Rewards are categories.

Erik J. Peterson
Dept. of Psychology
Colorado State University
Fort Collins, CO

Chapter 3 – fMRI analyses

In acquisition

fMRI data was acquired at the Intermountain Neuroimaging Consortium (INC) facility located at the University of Colorado at Boulder. All 18 right-handed participants were be pre-screened for the typical fMRI exclusion factors (e.g. metal implants, mental disorders, etc). Two sets of high resolution anatomical data were acquired. *TODO rest of scan details, cover both localizers and ana*

Following DICOM to nifiti-1 (4D) conversion using dicom2nii (`http://www.mccauslandcenter.sc.edu/mricro/mricron/dcm2nii.html`), each dataset was then subjected to the following preprocessing pipeline, carried out in SPM8 using that program's batch mode (`http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`). For complete code see, `https://github.com/andsoandso/fmri/tree/master/catreward/spm_m`). Anatomical data (MPRAGE) was first segmented in white and grey matter regions

(?, ?). Based on these segments, parameters necessary for normalization into T1 MNI-352 (1 mm) space were calculated. Anatomical data was then resampled from 1.27 to 1.00 mm^3 using fourth degree β-splines and finally normalized into MNI space. Normalization had two steps. The first is a Bayesian 12-parameter affine transformation (?, ?). The second is a set of nonlinear deformations, using a 1127 parameter discrete cosine transform (?, ?).

Movement regressors for all functional volumes were calculated. No participant moved more than 1.5 mm. Functional data was then slice-time corrected, using slice 13 (the middle slice from the descending acquisition) as the reference. Data was then coregistered with the pre-processed (native-space) anatomical data (?, ?), resampled to 3 mm^3 again using fourth degree β-splines, and normalized into MNI space using the anatomically-derived parameters above. Finally, the functional data was spatially smoothed using a 6 mm FWHM Gaussian, though a copy of the un-smoothed data was retained for the ROI analyses. Each voxel's time course was also low-pass filtered (using finite impulse response model, with a cutoff at 0.008 Hz, (?, ?)) prior to regression analysis. For all whole-brain analyses, the movement regressors were entered into every model as covariates, accounting for any head movement.

All statistical parametric maps presented below were derived from a Random Effects (RFX, or "second-level" in SPM8 jargon) analysis, multiple comparison corrected assuming a Gaussian Random Field using the Family Wise Error Rate (FWE) at the $p < 0.05$ (?, ?), and a minimum cluster size of 4 voxels), that is except for the raw, unthresholded, maps of t-values discussed in the next section.

In fMRI (and in time-series analysis in general) there is an intrinsic trade-off between simply detecting a signal in the presence of noise and then estimating the timecourse (i.e. shape) of that signal (Dale, 1999; Birn, Cox, & Bandettini, 2002; Liu, 2004). One way to optimize over both these objectives is to manipulate the trial order, inside a rapid event-related design (Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000). One state of the art method for setting the trial order is a genetic algorithm which uses two (wieghted) loss functions, one for signal detection and one for time-course estimation (Wager & Nichols, 2003). Kao, Mandal, Lazar, & Stufken, 2009, improved on this design, adding in psychological considerations, and greatly improving executation speed and documentation. As a result, Kao's method was used to optimize trial orders for part 1 and 2, along with the reward category (i.e. grating only) localizer scan.

