Valid and powerful statistical test for decoding accuracy—proposal of Permutation-based Information Prevalence Inference using the *i*-th order statistic

Satoshi Hirose[a,b]

[a]Center for Information and Neural Networks, National Institute of Information and Communications Technology, 1-4 Yamadaoka, Suita, Osaka 565-0871, Japan

[b]Graduate School of Medicine & Graduate School of Frontier Biosciences, Osaka University, 2-15 Yamadaoka, Suita, Osaka 565-0871, Japan

Corresponding author:
Satoshi Hirose
Center for Information and Neural Networks
National Institute of Information and Communications Technology
1-4 Yamadaoka, Suita, Osaka 565-0871, Japan
Tel: +81-80-9045-9529
Fax: +81-6-7174-8612
E-mail: satoshi.hirose@nict.go.jp

Declarations of interest:
None.

Abbreviations:

iPIPI: Permutation-based Information Prevalence Inference using the i-th order statistic

D-Acc: decoding accuracy

OS: order statistic

Abstract

In multivariate pattern analysis, such as fMRI decoding studies using pattern classification, a "second-level" group statistical test is typically performed after "first-level" decoding analyses for individual participants. Neuroscientists often test the mean decoding accuracy across participants against the chance level accuracy (e.g. Student *t*-test), to verify whether brain activation includes information about the label (i.e., cognitive content). However, because the decoding accuracy, as well as other information-like measures, can never be below the chance level, positive results for such tests only indicate that "*at least one person in the population has information about the label in the brain*." Thus, such tests cannot provide inferences about the trend of the majority. In this study, a second-level statistical test procedure is proposed and referred to as the **P**ermutation-based **I**nformation **P**revalence **I**nference using the ***i***-th order statistic (iPIPI). This procedure is robust against outliers and has high statistical power to provide an inference regarding major population trends. In iPIPI, the *i*-th worst samples decoding accuracy (i.e., the *i*-th order statistic) is compared to the null distribution estimated by the permutation tests, in order to test whether the proportion of the better-than-chance decoding accuracy in the population (information prevalence) is higher than the threshold. Thus, a significant result of the iPIPI can be interpreted as the "*majority of the population has information about the label in the brain*." By applying this approach to artificial data, the robustness of iPIPI against outliers was confirmed, along with its high statistical power. In addition, theoretical detail is provided, and the use of this method is demonstrated by reporting a real dataset measured in an fMRI decoding study. The program code for this iPIPI method is available at the following website: http://www2.nict.go.jp/bnc/hirose/iPIPI/index.html.

Keywords

fMRI decoding, group statistical test, information-like measure, i-th order statistic, population prevalence

1. Introduction

An increasing number of fMRI studies use multi-voxel pattern classification analysis, where the cognitive content (= label), such as visual input categories (Spiridon and Kanwisher 2002, Haxby et al., 2014, Nishida and Nishimoto 2018), types of movement (Gallivan et al., 2011, Nambu et al., 2015), and the effector of movement (Hirose et al., 2015, Hirose et al., 2018) is predicted (decoded) from fMRI signal patterns. The prediction's accuracy (decoding accuracy; D-Acc) is regarded as an index of the label information in the brain. In fMRI decoding studies, the experimental D-Acc obtained is often compared to the chance level accuracy obtained when the brain contains no information about the label (e.g., 50% for binary classifications). When the D-Acc is greater than the chance level, the brain, or a particular brain region, contains information regarding the cognitive content. In real experiments, the D-Acc for multiple participants is combined to infer the D-Acc's population.

The focus of this study concerns a statistical test for inferring the D-Acc's population from the D-Acc for multiple participants obtained in an experiment. In analogy with to conventional univariate analysis for voxel activity (Friston et al., 1995), the mean D-Acc across participants has been often compared to the chance level by using the Student's *t*-test (e.g. Gilbert and Fung, 2018, Gallivan et al., 2011) or the permutation tests (Gallivan et al., 2013, Nambu et al., 2015, Stelzer et al., 2013). The principle of these analyses is that the D-Acc sample mean is a good estimator of its population mean, and if the population mean is greater than chance, the brain activation is expected to typically contain cognitive content information. However, this is a misleading interpretation. Indeed, each participant's D-Acc should be equal to or better than the random estimator and cannot be less than the chance level. Thus, only one person in the population with a better-than-chance D-Acc is sufficient to make a population mean greater than chance. Therefore, the significant results of these tests cannot provide inferences regarding the population trend; they can only infer that "*there is at least one participant in the population that has a greater-than-chance D-Acc*" (Allefeld et al., 2016). Therefore, such tests fail to infer the population majority trend.

The above-mentioned problem is inevitable for statistical tests of the population mean. One solution is to assess the population information prevalence (i.e., the proportion of the population with a better D-Acc than chance level). Accordingly,

Allefeld et al. proposed a statistical test on the D-Acc, where the population information prevalence is evaluated instead of the population mean. This was inspired by other methods evaluating the proportion of the population for second-level (multi-subject) statistical test in fMRI data analyses, such as the dynamic causal model (Stephan et al., 2009), the second-level random effect of univariate analysis (Rosenblatt et al., 2014), and conjunction analysis (Friston et al., 1999). In this method, the null hypothesis is that the "*proportion of the population that has a greater-than-chance D-Acc is at or less than the predetermined prevalence threshold (e.g., 0.5).*" Thus, if the null hypothesis is rejected, it can be concluded that "greater than the prevalence threshold (e.g., more than half) of the population has a D-Acc greater-than-chance level". See Section 2.2.2 for the relationship between the methods proposed by Allefeld et al. (2016) and that proposed in this paper.

The method of Allefeld et al. (2016) theoretically solved the above concern and achieved a meaningful inference regarding the population trend. However, the method has a drawback in that it is weak against outlier results. Because the worst D-Acc among the participants is used as the test statistic, the method fails to capture the correct population characteristics, such as when a lower outlier exists in the experimental result. This is serious because empirical evidence has demonstrated the non-negligible possibility of a lower outlier in fMRI decoding (Deuker et al., 2013; Etzel et al., 2013; Fuentemilla et al., 2010; Noirhomme et al., 2014, Jamalabadi et al., 2016), and some theoretical studies have suggested that data with specific properties always yield below-chance D-Acc values (Kowalczyk, 2007). Indeed, such data can be observed in empirical biological experiments (e.g., Kowalczyk et al., 2007, Roadknight et al., 2012).

To address these issues in the present study, a statistical test on the information prevalence is developed by extending the existing method for robustness against outlier data. The idea is to generalize Allefeld's method to the i-th order statistic (i-th OS: i-th worst D-Acc). Namely, the 2nd OS (2nd worst D-Acc) or higher OS is used as the test statistic. In this manner, the method becomes robust against outliers because the i-th OS is immune to the existence of $i-1$ lower outliers.

The remainder of the paper is organized as follows: In the next section, the proposed statistical test on information prevalence is explained (Section 2). Then, the advantage of the method is demonstrated by providing numerical results on artificial

data (Section 3). Next, the application is demonstrated using a real fMRI dataset (Section 4). Finally, the results are summarized, and the key characteristics of the method are briefly discussed, including future directions and applications (Section 5).

2 Permutation-based Information Prevalence Inference using the *i*-th order statistic (iPIPI)

In this study, a novel procedure for evaluating information prevalence is proposed, the **P**ermutation-based **I**nformation **P**revalence **I**nference, using the ***i***-th order statistic (iPIPI).

In this section, the statistical model is clarified, the problem is defined, and the underlying theory is outlined (Section 2.0). Next, theoretical detail of iPIPI (Section 2.1) is explained, along with an introduction of the modified version of iPIPI, assuming an identical distribution among participants (Section 2.2). Finally, the practical procedure for performing iPIPI is explained (Section 2.3). Section 2.3 is self-contained, to some extent, and contains minimum equations, so that readers can skip 2.1 and 2.2 to see the practical procedure before the theoretical detail.

2.0 notations, problem definition, and outline of the calculation

Figure 1A shows the statistical model of the population (Ω) and how experimental results (sample D-Acc in a real experiment) are obtained. Population Ω can be partitioned into two subgroups, Ω_+ and Ω_-. People in Ω_+ have label information in their brain activity and, thus, greater D-Acc values than chance, while people in Ω_- do not. When a person with index n is randomly chosen from the population, he/she belongs to Ω_+ ($n \in \Omega_+$), with a probability γ, or otherwise belongs to Ω_- ($n \in \Omega_-$). The experiment with N participants is expressed as a two-step random sampling. First, N participants are randomly sampled from the population independently of each other ("Random sampling of participant"). Here each sampled participant is associated with the individual probability distribution (probability mass function) of the D-Acc ($p(a_n)$). Second, the sample D-Acc ($\hat{a}_n$) is randomly sampled from the distribution for each participant (Random sampling of D-Acc). The overall objective of iPIPI is to test whether γ is higher than the predetermined threshold γ_0 with the significance threshold α from the results of the experiment with N participants.

