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).

[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.

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

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.

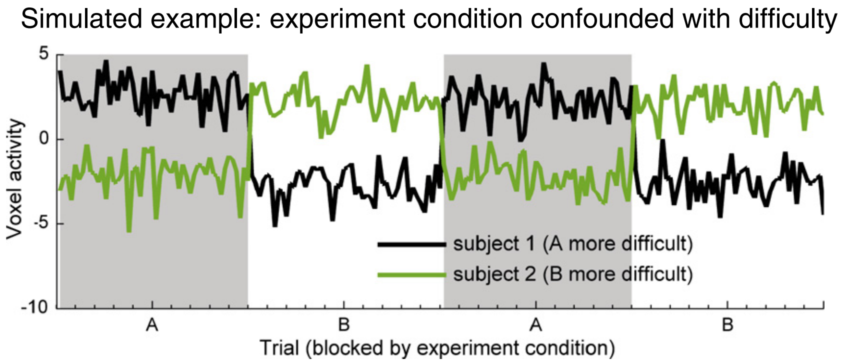
Simulations

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.



Individual-Subject Summary Statistics

Subject	Experimental Effect (GLM)	Discrimination Success (MVPA)
Subject 1	mean(A)-mean(B) = +4.75	classification accuracy = +13.15, within-minus-across = +3.826
Subject 2	mean(A)-mean(B) = -5.56	classification accuracy = +13.44, within-minus-across = +3.848

Group Test Statistics (two-tailed t-test)

Experimental Effect (GLM)	Discrimination Success (MVPA)
mean(A)-mean(B): t_1=-0.0780, p=0.9504, n.s.	classification accuracy: t_1=94.0, p<0.01, sig.
	within-minus-across: t_1=348, p<0.01, sig.

Fig. 1. Simulation 1. Experiment condition (rule A vs. rule B) does not affect voxel activity, but difficulty does. Moreover, subject 1 experiences condition A as more difficult, whereas subject 2 experiences condition B as more difficult. (Top) As shown in the activity of a single voxel (20 voxels were simulated for each subject), at the individual-subject level experiment effects are robust but differ in direction across subjects. (Middle) GLMA uses experiment effects as individual-subject summary statistics. These are positive for subject 1 and negative for subject 2. MVPA uses discrimination success (specifically, classification accuracy) as individual-subject summary statistics. These are positive for both subjects, reflecting only robustness and not direction of the underlying effects. (Bottom) A group test (two-tailed t-test) on GLMA summary statistics averages effect direction across subjects. Therefore, the confounding difficulty effects cancel out at the group level, the test statistic is approximately zero, and GLMA does not reject the null hypothesis. However, the same group test on MVPA summary statistics does *not* average effect direction across subjects, because MVPA summary statistics discard this information. Thus, confounded difficulty effects do not cancel out, the test statistic is not approximately zero, and MVPA rejects the null hypothesis erroneously, due to the confound. This invites the misinterpretation that MVPA is more sensitive than GLMA to the experiment effect, whereas in fact the MVPA result reflects a confound.

rule to be confounded with RT (which may be a proxy for difficulty or other variables) at the group level, and accordingly interprets any GLMA group results with caution. Critically, however, testing for confounds in the behavioral data at the *group* level is an insufficient diagnostic for group tests that do not average effect directions, as is often the case with MVPA. Instead, in such cases, confounds must be diagnosed at the *individual-subject* level.

For consistency with the case study presented below, we have described this simulated example in terms of random individual differences. However, this illustrates a more general point. By simply renaming the "difficulty" variable, this same simulation can illustrate how any variable that is confounded at the individual-subject level with the variable of interest remains confounded in group tests that rely on discrimination success rather than effects themselves. Importantly, this includes the large class of confounds arising from manipulated variables of no interest that are explicitly counterbalanced across subjects by the experimenter (e.g., condition presentation order, cue-to-condition mapping, etc.). Indeed, for explicitly counterbalanced variables, the discrepancy in experimental control afforded by the different approaches to group testing is greatest. This is because counterbalancing will ensure that the confound is perfectly controlled when the group test is applied to *effects*, whereas the confound remains fully uncontrolled when the group test is applied to discrimination success. In sum, this first simulation illustrates a general point that goes beyond random individual differences and includes any variable that is confounded at the individual-subject level. The observation that counterbalancing at the group level can maximize the disparity between GLMA and MVPA is striking, since this standard element of experiment design is often the only approach to controlling variables of no interest (e.g., time on task).

Simulation 2: controlling confounds with linear regression

The discussion above demonstrates that, unlike with GLMA, standard approaches to MVPA cause variables of no interest that are confounded at the individual-subject level but controlled at the group level (either by explicit counterbalancing or by random individual differences) to remain confounded in the final group test results. For such situations, linear regression can be used to control confounds after the dataset has been collected (cf. Cohen et al., 2003). Fig. 2 shows a further simulation of the example used above involving a confound of individual differences in rule difficulty.[3] This simulation illustrates how the confound can be mitigated by the use of linear regression prior to the application of MVPA. We note that linear regression is widely used to control for confounds in many types of analysis, including GLMA. Our point here, then, is twofold: first, that linear regression can be used (during preprocessing) in MVPA as well, and second that a wider range of confounds should be controlled in this fashion during MVPA as compared to GLMA, due to the fact

[3] Simulation 2 details: identical to Simulation 1, except that c_i now has the standard normal distribution across 500 simulated subjects. Thus some subjects find rule A more difficult, others find rule B more difficult, and still others find both rules to be of approximately equal difficulty. Twenty voxels are simulated for each subject.

