Feature selection and multivariate mapping in neuroimaging

CSML talk

September 2012

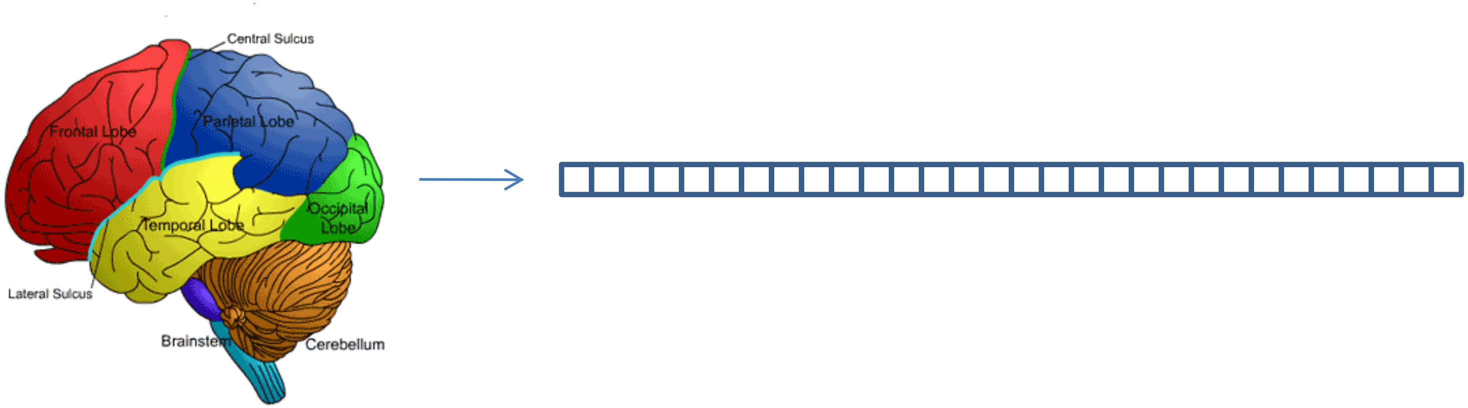
Jane Maryam Rondina

Supported by wellcome trust

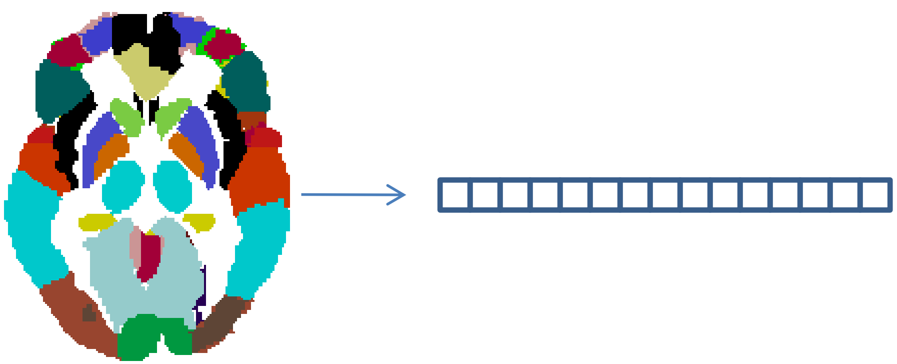
CAPES

Features extraction in neuroimaging

- Individual voxels constrained to a specific brain area

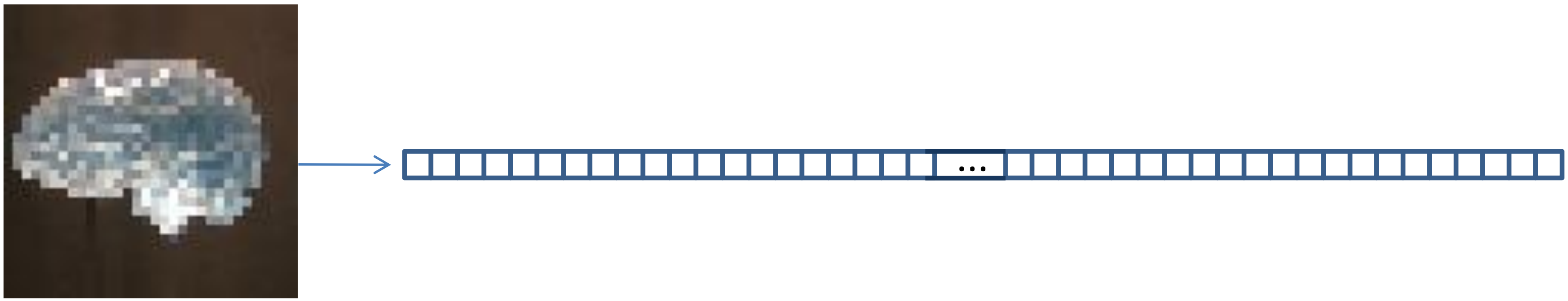


- Summarization of anatomical regions of interest



Features extraction in neuroimaging

- Individual voxels in the whole brain

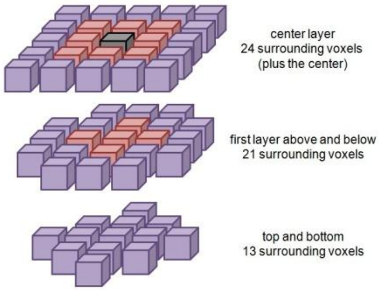


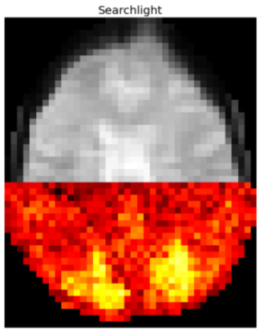
- Normalized images:
  - hundreds of thousands of voxels X few hundreds of examples;