Maps of blobs

Whole brain activity for the stimulus-response learning portion of the behavioral experiment (i.e. part 2) was examined first by comparing all trials to the baseline (rest) condition. This data is presented in two ways. First is a transparent overlay of the raw t-values. Second is the more typical statistically thresholded contrast image. The contrast map showed significant ($t(15) = 6.59$, $p < 0.05$) bilateral activity in the cerebellum, insula and anterior cingulate (Figure 1). Examination of the raw t-values confirms that observed significant effects were robust and widespread in their respective regions, but also allows for the analysis of overall and subthreshold patterns of activity. These raw data suggest near threshold levels of activity in the

head of the caudate, ventrol-medial, dorsal lateral frontal cortices as well as (weaker) activity in the occipital lobe (Figure **??**). And indeed in a two-way ANOVA looking for at that interaction between gains and losses significance clusters were observed in head and body of caudate, insula, posterior and anterior cingulate with the posterior activation extending into the precuneus, dorsal lateral (i.e middle frontal) and ventral medial cortex (Figure 3; $F(1, 270) = 30.76$, $p < 0.05$). When trials with gains and losses were examined separately, both showed activity in the same areas as when they were combined (not shown). Losses showed both increases and decreased of the BOLD signal compared to the rest condition, whereas gains exhibited only increases (not shown).

Regions and models

The right chunks. Following whole-brain analysis, regions of interest were selected using two methods, that were later compared. The first employed only regions from the Harvard-Oxford probabilistic anatomical atlas (?, ?), using the 50% cut-off. The second combined anatomical regions with functional clusters isolated using both sets of data collected during the second half of part 1 and from the reward-category localizer outlined above. Analyses showed that the clustered regions and entire anatomical regions displayed very similar model-fits. So to limit the complexity of the later analyses, and to possible increase power, functional analyses were dropped. Only anatomical based analyses are presented here. Anatomical regions of interest were selected *a priori* based on previous studies of reinforcement and category learning. Left and right subcortical regions of interest were the dorsal caudate,

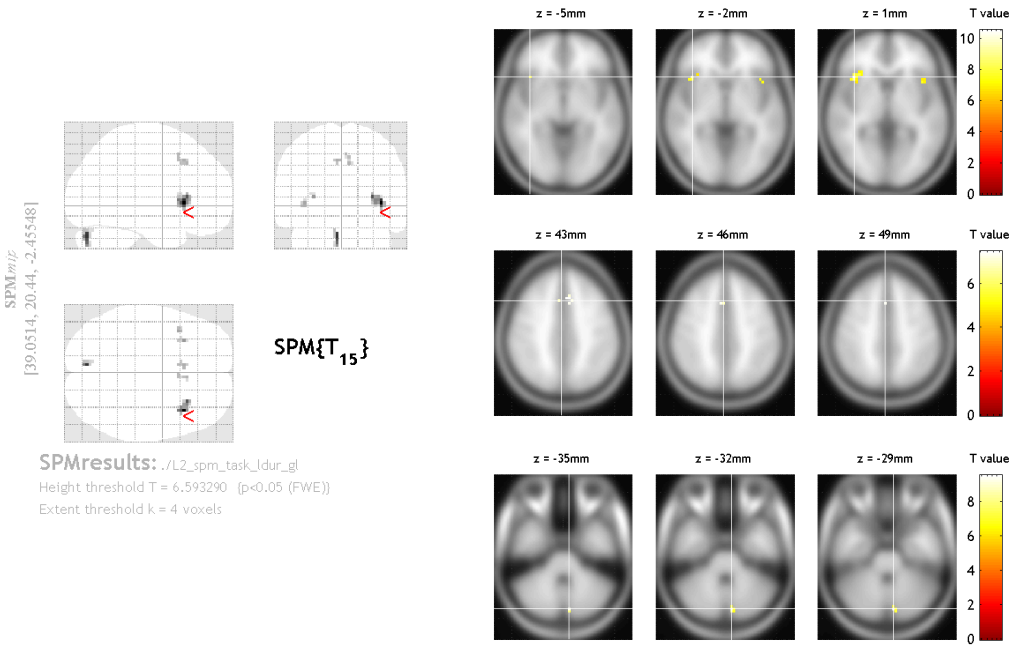

Figure 1. Statistical parametric map for all trials in the stimulus-response learning task (i.e. part 2), compared to the rest period. *Left* is a glass brain, showing all significant clusters mapped down to 3 two dimensional representations. *Right* is a set of axial slices highlighting strong areas of activity overlaid onto the T1 MNI-352 template. Z is the height of the axial slice in MNI space.

