Concept Classification with Bayesian Multi-task Learning

Marcel van Gerven
Radboud University Nijmegen
Intelligent Systems
Heyendaalseweg 135 6525 AJ
Nijmegen, The Netherlands
marcelge@cs.ru.nl

Irina Simanova
Max Planck Institute for Psycholinguistics
Wundtlaan 1 6525 XD
Nijmegen, The Netherlands
irina.simanova@mpi.nl

Abstract

Multivariate analysis allows decoding of single trial data in individual subjects. Since different models are obtained for each subject it becomes hard to perform an analysis on the group level. We introduce a new algorithm for Bayesian multi-task learning which imposes a coupling between single-subject models. Using the CMU fMRI dataset it is shown that the algorithm can be used for concept classification based on the average activation of regions in the AAL atlas. Concepts which were most easily classified correspond to the categories shelter, manipulation and eating, which is in accordance with the literature. The multi-task learning algorithm is shown to find regions of interest that are common to all subjects which therefore facilitates interpretation of the obtained models.

1 Introduction

Multivariate analysis allows decoding of neural representations at the single trial level in single subjects. Its introduction into the field of cognitive neuroscience has led to novel insights about the neural representation of cognitive functions such as language (Mitchell et al., 2008), memory (Hassabis et al., 2009), and vision (Miyawaki et al., 2008).

However, interpretation of the models obtained using a multivariate analysis can be hard due to the fact that different models are obtained for individual subjects. For example, when analyzing K separately acquired datasets, K sets of model parameters will be obtained which may or may not show a common pattern. In some sense, we are in need of a second-level analysis such that we can draw inferences on the group level, as in the conventional analysis of neuroimaging data using the general linear model. One way to achieve this in the context of multivariate analysis is by means of multi-task learning, a special case of transfer learning (Thrun, 1996) where model parameters for different tasks (datasets) are estimated simultaneously and no longer assumed to be independent (Caruana, 1997). In an fMRI context, multi-task learning has been explored using canonical correlation analysis (Rustandi et al., 2009).

In a Bayesian setting, multi-task learning is typically realized by assuming a hierarchical Bayesian framework where shared prior distributions condition task-specific parameters (Gelman et al., 1995). In this paper, we explore a new Bayesian approach to multi-task learning in the context of concept classification; i.e., the prediction of the semantic category of concrete nouns from BOLD response. Effectively, we are using a shared prior to induce parameter shrinkage. We show that Bayesian multi-task learning leads to more interpretable models, thereby facilitating the interpretation of the models obtained using multivariate analysis.

2 Bayesian multi-task learning

The goal of concept classification is to predict the semantic category y of a presented (and previously unseen) concrete noun from the measured BOLD response $\mathbf{x}$. In this paper, we will use Bayesian logistic regression as the underlying classification model. Let $\mathcal{B}(y;p) = p^y(1-p)^{1-y}$ denote the Bernoulli distribution and $l(x) = \log(x/(1-x))$ the logit link

Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics, pages 10–17,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

