

An Evaluation of Methods for Detecting Brain Activations from PET or fMRI Images¹

A.S. Lukic,² M.N. Wernick² and S.C. Strother³

²Illinois Institute of Technology, Chicago

³VA Medical Center and University of Minnesota, Minneapolis

Abstract – Brain activation studies based on PET or fMRI seek to explore neuroscience questions by using statistical techniques to analyze the acquired images. Currently, the predominant viewpoint toward quantifying the detection performance of these statistical methods is to model their output using random field theory, then to ascribe statistical significance (false-positive probability) based on the model. In this paper, we pursue instead an empirical strategy, based on receiver operating characteristics (ROC) analysis, as a first step toward a more-complete evaluation of the performance of brain-activation detection methods, including the power (true-positive probability) of various tests.

Using a phantom model derived from parameters measured from PET neuroimaging studies, we compare three methods for detecting brain activation. We consider one method based on pixel-by-pixel image comparisons (the t -test) and two methods based on pixel covariances (correlation thresholding and singular value decomposition (SVD) thresholding). The simple geometry of our phantom model allows us to construct an optimal detector, the generalized likelihood ratio test (GLRT), for comparison with the simpler detection procedures.

In this study the methods based on pixel covariances were found to perform better than the more widely used t -test. Among the covariance-based methods, none was found to be uniformly superior to the others. The performance of the GLRT served as an upper bound against which to compare the other methods. Our results suggest that correlation-based detectors are a promising direction for further investigation.

I. INTRODUCTION

The problem of quantifying brain function using detection theory is a difficult one because the true nature of the signal is unknown. Indeed, since the truth is unknown, it is a challenging problem even to quantify the performance of detection techniques for this problem.

Current brain-activation methodologies are based on thresholding of a test-statistic image. The predominant approach to characterizing the performance of these methods is to model their output by stationary random fields, using theorems based on topology to determine the false-positive probability [5].

A major weakness of the current approach is that it characterizes the results only in terms of false-positive

probability, saying nothing about the power (true-positive probability).

In this paper we use ROC curves to quantify detection performance. The ROC curves are derived from computer simulations based on a phantom, the parameters of which were measured from actual neuroimaging data. We compare the performance of the t -test [5], correlation thresholding [4], and singular value decomposition (SVD) thresholding [3]. The t -test, which is most commonly used in practice, is based on pixel-by-pixel comparison of images with a pooled variance

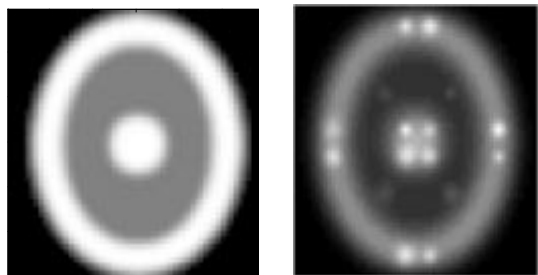


Fig 1. Baseline (left) and activated phantom (right) Brighter areas of the baseline represent gray matter; darker areas simulate white matter. The ratio of baseline activity is 4:1 between gray matter and white matter.

estimate. The other two methods are based on pixel covariances. To identify an upper limit for performance, and a reference point for comparison, we developed a generalized likelihood ratio test (GLRT) for our phantom model.

II. PHANTOM MODEL

To evaluate the detection methods, we developed the simple phantom model shown in Fig. 1. The values of all the parameters used to construct the phantom, given in Table I, are based on actual brain PET studies performed at the VA Medical Center, Minneapolis.

In the phantom, the ratio of baseline activity in “gray matter” to that in “white matter” is 4:1. We modeled the reconstructed images that contain no activations as the sum of a baseline (Fig. 1, left) and additive, colored, Gaussian noise. “Activated” brain images were obtained by introducing 16 two-dimensional Gaussian functions with a range of widths, and random, Gaussian-distributed amplitudes. A noise-free example is shown in Fig. 1 (right). For this phantom, the resulting hypothesis test can be written as:

$$\begin{aligned} H_0: \mathbf{x}_i &= \mathbf{b} + \mathbf{n}_i, \quad i = 1, \dots, N \\ H_1: \mathbf{x}_i &= \mathbf{b} + \mathbf{s}(\theta_i) + \mathbf{n}_i, \quad i = 1, \dots, N \end{aligned} \quad (1)$$

In (1), \mathbf{x}_i , \mathbf{b} , \mathbf{n}_i , and \mathbf{s} are $J \times 1$ image vectors representing the i th observed image, the baseline image (without activations), the i th realization of colored Gaussian noise, and

¹This research was supported by NIH grants NS34069 and MH57180.

the collection of activations, respectively. The variance of the noise in the j th pixel is

$$\sigma_j^2 = (0.05b_j)^2, \quad (2)$$

where b_j is the value of the j th pixel in \mathbf{b} .