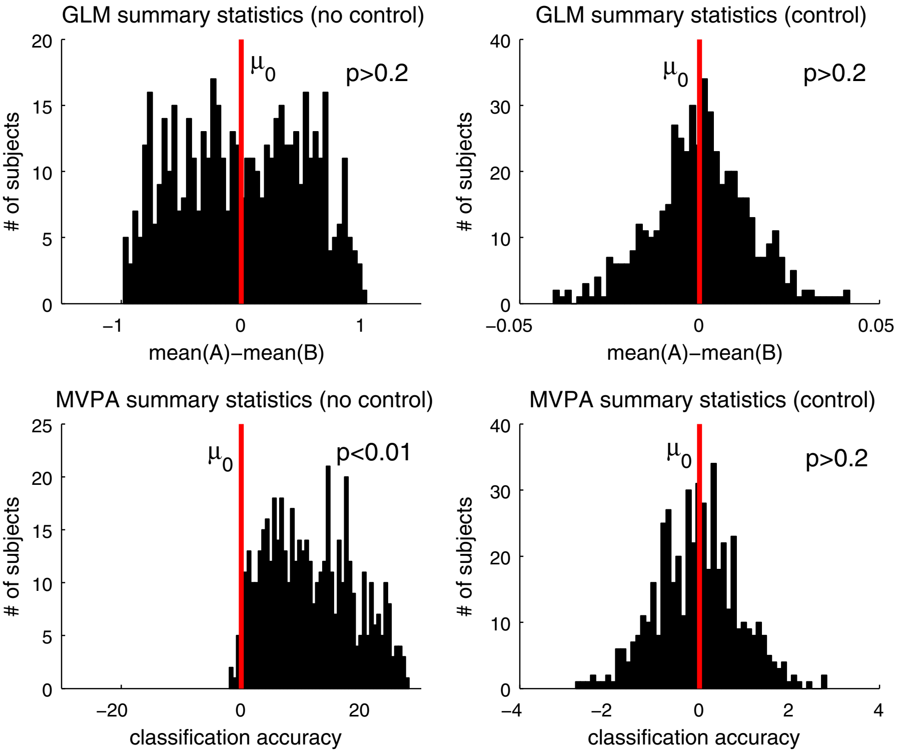


Fig. 2. Simulation 2. Experiment condition (rule A vs. rule B) does not affect voxel activity, but difficulty does. Moreover, experiment condition and difficulty are confounded at the individual-subject level in random directions across 500 subjects. Histograms depict the sample of individual-subject experiment effects (GLMA) or of classification accuracy (MVPA), and corresponding group test results (i.e., inset *p*-values reflect results of two-tailed group *t*-test). The GLMA test on experiment effects fails to reject the null hypothesis both before (Upper-Left) and after (Upper-Right) controlling for difficulty via linear regression. (Lower-Left) Before controlling for difficulty, the MVPA group test on classification accuracy rejects the null hypothesis, because classification accuracy tends to be positive at the individual-subject level. (Lower-Right) After removing difficulty from the voxel signal with linear regression prior to MVPA, the MVPA group test on classification accuracy fails to reject the null hypothesis, as appropriate.

that MVPA more often than GLMA would otherwise allow the confounds to survive in group test results.

Simulation 3: Representational similarity analysis

RSA is a newer form of MVPA that can be seen as a generalization of discrimination-based MVPA to allow the testing of more specific and informative research hypotheses regarding activity patterns. The relationship between RSA and discrimination-based MVPA is analogous to that between parametric and categorical GLMA. That is, RSA requires multiple (more than two) experiment conditions and the specification of a hypothesized ranking or ordering of a similarity measure (e.g., interpattern correlations) among condition pairs. A common RSA procedure would be to compute, for each subject, the z-scored pattern correlation between each pair of experiment conditions, and then form a weighted sum of these z-scored correlations, with each weight proportional to the hypothesized similarity of the two conditions in terms of a manipulated variable (plus an additive constant so that all weights sum to zero). This *weighted correlation sum* is the individual-subject summary statistic used by RSA. It has zero population mean in many cases, such as when pairwise correlations are all equivalent, or are modulated by a variable that is unrelated to the variable of interest (which was used to set the weights) at the individual-subject level. For example, Schapiro et al. (2012) used RSA to test whether training increased the similarity of frequently paired fractal image representations more than it increased the similarity of infrequently paired fractal image representations. This was accomplished by specifying weights of 1, −1, −1, and 1 for, respectively, the frequent-pair-post-training correlations, the frequent-pair-pre-training correlations, the infrequent-pair-post-training correlations, and the infrequent-pair-pre-training correlations.

Due to the increased specificity afforded by RSA as compared to discrimination-based MVPA, RSA may yield results in which it is more difficult to envision plausible confounds, as in the Schapiro et al. (2012) study. Nonetheless, the fundamental issue described in this article does generalize to RSA. Specifically, the weighted correlation sum summary statistic used in RSA will have a non-zero population mean (leading to rejection of the null hypothesis) whenever the pattern correlations are in fact modulated by a variable that is confounded at the individual-subject level with the variable of interest.

