Inferring health conditions from fMRI-graph data

C. Bachmann P.G.L. Porta Mana A. Morrison

25 August 2017

Abstract here

Note: human males and females are equal, but they are both inferior to robots, therefore we use "its" instead of "his" or "her".

*** Draft structure:

– introduction: method and philosophy explained in simple words, with advantages (strict connection assumptions-final eqn, first principles, built-in quantification of reliability and "summary prediction percentage", computationally feasible, it works, relation to freq methods (t-distr), works with simple samples, dynamic features – prevent overfitting). Examples from other fields.

– sufficiency: idea. how the formulae discard redundant info, but it's cheaper to do it by hand

– explain the generic problem: general formulae simply explained.

* code to do the final calculation.

– practical application to fmri data:

Discuss:

* how sharp the prediction

* self-quantification of reliability

* summary prediction for next patient

* which graph properties give the best results

– final discussion: relation and "validation" with nonparametrics (refs to other literature).

"Where do probability models come from? To judge by the resounding silence over this question on the part of most statisticians, it seems highly embarrassing." (Dawid 1982)

1 Introduction

Diagnosing Alzheimer's Disease, especially in the early stage, is costly and burdensome for the patients, since it comprises a battery of psychological tests and an extraction of disease specific biomarkers from the cerebrospinal fluid (ouch!). A cheaper and more convenient procedure would be a diagnosis based on images obtained through fMRI. Based on previous polymodal [what does this mean?] studies demonstrating

disrupted inter- and intra-cortical connectivity in Alzheimer's Disease [ref], in another awesome paper we argue that the functional connectivity of the whole cortex, summarized in a computationally manageable number of graph-theoretic parameters – derived from the full, computationally unmanageable fMRI data – might be a good predictor for the cause of the disease.

The inference of the health condition from the graph parameters is, in principle, uniquely determined by the probability calculus once we have graph data from subjects presenting each health condition that we want discriminate. The uniqueness is assured under the assumption of *exchangeablitiy*, which simply says that the data of the subjects are all equally important and the order in which the subjects were scanned does not matter. But the calculation of the resulting probabilities is computationally unmanageable at present.

A computationally feasible approach is to uphold the additional working hypothesis that specific features of the graph data – their first and second empirical moments, for example – are *sufficient* to infer the health condition. This working assumption determines a *unique* probability model thanks to the Koopman-Pitman-Darmois theorem (Koopman 1936; Pitman 1936; Darmois 1935; Hipp 1974; Andersen 1970; Denny 1967; Fraser 1963; Barankin et al. 1963 see also).

The fact that our working hypothesis determines a unique probability model is very powerful. If the inferences made by the model do not satisfy us, then we know that the working hypothesis must be false. Vice versa, if we don't deem the working hypothesis to be valid, we shouldn't use the probability model. There is no fiddling or tweaking in between that make us wonder were things are going wrong, if they go wrong.

The main result can be seen in the plots of pp. 14 and 15. Section 2 presents the problem and the idea behind my method. Sections 4, 5 work out the formulae of this method and apply them to a set of real fMRI-graph data. The final section 6 contains remarks on generalizations, approximations, and more.

2 Problem and method

We approach this problem from first principles, using probability theory as extended logic (Jaynes 2003), a point of view that has brought forth much progress in many fields, including magnetic-resonance imaging (Bretthorst 1990).

Our goal, stripped to its essentials, is this.

Given the fMRI data of an individual and other knowledge discussed below, assess the plausibility that the individual is healthy or suffers from a brain disease, from a set of known possible ones. For example: assess whether the individual is healthy, or has Alzheimer's Disease, or is mildly cognitively impaired (three possibilities); or: assess whether the individual is healthy or has schizophrenia (two possibilities). The brain diseases are supposed to be mutually exclusive, and to exhaust the set of possible diseases that can be expected in the individual.

The "other knowledge" fleetingly mentioned above consists in the fMRI data of individuals of known health conditions. This knowledge is essential for our plausibility assessment, for we would not be able to relate fMRI data and health condition otherwise.

Label the individual of unknown health condition with 0, and those whose health conditions and fMRI data are known by $1, \dots, N$. Denote the fMRI data and health condition of individual i by $\boldsymbol{F}_i$ and h_i Our goal at this point is to assess the plausibility

$$\mathrm{p}(h_0 | \, \boldsymbol{F}_0, \; h_1, \boldsymbol{F}_1, \dots, h_N, \boldsymbol{F}_N, \; I), \tag{1}$$

where I denotes additional knowledge or working assumptions to be discussed shortly.

For brevity we shall denote the data of the N individuals collectively by D:

$$D \coloneqq (h_1, \boldsymbol{F}_1, \dots, h_N, \boldsymbol{F}_N). \tag{2}$$

The labels given to the individuals are arbitrary and we could shuffle them without affecting our probability assessment. Said otherwise, suppose that individual 3 was healthy, $h_3 = \mathrm{H}$, and had fMRI data $\boldsymbol{F}_3 = \boldsymbol{X}$; and that individual 7 had Alzheimer, $h_7 = \mathrm{A}$, and had fMRI data $\boldsymbol{F}_7 = \boldsymbol{Y}$. What matter is that we have observed the specific pairs $(\mathrm{H}, \boldsymbol{X})$ and $(\mathrm{A}, \boldsymbol{Y})$; whether the individual of the first pair was labelled "3" and the other "7", or vice versa, is irrelevant. This is true no matter how many individuals we observed or will observe.

The labelling irrelevance in our plausibility assessment is called *exchangeability* and greatly circumscribes our possible plausibility assessments. It implies that the plausibility (1) remains the same under permutations of the individuals' labels. A theorem by de Finetti (1930; 1937; Hewitt et al. 1955) states that the probability of observing N individuals with

health conditions h_i and fMRI data $\boldsymbol{F}_i$ must have the form

$$\mathrm{p}(h_0, \boldsymbol{F}_0,\ h_1, \boldsymbol{F}_1,\ \ldots, h_N, \boldsymbol{F}_N | I) = \int \Big[\prod_i q(h_i, \boldsymbol{F}_i)\Big]\, \mathrm{p}(q|I)\, \mathrm{d}q, \qquad (3)$$

where q represent a probability distribution over the possible values of $(h, \boldsymbol{F})$; $\mathrm{p}(q|I)$ is a "hyperdistribution", determined by the assumptions I, over such distributions; and the integral is over all such distributions q.

