Automatic HyperParameter estimation in fMRI

David Afonso, Patrícia Figueired, and J. Miguel Sanches*

Institute For Systems and Robotics, Instituto Superior Técnico,
Lisboa, Portugal

Abstract. *Maximum a posteriori* (MAP) in the scope of the Bayesian framework is a common criterion used in a large number of estimation and decision problems. In image reconstruction problems, typically, the image to be estimated is modeled as a Markov Random Fields (MRF) described by a Gibbs distribution. In this case, the Gibbs energy depends on a multiplicative coefficient, called *hyperparameter*, that is usually manually tuned [13] in a trial and error basis.
In this paper we propose an automatic *hyperparameter* estimation method designed in the scope of *functional Magnetic Resonance Imaging* (fMRI) to identify activated brain areas based on *Blood Oxygen Level Dependent* (BOLD) signal.
This problem is formulated as classical binary detection problem in a Bayesian framework where the estimation and inference steps are joined together. The prior terms, incorporating the a priori physiological knowledge about the HRF, drift and spatial correlation across the brain (using edge preserving priors), are automatically tuned with the new proposed method.
Results on real and synthetic data are presented and compared against the conventional General Linear Model (GLM) approach.

Keywords: HyperParameter, Estimation, Bayesian, fMRI, HRF

1 Introduction

The detection of neuronal activation based on BOLD signals measured using fMRI is one of the most popular brain mapping techniques.

The data are classically analysed using the Statistical Parametric Mapping (SPM) technique [10, 9] where the *General Linear Model* (GLM) is used to describe the observations at each voxel and the corresponding coefficients are estimated by using the Least Square Method [13]. The detection of the activated regions is performed from the estimated *explanatory variables* (EV's) by using a second step of classical inference approach based on the Neyman-Pearson theorem.

Still, Bayesian approaches have been gaining popularity since they provide a formal method of incorporating prior knowledge in data analysis [5, 11]. In

* This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds.

[7], Groutte et al. propose a non-parametric approach where a *Finite Impulse Response* (FIR) filter is used to describe the Hemodynamic Response Function (HRF) and smoothing constraints are imposed at the solution by using a regularization matrix. Ciuciu et al. describe another non-parametric approach for the Bayesian estimation of the HRF in [3]. The authors make use of temporal prior terms to introduce physiological knowledge about the HRF. Basic and soft constraints are incorporated in the analysis, namely the considerations that the HRF starts and ends at zero and that the HRF is a smooth function. In [15] the authors propose a Bayesian approach in which the data noise is estimated using a spatio-temporal model and propose a half-cosine functions HRF model based on their experimental findings. In Yet another Bayesian approach based on the mathematical formalism of the GLM is proposed in [1]. The authors describe an SPM algorithm based on the maximum a posteriori (MAP) criterion to jointly estimate and detect the activated brain areas characterized by binary coefficients. The prior term introduced for these parameters comprises a bimodal distribution defined as the sum of two Gaussian distributions centered at zero and one.

In this paper we further improve this last method [1] by implementing an automatic HyperParameter estimation method to automatically set the prior strength in the Bayesian estimation, by constraining it to a probability density function. Additionally, data drift estimation is incorporated and the spatial correlation between neighbors is taken into account by using *edge preserving* priors that promote piecewise constant region solutions. The optimization of the overall energy function with respect to the activation binary variables is performed by using the *graph-cuts* (GC) based algorithm described in [2], which is computationally efficient and is able to find out the global minimum of the energy function.

2 Problem Formulation and Method

By making use of the problem formulation and variables defined in [1] we further incorporate a slow time data drift variable (a.k.a. baseline) $\mathbf{d_i}$ of time dimension N into the observation model, yielding eq. 1 at each i^{th} voxel, when L stimuli are applied simultaneously.

$$y_i(n) = \overbrace{h_i(m) * \underbrace{\sum_{k=1}^{L} \beta_i(k) p_k(n)}_{x_i(n)}}^{z_i(n)} + d_i(n) + \eta_i(n), \tag{1}$$

where $y_i(n)$ is the N length observed BOLD signal at the i^{th} voxel, $h_i(m)$ is the $M \leq N$ dimensional HRF, and $\eta_i(n)$ is noise signal. The activity unknowns $\beta_i(k) \in \{0, 1\}$ are binary with $\beta_i(k) = 1$ if the i^{th} voxel is activated by the k^{th} $p_k(n)$ stimulus. $\eta_i(n) \sim \mathcal{N}(0, \sigma_y^2)$ is assumed *Additive White Gaussian Noise*

(AWGN) which is an acceptable assumption mainly if a prewhitening preprocessing [8] of the data is performed.

