

A Statistical Approach to the Segmentation of MR Imagery and Volume Estimation of Stroke Lesions

Benjamin Stein and Joseph Horowitz

Dept. of Mathematics and Statistics, Univ. of Massachusetts,
Amherst, MA 01003

<http://www.math.umass.edu>

Abstract. We propose a three-dimensional method to segment magnetic resonance imagery (MRI) of ischemic stroke patients into lesion and background, and hence to estimate lesion volumes. It is a hierarchical, regularized method based on classical statistics that produces confidence regions for the lesion itself and a two-sided confidence interval for lesion volume. This approach requires a limited amount of user interaction to initialize. The procedure has been tested on real MR data, with volume estimates within 10% of those derived from doctors' hand segmentations. According to the physicians with whom we are working, these results are clinically useful to evaluate stroke therapies.

1 Introduction

In evaluating therapies for ischemic stroke patients, many physicians are interested in finding consistent, reliable estimates of lesion volume from magnetic resonance images (MRI). To do this, we present a common-sense algorithm, guided by classical statistical theory, called “packing.”

Several other research groups have used statistical approaches to segment tissue types in MRI (see [1, 4, 8, 9], e.g.), with varying degrees of user interaction, but with no single method emerging as superior. As in those methods, we are concerned with producing consistent estimates with limited user interaction, but our procedure goes beyond them in producing an assessment of the *error* of our estimate, in the form of an inner and outer confidence region for the lesion and a two-sided confidence interval for lesion volume. While this idea has been explored in [7] for near-infrared imaging, we do not know of any such results for MRI.

We have tested the packing method on real MRI and synthetic data, with promising results in all cases. For six sets of real imagery, the method consistently estimates lesion volumes to within 10% of those derived from doctors' hand segmentations. According to the doctors, these results are clinically useful.

In section 2 we outline the packing method and give its desirable statistical properties. We summarize the results for real and synthetic data in section 3. In section 4 we discuss the limitations of the packing method and the possibilities for future work.

2 The Packing Method

2.1 Initialization

We assume the lesion is brighter than its surroundings, i.e., the mean intensity of lesion voxels, μ_L , is higher than that of any other tissue type in the three-dimensional region of interest (ROI). The diffusion-weighted pulse sequence shows an ischemic stroke lesion as the only bright object, and thus the ROI can be the entire set of imagery. However, for pulse sequences such as T_2 -weighted imagery and FLAIR, the user must extract the ROI manually.

Aside from choosing a ROI, the only other manual step necessary to initialize the process is to sample the data. The user chooses a “base slice” from the stack of two-dimensional images, and two regions bounded by closed contours in that slice: one region completely inside the lesion and the other in the background (see figure 1). The pixels inside these contours constitute the lesion and background samples that our statistical analysis will be based upon. This step typically requires less than a minute of user interaction.

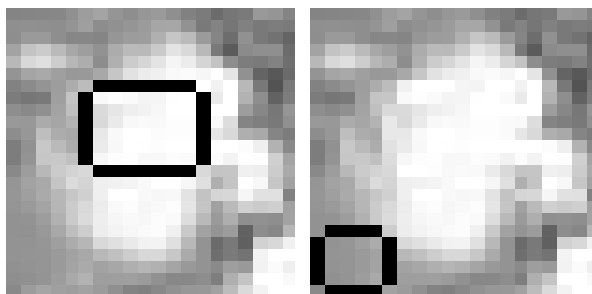


Fig. 1. Left: Lesion sample from the base slice. The boundary of the sample is superimposed onto the original subimage. Right: Background sample from the base slice.

2.2 Coarse-Grid Segmentation

After selecting the lesion and background samples, the method continues with no more outside assistance. We will explain this multi-step procedure by first explaining the mechanics of it at step i , $i = 1, \dots, D$. Usually $D = 2$ or 3 , but we give the results for general D for use in other applications. We cover the ROI by cubes of edge length d_i (cubes that can intersect one another), so that each cube contains d_i^3 voxels.

