

Adaptive hierarchical Bayesian mixture for sparse regression - An application to brain activity classification

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

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

² Université Paris-Sud 11, Orsay, France,

³ SELECT team, INRIA Saclay-Île-de-France, France

⁴ INSERM U562, Gif/Yvette, France

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

Abstract. In this article, we describe a novel method for regularized regression and apply it to the prediction of a behavioural variable from brain activation images. In the context of neuroimaging, regression or classification techniques are often plagued by the curse of dimensionality, due to the extremely high number of voxels and the limited number of activation maps. A commonly-used solution is regularization of the weights used in the parametric prediction function. To solve the difficult issue of choosing the correct amount of regularization in the model, we propose to use a Bayesian framework, but model specification needs a careful design to balance adaptiveness and sparsity. We introduce an adaptive mixture regularization that generalizes previous approaches. Based on a Variational Bayes estimation framework, our algorithm is robust to overfitting and more adaptive than other regularization methods. Results on both simulated and real data show the accuracy of the method in the context of brain activation images.

1 Introduction

A recent trend in neuroimaging [1] consists of inferring behavioral information or cognitive states from activation brain images such as those obtained with functional magnetic resonance imaging (fMRI). It can provide more sensitive analyses than standard statistical parametric mapping procedures [2]. Specifically, it can be used to check the involvement of one or several brain regions in specific cognitive or perceptual functions by evaluating the accuracy of the prediction of a behavioral or cognitive variable of interest (the *target*) when the classifier is instantiated on that particular brain region. Such an approach is particularly well suited for the investigation of population coding [3]: certain neuronal populations are thought to activate specifically when a certain perceptual or cognitive parameter reaches a given value. Inferring this parameter from the neuronal activity helps to *decode* the brain system.

The main difficulty in this procedure is the huge dimensionality of the data,

with far more features than samples. In this article, the samples (or activation maps) refer to the regression coefficients in the General Linear Model (GLM) analysis (i.e. beta maps), while the features correspond to the voxels. The large number of features leads to overfitting and thus a dramatic decrease in prediction accuracy. One common solution consists in working in the dual space using the kernel trick [4], but in the case of neuroimaging, one may prefer to use explicit loadings on brain regions, hence to work in the primal space. To deal with this dimensionality problem, some regularized regression techniques have been developed, forcing the majority of the features to have zero or close to zero loadings, such as Lasso [5] and elastic net [6]; however, these approaches require that the amount of regularization is fixed beforehand, and possibly optimized by cross-validation. By contrast, Bayesian methods (e.g. adaptive ridge regression [7] and Automatic Relevance Determination – ARD [8]) adapt the amount of regularization to the problem at hand. These regularized regression methods have already been used for predicting cognitive states. In [9], a model based on ARD has been proposed for weighting activity patterns in the case of logistic regression, but ARD can overfit in the case of very high dimension. Similarly, in [10] a Bayesian regression approach is used to classify brain states, but the construction relies on ad hoc voxel selection steps, so that there is no proof that the solution is optimal. In summary, Bayesian regression techniques have been developed in two contexts: on the one hand, adaptive ridge regression regularizes all the loadings with the same parameter, which is not well-suited for brain activity where only few clusters of voxels have task-related activity; on the other hand ARD regularizes separately each voxel, and is prone to overfitting when the model contains too many regressors.

In this article, we develop an intermediate approach for sparse regularized regression, which assigns each voxel to a class. Regularization is performed in each class separately, leading to a stable and adaptive regularization, while avoiding overfit – this approach is thus a compromise between ridge regression and ARD. The algorithm is based on a Variational Bayes (VB) approach which leads to a fast estimation of the weight distributions. The parameters-updating algorithm is no more complex than an Expectation Maximization algorithm, and it iteratively adapts the hyperparameters to the particular problem. Moreover, the VB approach has one important property for model selection : it contains a built-in criterion, the free energy of the model. After introducing our model and the VB approach, we show that the proposed algorithm performs better than reference methods on simulated data, and leads to promising results on real data.

