Multivariate pattern recognition analysis: brain decoding

Jessica Schrouff[1,2] and Christophe Phillips[1,2]

[1]Cyclotron Research Centre, University of Liège, Belgium

[2]Montefiore Institute of Electrical and Electronical Engineering, University of Liège, Belgium

Introduction

Two of the most fundamental questions in the field of neurosciences are how information is represented in different brain structures, and how this information evolves over time. Various tools, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), have been developed over the last few decades to record brain activity and investigate these questions. In particular, functional MRI (fMRI) tracks changes of the Blood Oxygenation Level-Dependent (BOLD) signal, which is a good indicator of brain activity (Ogawa, Lee, Kay, & Tank, 1990), with a spatial resolution of a few cubic millimeters and a typical temporal resolution in the order of 1 or 2 seconds.

Until recently, the methods used to analyze such data focused on characterizing the individual relationship between a cognitive or perceptual state and each image voxel, i.e. following a massively univariate statistical approach. A well-known univariate technique is Statistical Parametric Mapping (SPM) (Friston, Ashburner, Kiebel, Nichols, & Penny, 2007). SPM relies on the General Linear Model (Holmes, Poline, & Friston, 1997) to detect which voxels show a statistically significant response to the (combination of) experimental conditions of interest. However, there are limitations on what can be learned about the representation of information by examining voxels in a univariate fashion. For instance, spatially distributed sets of voxels considered as non-significant by a SPM analysis of one experimental condition might still carry information about the presence or absence of that condition. Furthermore, classic voxel-based analytic techniques are also mainly designed to perform group-wise comparisons and would therefore be unsuitable to evaluate the state of the disease of each individual.

On the other hand, Multi-Voxel Pattern Analyses (MVPA, see (Norman, Polyn, Detre, & Haxby, 2008), (Friston, et al., 2008) and (Haynes & Rees, 2006) for a review) allow an increased sensitivity to detect the presence of a particular mental representation. These multivariate methods, also known as brain decoding or mind reading, attempt to link a particular cognitive, behavioral, perceptual or medical state to specific patterns of voxels' activity. Application of these methods made it possible to decode the category of a seen object ((Spiridon & Kanwisher, 2002), (Cox & Savoy, 2003) and (Shinkareva, Mason, Malave, Wang, Mitchell, & Just, 2008)) or the orientation of a stripped pattern seen by the subject ((Kamitani & Tong, 2005) and (Haynes & Rees, 2005)) from the brain activation of the imaged subject. Advances in pattern-classification algorithms also allowed the decoding of less-controlled conditions such as memory retrieval tasks ((Polyn, Natu, Cohen, & Norman, 2005) and (Chadwick, Hassabis, Weiskopf, & Maguire, 2010)). Advanced mathematical tools are still under development to allow the classification of more complicated experimental data sets, such as examining the content of mind wandering or detecting the state of consciousness of a patient showing no response to a command.

Methodology

Multivariate pattern analysis derives from the fields of pattern recognition and machine learning, which are concerned with the automatic discovery of regularities in data. Those regularities then serve as the basis for the classification of new data. A classical example of pattern recognition is the automatic classification of handwritten digits (illustrated in fig.1): each digit is represented by a grey scale image of 28x28 pixels and the goal is to build an algorithm capable of classifying each image into the correct category (i.e. 0, 1,…, 9). We therefore need to build an "algorithmic machine" which will take images as inputs and produce their corresponding digit as outputs. Due to the large variability in handwritings, this operation is not trivial and the use of machine learning is necessary.

