2015

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

Philip T. Reiss

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

Philip T. Reiss[a,b,c,*]

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

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.

Keywords: brain decoding, cross-validation, likelihood ratio test, Neyman-Pearson Lemma, null hypothesis, permutation test

*Mailing address: Department of Child and Adolescent Psychiatry, New York University School of Medicine, 1 Park Ave., 7th floor, New York, NY 10016, USA. Phone: 646-754-5138. E-mail address: phil.reiss@nyumc.org.

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).

[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.

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 seek to assess whether a predictive model achieves better-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 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

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

If the observed value of the CV score (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 (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

[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 offer a favorable bias-variance tradeoff.

applying these permutations to the responses: the mth such data set is

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

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 $s_{\pi_m} = \sum_{i=1}^{n} (y_{\pi_m(i)} - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{-i}^{\pi_m})^2$ $(m = 1, \ldots, M)$ serves as a simulated null distribution of (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 (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) (herafter, 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-

[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).

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. In rule 6, the hypothetical reviewer questions the parametric assumptions underlying the analysis, and requires the authors to "repeat their analysis using nonparametric tests." Friston (2012) "praises" this suggestion (i.e., attacks it, in an ironic way) by noting that "the nonparametric tests will, by the Neyman-Pearson lemma, be less sensitive than the original likelihood ratio tests." Lindquist et al. (2013) retort that this "is only true if the exact parametric assumptions of the likelihood ratio test are valid, precisely the point the hypothetical reviewer sought to make." While this exchange arose in connection with rule 6, the point raised by Lindquist et al. (2013) is equally relevant to the present discussion of CV-based tests (rule 7), which are often applied in complex situations where it is difficult or unrealistic to specify a parametric model. Indeed, Golland and Fischl (2003), in an early formulation of a P^3 test for neuroimaging data, say explicitly that their test "does not assume a generative model for the data," and indeed view this as an advantage since it makes the test widely applicable.

But even if we do agree on a generative model for the data, there is a second fundamental problem: the NP Lemma does not, in most cases, provide that a likelihood ratio test (LRT) is the most powerful test—in either the "classical" (low-dimensional, or $p \ll n$) or the "modern" (high-dimensional, or $p > n$) setting.

The low-dimensional case

As already noted by Lindquist et al. (2013), the NP Lemma applies only when the null and alternative hypotheses are *simple*, i.e., each specifies a unique set of values for the parameters. Consider the above example, and suppose the error terms ε have the $N(0, \sigma^2)$ distribution (such a full specification of the error distribution is needed if we are to speak of likelihood). Letting $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$, the parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$, and the space of possible parameter values is $\boldsymbol{\Theta} = \{(\boldsymbol{\beta}, \sigma) : \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0\}$. If the null and alternative hypotheses were $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0, \sigma = \sigma_0$ and $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1, \sigma = \sigma_1$ for specific values $\boldsymbol{\beta}_0, \sigma_0, \boldsymbol{\beta}_1, \sigma_1$, the NP Lemma would tell us that the most powerful test of a given size α would reject H_0 for the highest values of the likelihood ratio $L(\boldsymbol{\beta}_1, \sigma_1)/L(\boldsymbol{\beta}_0, \sigma_0)$. Here $L(\boldsymbol{\beta}, \sigma)$ denotes the likelihood of the observed data, which is

$$\frac{\exp[-\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2/(2\sigma^2)]}{(2\pi)^{n/2}\sigma^n}. \tag{5}$$

In the problem stated above at (2), both H_0 and H_1 are not simple but *composite*: under H_0, $\boldsymbol{\theta}$ may take any value in

$$\boldsymbol{\Theta}_0 = \{(\boldsymbol{\beta}, \sigma) : \beta_0 \in \mathbb{R}, \beta_1 = \ldots = \beta_{p-1} = 0, \sigma > 0\},$$

while under H_1, $\boldsymbol{\theta}$ may take any value in $\boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. In general, with composite hypotheses, if the most powerful level-α test defined by the NP Lemma is the same for all $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$ and $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$ (it usually is not), we call this test uniformly most powerful (UMP). For composite hypotheses, an LRT rejects for large values of

$$\frac{\sup_{\theta \in \Theta \setminus \Theta_0} L(\boldsymbol{\beta}, \sigma)}{\sup_{\theta \in \Theta_0} L(\boldsymbol{\beta}, \sigma)}. \tag{6}$$

In general this test need not be UMP, although under certain regularity conditions it is asymptotically the most powerful test (Lehmann and Romano, 2005). For testing (2) with $p < n$, the F-test is traditionally viewed as optimal, and the LRT can be viewed as an approximation to the F-test (in a sense that is explained in Appendix A).

