

Multi-Class Sparse Bayesian Regression for Neuroimaging data analysis

Vincent Michel^{1,2,5}, Evelyn Eger^{3,5}, Christine Keribin^{2,4}, and Bertrand Thirion^{1,5}

¹ Parietal team, INRIA Saclay-Île-de-France, Saclay, France ,

² Université Paris-Sud 11, Orsay, France

³ INSERM U562, Gif/Yvette, France

⁴ Select team, INRIA Saclay-Île-de-France, France

⁵ CEA, DSV, I2BM, Neurospin, Gif/Yvette, France

Abstract. The use of machine learning tools is gaining popularity in neuroimaging, as it provides a sensitive assessment of the information conveyed by brain images. In particular, finding regions of the brain whose functional signal reliably predicts some behavioral information makes it possible to better understand how this information is encoded or processed in the brain. However, such a prediction is performed through regression or classification algorithms that suffer from the curse of dimensionality, because a huge number of features (i.e. voxels) are available to fit some target, with very few samples (i.e. scans) to learn the informative regions. A commonly used solution is to regularize the weights of the parametric prediction function. However, model specification needs a careful design to balance adaptiveness and sparsity. In this paper, we introduce a novel method, *Multi-Class Sparse Bayesian Regression (MCBR)*, that generalizes classical approaches such as Ridge regression and Automatic Relevance Determination. Our approach is based on a grouping of the features into several classes, where each class is regularized with specific parameters. We apply our algorithm to the prediction of a behavioral variable from brain activation images. The method presented here achieves similar prediction accuracies than reference methods, and yields more interpretable feature loadings.

1 Introduction

Machine learning approaches in neuroimaging have traditionally been limited to diagnostic problems, where patients were classified into different groups based on anatomical or functional data; by contrast, the standard framework for functional or anatomical brain mapping was based on mass univariate inference procedures. Recently, a new way of analyzing neuroimaging data has emerged, that consists in assessing how well behavioral information or cognitive states can be predicted from brain activation images such as those obtained with functional Magnetic Resonance Imaging (fMRI); see e.g. [5]. This approach opens new ways to understanding the mental representation of various perceptual and cognitive parameters. The accuracy of the prediction of the behavioral or cognitive target

variable, as well as the spatial layout of predictive regions, can provide valuable information about functional brain organization; in short, it helps to *decode* the brain system [6]. The main difficulty in this procedure is that there are far more features than samples, which leads to overfitting and poor generalization. In such cases, the use of the *kernel trick* is known to yield good performance, but the corresponding predictive feature maps are hard to interpret, because the predictive function is not sparse in the primal space (voxels space). Another way to deal with this issue is to use approaches such as feature selection or dimension reduction. However, it is suboptimal to perform feature selection and parameter estimation procedure separately, and there is a lot of interest in methods that perform both simultaneously, as sparsity inducing penalizations [12].

Let us introduce the following regression model :

$$y = \Phi w + \epsilon$$

where y represents the target data ($y \in \mathbb{R}^n$) and w the parameters ($w \in \mathbb{R}^m$). m is the number of features (or voxels) and Φ is the design matrix ($\Phi \in \mathbb{R}^{n \times m}$, each row is an m -dimensional sample). The crucial issue here is that $n \ll m$, so that estimating w is an ill-posed problem. One way to perform the estimation of w is to penalize the ℓ_2 norm of the weights. This requires the amount of penalization to be fixed beforehand, and possibly optimized by cross-validation. Bayesian regression techniques can be used instead to include regularization parameters in the estimation procedure, as penalization by weighted ℓ_2 norm is equivalent to setting Gaussian priors on the weights :

$$w \sim \mathcal{N}(0, A^{-1}), A = \text{diag}(\alpha_1, \dots, \alpha_m) \quad (1)$$

