Inferring health conditions from fMRI-graph data

P.G.L. Porta Mana [1], **C. Bachmann** [2,*], **A. Morrison** [2,3,4]

[1] *Independent researcher, Italy*
[2] *Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA BRAIN Institute I, Jülich Research Centre, Jülich, Germany*
[3] *Simulation Laboratory Neuroscience, Inst. for Advanced Simulation, JARA, Jülich Research Centre and JARA, Jülich, Germany*
[4] *Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr-University Bochum, Germany*

Correspondence*:
C. Bachmann
c.bachmann@fz-juelich.de

30 January 2018; updated 4 April 2018

ABSTRACT

Automated classification methods for disease diagnosis are currently in the limelight, especially for imaging data. Classification does not fully meet a clinician's needs, however: in order to combine the results of multiple tests and decide on a course of treatment, a clinician needs the likelihood of a given health condition rather than binary classification yielded by such methods. We illustrate how likelihoods can be derived step by step from first principles and approximations, and how they can be assessed and selected, using fMRI data from a publicly available data set containing schizophrenic and healthy control subjects, as a working example. We start from the basic assumption of partial exchangeability, and then the notion of sufficient statistics and the "method of translation" (Edgeworth, 1898) combined with conjugate priors. This method can be used to construct a likelihood that can be used to compare different data-reduction algorithms. Despite the simplifications and possibly unrealistic assumptions used to illustrate the method, we obtain classification results comparable to previous, more realistic studies about schizophrenia, whilst yielding likelihoods that can naturally be combined with the results of other diagnostic tests.

Keywords: disease diagnosis, decision theory, sufficient statistics, exchangeability, parametric statistical model, schizophrenia, fMRI, Bayesian probability theory

1 INTRODUCTION

Consider the following futuristic scenario. A 29-year-old man seeks medical advice because he finds himself in a very confused state: for the last weeks he has been hearing imaginary voices that offend and threaten him. The clinician, after listening to the complaints of the patient, suspects that the man might have developed schizophrenia. The clinician assigns an initial probability for the patient's being healthy or having schizophrenia considering age, sex, familial factors, symptoms, a psychological evaluation, the

incidence of the disease, and similar prior information:

$$\mathrm{P}(\text{health condition} \mid \text{prior info}). \tag{1}$$

Then she orders one or more diagnostic tests to make a better informed assessment of the probability for schizophrenia. Among these tests she orders a functional magnetic-resonance imaging (fMRI) scan. The advantage of fMRI lies in the non-invasive monitoring of brain structure and activity; the structural image is used to exclude morphological changes in the brain such as tumours, and the functional imaging can provide information about changes in brain activity associated with schizophrenia.

With the results of the tests and of the fMRI, the clinician updates her initial or prior probability to a "post-test" or posterior probability based on the results, according to Bayes's theorem:

$$\overbrace{\mathrm{P}(\text{health condition} \mid \text{results of all tests} \wedge \text{prior info})}^{\textit{post-test probability}} \propto \overbrace{\mathrm{P}(\text{health condition} \mid \text{prior info})}^{\textit{initial probability}} \times$$

$$\textit{likelihoods} \begin{cases} \mathrm{P}(\text{result of first test} \mid \text{health condition} \wedge \text{prior info}) \times \\ \mathrm{P}(\text{result of second test} \mid \text{health condition} \wedge \text{prior info}) \times \\ \qquad\qquad\qquad\qquad\qquad\qquad \cdots \times \\ \mathrm{P}(\text{result of fMRI} \mid \text{health condition} \wedge \text{prior info}) \end{cases} \tag{2}$$

where "$\wedge$" denotes logical conjunction ("and"), and we have reasonably assumed that the result of each test does not depend on those of the other tests, i.e. that their likelihoods are independent (Jaynes, 2003, § 4.2; Sox et al., 2013, § 4.7).

In the update formula above, the initial probability is assessed by the clinician. To calculate the post-test probability she needs the probabilities for each test result conditional on the health condition, either "healthy" or "schizophrenic". These probabilities are called the *likelihoods for the health condition* in view of each test. The term "likelihood" has its standard technical meaning in the present work: the probability of a proposition A given B is $\mathrm{P}(A \mid B)$, while the likelihood of A in view of B is $\mathrm{P}(B \mid A)$, i.e., A appears in the conditional (Jaynes, 2003, § 4.1; Good, 1950, § 6.1). A proposition can have high probability but low likelihood and vice versa. Probabilities, not likelihoods, are what we base our decisions upon.

The final, post-test probability is necessary to the clinician to decide upon a course of action (Sox et al., 2013, ch. 6; Goodman, 1999; Murphy, 2012, § 5.7); for example, to treat the patient according to one or another specific treatment, to dismiss him, or to order more tests. To make such decision the clinician will combine her post-test probabilities for the health conditions with a utility table (a reminder of decision theory is given in § 4.3).

The story above is futuristic because fMRI scans are not among the diagnostic tools for schizophrenia today. Our focus in the present work is not this particular brain disease, but we will use it as a working example. What we want to address is this question: *how can we assign a numerical value to the likelihood*

$$\mathrm{P}(\text{fMRI result} \mid \text{health condition} \wedge \text{prior info}). \tag{3}$$

of each health condition in view of the fMRI result? We will propose an answer that can be applied for any brain disease. Our study will use schizophrenia and real data connected to this disease as a concrete example.

To address the question above it is useful to mark out some features of the approach presented so far:

I. **Modularity.** The update formula (2) combines evidence from different tests, and this combination does not need to be done at once. The clinician can multiply her initial probability by the likelihood from the first test, normalize, and thus obtain a "post-first-test" probability. Later she can multiply this probability by the likelihood from the second test, normalize, and thus obtain a post-second-test probability; and so on with any number of other tests, a number that the clinician needs not fix in advance. She can therefore store the value of the likelihood from the fMRI result, to later combine it with new likelihoods from future tests to form a new, better-informed post-test probability.

II. **Decision-theoretic character.** The clinician's final goal is not simply a healthy/schizophrenic classification, but a *decision* upon a course of action about the patient (Sox et al., 2013, chs 6, 7; Jaynes, 2003, chs 13, 14; Raiffa and Schlaifer, 2000). This distinction is important: for example, a treatment without contraindications might be recommended even if there is only a 10% probability that the disease is present; or a dangerous treatment might be recommended only if there is a 90% probability that the disease is present.

The modularity of the present approach extends to the decision stage, because the post-test probability can be used with different decisions and utilities, which can also be updated later on. For example, after beginning a treatment the clinician happens to read about a new kind of treatment, having new benefits and contraindications. Using the post-test probabilities she already has, she may re-evaluate her decision using an updated utility table that includes the new treatment.

III. **Incomplete knowledge.** In general, we lack a complete biological understanding of the relation between brain activity and the health condition under study. In this case, the likelihoods can only be assessed by relying on examples of known *health condition–fMRI data* pairs, usually called a training or calibrating dataset. Moreover, this training dataset is often very small.

IV. **High dimensionality.** The fMRI data are positive-valued vectors with 10^7–10^8 components or more (Lindquist, 2008). This high dimension impacts the calculation of likelihoods and probabilities.

The first two points above are great advantages of the present approach, and also the reasons why it cannot be based on machine-learning algorithms or similar methods. Algorithms for deterministic classification would give the clinician a dichotomous, "healthy/schizophrenic" answer, with no associated uncertainty. Such an answer cannot be used by the clinician to weigh the benefits and risks of different courses of action, the assessment of which needs the probabilities of the health conditions. Probabilistic algorithms, on the other hand, are not flexible for combining evidence: they give a probability for the health condition, not a likelihood; and only the latter can be combined with the likelihoods from other tests, or stored for later reuse and combination.

We therefore approach the question of assigning the likelihoods (3) by means of the probability calculus, the same calculus from which eq. (2) is derived. What we will do is in essence no different from current Bayesian statistical analyses and modelling; but we would like to emphasize some aspects of this modelling that are usually left in the background. The probability calculus can be regarded as the extension of formal logic (truth calculus) to plausible inference (Jeffreys, 2003; Jaynes, 2003; Hailperin, 1996), a view also supported in medicine (Greenland, 1998; Maclure, 1998; Goodman, 1999), which has been proven with

increasing rigour by Koopman (1940b; 1940a; 1941), Cox (1946; 1961; 1979), Pólya (1949; 1968), and many others (Horvitz et al., 1986; Paris, 2006; Halpern, 1999; Snow, 1998; Dupré and Tipler, 2009; Terenin and Draper, 2017). The derivation of a probability proceeds much like an "axioms → logic rules → theorem" derivation in formal logic: one starts from the probabilities of some propositions, and by applying the probability rules, arrives at the probability of the desired proposition, eq. (3) in our case.

We would like to show this procedure step by step in the case of our problem,

in order to expose where assumptions and approximations enter the derivation. These may be improved by other researchers, or replaced by different ones when the method is applied to a different problem. Our discussion is inspired by Mosteller & Wallace's (Mosteller and Wallace, 1963) brilliant, thoughtful analysis of a statistically similar problem in a very different context.

The approach we follow deals naturally with the four points listed above. The small size of the training dataset, point III. above, is not an issue because the probability calculus allows for training datasets of any size. In fact, the calculus allows us to continuously update our inferences given new training data, making our inferences more and more precise and less likely to be affected by outliers.

The unmanageable size of our data space, point IV. above, will force us to make auxiliary assumptions that will translate into the choice of a reduced data space, discussed in § 2.3, and into the use of parametric statistical models, discussed in § 2.4. Regarding the latter, we will emphasize that assumptions about relevant and irrelevant information in our data may translate into mathematical statistical models. It is often difficult to relate biophysical considerations about quantities measured in the brain to the shape of a probability distribution, especially in multidimensional quantities. The notion of *sufficient statistics* (Dawid, 2013; Bernardo and Smith, 2000, ch. 4; Lindley, 2008, § 5.5; Diaconis and Freedman, 1981; Cifarelli and Regazzini, 1982; Lauritzen, 1988; Kallenberg, 2005), discussed in § 2.4.2, is a helpful bridge between biophysical considerations and probability distributions. The idea is that it may be easier for us to conceive a connection between biophysical considerations and some special statistics of our measurements, than between biophysical considerations and an abstract multidimensional distribution function. This "translation" is powerful, because if one finds the assumptions about relevance or irrelevance of some data unreasonable, one can then make different assumptions, resulting in a different statistical model.

Models inspired by sufficient statistics – especially their comparison and selection – can nevertheless be computationally demanding owing to the multidimensional integrals in their formulae, even when these are addressed by modern numerical methods such as Monte Carlo (MacKay, 2003, ch. IV; Murphy, 2012, chs 23–24). In the present study we shall use analytically tractable statistical models, but availing ourselves of Edgeworth's "Method of Translation" (1898; Johnson, 1949; Mead, 1965): the simple but potentially very fertile idea of transforming a quantity into a normally distributed one, discussed in § 2.4.3.

The possible combined choices of reduction of the data space, of sufficient statistics, and of transformations into normal variable, lead to a variety of possible models and likelihoods to be used by the clinician. Which is the "best" one? We discuss several criteria for choice in § 2.5, settling on one based on expected utility. We also briefly discuss the remarkable observation that common Bayesian criteria based on weight of evidence and Bayes factors (Jeffreys, 2003, chs V, VI, A; Good, 1950; MacKay, 1992a; Kass and Raftery, 1995) for the fMRI data gives results opposite to those of the expected-utility criterion.

In this article, we will calculate the likelihoods for the health conditions, eq. (3), and assess the models for so doing, in the following steps. First, in § 2.1, we briefly discuss schizophrenia and the use of fMRI to diagnose it, introducing a concrete dataset of fMRI data for schizophrenic and healthy patients. We then show that a simple and natural assumption, called *exchangeability*, would lead to a unique value of the

likelihoods (3) if the training dataset were large enough (§ 2.2). However, with a small training dataset we must face two problems: unmanageably large dimensions of the data space, and the need to specify prior beliefs, also involving functions on infinite-dimensional spaces.

To solve the first problem, in § 2.3 we assume that information adequate for our health inference can be found in a reduced data space of the fMRI, which we construct from time correlations between groups of voxels. To solve the second problem we introduce parametric statistical models in § 2.4 using the notions of *sufficient statistics* and of transformation into normal variables, mentioned above. We discuss how these models learn from the data and select three models as possible candidates. We then consider several criteria to select one of the three models against our data, as an example, and discuss how a more realistic assessment could be made in a real application (§ 2.5). We conclude with a discussion (§ 3) on how the choice of sufficient statistics and prior probabilities could be improved, and on the relation to machine-learning methods.

Our statistical terminology and notation follow ISO standards (ISO, 2009, 2006).

2 RESULTS

2.1 Selection of clinical use case and fMRI-data acquisition

Schizophrenia is a psychiatric disorder that comprises various symptoms that are categorized into positive (e.g. hallucinations), negative (e.g. loss of motivation) and cognitive (e.g. memory impairment) disease patterns. A common disease cause for all these widespread symptoms is still unknown. Functional magnetic resonance imaging (fMRI) has been used to gain insight into modifications in functional connectivity in this disease. In the resting state, functional connectivity is measured either by asking the subject to fulfil a certain task or at rest, instructing the subject to think about nothing specific but not fall asleep. In this condition, both increased and decreased functional connectivity have been reported in the default mode network, although the hyperactivity seems to be reported more often (Hu et al., 2017). Moreover, widespread connectivity changes in the dorsal attention network and the executive control network have been detected (Woodward et al., 2011; Yu et al., 2016).

Beyond these individual sub-networks, many studies have found profound changes in macroscopic brain structures, e.g. a shrinkage of whole brain and ventricular volume, reduced gray matter in frontal, temporal cortex and thalamus, and changes in white matter volume in frontal and temporal cortex (Shenton et al., 2010; Ellison-Wright and Bullmore, 2009, 2010). Since both gray matter loss and white matter changes are found, it is reasonable to conclude that not only the intrinsic activity of single areas is modified, but also the interplay of different brain areas, in particular in frontal and temporal cortex. It has been argued that these alterations in long range connectivity are responsible for a range of disease symptoms that are not attributable to single areas (Friston and Frith, 1995). Taking this disconnect hypothesis as a starting point, we can reach the working hypothesis that these changes are also reflected in the functional activity of the brain, and that fMRI images can be used to distinguish schizophrenic from healthy patients. We therefore conclude that schizophrenia is an appropriate condition to demonstrate our approach.

We requested data of schizophrenic and healthy patients from Schizconnect[1], a virtual database for public schizophrenia neuroimaging data. In our request we asked for resting state T2*-weighted functional (rfMRI) and T1-weighted structural magnet resonance images (MRI) from patients participating in the COBRE study either with no known disorder or diagnosed as schizophrenic according to the Diagnostic

[1] http://schizconnect.org/

and Statistical Manual of Mental Disorders (DSM) IV, excluding schizoaffective disorders. In the COBRE study, the voluntary and informed participation of the subjects was ensured by the institutional guidelines at the University of New Mexico Human Research Protections Office. The ensuing dataset comprised 91 healthy patients and 74 schizophrenic patients. Out of these we randomly selected 54 healthy and 49 schizophrenic subjects, to permit demonstration our method on a small dataset with unequal group size. A detailed description on the exact experimental design and the MRI scanning is provided by Çetin et al. (2014).

2.2 Calculation of probabilities: exchangeability

Let us describe our context more precisely and set up some mathematical notation. We have:

- A number of possible health conditions, in our example healthy (H) and schizophrenic (S). The variable c denotes health condition.
- A space of possible fMRI data. They are vectors with 10^7–10^8 or more positive components (Lindquist, 2008). The variable $\boldsymbol{f}$ denotes an fMRI result.
- A set of n patients, labelled in some way, the variable i denoting their labels. These labels may reflect information about the times the patients were examined, or about their geographical location. This possibility is important in the considerations to follow. In our study $n = 104$.
- Knowledge of the health condition and of the fMRI result of each patient. Let us use the propositions

$$\begin{aligned} C_i^c &:= \text{“Patient } i \text{ has health condition } c\text{”}, \\ F_i^{\boldsymbol{f}} &:= \text{“The fMRI of patient } i \text{ gives } \boldsymbol{f}\text{”}, \end{aligned} \tag{4}$$

the latter to be understood within a very small interval $(\boldsymbol{f}, \boldsymbol{f} + \mathrm{d}\boldsymbol{f})$. In our study we have $n_{\mathrm{H}} = 55$ healthy and $n_{\mathrm{S}} = 49$ schizophrenic patients.

