31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Bayesian spatio - temporal approach for the analysis of fMRI data with non - stationary noise

Vangelis P. Oikonomou, Evanthia E. Tripoliti and Dimitrios I. Fotiadis

Abstract— **In this work, the bayesian framework is used for the analysis of fMRI data. The novelty of the proposed approach is the introduction of a spatio - temporal model used to estimate the variance of the noise across the images and the voxels. The proposed approach is based on a spatio - temporal version of Generalized Linear Model (GLM). To estimate the regression parameters of the GLM as well as the variance components of the noise, the Variational Bayesian (VB) Methodology is employed. The use of VB methodology results in an iterative algorithm, where the estimation of the regression coefficients and the estimation of variance components of the noise, across images and across voxels, are alternated in an elegant and fully automated way. The proposed approach is compared with the Weighted Least Squares (WLS) approach and both methods are evaluated on a real fMRI experiment.**

I. INTRODUCTION

Functional magnetic resonance imaging (fMRI) has been established as a method for mapping sensory, motor and cognitive functions to specific areas of the brain. It is based on the blood-oxygen level dependent (BOLD) effect [1]. As the conditions of the paradigm are alternated, the signal in the activated voxels increases and decreases according to the paradigm. The mapping of the voxel's signal leads to the production of a set of 3-D fMR images (fMRI dataset). fMRI analysis aims to detect the activated regions of the brain. In order to be achieved, fMRI must be processed in two steps: preprocessing and statistical analysis [1]. The purpose of the preprocessing is to remove the artefacts in the data, and to condition (prepare) the data, in order to maximize the sensitivity of the statistical analysis. Preprocessing includes: slice timing, motion correction, spatial normalization and spatial smoothing [1].

The statistical analysis aims to produce a statistical map which indicates those points in the brain that have been activated in response to the stimulus. Statistical analysis is carried out in three steps: (a) modelling the fMRI data (modelling the response to the stimulus and the random error), (b) estimating the parameters of the model, and (c) detecting the effect of interest (activation or no) using the estimated parameters. The first two steps are referred as estimation step and the third as detection step. For the estimation of the parameters several methods have been proposed in the literature that can be grouped in two broad categories: classical and Bayesian inference. The methods based on classical inference include the least squares (LS) and the maximum likelihood (ML) approach. Both approaches produce the same estimators under Gaussian errors. The main difference between classical and Bayesian approach concerns both the estimation and the detection step. More specifically, methods based on classical inference do not use any prior knowledge in the estimation procedure and the detection is based on statistical parametric maps (SPMs) produced using a statistical test (e.g. t-test). On the contrary, methods based on Bayesian inference use prior knowledge of the parameters and they detect effects of interest using posterior probability maps (PPMs). For a comparison between the two general detection approaches, in the context of fMRI data analysis, the reader can look in [2], [3].

The bayesian framework [4], [5] has been used in many works addressing several issues of fMRI data analysis. In [6] it is used to estimate the regression parameters of the GLM as well as to determine the activated regions of the brain. For the estimation of the regression parameters an uniformative prior was used. In [7] it is presented a bayesian approach, based on the VB methodology, for the estimation of the regression parameters of the GLM as well as for the estimation of the noise. In [8], [9] the bayesian framework is used to determine the design matrix in a flexible manner. This is achieved using a particular prior, which is called Automatic Relevance Determination (ARD) [10], [11]. Also, the bayesian framework has been used in the analysis of fMRI data using a spatio - temporal linear model [12], [13]. These works are concentrated mostly to the spatial distribution of the regression parameters rather to the noise.

