Topics in Algorithmic Data Analysis
Assignment 1: The Scientific Discourse

Muhammad Yaseen - 2577833
s8muyase@stud.uni-saarland.de

> Language disguises the thought; so that from the external form of the clothes one cannot infer the form of the thought they clothe.
>
> *Tractatus Logico-Philosophicus, Ludwig Wittgenstein*

1 Introduction

As with any human activity there is no dearth of cases where scientific community has not held up to its ideals, with phrases 'novel method' and 'better than state of the art' essentially becoming a vacuous cliché. For example, there are well documented cases in Machine Learning [1] and the related field of Information Retrieval [2] where scientific rigour gave way to vested interests. In this report we examine one such instance of scientific jousting in the field of Data Mining.

2 Initial claim and the two critiques

Reshef et al. introduced a new dependency measure in 2011 which they termed Maximal Information Coefficient (MIC)[3]. In the beginning of their paper it was claimed that "**MIC satisfies these two heuristic properties**" i.e. **equitability** and **generality**. Later in the text while supporting the claim authors use a weaker language: "for a large collection of test functions with varied sample sizes, noise levels, and noise models, **MIC roughly equals** R^2" or "noiseless functional relationships **receive MIC scores approaching 1**". This equivocal language makes it harder to discern what the actual claim was and leads one to believe that the final conclusion they draw from their simulation results is – in a charitable reading – much weaker than the original (stated or implied) claim. Hence, their paper suffers from misuse of language problem identified by Lipton and Steinhardt [1].

Taking note of their work, Kinney and Atwal published a critique in which they claimed to show several deficiencies in the MIC [4]. They formalized equitability of MIC as

R^2-equitability and prove that it can't be satisfied by any non-trivial statistic, including MIC. Their proof depends on a justifiable interpretation, which according to them "follows naturally from text and figures" [5], however it nevertheless remains "an" interpretation and requires some leap of hermeneutics. Additionally, they posit a new definition which they call "self-equitability" and show that mutual information (MI) satisfies it and does better than MIC under this new definition.

In a rather interesting turn of events Kinney and Atwal are themselves faced with this definition-juggling when Murrell et al. [6] use a relaxed definition of R^2-equitability to prove that it is indeed satisfiable in their relaxed noise settings. They register their displeasure, both at the weaker definition and the paradoxically bold title implying a stronger claim [7].

Based on this background we make the following observations:

(i) Reshef et al. [3] should have been more careful and precise in their paper about their claim. Their ambiguous wording provided fertile ground for speculation.

(ii) We disagree with the claim that Kenny and Atwal "misrepresent" their work on the ground that it was the ambiguity in their language which provided room for alternative readings. They are however correct in claiming that Kinney and Atwal [4] "cite(d) but fail(ed) to address" their follow-up work [8] in which they took a detailed look at several things pointed to by Kinney and Atwal [4] e.g. the need for normalization and maximization. Judging by the dates of Kinney and Atwal [4] article and the time Reshef et al.'s article reached final form on arXiv (Aug 2013) we are inclined to conclude that they didn't have enough time left to go over it more critically and further refine their criticism as their manuscript was received for review by May 2013.

3 Rebuttal and Reply

For rebuttal by Reshef et al. [9] and Kinney and Atwal's reply [5] to it we make following observations:

(i) The fundamental problem with Reshef et al.'s initial claim is not that it was wrong, but rather that it was, as renowned physicist Wolfgang Pauli said: "**not even wrong**" [10]. As Kinney and Atwal point out that the implied claim was not falsifiable [5]. There was little one could do without providing an explicit definition and then arguing for or against it.

(ii) In the followup work they discuss the effect of normalization and maximization in MIC and do an ablation study without these properties [8]. Their simulations seem to show that these properties are necessary for their version of equitability. This, along with a formal definition, should have been included in the first published work. Kenny and Atwal are therefore correct in pointing out that in absence of a mathematical definition the additional simulations are irrelevant.

(iii) Kenney and Atwal could have been more careful in positing their interpretation as the central claim of Reshef et al. Reshef et al. in their reply to them however also fail to address some of the other valid criticism e.g. MIC reduces to simple MI under 2x2 grids, MIC's bias towards monotonic relationships, or how MIC seems insensitive to certain types of noise as demonstrated [4]. They also don't address the critisism of Simon and Tibshirani [11] on low power of MIC. Neither do they seem to object to "self-equitability" criterion.

