

Distributional Assumptions in Voxel-Based Morphometry

C. H. Salmond,^{*,†} J. Ashburner,[‡] F. Vargha-Khadem,^{*} A. Connelly,[†] D. G. Gadian,[‡] and K. J. Friston[‡]

^{*}Developmental Cognitive Neuroscience Unit, Institute of Child Health, University College London, London, United Kingdom; [†]Radiology and Physics Unit, Institute of Child Health, University College London, London, United Kingdom; and [‡]The Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College London, Queen Square, London, United Kingdom

Received October 30, 2001

In this paper we address the assumptions about the distribution of errors made by voxel-based morphometry. Voxel-based morphometry (VBM) uses the general linear model to construct parametric statistical tests. In order for these statistics to be valid, a small number of assumptions must hold. A key assumption is that the model's error terms are normally distributed. This is usually ensured through the Central Limit Theorem by smoothing the data. However, there is increasing interest in using minimal smoothing (in order to sensitize the analysis to regional differences at a small spatial scale). The validity of such analyses is investigated. In brief, our results indicate that non-normality in the error terms can be an issue in VBM. However, in balanced designs, provided the data are smoothed with a 4-mm FWHM kernel, nonnormality is sufficiently attenuated to render the tests valid. Unbalanced designs appear to be less robust to violations of normality: a significant number of false positives arise at a smoothing of 4 and 8 mm when comparing a single subject to a group. This is despite the fact that conventional group comparisons appear to be robust, remaining valid even with no smoothing. The implications of the results for researchers using voxel-based morphometry are discussed. © 2002 Elsevier Science (USA)

INTRODUCTION

Voxel-based morphometry was developed to characterize cerebral gray and white matter differences in structural MRI scans. In contrast to methods that frame the search in terms of regions of interest, voxel-based morphometry can detect structural differences throughout the brain. Voxel-based morphometry is essentially a technique that compares images (segments) of gray (or white) matter (obtained from segmented MRI images). This comparison uses statistical parametric mapping to identify, and make inferences about, regionally specific differences.

Voxel-based morphometry involves spatially normalizing all the images into the same stereotactic space, extracting the gray matter (or white matter) from the

normalized images, smoothing, and finally performing a statistical analysis to localize and make inferences about group differences. The output of the method is a statistical parametric map showing regions where gray matter density differs significantly among groups.

Normality of Residuals

The parametric statistical tests are carried out within the framework of the general linear model. In order for these tests to be valid, the errors must be normally distributed. Prior to smoothing, the segmented images may have a highly nonnormal density function, with most voxels having a value close to the extremes of the range of 0–1. This is because the voxel values in the segments correspond to the probability that the voxel is gray matter. The distribution of errors about any group mean will show a similar nonnormal distribution. However by appealing to the Central Limit Theorem, it is generally assumed that the errors are rendered normally distributed by spatial smoothing.

An important question in this context is what circumstances would result in deviations from normality sufficient to render the tests invalid? One potential situation is the comparison of a single subject and a group. In parametric statistics it is assumed that the group difference (more formally a contrast of parameter estimates) is normally distributed. Dividing this contrast by an estimate of its standard deviation gives the *t* statistic. Generally, this difference will be well-behaved because the group mean represents an average over many observations and will have a normal distribution by the central limit theorem. However, when one of the groups has only one subject the difference may be highly nonnormal and the distribution of the ensuing statistic will not conform to parametric assumptions. In short, for most designs, inferences are quite robust to violations of normality, but there are some (e.g., unbalanced) designs that may be less robust. In other words, there may be an interaction between the degree of nonnormality and experimental design that renders the tests invalid.

Recent applications of voxel-based morphometry have included analyses with small smoothing kernels (e.g., 4 mm (Gadian *et al.*, 2000)) and comparisons of an individual versus a group of controls (e.g., Woermann *et al.*, 1999a,b). Smaller Gaussian kernels are used to sensitize the analysis to a spatial scale equivalent to the structure of interest (e.g., the hippocampal formation). Investigations into the neuropathology of single cases are particularly important in clinical diagnosis and the field of clinical neuropsychology, where individual cognitive and behavioral profiles prevent the formation of a homogeneous clinical group. The validity of the parametric statistical tests in this context is investigated in this work.