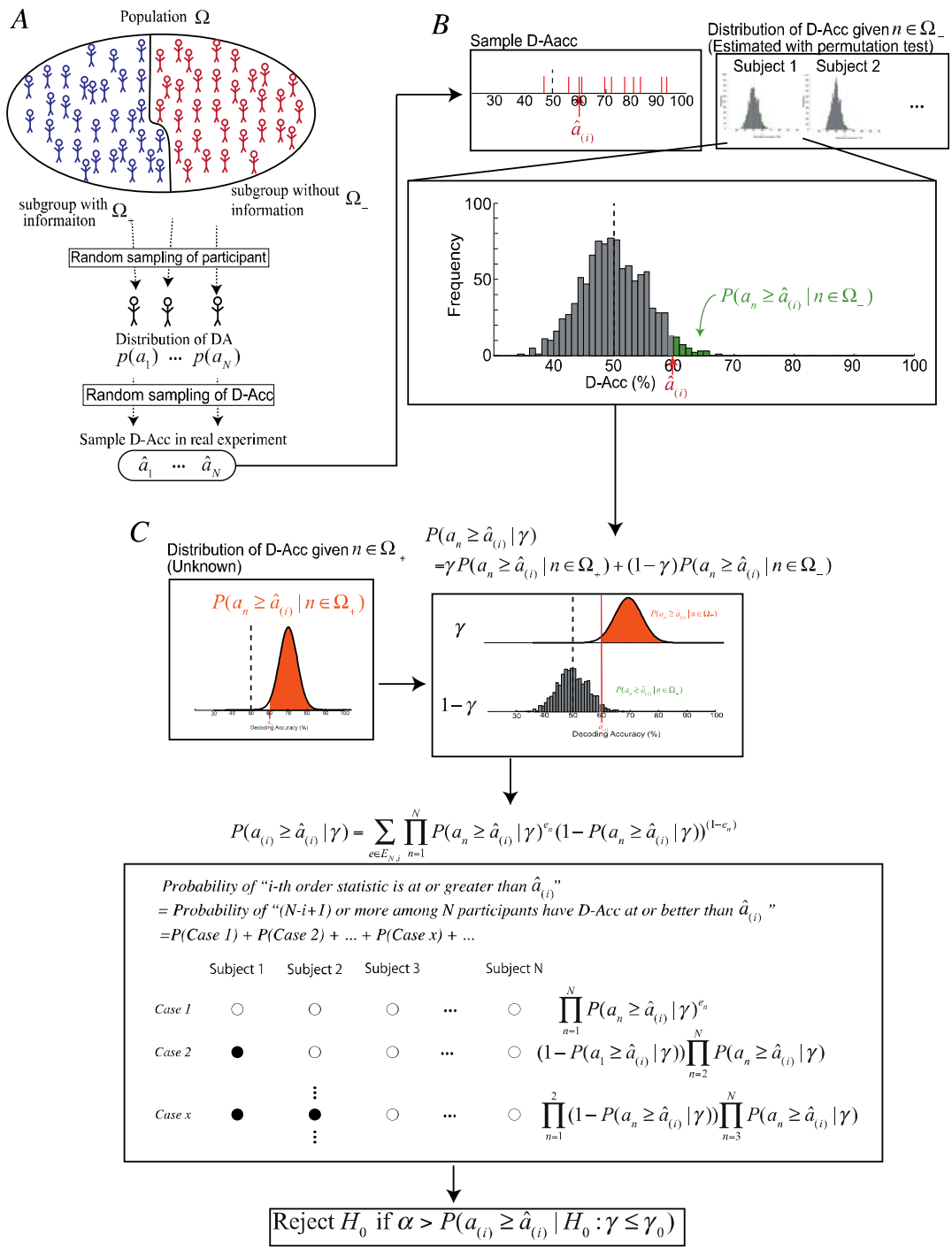


Figure 1: A statistical model and schematic design of iPIPI. **A:** Statistical model. The population (Ω) is composed of two subgroups, Ω_+ with label information, and Ω_- without label information. An experiment is expressed as two-step random sampling;

each participant was randomly chosen from the population (Random sampling of participant), and sample D-Acc in the experiment ($\hat{a}_n$) was randomly chosen from the D-Acc distribution for each chosen participant (Random sampling of D-Acc). **B:** Values obtained from experimental results. From the experimental results, i-th OS (i-th worst D-Acc; $\hat{a}_{(i)}$) was chosen as a test statistic (top left). Also, the permutation test was performed to estimate the probability that each participant's D-Acc was at or greater than $\hat{a}_{(i)}$ given that the participant did not have label information ($P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ (top right). Bottom: Example of the permutation test results for a participant. The histogram indicates the D-Acc distribution from the permutation tests. $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ is estimated by dividing the number of permutation tests with D-Acc at or greater than $\hat{a}_{(i)}$ (green bars) by a total number of permutation tests. **C:** Top left: The probability that each participant's D-Acc is greater than $\hat{a}_{(i)}$ given the participant has label information ($P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_+\right)$; unknown). Top right: the probability that a participant provides D-Acc at or greater than $\hat{a}_n$ can formulate as the weighted sum of $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_+\right)$ and $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ with the weight of the probability that a participant belongs to Ω_+ (γ) and to Ω_- ($1-\gamma$), respectively. Middle: The probability that "*i-th OS is at or greater than observed one* ($\hat{a}_{(i)}$)" is equivalent to the probability that "($N-i+1$) *or more among N participants have D-Acc at or greater than* $\hat{a}_{(i)}$." This is calculated by adding the probabilities of all the combinations of whether each participant's D-Acc was equal to or greater than $\hat{a}_{(i)}$ or not. In the table, the open and filled circles indicate the participants' D-Acc is better and worse than $\hat{a}_{(i)}$, respectively, such as, all D-Acc are at or better than $\hat{a}_{(i)}$ in Case 1, while D-Acc are at or better than $\hat{a}_{(i)}$ other than subject 1 in Case 2 and D-Acc for subject 1 and 2 are worse than $\hat{a}_{(i)}$ in case x. Bottom: The null hypothesis is rejected when $\alpha > P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0\right)$.

For the statistical test, we first choose the i-th OS (i-th worst D-Acc; $\hat{a}_{(i)}$) from the sample D-Acc observed in the real experiment (Figure 1B top left). In addition, the permutation tests estimate the probability that each participant's D-Acc is greater than $\hat{a}_{(i)}$ under the given condition that the participant does not have label information ($P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$). In the permutation test, we perform the decoding analysis with the "permuted" dataset, where the label for the trial is randomly permuted. By performing the permutation test multiple times, we achieve a reasonable estimate of $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ (Figure 1B bottom; see Allefeld et al., 2015, Stelzer et al., 2013 for empirical use in fMRI experiment; Good, 2000, Ojala and Garriga 2010 for theory). Without estimating the D-Acc probability distribution, when a person has label information ($p(a_n | n \in \Omega_+)$; Figure 1C top left), we formulate the probability that a participant provided a DA greater than $\hat{a}_n$ under a given γ ($P\left(a_n \geq \hat{a}_{(i)} | \gamma\right)$; Figure 1C top right) as a weighted sum of $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_+\right)$ and $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ with the weight of the probability that a participant belongs to Ω_+ (γ) and to Ω_+ ($1-\gamma$), respectively. Then, the probability that "*i-th OS is at or greater than observed one* ($\hat{a}_{(i)}$)" is formulated (1C middle) as a function of γ ($P\left(a_{(i)} \geq \hat{a}_{(i)} | \gamma\right)$; Figure 1C right). This is equal to the probability that "*$(N-i+1)$ or more among N participants have a D-Acc at or greater than $\hat{a}_{(i)}$*" and can be calculated by the summation of the probabilities for all combinations of whether each participant's D-Acc is equal to or greater than $\hat{a}_{(i)}$ or not. Although we cannot calculate the probability directly, we can calculate its upper bound under the null hypothesis: $\mathrm{H}_0: \gamma \leq \gamma_0$. Then, when the upper bound is lower than α, we can reject the null hypothesis and prove that γ is higher than the predetermined threshold γ_0 with the significance threshold α.

The notations are tabulated in Table 1.

Table 1: Notations used in the present study.

Predetermined Constant
N: Number of participants
γ_0: prevalence threshold See Eq. (6)
α: statistical threshold
i: index for order statistic
M: number of permutation tests for each participant
Probabilistic Variable
a_n: D-Acc of a participant $n,\ n = 1 \ldots N$
$a_{(i)}$: i-th order statistic of the N samples
Experimental results
$\hat{a}_n$: observed D-Acc of participants in the experiment ($n = 1 \ldots N$)
$\hat{a}_{(i)}$: the i-th order statistic of observed D-Acc in the experiment
$\hat{a}_{perm(n,m)}$: D-Acc obtained from the m-th permutation test for the participant with the index of n
Function
$BPDF(k, K, \rho) = {}_K C_k \rho^k (1-\rho)^{(K-k)}$: binomial probability mass function
$BCDF(k, K, \rho) = \sum_{h=0}^{k} BPDF(h, K, \rho)$: binomial cumulative probability mass function

2.1 Theoretical details

As noted above, the objective of the iPIPI is to test whether γ is higher than the predetermined threshold γ_0. Thus, the alternative hypothesis of $\mathrm{H}_1{:}\,\gamma > \gamma_0$ is set, and the null hypothesis as its negation $\mathrm{H}_0{:}\,\gamma \leq \gamma_0$ is set. Under this null hypothesis, we evaluate whether the i-th worst sample D-Acc $\hat{a}_{(i)}$ is so good that it rarely (with a probability less than the significance threshold α) occurre.

This calculation is conducted in four steps. First, the probability is formulated so that each participant's D-Acc is greater than $\hat{a}_{(i)}$ under a given γ ($P(a_n \geq \hat{a}_{(i)}|\gamma)$; Eq. (2.2)). Second, by combining the probability for each participant, the probability that "*the i-th worst D-Acc is better than the one actually observed* ($P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right)$"; Eq. (2.3)) is determined. Third, the upper limit of $P(a_n \geq \hat{a}_{(i)}|\gamma)$ under the null hypothesis ($P\left(a_n \geq \hat{a}_{(i)}|\mathrm{H}_0{:}\,\gamma \leq \gamma_0\right)$) is calculated using Eq. (2.5) and Eq. (2.6)). Fourth, the upper limit of $P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right)$ under the null hypothesis ($P\left(a_{(i)} \geq \hat{a}_{(i)}|\mathrm{H}_0{:}\,\gamma \leq \gamma_0\right)$) is calculated. Then, the upper limit of $P\left(a_{(i)} \geq \hat{a}_{(i)}|\mathrm{H}_0{:}\,\gamma \leq \gamma_0\right)$ is compared to α.

