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 27 November 2018 (first drafted 22 August 2018)

abstract

1 SNPs and insomnia

1.1 Introduction and goals

✠ 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. This is the approach of genetic association studies ✠ ***ref.

The present work has two distinct goals:

First, we present evidence of a strong association between some SNPs located in *** and the three main insomnia symptoms ✠ name the relevant SNPs here?, within population sampled in***. Our results involve associations between each symptom and several SNPs indiviudally, and also associations between each symptom and *pairs* of SNPs; the latter result shows different kinds of interaction between the alleles of a SNP pair.

Second, we give a detailed but intuitive discussion of the Bayesian method used to infer the associations described above. Similar methods have been used for other kinds of association studies, for example contextual text prediction (MacKay et al. 1995) and population-specific allele count (Lockwood et al. 2001). This method gives simple, clear, and intuitively understandable inferences about symptom-SNP associations; it is computationally fast; and it is easy to generalize to association studies of symptom *combinations* vs *multiple* SNPs, as we'll show in later sections.

The Method section of this work focus on explaining our Bayesian approach, but use the real data from which our results are derived. In the subsequent Result section we discuss the different relevant associations found.

1.2 The data

✠ description of our data here

2 Methods

2.1 Outline

We focus on the simplest kind of association: between one particular SNP with two alleles a, b, and one insomnia symptom. For example, we could be speaking of the SNP `rs875994` with alleles a = C, b = T, and onset insomnia O.

We imagine to have an arbitrarily large population from the same genetic pool as our sample. In this population we would like to know how large are the fraction $f_{|a}$ of individuals that show the symptom among those having allele a, and the fraction $f_{|b}$ that show the symptom among those having allele b. These two fractions are *conditional relative frequencies* (note that $f_{|a} + f_{|b} \neq 1$ in general, since we are not speaking about the frequencies of two mutually exclusive and exhaustive alternatives, but of frequencies *conditional* on such alternatives). ✠ Add some remarks about the role of 'limit frequencies' and 'arbitrarily large population'? Relation to partial exchangeability and belief about next individual in an endless sequence.

We are in particular interested in knowing how much these two conditional frequencies differ, that is, in $f_{|a} - f_{|b} =: \Delta f$. A larger difference may indicate some sort of biologic association between the SNP (or another

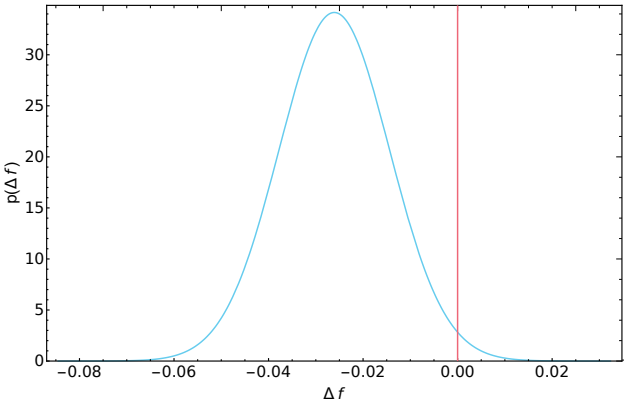

Figure 1 Example of uncertainty distribution about the conditional-frequency difference Δf

SNP linked to it) and the symptom. This difference in the large population is unknown, however; we can only make a plausible inference about its value from sampled data and other initial information. The most detailed statement we can make about it is by quantifying the distribution of our degree of belief $\mathrm{p}(\Delta f)$ like the one plotted in fig. 1. This distribution gives us many pieces of information: for example, a negative difference Δf is more plausible than a positive one; it is 8.3% plausible that $|\Delta f| < 0.01$, and therefore 91.7% plausible that $|\Delta f| > 0.01$; it is 90% plausible that $-0.045 < \Delta f < -0.007$; and so on. Several measures can be chosen to summarize the distribution with one number. For example, we could use the plausibility that $|\Delta f|$ exceeds a given value, say $\mathrm{p}(|\Delta f| \geqslant 0.01)$. A roughly equivalent but perhaps more preferable measure is the minimal frequency difference we are highly sure about:

$$x \quad \text{such that} \quad \mathrm{p}(|\Delta f| \geqslant x) = 0.9. \tag{1}$$

For the plot of fig. 1 such measure is 0.0112. ✠ add something about dependence of broadness on sample size, and 'smoothing' as discussed by MacKay & Bauman Peto (1995 § 2.6).