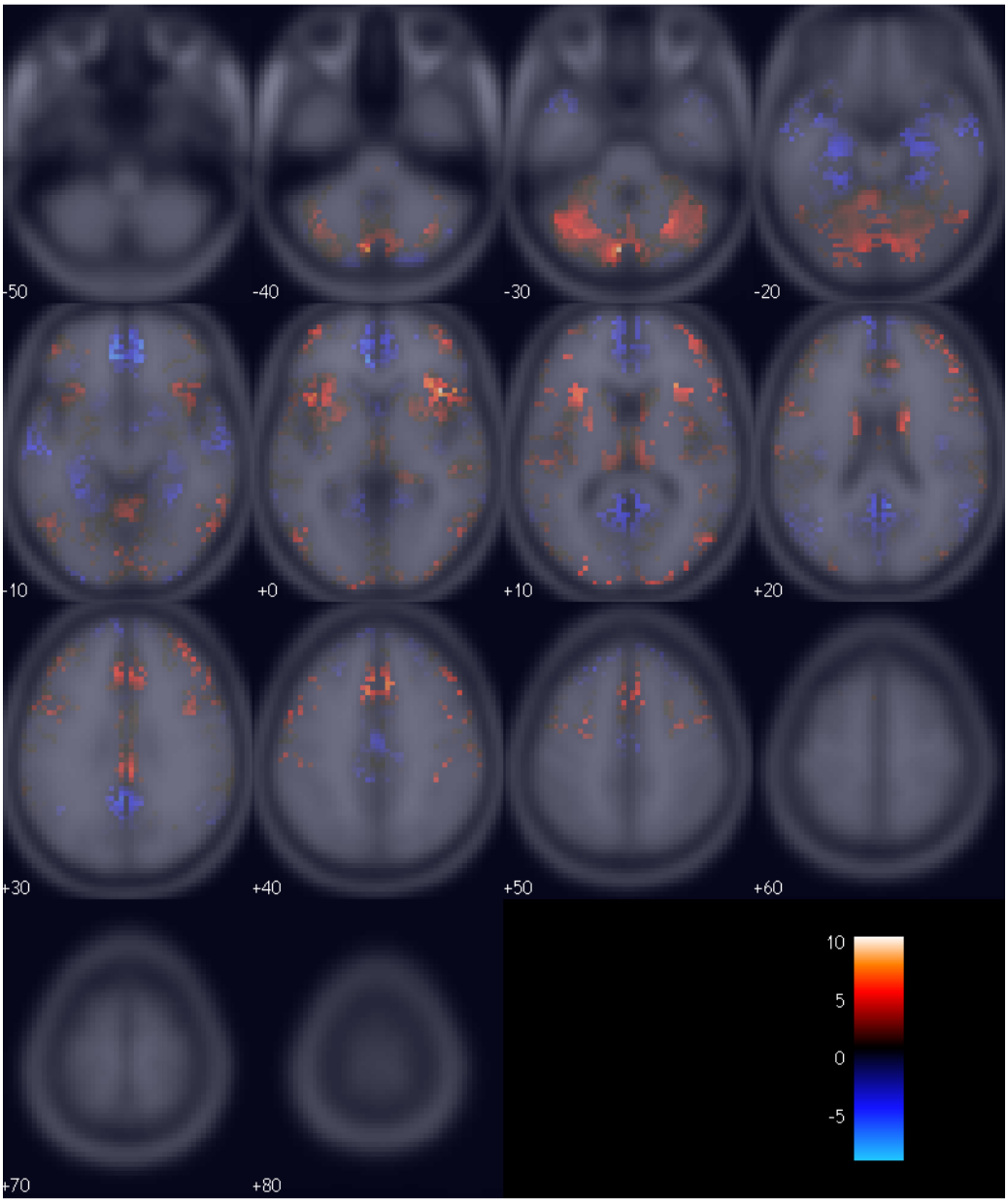

Figure 2. (Raw, that is unthresholded, t-values for all trials in the stimulus-response learning task (i.e. part 2), compared to the rest period, overlaid onto the T1 MNI-352 template. Each number is the height of the axial slice in MNI space.