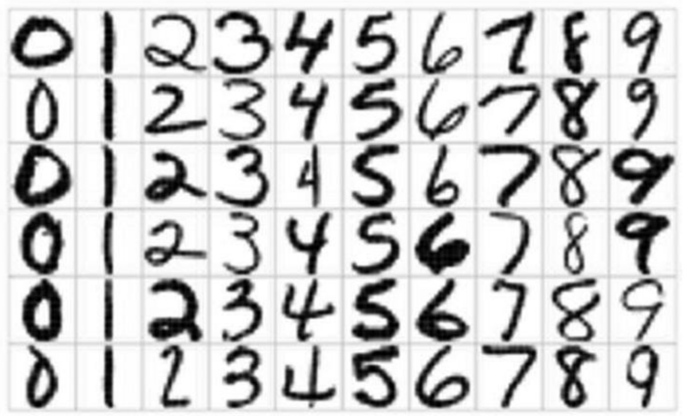

Figure 1. Examples of handwritten digits and their corresponding categories (0, 1,…, 9). Each digit is discretised as a 28x28 pixel grey scale image. Practically speaking many more than those 6 examples of handwritten digits would be necessary to build an efficient classifier.

This means that the computer has to learn which pattern in the images corresponds to which digit. This learning is achieved providing a "learning set", which is a set comprising both images (inputs) and corresponding digit (outputs). This is called "supervised learning". The machine can then build the required function using this learning set and finally assign outputs to new inputs. Similarly, instead of classifying data into discrete categories or outputs, a machine can be taught to regress out a continuous scalar from a series of inputs: after learning, the trained function can then predict the (continuous) outcome from a new input. For the rest of this chapter, we only focus on the discrete classification problem, typically into two categories.

The learning set is generally in the form of a matrix (illustrated in fig.2): each "data point" is represented by a vector, which is the collection of variables to feed in the machine, and a label, the output of the function. Usually, the data set comprises *n* observations or objects to classify, based on the values of *m* variables. The ratio between the number of objects and the number of variables that describe it plays an important role in the building of the function: if *m* is large (many variables) compared to *n* (few observations), then there is a risk of “over-fitting” (Pitt & Myung, 2002). If this dimensionality issue, also known as the “curse of dimensionality” (Bishop, 2006), is not accounted for properly, the optimized machine can achieve perfect classification on the training data but will classify poorly any new data point: the resulting classifier does not generalizes to similar but slightly different data points.

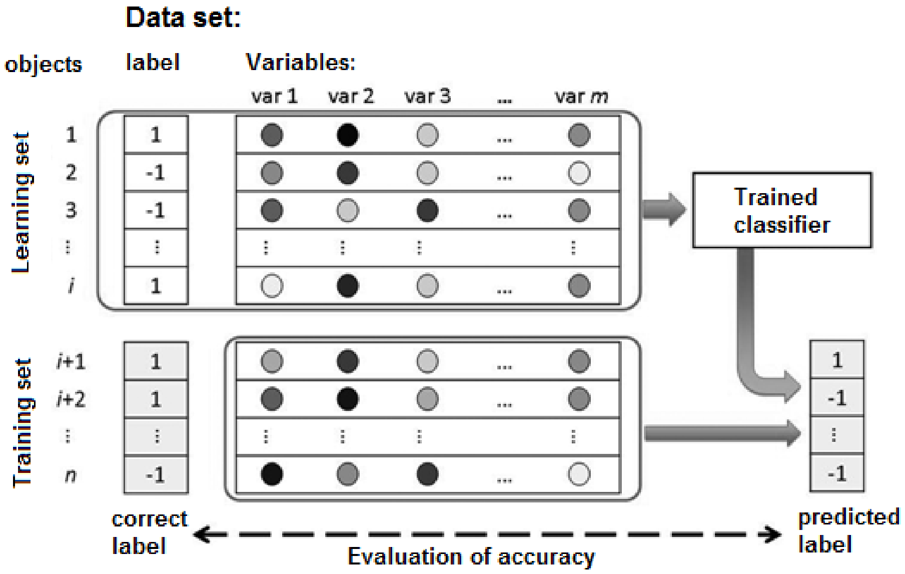

Figure 2. Graphical representation of the data set, split in training and test sets, and of the cross-validation procedure. Each of the *n* objects or data points is represented by a vector of *m* variables and its label. The training set, variables and label, is used to train the classifier. The trained classifier is then applied on the test set data and the predicted labels are compared to the true labels to assess the classifier.

