BAYESIAN REGRESSION OF FUNCTIONAL NEUROIMAGES

Dimitris G. Tzikas[1], *Aristidis Likas*[1], *Nikolas P. Galatsanos*[1], *Ana S. Lukic*[2], *and Miles N. Wernick*[2]

[1]Department of Computer Science, University of Ioannina, GREECE, 45110.[2]
Department. of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA.

ABSTRACT

A Bayesian approach is proposed for statistical analysis of fMRI data sets in a two state ("on-off") activation study. The approach is based on the Relevance Vector Machine (RVM) regression framework. According to this approach the shape of the activations is a superposition of kernel functions, one at each pixel of the image, and a hierarchical Bayesian model is employed which imposes a sparse representation by selecting a number relevant kernel functions. We have implemented an incremental method for constructing the RVM model and, in addition, we have employed a cross-validation criterion to deal with the problem of kernel width selection. The proposed method allows the accurate estimation of the activation locations when correlated noise is present even at low signal-to-noise ratios. We tested this method using an artificial phantom derived from a previous neuroimaging study with promising results compared with previous approaches.

1. INTRODUCTION

The aim of a two state neuroimaging study using fMRI is to compare two groups of images (acquired in two different conditions) in order to identify brain regions whose activity level changes in response to some task or drug. The result of the study is an activation map indicating these regions. Since the observed images are quite noisy, one of the most important components of a neuroimaging study is the statistical method used to detect the activation pattern.

In this paper we propose the use of RVM-based regression [7] to detect the location of the activation signal. According to this approach the activation signal is modeled as the superposition of kernel basis functions with unknown amplitudes. We associate one kernel function with unknown weight to each pixel of the image. Furthermore, prior knowledge that most of the amplitudes are zero is used. For this purpose a hierarchical Bayesian model is applied assuming a prior and a hyperprior for the unknown amplitudes. In this paper we consider Gaussian-shaped kernel functions, but any other kernel function could be used as well. The RVM learning algorithm is used to estimate the vector of amplitudes. Since RVM yields a sparse representation most obtained amplitudes are very small. Thus, the location of the activations is determined automatically as the location of largest amplitudes. In order to make the method efficient for large images we have implemented a fast incremental method for RVM modeling where kernels are added incrementally to the RVM model starting with one kernel [3].

We compared this method with previous methods in [4]. We found that this method outperformed all other methods and performed almost as well as the Reversible Jump Markov Chain Monte Carlo (RJMCMC) approach in [3]. The slight performance advantage of the RJMCMC seems to not justify its significantly larger computational cost. Furthermore, for MCMC techniques there is no universally accepted criterion or methodology to decide when to terminate [2].

2. THE IMAGING MODEL

Herein we consider a two-state ("on-off") activation study, in which activation- and control-state images are obtained, and used to identify activated regions. We model images of the brain in the control (c) and activation (a) states, respectively, as

$$g_j^{(c)}(x,y) = b(x,y) + n_j^{(c)}(x,y),$$
$$g_j^{(a)}(x,y) = b(x,y) + s(x,y) + n_j^{(a)}(x,y), \quad (1)$$

for $j = 1,\ldots,N$, where, at each spatial location (x,y) in the brain. The activation pattern we wish to determine is $s(x,y)$. The baseline value is $b(x,y)$ and $n_j^{(c)}(x,y)$ and $n_j^{(a)}(x,y)$ are the imaging noise contributions for the two cases.

We model the noise as additive colored Gaussian noise, the covariance of which is proportional to the value of the baseline image at the corresponding pixel and is assumed known [5]. This noise is given by

$n_j^{(c,a)}(x,y) \sim N(0, C_n)$.

We define the average activation image as

$$g^{(s)}(x,y)=\sum_{j=1}^{N}\left(g_j^{(a)}(x,y)-g_j^{(c)}(x,y)\right). \quad (2)$$

Then, from equation (1) follows

$$g^{(s)}(x,y)=s(x,y)+n^{(s)}(x,y),$$
$$n^{(s)}(x,y)\sim N(0,\frac{2}{N}C_n). \quad (3)$$

In the activation state, the image differs from the control state only by addition of the activation pattern. We decide to estimate the activation signal $s(x,y)$ using a RVM [7], and model it as the superposition of M kernel functions according to the equation

$$s(x,y)=\sum_{i=1}^{M}w_i K\left((x,y),(x_i,y_i)\right), \quad (4)$$

where M is the number of pixels of each image, $K(x,x_0)$ is some kernel function and w_i is the amplitude of the kernel centered at $[x_i,y_i]$. Though any kernel function can be used, we assume the commonly used Gaussian kernel functions, of the form

$$K(x,x_0)=e^{-\frac{1}{2}q\|x-x_0\|}. \quad (5)$$

The width of the kernel q is unknown, and must be specified through cross-validation.