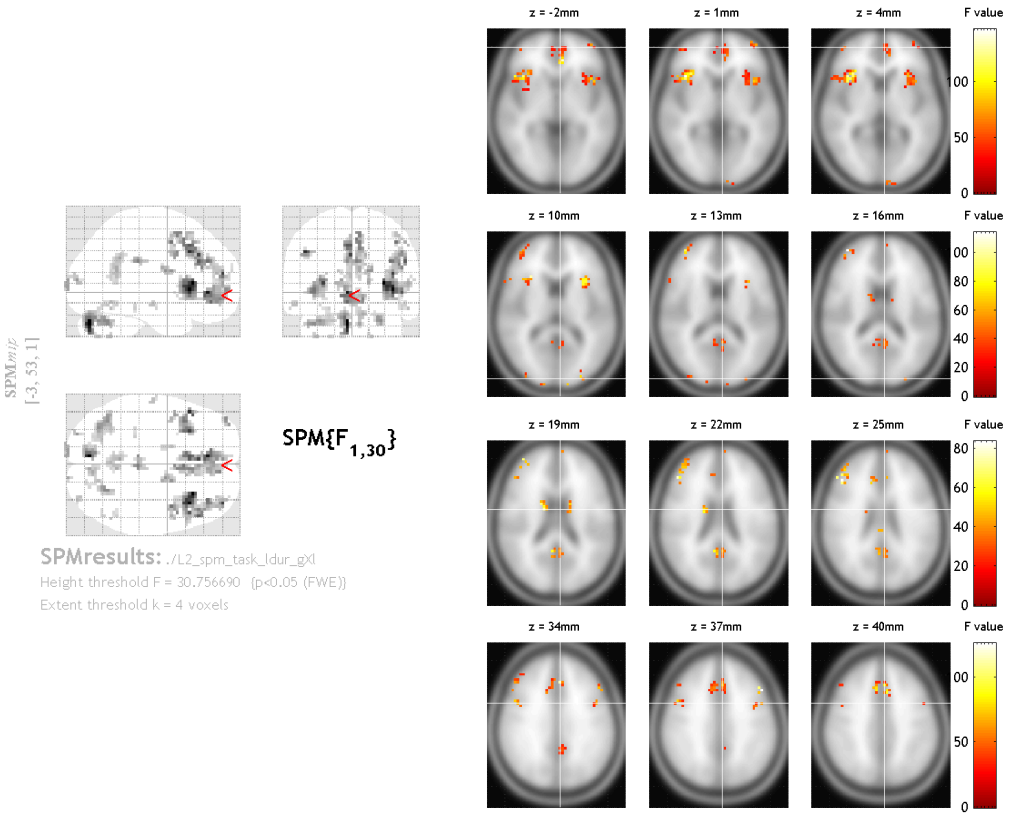

Figure 3. Statistical parametric map for all trials in the stimulus-response learning task (i.e. part 2) examining the interaction between gains and losses. *Left* is a glass brain, showing all significant clusters mapped down to 3 two dimensional representations. *Right* is a set of axial slices highlighting strong areas of activity overlaid onto the T1 MNI-352 template. Z is the height of the axial slice in MNI space.

putamen, ventral striatum/nucleus accumbens, hippocampus, and amygdala. Bilateral cortical areas were the middle frontal cortex (i.e. dorsal lateral PFC), superior frontal cortex (which contains ventral medial PFC), orbital frontal cortex, anterior and posterior cingulate.

