NeuroImage 98 (2014) 506–512

Contents lists available at ScienceDirect

NeuroImage

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

Comments and Controversies

Coping with confounds in multivoxel pattern analysis: What should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013

Alexandra Woolgar [a,b,*], Polina Golland [c], Stefan Bode [d]

[a] Perception in Action Research Centre & Department of Cognitive Science, Faculty of Human Sciences, Macquarie University, Sydney, Australia
[b] ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, Australia
[c] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[d] Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Accepted 22 April 2014
Available online 2 May 2014

Keywords:
Multivariate pattern analysis
fMRI
Reaction time
Modelling
Rule decoding
Frontoparietal cortex

ABSTRACT

Multivoxel pattern analysis (MVPA) is a sensitive and increasingly popular method for examining differences between neural activation patterns that cannot be detected using classical mass-univariate analysis. Recently, Todd et al. ("Confounds in multivariate pattern analysis: Theory and rule representation case study", 2013, NeuroImage 77: 157–165) highlighted a potential problem for these methods: high sensitivity to confounds at the level of individual participants due to the use of directionless summary statistics. Unlike traditional mass-univariate analyses where confounding activation differences in opposite directions tend to approximately average out at group level, group level MVPA results may be driven by any activation differences that can be discriminated in individual participants. In Todd et al.'s empirical data, factoring out differences in reaction time (RT) reduced a classifier's ability to distinguish patterns of activation pertaining to two task rules. This raises two significant questions for the field: to what extent have previous multivoxel discriminations in the literature been driven by RT differences, and by what methods should future studies take RT and other confounds into account? We build on the work of Todd et al. and compare two different approaches to remove the effect of RT in MVPA. We show that in our empirical data, in contrast to that of Todd et al., the effect of RT on rule decoding is negligible, and results were not affected by the specific details of RT modelling. We discuss the meaning of and sensitivity for confounds in traditional and multivoxel approaches to fMRI analysis. We observe that the increased sensitivity of MVPA comes at a price of reduced specificity, meaning that these methods in particular call for careful consideration of what differs between our conditions of interest. We conclude that the additional complexity of the experimental design, analysis and interpretation needed for MVPA is still not a reason to favour a less sensitive approach.

Introduction

In their 2013 paper "Confounds in multivariate pattern analysis: Theory and rule representation case study," Todd et al., make a highly relevant methodological point regarding multivoxel pattern analysis (MVPA) that had not been made explicit previously. They observe that counterbalancing variables-of-no-interest across participants does not control for the effect of those variables on group level MVPA results. For example, several studies that investigated the neural encoding of task rules using functional magnetic resonance imaging (fMRI) and MVPA have found systematically different activation patterns for different task rules in prefrontal cortex (PFC) and posterior parietal cortex (PPC) that allowed the rules to be decoded (Bode and Haynes, 2009; Haynes et al., 2007; Kastner and Ungerleider, 2000; Pessoa et al., 2003; Woolgar et al., 2011a,b). If some participants found one task rule more difficult than the other, however, differences in neural patterns in those participants could be caused by either differences in rule encoding or differences in rule difficulty. Using MVPA, these differences will also be seen at the group level, even if an equal number of participants found the opposite rule more difficult. This is because MVPA uses directionless summary measures (e.g. classification accuracy) which quantify how different the two conditions are in each participant, regardless of whether the voxel-by-voxel differences in individual participants are in the same or opposite directions. At the group level, the average accuracy is compared to chance. Thus, MVPA analyses are sensitive not only to differences in neural activation due to psychological effects of interest, but potentially also to any difference between conditions that exists

* Corresponding author.
E-mail address: Alexandra.woolgar@mq.edu.au (A. Woolgar).

http://dx.doi.org/10.1016/j.neuroimage.2014.04.059
1053-8119/

at the level of individual participants. By contrast, traditional univariate analyses (referred to by Todd et al. as "classic general linear approach, GLMA",[1] p.157) require also that the direction of any neural effect be consistent across individuals in order to have an effect at the group level. Thus, if some participants tend to show more activation for rule A, and others more activation for rule B, the differences will tend to cancel out at the group level. Univariate analyses are therefore *more specific* in the type of neural effects they reveal, at the expense of being *less sensitive* than MVPA, while MVPA is more sensitive to all neural differences at the expense of reduced specificity. In the hypothetical example given by Todd et al. in which "unknown to the experimenter, voxel activity is unresponsive to rule, but is responsive to difficulty" (page 158), the reason for observing activation differences between rule A and rule B using MVPA, was systematic differences in difficulty (in their example operationalized by response time, RT) between rules rather than differences in neural coding of rules per se. Based on these arguments, Todd et al. propose a method for removing potential RT confounds from multivoxel patterns, allowing us to increase the specificity of MVPA for effects of interest. They demonstrate that removing the effect of RT also removes the positive MVPA effects on group level in an fMRI case study. The paper raises two important questions: to what extent might previous reports of decoding be driven by RT confounds, and how can we reduce sensitivity to RT differences in future MVPA studies?

In this paper we further examine and compare methods for fMRI analysis and their sensitivity to confounds. First, we clarify the sensitivity and specificity trade-off inherent in these methods, highlighting that univariate analyses are also susceptible to confounds although they tend not to reveal them at typical sample sizes since the general sensitivity of these methods is low. Second, we present two methods for modelling the effect of RT on multivoxel patterns, one similar to that of Todd et al., the other a method used previously in our own work. Third, we test these models on our previously published empirical data, finding that the contribution of RT effects to rule encoding in our data was minimal. Finally, we discuss the meaning of confounds on a theoretical level, arguing that choice of method depends critically on the type of neural differences we wish to be sensitive for.