In particular, the rule-difficulty confound simulation described above for discrimination-based MVPA can be readily generalized into an RSA confound simulation.[4] Suppose that the experimenter now uses four "compound rules" (W vs. X vs. Y vs. Z), each composed of three "simple rules" (i.e., drawn from the set of A, B, C, D, E and F), similar to the designs used by Reverberi et al. (2011) and Cole et al. (2011). The design is such that each pair of compound rules has either two, one, or zero simple rules in common. The experimenter seeks to determine whether a neural signal (e.g., voxel activity) is driven by the simple rule manipulation. The experimenter can use RSA to determine whether the activity in a set of voxels exhibits a similarity structure (i.e., an ordering among pattern correlations) across compound rule pairs that is consistent with modulation by simple rules. However, unknown to the experimenter, voxel activity is not affected by simple rule, but is affected by difficulty. Moreover, assume that one subject

[4] Simulation 3 details: For the *i*-th subject on the *n*-th trial, difficulty is: $d_{i,n} = c_i^T x_n + \xi_{i,n}$, where $x_n = (A_n, B_n, C_n, D_n, E_n, F_n)^T$ is a vector of binary indicators for simple rule condition, $\xi_{i,n}$ is a noise term, and the confound weight vector, c_i, varies across subjects. In particular, $c_1 = (0.5, 0.25, 0, -0.25, -0.5, -0.75)^T$, and $c_2 = (-0.75, -0.5, -0.25, 0, 0.25, 0.5)^T$. 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. 3.

experiences the difficulty ranking of the simple rule conditions as $A > B > C > D > E > F$, whereas another subject experiences the difficulty ranking as $A < B < C < D < E < F$. The results of using RSA in this scenario are shown in Fig. 3. Briefly, a group test leads to spurious rejection of the null hypothesis with RSA due to the rule-difficulty confound. Moreover, a standard, group-level behavioral diagnostic (e.g., a 6-way ANOVA examining the effect of simple rule on RT) would fail to detect the confound, as the confound effects go in opposite directions across subjects.

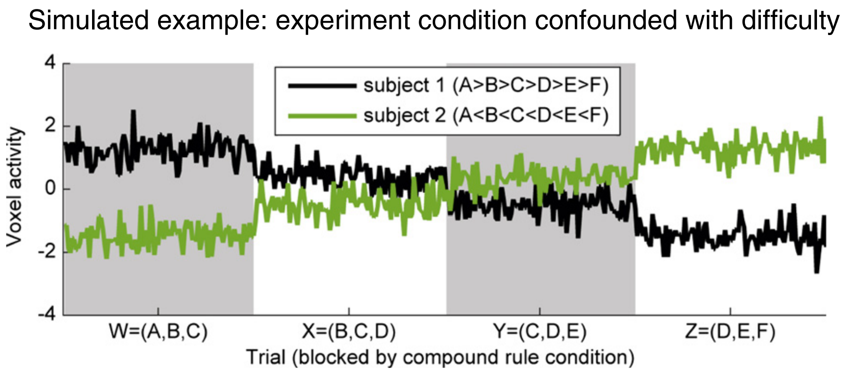
As with the previous simulations, this individual difference example was chosen for consistency with the case study described below, but the point about RSA extends to manipulated variables that are confounded with the variable of interest at the individual-subject level and counterbalanced at the group level (such as condition presentation order and cue-to-condition mapping). Note that, although parametric GLMA is somewhat analogous to RSA, there is no clear application for parametric GLMA in this example, because the simple rule conditions are purely categorical, and there is no a priori ranking of effects themselves associated with compound rules (in contrast to the clear a priori ranking of the similarity measures among compound rules). Accordingly we have not included a GLMA component in this simulation. However, for cases in which there is a clear a priori ranking of activation across conditions, it is again important that parametric GLMA applies group tests to effects themselves (rather than similarity measures), so that variables that are confounds only at the individual-subject level are controlled in group tests, whereas, as we have shown here, there is no corresponding control in RSA.

Below, we present a case study using discrimination-based MVPA on novel fMRI data that is similar to the simulated examples presented above: a confound between experiment condition (rule A vs. B) and RT (which serves as an observable proxy for difficulty or other unobservable variables) goes in random directions across subjects, leading to appropriately null GLMA results, but to misleading, positive MVPA results. We then demonstrate the efficacy and specificity of using linear regression to control for this confound.

Case study: flexible rule representations

An important goal of human brain imaging studies is to identify and track neural signals associated with internal mental states more precisely than can be achieved with behavior alone. One such class of mental states is rule representations; that is, representations of flexible stimulus–response mappings needed to perform a specified task. GLMA has not provided reliable results in identifying and tracking rule representations. However, there have been several reports using MVPA to identify rule representations. For example, Haynes et al. (2007) reported that (the intention to perform) addition versus subtraction is represented in medial frontal cortex. Bode and Haynes (2009) subsequently reported a within-trial temporal progression of rule representations from parietal to prefrontal cortex. More recently, Reverberi et al. (2011), Woolgar et al. (2011), and Cole et al. (2011) have all extended these initial findings. However, these MVPA studies may be vulnerable to the problem described above, given that factors such as difficulty, effort, and/or performance may be confounded with task rule. Moreover, these confounds may go in opposite directions across subjects, so that they are controlled in GLMA group tests and standard, group-level behavioral diagnostics, yet pose a problem for the interpretation of MVPA group test results.

Here we present an empirical case study in which such a confound is present in the data: a difference in reaction time (RT) between rules. For illustration purposes, we described our simulations in terms of a task difficulty variable. However, the use of RT, which is what we actually



Individual-Subject Summary Statistics

Subject	Weighted Correlation Sum (RSA)
Subject 1	0.66 r(W,X) - 0.33 r(W,Y) - 1.33 r(W,Z) + 0.66 r(X,Y) - 0.33 r(X,Z) + 0.66 r(Y,Z) = 4.34
Subject 2	0.66 r(W,X) - 0.33 r(W,Y) - 1.33 r(W,Z) + 0.66 r(X,Y) - 0.33 r(X,Z) + 0.66 r(Y,Z) = 4.09

Group Test Statistics (two-tailed *t*-test)

Weighted Correlation Sum (RSA)

$t_1 = 13.34$, p<0.05, sig.