Introducing the design matrix $\Phi=[\phi(x_1),\phi(x_2),...,\phi(x_N)]^T$, where $\phi(x_n)=[K(x_n,x_1),K(x_n,x_2),...,K(x_n,x_N)]^T$, and defining t as the vector of observation values of the average activation image $g^{(s)}(x,y)$ we compute the likelihood of the average activation image according to equation $p(t\mid w)=N\left(\Phi w,\frac{2}{N}C_n\right)$.

The model described requires the estimation of as many parameters as the available data points. This would lead to over fitting unless we impose some additional constraint on the parameters. Thus, we assume a Gaussian prior distribution over the amplitude vector $w=(w_1,w_2,...,w_M)$. This approach introduces our preference for smoother activation images. In other words we have

$$p(w_i\mid a_i)=N(0,a_i^{-1}), \quad (6)$$

where $a=[a_1,a_2,...,a_M]$ is a vector of M hyperparameters determining the strength of the prior distribution on each basis function's amplitude. The vector of hyperparameters, which is considered to be a random variable, is a scale parameter and as such it is assigned a gamma prior distribution given by

$$p(a_i)=Gamma(\alpha,\beta). \quad (7)$$

where the parameters α,β are set ($\alpha=\beta=10^{-4}$) to specify a non-informative prior. By integrating over the hyperparameters, we can compute the 'true' weight prior $p(w)=\int p(w\mid a)p(a)da$. This integral gives a Student-t prior, which is well known to give sparse representations since most of its mass is concentrated close to the origin of the axes of definition [7].

3. BAYESIAN INFERENCE

Estimation of the activation signal $s(x,y)$ is a Bayesian inference procedure. Because we cannot compute the posterior $p(w,a\mid t)$ directly, following the procedure described in [7] we decompose it using Bayes theorem as

$$p(w,a\mid t)=p(w\mid t,a)\,p(a\mid t). \quad (11)$$

Then, the posterior distribution over the weights can be analytically computed by

$$p(w\mid t,a)=\frac{p(t\mid w)\,p(w\mid a)}{p(t\mid a)}=N(\mu,\Sigma),$$
$$\Sigma=\left(\Phi^{\mathrm{T}}C_n\Phi+A\right)^{-1}, \quad (12)$$
$$\mu=\Sigma\Phi^{\mathrm{T}}C_n t,$$

with $A=diag(a_1,a_2,...,a_M)$.

On the other hand, the posterior over the hyperparameters $p(a\mid t)$ cannot be computed analytically and we approximate it by a delta function at its mode as

$$p(a\mid t)\approx\delta(a_{MP}),\ \text{where}$$
$$a_{MP}=\arg\max_a\left(p(a\mid t)\right). \quad (13)$$

This approximation is frequently used and in this problem it is very effective. .

Since $p(a\mid t)\propto p(t\mid a)\,p(a)$, we can find a_{MP} by maximizing $p(t\mid a)\,p(a)$

$$a_{MP}=\arg\max_a\left(p(t\mid a)\,p(a)\right) \quad (14)$$

$p(t\mid a)$ is known as marginal likelihood or type-*II* likelihood and is computed by marginalizing over the weights according to

$$p(t\mid a)=\int p(t\mid w)p(w\mid a)dw.$$

This gives

$$p(t\mid a)=N(0,C),\ \text{with}\ C=C_n+\Phi A^{-1}\Phi^T. \quad (15)$$

4. INCREMENTAL MARGINAL LIKELIHOOD MAXIMIZATION

Unfortunately a_{MP} cannot be computed analytically. Instead we can use iterative formula for its re-estimation:

$$a_i^{new} = \frac{1+2\alpha}{\mu_i^2 + \Sigma_{ii} + 2\beta}, \quad (17)$$

where μ_i is the i-th element of the posterior mean weight and Σ_{ii} is the i-th diagonal element of the posterior weight covariance [7].