Sensitivity for confounds

Compared to MVPA, traditional univariate analyses are more specific in the type of neural effects they detect, only revealing changes in activation levels that are in the same direction across participants. Traditionally, this feature of univariate analyses has been employed to discount differences in activation introduced by known confounds in experimental design. For example, if we have two experimental conditions, and we believe that participants will show greater activation for whichever condition is shown first, we counterbalance the order of presentation across individuals, such that the order effect will tend to cancel out at the group level. This cancellation is, of course, imperfect. To work perfectly, the effect of presentation order on brain activation would have to be of perfectly equal magnitude and in opposite directions in the corresponding voxels in the brains of the different individuals. Sensitivity to the order confound is reduced by counterbalancing, but is not completely eliminated. At typical fMRI sample sizes, however the residual differences do not usually produce significant effects. MVPA, by contrast, is more sensitive than univariate analyses. It owes this increased sensitivity both to its use of information that is distributed across multiple voxels (and could therefore be weak in any particular voxel), and to being blind to the direction of change in activation across individuals (see Cox and Savoy, 2003; Haxby et al., 2001; Haynes and Rees, 2006; Kriegeskorte and Bandettini, 2007; Kriegeskorte et al., 2006; Norman et al., 2006). This means that MVPA is more sensitive than univariate approaches both to effects of interest and to any other factors that drive a difference between conditions at the level of individual subjects (Todd et al., 2013). As Todd et al demonstrated, at typical fMRI sample sizes, MVPA is more likely than univariate approaches to produce significant effects due to confounds that cause activation in different directions in different subjects, in addition to detecting effects of interest, while univariate analyses are less likely to produce significant effects due to these types of confounds. Increasing the power of the analysis, e.g., including more participants, will increase the ability of univariate analyses to detect true effects, but it would also increase the power for detection of small differences introduced by confounds. Therefore, the theoretical argument in favour of univariate approaches only holds when the group size is large enough to detect the effect of interest but small enough not to reveal the effect of the confound.[2] More generally, the question is which combination of sensitivity to effects, and specificity for different types of effects, we prefer. If we are interested in effects that have a consistent direction across individuals we may prefer univariate techniques for their specificity.[3] On the other hand, if we are interested in effects that do not rely on one-to-one directional voxel correspondence between participants, such as differences in fine-grained patterns of activity where the patterns themselves are not required to be the same between participants, we prefer MVPA for its sensitivity to these types of effects.

Removing RT as a confound: possible models

Todd et al. propose an intriguing solution to the sensitivity vs. specificity trade-off. They suggest that we can increase the specificity of MVPA by explicitly modelling known confounds in our data. They focus in particular on the effect of RT, which, in Todd et al.'s data, appears to lead to false positive decoding results. Ideally, tasks would be designed so as not to have confounding behavioural differences in the first place, for example by titrating stimulus parameters for each condition for each participant separately, to ensure equal RT between conditions before scanning. Where this has not been achieved, however, Todd et al. suggest that we can at least reduce the sensitivity to RT differences in our analysis. This is a particular instance of a common practice in fMRI analysis, whereby we include covariates of no interest (e.g., linear trends, session means, motion parameters) in our first level model of the data in order to improve the detection of effects of interest. In the context of univariate analysis, this has been investigated in detail in the past (e.g. Buchel et al., 1998; Henson, 2007; Grinband et al., 2008; Yarkoni et al., 2009).

In principle, regressing any effect-of-no-interest will leave a more pure measure of the variance attributable to the effects of interest. However, in the case of RT in particular, it is unclear what the best model is. The answer depends on how we believe RT will be reflected in BOLD activation. RT may reasonably be thought to index, for example, task difficulty or effort. In line with the model proposed by Todd et al., this might be reflected in increased BOLD response (per unit time) for longer trials. Alternatively, we might take RT simply as a measure of time on task, in

[1] We prefer the terms mass-univariate or univariate analysis, since MVPA approaches also commonly use the general linear model (GLM) to estimate the BOLD response to different conditions (and perform classification on the resulting estimates of the regression coefficients).

[2] Note that this of course constitutes an extremely simplified example, as a full review of all factors that influence the outcome of mass-univariate and multivariate analyses is beyond the scope of this paper. The relationship between sample size, the detectable true effects and effects of potential confounds further depends on the complex interplay between multiple factors that we do not consider here, including the signal-to-noise ratio (SNR), experimental design (e.g. event-related designs, block design), the true effect size, the analysis performed, the model (e.g. classical GLM, FIR models), as well as the research question of interest (a more comprehensive overview of mass-univariate fMRI analysis can be found, e.g. in Brett et al., 2002; Faro and Mohamed, 2006; Friston et al., 1995, 2006; Henson, 2007). Simulation work, as conducted by Todd et al. (2013), is a first step towards specifying these relationships, but future work should incorporate multiple factors to yield a more comprehensive comparison of univariate and multivariate scenarios.