We note that (6) may be interpreted as the ratio of the maximized probabilities of the data under the alternative and null hypotheses. By contrast, the Bayes factor (Kass and Raftery, 1995), which has been widely used in neuroimaging (e.g. Penny, 2012), is the ratio of *integrals* of the data probability, with respect to the prior distributions for two models.

The high-dimensional case

Modern predictive analyses, in neuroimaging as in other fields, often involve data sets for which $p > n$. In this case, the likelihood ratio (6) not only does not converge in distribution to χ^2_{p-1}, but indeed does not exist. To see this, take $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^+\boldsymbol{y}$ where $\boldsymbol{X}^+$ is a generalized inverse of $\boldsymbol{X}$; then the residuals $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are all zero and the likelihood $L(\hat{\boldsymbol{\beta}}, \sigma)$ (see (5)) goes to infinity as $\sigma \to 0$.

If we cannot fit the model by maximizing the likelihood, what can we do? Since (assuming IID normal errors) the maximum likelihood estimate of $\boldsymbol{\beta}$ is also the least-squares estimate, a natural solution is to modify the sum of squared errors criterion, by adding a penalty. The popular Lasso method (Tibshirani, 1996) adds an ℓ_1 penalty, so that the estimate of $\boldsymbol{\beta}$ is the minimizer of

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{k=1}^{p-1}|\beta_k| \tag{7}$$

for some $\lambda > 0$. Minimizing this criterion yields a sparse estimate of $\boldsymbol{\beta}$, i.e., the estimates for some of the β_k's are zero, with higher λ implying more zero coefficients. (See Appendix B for a brief discussion of the Bayesian perspective, which views the penalty term as imposing a prior on $\boldsymbol{\beta}$.)

To illustrate how we might test the null hypothesis (2) for a high-dimensional model, we apply the Lasso to a portion of the data previously analyzed by Lindquist (2012). The original study examined fMRI measures of response to warm and hot stimuli applied to the left volar forearm in 20 volunteers. Here we consider only the hot (painful) stimulus trials, of which there were 11–24 per subject, with $n = 433$ trials in total. Each trial consisted of thermal stimulation for 18 seconds; then a 14-second interval, at the end of which the words "How painful?" appeared on a screen; then another 14-second interval after which the participant rated the overall pain intensity between 100 and 550 (with higher values indicating more pain). The BOLD signal was recorded in 21 pain-relevant regions at 23 2-second intervals. If we fit the pain prediction model (1), with y denoting log pain score and $\boldsymbol{x}$ denoting the fMRI measurements for the 21 regions $\times$ 23 time points along with a 1 for the intercept, we have $p = 1 + 21 \cdot 23 = 484$. For simplicity, I did not attempt to take within-subject correlation into account in the model.

Figure 1(a) and (b) show the estimates of $\beta_1, \ldots, \beta_{483}$, the "effects" on pain[4] of BOLD signal at each region and time point, based on Lasso fits with two values of λ. For a given data set, lower values of λ always imply a higher

[4]The scare quotes here are meant to avoid asserting that the BOLD signal truly causes pain. For lack of a better term, "effect" serves as shorthand for the increment in expected pain associated with a unit change in the BOLD signal.

likelihood. From that perspective, the estimate shown in Figure 1(a) is "better." But when overfitting is a concern—as it is here—the measure of a good model is not its likelihood, or ability to predict the sample responses, but rather its ability to predict future responses, which is captured by CV. Figure 1(c), produced with the `glmnet` package (Friedman et al., 2010) for R (R Core Team, 2014), shows that the expected mean squared error of prediction (based on 10-fold CV, the default in `glmnet`) is almost five times higher for $\log(\lambda) = -6$, the leftmost point, as for $\log(\lambda) = -1.84$, the CV-minimizing value. Unlike the overfitted estimate displayed in (a), the CV-optimal model fit shown in (b) is quite interpretable; its 55 nonzero coefficients are most highly concentrated between the end of the hot stimulus and the appearance of the question on the screen, suggesting that BOLD signal level in this time interval is most predictive of pain intensity. More specifically, the two coefficients of greatest magnitude are both at the 24-second point: higher BOLD signal at that time, in the lateral cerebellum (LCB) and right superior frontal gyrus, are associated with higher reported pain.