The i th realization of the activations $\mathbf{s}(\boldsymbol{\theta}_i)$ is parameterized by the vector

$$\boldsymbol{\theta}_i = \begin{bmatrix} \mathbf{a}_i^T & \mathbf{w}^T \end{bmatrix}^T, \quad (3)$$

which contains a vector of random activation amplitudes

$$\mathbf{a}_i = [a_1^i \quad a_2^i \quad \dots \quad a_{16}^i]^T, \quad (4)$$

and a vector consisting of the widths of the activations, which are assumed not to change from image to image:

$$\mathbf{w} = [FWHM_1 \quad FWHM_2 \quad \dots \quad FWHM_{16}]^T. \quad (5)$$

The activation amplitudes were assumed to obey a multivariate Gaussian distribution, *i.e.*, $\mathbf{a} \sim N(\boldsymbol{\mu}_a, \mathbf{C}_a)$. The mean amplitude (strength) of the activations was varied in the experiments to determine its effect on detection performance. We express the mean amplitude of each activation with respect to the local value of the baseline as follows:

$$\boldsymbol{\mu}_a = M [b_{j_1} \quad \dots \quad b_{j_{16}}]^T \quad (6)$$

where j_k is the pixel index of the center of activation k , and M is a parameter controlling the overall mean. In our experiments we studied the cases in which the amplitude means were 3% and 5% above the baseline values ($M = 0.03$ and $M = 0.05$).

The correlation matrix \mathbf{C}_a of the random amplitude vector was defined in the following way:

$$\begin{aligned} [\mathbf{C}_a]_{k,k} &= V\sigma_{j_k}^2 \\ [\mathbf{C}_a]_{p,k} &= \rho \sqrt{[\mathbf{C}_a]_{p,p} [\mathbf{C}_a]_{k,k}}, \quad 0 \leq \rho \leq 1 \end{aligned} \quad (7)$$

TABLE I
PARAMETERS OF THE PHANTOM

image resolution	60x60 pixels
pixel size	3.1mm
Phantom size	18cm x 15cm
FWHM of system blur	5.8mm
FWHM of noise blur	6.2mm
Standard deviation of the colored noise, relative to the background value	0.05
FWHM of 2-D circular Gaussian activations	6.5mm, 8.5mm, 10.5mm, 12.5mm
mean activation amplitude, M	3%, 5% above baseline
variance of activation amplitudes (relative to noise variance), V	0.1 - 2
correlation strength, ρ	0, 0.5, 0.99
total number of images, $2N$	10 - 100

where $\sigma_{j_k}^2$ is the variance of the noise at pixel j_k . We chose this simple covariance model so that we could study the effect on detection performance of the correlation coefficient ρ , which expresses the degree of correlation between the activations.

The coefficient of proportionality V (the relative variance of the activation amplitudes) was varied in the simulations between 0.1 and 2. The correlation coefficient ρ was varied in the simulations to test the effect of correlation strength on detection performance. Simulations were performed for three values of ρ : 0.0, 0.5, 0.99.

III. METHODS TESTED

The t -test was implemented as described in [5]. Two groups of images were simulated (control and test). The procedure was repeated twice to simulate the case when the test group does not exhibit activations (H_0), and the case when it does (H_1). The values at the centers of activations on the resulting t -test image were recorded.

To perform SVD thresholding we again simulated two groups of images (control and test). A data matrix consisting of images from both groups was doubly centered (mean-corrected) and SVD was performed. The values at the centers of the activations on the eigenimage corresponding to the largest eigenvalue were recorded.