Let $c_{im}, m = 1, \dots, M_i$, denote the cubes in the covering of the ROI at step i that are to be tested. For each cube c_{im} , define the region L_{im} as the set of voxels already classified as lesion; this constitutes the current lesion estimate,

and L_0 is the original lesion sample, as described above. For a cube c_{im} that borders but does not intersect L_{im} , we consider the following hypothesis test:

$$\begin{aligned} H_0 &: c_{im} \text{ is entirely inside the lesion} \\ H_A &: c_{im} \text{ is not entirely inside the lesion.} \end{aligned}$$

More precisely, the alternative H_A says that there is at least one voxel in c_{im} that is not completely inside the lesion.

Let S denote the (finite) ROI, and X_k be the intensity at $k \in S$. Given L_{im} , the mean of the intensities classified as lesion, call it $\bar{x}_{L_{im}}$, is a fixed constant. For regularization purposes, we will use for all tests in step i the mean of the lesion voxels at the end of step $i - 1$. Using the notation above, this mean is called \bar{x}_{L_{i-1}, M_i} ; for simplicity, we will call it $\bar{x}_{L_{i-1}}$.

We will test each cube c_{im} conditionally, given L_{im} . We can translate the above qualitative set of hypotheses into more precise language by introducing μ_{im} , the mean intensity of the cube being tested:

$$\begin{aligned} H_0 &: \mu_{im} \geq \bar{x}_{L_{i-1}} & (1) \\ H_A &: \mu_{im} < \bar{x}_{L_{i-1}} & (2) \end{aligned}$$

Note that the value $\bar{x}_{L_{i-1}}$ is a fixed constant for a conditional test, which we set up as follows. For a cube c_{im} that borders but does not intersect L_{im} , consider testing the hypotheses (1) and (2) according to the decision rule from the usual one-sample t -test: accept H_0 if

$$\bar{x}_{im} > \bar{x}_{L_{i-1}} - t_i\left(\frac{\alpha}{DN_i}\right) \sqrt{\frac{s_{im}^2}{d_i^3}}, \quad (3)$$

where \bar{x}_{im} is the sample mean of voxels in the cube, $t_i(\frac{\alpha}{DN_i})$ is the $1 - \frac{\alpha}{DN_i}$ quantile of a Student's t -distribution with $d_i^3 - 1$ degrees of freedom, α is a fixed constant between 0 and 1 (usually $\alpha = .05$ or $.1$), N_i is a deterministic upper bound of the number of cubes tested in step i , and s_{im}^2 is the sample variance of voxels in the cube. The t -quantile is chosen in accordance with the Bonferroni method of multiple comparisons [3].

When a cube is accepted into the lesion, we will classify those voxels as such and not test them again. We then update the candidate cubes to be tested in step i as those that abut the new lesion estimate, and use the decision rule in (3) for each such cube. The step terminates when all candidate cubes have been tested.

Denote by R_{im} the event that a cube c_{im} is not accepted into the lesion. Using (3), we find the probability of not accepting a cube into the lesion when indeed it should be accepted. Assuming that $\bar{x}_{L_{i-1}} \approx \mu_L$, we have (approximately) that

$$P(R_{im}) \leq \frac{\alpha}{DN_i}. \quad (4)$$

A more careful statement of this and its proof are in [6], as well as proofs of upcoming results.

Moreover, Bonferroni’s inequality [3] implies that the *overall* probability of a type I error for all the tests at step i can be controlled. That is, let \mathcal{L}_i denote the set of cubes $C_{im}, m = 1, \dots, M_i$, that are actually in the lesion. Then

$$P(\text{all } C_{im} \in \mathcal{L}_i \text{ are accepted into the lesion}) \geq 1 - \frac{\alpha}{D}. \quad (5)$$

2.3 Coarse-to-Fine Aspect

After step i is completed, we use L_{iM_i} as the lesion sample to begin step $i + 1$, which considers smaller cubes of edge length $d_{i+1} < d_i$. This coarse-to-fine approach allows us to update our segmentation into a more accurate one by “packing” it with smaller cubes (see figure 2).

