Exact MAP Activity Detection in *f*MRI Using a GLM with an Ising Spatial Prior

Eric R. Cosman, Jr.[1], John W. Fisher III[1], and William M. Wells III[1,2]

[1] Massachusetts Institute of Technology,
Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
ercosman@mit.edu, {fisher,sw}@csail.mit.edu
[2] Harvard Medical School, Brigham and Women's Hospital,
Department of Radiology, Boston, MA, USA

Abstract. Previous work [5] has shown how Ising spatial priors [1] can be incorporated into *f*MRI analysis in a principled manner by using Mutual Information as a statistic for protocol-related activity. The activation image with maximum *a posteriori* (MAP) probability can then be computed exactly in polynomial time by reduction to a Min-Cut/Max-Flow Problem [4]. In this work, we show that an Ising prior can be applied in the same manner using a standard, linear activation model.

1 Introduction

The functional imaging literature contains a number of methods aimed at limiting false detection of protocol-related brain activity in *f*MRI by taking advantage of the well-known fact that adjacent regions of the brain are likely to act in unison. These methods involve one or more of the following approaches: noise reduction by spatial smoothing of the *f*MRI time-series to "average out" spatially-white noise [2,6,8], regularization of voxel-specific activation statistics [2,5], and/or adjustment of voxel-independent activation statistics to reflect the size of apparent, surrounding activity clusters [6]. Specifically, [5] introduces a Bayesian approach for regularizing voxel-specific, non-parametric activation statistics in which an Ising spatial prior on protocol-dependent activity is integrated with an information-theoretic activity detector. By reduction to a Min-Cut/Max-Flow Problem [4], the maximum *a posteriori* (MAP) estimate of activity over the whole brain can be computed *exactly* in polynomial time by the Ford-Fulkerson method. This integration hinges on the interpretation of Mutual Information as an approximation of the log-likelihood ratio of a hypothesis test that assesses the statistical independence of a BOLD signal and an experimental protocol.

In this paper, we show that standard activation statistics, such as F-statistics, are derived from the log-likelihood ratio of a subset hypothesis test under classical, linear models of the BOLD signal. Consequently, the same exact MAP activity detection mechanism can be used with such General Linear Models (GLMs), thereby controling false positive rates in a principled, Bayesian manner.

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3217, pp. 703–710, 2004.

2 The General Linear Model

An *f*MRI experiment produces a set of time series $\{\boldsymbol{y}_i \in \Re^T : i = 1, \ldots, V\}$, each of which measures the BOLD signal over T epochs in one of the V voxels comprising the imaged brain volume. Under the General Linear Model (GLM), it is assumed that the BOLD signal is a linear combination of protocol-dependent components (the columns of matrix $\boldsymbol{H}$), confounding signals due to cardiopulmonary operations (the columns of matrix $\boldsymbol{D}$), and Gaussian noise [8]. For the special case of white noise, the GLM is written

$$\boldsymbol{y_i} = \boldsymbol{H}\boldsymbol{\eta}_i + \boldsymbol{D}\boldsymbol{\xi}_i + \boldsymbol{e}_i \qquad \boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \text{ i.i.d.} \qquad i = 1, \ldots, V \tag{1}$$

where $\boldsymbol{\eta}_i, \boldsymbol{\xi}_i$ are weight vectors on the columns of the *design matrix* $\boldsymbol{G} \equiv [\boldsymbol{H}\ \boldsymbol{D}]$. Under this model, classical activation statistics, such as the F statistic, can be derived from the log-likelihood ratio for a *two-sided, subset hypothesis test* $\{H_0 : \boldsymbol{\eta}_i = \boldsymbol{0},\ H_1 : \boldsymbol{\eta}_i \neq \boldsymbol{0}\}$, whereby we reject the null hypothesis (that there is no protocol-related neural activity) with an arbitrary threshold γ and the decision rule:

$$\lambda_i = \log \frac{\max_{\boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \sigma^2} \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{H}\boldsymbol{\eta}_i + \boldsymbol{D}\boldsymbol{\xi}_i, \sigma^2\boldsymbol{I})}{\max_{\boldsymbol{\xi}_i, \sigma^2} \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{D}\boldsymbol{\xi}_i, \sigma^2\boldsymbol{I})} \overset{\text{“}H_1\text{”}}{>} \gamma \tag{2}$$