Fig. 3. Simulation 3. Experiment condition (Simple rule: A vs. B vs. C vs. D vs. E vs. F) does not affect voxel activity, but difficulty does. Subjects 1 and 2 experience different rule difficulty rankings. (Top) As shown in the activity of a single voxel (20 voxels were simulated for each subject), at the individual-subject level experiment effects are robust but differ in direction across subjects. (Middle) RSA uses weighted correlation sum as the individual-subject summary statistic. The correlations are simply the Pearson product-moment correlation coefficient (z-scored using the Fisher transformation) between each pair of compound rule patterns. The weights are the number of simple rules that each compound rule pair shares in common, with the mean weight subtracted so that the sum over all weights is zero. Despite the fact that the confound between simple rule and difficulty takes different directions across subjects, the weighted correlation sum is positive for both subjects. (Bottom)A group test (two-tailed *t*-test) on these RSA statistics does *not* average effect direction across subjects. Thus, RSA rejects the null hypothesis due to the confound.

measured in this experiment, as a proxy for task difficulty *per se* is open to debate in this setting and we remain agnostic regarding which underlying variable (e.g., task difficulty, mental effort, or simply time on task) is driving RT and (as shall be shown), MVPA results. This does not impact our methodological point. Indeed, it is central to our point that the underlying variable cannot be known in the simple and standard rule representation design that we have used, which is why the conservative approach of removing RT is necessary. Since the rule that is performed more slowly varies randomly across subjects, the effect of this confound is approximately counterbalanced in the group, and is thus automatically eliminated during GLMA group tests. Accordingly, we show that GLMA fails to find rule-related neural activity. We then show that application of MVPA to the same data, using standard methods of group testing, appears to reveal patterns of activity that distinguish between rules, inviting interpretation of these as rule-related representations. Indeed, the results look strikingly similar to many of the reported MVPA-based rule representation findings in the literature. However, when we use linear regression to control for the RT confound prior to the application of MVPA, the findings disappear, commensurate with GLMA.

Materials and methods

Participants

Forty subjects (22 male; age M = 21.11 years) received $20/hour plus a performance-based bonus (approximately $10). The experiment was approved by the Princeton University Institutional Review Board for human subjects research. One subject's data were discarded due to equipment failure.

Task design & procedure

The task was a variant of the AX-CPT task (e.g., Barch et al., 1997; Bode and Haynes, 2009; Rosvold et al., 1956; Servan-Schreiber et al., 1996). Subjects pressed left, right, or down buttons on a right-hand response box. Trials were either cue trials or probe trials. On cue trials, participants saw a cue stimulus ('A' or 'B'), and pressed down to acknowledge the cue. On probe trials, participants saw a probe stimulus ('X' or 'Y'), and pressed left or right according to a rule designated by the most recent preceding cue: if A, then X = right, Y = left; if B, then X = left, Y = right. Since these rules specified incompatible probe-response mappings, correct responses required representation of the rule associated with the most recent cue. Each trial began with a 0.5 s stimulus presentation, and participants had 3 s to respond. On timeout and error trials only, participants received 1 s feedback screens. After the response (and any feedback), a blank screen was presented for the remainder of the intertrial interval (ITI: 10 ± 0.5 s between stimulus onsets). There were 54 trials (540 seconds) per block, and five experiment blocks. Each trial had a 50% chance of being either a new cue (i.e., a rule switch) or a probe stimulus, with the constraint that at least one probe always followed a cue. Therefore, rule, probe, and motor response were statistically balanced within-subject. Probe response RT was analyzed for the effects of rule and motor response.

fMRI data acquisition

We used a 3 T Siemens Allegra head-only MRI scanner. Anatomical images were acquired using a T1-weighted MPRAGE (FoV, 256 mm; matrix, 256 × 256; 176 1 mm sagittal slices). Functional images were acquired using a T2*-weighted echo-planar pulse sequence (TR, 2 s; echo time, 30 ms; flip angle, 75 degrees; in-plane resolution, 3 × 3mm; thickness, 3 mm; gap, 1 mm; interleaved slices; slices aligned to anterior-posterior commissural axis). Each experiment block corresponded to a single functional run.

fMRI analysis

Data were processed using SPM8 and custom MVPA code that implemented standard methods of both pattern classification and group testing (see below). Preprocessing included rigid-body realignment. The anatomical volume was transformed to standard space using SPM8's segmentation-based nonlinear warp. Prior to GLMA, functional volumes were warped and then spatially smoothed with an 8 mm FWHM Gaussian kernel. Prior to MVPA, each volume was spatially z-scored across voxels, and no smoothing was applied; the warp was applied later, to single-subject classification accuracy maps rather than to functional volumes.

We focused our analysis on probe trials immediately following a cue (i.e., in which the rule had just switched), since overall RT is longest on these trials (Monsell, 2003), and on the time of probe onset (when subjects had to use the rule to decide how to respond). Thus approximately 90 trials were analyzed per subject, 45 of each rule (or motor response). We searched for differences between rule (A vs. B) and motor response (right vs. left). The latter was included to demonstrate that controlling for RT had a specific effect on the results of the rule analysis, and not a more general effect of reducing statistical power (which should then have also diminished any observed effects of motor response). Error trials were excluded from all analyses.