- Challenges related to high dimensionality:
  - Model's performance
  - Interpretation

Feature selection methods in neuroimaging

- Univariate selection
  - *t*-test, F-test, ANOVA
- Recursive Feature Elimination (RFE)[1]
  - Recursive elimination of minimum weight voxel
  - Stepsize
- Searchlight[2]





1 Guyon and Elisseefi, 2003
2 Kriegeskorte, Goebel and Bandettini, 2005

Stability selection

- It is a general theory to address problems related to **variable selection** or estimation of discrete structure (as graphs or clusters);
- It relies on data perturbation (e.g. sub-sampling) in combination with high-dimensional selection algorithms;
- Its properties are promising to applications involving high dimensional data (specially the *p>>n* case)

Meinshausen, N., Buhlmann, P.: Stability selection. Journal of the Royal Statistical Society. 72 (2010) 417-473

Stability selection

$\{\hat{S}^{\lambda};\ \lambda \in \Lambda\}$, Variable selection in a traditional setting

Definition 1 (Selection probabilities) *Let I be a random subsample of $\{1,\ldots,n\}$ of size $\lfloor n/2 \rfloor$, drawn without replacement. For every set $K \subseteq \{1,\ldots,p\}$, the probability of being in the selected set $\hat{S}^{\lambda}(I)$ is*

$$\hat{\Pi}_K^{\lambda} = P^*(K \subseteq \hat{S}^{\lambda}(I)). \tag{5}$$

Definition 2 (Stable variables) *For a cutoff π_{thr} with $0 < \pi_{thr} < 1$ and a set of regularisation parameters Λ, the set of stable variables is defined as*

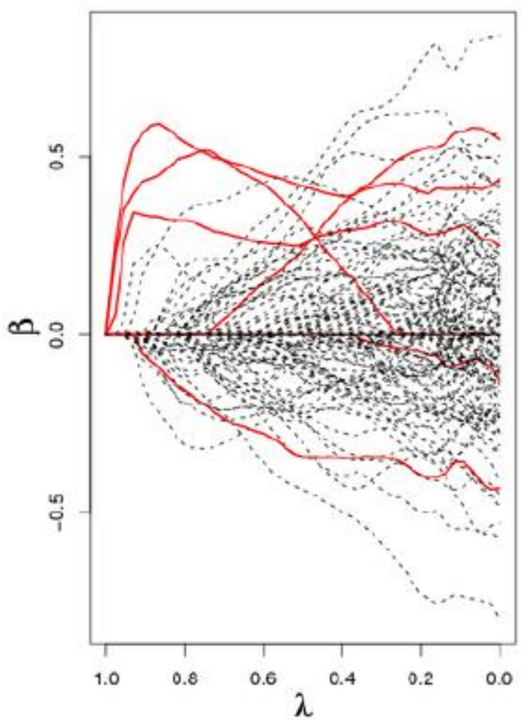
$$\hat{S}^{stable} = \{k : \max_{\lambda \in \Lambda} \hat{\Pi}_k^{\lambda} \geq \pi_{thr}\}. \tag{7}$$

Stability selection using LASSO

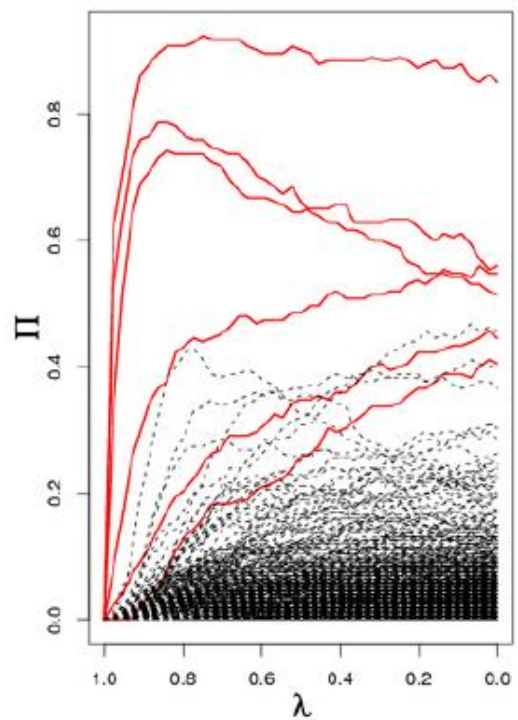
The Least Absolute Shrinkage and Selection Operator (LASSO) [1]

$$\hat{\beta}^{\lambda} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k|,$$

[1] Tibshirani R. Regression Shrinkage and Selection via the Lasso (1996). *Journal of the Royal Statistical Society*, vol 58(1), pgs 267-288.