The final segmentation of the lesion is L_D , where $d_D = 1$. Clearly, the t -quantile is too large for use, or does not exist at all, if d_i is small; at these steps we consider unions of cubes, rather than single cubes. This increases the degrees of freedom in the test while decreasing the number of tests necessary at each step. The result is a useful t -quantile and a more powerful set of tests.

Equation (5) controls the overall probability of a type I error at each step; similarly, for the entire multi-step procedure, we have

$$P\left(\bigcap_{i=1}^D [\text{all } C_{im} \in \mathcal{L}_i \text{ are accepted into the lesion}]\right) \geq 1 - \alpha. \quad (6)$$



Fig. 2. Coarse-to-fine aspect of packing. Left: Sample slice (restricted to a ROI). Center: Lesion segmentation after $4 \times 4 \times 4$ boxes are tested. Right: Final segmentation after $2 \times 2 \times 2$ cubes and single voxels are tested, which overestimates the lesion (see section 2.4).

2.4 One-sided Confidence Bound for Volume

Let L_F be the final union of voxels classified as lesion. The F stands for “forward” packing, which we will explain shortly. Let L be the true lesion; since L_F is a

union of the voxels that are, by (6), classified as lesion with high confidence, it follows that L_F gives an *outer confidence region* for L :

$$P(L \subset L_F) \geq 1 - \alpha. \quad (7)$$

Let $V_F = |L_F|$, the number of voxels classified as lesion. We can translate (7) into a statement about the true volume V of the lesion:

$$P(L \subset L_F) \leq P(V \leq V_F) \quad (8)$$

and hence

$$P(V \leq V_F) \geq 1 - \alpha. \quad (9)$$

Therefore, V_F can be regarded as the upper bound of a one-sided $(1 - \alpha)100\%$ confidence interval for V . We do not control the probability of admitting cubes that do not belong in the lesion. Cubes containing a mixture of lesion and background voxels are especially susceptible of being incorrectly admitted, resulting in an overestimate of V by V_F .

2.5 Two-Sided Confidence Interval and Point Estimate for Volume

If we pack the *background* in a similar way as above, we obtain another segmentation of the image. The complement of the final background region is another confidence region of the lesion, call it L_B , which satisfies

$$P(L_B \subset L) \geq 1 - \alpha.$$

We also obtain a lower confidence bound V_B for the lesion volume V :

$$P(V_B \leq V) \geq 1 - \alpha.$$

Combining these results with those from forward packing in section 2.4, we have a confidence set for the true lesion L :

$$P(L_B \subset L \subset L_F) \geq 1 - 2\alpha, \quad (10)$$

and the set $L_F \setminus L_B$ is a confidence set for the boundary of L . In terms of lesion volume,

$$P(V_B \leq V \leq V_F) \geq 1 - 2\alpha. \quad (11)$$

Therefore, (V_B, V_F) is a $(1 - 2\alpha)100\%$ confidence interval for the lesion volume. See figure 3 for an example of confidence regions for a slice of MR imagery.

3 Experiments and Results

The t -test (3) requires that intensity data be independent and normally distributed (*i.i.d. normal*). These assumptions are made in [2, 4, 9] without verification. A physical argument, by considering the image formation process, is made in [5], and some analysis of real MRI is presented in [6] for further evidence of i.i.d. normal intensities. Thus we assume the assumptions are correct.

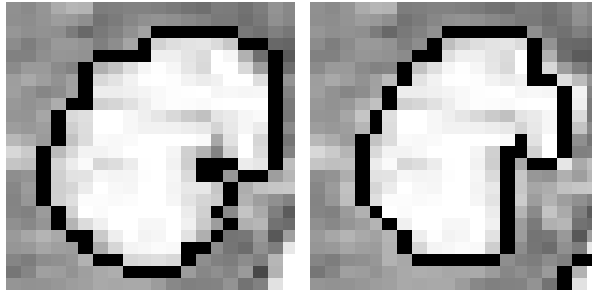


Fig. 3. Results superimposed onto subimage. Left: Lesion segmentation via forward packing (an outer confidence region). Right: Lesion segmentation using background packing (an inner confidence region).