[3] Because of the imperfect voxel-to-voxel correspondence between participants, mass-univariate analyses also tend to use spatially smoothed data, meaning that they are also specific for activation changes occurring on a fairly coarse spatial scale.

which case we would expect that activation is not higher per unit time, but simply that the activity is extended in time. This would result in an increased BOLD response per trial, but not per unit time. The chosen model strongly impacts on how RT is treated as a component to be removed, and, in turn, how the data is changed (rightly or wrongly) by its removal. To illustrate this point, we create three hypothetical voxel response profiles and three models designed to capture the response of these voxels, using the general linear model (GLM) framework. In the subsequent section we test these models with empirical data.

In response profile 1 ("No RT control", Fig. 1, second panel), voxel A responds only to rule A and is not affected by reaction time. This might be the case, for example, for a voxel which reflects activity of a fixed part of the processing pathway that does not vary from trial to trial (with variation in RT being driven by some other part of the processing pathway, perhaps reflected in the activity of a different voxel). The corresponding "No RT control" model (Fig. 1, fifth panel) simply models each trial as an event of zero duration. This model is similar to the first control analysis, "MVPA (no control)", by Todd et al.

In response profile 2 ("Additive RT" Fig. 1, third panel), voxel A again responds selectively to rule A (shows more activity for rule A than for rule B) but it also responds on every trial according to how effortful the task is, indexed by RT. This voxel is active whenever the participant performs either rule, with this component of its activity proportional to RT, and it shows additional activation for rule A which is not proportional to RT. In this situation, the "No RT control" model would misestimate the response of the voxel to rule A. Specifically, it would overestimate the voxel's response for longer RTs and underestimate its response for shorter RTs. A more appropriate "Additive RT" model for this scenario is shown in Fig. 1, sixth panel. Here, one regressor models the effect of RT, while a second regressor models the activity due to performing rule A. This model is similar to the "MVPA (after regression)" model used by Todd et al.

We also examine a third possibility. In response profile 3 ("Modulating RT", Fig. 1, fourth panel), voxel A responds only to rule A, but now it is active for as long as the person is performing rule A. In this scenario, RT modulates the response of the voxel to A. Note that there is no need to invoke the cognitive concept of difficulty as an explanation for this scenario, duration of activation simply depends on time on task. In this scenario, the "No RT control" model would overestimate the response of the voxel to rule A on longer relative to shorter trials. However, the "Additive RT" model is also inappropriate for this case, because the additive RT model predicts activity in this voxel even when the participant performs rule B. Since voxel A does not respond to rule B (and consequently its activity is not modulated by RT on B trials), the RT regressor in the "Additive RT" model will not fit well and the effect of RT will be underestimated. This means that the effect of RT on voxel A will only be partially removed by the RT regressor. Instead, in order to estimate the per-unit-time response correctly, trials of type A should be modelled as epochs of length RT (a varying-duration epoch model, Henson, 2007) as shown in the "Modulating RT" model, Fig. 1, bottom panel. This is the approach used in our previous MVPA work (Woolgar et al., 2011a,b) and elsewhere in univariate analysis (e.g. Christoff et al., 2001; Crittenden and Duncan, 2014; Grinband et al., 2008; Strand et al., 2008; Yarkoni et al., 2009). For example, Grinband et al. (2008) compared this model to other more commonly used models using simulated and empirical data in a univariate framework. They showed that for neural signals whose duration varies with a known psychophysical parameter, such as RT, the varying-duration epoch model provided greater statistical power and reliability for detecting neural activity. It is also in line with electrophysiological studies that indicate sustained neural activity up to the time of a behavioural response (e.g. Philiastides et al., 2006; Shadlen and Newsome, 2001).

Clearly, the choice of model to capture rule-related activity depends on which of the above hypothetical scenarios we believe applies to our voxel of interest (and we may have different predictions for different brain regions, e.g. Yarkoni et al., 2009).[4] In the next section, we compare the empirical fit of the three models above to one of our previous datasets.

Removing RT as a confound: empirical data

As described in the original paper (Woolgar et al., 2011b), participants viewed a blue square presented in one of four possible horizontal locations and applied one of two stimulus–response mapping rules in order to generate a button press response. On each trial, the current rule to use was cued by the background colour of the screen (two colours for each rule). The two rules involved an equal contribution from the 4 stimulus positions and 4 button press responses and were incompatible with each other, such that each rule required a different button press for each stimulus position. RT varied from trial to trial, between conditions, and between participants.

In the original paper (Woolgar et al., 2011b), RT was accounted for at the first level in MVPA analyses using the "Modulating RT" model described above. We prefer this model on theoretical grounds because it does not require cognitive concepts of difficulty or effort to account for activation, but simply assumes that voxels are active for longer on trials where the time on task (RT) is longer. Nonetheless, to address the empirical question of which model best fits our data, we carried out a re-analysis of our original data using the "No RT", "Additive RT" and "Modulating RT", models described above. We predicted that if RT differences between the two rules spuriously drove rule classification, as implied in the paper by Todd et al., decoding accuracy would be reduced in the "Additive RT" and/or "Modulating RT" models relative to the "No RT" control. If classification was comparable across models, however, this would indicate a minimal contribution of RT to rule classification. This analysis was carried out on an ROI basis, for 7 frontoparietal ROIs predicted to show rule decoding, and on a whole-brain basis using a roaming searchlight, as described in our original paper.

