Statistical bias in optimized VBM

Nicholas J. Tustison[a], Brian B. Avants[b], Philip A. Cook[b], James C. Gee[b], and James R. Stone[a]

[a]Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, Virginia, USA
[b]Penn Image Computing and Science Laboratory, University of Pennsylvania, Philadelphia, Pennsylvania, USA

ABSTRACT

The recent discovery of methodological flaws in experimental design and analysis in neuroscience research has raised concerns over the validity of certain techniques used in routine analyses and their corresponding findings. Such concerns have centered around selection bias whereby data is inadvertently manipulated such that the resulting analysis produces falsely increased statistical significance, i.e. type I errors. This has been illustrated recently in fMRI studies,[1] with excessive flexibility in data collection,[2] and general experimental design issues.[3] Current work from our group has shown how this problem extends to generic voxel-based analysis (and certain technique derivatives such as tract-based spatial statistics[4]) using fractional anisotropy images derived from diffusion tensor imaging.[5] In this work, we demonstrate how this circularity principle can potentially extend to the well-known optimized voxel-based morphometry technique[6] for assessing cortical density differences whereby the principal cause of experimental corruption is due to normalization strategy. Specifically, the popular sum-of-squared-differences (SSD) metric explicitly optimizes statistical findings potentially inflating type I errors. Additional experimentation demonstrates that this problem is not restricted to the SSD metric but extends to other commonly used metrics such as mutual information, neighborhood cross correlation, and Demons.

Keywords: circularity, cortical thickness, normalization

1. INTRODUCTION

Voxel-based morphometry (VBM)[7,8] has proven to be an invaluable technique for the characterization of cortical structural differences between populations using neuroimaging data. Subsequent work using VBM has explored fundamental neuroscience hypotheses and has, as a result, become a vital analysis technique in the neuroscientist's computational toolbox. Briefly, as originally proposed, the standard VBM preprocessing protocol aligns each image to a T1-weighted template prior to probabilistic segmentation which delineates the white matter, gray matter, and cerebrospinal fluid image regions. Spatial normalization is typically performed on each subject's T1-weighted image using a global affine normalization procedure to the template to account for global shape differences followed by deformable registration to minimize local anatomical differences. A modification of the standard VBM technique, known as "optimized VBM", was introduced with the work of Good et al.[6] Instead of spatial normalization to the T1-weighted brain template, an initial segmentation occurs in the space of each subject. The gray matter probability images are then normalized to a gray matter probabilistic template.* The normalization parameters are then applied to the corresponding T1-weighted brain images. The remaining portion of the workflow mirrors standard VBM.

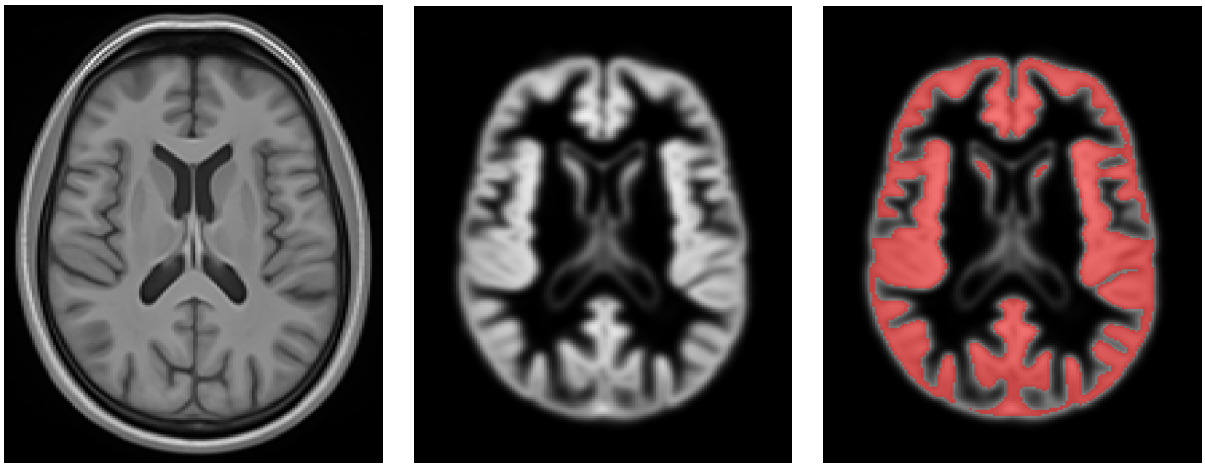
The crucial distinction between the two methods is the normalization step and the images used to find the transformation between each subject and the template. Implicitly quantifying the anatomical differences for many implementations is the SSD metric. In fact, the SSD is the metric originally proposed in both methods and is used in popular, publicly available methods such as SPM2† and FSL-VBM.‡

Further author information: (Send correspondence to N.J.T.)
N.J.T.: E-mail: njt4n@virginia.edu, Telephone: 1 434 924 7730

*Note that these steps are also applicable to analysis of the white matter.