For brevity we denote by C^c the conjunction of the propositions C_i^c for all patients having health condition c, i.e. our knowledge about which patients have that health condition; and analogously for F^c. By D^c we denote all data about patients with health condition c; by D we denote all our data.

- An imaginary patient, labelled “0”, whose fMRI result $\boldsymbol{f}$ is known, but whose health condition is not.
- Other pre-test information, denoted by I; for example the clinician's initial diagnosis of the health condition of patient 0, and the results of any other diagnostic tests.
- The probabilities $(\hat{p}_{\mathrm{H}}, \hat{p}_{\mathrm{S}})$ for the health condition of patient 0, conditional on the pre-test information, including the results from other tests. We call these *pre-test probabilities*. Note that they may differ from the *initial* probabilities of eqs (1)–(2), because they may include the likelihoods from other tests.
- A set of decisions about the patient 0 and their utilities conditional on the patient's health condition. We shall simply consider two decisions: dismiss (D) or treat (T). See § 4.3 for a summary of decision theory.

Our goal is to assign numerical values to the likelihoods for the health conditions (3): the conditional probability distribution that the fMRI result of patient “0” is $\boldsymbol{f}$ given the health condition of that patient and all other data. In our notation,

$$\begin{aligned} & \mathrm{p}(F_0^{\boldsymbol{f}} \,|\, C_0^c \,\wedge\, C_1^{c_1} \wedge F_1^{\boldsymbol{f}_1} \,\wedge\, \cdots \,\wedge\, C_n^{c_n} \wedge F_n^{\boldsymbol{f}_n} \,\wedge\, I)\,\mathrm{d}\boldsymbol{f} \\ \text{or just}\quad & \mathrm{p}(F_0^{\boldsymbol{f}} \,|\, C_0^c \wedge D \wedge I)\,\mathrm{d}\boldsymbol{f}. \end{aligned} \tag{5}$$

A natural assumption helps us restrict the values the distribution above may have. Within a group of patients *having the same health condition* we assume that the probability that a patient shows a particular fMRI $\boldsymbol{f}_i$ does not depend on the particular value of the patient's label i, no matter how many patients we have or may later add in that health group. This assumption is called *partial exchangeability* (de Finetti, 1938; Diaconis and Freedman, 1981; Aldous, 1985; Diaconis, 1988). If the labels carry e.g. temporal or geographical information, partial exchangeability means that we do not expect to observe particular kinds of fMRI results more often in the future than in the past, or more frequently in one location than another. As a concrete example: fix three possible fMRI results $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, $\boldsymbol{f}_3$ (each is a vector with 10^7–10^8 positive components) and consider the fMRI tests of three schizophrenic patients: say, one from five years ago in Germany, one from last week in Scotland, and one to be done six months from now in Italy. Partial exchangeability means that the probability that the German patient's test gave $\boldsymbol{f}_1$, the Scottish's gave $\boldsymbol{f}_2$, and the Italian's will give $\boldsymbol{f}_3$, is numerically equal to the probability that the German's gave $\boldsymbol{f}_2$, the Scottish's $\boldsymbol{f}_3$, and the Italian's will give $\boldsymbol{f}_1$; and likewise for all six possible permutations of the three results. Keeping the same fixed fMRI results $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, $\boldsymbol{f}_3$, we now consider three healthy patients instead, who may also live in different times and places. Partial exchangeability means that also in this case the values of the six possible joint probabilities obtained by permutation must all be equal – but this value can be different from the one for the schizophrenic patients considered before. Hence the term "partial": we can freely exchange the joint results within the schizophrenic group and within the healthy group without altering their probabilities, but not across groups. This assumption extends in an analogous way to more patients.

The assumption of partial exchangeability might not be completely true when we consider geographical or epochal differences, but we may still consider it as a good approximation. We are not making any exchangeability assumptions about the probabilities of the health conditions of our patients, though, because the incidence of a disease does often change with time and can depend heavily on geographical location.

To express partial exchangeability mathematically, suppose that the patients $i = 1, 2, 3, \dots$ have health condition $c = \mathrm{H}$ and the patients $i' = 1', 2', 3', \dots$ health condition $c = \mathrm{S}$. Then the joint distribution for their fMRI results satisfies

$$\mathrm{p}\Big(\bigwedge_i F_i^{\boldsymbol{f}_i} \bigwedge_{i'} F_{i'}^{\boldsymbol{f}_{i'}} \,\Big|\, \bigwedge_i C_i^{\mathrm{H}} \bigwedge_{i'} C_{i'}^{\mathrm{S}} \wedge I\Big) = \mathrm{p}\Big(\bigwedge_i F_i^{\boldsymbol{f}_{\pi(i)}} \bigwedge_{i'} F_{i'}^{\boldsymbol{f}_{\pi'(i')}} \,\Big|\, \bigwedge_i C_i^{\mathrm{H}} \bigwedge_{i'} C_{i'}^{\mathrm{S}} \wedge I\Big)$$
$$\text{for all permutations } \pi \text{ of } \{i\} \equiv \{1, 2, \dots\}, \text{ and all permutations } \pi' \text{ of } \{i'\} \equiv \{1', 2', \dots\}. \quad (6)$$

The assumption of partial exchangeability is simple and quite natural – and very powerful: it implies, as shown by de Finetti (1938; Diaconis, 1988, § 3; Bernardo and Smith, 2000, § 4.6), that the joint distributions above must have the form

$$\mathrm{p}\Big(\bigwedge_i F_i^{\boldsymbol{f}_i} \bigwedge_{i'} F_{i'}^{\boldsymbol{f}_{i'}} \,\Big|\, C \wedge I\Big) = \iint \Big[\prod_i q_{\mathrm{H}}(\boldsymbol{f}_i)\Big] \Big[\prod_{i'} q_{\mathrm{S}}(\boldsymbol{f}_{i'})\Big] \,\mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \,|\, I)\, \mathrm{d}q_{\mathrm{H}}\, \mathrm{d}q_{\mathrm{S}} \quad (7)$$

where q_{H}, q_{S} are distributions over the possible values of $\boldsymbol{f}$, and $\mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \,|\, I)$ is a "hyperdistribution" over such distributions, determined by the assumptions I. The double integral (which can be understood as a generalized Riemann integral: Lamoreaux and Armstrong, 1998; Swartz, 2001; Kurtz and Swartz, 2004), is over all distributions q_{H}, q_{S}. In other words, de Finetti's theorem say that the joint probability distribution for the fMRIs of healthy and schizophrenic patients can be seen as the product of independent distributions,

identical for healthy cases and identical for schizophrenic cases but different for the two cases, mixed over all possible such pairs of distributions with weight $\mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid I)$.

As a very cursory example, suppose we want the joint probability that a healthy patient has fMRI result $\boldsymbol{f}$ and a schizophrenic one $\boldsymbol{f}'$. De Finetti's formula first tells us to consider all possible distributions over positive vectors. As usual with infinities, "all" must be made precise by specifying a topology (for details see e.g. de Finetti, 1938; Diaconis and Freedman, 1981; Aldous, 1985; Diaconis, 1988); but intuitively these distributions comprise, e.g., multivariate truncated normals, gammas, exponentials, truncated Cauchys… and innumerable distributions that we can imagine and don't have a specific name for; all with their possible parameter values. De Finetti's formula tells us to choose one distribution q_{H}, from all those possible ones, for the healthy case and one q_{S} for the schizophrenic case, and to attach a weight to this pair, $\mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid I)$; then to calculate this pair at the values $\boldsymbol{f}$, $\boldsymbol{f}'$ and multiply them: $q_{\mathrm{H}}(\boldsymbol{f}) \times q_{\mathrm{S}}(\boldsymbol{f}')$. Then we consider a new pair of distributions, attach a weight to them, and again multiply their values at $\boldsymbol{f}$ and $\boldsymbol{f}'$. And so on, until all possible pairs are considered. Finally we calculate the sum of all such products, weighted accordingly: $\int q_{\mathrm{H}}(\boldsymbol{f}) q_{\mathrm{S}}(\boldsymbol{f}')\, \mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid I)\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{f}'$.

The generalization of the formulae above to more than two health conditions, or when only one health condition is considered, is straightforward.

As the cursory example above made quite clear, an integral over probability distributions is a mathematically complicated object (cf. Ferguson, 1974) and may not seem a great advancement in assigning a value to the distribution (5). In defence of de Finetti's formula we must say that it is completely manageable with discrete data spaces and provides a great insight in the way we reason about probability in relation to repeated events (de Finetti, 1937; Lindley and Phillips, 1976; Kingman, 1978; Koch and Spizzichino, 1982; Dawid, 2013; Bernardo and Smith, 2000, § 4.2). It also has several important consequences for our inference, which we now discuss.

Using de Finetti's formula and the definition of conditional probability we can rewrite our goal plausibility (5) as

$$\mathrm{p}(F_0^{\boldsymbol{f}} \mid C_0^c \wedge D \wedge I) = \int q_c(\boldsymbol{f})\, \mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid D \wedge I)\, \mathrm{d}q_{\mathrm{H}}\, \mathrm{d}q_{\mathrm{S}} \tag{8a}$$

$$\text{with} \quad \mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid D \wedge I) = \frac{\left[\prod_i q_{c_i}(\boldsymbol{f}_i)\right] \mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid I)}{\int \left[\prod_i q_{c_i}(\boldsymbol{f}_i)\right] \mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid I)\, \mathrm{d}q_{\mathrm{H}}\, \mathrm{d}q_{\mathrm{S}}}. \tag{8b}$$

The latter is called posterior distribution since it is conditional on all data D.

Excluding pathological prior distributions (Diaconis and Freedman, 1986), this posterior distribution becomes more and more concentrated on two particular distributions $\left(\hat{q}_{\mathrm{H}}, \hat{q}_{\mathrm{S}}\right)$, fully determined by the data, as our data D comprise a larger and larger number of patients. This concentration occurs independently of the original shape of the distribution $\mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}} \mid I)$. In this limit our probability distribution (5) becomes

$$\mathrm{p}(F_0^{\boldsymbol{f}} \mid C_0^c \wedge D \wedge I) \approx \hat{q}_c(\boldsymbol{f}) \tag{9}$$

with $\hat{q}_c$ completely determined by the data D. This would solve our plausibility assessment (5).

De Finetti's formula therefore tells us also the theoretical limit by which the pre-test probability for the health condition of the patient, $\mathrm{P}(C_0^c \mid D \wedge I)$, can be improved by the fMRI result. For example, if $\hat{q}_c(\boldsymbol{f})$ is more or less uniform in $\boldsymbol{f}$ or has the same peaks in $\boldsymbol{f}$ for each c, then the fMRI is of no use for

discriminating the health condition of the patient. This result follows mathematically from the assumption of exchangeability (6) and the rules of probability calculus, hence no amount of ingenuity could overcome this limit.

In our case the amount of data D is not enough to allow the use of the approximation (9). We should use the general formula (8), but it is unwieldy in two respects. First, the fMRI result $\boldsymbol{f}$ of a patient is a positive-valued vector with 10^7–10^8 components or more (Lindquist, 2008), so the distributions q_{H}, q_{S} are highly multidimensional. Second, the formula asks us to consider in principle *all* such distributions, as explained in the example above.

We tame this double unwieldiness in two ways. First, it is conceivable that not all information contained in the fMRI result $\boldsymbol{f}$ of a patient be relevant to discriminate the patient's health condition c. The integral in eq. (8), 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 in the space of fMRI data. Being unable to perform the integral, we must try to apply heuristics based on our understanding of the target conditions and perform such dimensional reduction by hand. For example, by employing the hypothesis that in schizophrenic patients the time correlation between brain regions is altered with respect to healthy ones. We can thus address the first problem by reducing fMRI data $\boldsymbol{f}$ to a manageable set of graph properties $\boldsymbol{f}$, and applying our inference directly on these, as explained in the next section.

Second, we entertain a working assumption about which features of our graph data $\{\boldsymbol{f}_i\}$ from a set of patients are relevant for inferences about new patients. We can, for example, assume that only the first and second moments are relevant for making predictions about the graph quantities to be observed in the new patients; these moments are then called *sufficient statistics*. Assumptions of this kind reduce the infinite-dimensional space all possible distributions $(q_{\mathrm{H}}, q_{\mathrm{S}})$ to a finite-dimensional space of a parametric family of distributions of exponential type, as explained in § 2.4.

2.3 Trimming the data space: functional connectivity

The preprocessed functional image in standard space, in which the activity for each voxel is recorded, consists of approximately 10^{17} time series of 140 time points. We reduce this huge data space by the following steps.

First, we consider only the activity of the voxels belonging to the 94 regions defined by the lateral cortical Oxford atlas (see § 2.1 and Desikan et al., 2006) and average the activity of all voxels in a region (details in § 4.1). Second, we measure the functional connectivity defined by the Pearson correlation coefficient between pairs of regions, obtaining $94 \times 93/2 = 4371$ connectivity weights. This is still a considerable data space for our computational resources, so we select $d = 40$ connectivity weights that exhibit the greatest difference in their connection weight average across the schizophrenic and healthy groups.

The resulting distributions for four of these connectivity weights are depicted in fig. 1. Note that for each considered brain connection, the histograms of the healthy group and the schizophrenic group display significant overlap, such that none of them could be used in isolation to reliably discriminate between two groups.

Our data space is therefore vastly reduced, from $[0, \infty[^{10^{17}}$

to $[-1, 1]^{40}$. We let the symbol $\boldsymbol{f}$ stand for the set of connectivity weights extracted from the fMRI data, rather than the full fMRI data themselves. With this new meaning of the symbol $\boldsymbol{f}$, the likelihoods for the

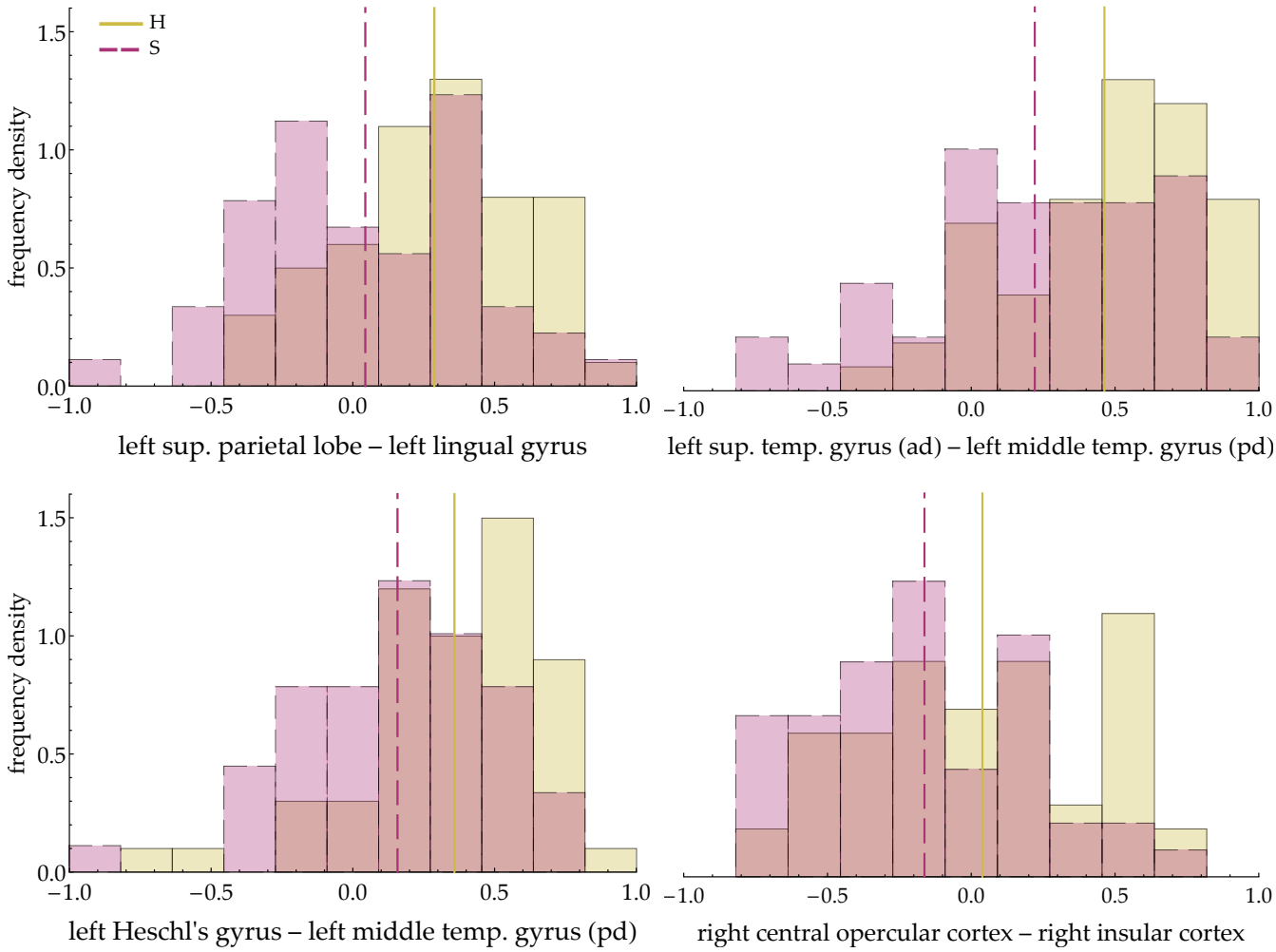

Figure 1. Distributions of functional connectivities of schizophrenic and healthy patients. Normalized density histograms of connectivity weights for healthy (H, yellow, solid) and schizophrenic (S, red, dashed) patients of four cortical connections selected to demonstrate our statistical framework. Empirical means are shown as vertical lines. All connectivities for healthy and schizophrenic patients have substantial overlap, evidenced by the darker regions in the histograms.

health conditions (5) and de Finetti's formulae (7)–(8) remain formally unchanged, but now involve a data space with much fewer dimensions.