A more complete quantification of our uncertainty is the distribution $\mathrm{p}(f_{|a}, f_{|b})$ among the possible joint values of the two conditional frequencies, from which $\mathrm{p}(\Delta f)$ can be calculated. This joint distribution is also the optimal starting point when frequencies conditional on allele combinations of several SNP are considered. Our goal is therefore to

quantify this joint degree of belief, given: (1) the conditional frequencies in a population sample, (2) our initial information or guesses about such frequencies. In formulae, we want to assign a numerical value to

$$\mathrm{p}(\text{conditional frequencies}|\,\text{sample data}, \text{initial information}).$$

In the rest of this section we shall calculate this degree of belief by methodically applying the probability calculus (Jeffreys 1983; Cox 1946; Jaynes 2003; Hailperin 1996).

First let's observe that the approach just outlined is not dichotomous, unlike a classical significance test. We are not asking whether the frequency difference is 'significant' or not. Rather, we find a gradation of cases: from conditional frequencies likely to be very distinct, to conditional frequencies likely to be very similar. These cases can be sorted, for example with the measure of formula (1), 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.

2.2 Inference: concrete calculation

First step is to use Bayes's theorem:

$$\mathrm{p}(\text{frequencies}|\,\text{data}, \text{initial info}) \propto$$
$$\mathrm{p}(\text{data}|\,\text{frequencies}, \text{initial info}) \times \mathrm{p}(\text{frequencies}|\,\text{initial info}). \quad (2)$$

The first degree of belief in the product above is given by a simple sampling formula, which we'll discuss shortly. The second can be modelled in several reasonable ways; we'll see, however, that they all lead to very similar conclusions about the conditional frequencies and their difference, owing to our large sample size.

In this section we do step by step the calculations outlined above. For definiteness we consider onset insomnia (s) and the SNP rs875994 with alleles a, b. The limit conditional frequencies are denoted f_a and f_b. The sample data, denoted by D, consist of the number of individuals F_a that show symptom s among the sampled individuals having allele a, and the number F_b showing the same symptom among those having allele b. Our initial information consists in the number N_a of sampled individuals

with allele a, and the number N_b with allele b. The total number of sampled individuals is therefore $N_a + N_b$. This initial information and our initial beliefs are denoted by I; the numbers N_a, N_b will often be indicated explicitly even though they're part of I.

Our belief about the joint conditional frequencies is expressed by the density function

$$\mathrm{p}(f_a, f_b | D, I). \tag{3}$$

According to Bayes's theorem (2), the belief above is proportional to the product of $\mathrm{p}(D | f_a, f_b, I)$ and our initial belief $\mathrm{p}(f_a, f_b | I)$. Let's consider these in turn.

In our hypothetical large population a fraction f_a of individuals having allele a shows symptom s, and a fraction $1 - f_a$ doesn't show that symptom. Then, upon sampling N_a individuals with allele a, our belief that a fraction F_a of these will show symptom s and a fraction $1 - F_a$ won't show this symptom is

$$\mathrm{p}(F_a | f_a, N_a, I) = \binom{N_a}{N_a F_a} {f_a}^{N_a F_a} (1 - f_a)^{N_a (1 - F_a)}, \tag{4}$$

and analogously for allele b. Our belief about obtaining data $D = (F_a, F_b)$ is therefore

$$\mathrm{p}(D | f_a, f_b, N_a, N_b, I) = \prod_{x=a,b} \binom{N_x}{N_x F_{s|x}} {f_{s|x}}^{N_x F_{s|x}} (1 - f_{s|x})^{N_x (1 - F_{s|x})}. \tag{5}$$

Our initial degree of belief about the limit conditional frequencies is based on the following main assumption: we expect the frequencies conditional on the two alleles, f_a and f_b, not to be wildly different. This is a conservative assumption. 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.