Lasso path



Stability path of Lasso

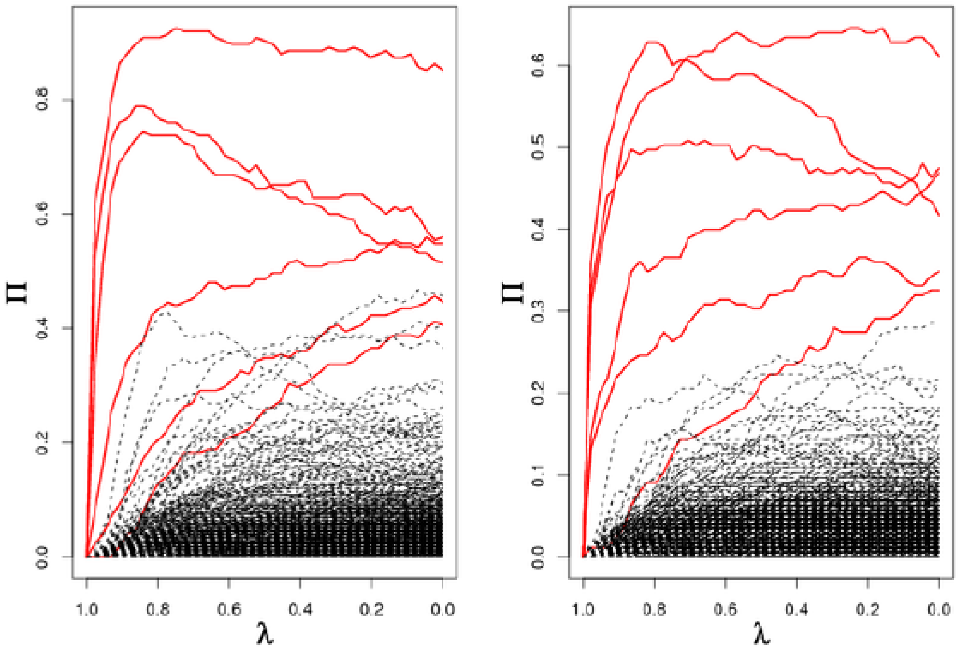
Meinshausen,& Buhlmann, 2010

Lasso and Randomised Lasso

Randomised Lasso with weakness $\alpha \in (0, 1]$:

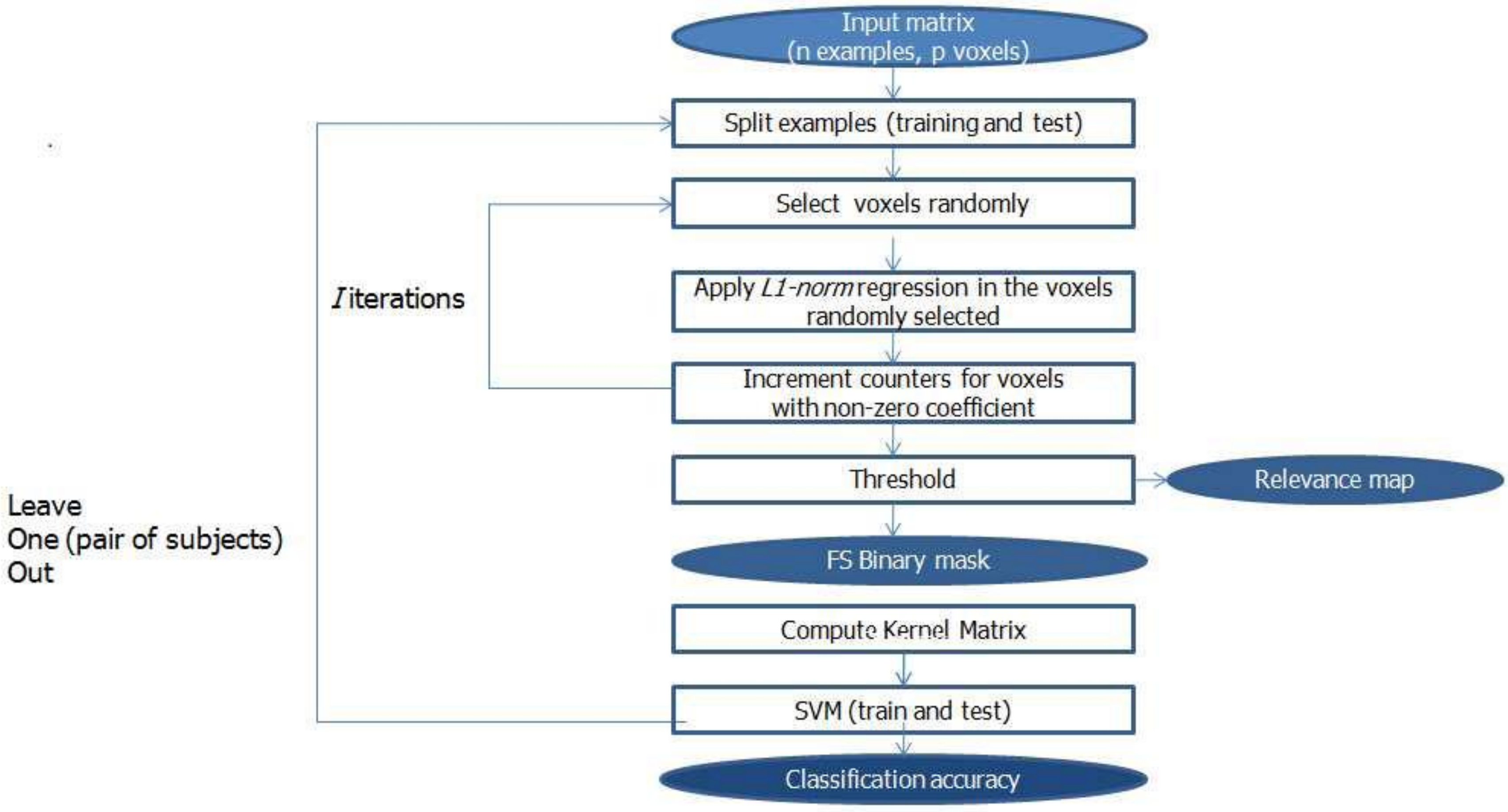
Let W_k be i.i.d. random variables in $[\alpha, 1]$ for $k = 1, \dots, p$. The randomised Lasso estimator $\hat{\beta}^{\lambda,W}$ for regularisation parameter $\lambda \in \mathbb{R}$ is then

