NeuroImage **14,** 1454–1462 (2001)
doi:10.1006/nimg.2001.0770, available online at http://www.idealibrary.com on IDEAL®

"Voxel-Based Morphometry" Should Not Be Used with Imperfectly Registered Images

Fred L. Bookstein

University of Michigan

Received July 27, 2000; published online May 11, 2001

John Ashburner and Karl Friston (2000) introduced a standardized method of "voxel-based morphometry" (VBM) for comparisons of local concentrations of gray matter between two groups of subjects. Segmented images of gray matter from grossly normalized high-resolution images are smoothed and their group differences analyzed by the now-conventional voxelwise Worsley approach to Gaussian random fields of differences. This comment concerns an unfortunate interaction between the algorithm's spatial normalization and voxelwise comparison steps, whereby several obvious quantitative confounds are injected at the core of the inference engine the authors put forward. Specifically, the statistics of the resulting voxelwise comparisons are uninformative about group differences wherever the spatial normalization algorithm has failed to register on any robustly appearing image gradient. The method of Ashburner and Friston is defensible only far from all image gradients.

INTRODUCTION

In a recent issue of this journal, John Ashburner and Karl Friston (2000) argued for a standardized method of "voxel-based morphometry" (VBM) for comparisons of local concentrations of gray matter between two groups of subjects. (This paper will be cited below as "AF.") In one version or another, the method has been exemplified in the peer-reviewed literature since at least the widely cited study of the medial thalamus in schizophrenia by Andreasen *et al.* (1994). In general, VBM methods combine spatial normalization with tissue classification and the analysis of the ensuing fields of gray level representing variously MR image intensity or estimated concentration of neural gray matter. Although the Andreasen group has not persisted in the use of VBM, the rate of appearance of empirical studies from other centers developing this method has begun to accelerate, including additional exemplars in the most widely browsed media (e.g., Paus *et al.*, 1999). The present comment argues that this diffusion is premature, owing to an unfortunate confusion at the foundation of the method.

The feature I am criticizing appears not to be part of the existing literature critical of VBM, which seems mainly concerned with statistical procedures for excursions of the resulting parametric fields (e.g., Worsley *et al.*, 1999; Bullmore *et al.*, 1999). Indeed one central concern of AF was a collection of implementation-specific performance issues such as segmentation under nonuniformity of intensity and problems with an earlier spatial extent statistic. The issue I am pointing out lies deeper, buried in the assumptions underlying the arithmetic by which those fields are produced from voxels originally arising at a great variety of locations. It is from the averaged images, not their statistical manipulation, that the more serious fallacies of the VBM method arise. Not only in the specific VBM implementation put forward in AF, but also in every other context in which it has been attempted, its two steps (spatial normalization and voxel-based analysis of gray scale) interact computationally in a manner that blocks all valid statistical inference wherever the spatial normalization fails to attend to registration locally. The paper I am criticizing couches VBM within a "continuum" of methods, with "tensor-based morphometry" (TBM) at the far end. I begin my argument by embracing that context but demonstrating that VBM actually does not.

ON THE "CONTINUUM" OF REGISTRATION METHODS

In the article on which I am commenting, the description of the spatial normalization step is not allocated much space, but instead the reader is referred to two earlier publications, Ashburner *et al.* (1997) and Ashburner and Friston (1999), which introduce least-squares methods for affine transformations and then linear combinations of nonlinear basis functions. It is sufficient for my argument here to go forward that these methods are least-squares in gray-level differences between the candidate image and a "template image" that "should be the average of a large number of

1053-8119/01 $35.00

MR images that have been registered," and that the basis functions used for the registration be at large scale and not be tuned to the detailed geometry of the template image. There seems to be nothing further that AF has to say about this normalization step. In particular, a paragraph on page 808 enumerating "a number of assumptions [that] need to hold in order for VBM to be valid" reminds the reader that "confounding effects must be eliminated or modeled as far as possible," but does not list systematic registration bias as such a potential confound—in fact, this discussion does not refer to the interaction of registration with voxelwise analysis in any way. In my view, this missing assumption is considerably more important for the "validity" of VBM than those that AF listed, particularly inasmuch as it cannot be examined in the VBM computing context itself, but requires a more sophisticated environment in which statistics of different registration rules are carried out in a consistent Euclidean framework.