A drawback of the above optimization method is the complexity of computing matrix Σ, if the number of basis functions is large. Some of these computations can be avoided by pruning basis functions whose amplitude is estimated to be zero. However, initially there are N basis functions, and computation of Σ is time consuming

One can bypass this difficulty by initially assuming only one basis function, and then adding or deleting basis functions at each iteration [6]. For the case of uniform prior over hyperparameter a, maximization of (14) is equivalent to maximizing

$$L(\alpha) = \log p(t \mid \alpha)$$

$$= -\frac{1}{2}\left[N \log 2\pi + \log|C| + t^T C^{-1} t \right]. \quad (18)$$

Given a single hyperparameter α_i we can decompose $L(\alpha)$ into two terms, one being independent of α_i:

$$L(\alpha) = L(\alpha_{-i}) + l(\alpha_i),$$

where $L(\alpha_{-i})$ is independent of α_i,

$$l(\alpha_i) = \frac{1}{2}\left[\log \alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i} \right], \quad (19)$$

$$s_i = \phi_i^T C_{-i}^{-1} \phi_i \text{ and } q_i = \phi_i^T C_{-i}^{-1} t. \quad (20)$$

C_{-i} is obtained from matrix C with the contribution of basis function ϕ_i removed:

$$C_{-i} = C - \alpha_i^{-1} \phi_i \phi_i^T \quad (21)$$

Analysis of $l(\alpha_i)$ shows that $L(\alpha)$ has a unique maximum with respect to α_i:

$$\alpha_i = \frac{s_i^2}{q_i^2 - s_i} \quad if\ q_i^2 > s$$

$$\alpha_i = \infty \quad if\ q_i^2 \le s \quad (22)$$

Thus we can find a_{MP} that maximizes the marginal likelihood (14) by iteratively:

- adding a basis function ϕ_i with $q_i^2 > s$,
- re-estimating hyperparameter α_i for a basis function already in the model,
- or deleting a basis function ϕ_i with $q_i^2 \le s$.

When adding a basis function or re-estimating the value of its hyperparameter we set $\alpha_i = \frac{s_i^2}{q_i^2 - s_i}$ which maximizes $L(\alpha)$. Thus at each step the marginal likelihood increases until it reaches a local minimum. Vectors s_i and q_i can be calculated using their values calculated at the previous iteration of the algorithm and thus computationally the algorithm is very efficient.

5. RESULTS

To evaluate the method we used a simple phantom whose properties were derived from a positron emission tomography (PET) neuroimaging study, but are also representative of whole-brain, blood-oxygenation-level-dependent (BOLD) functional magnetic resonance imaging (fMRI) studies that have been spatially smoothed [5]. Figure 1 (a) shows the average of ten simulated activation patterns. The location and amplitude of the activation were varied randomly from image to image to represent physiological variability between subjects or scans. Figure 1 (b) shows the average of ten simulated "activated" images. Figures 1 (c) and (d) show the activation pattern estimated by the RVM method for different values of the hyperparameters.

Two types of activation studies were performed, one where an activation was actually present, and one where there was no activation. The purpose of this was to simulate both null- and alternative-hypothesis conditions for purposes of measuring a ROC curve. Both types of activation studies were simulated 50 times to get a good statistical estimate of the ROC curve.

A statistical parametric map (SPM) was generated for each activation study and then the value of the pixel where the activation was sometimes present was thresholded to decide between the two hypotheses. The SPM was generated by computing the value of the likelihood ratio for the two hypotheses at each pixel of the image

$$\frac{p\left(g^{(s)}(x,y) \mid \hat{s}(x,y), H_1\right)}{p\left(g^{(s)}(x,y) \mid \hat{s}(x,y), H_0\right)} = \frac{\hat{s}(x,y)}{\sigma(x,y)/\sqrt{N}} \quad (23)$$

where $\hat{s}(x,y)$ is the estimate of the activation signal, and $\sigma(x,y)$ is the standard deviation of the imaging noise at location (x,y).

To evaluate and compare performance, we used the area under the ROC curve for false positive fraction between 0.0 and 0.1 because this is the most useful range

of operating points. A comparison is shown in TABLE I, with various methods listed in order of performance. These methods are described in detail in [4]. The RVM provided the second best performance in the simple case we considered and was only slightly worse than the RJMCMC approach in [3]. However, RJMCMC being a random sampling method is notoriously time consuming. Furthermore, it is difficult to establish when the chain has converged in order to stop sampling.