The key point that this example illustrates is that (out-of-sample) prediction error, rather than (in-sample) likelihood, is what ultimately governs the model fit in penalized regression, as in other predictive models for high-dimensional data. As observed above, a classical test such as LRT is unavailable here (and the recently proposed test of Lockhart et al., 2014, refers to individual coefficients rather than the global null hypothesis (2)). It seems natural, then, to turn to lower-than-chance prediction error, as opposed to higher-than-chance likelihood, as a guiding principle for testing (2); and this is what CV-based tests aim to do (cf. van de Wiel et al., 2009).

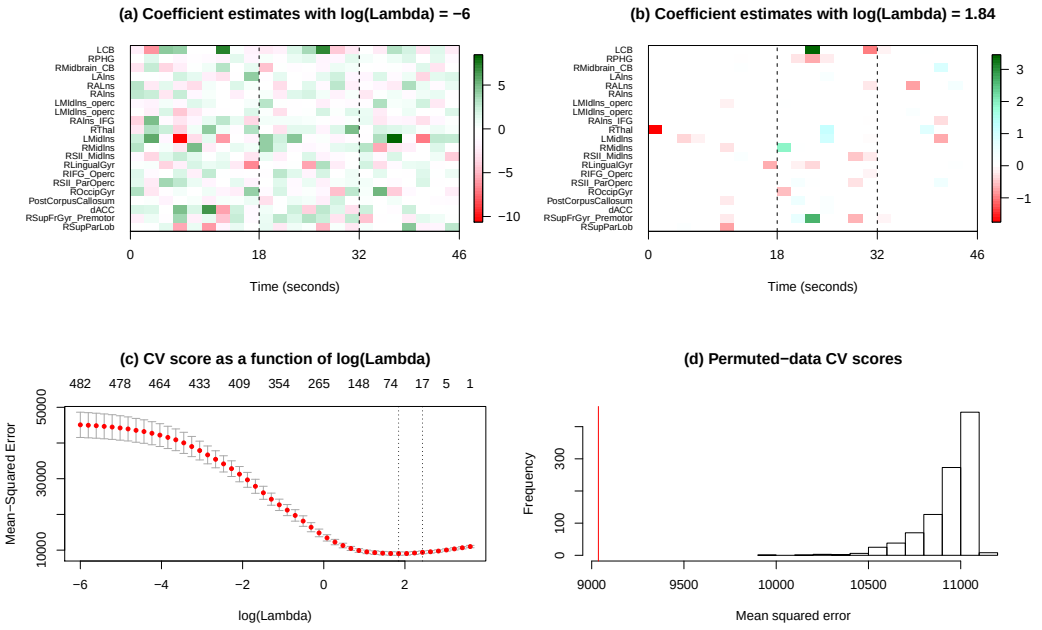

Figure 1: (a) Estimate of $\boldsymbol{\beta}$ obtained by minimizing the Lasso criterion (7) with $\log(\lambda) = -6$. The rows represent 21 pain-relevant regions, while the columns represents 23 time points. The first dashed line marks the end of the hot stimulus, while the second marks the time at which "How painful?" appeared on the screen. (b) Estimate based on $\log(\lambda) = -1.84$. (c) Ten-fold cross-validated mean squared error, $\pm$ 1 standard error (SE), for a range of $\log(\lambda)$ values. The first dotted line indicates where CV is minimized ($\log(\lambda) = -1.84$); the second indicates the largest $\log(\lambda)$ at which the CV score is within 1 SE of the minimum. The numbers along the top edge are counts of nonzero coefficient estimates, which decrease as λ increases. (d) Histogram of 999 permuted-data CV scores, with the real-data CV score indicated by the red line.

I calculated the minimum CV using 999 data sets in which the pain scores were permuted within subject. The real-data value is well to the left of the histogram shown in Figure 1(d), implying a p-value of .001—strong evidence that we can reject (2), i.e., that variations in the BOLD signal in certain regions at certain times are associated with reported pain in this task. Such a conclusion would not be attainable by a standard LRT (but see Appendix B).