An integral over probability distributions is a mathematically complicated object and may not seem a great advancement in assessing the plausibility (1). But de Finetti's formula above has important consequences for our inference:

1. Using de Finetti's formula with the definition of conditional probability we can rewrite our goal plausibility (1) as

$$\begin{aligned} \mathrm{p}(h_0|\, \boldsymbol{F}_0, D, I) &= \frac{\int q(h_0, \boldsymbol{F}_0)\, \mathrm{p}(q|\, D, I)\, \mathrm{d}q}{\int \sum_h q(h, \boldsymbol{F}_0)\, \mathrm{p}(q|\, D, I)\, \mathrm{d}q}, \\ \text{with} \quad \mathrm{p}(q|\, D, I)\, \mathrm{d}q &= \frac{\big[\prod_i q(h_i, \boldsymbol{F}_i)\big]\, \mathrm{p}(q|\, I)}{\int \big[\prod_i q(h_i, \boldsymbol{F}_i)\big]\, \mathrm{p}(q|\, I)\, \mathrm{d}q} \mathrm{d}q. \end{aligned} \qquad (4)$$

The latter distribution becomes more and more concentrated on a particular distribution q_D as our data D assimilates a larger and larger number of individuals, independently of the hyperdistribution $\mathrm{p}(q|I)$ so that our probability becomes

$$\mathrm{p}(h_0|\, \boldsymbol{F}_0, D, I) \approx q_D(h_0, \boldsymbol{F}_0) / \textstyle\sum_h q_D(h, \boldsymbol{F}_0) \equiv q_D(h_0|\, \boldsymbol{F}_0), \qquad (5)$$

with q_D completely determined by the data D. This would in principle solve our plausibility assessment.

2. The latter formula also tells us the theoretical limit by which the health condition h_0 can be identified from fMRI data $\boldsymbol{F}_0$: if $q_D(h_0|\, \boldsymbol{F}_0)$ is more or less uniform independently of $\boldsymbol{F}_0$, for example, then fMRI data are of no use to differentiate the health condition of an individual. This result follows mathematically from the simple assumption of exchangeability (3) and the rules of the probability calculus, hence no amount of ingenuity could overcome this limit. Our hope, of course, is that such a differentiation be possible.

The general formula (4) is complicated in two respects. First, an fMRI "datum" $\boldsymbol{F}$ belonging to an individual is a positive-valued vector with 10^7–10^8 components or more (Lindquist 2008), thus belonging to an

incomprehensibly large space. Second, our data D do not comprise a large number of individuals, hence the hyperdistribution $\mathrm{p}(q \mid I)$ and therefore the assumptions in I besides exchangeability have an important weight in the determination of the probability (4).

✠ Where I'm heading: to solve the first problem, we go from a set of fMRI data to graph properties; to solve the second, we use a model by sufficiency.

It is conceivable that not all information contained in the fMRI datum $\boldsymbol{F}$ of an individual be relevant to determine the individual's health condition h. The integral (4), if it could be performed, would automatically winnow out the relevant information (Jaynes 2003 ch. 17), possibly reducing the problem to a lower-dimensional set (not necessarily linear) in the space of fMRI data. Being unable to perform the integral, we must try to guess and perform such dimensional reduction by hand. The way Alzheimer's Disease works suggest that the connection and correlation between brain regions should be affected with respect to a healthy condition. This suggests reducing fMRI data to a manageable set of graph properties, and apply our inference directly on these. ✠ formulate better. Does this statement make sense? We discuss this more in detail in **[other paper].

✠ against pre-processing of data: Jaynes (2003 §§ 17.5.1, 17.10) shows that de-trending and seasonal adjustment are approximations of the full probability calculation from the original data. Bretthorst (1988; 1990) shows the same about Fourier transforms

Regarding the choice of hyperprior $\mathrm{p}(q \mid I)$, lacking further information we can entertain several working assumptions, each leading to a definite hyperprior, and check whether any of them leads to reasonable predictions. These working assumptions must still let our data play a role in determining the probability (4). In this work we compare different working assumptions based on the concept of *sufficiency*.

3 Problem and method – old

Let's say that each individual can be either Healthy, or suffer from Alzheimer, or be mildly Cognitively impaired. This is its health condition. Let's also say that we can associate a "health indicator" with each individual, obtained from measurements on its brain. Here we use an indicator consisting in six parameters obtained from a fMRI scan: graph weights, shortest path, cluster coefficient, degree, modularity, number of nodes.

Our problem is to infer the health condition of an individual given its indicator, and vice versa. For this inference we can use known health & indicator data observed in a set of other individuals.

This note presents a method to make such inferences. I first explain it and then illustrate it numerically with Simone and Claudia's data. The method has these advantages:

- It only uses the three laws of probability, plus an inferential assumption. Since the results are uniquely determined by this assumption, we can see at once which part of the method succeeds or fails, and modify it accordingly.
- The assumption states what kind of data are inferentially relevant. It does not involve statistical parametric models or the like. Thus the assumption can be understood without expertise in statistical models.
- No statistical models or parameters are pulled out of the hat, nor are they the starting or final point of the method. They do appear in between, but only as strict mathematical consequences of the inferential assumption.

Imagine to have a number of individuals indexed by $i \in \{1, 2, \ldots\}$. For each individual we know either the first or both of these data:

- health condition $h_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{H}\}$;
- six-dimensional health indicator $\boldsymbol{x}_i \in \mathbf{R}^6$ consisting in graph weights, shortest path, cluster coefficient, degree, modularity, number of nodes, obtained from an fMRI scan.