A Way To Many. In total there 6 models under evaluation – the three kinds of similarity adjustment, ("none", "exp", and "gauss"), with two possible reward codes ("acc" and "gl"). This will expand to include an absolute (positive-only) coding scheme (more on that later), bringing the total to 12 models, with 2 terms of interest (i.e. value and the reward prediction error) that is 24 comparisons. The are however a number of confounds to our signals of interest, the similarity metrics and the reward codes themselves, the grating parameters, as well as the selected responses. However as the models are not nested (i.e. they cannot be made identical by simply adding or subtracting parameters (CITE)) and so are not amenable to F-tests, the common statistical way to compare fits. Further complicating the issues is the fact that each of the models is covariate, if not collinear, with the others. To top it off, none of the three are statistically independent; Reinforcement learning can viewed as a regression of the reward code onto behavioral choices (CITE). All these factors combined would make statistical testing difficult. But fortunately finding *the* best model is not the goal.

The latest recordings of phasic (i.e. reward prediction) activity in the VTA/SNc suggests a complicated reward and prediction error coding scheme (see **??**), where several separate sets of calculations may be carried out independently (?, ?, ?, ?).

The observed BOLD signal is then an aggregate of these many activities, making it possible that more than one of the models under study is correct. Under this constraint null hypothesis significance tests are not the right choice, model selection is. Model selection is the process finding a *family* of models/variates that best predict a given dataset (?, ?), with most techniques trying to wisely balance parsimony with increasing fit (i.e. bias versus variance (?, ?)). Unfortunately most model selection techniques require assumptions our models cannot meet (e.g. requiring statistical independence). The few that can tend to be complex recent statistical inventions and rather than navigate the those troubled and unproven waters, I took a simpler approach.

A score (AIC, Akaike Information Criterion (?, ?)) was assigned to each of the models/codes for every participant and region of interest. Based on the average score across participants normalized based on the non-parametric (boxcar) model. The absolute AIC score will vary by participant, but only the relative values are of interest. Normalization based on the boxcar model, which serves somethings like a null hypothesis, provides a way to cast each participant's fits into relative terms suitable for averaging. Using the normalized average score each model's performance was then ranked by subtracting each score from the best (lowest) score (?, ?). The set of scores that accounted for 95% of the total possible information were retained (?, ?). The retained set was then transformed to Akaike Weights, a way to compare the conditional probabilities of each model being true (?, ?).

Information on (Akaike's) Information. AIC is a measure of loss; how much information is lost by substituting the model for the true distribution, i.e. the data. The lower the AIC score, the better the model. Unlike both the null hypothesis tests, and Bayesian measures, AIC based methods do not seek to find *a* truth, but instead serve to rank models. It offers then only relative insight, and is unable to make any claims about absolute significance. Significance is a separate question, one I'll return to later. Besides this limitation, AIC has some significant advantages. Five reason are reviewed below.

One, unlike maximum-likelihood AIC is designed to a be parsimonious score. It penalizes for additional parameters. It may therefore choose a overall worse model (as measured by likelihood of mean squared error) over a better but more complex one. This is the essence of Occam's razor[1].

Two, it fits with the process of science. When designing an experiment it is rare that there are only two possible outcomes, instead typically there are several competing hypothesis, some of which may not be mutually exclusive. AIC's focus on relative differences, and evidential weights meshes perfectly with the idea of multiple working hypotheses (?, ?).

Three, truth can remain elusive. A common alternative to AIC is BIC, the Bayesian Information Criterion. Like AIC, it is derived from the log-likelihood of a model, however its derivation requires a rather strict (and often unrealistic) assumption – that the true model is among the candidates (?, ?). And while it may be

[1]Famously and pithily expressed as, "Entities are not to be multiplied beyond necessity".

philosophically debatable whether any mathematical model can *completely* describe reality, in this study it is a known fact that my models are incomplete. The reinforcement learning literature contains several findings I (or anyone) can't yet account for (see the *Introduction* for a review).