GLMA included a 128 s high-pass filter, and the design matrix included an event-related regressor for each rule (or response), time-locked to probe onset and convolved with the canonical HRF in SPM. Covariates included the six head motion parameters that were estimated during realignment. Individual-subject A > B (or left > right) effect maps were entered into a group test.

For MVPA, we used a Gaussian Naïve Bayes (GNB) classifier (e.g., Pereira et al., 2009) with a 4-voxel radius searchlight approach (Kriegeskorte et al., 2006) for exploratory, whole-brain analysis, similar to MVPA analyses in recent reports (Bode and Haynes, 2009; Reverberi et al., 2011; Woolgar et al., 2011; Cole et al., 2011). The temporal unit for the classifier was single-trial BOLD signal. That is, BOLD signal was extracted at each probe onset time (plus 6 s for hemodynamic lag), labeled (i.e., A/B or left/right), and grouped into training and testing sets. We generated single-subject classification accuracy maps using leave-1-run-out cross-validation (e.g., Pereira et al., 2009). Maps were then warped into MNI space and entered into a group test. Prior to all MVPA analyses, we used linear regression to remove low frequency cosines (implementing a 128 s high-pass filter) as well as head motion parameters, similar to GLMA. Initially, no other variables were removed. In a second analysis, we tested the effect of subtracting the spatial mean from each searchlight, as suggested by Cole et al. (2011). In a third analysis, trial-by-trial RT was removed in two passes: simultaneously with the cosines and head motion parameters, and then again from the set of classifier examples immediately prior to MVPA.

Group tests were conducted and corrected for whole-brain cluster-based significance at the 0.05 level using threshold-free cluster enhancement (TFCE: Smith and Nichols, 2009), a parameter-free algorithm that achieves a cluster-corrected result analogous to a *t*-test.

Results

Behavioral data

There were no significant effects of rule or motor response on accuracy. For RT, a within-subjects ANOVA of rule (A, B) by motor response (left, right) revealed only a main effect of motor response, with left responses slower than right, $F(1,38) = 12.87$, $p = 0.0009$. Critically, there was no main effect of rule ($p = 0.5743$), indicating that any individual-subject effects of rule on RT were canceled out in the group analysis. However, separate *t*-tests for each subject revealed that such individual-subject effects were in fact present in a sizeable subset of subjects. RT differed significantly across rules for 13 of 39 subjects: eight subjects were significantly slower for the A rule, and five subjects were significantly slower for the B rule. A binomial test of this eight-to-five split failed to reject the null hypothesis that subjects in the population were equally likely to respond more slowly to one rule versus the other, $b(13,0.5) = 8$, $p = 0.1134$. Thus, although RT was confounded with rule in a subset of subjects, the confound took random and approximately balanced directions across subjects, and was therefore controlled at the group level by standard analysis of the behavioral data.

fMRI data

fMRI results are shown in Fig. 4. Consistent with the literature, GLMA revealed no significant effects of rule, whereas standard, discrimination-based MVPA showed significant results in bilateral middle frontal gyrus, left pre- and post-central gyrus, left caudate, and left superior parietal lobule. Subtracting the searchlight mean from each classifier example eliminated the superior parietal lobule cluster, but left all other MVPA results intact. However, removing RT by linear regression eliminated all significant results related to rule (corrected $p > 0.2$ in all searchlights). This suggests that the apparent effect of rule observed using standard MVPA methods was driven by a confound with RT. Both GLMA and MVPA revealed significant motor response (left vs. right) differences in left pre- and postcentral gyri, consistent with subjects' use of the right (i.e., contralateral) hand for all responses. Moreover, MVPA motor response results were unaffected by removal of the effects of RT by linear regression, demonstrating specificity of this method of control, and further implicating the confound with RT as the source of apparent rule-related results.

To further establish that a confound with RT drove the initial MVPA results, we conducted two auxiliary GLMA tests. Results are shown in Fig. 5. In one, we included RT as a parametric regressor, and tested this regressor against zero in a group *t*-test. It is clear that RT exhibited strong, uniformly positive correlations with BOLD signal throughout much of the brain. Indeed, over half of the voxels in the SPM8 probabilistic gray-matter mask (thresholded at 0.1) were found to be significantly positively correlated with RT after whole-brain correction. In a second auxiliary test, we determined which rule led to faster RT separately for each subject, and then tested the slow rule > fast rule contrast in the group. Given the strength and broad distribution of the RT correlation, we reasoned that even a weak relationship between rule and RT on a per-subject basis would lead to diffuse patterns of weak, positive activation in response to the slow rule > fast rule contrast. Indeed, a weak, diffuse, positive pattern is shown in Fig. 5B. Although this whole-brain, exploratory GLMA did not produce significant results after correction for multiple comparisons, a more focused ROI analysis confirmed that slow rule trials led to significantly more activation than fast rule trials within the regions where MVPA rule classification was initially successful, $t(38) = 1.726$, $p = 0.0462$. It is entirely plausible that this pattern, which is weak in each voxel but consistent across many voxels, would lead to robust exploratory results with MVPA but not GLMA, since MVPA can aggregate over weak signals in many voxels more effectively than GLMA. The spatial overlap between the slow > fast GLMA result and A versus B MVPA result suggests that the reason that MVPA succeeded in rule classification was the confound with RT.

