Model Interpretation

Janaina Mourao-Miranda,

Machine Learning and Neuroimaging Lab,

University College London, UK

Course 2021

September 15th- 17th

Interpretation: definition

- In machine learning:

Which features are driving the predictions?

- In neuroscience:

Which brain regions (or time windows) are driving the predictions?

Linear Predictive Models

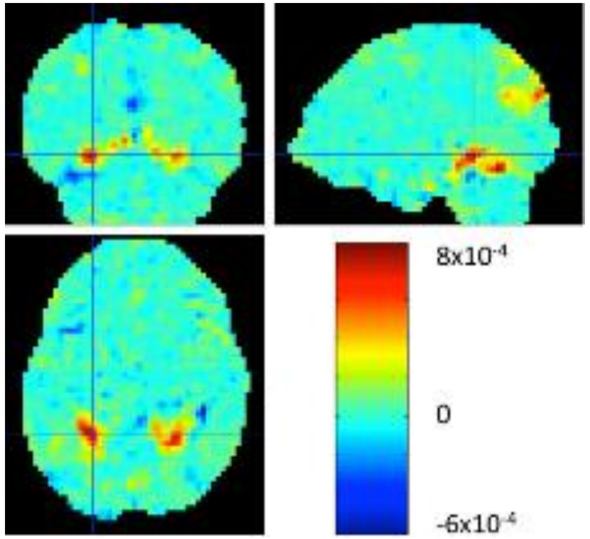
- Linear predictive models (classifier or regression) are parameterized by a weight vector **w** and a bias term *b*.

- The general equation for making predictions for a new test example $\mathbf{x}_*$ is:

$$f(\mathbf{x}_*) = \mathbf{w} \times \mathbf{x}_* + b$$

- **w** has the same dimensionality of the input data and can be plotted as an image.

Weight map or predictive pattern



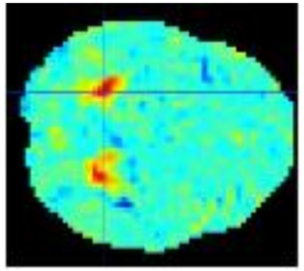
(Haxby dataset, S1, Faces vs Houses)

Predictive function

Linear predictive function: $f(\mathbf{x}_*) = \mathbf{w} \times \mathbf{x}_* + b$

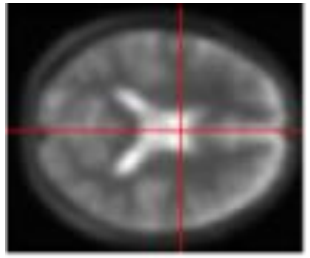
Estimated parameters:

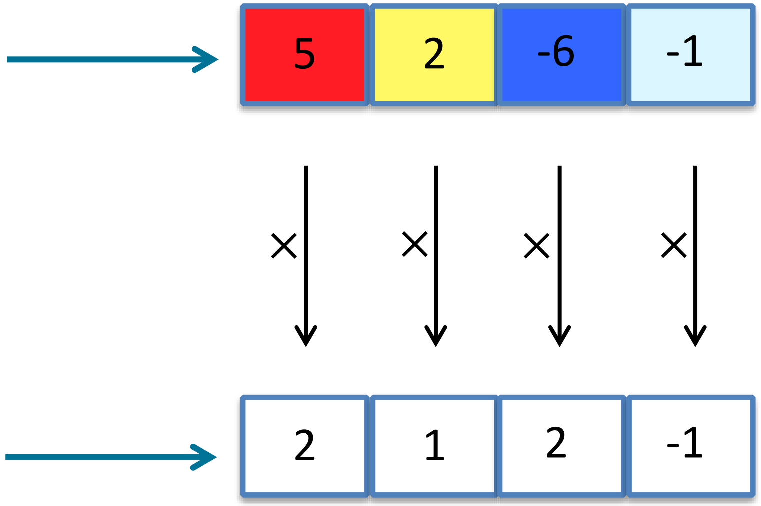
w:



b: 0

New example (**x***):





5	2	-6	-1
×	×	×	×
2	1	2	-1

$$f(\mathbf{x}_*) = (5 \times 2) + (2 \times 1) + (-6 \times 2) + (-1 \times -1) + 0$$

$$f(\mathbf{x}_*) = 10 + 2 - 12 + 1 = 1$$

$f(\mathbf{x}_*)$ is the predicted score for regression or the distance to the decision boundary for classification models.

How is w computed?

- **w** is estimated by solving an optimization problem consisting of a data fit term *E* and a penalty/regularization term *J*.

$$\min_{\mathbf{w}\hat{I}\ R^p}\{E(\mathbf{w})+/\,J(\mathbf{w})\}$$

$\downarrow$

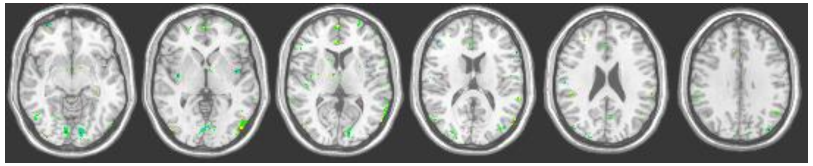
Regularization parameter

- Data fit term or loss function (*E*): denotes the price we pay when we make mistakes in the predictions (e.g. squared loss, Hinge loss).