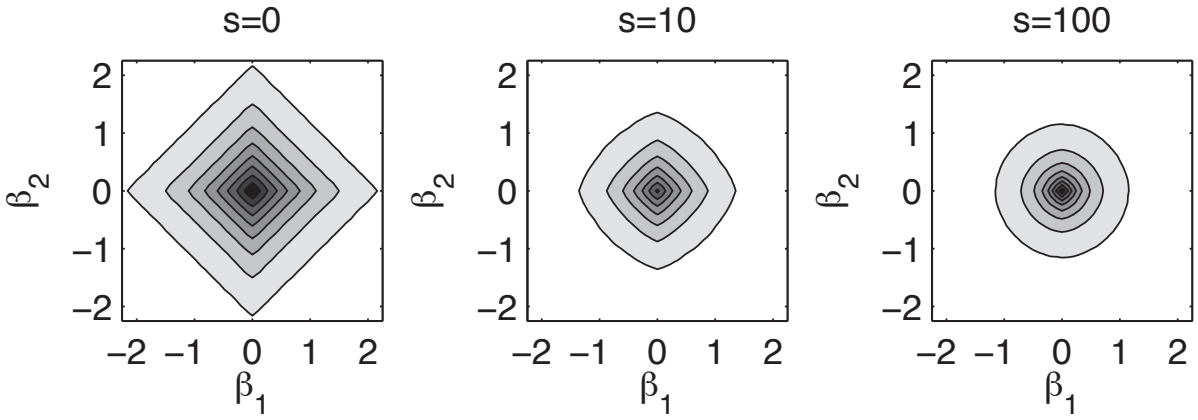


Figure 1: Contour plots of samples drawn from the prior for two regression coefficients β_1 and β_2 given three different values of the coupling strength s. For uncoupled covariates, the magnitude of one covariate has no influence on the magnitude of the other covariate. For strongly coupled covariates, in contrast, a large magnitude of one covariate increases the probability of a large magnitude in the other covariate.

function. We are interested in the following predictive density:

$$p(y \mid \mathbf{x}, \mathbf{D}, \boldsymbol{\Theta}) = \int \mathcal{B}(y; l^{-1}(\mathbf{x}^T \boldsymbol{\beta})) p(\boldsymbol{\beta} \mid \mathbf{D}, \boldsymbol{\Theta}) d\boldsymbol{\beta}$$

where we integrate out the regression coefficients $\boldsymbol{\beta}$ and condition on the response $\mathbf{x}$, observed training data $\mathbf{D} = (\mathbf{y}, \mathbf{X})$ and hyper-parameters $\boldsymbol{\Theta}$. Using Bayes rule, we can write the second term on the right hand side as

$$p(\boldsymbol{\beta} \mid \mathbf{D}, \boldsymbol{\Theta}) \propto p(\mathbf{D} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \boldsymbol{\Theta}) \tag{1}$$

where

$$p(\mathbf{D} \mid \boldsymbol{\beta}) = \prod_n \mathcal{B}(y_n; l^{-1}(\mathbf{x}_n^T \boldsymbol{\beta}))$$

is the likelihood term, which does not depend on the hyper-parameters $\boldsymbol{\Theta}$, and $p(\boldsymbol{\beta} \mid \boldsymbol{\Theta})$ is the prior on the regression coefficients.

Let $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Theta})$ denote a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Theta}$. In order to couple the tasks in a multi-task problem, we will use the multivariate Laplace prior, which can be written as a scale-mixture using auxiliary variables $\mathbf{u}$ and $\mathbf{v}$ (van Gerven et al., 2010):

$$p(\boldsymbol{\beta} \mid \boldsymbol{\Theta}) = \int \left(\prod_k \mathcal{N}(\beta_k; 0, u_k^2 + v_k^2) \right) \times \mathcal{N}(\mathbf{u}; \mathbf{0}, \boldsymbol{\Theta}) \mathcal{N}(\mathbf{v}; \mathbf{0}, \boldsymbol{\Theta}) d\mathbf{u}\, d\mathbf{v}$$

The multivariate Laplace prior allows one to control the prior variance of the regression coefficients $\boldsymbol{\beta}$ through the covariance matrix $\boldsymbol{\Theta}$ of the auxiliary variables $\mathbf{u}$ and $\mathbf{v}$. This covariance matrix is conveniently specified in terms of the precision matrix:

$$\boldsymbol{\Theta}^{-1} = \frac{1}{\theta} \mathbf{V} \mathbf{R} \mathbf{V}.$$

Here, θ is a scale parameter which controls regularization of the regression coefficients towards zero and $\mathbf{R}$ is a structure matrix where $r_{ij} = -s$ specifies a fixed coupling strength s between covariate i and covariate j. A negative r_{ij} penalizes differences between covariates i and j, see van Gerven et al. (2010) for details. $\mathbf{V}$ is a scaling matrix whose sole purpose is to ensure that the prior variance of the auxiliary variables is independent of the coupling strength.[1] Figure 1 shows the multivariate Laplace prior for two covariates and three different coupling strengths.

