Automatic Classifiers as Scientific Instruments: One Step Further Away from Ground-Truth

Jacob Whitehill [1] **Anand Ramakrishnan** [1]

Abstract

Automatic machine learning-based detectors of various psychological and social phenomena (e.g., emotion, stress, engagement) have great potential to advance basic science. However, when a detector d is trained to approximate an existing measurement tool (e.g., a questionnaire, observation protocol), then care must be taken when interpreting measurements collected using d since they are one step further removed from the underlying construct. We examine how the accuracy of d, as quantified by the correlation q of d's outputs with the ground-truth construct U, impacts the estimated correlation between U (e.g., stress) and some other phenomenon V (e.g., academic performance). In particular: (1) We show that if the true correlation between U and V is r, then the expected sample correlation, over all vectors $\mathcal{T}^n$ whose correlation with U is q, is qr. (2) We derive a formula for the probability that the sample correlation (over n subjects) using d is positive given that the true correlation is negative (and vice-versa); this probability can be substantial (around $20-30\%$) for values of n and q that have been used in recent affective computing studies. (3) With the goal to reduce the variance of correlations estimated by an automatic detector, we show that training multiple neural networks $d^{(1)}, \ldots, d^{(m)}$ using different training architectures and hyperparameters for the same detection task provides only limited "coverage" of $\mathcal{T}^n$.

1. Introduction

Automatic classifiers have the potential to advance basic research in psychology, education, medicine, and many other fields by serving as *scientific instruments* that can measure behavioral, medical, social, and other phenomena with higher temporal resolution, lower cost, and greater consistency than is possible with traditional methods such as human-coded questionnaires or observation protocols. The affective computing (Picard, 2010) community is starting to see some first fruits of this potential: Perugia, et al. (Perugia et al., 2017) used the Empatica E4 wristband sensor to explore the relationship between participants' ($n = 14$) electrodermal activity (EDA) and their emotional states when playing cognitive games. Parra, et al. (Parra et al., 2017) used the Emotient facial expression recognition software to identify a positive correlation ($r = 0.32$, $n = 59$ participants) between emotions and adult attachment (Collins & Read, 1990). Chen, et al. (Chen et al., 2014) used Emotient in a study of how facial emotion is associated with job interview performance among $n = 4$ participants.

In most empirical studies designed to measure the relationship between two phenomena U and V (e.g., engagement (Monkaresi et al., 2017), grit (Duckworth et al., 2007), stress, attachment (Collins & Read, 1990), academic performance, etc.), the investigator chooses a validated instrument for each phenomenon and records measurements of each variable for n participants. She/he then computes a statistic, such as the Pearson product-moment coefficient, that captures the magnitude and sign (as well as statistical significance) of the relationship between the two variables. Machine learning offers the potential to create a new array of scientific instruments with important advantages compared to standard measurement tools. However, they also bring a potential pitfall that – while not fundamentally new, i.e., there is always a separation between a construct and its measurement – is exacerbated compared to using standard measurements: If one creates a new scientific instrument by training an automatic detector d to mimic a standard instrument as closely as possible, then d is *one degree of separation further removed from the underlying phenomenon* U – i.e., it is an estimator of another estimator.

Motivating example: Suppose a behavioral scientist wishes to examine the relationship between stress (construct U) and academic performance (construct V). Using a traditional approach, she/he could conduct an experiment

[1]Department of Computer Science, Worcester Polytechnic Institute (WPI), MA, USA. Correspondence to: Jacob Whitehill <jrwhitehill@wpi.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

in which each participant completes some cognitively demanding task and then takes a test. To measure stress, the scientist could also ask each participant to complete an established survey, e.g., the Dundee State Stress Questionnaire (Matthews et al., 1999). The relationship between U and V could then be estimated as the correlation $r = \rho(\mathbf{u}, \mathbf{v})$ between the vector of test scores $\mathbf{v}$ and the corresponding vector of stress measurements $\mathbf{u}$ over all n participants.

However, suppose that the researcher also has access to an *automatic stress detector* d that uses the participant's face pixels to measure his/her stress level. Suppose that the accuracy of d was previously validated w.r.t. a standard stress questionnaire (like (Matthews et al., 1999)), and the validation showed that the outputs of d, which we denote with $\widehat{\mathbf{u}}$, have an expected correlation of q with the standard questionnaire. What could go wrong, in terms of spurious deductions, when the correlation between U and V is estimated as $\rho(\widehat{\mathbf{u}}, \mathbf{v})$ instead of $\rho(\mathbf{u}, \mathbf{v})$?

Figure 1 shows one hypothetical example of what can go wrong: vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ contain measurements from n participants of constructs U and V, respectively, where $\mathbf{u}$ is obtained through a standard instrument. The Pearson product-moment correlation between two vectors can be written:

$$\rho(\mathbf{u}, \mathbf{v}) = \frac{(\mathbf{u} - \mu_{\mathbf{u}})^\top (\mathbf{v} - \mu_{\mathbf{v}})}{\|\mathbf{u} - \mu_{\mathbf{u}}\|_2 \|\mathbf{v} - \mu_{\mathbf{v}}\|_2}$$

where $\mu_{\mathbf{u}}$ (or $\mu_{\mathbf{v}}$) is a vector whose elements equal the mean value of $\mathbf{u}$ (or $\mathbf{v}$). Defined in this way, the value of ρ is random if either of its two arguments is random. If $\mathbf{u}$ and $\mathbf{v}$ are both normalized to have 0-mean and unit-length, then their correlation depends only on the *angle* between them:

$$\rho(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} = \cos \angle(\mathbf{u}, \mathbf{v})$$