$$\hat{\beta}^{\lambda,W} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k}. \qquad (13)$$



Meinshausen,& Buhlmann, 2010

SCoRS (Survival Count on Random Subspace)



Exploring parameters space in SCoRS

$$S = \frac{p}{2^i * n}, \text{ where } i = 4, 3, 2, 1, 0, -1, -2, -3, -4 \tag{1}$$

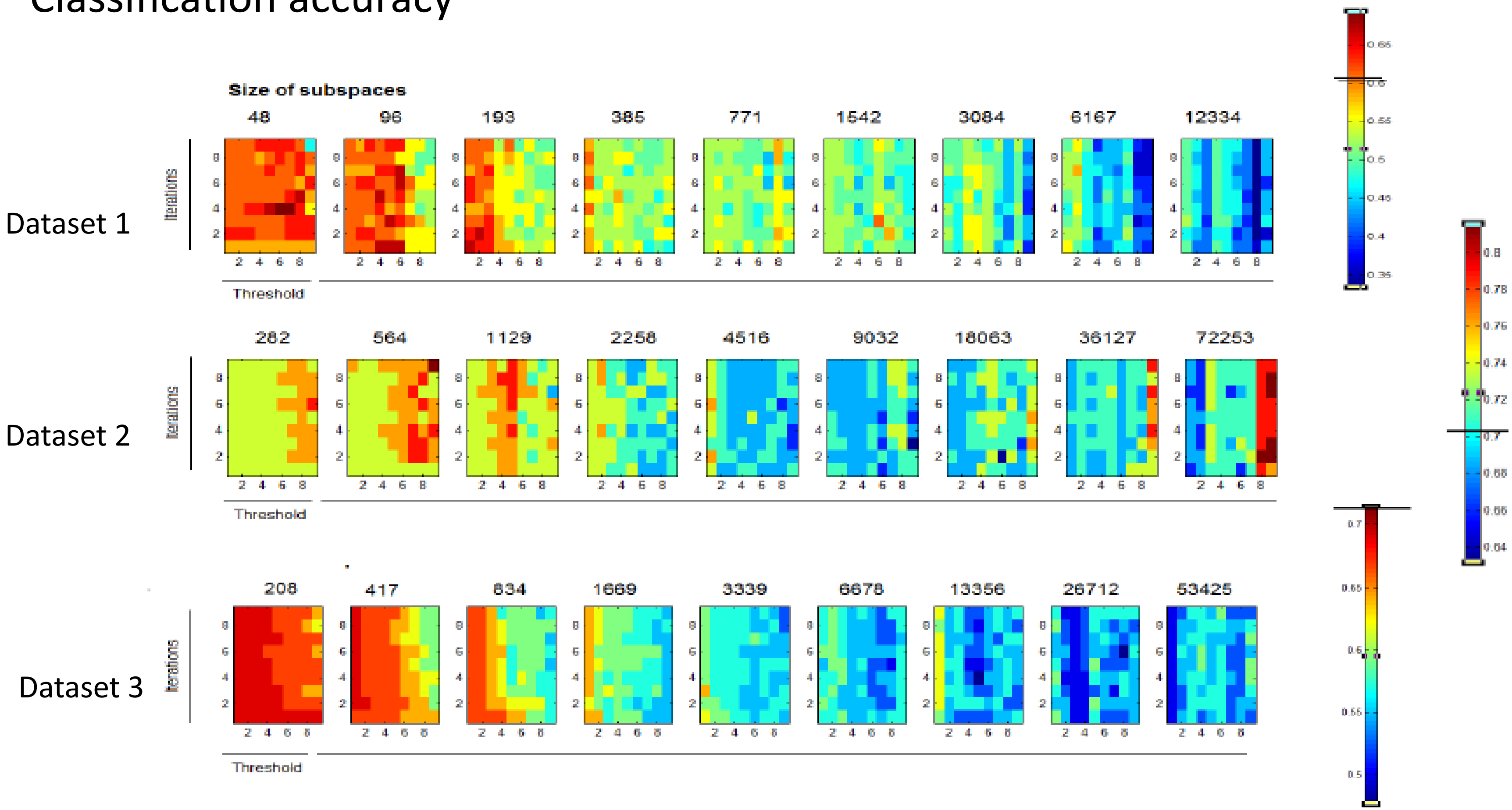
$$I = i * r, \text{where } i = 1:9 \text{ and } r \text{ was fixed as } 10^3 \tag{2}$$