As already mentioned, the bayesian framework can be used in the estimation step as well as in the detection step. In the proposed approach, the bayesian framework is used in both steps. First, in the estimation to obtain the posterior probability distribution of the parameters and then in the detection step the posterior distribution is used to produce the posterior probability maps (PPMs). The proposed work is concentrated mainly in the modelling of the noise which is observed in fMRI time series. The noise in a fMRI time series mainly consists of two components, the drift, which is a low frequency component, and a high frequency component. Many approaches have been proposed to deal with the drift component including temporal filtering [1], modelling the drift as an AR(1) process [7] or using an extended design matrix [8]. The extended design matrix contains the BOLD response and the first components from the Discrete Cosine Transform (DCT) [8] or polynomial

V. P. Oikonomou is with the Department of Computer Science, University of Ioannina, GR 45 110, Ioannina, Greece voikonom@cs.uoi.gr

E. E. Tripoliti is with the Department of Computer Science, University of Ioannina, GR 45 110 Ioannina, Greece evi@cs.uoi.gr

D. I. Fotiadis is with the Faculty of Materials Science and Engineering, University of Ioannina, GR 45 110, Ioannina, Greece fotiadis@cs.uoi.gr

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

terms [14]. The high frequency noise is assumed stationary and is usually modelled as white gaussian noise. However, in [15], [16] they have started to investigate the possibility that this component might be non stationary. This non - stationarity could have its origin in head movement as it is reported in [15].

II. METHODOLOGY

In fMRI, the signal intensity of N voxels is measured at time $t = 1, 2, \cdots, T$. Each voxel n is assumed to be a linear combination of effects (included in the design matrix) plus a noise term:

$$\mathbf{y}_n = \mathbf{X}\beta_n + \mathbf{e}_n, \tag{1}$$

where $\mathbf{y}_n$ is a Tx1 vector containing the fMRI time series, $\mathbf{e}_n$ is a Tx1 vector containing the noise, $\mathbf{X}$ is the Txp design matrix and β_n is a px1 vector of the regression parameters. In general, in fMRI analysis each voxel time series is analyzed independently from the others. We can collect all the voxels in one matrix NxT. This approach will help to use a spatio - temporal non - stationary model for the noise. Now the fMRI dataset can be described as:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{2}$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$ is a TxN matrix containing all the voxels, $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_N]$ is TxN matrix containing the noise and $\mathbf{B} = [\beta_1, \cdots, \beta_N]$ is a pxN matrix containing the regression parameters of all voxels. The design matrix $\mathbf{X}$ is the same for all voxels.

In our study we assume a spatio - temporal distribution for the noise. The non stationarity of the noise is expressed by two variance components, one variance component is responsible for the variance of each voxel and the other component is responsible for the variance of each image (slice of fMRI volume image). This spatio - temporal distribution results in a time varying variance model for the noise [15].

We assume a matrix Gaussian distribution for the noise given as:

$$p(\mathbf{E}) = N(0, \Omega^{-1}, \Sigma^{-1}). \tag{3}$$

The matrix Ω is a TxT diagonal precision matrix and each element in the main diagonal describes the precision (inverse variance) in each image (slice of fMRI volume image). The matrix Σ is a NxN diagonal precision matrix and each diagonal element describes the precision in each voxel. The distribution of the noise for the n-th voxel is a Guassian distribution given as:

$$p(\mathbf{e}_n) = N(0, (\sigma_n \Omega)^{-1}). \tag{4}$$

Each voxel is independent from the others, so the likelihood of the observations $\mathbf{Y}$ can be written as:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{B}, \Omega, \Sigma) = \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{X}, \beta_n, \sigma_n, \Omega). \tag{5}$$

In addition, the regression parameters are independent between the voxels. The probability distribution in that case is given as:

$$p(\mathbf{B}) = \prod_{n=1}^{N} p(\beta_n). \tag{6}$$

Each regression parameter in a voxel is independent to the others *a priori*. This assumption is included in the proposed model through the prior distribution, which is called the ARD prior [10] and is given as:

$$p(\beta_n|\mathbf{a}_n) = \prod_{p=1}^{P} p(\beta_{np}|a_{np}) = \prod_{p=1}^{P} N(0, a_{np}^{-1}). \tag{7}$$

The ARD prior is a hierarchical prior [5] and introduces into our model the parameters a_{np}. These parameters control the prior distribution of the regression parameters of the linear model. A gamma distribution [4] is used for each parameter a_{np}. So, the overall hyperprior is given as:

$$p(\mathbf{a}_n) = \prod_{p=1}^{P} \Gamma(a_{np}; b_{np}, c_{np}). \tag{8}$$

The gamma distribution is given as:

$$\Gamma(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\{-\frac{x}{b}\}, \tag{9}$$

where b and c are the shape and scale parameters of the gamma distribution.