To validate and assess the generalizability of the trained classifier, the data set is usually divided into two parts: the “learning set”, on which the classifier is trained, and the “testing set”, used to compute the accuracy of the classifier. Depending on the size of the data set available, this training-then-testing procedure can be repeated multiple times with different split of the data set. The final accuracy of the model is computed as the mean of the model accuracy at each division. This is referred to as a “cross-validation procedure”.

The accuracy of the model is generally computed in terms of correct or mis-classifications for each category (see table I and equation 1). For example with the handwritten digits, 80% classification accuracy for 1’s means that four times out of five the trained machine correctly recognizes the handwritten digit, and is incorrect in one case out of five, i.e. a ‘1’ is not recognized or another digit is recognized as a ‘1’. More specific parameters can also be derived, such as the sensitivity and specificity of the model (equations 2-3), which allows more insight into the characteristics of the classification algorithm. Still with the handwritten digits to classify, for example, a sensibility of 90% and specificity of 60% mean that: if a '1' is actually handwritten, then he's correctly recognized 9 times out of 10 (1 "false negative") but for any 10 other handwritten digits except a '1', 4 of them are erroneously classified as a '1' (4 "false positive").

This can be very important, for example, in the case of pathology classification: it is crucial to correctly detect all diseased patients and limit to a minimum the number of false negatives, i.e. maximize the sensitivity even if it reduces a bit the specificity.

	Class 1 (true label: -1)	**Class 2 (true label: +1)**
Class 1 (predicted label: -1)	True Negative (TN)	False Negative (FN)
Class 2 (predicted label: +1)	False Positive (FP)	True Positive (TP)

Table I: Different types of classification errors for the case of two classes. The '-1' and '+1' labels correspond to class 1 and class 2 respectively, for example the absence (-1) or presence (+1) of a disease for a test subject.

$$\text{Accuracy}= TP + TN / (TP+FP+FN+TN) \qquad \text{(eq.1)}$$

$$\text{Sensitivity}= TP/(TP+FN) \qquad \text{(eq.2)}$$

$$\text{Specificity}= TN/ (TN+FP) \qquad \text{(eq.3)}$$

There are many ways of building the input-output function, depending on how the data are mathematically modeled. Among the various existing techniques, common ones are the kernel-based approaches (Müller, Mika, Rätsch, Tsuda, & Schölkopf, 2001) such as Support Vector Machines (SVM, (Burges, 1998)), Relevant Vector Machine (RVM, (Tipping, 2001)) and Gaussian Processes (GP, (Rasmussen & Williams, 2006)), as well as other approaches like Linear Discriminant Analysis (Bishop, 2006) or Gaussian Naïve Bayes (Mitchell, et al., 2004; Friston, Ashburner, Kiebel, Nichols, & Penny, 2007).

When it comes to applying data mining tools to brain imaging data, SVM is one of the most widespread methods. This technique is based on a simple and logical idea: when discriminating between objects from different categories, the larger the distance between objects from different categories, the better the classification. SVM is mostly a binary classifier, i.e. it discriminates between only two categories of patterns, and thus aims at finding the best hyper-plane separating the data of the two categories: the margin between data points from each category and the hyper-plane is maximized (as illustrated in fig.3 for the case of 2 dimension data points). In practice, the data points located on the margin are the only ones defining the hyper-plane and are called "support vectors", hence the name "support vector machine". SVM (like RVM) is a sparse technique relying on a form of "automatic relevance determination" (Neal, 1996), i.e. the automatic selection of relevant or representative data points among the whole data set.