2.4 Trimming the distribution space: models by sufficiency and generalized normals

2.4.1 Parametric models

The integrals in de Finetti's formulae (7)–(8) still represent a mixture of all imaginable pairs of probability distributions $(q_{\mathrm{H}}, q_{\mathrm{S}})$ over the space $[-1,1]^d$, and are therefore extremely complex. We now examine two ways to reduce this integral to a manageable set of distributions and to obtain analytically tractable formulae.

In the Bayesian literature, the complication of considering all possible distributions $q_c(\boldsymbol{f})$ of the quantity $\boldsymbol{f}$ is typically sidestepped by restricting them to a finite-dimensional set of distributions $L_c(\boldsymbol{f}\,|\,\boldsymbol{\theta}_c)$, identified or indexed by a finite number of parameters $\boldsymbol{\theta}_c$. For this reason, such a set is called a parametric family of distributions. An example of parametric family is the set of d-variate normal distributions parameterized by their mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. With this restriction, the integrals in de Finetti's formulae (7) and

(8) represent mixtures of distributions within the parametric family, the weight for each distribution being represented by a weight for its parameters. That is, we are no longer considering mixtures of all possible multivariate truncated normals, gammas, exponentials, etc., as in the example of § 2.2, but only mixtures of truncated normals, say, with different means and covariance matrices. These integrals are thus ordinary finite-dimensional integrals. What happens to formulae (7)–(8) is that

$$q_c(\boldsymbol{f}) \text{ is replaced by } L_c(\boldsymbol{f}\,|\,\boldsymbol{\theta}_c), \qquad \mathrm{p}(q_{\mathrm{H}}, q_{\mathrm{S}}\,|\,I)\,\mathrm{d}q_{\mathrm{H}}\,\mathrm{d}q_{\mathrm{S}} \text{ is replaced by } \mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}\,|\,M, I)\,\mathrm{d}\boldsymbol{\theta}_{\mathrm{H}}\,\mathrm{d}\boldsymbol{\theta}_{\mathrm{S}}. \tag{10}$$

The distribution L_c is called the likelihood of the parameters $\boldsymbol{\theta}_c$, and $\mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}\,|\,M, I)$ is the prior parameter distribution. A parametric family and a prior distribution over its parameters are jointly called a parametric statistical model, which we denote by M. The term "model" is justly criticized by some probability theorists (see e.g. Besag & Kalman in Besag et al., 2002) but widely used, so we shall adopt it here.

In our present problem, the parametric statistical model needs not be the same for all health conditions: for example, we could use normal distributions for one condition and beta distributions for another, if that choice better reflected the distributions of connectivity weights under the two different health conditions. For this reason, we use the subscript "c" in the formulae above. The likelihood we want to determine, eq. (5), thus becomes, from eq. (8) with the replacements (10),

$$\mathrm{p}(F_0^{\boldsymbol{f}}\,|\,C_0^c \wedge D \wedge M \wedge I) = \int L_c(\boldsymbol{f}\,|\,\boldsymbol{\theta}_c)\,\mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}\,|\,D, M, I)\,\mathrm{d}\boldsymbol{\theta}_{\mathrm{H}}\,\mathrm{d}\boldsymbol{\theta}_{\mathrm{S}} \tag{11a}$$

$$\text{with} \quad \mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}\,|\,D, M, I) = \frac{\left[\prod_i L_{c_i}(\boldsymbol{f}_i\,|\,\boldsymbol{\theta}_{c_i})\right]\mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}\,|\,M, I)}{\int\left[\prod_i L_{c_i}(\boldsymbol{f}_i\,|\,\boldsymbol{\theta}_{c_i})\right]\mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}\,|\,M, I)\,\mathrm{d}\boldsymbol{\theta}_{\mathrm{H}}\,\mathrm{d}\boldsymbol{\theta}_{\mathrm{S}}}. \tag{11b}$$

2.4.2 Models by sufficient statistics

The choice of a statistical model often appears as an art and a matter of experience. Notable statisticians have voiced concerns over the esoteric character of this choice. Dawid (1982a, p. 220) says: "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". And Diaconis (1988, § 8, p. 121) remarks:

> de Finetti's alarm at statisticians introducing reams of unobservable parameters has been repeatedly justified in the modern curve fitting exercises of today's big models. These seem to lose all contact with scientific reality focusing attention on details of large programs and fitting instead of observation and understanding of basic mechanism. It is to be hoped that a fresh implementation of de Finetti's program based on observables will lead us out of this mess.

Authors like these have also tried to develop intuitive methods to choose a model, for example by proving that a parametric family can be uniquely determined by some symmetry assumptions about our inferences, or by other information-theoretical properties (Bernardo and Smith, 2000, ch. 4; Lindley, 2008, § 5.5; an enlightening discussion of this topic is given by Dawid, 2013). In the present study we want to emphasize, as Cifarelli & Regazzini (1982) did, that a statistical model can be chosen by selecting a *sufficient statistics* (Kolmogorov, 1942; Freedman, 1962; Diaconis and Freedman, 1980, 1981; Cifarelli and Regazzini, 1982; Lauritzen, 1988; Diaconis, 1992; Kallenberg, 2005, and the textbook references above). Here is an example.

Imagine that we have patients labelled $i' \in \{1', 2', \dots\}$, and n patients labelled $i \in \{1, 2, \dots\}$, all with the same health condition. Of the second set of patients we also know the connectivity weights $\{\boldsymbol{f}_i\}$

obtained by fMRI. We want to specify the joint probability distribution $\mathrm{p}(\{\boldsymbol{f}_{i'}\}|\,\{\boldsymbol{f}_i\}, I)$ that the fMRIs of the patients $\{i'\}$ yield connectivity weights $\{\boldsymbol{f}_{i'}\}$, conditional on our knowledge of the connectivity weights of the n patients $\{i\}$. Now assume that the probabilities for the fMRI results are exchangeable, so that de Finetti's formulae (7) and (8) hold. Also assume that in order to specify the joint distribution we do not need the full set of data $\{\boldsymbol{f}_i\}$, but only their number n and some statistics, e.g. the empirical mean and covariance matrix of these data,

$$\bar{\boldsymbol{f}} := \frac{1}{n}\sum\nolimits_i \boldsymbol{f}_i, \qquad \mathrm{Cov}(\boldsymbol{f}) := \frac{1}{n}\sum\nolimits_i (\boldsymbol{f}_i - \bar{\boldsymbol{f}})(\boldsymbol{f}_i - \bar{\boldsymbol{f}})^{\intercal}; \tag{12}$$

the rest of the details of the data $\{\boldsymbol{f}_i\}$ being irrelevant. In other words, we are assuming that the statistics above are *sufficient* for us to make predictions as if we had the full data. In symbols,

$$\mathrm{p}(\{\boldsymbol{f}_{i'}\}|\,\{\boldsymbol{f}_i\}, I) = \mathrm{p}(\{\boldsymbol{f}_{i'}\}|\,n, \bar{\boldsymbol{f}}, \mathrm{Cov}\,\boldsymbol{f}, I). \tag{13}$$

If this is true no matter the number of patients $\{i'\}$ and $\{i\}$, then these statistics are called *(predictive) sufficient statistics*. There are several notions of sufficiency, including the traditional one by Fisher (1922) and Neyman (1935), but they all are more or less equivalent (Bernardo and Smith, 2000, § 4.5.2).

The assumption of the existence of sufficient statistics has a very important consequence for de Finetti's formulae (7) and (8): the space of possible prior distributions is hugely reduced, constrained to be non-zero only over a parametric family of distributions that is determined by the sufficient statistics. The replacement (10) takes place, leading to the simpler formula (11) for the likelihoods of the health conditions. The number of parameters is equal to that of the sufficient statistics. The proof of this reduction was given by Pitman and Koopman (Koopman, 1936; Pitman, 1936; Darmois, 1935; for generalizations, e.g. to the discrete case, see Hipp, 1974; Andersen, 1970; Denny, 1967; Fraser, 1963; Barankin and Maitra, 1963). When the sufficient statistics are the mean and covariance matrix, as above, the likelihoods turn out to be (truncated) multivariate normal distributions.

In the slightly more complicated case of two or more health conditions and the assumption of partial exchangeability (6), this theorem leads to formula (11) with likelihoods L_c determined by the sufficient statistics that we have chosen for the different health conditions (Bernardo and Smith, 2000, § 4.6). The prior distribution over the parameters of the likelihoods, $\mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}|\, I)$, is not determined by the theorem, but has to be determined by other consideration that can again involve symmetry and information theory.

As we mentioned in the introduction, the notion of sufficient statistics can be a helpful bridge between biophysical considerations and the specification of probabilities. It may be easier to conceive and understand a connection between biophysical considerations and some statistics of our measurements, than between biophysical considerations and an abstract multidimensional distribution function. Once such statistics are selected, they in turn uniquely select a probability distribution for us. Vice versa, if a statistical model based on some sufficient statistics proves to be very reliable in its predictions, we may conclude that its sufficient statistics must have an important biological meaning.

In the rest of this study we shall use three statistical models determined by three different sufficient statistics. Our choice of statistics is unfortunately not biologically motivated, as such models have yet to be determined for fMRI data. However, they are adequate to demonstrate the approach and we hope that authors with more experience will pursue this line of thought and derive better-motivated sufficient statistics.

2.4.3 Edgeworth's "method of translation": generalized normal models

The assumption of a sufficient statistics makes the integrals in de Finetti's formulae (7) and (8) finite-dimensional, but these integrals and other expressions that depend on them, like the post-test probabilities, may still lack a closed form and be analytically intractable. In this case we could use numerical methods, a computationally costly possibility we consider in the Discussion, § 3.3. In the present work we choose models with closed-form formulae instead; their swift calculation facilitates the model comparison to be illustrated later.

Our starting point is an analytically tractable model by sufficient statistics that has been the subject of much study (Gelman et al., 2014, § 3.6; Minka, 2001; Murphy, 2007): it has a normal likelihood, with mean $\boldsymbol{\lambda}$ and covariance matrix $\boldsymbol{\Lambda}$ as parameters, and a normal-inverse-Wishart prior distribution over these parameters. This parameter prior maintains the same functional form when updated with training data; this kind of prior is called *conjugate* (DeGroot, 2004, ch. 9; Diaconis and Ylvisaker, 1979). This model is outlined more in detail in § 4.2.

We try to make the normal + normal-inverse-Wishart model more flexible by combining it with an idea that Edgeworth (1898) called "Method of Translation", discussed also by Johnson (1949) and Mead (1965): the transformation of a quantity into a normally distributed one. That is, instead of considering the connectivity weights $\boldsymbol{f}$, we consider transformed quantities $l(\boldsymbol{f})$, where l is a component-wise monotonic function, and suppose the latter quantities to be normally distributed. This leads to generalized-normal likelihoods of the form

$$\mathrm{N}[l(\boldsymbol{f})|\,\boldsymbol{\lambda},\boldsymbol{\Lambda}]\,l'(\boldsymbol{f})\,\mathrm{d}\boldsymbol{f} \tag{14}$$

where N is the normal distribution with mean $\boldsymbol{\lambda}$ and covariance matrix $\boldsymbol{\Lambda}$, and l' is the Jacobian determinant of l.

This simple idea has an amazing scope: it allows us to explore a vast range of non-normal likelihoods – in particular likelihoods defined on bounded domains such as $[-1,1]^d$ – and to keep the low computational costs of the conjugate prior. In our present problem it has also an additional convenient feature: in the calculation of the post-test probability, the Jacobian determinant l' disappears, as eq. (19) below shows.

These generalized-normal models, which we denote by M_l, are also determined by a choice of sufficient statistics, analogous to eq. (12): the mean and covariance matrix of the transformed data $\{l(\boldsymbol{f}_i)\}$,

$$\overline{l(\boldsymbol{f})}, \qquad \mathrm{Cov}[l(\boldsymbol{f})]. \tag{15}$$

In our partially exchangeable case, with formulae (11), we need to specify a likelihood L_c for each health condition c. We assume these two likelihoods L_{H}, L_{S} to be functionally identical generalized normals, i.e. the function l is the same for the healthy and the schizophrenic case; but their means and covariance matrices $(\boldsymbol{\lambda}_{\mathrm{H}},\boldsymbol{\Lambda}_{\mathrm{H}})$ and $(\boldsymbol{\lambda}_{\mathrm{S}},\boldsymbol{\Lambda}_{\mathrm{S}})$ can be different.

We also need to specify a joint parameter prior for $(\boldsymbol{\lambda}_{\mathrm{H}},\boldsymbol{\Lambda}_{\mathrm{H}};\boldsymbol{\lambda}_{\mathrm{S}},\boldsymbol{\Lambda}_{\mathrm{S}})$. To use the analytic advantage of the conjugate prior, we assume that the distribution for these parameters is a product of independent distributions:

$$\mathrm{p}(\boldsymbol{\lambda}_{\mathrm{H}},\boldsymbol{\Lambda}_{\mathrm{H}};\boldsymbol{\lambda}_{\mathrm{S}},\boldsymbol{\Lambda}_{\mathrm{S}}|\,M_l,I) = \mathrm{p}(\boldsymbol{\lambda}_{\mathrm{H}},\boldsymbol{\Lambda}_{\mathrm{H}}|\,M_l,I)\times\mathrm{p}(\boldsymbol{\lambda}_{\mathrm{S}},\boldsymbol{\Lambda}_{\mathrm{S}}|\,M_l,I), \tag{16}$$

each of them being a normal-inverse-Wishart distribution described in § 4.2. This independence assumption is quite strong and has an important consequence: *the likelihood for a health condition only depends on the data from previous patients having that same health condition*.