The *maximum a posteriori* (MAP) estimation of the unknown vectors $\mathbf{b}_i$, $\mathbf{h}_i$ and $\mathbf{d}_i$ is obtained, in matrix form, by minimizing the following energy function

$$\begin{aligned} E(\mathbf{y}_i, \mathbf{b}_i, \mathbf{h}_i, \mathbf{d}_i) &= \overbrace{E_y(\mathbf{y}_i, \mathbf{b}_i, \mathbf{h}_i, \mathbf{d}_i)}^{\text{Data fidelity term}} + \overbrace{E_h(\mathbf{h}_i) + E_d(\mathbf{d}_i)}^{\text{Prior terms}} \\ &= -\log(p(\mathbf{y}_i|\mathbf{h}_i, \mathbf{h}_i,)) - \log(p(\mathbf{h}_i)) - \log(p(\mathbf{d}_i)) \end{aligned} \tag{2}$$

where the prior terms incorporate the *a priori* knowledge [12] about the temporal behavior of $\mathbf{h}_i$ and $\mathbf{d}_i$ - the HRF ($C1$) starts and ends at 0; ($C2$) is smooth; ($C3$) has similar magnitude to the HRF one gamma function ($h^c(t)$) proposed in [4, 8]; and that the $\mathbf{d}_i$ is ($C4$) a slow varying signal with a smaller *bandwidth* than the one of $\mathbf{h}_i$.

By the *Hammersley-Clifford* theorem [6] and *Markov Random Fields* theory, these constraints may be imposed in the form of the following *Gibbs distributions*

$$p(\mathbf{h}_i) = \frac{1}{Z_h} e^{-\alpha U(\mathbf{h}_i)} \tag{3}$$

$$p(\mathbf{d}_i) = \frac{1}{Z_d} e^{-\gamma U(\mathbf{d}_i)} \tag{4}$$

where Z_h and Z_d are partition functions and the *Gibbs* energies $U(\mathbf{h}_i)$ and $U(\mathbf{d}_i)$ are designed in the following way, where $[\alpha, \gamma]$ are regularization parameters to tune the degree of smoothness of the estimated vectors.

$$U(\mathbf{h}_i) = \overbrace{w_h(1)}^{\text{C3}} \underbrace{h(1)^2}_{\text{C1}} + \overbrace{w_h(M)}^{\text{C3}} \underbrace{h(M)^2}_{\text{C1}} + \underbrace{\sum_{n=2}^{M-1} \overbrace{w_h(n)}^{\text{C3}} \left[\overbrace{(h_i(n+1) - h_i(n)) - (h_i(n) - h_i(n-1))}^{\text{Discrete version of the } 2^{nd} \text{ derivative}} \right]^2}_{\text{C2}} \tag{5}$$

$$U(\mathbf{d}_i)_i = \underbrace{\sum_{n=2}^{N} \left[\overbrace{d_i(n) - d_i(n-1)}^{\text{Discrete version of the } 1^{st} \text{ derivative}} \right]^2}_{\text{C4}} \tag{6}$$

Here the weigh coefficients $w_h(n) = 1/(|h^d(n)| + 10^{-6})^2$, where the discrete version of the HRF gamma function is $h^d(n) = h^c(t)|_{t=n \times TR}$, are used to compensate for the reduced prior strength when the second derivatives are small.