In the figure, this correlation is $\cos(105°) \approx -.259$, i.e., the data suggest that U is *negatively* correlated with V. Suppose instead that the researcher had used an automatic detector d to obtain $\widehat{\mathbf{u}}$, where prior analysis had established that the expected correlation of d's outputs and the standard instrument was $q = \cos(30°) \approx 0.866$. If the researcher uses the correlation $\rho(\widehat{\mathbf{u}}, \mathbf{v})$ to estimate the relationship between U and V, then she/he would obtain $\cos(135°) = -0.707$ – a much larger magnitude, but at least the same sign as, the -0.259 correlation obtained using a standard instrument for U. But the bigger problem is the following: $\widehat{\mathbf{u}}$ is not the only vector whose correlation with the "ground-truth" measurements $\mathbf{u}$ is q. Vector $\widehat{\mathbf{u}}'$ also has the same correlation. If the researcher obtained measurements $\widehat{\mathbf{u}}'$, then she/he would deduce a *positive* correlation of $\rho(\widehat{\mathbf{u}}', \mathbf{v}) = \cos(75°) \approx 0.259$ – this is opposite to the correlation obtained with a standard instrument.

In this paper we explore how the accuracy q of a scientific instrument d, as measured by the Pearson correlation with the ground-truth construct U, impacts the estimated correlation between constructs U and V. Although there are various ways of quantifying the relationship between two vectors of measurements (e.g., RMSE, MAE), the Pearson correlation is one of the most commonly used metrics. **Contributions**: (1) We prove that $\mathrm{E}[\rho(\widehat{\mathrm{u}}, \mathbf{v})] = qr$, where r is the true correlation between U and V and random vector $\widehat{\mathrm{u}}$ is sampled uniformly over the $(n-3)$-sphere $\mathcal{T}^n$ of 0-mean unit-vectors whose correlation with $\mathbf{u}$ is q. Next, as one of the most fundamental aspects of the relationship between two variables is whether they are positively or negatively correlated, (2) we derive a function h to compute the probability that the sample correlation (over n subjects) using d is positive, given that the true correlation between U and V is negative (and vice-versa). We also prove that h is monotonically decreasing in n and in q, i.e., the danger of a false correlation is mitigated by training a more accurate detector or collecting data from more participants. Finally, (4) we explore to what extent the sphere $\mathcal{T}^n$ can be "covered" by measurement vectors $\widehat{\mathrm{u}}^{(1)}, \ldots, \widehat{\mathrm{u}}^{(m)}$ obtained by training m different neural networks on the same dataset for the same detection task but using different configurations (e.g., architectures, hyperparameters, etc.). We also devise a novel technique to visualize this coverage of $\mathcal{T}^n$.

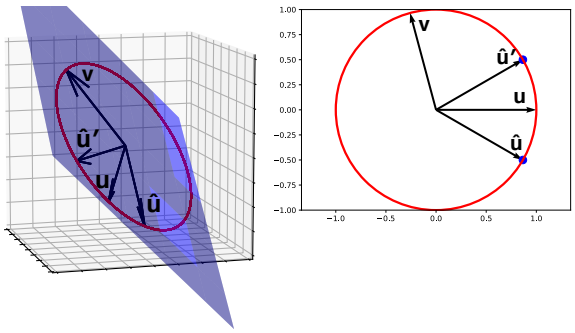

Figure 1. $\mathbf{u}$ and $\mathbf{v}$ are measurements of some behavioral, social, educational, (or other) phenomenoma for n participants in an experiment. $\widehat{\mathbf{u}}$ (or $\widehat{\mathbf{u}}'$) are proxy measurements of $\mathbf{u}$ that were obtained from an automatic detector. Both $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{u}}'$ have the same correlation ($r \approx 0.867$) with $\mathbf{u}$. However, depending on *which* vector is obtained from the detector, the estimated correlation can be very different: $\rho(\widehat{\mathbf{u}}, \mathbf{v}) < 0$, but $\rho(\widehat{\mathbf{u}}', \mathbf{v}) > 0$.

2. Related Work

The issue of how product-moment (Pearson) correlations among a subset of variables constrain the possible correlations among the remaining variables has interested statis-

ticians since the 1960s. While there has been significant prior work on the trivariate case in particular, we are not aware of any work that proves exactly the same results as what we present here. (Priest, 1968) showed a lower bound on the mean intercorrelation between variables. (Glass & Collins, 1970), and also (Leung & Lam, 1975), proved that, in trivariate distributions, there are range restrictions on the possible correlations between U and V when the correlations between V and W and between U and W are already known. (Olkin, 1981) extended this result to multivariate distributions beyond 3 variables.

More recent, and most similar to our work, is a study by (Carlson & Herdman, 2012) from the operations research community in 2012. They examined the methodological risk of using proxy measures to estimate the correlations between different constructs. Using analytical results by (Leung & Lam, 1975), they show how the observed correlations between $\mathbf{u}$ and $\widehat{\mathbf{u}}$ can vary substantially as a function of the reliability of a proxy measure $\widehat{\mathbf{u}}$ of $\mathbf{u}$. In contrast to our work, theirs is based on simulations and contains no formal proofs.

Finally, which vector of measurements $\widehat{\mathsf{u}} \in \mathcal{T}^n$ is obtained could potentially be related to subgroup membership – e.g., gender, ethnicity, age. From the perspective of *fairness* in machine learning (Barocas et al., 2017; Kearns et al., 2018), it could therefore be important to understand how much variation there is among different subgroup populations over the different $\widehat{\mathsf{u}}$ that are obtained from the scientific instrument.

3. Modelling assumptions

Notation: We typeset random variables in Futura font; all other variables are fixed (non-random). $\mathbf{u}$ and $\mathbf{v}$ are fixed n-vectors of ground-truth measurements of two phenomena U and V, respectively. $\hat{\mathsf{u}}$ is a random n-vector (sampled from $\mathcal{T}^n$), obtained from scientific instrument d, containing noisy measurements of $\mathbf{u}$.