Four, AIC values are easily interpretable once they're transformed to Akaike Likelihoods or Weights[2]. The likelihood is simply the likelihood the model is correct (based on the information loss associated with it), while the Akaike Weights are just normalized likelihoods. As the Weights sum to one, the conditional likelihood of one model compared to another is just the ratio of their weights (?, ?). For example, if the conditional likelihood of model A over model B is w_A/w_B. That is, the likelihoods and Akaike Weights are intrinsically measures of effect size (?, ?, ?). Despite the fact that it is often used to express the likelihood of correctly rejecting the null hypothesis, the p value is not a measure of effect, as p is contingent not just on effect size but on sample number.

Five, AIC has a history with models of categorization. ?, ?, ?, among several others, used AIC to compare behavioral results to several alternative models of categorization.

F-them. AIC ranks offer no information on significance, in the familiar null hypothesis sense, or on the absolute fit of the model. I addressed both of these in a series of $F-tests$ run prior to AIC analysis. These (fixed-effect, across participant)

[2]Likelihood for model k among K working hypotheses/models is given by $L_k = e^{-0.5(AIC_k - min_K(AIC))}$, which is then normalized, becoming an Akaike Weight by $w_k = L_k / \sum_{k=1}^{K} L_k$ (?, ?).

ominbus tests asked whether the total set of regression parameters for each linear model (described below) could explain the BOLD time series better than chance, i.e could the null hypothesis (of 0) be rejected. However in keeping with recommendations of ?, ?, ?, who argue that as AIC and significance tests are so dissimilar that direct comparison/interaction between them them will be at best misleading, the models are not discarded based on significance. Instead all models are retained, and later AIC ranked. The F-tests are a separate measure whose results are integrated during interpretation, not during model selection/analysis.

Code, BOLD, and models.. No available fMRI analysis package returns AIC scores (or measures that could be converted to such) and none allow for the efficient (i.e programmatic) analysis of many competing computational models. So I created a roi-focused fMRI data *analysis* tool in Python (v2.7.1) to meet those two needs. This module, simply named "roi", has since been release under the BSD license and is available for download at `https://github.com/andsoandso/roi`. It relies on the nibabel library to read the nifiti-1 files (v1.2.0; `http://nipy.org/nibabel`), nitime for timeseries analysis, (v0.4; `http://nipy.sourceforge.net/nitime/`) Numpy for generic numerical work (v1.6.1; `http://numpy.scipy.org/`), with the GLS function from the scikits.statsmodels module handling the regerssions (v0.40; `http://statsmodels.sourceforge.net/`). Model-to-BOLD fit parameters, as well as other useful metadata, was then extracted and stored in text files suitable for importing into R (v2.15.1; `http://www.r-project.org/`). All plotting and model ranking (as well as the F-tests) were carried out in R. For complete BSD licensed code see,

https://github.com/andsoandso/fmri/tree/master/catreward/roi/results.

Model Results

References

Birn, R. M., Cox, R. W., & Bandettini, P. A. (2002, Jan). Detection versus estimation in event-related fmri: choosing the optimal stimulus timing. *Neuroimage*, *15*(1), 252–64.

Dale, A. M. (1999, Jan). Optimal experimental design for event-related fmri. *Hum Brain Mapp*, *8*(2-3), 109–14.

Kao, M.-H., Mandal, A., Lazar, N., & Stufken, J. (2009, Feb). Multi-objective optimal experimental designs for event-related fmri studies. *Neuroimage*, *44*(3), 849–56.

Liu, T. T. (2004, Jan). Efficiency, power, and entropy in event-related fmri with multiple trial types. part ii: design of experiments. *Neuroimage*, *21*(1), 401–13.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000, Jun). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage*, *11*(6 Pt 1), 735–59.

Wager, T. D., & Nichols, T. E. (2003, Feb). Optimization of experimental design in fmri: a general framework using a genetic algorithm. *Neuroimage*, *18*(2), 293–309.