Are CV-based tests really suboptimal?

We have seen that the NP Lemma does not establish the LRT's supremacy for testing (2) in the low-dimensional case, and certainly not in the high-dimensional case. But this in itself does not disprove that CV-based tests are "suboptimal." Is it possible to do so?

To a degree, it is. In Appendix C we show that in the low-dimensional case, for a one-way ANOVA model, the leave-one-out CV test is exactly equivalent to a permutation F-test. As shown by Hoeffding (1952) and Robinson (1973), the latter is asymptotically as powerful as an ordinary F-test (i.e., one that rejects when the F-statistic is in the right tail of the F-distribution) under the assumption of normal IID errors, but more robust to departures from those assumptions. The ordinary F-test, in turn, is optimal in the sense described in Appendix A. Thus our CV test is optimal in this simple case, a result illustrated by the power curves in Figure 2. The top row shows estimates and 95% confidence intervals for the CV test's power to detect the difference between two groups at the 5% level, based on 1000 replications for each of six equally spaced R^2 values. For combined sample

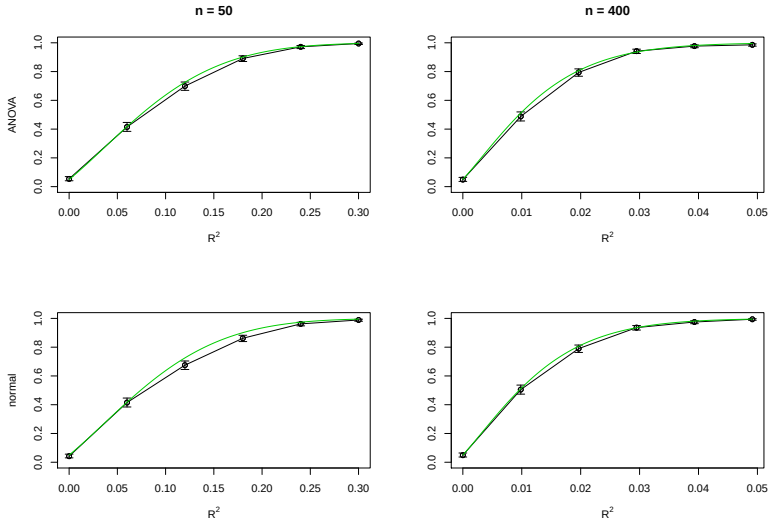

Figure 2: Power of the CV permutation test (black) compared with that of the benchmark F-test (green), for two cases with $p = 2$: a one-way ANOVA comparing two groups (above) and a normally distributed covariate (below).

size $n = 50$, and even more so for $n = 400$, the power is virtually indistinguishable from the benchmark F-test power (shown in green) given by Cohen (1988) for corresponding effect sizes $f^2 = R^2/(1 - R^2)$. For detecting the effect of a normally distributed covariate at the 5% level, the CV test is noticeably less powerful than the F-test for intermediate R^2 values with $n = 50$, but this difference essentially disappears for $n = 400$.

In practice, for testing (2) in the low-dimensional case, it seems best to use an ordinary or permutation F-test.[5] But the equivalence of the CV statistic and the F-statistic for (permutation tests of) (2) in the low-dimensional

[5]Cf. Pepe et al. (2013), who shows that for risk models, a classical test for zero effect of a predictor is preferable to testing that it improves predictive performance.

case suggests that for the high-dimensional case, in which classical tests are unavailable, a CV-based test is worth considering.

What is the null hypothesis?

Friston (2013) concludes his discussion of CV as follows: "I have mixed feelings about cross validation, particularly in the setting of multivariate pattern classification procedures. On the one hand, these procedures speak to the important issue of multivariate characterisation of functional brain architectures. On the other hand, their application to hypothesis testing and model selection could be viewed as a non-rigorous and slightly lamentable development."[6]

One aspect of CV-based testing that does seem quite vulnerable to lack of rigor is the specification of the null hypothesis. In some papers describing P^3 tests, the null hypothesis has been formulated along the lines of "there is no information in the data" or "we cannot predict the outcome from the data." Such formulations deviate from the classical notion of a statistical hypothesis. Traditionally, a statistical hypothesis is an assertion about the probability distribution generating the given data—e.g., for parametric distributions, an assertion about the values of the parameters—as opposed to a statement about our ability to learn something from the data. This point matters, because without a clear and objective formulation of the null hypothesis

[6]While my focus here is on hypothesis testing, I wish to note that Prof. Friston's suggestion that CV is also unsuited to *model selection* is rather surprising. CV is very widely used for tuning parameter selection and more generally for model selection, as surveyed by Arlot and Celisse (2010).

we are testing, we may have a hard time specifying or simulating the null distribution.