The results of the re-analyses are shown in Figs. 2 and 3. The three models yielded highly similar results. In the ROI analysis (Fig. 2), there was no significant difference between any of the three models in any region, suggesting that the contribution of RT to rule discrimination was minimal in this dataset. In particular, the difference between the "No RT" (white bars) and "Additive RT" (black bars) was not significant, suggesting that including an RT regressor in the GLM had a minimal effect. Moreover, the brain map of rule decoding revealed with a roaming searchlight (Fig. 3) was nearly identical in all three models, suggesting that there were no brain regions in which any one model lead to higher rule decoding results. These findings are a strong indicator that RT effects were not a driving factor for decoding in this dataset.

RT confounds in previous rule decoding studies

Todd et al. show that after removing RT as a potential confound in their data, their MVPA results are non-significant on a group level. They therefore conclude that RT differences, and not rule-specific

[4] A combination model (a varying-duration epoch model plus an additional single RT regressor) is not possible as the RT regressor would be the sum of the regressors A and B, such that the regression matrix would become singular. Intuitively, any contribution of the additional regressor cannot be distinguished from a sum of the contributions of the two original regressors. Nonetheless, other models are conceivable that we did not account for here. For example, another possible scenario is one in which activation is of fixed duration but its onset delayed on longer RT trials (e.g. Yarkoni et al., 2009). Alternatively, RT could be inversely related to activation per unit time. This would be true if shorter RT reflected the recruitment of more computational processes per unit time, while the overall computational resources devoted to the task were constant. In this scenario, a strong, short burst of neural activation would accompany short RT trials, whereas long RT trials would be characterised by weak but prolonged neural activation profile. The "No RT control" model may capture the net effect of this scenario reasonably well as it assumes constant total neural activity on each trial, but it would ignore its varying temporal spread.