For our theoretical results (Propositions 1 through 4), we assume the correlation between $\hat{\mathsf{u}}$ and $\mathbf{u}$ is exactly q; then, $\hat{\mathsf{u}}$ is sampled uniformly from the sphere $\mathcal{T}^n$ of all such vectors. However, for our empirical results regarding the probability of a "false correlation" (see Section 5), we relax this assumption: It is unlikely that instrument d (that we assume was previously estimated to have correlation q w.r.t. ground-truth) always produces a vector $\hat{\mathsf{u}}$ whose correlation with $\mathbf{u}$ is *exactly* q. Instead, the actual correlation $\hat{\mathsf{q}}$ of $\hat{\mathsf{u}}$ and $\mathbf{u}$ comes from a sampling distribution of Pearson correlations (with "true" correlation q and the number of subjects n as parameters; see Section 5.1). Then, given a fixed $\hat{\mathsf{q}}$, we sample $\hat{\mathsf{u}}$ from the corresponding $\mathcal{T}^n$ (which depends on $\hat{\mathsf{q}}$). When we compute the probability of a "false correlation" in our case studies (Section 5.2), we marginalize over $\hat{\mathsf{q}}$.

4. Expected Correlation of $\widehat{\mathsf{u}} \in \mathcal{T}^n$ with $\mathbf{v}$

When we use an automatic classifier d to obtain a vector of measurements, then we obtain a vector $\widehat{\mathbf{u}}$ whose correlation with the underlying construct U (e.g., stress) is q. However, as illustrated in the example above, there can be multiple such vectors, and which one is obtained can make a big difference on the estimated correlation. As we show below, the set of 0-mean unit-length vectors with a fixed correlation to another unit-vector is an $(n-3)$-sphere embedded in $\mathbb{R}^n$. If we sampled uniformly at random from this sphere, then what would be the expected sample correlation between $\widehat{\mathsf{u}}$ and some other vector $\mathbf{v}$ (e.g., academic performance)?

To simplify our analyses below, we assume $\mathbf{u}, \widehat{\mathbf{u}}, \mathbf{v}$ all have 0-mean and unit-length since Pearson correlation is invariant to these quantities. (Regarding the uniformity assumption: see Future Work in Section 7.)

Proposition 1. *Let $\mathbf{u}, \mathbf{v}$ be n-dimensional, 0-mean, unit-length vectors with a Pearson product-moment correlation $\rho(\mathbf{u}, \mathbf{v}) = r$. Then (1) the set $\mathcal{T}^n$ of 0-mean, unit-length vectors whose correlation with $\mathbf{u}$ is q is an $(n-3)$-sphere embedded in $\mathbb{R}^n$. Moreover, (2) if $\widehat{\mathsf{u}}$ is a random vector sampled uniformly from $\mathcal{T}^n$, then the expected sample correlation $E[\rho(\widehat{\mathsf{u}}, \mathbf{v})] = qr$.*

Proof. The set of all 0-mean n-vectors constitutes a hyperplane

$$\mathcal{H} \doteq \{\mathbf{x} \in \mathbb{R}^n : \mathbf{1}^\top \mathbf{x} = 0\}$$

that passes through the origin with normal vector $\mathbf{1} \doteq (1, \ldots, 1)$. The set of all unit-length vectors constitutes an $(n-1)$-sphere

$$\mathcal{S}^{n-1} \doteq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$$

embedded in $\mathbb{R}^n$. Therefore, $\mathbf{u}, \mathbf{v}, \widehat{\mathbf{u}} \in \mathcal{H} \cap \mathcal{S}^{n-1}$. Figure 1 (left) shows $\mathcal{H}$ in blue, as well as the intersection of $\mathcal{H}$ with $\mathcal{S}^{n-1}$ as a red circle. Since all three vectors have 0-mean and unit-length, then the correlations between these vectors depend only on the angles between them. Hence, w.l.o.g. we can rotate the axes so that $\mathcal{H}$ consists of all vectors whose first coordinate is 0, and all correlations will be preserved. After doing so, the only remaining constraint is that the projected vectors have unit-length.

More precisely, we can compute an orthonormal basis $\mathbf{B}$ such that the first coordinate of vector $\mathbf{Bx}$ is 0 for every $\mathbf{x} \in \mathcal{H}$, and such that

$$\mathbf{Bu} = (0, 1, \overbrace{0, 0, \ldots, 0}^{n-2}) \quad (1)$$

$$\mathbf{Bv} = (0, a, b, 0, \ldots, 0) \quad (2)$$

Geometrically, this means that we can define $\mathbf{B}$ so that the projected $\mathbf{u}, \mathbf{v}$ lie in the plane spanned by the second and third vectors in basis $\mathbf{B}$ (see Figure 1 (right)) – this makes the rest of the derivation much simpler. a represents the component of $\mathbf{v}$ parallel to $\mathbf{u}$, and b is the component orthogonal to $\mathbf{u}$. Since the correlation between $\mathbf{u}$ and $\mathbf{v}$ is r, and since $\mathbf{B}$ is orthonormal, then

$$\begin{aligned}(\mathbf{Bu})^\top(\mathbf{Bv}) &= \mathbf{u}^\top\mathbf{v} \\ &= r \\ &= 0\times 0 + 1\times a + 0\times b + 0 + \ldots + 0 \\ &= a\end{aligned}$$

and hence $a = r$. Since $\|\mathbf{v}\|_2 = \|\mathbf{Bv}\|_2 = 1$, then $b = \sqrt{1-r^2}$.