Correlation thresholding as described in [4] involves only one group of images. It requires calculation of the correlations according to:

$$R(\mathbf{x}(i), \mathbf{x}(j)) = \frac{\sum_{k=1}^{2N} \mathbf{x}_k(i) \mathbf{x}_k(j)}{\sqrt{\sum_{k=1}^{2N} \mathbf{x}_k^2(i) \sum_{p=1}^{2N} \mathbf{x}_p^2(j)}} \quad (8)$$

where $2N$ is number of images, and $\mathbf{x}_k(i)$ is the value of the i th pixel in the k th image. The maximum value for each pixel is recorded as:

$$\max_{j \in I_i} \{R(\mathbf{x}(i), \mathbf{x}(j))\} \quad (9)$$

where I_i is the set of indices of pixels separated from the i th pixel by more than the FWHM of the noise correlation function.

Owing to the specific properties of our simplified brain model, we can predict without numerical simulations which values in the sample correlation matrix will have the maximum value. Let us denote a pixel in an activation in the 'gray matter' as A_g , a pixel in the activation in the 'white matter' as A_w , and a noise-only pixel as A_n . The following inequalities will then hold:

$$\begin{aligned} \text{Var}(A_n) &\leq \text{Cov}(A_n, A_w) \leq \text{Var}(A_w) \\ &\leq \text{Cov}(A_w, A_g) \leq \text{Var}(A_g) \end{aligned} \quad (10)$$

Therefore, the maximum value in a row of the sample covariance matrix, corresponding to a pixel in the activation in the “gray matter” will be:

$$\max_{j \in I} \text{Cov}(A_g, A_j) = \text{Cov}(A'_g, A''_g) \quad (11)$$

where A_i is any other pixel and A'_g, A''_g are two pixels in two different activations in the “gray matter.” Hence, with our model the procedure can be interpreted simply as estimating the covariance between pixels belonging to activations in the “gray matter,” and making a decision based on comparison to a threshold. We utilized equation (11) to evaluate the

performance of the test numerically.

We applied this strategy to estimate the performance of the correlation thresholding method described in [4]. We also repeated the evaluation procedure for a slight variation of this method in which we defined the correlations as:

$$R(\mathbf{x}(i), \mathbf{x}(j)) = \sum_{k=1}^N \mathbf{x}_k(i) \mathbf{x}_k(j) \quad (12)$$

III. GENERALIZED LIKELIHOOD RATIO TEST

As a basis for comparison we designed a GLRT for this detection problem. The GLRT is performed by comparing the

