The similarity index problem

A collection of thoughts—not a paper draft.

Abstract. In neuroscience studies, it is common to compare the similarity between pairs of brain activation volumes, muscle responses, or brain networks. The idea is that greater similarity showcases greater interdependence between the two underlying processes. Similarity is often quantified using the Dice index (joint probability divided by average marginal probability) and/or the Jaccard index (joint probability divided by total probability). We show that this is problematic, because a greater Dice or Jaccard index does *not* imply greater interdependence. We argue for alternative measures from information theory.

Contents

1 Setup

In a probability context, the similarity indices are best described using a bivariate Bernoulli random vector $Y = (Y_1, Y_2)$ which can take on values $y = (y_1, y_2) \in \{0, 1\}^2$.

We denote $p_{y_1 y_2} \coloneqq \Pr(Y_1 = y_1 \wedge Y_2 = y_2)$, such that the probabilities read

	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	p_{00}	p_{01}
$Y_1 = 1$	p_{10}	p_{11}

and the probability mass function can be written as

$$P(Y = y) = p_{00}^{(1-y_1)(1-y_2)} p_{01}^{(1-y_1)(y2)} p_{10}^{(y_1)(1-y_2)} p_{11}^{(y_1)(y_2)} \tag{1}$$

with the necessary constraint

$$p_{00} + p_{01} + p_{10} + p_{11} = 0. \tag{2}$$

Now, we turn to the similarity indices. The Dice coefficient reads

$$d_{y_1 y_2} = 2 \cdot \frac{p_{y_1 y_2}}{\sum\limits_{y_1} p_{y_1 y_2} + \sum\limits_{y_2} p_{y_1 y_2}} \tag{3}$$

and the Jaccard coefficient reads

$$j_{y_1 y_2} = \frac{p_{y_1 y_2}}{\sum\limits_{y_1} p_{y_1 y_2} + \sum\limits_{y_2} p_{y_1 y_2} - p_{y_1 y_2}}. \tag{4}$$

We will also argue for alternative measures. The Forbes index reads

$$f_{y_1 y_2} = \frac{p_{y_1 y_2}}{\sum\limits_{y_1} p_{y_1 y_2} \sum\limits_{y_2} p_{y_1 y_2}} \tag{5}$$

and its logarithmic form is the pointwise mutual information

$$m_{y_1 y_2} = \log f_{y_1 y_2}. \tag{6}$$

2 Problem

Long story short, the problem is that the similarity indices are sensitive to the marginal probabilities *even if the joint probability remains fixed.* We first present the mathematical argument and then explain its meaning.

2.1 Maths

We can express the joint probability as a multiple of the product of marginal probabilities

$$p_{11} = \alpha(p_{01} + p_{11})(p_{10} + p_{11}), \tag{7}$$

where α is some non-negative scalar (the exact bounds of α depend on the probabilities, but are of no concern to us). Using Equation (2), we can rearrange this to

$$p_{01} = \frac{p_{11}(1 - \alpha p_{11} - \alpha p_{10})}{\alpha p_{10} + \alpha p_{11}}. \tag{8}$$

The Dice index then reads

$$d_{11} = 2 \cdot \frac{\alpha p_{11}(p_{10} + p_{11})}{\alpha(p_{10} + p_{11})^2 + p_{11}} \tag{9}$$

and the Jaccard index reads

$$j_{11} = \frac{\alpha p_{11}(p_{10} + p_{11})}{\alpha p_{10}(p_{10} + p_{11}) + p_{11}} \tag{10}$$

which readily shows that both similarity indices depend not only on α and p_{11}, but also on p_{10}. Put differently, for any given α and p_{11}, the similarity indices will vary with p_{10}.[1]

2.2 Meaning

We have shown that the indices depend on p_{01} and p_{10}. Put differently, they depend on the marginal probabilities $p_{01} + p_{11}$ and $p_{10} + p_{11}$ even if the joint probability p_{11} is fixed. For the sake of being super explicit:

1. As the joint probability increases, the similarity indices increase. This is desirable.
2. As the marginal probabilities increase, the joint probability increases and thus the similarity indices increase. This is undesirable.
3. As the marginal probabilities approach each other (the smaller one increases and the greater one decreases), *but the joint probability remains fixed*, the similarity indices increase. This is undesirable.

[1] More specifically, each similarity index is a concave function of p_{10} or p_{01} with its maximum at $p_{10} = p_{01} = -\alpha^{-1}(\sqrt{\alpha p_{11}} + \alpha p_{11})$.