Bayesian Ridge Regression (BRR) [1] corresponds to the particular case $\alpha_1 = \dots = \alpha_m$, i.e. all the weights are regularized identically. *BRR* is not well-suited for datasets where only few sets of features are truly informative. *Automatic Relevance Determination (ARD)* [10] is the particular case where $\alpha_i \neq \alpha_j$ if $i \neq j$, i.e. all the weights have a specific regularization parameter. However, by regularizing separately each feature, *ARD* is prone to overfitting when the model contains too many regressors [9]. In order to cope with the drawbacks of *BRR* and *ARD*, we can group the features into different classes, and thus regularize these classes differently. This is the main idea behind the *group Lasso* (ℓ_{21} norm) [13]. However, *group Lasso* needs pre-defined classes and is thus not applicable in most standard situations, in which classes are not available beforehand; defining them arbitrarily is not consistent with a bias free search of predictive features. Thus, the different classes underlying the regularization have to be estimated from the data. In this paper, we develop an intermediate approach for sparse regularized regression, which assigns voxels to one among K classes. Regularization is performed in each class separately, leading to a stable and adaptive regularization, while avoiding overfit. This approach, called *Multi-Class Sparse Bayesian Regression (MCBR)*, is thus an intermediate between *BRR* and *ARD*. It reduces the overfitting problem of *ARD* in large dimension settings without the use of kernels, and is far more adaptive than *BRR*. The closest work to our approach is the Bayesian regression detailed in [8], but the

construction relies on ad hoc voxel selection steps, so that there is no proof that the solution is optimal. After introducing our model and giving some details on the parameter estimation algorithm (Gibbs sampling procedure), we show that the proposed algorithm yields similar accuracy as reference methods, and provides more interpretable weights maps.⁶

2 Model and Algorithm

Multi-Class Sparse Bayesian Regression We use classical priors for regression, see[1, 10]. First, we model the noise as an i.i.d. Gaussian variable:

$$\epsilon \sim \mathcal{N}(0, \lambda^{-1} I_n) \quad (2)$$

$$p(\lambda) = \Gamma(\lambda_1, \lambda_2) \quad (3)$$

where Γ stands for the gamma density with two hyper-parameters λ_1, λ_2 . In order to combine the sparsity of *ARD* with the stability of *BRR*, we introduce an intermediate representation, in which each feature i belongs to one class among K indexed by a discrete variable z_i . All the features within a class $k \in \{1, \dots, K\}$ share the same precision parameter α_k . We use the following prior on the z variable :

$$p(z) = \prod_{i=1}^m \prod_{k=1}^K \pi_k^{\eta_{ik}} \text{ with } \begin{cases} \eta_{ik} = 0 & \text{if } z_i \neq k \\ \eta_{ik} = 1 & \text{if } z_i = k \end{cases} \quad (4)$$

We introduce an additional Dirichlet prior on π , $p(\pi) = Dir(\delta)$, with hyper-parameter δ . By updating at each step the probabilities π_k of each class, the sampling algorithm can prune classes. As in Eq. (1), we make use of an independent Gaussian prior for the weights :

$$w \sim \mathcal{N}(0, A^{-1}), A = \text{diag}(\alpha_{z_1}, \dots, \alpha_{z_m}) \quad (5)$$

$$p(\alpha_k) = \Gamma(\gamma_1^k, \gamma_2^k), k = 1, \dots, K \quad (6)$$

where $\alpha_k, k \in \{1, \dots, K\}$ are the precision parameters, each one having two hyper-parameters γ_1^k, γ_2^k . The complete generative model of *MCBR* is summarized in Fig.1. We have developed a Gibbs sampling procedure to estimate the parameters of our model (due to lack of space, the conditional distributions are not detailed in this paper). The link between this model and other regularization methods is obvious : with $K = 1$, we retrieve the model of the *BRR*, and with $K = m$ and fixing $p(z) = \prod_{i=1}^m \delta_{z_i, i}$, we retrieve *ARD* regularization.

Initialization and priors on the model parameters Our model needs few hyper-parameters; we choose here to use slightly informative and class-specific hyper-parameters in order to reflect a wide range of possible behaviors for the weights distribution. We set $K = 9$, with weakly informative priors

⁶ Supplementary material can be found at <http://parietal.saclay.inria.fr/research/decoding-and-modelling-of-brain-function-with-fmri/misc/supp-mat.pdf/view>

$\gamma_1^k = 10^{k-3}, k \in \{1, \dots, K\}$ and $\gamma_2^k = 10^{-2}, k \in \{1, \dots, K\}$. Moreover, we set $\lambda_1 = \lambda_2 = 1$. Starting with a given number of classes and letting the model automatically prune the classes, can be seen as a means to avoid costly model selection procedures. The number of iterations used in the Gibb sampling is fixed to 1000 in all our experiments. Results on both simulated and real data (not shown), show that this number allows the algorithm to reach a stationary distribution.