Now consider any vector $\widehat{\mathbf{u}}$ whose correlation with $\mathbf{u}$ is q. Let us define $(\hat{u}_1, \ldots, \hat{u}_n) \doteq \mathbf{B}\widehat{\mathbf{u}}$. By construction of $\mathbf{B}$, we already know that $\hat{u}_1 = 0$. We also have

$$\begin{aligned}(\mathbf{B}\widehat{\mathbf{u}})^\top(\mathbf{Bu}) &= \widehat{\mathbf{u}}^\top\mathbf{u} \\ &= q \\ &= \hat{u}_1\times 0 + \hat{u}_2\times 1 + \hat{u}_3\times 0 + \ldots + \hat{u}_n\times 0 \\ &= \hat{u}_2\end{aligned}$$

and hence $\hat{u}_2 = q$. Since $\mathbf{B}\widehat{\mathbf{u}}$ is a unit-vector, then

$$\mathbf{B}\widehat{\mathbf{u}} \in \left\{(0, q, \hat{u}_3, \ldots, \hat{u}_n) : \sum_{i=3}^{n}\hat{u}_i^2 = 1-q^2\right\}$$

This set is the surface of an $(n-3)$-sphere, with radius $\sqrt{1-q^2}$, embedded in $\mathbb{R}^n$. Since $\mathbf{B}$ simply rotates the axes, then $\mathcal{T}^n$ is likewise a $(n-3)$-sphere embedded in $\mathbb{R}^n$. This proves part 1.

For part 2: When sampling uniformly from $\mathcal{T}^n$, the distribution of $\hat{\mathrm{u}}_3$ on the $(n-3)$-sphere is symmetrical about 0. Then $\mathrm{E}[\hat{\mathrm{u}}_3] = 0$, and hence:

$$\begin{aligned}\mathrm{E}[\rho(\widehat{\mathrm{u}}, \mathbf{v})] &= \mathrm{E}[\widehat{\mathrm{u}}^\top\mathbf{v}] \\ &= \mathrm{E}[(\mathbf{B}\widehat{\mathrm{u}})^\top(\mathbf{Bv})] \\ &= \mathrm{E}[0 + q\times r + \hat{\mathrm{u}}_3\times\sqrt{1-r^2}] \\ &= qr + \mathrm{E}[\hat{\mathrm{u}}_3]\sqrt{1-r^2} \\ &= qr\end{aligned}$$

□

Example: For the case $n = 3$, consider the four vectors shown in Figure 1 (left) whose values are approximately:

$$\mathbf{u} = (.816, -.408, -.408) \quad \mathbf{v} = (-.211, -.577, .788)$$
$$\widehat{\mathbf{u}} = (.707, 0, -.707) \quad \widehat{\mathbf{u}}' = (.707, -.707, 0)$$

By construction, $\rho(\widehat{\mathbf{u}}, \mathbf{u}) = \rho(\widehat{\mathbf{u}}', \mathbf{u}) = \cos(30°) = q$, and $\rho(\mathbf{u}, \mathbf{v}) = \cos(105°) = r$. Via a change of basis $\mathbf{B}$, the vectors can be rotated so that

$$\begin{aligned}\mathbf{Bu} &= (0, 1, 0) \\ \mathbf{Bv} &= (0, \cos(105°), \sin(105°)) \\ \mathbf{B}\widehat{\mathbf{u}} &= (0, \cos(30°), -\sin(30°)) \\ \mathbf{B}\widehat{\mathbf{u}}' &= (0, \cos(30°), \sin(30°))\end{aligned}$$

The rotated vectors are shown in Figure 1 (right). The set $\mathcal{T}^3$ contains exactly two elements (since it is a 0-sphere): $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{u}}'$. If $\widehat{\mathrm{u}}$ is sampled uniformly at random from $\mathcal{T}^3$, then $\mathrm{E}[\rho(\widehat{\mathrm{u}}, \mathbf{v})] = qr \approx -.224$. This result agrees with

$$\begin{aligned}\frac{1}{2}\left[\rho(\widehat{\mathbf{u}}, \mathbf{v}) + \rho(\widehat{\mathbf{u}}', \mathbf{v})\right] &= \frac{1}{2}\left[\cos(135°) + \cos(75°)\right] \\ &\approx -.224\end{aligned}$$

5. Probability of false correlations

One of the most fundamental distinctions is whether two phenomena are positively or negatively correlated with each other (or neither). What is the probability that $\rho(\widehat{\mathrm{u}}, \mathbf{v}) \geq 0$ given that the correlation $r < 0$ (*false positive correlation*); or that $\rho(\widehat{\mathrm{u}}, \mathbf{v}) < 0$ given that the correlation $r \geq 0$ (*false negative correlation*)? How do these probabilities change as n increases or q increases? The proofs of the following propositions are given in the supplementary materials.

Proposition 2. *Let $q \in (0, 1]$ be the correlation between the detector's output $\widehat{\mathrm{u}}$ and ground-truth $\mathbf{u}$; let r be the correlation between $\mathbf{u}$ and $\mathbf{v}$; and let $\widehat{\mathrm{u}}$ be sampled uniformly from $\mathcal{T}^n$. If $r < 0$, then the probability of a* false positive *correlation (in the sense defined above) is given by the function*

$$h(n, q, r) = \begin{cases} \frac{1}{2}I[c^2 \leq 1-q^2] & n = 3 \\ \frac{1}{2}\int_0^\infty f_1(t)F_{n-3}\left(\frac{1-q^2-c^2}{c^2}t\right)dt & n > 3 \end{cases}$$

where $I[\cdot]$ is the 0/1 indicator function, f_k and F_k are the PDF and CDF of a χ^2-random variable with k degrees of freedom, and $c = |qr|/\sqrt{1-r^2}$. If $r > 0$, then the probability of a false negative *correlation is also given by h.*

Proposition 3. *For every fixed $c > 0$ and $q \in (0, 1]$, function h is monotonically decreasing in n.*

Proposition 4. *For every fixed $n > 3$, function h is monotonically decreasing in $q \in (0, 1]$.*

Propositions 3 and 4 imply that the probability of a false correlation diminishes as n increases or q increases.