Its can be shown that the overall energy eq. (2) is rewritten as follows

$$E_y = \frac{1}{2\sigma_y^2} \|(\mathbf{\Psi}_i \mathbf{b}_i + \mathbf{d}_i - \mathbf{y}_i)\|_2^2 + \alpha \mathbf{h}_i^T \mathbf{H}_0 \mathbf{D}_h \mathbf{h}_i + \gamma \mathbf{d}_i^T \mathbf{D}_d \mathbf{d}_i \tag{7}$$

where $\mathbf{H_0} = diag\{h^d(n)\}$ is a $M \times M$ diagonal matrix containing the HRF. $\mathbf{D}_h$ and $\mathbf{D}_d$ are $M \times M$ and $N \times N$ second and first order difference matrix operators [13], respectively. $\mathbf{\Psi}_i$ is a toeplitz $L \times N$ convolution matrix of $\mathbf{b}_i$ and $\mathbf{b}_i$ as defined in [1]

3 Optimization

The MAP estimation of the unknown vectors $\mathbf{b}_i$, $\mathbf{h}_i$ and $\mathbf{d}_i$ is obtained by minimizing the energy function (7) with respect to each vector, one step at a time. $\mathbf{b}_i$ is first estimated with the drift initialized with the mean of the TC ($\mathbf{d}_i^0 = \bar{\mathbf{y}})_i$ and $\mathbf{h}_i^0$ equal to $h^d(n)$ [4, 8].

3.1 Step One: b estimation

Its easily shown that the binary elements of $\hat{\mathbf{b}}_i^t = \{\hat{\beta}_{k,i}^t\}$ that leads to the minimization of (7) are a simple binarization by thresholding ($thrs = 0$) of the following fields, in matrix notation, where $\psi_i^{t-1}(k)$ is the k^{th} column of $\mathbf{\Psi}_i^{t-1}$:

$$\mathcal{B}_i^t(k) = -\psi_i^{t-1}(k)^T \left[\psi_i^{t-1}(k) + 2(\mathbf{d}^{t-1} - \mathbf{y})\right] \tag{8}$$

To solve this huge combinatorial problem, a fast and computationally efficient *graph-cuts* based algorithm [2] is used to binarize the fields $\mathcal{B}_{r,l}^t(k)$, defined in (8), at each (r, l) pixel location in the data slice, by minimizing the following energy function:

$$\mathbf{\Sigma}(\beta_{r,l}(k), \mathcal{B}_{r,l}(k)) = \overbrace{\sum_{r,l} \mathcal{B}_{r,l}(k)(1 - \beta_{r,l}(k))}^{\text{data fidelity term}} + \underbrace{\hat{\sigma_y^2} \sum_{r,l} \left[V_{r,l}^v(k) + V_{r,l}^h(k)\right] / \tilde{g}_{r,l}}_{\text{spatial regularization term}} \tag{9}$$

where $\hat{\sigma_y^2}$ is the observed signal variance ($var(\mathbf{y})$); $V_{r,l}^v(k)$ and $V_{r,l}^h(k)$ are XOR $\oplus$ operators between $\beta_{r,l}(k)$ and its causal vertical $\beta_{r,l+1}(k)$ and horizontal $\beta_{r+1,l}(k)$ neighbors, respectively; $\tilde{g}_{r,l}$ ($10^{-2} \leq \tilde{g}_{r,l} \leq 1$) is the normalized (smoothed) filtered gradient of $\mathcal{B}(k)$.

Non-uniform solutions to (9) have a higher cost due to the *spatial regularization term*. However, in order to preserve transitions, the division by $\tilde{g}_{r,l}$ reduces this non-uniform cost at locations where the gradient magnitude is large.

3.2 Step Two: h estimation

A new estimative of $\hat{\mathbf{h}}_i$ is calculated by finding the null derivative point of (2) with respect to $\mathbf{h}$, yelding:

$$\hat{\mathbf{h}}_i = \left[(\mathbf{\Phi}_i^t)^T \mathbf{\Phi}_i^t + 2\alpha\sigma_y^2 \mathbf{H_0} \mathbf{D}_h^T\right]^{-1} (\mathbf{\Phi}_i^t)^T (\mathbf{y}_i - \mathbf{d}_i^{t-1}) \tag{10}$$

where $\mathbf{\Phi}_i^t$ is calculated with the current $\mathbf{b}_i^t$ vector, estimated at the previous iteration step 3.1. However, the HRF is only estimated in the case of voxel activation by at least one paradigm, i.e., if $\exists_{(r,l)} : \hat{\beta}_{r,l}(k) > 0$.