Also, in the proposed model the precision component of each image $\{\omega_1, \omega_2, \cdots, \omega_T\}$ and the precision component of each voxel $\{\sigma_1, \sigma_2, \cdots, \sigma_N\}$ must be estimated. This means that we must place a prior distribution over each precision component. The prior distribution that is often used for a precision component is the gamma distribution. So, the prior over each precision component for each voxel is given as:

$$p(\sigma_n) = \Gamma(\sigma_n; b_{\sigma_n}, c_{\sigma_n}), n = 1, \cdots, N, \tag{10}$$

and the prior over each precision component for each image is given as:

$$p(\omega_t) = \Gamma(\omega_t; b_t^{(\omega)}, c_t^{(\omega)}), t = 1, \cdots, T. \tag{11}$$

The overall prior according to our assumptions is given as:

$$p(\mathbf{B}, \{\omega_t\}_{t=1}^{T}, \{\sigma_n\}_{n=1}^{N}, \{\mathbf{a}_n\}_{n=1}^{N}) = \prod_{n=1}^{N} p(\beta_n|\mathbf{a}_n) p(\mathbf{a}_n) \cdot \prod_{n=1}^{N} p(\sigma_n) \cdot \prod_{t=1}^{T} p(\omega_t). \tag{12}$$

The VB methodology is an approximate method for bayesian inference [17]. In the bayesian inference we are interested to obtain the posterior distribution of the parameters, in our study the posterior distribution of the regression parameters, the parameters a_{np} and the precision components. However, the posterior distribution is not always tractable and we need approximation techniques to obtain it. One such technique is the VB methodology. The general idea on the VB methodology is that the true posterior is difficult to be obtained in close form, so it is approximated with another distribution which produces a closed form solution. A usefull approximation is to assume that *a posteriori* each parameter

is independent of the others. In our study, the following approximation is used:

$$p(\mathbf{B},\{\omega_t\}_{t=1}^T,\{\sigma_n\}_{n=1}^N,\{\mathbf{a}_n\}_{n=1}^N|\mathbf{Y}) \approx \prod_{n=1}^N q(\beta_n|\mathbf{a}_n)q(\mathbf{a}_n)\cdot \prod_{n=1}^N q(\sigma_n)\cdot\prod_{t=1}^T q(\omega_t). \quad (13)$$

Applying the VB methodology we obtain the following solutions:

$$\begin{aligned}
q(\beta_n) &= N(\hat{\beta}_n, C_{\beta_n}), n=1,\cdots,N, && (14)\\
q(\mathbf{a}_n) &= \prod_{p=1}^P \Gamma(a_{np}; b'_{np}, c'_{np}), n=1,\cdots,N, && (15)\\
q(\sigma_n) &= \Gamma(\sigma_n; b'_{\sigma_n}, c'_{\sigma_n}), n=1,\cdots,N, && (16)\\
q(\omega_t) &= \Gamma(\omega_t; b'_{\omega_t}, c'_{\omega_t}), t=1,\cdots,T, && (17)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{C}_{\beta_n} &= (\hat{\sigma_n}\mathbf{X}^T\hat{\Omega}\mathbf{X}+\hat{\mathbf{A}}_n)^{-1}, && (18)\\
\hat{\beta}_n &= (\hat{\sigma_n}\mathbf{X}^T\hat{\Omega}\mathbf{X}+\hat{\mathbf{A}}_n)^{-1}\hat{\sigma}_n\mathbf{X}^T\hat{\Omega}\mathbf{y}_n, && (19)\\
\frac{1}{b'_{np}} &= \frac{1}{2}(\hat{\beta}^2_{np}+\mathbf{C}_{\beta_n}(p,p))+\frac{1}{b_{np}}, && (20)\\
c'_{np} &= \frac{1}{2}+c_{np}, && (21)\\
\hat{a}_{np} &= b_{np}c_{np}, && (22)\\
\frac{1}{b'_{\sigma_n}} &= \frac{1}{2}(\mathbf{y}_n-\mathbf{X}\beta_n)^T\hat{\Omega}(\mathbf{y}_n-\mathbf{X}\beta_n) \\
&\quad +tr(\mathbf{X}^T\hat{\Omega}\mathbf{X}\mathbf{C}_{\beta_n})+\frac{1}{b_{\sigma_n}}, && (23)\\
c'_{\sigma_n} &= \frac{T}{2}+c_{\sigma_n}, && (24)\\
\hat{\sigma}_n &= b'_{\sigma_n}c'_{\sigma_n}, && (25)\\
\frac{1}{b'_{\omega_n}} &= \frac{1}{2}(\mathbf{y}_t^T\hat{\Sigma}\mathbf{y}_t-2\mathbf{y}_t^T\hat{\Sigma}\hat{B}\mathbf{x}_t+\mathbf{x}_t^T G\mathbf{x}_t)+\frac{1}{b_{\omega_n}}, && (26)\\
c'_{\omega_t} &= \frac{N}{2}+c_{\omega_t}, && (27)\\
\hat{\omega}_t &= b'_{\omega_t}c'_{\omega_t}. && (28)
\end{aligned}$$

In the above equations the matrices $\hat{\mathbf{A}}_n, n=1,\cdots,N$ are pxp diagonal matrices having the parameters $\hat{a}_{n1},\hat{a}_{n2},\cdots,\hat{a}_{np}$ in the main diagonal. The matrix $\hat{\Sigma}$ is a NxN diagonal matrix containing in the main diagonal the mean of precision components for each voxel, $\hat{\sigma}_n, n=1,...,N$, and the matrix $\hat{\Omega}$ is a TxT diagonal matrix containing in the main diagonal the mean of precision components for each image, $\hat{\omega}_t, t=1,...,T$. The quantity G is calculated as follows:

$$G=\sum_{n=1}^N \sigma_n(\mathbf{C}_{\beta_n}+\beta_n\beta_n^T). \quad (29)$$

The algorithm consists of the iterative application of equations (18)-(29). First, the equations (18)-(25) and (29) are applied over all voxels to obtain the estimates of the regression parameters and the precision component for each voxel. Also, in this step the quantity G is calculated. Then, the equations (26)-(28) are applied to estimate the precision component of each image. In this step we use the estimated regression parameters obtained in the previous step.

The above algorithm is applied until convergence of the variational bound or until the convergence of the parameters. The variational bound can be calculated from:

$$\begin{aligned}
F(q,\theta) &= <\log p(\mathbf{Y}|\mathbf{X},\mathbf{B},\{\omega_t\}_{t=1}^T,\{\sigma_n\}_{n=1}^N>_{q(\theta)} \\
&\quad -KL(q(\theta)||p(\theta)), && (30)
\end{aligned}$$

where $\theta=\{\mathbf{B},\{\omega_t\}_{t=1}^T,\{\sigma_n\}_{n=1}^N,\{a_n\}_{n=1}^N\}$. The Variational bound is the difference between the expected likelihood and the KL divergence between the approximate posterior and the prior of parameters.

III. RESULTS

We compare the proposed approach with the Weighted Least Squares (WLS) approach using real fMRI data. The scaling matrix of the WLS approach is estimated using the residuals of the LS approach [15]. The fMRI dataset have been downloaded from http://www.fil.ion.ucl.ac.uk/spm/. A healthy volunteer participated in the fMRI experiment, where an auditory stimulus was used. More specifically, bi - syllabic words presented binaurally at a rate of 60 words per minute. During the experiment 96 3-D fMRI volume images were acquired. The acquisition was made in blocks of 6 and lasted 6.05sec. The blocks alternated between two conditions: rest and auditory stimulus. The size of the fMRI volume images was 64x64x64 and the voxel size was 3mm x 3mm x 3mm. The fMRI volume images were acquired on a 2T Siemens Magnetom. Due to T1 effect the first two blocks were discarded. After preprocessing, the size fMRI volume images was 79x95x68 and the voxel size was 2mm x 2mm x 2mm.