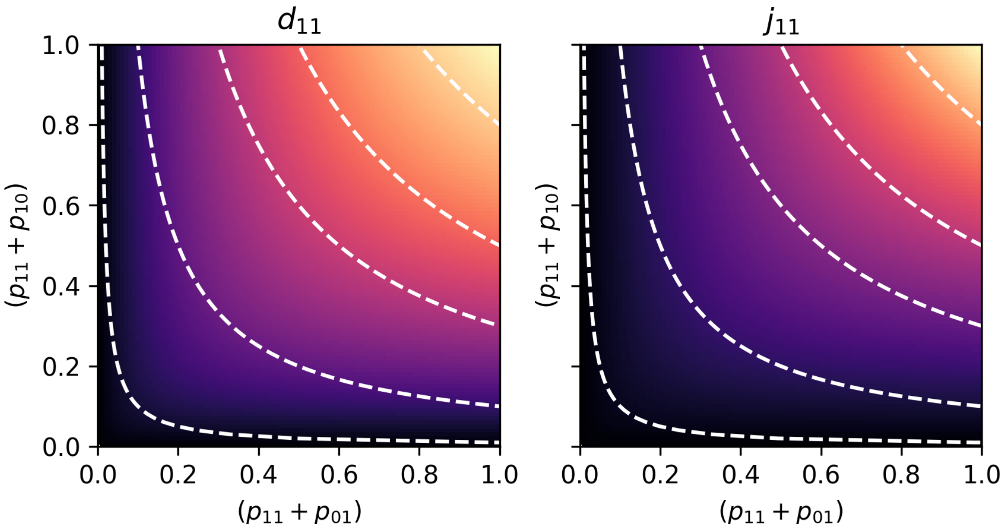

Figure 1: Illustration of how the similarity indices change with the marginal probabilities. This plot is for $\alpha = 1$ (i.e. independence) but the concept holds for any α. The square plot is the parameter space spanned by p_{01}, p_{10}, p_{11} given α. The dashed lines are the subspaces spanned by p_{01} and p_{10} given α and several p_{11}. Moving along the dashed lines means changing the marginal probabilities without changing the joint probability.

Note that the second point is fairly obvious and has been mentioned before in `Beware the Jaccard` (Salvatore et al., 2020) and also in some of the papers that employ either similarity index. The third point is a little less obvious.

3 Current Use

Both similarity indices are being used a lot in neuroscience research—mostly with fMRI but also with TMS and fNIRS. These indices are used to quantify overlap between two neural activation areas/volumes; to quantify co-occurence of muscle activations; or to quantify the similarity of two networks with common nodes. In either case, two binary vectors of paired observations are compared. The indices are then compared between different pairs.

3.1 Timeline

At first, mostly the Dice index was used. I could not find a neuroscience methods paper recommending the Dice index as an overlap measure, but it looks like its first use in an fMRI study was by Rombouts et al. (1997).

Overall, the Dice index has been used extensively with **fMRI** (Rombouts et al., 1997; Rombouts et al., 1998; Machielsen et al., 2000; Brannen et al., 2001; Quigley et al., 2001; Fernández et al., 2003; Raemaekers et al., 2007; Rau et al., 2007; Clément & Belleville, 2009; Meindl et al., 2009; Fesl et al., 2010; Maïza et al., 2011; Gorgolewski et al., 2013; Tie et al., 2013; Zhu et al., 2013; Gross & Binder, 2014; Kristo et al., 2014; Jann et al., 2015; Moessnang et al., 2016; Morrison et al., 2016; Wilson et al., 2016; Branco et al., 2018; Nettekoven et al., 2018; Fröhner et al., 2019; Ganzetti et al., 2019; Morales et al., 2019; Bach et al., 2021; Elin et al., 2021; Ibinson et al., 2022) and combined **TMS-fMRI** (Weiss Lucas et al., 2020).

The Jaccard index wasn't used much until it was recommended by by Maitra (2010) (see also Bennett & Miller, 2010) as a better alternative to the Dice index (in the fMRI context).

Overall, the Jaccard index has been used extensively with **fMRI** (Maldjian et al., 2002; Clément & Belleville, 2009; Crossley et al., 2013; Diederen et al., 2013; Karahanoğlu & Van De Ville, 2015; Ramezani et al., 2015; James et al., 2016; Marchitelli et al., 2016; Morrison et al., 2016; Marchitelli et al., 2017; Adrian et al., 2018; Fröhner et al., 2019; Jackson et al., 2019; Kampa et al., 2020; Mattioni et al., 2020; Bach et al., 2021; Ravindran et al., 2021), **TMS** (Melgari et al., 2008; Säisänen et al., 2019; DeJong et al., 2021; Nazarova et al., 2021; Tardelli et al., 2022), combined **TMS-fMRI** (McGregor et al., 2012), and **fNIRS** (Montero-Hernandez et al., 2018).[2]

4 Prior Work

Here, I will briefly summarize prior methodological work on the similarity indices.