Reference methods and evaluation procedure *Multi-Class Sparse Bayesian Regression* is compared to different methods :

- *Bayesian Ridge Regression* (or *BRR*), which is simply *MCBR* with $K = 1$.
- *ARD* regularization on regression. We work in the primal space, hence we do not use a kernel approach in our experiments. This method does not need any parameter optimization.
- the *Elastic net* (or *Enet*) approach [14, 2], which is a combined ℓ_1 and ℓ_2 regularization. This method requires a double optimization for the two parameters λ (amount of ℓ_2 regularization) and s (fraction of the ℓ_1 norm). We use a cross-validation loop within the training set to optimize them. The values are in the range 10^{-3} to 10^3 in multiplicative steps of 10 for λ , and in the range 0 to 1 in steps of 0.1 for s .
- *Support Vector Regression* (or *SVR*) with a linear kernel (see [4]), which is the reference method in neuroimaging, due to its robustness in large dimension. The C parameter is optimized by cross-validation in the range 10^{-3} to 10^3 in multiplicative steps of 10.

The performance of the different regression models is evaluated using ζ , the ratio of explained variance (or R^2 coefficient):

$$\zeta(\Phi^l, y^l, \Phi^t, y^t) = \frac{\text{var}(y^t) - \text{var}(y^t - \hat{y}^t)}{\text{var}(y^t)} \quad (7)$$

where Φ^l, y^l are a learning set, Φ^t, y^t a test set and \hat{y}^t refer to the target predicted using the learning set. This is the amount of variability in the response that can be explained by the model (perfect prediction yields $\zeta = 1$, while $\zeta < 0$ if prediction is worse than chance).

3 Experiments and Results

We have performed some simulations, where a combination of signals from several regions in smooth images is correlated to some target information. Due to lack of place, we do not show the results here, but provide them as supplementary material. We observed that:

- the *MCBR* outperforms other methods, and recovers correct feature maps.
- using informative and class-dependent priors yield higher accuracy than identical priors. A decrease of 0.3 in explained variance is observed when using identical priors for all the classes.

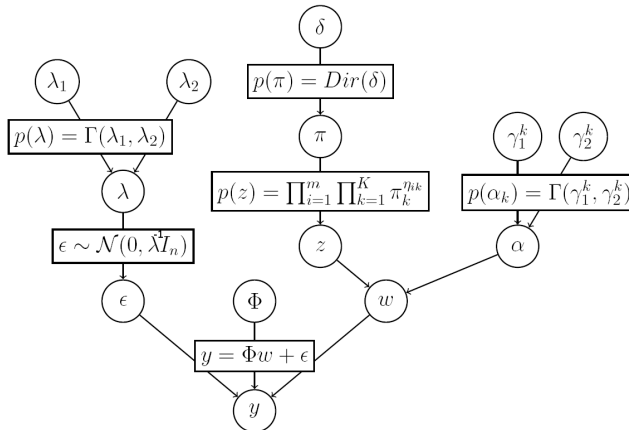


Fig. 1. Generative model of the *Multi-Class Sparse Bayesian Regression*.

Experiments on Real Data We used a real dataset related to an experiment on the representation of objects, described precisely in [7]. During the experiment, ten healthy volunteers viewed objects of three different sizes and four different shapes, with 4 repetitions of each stimulus in each one of 6 sessions, resulting in a total of $n = 72$ images by subject. Functional images were acquired on a 3-T MR system with eight-channel head coil (Siemens Trio, Erlangen, Germany) as T2*-weighted echo-planar image (EPI) volumes. Twenty transverse slices were obtained with a repetition time of 2 s (echo time, 30 ms; flip angle, 70°; $2 \times 2 \times 2$ -mm voxels; 0.5-mm gap). Realignment, normalization to MNI space and General Linear Model (GLM) fit were performed with the SPM5 software. For our analysis we used the resulting session-wise parameter estimate images. The four different shapes of objects are pooled across the three sizes, and we are interested in discrimination between sizes. This can be handled as a regression problem, where we aim at predicting the size of an object corresponding to an fMRI scan. We used parcellation as a preprocessing, which allows important unsupervised reduction of the feature space dimension. Our parcellation uses Ward’s hierarchical agglomerative clustering algorithm [11] to create groups of voxels that have similar activity across trials. Thus, the signal is averaged in each parcel. The number of parcels used here is fixed to 400 for the whole brain. Note that we do not focus on the influence of the parcellation on the results, but on the comparison of the results of different regression methods. The dimensions of the real data set are $m = 400$ and $n = 72$ (divided in 3 sizes). The prediction score is computed with a 4-folds cross-validation (i.e. a leave-one-object-out validation) for each subject in the intra-subject analysis, and with a 10-folds cross-validation (i.e. a leave-one-subject-out validation) for the inter-subject analysis. In that case, the procedure builds a predictor of object size that generalizes across subjects. The parameters of *Enet* and *SVR* are optimized with a 4-folds cross-validation in the ranges given before.