The aim of this paper was to establish a lower bound on the degree of smoothing applied during voxel-based morphometry that is imposed by considerations of statistical validity. The motivation for using low degrees of smoothing is to sensitize the analysis to small structures in accord with the matched filter theorem. However, we are not necessarily advocating the use of low degrees of smoothing to look at fine-grain anatomical differences. We are simply trying to establish the limits on smoothing that should be adopted in voxel-based morphometry. Fine-grain or high-resolution analyses of differences in anatomy would generally be finessed with deformation field and tensor-based morphometry. Voxel-based morphometry deliberately uses smooth deformation fields and smoothing of the gray-matter partitions to detect differences in the relative volumes of tissue partitions at a fairly low resolution.

As with all questions of model assumptions, the key one is not to prove that the assumptions are false. In reality, they must be false. Given enough data, one could always detect things like nonnormality. The important issue here is the degree of robustness to the effect that nonnormality has on the decisions or inferences made. In what follows we present a quantitative analysis of nonnormality using QQ plots. However, although the results from these analyses were reassuring, they do not directly address the issue of robustness. In a second step, we then move on to test robustness directly in terms of false positive inferences.

MATERIALS AND METHODS

In brief, to assess the lower bound on smoothing that is required to render VBM analyses valid, we adopted the following strategy. In the first instance we assessed the degree of nonnormality over all voxels included in the search volume using a metric of nonnormality. This metric comprised the correlation coefficient from a QQ plot. Although distributional approximations exist for this coefficient, we used simulations to establish its null distribution to avoid any assumptions about the behavior of our data. This entailed computing the non-normality coefficient for every voxel in real data and

comparing the number of nonnormal voxels, after thresholding, with the equivalent number based on simulated data that were exactly Gaussian. This is a somewhat descriptive exercise that allows us to assess, quantitatively, the degree of nonnormality expressed at low degrees of smoothing. However, the degree of nonnormality is operationally irrelevant in the sense that it is the impact of nonnormality on inference and false-positive rates that defines robustness. The second part of our analysis therefore compared the false-positive rates in a series of VBM analyses at different degrees of smoothness. Again, this involved comparing the expected and observed false-positive rates under a simple Poisson model. The expected false-positive rates were obtained using rerandomization strategies to provide surrogate data.

MRI Data Acquisition and Preprocessing

All subjects (20 children, mean age 13 years; 11 male and 9 female, with no known neurological or psychiatric history) were scanned on a 1.5 T Siemens Vision scanner, using a T1-weighted 3D MPRAGE sequence (Mugler and Brookeman, 1990) with the following parameters: TR, 9.7 ms; TE, 4 ms; TI, 300 ms; flip angle, 12°; matrix size, 256 × 256 × 128; field of view, 250 × 250 × 160 mm. The data were analyzed in SPM99 (Wellcome Department of Imaging Neuroscience, London, UK). Each scan was spatially normalized (Friston *et al.*, 1995; Ashburner and Friston, 1999). The images were then segmented using the Bayesian algorithm described in Ashburner and Friston (1997). This produced continuous probability maps where the values correspond to the posterior probability that a voxel belongs to the gray-matter partition. The gray-matter images were smoothed with 12, 8, 4, and 0 mm isotropic Gaussian kernels. This smoothing renders the voxel values an index of the amount of gray matter per unit volume under the smoothing kernel. The term “gray matter density” is generally used to refer to this measure.

Effects of Smoothing on Normality of Residuals

QQ plots. It is not possible to prove that data are normally distributed. However, it is possible to quantify the degree of nonnormality. One method is the QQ plot. A QQ plot is a plot of the sample quantile versus the sample quantile that would be expected if the data were normally distributed. For normally distributed data, the QQ plot of the data should be a straight line. A significant deviation from a straight line can be identified by computing the correlation coefficient of the plot (as described by Johnson and Wichern, 1998). If the correlation coefficient falls below a particular value, given a certain sample size, nonnormality can be inferred. For more information see Ashburner and Friston (1999).

Twelve sets of 10 scans were selected randomly, from the 20 possible scans, for three levels of smoothness (8, 4 and 0 mm). Correlation coefficients from a QQ plot were computed over all voxels where the mean intensity over all the images was greater than 0.05 (arbitrary units of probability). Voxels of low mean intensity were excluded as they would not be included in a conventional SPM analysis. The QQ plots were calculated using the residuals of a model that accounted for the confounding effects of age, sex, and total amount of gray matter in each volume.

This procedure was repeated but replacing the data with simulated Gaussian noise for comparison. The proportion of voxels where the correlation coefficient fell below the tabulated value (indicating residuals that were not normally distributed at $P < 0.05$) was used to quantify nonnormality by comparing this proportion in the real and simulated data.

Transforming the smoothed, segmented data with a "logit" transform prior to performing statistical tests may render the errors more normally distributed. This is because every voxel in the smoothed image segment has a value between 0 and 1. In order to assess the improvement the logit transform makes, the QQ analyses were calculated with and without the logit transform.