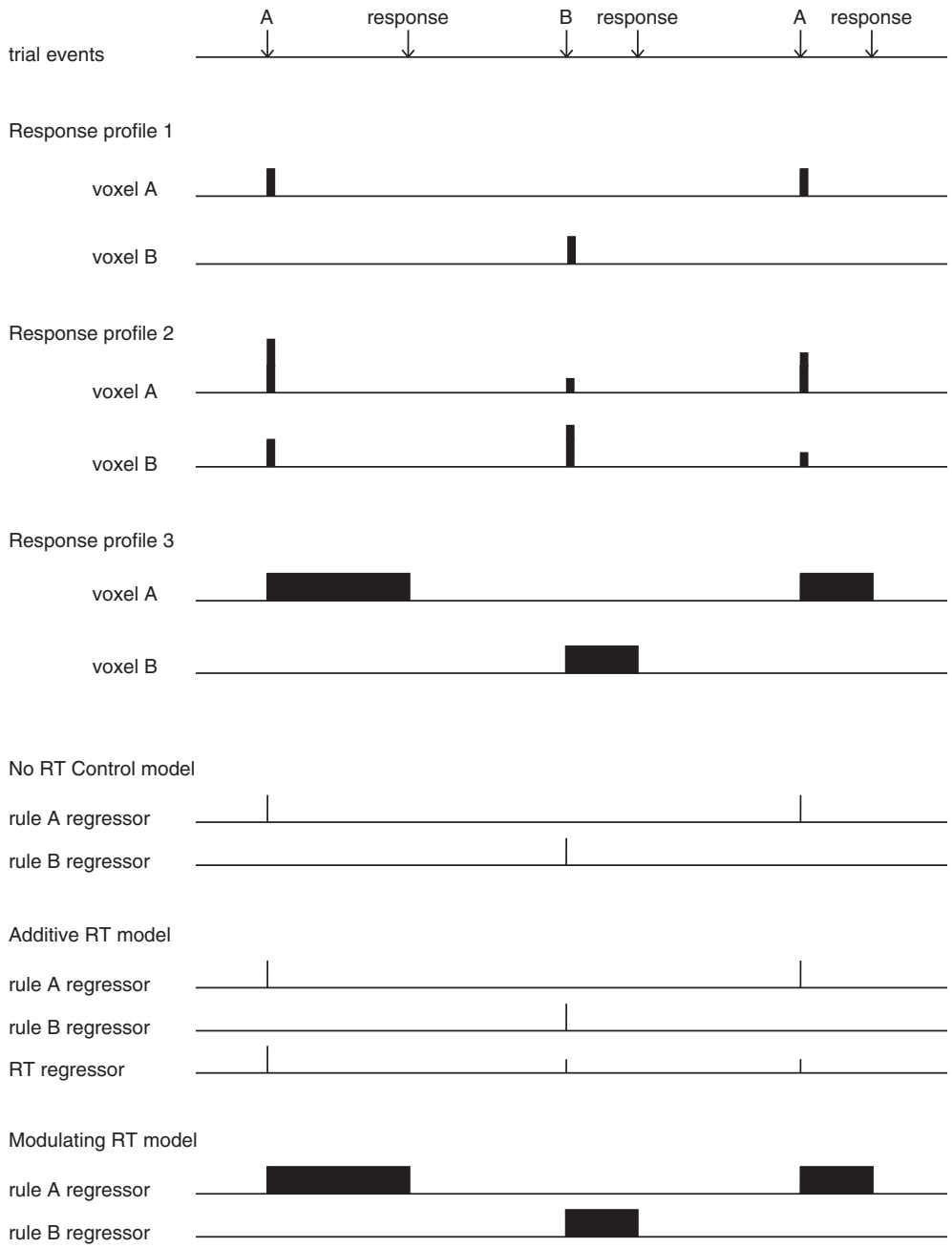

Fig. 1. Hypothetical voxel response profiles for rules A and B and first level models designed to capture rule related voxel activity. Top panel: timeline of stimulus events. Second panel: response of voxels A and B under response profile 1. Voxels respond to their preferred event (A or B) for a fixed time, independent of reaction time. Third panel: response of voxels A and B under response profile 2. Voxel activity is the sum of an RT-independent component selective for the preferred rule, and an RT-dependent component unselective for rule (e.g., difficulty). Fourth panel: response of voxels A and B under response profile 3. Voxels are activated by their preferred rule and remain active for the duration of the trial until a response is given. Fifth, sixth and bottom panels: "No RT Control", "Additive RT" and "Modulating RT" models proposed to capture activity under response profiles 1, 2, and 3 respectively.

coding, drove the observed effects in their data, raising the question of how widely RT differences might have driven rule decoding reported previously in the literature. Here, we argue that there is no reason to think that our previous demonstrations of rule decoding were driven by RT or difficulty effects. First, in our previous work cited by Todd et al. (Woolgar et al., 2011b), RT was accounted for at the first level in MVPA analyses using the "modulating RT" model above. This was achieved by modelling trials in each condition as boxcar regressors lasting for the length of the RT on each trial before convolution with the HRF. This model aims to avoid overestimating the BOLD response on longer relative to shorter trials (Henson, 2007), as would be needed for both univariate and multivariate analyses. Moreover, our re-analysis of the data here suggests that the effect of RT on rule decoding in this dataset was minimal. We also used the "Modulating RT" model to account for possible RT effects in another rule decoding study (Woolgar et al., 2011a) which replicated our original rule decoding results.

An alternative approach to reducing the effect of RT was demonstrated by Bode and Haynes (2009). In this study, a rule cue was presented first (1.4 s) followed by a 2.8 s delay during which only a fixation cross was shown. Then the stimulus was presented for 4.2 s, during which time participants had to withhold their rule-guided response until a response screen (fixation cross) was shown. Using an FIR model and performing rule decoding at each TR, rules could be decoded from patterns of brain activity throughout all task phases, including the first preparation phase in which the rule could be prepared but not applied, and the subsequent stimulus phase in which no action

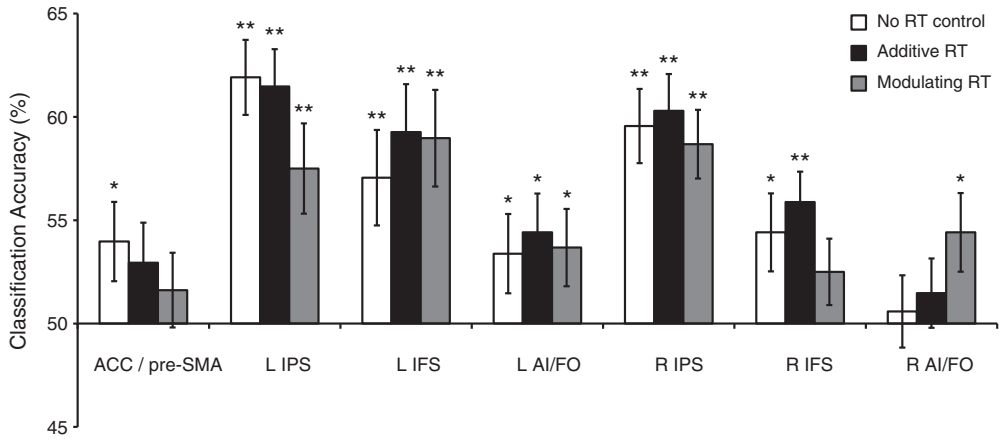

Fig. 2. Multivoxel classification of rule in the dataset of Woolgar et al. (2011b). Data were preprocessed (spatial realignment and slice timing correction) and high-pass filtered (128 s) as described in the original paper. Classification accuracy refers to the ability of a linear support vector machine (LinearCSVMC) to correctly classify regression coefficient estimates (betas) pertaining to each of the two stimulus–response mapping rules. Classification was carried out using a leave-one-block-out 10-fold splitter, as described in Woolgar et al. (2011b), implemented by wrapping the LIBSVM library (Chang and Lin, 2011). First level models consisted of 120 regressors-of-interest, modelling four stimuli positions, four button press responses and four background colours in each of ten blocks, with block and session means included as covariates-of-no-interest. In this study the four background colours dictated which of two rules to use, thus 40 betas (4 colours * 10 blocks) were used for classification, and the classifier was trained to distinguish the 20 betas pertaining to each of the two rules. Displayed are the classification results using three first level models (see main text for further detail). White bars: "No RT control" model, in which trials were modelled as events of zero duration. Black bars: "Additive RT" model in which an additional regressor was added to the GLM. This is similar to the "MVPA (after regression)" analysis of Todd et al. (2013), except that beta estimates (rather than selected scans) were used for classification, and regression with RT was only carried out once, at the level of GLM estimation, since mathematically a linear regression cannot remove more variance if carried out a second time. Grey bars: "Modulating RT" model, in which trials in each condition were modelled as events lasting from stimulus onset until response (i.e. for the duration of RT) as in Woolgar et al. (2011b). ROIs were taken from a prior review of imaging literature, as described in Woolgar et al. (2011b), selected for their response to a wide range of task demands. ACC/pre-SMA: Anterior cingulate cortex/pre-supplementary motor area; IPS: intraparietal sulcus; IFS: inferior frontal sulcus; AI/FO: anterior insular/frontal operculum; L: left; R: right. Error bars show standard error. Significance markings indicate significance of rule coding in each ROI compared to chance (50%), $^{*}p < 0.05$, $^{**}p < 0.01$, no correction for multiple comparisons. Decoding did not differ significantly between models in any ROI.

