NeuroImage 116 (2015) 248–254

Contents lists available at ScienceDirect

NeuroImage

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

Comments and Controversies

Cross-validation and hypothesis testing in neuroimaging: An irenic comment on the exchange between Friston and Lindquist et al.

Philip T. Reiss*

Department of Child and Adolescent Psychiatry and Department of Population Health, New York University School of Medicine, New York, NY, USA
Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA

ARTICLE INFO

Article history:
Received 24 November 2014
Accepted 16 April 2015
Available online 25 April 2015

Keywords:
Brain decoding
Cross-validation
Likelihood ratio test
Neyman–Pearson Lemma
Null hypothesis
Permutation test

ABSTRACT

The "ten ironic rules for statistical reviewers" presented by Friston (2012) prompted a rebuttal by Lindquist et al. (2013), which was followed by a rejoinder by Friston (2013). A key issue left unresolved in this discussion is the use of cross-validation to test the significance of predictive analyses. This note discusses the role that cross-validation-based and related hypothesis tests have come to play in modern data analyses, in neuroimaging and other fields. It is shown that such tests need not be suboptimal and can fill otherwise-unmet inferential needs.

Introduction

Friston (2012) lampoons hostile statistical reviews in neuroimaging by setting forth ten "ironic rules" that an imagined reviewer can follow to ensure a paper's rejection. The seventh of these is to question the validity of the analyses. A suggested example paragraph, by which a reviewer can implement this "rule," reads in part:

> *… the validity of the inference seems to rest upon many strong assumptions. It is imperative that the authors revisit their inference using cross validation and perhaps some form of multivariate pattern analysis.*

Friston notes, however, that the authors can counter with the following response, which he regards as "correct":[1]

> *… the inference made using cross validation accuracy pertains to exactly the same thing as our classical inference; namely, the statistical dependence (mutual information) between our explanatory variables and neuroimaging data. In fact, it is easy to prove (with the Neyman–Pearson Lemma) that classical inference is more efficient than cross validation.*

Lindquist et al. (2013) offer a thorough critique of the ten rules, and in a rejoinder, Friston (2013) graciously concedes many of the points raised in their paper and in a more narrowly focused comment by Ingre (2013). He does, however, expand on several points that remain in dispute, and prominent among these is the role of cross-validation as highlighted by rule 7.

This note aims (i) to clarify why cross-validation scores, and other measures of prediction accuracy, have come to play a role in hypothesis testing for predictive models in neuroimaging and other fields; and (ii) to show that tests constructed in this way need not be suboptimal as asserted by Friston, and indeed can fulfill inferential needs that are not met by classical methods. Friston has raised some concerns that are well worth discussing; even so, I hope to demonstrate that the more cogent argument in the above hypothetical exchange is that of the reviewer.

Whereas Prof. Friston's initial paper adopted the unusual device of an ironic presentation, I have aimed here for a discussion that is *irenic*, *i.e.*, seeking to reconcile differing viewpoints—an important desideratum in a multidisciplinary field such as neuroimaging. A clear understanding of the issues at hand requires not only that we bring together the viewpoints of statisticians and of neuroimagers who make heavy use of statistics; but also that we bring together classical, likelihood-oriented statistical theory and newer, prediction-oriented machine learning approaches.

A class of tests, and a simple example

While Friston's critique focused on cross-validation (CV), it seems reasonable to broaden the discussion somewhat. The class of tests in question seeks to assess whether a predictive model achieves better-

* Department of Child and Adolescent Psychiatry, New York University School of Medicine, 1 Park Ave., 7th floor, New York, NY 10016, USA.
E-mail address: phil.reiss@nyumc.org.

[1] Appendix 1 of Friston et al. (2007), elaborates on the first sentence of this response. The present note will focus more on the second sentence.

http://dx.doi.org/10.1016/j.neuroimage.2015.04.032
1053-8119/

than-chance performance (see Golub et al., 1999, for an early example). To do this, one needs (i) a measure of performance, and (ii) an estimate of the chance (null) distribution of this measure. The performance measure (i) is usually an estimate of prediction error, which is most often provided by a CV score, such as misclassification rate or area under the ROC curve for left-out data. But in some cases another score, such as the Akaike (1973) information criterion, might serve as the prediction error metric. For (ii), a binomial distribution is sometimes used as a null distribution for the number of misclassifications. This, however, may entail serious bias due to ignoring the dependence structure of the data (Noirhomme et al., 2014). This pitfall can be avoided by the more generally applicable approach of using permuted data sets to simulate the null distribution of the performance measure (see Nichols and Holmes, 2001, for an introduction to permutation testing in neuroimaging). In what follows, then, I will sometimes refer to a broader category of predictive performance permutation tests, or "P^3 tests," which may or may not adopt a CV score as the performance measure.