3.1 Synthetic Data

Based on the above assumption, we can develop synthetic imagery to test the packing method in a controlled environment in which the volume of a “lesion” is known. Each synthetic lesion is modeled as a union of connected spheres placed randomly into a $30 \times 30 \times 30$ ROI, a size comparable to real MR data. The mean intensities are chosen to be smaller for sites far away from the centers of the spheres, with a precipitous decrease in the mean when the sites change from lesion to background. These varying intensities are meant to reflect the inherent heterogeneity within tissue types in real MRI. Finally, in keeping with the i.i.d. normal assumption, a white noise process is added to the imagery. The means and standard deviations of intensities in each tissue type are in keeping with those of real imagery (see figure 4).

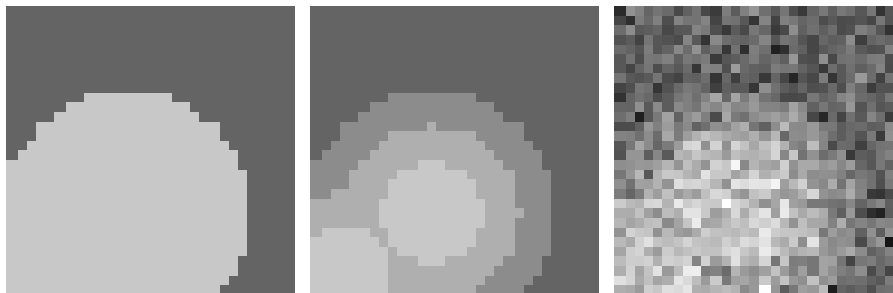


Fig. 4. Synthetic image construction. Left: Binary image. Center: Mean intensities adjusted to reflect the inhomogeneity of real MRI. Right: White noise added.

To test the packing method, we randomly select a union of spheres and corrupt it with an independent white noise process 50 times. Each time we segment the images with the packing method and record the resulting confidence bounds for “lesion” volume. We have chosen $\alpha = .05$, which gives a 90% confidence interval. A summary of the results for five synthetic data sets are in table 1.

Lower CB		Actual	Upper CB	
mean	sd		mean	sd
1801	158.3	2047	2286	113.6
2804	160.1	3209	3402	154.0
3575	170.1	3987	4295	181.2
6020	207.4	6432	6894	228.3
6291	235.0	6794	7317	230.3

Table 1. Volume estimates for synthetic images using the packing method. The actual volumes of the synthetic lesions are listed in the middle column. The left-hand columns are the average lower 90% confidence bounds, based on 50 runs of the white noise process, and the corresponding standard deviations. The right-hand columns are the averages and standard deviations of the upper 90% confidence bounds.

We see from the results in table 1 that the mean half-width of each interval is within 10% of the actual volume. This means that the midpoint of each interval will be accurate in estimating the actual volume to within 10%, approximately. As for the coverage rates of the individual intervals, we expect the actual volume to be captured by the intervals 90% of the time. Our intervals, in fact, perform better than the expected rate. Intervals for the two smallest “lesions” do not capture the actual volume five times, as expected, but for the two largest, the volume is not captured only once in each. We also see a decrease in standard deviation as the actual volume increases (as a percentage of the actual volume). Thus it appears that the larger the percentage of “lesion” voxels in the ROI, the more accurate and reliable our results.

3.2 Real MRI

The synthetic data indicates that the packing method is producing viable 90% confidence intervals, and point estimates that estimate the actual volume to within 10% of the actual volume. Now we will use the procedure to segment six real MR data sets, obtained from Baystate Medical Center (Springfield, MA) using a 1.5 T Picker Edge machine. Patient 1 was imaged using a T_2 -weighted spin-echo pulse sequence and has 1 mm slice thicknesses; the remaining five sets come from a FLAIR pulse sequence with 2.5 mm thicknesses.

For validation we obtained two hand segmentations from two physicians at Baystate, giving a total of four lesion volume estimates. Although a hand segmentation is considered the “gold standard” in MR image segmentation, there is still intra- and inter-observer variability.