Formulation of $P(a_n \geq \hat{a}_{(i)}|\gamma)$

By using the probability that a participant belongs to Ω_+ (γ) and the D-Acc probability distribution given a person has label information (($p(a_n|n \in \Omega_+)$) and the distribution given the person does not have label information ($p(a_n|n \in \Omega_-)$), $p(a_n|\gamma)$ is formulated as follows;

$$p(a_n|\gamma) = (1-\gamma)p(a_n|n \in \Omega_-) + \gamma p(a_n|n \in \Omega_+) \#(2.1).$$

Note that $p(a_n|n \in \Omega_-)$ and $p(a_n|n \in \Omega_+)$ is not necessarily identical among participants, while γ is common across participants. From Eq. (2.1), the probability that "*D-Acc of a participant is at or greater than* $\hat{a}_{(i)}$" is formulated as

$$\boldsymbol{P(a_n \geq \hat{a}_{(i)}|\gamma) = (1-\gamma)P(a_n \geq \hat{a}_{(i)}|n \in \Omega_-) + \gamma P(a_n \geq \hat{a}_{(i)}|n \in \Omega_+) \#(2.2)}\ .$$

Formulation of $P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right)$

Then, we formulate the probability that the "*i-th worst D-Acc is at or greater than* $\hat{a}_{(i)}$." This is equivalent to the probability that "$(N-i+1)$ *or more among* N

participants have a D-Acc at or greater than $\hat{a}_{(i)}$" and, thus, the probability can be calculated as a sum of the probability of all combinations of whether each participant's D-Acc is equal to or greater than $\hat{a}_{(i)}$ or not;

$$P\left(a_{(i)} \geq \hat{a}_{(i)}|\gamma\right) = \sum_{e \in \mathrm{E}_{N,i}} \prod_{n=1}^{N} P\left(a_n \geq \hat{a}_{(i)}|\gamma\right)^{e_n} \left(1 - P\left(a_n \geq \hat{a}_{(i)}|\gamma\right)\right)^{(1-e_n)} \#(2.3) \quad .$$

Here, $\mathrm{E}_{N,i}$ is the group of all N-dimensional vectors, with elements of 1 and 0, and the number of 0s is less than i (the number of 1s is more than $N - i$). When a vector e is an element of $\mathrm{E}_{N,i}$, e_n is the n-th element of e. The N-dimensional vector e represents each combination, and each element of e represents whether the D-Acc of the corresponding participant is better than $\hat{a}_{(i)}$ (1) or not (0). For example, when $N = 3$ and $i = 2$, $\mathrm{E}_{N,i} = \{[1\ 1\ 1], [0\ 1\ 1], [1\ 0\ 1], [1\ 1\ 0]\}$ and the first element of $\mathrm{E}_{N,i}$ ($[1\ 1\ 1]$), represents the case that the D-Acc of all three participants are greater than $\hat{a}_{(i)}$ and $[0\ 1\ 1]$ represents the case that the D-Acc of participant #1 is not greater than $\hat{a}_{(i)}$ but the remaining two (#2 and #3) are greater than $\hat{a}_{(i)}$. Because of each participant's independence, the probability of each case is the product of the probability of each participant's result, e.g., when $e = [0\ 1\ 1]$, $\prod_{n=1}^{3} P\left(a_n \geq \hat{a}_{(i)}|\gamma\right)^{e_n} \left(1 - P\left(a_n \geq \hat{a}_{(i)}|\gamma\right)\right)^{(1-e_n)} = \left(1 - P\left(a_1 \geq \hat{a}_{(i)}|\gamma\right)\right) P\left(a_2 \geq \hat{a}_{(i)}|\gamma\right) P\left(a_3 \geq \hat{a}_{(i)}|\gamma\right)$.

<u>Upper limit of $P(a_n \geq \hat{a}_{(i)}|\mathrm{H}_0\colon \gamma \leq \gamma_0)$</u>

We find the upper limit of $P(a_n \geq \hat{a}_{(i)}|\gamma)$ as

$$P\left(a_n \geq \hat{a}_{(i)}\middle|\gamma\right) \leq (1-\gamma) P\left(a_n \geq \hat{a}_{(i)}|n \in \Omega_-\right) + \gamma \#(2.4).$$

Eq. (2.4) can be easily verified from Eq. (2.2) by noting that $P\left(a_n \geq \hat{a}_{(i)}|n \in \Omega_+\right) \leq 1$, because of the basic nature of probability. Then, the right side of Eq. (2.4) is an increasing function of γ, $P\left(a_n > \hat{a}_{(i)}\middle|\gamma\right)$ is bounded above with Q_n under the null hypothesis $\mathrm{H}_0\colon \gamma \leq \gamma_0$;

$$Q_n \geq P\left(a_n \geq \hat{a}_{(i)}|\mathrm{H}_0\colon \gamma \leq \gamma_0\right) \#(2.5),$$

where

$$Q_n = (1-\gamma_0) P\left(a_n \geq \hat{a}_{(i)}|n \in \Omega_-\right) + \gamma_0 \#(2.6).$$

Note that the value of Q_n can be calculated after we estimate $P\left(a_n \geq \hat{a}_{(i)} | n \in \Omega_-\right)$ with the permutation tests.

<u>Upper limit of $P(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0 \colon \gamma \leq \gamma_0)$</u>

Because $P\left(a_{(i)} \geq \hat{a}_{(i)} | \gamma\right)$ in Eq. (2.3) monotonically increases with each $P\left(a_n \geq \hat{a}_{(i)} | \gamma\right)$ (see below), and $P\left(a_n \geq \hat{a}_{(i)} | \gamma\right)$ is bounded above with Q_n under the null hypothesis (Eq. (2.5)), we find that $P\left(a_{(i)} \geq \hat{a}_{(i)} | \gamma\right)$ is bounded above with L under the null hypothesis $\mathrm{H}_0 \colon \gamma \leq \gamma_0$ as follows;

$$L \geq P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0 \colon \gamma \leq \gamma_0\right) \#(2.7),$$

where

$$L = \sum_{e \in \mathrm{E}_{N,i}} \prod_{n=1}^{N} {Q_n}^{e_n} (1 - Q_n)^{(1-e_n)} \#(2.8) .$$

The monotonicity is obvious because when the probability that the "*D-Acc of one of the participants is equal to or greater than* $\hat{a}_{(i)}$" ($P\left(a_n \geq \hat{a}_{(i)} | \gamma\right)$) increases, the probability that "*more than* $N - i$ *participants have a D-Acc equal to or greater than* $\hat{a}_{(i)}$" ($P\left(a_{(i)} \geq \hat{a}_{(i)} | \gamma\right)$) should increase.

After estimating $P\left(a_n \geq \hat{a}_{(i)} | n \in \Omega_-\right)$ by means of the permutation tests, we can calculate Q_n with Eq. (2.6), and consequently, we can calculate L with Eq. (2.8). When L is smaller than the statistical threshold α, it is proven that $P\left(a_{(i)} \geq \hat{a}_{(i)} | \mathrm{H}_0\right)$ is smaller than α from Eq. (2.7). Thus, we can conclude that the event that the "*i-th worst sample D-Acc is at or greater than the observed one* $\hat{a}_{(i)}$" rarely (with the probability of less than α) occurs under the null hypothesis. Therefore, we can reject the null hypothesis and accept $\mathrm{H}_1 \colon \gamma > \gamma_0$, i.e., "the brain activation has information about the label in proportion with more than γ_0 of the population."

<u>*Predetermined parameters*</u>

iPIPI requires four predetermined parameters: N (number of participants), α (threshold for statistical significance), γ_0 (threshold for the information prevalence), and i (order of the test statistics). These should satisfy the following inequality;

$$\alpha > BCDF(i-1, N, 1-\gamma_0) \#(2.9)$$

where $BCDF$ is a cumulative distribution function of the binomial distribution (Table 1). This constraint is derived as follows: First, from Eq. (2.6), Q_n is obviously bounded below with γ_0 because of $P\left(a_n \geq \hat{a}_{(i)} | n \in \Omega_-\right) \geq 0$. Then, from Eq. (2.8), and because L is monotonically increased with Q_n, L is bounded below as

$$L > \sum_{e \in \mathrm{E}_{N,i}} \prod_{n=1}^{N} {\gamma_0}^{e_n} (1-\gamma_0)^{(1-e_n)} \ \#(2.10) \quad ,$$

From Lemma B1 (Appendix B1), the right side of Eq. (2.10) is $BCDF(i-1, N, 1-\gamma_0)$. Thus, Eq. (2.9) should be satisfied, otherwise, L is always larger than α and the statistical test never provides a positive result.

See Section 2.3 for the empirical guidelines for choosing particular parameter values within this constraint.

2.2 Modification with the assumption of identical distribution

2.2.1 iPIPI-id

In the original version of iPIPI explained above, we do not assume that $p(a_n)$ was identical among the participants. When the assumption of the identical distribution is accepted, Eq. (2.6) becomes

$$Q = (1-\gamma_0) P\left(a \geq \hat{a}_{(i)} | \Omega_-\right) + \gamma_0 \#(2.11),$$

where $P(a_n \geq \hat{a}_{(i)} | n \in \Omega_-)$ for all n is identical and it is written as $P\left(a \geq \hat{a}_{(i)} | \Omega_-\right)$. Then, since Q_n is unique among the participants ($Q_n = Q$), Eq. (2.8) can be reformulated by using Lemma B1 (Appendix B1) as

$$L = BCDF(i-1, N, 1-Q) \#(2.12) \quad .$$

We call this as iPIPI-id, which stands for the identical distribution (i.d.).

Merit of assuming i.d. is that it can make the estimate of the distribution of D-Acc under the given condition of $n \in \Omega_-$ precise because one value of $P\left(a \geq \hat{a}_{(i)} | \Omega_-\right)$ is estimated from $M \times N$ permutation tests, where M is the number of permutation tests for each participant. This is in contrast with the original iPIPI, where

M permutation tests are available to estimate $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$. In addition, when N and i increase, the number of components of $\mathrm{E}_{N,i}$ and the computational cost of Eq. (2.8) rapidly increase (Supplementary Figure 1). In such cases, the assumption of i.d. helps us to reduce the computational cost to evaluate L with Eq. (2.12). On the other hand, the assumption of the i.d. may deteriorate the model's flexibility. In the original iPIPI model, data characteristics, such as noise and signal variability across sessions, which affect D-Acc distribution ($p(a_n)$), do not need to be uniform. This may be an advantage in fMRI decoding because empirically the fMRI data characteristics affecting the D-Acc distribution often vary among participants. We discuss the merit and the demerit of this assumption in Section 5.1.

2.2.2 iPIPI-one

By assuming the i.d. and fixing $i = 1$, Eq. (2.12) becomes simply the N-th power of Q;

$$L = Q^N \#(2.13)$$

We refer to this as iPIPI-one. iPIPI-one is almost the same as the method previously proposed by Allefeld et al., (2016), referred to as the "Permutation-based information prevalence inference using the minimum statistic." The only difference is in the estimation of $P(a_n \geq \hat{a}_{(i)} | n \in \Omega_-)$ from the permutation tests. $P(a_n \geq \hat{a}_{(i)} | n \in \Omega_-)$ is estimated as the arithmetic mean of the estimate from each participants' permutation tests in the iPIPI-one (Eq. 2.15), while it is estimated as the geometric mean and evaluated via the second-level permutation (Stelzer et al., 2013) in Allefeld's method.

Because the arithmetic and geometric means are identical if the assumption of the i.d. is true, and the results of these two methods are identical in our artificial data (Section 3) as well as the empirical fMRI data (Section 4), we only report the results of the iPIPI-one in the present paper.

