OHBM 2017

Abstract #3778
Poster #1709

The Overfitting Toolbox (TOT):
Large-Scale Search in Model Space for Expected Neuroimaging Effects

Bernstein Center for Computational Neuroscience Berlin

Joram Soch[1,6,•], Carsten Allefeld[1,2], John-Dylan Haynes[1,2,3,4,5,6]

[1] Bernstein Center for Computational Neuroscience, Berlin, Germany [2] Berlin Center of Advanced Neuroimaging, Berlin, Germany
[3] Berlin School of Mind and Brain, Berlin, Germany [4] Excellence Cluster NeuroCure, Charité-Universitätsmedizin Berlin, Germany
[5] Department of Neurology, Charité-Universitätsmedizin Berlin, Germany [6] Department of Psychology, Humboldt-Universität zu Berlin, Germany
• Bernstein Center for Computational Neuroscience, Philippstraße 13, Haus 6, 10115 Berlin, Germany / **joram.soch@bccn-berlin.de**

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

"Isla de Muerta (…) cannot be found except by those who already know where it is." – Captain Jack Sparrow [4]

Introduction

A common problem in experimental science is if the analysis of a data set yields no significant result even though there is a strong prior belief that the effect exists. In this case, overfitting can help, a technique that has become common practice in psychology [1] and neuroimaging [2]. Functional magnetic resonance imaging (fMRI) is very suitable for overfitting, because general linear models (GLMs) allow to test a hypothesis at several ten thousand voxels, such that significant results are likely to be found. Furthermore, analysis pipelines have a high number of free parameters supporting a large model space. We present *The Overfitting Toolbox* (TOT), a set of computational tools that allow to systematically exploit multiple model estimation, parallel statistical testing, varying statistical thresholds and other techniques that allow to increase the number of positive inferences.

Features

The Overfitting Toolbox (TOT):

- assists in massive model set-up for a given fMRI data set;
- allows to circumvent the laborious burden of interrogating all these models;
- automatically searches through the model space for experimental effects;
- takes expected effect and desired brain regions as input parameters and identifies models which make this effect significant in these regions;
- can help if the effect is still not being observed despite large number of statistical tests by implementing different significance levels, extent thresholds and multiple comparison corrections.

Empirical Validation of the *The Overfitting Toolbox*

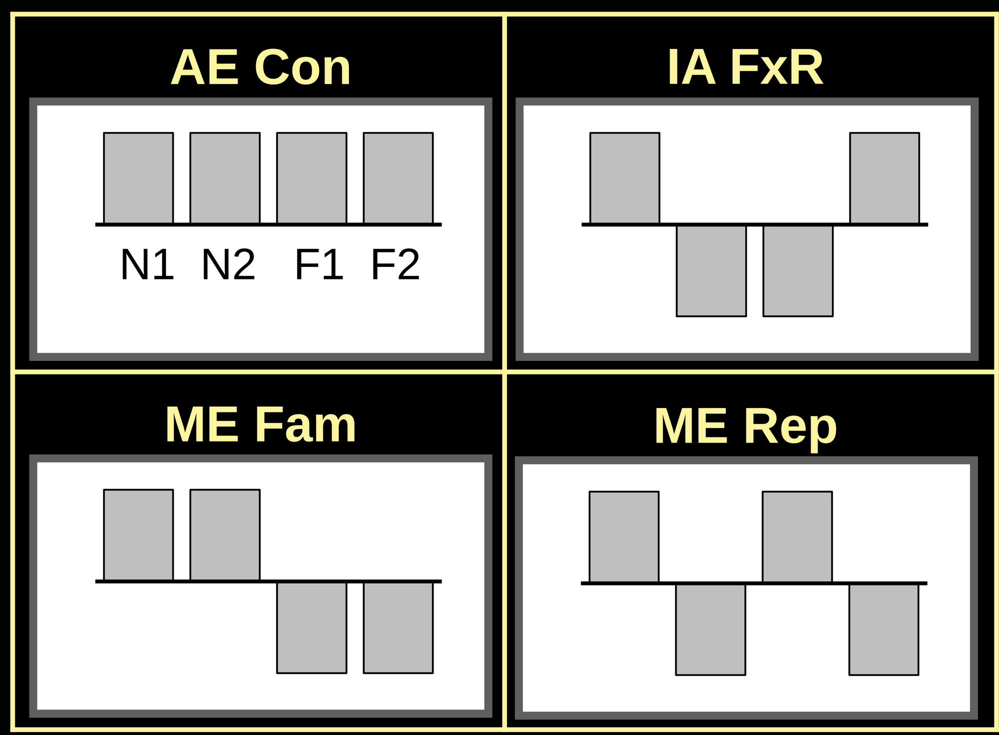
1 We analyze an SPM template data set [5,6] which was based on a 2 x 2 factorial design with 4 experimental conditions:

familiarity (Fam)		repetition (Rep): 1st presentation	repetition (Rep): 2nd presentation
	non-famous faces	N1	N2
	famous faces	F1	F2

2 We specify and estimate a very large model space consisting of 4,320 GLMs:

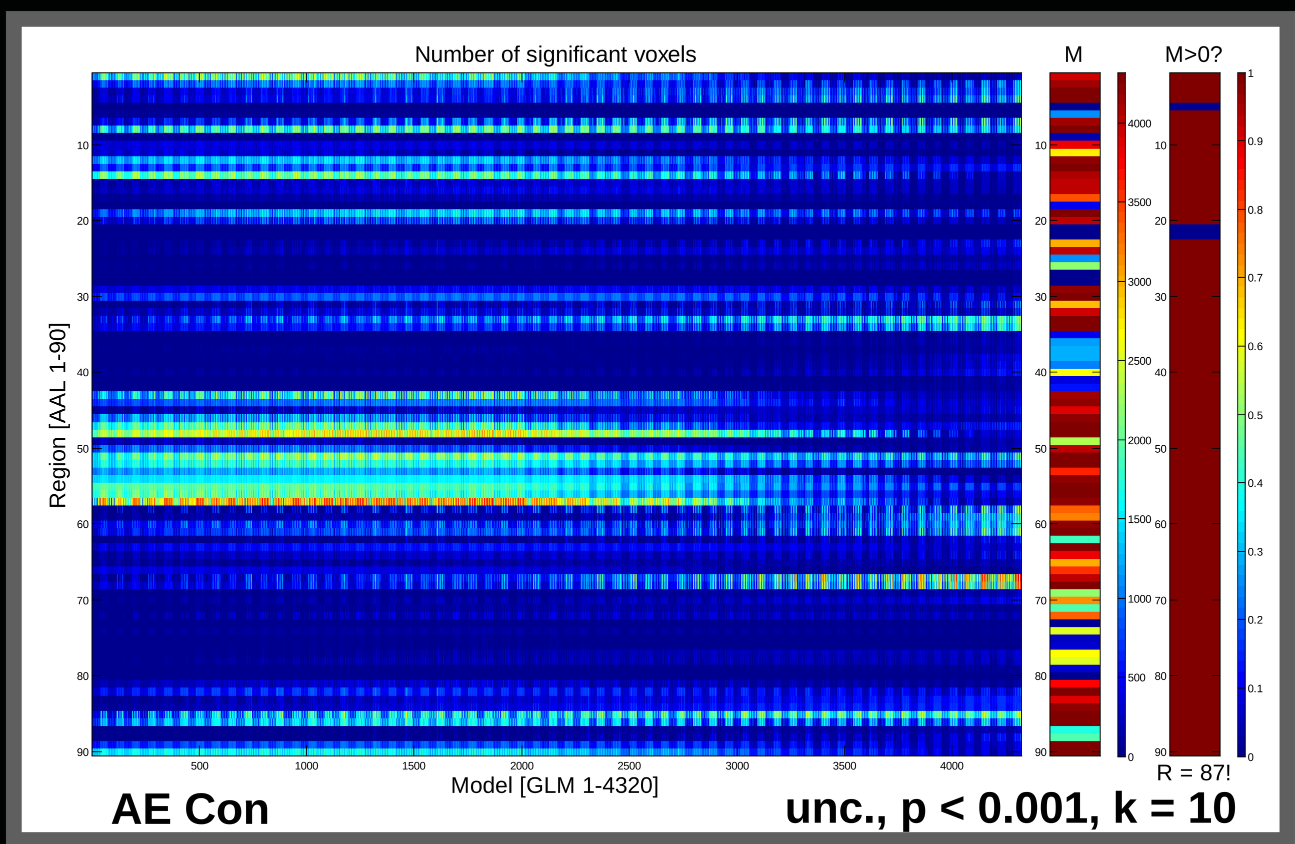
	Model space dimensions						
1	event onsets	−2s	−1s	±0s	+1s	+2s	5
2	event durations	0s, 0.5s	1s	1.5s	2s	2.5s	6
3	parametric regressors	0, 1	2	3	4	5	6
4	movement params	none	transl.	rotat.	all		4
5	hemodyn. derivatives	none	1st only	1st & 2nd			3
6	AR model	AR(0)	AR(1)				2
Π	Total number of models						4,320