There is a discussion late in the paper on "the effect of spatial normalization," but its concern seems limited to the effect on actual quantification of the amount of gray matter present, not on the voxelwise group differences that are the ultimate goal of analysis. AF avers that in the limit of a perfect registration, "all the information would be in the deformation fields and would be tested using TBM." That statement is incorrect—although indeed the information would thereby be in the deformation fields, it would not be appropriate to test for shape differences there using TBM methods, for reasons I have published elsewhere (Bookstein, 1999). But regarding the interaction of the normalization step and the voxelwise comparisons of gray, the present article says only, "It is envisaged that . . . a continuum will arise with simple VBM (with low-resolution spatial normalization) at one end of the methodology spectrum and statistical tests based on Jacobian determinants at the other (with high-resolution spatial normalization)."

The "continuum" metaphor here may have arisen from a thought experiment such as that in Fig. 1. The vertical axis here stands for the density of gray matter, the proportion of gray matter within a voxel, or the probability of gray matter along a transect through some point of a medical image. The multiple diagonal lines indicate the value of this density across the forms of a data set; that these lines come in two clusters hints at the presence of a group difference in this distribution. This figure captures the commonplace awareness that there are indeed two channels involved in any medical image analysis, the "vertical" (intensity of gray) and the "horizontal" (identification of matching voxels at which statistics of gray are to be compared), along with the intuition that even when the comparisons of gray exploit the same data that were already used for registration there ought to be some way of

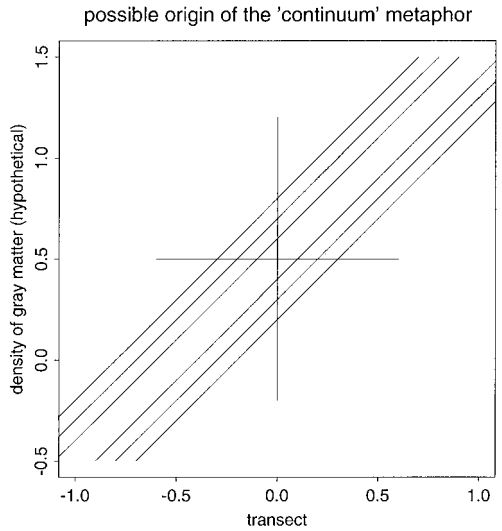

FIG. 1. The "continuum" metaphor of Ashburner and Friston (2000) may have presumed a trade-off of registration against image contents like this one. If all images had a perfectly linear edge-profile, differing only in position, then analysis of registration variability (along the horizontal transect) and analysis of voxelwise registered contents (along the vertical transect) would yield the same group difference signal.

combining the two approaches so that each helps circumvent the confounds built into the other.

Ashburner and Friston may have reasoned that in this setup, the ordinary *t* test for group difference is the same whether it is taken vertically (i.e., a voxelwise averaging of image contents) at the central point here, whatever registration happened to apply, or instead is taken horizontally, by asking what shift is required to register that central point at, say, 50% gray. In this sense, it looks like it does not matter whether one registers and then considers the registration function (their TBM would reduce to testing the derivative along the horizontal here) or instead tests the same group difference along the vertical (the VBM version of AF). Alas, in any context of actual anatomical imagery this tempting graphical metaphor is seriously misleading. To show this, we must build a mathematical setting that combines the two procedures, registration statistics and grayscale statistics, that AF places at the opposite ends of that "continuum."

THE SHIFT FUNCTIONAL

A useful way to begin exploring this mathematical context is to inquire about the nature of a grayscale basis capable of detecting the simplest nonlinear image registration, translation of an edge within a fixed frame, in a manner that imposes equivalence between registration and voxelwise approaches after the fond hope of Fig. 1. To ease this exposition the example deals with the case of a one-dimensional image, gray values along a line, but the formalism is similar in

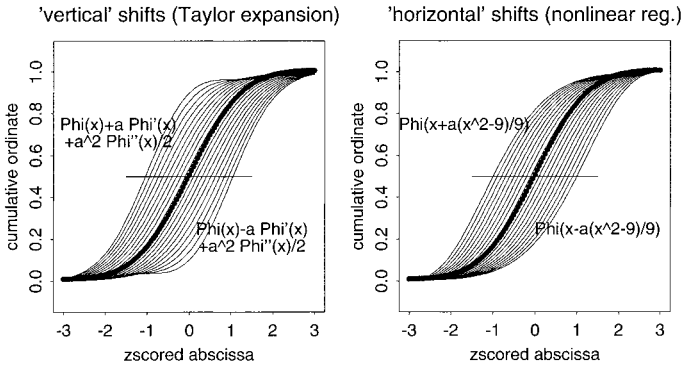

FIG. 2. Shifts of edges can be achieved either by least-squares superposition in gray scale (left), using basis functions carefully tuned to the image template and its derivatives, or instead by reregistration (right) at an appropriately small scale. In both panels the central, heavy curve is the standard cumulative Gaussian distribution.

higher dimensions. The scheme in Fig. 2 allows the mathematician to convert between "horizontal" and "vertical" findings in Fig. 1 with the necessary authority. Without a formalism like this, there is no way to talk about the effect of different registrations on the resulting voxel-based analyses.