The design matrix that was used in order to model the fMRI experiment consisted of 84 rows (one for each observation) and 5 columns. The first 3 columns contain the regressors that model the drift term in the fMRI time series using polynomial terms, and the other 2 columns contain the regressors for the BOLD response and a constant mean value, i.e. $\mathbf{X}=[\mathbf{t}\ \mathbf{t}^2\ \mathbf{t}^3\ \mathbf{s}\ \mathbf{1}]$, where $\mathbf{t}=[\frac{1}{T},\frac{2}{T},\cdots,1]$, $\mathbf{s}$ is the BOLD response of fMRI experiment and $\mathbf{1}$ is a Tx1 column vector of 1s to model the mean. For the initialization of the algorithm we set the shape and scale of each gamma distribution to 10^6 and 10^{-6}, respectively.

In Fig. 1, the parameter of the linear model responsible for the BOLD response is depicted using the proposed approach and the WLS approach, as well as the histograms of these. It is evident in this time serie the use of ARD prior. The ARD prior is a sparse prior, and this sparseness can be observed in the histogram of the regression parameter responsible for the BOLD response. In Fig. 2 we depict the posterior probability maps (PPMs) using the proposed approach and the WLS approach. To produce these maps the contrast vector $\mathbf{c}=[0\ 0\ 0\ 1\ 0]$ is used. Also, PPMs showing the presence of the drift in a particular voxel are depicted. To produce these maps the contrast vector $\mathbf{c}=[1\ 1\ 1\ 0\ 0]$ is used. The PPMs for the drift show significant difference between the two approaches. More specifically, the drift, using the proposed approach, is

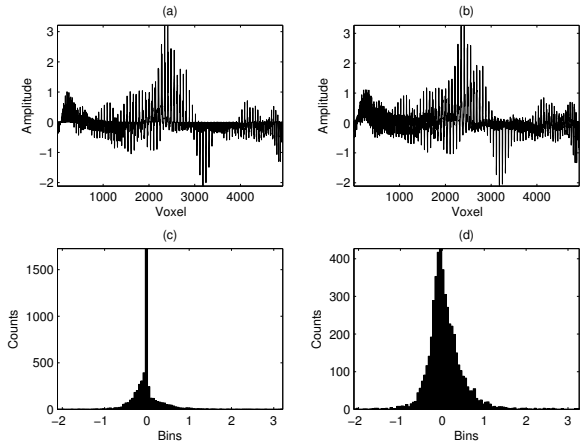


Fig. 1. Time series of the regression parameter for BOLD response using: (a) the proposed approach, and (b) WLS approach, and its histogram using: (c) the proposed approach, and (d) the WLS approach.

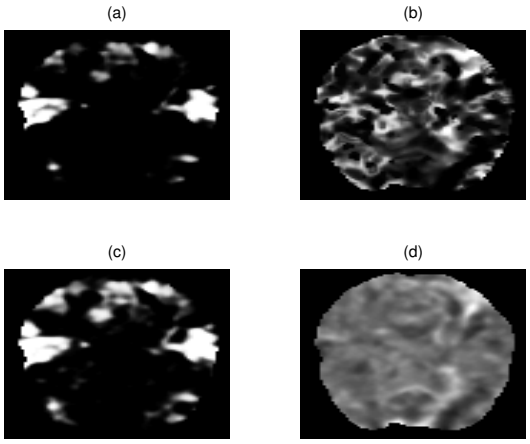


Fig. 2. PPMs of slice 28. (a) Activation of BOLD response using the proposed approach, (b) Drift presence using the proposed approach, (c) Activation of BOLD response using the WLS approach, (d) Drift presence using the WLS approach.

presented in areas of the occipital and temporal lobe while the drift, using WLS, is presented in areas of the occipital and frontal lobe. The presence of drift in a voxel means that there is high correlation for the samples of this voxel. In addition, in the proposed approach both brain hemispheres show high correlation. On the contrary, the WLS approach results in high correlation only in the right hemisphere.