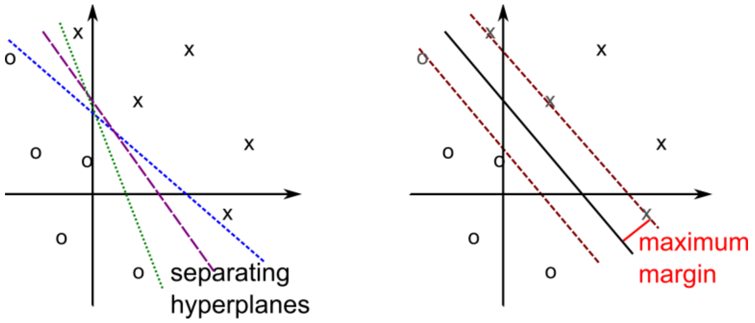

Figure 3. “Support Vector Machine” principle: The objects considered have two variables, their value being represented by both axes, and come from two categories (x and o here). The object families are separated by a hyper-plane. Among all the possible hyper-planes (3 are shown in the left image), the one that maximizes the margin (right image) is picked up by SVM. The data points on the margin are called “support vectors”.

Applications

Data mining tools can be applied to a large variety of data sets, from classification of handwritten digits to classification of brain activity induced by visualized images. The simple form of the required inputs, a vector of variables, leads to this large spectrum of applicability: all objects which can be characterized by a set of variables stacked in a vector can be classified. For example, an fMRI data set consists of a series of three-dimensional images. Each volume can therefore be considered as an observation and vectorized according to its voxels values. Volumes corresponding to different mental conditions, for example watching images of faces versus those of buildings, can then be classified. In a similar way, grey matter density extracted from structural MR images of patients and controls subjects can be classified (Vemuri, et al., 2008). Positron Emission Tomography (PET) images (Phillips, et al., 2011) and Electro- or Magneto-EncephaloGraphic (EEG or MEG) data (Chan, Halgren, Marinkovic, & Cash, 2011) have also been classified using this perspective.

Multivariate decoding of neuroimaging data can be used to achieve two different objectives: firstly and obviously predict the perceptual, cognitive or medical state of one or many subjects but also, secondly, reveal the pattern of voxels leading to the discrimination of these states. With a linear SVM (and other linear kernel machines), these two goals can be reached simultaneously: the estimated weight associated to each voxel reveals the patterns of voxels considered as important by the model to perform the classification. Or else, the trained machine can be treated as a ‘black box’ that predicts the category of any new data fed in. This last application can be viewed as a diagnostic tool in the case of a disease-versus-healthy classification (or any variation).

Diagnostic tool

The application of multivariate analysis as a diagnostic tool could become particularly useful, especially when the concerned disease is difficult to diagnose using classical clinical exams. Alzheimer’s disease (AD), one of the commonest causes of dementia, is a good example: a definitive diagnosis of AD can only be obtained using post-mortem histopathological analysis. Currently, AD is diagnosed using clinical exams, neuropsychological testing and manual measurements on brain images (MRI or PET), leading to time-consuming criteria and accuracies of the diagnosis around 80% at best (Knopman, et al., 2001). AD is therefore often misdiagnosed, although an early treatment would be more effective. Multivariate analyses performed on structural MR images of patients and controls allowed the construction of SVM, able to automatically distinguish between healthy subjects and AD patients with accuracies between 86% (Vemuri, et al., 2008) and 96% (Kloppel, et al., 2008), depending on the sample size and information used. Moreover, in Kloppel et al., 2008, data from different scanners were used, suggesting that a trained classifier could be applied across centers.

These studies relied on SVM and therefore led to binary 0/1 classifications: the input image is classified in one category or the other without further information about the classification reliability. Nevertheless, when using a multivariate technique as a diagnostic tool, knowing the probability associated with the prediction would be very valuable. Consider 2 new images, from 2 patients A and B, classified as “healthy” but the one from patient A has a probability of 99% of being so and the one from B only 51%. In both cases, under a binary classification scheme, the corresponding subjects could be considered as “healthy” but the classification probability of the B image is very close to the “50% chance threshold”, which might suggest further exams are needed to confirm or infirm the diagnostic of patient B.

