

# Exact MAP Activity Detection in fMRI using a GLM with an Ising Spatial Prior <sup>★</sup>

Eric R. Cosman, Jr.<sup>1</sup>, John W. Fisher III<sup>1</sup>, and William M. Wells III<sup>1,2</sup>

<sup>1</sup> Massachusetts Institute of Technology,  
Artificial Intelligence Laboratory, Cambridge, MA, USA  
{ercosman, fisher, sw}@ai.mit.edu

<sup>2</sup> Harvard Medical School, Brigham and Women's Hospital,  
Department of Radiology, Boston, MA, USA

**Abstract.** Previous work [4] has shown how Ising spatial priors [1] can be incorporated into fMRI 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 [3]. 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 fMRI 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 the fMRI time-series to “average out” spatially-white noise [2, 5, 7], regularization voxel-specific activation statistics [2, 4], and/or adjustment voxel-independent activation statistics to reflect the size of apparent, surrounding activity clusters [5].

Specifically, [4] 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 a Min-Cut/Max-Flow Problem [3], the maximum *a posteriori* (MAP) estimate of activity over the whole brain can be computed *exactly* in polynomial time by the Ford-Fulkerson method. The integration hinges on the interpretation of Mutual Information as an approximation of the log-likelihood ratio of a hypothesis test for statistical independence of the BOLD signal and experimental protocol.

In this paper, we show that standard activation statistics, e.g. 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

---

<sup>★</sup> E. Cosman was supported under the NSF award #IIS-9610249. W. Wells was supported by the same ERC grant, and by NIH 1P41RR13218.

detection mechanism can be used with such General Linear Models (GLMs), thereby controlling false positive rates in a principled, Bayesian manner.

## 2 The General Linear Model

An fMRI experiment produces a set of time series  $\{\mathbf{y}_i \in \mathbb{R}^T : i = 1, \dots, V\}$ , each of which measures the BOLD signal over  $T$  epochs in one of the  $V$  voxels comprising the imaged brain volume. Under a General Linear Model (GLM), it is assumed that the BOLD signal is a linear combination of protocol-dependent components (the columns of matrix  $\mathbf{H}$ ), confounding signals due to cardio-pulmonary operations (the columns of matrix  $\mathbf{D}$ ), and Gaussian noise [7]. For the special case of white noise, the GLM is written

$$\mathbf{y}_i = \mathbf{H}\boldsymbol{\eta}_i + \mathbf{D}\boldsymbol{\xi}_i + \mathbf{e}_i \quad \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \text{ i.i.d.} \quad i = 1, \dots, V \quad (1)$$

where  $\boldsymbol{\eta}_i, \boldsymbol{\xi}_i$  are weight vectors on the columns of the *design matrix*  $\mathbf{G} \equiv [\mathbf{H} \ \mathbf{D}]$ .

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*  $\{H_0 : \boldsymbol{\eta}_i = \mathbf{0}, H_1 : \boldsymbol{\eta}_i \neq \mathbf{0}\}$ , whereby we reject the null hypothesis (that there is no protocol-related neural activity) with an arbitrary threshold  $\gamma$  and the decision rule:

$$\lambda_i = \log \frac{\max_{\boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \sigma^2} \mathcal{N}(\mathbf{y}_i; \mathbf{H}\boldsymbol{\eta}_i + \mathbf{D}\boldsymbol{\xi}_i, \sigma^2 \mathbf{I})}{\max_{\boldsymbol{\xi}_i, \sigma^2} \mathcal{N}(\mathbf{y}_i; \mathbf{D}\boldsymbol{\xi}_i, \sigma^2 \mathbf{I})} \stackrel{\text{"}H_1\text{"}}{>} \gamma \quad (2)$$

$$\lambda_i - \gamma \stackrel{\text{"}H_1\text{"}}{>} 0 \quad (3)$$

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

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

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

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

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

$$F_i = \frac{\mathbf{y}'_i(\mathbf{P}_G - \mathbf{P}_D)\mathbf{y}_i/(g-d)}{\mathbf{y}'_i(\mathbf{I} - \mathbf{P}_G)\mathbf{y}_i/(T-g)} \sim F_{g-d, T-g} \text{ under } H_0 \text{ [6]} \quad (6)$$

$$F_i = \left(\frac{T-g}{g-d}\right) \frac{-\mathbf{y}'(\mathbf{I} - \mathbf{P}_G)\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{P}_D)\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{P}_G)\mathbf{y}}$$