†http://www.fil.ion.ucl.ac.uk/spm/software/spm2/

‡http://www.fmrib.ox.ac.uk/fsl/fslvbm/index.html



(a) (b) (c)

Figure 1. (a) The T1 template created from the set of $56+29=85$ T1 IXI images used in this study. (b) The gray matter probability template created from the gray matter probability maps. (c) Mask region defining the domain of the VBM calculations (average gray matter probability ≥ 0.5).

Closer inspection of the optimization problem of finding the set of transformations mapping the entire set of images to the gray matter template using SSD reveals the circularity of the optimized approach. Optimization entails finding the set of M optimal transforms, $\mathcal{T}^* = \{\mathcal{T}_1, \ldots, \mathcal{T}_M\}$ which maps the M gray matter probability images $\{\mathcal{I}_1^{gm}, \ldots, \mathcal{I}_M^{gm}\}$ to the gray matter template $\mathcal{J}^{gm}$, i.e.

$$\mathcal{T}^* = \underset{\mathcal{T}_1,\ldots,\mathcal{T}_M}{\operatorname{argmin}} \frac{1}{N} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(\mathcal{I}^{gm}(\mathcal{T}_m(x_n)) - \mathcal{J}^{gm}(x_n)\right)^2 . \tag{1}$$

Switching the summations,

$$\mathcal{T}^* = \underset{\mathcal{T}_1,\ldots,\mathcal{T}_M}{\operatorname{argmin}} \sum_{n=1}^{N} \underbrace{\frac{1}{N} \sum_{m=1}^{M} \left(\mathcal{I}^{gm}\left(\mathcal{T}_m(x_n)\right) - \mathcal{J}^{gm}(x_n)\right)^2}_{\approx average\ voxelwise\ variance}, \tag{2}$$

and recognizing that the template gray matter probability image is a good approximation of the average gray matter probability image, it becomes apparent that the optimization intended to minimize local shape differences instead, and more problematically, explicitly minimizes the average variance which directly maximizes Student's t-test. Thus, standard preprocessing normalization in optimized VBM, instead of maximizing anatomical alignment, directly manipulates the data such that the statistical testing results are maximized which, by definition,[3] is circular.

2. MATERIALS AND METHODS

Although Eq. (2) demonstrates the susceptibility of the SSD difference metric to producing false positives, we illustrate such effects on public data for mutual information (MI), neighborhood cross correlation (CC), Demons, and SSD. The relationship of the Demons metric to SSD would give prior indication that similar effects should be seen whereas the effect should be lesser for MI and CC given that the similarity measures are less localized in addition to not being based on absolute intensity differences.

2.1 Images

Although any of the several high quality public data sets could be used to demonstrate the presence of circular analysis in optimized VBM, we use the IXI data set consisting of approximately 600 subjects from three different

Table 1. Change in volume of significant regions relative to initial alignment

	0 mm	4 mm	8 mm	16 mm
MI	-62%	-87%	-91%	-90%
CC	-46%	-58%	-46%	-33%
SSD	26%	14%	-8%	10%
Demons	16%	27%	-3%	34%

locations. Demographic information including gender and age is also available for each subject. Further details can be found on the IXI website.[§] For the study in the paper, we used the T1 images from two male subgroups in the age ranges of 20 to 30 years and 60 to 80 years. The former subgroup consists of 29 total images and the latter consists of 56 images.

2.2 Generation of Aligned Cohort

The combined set of T1 images were first used to create an average unbiased shape and intensity atlas[9] (see Fig. 1(a)). This process also produced a set of transformations which warps each image to the template space. Each T1 was then segmented in its native space using the Atropos[10] tool in ANTs. The resulting gray matter probability maps were then warped to the template space where they were averaged to create an average gray probability map (denoted as the "gray matter probability template" shown in Fig. 1(b)). The two aligned cohorts were then used to create two sets of simulated, aligned gray matter probability maps reflecting the characteristic intensities of group. These probability maps were created by calculating the voxelwise mean and standard deviation of the aligned gray matter probability maps within a given cohort. The voxel intensity value of the simulated gray matter probability was then randomly generated using these voxelwise Gaussian parameters. For our experiments, we generated 21 synthetic gray matter probability maps from the 20-to-30 group and 20 synthetic gray matter probability maps from the 60-to-80 group.

Using the four similarity metrics described above, we then spatially normalized each synthetic gray matter probability map to the gray matter probability template.[¶] Despite the fact that each synthetic image was generated in situ in the coordinate system of the template, according to the discussion above in deriving Eq. (2), optimizing the different similarity metrics will generate different alignments yielding different VBM results with the SSD and Demons metric falsely increasing the volume of statistically significant regions.