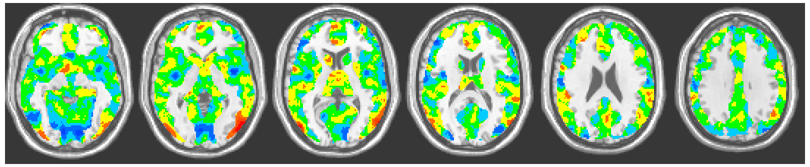
- Regularization term (*J*): favours certain properties (e.g. sparsity) and improves the generalisation over unseen examples (e.g. L2-norm, L1-norm).

- Many machine learning algorithms are particular choices of *E* and *J.*

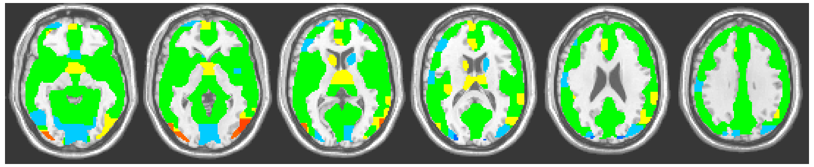
Impact of the regularization on w



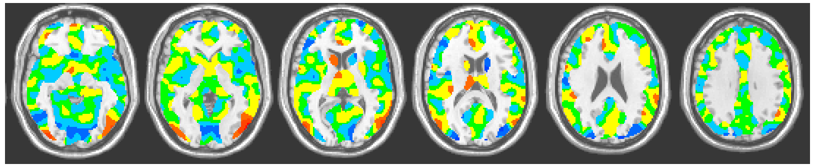
LASSO
86.31%



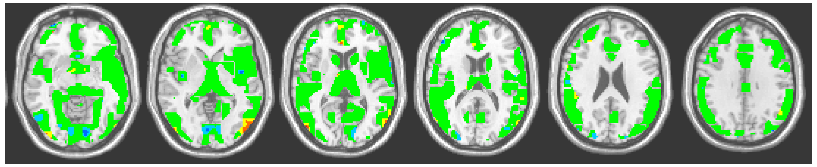
Elastic Net
88.02%



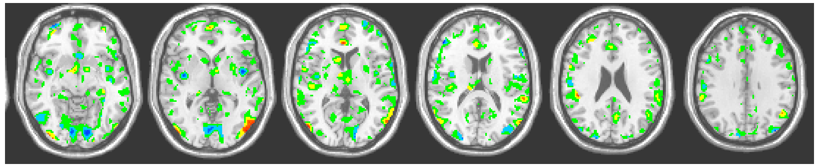
Total Variation (TV)
85.79%



Laplacian (LAP)
83.71%



Sparse TV
85.86%



Sparse LAP
87.05%

- Weight maps for classifying fMRI images during visualization of pleasant vs. unpleasant pictures.

- **All models used a square loss + regularization.**

Baldassarre L, Pontil M, Mourao-Miranda J (2017)

Potential strategies to improve model interpretability

1. Feature selection (e.g. DeMartino et al. 2008; Chu et al. 2012; Rondina et al. 2013)
2. Searchlight mapping (e.g. Krigeskorte et al, 2006; Allefed and Haynes 2014)
3. Sparse algorithms (e.g. Gramfort et al. 2013; Grosenick et al. 2013; Baldassarre et al. 2017)
4. Atlas based weight summarization (e.g. Schrouff et al. 2013)
5. Multiple Kernel Learning (e.g. Schrouff et al. 2018)
6. Permutation test (e.g. Mourao-Miranda et al, 2005; Gaonkar and Davatzikos, 2013)
7. Transforming weights into activation patterns (e.g. Heufe at al., 2014)

Potential strategies to improve model interpretability in PRoNTo

- Feature selection (e.g. DeMartino et al. 2008; Chu et al. 2012; Rondina et al. 2013)

- Searchlight mapping (e.g. Krigeskorte et al, 2006; Allefed and Haynes 2014)