The specification of the prior in terms of θ and $\mathbf{R}$ promotes sparse solutions and allows the inclusion of prior knowledge about the relation between covariates. The posterior marginals for the latent variables $(\boldsymbol{\beta}, \mathbf{u}, \mathbf{v})$ can be approximated using expectation propagation (Minka, 2001) and the posterior variance of the auxiliary variables u_i (or v_i by symmetry) can be interpreted as a measure of im-

[1] $\mathbf{V}$ is a matrix with $\sqrt{\text{diag}(\mathbf{R}^{-1})}$ on the diagonal.

portance of the corresponding covariate x_i since it eventually determines how large the regression coefficients β_i can become.

Interpretation becomes complicated whenever we have collected multiple datasets for the same task since each corresponding model may give different results regarding the importance of the covariates used when solving the classification problem. Multi-task learning presents a solution to this problem by dropping the assumption that datasets $\{\mathbf{D}_1, \ldots, \mathbf{D}_K\}$ are independent. Here, this is easily realized using the multivariate Laplace prior by working with the augmented dataset

$$\mathbf{D}^* = \left(\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{bmatrix}, \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_K \end{bmatrix} \right)$$

and by assuming that each covariate is coupled between datasets. I.e., the structure matrix is given by elements

$$r_{ij} = \begin{cases} -s & \text{if } i \neq j \text{ and} \\ & (i-j) \bmod P = 0 \\ 1 + (K-1) \cdot s & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where P stands for the number of covariates in each dataset. In this way, we have coupled covariates over datasets with coupling strength s. Note that this coupling is realized on the level of the auxiliary variables and not on the regression coefficients. Hence, coupled auxiliary variables control the magnitude of the regression coefficients $\boldsymbol{\beta}$ but the $\boldsymbol{\beta}$'s themselves can still be different for the individual subjects.

3 Experiments

In order to test our approach to Bayesian multi-task learning for concept classification we have made use of the CMU fMRI dataset[2], which consists of sixty concrete concepts in twelve categories. The dataset was collected while nine English speakers were presented with sixty line drawings of objects with text labels and were instructed to think of the same properties of the stimulus object consistently during each presentation. For each concept there are six instances per subject for which BOLD response in multiple voxels was measured.

In our experiments we assessed whether previously unseen concepts from two different categories (e.g., *building-tool*) can be classified correctly based on measured BOLD response. To this end, all concepts belonging to two out of the twelve semantic categories were selected. Subsequently, we trained a classifier on all concepts belonging to these two categories save one. The semantic category of the six instances of the left-out concept were then predicted using the trained classifier. This procedure was repeated for each of the concepts and classification performance was averaged over all concepts. This performance was computed for all of the 66 possible category pairs.

In order to determine the effect of multi-task learning, results were obtained when assuming no coupling between datasets ($s = 0$) as well as when assuming a very strong coupling between datasets ($s = 100$). The scale parameter was fixed to $\theta = 1$. In order to allow the coupling to be made, all datasets are required to contain the same features. One way to achieve this is to warp the data for each subject from native space to normalized space and to perform the multi-task learning in normalized space. Here, in contrast, we computed the average activation in 116 predefined regions of interest (ROIs) using the AAL atlas (Tzourio-Mazoyer et al., 2002). ROI activations were used as input to the classifier. This considerably reduces computational overhead since we need to couple just 116 ROIs instead of approximately 20000 voxels between all nine subjects.[3] Furthermore, it facilitates interpretation since results can be analyzed at the ROI level instead of at the single voxel level. Of course, this presupposes that category-specific information is captured by the average activation in predefined ROIs, which is an important open question we set out to answer with our experiments.