3 The experimental design implies that 4 different effects can be tested:

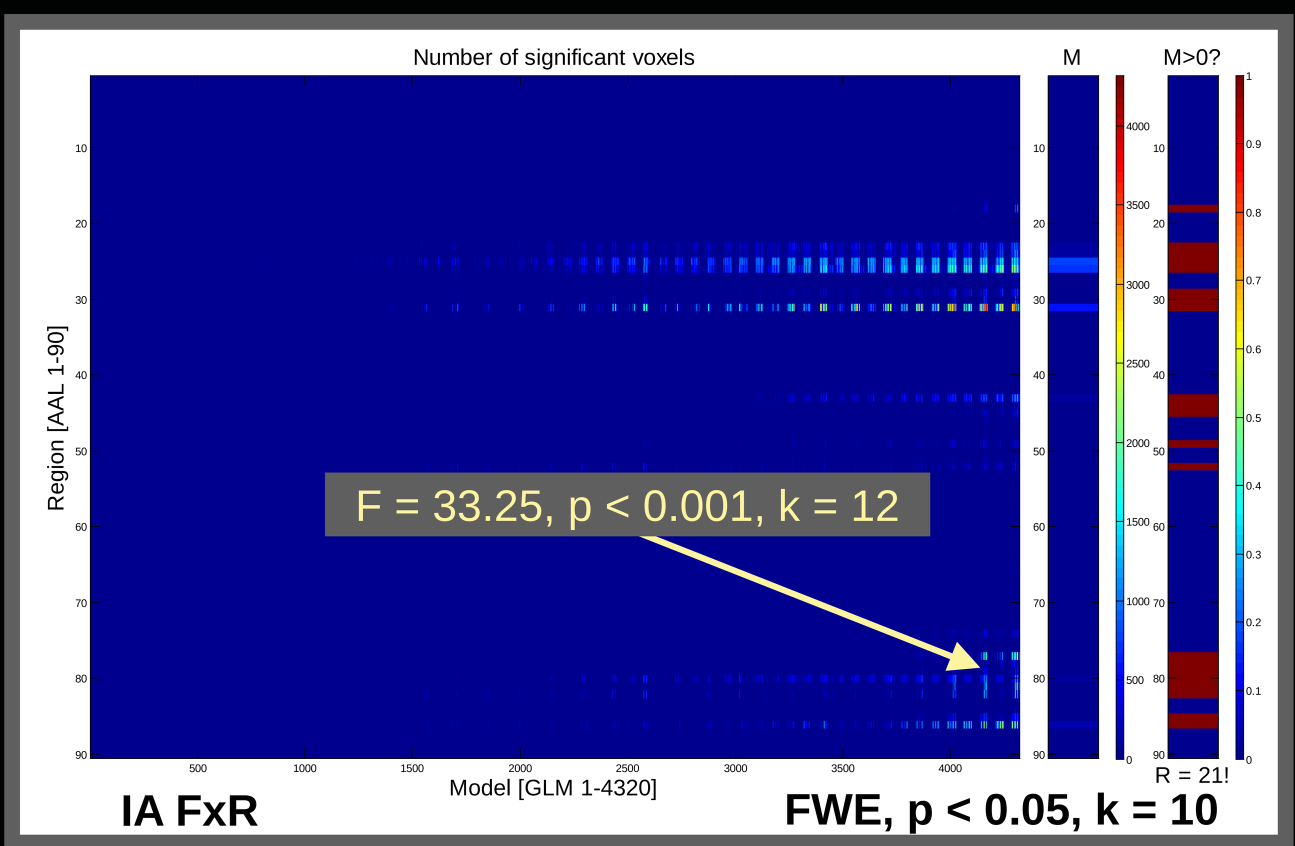


4 We look for these effects in 90 different regions [7]:

5 When **R** = number of regions and **M** = number of models, this gives us an R x M matrix for each contrast indicating how many voxels are activated when testing for this *effect* in a specific *region* using a specific *model*. Here's this matrix for the average effect of condition (**AE Con**) when not correcting for multiple comparisons (**unc., p < 0.001, k = 10**).



6 Let's say we „know" that there is an interaction of familiarity and repetition (**IA FxR**) in left auditory cortex (**AAL 79**). The toolbox can help us to identify a *model* for which this *effect* becomes significant in that *region*. In fact, such a model exists, even when correcting for multiple comparisons (**FWE, p < 0.05, k = 10**).



7 With TOT, we detected experimental effects in *almost every region* using *at least one model*. This was not the case when controlling for multiple analyses using model selection [8] or model averaging [9]. In particular, the demonstrated interaction of familiarity and repetition (**IA FxR**) in left auditory cortex (**Step 6**) was not significant with cvBMS [8] and cvBMA [9]:

Proportion of regions with significant voxels	TOT unc.	TOT FWE	cvBMS unc.	cvBMS FWE	cvBMA unc.	cvBMA FWE
AE Con	97%	81%	82%	43%	81%	40%
ME Fam	88%	27%	28%	0%	22%	0%
ME Rep	89%	22%	14%	0%	11%	0%
IA FxR	80%	23%	52%	1%	51%	1%

8 Your turn! Ask the presenter to identify a GLM that makes your favorite effect significant! Input the expected effect (**Step 3**) and the desired region (**Step 4**) as well as statistical thresholds (**Step 5/6**) and TOT outputs *models* that allow to detect this *effect* in that *region*.

Discussion