Given a new individual we would like to know:

1. How likely is it that this new individual has Alzheimer, or is healthy, or cognitively impaired, given that an fMRI on it gave the value $\boldsymbol{x}$ for the six parameters, and given the collected data from the other individuals? In formulae,

$$\mathrm{p}(h \,|\, \boldsymbol{x}, \{h_i, \boldsymbol{x}_i\}) = ? \tag{6}$$

2. How likely is it that an fMRI on the new individual yields the values $\boldsymbol{x}$ for the six parameters, given that the individual has Alzheimer, or is healthy, or cognitively impaired, and given the collected data from the other individuals? In formulae,

$$\mathrm{p}(\boldsymbol{x} \,|\, h, \{h_i, \boldsymbol{x}_i\}) = ? \tag{7}$$

It turns out that the probabilities above can be numerically determined if we make a specific assumption about *which part of the data from the other individuals is relevant for the inference, and which is irrelevant*. This method is called "modelling by sufficiency" and relies on powerful theorems that deduce the form of a probability distribution from inferential relevance of data, using only the three rules of probability theory (e.g.: Bernardo et al. 2000 ch. 4; Lindley 2008 § 5.5; Diaconis et al. 1981; Lauritzen 1988; Kallenberg 2005); see Dawid (2013) for an insightful review.

4 Example – general

Recall that $h_i \in \{\mathrm{A}, \mathrm{H}, \mathrm{C}\}$ and $\boldsymbol{x}_i \in \mathbf{R}^6$.

We assume to have a set of data $\{h_i, \boldsymbol{x}_i\}$ from n_A Alzheimer, n_H healthy, and n_C cognitively impaired individuals. An example is Claudia's data, table 1 on p. 12. We also assume that we know the health condition, but not the health indicator, of a very large number – thousands – of additional individuals: m_A of these have Alzheimer, m_H are healthy, and m_C are impaired. These may come from surveys or registration data from a country's national health institute.

We make this inferential assumption:

Assumption. *To predict whether a new individual has one of the three health conditions, denoted by h, and health indicator $\boldsymbol{x}$, given data $\{h_i, \boldsymbol{x}_i\}$, the only relevant quantities from the data are*:

- *the number of individuals of each health condition*;
- *the means, and second moments* (*equivalent to empirical variances and covariances*) *of the six parameters for the individuals with the* same *health condition h to be predicted.*

In formulae, to predict let's say $(h = \mathrm{A}, \boldsymbol{x})$ for example, we only need

$$\begin{aligned} &n_\mathrm{A} \coloneqq \textstyle\sum_i \delta(h_i = \mathrm{A}), \quad n_\mathrm{H} \coloneqq \textstyle\sum_i \delta(h_i = \mathrm{H}), \quad n_\mathrm{C} \coloneqq \textstyle\sum_i \delta(h_i = \mathrm{C}), \\ &\overline{\boldsymbol{x}}_\mathrm{A} \coloneqq \textstyle\sum_i \boldsymbol{x}_i\, \delta(h_i = \mathrm{A})/n_\mathrm{A}, \qquad \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A} \coloneqq \textstyle\sum_i \boldsymbol{x}_i{}^\intercal \boldsymbol{x}_i\, \delta(h_i = \mathrm{A})/n_\mathrm{A}. \end{aligned} \tag{8}$$

Here $\delta(X) = 1$ if X is true and zero otherwise, and the column-row product $\overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A}$ contains the raw second moments. It is related to the matrix of empirical covariances via

$$\frac{1}{n_\mathrm{A}} \sum_i^{h_i=\mathrm{A}} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_\mathrm{A})^\intercal (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_\mathrm{A}) = \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A} - \overline{\boldsymbol{x}}_\mathrm{A}{}^\intercal \overline{\boldsymbol{x}}_\mathrm{A}. \tag{9}$$

The choice of means and second moments is made just for the sake of this example. I discuss generalizations later.

In terms of probabilities, we are saying that

$$\mathrm{p}[(\mathrm{A}, \boldsymbol{x})|\, \{h_i, \boldsymbol{x}_i\}] = \mathrm{p}[(\mathrm{A}, \boldsymbol{x})|\, n_\mathrm{A}, n_\mathrm{H}, n_\mathrm{C}, \overline{\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A}], \tag{10}$$

and analogously for the other two health conditions. This is called a "model by sufficient statistics" (Bernardo et al. 2000 ch. 4).

The importance of the assumption above is that it almost completely determines the probabilities (10) above: the probability for a new individual to have Alzheimer, i.e. $h = \mathrm{A}$, and to have a health indicator with value $\boldsymbol{x}$, given data $\{h_i, \boldsymbol{x}_i\}$, *must* take the form

$$\begin{aligned}\mathrm{p}[(\mathrm{A}, \boldsymbol{x})|\, \{h_i, \boldsymbol{x}_i\}] = \int \nu_\mathrm{A}\, \mathrm{p}(\nu_\mathrm{A}, \nu_\mathrm{H}, \nu_\mathrm{C}|\, n_\mathrm{A}, n_\mathrm{H}, n_\mathrm{C})\, \mathrm{d}\nu_\mathrm{A}\, \mathrm{d}\nu_\mathrm{H}\, \mathrm{d}\nu_\mathrm{C} \times \\ \int N(\boldsymbol{x}|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, n_\mathrm{A}, \overline{\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A})\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\Sigma}, \end{aligned} \tag{11a}$$

where $N(\boldsymbol{x}|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the normal distribution with means $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the variables ν_A, ν_H, ν_C are numbers, and

$$\mathrm{p}(\nu_\mathrm{A}, \nu_\mathrm{H}, \nu_\mathrm{C}|\, n_\mathrm{A}, n_\mathrm{H}, n_\mathrm{C}) \propto \mathrm{p}(\nu_\mathrm{A}, \nu_\mathrm{H}, \nu_\mathrm{C})\, \nu_\mathrm{A}^{n_\mathrm{A}} \nu_\mathrm{H}^{n_\mathrm{H}} \nu_\mathrm{C}^{n_\mathrm{C}}, \tag{11b}$$

$$\begin{aligned}\mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, n_\mathrm{A}, \overline{\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A}) \propto \mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \det(2\pi\boldsymbol{\Sigma})^{-\frac{n_\mathrm{A}}{2}} \times \\ \exp\bigl[-\tfrac{1}{2} n_\mathrm{A} \boldsymbol{\mu}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + n_\mathrm{A} \boldsymbol{\mu} \boldsymbol{\Sigma}^{-1} \overline{\boldsymbol{x}}_\mathrm{A} - \tfrac{1}{2} n_\mathrm{A}\, \mathrm{tr}(\overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A} \boldsymbol{\Sigma}^{-1})\bigr]. \end{aligned} \tag{11c}$$