4 Results

4.1 Classification of category pairs

We achieved good classification performance for many of the category pairs both with and with-

[2] http://www.cs.cmu.edu/∼tom/science2008

[3] The efficiency of our algorithm depends on the sparseness of the structure matrix $\mathbf{R}$.

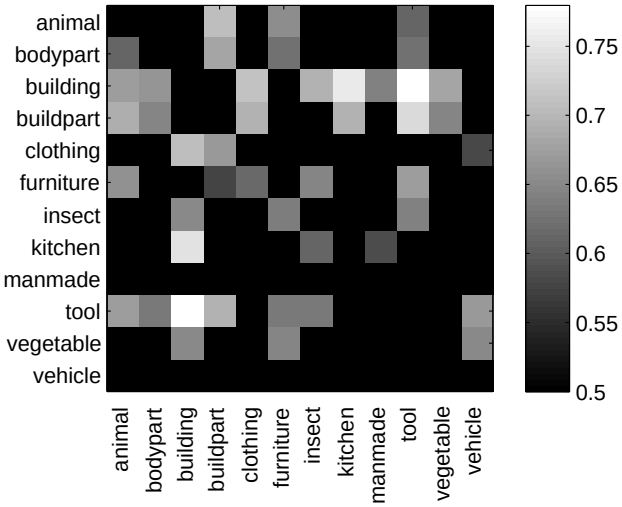


Figure 2: Accuracies for concept classification of the 66 category pairs. The upper triangular part shows the results of multi-task learning whereas the lower triangular part shows the results of standard classification. Non-significant outcomes have been masked (Wilcoxon rank sum test on outcomes for all nine subjects, p=0.05, Bonferroni corrected).

out multi-task learning. Figure 2 shows these results where non-significant outcomes have been masked. Interestingly, outcomes for all subjects showed a preference for particular category pairs. The concepts from *building-tool*, *building-kitchen* and *buildpart-tool* had the highest mean classification accuracies (proportion of correctly classifier trials) of 0.78, 0.76 and 0.74, closely followed by concepts from *building-clothing* and *animal-buildpart* with a mean classification accuracy of 0.71.

This result bears a strong resemblance to the recent work of Just et al. (2010). The authors conducted a factor analysis of fMRI brain activation in response to presentations of written words of different categories and discovered three semantic factors with the highest predictive potential: manipulation, eating and shelter-entry. They subsequently used these factors to select voxels for a features set and were able to accurately identify the activation generated by concrete word using multivariate learning methods on the basis of selected voxels. Moreover, using the factor-related activation profiles they were able to identify common neuronal signatures for particular words across participants. The authors suggest the revealed factors to represent major semantic dimensions that relate to the ways the human being can interact with an object. Although they assume the existence of other semantic attributes that determine conceptual representation, the factors shelter, manipulation and eating are proposed to be dominant for the particular set of nouns. It is easy to draw an analogy as the set of words used by Just and colleagues was exactly the same as in the current study. Although the taxonomic categorization used in our study does not exactly match the factor-based categorization, most of the items from categories *building*, *buildpart*, *tool* and *kitchen* show a strong correspondence with one of the semantic factors and are listed among ten words with highest factor scores according to Just et al. (2010) (see Table 1).

Table 1: Stimulus words from the semantic categories that showed best classification accuracies. Superscripts indicate the words belonging to the list of ten words with highest factor scores in the study by Just et al. (Just, 2010). We use the following abbreviations: s = shelter, m = manipulation, e = eating.