2 Methods

We introduce the following regression model :

$$y = \Phi w + \epsilon \tag{1}$$

where y represents the behavioural data to be fitted ($y \in \mathbb{R}^n$) and w the parameters ($w \in \mathbb{R}^m$). n is the number of beta maps obtained with a GLM (each

image corresponds to one stimulus presentation) and depends on the number of blocks in the paradigm; m is the number of voxels and Φ is the design matrix ($\Phi \in \mathbb{R}^{n \times m}$, each row is an m -dimensional activation map). The crucial issue here is that $n \ll m$, so that estimating w is an ill-posed problem. A solution is to introduce some priors over the parameter distribution.

Priors on regression and adaptive relevance determination (ARD) -

Regularized regression can be used to solve this ill-posed problem, by imposing a prior on the weights, hence possibly a sparse feature weighting. First, we model the noise with a Gaussian density:

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

$$p(\sigma^2) = \Gamma^{(-1)}(\lambda_1, \lambda_2) \quad (3)$$

with two hyperparameters λ_1, λ_2 . $\Gamma^{(-1)}$ stands for the inverse gamma density. For mathematical convenience, we make use of conjugate Gaussian prior for the weights, leading to an L_2 penalty :

$$w \sim \mathcal{N}(0, A^{-1}), A = \text{diag}(\alpha_1, \dots, \alpha_m) \quad (4)$$

$$p(\alpha_i) = \Gamma(\gamma_1^i, \gamma_2^i) \quad (5)$$

where $\alpha_i, i \in [1, m]$ are the precision parameters, and Γ is the gamma density. Two important cases correspond to adaptive ridge regression ($\alpha_1 = \dots = \alpha_m$) and ARD ($\alpha_i \neq \alpha_j$ if $i \neq j$). Still, the highly adaptive regularization of the ARD can lead to severe overfitting if $n \ll m$.

Mixture model (VBK) - In order to accommodate for the sparsity of ARD with the stability of adaptive ridge regression, we introduce an intermediate representation, in which each voxel i belongs to one class among K indexed by the discrete variable z_i . Thus, all the features within a class $k \in [1, \dots, K]$ share the same precision parameter α_k . Next, we introduce a prior on z :

$$p(z) = \prod_{i=1}^m \prod_{k=1}^K \pi_k^{\eta_{ik}} \text{ with } \begin{cases} \eta_{ik} = 0 & \text{if } z_i \neq k \\ \eta_{ik} = 1 & \text{if } z_i = k \end{cases} \quad (6)$$

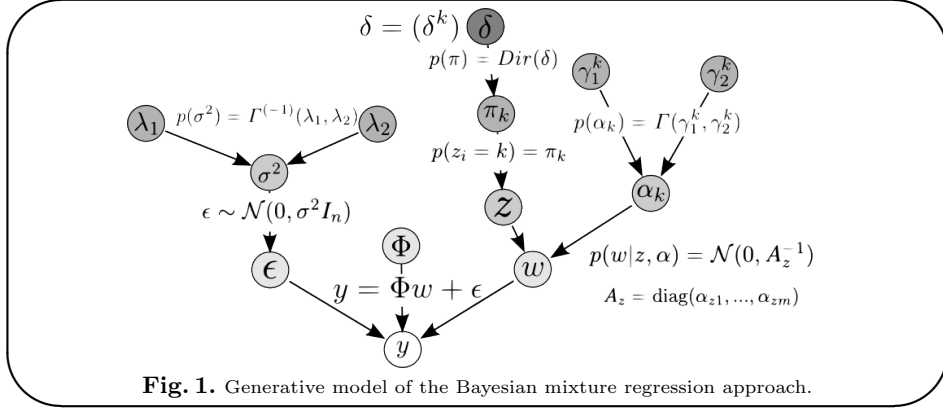
and a Dirichlet prior on π_k with hyperparameter δ : $p(\pi) = \text{Dir}(\delta)$. The complete generative model is summarized in Fig.1. This model has no spatial constraints, and thus is not spatially regularized.

