DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Cleaning up the record on the maximal information coefficient and equitability

The Harvard community has made this article openly available. **Please share** how this access benefits you. Your story matters

Citation	Reshef, D. N., Y. A. Reshef, M. Mitzenmacher, and P. C. Sabeti. 2014. "Cleaning up the Record on the Maximal Information Coefficient and Equitability." Proceedings of the National Academy of Sciences 111 (33) (August 19): E3362–E3363. doi:10.1073/pnas.1408920111. http://dx.doi.org/10.1073/pnas.1408920111.
Published Version	doi:10.1073/pnas.1408920111
Citable link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:26569221
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Cleaning up the record on the maximal information coefficient and equitability

Although we appreciate Kinney and Atwal's interest in equitability and maximal information coefficient (MIC), we believe they misrepresent our work. We highlight a few of our main objections below.

Regarding our original paper (1), Kinney and Atwal (2) state "MIC is said to satisfy not just the heuristic notion of equitability, but also the mathematical criterion of R^2 equitability," the latter being their formalization of the heuristic notion that we introduced. This statement is simply false. We were explicit in our paper that our claims regarding MIC's performance were based on large-scale simulations: "We tested MIC's equitability through simulations....[These] show that, for a large collection of test functions with varied sample sizes, noise levels, and noise models, MIC roughly equals the coefficient of determination R^2 relative to each respective noiseless function." Although we mathematically proved several things about MIC, none of our claims imply that it satisfies Kinney and Atwal's R^2 equitability, which would require that MIC exactly equal R^2 in the infinite data limit. Thus, their proof that no dependence measure can satisfy R^2 equitability, although interesting, does not uncover any error in our work, and their suggestion that it does is a gross misrepresentation.

Kinney and Atwal seem ready to toss out equitability as a useful criterion based on their theoretical result. We argue, however, that regardless of whether "perfect" equitability is possible, approximate notions of equitability remain the right goal for many data exploration settings. Just as the theory of NP completeness does not suggest we stop thinking about NP complete problems, but

instead that we look for approximations and solutions in restricted cases, an impossibility result about perfect equitability provides focus for further research, but does not mean that useful solutions are unattainable. Similarly, as others have noted (3), Kinney and Atwal's proof requires a highly permissive noise model, and so the attainability of R^2 equitability under more limited noise models such as those in our work remains an open question.

Finally, the authors argue that mutual information is more equitable than MIC. However, they provide as justification only a single noise model, only at limiting sample sizes ($n \geq 5,000$). As we've shown in follow-up work (4), which they themselves cite but fail to address, MIC is more equitable than mutual information estimation under many other realistic noise models even at a sample size of 5,000. Kinney and Atwal have stated, "...it matters how one defines noise" (5), and a useful statistic must indeed be robust to a wide range of noise models. Equally importantly, we've established in both our original and follow-up work that at sample size regimes less than 5,000, MIC is more equitable than mutual information estimates across all noise models tested. MIC's superior equitability in these settings is not an "artifact" we neglected—as Kinney and Atwal suggest—but rather a weakness of mutual information estimation and an important consideration for practitioners.

We expect that the understanding of equitability and MIC will improve over time and that better methods may arise. However, accurate representations of the work thus

far will allow researchers in the area to most productively and collectively move forward.

David N. Reshef[a,b,1,2], **Yakir A. Reshef**[b,1,2], **Michael Mitzenmacher**[c,3], **and Pardis C. Sabeti**[d,e,3]

[a]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Boston, MA 02139;* [b]*Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School and MIT, Cambridge, MA 02139;* [c]*School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138;* [d]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; and* [e]*Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142*

