the moment the simplest reduction of such a graph, with only one particular SNP with two alleles a, b, and one symptom S. In an arbitrarily large population, consider the fraction $f_{S|a}$ of individuals that show symptom S among those having allele a, and the fraction $f_{S|b}$ among those having allele b. These are *conditional relative frequencies*. If these conditional frequencies are markedly different, then we can conclude that the SNP must have some causal relevance, however indirect, for the symptom.

These conditional frequencies are far easier to study than causal relations, 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 (1) the conditional frequencies in a population sample, (2) a representation of our initial information about such frequencies. The idea behind the calculation is 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 an 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.

✠ 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, unlike a 'significance' test. Rather, we will find a graduation of cases: from frequencies predicted to be clearly distinct, to frequencies 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.

In the next sections we use formula (1) to estimate the limit frequencies given a sample of 6029 individuals from ✠ *** details here. We shall first focus on the conditional frequencies of each insomnia symptom given one SNP at a time. The inferences in this case are very simple, intuitive,

and easily visualizable. In the subsequent sections we shall consider the study of each symptom given *pairs* of SNPs, and the study of all possible eight *symptom combinations* given one SNP at a time.

2 Inference

✠ Add some remarks about the role of these 'limit frequencies'. Relation to partial exchangeability and belief about next individual in an endless sequence.

Consider for definiteness symptomn A and a particular SNP with alleles a, b. The limit conditional frequencies are denoted $f_{A|a}$ and $f_{A|b}$. Denote the sample data by D and our initial beliefs by I. We want to quantify our joint belief about these frequencies, expressed by the density function

$$\mathrm{p}(f_{A|a}, f_{A|b} \,|\, D, I). \tag{2}$$

According to Bayes's theorem (1), the belief above is proportional to the product of $\mathrm{p}(D|\, f_{A|a}, f_{A|b}, I)$ and our initial belief $\mathrm{p}(f_{A|a}, f_{A|b} \,|\, I)$. Let's consider these in turn.

In our hypothetical large population a fraction $f_{A|a}$ of individuals with allele a shows symptom A, and a fraction $1 - f_{A|a}$ doesn't show that

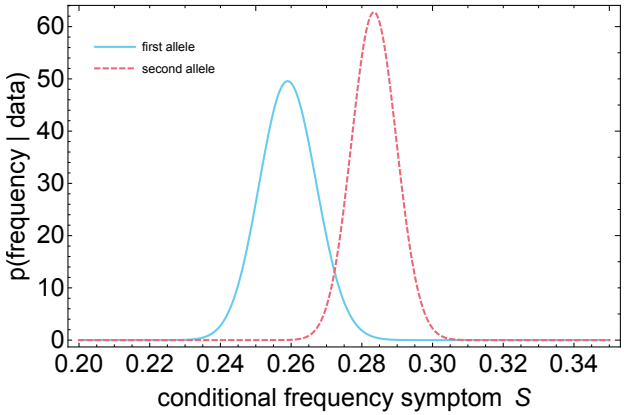

Figure 1 Example of distributions of belief

symptom. Then, upon sampling N_a individuals with allele a, our belief that a fraction $n_{A|a}$ of these will show symptom A and a fraction $1 - n_{A|a}$ won't show this symptom is

$$\mathrm{p}(n_{A|a} \,|\, N_a, f_{A|a}, I) = {f_{A|a}}^{N_a\, n_{A|a}}\, (1 - f_{A|a})^{N_a\,(1-n_{A|a})}, \tag{3}$$

and analogously for allele b. Our belief about obtaining data $D = (n_{A|a}, n_{A|b})$ is therefore

$$\mathrm{p}(D \,|\, N_a, N_b, f_{A|a}, f_{A|b}, I) = \prod_{x=a,b} {f_{A|x}}^{N_x\, n_{A|x}}\, (1 - f_{A|x})^{N_x\,(1-n_{A|x})}. \tag{4}$$

Our initial degree of belief about the limit conditional frequencies is based on these assumptions:

(1) we expect the frequencies conditional on the two alleles not to be wildly different in comparison to the full range $[0, 1]$;

(2) besides the point above, we don't want to make any special assumptions about the values of the two conditional frequencies.

The first assumption is conservative; therefore, if our updated belief conditional on the data will show clearly distinct conditional frequencies, it will be because the data have given enough evidence to overwhelm our initial conservative belief.