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

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

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

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

Furthermore, the threshold  $\gamma$  for the classical likelihood ratio test can be interpreted in a Bayesian framework as the prior log-odds of detection, by putting a simple prior  $p(H_0) = 1 - p(H_1)$  on the competing hypotheses, a maximum-likelihood prior  $p(\theta_k|H_k) = \delta(\theta_k - \arg \max_{\theta_k} p(\mathbf{y}_i|\theta_k, H_k))$  on unknown parameters  $\theta_k$ , and using the MAP decision rule:

$$p(H_1|\mathbf{y}_i) \stackrel{\text{“}H_1\text{”}}{>} p(H_0|\mathbf{y}_i) \quad (10)$$

$$\lambda_i = \log \frac{\max_{\theta_1} p(\mathbf{y}_i|\theta_1, H_1)}{\max_{\theta_0} p(\mathbf{y}_i|\theta_0, H_0)} \stackrel{\text{“}H_1\text{”}}{>} \log \frac{p(H_0)}{p(H_1)} \equiv \gamma \quad (11)$$

### 3 An Ising Model for Spatial Correlation

We are motivated by the fact that neural activity and its sequel, the activity-dependent BOLD signal, are spatially correlated to use the Ising Markov Random Field [1] as prior on assessments of protocol-related neural activity from fMRI.

We refer to  $h \equiv h_1, \dots, h_V$  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:

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

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

where  $\text{NHC}(h)$  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  $\mathbf{y}_i$  are mutually independent across voxels. Therefore, the likelihood of the data  $\mathbf{y}$  is

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

Choosing an uninformative prior  $p(\theta_0|h)$ ,  $p(\theta_1|h)$  on the configuration of GLM parameters under each hypothesis, and taking  $\gamma$  and  $\beta$  as known hyperparameters, we get the following MAP estimation criteria:

$$\hat{h}, \hat{\theta}_0, \hat{\theta}_1 = \arg \max_{h, \theta_0, \theta_1} \log p(h, \theta_0, \theta_1|\mathbf{y}) \quad (15)$$

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

$$(17)$$

Since  $h_i$  is binary-valued, it is clear from Equation 16 that the posterior is increased by maximizing  $\theta_{0i}$  and  $\theta_{1i}$  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

$$\hat{h} = \arg \max_h \sum_{i=1}^V h_i \left( \log \frac{p(\mathbf{y}_i|\hat{\theta}_{1i}, h_i = 1)}{p(\mathbf{y}_i|\hat{\theta}_{0i}, h_i = 0)} - \gamma \right) + \beta \cdot \text{NHC}(h) \quad (18)$$

$$= \arg \max_h \sum_{i=1}^V \left( h_i (\lambda_i - \gamma) + \beta \sum_{i \sim j} \delta(h_i - h_j) \right) \quad (19)$$

## 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 intractable. However, Greig *et al.* [3] 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 for further discussion.