2.3 Practical Procedure

The procedure for performing the iPIPI is constructed with 1) set four predetermined parameters: α (statistical threshold), γ_0 (prevalence threshold), N (number of

participants), i (rank of the order statistic), 2) perform the permutation tests to estimate the probability for each participant that the "*D-Acc is at or better than* $\hat{a}_{(i)}$ *given the participant does not have label information*," ($P\left(a_n \geq \hat{a}_{(i)} | n \in \Omega_-\right)$), and 3) calculate the upper limit of the "*i-th worst sample D-Acc is at or better than observed one* $\hat{a}_{(i)}$*;*" $P\left(a_{(i)} \geq \hat{a}_{(i)} | H_0\right)$, and compare it with α.

<u>Step 1: Determine the values of the four predetermined parameters</u>

The iPIPI objective is to test whether the population proportion (prevalence) of label information is larger than the threshold. Therefore, the iPIPI's positive results can infer that "the brain activation has information about the label in proportion more than the γ_0 of the population with a significant threshold of α." The iPIPI test statistic is the i-th worst D-Acc obtained from the results of the experiment with N participants. In the first step of the iPIPI, we determined that the four parameters show up in the above sentences, i.e., α (the threshold for statistical significance) and γ_0 (the threshold for the information prevalence), as well as N (number of participants) and i (order of statistic for test statistics).

The combination of parameters is constrained by Eq. (2.9), which determines the upper limit of i when the other parameters N, α, and γ_0 are fixed.

Under this constraint, there is no universal optimal set of parameter values because it depends on an unknown data structure. Here, a guideline for selecting the parameters is provided. This is not a solid recommendation but suggested as a first choice. First, $\gamma_0 = 0.5$ may be intuitively acceptable, because we can infer that "*a majority (more than half) of the population has label information,*" and it is, at present, the standard for existing statistical methods on population prevalence (Friston et al., 1999; see also Discussion in Allefeld et al., 2016). In addition, $\alpha = 0.05$ before the correction for multiple comparisons may be the primary choice because it is the most commonly used in many statistical tests published in neuroscience papers.

As for the values of i and N, the robustness against outliers of the iPIPI increases as i increases, because the iPIPI is immune to the existence of $(i-1)$ of the lower outliers (Section 3.1 and Appendix A1). Thus, a larger i can provide better statistical power when outliers occur with a certain probability (see Section 3). Thus, we

recommend the largest available i. Roughly, the available i becomes large when N becomes large (see Figure 2B). Thus, a larger N is considered better.

In summary, our suggestion for the parameters is to use $\gamma_0 = 0.5$, $\alpha = 0.05$ and an N as large as possible. Then, it is recommended to determine i as the maximum available value satisfying the constraint Eq. (2.9).

<u>Step 2: Identify $\hat{a}_{(i)}$ and permutation test</u>

After finishing the experiment, the main analysis identifies a particular value of $\hat{a}_{(i)}$. Also, $P(a_n \geq \hat{a}_{(i)} | n \in \Omega_-)$ is estimated with the permutation tests (Good, 2000, Ojala and Garriga 2010). When M permutation tests are performed for each participant, we can calculate D-Acc with each of the permuted data. Then, we estimate $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ for $n = 1 \ldots N$ by calculating the proportion of permutation tests that provide performance better than $\hat{a}_{(i)}$,

$$P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{a}_{perm(n,m)} \geq \hat{a}_{(i)}\right] \#(2.14),$$

where $\hat{a}_{perm(n,m)}$ is the D-Acc obtained from the m-th permutation test of the n-th participant and Iverson bracket $[\cdot]$ has the value 1 for a true condition and the value 0 for a false condition.

When we assume an identical distribution (i.d.) of $p(a_n)$ among the participants, (iPIPI-id), all permutation results $\hat{a}_{perm(n,m)}$ come from the identical distribution. Thus, we pooled the permutation results from all the participants and estimate one value:

$$P\left(a \geq \hat{a}_{(i)} \middle| \Omega_-\right) = \frac{1}{M \times N} \sum_{n=1}^{N} \sum_{m=1}^{M} \left[\hat{a}_{perm(n,m)} \geq \hat{a}_{(i)}\right] \#(2.15).$$

<u>Step 3: Calculate L and compare it with α</u>

Finally, by substituting the estimated values of $P(a_n \geq \hat{a}_{(i)} | n \in \Omega_-)$ to Eq. (2.6) and (2.8), we can calculate the upper limit (L) of the probability that the "*i-th OS is at or better than the observed one* ($\hat{a}_{(i)}$)." If $L < \alpha$, then we reject the null hypothesis and conclude that $\gamma \geq \gamma_0$.

When we assume i.d. of $p(a_n)$ (iPIPI-id), L can be calculated with Eq. (2.11) and Eq. (2.12).

3 Numerical results from artificial data

We designed the iPIPI to improve robustness against outliers compared to when we used the minimum statistic ($\hat{a}_{(1)}$) as proposed by Allefeld et al., (2016). If one lower outlier exists in the dataset, the test statistic $\hat{a}_{(1)}$ becomes the outlier, and the statistical test fails to achieve a significant result even when most of the population has brain activity associated with the label ($n \in \Omega_+$). Thus, we extended the method so that the 2nd and higher OS are candidates for the test statistic to make iPIPI tolerant of lower outliers. In this section, we provide examples demonstrating the benefit of this extension.

In the first example (Section 3.1), we demonstrate that, when the lower outlier occurs with a certain probability, iPIPI-one decreases its statistical power when the number of participants increases. This counter-intuitive drawback of iPIPI-one can be solved by using iPIPI with the maximum available value of i (i_{max}). Then, in the second example (Section 3.2), we demonstrate that, even when outliers are not present, iPIPI-one cannot increase the statistical power when the number of participants increases. This is also solved by using iPIPI with i_{max}.

In the examples below (Sections 3.1 and 3.2), we consider artificial data, where each participant performed 100 trials with a binary choice (chance level = 50%). For simplicity, we assume that the decoder for the participant with label information ($n \in \Omega_+$) always correctly predicts the label (i.e., DA = 100%), while the decoder for the participant without label information ($n \in \Omega_-$) predicted the label in each trial randomly and thus, DA for $n \in \Omega_-$ follows a binomial distribution ($p(a_n|n \in \Omega_-) = BPDF(a_n, 100, 0.5)$). In addition, we assume that the permutation tests can correctly estimate the distribution of $p(a|\Omega_-)$ and therefore $P\left(a \geq \hat{a}_{(i)} \middle| \Omega_-\right)$ is correctly estimated as

$1 - BCDF(\hat{a}_{(i)}, 100, 0.5)$.

We compare the probability of the significant result (P_{sig}) between $i = 1$ (iPIPI-one) and iPIPI with a maximum value of i that satisfies Eq. (2.9) (i_{max}). The threshold for statistical significance and the prevalence threshold are then fixed at 0.05 and 0.5, respectively ($\alpha = 0.05, \gamma_0 = 0.5$).

In this section, we only provide the numerical calculation results. See

Appendix A1 and A2 for details of the calculation and the MATLAB source codes for the calculation can be downloaded at http://www2.nict.go.jp/bnc/hirose/iPIPI/index.html. Note that iPIPI-id provides identical results to iPIPI in these examples because the D-Acc distribution is identical across participants.

3.1 Robustness against outliers

Here, we consider the situation that a person has label information ($n \in \Omega_+$; D-Acc = 100%), with a probability of 0.9, while the person does not ($n \in \Omega_-$) with a probability of 0.1. In addition, we assume that a participant in Ω_- always provides outlier results of the D-Acc = 30%. We emphasize that the alternative hypothesis $\mathrm{H}_1 : \gamma > \gamma_0$ is always true ($\gamma = 0.9, \gamma_0 = 0.5$) in this situation; thus, P_{sig} indicates statistical power, which is ideally close to 1.

We increased the number of participants from 5 to 50 and evaluated the probability of significant results P_{sig}. When $N < 5$, the predetermined parameters cannot satisfy the constraint (Eq. (2.10)); therefore, we cannot apply iPIPI.

The probability of a significant result of iPIPI-one ($i = 1$) gradually decreases as the number of participants (N) increases from 5 to 50 (solid blue line in Figure 2A), because the probability that at least one outlier exists increases (see Appendix A1). Furthermore, when i is fixed at a certain value ($i \geq 2$), the probability of a significant result gradually decreases (dotted colored lines). However, the maximum available value of i (i_{max}) also increases when N increases (Figure 2B). Thus, by using i_{max}, iPIPI can resist the decrease of the D-Acc (solid red line in Figure 2A). This is because the iPIPI is immune to the existence of $(i - 1)$ lower outliers (Appendix A1).