5.1. Marginalizing over the sampling distribution of $\hat{\mathrm{q}}$

Up to now we have glossed over the important detail that, when a scientific instrument with *average* accuracy (Pearson correlation) of q is used to obtain measurements for n

subjects, the correlation of the sampled $\hat{\mathbf{u}}$ with their ground-truth values need not be exactly q; rather, the actual correlation $\hat{\mathsf{q}}$ is drawn from a sampling distribution $\Pr(\hat{\mathsf{q}} \mid q, n)$. Particularly for small n, $\hat{\mathsf{q}}$ can deviate substantially from q. Hence, to compute the probability of a false correlation, it is necessary to marginalize (via numeric integration) over $\hat{\mathsf{q}}$:

$$\int_{\hat{\mathsf{q}}} h(n, \hat{\mathsf{q}}, r) \Pr(\hat{\mathsf{q}} \mid q, n) d\hat{\mathsf{q}}$$

The sampling distribution of $\hat{\mathsf{q}}$ can be estimated using the formula derived by (Soper, 1913); see the supplementary materials for more details.

5.2. Case Studies

To put these theoretical results into perspective, we conducted simulations based on two recent affective computing studies that used automated detectors as scientific instruments. The first study ($n = 14$), by (Perugia et al., 2017), used an Empatica E4 wrist sensor to investigate how electrodermal activity (EDA) (U) is correlated with the subjects' emotions (V). The second study ($n = 20$), by (Whitehill et al., 2014), explored the relationship between student engagement (U), as measured by an engagement detector that analyzes static images of students' faces, and test performance (V) in a cognitive skills training task. In order to estimate the probability of a false correlation (in the sense described above), we need to know the accuracy of the automatic detector – i.e., the correlation q between the automatic measurements and ground-truth of construct U – as well as the *true* correlation r between constructs U and V.

Estimating q and r: The value of q can easily be estimated using cross-validation or other standard procedures. For the first study (EDA), we use the value $q = 0.57$ reported in (Poh et al., 2010) for cognitive tasks with a distill forearm sensor of EDA. The value of r is not knowable without access to ground-truth measurements; instead, we hypothesize that the ground-truth correlation between U and V is exactly what was estimated by the authors (Perugia et al., 2017) using the E4 sensor and emotion survey instruments: $r = -.497$. For the second study (Engagement), we use the value $q = 0.50$ reported in (Whitehill et al., 2014) that was obtained using subject-independent cross-validation. For r, we use the correlation obtained by the authors ($r = 0.37$) when correlating test performance with *human*-labeled student engagement.

Results: Plots of the probability of a false correlation (obtained from function h derived above) as a function of the number of participants n are shown for each study (with their associated q and r values) in Figure 2. The red dot in each graph shows the actual number of participants from each experiment. Even for $n \geq 100$ subjects, the probability is non-trivial. For the values $n = 14$ and $n = 20$, these probabilities are substantial – around 20% for the EDA study and around 30% for the Engagement study. **The possibility of a false correlation is *not* protected against by statistical significance testing** – it is possible for the estimated correlation between constructs U and V to be highly significant and yet have the wrong sign compared to the ground-truth correlation. While this is almost always theoretically possible due to the inherent separation between a construct and its measurement, the use in basic research of automatic detectors that are trained to estimate another estimator can make this problem worse.

6. Coverage of $\mathcal{T}^n$ when training a detector

Given that *which* vector $\hat{\mathbf{u}} \in \mathcal{T}^n$ is obtained from an automatic detector d can substantially impact the estimated correlation $\rho(\hat{\mathbf{u}}, \mathbf{v})$ between constructs U and V, it could be useful to *average* the sample correlations over *many* vectors $\hat{\mathbf{u}}^{(1)}, \ldots, \hat{\mathbf{u}}^{(m)}$ from $\mathcal{T}^n$, i.e., to compute $\frac{1}{m} \sum_{i=1}^{m} \rho(\hat{\mathbf{u}}^{(i)}, \mathbf{v})$. This could help to reduce the variance of the estimator. In this section we explore whether it is feasible to generate many different $\hat{\mathbf{u}}^{(1)}, \ldots, \hat{\mathbf{u}}^{(m)}$, all with similar correlation q with ground-truth $\mathbf{u}$, by training a set of automatic detectors $d^{(1)}, \ldots, d^{(m)}$ using slightly different training configurations. In particular, we varied: (1) the architecture (VGG-16 (Simonyan & Zisserman, 2014) versus ResNet-50 (He et al., 2016)) and (2) the random seed of weight initialization. Inspired by recent work by Huang, et al. (Huang et al., 2017) on how an entire ensemble of detectors can be created during a *single* training run, we also varied (3) the number of training epochs, and saved snapshots of the trained detectors at regular intervals.

During training, each detector's estimates $\hat{\mathbf{u}}$ of the test labels evolves, and so does the correlation between $\hat{\mathbf{u}}$ and the ground-truth labels $\mathbf{u}$. However, for the tasks we examined (described below), the test correlations tend to stabilize over time, and they converge to roughly the same value even across different training runs and detection architectures. Given a set of measurement vectors $\hat{\mathbf{u}}^{(1)}, \ldots, \hat{\mathbf{u}}^{(m)}$ (produced by detectors $d^{(1)}, \ldots, d^{(m)}$) whose accuracies (Pearson correlations with ground-truth) are all approximately q, we can project them onto the $(n-3)$-sphere $\mathcal{T}^n$ of 0-mean, unit-length vectors whose correlation with $\mathbf{u}$ is q. We can then visualize the "coverage" of this sphere by projecting it onto a 2-D plane, and also compare the coverage to a random sample of m elements of $\mathcal{T}^n$.