At the center of either panel in Fig. 2 is the function $\Phi(x) = 1/\sqrt{2\pi}\int_{-\infty}^{x} e^{-x^2/2}dx$, the standard Gaussian ogive (cumulative probability distribution). We can imagine it to indicate a registerable structure somewhere in the middle of an image, that is, a true (discrete, local) feature in the intensity profile or density of gray matter along some one-dimensional cut. Suppose there is a family of images that all arise from this one by reregistration of this shape within the image boundaries (here set arbitrarily at ± 3): the set of images $y_a(x) = \Phi(x - a)$, where a is a parameter for the shift of this structure. In the research context intended by AF, the parameter a is different for different subjects; indeed, it may vary systematically by patient group.

We can imagine two ways to proceed with the analysis of such a data set. In one approach, we leave the images unregistered except for some "global normalization," and examine the variation in the domain of grayscale functions $y(x)$. That is, we pursue the least-squares "prediction" of values $y_a(x)$ by coefficients in a multiplying functions of x derived from Φ. The predictive representation is a familiar tactic, the expansion of $y_a(x)$ in powers of a: the Taylor series

$$y_a(x) = y_0(x - a) = y_0(x) - ay_0'(x) + \frac{a^2}{2}y_0''(x) - \ldots,$$

where $y_0(x) = \Phi(x)$, the actual edge shape for this example, and the primes indicate differentiation with respect to a at $a = 0$, where $y_a(x - a)$ becomes a function of x only. Specifically, we have

$$y_a(x) = \Phi(x) - \frac{a}{\sqrt{2\pi}}e^{-x^2/2} + \frac{a^2}{2\sqrt{2\pi}}(-xe^{-x^2/2}) - \ldots,$$

the usual series of Gaussian derivative kernels scaled by $(-)^k a^k/k!$. At the left in Fig. 2 is shown the sum of terms through a^2 in this series for values of a between -1 and 1; the approximation to the shifted edge appears satisfactory enough throughout this range. Although this demonstration has varied the parameter a *a priori*, in practice it would be derived empirically, by actually fitting the Taylor series as shown, subject by subject, in order to retrieve the individual values of a encoding this residual misregistration.

In the other approach, the parameter a is extracted explicitly, as by defining a "landmark point" where this edge has its point of inflection that is then used to standardize the image geometry by image warping (which is to say, by voxel relabeling). At the right in Fig. 2 is one inverse of this normalization, the transformation $x \to x - a(x^2 - 9)/9$ that fixes the endpoints ± 3 but shifts the inflection of that standard Gaussian ogive from abscissa 0 to abscissas a from -1 to 1. The actual "large-scale registration" that measures this shift might involve the inverse function $x \to -9/2a\,(-1 + \sqrt{1 - (4a/9)(x - a)})$.

This representation is the only stratagem known to me that permits the interchange of "horizontal" and "vertical" statistical tactics, the metaphor of Fig. 1, in the sense that the same signal is detected in either formalism. In morphometric summaries of samples of curves like these for which the parameter a varies, the "horizontal" statistic that tests for group mean differences in the location of that inflection point will be identical, if correctly implemented, with the "vertical" statistic that uses least squares to fit the corresponding curve all along its length as a linear combination $\Phi(x) - c_1\Phi'(x) + c_2\Phi''(x) - \ldots$ with the appropriate constraint $(-)^k k! c_k^{1/k} = -c_1$, the shift we seek. The derivatives of the template supply the conversion between the two units (centimeters, gray levels) in which that shift might have been observed.

Furthermore, the same net signal can be gotten by combining the two approaches after each has been assessed in its own domain, vertical or horizontal, after any extent of misregistration. That is to say, suppose we have a set of true images $y_{a^i}(x)$ as in Fig. 2, where a^i is a patient-specific shift, and that we have registered them by a procedure that used the information in this edge to some extent, but only imperfectly. Specifically, suppose that their horizontal shifts are estimated in some manner that resulted in reassigning that inflectional "landmark" to the point a_l^i, which is correlated with a^i but somewhat attenuated—say, by half—and then vertical shifts are fitted to these imperfectly registered images as $y(x) \sim \Phi(x) - a_v^i\Phi'(x) + \ldots$ by the usual constrained least squares. (The subscript l is for

"landmark," and *v* is for "voxel.") In the resulting parameter space (a_l^i, a_v^i), which straddles the ends of the AF "continuum," we will have $a_l^i + a_v^i = a^i$, the correct patient-specific value.