Results on a real functional neuroimaging dataset The results of the different methods (mean and standard deviation of ζ across 10 subjects) with fMRI data are shown Tab.1 for the intra-subject analysis, and Tab.2 for the inter-subject analysis. The proposed algorithm yields equivalent results to *Enet* in the intra-subject case, but 8% increase of the explained variance in the inter-subject case. Moreover, the *MCBR* algorithm is almost as good as the *SVR* in both cases. The histograms of the (voxel-level) weights averaged across subjects are given in Fig.2 for *Enet*, *MCBR* and *SVR* algorithms. We can see that the feature maps obtained in the *Enet* method are less sparse than those obtained with the *MCBR* method. Indeed, our algorithm regularizes more strongly uninformative features, and more weakly the weights of informative features.

	BRR	ARD	Enet	SVR	<i>MCBR</i>
Mean ζ	-0.15	0.85	0.89	0.91	0.89
Std ζ	0.51	0.08	0.05	0.03	0.04

Table 1. Intra-subject analysis - Mean and standard deviation of ζ averaged across 10 subjects.

	BRR	ARD	Enet	SVR	<i>MCBR</i>
Mean ζ	0.01	0.7	0.71	0.8	0.79
Std ζ	0.37	0.15	0.16	0.13	0.05

Table 2. Inter-subject analysis - Mean and standard deviation of ζ averaged across 10 subjects.

The averaged weights of the parcels across subjects in the intra-subject analysis are shown in Fig.2 for *Enet* (a), *MCBR* (b) and *SVR* (c) algorithms. The *MCBR* algorithm finds the relevant regions of interest in the occipital region, as expected, while leaving the remainder of the brain with null weights. Starting from the whole brain, *MCBR* selects very few parcels in the occipital cortex, corresponding to visual areas (V2-V3) and a part of the posterior-dorsal lateral occipital region of the lateral occipital complex. This is consistent with the fact that lateral visual cortex contains highly reliable signals discriminative of size differences between object exemplars. The *Enet* method finds a relevant region in the lateral occipital complex too, but selects also more questionable regions (e.g. in the temporal lobe), yielding less interpretable activation maps. The results of the *SVR* algorithm are very difficult to interpret.

4 Discussion

Regularization of voxels loadings significantly increases the generalization ability of the predictive model. However, this regularization has to be adapted to each particular dataset. In place of costly cross-validation procedures, we cast regularization in a Bayesian framework and treat the regularization weights as hyper-parameters. This approach yields an adaptive and efficient regularization, and can be seen as a compromise between a global regularization (*BRR*) which does not take into account the sparse or focal distribution of the information, and *ARD*, that is subject to overfit in high-dimensional feature spaces.

Results on real data show that our algorithm gives access to interpretable feature maps which are a powerful tool for understanding brain activity. Moreover, the *MCBR* algorithm yields more accurate predictions than other regularization