We used the packing method to segment each set of MRI. Each initialization of the method, which includes choosing a base slice and taking samples of homogeneous tissue intensities from it, produces different confidence regions and intervals. We have chosen four such initializations (two samples each from two different base slices) for comparison with the hand segmentations. The results for all six data sets are in table 2.

Patient	Lower CB		Hand Segs		Upper CB	
	mean	sd	mean	sd	mean	sd
1	1869	74.9	2008	122.2	2182	89.2
2	2432	152.8	3517	176.8	3979	286.5
3	20230	1125.8	22457	1389	25277	1601.2
4	9654	340.1	10375	543.7	11421	582.8
5	8376	353.5	9754	865.9	10873	755.9
6	2892	269.4	3478	99.6	3932	237.8

Table 2. Means and standard deviations of confidence bounds for six real MR data sets, based on four initializations of the packing method for each patient. The center columns show the grand means and standard deviations of volumes from the doctors' hand segmentations.

Note that the standard deviations for the confidence bounds are higher than for the synthetic imagery, except for Patient 1, which is the high resolution T_2 data with a 1 mm slice thickness. The variability in the FLAIR imagery (with 2.5 mm slice thickness) could be due to the voxel effect, which creates intensity inhomogeneity. Also, lesions with irregular shapes can be difficult to sample from; this is the primary reason why we see higher standard deviations in the upper confidence bounds, which require lesion samples for initialization.

As before, the intervals capture the average of the hand segmentations for each patient. Looking at individual estimates, of the four intervals for the Patient 1 data, one interval does not capture one of the volumes arising from the hand segmentations; this occurs for Patient 6 as well. As with the synthetic data, it appears that the method is more reliable for the larger lesions.

The widths of the intervals are more variable than with the synthetic data. The Patient 2 data set produces an average interval half-width that is over 20% of the mean of the hand segmentations. This set features a small, irregular lesion that is especially difficult to consistently classify. Similarly, we note a nearly 20% underestimate in Patients 5 and 6. See figures 5–7 for sample slices and confidence regions for Patients 1, 2, and 6.

Despite some dissatisfying results for interval widths, we can still derive useful information from them. Using the midpoint as a point estimator for actual lesion volume, the worst error, as a percentage of the mean of the hand segmented volumes, is 9% (Patient 2). The physicians required that our method be accurate to within 20% of the actual volume, so it appears that the packing method produces clinically useful volume estimates.

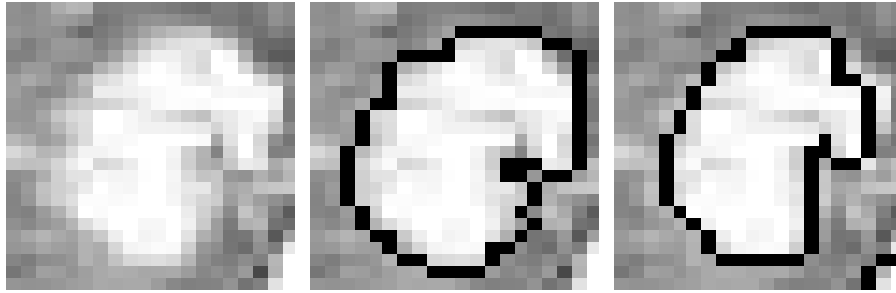


Fig. 5. Left: Sample subimage of Patient 1. Center: Outer confidence region. Right: Inner confidence region.

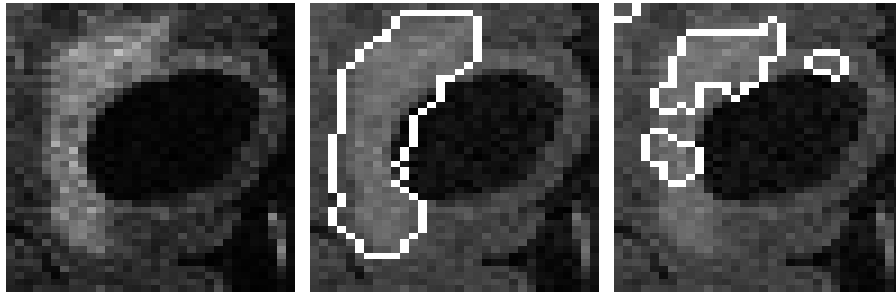


Fig. 6. Left: Sample subimage of Patient 2. Center: Outer confidence region. Right: Inner confidence region.

