CONTROLLING THE ERROR IN FMRI: HYPOTHESIS TESTING OR SET ESTIMATION?

*Zachary Harmany and Rebecca Willett**

Duke University
Electrical and Computer Engineering
Durham, NC

Aarti Singh and Robert Nowak†

University of Wisconsin-Madison
Electrical and Computer Engineering
Madison, WI

ABSTRACT

This paper describes a new methodology and associated theoretical analysis for rapid and accurate extraction of activation regions from functional MRI data. Most fMRI data analysis methods in use today adopt a hypothesis testing approach, in which the BOLD signals in individual voxels or clusters of voxels are compared to a threshold. In order to obtain statistically meaningful results, the testing must be limited to very small numbers of voxels/clusters or the threshold must be set extremely high. Furthermore, voxelization introduces partial volume effects (PVE), which present a persistent error in the localization of activity that no testing procedure can overcome. We abandon the multiple hypothesis testing approach in this paper, and instead advocate a new approach based on set estimation. Rather then attempting to control the probability of error, our method aims to control the spatial volume of the error. To do this, we view the activation regions as level sets of the statistical parametric map (SPM) under consideration. The estimation of the level sets, in the presence of noise, is then treated as a statistical inference problem. We propose a level set estimator and show that the expected volume of the error is proportional to the sidelength of a voxel. Since PVEs are unavoidable and produce errors of the same order, this is the smallest error volume achievable. Experiments demonstrate the advantages of this new theory and methodology, and the statistical reasonability of controlling the volume of the error rather than the probability of error.

***Index Terms*—** Magnetic resonance imaging, Signal detection, Neuroimaging, fMRI

1. FMRI ANALYSIS

First we review the standard hypothesis testing approach used in most studies today, and we point out the limitations and flaws of such methods. Then we formulate the fMRI analysis problem as a level set estimation task, and discuss the virtues of this perspective. Before moving on, we establish basic notation that will be used throughout the paper. Let y_i, $i = 1, \ldots, n$, denote the elements of the statistical parameter map (SPM) under consideration, n being the number of voxels. Elements of the SPM may be t-statistics, cross-correlation coefficients, z-statistics, or one of a variety of other measures, as described in [1]; in this paper, for simplicity of presentation, we will assume $y_i \in [-1, 1]$. In general, each statistic is modeled as the sum of a deterministic and stochastic component:

$$y_i = \bar{f}_i + \epsilon_i \, .$$

The deterministic component $\bar{f}_i$ is the mean value of y_i, and is viewed as the average of an underlying continuous activation function over the i-th voxel. The stochastic component ϵ_i is assumed to have a mean value of zero. Assume that the volume of the brain is embedded into the unit cube $[0, 1]^3$, and that each voxel corresponds to $1/n$ of this volume. Let f denote the continuous activation function defined on $[0, 1]^3$. Then $\bar{f}_i = n \int_{V_i} f(x)\, dx$, where V_i is the subcube that is the i-th voxel (the factor of n accounts for normalization by the voxel volume).

1.1. Hypothesis Testing for fMRI

Broadly speaking, hypothesis testing procedures are based thresholding the SPM, or functions of it, at a certain level. Points exceeding the threshold are declared as active. In voxel-wise testing, it is usually assumed that the ϵ_i are zero-mean Gaussian errors. Inactive voxels are assumed to have $\bar{f}_i = 0$, and the Gaussian distribution of ϵ_i then provides a principled means for selecting an appropriate threshold. A Bonferroni correction (BC) or sequential p-value method can be used to determine a threshold that controls the family-wise error rate (probability that one or more voxels is falsely detected) or the false discovery rate (FDR). Because these both control error rates over large numbers of voxels, the thresholds they prescribe are extremely conservative; typically very few voxels exceed these stringent, albeit statistically sound, thresholds. One way to circumvent this problem is to perform a much smaller number of tests, say on a subset of voxels or larger clusters/groups of voxels [2]. Of course, this may significantly sacrifice the resolution, or regional specificity, of fMRI analysis. An excellent discussion of this tradeoff may be found in [3], which also describes a variety of testing approaches based on the theory of Gaussian fields. Also, we mention that other alternatives to voxel-based testing have been proposed. For example, [4] uses hypothesis testing based on the wavelet transform of the SPM to detect activity. This offers another approach to trading regional specificity in return for enhanced SNR.

