NeuroImage 77 (2013) 157–165

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Confounds in multivariate pattern analysis: Theory and rule representation case study

Michael T. Todd *, Leigh E. Nystrom, Jonathan D. Cohen

Department of Psychology and Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA

ARTICLE INFO

Article history:
Accepted 15 March 2013
Available online 2 April 2013

Keywords:
fMRI
MVPA
GLM
Rule representations
Prefrontal cortex

ABSTRACT

Multivariate pattern analysis (MVPA) is a relatively recent innovation in functional magnetic resonance imaging (fMRI) methods. MVPA is increasingly widely used, as it is apparently more effective than classical general linear model analysis (GLMA) for detecting response patterns or *representations* that are distributed at a fine spatial scale. However, we demonstrate that widely used approaches to MVPA can systematically admit certain confounds that are appropriately eliminated by GLMA. Thus confounds rather than distributed representations may explain some cases in which MVPA produced positive results but GLMA did not. The issue is that it is common practice in MVPA to conduct group tests on single-subject summary statistics that discard the sign or direction of underlying effects, whereas GLMA group tests are conducted directly on single-subject effects themselves. We describe how this common MVPA practice undermines standard experiment design logic that is intended to control at the group level for certain types of confounds, such as time on task and individual differences. Furthermore, we note that a simple application of linear regression can restore experimental control when using MVPA in many situations. Finally, we present a case study with novel fMRI data in the domain of rule representations, or flexible stimulus–response mappings, which has seen several recent MVPA publications. In our new dataset, as with recent reports, standard MVPA appears to reveal rule representations in prefrontal cortex regions, whereas GLMA produces null results. However, controlling for a variable that is confounded with rule at the individual-subject level but not the group level (reaction time differences across rules) eliminates the MVPA results. This raises the question of whether recently reported results truly reflect rule representations, or rather the effects of confounds such as reaction time, difficulty, or other variables of no interest.

Introduction

Analysis of functional magnetic resonance imaging (fMRI) data can be characterized in terms of two broad approaches: general linear model analysis (GLMA) and multivariate pattern analysis (MVPA). GLMA assesses, on a voxel-by-voxel basis, the mean difference in activity between, or effect of, experiment conditions (e.g., Friston et al., 1995). This leads to a voxel-by-voxel map of effects (i.e., GLMA summary statistics). At the group level, a test can be conducted at each voxel to determine whether the effect is consistent across subjects. In the MVPA methods that we consider here, developed to detect a type of "distributed representations" as discussed further below, a classifier is trained to discriminate between multivoxel patterns of activity from different experiment conditions. Here, the summary statistic is discrimination success, which is akin to *significance* of the effect, rather than the effect itself. A classifier can be trained once, on the whole brain or a particular region of interest, or many times, in small "searchlight regions centered on each voxel (e.g., Kriegeskorte et al., 2006). We focus on the latter, searchlight analysis because this approach is most comparable to GLMA, although the point of this article holds equally for whole-brain MVPA (see Discussion). Searchlight MVPA leads to a voxel-by-voxel map of local (searchlight) discrimination success (i.e., MVPA summary statistics). At the group level, a test can be conducted at each voxel to determine whether discrimination success in the surrounding searchlight is consistent across subjects.[1]

MVPA is increasing in popularity, because its use of information combined across multiple voxels makes it more sensitive than GLMA to certain types of "distributed representation." Accordingly, MVPA has successfully characterized the neural substrates of many representations that have eluded GLMA, ranging from low-level perceptual features to abstract memories or task rules (e.g., Bode and Haynes, 2009; Carlin et al., 2011; Cole et al., 2011; Haynes and Rees, 2005; Haynes et al., 2007; Kamitani and Tong, 2005; Peelen et al., 2010; Polyn et al., 2005; Reverberi et al., 2011; Vickery et al., 2011;

* Corresponding author.
E-mail addresses: mttodd@berkeley.edu, mttodd@gmail.com (M.T. Todd).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.03.039

[1] More precisely, there are two commonly used types of summary statistics in discrimination-based MVPA. The first is "classification accuracy," a function of the number of trials for which experiment condition can be correctly identified from patterns of voxel activity (e.g., Haynes et al., 2007). The second is "within-minus-across pattern similarity," or the difference between within-condition and across-condition pattern correlations (e.g., Haxby et al., 2001; Peelen et al., 2009). Both types are zero-centered under the null hypothesis of no discriminability, and both behave similarly with regard to the issue raised here. Note that either type of summary statistic can be used to characterize discrimination success of classifiers trained on searchlights or the whole brain, and thus both searchlight and whole-brain discrimination-based MVPA are subject to the concern discussed here.

Woolgar et al., 2011). As described above, the robustness of MVPA results is often established by conducting group (i.e., across-subject) tests on discrimination success. However it is not generally recognized, and is the point of this article, that group tests on discrimination success can preserve confounds that are controlled when group tests are conducted on effects themselves, as in GLMA. This is because, as noted above, discrimination success is akin to effect significance rather than to effects themselves. Since effect significance discards the sign or *direction* of the underlying effect, this information is not exposed to across-subject averaging in the group test. This discarding of direction information seems innocuous, but in fact undermines a key element of standard group test logic, which assumes that such across-subject averaging of direction occurs. As a consequence, the interpretability of such group test results may be compromised by common confounds over which control has been inadvertently loosened.