“Relevance Vector Machine” (RVM), akin to SVM, provides such probabilistic prediction by returning the posterior probability of being in one category versus the other, thanks to its Bayesian formulation. RVM was recently applied on fluorodeoxyglucose PET (FDG-PET) data, i.e. images of cerebral glucose metabolism, of patients (Vegetative State, VS, and Locked-In Syndrome, LIS) and healthy controls (Phillips, et al., 2011).

First, an RVM was trained to discriminate VS patients (13 patients) from healthy subjects (37) using their FDG-PET images. With these well defined and separate categories, i.e. unconscious versus conscious subjects, cross-validation of this “consciousness classifier” showed 100% accuracy. Then the trained RVM was applied on the FDG-PET images of 8 LIS patients: the probabilities returned ranged between 61% and 100% to be in the “conscious” category. This suggests that LIS patients could be automatically and correctly classified as conscious, contrary to VS patients, based only on the patterns of their cerebral metabolism.

Relevant patterns of voxels

In the case of a linear kernel classifier, the relevance of each voxel can be estimated as a weighted linear combination of the same voxel values of the images used for training. This “voxel relevance” is not a statistical value per se but simply reflects how much any voxel contributes to the classification of the input image in to one output category or the other: a large (respectively small) value (in absolute value) indicates that this voxel has relatively much (respectively little) influence on the classification. Since there is a value per voxel, it is possible to present a “relevance map” as an image in brain space (see fig. 4 for the relevance maps of the control versus VS classification, Phillips et al., 2011). Such a map can bring insight on the location of the discriminating areas and therefore help neuroscientists build more efficient criteria of diagnosis but also orient them during the elaboration of new hypotheses concerning the origins or evolution of a certain disease or disorder.

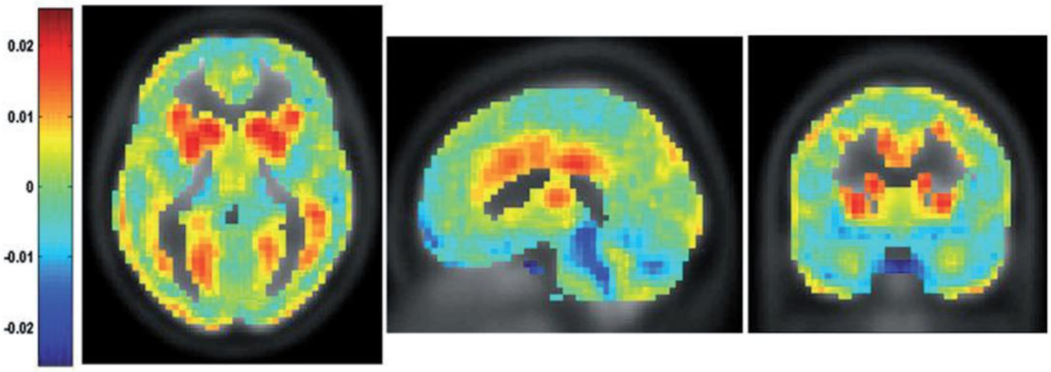

Figure 4. (Adapted from Phillips et al., 2011). Distribution over the brain volume of the voxel relevance for a « Relevance Vector Machine » trained to discriminate between FDG-PET images of VS patients and healthy subjects. A positive value (yellow-red) indicates that relatively large metabolic activity in those voxels will drive the classification towards the “healthy subject” category. Conversely, negative values (blue-purple) pushes towards the “VS patient” category. The voxels with little relevance (green) hardly contribute to the classification of data.

Perspectives

In view of the recent advances in multivariate pattern analysis, these techniques certainly will become more common to study consciousness. Their application as a diagnostic tool to differentiate patients presents obvious advantages such as objectiveness, automation and the fact that a posterior probability can be provided with the final prediction, compared to current time-consuming and subjective criteria.