For permutation tests in general, the null hypothesis is an instance of the generic "randomization hypothesis" (Hoeffding, 1952): there is a group $\mathcal{G}$ of transformations of the data $\boldsymbol{Z}$ such that for each $g \in \mathcal{G}$, $g\boldsymbol{Z}$ has the same distribution as $\boldsymbol{Z}$. This idea encompasses not only permutations but other transformations such as sign flipping (Winkler et al., 2014).

Often a randomization hypothesis is natural in a nonparametric setting, where we do not specify a parametric model but wish to test the null hypothesis that, say, some group of permutations of the responses leaves the data distribution unchanged. But sometimes a parametric null hypothesis implies a randomization hypothesis. For instance, for linear regression with IID errors, we justified using $s_{\pi_1}, \ldots, s_{\pi_M}$ to simulate the null distribution by arguing that under null hypothesis (2), transformations of the form (4) do not change the data distribution. For the pain data example, the null hypothesis is again (2); but we permuted observations only within participants, since observations for different participants are not exchangeable.

In practice, it is common to test the effect of, say, imaging-based variables while adjusting for nuisance variables such as age and sex. The null hypothesis then posits exchangeability not for the errors, but (approximately) for the residuals (see Winkler et al., 2014, for a detailed discussion).

Is statistical significance the appropriate aim?

Many researchers, including Prof. Friston, have grave concerns about the entire statistical paradigm of hypothesis testing. But even those of us who

are less troubled by such fundamental concerns should acknowledge that P^3 tests will be quite useless in some contexts.

Here is analogy from a more classical setting. A standard test statistic can be used to assess whether an intraclass correlation coefficient (ICC) equals a given number, such as zero. But publications that use the ICC to describe the test-retest reliability of a psychometric measure do not usually report a p-value for rejecting the null hypothesis ICC=0. Why not? Because it would be superfluous. Any psychometric measure worth its salt has an ICC well above 0; the question is how high the ICC is, and this is answered by a point estimate and confidence interval.

Similarly, with neuroimaging-based predictive analyses, testing what amounts to a null hypothesis of zero predictive value may or may not be appropriate, depending on the application. For a preliminary proof of concept or of relevance, such a test may make sense. But for applications in which the utility of image-based prediction has already been established, the principal question is not *whether* meaningful prediction is possible, but how well one can predict. A significant p-value should then go without saying, and a more meaningful measure might be, say, a cross-validated area under the ROC curve (see Hsing et al., 2003).

One class of applications in which hypothesis testing does seem to be needed has to do with showing that images provide predictive value beyond what is available from other variables, which are likely to be easier and less expensive to collect (see Boulesteix and Sauerbrei, 2011). For example, consider the recent ADHD-200 Global Competition in which teams developed diagnostic classifiers for attention deficit/hyperactivity disorder based

on multimodal brain images, which they then applied to test data for which diagnoses were not provided. The approach of Eloyan et al. (2012) predicted the test data diagnoses most accurately—among teams who used the imaging data. Surprisingly, however, Brown et al. (2012) attained somewhat better predictive accuracy without using the images at all. My colleagues and I (Reiss et al., 2015) have shown how CV-based testing can be modified to assess whether image data adds predictive value beyond that offered by non-imaging predictors. In our analysis of a portion of the ADHD-200 data, images derived from subjects' resting state fMRI scans appeared significantly predictive of diagnosis, but this result vanished upon adjusting for covariates such as age and sex.

Conclusion

While P^3 testing may lack the elegance of classical optimal testing theory, it is becoming increasingly pervasive. Indeed, P^3 tests are beginning to be employed clinically to detect awareness in vegetative states. The importance of getting the methodology right in such applications is clear (Goldfine et al., 2013). While not all examples will be this dramatic, I believe that, rather than trying to delegitimize such tests, we should devote more effort to understanding how they work and how they can be improved. There is plenty of work to do along these lines.