We explored the coverage of $\mathcal{T}^n$ for two automatic face analysis problems: student engagement recognition and age estimation using the HBCU (Whitehill et al., 2014) (Engagement) and GENKI (Lab) (Age) datasets, respectively; see Figure 3 for labeled examples.

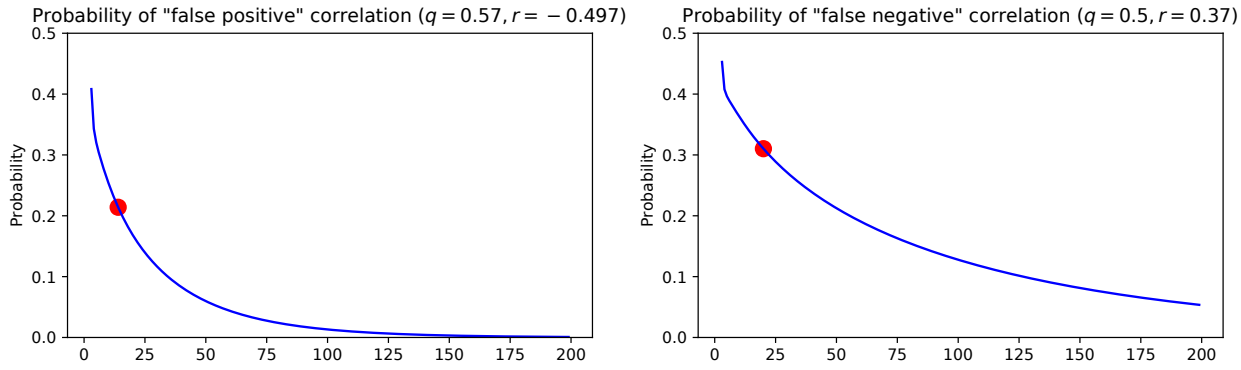

Figure 2. Probability of false correlation, for fixed q and r, as a function of n. The probability decreases as n grows, but for small n it can still be substantial. The red dots indicate the n from two recent behavioral studies that used an automatic detector of affective state as a scientific instrument. **Left**: Example inspired by a study on electrodermal activity (Perugia et al., 2017). **Right**: Example inspired by a study on student engagement (Whitehill et al., 2014).

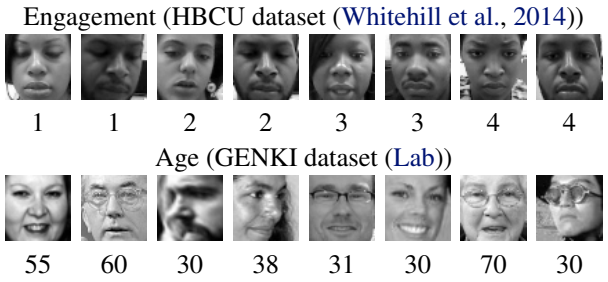

Figure 3. Examples of face images used in the experiments in Section 6.5. **Top**: student engagement dataset (Whitehill et al., 2014), in which engagement is rated on a 1-4 scale. **Bottom**: images from the GENKI (Lab) dataset that were labeled with their perceived age (in years).

6.1. Detection architectures

We examined two modern deep learning-based visual recognition architectures – VGG-16 (Simonyan & Zisserman, 2014) and ResNet-50 (He et al., 2016) – to assess how much "coverage" of $\mathcal{T}^n$ each architecture can produce, as well as to compare the variability of the resulting measurement vectors to each other.

6.2. Training procedures

Engagement detector: We performed 25 training runs for each of the two network architectures (VGG-16, ResNet-50). Training data consisted of 7629 face images from 15 subjects of HBCU (Whitehill et al., 2014), and testing data were 500 images from the remaining 5 subjects. (This corresponds to just one cross-validation fold from the original study (Whitehill et al., 2014).) Optimization was performed using SGD for 10000 iterations, and the network weights were saved every 1000 iterations. In total, this produced 250 detectors. The average correlation (over all 250 detectors) between the detectors' test outputs $\widehat{\mathbf{u}}$ and ground-truth $\mathbf{u}$ was 0.61 (s.d. 0.081) for VGG-16 and 0.64 (s.d. 0.009) for ResNet-50. Inspired by (Huang et al., 2017), we also tried both cosine and triangular (Smith, 2017) learning rates. However, in pilot testing we found that these delivered worse accuracy than exponential learning rate decay and we abandoned the approach.

Age detector: We performed 25 training runs for VGG-16 and ResNet-50 using SGD for 10000 with snapshots every 1000 iterations, as for engagement recognition. This produced 250 detectors. Training data consisted of 31040 face images of the GENKI dataset (Lab), and testing data consisted of 500 face images. The average correlation of the automatic measurements with ground-truth on the test set was 0.595 (s.d. 0.036) for VGG-16 and 0.60 (s.d. 0.014) for ResNet-50.