Construct a capacitated network with  $V+2$  vertices, comprising  $i=1, \dots, V$  voxels, a source  $s$ , and a sink  $t$ . Let the graph have the following edges and corresponding capacities:

$$\begin{array}{lll} (s, i) & c_{si} = \lambda_i - \gamma & \text{if } \lambda_i - \gamma > 0 \\ (i, t) & c_{it} = \gamma - \lambda_i & \text{if } \lambda_i - \gamma \leq 0 \\ (i, j) \text{ and } (j, i) & c_{ij} = c_{ji} = \beta & \text{if } i \sim j \end{array} \quad (20)$$

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 vertices. The set of edges with a vertex in  $A$  and a vertex in  $I$  is called a *cut* and the capacity  $C(h)$  can be written as follows since  $h$  is binary-valued:

$$\begin{aligned} C(h) &= \sum_{k \in A} \sum_{l \in I} c_{kl} & (21) \\ &= \sum_{i=1}^V \left( h_i \max(0, \gamma - \lambda_i) + (1 - h_i) \max(0, \lambda_i - \gamma) + \beta \sum_{i \sim j} 1 - \delta(h_i - h_j) \right) & (22) \end{aligned}$$

This expression differs from the posterior  $\log p(h, \hat{\theta}_0, \hat{\theta}_1 | \mathbf{y})$  (Equation 19) by a term which does not depend on  $h$ . Therefore, the MAP estimation is equivalent to finding the minimum cut in the network: MAP estimate voxels are active if they are on the source side of the minimum cut, and are inactive otherwise.

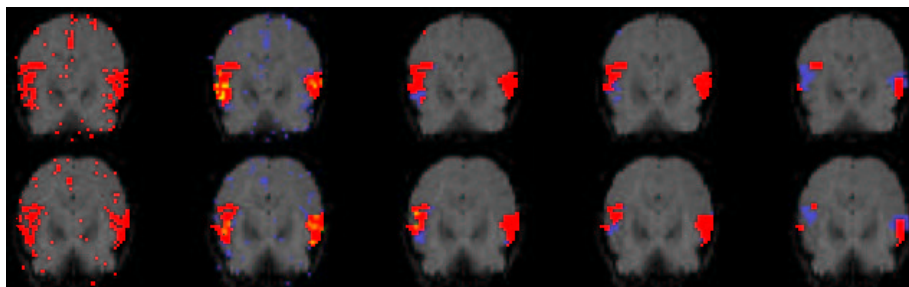
## 5 Discussion

Inspection of the reduction in Section 4 clarifies the relationship between classical, voxel-independent fMRI analysis, and Bayesian analysis with a Ising prior. In both approaches, the log-likelihood ratio  $\lambda_i$  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 a decision rule  $\lambda_i - \gamma > 0$  (Equation 20) equivalent to that from the classical likelihood ratio test (Equation 3). Therefore, MAP estimation proceeds by first partitioning the data according to the classical, likelihood-ratio test decision rule with threshold  $\gamma$ , and then adjusting the partition to account for the Ising prior by finding the minimum cut. Moreover, the hyperparameter  $\gamma$  has a number of interpretations: (1) a penalty for declaring a voxel active, (2) as corresponding to the size  $\alpha$  of the classical, voxel-independent test (Equation 9), and (3) as the prior log-odds of detection in a simple, voxel-independent, Bayesian framework (Equation 11).

Figure 1 shows the effect of varying the strength  $\beta$  of the spatial prior, for a given threshold  $\gamma_\alpha$ . 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  $\mathbf{H}$  is an encoding of the protocol, and the confounder subspace is

empty  $\mathbf{D} = \mathbf{0}$ . The equivalent test size for the threshold  $\gamma_\alpha$  is  $\alpha = 1 \times 10^{-7}$ . For each  $\beta$ , voxels declared active in the MAP activation map are colored red and yellow. Voxels colored yellow are newly detected relative the MAP activation map for the previous, smaller value of  $\beta$ . Similarly, voxels colored blue are newly inactive relative to the MAP activation map for the previous, smaller  $\beta$  value.

The presence of both yellow and blue voxels highlights the fact that the application of the Ising spatial prior is not simply a statistically-principled, erosion operation. With increasing  $\beta$ , voxels which might be rejected at level  $\alpha$  in a classical, voxel-independent test, may be declared active due to their proximity to other strongly active voxels. Of course, for typical thresholds, at which there is a relative paucity of initially active voxels, the primary effect of the Ising prior is to control the number of false detections by removing spatially-isolated activations.