Beyond these difficulties associated with hypothesis testing approaches to fMRI, there is also the error due to partial volume effects (PVE). PVEs are always present in fMRI and place a limit on the regional specificity (i.e., accuracy with which activation can be localized). The localization accuracy is proportional to the sidelength of a voxel. For example, suppose that a certain voxel contains a volume of the brain that is activated in one half of the voxel, but not the other. It is quite possible that the SPM value for this voxel could exceed a detection threshold, but subsequently inferring that this implies this entire voxel volume is active is obviously incorrect. One can only safely say that some part of the voxel volume is active. Similar problems arise in cluster- and wavelet-based testing.

*R. Willett was partially supported by NSF Award No. CCF-06-43947.
†R. Nowak was partially supported by NSF Award No. CCR-0310889

1.2. Level Set Estimation for fMRI

In light of the challenges associated with controlling the probability of error in fMRI, and the fact that even under such control errors in regional specificity persist at the voxel level due to PVE, we advocate an alternative to hypothesis testing. The best we can hope for is an localization error whose volume is proportional to voxel side-length, so we aim to localize activation to within this accuracy.

The ideal goal of fMRI is to determine the subset of $[0,1]^3$ where the activation function f exceeds a certain positive level, indicating regions where the BOLD response is especially strong.

Aim 1 *(Ideal): For a given level $\gamma > 0$, determine the level set*

$$S_\gamma \equiv \{x \in [0,1]^d : f(x) \geq \gamma\} \subset [0,1]^3.$$

This is similar in spirit to testing approaches based on the theory of Gaussian random fields [3, 5]. In those approaches, the SPM is viewed as a voxelated representation of an underlying continuous Gaussian field. The function f in our set-up would be the mean of such a field. Based on the Gaussian field assumption, one can probabilistically characterize the *excursion set*, which is closely related to our notion of a level set. The excursion set is a level set of the SPM, whereas S_γ is the gamma-level set of only the *deterministic* component of the SPM.

The ideal aim is unachievable for two reasons. First, the PVE artifact limits the accuracy of any approach to this problem to $n^{-1/3}$, the side-length of a voxel. Second, the stochastic component of the SPM introduces another source of error. Remarkably, a careful statistical analysis shows that the effect of the stochastic error can be controlled to order $n^{-1/3}$ as well. In anticipation of this, we pose a second, more useful, aim:

Aim 2 *(Practical): For a given level $\gamma > 0$, construct an estimator of the γ-level set, denoted $\widehat{S}$, that satisfies the following error bound:*

$$\mathbb{E}\left[\text{vol}\left(\Delta(S_\gamma, \widehat{S})\right)\right] \propto n^{-1/3}, \tag{1}$$

where $\mathbb{E}$, above and throughout the paper, denotes the expectation over the random stochastic errors, vol *stands for volume, S_γ^c is the complement of S_γ, and*

$$\Delta(S_\gamma, \widehat{S}) \equiv (S_\gamma \cap \widehat{S}^c) \cup (S_\gamma^c \cap \widehat{S})$$

is the so-called the symmetric difference between the sets S_γ and $\widehat{S}$.

The symmetric difference of two sets is simply the total of all points included in one set and not the other, and vice-versa. The volume of this set is the overall error in the estimation of the activation region(s). Because the level set estimator is a random variable, we consider the expected value of the volume error. Similar quantifications of the size of the error can be made in terms of probability, rather than expectation, but the expectation bound most clearly illustrates the performance.

1.3. The Level Set Approach vs. Gaussian Field Approaches

As pointed out above, there is a close connection in the formulation of level set estimation and testing based on Gaussian random field models. Our main criticism of the latter is that they are based on a strong assumption about the spatial dependencies in the BOLD signal, namely that the SPM is a realization of a smooth Gaussian process. While there is little doubt that dependencies exist, it is very unclear that a Gaussian process is a reasonable model for them. The assumed smoothness of the field, W in [3], strongly influences the shapes, sizes and boundary smoothness of the excursion sets. Even the best choice of W may not yield excursion sets that are good models for real activation patterns. The Gaussian field models are isotropic processes, and thus favor isotropic excursion sets. Actual activation patterns may be highly non-isotropic. In contrast, the level set approach assumes only that the boundary of S_γ is lower dimensional (e.g., a two dimensional surface when studying a three dimensional volume). No further assumptions are placed on the shape or smoothness of the level set.