Other applications could also be envisaged, for example, classification techniques could be applied to ‘response to command’ experiences in fMRI (Monti, et al., 2010), leading to reproducible user-independent and possibly more accurate results than the current “General Linear Model” analysis used. Furthermore, multivariate analysis could be applied on-line as the data are acquired, i.e. during the recording, the model is updated in real time with each new image. On-line processing of ‘response to command’ fMRI experiments offers a new communication channel relying solely on brain activation (Sorger, et al., 2009). Such “Brain Computer Interface” schemes could certainly benefit, in term of accuracy and speed, from more advanced “brain decoding” tools.

Finally, “brain reading” could maybe one day offer a partial view on the mental content of patients in altered states of consciousness. First a model would be trained with data acquired from healthy subjects thinking about different semantic categories of pictures and words, for example faces, buildings, animals or emotions (Mitchell, et al., 2008). Then this model would be applied on data from Minimally Conscious State or VS patients.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine learning.* (M. Jordan, J. Kleinberg, & B. Schölkopf, Eds.) Springer.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery , 2*, 121-167.

Chadwick, M., Hassabis, D., Weiskopf, N., & Maguire, E. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology , 20*, 1-4.

Chan, A. M., Halgren, E., Marinkovic, K., & Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *NeuroImage , 54*, 3028-3039.

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*, 261-270.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., & Penny, W. (2007). *Statistical Parametric Mapping: the analysis of functional brain images.* (K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny, Éds.) Elsevier Academic Press.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., et al. (2008). Bayesian decoding of brain images. *NeuroImage , 39*, 181-205.

Haynes, J., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci , 8*, 686-691.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci , 7*, 523-534.

Holmes, A., Poline, J.-B., & Friston, K. J. (1997). Characterizing brain images with the general linear model. Dans R. Frackowiak, K. Friston, C. Frith, R. Dolan, & J. Mazziotta, *Human Brain Function,* (pp. 59-84). Academic Press USA.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience , 8*, 679-685.

Kloppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer’s disease. *Brain , 131*, 681-689.

Knopman, D., DeKosky, S., Cummings, J., Chui, H., Corey-Bloom, J., Relkin, N., et al. (2001). Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology , 56*, 1143-1153.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., et al. (2004). Learning to Decode Cognitive States from Brain Images. *Machine Learning , 57*, 145-175.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science* , 1191-1195.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., et al. (2010). Willful modulation of brain activity in disorders of consciousness. *N Engl J Med , 362*, 579-589.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw. , 12*, 181-202.

Neal, R. (1996). *Bayesian Learning for Neural Networks.*

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2008). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *TRENDS in Cognitive Sciences , 10*, 424-430.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on bloodoxygenation. *Proc. Natl. Acad. Sci. USA , 87*, 9868-9872.

Phillips, C. L., Bruno, M.-A., Maquet, P., Boly, M., Noirhomme, Q., Schakers, C., et al. (2011). “Relevance vector machine” consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. *NeuroImage , 56*, 797-808.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *TRENDS in Cognitive Sciences , 6*, 421-425.

Polyn, S., Natu, V., Cohen, J., & Norman, K. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science , 310*, 1963-1966.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian Processes for Machine Learning.* (T. Dietterich, Ed.) the MIT Press.

Shinkareva, S., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., & Just, M. A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS one , 3*, 3:e1394.

Sorger, B., Dahmen, B., Reithler, J., Gosseries, O., Maudoux, A., Laureys, S., et al. (2009). Another kind of 'BOLD Response': answering multiple-choice questions via online decoded single-trial brain signals. In S. Laureys, *Progress in Brain research* (Vol. 177, pp. 275-292). Elsevier.

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron , 35*, 1157-1165.

Tipping, M. 2.–2. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res. , 1*, 211-244.

Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., et al. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage , 39*, 1186-1197.