Tilting at Algorithmic Windmills

Nicholas J. Tustison[a,1], Brian B. Avants (combined first author, second author or senior author?)[b]

[a]*Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA*
[b]*Penn Image Computing and Science Laboratory, University of Pennsylvania, Philadelphia, PA*

Abstract

Exploration of neuroscience hypotheses have been greatly enhanced by the increased availability of high-performance computational resources, large-scale communal efforts such as the Insight Toolkit and the R statistical project and algorithmic advancements for data transformations and processing. An integral component of the vetting process for novel algorithmic techniques includes comparison with other methods previously established within the community. Public availability, including open source distribution, of established packages has significantly facilitated these comparative evaluations. However, complementing the recent set of papers pointing to serious methodological and statistical bias considerations in neuroimaging research, we point out potential issues associated with a type of measurement bias in comparative algorithmic evaluations and propose a set of guidelines for authors and reviewers to minimize this confound.

Keywords: open science, reproducibility

1. Introduction

2. Notes/Thoughts

- Several kinds of bias have cropped up in the neuroimaging community recently:
  - methodological bias: Kreikesgorte
  - statistical bias: Vul2009, Tustison2012
  - What about the dead salmon fmri study—multiple comparisons correction?
  - The single subject VBM study (assuming the single subject is representative of the population).
  - overview: 9 circles of scientific hell —neuroskeptic
  - Types of measurement bias: [1]
    - Instrument bias (code as a type of instrument which needs to be calibrated correctly).
    - Expectation bias (i.e. confirmation bias)

  We are concerned with the latter two.

- How does one avoid confirmation bias? If I'm comparing my algorithm to somebody else's algorithm, I'm going to be naturally predisposed to confirmatory evidence that my algorithm is better while neglecting (most likely not with intent of doing so) evidence that disconfirms my original belief (that my algorithm is better). Richard Feynman quote (Cargo Cult Science):

  > The first principle [of science] is that you must not fool yourself–and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.

- One particular aspect of the previous item is that it needs to be highlighted if somebody codes up the algorithm to be compared themselves. Writing bug-free code is difficult. Look at how many bug fixes are provided by the ITK community on a daily basis.

- Also, is simply saying that one algorithm is better than another sufficient? Should there be some theoretical explanation for it. For example, when I wrote my DMFFD image registration paper, the comparison wasn't IRTK vs. DMFFD but rather FFD vs. DMFFD and there were theoretical reasons why performance should be better with the latter. Also, I insisted to the reviewer on not using IRTK for comparison since there are also so many other performance-related issues (type of interpolation, gradient step, metric, metric implementation, etc.) which would confound comparisons. Also, it probably helped that the theme had a more specific, more verifiable focus, i.e. DMFFD produces a more efficient energy minimizer since it acts as a preconditioner in the standard gradient descent optimization.

- Comparisons are best made with publicly available data and results of the comparisons should be made available.

- Operating system needs to be defined. Freesurfer results varied with MacOSx.

[1]Corresponding author: PO Box 801339, Charlottesville, VA 22908; T: 434-924-7730; email address: ntustison@virginia.edu.

- Ideally, the authors of the new algorithm would work with the authors of the compared algorithm. Given human nature, this type of cooperation is not guaranteed. However, a minimum set of information is needed.
- Arno Klein as an example of somebody who did it right. Takes a lot of work.
-
- The source of the package software needs to be defined. For example, N4 has been instantiated in Slicer, ANTs, and c3d (also might be implemented in one of Styner's NITRC projects). However, each might use different default parameters and have other tweaks which effects performance. In Vladimir Fonov's github repository containing various processing scripts for MINC, one can see from the history how N4 was used but then the users switched back to N3MNI (due to performance issues?). However, they used N4 out of the c3d package which the original authors (N.T./B.A.) haven't touched in three years so the parameters aren't optimal (shrink factors = 4, spline distance = 100 with 3 levels: $100 \times 50 \times 50$). Changing these parameters (which are crucial to performance) isn't accessible to the user from c3d.
- Great resources for code-sharing
  - github
  - NITRC
  - ITK

References

[1] Sackett, D. L., 1979. Bias in analytic research. J Chronic Dis 32 (1-2), 51–63.