With the assumptions above, the likelihood for the health condition needed by the clinician has a closed form for this model (see § 4.2). The likelihood for patient 0's being healthy, in view of the patient's measured connectivity weights $\boldsymbol{f}$, and given the data $(\boldsymbol{f}_i)$ from previous n_{H} healthy patients, is

$$\mathrm{p}(\text{fMRI result } \boldsymbol{f} \mid \text{ healthy } \wedge \text{ prior info}) = \mathrm{t}\Big[l(\boldsymbol{f}) \Big|\ \nu_{\mathrm{H}} - d + 1, \boldsymbol{\delta}_{\mathrm{H}}, \tfrac{\kappa_{\mathrm{H}}+1}{\kappa_{\mathrm{H}}\,(\nu_{\mathrm{H}}-d+1)}\boldsymbol{\Delta}_{\mathrm{H}}\Big] \prod_k l'(f_k), \tag{17}$$

where t is a multivariate t distribution with $\nu_{\mathrm{H}} - d + 1$ degrees of freedom, mean $\boldsymbol{\delta}_{\mathrm{H}}$, and scale matrix $\frac{\kappa_{\mathrm{H}}+1}{\kappa_{\mathrm{H}}\,(\nu_{\mathrm{H}}-d+1)}\boldsymbol{\Delta}_{\mathrm{H}}$. This distribution has covariance matrix $\frac{\kappa_{\mathrm{H}}+1}{\kappa_{\mathrm{H}}\,(\nu_{\mathrm{H}}-d-1)}\boldsymbol{\Delta}_{\mathrm{H}}$ (note the different denominator from the scale matrix), and approaches a generalized normal for large ν_{H}. The final factor is the Jacobian determinant of l.

The most important feature of this likelihood is the dependence of the coefficients $(\kappa_{\mathrm{H}}, \boldsymbol{\delta}_{\mathrm{H}}, \nu_{\mathrm{H}}, \boldsymbol{\Delta}_{\mathrm{H}})$ on the data $(\boldsymbol{f}_i)$ of the previous n_{H} healthy patients:

$$\begin{aligned} &\kappa_{\mathrm{H}} = \kappa_0 + n_{\mathrm{H}}, && \nu_{\mathrm{H}} = \nu_0 + n_{\mathrm{H}}, \\ &\boldsymbol{\delta}_{\mathrm{H}} = \frac{\kappa_0\,\boldsymbol{\delta}_0 + n_{\mathrm{H}}\,\overline{l(\boldsymbol{f})}}{\kappa_0 + n_{\mathrm{H}}}, && \boldsymbol{\Delta}_{\mathrm{H}} = \boldsymbol{\Delta}_0 + n_{\mathrm{H}}\ \mathrm{Cov}[l(\boldsymbol{f})] + \frac{\kappa_0\,n_{\mathrm{H}}}{\kappa_0 + n_{\mathrm{H}}}\big[\overline{l(\boldsymbol{f})} - \boldsymbol{\delta}_0\big]\big[\overline{l(\boldsymbol{f})} - \boldsymbol{\delta}_0\big]^{\intercal}, \end{aligned} \tag{18}$$

where $(\kappa_0, \boldsymbol{\delta}_0, \nu_0, \boldsymbol{\Delta}_0)$ are prior coefficients that represent the clinician's knowledge before any patients were observed. As the number n_{H} of observed healthy patients increases, the probability for the transformed data $l(\boldsymbol{f})$ tends to a normal distribution with mean and covariance matrix equal to the empirical average and covariance matrix of the transformed data. The formulae above show that previous data enter only through the sufficient statistics $\overline{l(\boldsymbol{f})}$ and $\mathrm{Cov}[l(\boldsymbol{f})]$.

An analogous formula holds for the likelihood for the patient's being schizophrenic, with coefficients $(\kappa_{\mathrm{S}}, \boldsymbol{\delta}_{\mathrm{S}}, \nu_{\mathrm{S}}, \boldsymbol{\Delta}_{\mathrm{S}})$ that depend on the data of previous schizophrenic patients and some initial coefficients. The function l and the prior coefficients $(\kappa_0, \boldsymbol{\delta}_0, \nu_0, \boldsymbol{\Delta}_0)$ could be different for the healthy and schizophrenic cases, but for simplicity we assume them identical for both health conditions.

If $(\hat{p}_{\mathrm{H}}, \hat{p}_{\mathrm{S}})$ is the pre-test probability distribution for the health condition of patient 0, his post-test probability to be healthy is

$$\mathrm{P}(\text{healthy} \mid \text{ fMRI result } \wedge \text{ prior info}) =$$

$$\frac{\mathrm{t}\Big[l(\boldsymbol{f}) \big|\ \nu_{\mathrm{H}} - d + 1, \boldsymbol{\delta}_{\mathrm{H}}, \frac{\kappa_{\mathrm{H}}+1}{\kappa_{\mathrm{H}}\,(\nu_{\mathrm{H}}-d+1)}\boldsymbol{\Delta}_{\mathrm{H}}\Big]\,\hat{p}_{\mathrm{H}}}{\mathrm{t}\Big[l(\boldsymbol{f}) \big|\ \nu_{\mathrm{H}} - d + 1, \boldsymbol{\delta}_{\mathrm{H}}, \frac{\kappa_{\mathrm{H}}+1}{\kappa_{\mathrm{H}}\,(\nu_{\mathrm{H}}-d+1)}\boldsymbol{\Delta}_{\mathrm{H}}\Big]\,\hat{p}_{\mathrm{H}} + \mathrm{t}\Big[l(\boldsymbol{f}) \big|\ \nu_{\mathrm{S}} - d + 1, \boldsymbol{\delta}_{\mathrm{S}}, \frac{\kappa_{\mathrm{S}}+1}{\kappa_{\mathrm{S}}\,(\nu_{\mathrm{S}}-d+1)}\boldsymbol{\Delta}_{\mathrm{S}}\Big]\,\hat{p}_{\mathrm{S}}}. \tag{19}$$

Note that the Jacobian determinants l' do not appear in this formula.

2.4.4 Generalized normal models in our study

In the rest of our study we compare three different transformations l of the connectivity weights $\boldsymbol{f}$, with one set of prior coefficients $(\kappa_0, \boldsymbol{\delta}_0, \nu_0, \boldsymbol{\Delta}_0)$ each:

Logit-normal model: A slightly modified logit transformation

$$l(f_k) \coloneqq \ln\frac{1+f_k}{1-f_k}, \qquad l'(f_k) = \frac{2}{1-f_k^2}, \tag{20}$$

with prior coefficients

$$\kappa_0 = 1, \quad \boldsymbol{\delta}_0 = 0, \quad \nu_0 = d+1, \quad \boldsymbol{\Delta}_0 = (d+2)\boldsymbol{I}_d. \tag{21}$$

Tangent-normal model: A tangent transformation

$$l(f_k) \coloneqq \tan\frac{\pi f_k}{2}, \qquad l'(f_k) = \frac{\pi}{1+\cos\pi f_k}, \tag{22}$$

with prior coefficients

$$\kappa_0 = 1, \quad \boldsymbol{\delta}_0 = 0, \quad \nu_0 = d+1, \quad \boldsymbol{\Delta}_0 = \frac{d+2}{4}\boldsymbol{I}_d. \tag{23}$$

Normal model: An identity transformation (that is, no transformation at all)

$$l(f_k) \coloneqq f_k, \qquad l'(f_k) = 1, \tag{24}$$

with prior coefficients

$$\kappa_0 = 1, \quad \boldsymbol{\delta}_0 = 0, \quad \nu_0 = d+1, \quad \boldsymbol{\Delta}_0 = 10\boldsymbol{I}_d. \tag{25}$$

For brevity we shall denote $l(\boldsymbol{f}) \coloneqq \big(l(f_i)\big)$.

The first two transformations, plotted in the upper panel of fig. 2, map the bounded domain $]-1, 1[$ of the connectivity weights into the reals, and thus restrict the generalized-normal likelihood to meaningful values of the connectivity weights. The last model instead allows for connectivity weights outside their meaningful bounds. It can be conceived as the model of a person who has no precise knowledge of what

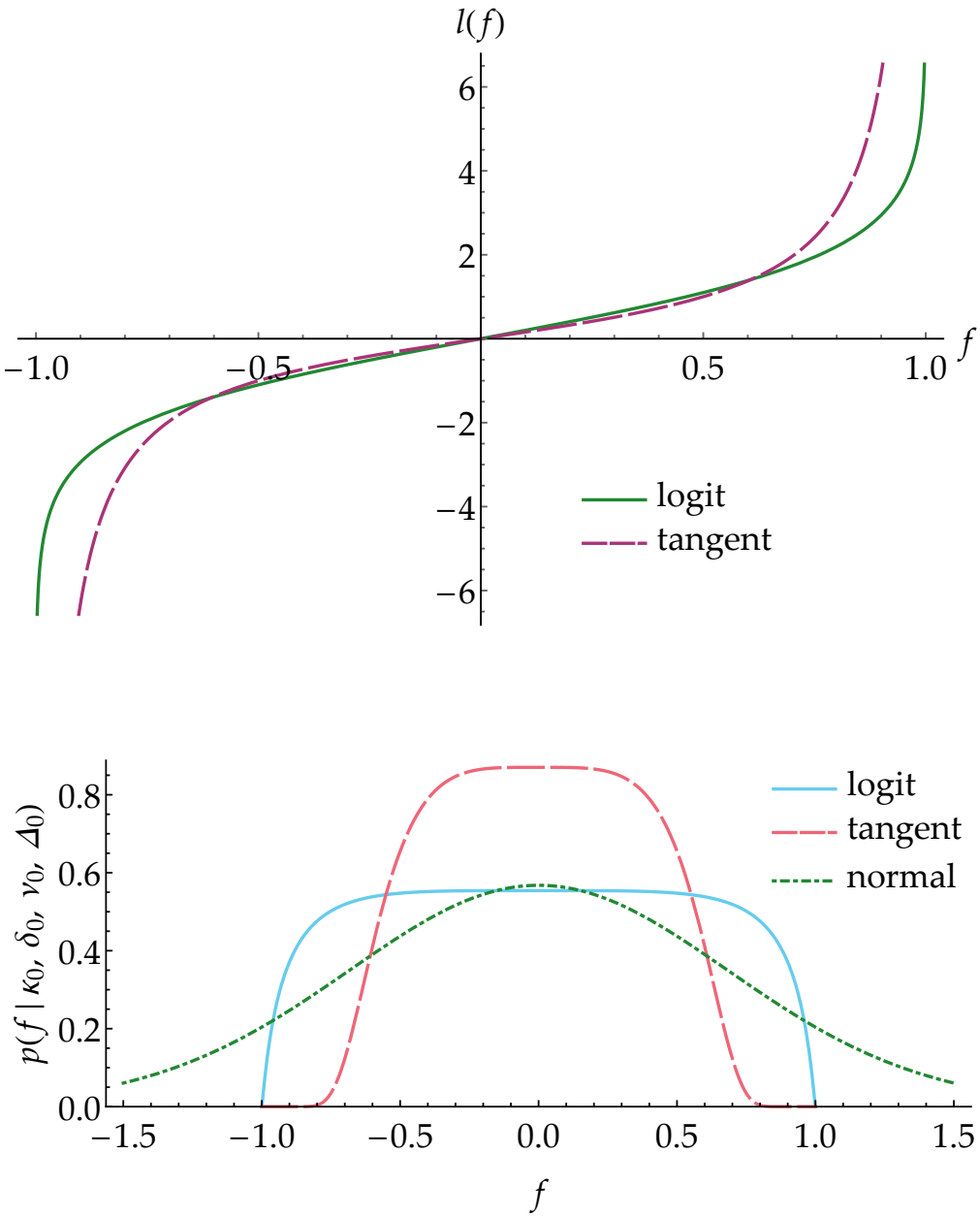

Figure 2. General normalized models. Upper panel: Generalized logit and tangent transformation functions as defined in the text. Lower panel: Prior distributions for any connectivity weight f conditional on the three generalized normal models defined in the text

the quantities $\boldsymbol{f}$ are. Since the clinician's final predictions concern health conditions given data $\boldsymbol{f}$, not the data $\boldsymbol{f}$ themselves, this model can still be meaningfully used. The probabilities for the connectivity weight f_i conditional on the prior coefficients above are shown in the lower panel of fig. 2.

The prior coefficients are chosen by the following criteria: uniform marginal distributions of the correlations between connectivity weights (leading to $\nu_0 = d + 1$ as explained before); large uncertainty in the location parameters ($\kappa_0 = 1$); symmetry with respect to the origin ($\boldsymbol{\delta}_0 = 0$); a prior distribution for the connectivity as flat as possible (its second derivative vanishes at the origin, leading to the values of $\boldsymbol{\Delta}_0$ above). In the case of the identity transformation we have chosen a $\boldsymbol{\Delta}_0$ that somewhat concentrates the prior around the true range of the connectivity weights, $[-1, 1]$.

The numerical values of the main quantities used throughout this study are summarized in table 1.

$n_{\mathrm{H}} = 55$	healthy patients
$n_{\mathrm{S}} = 49$	schizophrenic patients
$d = 40$	graph parameters
$\kappa_0, \boldsymbol{\delta}_0, \nu_0, \boldsymbol{\Delta}_0$	prior coefficients: see eqs (21), (23), (25)

Table 1. Numerical values in our study

2.5 Model comparison and selection

2.5.1 Criteria for model comparison

Any two statistical models differ in two main characteristics: their predictive power and their learning speed. Predictive power is a model's capacity to give high probability to propositions that turn out to be true, during and especially after its learning phase (cf. Dawid, 1982b). Learning speed is how quickly a model reaches unchanging, stable predictive probabilities as it gets updated with new data; note, however, that a model may also never reach stable probabilities (see e.g. Bruno, 1964; Berk, 1966); "Alas, this seems like a model of the way things work in practical inference – as more data comes in, one admits a richer and richer variety of explanatory hypothesis" (Diaconis, 1988, § 3, p. 113). Thus, our choice of a model depends on the relative importance we give to these two characteristics.

These two characteristics need not go hand in hand: a model can quickly learn with very little data but settle on probabilities with poor predictive power; conversely, it can reach great predictive power but only after a long learning phase with a huge amount of data. A model can also "unlearn", i.e. its predictive power initially increases and then decreases before stabilizing. The latter phenomenon can happen because every model initially makes its prediction through a mixture of likelihoods – the integral (11), in our case the t distribution (36) – but as the training continues the mixture is replaced by a single likelihood, in our case a generalized normal (14), as explained in § 2.4.3. It may happen that the mixture of likelihoods has higher predictive power than a single likelihood, and in this case a decline in predictive power will be observed. A model with such behaviour is obviously unfit for the clinician's goal; this also means that the sufficient statistics on which it is based capture very poorly the differences in connectivity weights between health conditions.

When only a small amount of training data is available, it can be difficult to assess which of two models has or will have the greater predictive power. The first model can initially reach a greater predictive power than the second, but the second model may eventually reach greater predictive power than the first, with further training data. Models having the same likelihood, however, have the same final predictive power; their learning speeds depend on their parameter priors.

In a diagnostic problem like that faced by our clinician, the choice of a model is ultimately dictated by the predictive power of the post-test probabilities given by the model; but if we have little training data, the learning speed of the model is also of some importance. Several quantitative criteria can be conceived to assess these two characteristics:

1. the post-test probabilities the model gives to the correct health conditions for all training data, i.e. $\mathrm{P}(c_1, c_2, \dots)$;

2. the post-test probabilities the model gives to the correct health conditions in the final phase of the training only, i.e. $\mathrm{P}(c_{\text{last}} \mid c_1, c_2, \dots)$;
3. the expected utility the model yields for all training data;
4. the expected utility the model yields in the final phase of the training only.

Post-test probabilities (criteria 1, 2) are important for obvious reasons, and utilities (criteria 3, 4) are important because the clinician's overall problem is one of decision, as emphasized in the Introduction. Consideration of all data (criteria 1, 3) is important if we are interested in the performance of the model for the whole set of patients; but consideration of the final data only, conditional on the previous ones (criteria 2, 4), tells us how much the model has learned (compare with a similar remark in model comparison using Bayes factors by Berger and Pericchi, 1996).

We must keep in mind that these criteria assess a statistical model not by itself but in combination with other factors, since they also depend on pre-test probabilities (which can be influenced by other diagnostic tests) or utilities.