1. Sparse algorithms (e.g. Gramfort et al. 2013; Grosenick et al. 2013; Baldassarre et al. 2017)

2. Atlas based weight summarization (e.g. Schrouff et al. 2013)

3. Multiple Kernel Learning (e.g. Schrouff et al. 2018)

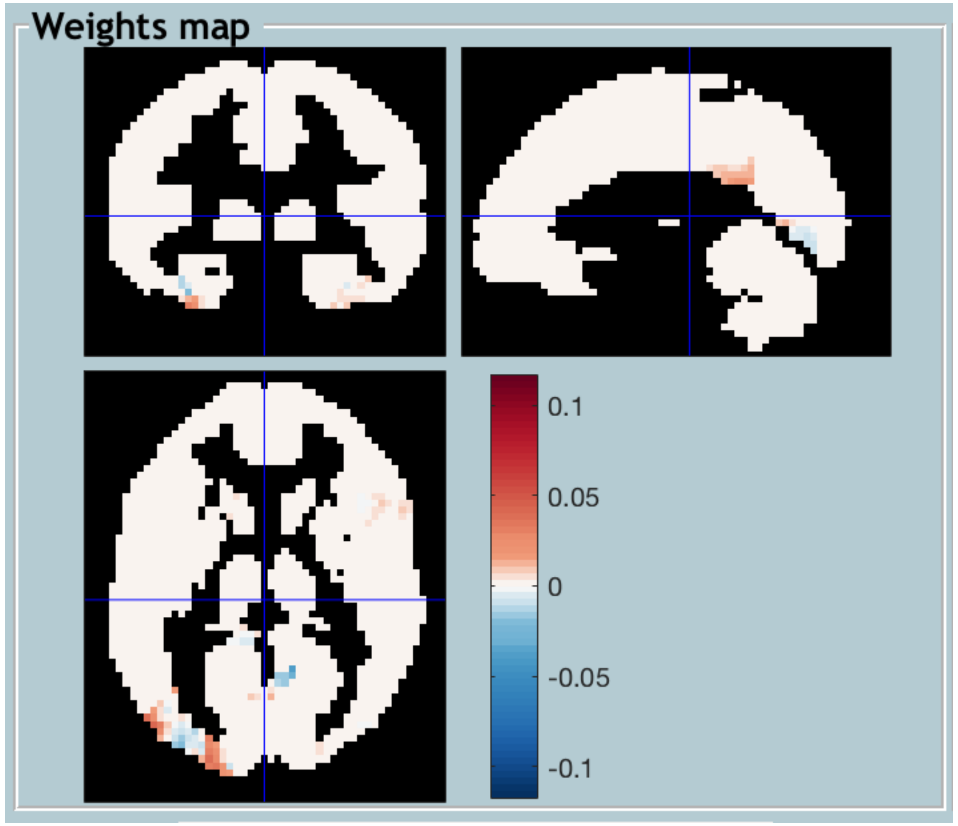
- Permutation test (e.g. Mourao-Miranda et al, 2005; Gaonkar and Davatzikos, 2013) with extra code

- Transforming weights into activation patterns (e.g. Heufe at al., 2014)

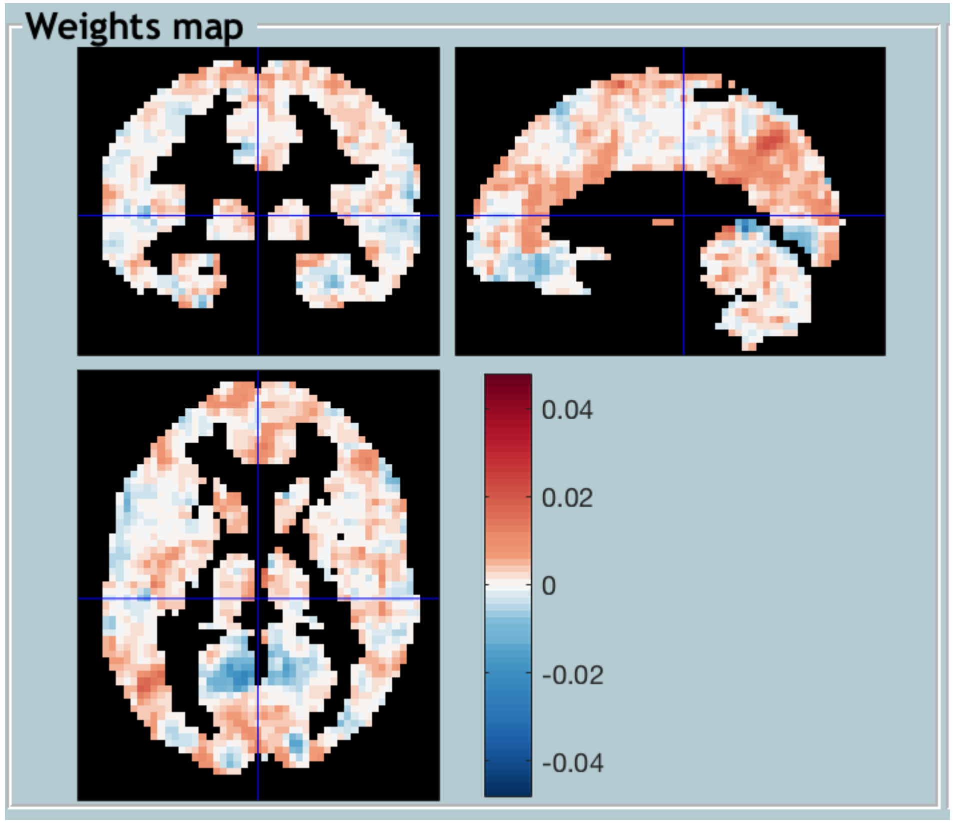
1- Sparse algorithms

- Include a regularization term that enforces sparsity (e.g. LASSO, Elastic-net).
- Examples of sparse algorithms in PRoNTo:
  - ✓ L1- Multi-kernel Learning
  - ✓ L1- Support Vector Machine (non-kernel)
  - ✓ L1- Logistic regression (non-kernel)
- Other sparse models can be added to PRoNTo using custom machine option (see Chapter 22 in the PRoNTo manual).