2. LEVEL SET ACTIVATION DETECTION IN FMRI

2.1. Error metrics

Careful selection of an error metric is the first step in designing an effective level set estimator. In particular, the method presented in this paper is designed to minimize the *weighted* symmetric difference

$$\mathcal{E}(S, S_\gamma) \equiv \int_{\Delta(S_\gamma, S)} |\gamma - f(x)| dx. \tag{2}$$

This is a useful metric for broad classes of activation functions f. In particular, consider activation functions with $f(x) < \gamma_{\min}$ in inactive regions and $f(x) > \gamma_{\max}$ in active regions. Note that if $\gamma_{\min} < \gamma < \gamma_{\max}$, then there exist constants $c, C > 0$ such that $c \leq |\gamma - f(x)| \leq C$. It follows that

$$c\,\text{vol}(\Delta(S_\gamma, S)) \leq \int_{\Delta(S_\gamma, S)} |\gamma - f(x)| dx \leq C\,\text{vol}(\Delta(S_\gamma, S)),$$

and therefore minimizing $\mathcal{E}(S, S_\gamma)$ minimizes the volume of the erroneous region, as desired.

The quantity $\mathcal{E}(S, S_\gamma)$ cannot be directly evaluated without knowledge of S_γ (the level set we wish to estimate.) However, note that the weighted symmetric difference can be decomposed as $\mathcal{E}(S, S_\gamma) = \mathcal{R}(S) - \mathcal{R}(S_\gamma)$, where

$$\mathcal{R}(S) \equiv \int \frac{\gamma - f(x)}{2} \left[\mathbb{I}_{\{x \in S\}} - \mathbb{I}_{\{x \in S^c\}}\right] dx \tag{3}$$

and $\mathbb{I}_{\{A\}} = 1$ if event A is true and 0 otherwise. Thus minimizing $\mathcal{E}(S, S_\gamma)$ is equivalent to minimizing $\mathcal{R}(S)$, since $\mathcal{R}(S_\gamma)$ is a constant, and the value of $\mathcal{R}(S)$ is independent of S_γ. Furthermore, $\mathcal{R}(S)$ has an empirical counterpart that can be computed from the SPM:

$$\widehat{\mathcal{R}}_n(S) = \frac{1}{n} \sum_{i=1}^{n} \frac{\gamma - y_i}{2} \left[\mathbb{I}_{\{x_i \in S\}} - \mathbb{I}_{\{x_i \in S^c\}}\right]. \tag{4}$$

Note that $\mathcal{R}(S) = \mathbb{E}\left[\widehat{\mathcal{R}}_n(S)\right]$.

2.2. Estimation via Trees

Building upon the work in [6], we can show that level set formulation presented above leads to theoretically optimal estimators. In particular, we propose to estimate the level set of the activation function f based on the SPM by using a tree-pruning method akin to CART [7] or dyadic decision trees [8]. Trees are utilized for a couple of reasons. First, they provide a simple means of generating a spatially adaptive partition, yielding an automatic data aggregation in regions estimated to be strictly above or below the γ level. This adaptive

and automatic aggregation effectively boosts the signal-to-noise ratio. Second, the optimal partition can be computed very rapidly using a simple bottom-up pruning scheme.

Let $\mathcal{T}_n$ denote the collection of all 8-ary trees in three dimensions (8-ary trees are based on recursively partitioning cubic volumes into 8 sub-cubes). For example, consider a $64 \times 64 \times 64$ voxel volume. The voxels represent the limit of the 8-ary partition process, generated by a 6 level 8-ary tree ($2^6 = 64$). In this example, $n = 64^3$ and the side-length of each voxel is $n^{-1/3} = 1/64$ (assuming the brain volume is normalized to be the unit cube). Note, however, that is is not necessary to complete the partitioning process to the voxel level; using less than 6 levels will result in partition cells composed of groups of voxels. Moreover, the level of the tree can vary spatially, yielding smaller cells in certain areas and larger cells in others. This is what we call a spatially adaptive partition. Spatial adaptivity is crucial in level set estimation, since ideally we wish to aggregate the SPM wherever $f(x)$ strictly exceeds or falls below the target level of γ.