4.1 Salvatore et al. (2020)

`Beware the Jaccard...` This paper highlights a different shortcoming of the Jaccard index and argues for the Forbes index. Consider three Bernoulli random variables y, r, q. In their terminology, y is the true underlying binary reality (which we do not observe), r is the reference track (our imperfect observation of y), and q is the query track. We wish to compare q with y, however, we can only compare q with r.

[2] McGregor et al. (2012) used a ”modified Jaccard index”.

The authors make two important assumptions. For these assumptions, we quickly need to introduce the false positive rate (FPR), false negative rate (FNR), true positive rate (TPR) and true negative rate (TNR).

$$\text{FPR} = \Pr(r = 1|y = 0) = \frac{\Pr(y = 0, r = 1)}{\Pr(y = 0)} = 1 - \text{TNR} \tag{11}$$

$$\text{FNR} = \Pr(r = 0|y = 1) = \frac{\Pr(y = 1, r = 0)}{\Pr(y = 1)} = 1 - \text{TPR} \tag{12}$$

$$\text{TPR} = \Pr(r = 1|y = 1) = \frac{\Pr(y = 1, r = 1)}{\Pr(y = 1)} = 1 - \text{FNR} \tag{13}$$

$$\text{TNR} = \Pr(r = 0|y = 0) = \frac{\Pr(y = 0, r = 0)}{\Pr(y = 0)} = 1 - \text{FPR} \tag{14}$$

In an ideal scenario, the authors assume that FPR $= 0$ and FNR > 0.[3] In other words, TNR $= 1$ and TPR < 1.

$$\Pr(r = 1|y = 0) = 0 \quad \Longrightarrow \quad \Pr(r = 0|y = 0) = 1 \tag{15}$$

$$\Pr(r = 0|y = 1) > 0 \quad \Longrightarrow \quad \Pr(r = 1|y = 1) < 1 \tag{16}$$

Furthermore, they assume that false positives are random, which means that

$$\Pr(r = 1|y = 1, q = 1) = \Pr(r = 1|y = 1, q = 0) = \Pr(r = 1|y = 1) \tag{17}$$

Informally, these assumptions mean: all 0's (TNR $= 1$) are preserved from y to r; however, only a fraction of 1's (TPR < 1) are preserved from y to r while the remaining 1's (FNR > 0) are turned into 0's. Furthermore, which of the 1's are preserved is random.

From a set theoretic perspective, one could say that r is a random subset of y. This is why the authors call the fraction of 1's that are preserved from y to r the "subsetting rate", which they denote k. I will from now on adopt their notation, so $k = \text{TPR} = \Pr(r = 1|y = 1)$.

Denoting $p_{ij} \coloneqq \Pr(y = i, q = j)$, we have the contingency tables:

	$q = 0$	$q = 1$
$y = 0$	p_{00}	p_{01}
$y = 1$	p_{10}	p_{11}

	$q = 0$	$q = 1$
$r = 0$	$p_{00} + (1 - k)p_{10}$	$p_{01} + (1 - k)p_{11}$
$r = 1$	$p_{10}k$	$p_{11}k$

[3] Note that in the Supplement, they give their argument in terms of FPR $= 0$. However, in the main text, they give it in terms of the false discovery rate FDR $= 0$. These two are equivalent, since FDR $= \Pr(y = 0|r = 1) = \frac{\Pr(y=0,r=1)}{\Pr(r=1)} = \frac{\text{FPR}\Pr(y=0)}{\Pr(r=1)}$. Also note that the false discovery rate can also be expressed FDR $= 1 - \Pr(y = 1|r = 1) = 1 - \frac{\Pr(y=1,r=1)}{\Pr(r=1)}$, which is the formulation they use in the main text.