In Fig. 3 the variance in each image, produced by the proposed approach and by the residuals of the LS approach, is depicted. It is obvious that there is high correlation between the beginning and the ending of the stimulus and the increase of variance in those images. For example we can observe that the beginning of the stimulus at time $t = 49$ produces an increase of the variance at the particular image. This finding justifies the assumption of non stationary noise.

In Fig. 4 the results of the clustering the PPMs is demonstrated. The clustering was performed using the k-means algorithm, and the number of clusters was set equal to two. In these images the white regions show the activated areas of the brain, while the gray regions show the non activated areas of the brain. Both approaches present significant activation on the auditory cortex. However, we observe that the WLS presents extended activation in brain areas which are not related to the experiment. Also, based on the clustering procedure we show in Fig. 5, the averaged reconstructed signal and the averaged acquired signal for both approaches and both situations (activated, non - activated). We can observe that the temporal courses of these signals show a clear activation.

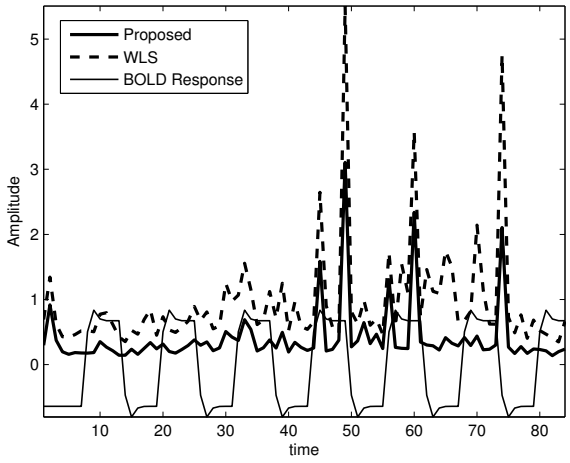


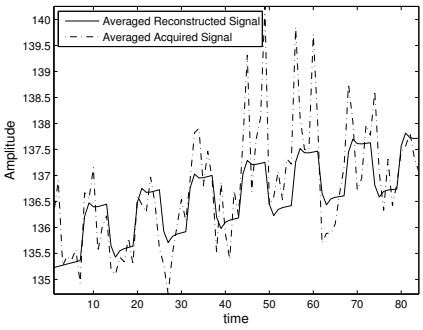
Fig. 3. Variance components for each image and the BOLD response.

(a) Clustering results using the proposed approach (b) Clustering results using the WLS approach

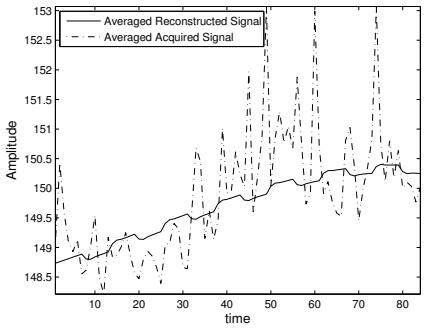
Fig. 4. Grayscale images showing the clusters of PPMs. The white color shows the activated areas and the gray color shows the non activated areas.

IV. CONCLUSIONS

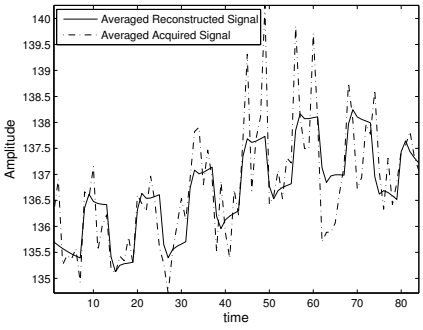
We have presented a bayesian approach for the statistical analysis of fMRI data. The proposed approach concentrates mostly on modelling the non-stationary nature of the noise. The non-stationarity of the noise is due to the head movement or due to the reaction of the subject to a stimulus. The noise is modeled using a spatio - temporal model. In this model