I completely agree with Prof. Friston that "the role of cross validation in neuroimaging deserves further discussion." And his ironic contribution has done a valuable service by stimulating discussion of this and other statistical issues. I would submit, however, that in our collective quest for much-needed

advances in statistics for neuroimaging data, we will generally be best served by less irony—and more irenics.

Acknowledgments

My sincere thanks are extended to the referees for their invaluable feedback; to Christos Davatzikos for helpful advice; to Martin Lindquist for providing the fMRI data; to Lan Huo and Pei-Shien Wu for their work on the power simulations; and to Brian Caffo, Xavier Castellanos, Ani Eloyan, Pei-Shien Wu and Yuliya Yoncheva for helpful feedback on the manuscript. This work was partially supported by National Institutes of Health grant 1R01MH095836-01A1, and indirectly through grants R01MH076136-06 and R01DA035484-01 supporting the original pain study.

Appendix A. The LRT as an approximation to the F-test for null hypothesis (2)

Here we show why for the linear model with IID normal errors,

$$y = \beta_0 + \sum_{k=1}^{p-1} \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

even when $p < n$, the LRT is not (quite) optimal for testing $H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$ versus $H_1 : \beta_k \neq 0$ for some $k \in \{1, \ldots, p-1\}$.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and define the "hat matrices" $\boldsymbol{H}_0 = n^{-1}\mathbf{1}_n\mathbf{1}_n^T$, $\boldsymbol{H}_1 = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ where $\boldsymbol{X}_{n\times p}$ has ith row $(1, x_{i1}, \ldots, x_{i,p-1})$. The hat matrices are so named because they transform the response vector $\boldsymbol{y}$ to the vector $\hat{\boldsymbol{y}}$ of fitted values, under the null and alternative models respectively—i.e., $\hat{\boldsymbol{y}} = \boldsymbol{H}_0\boldsymbol{y}$ under H_0, $\hat{\boldsymbol{y}} = \boldsymbol{H}_1\boldsymbol{y}$ under H_1. The sum of squared residuals

under the null and alternative models are $\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_0)\boldsymbol{y}$ and $\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{y}$, respectively.

There is no UMP test here, but Lehmann and Romano (2005) list three forms of invariance, i.e., transformations of the data (e.g., rescaling) that in a technical sense have no bearing on the evidence against H_0, and show that among tests that are invariant to such transformations, the F-test, which rejects if

$$F = \frac{\left[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_0)\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{y}\right]/(p-1)}{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{y}/(n-p)}. \tag{A.1}$$

exceeds the $1-\alpha$ quantile of the F-distribution with $p-1$ and $n-p$ df, is the most powerful; that is, the F-test is "UMP invariant."

An LRT statistic for composite hypotheses is twice the log of (6), or

$$2(\ell_1 - \ell_0) \tag{A.2}$$

where ℓ_1 and ℓ_0 denote the numerator and denominator of (6), respectively. For these particular hypotheses, the LRT statistic (A.2) is a monotonically increasing function of the F-statistic, since the former equals

$$n\log\left[\frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_0)\boldsymbol{y}}{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{y}}\right] = n\log\left(1+\frac{p-1}{n-p}F\right). \tag{A.3}$$

The two *tests* are not exactly equivalent, however, in that whereas the F-test is based on an exact distribution, the LRT is based on a large-sample approximation due to Wilks (1938): it rejects when (A.2) exceeds the $1-\alpha$ quantile of the χ^2_{p-1} distribution, to which (A.3) converges in distribution as $n \to \infty$. Interestingly, since the two tests are not equivalent and the NP Lemma is used indirectly to prove that the F-test is UMPI (Lehmann and Romano, 2005), it could be said that in this case the NP Lemma shows

the LRT *not* to be the optimal test—although for large n the two tests are virtually the same.[7]

Appendix B. Likelihood-based tuning parameter selection

In the main text we argued that prediction error, rather than likelihood, typically serves as the guiding principle for tuning parameter selection in high-dimensional penalized regression. This can be seen as part of the increased attention to prediction error in the statistical literature since the work of Stone (1974) on CV and that of Akaike (1973) on information criteria—a shift that was championed by Breiman (2001).