Estimation and Selection of the model by Variational Bayes -

To select a model among several alternatives, it is natural to keep the model that yields the largest data evidence $p(y)$. Thus, we use the variational approach that provides a closed-form approximation $q(\theta)$ of $p(\theta|y)$, where $q(\theta)$ is in a given family of distributions and $\theta = [\sigma^2, z, \alpha, w, \pi]$ are the parameters of the model. We have:

$$q(\theta) = q(w) \left(\prod_{i=1}^m q(z_i) \right) \left(\prod_{k=1}^K q(\alpha_k) q(\pi_k) \right) q(\sigma^2) \quad (7)$$

By using conjugate priors, this variational scheme provides closed form for the update rules. We can then decompose $\log p(y)$ as the sum of the free energy \mathcal{F}



and the Kullback-Leibler divergence between the true posterior $p(\theta|y)$ and the variational approximation $q(\theta)$:

$$\log p(y) = \mathcal{F}(q(\theta)) + D_{KL}(q(\theta)||p(\theta|y)) \quad (8)$$

$$\mathcal{F}(q) = \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta \quad (9)$$

The free energy \mathcal{F} is thus a lower bound on $\log p(y)$ with equality iff $q(\theta) = p(\theta|y)$, and inferring the density q of the parameters corresponds to maximizing \mathcal{F} , which we do not detail here. Moreover, free energy is a measure of the quality of the model and can be used in a model-selection scheme. In which case, the global time-consuming cross-validation-based optimization of K is avoided.

Initialization and validation - The initialization is set as in [8], with weakly-informative priors, $\lambda_1 = \lambda_2 = \gamma_1 = \gamma_2 = 10^{-6}$ and $\delta_k = 5, \forall k \in [1, \dots, K]$ (see [11]). The initialization of z is performed by using a K-Means on the F-statistics of the features. Since the estimation algorithm converges to a local maximum of \mathcal{F} , it is very sensitive to this initialization.

The performance of the competing models is assessed using the ratio of explained variance ζ . Let Φ^l, y^l be a learning set, Φ^t, y^t a test set, and $\hat{y}^t(\Phi^l, y^l, \Phi^t)$ the prediction obtained with a model trained on Φ^l, y^l and tested with Φ^t .

$$\zeta(\Phi^l, y^l, \Phi^t, y^t) = \frac{\text{var}(y^t) - \text{var}(y^t - \hat{y}^t(\Phi^l, y^l, \Phi^t))}{\text{var}(y^t)} \quad (10)$$

ζ is the amount of variability in the response that can be explained by the model (prediction is perfect if $\zeta = 1$, and is worst than chance if $\zeta < 0$).

3 Experiments and Results

Simulated Data - We have tested our algorithm on a simulated dataset X of n images with squared Regions of Interest (ROIs) \mathcal{R} (defined by a position and a width). We note b the background (i.e. outside the ROIs). The signal in the (i, j) voxel of the k^{th} image is simulated as :

$$X_{i,j,k} = \sum_{r \in \mathcal{R}} \mathbb{I}_r(i,j) \alpha_{r,k} u_{i,j,k} + \mathbb{I}_b(i,j) u_{i,j,k} + \epsilon_{i,j,k} \quad (11)$$