(iv) Also, instead of clearly defining what is meant with equitability they again offer more simulation evidence which can't be compared against a definition. Refuting simulation with another simulation is not very productive if the two sides are working with a different underlying definition. We contrast this with the proofs provided by Kenny and Atwal [4] and Murrell et al. [6] on which principled disagreement is possible because of explicitly stated assumptions.

4 The Last Stand

In this section we address the latest preprint by Reshef et al. [12] in which they "formally present and characterize equitability", nearly four years after the initial publication and admit that the "definition is imprecise in that it does not specify what is meant by 'noisy'or 'similar'" in their initial publication of 2011. They provide precise definitions, which has the benefit that they can be analyzed and any deficiencies in them can be argued for just as with Kenny and Atwal's definition of R^2-equitability which Murrell et al. [6] disputed as being too stringent. In this latest work they also address directly the impossibility proof provided by Kenny and Atwal and properly contextualize it in their framework and how it did not apply to their noise setting.

Addressing the statement on lack of power of MIC made by Simon and Tibshirani [11] they also formalize the relationship between power and equitability and come to the conclusion that "if we want to achieve higher power against this larger set of null hypotheses, we may need to give up some power against independence" and "there are situations in which it may be desirable to give up some power against independence in exchange for a degree of equitability."

Further explaining about these cases they state: "primary motivation given for equitability is that often data sets contain so many relationships that we are not interested in all deviations from independence but rather only in the strongest few relationships" which echoes the language of their followup work "MIC is a more appropriate measure of dependence in a situation in which there are likely to be an overwhelming number of significant relationships in a data set, and there is a need to automatically find the strongest ones." and distinguish between "testing for the presence of statistical dependence" and "quantifying the strength of a dependence". However, no such distinction was made in the original work which plainly stated "MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships." or "MINE is useful for identifying and characterizing structure in data." The only vague reference to it may be found it the phrase "rank the pairs by their scores, and examine

the top-scoring pairs". If this was in fact their intention since the beginning they should have made it clear, as they do in this much later work.

In summary, this work addresses most of the criticisms and more importantly formalized the notion of equatibility which makes further principled argument in favor or against the definitions and statistic itself possible. It also specifies the use cases for which MIC is better than traditional dependence measures and mentions potential trade-offs when using it. Naturally, this is a much more balanced statement than ones found in their earlier work.

5 Usefulness of Equitability and *MIC*

Equitability (and generality for that matter) is indeed a desirable property in any dependence measure. This way one could be confident that they are not missing out on any interesting relationship due to the bias of the measure towards type of signal e.g. Linear vs. Sinusoidal and removes any unstated assumption on the type of relationships. All kinds of relationships (where they exist) are information whereas noise quantifies the deviation or degradation. Therefore it makes sense to only penalize the strength of a relationship based on amount of noise and not on the type of relationship itself.

MIC however, seems to have too many knobs (k in estimation algorithm, grid size, optimization) which need to be adjusted without clear guidelines on how to do it or implications thereof. This in our opinion makes it less reliable and not mature enough yet for general use. Its theoretical properties i.e. the precise sense in which it is equitable and practical limitations i.e power and estimation quality need to be well-understood before using it to draw inferences in real world settings.

6 Peer Review: Theory and Practice

In theory, publications are supposed to be judged by relevant experts in the (sub)field. However the match between paper's content and reviewer's area of expertise may not be perfect so reviewers defer their judgement and give authors a benefit of doubt owing to their own lack of sufficient expertise to critically evaluate the submission i.e. they may acquiesce to "Mathiness" [1]. The opposite may also happen, where reviewers fail to understand the significance of work and reject it. We also note that reviewing is a voluntary process and reviewers receive no compensation for the intellectual labor and thus have no real incentive to put in their time which would otherwise be spent on more productive pursuits.

7 On the value of preprint servers

Value of preprint servers like arXiv can be understood if we notice that the faster and wider dissemination of research carries obvious benefits to the researchers. One of them

is that it allows people to claim precedence which is a hugely contested prize in academia. On the other hand, the review and publication cycle in conferences and journals which cater to this purpose can easily take several months in former and even longer than a year in the latter to finally give a decision on publication.

This can also be used to publish comments on earlier works which don't necessarily qualify as a complete work on their own e.g Simon and Tibshirani [11] and work in progress e.g. Reshef et al. [12]. Additionally, it can also be used to published theses and technical reports which don't fit the journals. This provides a more stable, timestamped, and versioned reference to cite than for example a blog post or a PDF on one's personal website.

