NeuroImage 141 (2016) 378–392

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Valid population inference for information-based imaging: From the second-level *t*-test to prevalence inference

Carsten Allefeld[a,*], Kai Görgen[a,1], John-Dylan Haynes[a,b,1]

[a] Bernstein Center for Computational Neuroscience, Berlin Center of Advanced Neuroimaging, Department of Neurology, and Excellence Cluster NeuroCure, Charité – Universitätsmedizin Berlin, Germany
[b] Berlin School of Mind and Brain and Department of Psychology, Humboldt-Universität zu Berlin, Germany

ARTICLE INFO

Article history:
Received 2 December 2015
Accepted 18 July 2016
Available online 20 July 2016

Keywords:
Information-based imaging
Multivariate pattern analysis
t-Test
Population inference
Effect prevalence

ABSTRACT

In multivariate pattern analysis of neuroimaging data, 'second-level' inference is often performed by entering classification accuracies into a *t*-test vs chance level across subjects. We argue that while the random-effects analysis implemented by the *t*-test does provide population inference if applied to activation differences, it fails to do so in the case of classification accuracy or other 'information-like' measures, because the true value of such measures can never be below chance level. This constraint changes the meaning of the population-level null hypothesis being tested, which becomes equivalent to the global null hypothesis that there is no effect in any subject in the population. Consequently, rejecting it only allows to infer that there are some subjects in which there is an information effect, but not that it generalizes, rendering it effectively equivalent to fixed-effects analysis. This statement is supported by theoretical arguments as well as simulations. We review possible alternative approaches to population inference for information-based imaging, converging on the idea that it should not target the mean, but the prevalence of the effect in the population. One method to do so, 'permutation-based information prevalence inference using the minimum statistic', is described in detail and applied to empirical data.

Introduction

Since the seminal work of Haxby et al. (2001), an increasing number of neuroimaging studies have employed multivariate methods to complement the established mass-univariate approach (Friston et al., 1995) to the analysis of functional magnetic resonance imaging (fMRI) data, a field now known as multivariate pattern analysis (MVPA; Norman et al., 2006). Most MVPA studies use classification (Pereira et al., 2009) to examine activation patterns; the accuracy of a classifier in distinguishing activation patterns associated with different experimental conditions serves as a measure of multivariate effect strength. Since the target of MVPA is not a generally increased or decreased level of activation but the *information content* of activation patterns (cf. Pereira and Botvinick, 2011), it has also been characterized as information-based imaging and distinguished from traditional activation-based imaging (Kriegeskorte et al., 2006).

Many methodological aspects of MVPA have already been discussed in detail: what kind of classifier to use (Cox and Savoy, 2003; Norman et al., 2006), whether to adapt parametric multivariate statistics instead of classifiers (Allefeld and Haynes, 2014; Nili et al., 2014), how to understand searchlight-based accuracy maps (Etzel et al., 2013), or how classifier weights can be made interpretable (Haufe et al., 2014; Hoyos-Idrobo et al., 2015). By contrast, the topic of population inference based on per-subject measures of information content, i.e. the question whether an information effect observed in a sample of subjects generalizes to the population these subjects were recruited from, has not yet received sufficient attention (but see Brodersen et al., 2013).

In univariate analysis of multi-subject fMRI studies, the standard way to achieve population inference is to perform a 'second-level' null hypothesis test (Holmes and Friston, 1998). For each subject, a 'first-level' contrast (activation difference) is computed, and this contrast enters a second-level analysis, a *t*-test or an ANOVA. Specifically for a simple one-sided *t*-test vs 0, reaching statistical significance allows to infer that the experimental manipulation is associated with an increase of activation on average in the population of subjects.

* Corresponding author.
E-mail addresses: carsten.allefeld@bccn-berlin.de (C. Allefeld), haynes@bccn-berlin.de (J. Haynes).
[1] Charité-Campus Mitte, Philippstr. 13, Haus 6, Berlin 10115, Germany.

http://dx.doi.org/10.1016/j.neuroimage.2016.07.040
1053-8119/

This is interpreted in such a way that the effect is 'common' or 'stereotypical' in that population (Penny and Holmes, 2007, p. 156).

With the adoption of information-based imaging, it has become accepted practice to apply the same second-level inferential procedures to the results of first-level multivariate analyses, in particular classification accuracy (see e.g. Haxby et al., 2001, Haynes et al., 2007, Spiridon and Kanwisher, 2002): A classifier is trained on part of the data and is tested on another part, using each part for testing once (cross-validation), and the classification performance is quantified in the form of an accuracy, the fraction of correctly classified test data points. Applied for example to two different experimental conditions, if there was no multivariate difference in the data between conditions, the classifier would operate at 'chance level', i.e. it would on average achieve a classification accuracy of 50%. At the second level, accuracies from different subjects are then entered into a one-sided one-sample *t*-test vs 50%, in order to show that the ability to classify above chance and therefore the presence of an information effect is typical in the population the subjects were recruited from.