While classification problems seem to be the most popular class of predictive or "decoding" analyses in neuroimaging, analyses with continuous outcomes have become increasingly popular (Cohen et al., 2011) and will serve here as a running example. Consider n observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, with the responses y_i generated from

$$y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_{ik} + \varepsilon_i, \tag{1}$$

where the ε_i's are independent and identically distributed (IID) with zero mean and finite variance. Alternatively we can write $y_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$. For example, y may denote a pain score. In a classical, low-dimensional setting, the predictors $\boldsymbol{x}$ may be demographic factors such as age and sex. In a high-dimensional scenario of a sort that is increasingly popular in neuroimaging, the predictor vector refers to a quantity, measured by an imaging modality at each of a set of regions of interest, which may predict or "encode" the response (pain). Either way, we wish to test the null hypothesis

$$H_0 : \beta_1 = \ldots = \beta_{p-1} = 0 \tag{2}$$

versus the alternative $H_1 : \beta_k \neq 0$ for some $k \in \{1, \ldots, p-1\}$.

A CV-based P^3 test might proceed as follows. Intuitively, if we have a good procedure for estimating $\boldsymbol{\beta}$, then if we apply this procedure to the entire data set except for one observation, then the resulting estimate will do a good job of predicting the left-out response. Let $\hat{\boldsymbol{\beta}}_{-i}$ be the estimate obtained with the ith observation $(\boldsymbol{x}_i, y_i)$ excluded; the ensuing predicted value for the ith response is $\hat{y}_{i;-i} \equiv \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{-i}$. The overall quality of such predictions can be gauged by the cross-validated sum of squared residuals

$$s = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i;-i}\right)^2 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{-i}\right)^2. \tag{3}$$

If the observed value of the CV score (Eq. (3)) is smaller than we would expect under H_0—in other words, if it lies in the left tail of the null distribution of Eq. (3)—then this constitutes evidence against H_0.[2]

To simulate the null distribution, we can choose a large number of permutations, say $\pi_1, \ldots, \pi_M$, of $\{1, \ldots, n\}$, and create artificial data sets by applying these permutations to the responses: the mth such data set is

$$\left(\boldsymbol{x}_1, y_{\pi_m(1)}\right), \ldots, \left(\boldsymbol{x}_n, y_{\pi_m(n)}\right). \tag{4}$$

Let $\hat{\boldsymbol{\beta}}_{-i}^{\pi_m}$ be the estimate obtained from the mth transformed data set with its ith observation left out. The observed distribution of the permuted-data CV score $s_{\pi_m} = \sum_{i=1}^{n}\left(y_{\pi_m(i)} - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{-i}^{\pi_m}\right)^2$ $(m = 1, \ldots, M)$ serves as a simulated null distribution of Eq. (3), and the p-value is given by

$$\frac{\#\{m : s_{\pi_m} < s\} + 1}{M + 1}.$$

Adding 1 to the numerator and denominator is equivalent to including the original statistic value in the permutation distribution, as required to obtain a valid test (see Phipson and Smyth, 2010).

To see why the empirical distribution of $s_{\pi_1}, \ldots, s_{\pi_M}$ mirrors the null distribution, observe that if H_0 is true, then $y_1, \ldots, y_n$ are simply IID with mean β_0 and variance σ^2. Thus under H_0, the permuted data (Eq. (4)) arise from the same distribution as the original data, and hence $s_{\pi_1}, \ldots, s_{\pi_M}$ arise from the same distribution as s.

The above is just one simple example of a very general technique. In other P^3 tests, linear regression might be replaced by support vector machines or other predictive algorithms; and the squared error loss could be replaced by other loss functions, or more general measures of performance on left-out data. More general treatments can be found in Golland and Fischl (2003) and Ojala and Garriga (2010).

Why not just use a likelihood ratio test?

As we saw in the Introduction, Friston (2012) appeals to the fundamental lemma of Neyman and Pearson (1933) (hereafter, the NP Lemma) to argue against CV-based tests. Prof. Friston has provided two explanations of how the NP Lemma applies. In a footnote to the above-cited remark on rule 7 (Friston, 2012), he writes: "Inferences based upon cross validation tests (*e.g.*, accuracy or classification performance) are not likelihood ratio tests because, by definition, they are not functions of the complete data whose likelihood is assessed. Therefore, by the Neyman–Pearson Lemma, they are less powerful."[3]

In his rejoinder, Friston (2013) elaborates on how CV is used for hypothesis testing in neuroimaging, and then offers a somewhat different explanation of how the NP Lemma applies: "For example, do the voxels in my hippocampal volume of interest encode the novelty of a particular stimulus? To answer this question one has to convert the cross validation scheme into a hypothesis testing scheme—generally by testing the point null hypothesis that the classification accuracy is at chance levels. It is this particular application that is suboptimal. The proof is straightforward: if a test of classification accuracy gives a different p-value from the standard log likelihood ratio test then it is—by the Neyman–Pearson Lemma—suboptimal. In short, a significant classification accuracy based upon cross validation is not an appropriate proxy for hypothesis testing. It is in this (restricted) sense that the Neyman–Pearson Lemma comes into play."

There are two fundamental problems with these appeals to the NP Lemma. The first problem was pointed out by Lindquist et al. (2013) and acknowledged by Friston (2013), but calls for further elaboration.

[2] In practice, rather than leave-one-out CV as described here and in Appendix C, K-fold CV is typically used—resulting in computational savings that are particularly helpful when CV is combined with permutation. Hastie et al. (2009) recommend $K = 5$ or 10, which offers a favorable bias-variance tradeoff.

[3] Note that the CV-based test developed above *does* use all the data for model fitting (although each training set fit does not). This advantage of CV over reserving part of the data solely for validation was noted by Simon et al. (2003).