$$T = i * r, \text{where } i = 1:9 \text{ and } r \text{ was fixed as } 10^{-1} \tag{3}$$

S – subspace size
I – number of iterations
T - threshold

Exploring parameters space in SCoRS

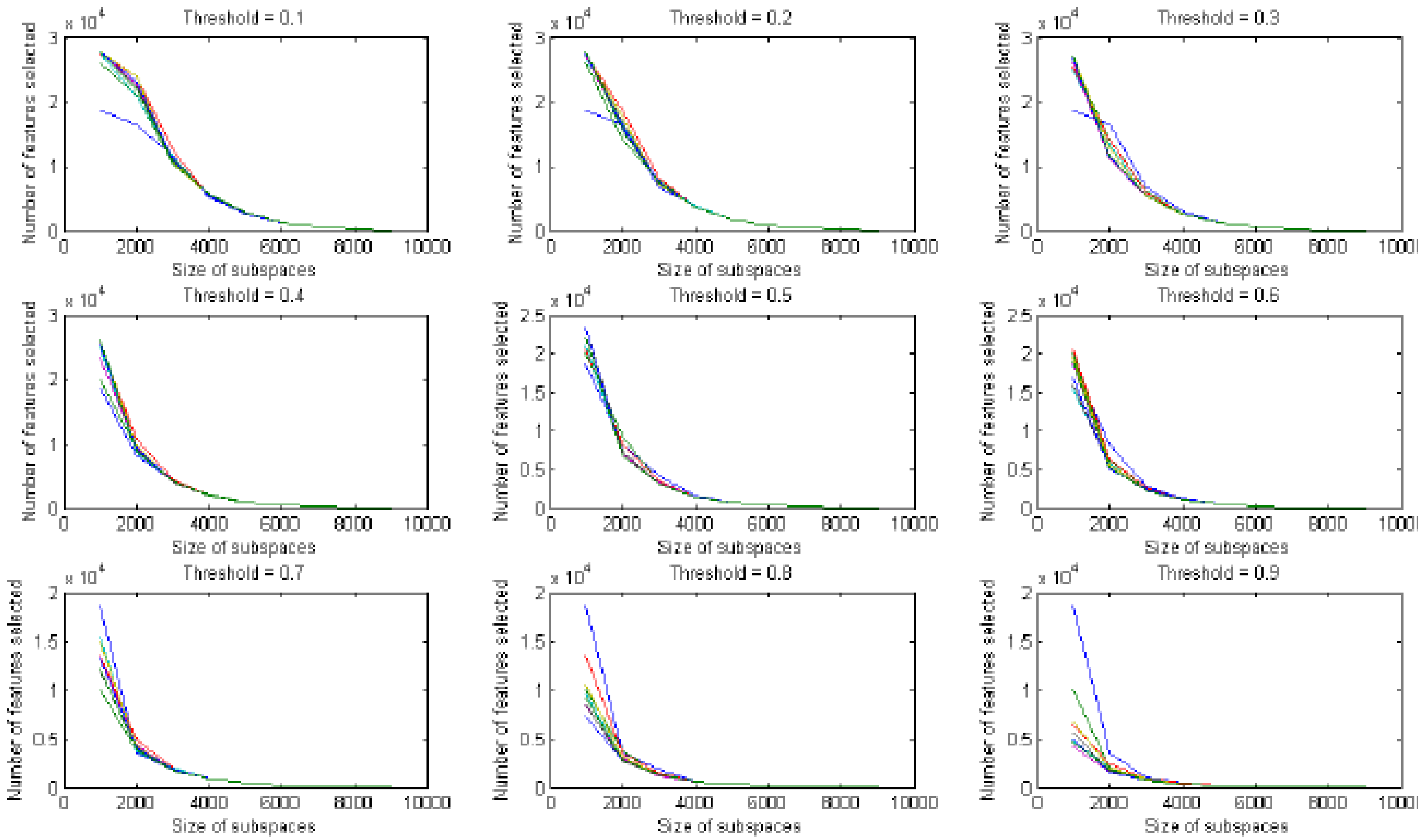
Classification accuracy



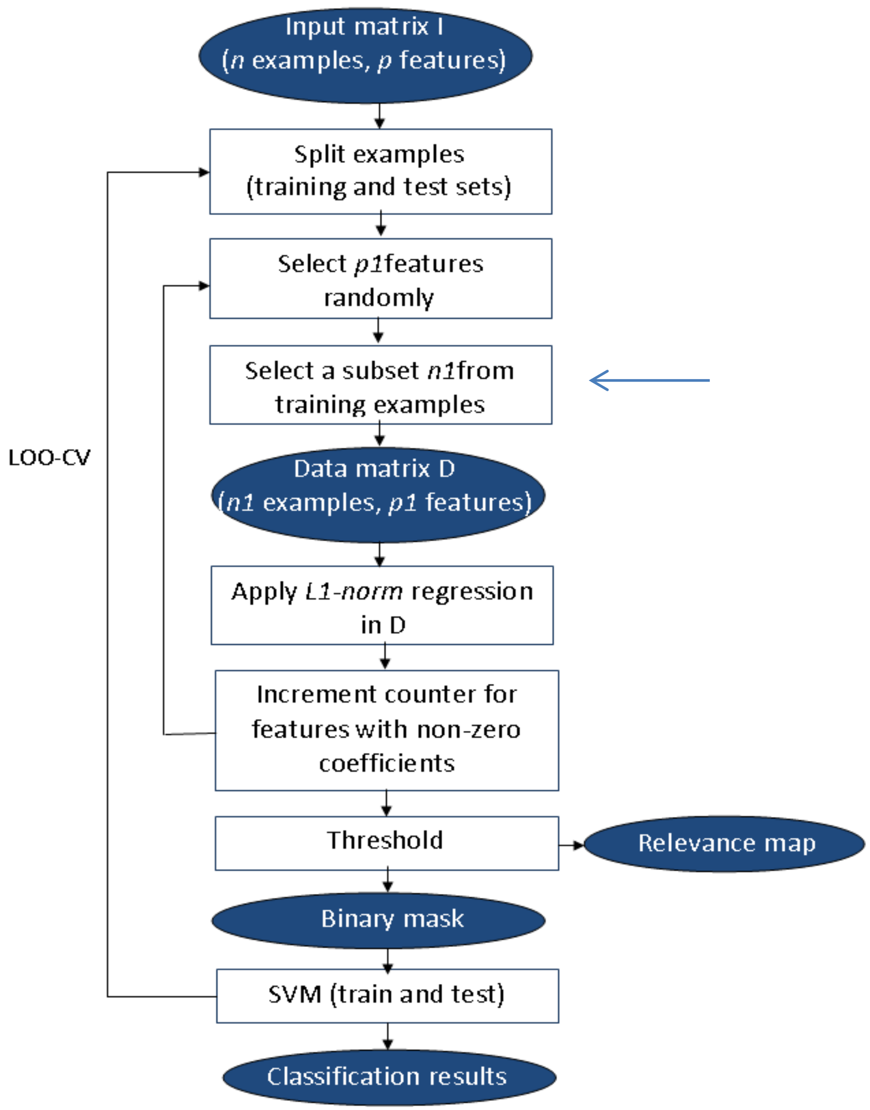
J. Rondina, J. Shawe-Taylor, and J. Mourao-Miranda, "A new feature selection method based on stability theory - exploring parameters space to evaluate classification accuracy in neuroimaging data," LNAI Survey of the state of the art Machine Learning and Interpretation of Neuroimaging, vol. 7263, pp. 58–66, 2012.

Exploring parameters space in SCoRS
(Number of features through threshold levels)