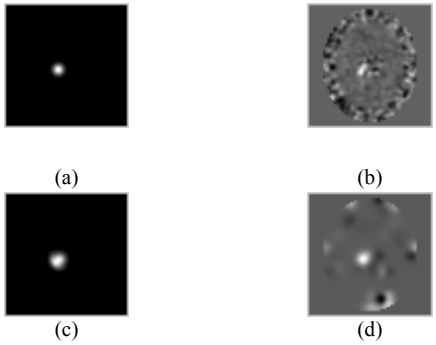

Figure 1. (a) - the average of 10 simulated activation patterns, (b) - the average of 10 "activated" images, (c) - the activation pattern estimated by RVM with α=1, β=0, (d) – the activation pattern estimated by RVM with $\alpha=10^{-2}$, β=0

An important problem with the application of the RVM model is the specification of the kernel characteristics (in our case the width of the Gaussian kernel) which are assumed to be known a priori. In order to specify an appropriate value for kernel width, we have used the leave-one-out cross validation criterion

$$\hat{\sigma}^2_{LOO} = \frac{\hat{y}^T P(diag(P))^{-2} P\hat{y}}{N}, \qquad (24)$$

where $P = I_N - \Phi\Sigma\Phi^T$ is known as the projection matrix. Experiment for several width values have shown that the use of cross-validation to select the appropriate kernel width in each experiment, leads to a slight increase in the measured ROC area, compared to using the same width for all experiments.

In the experimental results shown here, C_n was assumed known. We also found that the best detection performance occurred when using a completely uninformative prior, obtained by setting the hyperparameters $\alpha = \beta = 0$. However the resulting RVM fit is quite stable over a large range of values for the hyperparameters, see for example Fig. 1c and 1d.

6. CONCLUSIONS

In this paper we used the Relevance Vector Machine (RVM) methodology to estimate the activation signal in functional neuroimages. This approach is based on a fast incremental method that requires significantly less time compared to the original RVM formulation [6]. Initial experiments with simulated phantom data that the proposed algorithm outperforms all methods reported in [4] and yields very close performance to the RJMCMC based method reported in [3]. However, unlike RJMCMC where the number of activations is determined by searching, RVM provides a method for their incremental determination, therefore it is extremely faster.

TABLE I

PERFORMANCE COMPARISON

Method	Area under the ROC curve*
RJMCMC	0.0818
RVM	0.0732
SVD thresholding, column centering	0.0624
t-test, pooled variance estimate	0.0439
SVD thresholding, Fisher, row centering	0.0387
SVD thresholding, Fisher, column centering	0.0318
SVD thresholding, Fisher, double centering	0.0311
SVD thresholding, row centering	0.0252
t-test, single-pixel variance estimates	0.0242
SVD thresholding, double centering	0.0160

* For false positive fraction between 0 and 0.1.

REFERENCES

[1] Bowman A., Hall P. and Prvan T. "Bandwith selection for the smoothing of distribution functions", *Biometrika*, Volume 85, Issue 4, pp. 799-808, 1998.

[2] Carlin B., and Louis T., *Bayes and Empirical Bayes Methods for Data Analysis*, CRC Press; 2nd edition, 2000

[3] Lukic A. S., Wernick M. N., Galatsanos N. P., Yang Y., Strother S. C. "A Reversible Jump Markov Chain Monte Carlo Algorithm for Analysis of Functional Neuroimages" *IEEE International Conference on Image Processing*, Rochester, NY, September 2002.

[4] Lukic A. S., Wernick M. N., Strother S. C. "An evaluation of methods for detecting brain activations from functional neuroimages" *Artificial Intelligence in Medicine* 25(2002) 69-88.

[5] Strother S. C., Wernick M. N. "Technical Report: Deducing the Statistical Properties of Brain activation from Real Data for Use in Constructing Phantoms", www.ipl.iit.edu/IPL_papers/ipl_iit_101.pdf

[6] Tipping M. E., Faul A. "Fast Marginal Likelihood Maximisation for Sparse Bayesian Models" *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Jan 3-6, 2003.

[7] Tipping M. E. "Sparse Bayesian Learning and the Relevance Vector Machine" *Journal of Machine Learning Research* 1 (2001) 211-244.