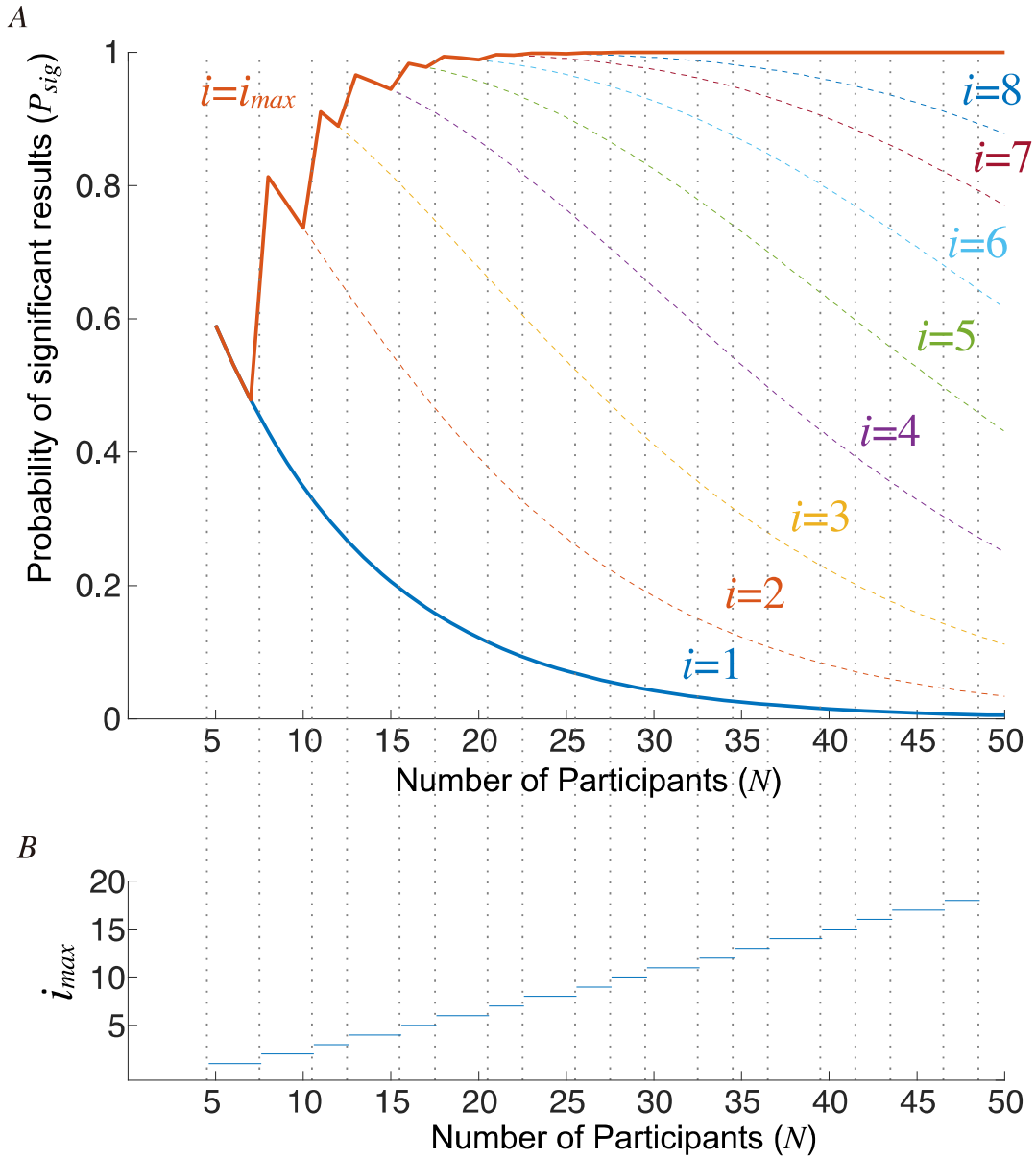


Figure 2: Numerical result from artificial data with outliers. **A**: The probability of a significant result (P_{sig}) is plotted against the number of participants (N). The blue solid line indicates the result of iPIPI-one ($i = 1$). The color dotted lines indicate the result from fixed i for $i = 1 \ldots 8$. The red solid line indicates the result from iPIPI with a maximum available value of i (i_{max}). The vertical black dotted lines indicate the point of i_{max} change. When $N < 5$, the iPIPI is not available because Eq. (2.9) cannot be satisfied with any i. **B**: The value of i_{max} is plotted against N.

3.2 Detectability of significance

In the next example, we did not include outliers explicitly, and the observed probability in the experiment was randomly chosen from the population distribution. Namely, participants have label information ($n \in \Omega_+$), with a certain probability γ and, if so, D-Acc = 100%. Participants do not have label information ($n \in \Omega_-$), with the probability of $(1-\gamma)$ and, if so, D-Acc follows $BPDF(a, 100, 0.5)$. We varied the true probability that a participant has label information (γ) from 0 to 1 with 0.01 steps, and varied the number of participants from 10 to 100 ($N = 10, 30, 50, 100$). When $\gamma > 0.5$ the alternative hypothesis $\mathrm{H}_1: \gamma > \gamma_0$ is true and, thus, P_{sig} indicates statistical power, which is ideally close to 1. In contrast, when $\gamma \leq 0.5$, the null hypothesis $\mathrm{H}_0: \gamma \leq \gamma_0$ is true and thus, P_{sig} indicates that the false alarm ratio should be smaller than $\alpha = 0.05$.

In Figure 3, P_{sig} (vertical axis) is plotted against true γ (horizontal). The vertical black dotted line indicates the threshold for population prevalence γ_0 (0.5), while the horizontal dotted black line indicates the statistical threshold α (0.05). As we observed in the figure, P_{sig} for iPIPI-one does not change when N increases ($i = 1$; Figure 3 blue lines). However, when we use iPIPI with i_{max}, P_{sig} (statistical power) increases with an increase of N when $\gamma > 0.5$, but remains lower than α when $\gamma \leq 0.5$. Consequently, the iPIPI with i_{max} provides a high statistical power while keeping the false alarm ratio lower than the statistical threshold.

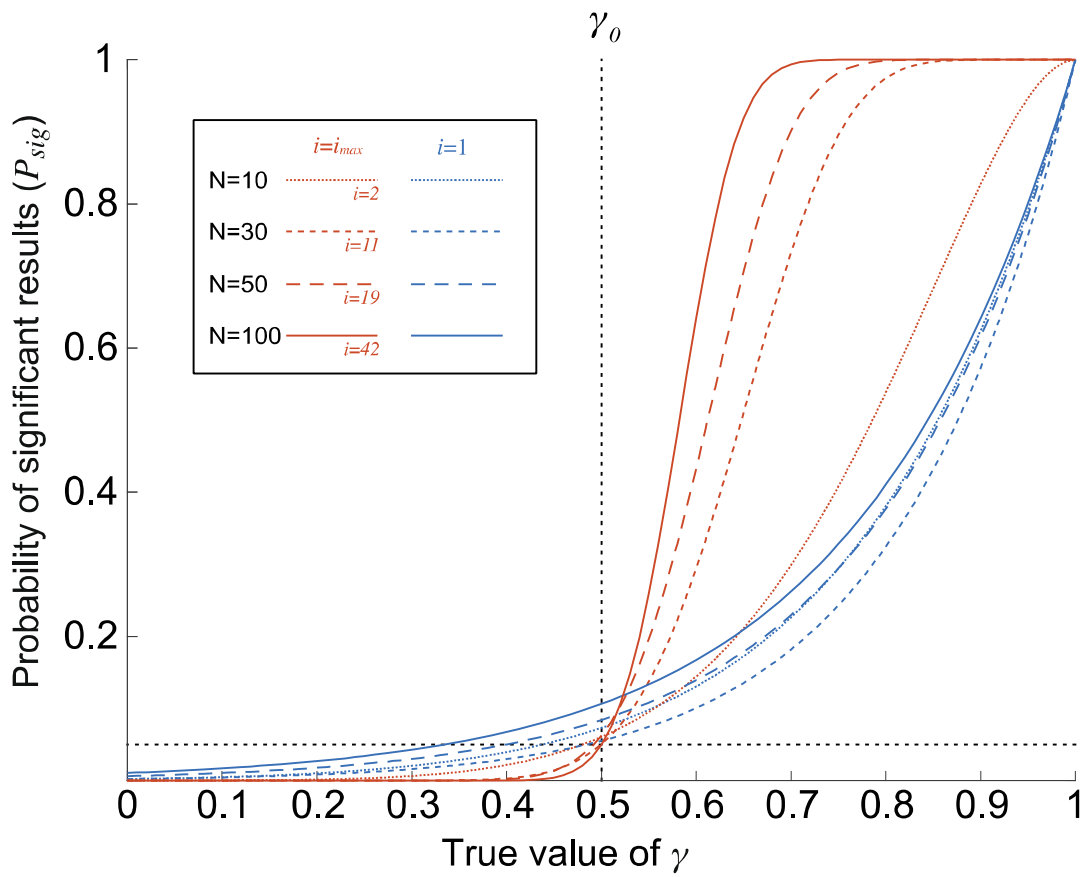


Figure 3: Numerical results from artificial data without outliers. The probability of a significant result (P_{sig}) from iPIPI with a maximum available value of i ($i = i_{max}$; red lines) and iPIPI-one ($i = 1$; Blue lines) with $N = 10,30,50,100$ is plotted against the true value of γ. The vertical and horizontal dotted black lines indicate the prevalence threshold $\gamma_0 = 0.5$ and the statistical threshold $\alpha = 0.05$. When $\gamma > 0.5$ (right side from the vertical dotted line in the graph), the alternative hypothesis $\mathrm{H}_1{:}\,\gamma > \gamma_0$ is true and thus, P_{sig} indicates statistical power and is ideally close to 1, while $\gamma \leq 0.5$, the null hypothesis $\mathrm{H}_0{:}\,\gamma \leq \gamma_0$ is true and thus, P_{sig} indicates a false alarm ratio and should be lower than α.

4 Application to empirical fMRI data

Here, to illustrate the practical application of iPIPI, we provide an example of its application to a real fMRI dataset, which is reported in Hirose et al., 2018.

4.1 Experiment and analyses procedures

For experimental details, please refer to the original paper (Hirose et al., 2018). Briefly, 12 participants performed a finger-tapping task using either his/her left- or right-hand finger, whichever they chose. Each participant underwent 10 sessions, and each session consisted of 15 trials. Each trial started with a 4-sec decision period, where participants chose either the left or right hand to be moved. Then, after a 4-sec delay, the participants moved the chosen hand finger. Due to ethical issues, the data cannot be open access. We predicted the chosen hand from brain activity measured before the movement execution.

Although each participant underwent 150 trials in total, there were error trials, in which no finger movement occurred or the participant moved multiple fingers. The number of error trials was 11 for a participant, five for another participant, and one for the other two participants. No error trials were found in the other eight participants. The error trials were excluded from the decoding analysis.

The preprocessed fMRI volume, which was measured before using an action initiation to predict the chosen hand. The classifier was trained by the iSLR algorithm (Hirose et al., 2015) on voxel activities in the whole brain. Classification accuracy was evaluated by the leave-one-session-out cross-validation procedure, in which we trained the classifiers using the trials of the nine sessions (training dataset) and used the trials of the remaining session to evaluate the prediction accuracy (test dataset). The balanced accuracy (Brodersen et al., 2010, Brodersen et al., 2012) was calculated from the cross-validation results.

The preprocessing, first-level decoding analyses and permutation test for each participant was performed by Multi-Voxel Pattern Classification toolbox (http://www2.nict.go.jp/bnc/hirose/mvpc/index.html).

Then, we applied the original iPIPI, iPIPI-id, and iPIPI-one to the results from the experiment and permutation tests for the statistical test. For all tests, we set the predetermined thresholds as $\alpha = 0.05$ and $\gamma_0 = 0.5$. We used $i = 3$, which is the

maximum available value ($N = 12$) that satisfies Eq. (2.9), for iPIPI and iPIPI-id.

In the permutation test, the labels (left or right hand) were randomly permuted within each session, and we performed leave-one-session-out cross-validation to evaluate the D-Acc for the permuted data. This was repeated 1,000 times for each participant. Then, we estimated $P\left(a_n \geq \hat{a}_{(i)} \middle| n \in \Omega_-\right)$ from the permutation results (Eq. (2.12)). For iPIPI-id and iPIPI-one, we pooled the results of the permutation tests for all participants (Eq. (2.13)).