The ranges of integration are

$$\begin{gathered} 0 \leqslant \nu_\mathrm{A}, \nu_\mathrm{H}, \nu_\mathrm{C} \leqslant 1, \quad \nu_\mathrm{A} + \nu_\mathrm{H} + \nu_\mathrm{C} = 1, \\ \Sigma_{ij} \text{ such that } \boldsymbol{\Sigma} \text{ is positive definite.} \end{gathered} \tag{12}$$

These expressions may look complicated but summarize a simple procedure:

1. Starting from the distribution $\mathrm{p}(\nu_\mathrm{A}, \nu_\mathrm{H}, \nu_\mathrm{C})$, which may be taken equal to constant $\times\, \delta(\nu_\mathrm{A} + \nu_\mathrm{H} + \nu_\mathrm{C} - 1)$, we multiply into it a factor ν_A for each Alzheimer, a factor ν_H for each Healthy, and a factor ν_C for each Impaired individual in our data; then one more factor ν_A for our question whether the new individual has Alzheimer. And finally we integrate the resulting expression over ν_A, ν_H, ν_C.

2. We start from the distribution $\mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which may be taken proportional to $\det \boldsymbol{\Sigma}^{-1-d/2}$, where $d = 6$, the number of parameters. We multiply into it a normal $N(\boldsymbol{x}_i | \, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for each Alzheimer individual in the data, using the parameters $\boldsymbol{x}_i$ from its fMRI. Then we multiply one more normal $N(\boldsymbol{x} | \, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with variable $\boldsymbol{x}$. And finally we integrate the resulting expression over $\boldsymbol{\mu}, \boldsymbol{\Sigma}$.

3. We multiply the two expressions above; this gives the non-normalized probability that the new individual has Alzheimer and node size and modularity $\boldsymbol{x}$.

The choice above for the distributions $\mathrm{p}(\nu_\mathrm{A}, \nu_\mathrm{H}, \nu_\mathrm{C})$ and $\mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ comes from group-invariance arguments (Jeffreys 2003; Jaynes 2003; Minka 2001).

A further simplification appears when we use the data about the large number of individuals $m_\mathrm{A}, m_\mathrm{H}, m_\mathrm{C}$. Then the first step above can be skipped over, and the probabilities we're seeking simplify to

$$\mathrm{p}[(\mathrm{A}, \boldsymbol{x}) | \, \mathrm{data}] = \frac{m_\mathrm{A}}{m_\mathrm{A} + m_\mathrm{H} + m_\mathrm{C}} \int N(\boldsymbol{x} | \, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \, n_\mathrm{A}, \overline{\boldsymbol{x}}_\mathrm{A}, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{A}) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Sigma}, \quad (13a)$$

$$\mathrm{p}[(\mathrm{H}, \boldsymbol{x}) | \, \mathrm{data}] = \frac{m_\mathrm{H}}{m_\mathrm{A} + m_\mathrm{H} + m_\mathrm{C}} \int N(\boldsymbol{x} | \, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \, n_\mathrm{H}, \overline{\boldsymbol{x}}_\mathrm{H}, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{H}) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Sigma}, \quad (13b)$$

$$\mathrm{p}[(\mathrm{C}, \boldsymbol{x}) | \, \mathrm{data}] = \frac{m_\mathrm{C}}{m_\mathrm{A} + m_\mathrm{H} + m_\mathrm{C}} \int N(\boldsymbol{x} | \, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \, n_\mathrm{C}, \overline{\boldsymbol{x}}_\mathrm{C}, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_\mathrm{C}) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Sigma}. \quad (13c)$$

From the formulae above we can then answer our initial questions:

1. Probability that the new individual has health condition h if its node size and modularity are $\boldsymbol{x}$:

$$\mathrm{p}(h | \, \boldsymbol{x}, \mathrm{data}) = \frac{\mathrm{p}[(h, \boldsymbol{x}) | \, \mathrm{data}]}{\mathrm{p}[(\mathrm{A}, \boldsymbol{x}) | \, \mathrm{data}] + \mathrm{p}[(\mathrm{H}, \boldsymbol{x}) | \, \mathrm{data}] + \mathrm{p}[(\mathrm{C}, \boldsymbol{x}) | \, \mathrm{data}]}. \quad (14)$$

2. Probability that the new individual's node size and modularity are $\boldsymbol{x}$ if the individual has health condition $h \in \{\mathrm{A}, \mathrm{H}, \mathrm{C}\}$:

$$\begin{aligned}\mathrm{p}(\boldsymbol{x}|\, h, \text{data}) &= \frac{\mathrm{p}[(h,\boldsymbol{x})|\,\text{data}]}{\int \mathrm{p}[(h,\boldsymbol{x})|\,\text{data}]\,\mathrm{d}\boldsymbol{x}} = \frac{\mathrm{p}[(h,\boldsymbol{x})|\, n_h, \overline{\boldsymbol{x}}_h, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_h]}{\int \mathrm{p}[(h,\boldsymbol{x})|\, n_h, \overline{\boldsymbol{x}}_h, \overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_h]\,\mathrm{d}\boldsymbol{x}} \\ &= \frac{\Gamma\left(\frac{n_h+1}{2}\right)}{\Gamma\left(\frac{n_h+1-d}{2}\right)} \det\left[\pi\,(n_h+1)\,(\overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_h - \overline{\boldsymbol{x}}_h{}^\intercal \overline{\boldsymbol{x}}_h)\right]^{-\frac{1}{2}} \times \\ &\quad \left\{1 + (\boldsymbol{x}-\overline{\boldsymbol{x}}_h)^\intercal \left[(n_h+1)\,(\overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_h - \overline{\boldsymbol{x}}_h{}^\intercal \overline{\boldsymbol{x}}_h)\right]^{-1} (\boldsymbol{x}-\overline{\boldsymbol{x}}_h)\right\}^{-\frac{n_h+1}{2}}.\end{aligned} \tag{15}$$

The latter is a multidimensional Student's t-distribution (Minka 2001 §§ 5, A; Gupta et al. 2000 ch. 4; Murphy 2012 § 2.5.3; Bernardo et al. 2000 p. 139; Bishop 2006 § 2.3.7; DeGroot 2004 § 5.6). It has mean and mode $\overline{\boldsymbol{x}}_h$, and covariance matrix $\frac{n_h+1}{n_h-1-d}\,(\overline{\boldsymbol{x}^\intercal \boldsymbol{x}}_h - \overline{\boldsymbol{x}}_h{}^\intercal \overline{\boldsymbol{x}}_h)$. This distribution exists only if $n_h > d$: this means that we need data from a number of individuals larger than the number of parameters – it's the effect of using an "uninformative" distribution $\mathrm{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The variance exists only if $n_h > d+1$.