This is the proper representation of the "continuum" about which AF was speculating. Imperfect registration has weakened either of the two signals a_l^i and a_v^i when computed separately but has not altered their sum, which is what ought to have been computed in the first place. Any claimed "continuum" running from VBM to TBM (that is, from gray-level variation to deformation variation) must be managed so as to produce the same signal regardless of how a comparison is divided up between registration and gray-level processing. But this is not possible for the VBM method as introduced in AF, as the registration variation is not preserved in any units commensurate with the grayscale analysis, and so the tradeoff between better registration and better gray-scale analysis simply cannot be managed. In the VBM analysis of the "partially registered image," whenever $a_l^i \neq 0$, the value of $a_l^i + a_v^i = a^i$ is inaccessible.

If the phrase "voxel-based morphometry" be reserved for the Ashburner-Friston method, then the continuum method I am recommending, even though it includes the AF term a_v^i as one component, can't be called VBM any longer. Perhaps it could be named "registration-commensurate voxel morphometry," RCVM. There is an example (involving the splenium of the midline corpus callosum) in Bookstein (1999), although neither the name nor the acronym appears there. The version of VBM introduced by Ashburner and Friston would acquire an alternate characterization as "RCVM without the registration signal." I would welcome pointers to other anticipations of RCVM in any responses to this Comment.

Arriving at AF's VBM approach by removing the a_l^i signal from RCVM in this way can be rephrased from the standpoint of spatial normalization as "regressing out" a set of basis functions. Any spatial normalization is a regression that stabilizes some edge information (for instance, the extreme extents of the brain volume) quite enthusiastically, but that responds to others with much greater attenuation. The stronger the tuning of an edge to the normalization basis, the weaker the image variation that remains: but VBM incorporates no formalism for the "strength of a regression" voxelwise, the term at left in Fig. 2. There exist good methods for expressing group differences in the information used for edge-based registration (the methodology applying at the right in the same figure). If this channel of information were restored to the VBM output flow, we would be back at an approximation of a^i, the "continuum" that AF appropriately acknowledges. But that is not what the published algorithm does.

The spatial normalization underlying the present VBM algorithm, we are told on page 807, "merely corrects for global brain shape differences." But the correction is not based on any model for what is "global," merely on an arbitrary stopping criterion for a space of "smooth spatial basis functions." The claim that this algorithm has " discounted global shape differences" is wholly metaphorical. There is no model for the quantification of the "discount," and hence no way to stabilize VBM output against reasonable disagreements about the detail with which to apply specific suites of basis functions, the differences among different sets of basis functions (trigonometric, polynomial, radial), and even the differences among the estimations of the uniform (global) registration underlying these and all other possibilities. The appropriate basis set is rather, by analogy with Fig. 2, the set consisting of the average template and its changes under large-scale deformation. In Fig. 2, where only one landmark was considered variable, the additional basis for the expression of deformation via grayscale regressions consisted of the first and second derivatives of the template. By analogy, in more general contexts it might comprise an orthogonalization of the first- and second-order changes in the mean image under realistic models of image deformation as it is actually encountered. These will continue to be expressible either in the "horizontal" (deformation) metric or in the "vertical" (grayscale) metric, just as laid out for the simplest case in Fig. 2. Note that the regression is "multiple multiple" in form, regressing all voxels simultaneously on all the basis functions jointly, with constraints on the coefficients as already noted for the simple shift of the Φ waveform in the example.

Without a protocol for monitoring the tradeoff of a_l against a_v sketched in Fig. 2, there is no way for the user to accommodate differences in "findings" that owe purely to differences in the minutiae of registration (e.g., the number of basis elements for spatial normalization and their functional form). The actual effect of registration upon an edge is a highly nonstationary function of the location and orientation of the edge. In one pose, it may be almost completely stabilized by the nonlinear basis chosen; in another pose, a few millimeters away or rotated by 45°, it may fall within the null space of the same normalization basis. A nonsignificant finding by VBM might mean absence of signal or instead presence of signal that is tuned to the spatial normalization.