To further clarify, we note that although the GLMA A > B rule contrast produced different results than the GLMA slow > fast rule contrast, the analogous MVPA classification results are identical to each other. This follows from the fact that the slow/fast variable was defined to be perfectly collinear or confounded with the A/B variable at the individual-subject level, but not necessarily at the group level. Thus MVPA classification on either A versus B or fast versus slow would produce identical results, whereas, because the fast rule was A

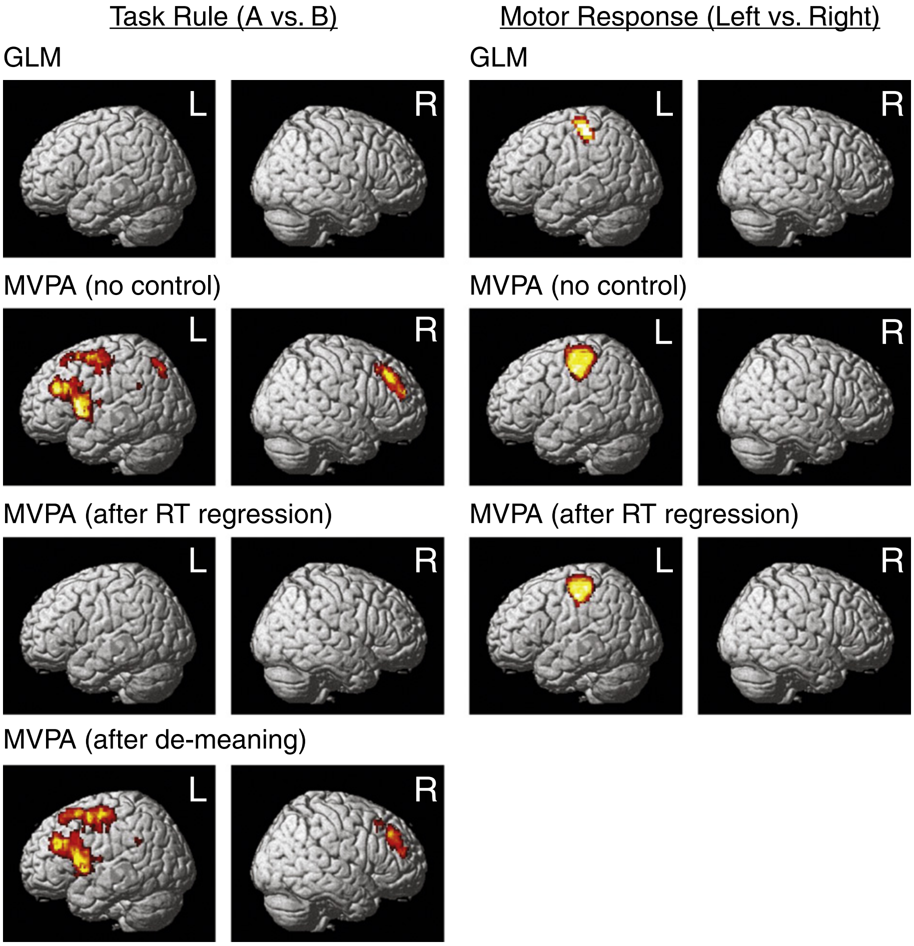


Fig. 4. Results of rule (A/B) and motor response (Left/Right) analyses. Overlays depict inverse corrected *p*-values, using TFCE whole brain cluster correction (Smith and Nichols, 2009).

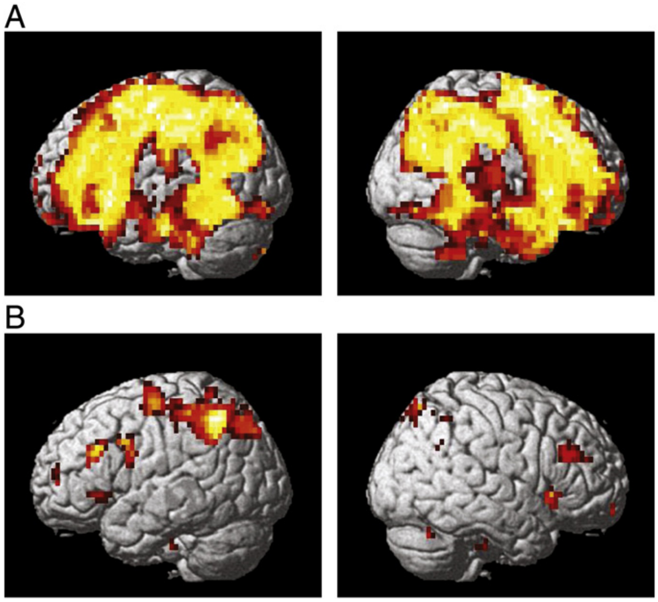


Fig. 5. GLMA reveals RT effects. (A) Most cortical voxels are significantly correlated with RT. Overlay depicts inverse corrected p-values, thresholded at 0.05, using TFCE whole-brain cluster correction (Smith and Nichols, 2009). (B) Slow rule > fastrule contrast produces weak effects, reflecting the strong relationship between RT and voxels, and the weak relationship between rule and RT. Overlay depicts group t-statistics, thresholded at $p = 0.01$, uncorrected, for illustration purposes.

in some subjects but B in others, the slow > fast effects cancel out in the GLMA A > B group result. This is another way of illustrating our overall point, and of motivating these auxiliary GLMA tests in terms of determining which difference, fast versus slow or A versus B, drove the initially positive MVPA result.

Finally, although not presented here, we conducted a follow up experiment with 40 new subjects. The design was similar to that described above, but response deadline was reduced from three seconds to one second, reducing RT variability across rules. MVPA on this follow-up dataset did not identify rule representations anywhere in the brain, although a motor response (left versus right) analysis produced results that were essentially identical to those reported here. This follow-up is consistent with the interpretation that the initially positive MVPA results reported here reflected the confound with RT rather than rule representations.