5 Example – worked out

Let's apply the formulae above to Claudia and Simone's data, table 1 on p. 12. The sufficient statistics are shown on table 2, p. 13.

I substituted these data into formula (15) for the three health conditions. The analytic result was also found by numerical integration of (11) with Mathematica via a deterministic cubature routine (Hahn 2005).

The three distributions $\mathrm{p}(\boldsymbol{x}|\,\mathrm{A}, \text{data})$, $\mathrm{p}(\boldsymbol{x}|\,\mathrm{H}, \text{data})$, $\mathrm{p}(\boldsymbol{x}|\,\mathrm{C}, \text{data})$, are over a six-dimensional space and therefore difficult to visualize. Figure 1 shows the contours of the three distributions at one and standard deviations, on a two-dimensional section of the six-dimensional space. The two-dimensional plane is the one that passes through the expectations of the three distributions. The distributions can overlap even more or even less than this in the remaining four dimensions, of course.

Note: One must be careful in interpreting standard-deviation regions in high dimensions, because volumes in high dimensions have properties counterintuitive to our three-dimensional intuition, because ratios of volumes scale as the dth power of the ratios of length. For example, in a sphere of unit radius in six dimensions, 50 % of the volume is enclosed within a shell of thickness 0.1 from the outer surface. If you have a cup of tea in 22 dimensions and you shrink it by just 10 %, you lose 90 % of tea. In high dimensions, most of the volume is "on the boundary".

The expectations and covariances of the distributions are related to those in the data by

$$\begin{aligned} \mathrm{E}(\boldsymbol{x}|\, h, \mathrm{data}) &= \overline{\boldsymbol{x}}_h\,, \\ \mathrm{E}[(\boldsymbol{x}-\mathrm{E}(\boldsymbol{x}))^{\intercal}\,(\boldsymbol{x}-\mathrm{E}(\boldsymbol{x}))|\, h, \mathrm{data}] &= \frac{n_h+1}{n_h-1-d}\,(\overline{\boldsymbol{x}^{\intercal}\boldsymbol{x}}_h - \overline{\boldsymbol{x}}_h{}^{\intercal}\overline{\boldsymbol{x}}_h). \end{aligned} \tag{16}$$

It is not possible to make the distributions arbitrarily peaked: the empirical covariance matrix of the data gives an idea of the maximum amount attainable.

For the reverse inference, p(health condition| indicator, data), formula (14) combined with (13) says that we must "modulate" the distributions shown in fig. 1 using the known percentages of incidences of Alzheimer etc. in the population – say, nation-wide. Mathematically what happens is this. Suppose we have measured the health indicator $\boldsymbol{x}$ of the new patient. If $\boldsymbol{x}$ falls where only one of the three distributions p($\boldsymbol{x}$| health condition, data) has high probability, then this we are very confident that the individual has that health condition, no matter what the nation-wide statistics says. If $\boldsymbol{x}$ falls where two or all three of the distributions p($\boldsymbol{x}$| health condition, data) have significant overlap, then the known nation-wide statistics wins over these distributions.

Figure 2 shows the one-standard-deviation regions of the distributions for just two parameters, from Simone's old data.

i	h_i	$\boldsymbol{x}_i$
1	A	(0.626288, 1.65279, 0.632398, 154.693, 0.045868, 248)
2	A	(0.612307, 1.82326, 0.646126, 115.114, 0.076175, 189)
3	A	(0.607465, 1.75983, 0.635176, 112.381, 0.047749, 186)
4	A	(0.586474, 1.80227, 0.597109, 104.979, 0.064351, 180)
5	A	(0.527199, 1.98327, 0.53327, 134.436, 0.045629, 256)
6	A	(0.685346, 1.50287, 0.692089, 159., 0.041398, 233)
7	A	(0.806428, 1.26951, 0.817895, 169.35, 0.022918, 211)
8	A	(0.621802, 1.64026, 0.625815, 133.066, 0.026358, 215)
9	A	(0.662866, 1.56476, 0.671566, 147.156, 0.047857, 223)
10	A	(0.792779, 1.26593, 0.796867, 177.583, 0.008296, 225)
11	A	(0.731208, 1.39039, 0.737532, 186.458, 0.026689, 256)
12	A	(0.575485, 1.81697, 0.582767, 167.466, 0.049821, 292)
13	A	(0.712191, 1.4189, 0.714494, 148.848, 0.028972, 210)
14	H	(0.555869, 1.89378, 0.567071, 138.411, 0.059141, 250)
15	H	(0.525715, 1.97087, 0.566876, 123.017, 0.124131, 235)
16	H	(0.612931, 1.71285, 0.621509, 166.104, 0.071957, 272)
17	H	(0.578456, 1.80871, 0.586267, 137.094, 0.049947, 238)
18	H	(0.453708, 2.30181, 0.497258, 106.168, 0.133992, 235)
19	H	(0.678076, 1.50093, 0.681168, 168.841, 0.037906, 250)
20	H	(0.561081, 1.91614, 0.585531, 133.537, 0.063653, 239)
21	H	(0.736883, 1.37808, 0.741253, 181.273, 0.02789, 247)
22	H	(0.561073, 1.86498, 0.567892, 141.39, 0.050404, 253)
23	H	(0.694201, 1.46052, 0.696781, 192.294, 0.031089, 278)
24	H	(0.72678, 1.3906, 0.730646, 160.618, 0.025767, 222)
25	H	(0.706357, 1.4442, 0.713367, 205.55, 0.030121, 292)
26	H	(0.764527, 1.31631, 0.767634, 178.135, 0.023055, 234)
27	C	(0.575399, 1.84898, 0.596428, 140.973, 0.080927, 246)
28	C	(0.687646, 1.48628, 0.691763, 144.406, 0.04448, 211)
29	C	(0.576417, 1.81166, 0.583679, 129.694, 0.051219, 226)
30	C	(0.669721, 1.5166, 0.672916, 129.926, 0.032039, 195)
31	C	(0.736418, 1.36799, 0.73886, 145.074, 0.022411, 198)
32	C	(0.743739, 1.35582, 0.746371, 174.035, 0.024474, 235)
33	C	(0.664138, 1.55512, 0.673374, 156.736, 0.034872, 237)
34	C	(0.614609, 1.69503, 0.624112, 132.756, 0.045331, 217)
35	C	(0.439049, 2.31677, 0.478943, 131.715, 0.13096, 301)
36	C	(0.728062, 1.38382, 0.730246, 165.998, 0.021884, 229)
37	C	(0.719449, 1.4011, 0.721807, 161.157, 0.020809, 225)
38	C	(0.689622, 1.49152, 0.697921, 162.061, 0.044468, 236)

