NeuroImage 103 (2014) 401–410

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Selective correlations; not voodoo

J.D. Rosenblatt [a], Y. Benjamini [b,c,*]

[a] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
[b] Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel Aviv University, Israel
[c] The Sagol School of Neurosciences, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Accepted 11 August 2014
Available online 19 August 2014

Keywords:
Social neuroscience
Selection bias
Effect size
Correlations
Voodoo

ABSTRACT

The problem of "voodoo" correlations–exceptionally high observed correlations in selected regions of the brain–is well recognized in neuroimaging. It arises when quantities of interest are estimated from the same data that was used to select them as interesting. In statistical terminology, the problem of inference following selection from the same data is that of *selective inference*. Motivated by the unwelcome side-effects of splitting the data–the recommended remedy–we adapt the recent developments in selective inference in order to construct confidence intervals (CIs) with good reproducibility prospects, even if selection and estimation are done with the same data. These intervals control the expected proportion of non-covered correlations in the selected voxels—the False Coverage Rate (FCR). They extend further toward zero than standard intervals, thus attenuating the impression made by highly biased observed correlations. They do so adaptively, in that they coincide with the standard CIs when far away from the selection point. We complement existing analytic proofs with a simulation, showing that the proposed intervals control the FCR in realistic social neuroscience problems. We also suggest a "confidence calibration plot", to allow the intervals to be reported in a clear and interpretable way. Applying the proposed methodology on a loss-aversion study, we demonstrate that with the sample size and selection type employed, selection bias is considerable. Finally, selective intervals are compared to the currently recommended data-splitting approach. We discover that our approach has more power and typically more informative, as no data is discarded.
Computation of the intervals is implemented in an accompanying software package.

Introduction

In the pursuit of brain regions that are highly correlated with behavioral measures (neural correlates), past practice has been to report correlations between the imaging measurements and behavioral attributes only in selected regions. These may have been selected based on the same correlation that will be reported. This practice has attracted condemnation for some time: Cureton (1950) refers to correlations reported in this manner as "baloney", with no hope of any meaningful interpretation (cited by Vul et al., 2009b). These early warnings were not echoed in the neuroimaging community until recently.

The implications of such uncontrolled *selective estimation* have been raised more recently by two provocative papers: Vul et al. (2009a) and Button et al. (2013). The problem raised by Vul et al. (2009a) is essentially that reported correlations between imaging attributes and behavioral attributes are "puzzlingly high". Using meta-analysis augmented with questionnaires, the researchers found that many published studies were likely to have applied selective estimation: the correlations reported are in locations selected based on these same correlations, thus justifying the names circular inference and double-dipping. These papers raised the awareness of the matter, not only through impressive meta-analysis, but also by provocative rhetoric. They were so influential that the original title of the former paper–"Voodoo Correlations"–has become an unofficial term for selection bias.

Initially, the many comments on Vul et al. (2009a), in the blogosphere and in the scientific literature (Fiedler, 2011; Poldrack and Mumford, 2009; Lazar, 2009; Lindquist and Gelman, 2009; Nichols and Poline, 2009; Yarkoni, 2009; Lieberman et al., 2009), were not in agreement on the source of the problem and the necessary course of action. Proposed causes included multiplicity control, reporting standards, sample size, sampling bias, and others.

To show the contribution of various different factors to this selection bias, we perform a simulation study. Fig. 1 reports the average estimated correlation following the selection stage. It demonstrates that selection bias is present whenever non-independent selection occurs, and that it can be quite considerable. Even large observed correlations, say $r = 0.8$, can stem from non-existing ones merely due to selection bias. Bias occurs in the presence or in the absence of a true effect, and will be present even if flawless control of multiplicity is performed

* Corresponding author at: Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel Aviv University, Israel.
E-mail address: ybenja@post.tau.ac.il (Y. Benjamini).