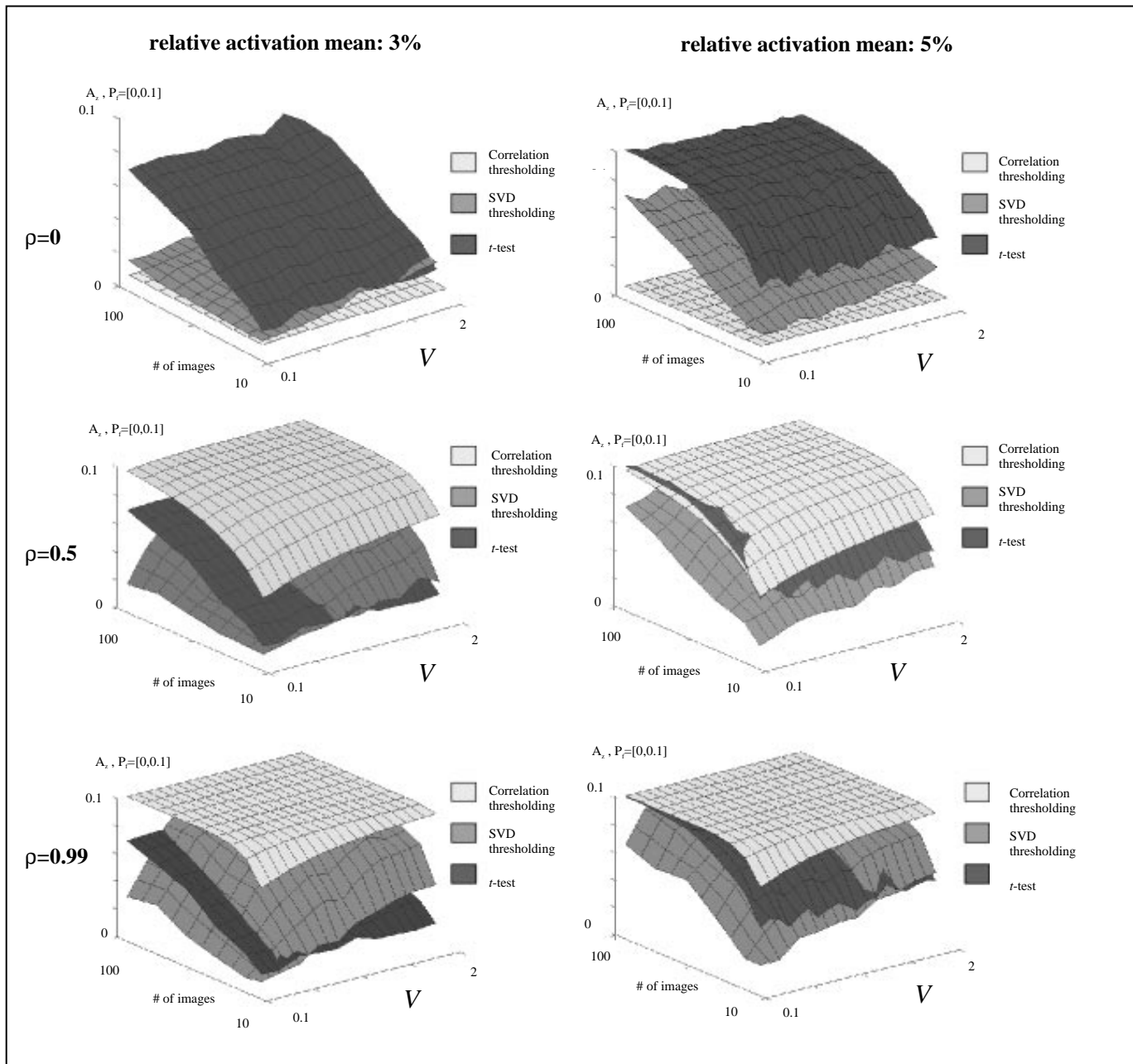


Fig. 2. Areas under the region of the ROC where false-positive probability is less than 0.1 for three methods tested as a function of V (the variance of the activation amplitude relative to that of the noise) and the number of images. In SVD thresholding and the t -test, the number of images is $2N$ (N in the control group; N in the test group); in correlation thresholding, only one group is needed, of size $2N$. In each graph the activation mean and correlation coefficient were kept constant.

likelihood ratio to a threshold, i.e.,

$$\frac{p(\mathbf{x}|H_1, \hat{\theta})}{p(\mathbf{x}|H_0)} \geq T \quad (13)$$

in which \mathbf{x} is a concatenation of all the observed images, θ is a concatenation of the vectors of parameters of activations in all the observed images, d_j denotes decision in favor of H_j , and $\hat{\theta}$ denotes the maximum *a posteriori* (MAP) estimator for θ , i.e.,

$$\hat{\theta} = \arg \max_{\theta} p(\mathbf{x}|\theta)p(\theta) \quad (14)$$

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N \{ [\mathbf{x}_i - \mathbf{b} - \mathbf{s}(\theta_i)]^T \mathbf{C}^{-1} [\mathbf{x}_i - \mathbf{b} - \mathbf{s}(\theta_i)] + (\mathbf{a}_i - \boldsymbol{\mu}_a)^T \mathbf{C}_a^{-1} (\mathbf{a}_i - \boldsymbol{\mu}_a) \} \quad (15)$$

where \mathbf{x}_i is the vector of the pixel values for the i th realization of the reconstructed image, and \mathbf{C} and \mathbf{C}_a are the covariance matrices of the noise and activation amplitudes, respectively. The MAP estimator $\hat{\theta}$ was computed by using the conjugate gradient algorithm. Pixels outside the phantom were not included in \mathbf{x}_i . The computation is quite tractable because \mathbf{C} is symmetric and the gradient of the cost function can be described analytically. In this way we estimated the sizes and