3. RESULTS

Following basic VBM protocol, voxelwise t-tests were performed between each synthetic group over all four metrics and for different levels of smoothing (0 mm, 4 mm, 8 mm, and 16 mm given in full width at half maximum (FWHM)). We then thresholded the resulting p-values at $p \leq 0.05$ to yield significant regions. We then calculated the volume of significant regions for each metric case and compared the volume to the results of no normalization with the same smoothing level. The results are shown in Table 1. Positive values are indicative of increased statistical significance after spatial normalization. Whereas the less localized metrics which are not dependent on absolute intensity differences, i.e. CC and MI, demonstrate a decrease in statistical significance with normalization, the Demons and SSD metrics trend towards a substantial increase in statistically significant volume as predicted by Eq. (2).

4. CONCLUSIONS

Spatial normalization is a crucial preprocessing step in many fundamental analysis techniques, including optimized VBM. However, as we point out, certain normalization strategies (which are commonly used in various analyses) are prone to statistical bias by falsely elevating statistical significance. As we have shown, certain metrics such as Demons and SSD, explicitly optimize the average voxelwise t-statistic instead of maximizing

[§]http://biomedic.doc.ic.ac.uk/brain-development/

[¶]We used the `antsRegistration` tool also available in ANTs. Specific common command line parameters were: `--convergence [10,0,10] --shrink-factors 1 --smoothing-sigmas 0 --transform SyN[0.5,3,0]`.

anatomical alignment. We encourage researchers to reexamine their own workflow to avoid such problematic practices.

APPENDIX

Given that the Student's t-statistic is estimated from both the group variance and the difference in group means, an immediate question might concern the possibility of a concomitant decrease in group mean difference thereby negating the decrease in group difference caused by group normalization. It is important to note that it is the *average* group variance that goes down and not necessarily the variance at a single voxel site. Additionally, we can look at the average group mean difference, $\overline{\Delta\mu}$, over the entire image domain of interest and manipulate the summation ordering which yields

$$\overline{\Delta\mu} = \frac{1}{N}\sum_{n=1}^{N}\left(\mu_1(x_n) - \mu_2(x_n)\right) \tag{3}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M_1}\sum_{m_1=1}^{M_1}\mathcal{I}_{m_1}^{gm}(T_{m_1}(x_n)) - \frac{1}{M_2}\sum_{m_2=1}^{M_2}\mathcal{I}_{m_2}^{gm}(T_{m_2}(x_n))\right) \tag{4}$$

$$= \left(\frac{1}{M_1}\sum_{m_1=1}^{M_1}\frac{1}{N}\sum_{n=1}^{N}\mathcal{I}_{m_1}^{gm}\left(T_{m_1}(x_n)\right)\right) - \left(\frac{1}{M_2}\sum_{m_2=1}^{M_2}\frac{1}{N}\sum_{n=1}^{N}\mathcal{I}_{m_2}^{gm}\left(T_{m_2}(x_n)\right)\right). \tag{5}$$

The inner summations in the two terms make apparent that the average group mean difference is dependent solely on how the pixel intensity values are mapped within the coordinate space defined by the template. In other words, $\overline{\Delta\mu}$ depends on the difference between the aggregate intensities of group 1 and group 2 which is not being directly optimized via minimization of the similarity metric (unlike the average group variance). For example, if one combined an intensity-preserving transformation model with the SSD metric, although the average variance is guaranteed to decrease, the average group mean difference will remain constant throughout the spatial normalization optimization.

REFERENCES

1. E. Vul, C. Harris, P. Winkielman, and H. Pashler, "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition," *Perspectives on Psychological Science* **4**, pp. 274–290, May 2009.
2. J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychol Sci* **22**, pp. 1359–66, Nov 2011.
3. N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nat Neurosci* **12**, pp. 535–40, May 2009.
4. S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews, and T. E. J. Behrens, "Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data.," *Neuroimage* **31**, pp. 1487–1505, Jul 2006.
5. N. J. Tustison, B. B. Avants, P. A. Cook, J. Kim, J. Whyte, J. C. Gee, and J. R. Stone, "Logical circularity in voxel-based analysis: Normalization strategy may induce statistical bias," *Hum Brain Mapp* , Nov 2012.
6. C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains.," *Neuroimage* **14**, pp. 21–36, Jul 2001.
7. I. C. Wright, P. K. McGuire, J. B. Poline, J. M. Travere, R. M. Murray, C. D. Frith, R. S. Frackowiak, and K. J. Friston, "A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia," *Neuroimage* **2**, pp. 244–52, Dec 1995.
8. J. Ashburner and K. J. Friston, "Voxel-based morphometry–the methods.," *Neuroimage* **11**, pp. 805–821, Jun 2000.
9. B. B. Avants, P. Yushkevich, J. Pluta, D. Minkoff, M. Korczykowski, J. Detre, and J. C. Gee, "The optimal template effect in hippocampus studies of diseased populations.," *Neuroimage* **49**, pp. 2457–2466, Feb 2010.
10. B. B. Avants, N. J. Tustison, J. Wu, P. A. Cook, and J. C. Gee, "An open source multivariate framework for n-tissue segmentation with evaluation on public data.," *Neuroinformatics* , Mar 2011.