was required. Although there could still have been differences in internal processing time for different task components, the total time on task was kept constant such that no RT differences between rules emerged for any participant (re-analyses of individual RT data: all $p > .05$). In addition, there was no difference in behavioural accuracy for the two rules for any individual participants (all $p > .05$), suggesting that the two rule conditions were well matched for difficulty. These considerations make it highly unlikely that a difference in RT drove the rule classification seen in this study.

Taken together, differences in RT between conditions appear to have been a driving factor in Todd et al.'s case study, perhaps reflecting the relatively large behavioural difference between conditions in their sample. However, rule decoding results from other studies (Bode and Haynes, 2009; Woolgar et al., 2011a,b) are unlikely to only reflect individual differences in RT. Perhaps, the lack of residual rule decoding seen in the data of Todd et al. after RT regression might reflect the relatively small number of data samples included after the authors selected a small proportion of scans (45 scans per condition per participant) for inclusion in the analysis.

The meaning of confounds

One question that warrants further discussion is under what circumstances the higher sensitivity/lower specificity of MVPA is most problematic for our field. Certainly, it would be unhelpful if reportedly "new" effects of interest are in fact driven by some other factor(s) for which effects are already well documented. For example, classifying neural differences that are simply due to different levels of difficulty would be uninteresting in brain regions such as frontoparietal cortices that are already known to show increased activity for increased cognitive demand (Duncan and Owen, 2000). To some extent, this question has been addressed in previous studies through comparison of univariate and multivariate effects (e.g. Coutanche, 2013). Univariate effects are thought to index broad differences between the conditions (e.g., increased difficulty/effort/attention) while multivoxel phenomena are thought to index within-condition information content (e.g. Mur et al., 2009). Indeed, many MVPA studies either report univariate and multivariate effects in parallel or explicitly avoid detection of unspecific effects reflected in the average signal, for example by subtracting the mean response across-voxels. On the other hand, it might be overly conservative to require that all activation differences between conditions are zero on average, producing no univariate difference at all, in order to be interesting. It is possible that within a neural population showing a distributed representation of two rules, there may still be more neurons dedicated to one rule as opposed to another, leading to slight differences in population means. These small differences in neural activity between conditions within single voxels may form the biases that MVPA is sensitive for (Kamitani and Tong, 2006), and a perfect overall balance in activation across all relevant voxels seems unlikely. The superior sensitivity of MVPA allows us to discover even these small activation differences in order to characterise the types of distinctions relevant for neural processing. Requiring activation patterns to be perfectly weighted may miss important distinctions that neural circuits are making.

In other cases, the higher sensitivity/lower specificity of MVPA may cause us to be sensitive to distinctions that are intrinsically relevant for the aspect of processing under investigation. For example, it might be true for rules, as well as for all other kinds of representations, that different conditions are intrinsically associated with differences in behaviour. It seems unlikely that any two conditions exist that are exactly the same on a behavioural level and only differ in the neural. Of course, such behavioural differences may not be reflected in RT and may be difficult to measure. For example simply having – and always developing – *preferences* for one alternative in an arbitrary laboratory task would be sufficient to create such an intrinsic and unavoidable "confound" between conditions. Consider, for example, a study by Bode et al. (2013) in which participants had to guess which object category (a piano or a chair) they saw, even though (unknown to them), no object had been presented. Simply liking one object category more than the other, or developing a preference for one object category, may lead to small but systematic differences in choices and, in turn, in the strength of activation patterns. This effect is similar to what Todd et al. propose for RT: decoding is driven by preferences as well as stimulus distinctions. In this example, however, preferences are not a confound, but are part of