For example, across-subject counterbalancing works by ensuring that effects due to the counterbalanced variable (e.g., presentation order) are confounded with experiment condition in different directions across subjects. When experiment effects themselves are used as summary statistics, as in GLMA, effect direction is duly averaged across subjects in the group test. Because counterbalanced effects take opposite directions across subjects, the averaging process in the group test cancels these out. However, when summary statistics are akin to effect significance, as with MVPA, then effect direction is not averaged across subjects in the group test and counterbalanced confound effects do not cancel out, producing potentially spurious results. Note that the issue we describe here is not specific to confounds that are explicitly counterbalanced. In general, confounds that go in different and approximately balanced directions across subjects (e.g., random individual differences in experiment condition preference, familiarity, or difficulty) will be approximately controlled in group tests based on effects (as in GLMA), but will survive in group tests based on effect significance (as in discrimination-based MVPA). The difference between types of confounds with regard to this issue is further discussed below.

It is important to specify our particular definitions of both MVPA and "distributed representations" in order to clarify the scope of the problem that we describe. By MVPA, we refer in this article specifically to that family of methods that was developed following Haxby et al. (2001). This family of methods is unified by a particular definition of "distributed representation." Specifically, in this definition, distributed representations are those in which voxelwise effects are uncorrelated, even taking opposite directions, across neighboring voxels within a brain region (e.g., Boynton, 2005; Haxby et al., 2001; Haynes and Rees, 2005; Kamitani and Tong, 2005; Norman et al., 2006). That is, this definition refers specifically to the presence of fine-grained spatial structure within each brain region in which activity is observed. "MVPA" has then been used in this literature to refer to the family of methods that have been used to detect such distributed representations. Due to the complex, fine-grained structure of across-voxel patterns in these types of distributed representations, aggregating directional voxelwise statistics (e.g., effects) at the group level based on spatial alignment is unlikely to be fruitful. This is due to the fact that across-subject alignment is unlikely to be sufficient to align patterns with such fine spatial scale. Thus, researchers have turned to aggregating *directionless* statistics (e.g., classifier output) at the group level when using MVPA methods within this literature. However, this practice of aggregating directionless statistics leads to the problem that we describe in this article. We emphasize that the use of pattern classifiers is not the defining characteristic of MVPA as discussed in this article: there are other applications of pattern classification techniques in which directional voxelwise statistics can be appropriately aggregated at the group level (e.g., Mourão-Miranda et al., 2006). Such applications are unlikely to be able to detect the type of fine-scaled "distributed representations" addressed by Haxby et al. (2001), and are thus outside the definition of MVPA used in this article. Such methods also avoid the particular methodological problem that we describe.

Representational similarity analysis (RSA: Kriegeskorte et al., 2008) is a newer form of MVPA that is growing in popularity. The relationship between RSA and discrimination-based MVPA is analogous to that between parametric GLMA and categorical GLMA. Although RSA uses different summary statistics than discrimination-based MVPA, its summary statistics share the critical property of discarding the sign of underlying effects, so that RSA can still be susceptible to the confound issue described here (described further below).

Indeed, the issue that we have introduced here is theoretically general in that it affects any analysis that applies standard group test logic to individual summary statistics that discard the sign of underlying effects. Although it is beyond the scope of this article to thoroughly survey all such methods, our point is to recognize that this class includes many applications of MVPA as defined above. We further consider the generality of the issue in the Discussion. To illustrate the problem concretely, we present several simulated examples in the next section.

Simulations

Simulation 1: individual differences and manipulated variables of no interest

In the first simulation (Fig. 1), effects due to random individual differences are controlled in GLMA but not MVPA.[2] An experimenter seeks to determine whether a neural signal (e.g., voxel activity) differs across two experiment conditions (e.g., use of rule A vs. B to perform a task). Unknown to the experimenter, voxel activity is unresponsive to rule, but is responsive to difficulty. Furthermore, it happens incidentally that rule A is more difficult than rule B for some subjects, whereas the reverse is true for other subjects. Thus, at the individual-subject level, rule and difficulty are confounded and experiment effects (i.e., mean A activity minus mean B activity) appear robust due to the confound. However, the task that is more difficult varies randomly across subjects, and therefore difficulty effects are approximately counterbalanced across subjects, and will cancel out when experiment effects are averaged in a group test, as in GLMA. Accordingly, GLMA group tests are not affected by the task difficulty confound (and should produce a null result in this example, since we assumed that there was no actual effect of the rule). In contrast, group tests that average discrimination success, as in discrimination-based MVPA, fail to mitigate this same confound. The problem is that discrimination success reflects only the robustness of the individual experiment effect, and is therefore positive whenever individual-subject level experiment effects are significant, irrespective of effect direction. Thus, confounding difficulty effects will not cancel out when A vs. B discrimination success is averaged at the group level. This leaves open the opportunity to misinterpret MVPA results as evidence for neural differences due to rule, rather than to task difficulty (which, in this example, is the actual cause of the observed neural effect). Thus GLMA mitigated the difficulty confound, whereas MVPA did not.

It is important to note that, if most subjects in a sample experience the *same* rule condition as more difficult than the other (e.g., if most subjects experience rule A as more difficult than rule B), then this confound will be reflected in GLMA as well as MVPA group tests. This is widely recognized, and is a motivation for the standard practice of conducting group tests on behavioral measures as a complement to analyses of the imaging data (e.g., a group *t*-test on the effect of rule on RT). If such a test is significant then the experimenter considers

[2] Simulation 1 details: For the i-th subject on the n-th trial, difficulty is: $d_{i,n} = c_i(A_n - B_n) + \xi_{i,n}$, where A_n and B_n are binary indicators for experiment condition, $\xi_{i,n}$ is a noise term, and the confound weight, c_i, is positive for subject 1 but negative for subject 2. Then, activity of the j-th voxel is simply positively weighted difficulty plus noise: $v_{i,j,n} = b_{i,j}d_{i,n} + \ _{i,j,n}$. Twenty voxels were simulated for each subject, but just one voxel is illustrated for each subject in Fig. 1. The Gaussian Naïve Bayes (GNB) classifier was used for classification.