Building	**Buildpart**	**Tool**	**Kitchen**
apartment^{s}	window	chisel^{m}	glass^{e}
barn	door^{s}	hammer^{m}	knife^{m}
house^{s}	chimney	screwdriver^{m}	bottle
church^{s}	closet^{s}	pliers^{m}	cup^{e}
igloo	arch	saw^{m}	spoon^{m}

The subsets of items that are set far apart in the suggested semantic dimensions appear to be preferred by the classifier in our study. The classifier was not able to identify the category of an unseen concept in pairs *building-buildpart* and *tool-kitchen*, possibly since they these categories shared the same semantic features. Thus, the current study brings an independent corroboration for the finding on the semantic dimensions underlying concrete noun representation.

4.2 Single versus multi-task learning

The use of AAL regions instead of native voxel activity patterns allowed efficient multi-task learning by coupling each region between nine subjects. Reliable classification accuracies were obtained for all the participants, although there were strong differences in individual performances (Fig. 3). The move

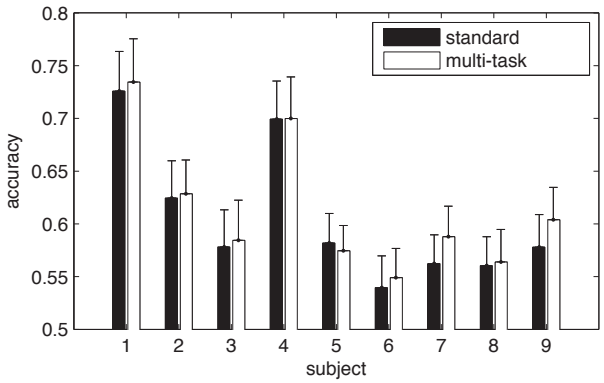


Figure 3: Classification performance per subject averaged over all category pairs for standard classification and multi-task learning (error bars show standard error of the mean).

to multi-task learning seems to improve classification results slightly in most of subjects, although the improvement is not significant.

The main outcome and advantage of our approach to multi-task learning is the convergence of models obtained from different subjects. Figure 4 shows that the subject-specific models become strongly correlated when they are obtained in the multi-task setting, even for weak coupling strengths. For strong coupling strengths, the models are almost perfectly correlated, resulting in identical models for all the nine subjects as shown in Fig. 4 for the category pair *building-tool*. It is important to realize here that the model is defined in terms of the variance of the auxiliary variables, which acts as a proxy to the importance of a region. At the level of the regression coefficients β, the model will still find subject-specific parameters due to the likelihood term in Eq. (1). Even though the contribution of each brain region is constrained by the induced coupling, this does not impede but rather improve classification performance. This fact entitles us to believe that our approach to multi-task learning tracks down the common task-specific activations while ignoring background noise.

Our study demonstrates that Bayesian multi-task learning allows generalization across subjects. Our algorithm identifies identical cortical locations as being important in solving the classification problem for all individuals within the group. The identified regions agree with previously published results on concept encoding. For example, the regions which were considered important for the category pair *building-tool* (Fig. 5) are almost indistinguishable from those described in a recent study by Shinkareva et al. (2008). These are regions that are traditionally considered to be involved in reading, objects meaning retrieval and visual semantic tasks (Vandenberghe et al., 1996; Phillips et al., 2002).

Strikingly, very similar regions were picked by the classifier for the other two category pairs with high classification accuracy, i.e., *building-kitchen* and *buildpart-tool*. This fact brings back the issue about the semantic factors relevant for the discrimination of the entities from these categories. The factors shelter, manipulation and eating are associated with the concepts from the first three addressed category pairs. The locations of voxel clusters associated with the semantic factors in (Just et al., 2010) match the brain regions that contributed to the classification for the three most optimal pairs in our experiment. In the Just et al. study these were left and right fusiform gyri, left and right precuneus and left inferior temporal gyrus for shelter, left supramarginal gyrus, left postcentral gyrus and left inferior temporal gyrus for manipulation and left inferior frontal gyrus, left middle/inferior frontal gyri, and left inferior temporal gyrus for eating. The occipital lobes detected exclusively in our experiment might be explained by the fact that in our experiment the subjects were viewing picture-text pairs in contrast to only text in (Just et al., 2010).

