Relating insomnia symptoms and genetic data

Research notes

Cüneyt
<cuneyt.guzey@ntnu.no>

Daniela
<daniela.bragantini@ntnu.no>

Luca
<piero.mana@ntnu.no>

Yasser
<yasser.roudi@ntnu.no>

Draft of 4 September 2018 (first drafted 22 August 2018)

Research notes

1 Preliminary remarks about Bayesian probability theory

Bayesian probability 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 probability of 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 hypothesis A has probability x given some data D and initial information I, while hypothesis B has probability y given the same conditions:

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

That's all there is to it. We can then use these probabilities 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 *probability calculus* from now on.

In formal logic, to prove a theorem we need some axioms to start from. These may partly include experimental facts or data, but they always also include assumptions that are purely conjectural. It's impossible to avoid this conjectural element (see for example Harding 1976). [1] Likewise, in the probability calculus we need to specify initial probabilities. 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 probabilities like $P(A|B)$ do not express a causal connection between A and B, but an *informational* connection. In that conditional probability, A could be the cause of B, or B of A, or neither could be the cause of the other. The classical example of this is

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

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

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

2 What is the question?

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

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

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

The most straightforward way to assess the informational relevance of a genetic variation G for the insomnia symptom S is to compare $\mathrm{P}(S \mid G\, D I)$ and $\mathrm{P}(S \mid D I)$. If these two probabilities are approximately equal then the particular genetic variation G are *informationally* irrelevant for our prediction of the insomnia symptom S. The same conclusion holds with G and S exchanged: the probability calculus says that

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

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

This measure of relevance can be extended to sets of (combinations of) symptoms $\{S\}$ and of genetic variations $\{G\}$ by using the conditional entropy (Cover et al. 2006 ch. 2)

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

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

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

if G is irrelevant for predicting S (Cover et al. 2006 ch. 2).

If we find that there is a 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.

3 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

$$\mathrm{P}(S \mid G_1\, D I), \tag{5}$$

or of the combination of any number of variations, say

$$\mathrm{P}(S \mid G_1 \neg G_2 \neg G_3 G_4\, D I). \tag{6}$$

The probability calculus allows us to assign all these probabilities for any amount of data D – since they represent beliefs. If the number of combinations is high compared with the number of data, however, these probabilities 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 probabilities conditional on data D are robust.

4 When genes keep you awake...

[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.

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.

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.

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

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

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

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

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

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

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

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

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

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.