where $u_{i,j,k}$ is a random value from an uniform distribution in $[0, 1]$, $\epsilon_{i,j,k}$ a random value from a Gaussian distribution $\mathcal{N}(0, 1)$ smoothed with a parameter of 2 voxels to mimic the correlation structure observed in real fMRI datasets, $\alpha_{r,k} \sim \mathcal{U}[0, 1]$ for ROI r and image k . $\mathbb{I}_r(i, j) = 1$ (resp. \mathbb{I}_b) if the (i, j) voxel is in r (resp. b), and $\mathbb{I}_r(i, j) = 0$ (resp. \mathbb{I}_b) elsewhere. We simulate the target Y as : $Y_k = \sum_{r \in \mathcal{R}} \alpha_{r,k}$. We generate a dataset of 250 images, and split it into a learning set of $n = 200$ images and validation set of 50 images. The images have a size of 20×20 , with two non-overlapping ROIs of width 2 pixels. An example is given in Fig.3. We compare our algorithm with three other methods : a bilinear kernel-based variational ARD regression (also called Relevance Vector Machine *RVM* [8]), an elastic net regularization procedure (called *Enet* [6] with parameters $s = 0.5$ and $\lambda = 0.1$), and a Support Vector Regression procedure (called *SVR* [12] with a linear kernel and $C = 1$). The regularization parameters for the RVM, SVR and *Enet* methods are fixed using an ad hoc calibration (by selecting the ones that yield the best accuracy on simulated datasets).

Results on Simulated Data - In a first experiment, we report the average results obtained for the different methods for 40 tests, and K equals to 1, 2 or 3. See the results Fig.3: the *VBK* algorithm outperforms the other methods for $K > 1$ (c). Moreover, the *VBK* method finds very low and stable weights outside the ROIs (a,b), where *Enet* leads to a sparse (many weights are closed to zero) but less stable (higher standard deviation) regularized solutions. Both *RVM* and *SVR* yield a poorly regularized solution : many irrelevant voxels have a significant weight. In a second experiment, we compute the explained variance and the free energy for different models with $K \in [1, 2, 3, 4, 5]$ for 20 samples (see Fig. 2). We can see that the free energy is strongly correlated with the explained variance, and yields the same optimum value $K = 3$. Moreover, the *VBK* classifier has a standard deviation which increases with K : being more adaptive, the classifier yields less stable prediction.

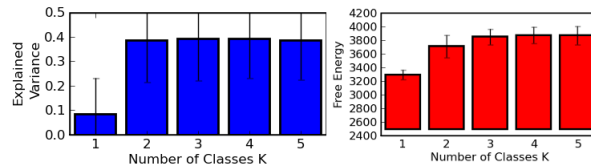


Fig. 2. Results of the model selection procedure in the simulation experiment. The free energy (red) and the explained variance (blue) are averaged over 20 simulations. They are strongly correlated, with a maximum reached for $K = 3$. Thus, the free energy of the *VBK* model can be used for the selection of the model.

Real data - We use a real dataset related to a *numerotopy* (mental representations of quantities) experiment. During the experiment, ten healthy volunteers view dot patterns with different quantities of dots ($\nu = 2, 4, 6$ and 8 ; we take $y = \log(\nu)$), with 8 repetitions of each stimulus : so that we have a total of $n = 32$