5 Discussion

We have demonstrated that Bayesian multi-task learning can be realized through Bayesian logistic regression when using a multivariate Laplace prior that couples features between multiple datasets. This approach has not been used before and yields promising results. As such it complements other Bayesian and non-Bayesian approaches to multi-task learning such as those reported in (Yu et al., 2005; Dunson et al., 2008; Argyriou et al., 2008; van Gerven et al., 2009; Obozinski et al., 2009; Rustandi et al., 2009).

Results show that many category pairs can be classified based on the average activation of regions

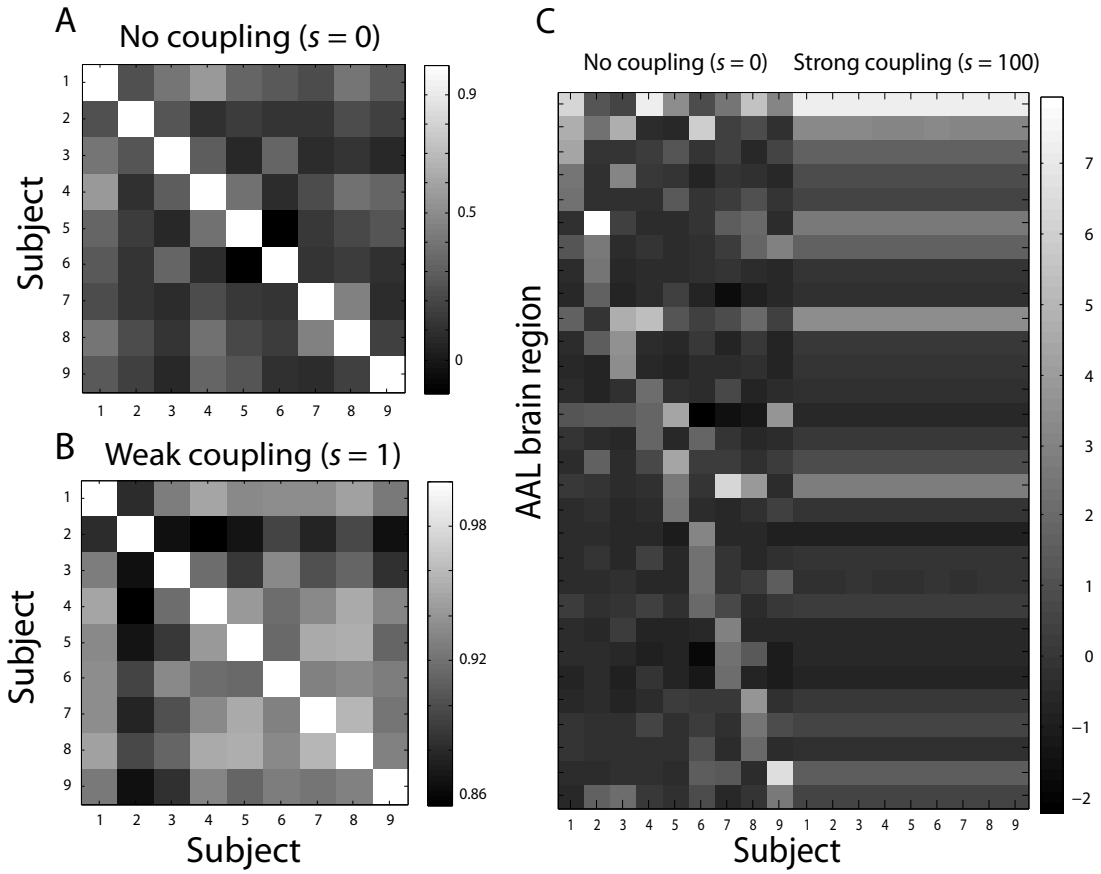


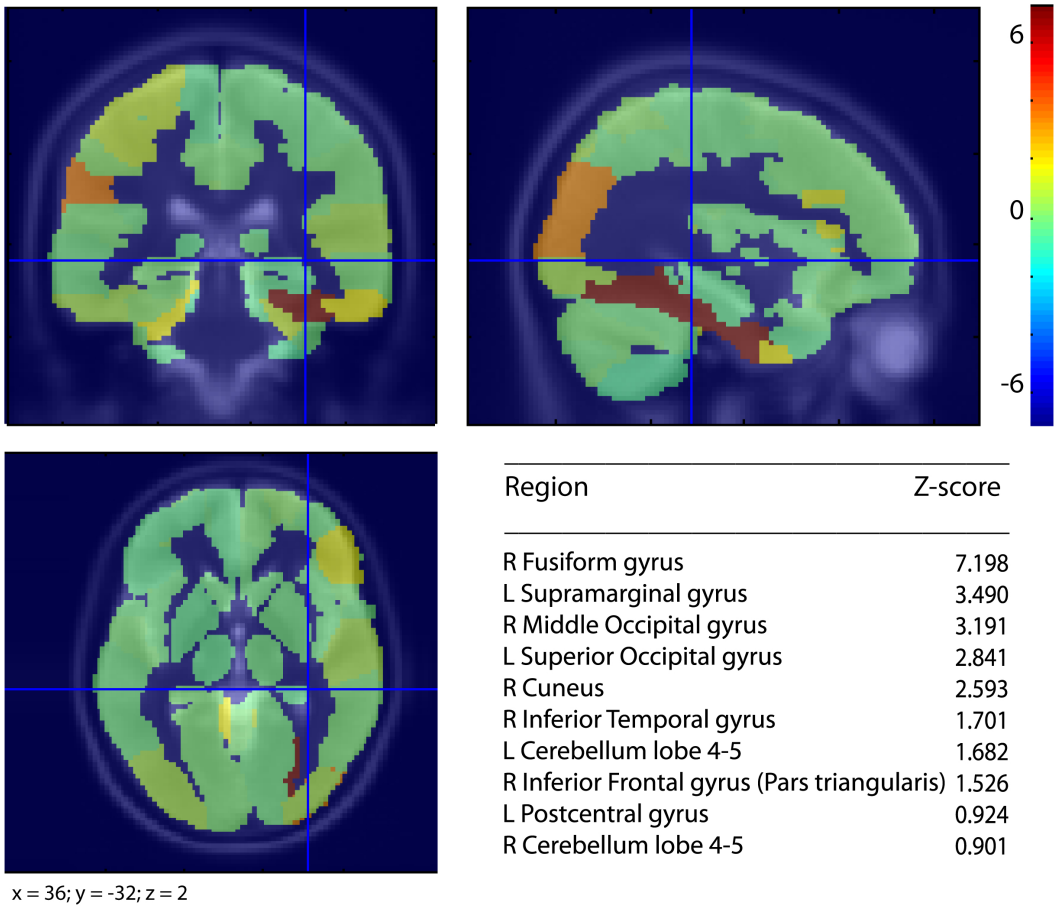
Figure 4: Correlation matrices for subject-specific models for standard classification (A) and multi-task learning (B) with weak coupling (s=1) for *building* versus *tool*. The right panel (C) shows the difference between the obtained models for standard classification and strong coupling (s=100) for the thirty most important AAL regions.

in the AAL template. Although obtained accuracies are lower than those which would have been obtained using single-voxel activations, it is interesting in its own right that the activation in just 116 predefined regions still allows concept decoding. However, it remains an open question to what extent classifiability truly reflects semantic processing instead of sensory processing of words and/or pictures.

The coupling induced by multi-task learning leads to interpretable models when using auxiliary variable variance as a measure of importance. The obtained models for the pairs which were easiest to classify corresponded well to the results reported in (Shinkareva et al., 2008) and mapped nicely onto the semantic features *shelter*, *manipulation* and *eating* identified in (Just et al., 2010).