Thus it is at the edges of regions, where the difference between adequately and inadequately aligned voxels bears the greatest import for the grayscale statistics that ensue, that VBM is most vulnerable to the cryptic effects of partial registration. In the absence of any explicit representation of that spatial normalization, this degree of ambiguity should be judged intolerable in any applied scientific context. In the alternative normalization model sketched in Fig. 2, for which the basis set consists of the mean image and an appro-

priate number of its derivatives, the problem is obviated. It is this sort of basis, not the registration "at large scale" recommended in AF and other tutelary publications on this method, that supports the idea of a "continuum" between VBM and the better deformation-based statistical methods.

In practice, unless the basis for normalization is tuned to the typical image (Φ and its derivatives, in the example of Fig. 2), the consumer of VBM output has no way to discern what part of a particular normalized image contrast is robust against moderate changes of registration rule, nor which features, contrariwise, have already been attenuated by the "large-scale registrations" (which, of course, entail small-scale implications) applied hitherto. If Fig. 2 is understood as the small-scale variation remaining after "global registration" of whatever nature, then group differences in these features will be visualized by VBM as a relief map of the edges at which the registration has failed in whole or in part. Such registration errors may well rise above the Worsley-style threshold for significance and thus be reported as differences of gray "at" the voxels underneath. But in fact they would instead be reporting only differences between groups in the meaning of the word "at" according to which the voxel-based statistics were accrued. Meanwhile, other differences of equal or greater geometric amplitude might have been absorbed in the normalization and thereby would have disappeared.

Because global registrations exploit local features like these to different extents according to their position and direction within the image; furthermore, the visualization of peaks like these need not obviously match the researcher's pre-existing mental map of what the underlying boundary looks like as a whole. That is, systematic spatially partial misregistrations will be treated as gray-level signal by the VBM method. Yet such misregistrations are typical consequences of disease-specific deformations of normal anatomy—diseases cannot be expected to align their dysmorphy squarely either with the large-scale spatial basis functions or with their null space: see, again, Bookstein (1999). A VBM ridge of cortical displacement owing to an abnormal corpus callosum, for instance, will not necessarily trace the entire anterior cingulate, but only a few of its subarcs, and thus might pretend to be a legitimate spatially concentrated finding. Simultaneously, the presence of these fallacious signals arising from misregistration must attenuate the sampling distribution of true differences in gray scale at correctly registered voxels and thus must lower the power with which the statistical step can detect those differences that might actually be present. This argument applies particularly to the resampling version of the statistical step. Here intragroup differences in misregistration feed directly into the reference distribution, so that the detection of "true" groupwise voxel differences (that is, those that appear even in registrations according with arbitrarily small neighborhoods of the target voxel) must necessarily suffer a drop in efficiency. There remain the *inter*group differences that depend on group differences in misregistration. Although these will be tested correctly for significance, as misregistered, nevertheless in the ensuing interpretation the registration-dependent part a_I cannot be separated from the voxel-specific a_v remaining.

WHY NOT TO USE VOXELWISE STATISTICS IN ANY EVENT

Notice that the analysis of gray-scale imagery that produces the shift function expansion $\Sigma(-)^k(a^k/k!)\Phi^{(k)}(x)$ is not computed voxel by voxel. In the model of Fig. 2, which embodies many of the most important scientific applications of voxelwise gray matter analysis (shifts of relative cortical compartment volumes, atrophy, etc.), the local image surface is fitted by a constrained superposition of Gaussians and derivatives that explicitly accounts for residual registration error.

But perhaps the voxels to which we are attending are not at the center of a Gaussian edge in the way Fig. 2 is suggesting. Perhaps instead they lie some distance from this edge structure, past the abscissas ± 2 in Fig. 2 at which the function Φ appears to stop curving (that is, for which the shift of edge appears no longer to affect the encoding of intensity for the vertical comparisons to come). Let us ask, indeed, what effect misregistration of edges at these greater distances has upon the voxelwise statistics that correspond to residual misregistration—precisely the circumstance for which AF has declared VBM to be the method of choice. The situation is as in Fig. 3: a misregistered edge lies at some distance, so that we are comparing image contents in the far tails of our underlying edge model. (Note how the vertical scales diminish as we move away from the center of the edge.) Surely there is no remaining effect of the registration error on voxelwise tests of image contents?—the signal from transects like that at the upper left goes away immediately?