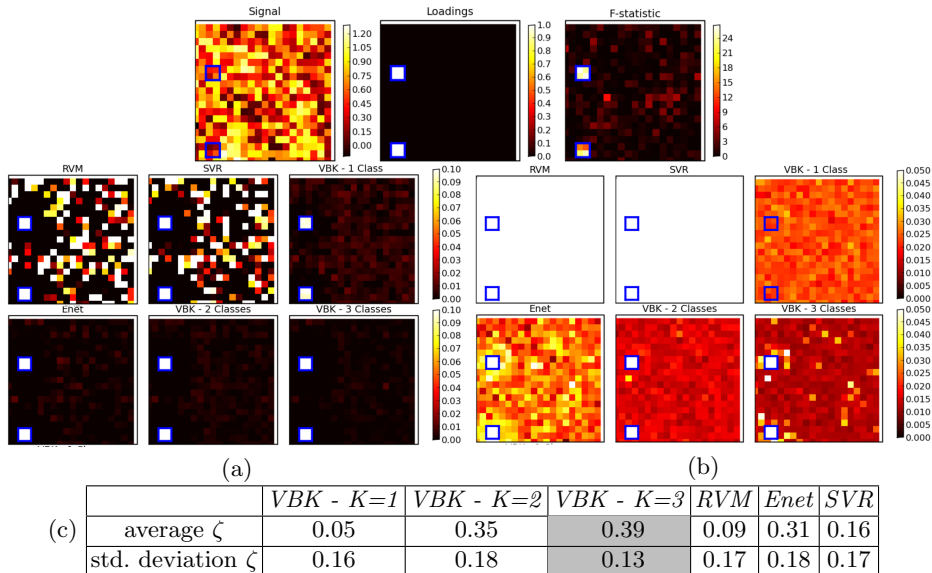


Fig. 3. Results of the simulation experiment. ROIs are outlined by blue squares. Top - Example of simulated data: amplitude of the signal of an image (left), simulated loadings (middle) and F-statistic (right). Mean **(a)** and standard deviation **(b)** for the weights found with different methods. The *VBK* approach gives weights similar to those of the *Enet* method, but with more stable estimation outside the ROIs. The *RVM* and *SVM* approaches lead to non-zero weights outside the ROIs, and weights estimation is not stable across trials. **(c)** Ratio and standard deviation of ζ (see Eq.10) for different methods averaged on 40 simulations. The *VBK* algorithm outperforms all the other techniques and yields less variable results (when $K > 1$).

images by subject. Functional images were acquired on a 3 Tesla MR system with 12-channel head coil (Siemens Trio TIM) as T2* weighted echo-planar image (EPI) volumes using a high-resolution EPI-sequence. 26 oblique-transverse slices covering parietal and superior parts of frontal lobes were obtained in interleaved acquisition order with a TR of 2.5 s (FOV 192 mm, fat suppression, TE 30 ms, flip angle 78° , $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ voxels). The slice-timing, realignment, co-registration and the fit of the GLM have been performed with the SPM5 software. For our analysis we use the resulting activation maps (each image is standardized by removing the mean and dividing by the standard deviation). We select 1000 voxels included in the main ROI, i.e. the Intra-Parietal Sulcus (IPS), which has been manually delineated in all the available datasets prior to fMRI data analysis. We divide this region into 200 parcels with an alternative of Ward’s algorithm [13], and we average the values of the beta maps over voxels within each parcel. The signal is thus more stable and the computation faster.

Results on Real data - We compute the explained variance obtained in a leave-one-session-out procedure for different methods and we compare the results with the *VBK* algorithm ($K = 3$). The averaged results across 10 subjects. are given in Fig.4: the *Enet* algorithm outperforms the other methods, but yields less stable predictions. The *VBK* method performs better than *ARD* but worst than the adaptive *Ridge* algorithm. More importantly, the *VBK* algorithm pro-

vides maps of probabilistic belonging. We perform a two-class study (the binary case yields to more interpretable maps) on real data. Fig.5 (a) shows the average loadings w found by the *VBK* algorithm ($K = 2$) across subjects superimposed on the anatomical image of one subject. Fig.5 (b) and Fig.5 (c) give the average probability of each voxel belonging to the low-weight or high-weight class. We can see that the *VBK* provides explicit classification maps that allow to understand the anatomical organization of discriminant brain activity.

Fig. 4. Results on real data: ratio and standard deviation of ζ (see Eq.10) for different methods averaged across 10 subjects. The *VBK* ($K = 3$) method performs better than the *ARD* and is respectively 2% and 4% below the *Ridge* and *Enet* algorithms. The *Enet* algorithm yields to less stable predictions.