Discussion

We have pointed out that standard MVPA group tests can admit confounds that are appropriately controlled in GLMA group tests. The issue is that in MVPA, group tests are commonly applied to individual-subject summary statistics (e.g., discrimination success) that discard the sign or direction of underlying effects. This violates the ubiquitous assumption that individual-subject effect directions are averaged in group tests and exposes group test results to confounds that are usually controlled by standard design and diagnostic practices. We illustrated this issue in simulations for both discrimination-based MVPA and RSA. We further explored this issue in an experimental case study designed to identify rule representations. In our experimental data, GLMA produced null results while, similar to recent rule studies (Bode and Haynes, 2009; Cole et al., 2011; Haynes et al., 2007; Reverberi et al., 2011; Woolgar et al., 2011), MVPA results were positive. However, use of linear regression to control for RT (which may be a proxy for difficulty or effort) eliminated the MVPA results. Moreover, a follow-up experiment with restricted RT variability yielded null MVPA results. This raises the question of whether recent results using MVPA to identify rule representations in fact reflect the influence of confounds in the manner that we identify here.

Given this issue, it may prove difficult to redesign experiments to control for confounds such as time on task or individual differences in MVPA group test results. For cases in which confounds cannot be removed by experiment design, linear regression is a standard tool for restoring control (Cohen et al., 2003). In our experiment, linear regression produced effective and specific confound control. In contrast, the method of subtracting the mean from each classifier example prior to MVPA (e.g., Cole et al., 2011) did not effectively control the confound.

It is important to characterize the scope of the concerns raised here, with regard to how the results of MVPA are interpreted. The concern is limited to cases in which a brain region revealed by MVPA is interpreted as representing a particular cognitive variable or type of information (e.g., “Brain region A represents information X”). In contrast, claims of the following form are not problematic: “Brain region A can predict behavior Y,” or, “The relationship between brain region's A and behavior Y follows model Z.” These latter claims are permissible because they do not depend on interpreting the brain regions activity as representing a particular type of information. For region's Polyn et al. (2005) reported that MVPA can extract information that can be used to predict the semantic category (i.e., face, location, or object) of a recalled item several seconds before the recall behavior occurs. It is sufficient for this claim that *any* neural signal predicts the recalled item substantially prior to the recall behavior, whether it actually represents the recalled item or some other variable that is correlated (i.e., confounded) with it. Similarly, Haynes et al. (2007) reported that MVPA can extract information that predicts subjects' intentions (i.e., to use one or the other of two arithmetic rules), prior to the time at which subjects behaviorally report awareness of the intention. Again, concerns about confounds are irrelevant to this claim, because it does not interpret the brain activity as representing a particular cognitive variable. However, both Polyn et al. (2005) and Haynes et al. (2007) make other claims that do interpret the representations revealed by the MVPA results. These interpretations may be more problematic given the issue raised here.

The nature of the interpretation drawn from MVPA results is related to the issue of whole-brain versus searchlight MVPA. Whole-brain MVPA can be considered a special case of searchlight MVPA, in which the entire brain (or a single, pre-selected, large subset of voxels) is analyzed as a single large searchlight. Accordingly, it is subject to the same concerns discussed here with respect to searchlight-based methods. In practice however, whole-brain MVPA is more often used to make claims such as, “The relationship between pattern of brain activity A and behavior Y follows model Z.” As noted above, such claims are not subject to the issue raised in this article.

Although the issue raised here only affects certain types of claims, it does affect any analysis method that applies the logic of a group test to individual-subject summary statistics, such as discrimination success, that discard the sign or direction of underlying effects. Although a thorough survey is beyond the present scope, we suspect that this includes all methods currently in use to detect “distributed” representations — i.e., those in which positively and negatively signed voxelwise effects are randomly intermixed at a fine spatial scale. This class of methods includes discrimination-based MVPA as well as RSA (both whole-brain and searchlight). Also included are methods that conduct cross-validation at the group level on transformed individual-subject data, when the transformation that is used discards the signs of the underlying voxel signals, as is the case with the Procrustean transform-based “hyperalignment” method introduced by Haxby et al. (2011). There are forms of MVPA that do not discard the sign or direction of individual effects prior to aggregating at the group level — e.g., MVPA using cross-validation at the group level (Mourão-Miranda et al., 2006). However, such methods are conceptually and theoretically quite similar in functionality to GLMA in that they average something akin to effects themselves (i.e., summary statistics that include directional information) across subjects, and are thus unlikely to be able to detect distributed representations at the group level. The reasoning behind this assertion is that, as is well known from nearly two decades of GLMA fMRI studies, existing alignment algorithms are unlikely to provide sufficient intersubject registration for distributed patterns to be aligned across subjects. Thus, finely distributed patterns will tend to "wash out" in any group test that does *not* discard effect sign or direction. Thus, we suspect that future work may show that the problem of detecting distributed patterns at the group level must be fundamentally reframed before it can be solved while avoiding the confound issue that we have described here.

We have focused on the specific example of a confound between rules and individual differences in task performance (e.g., difficulty and/or RT), as such confounds are a common problem in research on cognitive control, a domain for which we had relevant case study data. However, the issue generalizes to other representational domains as well. For example, Oruç and Barton (2011) used discrimination-based MVPA to distinguish between different exemplars of faces as well as different exemplars of objects (i.e., for within-category discrimination). Their goal was to determine which brain regions, of those that are known to discriminate between faces and objects, represent the features of faces or objects *per se*. They concluded that within-category discrimination succeeded in certain areas, and interpreted those areas as representing within-category features. However, it is possible that individual differences in other factors – for example, which faces were perceived as more familiar, or which cars were experienced as more desirable – drove an attentional mechanism that in turn drove the MVPA result. Again, while such factors may have been approximately counterbalanced at the group level, MVPA results based on group tests of discrimination success would still be subject to such confounds which could, in turn, compromise interpretation of the findings in terms of category-representation.