http://dx.doi.org/10.1016/j.neuroimage.2014.08.023
1053-8119/

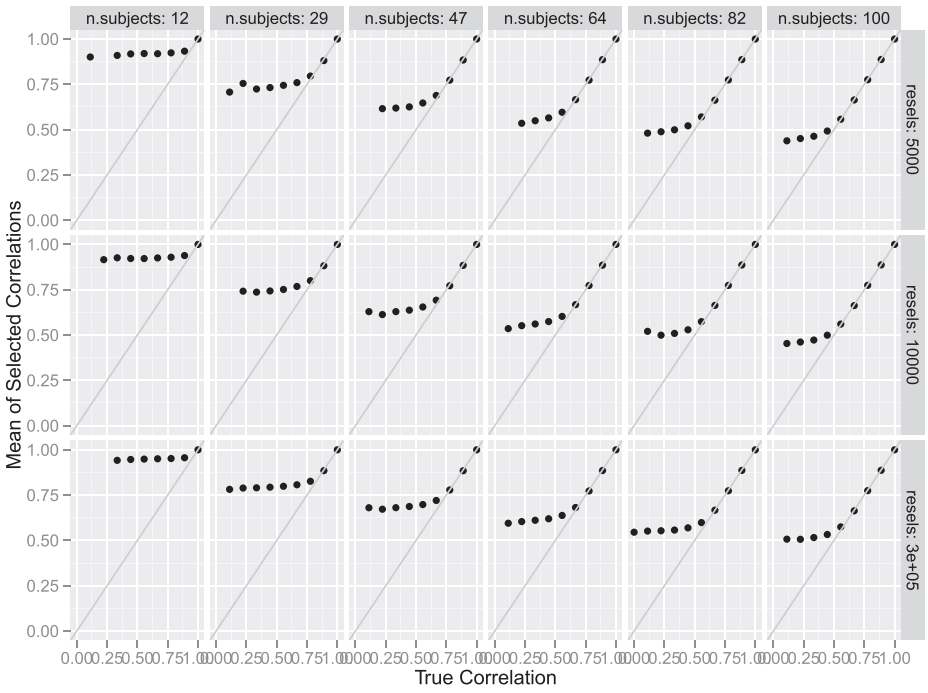

Fig. 1. The mean correlation surviving a selection stage. The figure demonstrates that selection bias is present whenever a non-independent data-driven parameter selection has been performed. The number of resolution elements (resels) varies from 5000 to $3 \cdot 10^5$ (across rows). The true underlying correlation varies from 0 to 1 (x axis), and is the same for all observations. The number of subjects underlying each observed correlation varies from 12 to 100 (across columns). Selection was performed so that the FWER is controlled at 0.05 using the Bonferroni procedure. See Appendix C.1 for more details. Simulation standard errors are nowhere above 0.032. Also note that the most extreme observed values are unbiased. This fact will be revisited when discussing the desired properties of our solutions.

(Bonferroni in our example). In fact, the more conservative the multiplicity control, the higher the selection threshold, so that only extreme correlations survive it. Finally, bias will occur even in very large samples, although it does decrease with sample size: the larger the sample size, the smaller the standard errors of the estimated correlation, the lower the selection threshold, and therefore the milder the selection bias. For a rough intuition regarding the effect of different parameters on the magnitude of the bias we refer the reader to Appendix A.

Imposing independence by splitting the data was the recommended remedy in Vul et al. (2009a), and shared by almost all commentators (Kriegeskorte et al., 2010; Fiedler, 2011; Poldrack and Mumford, 2009; Lazar, 2009; Lindquist and Gelman, 2009; Nichols and Poline, 2009; Yarkoni, 2009) While remedying bias, splitting the data introduces variance effects, making it an unattractive method when dealing with small samples. This matter is elaborated in the Splitting the data section.

Another unbiased approach is that of selecting parameters using the same data, but with a statistically independent criterion. This is implied in Kriegeskorte et al. (2010), and several examples of candidate statistics (albeit in a genetic setup) are suggested by Bourgon et al. (2010). If voxel-wise bias can be sacrificed for the sake of global accuracy, spatial priors in a Bayesian framework allow the spatial pooling of information for improved accuracy. We briefly comment on this view in Appendix B.

Here, we choose a different path, and demonstrate that it is possible to explicitly account for the selection stage at the estimation stage. Ultimately, we will show that:

- The bias introduced by circular inference can be accounted for by more than one way.
- It is typically preferable to account for the inherent bias in circular inference, rather than splitting a small sample to avoid it.

Our methods rest on confidence intervals (CIs) that offer coverage of population parameters, even after a biasing voxel/parameter selection. The Methods section presents two methods of selective confidence interval construction, which are directly relevant to voxel-based analysis. Both methods are sketched in the Overview subsection, leaving technical detail to following subsections, which can be skipped upon a first read. In the Results section we demonstrate the application of our intervals to the loss-aversion study by Tom et al. (2007). The Discussion section deals with shortcomings and possible extensions of the method: point estimates, cluster inference, choice of method, and duality between selection and estimation. Finally, and no less importantly, we identify areas for future research effort, which will be required in order to make selective estimation a readily available tool in researcher's arsenal. We believe this will be worthwhile because "voodoo correlations are everywhere—not only in neuroscience", as the title of Fiedler (2011) states. We could not agree more.

Methods

Overview

A $(1-\alpha)$% CI means that the population parameter will not be covered by the interval with a frequency of α%, over repeated experiments. When generalizing this error criterion to many parameters such as many voxel-wise correlations, or region-wise correlations, several candidate generalizations come to mind. The most natural candidates being control of the frequency of experiments where a parameter is not covered, and control of the expected proportion of non-covered parameters.[1]

The error measure we seek not only deals with a multitude of parameters, but also deals with the effect of selecting a subset of these. Fig. 2.1 depicts a case where 3 out of 20 candidate parameters were selected by a hypothesis test. When constructing 90% confidence intervals on all 20 parameters, 2 fail to cover, as expected. If focusing on the 3 parameters selected, 2 out of the 3 do not cover their underlying population parameter. This coverage is clearly worse than the 1 out of 10 error implied by the confidence level.

[1] The former leads to simultaneous coverage, and the latter is trivially satisfied by controlling the classical confidence level.