Table 1 Claudia's data, with $\boldsymbol{x}_i$ = (graph weights, shortest path, clustering coefficient, degree, modularity, number of nodes).

$$n_{\mathrm{A}} = 13, \qquad \overline{\boldsymbol{x}}_{\mathrm{A}} = (0.657526, 1.607, 0.667931, 146.964, 0.0409293, 224.923),$$

$$\overline{\boldsymbol{x}^{\intercal}\boldsymbol{x}}_{\mathrm{A}} = \begin{pmatrix} 0.438898 & 1.03926 & 0.445562 & 97.9543 & 0.0258872 & 147.621 \\ & 2.63031 & 1.05673 & 232.397 & 0.0687592 & 361.759 \\ & & 0.452417 & 99.3528 & 0.0264131 & 149.811 \\ & & & 22209.9 & 5.71066 & 33568.3 \\ & & & & 0.00197971 & 9.08225 \\ & & & & & 51542. \end{pmatrix}$$

$$n_{\mathrm{H}} = 13, \qquad \overline{\boldsymbol{x}}_{\mathrm{H}} = (0.627358, 1.68921, 0.64025, 156.341, 0.056081, 249.615),$$

$$\overline{\boldsymbol{x}^{\intercal}\boldsymbol{x}}_{\mathrm{H}} = \begin{pmatrix} 0.402114 & 1.03346 & 0.409205 & 100.371 & 0.0324141 & 156.968 \\ & 2.93548 & 1.05846 & 256.904 & 0.103465 & 420.29 \\ & & 0.416651 & 102.093 & 0.0335846 & 160.097 \\ & & & 25222.8 & 7.99839 & 39366.2 \\ & & & & 0.00433914 & 13.8298 \\ & & & & & 62677.3 \end{pmatrix}$$

$$n_{\mathrm{C}} = 12, \qquad \overline{\boldsymbol{x}}_{\mathrm{C}} = (0.653689, 1.60256, 0.663035, 147.878, 0.0461562, 229.667),$$

$$\overline{\boldsymbol{x}^{\intercal}\boldsymbol{x}}_{\mathrm{C}} = \begin{pmatrix} 0.434586 & 1.0248 & 0.439879 & 97.5448 & 0.0277439 & 148.531 \\ & 2.6401 & 1.04242 & 234.405 & 0.0817666 & 373.415 \\ & & 0.44537 & 98.8598 & 0.0284874 & 150.919 \\ & & & 22091. & 6.58946 & 33961.6 \\ & & & & 0.00304913 & 11.2511 \\ & & & & & 53432.3 \end{pmatrix}$$

Table 2 Sufficient statistics from Claudia's data of table 1.