In this paper we used the multivariate Laplace prior to induce a coupling between tasks. It is straightforward to combine this with other coupling constraints such as coupling nearby regions within subjects. Our algorithm also does not preclude multi-task learning on thousands of voxels. Computation time depends on the number of non-zeros in the structure matrix $\mathbf{R}$ and matrices containing hundreds of thousands of non-zero elements are still manageable with computation time being in the order of hours.

Another interesting application of multi-task learning in the context of concept learning is to couple the datasets of all condition pairs within a subject. This effectively tries to find a model where used regions of interest can predict multiple condition pairs. The correlation structure between the models for each condition pair then informs about their sim-



Region	Z-score
R Fusiform gyrus	7.198
L Supramarginal gyrus	3.490
R Middle Occipital gyrus	3.191
L Superior Occipital gyrus	2.841
R Cuneus	2.593
R Inferior Temporal gyrus	1.701
L Cerebellum lobe 4-5	1.682
R Inferior Frontal gyrus (Pars triangularis)	1.526
L Postcentral gyrus	0.924
R Cerebellum lobe 4-5	0.901

Figure 5: The brain regions contributing to the identification of *building* versus *tool* categories.

ilarity. An interesting direction for future research is to perform multi-task learning on the level of the semantic features that define a concept instead of on the concepts themselves. If we are able to predict the semantic features reliably then we may be able to predict previously unseen concepts from their constituent features (Palatucci et al., 2009).

References

A. Argyriou, T Evgeniou, and M. Pontil. 2008. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272.

R. Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

D. Dunson, Y. Xue, and L. Carin. 2008. The matrix stick-breaking process: flexible Bayes meta analysis. *Journal of the American Statistical Association*, 103(481):317–327.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian Data Analysis*. Chapman and Hall, London, UK, 1st edition.

D. Hassabis, C. Chu, G. Rees, N. Weiskopf, P. D. Molyneux, and E. A. Maguire. 2009. Decoding neuronal ensembles in the human hippocampus. *Current Biology*, 19:546–554.

M. A. Just, V. L. Cherkassky, S. Aryal, and T. M. Mitchell. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, 5(1):e8622.

T. Minka. 2001. Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller, editors, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani.

2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929.

G. Obozinski, B. Taskar, and M. I. Jordan. 2009. Joint covariate selection and joint subspace selection for multiple classification problems. In *Statistics and Computing*. Springer.

M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. 2009. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Neural Information Processing Systems*, pages 1410–1418.

J. A. Phillips, U. Noppeney, G. W. Humphreys, and C. J. Price. 2002. Can segregation within the semantic system account for category-specific deficits? *Brain*, 125(9):2067–2080.

I. Rustandi, M. A. Just, and T. M. Mitchell. 2009. Integrating multiple-study multiple-subject fMRI datasets using canonical correlation analysis. In *Proceedings of the MICCAI 2009 Workshop*.

S. V. Shinkareva, R. A. Mason, V. L. Malave, W. Wang, and T. M. Mitchell. 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE*, 3(1):e1394.

S. Thrun. 1996. Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems*, pages 640–646. The MIT Press.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289.

M. A. J. van Gerven, C. Hesse, O. Jensen, and T. Heskes. 2009. Interpreting single trial data using groupwise regularisation. *NeuroImage*, 46:665–676.

M. A. J. van Gerven, B. Cseke, F. P. de Lange, and T. Heskes. 2010. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50(1):150–161.

R. Vandenberghe, C. Price, R. Wise, O. Josephs, and R. S. Frackowiak. 1996. Functional anatomy of a common semantic system for words and pictures. *Nature*, 383(6597):254–256.

K. Yu, V. Tresp, and A. Schwaighofer. 2005. Learning Gaussian processes from multiple tasks. In *International Conference on Machine Learning*, pages 1012–1019.