There is, however, a way to reformulate the likelihood and thereby recover its central role: treating the coefficients as random effects and the tuning parameter (here λ) as a bona fide parameter, proportional to the random effects variance. One can then maximize the likelihood over λ and the other parameters. The same approach has a Bayesian formulation, in which the penalty term represents a prior distribution and the estimate maximizes the posterior. Likelihood or (empirical) Bayesian methods have been applied successfully to spline smoothing (Ruppert et al., 2003; Reiss and Ogden, 2009; Wood, 2011),

[7]Here I have used the term "LRT" in its colloquial sense, to refer to the χ^2 test of Wilks (1938). This usage seems appropriate for examining the claim (Friston, 2013) that the "standard" LRT is optimal. In the more precise nomenclature of Abramovich and Ritov (2013), (6) is a *generalized* likelihood ratio test statistic (since it generalizes the simple-vs.-simple LRT statistic to which the NP Lemma applies); and for nested linear models, the above argument shows that the F-test *is* the (exact) generalized LRT, whereas what I have called "the LRT" is the asymptotic generalized LRT.

and moreover can be used to derive an LRT for a null hypothesis such as (2) (Crainiceanu and Ruppert, 2004). An analogous Laplace prior approach has been developed for Lasso estimation (Park and Casella, 2008) and could perhaps form the basis of an LRT for (2); but it cannot be asserted *a priori* that such an LRT would outperform CV-based tests. See also Goeman et al. (2006), who pursue an empirical Bayes approach to develop a score test, which is *locally* most powerful, for high-dimensional alternatives. Empirical Bayes estimation has been particularly influential in neuroimaging (Friston et al., 2002).

Appendix C. Asymptotic optimality of a CV permutation test for null hypothesis (2)

To relate our leave-one-out CV test statistic (3) to the F-statistic, we use the identity

$$y_i - \hat{y}_{i;-i} = \frac{y_i - \hat{y}_i}{1 - h_{ii}},$$

where $\hat{y}_i$ is the ith (full-data) fitted response and h_{ii} is the ith diagonal element of the hat matrix $\boldsymbol{H}_1$, to re-express (3) as

$$\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y} \tag{C.1}$$

where $\boldsymbol{D} = \text{Diag}\left\{\frac{1}{(1-h_{11})^2}, \ldots, \frac{1}{(1-h_{nn})^2}\right\}$. In the special case of a balanced one-way (fixed-effects) ANOVA design, $h_{ii} = p/n$ for each i, so (C.1) reduces to

$$(1 - p/n)^{-2}\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y}. \tag{C.2}$$

The $\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y}$ term in the numerator of the F-statistic (A.1) is invariant to permutation. Thus when $\boldsymbol{y}$ is replaced by a permuted version $\boldsymbol{y}_\pi$,

the resulting F-statistic is

$$F_\pi = \frac{\left[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_0)\boldsymbol{y} - \boldsymbol{y}_\pi^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{y}_\pi\right]/(p-1)}{\boldsymbol{y}_\pi^T(\boldsymbol{I}-\boldsymbol{H}_1)\boldsymbol{y}_\pi/(n-p)},$$

which is a strictly decreasing function of the permuted-data CV test statistic, i.e., of (C.2) with $\boldsymbol{y}$ replaced by $\boldsymbol{y}_\pi$. Hence a permutation test that rejects when the CV criterion is in the left tail of the permutation distribution is equivalent to one that rejects when F-statistic (A.1) is in the right tail of the distribution of F_π values—that is, a permutation F-test.

References

Abramovich, F., Ritov, Y., 2013. Statistical Theory: A Concise Introduction. CRC Press, Boca Raton, FL.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, in: Second International Symposium on Information Theory, Akademiai Kiado. pp. 267–281.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics Surveys 4, 40–79.

Boulesteix, A.L., Sauerbrei, W., 2011. Added predictive value of high-throughput molecular data to clinical data and its validation. Briefings in Bioinformatics 12, 215–229.

Breiman, L., 2001. Statistical modeling: The two cultures (with discussion). Statistical Science 16, 199–231.

Brown, M.R.G., Sidhu, G.S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P.H., Greenshaw, A.J., Dursun, S.M., 2012. ADHD-200 Global Competition: Diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. Frontiers in Systems Neuroscience 6.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y., Knowlton, B.J., Poldrack, R.A., 2011. Decoding continuous behavioral variables from neuroimaging data. Frontiers in Neuroscience 5.