$$\lambda_i - \gamma \overset{\text{“}H_1\text{”}}{>} 0 \tag{3}$$

We optimize the numerator first, stacking $\boldsymbol{\eta}_i, \boldsymbol{\xi}_i$ into a single weight vector $\boldsymbol{\zeta}_i$:

$$\begin{array}{rl} 0 = & \frac{d}{d\boldsymbol{\zeta}_i} \log \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{G}\boldsymbol{\zeta}_i, \sigma^2\boldsymbol{I}) \\ 0 = & \frac{d}{d\boldsymbol{\zeta}_i} ||\boldsymbol{y}_i - \boldsymbol{G}\boldsymbol{\zeta}_i||^2 \\ 0 = & \frac{d}{d\boldsymbol{\zeta}_i}(-2\boldsymbol{y}_i'\boldsymbol{G}\boldsymbol{\zeta}_i + \boldsymbol{\zeta}_i'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{\zeta}_i) \\ 0 = & -2\boldsymbol{G}'\boldsymbol{y}_i + (\boldsymbol{G}'\boldsymbol{G} + \boldsymbol{G}\boldsymbol{G}')\boldsymbol{\zeta}_i \\ \hat{\zeta}_i = & (\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{y} \end{array} \qquad \begin{array}{rl} 0 = & \frac{d}{d\sigma^2} \log \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{G}\hat{\boldsymbol{\zeta}}_i, \sigma^2\boldsymbol{I}) \\ 0 = & \frac{d}{d\sigma^2}\left(-\frac{1}{2} \log |\sigma^2\boldsymbol{I}| - \frac{||\boldsymbol{y}_i - \boldsymbol{G}\hat{\boldsymbol{\zeta}}_i||^2}{2\sigma^2} \right) \\ 0 = & \frac{d}{d\sigma^2}\left(\frac{n}{2} \log \sigma^2 + \frac{||\boldsymbol{y}_i - \boldsymbol{G}\hat{\boldsymbol{\zeta}}_i||^2}{2\sigma^2} \right) \\ 0 = & \frac{n}{2\sigma^2} - \frac{||\boldsymbol{y}_i - \boldsymbol{G}\hat{\boldsymbol{\zeta}}_i||^2}{2\sigma^4} \\ \hat{\sigma}^2 = & \frac{||\boldsymbol{y}_i - \boldsymbol{G}\hat{\boldsymbol{\zeta}}_i||^2}{n} \end{array} \tag{4}$$

By analogous optimization of the denominator, we get the following expression for the log-likelihood ratio, in which $\boldsymbol{P_X} \equiv \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (idempotent and symmetric) denotes a projection onto the column space of a matrix $\boldsymbol{X}$:

$$\begin{aligned} \lambda_i &= \log \frac{\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{G}\hat{\boldsymbol{\zeta}}_i, \hat{\sigma}_1^2\boldsymbol{I})}{\mathcal{N}(\boldsymbol{y}_i; \boldsymbol{D}\hat{\boldsymbol{\xi}}_i, \hat{\sigma}_0^2\boldsymbol{I})} = \frac{(T/2\pi e)^{T/2}}{||\boldsymbol{y}_i - \boldsymbol{G}\hat{\boldsymbol{\zeta}}_i||^T} \Bigg/ \frac{(T/2\pi e)^{T/2}}{||\boldsymbol{y}_i - \boldsymbol{D}\hat{\boldsymbol{\xi}}_i||^T} \\ &= \frac{T}{2} \log \frac{\boldsymbol{y}_i'(\boldsymbol{I} - \boldsymbol{P_D})\boldsymbol{y}_i}{\boldsymbol{y}_i'(\boldsymbol{I} - \boldsymbol{P_G})\boldsymbol{y}_i} \end{aligned} \tag{5}$$