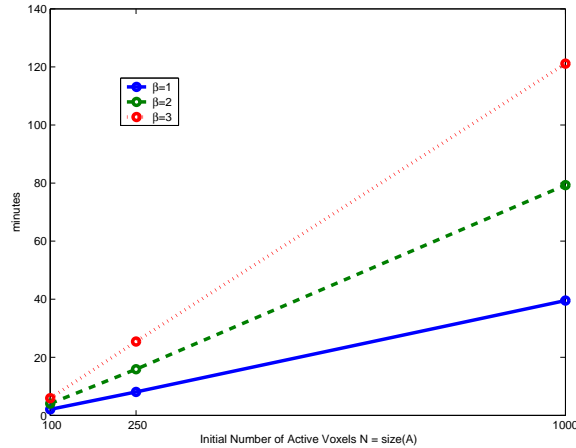
**Fig. 1.** Activation maps overlay two axial slices (at the level of the Sylvian fissure) from a word-association task, where the strength of a spatial prior  $\beta = 0, 0.5, 1, 2, 3$  increases from left to right, and the equivalent test size for the threshold  $\gamma_\alpha$  is  $\alpha = 1 \times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

## 5.1 Running Time

Figures 2 and 3 show running time for MAP estimation of the activation map on a particular *fMRI* dataset varies with the hyperparameters. The estimation was performed with a MATLAB implementation of the Ford-Fulkerson method (Edmonds-Karp algorithm, using depth-limited, depth-first search to find the shortest, feasible augmenting paths). The *fMRI* was acquired during a motor protocol and contains  $V = 23187$  voxels.

Figure 2 shows how running time increases with the threshold  $\gamma$ , 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  $\gamma$ , which respectively corresponded to classical tests of size  $\alpha = 6 \times 10^{-10}, 3 \times 10^{-7}, 6 \times 10^{-5}$ . For this and other *fMRI* datasets, we found the running time to vary approximately linearly with  $N$  over this range. This makes sense in light of the

fact that the Ford-Fulkerson method proceeds by sequentially augmenting *feasible paths* (i.e. those which can accommodate more flow) from the source  $s$  to the sink  $t$ . Since the number of above-threshold voxels  $N$ , which is small relative to the total number of voxels for typical thresholds, determines the number of edges emanating from the source  $s$ , the number of augmenting steps that can be performed before the Maximum Flow is achieved is roughly proportional to  $N$ .

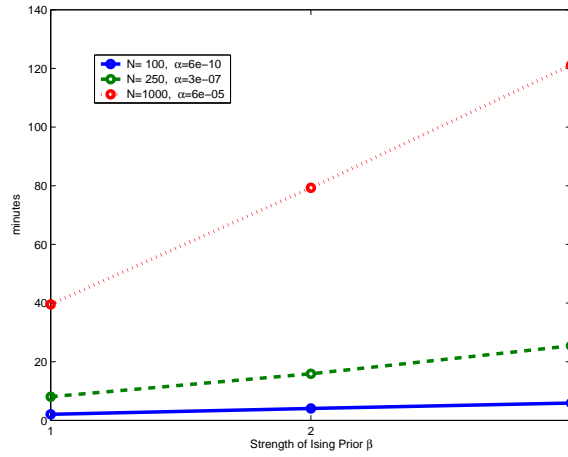


**Fig. 2.** Running Time as a function of the number of above-threshold voxels  $N = \text{size}(A) = 100, 250, 1000$ , for  $\beta = 1, 2, 3$

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 fMRI datasets. Naturally, the network capacity increases monotonically with increasing  $\beta$ , and with it the number of long-range interactions and flow-augmenting steps.

## 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. 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.
4. 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.



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

5. 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.
6. A. C. Rencher. *Methods of Multivariate Analysis*. Wiley, 2002.
7. K. J. Worsley and K. J. Friston. Analysis of fmri time series revisited – again. *Neuroimage*, 2:173–181, 1995.