4.2 Results

Original iPIPI: Figure 4A shows the distribution of the actual results (D-Acc) from all participants and the permutation results from an example participant. We set the 3rd OS (third worst observed D-Acc) as the test statistic $\hat{a}_{(3)} = 60.0\%$. Then, we estimated $P\left(a_n \geq \hat{a}_{(3)} \middle| n \in \Omega_-\right)$ for each participant by using the permutation tests. For example, since the D-Acc was better than 60.0% in 33 of 1,000 permutation tests for a participant presented in the figure (green bars in Figure 4A), we estimated $P\left(a_n \geq \hat{a}_{(3)} \middle| n \in \Omega_-\right) = 0.033$ for this participant. Likewise, we estimated $P\left(a_n \geq \hat{a}_{(3)} \middle| n \in \Omega_-\right)$ for each of the other participants. The estimated values ranged from 0.017 to 0.037. By substituting these estimated values into Eq. (2.6) and (2.8), we found $L = 0.024$. Since this was smaller than $\alpha = 0.05$, we rejected the null hypothesis. Since $\gamma_0 = 0.5$, we concluded that "the brain activation has information about the label in proportion to more than half of the population."

iPIPI-id: In iPIPI-id, we pooled the results of the permutation tests to estimate $P\left(a \geq \hat{a}_{(i)} | n \in \Omega_-\right)$. That is, we counted the total number of permutation tests with D-Acc better than $\hat{a}_{(3)}$ among 12,000 (12 x 1,000) permutation tests (Figure 4B). We found that the D-Acc was better than 60.0% in 316 of the 12,000 permutation tests (green bars in Figure 4B). Therefore, we estimated $P\left(a \geq \hat{a}_{(i)} | n \in \Omega_-\right)$ as 0.026 (316/12,000, Eq. (2.13)). By substituting this value into Eq. (2.7) and Eq. (2.8), we found $L = 0.024$. Thus, the null hypothesis is rejected also with iPIPI-id. In the current example, the statistical test provided virtually the same result as the original iPIPI (difference of values of $L < 0.000002$).

iPIPI-one: We set the worst observed D-Acc as the test statistic, $\hat{a}_{(1)} = 47.6\%$. From the permutation results, we found that the D-Acc was better than 47.6% in 7,998 of the 12,000 permutation tests (blue bars and green bars in Figure 4B). Thus, we estimated $P\left(a > \hat{a}_{(1)} | n \in \Omega_{-}\right)$ as 0.67. By using Eq. (2.11) and (2.13), we found $L = 0.11$. Thus, iPIPI-one could not reject the null hypothesis.

We cannot know the true population in the empirical data. However, when we look at the distribution of the obtained D-Acc (Red bars in Figure 4), the value of $\hat{a}_{(1)}$ appears to come from the population without information Ω_{-} rather than Ω_{+}, because it is below chance. Thus, we may assume that iPIPI-one failed to achieve significant results because one participant belonging to Ω_{-} provides low $\hat{a}_{(1)}$.

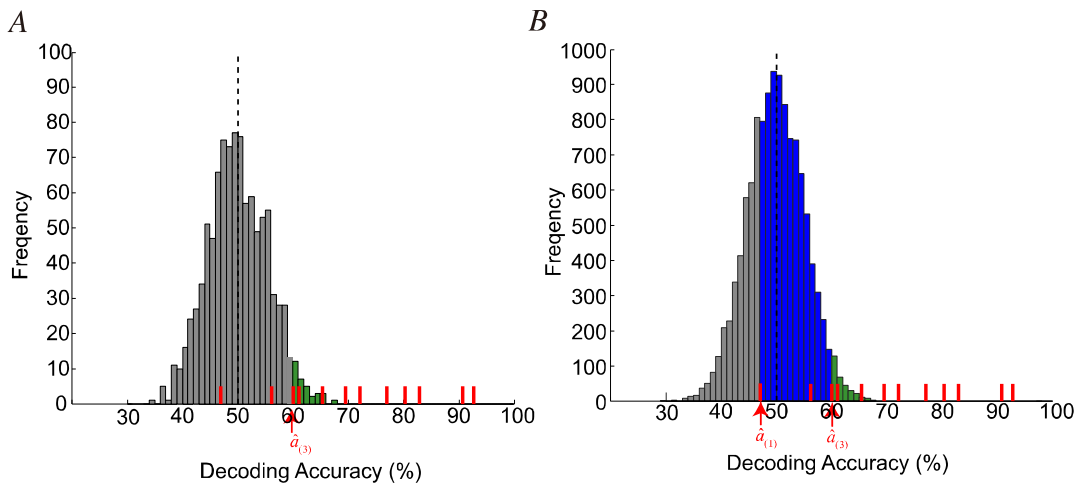


Figure 4: Permutation test results in the empirical experiment for the original iPIPI (**A**) and iPIPI-id and iPIPI-one (**B**). **A:** The histogram indicates the frequency of the D-Acc a_n given $n \in \Omega_-$, from 1,000 permutation tests for a representative participant. The red lines indicate the actual D-Acc obtained from all 12 participants, while the black dashed line indicates the chance level (50%). The arrow indicates the 3rd OS (3rd worst D-Acc) and green bars indicate the area that satisfies $a_n \geq \hat{a}_{(3)}$. **B:** The histogram indicates the frequency of D-Acc a_n given $n \in \Omega_-$ from 12,000 permutation tests for all 12 participants. The green bars indicate the area that satisfies $a_n \geq \hat{a}_{(3)}$, while $a_n \geq \hat{a}_{(1)}$ is indicated by the blue area + the green area.

5. Discussion

In this paper, we proposed a novel statistical test for decoding accuracy and named it iPIPI, which is an extension of the method proposed by Allefeld et al., (2016). In contrast to the commonly used statistical test, which compares the population mean D-Acc with the chance level, our method focuses on population information prevalence, i.e., the proportion of the population that has a better-than-chance D-Acc. The positive result of the statistical test on population information prevalence can infer major population trends as "*more than* γ_0 *(e.g., more than half) of the population has D-Acc better-than-chance level*." This is in contrast to the fact that the positive result of the statistical test on the population mean D-Acc can only infer that "*there is at least one participant in the population that has a better-than-chance D-Acc*," because participants' D-Acc should be equal to or better-than-chance but never worse than the chance level.

A statistical test proposed by Allefeld et al., 2016 also evaluates the population prevalence rather than the population mean. In their method, the minimum statistic is necessarily the test statistic. However, as shown in Section 3, using the minimum statistic leads to weakness against outliers. This weak point leads to a counter-intuitive drawback: the iPIPI-one can reduce its statistical power as the participants increase, in the situation where lower outliers appear with a certain probability (Figure 2A). This drawback is solved by using a higher OS. In fact, in iPIPI, a second or higher order statistic can be the test statistic. By choosing the highest order statistic, we effectively improved the robustness against outliers (Section 3.1). As a result, the iPIPI can increase the statistical power as the number of participants increase (Figure 2A). This characteristic can also be advantageous when outliers are not explicitly implemented. We have shown that the iPIPI can improve the statistical power when there are more participants, while iPIPI-one cannot (Section 3.2, Figure 3).

5.1 iPIPI and iPIPI-id

A different version of iPIPI was also introduced, namely iPIPI-id, in which the identical distribution of the D-Acc among the participants is assumed. The merit of this assumption is fewer parameters are estimated by the permutation tests and therefore more permutation tests can be used to estimate one parameter. In particular, we estimated one value of

$P(a \geq \hat{a}_{(i)}|\Omega_-)$ in the iPIPI-id from $M \times N$ permutation tests, while each of the N values of $P(a_n \geq \hat{a}_{(i)}|n \in \Omega_-)$ should be estimated from M permutation tests in the original iPIPI. This may become advantageous when the number of trials for each participant is low and, therefore, we cannot perform many permutation tests for each participant (small M), but many participants are recruited in the experiment (large N).

Furthermore, iPIPI-id has the advantage of computational cost. When both N and i increase, the computational cost for Eq. (2.8) in the original iPIPI rapidly increases, while its corresponding computation in iPIPI-id (Eq. (2.12)) increases more slowly (see Supplementary Figure 1A and B). The empirical evaluation found that iPIPI's computational cost increased exponentially when we used the maximum available i (Supplementary Figure 1A). The computation was completed within 1 second when $N < 23$ and was completed within 1 minute when $N < 28$, but it took about 5.5 minutes for $N = 30$. Thus, the empirical upper limit of N may be about 30 when we use the iPIPI with the maximum available i with the current implementation. When N is larger, we should select a smaller i with a sacrifice of statistical power (Supplementary Figure 1C). In contrast, the iPIPI-id with the maximum available i requires computational time that linearly increases with N and requires less than 0.03 seconds for $N = 1{,}000$ (Supplementary Figure 1B).

As for the demerits of assuming the identical distribution, assuming i.d. may not fit the real situation in some cases. Imagine that a participant did not perform the tasks in half of the trials. Then, the probability distribution of $p(a_n|n \in \Omega_-)$ of the participant may be distributed more widely than another participant without error. In such a situation, assuming an identical distribution may not be appropriate and, thus, the original iPIPI may provide a good fit to the true population.

Although such theoretical differences exist, there were few differences in our empirical experiment (Section 4). This is probably because the probability distribution did not largely differ among the participants. In our experiment, the difference in the number of trials between the participants was less than 10% (139-150) and all data was measured by the same MRI apparatus. Thus, we can speculate that the discrepancy among participants in our experiment was small enough to ignore for iPIPI-id. Although we cannot conclude how much, and what kind of, variation among the participant

results affected the discrepancy between the original iPIPI and iPIPI-id, the iPIPI-id may fail to capture the real population characteristics when there is a larger divergence among the participants. Therefore, in such a situation, we should choose the original iPIPI, in spite of its computational cost.

It may also be possible to aggressively utilize the unique feature of the original iPIPI that does not require assuming an identical distribution. For instance, we may be able to make the participant attend as many sessions as possible without determining the end of the experiment beforehand. We can combine the datasets measured by different MRI apparatus with no regard to the difference in signal and noise characteristics. Further algorithm development and empirical evaluations are needed to reduce the computational cost and evaluate the effect of this theoretical advantage in empirical studies.

6. Conclusions

In this paper, the iPIPI is proposed, along with a novel group-level statistical test for the decoding analysis. The advantages of the iPIPI are that the iPIPI 1) can provide the mathematically guaranteed, meaningful population inference; 2) is robust against outliers and can provide high statistical power; and 3) requires neither assuming a distribution nor an identical distribution among participants.