Dataset 1



B-SCoRS

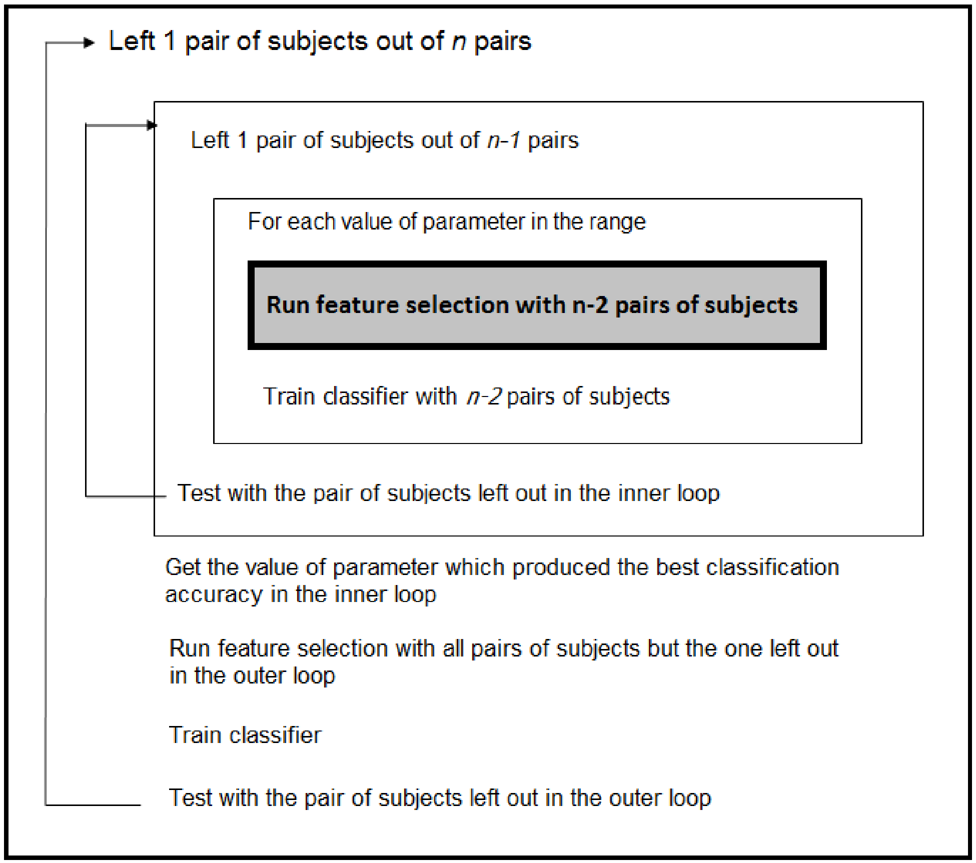

Input matrix I
(n examples, p features)
Split examples
(training and test sets)
Select p1features
randomly
Select a subset n1from
training examples
LOO-CV
Data matrix D
(n1 examples, p1 features)
Apply L1-norm regression
in D
Increment counter for
features with non-zero
coefficients
Threshold
Relevance map
Binary mask
SVM (train and test)
Classification results

Classification accuracy

Data matrix:
219,727 voxels
240 examples

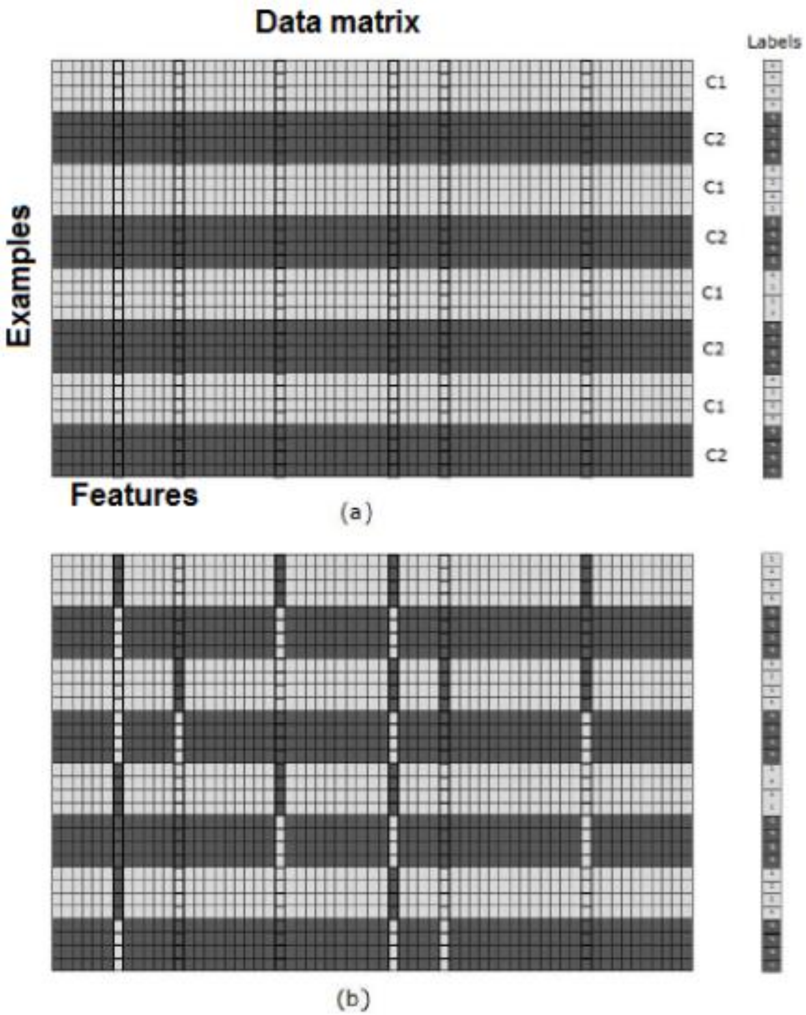
	N features	TP	TN	Acc
Whole brain	219727	0.63	0.70	0.67
No threshold	210922	0.63	0.70	0.67
T = 0.1	98738	0.67	0.73	0.70
T = 0.2	51094	0.63	0.73	0.68
T = 0.3	29958	0.63	0.73	0.68
T = 0.4	18046	0.63	0.80	0.72
T = 0.5	10704	0.67	0.80	0.74
T = 0.6	6170	0.67	0.77	0.72
T = 0.7	3265	0.67	0.77	0.72
T = 0.8	1473	0.67	0.73	0.70
T = 0.9	461	0.67	0.70	0.68