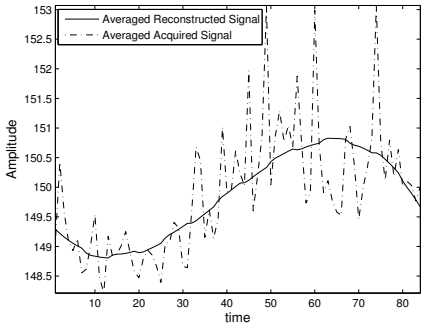
(a) Averaged Reconstructed signal using the proposed approach and the Averaged Acquired signal (activation case).



(b) Averaged Reconstructed signal using the proposed approach and the Averaged Acquired signal (deactivation case).



(c) Averaged Reconstructed signal using the WLS approach and the Averaged Acquired signal (activation case).



(d) Averaged Reconstructed signal using the WLS approach and the Averaged Acquired signal (deactivation case).

Fig. 5. Averaged Reconstructed and Acquired signals for activation and non activation cases.

the noise is decomposed in two component, one component is responsible for the variance of noise in each voxel and the other component is responsible for the variance of noise in each image (slice of fMRI volume). To analyze the fMRI data a spatio - temporal version of GLM is used. In the regression parameters we use as prior the ARD prior which tends to give us sparse estimates. The assumption of sparseness is valid if we expect during an experiment the activated areas of the brain spread in small regions and not into the whole brain.

REFERENCES

[1] P. Jezzard, P. M. Matthews, and S. M. Smith, *Functional MRI: An Introduction to Methods*. Oxford University Press, USA, 2001.

[2] K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and bayesian inference in neuroimaging: Theory," *NeuroImage*, vol. 16, pp. 465–483, June 2002.

[3] K. J. Friston and W. Penny, "Posterior probability maps and SPMs," *NeuroImage*, vol. 19, pp. 1240–1249, July 2003.

[4] G. Box and G. Tiao, *Bayesian inference in statistical analysis*. John Wiley and Sons, Inc, 1973.

[5] B. Charlin and T. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*. New York, NY: CRC Press, 2000.

[6] J. Kershaw, B. Ardekani, and I. Kanno, "Application of bayesian inference to fmri data analysis," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 1138–1153, Dec 1999.

[7] W. Penny, S. Kiebel, and K. Friston, "Variational bayesian inference for fmri time series," *NeuroImage*, vol. 19, pp. 727–741, July 2003.

[8] H. Luo and S. Puthusserypady, "A sparse bayesian method for determination of flexible design matrix for fmri data analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, pp. 2699–2706, Dec. 2005.

[9] V. P. Oikonomou, E. E. Tripoliti, and D. I. Fotiadis, "A sparse variational bayesian approach for fmri data analysis," *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, pp. 1–6, Oct. 2008.

[10] D. J. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, pp. 415–447, 1992.

[11] C. Bishop and M. Tipping, "Variational relevance vector machines," *Proc. 16th Conf. Uncertainty in Artificial Intelligence*, pp. 46–53, 2000.

[12] M. Woolrich, M. Jenkinson, J. Brady, and S. Smith, "Fully bayesian spatio-temporal modeling of fmri data," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 213–231, Feb. 2004.

[13] W. D. Penny, N. J. Trujillo-Barreto, and K. J. Friston, "Bayesian fmri time series analysis with spatial priors," *NeuroImage*, vol. 24, pp. 350–362, Jan. 2005.

[14] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales, and A. C. Evans, "A general statistical analysis for fmri data," *NeuroImage*, vol. 15, no. 1, pp. 1 – 15, 2002.

[15] J. Diedrichsen and R. Shadmehr, "Detecting and adjusting for artifacts in fmri time series data," *NeuroImage*, vol. 27, no. 3, pp. 624 – 634, 2005.

[16] H. Luo and S. Puthusserypady, "fMRI data analysis with nonstationary noise models: A bayesian approach," *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 1621–1630, Sept. 2007.

[17] M. Beal, *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, Univ. College London, London, U.K., 2003.