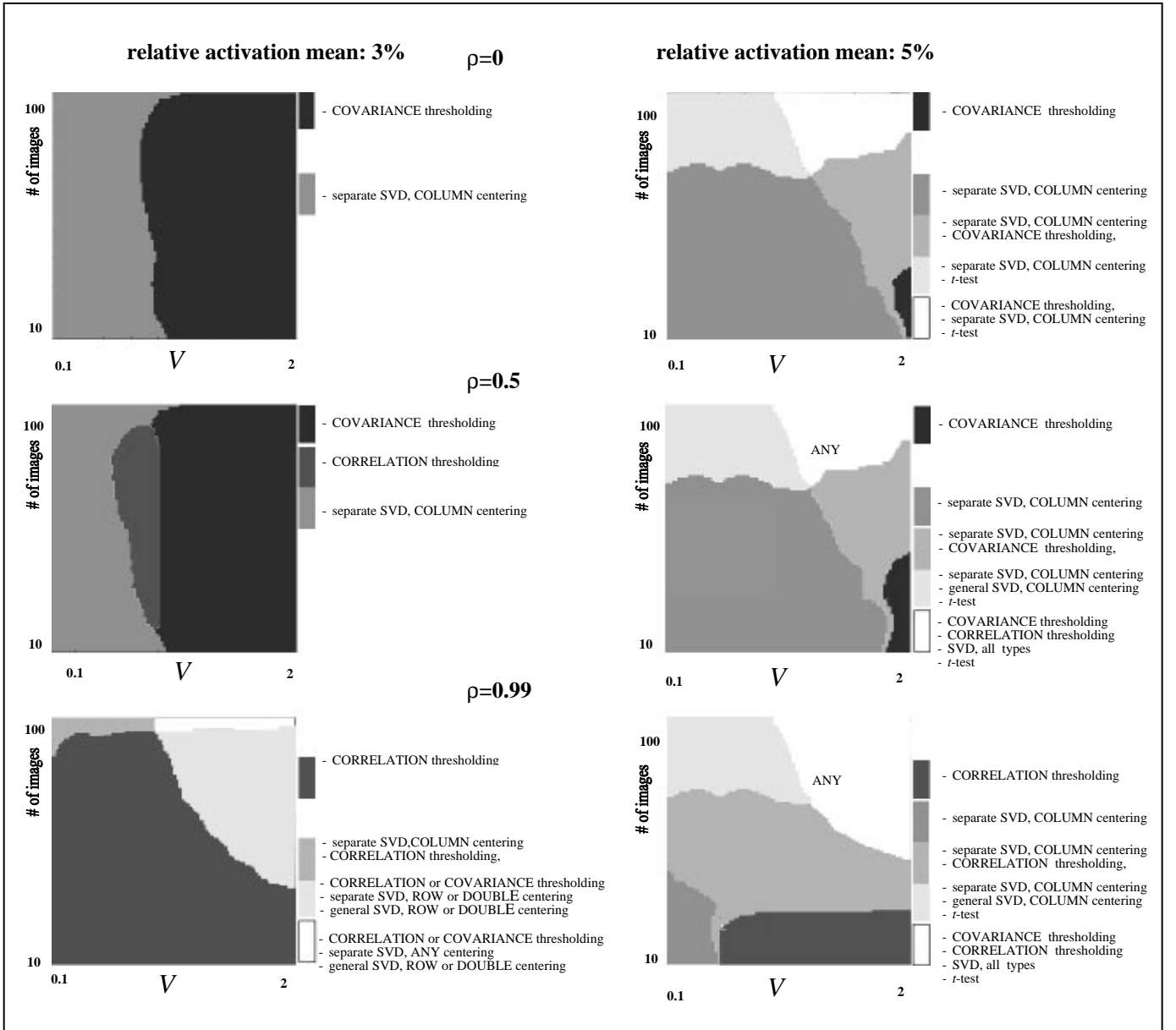


Fig. 3. Comparison of all methods tested. Regions indicate which methods work best for these combinations of parameter values. CORRELATION thresholding denotes the method described in [4], COVARIANCE thresholding denotes the modification of this method described in the text. Separate SVD indicates SVD performed on each of two groups of images separately and general SVD assumes SVD performed on the matrix formed from images from both groups. COLUMN centering results in the sum of pixels in each image being equal to zero, ROW centering results in the mean value of each pixel across all the images being equal to zero. DOUBLE centering denotes applying both COLUMN and ROW centering. For each graph relative activation mean and correlation coefficient were held constant.

amplitudes of all 16 Gaussian-shaped activations in every image. We then used these estimates to calculate the generalized likelihood ratio and perform the test in (13).

V. PERFORMANCE EVALUATION AND CONCLUSIONS

To compare the detection performance of the various methods, we performed 500 experiments for each case, recorded the values of the resulting images at the center of each activation and used the LABROC1 program [1] to estimate the ROC curve. The results shown in this paper are for a particular activation in the central region of the phantom. The results were similar for the other “gray matter” activations.