Rate of false positives. While QQ analyses can determine whether the data are not normally distributed, they cannot demonstrate the influences that any non-normality may have on subsequent statistical inference. One way of assessing this is to look at the false positive rate. The rate of false positives was assessed by randomly assigning the 20 children into two groups (each of size 10). Confounding factors of age, sex, and total amount of gray matter were included in the model. This was repeated a total of 10 times at three levels of smoothing (8, 4, and 0 mm). Significant increases and decreases in gray matter density were assessed, resulting in a total of 20 SPMs of the t statistic, for each smoothing level. The number of analyses with one or more false positives (at $P = 0.05$ corrected) was assessed.

Assuming false-positive SPMs are encountered like "rare events," we used the Poisson distribution to compare the probability of obtaining the observed number of SPMs with one or more maxima at a corrected level of significance. This probability assumes the tests are exact and normality has not been violated. Although these P values do not establish that VBM is valid, they do allow us to say that the tests are invalid if the P value falls below a critical threshold (i.e., $P = 0.05$).

Effects of Experimental Design on Robustness

The data from 17 of the 20 children were used in this investigation. One child was randomly chosen and compared against the remaining 16 children. This was

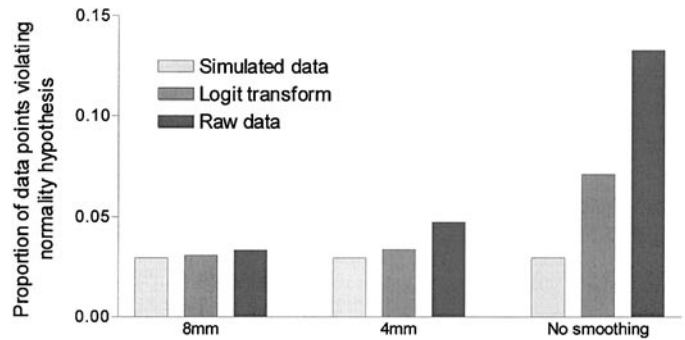


FIG. 1. QQ plots: The proportion of data points significantly violating the assumptions of normality at 8, 4, and 0 mm.

repeated a total of 10 times at all three levels of smoothing. Confounding factors of age, sex, and total amount of gray matter were included in the model. Significant increases and decreases in gray matter density were assessed resulting in a total of 20 SPMs of the t statistic, at three smoothing levels (12, 8, and 4 mm). The number of SPMs with one or more false positives (at $P = 0.05$ corrected) was assessed, and the probability of getting this number or more was computed as above with reference to the Poisson distribution.

RESULTS

Effects of Smoothing on Normality of Residuals

The effect of smoothing on balanced designs is shown in Fig. 1. It demonstrates that the proportion of voxels violating an assumption of normality (based on QQ plots) is below the expected limit (≤ 0.05) for smoothing kernels of 4 and 8 mm. The proportion under no smoothing (0 mm) is higher than the simulated data in both the logit transformed data and the raw data. The logit transform does reduce the proportion at all smoothing levels, suggesting any excess is indeed due to violations of the assumption of normality.

There were no analyses with one or more false positives at $P \leq 0.05$ (corrected) at any of the three smoothing levels (8, 4, and 0 mm).

Effects of Experimental Design on Robustness

The effect of unbalanced designs on robustness is shown in Fig. 2. It demonstrates that the number of SPMs with one or more false positives at $P \leq 0.05$ (corrected) decreases with smoothing. The false positive rate was within the expected range for 12 mm ($P = 0.4$). There were significantly more analyses with false positives than would be expected by chance at 4 and 8 mm (8 mm, $P = 0.01$; 4 mm, $P = 0.00001$). Figure 2 also demonstrates that no false positives were found in those that looked at increases in gray matter in the individual versus the group. As expected, the number

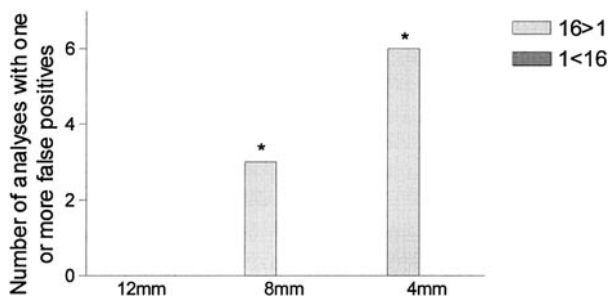


FIG. 2. Number of analyses with one or more false positives at 12, 8, and 4 mm: unbalanced design. $16 > 1$ refers to the contrast examining decreases in the individual versus the group, while $16 < 1$ refers to that examining increases in the individual versus the group. *Significantly more analyses meeting criteria than would be expected by chance.

of false positives was greater for the unbalanced design (1 vs 16) relative to the balanced design (10 vs 10).