Life the metaphor of Fig. 1, this intuition, too, is misleading. The effect of registration error persists to a considerable distance outward along the tail of the Gaussian. We can assess it by modeling the registration error as a standard Gaussian of its own, located at or near the inflection point and thus far from the voxels in question. Specifically, our interest is in the cumulative Gaussian tail $\Phi(x - a)$, where a is in the general range of 3 or so, so that to the untrained eye the image appears to have plateaued, and where x represents registration uncertainty as a Gaussian of its own for the variability of the actual value of a. I will take this edge location uncertainty x as having the usual mean of 0 and variance of 1. Thus we seek the mean and

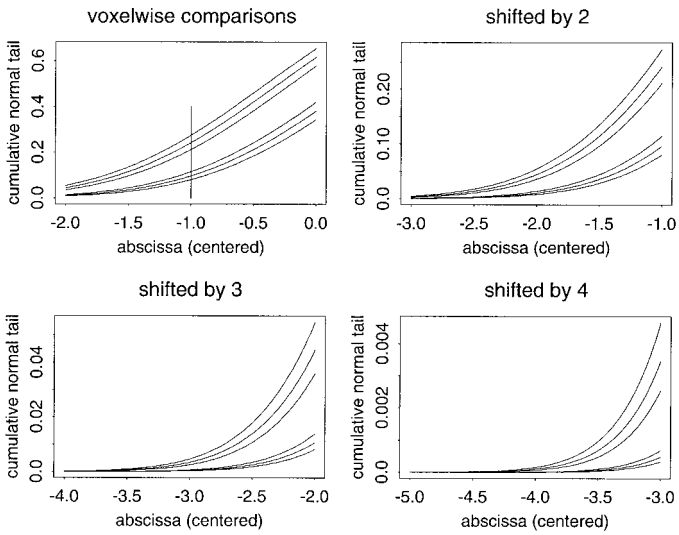

FIG. 3. The voxelwise methods may be intended to apply to image plateaux far from an edge. Here we explore the effect of edge location variation at 1, 2, 3, or 4 standard deviations from the voxel under study for two groups of three specimens with edge centers differing by twice within-group range. The decline in central tendency of the ogive at more distant center, relative to that at less distant center, increases with distance, but also the heterogeneity of these within-group variances. Note the different vertical scales of the four panels.

variance of $\Phi(x - a)$ for $x \sim N(0, 1)$ and a moderately large.

The expected value of $\Phi(x - a)$ can be expressed in closed form. By definition,

$$\mathrm{E}\Phi(x-a) = \frac{1}{\sqrt{2\pi}} \int_{x=-\infty}^{\infty} \Phi(x-a) e^{-x^2/2} dx$$

$$= \frac{1}{2\pi} \int_{x=-\infty}^{\infty} e^{-x^2/2} \int_{y=-\infty}^{x-a} e^{-y^2/2} dx dy.$$

Change to new variables $x' = (y + x)/\sqrt{2}$, $y' = (y - x)/\sqrt{2}$, an orthogonal rotation leaving both the integrand $e^{(-x^2-y^2)/2}$ and the area element $dxdy$ unchanged. The region $y < x - a$ of the xy-plane is the same as the region $y' < -a/\sqrt{2}$ of the $x'y'$-plane. Then

$$\mathrm{E}\Phi(x-a) = \frac{1}{2\pi} \int_{x'=-\infty}^{\infty} \int_{y'=-\infty}^{-a/\sqrt{2}} e^{(-x'^2-y'^2)/2} dx' dy'$$

$$= \frac{1}{\sqrt{2\pi}} \int_{y'=-\infty}^{-a/\sqrt{2}} e^{-y'^2/2} dy' = \Phi(-a/\sqrt{2}).$$

For the expected value of the square of Φ there seems to be no such exact expression, but we can exploit a very useful approximation to Φ known to statisticians as *Mills's Ratio* (Stuart and Ord (1994), 193–195). It expresses the tail-area of the standard Gaussian as a multiple of the ordinate at that point:

$$R(x) \equiv \Phi(-x)/\Phi'(-x) = \frac{1}{x} - \frac{1}{x^3} + \frac{1 \cdot 3}{x^5} - \frac{1 \cdot 3 \cdot 5}{x^7} + \ldots.$$

The series does not converge, but the remainder at any step is less than the last term used, and in any case it varies rather slowly in x by comparison with $e^{-x^2/2}$.