This initial belief is represented by a density qualitatively shown in fig. 2. Note how most of our belief's mass is concentrated along the diagonal of the coordinates (f_a, f_b). ✠ Say something about the rises at

the edges of the diagonal. Mathematically we write this density as an integral:

$$\mathrm{p}(f_a, f_b \mid I) = \int_0^\infty \mathrm{d}u \int_0^\infty \mathrm{d}v \; \mathrm{p}(f_a, f_b \mid u, v, I) \; \pi(u, v \mid I) \tag{6}$$

with

$$\mathrm{p}(f_a, f_b \mid u, v, I) \coloneqq \beta(f_a \mid u, v) \; \beta(f_b \mid u, v) \tag{7}$$

where $\beta(f \mid u, v)$ is a beta density with shape parameters u and v:

$$\beta(f \mid u, v) \coloneqq \frac{\Gamma(u+v)}{\Gamma(u)\,\Gamma(v)} \; f^{u-1} \, (1-f)^{v-1}, \qquad u, v > 0, \tag{8}$$

and π is a normalized density that we leave unspecified for the moment: several different choices for π will be used, to test how much our inferences depend on our initial belief. We discuss these choices in §*** together with the meaning of the parameters u, v.

We express $\mathrm{p}(f_a, f_b \mid I)$ as an integral to emphasize that our initial beliefs about f_a and f_b are *not* disjoint, and to show how their mutual dependence comes about: by first considering two independent densities

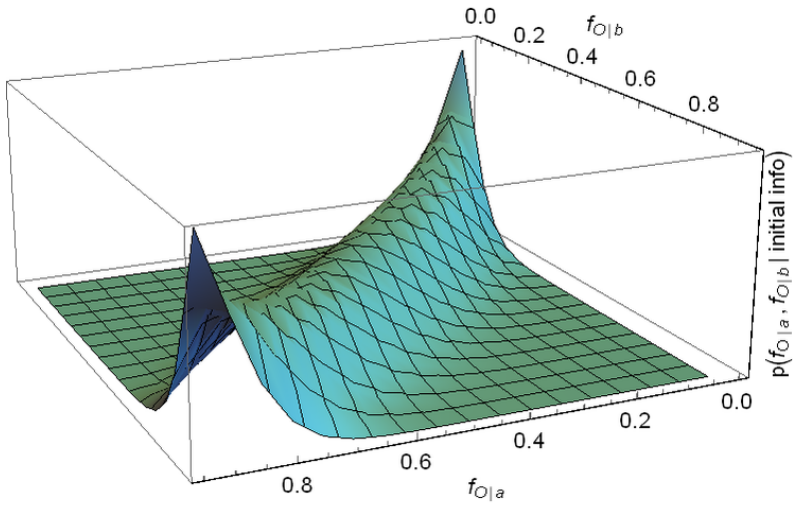

Figure 2 Qualitative plot of our initial belief

having the same parameters, and then mixing their product over these parameters.

This construction has the interesting consequence for our updated belief $\mathrm{p}(f_a, f_b \mid D, I)$. According to Bayes's theorem (2) combined with our initial belief (6) it is given by

$$\mathrm{p}(f_a, f_b \mid D, I) \propto \mathrm{p}(D \mid f_a, f_b \mid I)\, \mathrm{p}(f_a, f_b \mid I), \tag{9}$$

but, as shown in appendix A, it can also be written in the following way:

$$\mathrm{p}(f_a, f_b \mid D, I) = \int \mathrm{d}u \int \mathrm{d}v\; \mathrm{p}(f_a, f_b \mid D, u, v, I)\; \pi(u, v \mid D, I), \tag{10}$$

with the normalized density $\pi(u, v \mid D, I)$ defined by

$$\pi(u, v \mid D, I) \propto \pi(u, v \mid I) \times \underbrace{\int \mathrm{d}f_a \int \mathrm{d}f_b\; \mathrm{p}(D \mid f_a, f_b, I)\; \mathrm{p}(f_a, f_b \mid u, v, I)}_{=:\, \mathrm{p}(D \mid u, v, I)}. \tag{11}$$