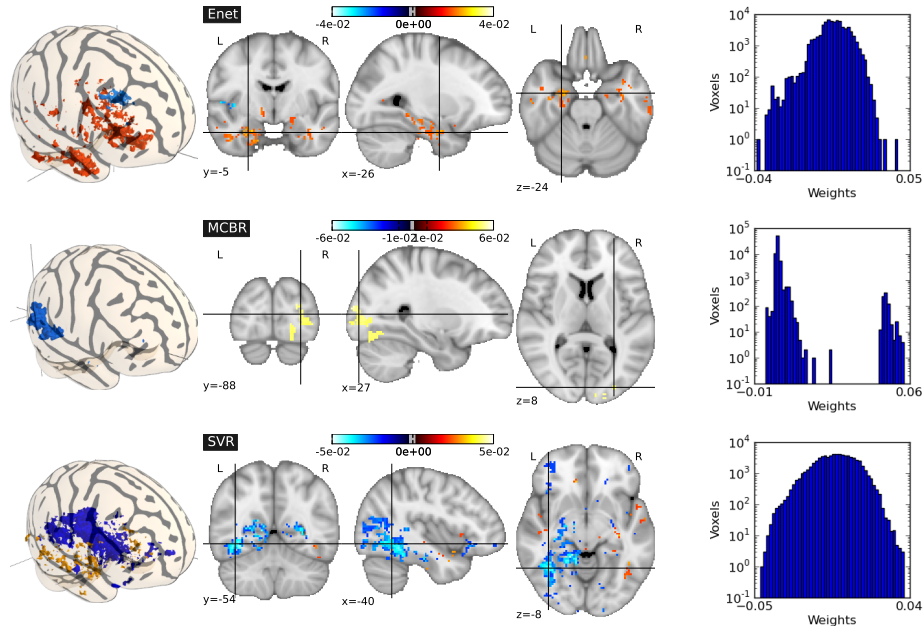


Fig. 2. Intra-subject analysis - Results obtained with real data in a whole brain analysis. Representation of the average weights across subjects superimposed on the anatomical image of one particular subject (left), and corresponding histograms of the averaged weights (right) for *Enet* (top), *MCBR* (middle) and *SVR* (bottom). With *Enet*, there are a lot of parcels with non-null weight. For the *MCBR* algorithm, starting from a whole-brain analysis, very few parcels have a non-null weight, yielding an interpretable predictive pattern: these parcels are embedded in the occipital region (V1-V3) and extend laterally. Finally, the weights for the voxels found by the *SVR* algorithm are less sparse, and spread throughout the whole brain, so that the interpretation of such a map is challenging.

methods (*BRR*, *ARD* and *Enet*). The standard method *SVR* performs slightly better than the *MCBR* algorithm (yet, the difference is not significant), probably due to the fact that the kernel helps to deal with the high dimensionality of the data. However, *SVR* does not yield meaningful feature maps, since it enforces sparsity in the dual space and not in the primal space.

The question of model selection (i.e. the number of classes K) has not been addressed in this paper, but the method detailed in [3] can be used within our framework. Here, model selection is performed implicitly by emptying classes that do not fit the data well. In that respect, the choice of heterogeneous priors for different classes is crucial: replacing our priors with class-independent priors yields a decrease of 0.3 in explained variance on simulated data. Moreover, our results are insensitive to the particular numerical choice on hyper-priors (data not shown), provided that the associated distributions cover the range of relevant parameter distributions. Crucially, the priors used here can be used in any

regression problem, provided that the target data is approximately scaled to the range of values used in our experiments. In that sense, the present choice of priors can be seen as universal.

Conclusion We have presented a multi-class regularization approach that includes adaptive ridge regression and automatic relevance determination as limit cases. Experiments on real data show that our approach is well-suited for neuroimaging, as it yields accurate predictions and also stable and interpretable feature loadings.

Acknowledgments: The authors acknowledge support from the ANR grant ViMAGINE ANR-08-BLAN-0250-02.

References

1. Bishop, C.M., Tipping, M.E.: Variational relevance vector machines. In: UAI '00: 16th Conference on Uncertainty in Artificial Intelligence. pp. 46–53 (2000)
2. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage* 44(1), 112 – 122 (2009)
3. Chib, S., Jeliazkov, I.: Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association* 96, 270–281 (2001)
4. Cortes, C., Vapnik, V.: Support vector networks. In: *Machine Learning*. vol. 20, pp. 273–297 (1995)
5. Cox, D., Savoy, R.: Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19(2), 261–270 (Jun 2003)
6. Dayan, P., Abbott, L.: *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press (2001)
7. Eger, C., Kell, A., Kleinschmidt, A.: Graded size sensitivity of object exemplar evoked activity patterns in human loc subregions. *Journal of Neurophysiology* 100(4):2038–47 (2008)
8. Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J.: Bayesian decoding of brain images. *NeuroImage* 39, 181–205 (2008)
9. Qi, Y., Minka, T.P., Picard, R.W., Ghahramani, Z.: Predictive automatic relevance determination by expectation propagation. In: *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. ACM Press (2004)
10. Tipping, M.: The relevance vector machine. In: *Advances in Neural Information Processing Systems*, San Mateo, CA (2000)
11. Ward, J.H.: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58(301), 236–244 (1963)
12. Yamashita, O., Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage* 42(4), 1414 – 1429 (2008)
13. Yuan, M., Yuan, M., Lin, Y., Lin, Y.: Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68, 49–67 (2006)
14. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* 67, 301–320 (2005)