Although the iPIPI is being introduced as a statistical method for the DA, it is also applicable to other “information-like” measures, such as Mahalanobis distance (Kreigstore et al., 2006, Nili et al., 2014), linear discriminant t (Nili et al., 2014) or pattern distinctness D (Allefeld and Haynes, 2014). We expect this study will provide a robust, second-level statistical test for information-based neuroimaging studies.

Appendix A1:

Here, we describe the calculation of the probability of the significant result (P_{sig}) reported in Section 3.1. Note that the iPIPI is identical to the iPIPI-id here because individual differences are not implemented. In Appendix A.1 and A.2, the equation for iPIPI-id (i.e., Eq. (2.11) and Eq. (2.12)) is used because of its simplicity. By definition, the probability of the significant result is determined as follows:

$$P_{sig} = P(\alpha > L)\#(A1.1).$$

By remembering that the possible values of $\hat{a}_{(i)}$ are 30% and 100%, the P_{sig} can separate into two cases and is formulated as follows.

$$P_{sig} = P(a_{(i)} = 30\%)P\left(\alpha > L \middle| a_{(i)} = 30\%\right) + P(a_{(i)} = 100\%)P(\alpha > L|a_{(i)} = 100\%)\# \quad (A1.2).$$

$P(\hat{a}_{(i)} = 100\%)$ is equivalent to the probability that "number of outliers (i.e., observation of $DA = 30\%$) are fewer than i among N participants," and $P\left(a_{(i)} = 30\%\right) = 1 - P(a_{(i)} = 100\%)$. Thus, they are formulated as follows.

$$P\left(a_{(i)} = 100\%\right) = BCDF(i-1, N, 0.1)\#(A1.3).$$

$$P\left(a_{(i)} = 30\%\right) = 1 - BCDF(i-1, N, 0.1)\#(A1.4).$$

Then, $P\left(\alpha > L \middle| a_{(i)} = 100\%\right)$ and $P\left(\alpha > L \middle| a_{(i)} = 30\%\right)$ is calculated as follows. First, because $p(a|\Omega_-) = BPDF(a, 100, 0.5)$, $P\left(a \geq a_{(i)}|\Omega_-\right) = 1 - BCDF((a_{(i)} - 1), 100, 0.5)$. By substituting value of $a_{(i)} = 100\%\ or\ 30\%$, we evaluate the value of $P\left(a \geq a_{(i)}|\Omega_-\right)$. Then, by substituting this with $\gamma_0 = 0.5$ into Eq. (2.11), we evaluate the value of Q_n. Then, by substituting the value of Q_n with values of N and i into Eq. (2.12), we evaluate the value of L. If the value of L is smaller than α, $P\left(\alpha > L \middle| a_{(i)}\right) = 1$ and if not, $P\left(\alpha > L \middle| a_{(i)}\right) = 0$.

In particular, $P\left(a \geq a_{(i)}|\Omega_-\right)$ was almost 0 (1.7×10^{-18}) and almost 1

$(1 - 3.9 \times 10^{-5})$ when $a_{(i)} = 100\%$ and 30%, respectively. Therefore, Q_n was almost 0.5 ($0.5 + 8.3 \times 10^{-19}$) and almost 1 ($1 - 2.0 \times 10^{-5}$) given $a_{(i)} = 100\%$ and 30%. By substituting these values and the value of N and i into Eq. (2.12), we found that $\alpha > L$ when $a_{(i)} = 100\%$, while $\alpha \leq L$ when $a_{(i)} = 30\%$ at all tested combinations of N and i. This means that the iPIPI provides significant results when $\hat{a}_{(i)}$ is 100% (not an outlier), while it provides non-significant results when $\hat{a}_{(i)}$ is 30% (outlier). Therefore, when the number of participants with an outlier D-Acc (30%) is smaller than $(\mathrm{i} - 1)$, iPIPI provides a significant result.

To calculate P_{sig}, Eq. (A1.3), Eq. (A1.4), $P\left(\alpha > L \middle| a_{(i)} = 30\%\right) = 0$ and $P\left(\alpha > L \middle| a_{(i)} = 100\%\right) = 1$ is substituted into Eq. (A1.2). Then, P_{sig} is obtained as

$$P_{sig} = BCDF(i - 1, N, 0.1)\#(A1.5).$$

As shown in Figure 2, the value of P_{sig} decreases when the number of participants increases with fixed i, because the probability of i or more outliers increases.

Appendix A2:

Calculating the probability of a significant result P_{sig} is described in sections 3.2. Like Eq. (A1.2), P_{sig} is separated according to the value of $a_{(i)}$,

$$P_{sig} = \sum_{a_{(i)}=0}^{100} p\left(a_{(i)}\right) P\left(\alpha > L \middle| a_{(i)}\right) \#(A2.1),$$

Then, $p\left(a_{(i)}\right)$ can be calculated by separating according to the number of participants belonging to Ω_+ (N_+) as

$$p\left(a_{(i)}\right) = \sum_{N_+=0}^{N} p(N_+) p\left(a_{(i)} | N_+\right) \#(A2.2),$$

where $p(N_+)$ is the probability that the number of participants belonging to Ω_+ is N_+. Because the probability that each participant belongs to Ω_+ is γ, N_+ follows a binomial distribution, whose sample size is N and success ratio is γ, i.e.,

$$p(N_+) = BPDF(N_+, N, \gamma) \#(A2.3).$$

Finally, to find the probability mass function of $a_{(i)}$ under given N_+ ($p\left(a_{(i)} | N_+\right)$), the cumulative probability mass function $P\left(a_{(i)} \leq \bar{a}_{(i)} | N_+\right)$ was formulated. This is the probability that the "*i-th worst D-Acc is worse than $\bar{a}_{(i)}$ when N_+ participants belong to Ω_+.*" Because the probability that "*the D-Acc of a participant belonging to Ω_+ (100%) is worse than $\bar{a}_{(i)}$*" is always 0, we evaluate the probability that "among $N - N_+$ participants belonging to Ω_-, i or more participants' D-Acc is at or worse than $\bar{a}_{(i)}$." Using $P\left(a \leq \bar{a}_{(i)} | \Omega_-\right)$, this can be formulated as

$$P\left(a_{(i)} \leq \bar{a}_{(i)} | N_+\right) = 1 - BCDF\left((i-1), (N - N_+), P\left(a \leq \bar{a}_{(i)} | \Omega_-\right)\right) \#(A2.4).$$

Recall that $P\left(a \leq \bar{a}_{(i)} | \Omega_-\right) = BCDF(\bar{a}_{(i)}, 100, 0.5)$, this can formulate as

$$P\left(a_{(i)} \leq \bar{a}_{(i)} | N_+\right) = 1 - BCDF\left((i-1), (N - N_+), BCDF(\bar{a}_{(i)}, 100, 0.5)\right) \#(A2.5).$$

Because the probability that $a_{(i)}$ is at certain value ($\bar{a}_{(i)}$) is the difference between

$P\left(a_{(i)} \leq \bar{a}_{(i)} | N_{+}\right)$ and $P\left(a_{(i)} \leq (\bar{a}_{(i)} - 1) | N_{+}\right)$, $P\left(a_{(i)} = \bar{a}_{(i)} | N_{+}\right)$ can be formulated as

$$P\left(a_{(i)} = \bar{a}_{(i)} | N_{+}\right) = P\left(a_{(i)} \leq \bar{a}_{(i)} | N_{+}\right) - P\left(a_{(i)} \leq (\bar{a}_{(i)} - 1) | N_{+}\right) \#(A2.6).$$

Also, by using Eq. (A2.5), we obtain

$$P\left(a_{(i)} = \bar{a}_{(i)} | N_{+}\right) =$$

$$BCDF\left(i, N - N_{+}, BCDF\left((\bar{a}_{(i)} - 1), 100, 0.5\right)\right) - BCDF\left(i, N - N_{+}, BCDF\left(\bar{a}_{(i)}, 100, 0.5\right)\right) \#(A2.7).$$

By definition, the right hand side of Eq. (A2.7) becomes the probability mass function of $a_{(i)}$ under given N_{+} ($p\left(a_{(i)} | N_{+}\right)$) by replacing the $\bar{a}_{(i)}$ with $a_{(i)}$; therefore,

$$(p\left(a_{(i)} | N_{+}\right) =$$

$$BCDF\left(i, N - N_{+}, BCDF\left((a_{(i)} - 1), 100, 0.5\right)\right) - BCDF\left(i, N - N_{+}, BCDF\left(a_{(i)}, 100, 0.5\right)\right) \#(A2.8).$$

By substituting Eq. (A2.3) and (A2.8) into (A2.2), we evaluate $p\left(a_{(i)}\right)$.

$P\left(\alpha > L \middle| a_{(i)}\right)$ is evaluated for each value of $a_{(i)}$ as done in Appendix A.1. That is, first, we evaluated $P\left(a \geq a_{(i)} | \Omega_{-}\right) = BCDF(a_{(i)}, 100, 0.5)$ with fixed value of $a_{(i)}$, then, we evaluated Q with Eq. (2.11) and then, we evaluated L with Eq. (2.12). If $\alpha > L$, we found $P\left(\alpha > L \middle| a_{(i)}\right) = 1$ and $P\left(\alpha > L \middle| a_{(i)}\right) = 0$ otherwise. Finally, by substituting them with values of $p\left(a_{(i)}\right)$ into Eq. (A2.1), thus the value of P_{sig} (Figure 3) can be derived.

Appendix B:

The following lemma was used to convert from the right side of Eq. (2.10) to the right side of Eq. (2.9) and from Eq. (2.8) to Eq. (2.12).

Lemma B.1: When p is a probability ($0 \leq p \leq 1$) and the definition of $\mathrm{E}_{N,i}$ and e is the same as in Eq. (2.8), the following equation holds;

$$\sum_{e \in \mathrm{E}_{N,i}} \prod_{n=1}^{N} p^{e_n} (1-p)^{(1-e_n)} = BCDF(i-1, N, 1-p) \#(B.1) \quad .$$

Proof:

Recall that $\mathrm{E}_{N,i}$ is defined as the group of all N-dimensional vectors, with elements of 1 and 0, and the number of 0s is lower than i (the number of 1s is more than $N-i$). Thus, the number of elements that include h 0s is the number of possible groups of h from N (i.e., ${}_N C_h$) if h is less than i, while 0 otherwise. Thus, the right side of Eq. (B.1) can be reformulate as

$$\sum_{h=0}^{i-1} {}_N C_h p^{(N-h)} (1-p)^h \#(B.2) \quad .$$

By definition of BCDF and BPDF (Table 1), this is further reformulated as follows;

$$\sum_{h=0}^{i-1} BPDF\big(h, N, (1-p)\big) = BCDF\big(i-1, N, (1-p)\big) \#(B.3).$$

Thus, Lemma is proven.