By completing the square, then, we have

$$\mathrm{E}\Phi^2(x-a) = \frac{1}{\sqrt{2\pi}} \int_x (\Phi(x-a))^2 e^{-x^2/2} dx$$

$$= \frac{1}{(2\pi)^{3/2}} \int_x R^2(a-x) e^{-x^2/2} e^{-(x-a)^2} dx$$

$$= \frac{1}{(2\pi)^{3/2}} e^{-a^2/3} \int_x R^2(a-x) e^{-(3/2)(x-(2a/3))^2} dx$$

$$\sim \frac{1}{(2\pi)^{3/2}} R^2(a/3) e^{-a^2/3} \int_x e^{-(3/2)(x-(2a/3))^2} dx$$

$$= \frac{1}{2\pi\sqrt{3}} R^2(a/3) e^{-a^2/3},$$

because the integrand is nonnegligible only for x in the vicinity of $2a/3$. Hence the variance of $\Phi(x - a)$ is approximately

$$\frac{1}{2\pi\sqrt{3}} R^2(a/3) e^{-a^2/3} - \Phi^2(-a/\sqrt{2})$$

$$\sim \frac{1}{2\pi\sqrt{3}} R^2(a/3) e^{-a^2/3} - \frac{1}{2\pi} R^2(a/\sqrt{2}) e^{-a^2/2}.$$

A useful comparative statistic is the coefficient of variation, which is the square root of the variance divided by the mean. From the first term in Mills's Ratio, the leading term of the coefficient of variation turns out to be $(3/2)^{1/4} e^{a^2/12}$, which increases considerably more slowly than $e^{-a^2/2}$ falls. In short, as a increases—as the center of the edge moves farther and farther from the voxel at which we are looking—the signal-to-noise ratio of an actual edge shift falls inexorably to zero, indeed, but only quite gradually. To fall faster toward zero, the effect of a distant edge shift would have to presume an edge gradient shape that approaches its asymptote faster than e^{-cx^2}, a gradient "sharper than diffusion;" such gradients seem unlikely

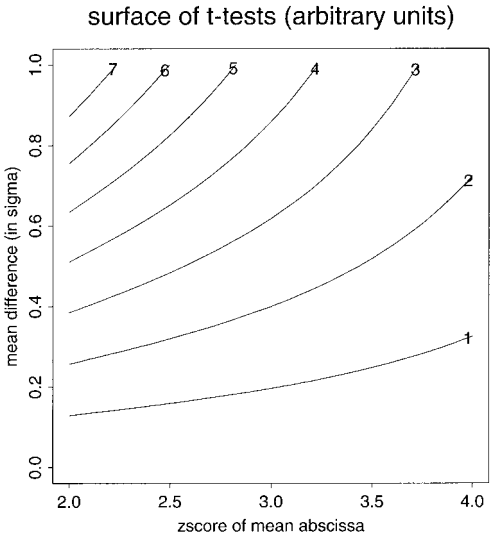

FIG. 4. The effect of an edge shift on voxelwise t fields is discernible quite far from the edge, until it is masked either by true intragroup grayscale variation or by running out of the bits coding voxel contents. The surface plotted here is proportional to the group difference signal from the previous figure for normally distributed edge-locations differing by fixed multiples of the edge width (vertical axis) over a range of distances from the pooled mean edge center (horizontal axis).

to arise from any real physical imaging process or statistical tissue classification (to say nothing of the smoothing steps built into the subsequent statistical processing).

Using the approximation var $\Phi \sim (2\pi\sqrt{3})^{-1}R^2(a/3)e^{-a2/3} - \Phi^2(-a/\sqrt{2})$, Fig. 4 plots the fraction

$$\frac{\mathrm{E}\Phi(x-(a+b)) - \mathrm{E}\Phi(x-(a-b))}{(\mathrm{var}\,\Phi(x-(a+b)) + \mathrm{var}\,\Phi(x-(a-b)))^{1/2}},$$

proportional to the t field for a vertical difference of average cumulative Gaussians like these, for edges a at from 2.0 to 4.0 standard deviations' remove and for edge shifts $2b$ between 0.1 and 1.0 in the same standard deviation units. (The factor for sample size has been suppressed.) It seems that until within-group variance swamps these tail effects, any edge will induce a bias to voxelwise group mean comparisons that is neither flat nor negligible for a considerable distance. Smoothing the image, as recommended by AF, has two effects both of which make the problem worse. By reducing the within-group fluctuations of the image plateaux, they allow the bias from edge geometry to extend to a longer distance; but, also, the smoothing of any registered image smooths the edge as well, increasing the effective standard deviation of the edge-gradient and again widening the region to which this counterintuitive bias pertains. For eight-bit images, the last bit ceases to be informative (that is, the image content is rounded to 0 or 1) at 2.88 standard deviations out; for ten-bit images, at 3.29 standard deviations; and for images that are classification probabilities, the preferred representation, the last bit fails still farther out, at 5.41 standard deviations for a 24-bit mantissa. The smoothing recommended in AF will increase all of these. Notice, also, that the presence of the edge injects a long-range order to voxelwise comparisons that vitiates any claims of veridicality for the random field assumptions that otherwise underlie the familiar Worsley excursion tests.