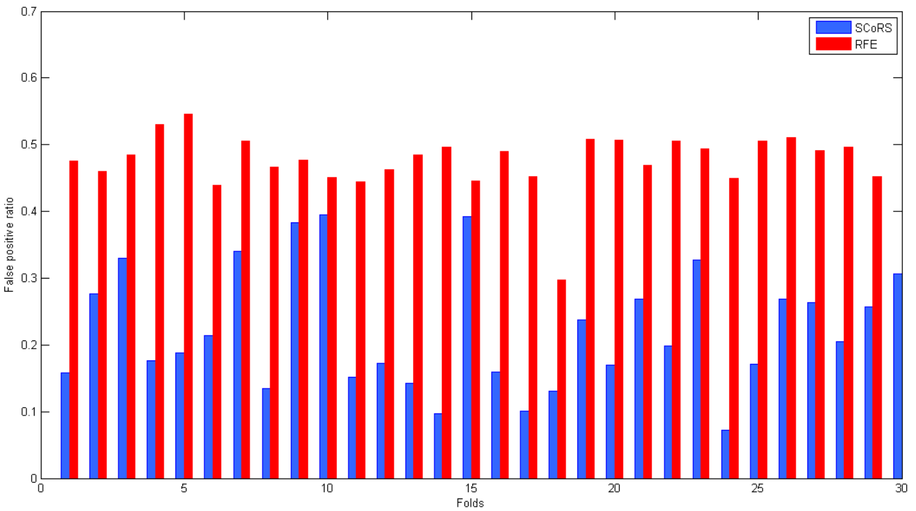
Nested CV for optimizing the number of features



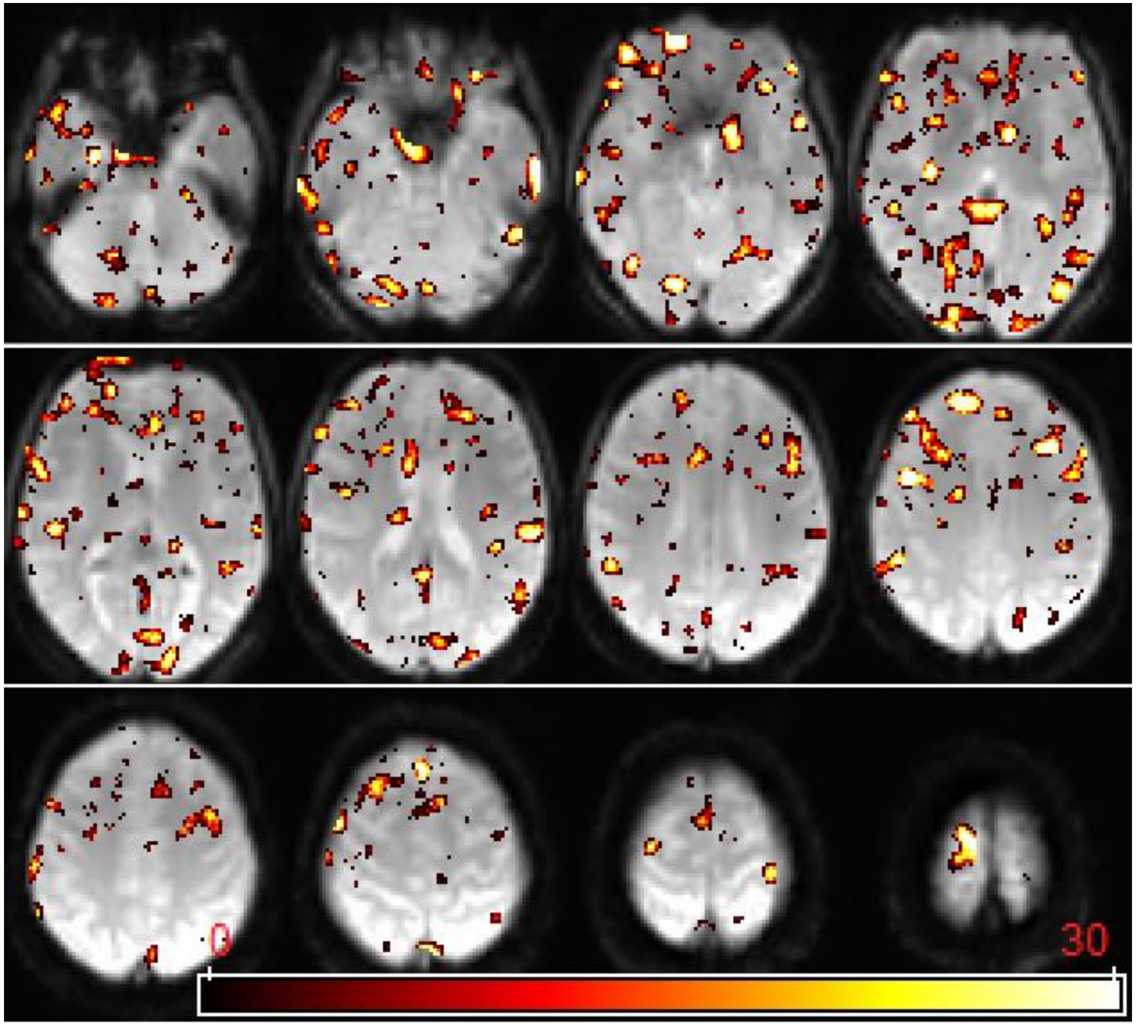
Method	N features	TP	TN	Acc
SCoRS	12006	0.67	0.77	0.72
RFE	32077	0.73	0.60	0.67

False positive selection





Spatial mapping



Thanks!