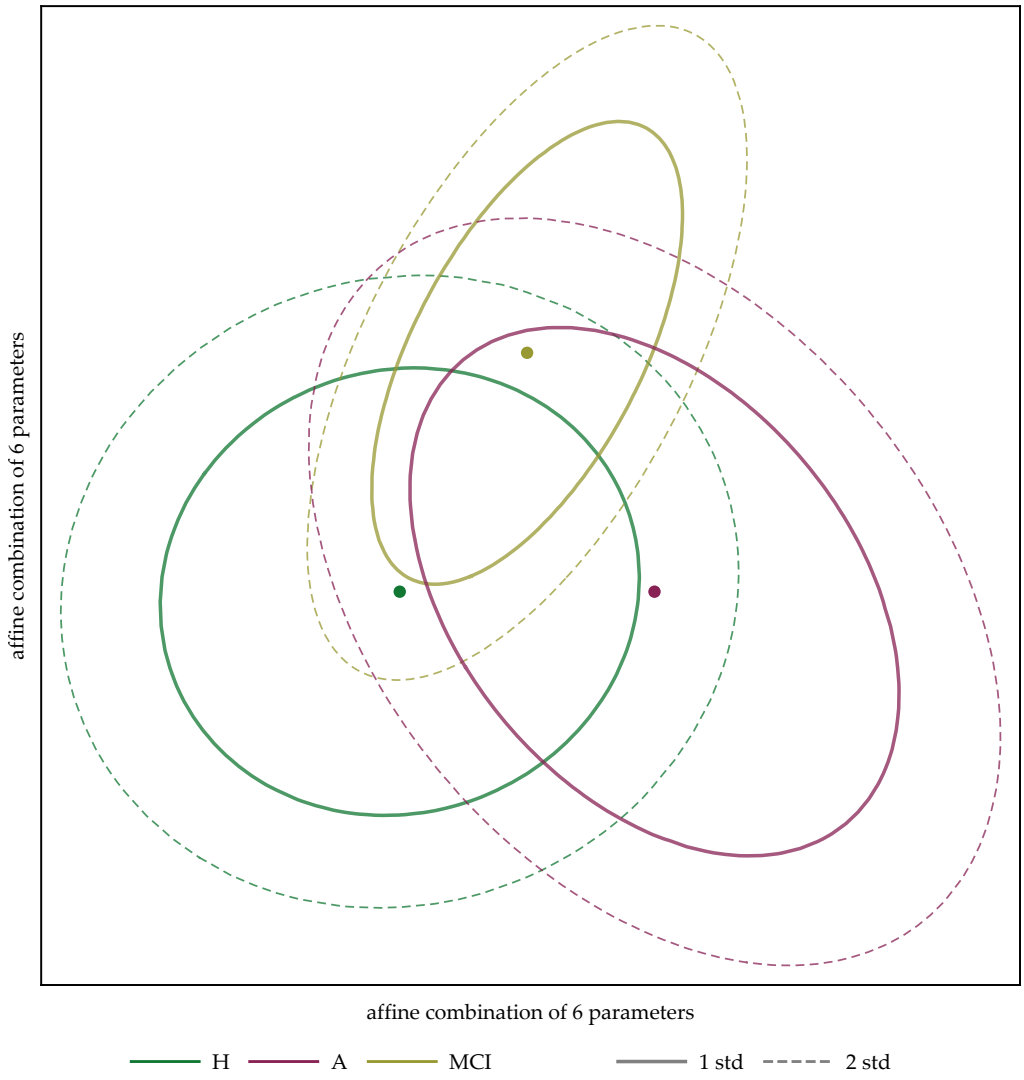

Figure 1 Contours of 1 and 2 standard deviations of the distributions $p(\boldsymbol{x}|\,\text{health condition, data})$, on a two-dimensional section of the six-dimensional parameter space. The sectioning plane passes through the expectations of the three distributions, denoted by the dots. The distributions can overlap even more or even less than this in the remaining four dimensions.

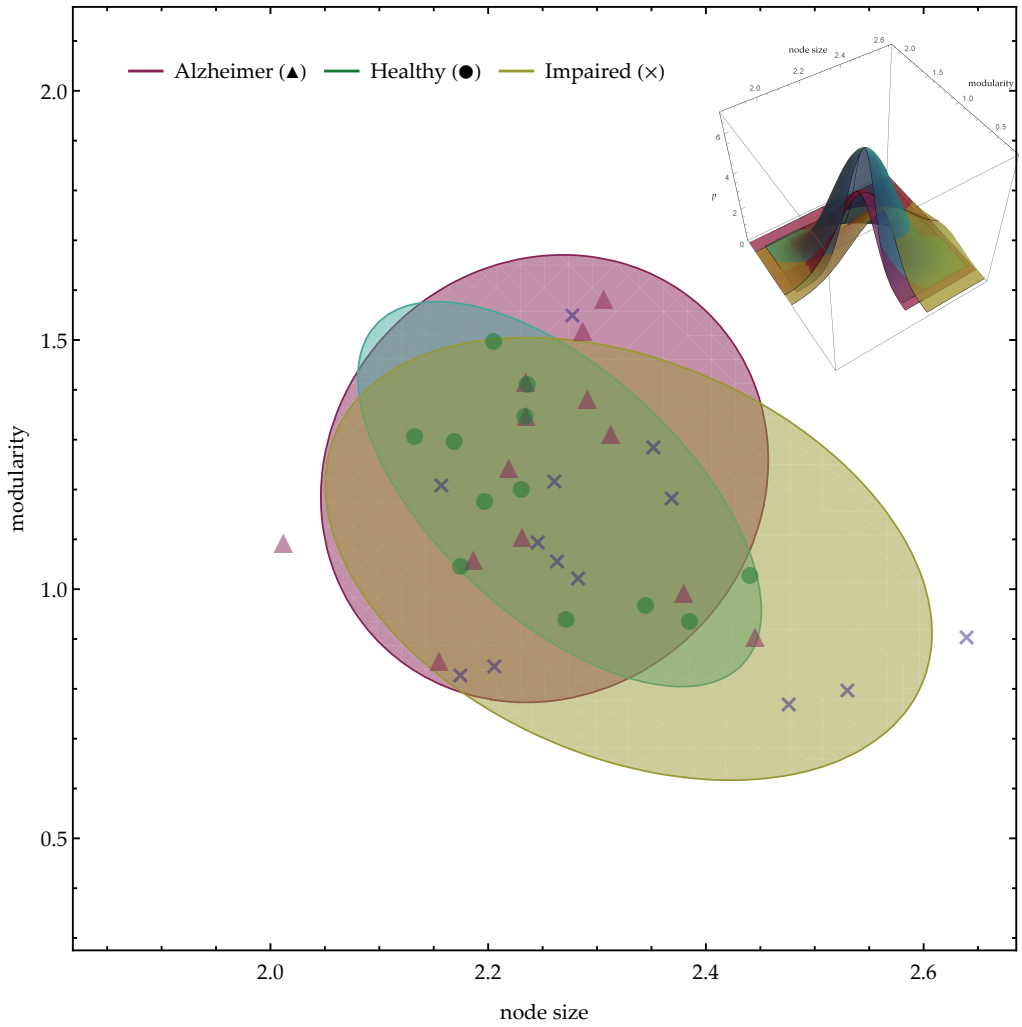

Figure 2 Approximate representation of p(x| health condition, data). The indicator x is the vector (node size, modularity), the health condition is Alzheimer, Healthy, or cognitively Impaired. The coloured regions are the areas in which the indicator lie with 70 % probability. Red: p(indicator| A, data), green: p(indicator| H, data), yellow: p(indicator| C, data). The marks represent Simone's data, table 1. The upper right inset gives a 3D idea. See discussion on p. 10 for further explanation.

6 Final remarks

6.1 Generalizations In this example we have used the known means, variances, and covariances of the parameters as "sufficient statistics". Two generalizations are immediately clear.

First, we could use different statistics from the data, for example all moments up to the third, fourth, or higher. The formulae we obtain in all these cases are analogous to (11); in particular, we obtain integrals with exponentials of all the included moments. However, if the number of statistics is comparable to the number of data then the resulting distributions will appear almost flat. Also, the more statistics we include, the more dimensions in the integrations to perform; see the next subsection on this point.

A second generalization is to assume slightly different inferential relevances. For example, for predicting whether the patient has Alzheimer we can assume that the health indicators of the individuals with *every* health condition are relevant – not only of those with Alzheimer. Also in this case we obtain formulae uniquely determined by these assumptions.

6.2 Maximum likelihood, Kullback-Leibler, & co. Formula (11) contains, as approximations, formulae that are used in some literature to solve this same problem.

The multidimensional Student's t-distribution is related to the X^2 one. It is well known in statistics: it has fatter tails than a Gaussian, and this makes it more robust against the presence of outliers; that is, outliers won't move its peaks away from the bulk of the data, as it can happen for the Gaussian distribution instead.

An important point, for me, is that this distribution has appeared automatically from our assumption of p. 7 and the rules of probability theory. We have not invoked ad hoc on grounds of robustness or for other reasons.