6.3. Visualizing elements of the $(n-3)$-sphere $\mathcal{T}^n$

Given a set of m trained detectors, we can sample vectors of age/engagement estimates $\widehat{\mathbf{u}}^{(1)}, \ldots, \widehat{\mathbf{u}}^{(m)} \in \mathbb{R}^n$ whose correlation with ground-truth $\mathbf{u}$ is approximately q. Then we can visualize how these vectors "cover" the $(n-3)$-sphere $\mathcal{T}^n$ using the following procedure:

1. Normalize $\mathbf{u}$, as well as each $\widehat{\mathbf{u}}^{(j)}$, to have 0-mean and unit-length.

2. Compute an orthonormal basis $\mathbf{B}$ (e.g., using a QR decomposition) so that (a) the first component of $\mathbf{Bx}$ is 0 for every $\mathbf{x} \in \mathcal{H}$, and (b) $\mathbf{Bu} = (0, 1, 0, 0, \ldots, 0)$ (see Equation 1).

3. Project each $\widehat{\mathbf{u}}^{(j)}$ onto the new basis $\mathbf{B}$. By construction, the first component of each projection will be 0 and the second component will be $q = \rho(\widehat{\mathbf{u}}^{(j)}, \mathbf{u})$.

4. Define each $\mathbf{x}^{(j)}$ to be the last $n-2$ components of vector $\mathbf{B}\widehat{\mathbf{u}}^{(j)}$.

5. Project the $\{\mathbf{x}^{(j)}\}$ onto the two principal axes obtained from principal component analysis (PCA).

Since the 2-D projection of a 0-centered sphere onto any orthonormal projection is a disc, the output of the procedure above is a set of points that lie on a disc of radius $\sqrt{1-q^2}$.

6.4. Generating random vectors on $\mathcal{T}^n$

In order to assess how evenly the sphere $\mathcal{T}^n$ is "covered" by the vectors obtained from the automatic detectors, we can generate random vectors of $\mathcal{T}^n$ and likewise project them onto a 2-D disc. We generate each such vector as follows:

1. Sample each $\mathbf{z}_i$ ($i = 1, \ldots, n-2$) from a standard normal distribution.

2. For $i = 1, \ldots, n-2$, set $\hat{\mathbf{u}}_i = \frac{\sqrt{1-q^2} \times \mathbf{z}_i}{\sqrt{\sum_{i'=1}^{n-2} \mathbf{z}_{i'}^2}}$.

We then project the vectors in the set $\{\hat{\mathbf{u}}^{(j)}\}$ onto the two principal axes obtained from PCA. To enable a fair comparison between the variances of the randomly generated elements of $\mathcal{T}^n$ and those obtained from the trained detectors, we run PCA *separately* for each set.

6.5. Results

We projected all vectors $\{\widehat{\mathbf{u}}^{(j)}\}$ whose correlation with $\mathbf{u}$ was between 0.575 and 0.625 for Engagement recognition and 0.625 and 0.65 and Age estimation; this amounted to 50% of the engagement detectors and 45% of the age detectors. The projections are shown for each task in Figure 4. First, we observe that there is some "spread" – the measurement vectors occupy different clusters on the sphere. This indicates that the same training data can still yield automatic measurements $\widehat{\mathbf{u}}$ on testing data whose correlations with each other is far less than 1. In fact, for engagement recognition, the minimum correlation, over all pairs $(\widehat{\mathbf{u}}^{(j)}, \widehat{\mathbf{u}}^{(j')})$, was 0.64.

For engagement recognition, the VGG-16 based measurements and the ResNet-50 based measurements each resided within their own clusters on the sphere, and these clusters did not overlap. This suggests that, even though both architectures yielded similar overall accuracies, they are making different kinds of estimation errors on the test set. Interestingly for both age estimation and engagement recognition we can see that VGG-16 has a bigger "spread" compared to ResNet-50. We speculate this might be due to VGG-16 (138 million) having significantly more parameters compared to ResNet-50 (25 million), thus enabling it to learn more varied features.

Finally, a comparison of the variance between the automatic measurements $\widehat{\mathbf{u}}^{(1)}, \ldots, \widehat{\mathbf{u}}^{(m)}$ and random samples from $\mathcal{T}^n$ indicates that varying the training configuration (architecture, hyperparameters) provides only limited ability to cover the sphere: the variance in the vectors, as quantified as the sum of the trace of their covariance matrix, was statistically significantly less compared to randomly sampled points on $\mathcal{T}^n$ ($p < 0.01$, 1-tailed, Monte Carlo simulation).

7. Conclusions

Advances in machine perception present a powerful opportunity to create new scientific instruments that can benefit basic research in sociobehavioral sciences. However, since detectors are often trained to estimate existing measures, which are already only an estimate of underlying constructs, then these instruments are essentially one step further removed from ground-truth. For this reason, it is important to interpret results obtained with them with care.

In this paper, we investigated how measurements of construct U obtained with an automatic detector can impact the estimated correlation between U and another construct V. We showed that: (1) The set of 0-mean unit-length n-vectors with a fixed Pearson product-moment correlation q to vector $\mathbf{u}$ is a $(n-3)$-sphere $\mathcal{T}^n$ embedded in $\mathbb{R}^n$. (2) If the correlation between automatic measurements $\widehat{\mathrm{u}}$ and the ground-truth measurements is q; if the true correlation between U and V is r; and if $\widehat{\mathrm{u}}$ is sampled uniformly from $\mathcal{T}^n$; then the expected sample correlation obtained with the automatic detector is qr. (3) The probability of a "false correlation", i.e., a sample correlation between constructs U and V whose sign differs from the true correlation, is monotonically decreasing in n (number of participants) and also monotonically decreasing in q (accuracy of the detector). These probabilities can be non-trivial for small values of n that are nonetheless sometimes found in contemporary research using automatic facial expression and affect detectors. Moreover, the danger of a false correlation is not eliminated through statistical significance testing. (4) We explored empirically how efficiently multiple neural network-based detectors of age and student engagement, when trained using different architectures and hyperparameters but the same training data, can "cover" the sphere $\mathcal{T}^n$.