It is as if u, v were unknown parameters with initial belief density $\pi(u, v \mid I)$, and our update for the frequencies proceeded by first updating our belief about the parameters, eq. (11), and then marginalizing them out, eq. (10). This is a so-called *hierarchic* model (Good 1980). This hierarchic way of thinking 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. A point rarely emphasized in the literature, though, is that there is no mathematical difference between a hierarchic and a non-hierarchic model: we could forget about the integrals in formula (6) and about the update formula (10), and simply treat $\mathrm{p}(f_a, f_b \mid I)$ as the density depicted in fig. 2, with update (2). The results would be the same. Further discussion about the formulae above is given in §***.

With large sample sizes the density $\pi(u, v \mid D, I)$ turns out to be so peaked with respect to $\mathrm{p}(f_a, f_b \mid u, v, I)$ that it can be considered as a Dirac delta centred on the parameters u_{M}, v_{M} that maximize it. We thus obtain a good approximation of the updated belief (10) that doesn't involve parameter integration:

$$\mathrm{p}(f_a, f_b \mid D, I) \approx \mathrm{p}(f_a, f_b \mid u_{\mathrm{M}}, v_{\mathrm{M}}, I) \quad \text{with} \quad (u_{\mathrm{M}}, v_{\mathrm{M}}) \coloneqq \arg\max_{u,v} \pi(u, v \mid D, I). \tag{12}$$

The maximum of $\pi(u, v|\, D, I)$, or better of its logarithm, can easily be found with most optimization methods. The explicit expression to be optimized, discussed in appendix***,

$$k\,[\ln\Gamma(\textstyle\sum_s \boldsymbol{u}_s) - \sum_s \ln\Gamma(\boldsymbol{u}_s)] - \sum_{x=a,b} [\ln\Gamma(N_x + \textstyle\sum_s \boldsymbol{u}_s) - \sum_s \ln\Gamma(N_x F_{s|x} + \boldsymbol{u}_s)] \tag{13}$$

Combininig eqs (12), (7), (8) we thus arrive at the following explicit approximate formula for our belief about the conditional frequencies given the data:

$$\mathrm{p}(f_a, f_b|\, D, I) \approx \beta(f_a|\, u_\mathrm{M}, v_\mathrm{M})\ \beta(f_b|\, u_\mathrm{M}, v_\mathrm{M}) \qquad \text{with } (u_\mathrm{M}, v_\mathrm{M}) \text{ maximizing (13).} \tag{14}$$

✠ ref here to (MacKay 1996)

2.3 Generalization to symptom combinations

2.4 Choices of initial beliefs and meaning of the parameters

The parameters $\boldsymbol{u}$ of the Dirichlet density for the frequency distribution $\boldsymbol{f}$ have very intuitive meanings, even more evident if we rewrite them as their sum and a pair of normalized parameters:

$$\alpha \coloneqq \textstyle\sum \boldsymbol{u}, \qquad \boldsymbol{\nu} \coloneqq \boldsymbol{u} / \sum \boldsymbol{u}. \tag{15}$$

The normalized parameters $\boldsymbol{\nu}$ are the expected frequency distribution:

$$\mathrm{E}(\boldsymbol{f}|\, \boldsymbol{u}) = \boldsymbol{\nu} \tag{16}$$

The sum of the parameters α expresses the sharpness of the distribution, as seen from the covariance matrix:

$$\mathrm{var}(f_i|\, \boldsymbol{u}) = \frac{\nu_i\,(1-\nu_i)}{\alpha+1}, \qquad \mathrm{cov}(f_i\, f_j|\, \boldsymbol{u}) = -\frac{\nu_i\,\nu_j}{\alpha+1}. \tag{17}$$