HyperParameter estimation In this method the regularization parameter α is not constant but is automatically and adaptively estimated along the iterative process as follows. Considering (5), we can rewrite eq. (3) as

$$p(\mathbf{h}_i) = \prod_n^N \underbrace{\frac{1}{Z_h} e^{-\alpha w(n)\delta(n)^2}}_{p(\delta)} \tag{11}$$

where δ^2 is the 2^{nd} derivative operator in (5). By assuming $p(\delta(n))$ to be a probability density function (of unitary area) and $\alpha w(n) = \frac{1}{2\sigma(n)^2}$ we get

$$\frac{\sqrt{2\pi\sigma(n)^2}}{Z_h(n)} \underbrace{\int_{-\inf}^{+\inf} \frac{1}{\sqrt{2\pi\sigma(n)^2}} e^{\frac{\delta(n)^2}{2\sigma(n)^2}} d\delta(n)}_{=1} = 1 \tag{12}$$

which implies that $Z_h(n) = \sqrt{\frac{\pi}{\alpha w(n)}}$, hence the energy term of (2) with respect to h can be rewritten as

$$\mathbf{E}_h(h_i) = -\log(h) = \frac{N}{2}\log\pi - \frac{1}{2}\sum_n^N \log w(n) - \frac{N}{2}\log\alpha + \alpha\sum_n^N w(n)\delta(n)^2 \tag{13}$$

By finding the null derivative of (13) we obtain the automatic HyperParameter estimation that is, in each iteration, dependent on the initialization and current estimate of the HRF.

$$\alpha^t = \frac{N}{2U(h)} = \frac{\frac{N}{2}}{(\mathbf{h}_i^{t-1})^T(\mathbf{H_0}\mathbf{D}_h)\mathbf{h}_i^{t-1}} \tag{14}$$

3.3 Step Three: d estimation

A new estimative of $\hat{\mathbf{d}}_i$ is calculated by finding the null derivative point of (2) with respect to $\mathbf{d_i}$, yelding:

$$\hat{\mathbf{d}}_i = \left[\mathbf{I} + 2\gamma\sigma_y^2\mathbf{D}_d^T\right]^{-1}(\mathbf{y}_i - \mathbf{\Psi}_i^t\mathbf{b}_i^t) \tag{15}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{\Psi}_i^t$ is computed by using the current $\hat{\mathbf{h}}_i^t$ vector, obtained in the previous iteration step.

Since γ is a regularization parameter associated with the drift signal, a much slower frequency signal than HRF, then γ should be higher than α, i.e., $\gamma \gg \alpha$ [14]. Here $\gamma = 100\alpha$.

4 Experimental Results

In this section tests with synthetic and real data are presented to illustrate the application of the algorithm and evaluate its performance. The real data, acquired in ?????, were kindly provided by Prof[a]. Patrcia Figueiredo from Institute for Systems and Robotics at the Instituto Superior Técnico.

4.1 Synthetic Data

The synthetic data is based on the well known *Shepp-Logan* image phantom with 256×256 pixels. The paradigm was generated in a block-design basis with 4 epochs, 60 *sec* each (30 *sec* of activation and 30 *sec* of rest) with $TR = 3$ *sec*.

Making use of (1) for generating the $\mathbf{y}_i$ observed data, a Monte Carlo experiment with a total of $3,276,800$ runs was performed with several different noise levels (see

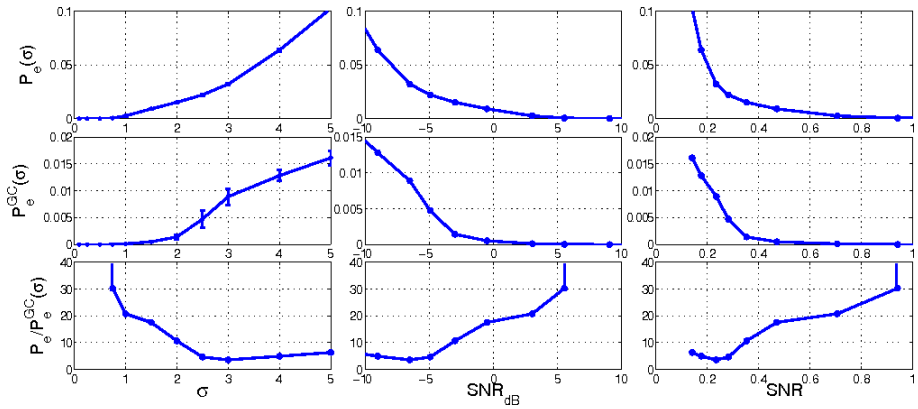

Fig. 1. Monte Carlo results for 50 runs on 256×256 pixels, for $\sigma = \{0.1,\ 0.25,\ 0.5,\ 0.75,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 4,\ 5\}$. The mean and standard deviation (error bars) values of P_e, as a function of σ, SNR_{dB} and SNR, are presented on the first, second and third columns, respectively. The results with and without GC are displayed on the top and middle rows, respectively, and the ratio $[P_e(GC)/P_e(w/GC)]$ (σ is displayed on the bottom row.