Finally, we note that although the logic of the issue described here generalizes broadly, its impact in any particular case depends on the existence of plausible confounds. For example, we showed via simulation that newer RSA forms of MVPA are theoretically susceptible to the same problem of confounds that affects discrimination-based MVPA. However, RSA offers the ability to specify hypotheses that are substantially more specific, and this may make it possible to rule out plausible confounds in a given study. For example, it is not immediately obvious that there are plausible confounds in the RSA study of Schapiro et al. (2012). Unfortunately, taking advantage of this property of RSA requires more elaborate experimental designs, and a thorough analysis of the strategy of using such designs to mitigate the confound issue raised here is beyond the scope of this article. However, RSA is clearly an attractive avenue for advancing MVPA methods for multiple reasons, including the potential to mitigate the confound issue described here.

In conclusion, MVPA is a powerful, emerging set of tools for fMRI research. However, these tools are still evolving, and we have demonstrated one way in which they may systematically admit confounds that can lead to misinterpretation of results. Newer forms of MVPA, such as RSA, may help to reduce this problem. For standard MVPA decoding analysis, linear regression can sometimes be used to control for confounds.

Conflict of interest

The authors state that there is no conflict of interest regarding the publication of this work.

References

Barch, D.M., Braver, T.S., Nystrom, L.E., Forman, S.D., Noll, D.C., Cohen, J.D., 1997. Dissociating working memory from task difficulty in human prefrontal cortex. Neuropsychologia 35 (10), 1373–1380 (Oct).

Bode, S., Haynes, J.-D., 2009. Decoding sequential stages of task preparation in the human brain. NeuroImage 45 (2), 606–613 (Apr).

Boynton, G.M., 2005. Imaging orientation selectivity: decoding conscious perception in v1. Nat. Neurosci. 8 (5), 541–542 (May).

Carlin, J., Calder, A., Kriegeskorte, N., Nili, H., Rowe, J., 2011. A head view-invariant representation of gaze direction in anterior superior temporal sulcus. Curr. Biol. 21 (21), 1817–1821.

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd edition. Lawrence Erlbaum Associates.

Cole, M., Etzel, J., Zacks, J., Schneider, W., Braver, T., 2011. Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex. Front. Hum. Neurosci. 25 (3), 607–623.

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., Frackowiak, R., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2 (4), 189–210.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex — supplemental material. Science 293 (5539), 2425–2430 (Sep, notes supplemental material (online)).

Haxby, J., Guntupalli, J., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, M., Hanke, M., Ramadge, P., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72 (2), 404–416.

Haynes, J.-D., Rees, G., 2005. Predicting the stream of consciousness from activity in human visual cortex. Curr. Biol. 15 (14), 1301–1307 (Jul).

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. 17 (4), 323–328 (Feb).

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.

Kriegeskorte, N., Goebel, R., Bandettini, P.A., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103 (10), 3863–3868 (Mar, URL http://www.pnas.org/cgi/content/full/103/10/3863).

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis — connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2 (Article 4), 1–28 (Jan, URL http://www.frontiersin.org/systemsneuroscience/paper/10.3389/neuro.06/004.2008/).

Monsell, S., 2003. Task switching. Trends Cogn. Sci. 7 (3), 134–140.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M.J., 2006. The impact of temporal compression and space selection on svm analysis of single-subject and multi-subject FMRI data. NeuroImage 33 (4), 1055–1065 (Dec).

Norman, K.A., Newman, E., Detre, G.J., Polyn, S.M., 2006. How inhibitory oscillations can train neural networks and punish competitors. Neural Comput. 18 (7), 1577–1610 (Jul).

Oruç, I., Barton, J., 2011. Multi-voxel pattern analysis of face and object exemplar discrimination in occipital cortex. J. Vis. 11 (11), 653-653.

Peelen, M., Fei-Fei, L., Kastner, S., 2009. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460 (7251), 94–97.

Peelen, M.V., Atkinson, A.P., Vuilleumier, P., 2010. Supramodal representations of perceived emotions in the human brain. J. Neurosci. 30 (30), 10127–10134 (Jul).

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45 (1), S199–S209.

Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A., 2005. Category-specific cortical activity precedes retrieval during memory search. Science 310 (5756), 1963–1966 (Dec).

Reverberi, C., Görgen, K., Haynes, J., 2011. Compositionality of rule representations in human prefrontal cortex. Cereb. Cortex 22 (6), 1237–1246.

Rosvold, H., Mirsky, A., Sarason, I., Bransome Jr., E., Beck, L., 1956. A continuous performance test of brain damage. J. Consult. Psychol. 20 (5), 343–350.

Schapiro, A., Kustner, L., Turk-Browne, N., 2012. Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr. Biol. 22, 1622–1627.

Servan-Schreiber, D., Cohen, J.D., Steingard, S., 1996. Schizophrenic deficits in the processing of context. A test of a theoretical model. Arch. Gen. Psychiatry 53 (12), 1105–1112 (Dec).

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44 (1), 83–98 (Jan).

Vickery, T., Chun, M., Lee, D., 2011. Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron 72 (1), 166–177.

Woolgar, A., Thompson, R., Bor, D., Duncan, J., 2011. Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. NeuroImage 56 (2), 744–752.