1- Sparse algorithms



Sparse model – L1MKL
Haxby data, S1, faces vs. houses

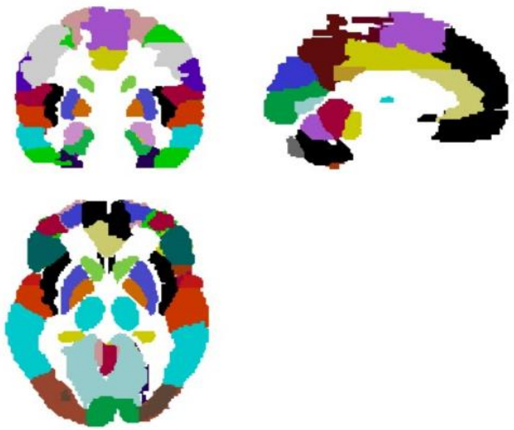


Non sparse model – SVM
Haxby data, S1, faces vs. houses

2- Atlas based summarization

- Summarize whole brain weights using a pre-defined atlas (Schrouff et al, 2013).

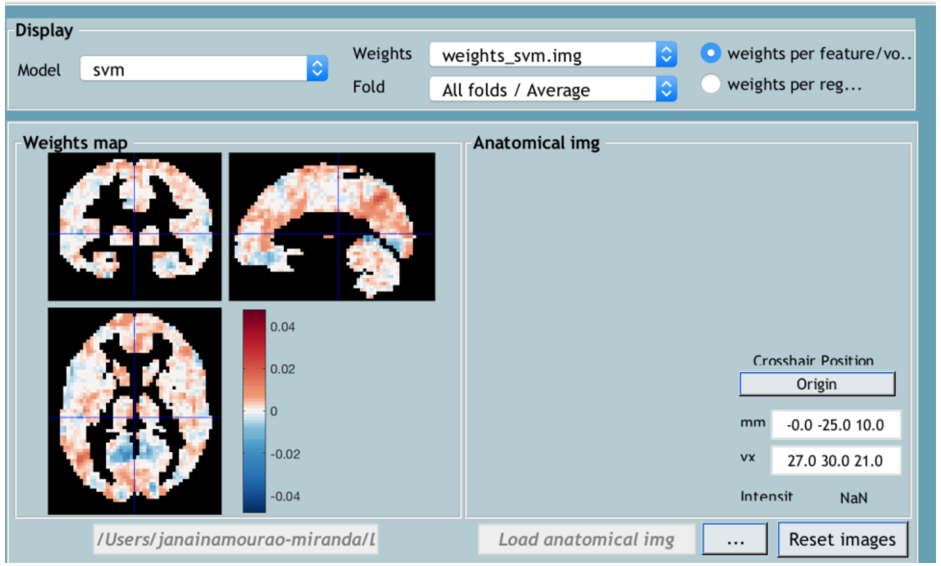
- Average the absolute value of the weights within each regions to create a rank according to their contribution to the decision function.



AAL atlas (Tzourio-Mazoyer, 2002)

$$NW_{ROI} = \frac{\underset{v\text{Î}\ ROI}{\text{å}} \left|w_v\right|}{\#\{v\,\text{Î}\ \ ROI\}}$$

2- Atlas based summarization



Display

Model: svm

Weights: weights_svm.img

Fold: All folds / Average

● weights per feature/vo..

○ weights per reg...