In this paper we argue that despite of the seemingly analogous statistical procedure, a *t*-test vs chance level applied to accuracies cannot provide evidence that the corresponding effect is typical in the population. In contrast to other criticisms of this use of the *t*-test (see below), in our view the problem is not so much that the estimation distribution of cross-validated accuracies is not normal or even symmetric, or that a normal distribution model is generally inadequate for a quantity bounded to an interval [0%, 100%]. Rather, the problem is that other than estimated accuracies, the *true single-subject accuracy can never be below chance level* because it measures an amount of information.[2] We will show that this restriction changes the meaning of the *t*-test: It now tests the global null hypothesis (Nichols et al., 2005) that there is *no information in any subject in the population*. As a consequence, achieving a significant test result allows us only to infer that *there are people in which there is an effect*, but not that the presence of information generalizes to the population. The argument does not only hold for classification accuracy, but also for other 'information-like' measures.

The *t*-test on accuracies has been criticized before (Brodersen et al., 2013; Stelzer et al., 2013) on the grounds that its distributional assumptions are not fulfilled for cross-validated classification accuracies. Such a distributional error invalidates the calculation of critical values for the *t*-statistic and can therefore lead to an increased rate of false positives. This problem may be solved by better distribution models (Brodersen et al., 2013) or the use of non-parametric statistics (Stelzer et al., 2013). Our criticism goes significantly beyond that: Not only is the *t*-test quantitatively wrong, but it effectively tests a null hypothesis that is qualitatively different from its use with univariate statistics, with the consequence that rejection of this null hypothesis no longer supports population inference.

Please note that our criticism pertains specifically to a second-level *t*-test applied to per-subject classification accuracies or similar measures. It does not apply to the classification *of* subjects, e.g. into different patient groups in medical applications (Sabuncu, 2014; Sabuncu and Van Leemput, 2012), or to the classification of condition-specific patterns *across* subjects (Mourao-Miranda et al., 2005). Moreover, it only concerns quantities that measure the information content of data, but not related quantities like classifier weights (Gaonkar and Davatzikos, 2013; Gaonkar et al., 2015; Wang et al., 2007, see below).

The organization of the paper is as follows: In the section The problem with the *t*-test on accuracies we detail how a second-level *t*-test achieves population inference for univariate contrasts. We then explain that MVPA measures are 'information-like' and show, both theoretically and using simulations, that for such measures the *t*-test effectively tests the global null hypothesis that there is no effect in any subject. The section An alternative: information prevalence inference reviews possible alternatives to the *t*-test on accuracies, converging on the idea that population inference for information-based imaging should target the proportion of subjects in the population with an effect. One way to implement such an 'information prevalence inference' is described in detail in the section Permutation-based information prevalence inference using the minimum statistic, and results of its application to real data are compared with those of the *t*-test. We conclude with the discussion of a number of questions surrounding the problem of population inference for information-based imaging.

The problem with the *t*-test on accuracies

Population inference in univariate fMRI analysis

To see why the *t*-test on accuracies cannot provide population inference, we briefly recapitulate how standard univariate analysis does achieve it. In a single subject, an activation difference or contrast $\Delta\beta$ is estimated based on the general linear model (GLM; Friston et al., 1995). Because it is obtained from noisy data, the estimate is itself noisy,

$$\hat{\Delta\beta} \sim \mathcal{N}(\Delta\beta, \sigma_1^2), \tag{1}$$

where σ_1^2 denotes the *estimation variance* of the contrast (cf. Fig. 1a). If several subjects are included in a study, the true activation difference $\Delta\beta$ varies across subjects:

$$\Delta\beta_k \sim \mathcal{N}(\Delta\mu, \sigma_2^2) \tag{2}$$

where $\Delta\mu$ is the average true activation difference in the population of subjects and σ_2^2 the *population variance* of the effect (Fig. 1b). The added subscript k indicates that we now consider the subject as randomly sampled from the population. The estimated contrast in several subjects therefore shows variation for two reasons — they are noisy estimates (σ_1^2), and different subjects respond differently (σ_2^2):

$$\hat{\Delta\beta}_k \sim \mathcal{N}(\Delta\mu, \sigma_1^2 + \sigma_2^2). \tag{3}$$

The symbol $\hat{\Delta\beta}_k$ indicates that this contrast is both estimated and sampled.

A one-sided *t*-test applied to the $\hat{\Delta\beta}_k$ from a sample of subjects $k = 1 \ldots N$ has the null hypothesis $\Delta\mu = 0$. If it can be rejected ($\Delta\mu > 0$), this allows us to make a statement about the population of subjects because $\Delta\mu$ is a parameter of a *population model* (Eq. (2)). And this statement concerns a typical effect because $\Delta\mu$ is the mean, median, and mode of the assumed normal distribution. This kind of test is also called *random-effects analysis* (RFX) because it treats subjects as randomly sampled from a population (Searle et al., 1992). It

[2] Note that in this paper we only discuss the standard case of MVPA where the pair of experimental conditions is the same for training and test data. In 'cross-decoding' (cf. Haynes and Rees, 2005a), where it is tested whether a classifier trained on one pair of conditions is able to extract information corresponding to another pair of conditions, below-chance true accuracies may be possible. Cross-decoding targets not just the presence of information, but also the degree to which its neurophysiological representation is invariant with respect to another experimental manipulation.