In fact, from the update formula*** we see that α quantifies the amount of data necessary to modify our initial belief.

$$\frac{\mathrm{E}(f_{\mathrm{O}|a} - f_{\mathrm{O}|b}|\, D, I)}{\sigma(f_{\mathrm{O}|a} - f_{\mathrm{O}|b}|\, D, I)}$$

3 Results

Appendices

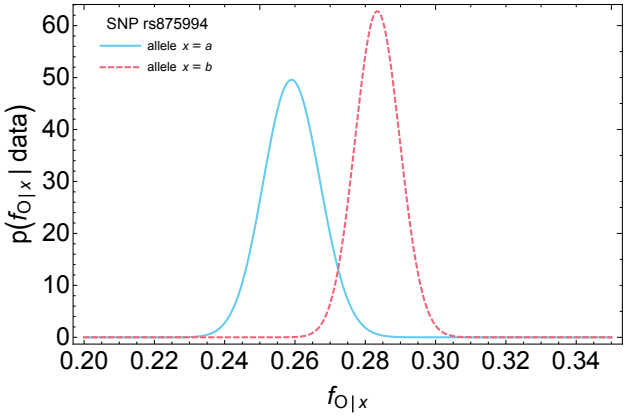

Figure 3 Example of distributions of belief

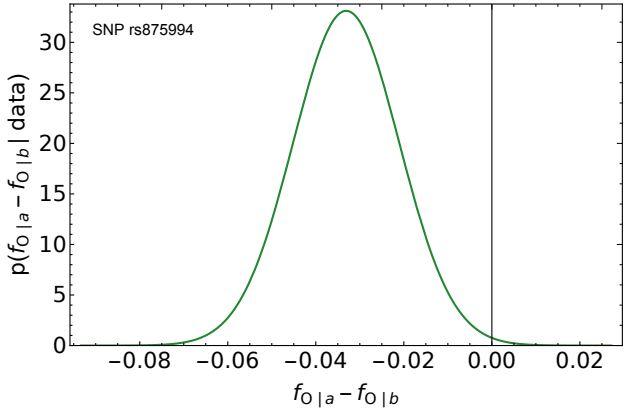

Figure 4 Distribution for the difference between the frequencies $f_{O|a}$, $f_{O|b}$ of onset insomnia (O) conditional on the two alleles of SNP rs875994

A Derivation of Bayes's theorem in hierarcic form

We write Bayes's theorem (9) with our initial belief (6) written in full:

$$\mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b}|\, D, I) \propto \int \mathrm{d}\alpha \int \mathrm{d}\boldsymbol{\nu}\; \mathrm{p}(D|\, f_{\mathrm{O}|a}, f_{\mathrm{O}|b}|\, I)\; \mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b}|\, \alpha, \boldsymbol{\nu}, I)\; \pi(\alpha, \boldsymbol{\nu}|\, I). \tag{18}$$

Multiplying and dividing within the integral with the expression

$$\mathrm{p}(D|\, \alpha, \boldsymbol{\nu}, I) \coloneqq \int \mathrm{d}f_{\mathrm{O}|a} \int \mathrm{d}f_{\mathrm{O}|b}\; \mathrm{p}(D|\, f_{\mathrm{O}|a}, f_{\mathrm{O}|b}, I)\; \mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b}|\, \alpha, \boldsymbol{\nu}, I) \tag{19}$$

we obtain the alternative form (10)