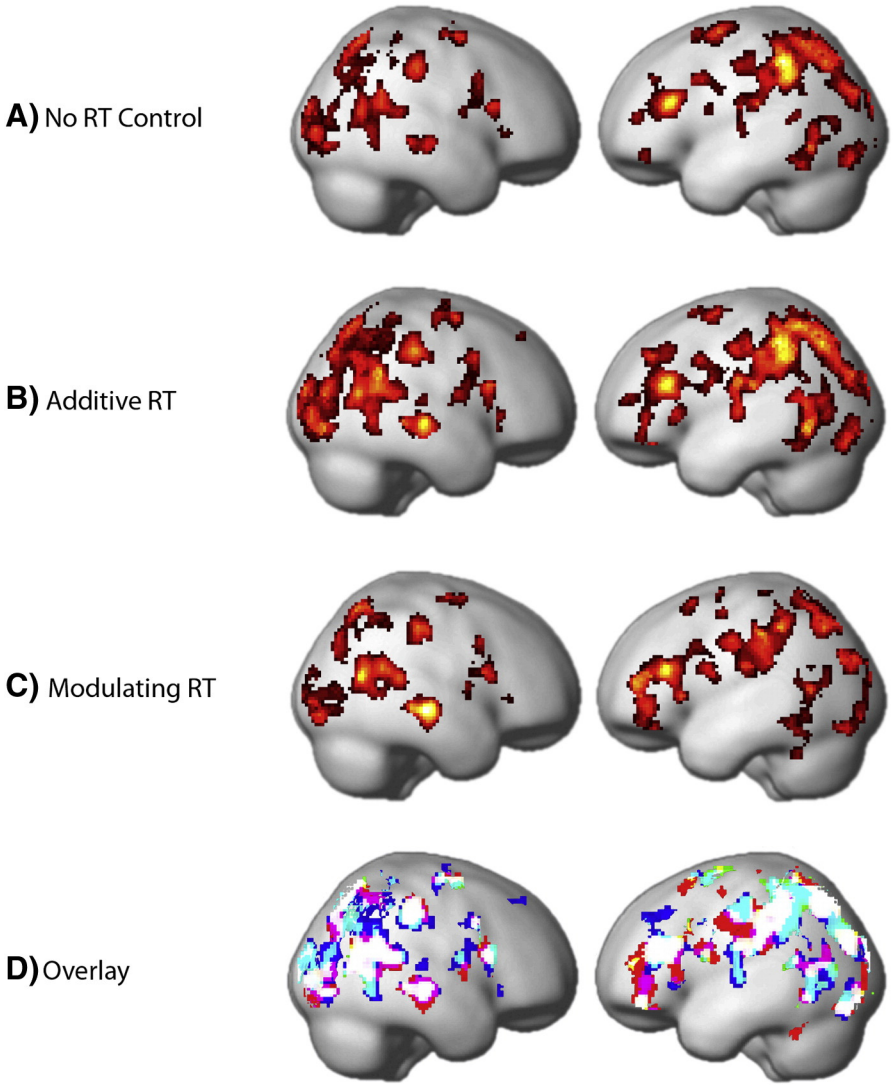

Fig. 3. Multivoxel classification of rule in the dataset of Woolgar et al. (2011b): searchlight analyses. Panels depict regions of above chance classification using the "No RT control" model (panel A), "Additive RT" model (panel B), and the "Modulating RT" model (panel C) described in the text. Panel D re-depicts the results from panels A to C overlaid on a single rendering with the following colour scheme: green: supra-threshold voxels unique to the "No RT control" model; blue: supra-threshold voxels unique to the "Additive RT" model; red: supra-threshold voxels unique to the "Modulating RT" model; cyan: overlapping voxels in the "No RT control" and "Additive RT" models only; yellow: overlapping voxels in the "No RT control" and the "Modulating RT" models only; purple: overlapping voxels in the "Additive RT" and "Modulating RT" models only; white: overlapping voxels in all 3 analyses. Analyses were performed as described in the legend for Fig. 2, using a 5 mm roaming searchlight centred in turn on each voxel in the brain in order to derive individual whole brain classification maps which were subsequently normalised and smoothed (8 mm FWHM) before comparing to chance (50%) in a second level random effects analysis using SPM5 (for more detail see Woolgar et al., 2011b). T-maps are thresholded at $t = 2.84$, equivalent to $p < 0.05$ with false discovery rate (FDR) correction in the "No RT control" analysis.

the basic mechanism underlying decision formation. In consequence, removing participant preference would eliminate a defining feature of the variable of interest itself. Similarly in the case of rule decoding, representing two rules might involve choosing one rule as the "default" or preferred rule and the other as the "alternative" rule, and such a process might constitute an intrinsic part of rule processing. Thus, Todd et al.'s results remind us of the need for more sophisticated interpretations which consider carefully the multiple factors contributing to our variables of interest. Hopefully, this will facilitate a more thorough understanding of cognitive concepts such as rules and decisions.

As a consequence, studies using MVPA should follow some practical guidelines. First, one should carefully examine whether differences in behaviour (such as RT) between experimental conditions are likely to occur at the level of individual participants, considering the specific experimental task. Where possible pre-experimental training should be used and/or the stimulus set should be titrated to limit the behavioural differences. If training is not feasible or not a desired option (e.g. in learning paradigms), one should consider whether behavioural differences are meaningful and should be explicitly incorporated into the interpretation of the results. To allow readers to assess the importance of potential confounds authors should report behavioural differences between conditions at the level of individual participants (e.g. report whether any participants show a significant behavioural difference between conditions). Moreover, if potential behavioural differences are regarded as confounds, they should be accounted for in the GLM. In doing so, it has to be decided which model is most appropriate to account for the confound. As we have discussed, this question is not trivial and depends on how the particular behavioural effect is thought to impact brain activity and which may vary between brain regions. There may be some value in running multiple models, as demonstrated here, in order to assess the impact of different models on the results. Clearly, authors should always describe their rationale for the chosen model and report the results of all the different approaches if more than one is used (e.g. Carp, 2012).

Conclusions

Compared to traditional univariate analyses, MVPA is more sensitive to all neural differences at the level of individual participants. As such, additional controls are required to account for known confounds in our data. In the case of RT, there are multiple possible modelling variants that may be used to separate the effect of RT from the factor of interest. The appropriate model strongly depends on our assumptions for how the confound affects neural activation in the brain region of interest, and is not a trivial problem. Reassuringly, in our empirical data, the choice of model made very little difference to multivoxel pattern classification. In contrast to the case study presented by Todd et al., we find no evidence that our previous reports of rule coding in frontoparietal cortex were driven by RT effects. Nonetheless, their paper provides a timely reminder that counterbalancing or equating factors over a group of participants does not suffice for eliminating their contribution to MVPA. With more sophisticated analysis techniques becoming increasingly available, we must think carefully about how we define cognitive concepts such as rules and what features contribute to the way the brain distinguishes between them. More generally, it reminds us that the details of our experimental design, imaging methods and analysis must depend on the precise question we aim to answer. Since all techniques have both strengths and weaknesses, fMRI-based findings must be evaluated in the light of converging evidence from complimentary sources, including behavioural measures, single-unit recordings and alternative neuroimaging techniques.