Weights map

Anatomical img

Crosshair Position

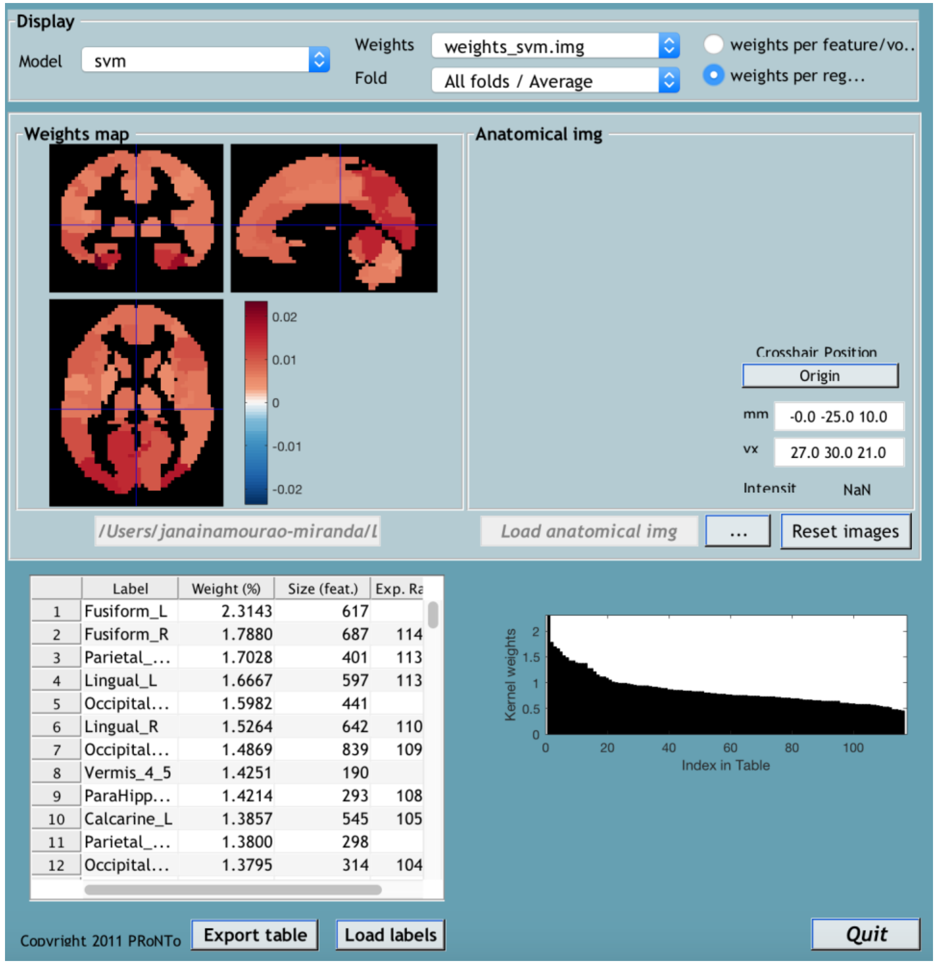
mm: -0.0 -25.0 10.0

vx: 27.0 30.0 21.0

Intensit: NaN

/Users/janainamourao-miranda/L

Non sparse model – SVM
Haxby data, S1, faces vs. houses



Display

Model: svm

Weights: weights_svm.img

Fold: All folds / Average

○ weights per feature/vo..

● weights per reg...

Weights map

Anatomical img

Crosshair Position

mm: -0.0 -25.0 10.0

vx: 27.0 30.0 21.0

Intensit: NaN

/Users/janainamourao-miranda/L

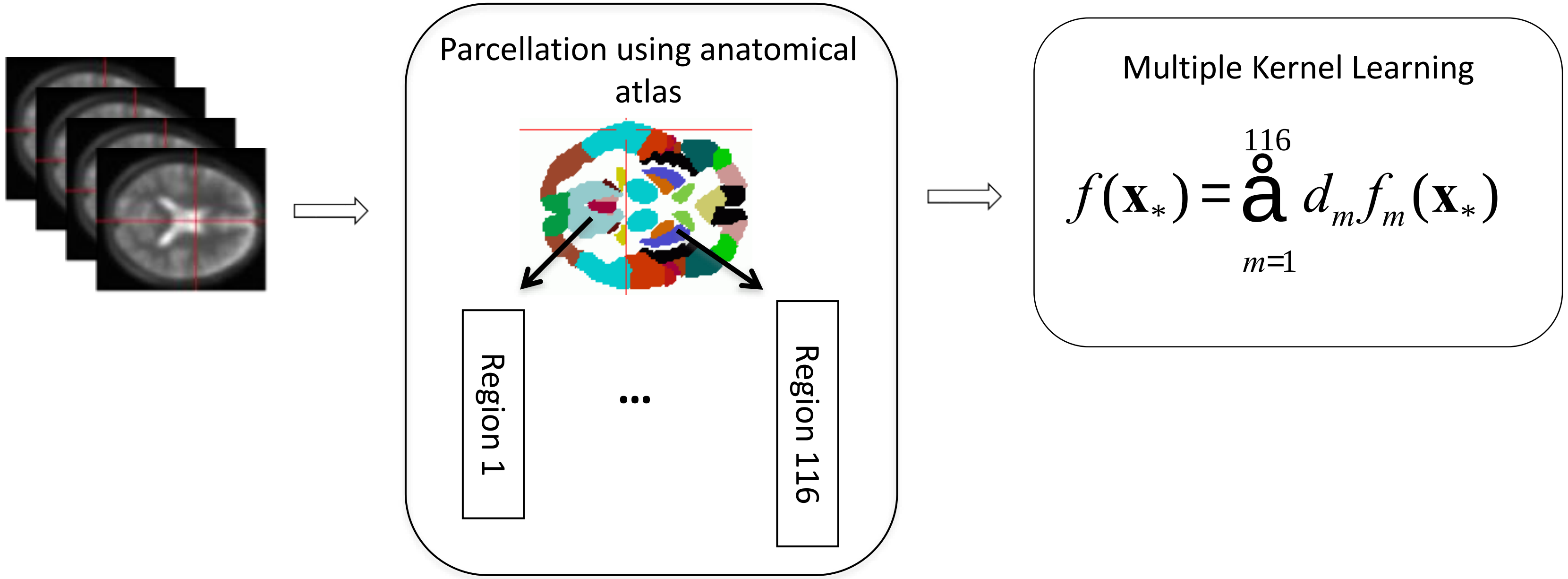
	Label	Weight (%)	Size (feat.)	Exp. Ra
1	Fusiform_L	2.3143	617	
2	Fusiform_R	1.7880	687	114
3	Parietal_...	1.7028	401	113
4	Lingual_L	1.6667	597	113
5	Occipital...	1.5982	441	
6	Lingual_R	1.5264	642	110
7	Occipital...	1.4869	839	109
8	Vermis_4_5	1.4251	190	
9	ParaHipp...	1.4214	293	108
10	Calcarine_L	1.3857	545	105
11	Parietal_...	1.3800	298	
12	Occipital...	1.3795	314	104

