LETTER

R^2-equitability is satisfiable

Kinney and Atwal (1) make excellent points about mutual information, the maximal information coefficient (2, 3), and "equitability." One of their central claims, however, is that, "No nontrivial dependence measure can satisfy R^2-equitability." We argue that this is the result of a poorly constructed definition, which we quote:

> "A dependence measure $D[X;Y]$ is R^2-equitable if and only if, when evaluated on a joint probability distribution $p(X,Y)$ that corresponds to a noisy functional relationship between two real random variables X and Y, the following relation holds:
>
> $$D[X;Y] = g\left(R^2[f(X);Y]\right).$$
>
> Here, g is a function that does not depend on $p(X,Y)$ and f is the function defining the noisy functional relationship, i.e.,
>
> $$Y = f(X) + \eta,$$
>
> for some random variable η. The noise term η may depend on $f(X)$ as long as η has no additional dependence on X. . ."

This definition is undone by the unconventional specification of the noise term. Specifically, allowing η to depend arbitrarily on $f(X)$ lets many different combinations of f and η result in the same $p(X,Y)$. For example, consider $f_1(X) = X^2$ and $\eta_1 = \mathcal{N}(0,1)$, against $f_2(X) = X$ and $\eta_2 = -f_2(X) + f_2(X)^2 + \mathcal{N}(0,1)$. The resulting $p(X,Y)$ distributions are identical, but $R^2[f_1(X);Y] \neq R^2[f_2(X);Y]$—a consequence of the deterministic trend embedded in η_2.

We emphasize the cause of the definitional deficiency (which the authors exploit to demonstrate unsatisfiability) because it suggests an immediate fix: make η trendless. By constraining the expectation $E[\eta|f(X)] = 0$, the identifiability issue is resolved without limiting expressive power: any trend removed from η can, and should, be included in $f(X)$ instead. Under this formulation, we also see no reason to restrict the dependence of η to $f(X)$ alone; it can depend arbitrarily on X, as long as $E[\eta|X] = 0$.

Without a trend in η, not only does the resulting definition of R^2-equitability escape Kinney and Atwal's *reductio*, but it is demonstrably satisfiable. Because $E[\eta|X] = 0 \Rightarrow f(X) = E[Y|X]$, $R^2[f(X);Y]$ is determined by $p(X,Y)$, satisfying the modified definition with g as the identity function. Further, in the large sample limit (for nonpathological functions), $\hat{f}(X) \approx f(X)$ is estimable from X, Y, yielding increasingly accurate approximations of $R^2[\hat{f}(X);Y] \approx R^2[f(X);Y]$, suggesting a family of schemes for nonparametric estimation of $D[X;Y]$ that satisfy R^2-equitability.

R^2-equitable measures of dependence care only about how accurately Y can be predicted—under a quadratic loss function—by X and are thus sensitive to nonlinear transformations of Y and not symmetric ($D[X;Y] \neq D[Y;X]$), in contrast to any dependence measure satisfying Kinney and Atwal's self-equitability (1). These two distinct notions of equitability are useful in different circumstances: R^2-equitability should be preferred when quantifying how well you can predict an outcome in expectation (measuring your least-squares predictive accuracy), and measures satisfying self-equitability (exemplified by mutual information) may be more appropriate when quantifying how well you can predict Y in probability, being sensitive to how the distribution $p(Y|X)$ varies with X.

Thus, a simple modification of Kinney and Atwal's definition renders a satisfiable notion of R^2-equitability that is usefully distinct from the notion of self-equitability the authors propose (1). Both can coexist.

ACKNOWLEDGMENTS. B.M. is supported by Center for AIDS Research Translational Virology Core Grant P30 AI036214 and Molecular Epidemiology Avant Garde Grant DP1 DA034978.

***Ben Murrell*[a,1], *Daniel Murrell*[b], *and Hugh Murrell*[c]**

[a]Department of Medicine, University of California, San Diego, La Jolla, CA 92093; [b]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom; and [c]Computer Science, University of KwaZulu-Natal, Pietermaritzburg 3201, South Africa

1 Kinney JB, Atwal GS (2014) Equitability, mutual information, and the maximal information coefficient. *Proc Natl Acad Sci USA* 111(9): 3354–3359.

2 Reshef DN, et al. (2011) Detecting novel associations in large data sets. *Science* 334(6062):1518–1524.

3 Reshef DN, Reshef Y, Mitzenmacher M, Sabeti P (2013) Equitability analysis of the maximal information coefficient with comparisons. arXiv:1301.6314v1 [cs.LG].

Author contributions: B.M., D.M., and H.M. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: bmurrell@ucsd.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1403623111