Since the F-statistic F_i typically used for this test is a monotonic function of λ_i, the likelihood ratio test and F-test are equivalent:

$$F_i = \frac{\boldsymbol{y}_i'(\boldsymbol{P_G} - \boldsymbol{P_D})\boldsymbol{y}_i/(g-d)}{\boldsymbol{y}_i'(\boldsymbol{I} - \boldsymbol{P_G})\boldsymbol{y}_i/(T-g)} \sim F_{g-d,T-g} \text{ under } H_0 \text{ [7]} \tag{6}$$

$$F_i = \left(\frac{T-g}{g-d}\right) \frac{-\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P_G})\boldsymbol{y} + \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P_D})\boldsymbol{y}}{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P_G})\boldsymbol{y}}$$

$$F_i = \frac{T-g}{g-d}\left(\exp\left\{\frac{2\lambda_i}{T}\right\} - 1\right) \tag{7}$$

$$\lambda_i = \frac{T}{2}\log\left(\frac{g-d}{T-g}F_i + 1\right) \tag{8}$$

where $g = \text{rank}(\boldsymbol{G})$ and $d = \text{rank}(\boldsymbol{D})$. We can use Equation 8 to compute the threshold γ_α on the log-likelihood ratio λ_i corresponding to a test of size α (or vice versa):

$$\gamma_\alpha = \frac{T}{2}\log\left(\frac{g-d}{T-g}F_{\alpha;g-d,T-g} + 1\right) \tag{9}$$

Furthermore, the threshold γ for the classical likelihood ratio test can be interpreted in a Bayesian framework as the prior log-odds of detection. Using a simple prior $p(H_0) = 1 - p(H_1)$ on the competing hypotheses, and assuming that parameters θ_k are fixed but unknown with flat priors $p(\theta_k \mid H_k) \propto c$, we get the following MAP decision rule:

$$\max_{\theta_1} p(\theta_1, H_1 \mid \boldsymbol{y}_i) \overset{\text{“}H_1\text{”}}{>} \max_{\theta_0} p(\theta_0, H_0 \mid \boldsymbol{y}_i) \tag{10}$$

$$\lambda_i = \log\frac{\max_{\theta_1} p(\boldsymbol{y}_i \mid \theta_1, H_1)}{\max_{\theta_0} p(\boldsymbol{y}_i \mid \theta_0, H_0)} \overset{\text{“}H_1\text{”}}{>} \log\frac{p(H_0)}{p(H_1)} \equiv \gamma \tag{11}$$

3 An Ising Model for Neural Activity

We are motivated to use an Ising Markov Random Field [1] as prior on assessments of neural activity, by the fact that neural activity and its sequel, the activity-dependent BOLD signal. We refer to $h \equiv h_1, \ldots, h_V$ as an *activation map*, where $h_i \in \{0, 1\}$ is the assessment of (in)activity at voxel i, such that $h_i = 0$ and $h_i = 1$ correspond to hypotheses H_0 and H_1, respectively, as defined using a GLM as in Section 2. An Ising prior on the activation map h quantifies the notion that adjacent voxels are likely to act in unison by assigning greater probability to configurations with a greater number of homogeneous second-order cliques (since adjacent voxels are defined to be neighboring). In this work, we augment the prior with singleton clique potentials that penalize the total number of voxels declared active:

$$p(h|\gamma,\beta) = \frac{1}{Z(\gamma,\beta)} \exp\left\{-\gamma\sum_{i=1}^{V} h_i + \beta\sum_{i=1}^{V}\sum_{j\sim i}\delta(h_i - h_j)\right\} \tag{12}$$

$$= \frac{1}{Z(\gamma,\beta)} \exp\{-\gamma\cdot\#\{h_i = 1\} + \beta\cdot \mathrm{NHC}(h)\} \tag{13}$$

where NHC(h) gives the number of homogeneous cliques in configuration h, $Z(\gamma,\beta)$ is the partition function, and $j \sim i$ denotes that voxel j is a neighbor of voxel i. Conditioned on the activation map, the BOLD signals $\boldsymbol{y}_i$ are mutually independent across voxels. Therefore, the likelihood of the data $\boldsymbol{y}$ is

$$p(\boldsymbol{y}|\theta_0,\theta_1,h) = \prod_{i=1}^{V} p(\boldsymbol{y}_i|\theta_{0i},\theta_{1i},h_i) = \prod_{i=1}^{V} \frac{p(\boldsymbol{y}_i|\theta_{1i},h_i=1)^{h_i}}{p(\boldsymbol{y}_i|\theta_{0i},h_i=0)^{h_i-1}} \tag{14}$$