Applied to our models, each of these four criteria gives a very similar picture. We shall calculate the results for criteria 2 and 4, the latter with two different utility tables. This calculation can be explained in very intuitive terms:

Imagine that the $n_{\text{H}} = 55$ healthy and $n_{\text{S}} = 49$ schizophrenic patients visit the clinician in turn, in an unknown order. For each patient, let us further assume that the clinician has pre-test probabilities $(\hat{p}_{\text{H}}, \hat{p}_{\text{S}}) = (0.5, 0.5)$, i.e. she is completely uncertain about the patient's health condition. The incidence in the population is much lower than 50%, of course, but the patients presenting themselves for diagnosis are not representative of the full population.

As stated in § 2.2, pre-test probabilities represent the clinician's uncertainty before the fMRI test is made; they can be based for example on a first diagnosis considering symptoms and medical history of the patient and of the patient's family, on psychological evaluations, and on other diagnostic tests. Here we assume complete uncertainty for demonstration purposes.

The clinician acquires the fMRI result $\boldsymbol{f}$ for that patient, and uses the statistical model, trained with the data from all the patients that previously visited her, to update to a post-test probability for schizophrenia p_{S}, given by eq. (19); obviously $p_{\text{H}} = 1 - p_{\text{S}}$. Now the clinician must make a decision – say, treat or dismiss – based on the expected utilities of the decisions available. Each decision has a different utility depending on the patient's true health condition, as summarized by a table. We consider two tables: a symmetric one

(symmetric)	healthy	schizophrenic
dismiss	1	0
treat	0	1

(26)

and an asymmetric one

(asymmetric)	healthy	schizophrenic
dismiss	1	−2
treat	−1	2

(27)

In order to maximize the expected utility, as explained in § 4.3, the clinician dismisses the patient if $p_{\text{S}} < 1/2$ in the case of the symmetric utility table, and if $p_{\text{S}} < 1/3$ in the case of the asymmetric one; and

treats the patient otherwise. After the clinician's decision is made, the patient's true health condition is revealed and we record the actual utility gained for each table, and the post-test probability the clinician assigned to the true health condition, or its logarithm, usually called "negative surprise" (Bartlett, 1952; Good, 1956, 1957a). The health condition and fMRI data of this patient are used to update the model. The next patient is received, and so on, until all patients have been examined.

The particular sequence of utilities and log-probabilities recorded in the manner just described depends on the exact order by which the 104 patients visit the clinician. We take an approximate average over all possible $104! \approx 10^{166}$ orders by randomly sampling $520\,000$ of them. The values of these averages for the final patient constitute the quantitative criteria 2 and 4.

With the symmetric utility table (26), the average utility is also the average number of schizophrenic patients for which the model yields $p_{\mathrm{S}} > 1/2$. The average utility for the last patient, when the model has been trained with the rest of the patients, is therefore a form of leave-one-out cross-validation (Allen, 1974; Stone, 1974; Kotz et al., 2006, vol. 2, pp. 1454–1458). The asymmetric table (27), slightly more realistic, tells us that dismissing a schizophrenic patient has worse consequences than treating a healthy one, and treating a schizophrenic patient has better consequences than dismissing a healthy one (McKenzie, 2014; Ho et al., 2000). For this reason the patient is dismissed, more conservatively, only if $p_{\mathrm{S}} < 1/3$. Note that scaling a utility table by a positive factor or shifting its values by a constant represent changes in the unit of measure and in the zero of utilities, and therefore do not affect our relative comparison of the statistical models.

2.5.2 Results for our three models

The averaged sequences of utilities and log-probabilities calculated as in § 2.5.1 are shown in fig. 3. The R code for the calculation is publicly available (Porta Mana et al., 2018). The results, summarized in table 2, are qualitatively identical by the two criteria and two utility tables we chose: the normal model gives the best values for the final patient, followed by the logit-normal model; the tangent-normal model has the worst final predictive power, at or below chance level.

	symmetric utility (26)	asymmetric utility (27)	log-probability
normal model	0.80	0.70	−0.29
logit-normal model	0.75	0.66	−0.34
tangent-normal model	0.56	0.03	−0.60
chance	0.50	0.41	−0.69

Table 2. Final results for the three models

The plots of fig. 3 illustrate the points made at the beginning of § 2.5.1: one model can initially learn faster than another and yet be overtaken in the later stages of learning; this is the case for the logit-normal and normal models. The tangent-normal model shows strong unlearning; this means that a tangent-normal likelihood and its sufficient statistics do not distinguish well between healthy and schizophrenic conditions. The slight downward bends at the final stages of the logit-normal and normal models raise the suspicion that they might also show some unlearning if further training data were supplied.

The trends of the logit-normal and normal models suggest that the learning phase is not finished: more patients are needed before their predictive probabilities become stable. This is also evident from the updated

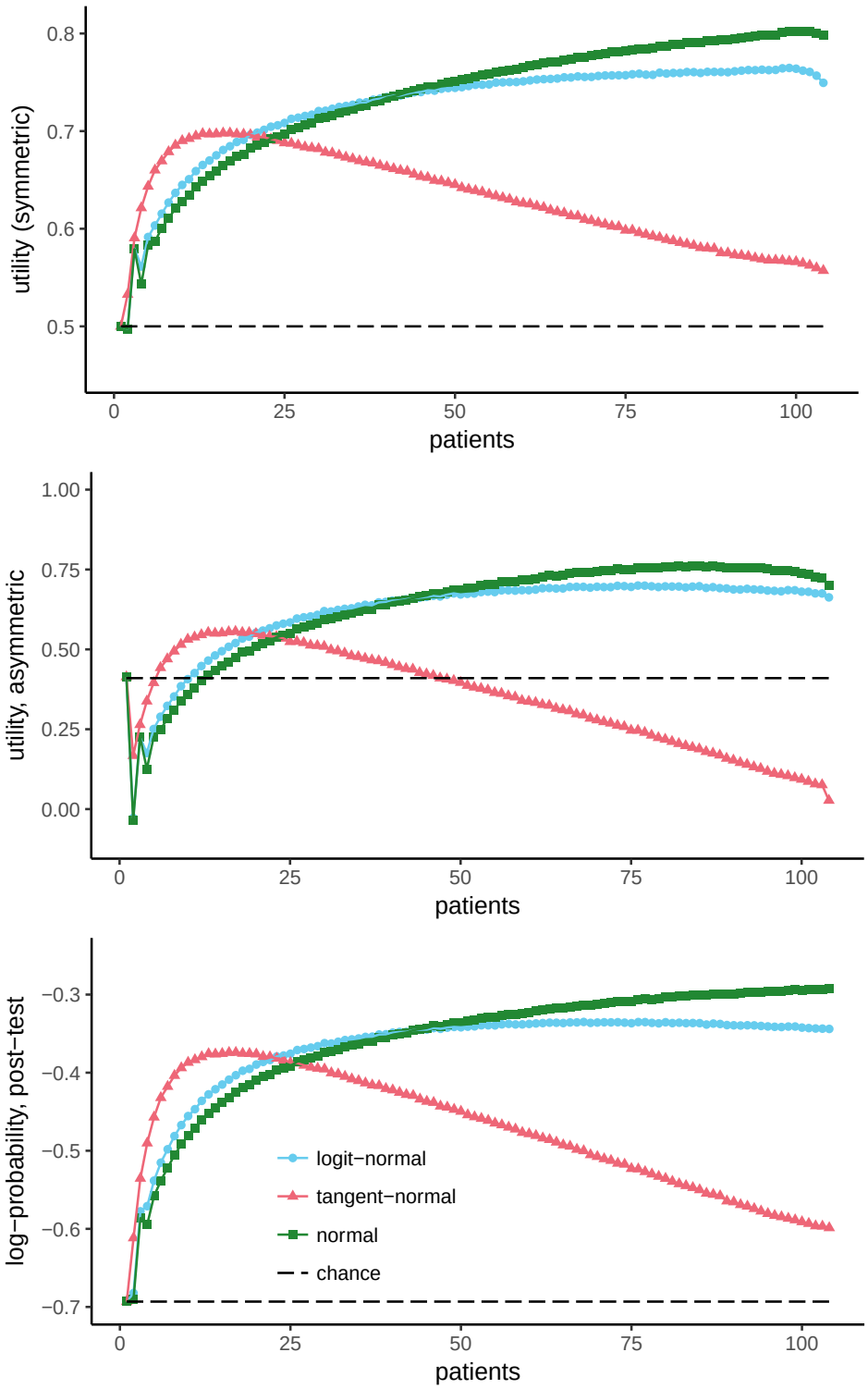

Figure 3. Averaged sequence of utilities, with utility tables (26) and (27), and of log-probabilities for our three models. The standard deviations of the averages are smaller than the markers' size. The average values for the first and last patients are exact.

marginal distributions of their parameters $(\boldsymbol{\lambda}_{\mathrm{H}}, \boldsymbol{\Lambda}_{\mathrm{H}}; \boldsymbol{\lambda}_{\mathrm{S}}, \boldsymbol{\Lambda}_{\mathrm{S}})$, for example those for the connectivity weight f between the left superior parietal lobule and left lingual gyrus, shown in fig. 4 for the logit-normal model. The distributions of the location parameters $(\boldsymbol{\lambda}_{\mathrm{H}}, \boldsymbol{\lambda}_{\mathrm{S}})$ have reached the empirical means of the data, but

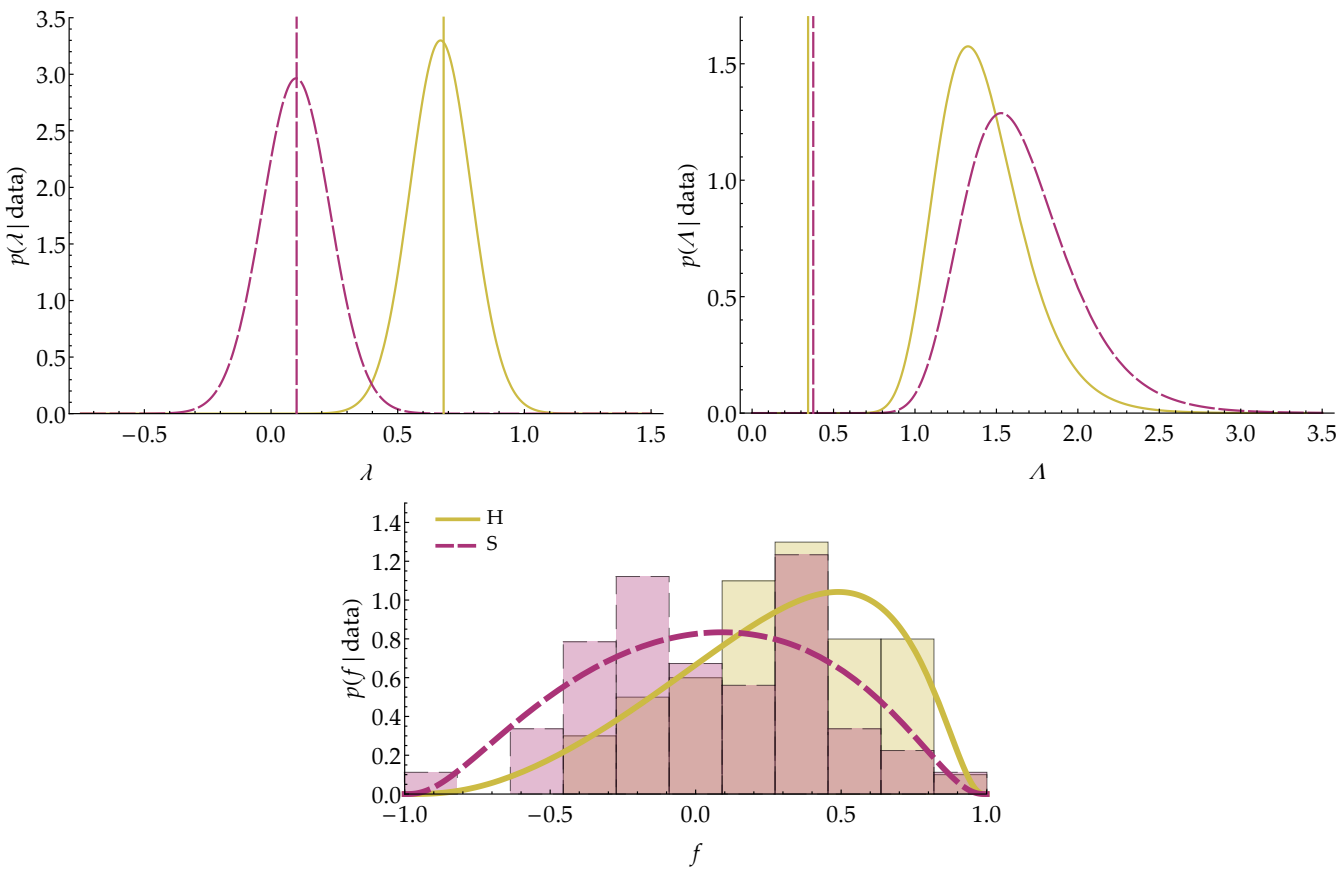

Figure 4. Updated distributions of the logit-normal model for the location parameters $(\boldsymbol{\lambda}_{\mathrm{H}}, \boldsymbol{\lambda}_{\mathrm{S}})$ (left), scale parameters $(\boldsymbol{\Lambda}_{\mathrm{H}}, \boldsymbol{\Lambda}_{\mathrm{S}})$ (right), and connectivity weights f_{H}, f_{S} (bottom, superposed on the empirical distributions) for the connectivity between left superior parietal lobule and left lingual gyrus, corresponding to the bottom left panel of fig. 1. The vertical lines in the first two plots indicate the corresponding empirical statistics from the data.

those of the scale parameters $(\boldsymbol{\Lambda}_{\mathrm{H}}, \boldsymbol{\Lambda}_{\mathrm{S}})$ are still very far away from the empirical variances. The reason is that the prior for the scale parameters had a peak at a very large value of $\Lambda \approx 20$. The 55 data for healthy patients and 49 for schizophrenic ones have shifted this peak to $\Lambda_{\mathrm{H}} = 1.3$ and $\Lambda_{\mathrm{S}} = 1.5$, but more data are needed to shift these peaks to even smaller values – provided that in the meantime the empirical values do not change too much as new data are gathered.

The peak at high values of $\boldsymbol{\Lambda}$ is a known inconvenient feature of the normal-inverse-Wishart conjugate prior, related to the correlation between correlation and variance components of $\boldsymbol{\Lambda}$ characteristic of this prior (e.g. Barnard et al., 2000).

2.5.3 Contrast with other model-comparison criteria

Common Bayesian model-comparison criteria are based on the joint probability that the model gives to training data; especially its logarithm, called "weight of evidence" or "marginal log-likelihood", or the ratio of such logarithms, called Bayes factors (Jeffreys, 2003, chs V, VI, A; Good, 1950; MacKay, 1992a; Kass and Raftery, 1995). The simple reason is Bayes' theorem:

$$\mathrm{P}(\text{model} \mid \text{ data } \wedge \text{ prior info}) \propto \mathrm{P}(\text{data} \mid \text{ model } \wedge \text{ prior info}) \times \mathrm{P}(\text{model} \mid \text{ prior info}), \qquad (28)$$

the latter probability usually assumed the same for all models (but see Porta Mana, 2017). A higher weight of evidence means that the model is more probable.

In our study, however, we have two kinds of data: health conditions and fMRI results. Since the likelihood for the health condition, used by the clinician to arrive at a post-test probability, gives the probability for the fMRI results $\{\boldsymbol{f}_i\}$, it seems intuitive to calculate the weights of evidence of our models based on these data. The result is the opposite of what we obtain with the averaged-utility criterion or any of other three mentioned above. We obtain:

$$\begin{aligned}
&\ln \mathrm{p}(\text{fMRI results} \mid \text{logit-normal model} \wedge \text{health conditions}) = -1913,\\
&\ln \mathrm{p}(\text{fMRI results} \mid \text{tangent-normal model} \wedge \text{health conditions}) = -1858,\\
&\ln \mathrm{p}(\text{fMRI results} \mid \text{normal model} \wedge \text{health conditions}) = -2488,
\end{aligned} \tag{29}$$

which gives the normal model a much smaller probability than the other two, and the tangent-normal model the highest.