Combining together the sampling formula (5), the expression of the beta density (8), and the update formula (11) we obtain

$$\pi(\alpha, \boldsymbol{\nu}|\, D, I) \propto \pi(\alpha, \boldsymbol{\nu}) \times \left[\int \mathrm{d}f_{\mathrm{O}|a} \int \mathrm{d}f_{\mathrm{O}|b}\; \beta\Big(f_{\mathrm{O}|a}\big|\; \alpha + N_a, \tfrac{\alpha\boldsymbol{\nu} + N_a F_{\mathrm{O}|a}}{\alpha + N_a}\Big)\; \beta\Big(f_{\mathrm{O}|b}\big|\; \alpha + N_b, \tfrac{\alpha\boldsymbol{\nu} + N_b F_{\mathrm{O}|b}}{\alpha + N_b}\Big)\right] \tag{20}$$

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

$$\mathrm{p}(f_1|\, n_0, n_1, \theta_0, \theta_1)\, \mathrm{d}f_1 = \frac{\Gamma(n + \theta)}{\Gamma(n_0 + \theta_0)\, \Gamma(n_1 + \theta_1)}\, (1 - f_1)^{n_0 + \theta_0 - 1}\; {f_1}^{n_1 + \theta_1 - 1}\, \mathrm{d}f_1 \tag{21}$$

This distribution has expected value and variance

$$\begin{aligned} \mathrm{E}(f_1|\, n_0, n_1, \theta_0, \theta_1) &= \frac{n_1 + \theta_1}{n + \theta}, \\ \mathrm{var}(f_1|\, n_0, n_1, \theta_0, \theta_1) &= \frac{(n_0 + \theta_0)\, (n_1 + \theta_1)}{(n + \theta)^2\, (n + \theta + 1)}. \end{aligned} \tag{22}$$

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

$$
\begin{aligned}
\mathrm{p}(f|D) &\propto \mathrm{p}(D|f)\,\mathrm{p}(f|I) \\
&\propto \mathrm{p}(D|f) \int \mathrm{d}u\, \mathrm{p}(f|u)\,\mathrm{p}(u|I) \\
&\propto \int \mathrm{d}u\, \mathrm{p}(D|f)\,\mathrm{p}(f|u)\,\mathrm{p}(u|I) \\
&\propto \int \mathrm{d}u\, \mathrm{p}(D|f,u)\,\mathrm{p}(f|u)\,\mathrm{p}(D|u)\,\mathrm{p}(u|I) \\
&\propto \int \mathrm{d}u\, \mathrm{p}(f|D,u)\,\mathrm{p}(u|D)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{p}(u|D) &\propto \int \mathrm{d}f\, \mathrm{p}(D|f)\,\mathrm{p}(f|u)\,\mathrm{p}(u|I) \\
&\propto \prod_{x=a,b} \left[\frac{\Gamma(\sum_s u_s)}{\prod_s \Gamma(u_s)} \, \frac{\prod_s \Gamma(N_x F_{s|x} + u_s)}{\Gamma(N_x + \sum_s u_s)} \right]
\end{aligned}
$$

$$
\mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b}|D,I) \approx \mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b}|u_{\mathrm{M}}, v_{\mathrm{M}}, I)
$$

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.)

Cox, R. T. (1946): *Probability, frequency, and reasonable expectation*. Am. J. Phys. **14**[1], 1–13. http://algomagic.org/ProbabilityFrequencyReasonableExpectation.pdf.

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).

Hailperin, T. (1996): *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. (Associated University Presses, London).

Heidbreder, G. R., ed. (1996): *Maximum Entropy and Bayesian Methods. Santa Barbara, California, U.S.A., 1993*. (Springer, Dordrecht).

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.

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

Lockwood, J. R., Roeder, K., Devlin, B. (2001): *A Bayesian hierarchical model for allele frequencies*. Genet. Epidemiol. **20**[1], 17–33. http://www.stat.cmu.edu/~roeder/publications/lrd2001.pdf.

MacKay, D. J. C. (1996): *Hyperparameters: optimize, or integrate out?* In: Heidbreder (1996), 43–59. https://bayes.wustl.edu/MacKay/alpha.pdf.

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.