Acknowledgments

AW was supported by an Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA, DE120100898) and by ARC Discovery Projects grant DP12102835. PG was supported by NSF IIS/CRCNS 0904625, NIH NIBIB NAMIC U54-EB005149, and NIH NIBIB NAC P41-EB015902. SB was supported by a University of Melbourne Early Career Researcher Grant, a Melbourne Neuroscience Institute (MNI) project grant, and an ARC Discovery Early Career Researcher Award (DECRA, DE140100350). We thank John Duncan, Mark Williams and Kiley Seymour for helpful comments on this article.

References

Bode, S., Haynes, J.D., 2009. Decoding sequential stages of task preparation in the human brain. NeuroImage 45 (2), 606–613.
Bode, S., Bogler, C., Haynes, J.D., 2013. Similar neural mechanisms for perceptual guesses and free decisions. NeuroImage 65, 456–465.
Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. Nat. Rev. Neurosci. 3, 243–249.
Buchel, C., Holmes, A.P., Rees, G., Friston, K.J., 1998. Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. NeuroImage 8, 140–148.
Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front. Neurosci. 6, 149.
Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 27.
Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J.K., Holyoak, K.J., Gabrieli, J.D.E., 2001. Rostrolateral prefrontal cortex involvement in relational integration during reasoning. NeuroImage 14 (5), 1136–1149.
Coutanche, M.N., 2013. Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? Cogn. Affect. Behav. Neurosci. 13 (3), 1–7.
Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging fMRI "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19 (2 pt 1), 261–270.
Crittenden, B.M., Duncan, J., 2014. Task difficulty manipulation reveals multiple demand activity but no frontal lobe hierarchy. Cereb. Cortex 24 (2), 532–540.
Duncan, J., Owen, A.M., 2000. Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends Neurosci. 23 (10), 475–483.
Faro, S.H., Mohamed, F.B., 2006. Functional MRI. Basic principles and clinical applicationsSpringer, New York, NY, USA.
Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C., Frackowiak, R.S., Turner, R., 1995. Analysis of fMRI time-series revisited. NeuroImage 2 (1), 45–53.
Friston, K.J., Ashburner, J., Kiebel, S., Nichols, T.E., Penny, W.D., 2006. Statistical Parametric Mapping: The analysis of functional brain images. Elsevier, London, UK.
Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., Hirsch, J., 2008. Detection of time-varying signals in event-related fMRI designs. NeuroImage 43 (3), 509–520.
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. Science 293 (5539), 2425–2430.
Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7 (7), 523–534.
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.
Henson, R.N., 2007. Efficient experimental design for fMRI. In: Frackowiak, R.S., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., Penny, W.D. (Eds.), Statistical parametric mapping. The analysis of functional brain images. Academic Press, London, pp. 193–210.
Kamitani, Y., Tong, F., 2006. Decoding seen and attended motion directions from activity in the human visual cortex. Curr. Biol. 16 (11), 1096–1102.
Kastner, S., Ungerleider, L.G., 2000. Mechanisms of visual attention in the human cortex. Annu. Rev. Neurosci. 23, 315–341.
Kriegeskorte, N., Bandettini, P., 2007. Analyzing for information, not activation, to exploit high-resolution fMRI. NeuroImage 38 (4), 649–662.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103 (10), 3863–3868.
Mur, M., Bandettini, P.A., Kriegeskorte, N., 2009. Revealing representational content with pattern-information fMRI — an introductory guide. Soc. Cogn. Affect. Neurosci. 4 (1), 101–109.
Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10 (9), 424–430.
Pessoa, L., Kastner, S., Ungerleider, L.G., 2003. Neuroimaging studies of attention: from modulation of sensory processing to top-down control. J. Neurosci. 23 (10), 3990–3998.
Philiastides, M.G., Ratcliff, R., Sajda, P., 2006. Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. J. Neurosci. 26, 8965–8975.
Shadlen, M.N., Newsome, W.T., 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J. Neurophysiol. 86, 1916–1936.
Strand, F., Forssberg, H., Klingberg, T., Norrelgen, F., 2008. Phonological working memory with auditory presentation of pseudo-words — an event related fMRI Study. Brain Res. 1212, 48–54.
Todd, M.T., Nystrom, L.E., Cohen, J.D., 2013. Confounds in multivariate pattern analysis: theory and rule representation case study. NeuroImage 77, 157–165.
Woolgar, A., Hampshire, A., Thompson, R., Duncan, J., 2011a. Adaptive coding of task relevant information in frontoparietal cortex. J. Neurosci. 31 (41), 14592–14599.
Woolgar, A., Thompson, R., Bor, D., Duncan, J., 2011b. Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. NeuroImage 56 (2), 744–752.
Yarkoni, T., Barch, D.M., Gray, J.R., Conturo, T.E., Braver, T.S., 2009. BOLD correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI analysis. PLoS ONE 4 (1), e4257.