This discrepancy with the averaged-utility criterion is not completely surprising, though. Imagine a disease that leads to no differences at all between the fMRI results of patients with the disease and those of healthy controls. If we found a statistical model that predicted the fMRI results with certainty, this model would thus have a the highest weight of evidence (zero), and yet its final average utility would be at chance level, since it could not help us at all in telling healthy from diseased patients. For our problem the right comparison and selection criterion is the utility or one of the other three criteria previously listed.

2.5.4 Final assessment of models

The average-utility criterion clearly excludes the tangent-normal model, which even shows a rapid unlearning. The logit-normal and normal models have almost similar performances, at around 75–80% of final patients correctly treated. We can also plot the sequence of utilities averaged over healthy and schizophrenic patients separately, as in fig. 5, which gives us an idea of the ratio between true and false negatives, and true and false positives. Both models show around 35% false positives (dismissed schizophrenic patients), with the normal model giving slightly higher final rates of true negatives, 93%, and true positives, 65%, than the normal, 85% and 63%.

However, we emphasize that the features used as inputs to the model were selected using a very simple heuristic (maximum difference in means between the two groups, see § 2.3), and the statistical model was selected for its computational properties rather than its fit to the distributions of connection weights derived from fMRI data. This notwithstanding, we believe that with further training, the logit-normal model could reach a higher predictive power. The reason is that some empirical distributions of connectivity weights, like the one for schizophrenic patients shown in red in fig. 4, seem to be bimodal; and the logit-normal likelihood, unlike the normal one, is capable of bimodality, thus fitting these distributions better.

Our assessment, however, is just an illustrative example for the general method discussed in this paper, and we are not earnestly proposing the logit-normal model (nor any other specific model) as the optimal one to use in the problem of diagnosing schizophrenia. We note that any model assessment and selection using the average-utility metric depends on several important quantities and assumptions:

A. the pre-test probabilities given by the clinician; we assumed these to be $(0.5, 0.5)$;
B. the clinician's range of decisions; we assumed it to be simply "treat or dismiss", but it could comprise several different kinds of treatments;

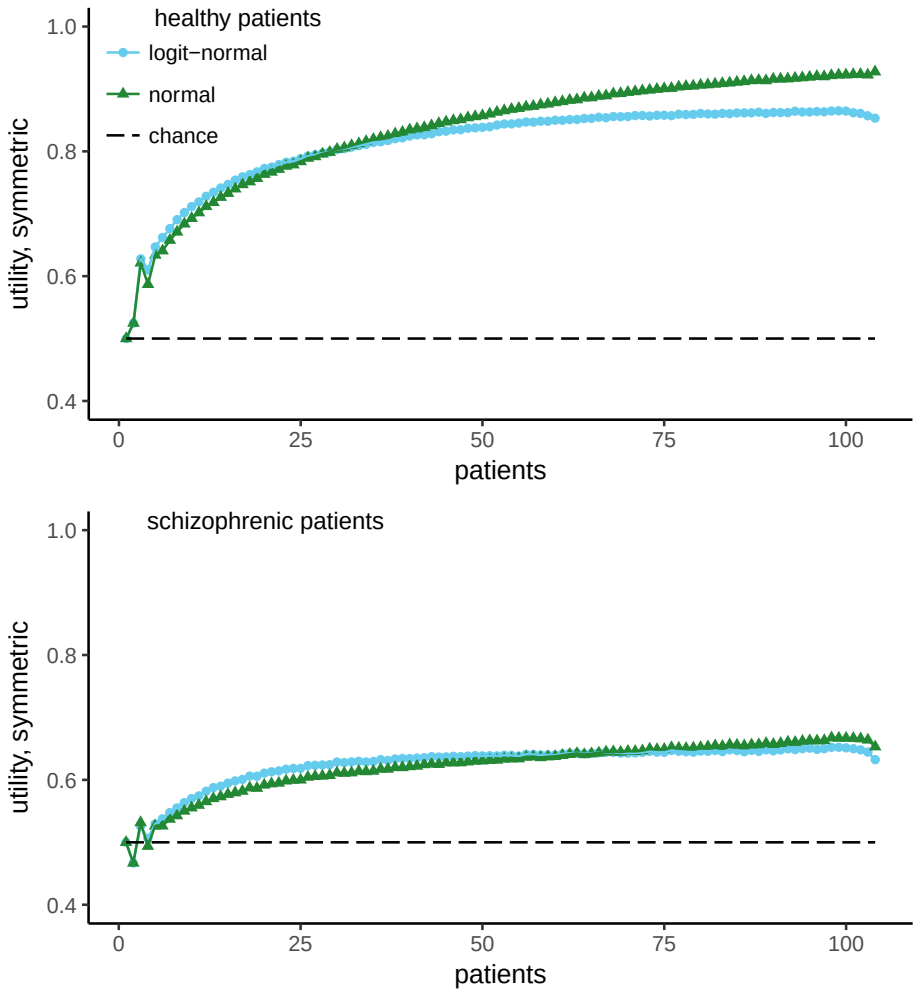

Figure 5. Sequence utilities averaged over healthy (left) and schizophrenic (right) patients separately, for the logit-normal and normal models

C. the utilities of the clinician's decisions; we assumed these to be as in formula (37);

D. the ratio between the numbers of healthy and schizophrenic training data; $55/49 = 1.12$ in our case;

E. the clinician diagnoses one patient at a time.

For a proper model assessment we should therefore investigate and consider more realistic rates of healthy vs schizophrenic cases that visit a particular clinician, in order to have better-informed pre-test probabilities; and we should consider more realistic decisions available to the clinician, together with realistically quantified utilities.

Assumption E. deserves some explanation as it may mistakenly appear that it doesn't matter whether patients visit the clinician simultaneously or one at a time. Suppose two patients, Tom and Joe, visit the clinician together, and the clinician obtains fMRI data for both, $\boldsymbol{f}_{\mathrm{T}}$ and $\boldsymbol{f}_{\mathrm{J}}$. The joint post-test probability for their health conditions c_{T} and c_{J} is different from the one obtained first calculating Joe's one, say, and then Tom's using Joe's results:

$$\mathrm{p}(c_{\mathrm{T}}, c_{\mathrm{J}} \mid \boldsymbol{f}_{\mathrm{T}}, \boldsymbol{f}_{\mathrm{J}}, D, M_l) \neq \mathrm{p}(c_{\mathrm{T}} \mid \boldsymbol{f}_{\mathrm{T}}, c_{\mathrm{J}}, \boldsymbol{f}_{\mathrm{J}}, D, M_l) \times \mathrm{p}(c_{\mathrm{J}} \mid \boldsymbol{f}_{\mathrm{J}}, D, M_l). \tag{30}$$

This inequality can be easily verified by applying the probability product rule to the left side, and can be understood as follows. Suppose the clinician first wants to calculate the likelihood for Joe's being healthy. If Tom is schizophrenic then his fMRI result is unimportant for Joe's likelihood, owing to our assumption of independent priors (16). But if Tom is healthy, then his fMRI result, *which is known to the clinician*, should lead to an updated model for Joe's likelihood. The likelihood for Joe's being healthy is therefore a mixture of these two possible likelihoods, with weights proportional to the post-test probability for Tom's health condition. Thus, Joe's likelihood is affected by Tom's fMRI result even if Joe's is calculated first and Tom's health condition is not yet known. More generally, if several patients visit the clinician simultaneously, she should order diagnostic fMRI tests for all of them at once and calculate a joint post-test probability for them, in order to make the best-informed prediction for each.

3 DISCUSSION

3.1 Summary

The diagnosis of a medical condition is a complex process that takes in a variety of judgements and evidence from the clinician and from any diagnostic tests available to her. Bayesian probability theory has found wider acceptance in medicine because it can consistently combine and frame such judgements and evidence (Goodman, 1999; Davidoff, 1999; Greenland, 1998). Formulae (1)–(3) show the basic scheme of how the clinicians' judgements and the results of diagnostic tests are combined (Sox et al., 2013). The role of a diagnostic test is not simply to give a dichotomic answer, e.g. healthy/ill, but a *likelihood* for each health condition, to be combined into this scheme together with the likelihoods from other tests. The final probability obtained from these likelihoods is finally used by the clinician to decide upon a course of action, e.g. dismiss/treat (§ 4.3).

The values of likelihoods from such tests need to be determined from a set of training set of data for each health condition. In this work we have discussed how to determine the likelihoods when the diagnostic test and training set consist of fMRI data, considering for concreteness the case of schizophrenia as the disease in question, and by using real fMRI data from healthy and schizophrenic patients (§ 2.1). We derived them step by step from first principles through a sequence of assumptions:

(a) **Partial exchangeability** with respect to the health conditions, explained in § 2.2. We believe this assumption to be very natural in medical diagnostics. By itself it already leads to a specific expression for the likelihood, eq. (8), although this expression is very difficult to compute.

(b) **Sufficiency** of an empirical statistics of a reduced set of fMRI data (§§ 2.3–2.4). We believe such kind of assumption to be sound and at least approximately true when neurologically motivated, and moreover it provides a bridge between biophysical considerations and the specification of probabilities. It leads to a likelihood, eq. (11), amenable to numerical or analytic computation.

(c) **Prior knowledge** of the empirical statistics for the different health conditions (§ 2.4.4). An assumption of this kind is always necessary, especially with small training data sets; but it affects our inferences less and less as our training data accumulate.

In our study, for (b) we specifically assumed the sufficiency of the first and second moments of some functions (§ 2.4) of the functional-connectivity weights obtained from the fMRI data (§ 2.3). Our focus on connectivities was neurologically motivated, but our choice of first and second moments of particular functions was made for mathematical simplicity, in order to illustrate the method. Our choice of roughly flat,

conjugate priors for assumption (c) was also made for the sake of mathematical simplicity and illustration (§ 2.4.4).

Notwithstanding our simple, mainly illustrative choices, we obtained good diagnostic performances, briefly discussed in the next section. To assess these performances and the relative predictive power of different choices in assumption (b) above, we presented several criteria, based on decision theory (§ 2.5). These criteria also help in understanding whether the training of the likelihoods has stabilized (§ 2.5). We observed that in this kind of study decision-theoretical criteria can yield results in seeming contrast with Bayesian model-comparison criteria, like weights of evidence and Bayes factors (§ 2.5.3); but this contrast is understandable and is not a sign of inconsistency.

We emphasize that the analysis and conclusions presented here are just illustrations of the general method, and are not meant to be used for real diagnoses: as explained in §§ 2.4.2 and 2.5.4, for a real application we would first need to investigate better-informed sufficient statistics, realistic decisions available to the clinician, the latter's utilities, and realistic pre-test probabilities. In this study we used simple default values for illustrative purposes only.

3.2 Comparison with other studies and methods

In this study we used a rather naive approach, explained in § 2.3, to select a subset of brain connections for our analysis. That approach can be criticized in two different ways. First, it ignores the fact that if distributions are narrow, the means can be close together without much overlap, and that such distributions are likely to be more informative for the purposes of discrimination. This could be improved, for example, by taking the minimal area of the distributions' overlap (dark red area in fig. 1) instead. Second, we did not restrict our search for suitable connections and associated brain areas to the areas that are known to be part of resting-state networks identified in previous studies, like the default mode network. Our lack of specificity, though, was motivated by previous studies which have demonstrated that functional connectivity can be also measured in other task-related networks, induced by spontaneous activity (He et al., 2009), and that in resting state different activity patterns can appear in schizophrenic patients, owing e.g. to hallucinations during the scan (e.g. Shergill et al., 2000).

Despite our simplified and possibly unrealistic choices of connectivity weights, sufficient statistics, parameter priors, pre-test probabilities, and utility functions, we obtained a diagnostic performance of 80%, measured by leave-one-out cross-validation (§ 2.5), comparable to classification results based on fMRI data reported in previous studies (e.g., Venkataraman et al., 2012, 18 healthy + 18 schizophrenic patients; Cetin et al., 2016, 45 + 46 patients; Demirci et al., 2008, 36 + 34 patients).

Our results also compare with classification results that we achieved using machine-learning methods. Cross-validation tests using support-vector machines (using 80% of the data for training and 20% for testing in a randomized iterative way) also yielded around 80% of correct diagnoses (data not shown).

But, as explained in the Introduction (points I., II.), methods like these, which simply classify or are deterministic, do not fit the clinician's decision-theoretical problem: they cannot be combined with other diagnostic tests and do not fit a general decision-theoretic approach – cf. §§ 2.5.1, 4.3. Most machine-learning methods (Bishop, 2006; Murphy, 2012) are thus ruled out.

There is no real contrast, however, between machine-learning methods and the method presented here: machine-learning algorithms can be interpreted as convenient, fast approximations of Bayesian methods (see e.g. the explicative image in Huszár, 2017), often combined with default utility functions and decision

rules (Murphy, 2012; MacKay, 2003, 1992a,e,d,c,b). A machine-learning classification algorithm that gives a good performance can suggest good likelihoods or parameter priors to be used in a statistical model. For example, the simplest version of a support-vector machine (Bishop, 2006, ch. 7; Murphy, 2012, § 14.5) can be interpreted as a model where the likelihood $L_{\mathrm{H}}(\boldsymbol{f}\,|\,\boldsymbol{\theta}_{\mathrm{H}})$ for one health condition is very large in a half of the dataspace $\boldsymbol{f} \in [-1,1]^d$ and zero in the rest, the hyperplane separating these half-spaces being determined by the parameter $\boldsymbol{\theta}_{\mathrm{H}}$. The likelihood $L_{\mathrm{S}}(\boldsymbol{f}\,|\,\boldsymbol{\theta}_{\mathrm{S}})$ for the other health condition is likewise large or zero in two half-spaces separated by a hyperplane determined by $\boldsymbol{\theta}_{\mathrm{S}}$; see eqs (11) in § 2.4.1. In this model the parameter prior $\mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}}\,|\,M, I)$ correlates the two parameters $(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}})$ in such a way that the two hyperplanes coincide and the two likelihoods have support on opposite sides. As the model is trained and the parameter prior is updated, eq. (11b), the hyperplane moves in space in a way to maximize the product of the likelihoods. This corresponds to the search of an optimal separation hyperplane by the support-vector machine.

3.3 Possible improvements

Besides using more realistic pre-test probabilities, and utility values, the method presented here could be improved in several other respects, especially with regard to assumptions (b) and (c) summarized in § 3.1.

In point (b) we assumed that particular connectivity weights are sufficient to distinguish among schizophrenic and healthy patients. These connectivities were calculated as in §§ 2.3 and 4.1. More sophisticated choices of Regions Of Interests and functional-connectivity measures (Marrelec and Fransson, 2011; Smith et al., 2011; Wang et al., 2014; Gheiratmand et al., 2017; Demirci et al., 2008) or even integration of functional and structural imaging (Michael et al., 2010) could obviously lead to an increased predictive power. A different choice of sufficient statistics could also improve the performance. With increased computational power it would even be possible to use the full set of connectivities rather than sufficient statistics thereof – so-called "nonparametric" models (cf. Zhang et al., 2014, 2016; Nielsen et al., 2016; Kook et al., 2017).

Regarding assumption (c), in § 2.4.3 we mentioned that parametric models may lead to probabilities that lack a closed form and are analytically intractable. The models we chose, with conjugate prior, have the advantage of having closed-form formulae, but they also restrict our choice of sufficient statistics and parameter priors. Higher predictive power could be achieved by using other kinds of sufficient statistics, leading to likelihoods that are not generalized normals, or by using non-conjugate priors, e.g. treating means, correlations, and variances independently, as discussed by Barnard et al. (2000). In this case numerical methods are needed, such as Markov-Chain Monte Carlo sampling and integration (MacKay, 2003, ch. IV; Murphy, 2012, chs 23–24). It must be kept in mind, though, that numerical methods can be computationally vastly more expensive than analytic ones. In preliminary studies that led to our present work we considered a couple of statistical models that require Monte Carlo sampling: a truncated normal and a product of beta distributions among them. The calculation of the relevant integrals for these models has vastly higher time costs than for the closed-form models presented here. For example, calculation of a posterior parameter distribution for the truncated-normal model required $17\ \mathrm{h}$ on a computer cluster; whereas the corresponding calculation for the models in the present work takes a fraction of a second on a laptop. To assess and select a model for our clinician to use, such integrals need to be computed over and over, as we explained in § 2.5.1. The assessment of statistical models that require Monte Carlo methods can therefore lead to months of computation. Further explorations are needed in this direction.