For example, in the second table, we can get the right column by starting with the bottom cell

$$\begin{aligned}
&\Pr(r=1,q=1)\\
&=\Pr(r=1,q=1|y=1)\Pr(y=1)+\underbrace{\Pr(r=1,q=1|y=0)\Pr(y=0)}_{\text{0 since }r=1\text{ is impossible given }y=0}\\
&=\Pr(r=1|q=1,y=1)\Pr(q=1|y=1)\Pr(y=1)\\
&\overset{(17)}{=}\Pr(r=1|y=1)\Pr(q=1|y=1)\Pr(y=1)\\
&=\Pr(r=1|y=1)\Pr(q=1,y=1)\\
&=\text{TPR }\ \Pr(q=1,y=1)\\
&=k\ \ \Pr(q=1,y=1),
\end{aligned}$$

and then subtracting this from the marginal probability to get the top cell $p_{01}+p_{11}-kp_{11}=p_{01}+(1-k)p_{11}$. Equivalently for the left column.

Given these tables, we can then write the Jaccard index as

$$\begin{aligned}
j_{\{q=1,r=1\}}&=\frac{p_{11}k}{p_{11}k+p_{10}k+p_{01}+(1-k)p_{11}}\\
&=\frac{p_{11}}{p_{10}+\frac{p_{11}+p_{01}}{k}}\neq j_{\{q=1,y=1\}}
\end{aligned}\tag{18}$$

which is clearly biased. This is precisely the argument they give in the main text and supplement B.1.2, using set notation.[4] In contrast, the Forbes index

$$\begin{aligned}
f_{\{q=1,r=1\}}&=\frac{p_{11}k}{(p_{11}k+p_{10}k)(p_{11}k+p_{01}+(1-k)p_{11})}\\
&=\frac{p_{11}}{(p_{10}+p_{11})(p_{01}+p_{11})}=f_{\{q=1,y=1\}}
\end{aligned}\tag{19}$$

is unbiased. This is their main argument for dropping the Jaccard index and using the Forbes index. This argument is clearly different from ours. In supplement B.1.3, they also mention that ”[...]even in case of independence between the tracks, the Jaccard index tends to increase when the marginal probabilities increase (the tracks are more expressed) independently by the co-occurrence between them.” In other words, they say that as the marginal probabilities increase, the joint probability increases and thus also the Jaccard index. In contrast, we show that as the marginal probabilities change while the joint probability remains fixed, the Jaccard index changes.

[4] Equation (18) can be rewritten as $[p_{11}]\,/\,\left[p_{11}+p_{01}+p_{10}+\left(\frac{1}{k}-1\right)(p_{11}+p_{01})\right]$ to mimic their equation more closely.

5 Solution?

5.1 Forbes Index and Pointwise Mutual Information

Going back to equation 7, reprinted here for convenience,

$$p_{11} = \alpha(p_{01} + p_{11})(p_{10} + p_{11}) \tag{7}$$

it is of course obvious that the Forbes index and the pointwise mutual information read

$$f_{11} = \alpha \tag{20}$$

$$m_{11} = \log \alpha, \tag{21}$$

meaning that they are completely independent of the marginal probabilities $p_{01} + p_{11}$ and $p_{10} + p_{11}$.

We also note that the Forbes index and the pointwise mutual information can be expressed in terms of the covariance

$$\mathrm{Cov}(Y_1, Y_2) = p_{00}p_{11} - p_{01}p_{10}, \tag{22}$$

such that they read

$$f_{11} = \frac{p_{11}}{p_{11} - \mathrm{Cov}(Y_1, Y_2)} \tag{23}$$

$$m_{11} = -\log\left(1 - \frac{\mathrm{Cov}(Y_1, Y_2)}{p_{11}}\right). \tag{24}$$

In the bivariate Bernoulli case, zero-covariance implies independence (Dai et al., 2013). Thus, this form again highlights that, in the case of independence, the Forbes index equals one and the pointwise mutual information is zero.

6 Maths

6.1 Problem

$$\begin{aligned}
p_{11} &= \alpha(p_{11}+p_{01})(p_{11}+p_{10}) \\
&= \alpha(p_{11}^2+p_{11}p_{01}+p_{11}p_{10}+p_{01}p_{10}) \\
&= \alpha(p_{11}(p_{11}+p_{01}+p_{10})+p_{01}p_{10}) \\
&= \alpha(p_{11}(1-p_{00})+p_{01}p_{10}) \\
\alpha p_{01}p_{10} &= p_{11}-\alpha p_{11}(1-p_{00}) \\
&= p_{11}-\alpha p_{11}+\alpha p_{11}p_{00} \\
&= p_{11}-\alpha p_{11}+\alpha p_{11}(1-p_{11}-p_{01}-p_{10}) \\
&= p_{11}-\alpha p_{11}+\alpha p_{11}-\alpha p_{11}^2-\alpha p_{11}p_{01}-\alpha p_{11}p_{10} \\
\alpha p_{01}p_{10}+\alpha p_{11}p_{01} &= p_{11}-\alpha p_{11}^2-\alpha p_{11}p_{10} \\
p_{01} &= \frac{p_{11}(1-\alpha p_{11}-\alpha p_{10})}{\alpha p_{10}+\alpha p_{11}}
\end{aligned}$$