Fig. 1 captation). The resulting mean and standard deviation (error bars) values of probability of error (P_e), as a function of σ, SNR_{dB} and SNR, are presented in Fig.1. The results with and without GC are shown, as well as the ratio $P_e(GC)/P_e(w/GC)$.

These results demonstrate the activity detection robustness of the method, even in highly noisy data. They also show that taking into account the spatial correlation among neighboring voxels leads to a significant decrease in P_e. As expected, the improvement increases when the amount of noise increases or, equivalently, the SNR decreases. This is observed by the monotonic increasing behavior of the $[P_e(GC)/P_e(w/GC)]$ (σ).

4.2 Real Data

Two volunteers with no history of psychiatric or neurological diseases participated in a visual stimulation and a motor task fMRI experiment. Functional images were obtained using echo-planar imaging (EPI) with $TR/TE = 2000\,ms/50\,ms$. Datasets were pre-processed and analyzed using the FSL software (http://www.fmrib.ox.ac.uk/fsl) for: motion correction; non-brain removal and mean-based intensity normalization. The data used in the standard SPM-GLM analysis using FSL (not for the proposed SPM-Drift-GC) was further pre-processed with spatial smoothing (Gaussian kernel, $5\,mm$ FWHM) and high-pass temporal filtering (Gaussian-weighted least squares straight line fitting, $50\,sec$ cut-off).

For the FSL processing a GLM approach with local autocorrelation correction was used on square stimulus functions convolved with the canonical Gamma HRF and it's first derivative [4, 8]. Linear stimulus/baseline contrast analysis and t-tests are applied to obtain the SPM, followed by cluster thresholding by the Gaussian Random Fields (GRF) theory. Since the results provided by this "'standard"' method are depend on the inference p-value and clustering Z-score threshold values used, our experienced

experimentalist provided two results of SPM-GLM: a relatively *strict* result and a more *loose* result, displayed on columns *d*) and *e*) of Fig. 2 respectively. The proposed SPM-Drift-GC method results are displayed on columns *a*), *b*) and *c*) of Fig. 2

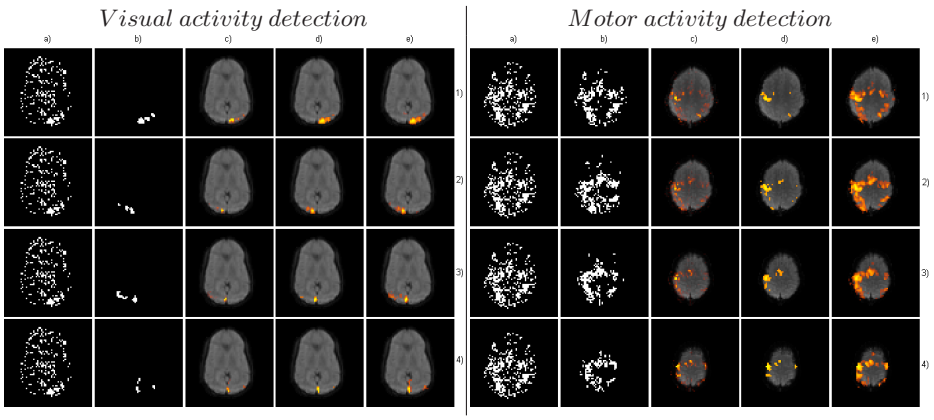

Fig. 2. Activated regions obtained by the new SPM-Drift-GC (a-b-c) and standard SPM-GLM (d-e) methods, on the visual (lef) and motor (right) real data, where each row *(1), 2), 3)* and *4))* corresponds to a different stimulus. Left to right: *a*) Binary SPM-Drift algorithm results without *GraphCuts*; *b*) Binary SPM-Drift-GC algorithm results; *c*) Weighted SPM-Drift-GC algorithm results; *d*) SPM-GLM algorithm *Strict* results; *e*) SPM-GLM algorithm *Loose* results. Activation intensity is color coded from red (0) to yellow (1) and is overlaid on the EPI brain image with linearly decreasing transparency from 100% ($activity = 0$) to 0% ($activity \geq 0.5$).