To obtain a single figure of merit, we calculated the area under the portion of ROC curve in the region where the false-positive probability is less than 0.1 [2]. This is more meaningful than the area under the entire ROC curve because it takes into account only the useful range of thresholds.

Our results are shown in Figs. 2 and 3. Figure 2 shows the performance only of the three detection methods exactly as they are proposed in [3-5]. In Fig. 3 we show a more extensive comparison that includes a number of modifications to these methods based on various ways to preprocess the data matrix, and different ways to define correlation. In both figures we evaluate the methods as a function of two important experimental parameters: the number of acquired images and the overall strength of the activations. The correlation coefficient ρ and activation means were held constant in each plot.

We draw the following conclusions from Figure 2 about the three main methods, as conventionally defined.

- *As expected, the correlation-based methods outperform the pixel-based t -test if there is sufficient correlation among the activations. As usually defined, these correlation-based methods fail when $\rho=0$. This is not the case for their modified versions (evaluated in Fig. 3) for which we find that there is a correlation-based method that surpasses the t -test in all cases.*

- *Correlation thresholding performs better in all cases than SVD thresholding when correlations are present. The difference in performance is more pronounced when the activations are weak.*

Figure 3 shows the parameter space divided into regions in which one or more methods perform better than all others. In addition to the three basic methods evaluated in Fig. 2, Fig. 3 compares a number of additional, modified versions: 1) variance thresholding, 2) row, column and double centering of the data matrix in SVD thresholding, and 3) SVD thresholding using separate data matrices, consisting only of “activated” or “non-activated” images. From Fig. 3 we draw the following conclusions.

- *The t -test never performs best, even when there is no correlation among the activations*

- *The performance of the t -test approaches that of correlation-based methods only when the activation means and the number of images are high.*

- *Covariance thresholding performs better than correlation thresholding for small values of ρ .*

- *In most cases some type of SVD thresholding gives the best performance. The principal exception is when the activation means and correlation are small, and the activation variance is high.*

As expected, the performance of the GLRT was almost always nearly perfect. A very slight decrease is noticed with the increase of the magnitude variances and small number of images. The performance of the correlation thresholding method approaches that of the GLRT when the activation correlation is very high. A more-realistic GLRT was also implemented, which sought to detect just one activation, assuming unknown size and amplitude, but known position. This GLRT always performed better than the t -test, and almost always better than SVD thresholding, but never better than correlation thresholding. SVD thresholding performs better than the one-activation GLRT when the activation correlation is extremely strong, the mean is low, and the variance is high.

The superiority of correlation thresholding and SVD thresholding (in some cases) over the one-activation GLRT is not surprising, because correlation-based methods exploit additional information conveyed by the correlations.

VI. SUMMARY

In this paper we used simulations based on a simple phantom model to test and compare the performance of three activation detection methods and many variations on them. We varied the parameters of our model to study their effect on detection performance. We also developed an optimal detection method for this model based on the GLRT.

Our study showed that, the widely used t -test is always outperformed by some correlation-based method, regardless of the number of images, and the mean, variance, and correlation of the activation amplitudes.

VII. REFERENCES

- [1] C. E. Metz, B. Herman, P.-L. Wang, J.-H. Shen, and B. Kronman, “LABROC1,” Chicago: Univ. of Chicago, 1993.
- [2] P. Skudlarski, T. R. Constable, J. C. Gore, “ROC Analysis of Statistical Methods Used in Functional MRI: Individual Subjects,” *Neuroimage*, vol. 9, pp. 311-329, 1999.
- [3] S. C. Strother, J. R. Anderson, K. A. Schaper, J. S. Siditis, and D. A. Rottenberg, “Linear Models of Orthogonal Subspaces & Networks from Functional Activation PET Studies of the Human Brain,” in *Information Processing in Medical Imaging*, vol. 3, Y. Bizais, C. Barillot, and R. Di Paola, Eds.: Kluwer, 1995, pp. 299-310.
- [4] K. J. Worsley, J. Cao, T. Paus, M. Petrides, A. C. Evans, “Applications of Random Field Theory to Functional Connectivity,” *Human Brain Mapping*, vol. 6, pp. 364-367, 1998.
- [5] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, “A Three Dimensional Statistical Analysis for CBF Activation Studies in Human Brain,” *J Cereb Blood Flow Metab*, vol. 12, pp. 900-918, 1992.