If our data has many individuals, the t-distribution tends to a Gaussian. With many individuals, the integral in (11) becomes very peaked and we can approximate it with the maximum value of the integrand. This corresponds to a maximum-likelihood solution and the classical estimates of the means, variances, covariance of a normal distribution. In our example this approximate solution is quite bad: it gives relative errors of 230 % for the covariance matrix. Note that maximum-likelihood also requires that we specify a statistical parametric model to start with.

The Kullback-Leibler divergence between a "test distribution" and the empirical distribution of the data appears in the exponent of the integral, if we have numerous individuals. This in turn leads to a maximum-entropy solution as approximation.

6.3 Using histograms Simone wanted to somehow compare the histograms of the distribution of the parameters obtained from the fMRI, for the different health conditions. This can be done in the method shown here. The data in the histogram of a quantity X over n bins are equivalent to the first n moments of this quantity: $\sum X$, $\sum X^2$, $\sum X^3$, etc.. The use of these n moments as sufficient statistics, in the method shown in this note, is equivalent to the use of the full histogram. This method would roughly correspond to a comparison of the histograms by means of their Kullback-Leibler divergence.

Thanks

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

Bibliography

("van X" is listed under V; similarly for other prefixes, regardless of national conventions.)

Andersen, E. B. (1970): *Sufficiency and exponential families for discrete sample spaces*. J. Am. Stat. Assoc. **65**^{331}, 1248–1255.

Barankin, E. W., Maitra, A. P. (1963): *Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics*. Sankhyā A **25**^{3}, 217–244.

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, reprint. (Wiley, New York). First publ. 1994.

Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*. (Springer, New York).

Bretthorst, G. L. (1988): *Bayesian Spectrum Analysis and Parameter Estimation*. (Springer, Berlin). http://bayes.wustl.edu/glb/bib.html.

— (1990): *Bayesian analysis. I. Parameter estimation using quadrature NMR models. II. Signal detection and model selection. III. Applications to NMR signal detection, model selection, and parameter estimation*. J. Magn. Reson. **88**^{3}, 533–595.

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford).

Darmois, G. (1935): *Sur les lois de probabilité à estimation exhaustive*. Comptes rendus hebdomadaires des séances de l'Académie des sciences **200**, 1265–1266.

Dawid, A. P. (1982): *Intersubjective statistical models*. In: (Koch, Spizzichino 1982), 287–232.
— (2013): *Exchangeability and its ramifications*. In: (Damien, Dellaportas, Polson, Stephens 2013), ch. 2, 19–29.
de Finetti, B. (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. **IV**[5], 86–133. http://www.brunodefinetti.it/Opere.htm.
— (1937): *La prévision : ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**[1], 1–68. Transl. as (de Finetti 1964).
— (1964): *Foresight: its logical laws, its subjective sources*. In: (Kyburg, Smokler 1980), 53–118. Transl. of (de Finetti 1937) by Henry E. Kyburg, Jr.
DeGroot, M. H. (2004): *optimal statistical decisions*, reprint. (Wiley, New York).
Denny, J. L. (1967): *Sufficient conditions for a family of probabilities to be exponential*. Proc. Natl. Acad. Sci. (USA) **57**[5], 1184–1187.
Diaconis, P., Freedman, D. (1981): *Partial exchangeability and sufficiency*. In: (Ghosh, Roy 1981), 205–236. Also publ. 1982 as technical report https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis_freedman_PES.pdf, http://statweb.stanford.edu/~cgates/PERSI/year.html.
Fraser, D. A. S. (1963): *On sufficiency and the exponential family*. J. Roy. Stat. Soc. B **25**[1], 115–123.
Ghosh, J. K., Roy, J., eds. (1981): *Statistics: Applications and New Directions*. (Indian Statistical Institute, Calcutta).
Gupta, A. K., Nagar, D. K. (2000): *Matrix Variate Distributions*. (Chapman & Hall/CRC, Boca Raton, USA).
Hahn, T. (2005): *CUBA – a library for multidimensional numerical integration*. Comput. Phys. Comm. **168**[2], 78–95. arXiv:hep-ph/0404043, library at http://www.feynarts.de/cuba. See also (Hahn 2014).
— (2014): *Concurrent Cuba*. arXiv:1408.6373. See (Hahn 2005).
Hewitt, E., Savage, L. J. (1955): *Symmetric measures on Cartesian products*. Trans. Am. Math. Soc. **80**[2], 470–501.
Hipp, C. (1974): *Sufficient statistics and exponential families*. Ann. Stat. **2**[6], 1283–1292.
Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst; http://omega.albany.edu:8008/JaynesBook.html, http://omega.albany.edu:8008/JaynesBookPdf.html, http://www-biba.inrialpes.fr/Jaynes/prob.html. First publ. 1994.
Jeffreys, H. (2003): *Theory of Probability*, 3rd ed. (Oxford University Press, London). First publ. 1939.
Kallenberg, O. (2005): *Probabilistic Symmetries and Invariance Principles*. (Springer, New York).
Koch, G., Spizzichino, F., eds. (1982): *Exchangeability in Probability and Statistics*. (North-Holland, Amsterdam).
Koopman, B. O. (1936): *On distributions admitting a sufficient statistic*. Trans. Am. Math. Soc. **39**[3], 399–409.
Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.
Lauritzen, S. L. (1988): *Extremal Families and Systems of Sufficient Statistics*. (Springer, Berlin). First publ. 1982.
Lindley, D. V. (2008): *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*, reprint. (Cambridge University Press, Cambridge). First publ. 1965.

Lindquist, M. A. (2008): *The statistical analysis of fMRI data*. Stat. Sci. **23**[4], 439–464. arXiv: 0906.3662.

Minka, T. (2001): *Inferring a Gaussian distribution*. Tech. rep. (MIT media Lab, Cambridge, USA). http://research.microsoft.com/en-us/um/people/minka/papers/. First publ. 1998.

Murphy, K. P. (2012): *Machine Learning: A Probabilistic Perspective*. (MIT Press, Cambridge, USA).

Pitman, E. J. G. (1936): *Sufficient statistics and intrinsic accuracy*. Math. Proc. Camb. Phil. Soc. **32**[4], 567–579.