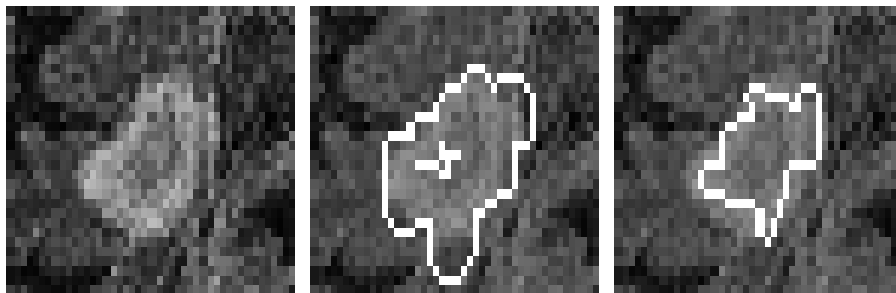


Fig. 7. Left: Sample subimage of Patient 6. Center: Outer confidence region. Right: Inner confidence region.

4 Conclusion

We see from the results in section 3 that the packing method is performing well across a diverse set of synthetic and real MR imagery. The confidence intervals for lesion volume, developed in section 2, capture the actual volume at a better rate than we expect; in practice we use 90% intervals, but the method captures the actual volume more often than 90% of the time.

The interval widths are high in some circumstances, particularly for low contrast imagery. However, even in the extreme cases, the midpoint of the interval still estimates to within 10% of the actual lesion volume, which is a clinically useful result. The error rate is substantially lower for many of the synthetic and real data sets.

Certainly these results are encouraging, and a good starting point for further research. The packing method is a new approach to lesion volume estimation, and several avenues for improvement remain, including the need for faster initialization, systematic data pre-processing, smaller confidence bounds, and better point estimation. Also, the packing method is “soft” in that it does not use the specifics of the MR image formation process to segment the imagery; an approach which integrates packing with this information could produce a more reliable lesion estimate.

Acknowledgements

Thanks to Richard Hicks (Radiology) and George Howard (Neurology) of Baystate Medical Center, Springfield, MA, for the imagery and the hand segmentations used in this report.

References

1. R. Adams and L. Bischof. Seeded region growing. *IEEE Trans on Pattern Analysis and Machine Intelligence* 16:641–647, 1994.
2. H. Cline, W.E. Lorensen, R. Kikinis, F. Jolesz. Three-dimensional segmentation of MR images of the head using probability and connectivity. *Journal of Computer Assisted Tomography*, 14(6):1037–1045, 1990.
3. R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, New Jersey, 1992.
4. A. Martel, S. Alder, G. Delay, P. Morgan, and A.R. Moody. Measurement of infarct volume in stroke patients using adaptive segmentation of diffusion weighted MR images. *Proceedings of the Medical Image Computing and Computer-Assisted Intervention Conference*, 1999.
5. J. Sijbers, A.J. Den Dekker, P. Scheunders, and D. Van Dyck. Maximum likelihood estimation of Rician distribution parameters. *IEEE Trans on Medical Imaging*, 17:357–361, 1998.
6. B. Stein. *Signal Formulation, Segmentation, and Lesion Volume Estimation in Magnetic Resonance Images*. Ph.D. dissertation, University of Massachusetts, 2001.

7. T. Tosteson, B. Pogue, E. Demidenko, T. McBride, and K. Paulsen. Confidence maps and confidence intervals for near infrared images in breast cancer. *IEEE Trans on Medical Imaging*, 18:1188–1193, 1999.
8. C. Watson, C. Jack Jr., and F. Cendes. Volumetric magnetic resonance imaging. *Archives of Neurology*, 54:1521–1531, 1997.
9. W. Wells, R. Kikinis, W. Grimson, and F. Jolesz. Adaptive segmentation of MRI data. *IEEE Trans on Medical Imaging*, 15:429–442, 1996.