Let $\pi(T)$ denote the partition induced on $[0,1]^3$ by the 8-ary tree T. Each leaf node of the tree corresponds to a cell of the partition. A zero or one is assigned to each leaf node of T (equivalently, to each cell $L \in \pi(T)$), and the union of leafs with label one form a set denoted S_T. Let $|L|$ denote the volume of the leaf L. Based the theory of tree-based level set estimators developed in [6], we have the following result.

Theorem 1 *Let*

$$\Phi_n(T) \equiv \sum_{L \in \pi(T)} \sqrt{\frac{2|L|}{n} \log\left(\frac{4n}{|L|^{4/3}}\right)}. \tag{5}$$

With probability at least $1 - 1/n$,

$$\mathcal{R}(S_T) \leq \widehat{\mathcal{R}}_n(S_T) + \Phi_n(T) \quad \forall\, T \in \mathcal{T}_n.$$

The proof of this theorem utilizes a union bound over the leaves coupled with Hoeffding's inequality (a concentration of measure inequality for bounded random variables), which bounds the contribution of the risk from each individual leaf. Theorem 1 shows that the ideal quantity, $\mathcal{R}(S_T)$, that we wish to minimize is upper bounded by the sum of empirical version $\widehat{\mathcal{R}}_n(S_T)$ and $\Phi_n(T)$, both of which are easily computed; i.e., although $\mathcal{R}(S_T)$ requires exact knowledge of $f(x)$, the upper bound does not. Thus, the set that minimizes $\widehat{\mathcal{R}}_n(S_T) + \Phi_n(T)$ also makes $\mathcal{R}(S_T)$ (and hence the volume error) small with very high probability. Intuitively, we can interpret $\Phi_n(T)$ as a regularization term which discourages excessively complex partitions.

Theorem 1 leads directly to the following level set estimator:

$$\widehat{S}_T = \underset{S_T : T \in \mathcal{T}_n}{\arg\min} \left\{ \widehat{\mathcal{R}}_n(S_T) + \Phi_n(T) \right\}. \tag{6}$$

In addition, the estimator defined by (6) and (5) is rapidly computable. In fact, a translation-invariant estimator can be computed in $O(n \log n)$ operations [6]. Furthermore, as shown in the following section, the estimator is nearly optimal.

2.3. Performance Analysis

Not only does the above framework give us a principled way to select a level set estimator, but it also allows us to bound the expected error volume. In particular, we have the following theorem, which also follows from the work in [6].

Theorem 2 *Let* $\widehat{S}_T$ *be as in (6) with* $\Phi_n(T)$ *as in (5). Then*

$$\mathbb{E}\left[\mathcal{E}(\widehat{S}_T, S_\gamma)\right] \leq \min_{S_T : T \in \mathcal{T}_n} \left\{\mathcal{E}(S_T, S_\gamma) + 2\Phi_n(T)\right\} + \frac{4}{n}.$$

The bound on the expected error in Theorem 2 allows us to bound the expected volume of the erroneous region (total volume of missed activation and falsely detected activation). In particular, the following theorem shows that for a broad class of functions f, the proposed method is nearly optimal. Again, assume that the activation function satisfies $f(x) < \gamma_{\min}$ in inactive regions, $f(x) > \gamma_{\max}$ in active regions, and $\gamma_{\min} < \gamma < \gamma_{\max}$. Furthermore, we assume that the boundaries of S_γ are two-dimensional surfaces (i.e., the boundary of a volume is a surface). Under these mild assumptions, we have the following theorem.

Theorem 3

$$\mathbb{E}\left[\text{vol}\left(\Delta(\widehat{S}_T, S_\gamma)\right)\right] \propto \mathbb{E}\left[\mathcal{E}(\widehat{S}_T, S_\gamma)\right] \propto \left(\frac{\log n}{n}\right)^{1/3}.$$

This shows that the expected volume of the error is nearly equal to the minimum dictated by partial volume effects, $n^{-1/3}$. The extra logarithmic factor is relatively negligible. Furthermore, the estimator automatically adapts to the form and extent of the level set, S_γ, without prior constraints on shape or size of the activation regions. The level set estimator is also computationally efficient.