$$\begin{aligned}
d_{11} &= \frac{2p_{11}}{2p_{11}+p_{10}+p_{01}} \\
&= \frac{2p_{11}}{2p_{11}+p_{10}+\frac{p_{11}(1-\alpha p_{11}-\alpha p_{10})}{\alpha p_{10}+\alpha p_{11}}} \\
&= \frac{2p_{11}(\alpha p_{10}+\alpha p_{11})}{(2p_{11}+p_{10})(\alpha p_{10}+\alpha p_{11})+p_{11}(1-\alpha p_{11}-\alpha p_{10})} \\
&= \frac{2p_{11}(\alpha p_{10}+\alpha p_{11})}{(2p_{11}+p_{10})(\alpha p_{10}+\alpha p_{11})+p_{11}(1-(\alpha p_{11}+\alpha p_{10}))} \\
&= \frac{2p_{11}(\alpha p_{10}+\alpha p_{11})}{(2p_{11}+p_{10})(\alpha p_{10}+\alpha p_{11})+p_{11}-p_{11}(\alpha p_{11}+\alpha p_{10})} \\
&= \frac{2p_{11}(\alpha p_{10}+\alpha p_{11})}{(p_{11}+p_{10})(\alpha p_{10}+\alpha p_{11})+p_{11}} \\
&= \frac{2\alpha p_{11}(p_{10}+p_{11})}{\alpha(p_{11}+p_{10})(p_{10}+p_{11})+p_{11}} \\
&= \frac{2\alpha p_{11}(p_{10}+p_{11})}{\alpha(p_{10}+p_{11})^2+p_{11}}
\end{aligned}$$

$$
\begin{aligned}
j_{11} &= \frac{p_{11}}{p_{11}+p_{10}+p_{01}} \\
&= \frac{p_{11}}{p_{11}+p_{10}+\frac{p_{11}(1-\alpha p_{11}-\alpha p_{10})}{\alpha p_{10}+\alpha p_{11}}} \\
&= \frac{p_{11}(\alpha p_{10}+\alpha p_{11})}{(p_{11}+p_{10})(\alpha p_{10}+\alpha p_{11})+p_{11}(1-\alpha p_{11}-\alpha p_{10})} \\
&= \frac{p_{11}(\alpha p_{10}+\alpha p_{11})}{(p_{11}+p_{10})(\alpha p_{10}+\alpha p_{11})+p_{11}(1-(\alpha p_{11}+\alpha p_{10}))} \\
&= \frac{p_{11}(\alpha p_{10}+\alpha p_{11})}{(p_{11}+p_{10})(\alpha p_{10}+\alpha p_{11})+p_{11}-p_{11}(\alpha p_{11}+\alpha p_{10})} \\
&= \frac{p_{11}(\alpha p_{10}+\alpha p_{11})}{p_{10}(\alpha p_{10}+\alpha p_{11})+p_{11}} \\
&= \frac{\alpha p_{11}(p_{10}+p_{11})}{\alpha p_{10}(p_{10}+p_{11})+p_{11}}
\end{aligned}
$$

6.2 Solution?

$$
\begin{aligned}
f_{11} &= \frac{p_{11}}{(p_{01}+p_{11})(p_{10}+p_{11})} \\
&= \frac{p_{11}}{p_{11}(p_{11}+p_{01}+p_{10})(p_{01}p_{10})} \\
&= \frac{p_{11}}{p_{11}(1-p_{00})+p_{01}p_{10}} \\
&= \frac{p_{11}}{p_{11}-\mathrm{Cov}(Y_1,Y_2)} \\
m_{11} &= \log\frac{p_{11}}{p_{11}-\mathrm{Cov}(Y_1,Y_2)} \\
&= \log p_{11} - \log(p_{11}-\mathrm{Cov}(Y_1,Y_2)) \\
&= \log p_{11} - \left(\log p_{11} + \log\left(1-\frac{\mathrm{Cov}(Y_1,Y_2)}{p_{11}}\right)\right) \\
&= -\log\left(1-\frac{\mathrm{Cov}(Y_1,Y_2)}{p_{11}}\right)
\end{aligned}
$$