Crainiceanu, C.M., Ruppert, D., 2004. Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society: Series B 66, 165–185.

Eloyan, A., Muschelli, J., Nebel, M., Liu, H., Han, F., Zhao, T., Barber, A., Joel, S., Pekar, J., Mostofsky, S., Caffo, B., 2012. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. Frontiers in Systems Neuroscience 6.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1.

Friston, K., 2012. Ten ironic rules for non-statistical reviewers. NeuroImage 61, 1300–1310.

Friston, K., 2013. Sample size and the fallacies of classical inference. NeuroImage 81, 503–504.

Friston, K.J., Ashburner, J., Kiebel, S.J., Nichols, T., Penny, W., 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, London.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16, 465–483.

Goeman, J.J., Van De Geer, S.A., Van Houwelingen, H.C., 2006. Testing against a high dimensional alternative. Journal of the Royal Statistical Society: Series B 68, 477–493.

Goldfine, A.M., Bardin, J.C., Noirhomme, Q., Fins, J.J., Schiff, N.D., Victor, J.D., 2013. Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study". Lancet 381, 289–291.

Golland, P., Fischl, B., 2003. Permutation tests for classification: towards statistical significance in image-based studies, in: Taylor, C.J., Noble, J.A. (Eds.), Information Processing in Medical Imaging: Proceedings of the 18th International Conference, Springer, Berlin. pp. 330–341.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Hastie, T.J., Tibshirani, R.J., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer, New York.

Hoeffding, W., 1952. The large-sample power of tests based on permutations of observations. Annals of Mathematical Statistics 23, 169–192.

Hsing, T., Attoor, S., Dougherty, E., 2003. Relation between permutation-test p values and classifier error estimates. Machine Learning 52, 11–30.

Ingre, M., 2013. Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: Comment on Friston (2012). NeuroImage 81, 496–498.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the american statistical association 90, 773–795.

Lehmann, E.L., Romano, J.P., 2005. Testing Statistical Hypotheses. 3rd ed., Springer, New York.

Lindquist, M.A., 2012. Functional causal mediation analysis with an application to brain connectivity. Journal of the American Statistical Association 107, 1297–1309.

Lindquist, M.A., Caffo, B., Crainiceanu, C., 2013. Ironing out the statistical wrinkles in "ten ironic rules". NeuroImage 81, 499–502.

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R., 2014. A significance test for the lasso. Annals of Statistics 42, 413–468.

Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society, Series A 231, 289–337.

Nichols, T.E., Holmes, A.P., 2001. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping 15, 1–25.

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S., 2014. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage: Clinical 4, 687–694.

Ojala, M., Garriga, G., 2010. Permutation tests for studying classifier performance. Journal of Machine Learning Research 11, 1833–1863.

Park, T., Casella, G., 2008. The Bayesian Lasso. Journal of the American Statistical Association 103, 681–686.

Penny, W., 2012. Comparing dynamic causal models using aic, bic and free energy. Neuroimage 59, 319–330.

Pepe, M.S., Kerr, K.F., Longton, G., Wang, Z., 2013. Testing for improvement in prediction model performance. Statistics in Medicine 32, 1467–1482.

Phipson, B., Smyth, G.K., 2010. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Statistical Applications in Genetics and Molecular Biology 9.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Reiss, P.T., Huo, L., Zhao, Y., Kelly, C., Ogden, R.T., 2015. Wavelet-domain regression and predictive inference in psychiatric neuroimaging. Invited revision.

Reiss, P.T., Ogden, R.T., 2009. Smoothing parameter selection for a class of semiparametric linear models. Journal of the Royal Statistical Society: Series B 71, 505–523.

Robinson, J., 1973. The large-sample power of permutation tests for randomization models. Annals of Statistics 1, 291–296.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, New York.

Simon, R., Radmacher, M.D., Dobbin, K., McShane, L.M., 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 95, 14–18.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society: Series B 36, 111–147.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58, 267–288.

van de Wiel, M.A., Berkhof, J., van Wieringen, W.N., 2009. Testing the prediction error difference between 2 predictors. Biostatistics 10, 550–560.

Wilks, S.S., 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics 9, 60–62.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B 73, 3–36.