Copyright 2011 PRoNTo

Non sparse model – SVM with atlas based summarization
Haxby data, S1, faces vs. houses

3- Multiple Kernel Learning

- Learn simultaneously the decision function and the relative contribution of the different kernels (e.g. brain regions or time windows, Schrouff et al, 2018).
- L1-MKL (SimpleMKL, Rakotomamonjy, et al. 2008): sparsity on the kernel combination -> selects a subset of kernels (e.g. brain regions or time windows) that are relevant to the model.



Parcellation using anatomical atlas

Region 1

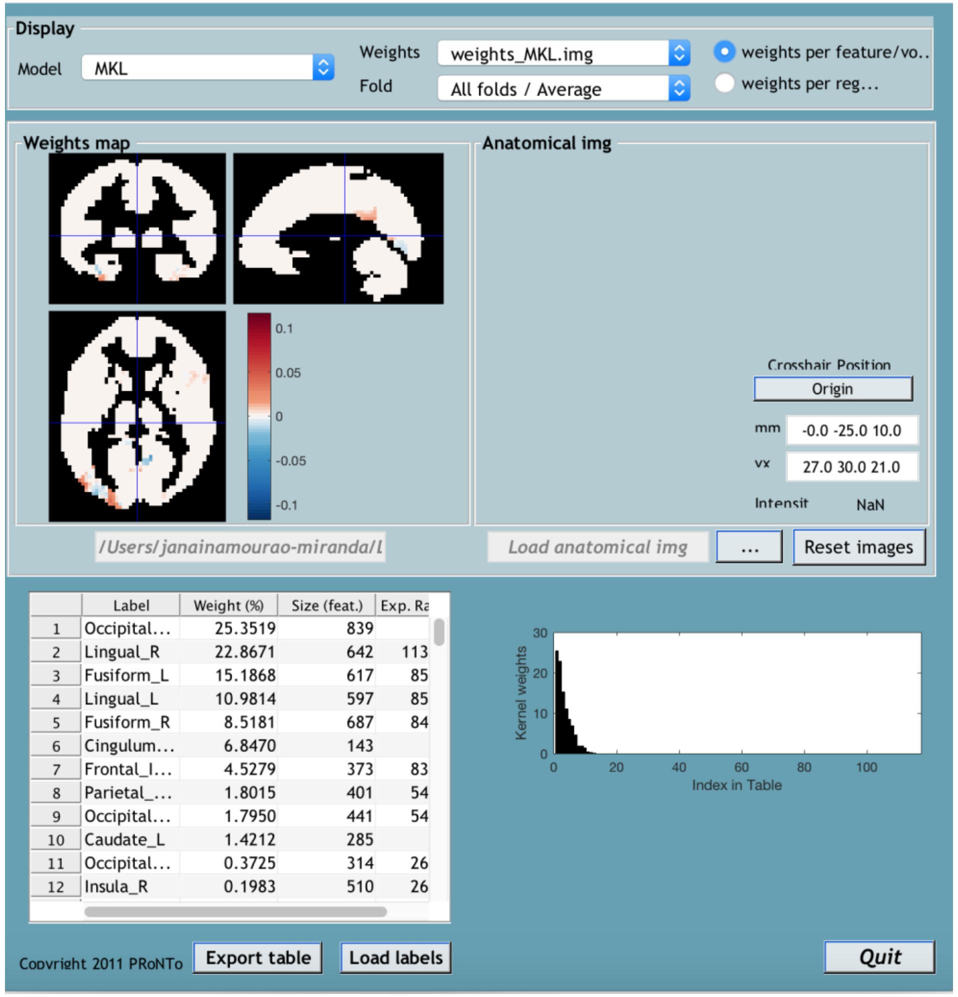
...

Region 116

Multiple Kernel Learning

$$f(\mathbf{x}_*) = \sum_{m=1}^{116} d_m f_m(\mathbf{x}_*)$$

3- Multiple Kernel Learning



Model: MKL

Weights: weights_MKL.img

Fold: All folds / Average

● weights per feature/vo..