Choosing a flat prior $p(\theta_0,\theta_1\,|\,h) \propto c$ on the configuration of GLM parameters under each hypothesis, and taking γ and β as known hyperparameters, we get the following MAP estimation criteria:

$$\hat{h},\hat{\theta}_0,\hat{\theta}_1 = \arg\max_{h,\theta_0,\theta_1} \log p(h,\theta_0,\theta_1\,|\,\boldsymbol{y},\gamma,\beta) \tag{15}$$

$$= \arg\max_{h,\theta_0,\theta_1} \log\prod_{i=1}^{V} \frac{p(\boldsymbol{y}_i\,|\,\theta_{1i},h_i=1)^{h_i}}{p(\boldsymbol{y}_i\,|\,\theta_{0i},h_i=0)^{h_i-1}} + \log p(h\,|\,\gamma,\beta) \tag{16}$$

Since h_i is binary-valued, it is clear from Equation 16 that the posterior is increased by maximizing θ_{0i} and θ_{1i} for each voxel independently. Therefore, $\hat{\theta}_0,\hat{\theta}_1$ are the maximum likelihood estimates derived as in Equation 4, and the MAP estimate for the activation map is given by

$$\hat{h} = \arg\max_{h} \sum_{i=1}^{V}\left(h_i\left(\log\frac{p(\boldsymbol{y}_i|\hat{\theta}_{1i},h_i=1)}{p(\boldsymbol{y}_i|\hat{\theta}_{0i},h_i=0)} - \gamma\right) + \beta\sum_{i\sim j}\delta(h_i - h_j)\right) \tag{17}$$

$$= \arg\max_{h} \sum_{i=1}^{V}\left(h_i\left(\lambda_i - \gamma\right) + \beta\sum_{i\sim j}\delta(h_i - h_j)\right) \tag{18}$$

4 Reduction to the Minimum-Cut Problem

Since the activation map h can assume 2^V values, direct search for the optimal configuration $\hat{h}$ is computationally intractible. However, Greig *et al.* [4] showed that the search can be reduced to the Minimum-Cut/Maximum-Flow Network Problem, which can be solved in polynomial time by the Ford-Fulkerson method (or Preflow Push algorithms). We review this reduction with minor modification. Construct a capacitated network with V+2 vertices, comprising i=1, . . . ,V voxels, a source s, and a sink t. Let the graph have the following edges and corresponding capacities:

$$\begin{array}{ccl} (s,i) & c_{si} = \lambda_i - \gamma & \text{if } \ \lambda_i - \gamma > 0 \\ (i,t) & c_{it} = \gamma - \lambda_i & \text{if } \ \lambda_i - \gamma \leq 0 \\ (i,j) \ \text{ and } \ (i,j) & c_{ij} = c_{ji} = \beta & \text{if } \quad i \sim j \end{array} \tag{19}$$

For any activation map h, let $A = \{s\} \cup \{i : h_i = 1\}$ and $I = \{t\} \cup \{i : h_i = 0\}$ define a two-set partition of the network verticies. The set of edges with a vertex in A and a vertex in I is called a *cut*, and its *capacity* $C(h)$ can be written as follows:

$$C(h) = \sum_{k \in A} \sum_{l \in I} c_{kl} \tag{20}$$

$$= \sum_{i=1}^{V} \left(h_i \max(0, \gamma - \lambda_i) + (1 - h_i) \max(0, \lambda_i - \gamma) + \beta \sum_{i \sim j} 1 - \delta(h_i - h_j) \right) \tag{21}$$

This expression differs from the log-posterior $\log p(h, \hat{\theta}_0, \hat{\theta}_1 | \boldsymbol{y}, \gamma, \beta)$ (Equation 18) by a term which does not depend on h. Therefore, the MAP esimation is equivalent to finding the minimum cut in the network. Voxels are active in MAP estimate if they are on the source size of the minimum cut. Otherwise, they are inactive.

5 Experiments

Figure 1 shows the effect of varying the strength β of the spatial prior, for a given threshold γ_α. Activation maps are shown overlaying two axial slices (at the level of the Sylvian fissure) from a word-association task, where the strength of a spatial prior $\beta = 0, 0.5, 1, 2, 3$ increases from left to right. A simple GLM was used in which $\boldsymbol{H}$ is an encoding of the protocol, and the confounder subspace is empty $\boldsymbol{D} = \boldsymbol{0}$. The equivalent test size for the threshold γ_α is $\alpha = 1 \times 10^{-7}$. For each β, voxels declared active in the MAP activation map are colored white. The fMRI data were not pre-processed or pre-smoothed, so that the effect of the spatial prior could be observed in isolation. Figures 2 and 3 show how the running time for MAP estimation varies with the hyperparameters. The estimation was performed with a MATLAB implementation of the Ford-Fulkerson method. Specificially, we implemented the Edmonds-Karp algorithm, using depth-limited, depth-first search to find the shortest, feasible augmenting paths). The fMRI was aquired during a motor and auditory protocol and contains $V = 23187$ voxels.