8 Editorial Responsibility of *Science*

Science is a well-reputed journal and a publication in it carries significant prestige for authors. This is coupled with the expectation that the review process would be correspondingly rigorous. In light of this, we maintain that the decision to publish Reshef et al.'s work in its initial form was not the right one. Especially, looking at the follow-up work [8] which includes a broader discussion on hyper-parameters and ablation studies which would have made for a more transparent publication. Additionally, reviewers should have asked to clarify the statements made in the paper.

At the same time, we believe that retracting the paper after-the-fact would not be beneficial for the following reasons: (1) retraction on part of journal is an extreme measure and carries the implication of serious scientific misconduct (fraud, fabrication, plagiarism etc.) which is not the case in their work (2) it also furthers the conception that judging the accuracy of a scientific work is solely the responsibility of reviewers, and once a paper is past review process it is "infallible". After all peer-review does not end at the publication stage but continues indefinitely based on the principle of faslifiability. Label of "peer reviewed" must not be used as a license to suspend one's own critical judgement when reading a published work.

9 Comment on Role of *PNAS*

As noted earlier journals and conference proceedings serve as a medium of communication. Thus, we think that it was correct to publish Kenny and Atwals work [4]. While that work has its flaws as mentioned earlier, nevertheless it presents some valid criticism of Reshef et al. Similarly we hold that it was right to publish the rebuttal of Reshef et al. as well as Kenny and Atwal's response to it. By doing this, they invite the attention of larger community and generate useful debate on the issue, which is the very purpose of the journal.

References

[1] Zachary C. Lipton and Jacob Steinhardt. "Troubling Trends in Machine Learning Scholarship". In: *Queue* 17.1 (Feb. 2019), pp. 45–77. ISSN: 1542-7730. DOI: 10.1145/3317287.3328534. URL: https://doi.org/10.1145/3317287.3328534.

[2] Jimmy Lin. "The Neural Hype and Comparisons Against Weak Baselines". In: *SIGIR Forum* 52.2 (Jan. 2019), pp. 40–51. ISSN: 0163-5840. DOI: 10.1145/3308774.3308781. URL: https://doi.org/10.1145/3308774.3308781.

[3] D.N. Reshef et al. "Detecting Novel Associations in Large Data Sets". In: *Science* (2011), pp. 1518–1524.

[4] J.B Kinney and G.S. Atwal. "Equitability, mutual information, and the maximal information coefficient". In: *Proc. Natl. Acad. Sci.* (2014), pp. 3354–3359.

[5] J.B Kinney and G.S Atwal. "Reply to Reshef et al.: Falsifiability or bust". In: *PNAS* (2014), E3364–E3364.

[6] Ben Murrell, Daniel Murrell, and Hugh Murrell. "R2-equitability is satisfiable". In: *Proceedings of the National Academy of Sciences* 111.21 (2014), E2160–E2160. ISSN: 0027-8424. DOI: 10.1073/pnas.1403623111. eprint: https://www.pnas.org/content/111/21/E2160.full.pdf. URL: https://www.pnas.org/content/111/21/E2160.

[7] Justin B. Kinney and Gurinder S. Atwal. "Reply to Murrell et al.: Noise matters". In: *Proceedings of the National Academy of Sciences* 111.21 (2014), E2161–E2161. ISSN: 0027-8424. DOI: 10.1073/pnas.1404661111. eprint: https://www.pnas.org/content/111/21/E2161.full.pdf. URL: https://www.pnas.org/content/111/21/E2161.

[8] David Reshef et al. *Equitability Analysis of the Maximal Information Coefficient, with Comparisons*. 2013. arXiv: 1301.6314 [cs.LG].

[9] D.N. Reshef et al. "Cleaning up the record on the maximal information coefficient and equitability". In: *PNAS* (2014), E3362–E3363.

[10] Oliver Burkeman. *The Guardian: Not Even Wrong*. URL: https://www.theguardian.com/science/2005/sep/19/ideas.g2. (accessed: 12.05.2020).

[11] Noah Simon and Robert Tibshirani. *Comment on "Detecting Novel Associations In Large Data Sets" by Reshef Et Al, Science Dec 16, 2011*. 2014. arXiv: 1401.7645 [stat.ME].

[12] Yakir A. Reshef et al. *Equitability, interval estimation, and statistical power*. 2015. arXiv: 1505.02212 [math.ST].