○ weights per reg...

mm: -0.0 -25.0 10.0

vx: 27.0 30.0 21.0

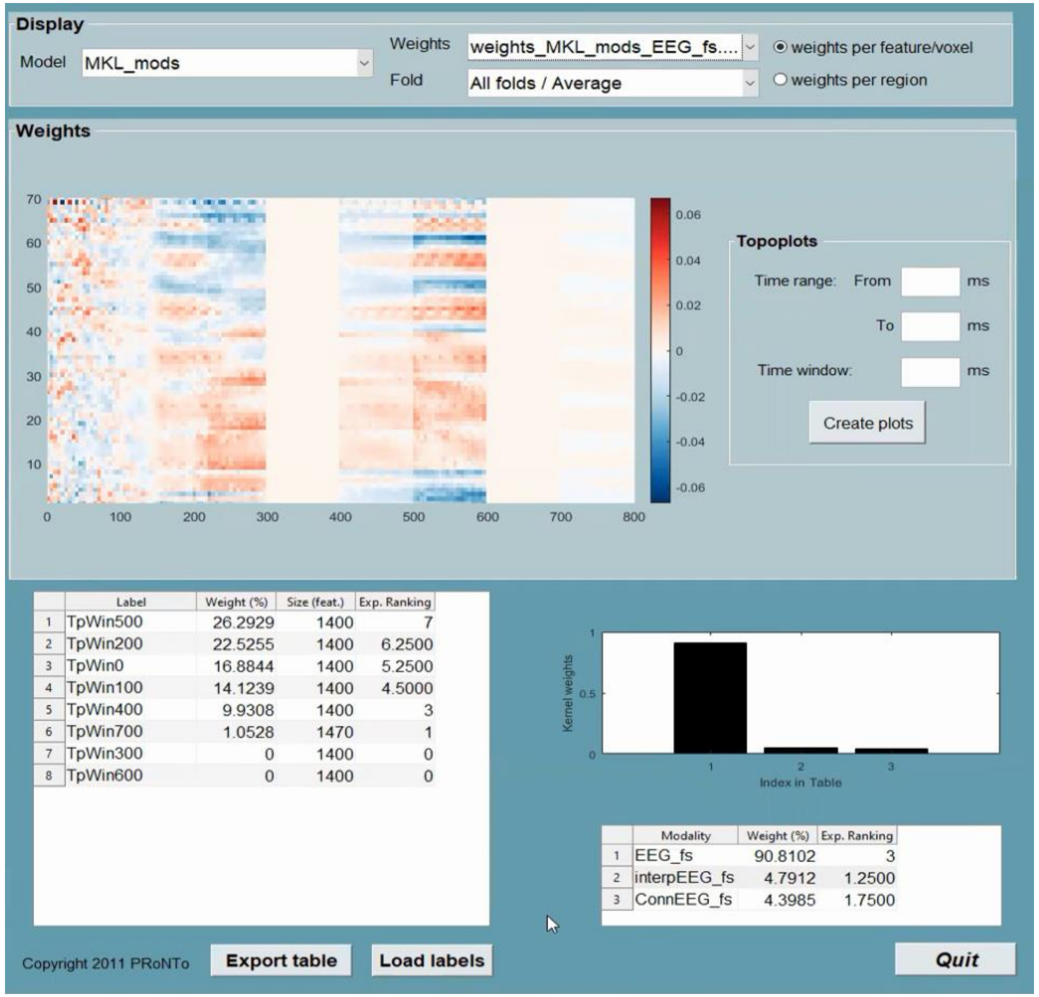
Intensit: NaN

/Users/janainamourao-miranda/L

	Label	Weight (%)	Size (feat.)	Exp. Ra
1	Occipital...	25.3519	839	
2	Lingual_R	22.8671	642	113
3	Fusiform_L	15.1868	617	85
4	Lingual_L	10.9814	597	85
5	Fusiform_R	8.5181	687	84
6	Cingulum...	6.8470	143	
7	Frontal_I...	4.5279	373	83
8	Parietal_...	1.8015	401	54
9	Occipital...	1.7950	441	54
10	Caudate_L	1.4212	285	
11	Occipital...	0.3725	314	26
12	Insula_R	0.1983	510	26

Copyright 2011 PRoNTo

Sparse model – L1MKL
Haxby data, S1, faces vs. houses



Model: MKL_mods

Weights: weights_MKL_mods_EEG_fs....