In practice, our results suggest that, particularly when the number of participants is small and/or the accuracy of the detector is modest, it is important to consider the possibility of a false correlation, or at least a skewed correlation (by factor q), when drawing scientific conclusions.

Limitation and future work: In our study we assumed that $\widehat{\mathrm{u}}$ is a random sample from the uniform distribution over $\mathcal{T}^n$ – this expresses the idea that *a priori* we may have no idea which *particular* element of $\mathcal{T}^n$ detector d will return. In

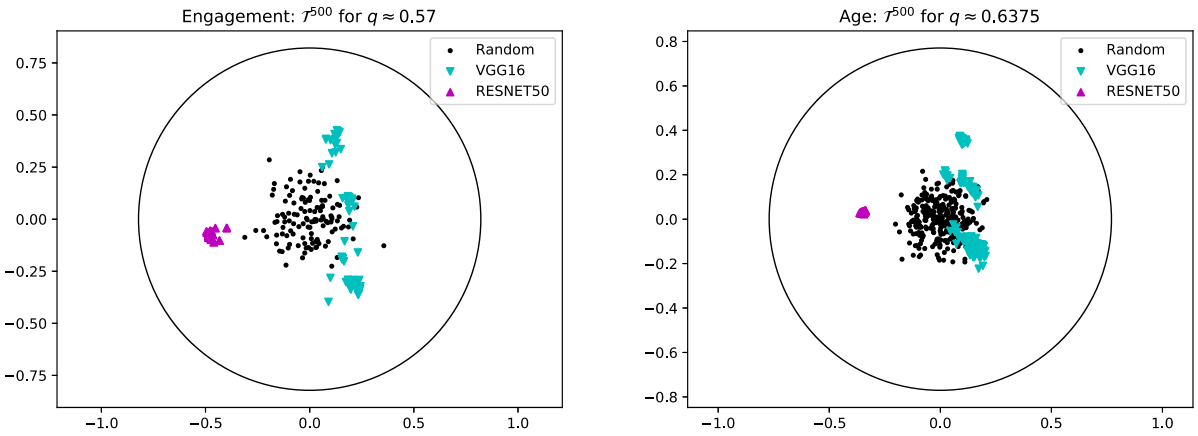

Figure 4. Coverage of the $\mathcal{T}^n$ sphere (projected onto 2 dimensions using PCA) from different neural networks trained to predict student engagement (**left**) and age (**right**) from face images. We used either VGG-16 or ResNet-50 neural networks. For comparison, we also sampled random vectors using the procedure from Section 6.4.

reality, however, detectors have biases – e.g., due to head pose, lighting conditions, training set composition, etc. – and these can affect which element of $\mathcal{T}^n$ is obtained.

Acknowledgements: This research was supported by a Cyberlearning grant from the National Science Foundation (grant no. #1822768).

References

Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. In *Conference on Neural Information Processing Systems, Long Beach, CA*, 2017.

Carlson, K. D. and Herdman, A. O. Understanding the impact of convergent validity on research results. *Organizational Research Methods*, 15(1):17–32, 2012.

Chen, L., Yoon, S.-Y., Leong, C. W., Martin, M., and Ma, M. An initial analysis of structured video interviews by using multimodal emotion detection. In *Proceedings of the 2014 workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems*, pp. 1–6. ACM, 2014.

Collins, N. L. and Read, S. J. Adult attachment, working models, and relationship quality in dating couples. *Journal of personality and social psychology*, 58(4):644, 1990.

Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92(6):1087, 2007.

Glass, G. V. and Collins, J. R. Geometric proof of the restriction on the possible values of rxy when r xz and ryz are fixed. *Educational and Psychological Measurement*, 30(1):37–39, 1970.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. An empirical study of rich subgroup fairness for machine learning. *arXiv preprint arXiv:1808.08166*, 2018.

Lab, M. P. The MPLab GENKI Database. `http://mplab.ucsd.edu`.

Leung, C.-K. and Lam, K. A note on the geometric representation of the correlation coefficients. *The American Statistician*, 29(3):128–130, 1975.

Matthews, G., Joyner, L., Gilliland, K., Campbell, S., Falconer, S., and Huggins, J. Validation of a comprehensive stress state questionnaire: Towards a state big three. *Personality psychology in Europe*, 7:335–350, 1999.

Monkaresi, H., Bosch, N., Calvo, R. A., and D'Mello, S. K. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2017.

Olkin, I. Range restrictions for product-moment correlation matrices. *Psychometrika*, 46(4):469–472, 1981.

Parra, F., Miljkovitch, R., Persiaux, G., Morales, M., and Scherer, S. The multimodal assessment of adult attachment security: developing the biometric attachment test. *Journal of medical Internet research*, 19(4), 2017.

Perugia, G., Rodríguez-Martín, D., Boladeras, M. D., Mallofré, A. C., Barakova, E., and Rauterberg, M. Electrodermal activity: explorations in the psychophysiology of engagement with social robots in dementia. In *Robot and Human Interactive Communication (RO-MAN), IEEE International Symposium on*, pp. 1248–1254, 2017.

Picard, R. W. Affective computing: from laughter to ieee. *IEEE Transactions on Affective Computing*, 1(1):11–17, 2010.

Poh, M.-Z., Swenson, N. C., and Picard, R. W. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE transactions on Biomedical engineering*, 57(5):1243–1252, 2010.

Priest, H. F. Range of correlation coefficients. *Psychological reports*, 22(1):168–170, 1968.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Smith, L. N. Cyclical learning rates for training neural networks. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pp. 464–472, 2017.

Soper, H. On the probable error of the correlation coefficient to a second approximation. *Biometrika*, 9(1/2):91–115, 1913.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., and Movellan, J. R. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.