DISCUSSION

Effects of Smoothing on Normality of Residuals

When using balanced group comparisons, smoothing at 4 mm appears to be sufficient to ensure that any nonnormality has little effect both as assessed using the QQ plots and the false-positive rate. This is consistent with, and extends the results of, Ashburner and Friston, who found that 12 mm was sufficient. The nonnormality detected by the QQ plots for unsmoothed data does not markedly inflate the rate of false positives. While this conclusion should be moderated by the limited number of SPMs used, it is not necessarily surprising. Despite the lack of spatial smoothing, the group analysis is rendered robust by averaging over subjects. By the central limit theorem this renders the contrasts normally distributed.

Effects of Experimental Design on Robustness

When comparing a single subject to a group, the false-positive rates were only within the expected range for smoothing kernels of 12 mm; i.e., we found no evidence that these analyses are invalid. Reducing the smoothing kernel to 8 or 4 mm does render the analyses significantly prone to false positives, suggesting that the tests are no longer robust. The implicit interaction between group size and smoothing on false positives is to be expected: (as discussed in the Introduction) group size (i.e., the design) may influence the robustness to violations of normality at low levels of smoothing. Reducing the smoothing kernel size to 4 mm when investigating an individual's neuropathology should therefore be avoided.

Increases versus Decreases in Gray Matter

An interesting observation was that the false positive rate appears to be systematically higher in the

tests for decreases in gray matter in an individual versus a group, compared to increases. This suggests a "skew" in the distribution of differences between a single subject and a group. In other words, the probability that a small region contains gray matter is skewed toward high values. This is intuitively sensible, even after smoothing, because the probability of a region being void of gray matter is much greater than the probability that it is all gray matter. This is because gray matter conforms to a sheet or manifold embedded in a volume of nongray matter.

In conclusion, our results indicate that nonnormality in the error terms can be an issue in VBM. However, provided the data are smoothed with a 12-mm FWHM kernel, nonnormality is sufficiently attenuated to render the tests valid in all situations. An important caveat, however, is that unbalanced designs appear to be less robust to violations of normality: a significant number of false positives arise at a smoothing of 4 and 8 mm when comparing a single subject to a group. This is in contrast to the observation that conventional group comparisons appear to be robust, remaining valid even with no smoothing.

ACKNOWLEDGMENTS

This work was funded by the Wellcome Trust and the Medical Research Council.

REFERENCES

- Ashburner, J., and Friston, K. 1997. Multimodal image coregistration and partitioning—A unified framework. *Neuroimage* **6**: 209–217.
- Ashburner, J., and Friston, K. 1999. Voxel based morphometry—The methods. *Neuroimage* **11**: 205–821.
- Friston, K. J., Ashburner, J., Frith, C. D., Poline, J. B., Heather, J. D., and Frackowiak, R. S. J. 1995. Spatial Registration and Normalisation of Images. *Hum. Brain Map.* **2**: 165–189.
- Gadian, D. G., Aicardi, J., Watkins, K. E., Porter, D. A., Mishkin, M., and Vargha-Khadem, F. 2000. Developmental amnesia associated with early hypoxic-ischaemic injury. *Brain* **123**: 499–507.
- Johnson, R. A., and Wichern, D. W. 1998. *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River, NJ.
- Mugler, J. P., and Brookeman, J. R. 1990. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magn. Reson. Med.* **15**: 152–157.
- Salmond, C. H., Ashburner, J., Vargha-Khadem, F., Gadian, D. G., and Friston, K. J. 2000. Detecting bilateral abnormalities with voxel-based morphometry. *Hum. Brain Map.* **11**: 223–232.
- Woermann, F. G., Free, S. L., Koepp, M. J., Ashburner, J., and Duncan, J. S. 1999a. Voxel-by-voxel comparison of automatically segmented cerebral gray matter—A rater-independent comparison of structural MRI in patients with epilepsy. *Neuroimage* **10**: 373–384.
- Woermann, F. G., Free, S. L., Koepp, M. J., Sisodiya, S. M., and Duncan, J. S. 1999b. Abnormal cerebral structure in juvenile myoclonic epilepsy demonstrated by voxel-based analysis of MRI. *Brain* **122**: 2101–2107.