We have demonstrated the potential of overfitting in fMRI data analysis and how to turn it from a subjective enterprise into an objective procedure. An important advantage over previous manual overfitting approaches is that TOT allows to automatically search through a large model space. These methods could have improved some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results [3]. As a next step, it would be desirable to reanalyze the entire amount of previous fMRI studies to harvest the false-positive effects that might have been missed using conventional statistical techniques [1,2]. Widespread use of *The Overfitting Toolbox* (TOT) will allow researchers to uncover literally unthinkable sorts of effects and lead to more spectacular findings and news coverage for the entire fMRI community [3].

References

(1) Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, vol. 349, iss. 6251, art. aac4716.
(2) Szucs D, Ioannidis JPA (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS ONE, vol. 15, no. 3, art. e2000797.
(3) Eklund A, Nichols TE, Knutsson H (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. PNAS, vol. 113, no. 28, pp. 7900-7905.
(4) Verbinski G, Bruckheimer J (2003). Pirates of the Caribbean: The Curse of the Black Pearl, Walt Disney Studios, July 9, 2003.
(5) Henson RNA, Shallice T, Gorno-Tempini ML, Dolan RJ (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cerebral Cortex, vol. 12, pp. 178-186.
(6) Ashburner J et al. (2013). SPM8 Manual. Chapter 29: "Face fMRI data", pp. 233-260; URL: http://www.fil.ion.ucl.ac.uk/spm/doc/spm8_manual.pdf.
(7) Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Juliot M (2002). Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation, NeuroImage, vol. 15, iss. 1, pp. 273-289.
(8) Soch J, Haynes JD, Allefeld C (2016). How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection. NeuroImage, vol. 141, pp. 469-489.
(9) Soch J, Meyer AP, Haynes JD, Allefeld C (2017). How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging. NeuroImage, in review.