Figure 2 shows how running time increases with the threshold γ, which we varied such that the number of above-threshold voxels $N = \text{size}(A) = 100, 250, 1000$ across runs. We also varied $\beta = 1, 2, 3$ for each setting of γ, which respectively corresponded to classical tests of size $\alpha = 6 \times 10^{-10}, 3 \times 10^{-7}, 6 \times$

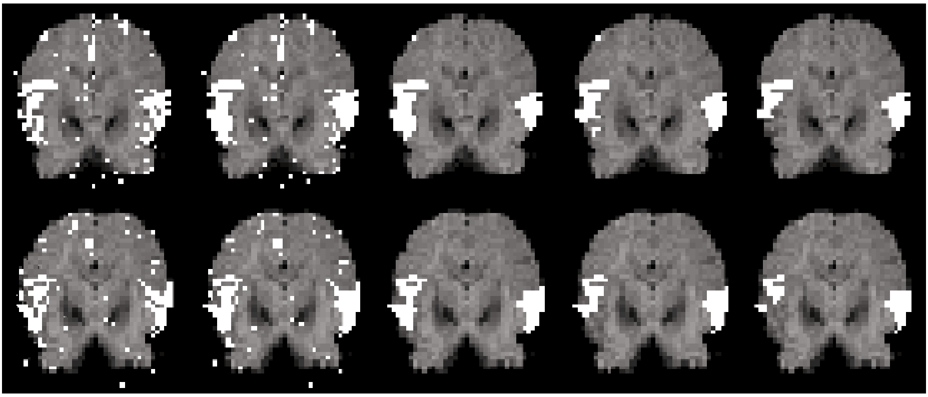

Fig. 1. Activation maps overlay two axial slices (at the level of the Sylvian fissure) from a word-association task, where the strength of a spatial prior $\beta = 0, 0.5, 1, 2, 3$ increases from left to right, and the equivalent test size for the threshold γ_α is $\alpha = 1 \times 10^{-7}$.

10^{-5}. For this and other datasets, the running time varied approximately linearly with N over this range. This is related to the fact that the Ford-Fulkerson method proceeds by sequentially augmenting *feasible paths* (i.e. those which can accomodate more flow) from the source s to the the sink t. Since the number of above-threshold voxels N (typically small relative to the total number of voxels) determines the number of edges emanating from the source s, the number of augmenting steps is roughly proportional to N.

Figure 3 shows that the same running time data varies roughly linearly as a function of $\beta = 1, 2, 3$. Again, this was typical over a number of *f*MRI datasets. Naturally, for increasing β, the network capacity increases monotonically, and with it, the number of long-range interactions and flow-augmenting steps.

6 Discussion

Inspection of the reduction in Section 4 clarifies the relationship between classical, voxel-independent *f*MRI analysis, and Bayesian analysis with an Ising prior. In both approaches, the log-likelihood ratio λ_i is computed at each voxel independently. Furthermore, in the Bayesian approach, voxels are initially partitioned into sets A and I (**A**ctive and **I**nactive) according to the decision rule $\lambda_i - \gamma > 0$, which is equivalent to that from the classical likelihood ratio test. Therefore, MAP estimation proceeds by first partitioning the data according to the classical, likelihood-ratio test, decision rule with threshold γ, and then adjusting the partition to account for the Ising prior. Moreover, the hyperparameter γ has a number of interpretations: (1) as a penalty for declaring a voxel active, (2) as corresponding to the size α of the classical, voxel-independent test, and (3) as the prior log-odds of detection in a simple, voxel-independent, Bayesian framework.