Fold: All folds / Average

● weights per feature/voxel

○ weights per region

Time range: From ms

To ms

Time window: ms

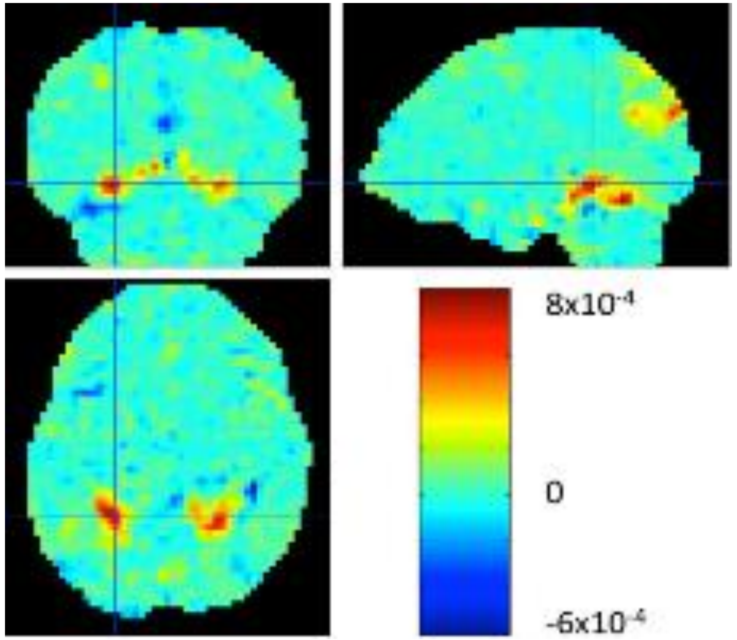
	Label	Weight (%)	Size (feat.)	Exp. Ranking
1	TpWin500	26.2929	1400	7
2	TpWin200	22.5255	1400	6.2500
3	TpWin0	16.8844	1400	5.2500
4	TpWin100	14.1239	1400	4.5000
5	TpWin400	9.9308	1400	3
6	TpWin700	1.0528	1470	1
7	TpWin300	0	1400	0
8	TpWin600	0	1400	0

	Modality	Weight (%)	Exp. Ranking
1	EEG_fs	90.8102	3
2	interpEEG_fs	4.7912	1.2500
3	ConnEEG_fs	4.3985	1.7500

Copyright 2011 PRoNTo

Sparse model – L1MKL
Multimodal face data, S1, famous vs. scrambled

Summary: weight maps interpretation



- ✓ Spatial representation of the predictive function.
- ✓ Show the contribution of each feature/voxel for the prediction.
- ✓ Multivariate pattern -> All voxels with weights different from zero contribute to the final prediction (no arbitrary threshold should be applied).
- ✓ The choice of regularization affects the sparseness and smoothness of **w.**
- ✓ Different strategies have been proposed to improve their interpretation.

Recommended reading

- Schrouff J, Mourao-Miranda J. ***Interpreting weight maps in terms of cognitive or clinical neuroscience: nonsense?*** International Workshop on Pattern Recognition in Neuroimaging (PRNI) (2018b).
- Schrouff J, Monteiro JM, Portugal L, Rosa MJ, Phillips C, Mourao-Miranda J. ***Embedding anatomical or functional knowledge in whole-brain multiple kernel learning models.*** Neuroinformatics (2018a).
- Baldassarre, L., Pontil, M. & Mourão-Miranda, J. **Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding.** Frontiers in Neuroscience: Brain Imaging Method (2017).
- Kia, S.M., Vega-Pons, S., Weisz, N. & Passerini, A. **Interpretability of Multivariate Brain Maps in Linear Brain Decoding: Definition, and Heuristic Quantification in Multivariate Analysis of MEG Time-Locked Effects.** Frontiers in Neuroscience (2017).
- Rondina J., Hahn T., de Oliveira L., Marquand A., Dresler T., Leitner T., Fallgatter A., Shawe-Taylor J. & Mourao-Miranda J. **SCoRS - a method based on stability for feature selection and mapping in neuroimaging.** IEEE Trans Med Imaging (2014).
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B. & Bießmann, F. **On the interpretation of weight vectors of linear models in multivariate neuroimaging.** NeuroImage (2014).
- Allefeld C, Haynes J-D. **Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA.** Neuroimage (2014).
- Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B. & Taylor, J.E. **Interpretable whole-brain prediction analysis with GraphNet.** NeuroImage (2013)

Recommended reading

- Schrouff, J., Cremers, J., Garraux, G., Baldassarre, L., Mourão-Miranda, J. & Phillips, C. **Localizing and comparing weight maps generated from linear kernel machine learning models.** International Workshop on Pattern Recognition in Neuroimaging (PRNI) (2013).
- Gaonkar, B. & Davatzikos, C. **Analytic estimation of statistical significance maps for support vector machine based multivariate image analysis and classification.** NeuroImage (2013).
- Gramfort, A., Thirion, B., and Varoquaux, G. **Identifying predictive regions from fMRI with TV-l1 prior.** International Workshop on Pattern Recognition in Neuroimaging (PRNI) (2013).
- Rakotomamonjy, A., Bach, F., Canu, S. & Grandvalet, Y. **SimpleMKL.** Journal of Machine Learning (2008).
- De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. **Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns.** NeuroImage (2008).
- Kriegeskorte, N., Rainer, G. & Bandettini, P. **Information-based functional brain mapping.** *PNAS* 103 (2006).
- Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. **Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data**. NeuroImage (2005).

Thank you!

Questions?