3. IMPROVED ESTIMATION VIA CROSS-VALIDATION

The above theoretical analysis demonstrates that the estimator in (6) controls the expected volume of the error. However, in practice, because the regularization term $\Phi_n(T)$ is based on worst-case analysis techniques, the resulting level set estimate can be somewhat over-regularized (*i.e.* over-smoothed). Attenuating the regularization term can result in improved performance. It can be shown that choosing this attenuation factor using a cross-validation procedure both is empirically effective and retains the optimality of the original estimator.

In particular, we assume that we have access to two data sets or runs collected under the same conditions, and we denote the two resulting SPMs as $\{y_i^T\}_{i=1}^n$ for training data and $\{y_i^V\}_{i=1}^n$ for validation data. Such data is often available in fMRI studies [2]. For an attenuation factor $\lambda \in [0,1]$, we have the estimator

$$\widehat{S}_\lambda = \underset{S_T : T \in \mathcal{T}_n}{\arg\min} \left\{ \widehat{\mathcal{R}}_n^T(S_T) + \lambda \Phi_n^T(T) \right\}, \tag{7}$$

where $\widehat{\mathcal{R}}_n^T$ is (4) evaluated using the training data. We then choose the optimal attenuation λ as $\widehat{\lambda} = \arg\min_{\lambda \in [0,1]} \widehat{\mathcal{R}}_n^V(\widehat{S}_\lambda)$, where $\widehat{\mathcal{R}}_n^V$ is (4) evaluated using the validation data, and set the final estimate to be $\widehat{S} = \widehat{S}_{\widehat{\lambda}}$. From here it is possible to derive the following theorem.

Theorem 4 *With probability at least* $1 - 1/n$,

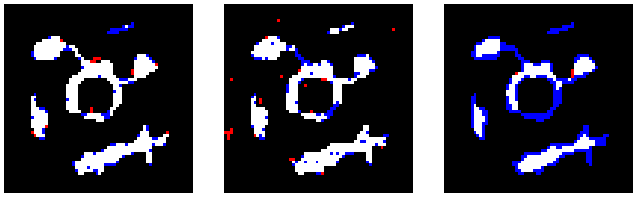
$$\mathbb{E}\left[\text{vol}\left(\Delta(\widehat{S}_T, S_\gamma)\right)\right] \leq \min_{\lambda \in [0,1]} \mathbb{E}\left[\mathcal{E}(\widehat{S}_\lambda, S_\gamma)\right] + \sqrt{\frac{\log n}{2n}} + \frac{1}{n} \propto \left(\frac{\log n}{n}\right)^{1/3}.$$

Theorem 4 indicates that the error volume of the cross-validated estimator is also near-optimal. The proof of the theorem follows through a straightforward application of the results in [9].

4. EXPERIMENTAL RESULTS

4.1. Simulated Phantom

We first use simulated data to perform our holdout method for parameter tuning. The phantom consists of regions of different activation level, representing levels of neurological activity. Additive unit-variance zero-mean Gaussian noise is used to corrupt this underlying activity to form our simulated data. The SPM is then computed using this known $\mathcal{N}(0,1)$ noise distribution. We compare the level set estimate $\widehat{S}_\gamma$ with the true S_γ. In Fig. 1, we compare our results to those obtained by traditional pre-smoothing and thresholding. We use a Gaussian smoothing kernel where the smoothing bandwidth of 1.2 voxels FWHM was chosen *clairvoyantly* via directly minimizing $\text{vol}(\Delta(S_\gamma, \widehat{S}))$. For this reason, the performance of this conventional approach will exceed what is possible in practice when the true S_γ is unknown. We also show a more typical example where a bandwidth of 3 voxels FWHM is chosen, demonstrating a severe degradation in performance when one over-smoothes the SPM. Note that in comparison to simple smoothing and thresholding, our errors are well-localized along the boundary of the level set, as the theory predicts. Averaging over 100 realizations of the SPM, we have an average symmetric difference volume of 0.023 using level sets, 0.030 using thresholding with clairvoyant pre-smoothing, and 0.077 using thresholding with a fixed typical amount of pre-smoothing. This means that our level set estimation procedure outperforms pre-smoothing and thresholding estimate *even with the best clairvoyant choice of the smoothing bandwidth*.