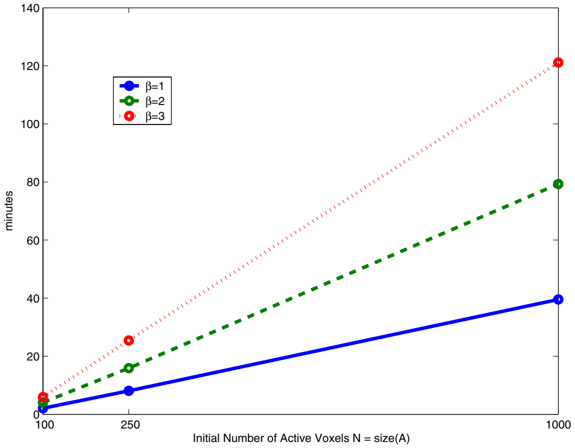

Fig. 2. Running Time as a function of the number of above-threshold voxels $N = \text{size}(A) = 100, 250, 1000$, for $\beta = 1, 2, 3$

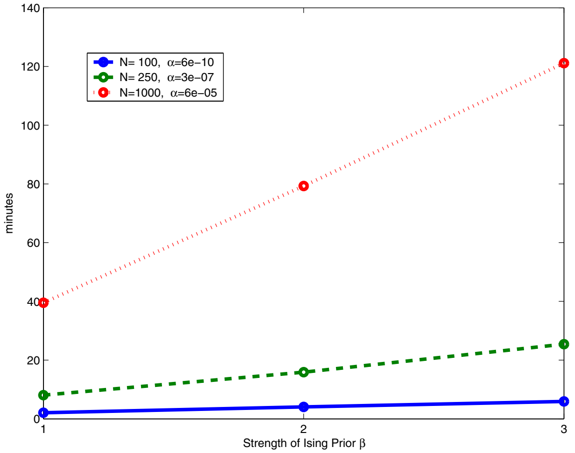

Fig. 3. Running Time as a function of strength of the Ising prior $\beta = 1, 2, 3$, for $N = \text{size}(A) = 100, 250, 1000$

The results of varying β (Figure 1) highlight the fact that the application of the Ising spatial prior is not simply a statistically-principled erosion operation. With increasing β, voxels which might be rejected at level α in a classical, voxel-independent test, may be declared active due to their proximity to other strongly active voxels. Of course, since the theshold typically exceeds the log-likelihood of most voxels, the primary effect of the Ising prior is to control the number of false detections by removing spatially-isolated activations.

Estimation of hyperparameters γ and β is complicated by the absence of ground truth activation maps. The MCMC-ML sampling approach of [3] could be adapted (in part) to find ML estimates of these hyperparameters using exact MAP estimates of the activation maps. However, the computational expense of such an approach is restrictive, as evaluation of optimality for each setting of (γ, β) involves running a min-cut computation and an MCMC simulation. Therefore, choice of optimal hyperparameters remains an open issue.

Finally, we note some possible extensions to this work. First, other classical activation statistics, such as the t-statistics, can also be derived from a likelihood ratio test and can thus be integrated into this framework. Furthermore, one can employ more specialized neighborhoods and clique potentials than we have shown. For instance, the coefficient β_{ij} might vary spatially according to prior beliefs about differences in regularity within and across anatomical boundaries derived from co-registered segmentations.

Acknowledgements. This research was supported in part by the grants NIH 5 P41 RR13218 and NCI 1 R21 CA89449-01. Further support was provided by FIRST BIRN, The Harvard Center for Neurodegeneration and Repair, and the NSF ERC grant, Johns Hopkins Agreement #8810274.

References

1. J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48:259–302, 1986.
2. X. Descombes, F. Kruggel, and D. Y. von Cramon. Spatio-temporal fmri analysis using markov random fields. *IEEE Transactions on Medical Imaging*, 17(6):1028–1039, December 1998.
3. X. Descombes, R. D. Morris, J. Zerubia, and M. Berthold. Estimation of markov random field prior parameters using markov chain monte carlo maximum likelihood. *IEEE Transactions of Image Processing*, 8(7), 1999.
4. D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(2):271–279, 1989.
5. J. Kim, J. W. F. III, A. Tsai, C. Wible, A. Willsky, and W. M. W. III. Incorporating spatial priors into an information theoretic approach for fmri data analysis. *Third International Conference on Medical Image Computing and Computer-Assisted Intervention*, 1935:62–71, October 2000.
6. J.-B. Poline, K. J. Worsley, A. C. Evans, and K. J. Friston. Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage*, 5:83–96, 1997.
7. A. C. Rencher. *Methods of Multivariate Analysis.* Wiley, 2002.
8. K. J. Worsley and K. J. Friston. Analysis of fmri time series revisited – again. *Neuroimage*, 2:173–181, 1995.