The comparison with support-vector machines sketched at the end of the previous section shows one important assumption of our statistical models: the independence of the prior parameter distributions for

the two health conditions: $\mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}}, \boldsymbol{\theta}_{\mathrm{S}} \mid I) = \mathrm{p}(\boldsymbol{\theta}_{\mathrm{H}} \mid I) \times \mathrm{p}(\boldsymbol{\theta}_{\mathrm{S}} \mid I)$, see eq. (16). Statistical models of which support-vector machines are approximations clearly cannot make this assumption. The performance of the model might improve by using a non-independent joint prior distribution, which allows training data for one health condition to influence the parameter distribution for the other. In the case of the generalized-normal models we examined, this can be achieved – whilst preserving their computational convenience – by taking a weighted average of several values of prior coefficients (33). This is known as a hierarchical model (Bernardo and Smith, 2000, § 4.6.5).

The possibilities for improvement listed in these last paragraphs suggest that the method we have presented here, hinging on first principles, has great potential for applications and for development in different directions; moreover this is not to the exclusion of other methods, but assimilating their principles and advantages.

4 METHODS

4.1 Data preprocessing

Preprocessing of the rfMRI images is carried out using the FMRIB Software Library tools (FSL, v5.08: Jenkinson et al., 2012; Smith et al., 2005) and consists of the following steps: removal of the first ten image volumes, leaving the remaining 130 volumes for further data processing; removing non-brain tissue (BET: Smith, 2002); motion correction (MCFLIRT: Jenkinson et al., 2002); spatial smoothing with a $6\ \mathrm{mm}$ full width at half maximum normal kernel; temporal low-pass filtering with a cut-off frequency of $0.009\ \mathrm{Hz}$; white matter and cerebrospinal fluid regression (FSL regfilt, MELODIC).

For each subject we first linearly register the rfMRI image first to the structural, skull-removed image (image segmentation for skull removing with SPM8, Wellcome Department of Cognitive Neurology, London, UKFSL; linear registration with FSL/FLIRT: Jenkinson and Smith 2001; Jenkinson et al. 2002) and then, through a non-linear mapping, to the MNI standard brain (non-linear registration with Advanced Normalization Tools (ANTS: Avants et al., 2011); MNI 152 standard brain, non-linear 6th generation (Grabner et al., 2006). Regions of interest (ROIs) of the resulting functional image in standard space are extracted such that they match the 94 regions identified by the Oxford lateral cortical atlas with a probability above 50% (Desikan et al., 2006). The temporal mean signals across the voxels in each ROI are used to calculate the functional connectivity measured based on the Pearson correlation coefficient.

4.2 The normal model with conjugate prior

This statistical model, denoted in this section by M, is amply discussed in the literature (Gelman et al., 2014, § 3.6; Minka, 2001; Murphy, 2007); here we only give a summary.

Its likelihood is a normal distribution

$$L(\boldsymbol{f} \mid \boldsymbol{\lambda}, \boldsymbol{\Lambda}, M, I) = \mathrm{N}[l(\boldsymbol{f}) \mid \boldsymbol{\lambda}, \boldsymbol{\Lambda}]\, l'(\boldsymbol{f}) \tag{31}$$

with mean $\boldsymbol{\lambda}$ and covariance matrix $\boldsymbol{\Lambda}$.

The prior distribution for the parameters $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is a normal-inverse-Wishart distribution, i.e. the product of a normal distribution for $\boldsymbol{\lambda}$ and an inverse-Wishart matrix distribution (Gupta and Nagar, 2000, § 3.4;

Tiao and Zellner, 1964; Bernardo and Smith, 2000, § 3.2.5) for $\boldsymbol{\Lambda}$:

$$\mathrm{p}(\boldsymbol{\lambda}, \boldsymbol{\Lambda} \mid \kappa_0, \boldsymbol{\delta}_0, \nu_0, \boldsymbol{\Delta}_0, M_l, I) = \mathrm{p}(\boldsymbol{\lambda} \mid \boldsymbol{\Lambda}, \kappa_0, \boldsymbol{\delta}_0, M_l, I) \times \mathrm{p}(\boldsymbol{\Lambda} \mid \nu_0, \boldsymbol{\Delta}_0, M_l, I), \tag{32}$$

with

$$\begin{aligned} \mathrm{p}(\boldsymbol{\lambda} \mid \boldsymbol{\Lambda}, \kappa_0, \boldsymbol{\delta}_0, M_l, I) &= \mathrm{N}(\boldsymbol{\lambda} \mid \boldsymbol{\delta}_0, \boldsymbol{\Lambda}/\kappa_0), \\ \mathrm{p}(\boldsymbol{\Lambda} \mid \nu_0, \boldsymbol{\Delta}_0, M_l, I) &= \mathrm{Wishart}^{-1}(\boldsymbol{\Lambda} \mid \nu_0, \boldsymbol{\Delta}_0) \propto \det(\boldsymbol{\Lambda})^{-\frac{\nu_0+d+1}{2}} \exp\bigl(-\tfrac{1}{2}\operatorname{tr} \boldsymbol{\Delta}_0 \boldsymbol{\Lambda}^{-1}\bigr). \end{aligned} \tag{33}$$

It should be noted how $\boldsymbol{\Lambda}$ appears as parameter in the distribution for $\boldsymbol{\lambda}$, so their distributions are not independent. The composite distribution depends on two scalar, one vector, and one matrix coefficients $(\kappa_0, \boldsymbol{\delta}_0, \nu_0, \boldsymbol{\Delta}_0)$.

This prior parameter distribution retains the same form when it is conditioned on the data $(\boldsymbol{f}_i)$ of n patients, becoming a posterior parameter distribution with updated coefficients $(\kappa, \boldsymbol{\delta}, \nu, \boldsymbol{\Delta})$ depending on the prior ones and on the sufficient statistics:

$$\begin{aligned} \kappa &= \kappa_0 + n, & \nu &= \nu_0 + n, \\ \boldsymbol{\delta} &= \frac{\kappa_0\, \boldsymbol{\delta}_0 + n\, \bar{\boldsymbol{f}}}{\kappa_0 + n}, & \boldsymbol{\Delta} &= \boldsymbol{\Delta}_0 + n\, \mathrm{Cov}(\boldsymbol{f}) + \frac{\kappa_0\, n}{\kappa_0 + n} (\bar{\boldsymbol{f}} - \boldsymbol{\delta}_0)(\bar{\boldsymbol{f}} - \boldsymbol{\delta}_0)^{\intercal}. \end{aligned} \tag{34}$$

The main features of the normal-inverse-Wishart distribution for $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ are these:

$$\begin{aligned} &\boldsymbol{\lambda}: \quad \text{mean \& mode} = \boldsymbol{\delta}, \qquad \text{covariances} = \frac{1}{\kappa\,(\nu - d - 1)} \boldsymbol{\Delta}; \\ &\boldsymbol{\Lambda}: \quad \begin{cases} \text{mean} = \dfrac{1}{\nu - d - 1} \boldsymbol{\Delta}, \qquad \text{mode} = \dfrac{1}{\nu + d + 2} \boldsymbol{\Delta}, \\ \text{diagonal variances} = \dfrac{2}{(\nu - d - 3)(\nu - d - 1)^2} (\boldsymbol{\Delta})_{kk}{}^2. \end{cases} \end{aligned} \tag{35}$$

These formulae above say that the uncertainty in the location parameter $\boldsymbol{\lambda}$ decreases as κ and ν increase for fixed $\boldsymbol{\Delta}$, and the uncertainty in the matrix scale parameter $\boldsymbol{\Lambda}$ decreases with increasing ν. When $\nu = d + 1$ the marginal distributions for the correlations of $\boldsymbol{\lambda}$ are uniform (Gelman et al., 2014, § 3.6; Barnard et al., 2000, § 2.2). Because of these properties, a "vaguely informative" parameter distribution should have small κ_0 and ν_0 (Minka, 2001; Murphy, 2007).

When the likelihood (31) and the parameter prior (32), updated with (34), are multiplied and the parameters are integrated, the resulting distribution for $\boldsymbol{f}$ is a multivariate t distribution (Kotz and Nadarajah, 2004; Minka, 2001; Murphy, 2007)

$$\mathrm{p}[\boldsymbol{f} \mid (\boldsymbol{f}_i), \kappa_0, \boldsymbol{\delta}_0, \nu_0, \boldsymbol{\Delta}_0, M_l, I] \equiv \mathrm{p}(\boldsymbol{f} \mid \kappa, \boldsymbol{\delta}, \nu, \boldsymbol{\Delta}, M_l, I) = \mathrm{t}\Bigl[l(\boldsymbol{f}) \bigm| \nu - d + 1, \boldsymbol{\delta}, \tfrac{\kappa+1}{\kappa\,(\nu-d+1)} \boldsymbol{\Delta}\Bigr] \tag{36}$$

with $\nu - d + 1$ degrees of freedom, mean $\boldsymbol{\delta}$, scale matrix $\frac{\kappa+1}{\kappa\,(\nu-d+1)} \boldsymbol{\Delta}$, and covariance matrix $\frac{\kappa+1}{\kappa\,(\nu-d-1)} \boldsymbol{\Delta}$.

4.3 Decision theory and utility

Once we have the post-test probabilities $(p_{\mathrm{H}}, p_{\mathrm{S}})$ for the possible health conditions of a patient given the fMRI data, there remains to decide upon a course of action. This is the domain of decision theory (Raiffa and Schlaifer, 2000; Jaynes, 2003, chs 13, 14; Sox et al., 2013, chs 6, 7).

Suppose we have only two courses of action: treat T or dismiss D. A decision-theoretical analysis needs, besides the probabilities for the health conditions, also the utilities (or costs) of choosing an action given the patient's true health condition. For example, treatment of a healthy patient could harm the latter, or it could be innocuous. With two courses of action and two health conditions we have four utilities $u_{\text{decision}|\text{condition}}$:

	healthy	schizophrenic
dismiss	$u_{\mathrm{D}\mid\mathrm{H}}$	$u_{\mathrm{D}\mid\mathrm{S}}$
treat	$u_{\mathrm{T}\mid\mathrm{H}}$	$u_{\mathrm{T}\mid\mathrm{S}}$

$$u_{\mathrm{D}\mid\mathrm{H}} > u_{\mathrm{T}\mid\mathrm{H}}, \qquad u_{\mathrm{T}\mid\mathrm{S}} > u_{\mathrm{D}\mid\mathrm{S}}. \tag{37}$$

Typically $u_{\mathrm{D}\mid\mathrm{H}} > u_{\mathrm{T}\mid\mathrm{H}}$ and $u_{\mathrm{T}\mid\mathrm{S}} > u_{\mathrm{D}\mid\mathrm{S}}$, and $u_{\mathrm{D}\mid\mathrm{H}}$, $u_{\mathrm{T}\mid\mathrm{S}}$ are positive and $u_{\mathrm{T}\mid\mathrm{H}}$, $u_{\mathrm{D}\mid\mathrm{S}}$ negative if we appropriately shift the zero of our measurement units.

The expected utilities for dismissal and treatment are therefore

$$\mathrm{E}(u_{\mathrm{D}}) = u_{\mathrm{D}\mid\mathrm{H}}\, p_{\mathrm{H}} + u_{\mathrm{D}\mid\mathrm{S}}\, p_{\mathrm{S}}, \qquad \mathrm{E}(u_{\mathrm{T}}) = u_{\mathrm{T}\mid\mathrm{H}}\, p_{\mathrm{H}} + u_{\mathrm{T}\mid\mathrm{S}}\, p_{\mathrm{S}}. \tag{38}$$

Decision theory says the clinician ought to chose the action having maximum expected utility. For example, she dismisses the patient if $\mathrm{E}(u_{\mathrm{D}}) > \mathrm{E}(u_{\mathrm{T}})$, that is if

$$p_{\mathrm{S}} < \frac{u_{\mathrm{D}\mid\mathrm{H}} - u_{\mathrm{T}\mid\mathrm{H}}}{u_{\mathrm{D}\mid\mathrm{H}} - u_{\mathrm{T}\mid\mathrm{H}} + u_{\mathrm{T}\mid\mathrm{S}} - u_{\mathrm{D}\mid\mathrm{S}}}. \tag{39}$$

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

CB constructed and analysed the graph data from fMRI scans. PGLPM developed the statistical model. They both made the statistical analysis of the data from the model. The manuscript was written PGLPM, CB, AM.

FUNDING

We acknowledge partial support by the Helmholtz Alliance through the Initiative and Networking Fund of the Helmholtz Association and the Helmholtz Portfolio theme "Supercomputing and Modeling for 830 the Human Brain".

ACKNOWLEDGMENTS

PGLPM thanks Mari & Miri for continuous encouragement and affection; the kind staff at Iris, where part of this work was done; Buster Keaton and Saitama for filling life with awe and inspiration; and the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, biorXiv, Hal archives, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible. We thank Alper Yegenoglu and Jakob Jordan for support and advice.

REFERENCES

Aldous, D. J. (1985). Exchangeability and related topics. In Aldous et al. (1985). VII, 1–198. https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/papers.html

Aldous, D. J., Ibragimov, I. A., and Jacod, J. (1985). *École d'Été de Probabilités de Saint-Flour XIII – 1983*, vol. 1117 of *Lecture notes in mathematics* (Berlin: Springer). Édité par P. L. Hennequin

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* 16, 125–127

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

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *NeuroImage* 54, 2033–2044

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

Barnard, G. A., Atkinson, A. C., Chan, L. K., Dawid, A. P., Downton, F., Dickey, J., et al. (1974). Discussion [Cross-Validatory choice and assessment of statistical predictions]. *J. Roy. Stat. Soc. B* 36, 133–147. See Stone (1974)

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sinica* 10, 1281–1311

Bartlett, M. S. (1952). The statistical significance of odd bits of information. *Biometrika* 39, 228–237

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.* 91, 109–122

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Stat.* 37, 51–58

Bernardo, J.-M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.) (1988). *Bayesian Statistics 3* (Oxford: Oxford University Press)

Bernardo, J.-M. and Smith, A. F. (2000). *Bayesian Theory*. Wiley series in probability and mathematical statistics (New York: Wiley), reprint edn. First publ. 1994

Besag, J., Bickel, P. J., Brøns, H., Fraser, D. A. S., Reid, N., Helland, I. S., et al. (2002). What is a statistical model?: Discussion and Rejoinder. *Ann. Stat.* 30, 1267–1310. http://www.stat.uchicago.edu/~pmcc/publications.html. See McCullagh (2002)

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

Browder, F. E. (ed.) (1992). *Mathematics into the Twenty-first Century: 1988 Centennial Symposium August 8–12*, vol. II of *American Mathematical Society centennial publications* (Providence, USA: American Mathematical Society)

Bruno, A. (1964). On the notion of partial exchangeability. *Giorn. Ist. Ital. Att.* 27, 174–196. Transl. in de Finetti (1972), ch. 10, pp. 229–246

Çetin, M. S., Christensen, F., Abbott, C. C., Stephen, J. M., Mayer, A. R., Cañive, J. M., et al. (2014). Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia. *NeuroImage* 97, 117–126

Cetin, M. S., Houck, J. M., Rashid, B., Agacoglu, O., Stephen, J. M., Sui, J., et al. (2016). Multimodal classification of schizophrenia patients with MEG and fMRI data using static and dynamic connectivity measures. *Front. Neurosci.* 10, 466

Cifarelli, D. M. and Regazzini, E. (1982). Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability. In Koch and Spizzichino (1982). 185–205

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

Cox, R. T. (1961). *The Algebra of Probable Inference* (Baltimore: The Johns Hopkins Press)

Cox, R. T. (1979). Of inference and inquiry, an essay in inductive logic. In Levine and Tribus (1979). 119–167

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

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

Davidoff, F. (1999). Standing statistics right side up. *Ann. Intern. Med.* 130, 1019–1021. http://www.perfendo.org/docs/bayesprobability/5.3_goodmanannintmed99all.pdf. See Goodman (1999); Sulmasy et al. (2000)

Dawid, A. P. (1982a). Intersubjective statistical models. In Koch and Spizzichino (1982). 217–232

Dawid, A. P. (1982b). The well-calibrated Bayesian. *J. Am. Stat. Assoc.* 77, 605–610

Dawid, A. P. (2013). Exchangeability and its ramifications. In Damien et al. (2013), chap. 2. 19–29

de Finetti, B. (1937). La prévision : ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* 7, 1–68. Transl. in Kyburg and Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

de Finetti, B. (1938). Sur la condition d'équivalence partielle. In *Colloque consacré à la théorie des probabilités. VI : Conceptions diverses*, eds. B. de Finetti, V. Glivenko, and G. Neymann (Paris: Hermann), no. 739 in Actualités scientifiques et industrielles. 5–18. Transl. in Jeffrey (1980), pp. 193–205, by P. Benacerraf and R. Jeffrey

de Finetti, B. (1972). *Probability, Induction and Statistics: The art of guessing* (London: Wiley)

DeGroot, M. H. (2004). *optimal statistical decisions*. Wiley classics library (New York: Wiley), reprint edn.