	<i>VBK</i>	<i>Ridge</i>	<i>ARD</i>	<i>Enet</i>
Average ζ	39%	41%	17%	43%
Std. dev. ζ	25%	20%	103%	60%

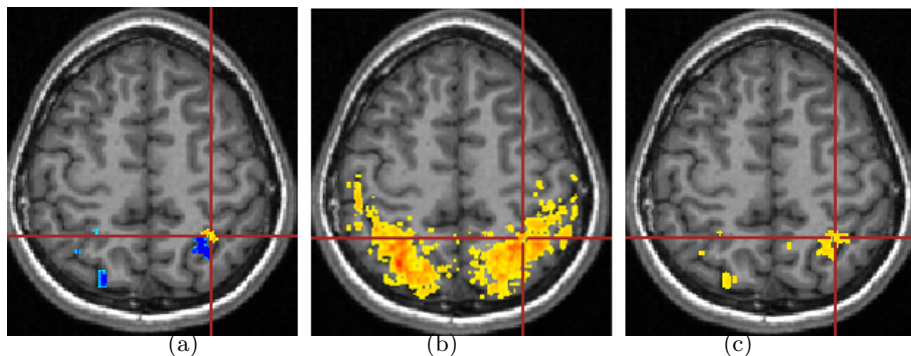


Fig. 5. Results on the real dataset with $K = 2$. (a) Average weights of the parcels across subjects (between -0.1 - blue - and 0.1 - red), found by the *VBK* algorithm and superimposed on the anatomical image of one subject. The loadings with low magnitudes are not shown. Average probability (between 0 -yellow - and 0.5 - red) of each voxel to belong in the low-weight class (b) or high-weight class (c).

4 Discussion

Regularization of voxels loadings significantly increases the generalization ability of the regression. However, this regularization has to be adapted to each particular fMRI dataset, which is done in this article by introducing a Bayesian mixture framework. Our approach performs an adaptive and efficient regularization, and is a compromise between a global regularization (ridge regression) which does not take into account the region-based structure of the information, and *ARD*, that is subject to overfit in large dimension.

On simulated data, our approach performs better than other classical methods such as *SVR*, *RVM* and *Enet*. Besides an increase of the explained variance which shows that the *VBK* approach extracts more information from the data, the loadings are less noisy and more stable, leading to more interpretable activation maps. The correlation between the free energy and the prediction accuracy confirms that free energy is a valuable model selection tool that furthermore avoids time-consuming optimization by cross-validation.

Results on real data show that the *VBK* algorithm gives access to interpretable loading maps which are a powerful tool for understanding brain activity. The

VBK algorithm yields to less accurate predictions than other regularization methods, which can be explained by the difficulty of initializing the variable z in the study of real data. We expect that alternative solution based on Gibb's sampling will lead to more accurate predictions.

A future direction of our work is to optimize the spatial model used in our framework (here we simply use a prior parcellation of the search volume) in relationship with the prediction function that is used. In parallel, we will develop non-linear versions (e.g. logistic/probit) of this model for classification problems.

Conclusion - We have presented a multi-class regularization approach that includes adaptive ridge regression and automatic relevance determination as limit cases; the ensuing problem of optimizing the number of classes is easily dealt with in the Variational Bayes framework. Our simulations and real experiments show that our approach is well-suited for neuroimaging, as it yields a powerful framework and also reliable and interpretable feature loadings.

References

1. Cox, D.D., Savoy, R.L.: Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* **19**(2) (2003) 261–270
2. Kamitani, Y., Tong, F.: Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* **8**(5) (April 2005) 679–685
3. Dayan, P., Abbott, L.F.: *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press (2001)
4. Aizerman, A., Braverman, E.M., Rozoner, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* **25** (1964) 821–837
5. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** (1996) 267–288
6. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* **67** (2005) 301–320
7. Bishop, C.M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer (2006)
8. Bishop, C.M., Tipping, M.E.: Variational relevance vector machines. In: *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. (2000)
9. Yamashita, O., aki Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage* **42** (2008)
10. Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J.: Bayesian decoding of brain images. *NeuroImage* **39** (2008) 181–205
11. Penny, W., Roberts, S.: Variational bayes for 1-dimensional mixture models. (2000)
12. Cortes, C., Vapnik, V.: Support vector networks. In: *Machine Learning*. Volume 20. (1995) 273–297
13. Ward, J.H.: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**(301) (1963) 236–244