The kind of density representing this initial belief is qualitatively shown in fig. 2. Mathematically we write it as an integral:

$$\mathrm{p}(f_{A|a}, f_{A|b} \,|\, I) = \int_0^\infty \mathrm{d}\alpha \int_0^1 \mathrm{d}\nu\; \beta(f_{A|a} \,|\, \alpha, \nu)\, \beta(f_{A|b} \,|\, \alpha, \nu)\, \pi(\alpha, \nu), \tag{5}$$

where $\beta(\,\cdot \,|\, \alpha, \nu)$ is a beta density with shape parameters $\alpha\nu$ and $\alpha\,(1-\nu)$:

$$\beta(f \,|\, \alpha, \nu) \coloneqq \frac{\Gamma(\alpha)}{\Gamma(\alpha\nu)\,\Gamma[\alpha\,(1-\nu)]}\; f^{\alpha\nu}\, (1-f)^{\alpha\,(1-\nu)}, \tag{6}$$

and π is a density that we leave unspecified for the moment: several different expressions for π will be used, to test how much our inferences depend on our initial belief.

We write $\mathrm{p}(f_{A|a}, f_{A|b} \,|\, I)$ in the above fashion, with integrals, only to emphasize that our beliefs about $f_{A|a}$ and $f_{A|b}$ are not disjoint, and to show how their mutual dependence comes about: by first considering two independent densities having the same parameters, and then mixing over these parameters.

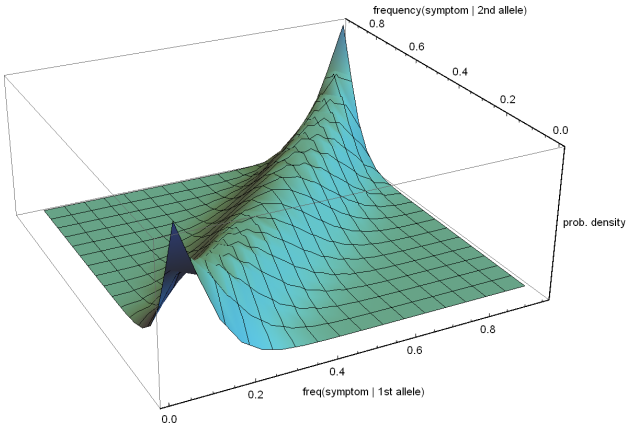

Figure 2 Initial belief

This construction has the interesting consequence that the our updated belief $\mathrm{p}(f_{A|a}, f_{A|b}|\, D, I)$ can be formally written in the following way:

$$\mathrm{p}(f_{A|a}, f_{A|b}|\, D, I) = \int \mathrm{d}\alpha \int \mathrm{d}\nu\; \mathrm{p}(f_{A|a}, f_{A|b}|\, \alpha, \nu, D, I)\; \mathrm{p}(\alpha, \nu|\, D, I), \quad (7)$$

with

$$\mathrm{p}(\alpha, \nu|\, D, I) \propto \int \mathrm{d}f_{A|a} \int \mathrm{d}f_{A|b}\; \mathrm{p}(D|\, f_{A|a}, f_{A|b}, I)\; \mathrm{p}(f_{A|a}, f_{A|b}|\, \alpha, \nu, I). \quad (8)$$

This representation is discussed in appendix***. It is as if α, ν were unknown parameters, and our update for the frequencies proceeded by first updating our belief about the parameters and then marginalizing them out. This is a so-called *hierarchic* model (Good 1980). A point rarely emphasized in the literature is that there is no mathematical difference between a hierarchic and a non-hierarchic model: we could forget about the integrals in formula (5) and about the update formula (7), and simply treat $\mathrm{p}(f_{A|a}, f_{A|b}|\, I)$ as the density depicted in fig. 2, with update (1). The results would be the same. The hierarchic way of thinking, though, often helps in constructing densities that better represent our initial beliefs, and also leads to formulae that can be better approximated when exact computation is unfeasible.

3 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 $\theta \coloneqq \theta_0 + \theta_1$:

$$\mathrm{p}(f_1 | \, n_0, n_1, \theta_0, \theta_1) \, \mathrm{d}f_1 = \frac{\Gamma(n+\theta)}{\Gamma(n_0+\theta_0)\,\Gamma(n_1+\theta_1)} \, (1-f_1)^{n_0+\theta_0-1} \, {f_1}^{n_1+\theta_1-1} \, \mathrm{d}f_1 \quad (9)$$

This distribution has expected value and variance

$$\begin{aligned} \mathrm{E}(f_1 | \, n_0, n_1, \theta_0, \theta_1) &= \frac{n_1+\theta_1}{n+\theta}, \\ \mathrm{var}(f_1 | \, n_0, n_1, \theta_0, \theta_1) &= \frac{(n_0+\theta_0)\,(n_1+\theta_1)}{(n+\theta)^2\,(n+\theta+1)}. \end{aligned} \quad (10)$$

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

Thanks

PGLPM thanks Mari & Miri for continuous encouragement and affection, and to Buster Keaton and Saitama for filling life with awe and inspiration. To the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Bibliography

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

Good, I. J. (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).

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

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