In general, visual inspection of the activation brain maps suggests good agreement between the methods, although the SPM-Drift-GC also detects some regions not present in the *strict* results, but present, most of them, in the *loose* results. However, in some brain slices, there are areas only detected as active by SPM-Drift-GC that correspond to low energy estimated HRF's (coded in transparent red) and somewhat deviant shaped HRF's from the rigid HRF restrictions of SPM-GLM.

5 Conclusions

In this paper, a new Bayesian parameter-free method to detect activated brain areas in fMRI is proposed where the*hyperparameters* are automatically estimated from the observations. Estimation and inference are joined together and the drift and HRF estimation and iteratively estimated by taking into account the spatial correlation.

Monte Carlo tests with synthetic data are presented to characterize the performance of the algorithm in terms of error probability. The introduction of the final step with *graph-cuts* greatly improves the accuracy of the algorithm, yielding an error probability that is close to zero even at the high noise levels observed in real data.

Real data activation results are consistent with a standard GLM approach, and most importantly, the activation clusters are best matched with the ones obtained at a significance threshold validated by the specialist, but with the advantage that the specification of user-defined subjective thresholds are not required. With the proposed method it also becomes unnecessary to apply spatial smoothing an high-pass temporal filtering as pre-processing steps, while accounting for important physiological properties of the data by estimating the HRF.

References

1. Afonso, D., Sanches, J., Lauterbach, M.: Joint bayesian detection of brain activated regions and local hrf estimation in functional mri. In: Proceedings IEEE ICASSP 2008. 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, USA (March 30 - April 4 2008)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23(11), 1222–1239 (2001)
3. Ciuciu, P., Poline, J.B., Marrelec, G., Idier, J., Pallier, C., Benali, H.: Unsupervised robust non-parametric estimation of the hemodynamic response function for any fmri experiment. IEEE Trans. Med. Imaging 22(10), 1235–1251 (2003)
4. Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R.: Event-related fMRI: characterizing differential responses. Neuroimage 7(1), 30–40 (Jan 1998)
5. Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J.: Classical and Bayesian inference in neuroimaging: Theory. NeuroImage 16, 465–483 (2002)
6. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Analysis and Machine Intelligence 6, 721–741 (November 1984)
7. Goutte, C., Nielsen, F.Å., Hansen, L.K.: Modelling the haemodynamic response in fmri with smooth fir filters. IEEE Trans. Med. Imaging 19(12), 1188–1201 (2000)
8. Jezzard, P., Matthews, P.M., Smith, S.M.: Functional magnetic resonance imaging: An introduction to methods. Oxford Medical Publications (2006)
9. K. J. Friston: Analyzing brain images: Principles and overview. In: R.S.J. Frackowiak and K.J. Friston and C. Frith and R. Dolan and J.C. Mazziotta (ed.) Human Brain Function, pp. 25–41. Academic Press USA (1997)
10. K. J. Friston and A. P. Holmes and K. J. Worsley and J. B. Poline and C. Frith and R. S. J. Frackowiak: Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Human Brain Mapping 2, 189–210 (1995)
11. Makni, S., Ciuciu, P., Idier, J., Poline, J.B.: Joint detection-estimation of brain activity in functional mri: A multichannel deconvolution solution. Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on] 53(9), 3488–3502 (2005)
12. Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., Poline, J.B.: Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. Human Brain Mapping 19, 1–17 (2003)
13. Moon, T.K., Stirling, W.C.: Mathematical methods and algorithms for signal processing. Prentice-Hall (2000)
14. Sanches, J., Marques, J.S.: A map estimation algorithm using IIR recursive filters. In: Proceedings International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Lisbon, Portugal (July 2003)

15. Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M.: Fully bayesian spatio-temporal modeling of fmri data. IEEE Trans. Med. Imaging 23(2), 213–231 (2004)