Acknowledgments

I would like to acknowledge Dr. Atsushi Yokoi, Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, and Dr. Isao Nambu, Nagaoka University of Technology, for helpful suggestions on earlier versions of the manuscript. This work was supported by Grant-in-Aid for Young Scientists (B) (JSPS KAKENHI 16K16649) to the author.

References

Allefeld, C., Görgen, K., Haynes, J. D., 2016. Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. NeuroImage. 141, 378–392. DOI: 10.1016/j.neuroimage.2016.07.040

Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M. The balanced accuracy and its posterior distribution. In Pattern Recognition (ICPR), 2010 20th international conference on (pp. 3121–3124). IEEE. DOI: 10.1109/ICPR.2010.764

Brodersen, K.H., Mathys, C., Chumbley, J.R., Daunizeau, J., Ong, C.S., Buhmann, J.M., Stephan, K.E., 2012. Bayesian mixed-effects inference on classification performance in hierarchical data sets. J. Mach. Learn. Res. 13, 3133–3176.

Deuker, L., Olligs, J., Fell, J., Kranz, T. A., Mormann, F., Montag, C., Reuter, M., Elger, C.E., Axmacher, N., 2013. Memory consolidation by replay of stimulus-specific neural activity. J. Neurosci. 33(49), 19373–19383. DOI: 10.1523/JNEUROSCI.0414-13.2013

Etzel, J. A., Zacks, J. M., Braver, T. S. 2013. Searchlight analysis: promise, pitfalls, and potential. Neuroimage, 78, 261–269. DOI: 10.1016/j.neuroimage.2016.07.040

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S., Turner, R., 1995. Analysis of fMRI time-series revisited. Neuroimage, 2(1), 45–53. DOI: 10.1006/nimg.1995.1023

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., Worsley, K. J., 1999. Multisubject fMRI studies and conjunction analyses. Neuroimage, 10(4), 385–396. DOI: 10.1006/nimg.1999.0484

Fuentemilla, L., Penny, W. D., Cashdollar, N., Bunzeck, N., Düzel, E., 2010. Theta-coupled periodic replay in working memory. Curr. Biol. 20(7), 606–612. DOI: 10.1016/j.cub.2010.01.057

Gallivan, J.P., McLean, D.A., Valyear, K.F., Pettypiece, C.E., Culham, J.C., 2011. Decoding action intentions from preparatory brain activity in human parieto-frontal networks. J. Neurosci. 31, 9599–9610. DOI: 10.1523/JNEUROSCI.0080-11.2011

Gallivan, J.P., McLean, D.A., Flanagan, J.R., Culham, J.C., 2013. Where one hand meets the other: limb-specific and action-dependent movement plans decoded from preparatory signals in single human frontoparietal brain areas. J. Neurosci. 33, 1991–2008. DOI: /10.1523/JNEUROSCI.0541-12.2013

Gilbert, S. J., Fung, H., 2018. Decoding intentions of self and others from fMRI activity patterns. NeuroImage, 172, 278-290.

Haxby, J. V., Connolly, A. C., Guntupalli, J. S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 37, 435–456. DOI: 10.1146/annurev-neuro-062012-170325

Hirose, S., Nambu, I., Naito, E., 2015. An empirical solution for over-pruning with a novel ensemble-learning method for fMRI decoding. J. Neurosci. Methods. 239, 238–245. DOI: 0.1016/j.jneumeth.2014.10.023

Hirose, S., Nambu, I., Naito, E., 2018. Cortical activation associated with motor preparation can be used to predict the freely chosen effector of an upcoming movement and reflects response time: An fMRI decoding study. NeuroImage, 183, 584–596. DOI: 10.1016/j.neuroimage.2018.08.060

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., Gais, S., 2016. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum. Brain. Mapp. 37(5), 1842–1855. DOI: 10.1002/hbm.23140

Kowalczyk, A. Classification of anti-learnable biological and synthetic data. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 176–187). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-74976-9_19

Kowalczyk, A., Greenawalt, D. M., Bedo, J., Duong, C., Raskutti, G., Thomas, R. J., Phillips, W. A., 2007. Large Validation of Anti-learnable Signature in Classification of Response to Chemoradiotherapy in Esophageal Adenocarcinoma Patients. The First International Symposium on Optimization and Systems Biology (OSB’07)

Nambu, I., Hagura, N., Hirose, S., Wada, Y., Kawato, M., Naito, E., 2015. Decoding sequential finger movements from preparatory activity in higher-order motor regions: a functional magnetic resonance imaging multi-voxel pattern analysis. Eur. J. Neurosci. 42, 2851–2859. DOI: 10.1111/ejn.13063

Nishida, S., Nishimoto, S., 2018. Decoding naturalistic experiences from human brain activity via distributed representations of words. Neuroimage, 180, 232-242. DOI: 10.1016/j.neuroimage.2017.08.017

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S., 2014. Biased binomial assessment of cross-validated

estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage: Clinical, 4, 687–694. DOI: 10.1016/j.nicl.2014.04.004

Ojala, M., Garriga, G. C., 2010. Permutation tests for studying classifier performance. J. Mach. Learn. Res. 11(Jun), 1833–1863.

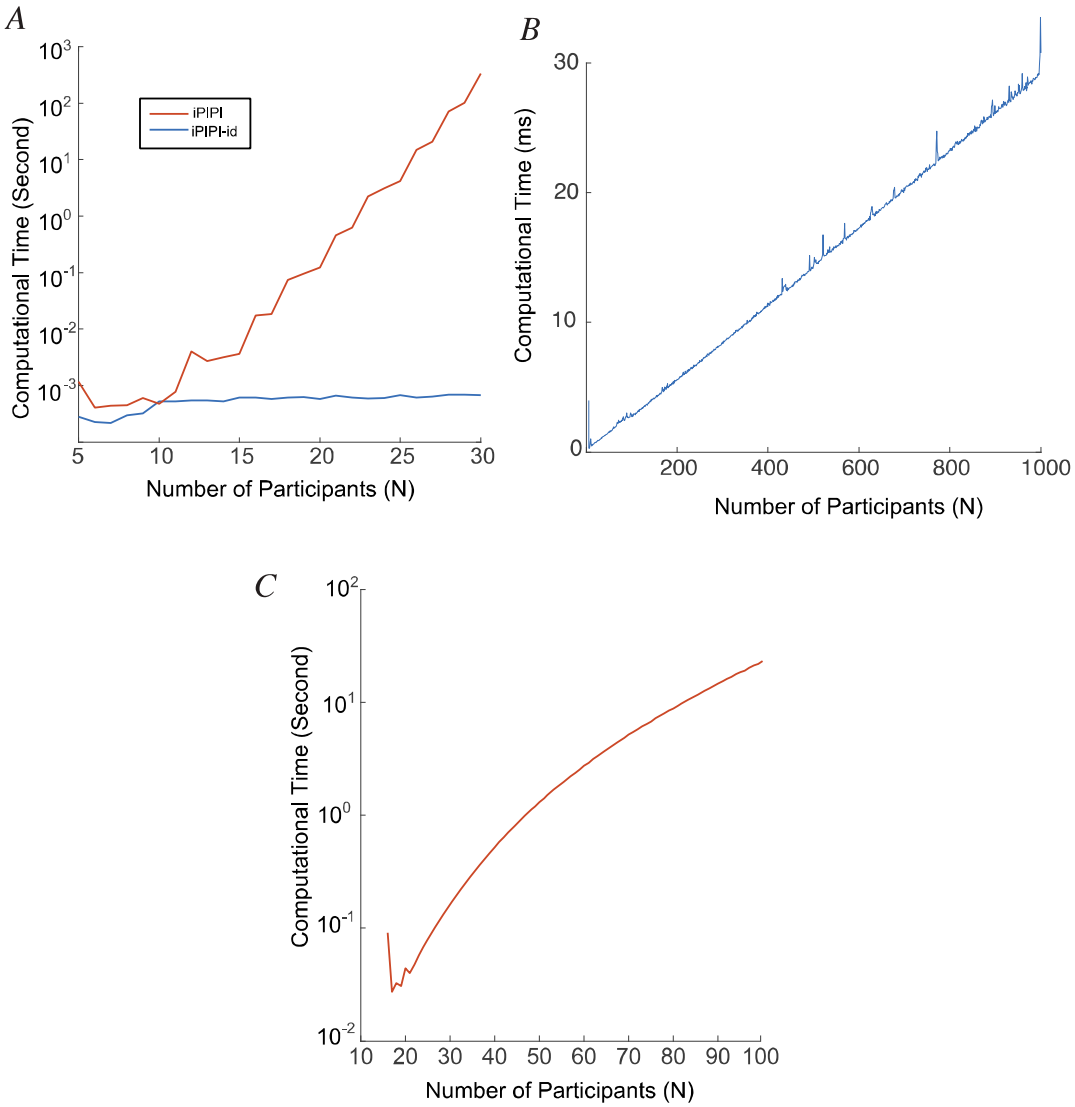
Roadknight, C., Aickelin, U., Qiu, G., Scholefield, J., Durrant, L. Supervised learning and anti-learning of colorectal cancer classes and survival rates from cellular biology parameters. In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference. DOI: 10.1109/ICSMC.2012.6377825

Rosenblatt, J. D., Vink, M., Benjamini, Y., 2014. Revisiting multi-subject random effects in fMRI: Advocating prevalence estimation. NeuroImage, 84, 113–121. DOI: 10.1016/j.neuroimage.2013.08.025

Spiridon, M., Kanwisher, N., 2002. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron, 35(6), 1157–1165. DOI: 10.1016/S0896-6273(02)00877-2

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage, 65, 69–82. DOI: 10.1016/j.neuroimage.2012.09.063

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage, 46 (4), 1004-1017. DOI: 10.1016/j.neuroimage.2009.03.025



Supplementary Figure 1: Benchmark test results on MATLAB 2018a, 64 bit, Linux, with 4 core of 3.5GHz CPU (Intel Xeon). **A**: Computational time for original iPIPI with maximum i (Red) and iPIPI-id with maximum i (blue) is plotted against the number of participants N. The vertical axis is log-scaled. **B**: Computational time for iPIPI-id with maximum i. **C**: Computational time for iPIPI with $i = 5$. These benchmark test results are roughly consistent with the fact that the order of numerical computational

complexity for original iPIPI with maximum i, iPIPI-id with maximum i and original iPIPI with fixed i are $O(N!)$, $O(N)$ and $O(N^i)$, respectively.