1 Reshef DN, et al. (2011) Detecting novel associations in large data sets. *Science* 334(6062):1518–1524.
2 Kinney JB, Atwal GS (2014) Equitability, mutual information, and the maximal information coefficient. *Proc Natl Acad Sci USA* 111(9): 3354–3359.
3 Murrell B, Murrell D, Murrell H (2014) R^2-equitability is satisfiable. *Proc Natl Acad Sci USA* 111(21):E2160.
4 Reshef DN, Reshef YA, Mitzenmacher M, Sabeti PC (2013) Equitability analysis of the maximal information coefficient with comparisons. arXiv:1301.6314v2 [cs.LG].
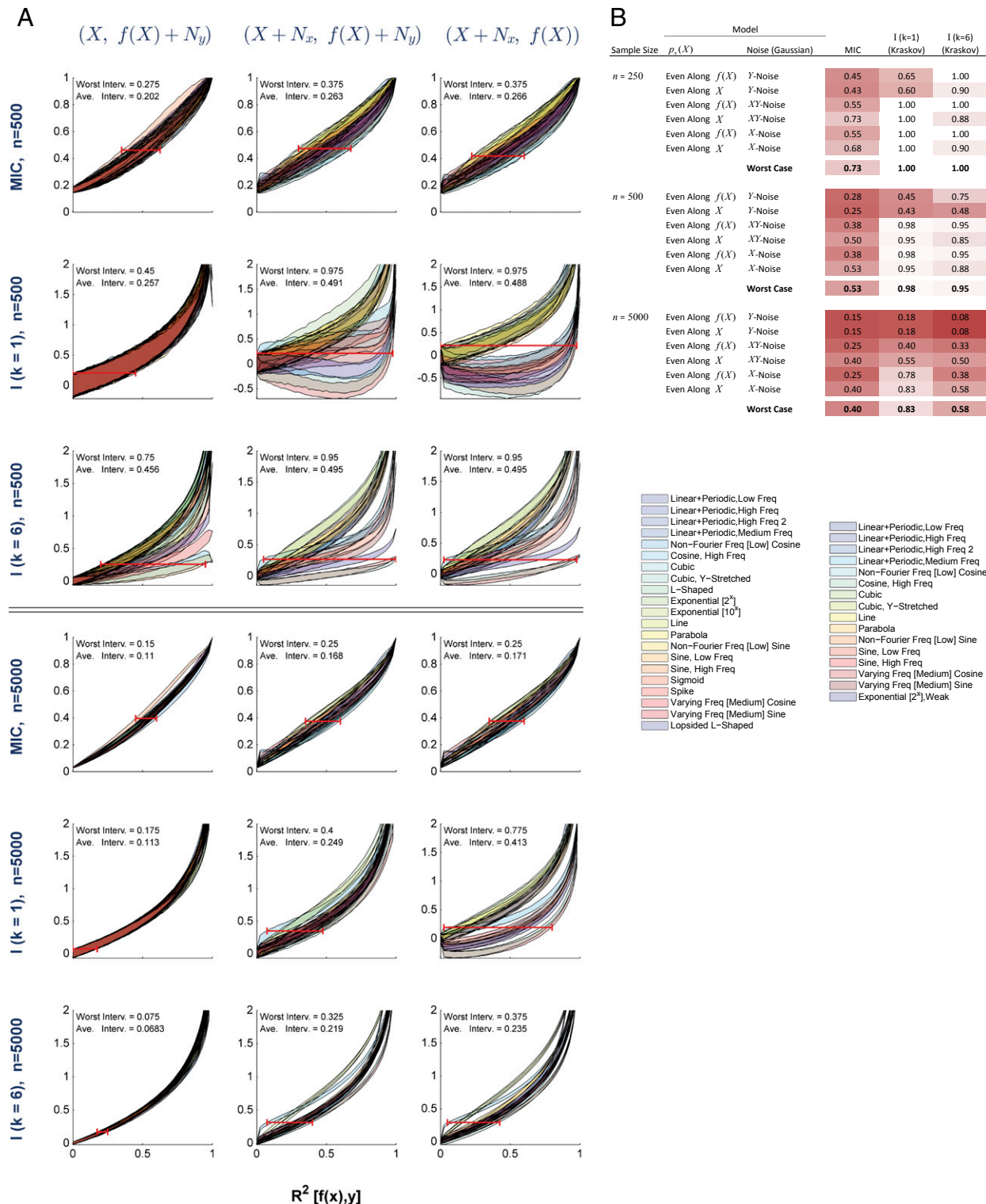5 Kinney JB, Gurinder SA (2014) Reply to Murrell et al.: Noise matters. *Proc Natl Acad Sci USA* 111(21):E2161.

Fig. 1. Equitability of MIC and mutual information under a range of noise models. The equitability of MIC and mutual information across a subset of noise models analyzed in refs. 1 and 4. For each noise model, the relationships tested are as in ref. 4. In each plot in *A*, each shaded region denotes 90% probability intervals based on 500 trials of a given relationship at each of 40 noise levels. In the noise models in *A*, N_x and N_y represent Gaussians, and *X*-values are chosen so that the noiseless data points are spaced uniformly along the graph of $f(X)$. The intervals plotted in red for each noise model in *A* represent the largest range of R^2 values that correspond to a single value of the statistic in question. This provides a quantitative measure of the equitability of each statistic (the shorter the interval, the more equitable the statistic). The values in *B* correspond to the lengths of these intervals across a larger range of sample sizes and the noise models found in ref. 4, and table cells are colored proportionally (red = interval of length 0; white = interval of length 1). In *A*, both the worst and average interval lengths are reported. As in ref. 2, results for the Kraskov et al. mutual information estimator are presented for both $k = 1$ and $k = 6$. The left plot legend applies to the leftmost noise model, and the right legend to the other two as in refs. 1 and 4. In almost every noise model tested, MIC is more equitable than mutual information, consistent with results reported in refs. 1 and 4. To ensure proper comparison, MIC was estimated as in ref. 1; however, we expect that as better estimators of MIC become available they will lead to further superior equitability over mutual information estimators and the MIC estimator used here.