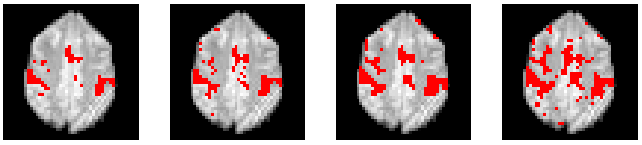


(a) (b) (c)

Fig. 1. Comparison of $\gamma = 0.70$ level set estimation, thresholding with clairvoyant pre-smoothing, and thresholding with typical pre-smoothing for a simulated phantom study. Red regions are false positives, blue regions indicate false negatives, their union comprises $\Delta(S_\gamma, \widehat{S})$. (a) Level set estimate; $\text{vol}(\Delta(S_\gamma, \widehat{S})) = 0.024$. (b) Voxel-wise threshold estimate with clairvoyant pre-smoothing; $\text{vol}(\Delta(S_\gamma, \widehat{S})) = 0.030$. (c) Voxel-wise threshold estimate with typical pre-smoothing; $\text{vol}(\Delta(S_\gamma, \widehat{S})) = 0.075$.

4.2. fMRI Data

We now turn our attention to fMRI data. This data consists of 64×64 axial slices with 122 time samples taken during a finger-tapping experiment. By splitting the data into two temporal halves, we generate two independent sets of data for the cross-validation method: the first 61 time samples for training, and the remaining 61 for validation. Shown in Fig. 2 are the results of our holdout procedure using two different levels of γ. By picking γ, we may examine different levels of activation in fMRI studies. We compare our results to voxel-wise thresholding with pre-smoothing. We used a Gaussian smoothing kernel with a 1.2 voxel FWHM, corresponding to the best choice of bandwidth in the simulated phantom experiment. The detected regions using the proposed level set estimation method show significantly fewer spurious detected areas compared to conventional methods, yielding better localization of neural activity.



(a) (b) (c) (d)

Fig. 2. Estimates of neural activation regions. The red regions overlaid onto the anatomic grayscale image represent the declared activity. With $\gamma = 0.95$: (a) level set estimate, (b) voxel-wise threshold estimate with pre-smoothing. With $\gamma = 0.90$: (c) level set estimate, (d) voxel-wise threshold estimate with pre-smoothing.

5. REMARKS AND CONCLUSIONS

In this paper, we abandoned standard hypothesis testing fMRI analysis in favor of a level set approach that aims to control the volume of the erroneous region. We argued that controlling the error volume is more natural in light of partial volume effects, which result in an unavoidable error in any event. The proposed level set estimator produces an estimate of active regions that is guaranteed to have an error that is commensurate with partial volume effects. In other words, the error volume, and hence the error in regional specificity, is controlled to nearly the absolute minimum. One matter that we have not investigated here is the choice of level, γ. The higher the level γ, the smaller the level set S_γ. In effect, γ gauges the strength of the activation in the set S_γ. The selection of γ could be based on a secondary criterion, such as the false discovery rate. Alternatively, it may be informative to examine several different level sets. These sets will be nested, with lower level sets containing higher level sets, providing a more descriptive view of activation patterns. Our ongoing work aims to address these matters.

6. REFERENCES

[1] J. Xiong, J. Gao, J. Lancaster, and P. Fox, "Assessment and optimization of functional mri analyses," *Human Brain Mapping*, vol. 4, pp. 153–167, 1996.

[2] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini, "Cluster-based analysis of fmri data," *NeuroImage*, vol. 33, no. 2, pp. 599–608, 2006.

[3] K. J. Friston, A. Holmes, J.-B. Poline, C. J. Price, and C. D. Frith, "Detecting activations in pet and fmri: Levels of inference and power," *Neuroimage*, vol. 40, pp. 223–235, 1996.

[4] D. Van De Ville, T. Blu, and M. Unser, "Integrated wavelet processing and spatial statistical testing of fmri data," *NeuroImage*, vol. 23, no. 4, pp. 1472–1485, 2004.

[5] D. O. Siegmund and K. J. Worsley, "Testing for a signal with unknown location and scale in a stationary gaussian random field," *Ann. Stat.*, vol. 23, pp. 608–639, 1994.

[6] R. Willett and R. Nowak, "Minimax optimal level set estimation," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2965–2979, 2007.

[7] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1983.

[8] C. Scott, *Dyadic Decision Trees*, Ph.D. thesis, Rice University, 2004.

[9] P. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Journal of Machine Learning*, vol. 48, pp. 85–113, 2002.