Demirci, O., Clark, V. P., and Calhoun, V. D. (2008). A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia. *NeuroImage* 39, 1774–1782

Denny, J. L. (1967). Sufficient conditions for a family of probabilities to be exponential. *Proc. Natl. Acad. Sci. (USA)* 57, 1184–1187

Desikan, R. S., S??gonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31, 968–980

Diaconis, P. (1988). Recent progress on de Finetti's notions of exchangeability. In Bernardo et al. (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. (1992). Sufficiency as statistical symmetry. In Browder (1992). 15–26. First publ. 1991 as technical report https://statistics.stanford.edu/research/sufficiency-statistical-symmetry

Diaconis, P. and Freedman, D. (1980). De Finetti's generalizations of exchangeability. In Jeffrey (1980). 233–249

Diaconis, P. and Freedman, D. (1981). Partial exchangeability and sufficiency. In Ghosh and Roy (1981). 205–236. http://statweb.stanford.edu/~cgates/PERSI/year.html. Also publ. 1982 as technical report https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis_freedman_PES.pdf

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Stat.* 14, 1–26

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Stat.* 7, 269–281

Dupré, M. J. and Tipler, F. J. (2009). New axioms for rigorous Bayesian probability. *Bayesian Anal.* 4, 599–606

Edgeworth, F. Y. (1898). On the representation of statistics by mathematical formulæ (Part I). *J. Roy. Stat. Soc.* 61, 670–700. See also Edgeworth (1899, 1900)

Edgeworth, F. Y. (1899). On the representation of statistics by mathematical formulæ (Part II & III). *J. Roy. Stat. Soc.* 62, 125–140, 373–385. See also Edgeworth (1898, 1900)

Edgeworth, F. Y. (1900). On the representation of statistics by mathematical formulæ (Supplement). *J. Roy. Stat. Soc.* 63, 72–81. See also Edgeworth (1898, 1899)

Ellison-Wright, I. and Bullmore, E. (2009). Meta-analysis of diffusion tensor imaging studies in schizophrenia. *Schizophrenia Research* 108, 3–10

Ellison-Wright, I. and Bullmore, E. (2010). Anatomy of bipolar disorder and schizophrenia: A meta-analysis. *Schizophrenia Research* 117, 1–12

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Stat.* 2, 615–629

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A* 222, 309–368. http://www.stats.org.uk/statistical-inference/Fisher1922.pdf

Fraser, D. A. S. (1963). On sufficiency and the exponential family. *J. Roy. Stat. Soc. B* 25, 115–123

Freedman, D. A. (1962). Invariants under mixing which generalize de Finetti's theorem. *Ann. Math. Stat.* 33, 916–923

Friston, K. J. and Frith, C. D. (1995). Schizophrenia: A disconnection syndrome? *Clinical Neuroscience* 3, 89–97

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Texts in statistical science (Boca Raton, USA: Chapman & Hall/CRC), 3 edn. First publ. 1995

Gheiratmand, M., Rish, I., Cecchi, G. A., Brown, M. R. G., Greiner, R., Polosecki, P. I., et al. (2017). Learning stable and predictive network-based patterns of schizophrenia and its clinical symptoms. *NPJ Schizophr.* 3, 22

Ghosh, J. K. and Roy, J. (eds.) (1981). *Statistics: Applications and New Directions* (Calcutta: Indian Statistical Institute)

Good, I. J. (1950). *Probability and the Weighing of Evidence* (London: Griffin)

Good, I. J. (1956). The surprise index for the multivariate normal distribution. *Ann. Math. Stat.* 27, 1130–1135. See also corrections Good (1957b)

Good, I. J. (1957a). The appropriate mathematical tools for describing and measuring uncertainty. In Good (1983), chap. 16. 173–177. First publ. 1957

Good, I. J. (1957b). Corrections to "The surprise index for the multivariate normal distribution". *Ann. Math. Stat.* 28, 1055. See Good (1956)

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

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. 2: The Bayes factor. *Ann. Intern. Med.* 130, 995–1013. http://www.perfendo.org/docs/bayesprobability/5.3_goodmanannintmed99all.pdf. See also comments and reply Davidoff (1999); Sulmasy et al. (2000)

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., and Collins, D. L. (2006). Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv* 9, 58–66

Greenland, S. (1998). Probability logic and probabilistic induction. *Epidemiology* 9, 322–332. See also Maclure (1998)

Gupta, A. K. and Nagar, D. K. (2000). *Matrix Variate Distributions*, vol. 104 of *Monographs and surveys in pure and applied mathematics* (Boca Raton, USA: Chapman & Hall/CRC)

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

Halpern, J. Y. (1999). Cox's theorem revisited. *J. Artif. Intell. Res.* 11, 429–435. See also Snow (1998)

He, Y., Wang, J., Wang, L., Chen, Z. J., Yan, C., Yang, H., et al. (2009). Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLOS ONE* 4, 1–18

Hipp, C. (1974). Sufficient statistics and exponential families. *Ann. Stat.* 2, 1283–1292

Ho, B.-C., Andreasen, N. C., Flaum, M., Nopoulos, P., and Miller, D. (2000). Untreated initial psychosis: Its relation to quality of life and symptom remission in first-episode schizophrenia. *Am. J. Psychiatry* 157, 808–815

Horvitz, E. J., Heckerman, D. E., and Langlotz, C. P. (1986). A framework for comparing alternative formalisms for plausible reasoning. *Proc. AAAI* 5, 210–214

Hu, M.-L., Zong, X.-F., Mann, J. J., Zheng, J.-J., Liao, Y.-H., Li, Z.-C., et al. (2017). A review of the functional and anatomical default mode network in schizophrenia. *Neuroscience Bulletin* 33, 73–84

Huszár, F. (2017). Everything that works works because it's Bayesian: Why deep nets generalize? http://www.inference.vc/everything-that-works-works-because-its-bayesian-2/

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science* (Cambridge: Cambridge University Press). Ed. by G. Larry Bretthorst; http://www-biba.inrialpes.fr/Jaynes/prob.html, http://omega.albany.edu:8008/JaynesBook.html, http://omega.albany.edu:8008/JaynesBookPdf.html. First publ. 1994

Jeffrey, R. C. (ed.) (1980). *Studies in inductive logic and probability. Vol. II* (Berkeley: University of California Press)

Jeffreys, H. (2003). *Theory of Probability* (London: Oxford University Press), 3 edn. First publ. 1939

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17, 825–841

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *NeuroImage* 62, 782–790

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images, medical image analysis. *Medical Image Analysis* 5, 143–156

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* 36, 149–176

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Probability and its applications (New York: Springer)

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf; https://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf

Kingman, J. F. C. (1978). Uses of exchangeability. *Ann. Prob.* 6, 183–197

Koch, G. and Spizzichino, F. (eds.) (1982). *Exchangeability in Probability and Statistics* (Amsterdam: North-Holland)

Kolmogorov, A. N. (1942). Definition of center of dispersion and measure of accuracy to form a finite number of observations [Sur l'éstimation statistique des paramètres de la loi de Gauss]. *Izv. Akad. Nauk SSSR Ser. Mat.* 6, 3–32. In Russian

Kook, J. H., Guindani, M., Zhang, L., and Vannucci, M. (2017). *NPBayes-fMRI*: Non-parametric Bayesian general linear models for single- and multi-subject fMRI data. *Stat. Biosci.* 2017, 1–19

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Trans. Am. Math. Soc.* 39, 399–409

Koopman, B. O. (1940a). The axioms and algebra of intuitive probability. *Ann. Math.* 41, 269–292

Koopman, B. O. (1940b). The bases of probability. *Bull. Am. Math. Soc.* 46, 763–774. Repr. in Kyburg and Smokler (1980), pp. 159–172

Koopman, B. O. (1941). Intuitive probabilities and sequences. *Ann. Math.* 42, 169–187

Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B., and Johnson, N. L. (eds.) (2006). *Encyclopedia of Statistical Sciences* (Hoboken, USA: Wiley), 2 edn. First publ. 1982

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications* (Cambridge: Cambridge University Press)

Kurtz, D. S. and Swartz, C. W. (2004). *Theories of Integration: The Integrals of Riemann, Lebesgue, Henstock-Kurzweil, and Mcshane*, vol. 9 of *Series in real analysis* (Singapore: World Scientific)

Kyburg, H. E., Jr. and Smokler, H. E. (eds.) (1980). *Studies in Subjective Probability* (Huntington, USA: Robert E. Krieger), 2 edn. First publ. 1964

Lamoreaux, J. and Armstrong, G. (1998). The fundamental theorem of calculus for gauge integrals. *Math. Mag.* 71, 208–212

Lauritzen, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics*, vol. 49 of *Lecture notes in statistics* (Berlin: Springer). First publ. 1982

Levine, R. D. and Tribus, M. (eds.) (1979). *The Maximum Entropy Formalism: A Conference Held at the Massachusetts Institute of Technology on May 2–4, 1978* (Cambridge, USA: MIT Press)

Lindley, D. V. (2008). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference* (Cambridge: Cambridge University Press), reprint edn. First publ. 1965

Lindley, D. V. and Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *American Statistician* 30, 112–119

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Stat. Sci.* 23, 439–464

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Comp.* 4, 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html

MacKay, D. J. C. (1992b). *Bayesian Methods for Adaptive Models*. Ph.D. thesis, California Institute of Technology, Pasadena, USA. http://www.inference.phy.cam.ac.uk/mackay/PhD.html

MacKay, D. J. C. (1992c). The evidence framework applied to classification networks. *Neural Comp.* 4, 720–736. http://www.inference.phy.cam.ac.uk/mackay/PhD.html

MacKay, D. J. C. (1992d). Information-based objective functions for active data selection. *Neural Comp.* 4, 590–604. http://www.inference.phy.cam.ac.uk/mackay/PhD.html

MacKay, D. J. C. (1992e). A practical bayesian framework for backpropagation networks. *Neural Comp.* 4, 448–472. http://www.inference.phy.cam.ac.uk/mackay/PhD.html

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms* (Cambridge: Cambridge University Press). http://www.inference.phy.cam.ac.uk/mackay/itila/. First publ. 1995

Maclure, M. (1998). How to change your mind. *Epidemiology* 9, 233. See Greenland (1998)

Marrelec, G. and Fransson, P. (2011). Assessing the influence of different ROI selection strategies on functional connectivity analyses of fMRI data acquired during steady-state conditions. *PLoS One* 6, e14788

McCullagh, P. (2002). What is a statistical model? *Ann. Stat.* 30, 1225–1267. http://www.stat.uchicago.edu/~pmcc/publications.html. See also the following discussion and rejoinder Besag et al. (2002)

McKenzie, K. J. (2014). How does untreated psychosis lead to neurological damage? *Can. J. Psychiatry* 59, 511–512

Mead, R. (1965). A generalised logit-normal distribution. *Biometrics* 21, 721–732

Michael, A. M., Baum, S. A., White, T., Demirci, O., Andreasen, N. C., Segall, J. M., et al. (2010). Does function follow form?: Methods to fuse structural and functional brain images show decreased linkage in schizophrenia. *NeuroImage* 49, 2626–2637

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

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed *federalist* papers. *J. Am. Stat. Assoc.* 58, 275–309. https://www.stat.cmu.edu/Exams/mosteller.pdf

Murphy, K. P. (2007). Conjugate bayesian analysis of the Gaussian distribution. http://thaines.com/content/misc/gaussian_conjugate_prior_cheat_sheet.pdf

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine learning series (Cambridge, USA: MIT Press)

Neyman, J. (1935). Su un teorema concernente le cosiddette statistiche sufficienti. *Giorn. Ist. Ital. Att.* VI, 320–334

Nielsen, S. F. V., Madsen, K. H., Røge, R., Schmidt, M. N., and Mørup, M. (2016). Nonparametric modeling of dynamic functional connectivity in fMRI data. arXiv:1601.00496

Paris, J. B. (2006). *The Uncertain Reasoner's Companion: A Mathematical Perspective*. No. 39 in Cambridge tracts in theoretical computer science (Cambridge: Cambridge University Press), reprint edn. Concerning Cox's theorem see also Snow (1998)

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

Pólya, G. (1949). Preliminary remarks on a logic of plausible inference. *Dialectica* 3, 28–35

Pólya, G. (1968). *Mathematics and Plausible Reasoning: Vol. II: Patterns of Plausible Inference* (Princeton: Princeton University Press), 2 edn. First publ. 1954

Porta Mana, P. G. L. (2017). Model comparison and Bayes factors: what is a model? In preparation

Porta Mana, P. G. L., Bachmann, C., and Morrison, A. (2018). Inferring health conditions from fMRI-graph data. Open Science Framework doi:10.17605/osf.io/r2huz

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

Shenton, M. E., Whitford, T. J., and Kubicki, M. (2010). Structural neuroimaging in schizophrenia from methods to insights to treatments. *Dialogues in Clinical Neuroscience* 12, 317–332

Shergill, S., Brammer, M., Williams, S., Murray, R., and McGuire, P. (2000). Mapping auditory hallucinations in schizophrenia using functional magnetic resonance imaging. *Archives of General Psychiatry* 57, 1033–1038

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., et al. (2005). Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage* 23, S208–219

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., et al. (2011). Network modelling methods for fMRI. *NeuroImage* 54, 875–891

Snow, P. (1998). On the correctness and reasonableness of Cox's theorem for finite domains. *Comput. Intell.* 14, 452–459

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

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc. B* 36, 111–133. See also discussion Barnard et al. (1974)

Sulmasy, D. P., Morgan, T., Caubet, J.-F., and Goodman, S. (2000). Toward evidence-based statistics [comments and response]. *Ann. Intern. Med.* 132, 507–508. See Goodman (1999); Davidoff (1999)

Swartz, C. (2001). *Introduction to Gauge Integrals* (Singapore: World Scientific)

Terenin, A. and Draper, D. (2017). Cox's theorem and the Jaynesian interpretation of probability. `arXiv:1507.06597`. First publ. 2015

ISO (2006). *ISO 3534-1:2006: Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization, Geneva

ISO (2009). *ISO 80000:2009: Quantities and units*. International Organization for Standardization, Geneva. First publ. 1993

Tiao, G. C. and Zellner, A. (1964). On the Bayesian estimation of multivariate regression. *J. Roy. Stat. Soc. B* 26, 277–285

Venkataraman, A., Whitford, T. J., Westin, C.-F., Golland, P., and Kubicki, M. (2012). Whole brain resting state functional connectivity abnormalities in schizophrenia. *Schizophr. Res.* 139, 7–12

Wang, H. E., Bénar, C. G., Quilichini, P. P., Friston, K. J., Jirsa, V. K., and Bernard, C. (2014). A systematic framework for functional connectivity measures. *Front. Neurosci.* 8, 405. doi:10.3389/fnins.2014.00405

Woodward, N. D., Rogers, B., and Heckers, S. (2011). Functional resting-state networks are differentially affected in schizophrenia. *Schizophrenia research* 130, 86–93

Yu, Q., Allen, E. A., Sui, J., Arbabshirani, M. R., Pearlson, G., and Calhoun, V. D. (2016). Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Current Topics in Medicinal Chemistry* 12, 2415–2425

Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., and Vannucci, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *Ann. Appl. Stat.* 10, 638–666

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage* 95, 162–175