That registration is "at large scale" and thus can be expected to misalign the details of most edges across subjects does not protect the user from any of these paradoxes. Whether or not a large-scale registration happens to overlay edges across subjects, it will typically preserve their separate orientations. The effect is to broaden the standard deviation of the Gaussian model for x, the true variation of edge centers, in both of the exegeses preceding. As a consequence, more terms are required in the Taylor series driving the interchange of horizontal and vertical analyses, Fig. 2, and also, yet again, the effective window of the asymptotic problem is widened, that is, made worse. In either case, the regression is confounded with anatomical difference to varying extents across the image, invalidating any use of voxelwise or cluster tests in the sequel. Notice, too, that this effect modifies the signal in a_v, the registered voxel value, independently of the signal a_l bearing the information about misregistration. Any fix, in the course of establishing the valid continuum-based RCVM method, would require that the misregistration be explicitly extracted by the Taylor method of Fig. 2, not this vertical computation (see, again, the little example in Bookstein, 1999).

DISCUSSION

Whether "at" an edge or within a surprisingly large multiple of the underlying edge width, voxelwise statistics about group differences are biased by registration failure in systematic ways that the AF implementation of VBM seems to have taken every opportunity to aggrandize. Voxelwise comparisons escape registration problems only when edges are known not to adjoin the voxels in question. For instance, we could test voxelwise with confidence everywhere except right atop edge points whenever registration is "perfect"—the TBM end of the AF "continuum"—but as the authors have clearly declared that VBM is intended for application at the other end, this convenient assumption cannot apply. Yet from the nature of the crucial discussions that were omitted from AF, the developers of VBM seem to have presumed that however imperfectly the images were registered, the registration error does not matter for the study of group differences pursuant to either of the fallacies I have explored above—that all effects of edge geometry, near or far,

have been swamped within subject-to-subject difference in image contents, or vice versa.

It follows, in the absence of any mechanism for examining this confound, that VBM can be applied responsibly only near the center of relatively uniform regions. The VBM protocol described in Ashburner and Friston (2000) ought to be used for empirical inferences about scientific questions only where image gradients are very low on average, and to that end the *t* fields of classification probability or any other statistical summaries it supplies need to be colorcoded for local average absolute gradient magnitude whenever they are displayed. In practice, this would suggest that the VBM output as described in AF be masked so that only those voxels are displayed that lie far from edges in almost every registered subject. Offering this masking field, keyed to the average image gradient underlying a normalized voxel, would be a most helpful service to the SPM community.

For those remaining tests, for voxels far from any information that might be registration-relevant, one gains power to the extent that the data arising within those regions have been smoothed within the maximum possible smoothing window. Thus the "properly masked AF method" I seem to be recommending reduces to *t* tests between gray levels over the interiors of a list of regions, together with a statistical analysis of the spatial normalization rules themselves: their group mean differences, their within-group variances, and their covariances with the normalized signal remaining. The multivariate machinery of these quantities is actually more accessible than that of the normalized images themselves (Dryden and Mardia, 1998; Bookstein, 1999).

That the registration rules underlying VBM remain obscure to the user bears substantial implications for the research communities within which VBM would typically be applied on a routinely uncritical basis. It is not enough that articles declare analyses to have been carried out using such-and-such a published package with the default parameter settings, as if the VBM implementation was something like a mass spectrometer with a published calibration and a parts number. "Findings" as sensitive as VBM's to deeply buried details of between-subject registration have an unacceptably tenuous relation to veridicality claims. The analysis of a single data set can be stable over variously normalized images only if the continuum delimited by the two panels of Fig. 2 is respected—only if the terms VBM annihilates, for mean differences and variance in spatial normalization viewed horizontally, are appended to the visualizations that the user must consider and the statistical computations must accommodate prior to hypothesis-testing. Until then, no VBM analysis should be published unless the authors have specified precisely how they have registered images over intersubject variation, precisely which voxels that VBM has reported to be interesting are in the vicinity of strong gradients that may have affected the spatial normalization computations, and precisely how that variation was, or was not, attenuated in the registration applied. It is not enough that findings appear to "replicate" on a new sample, as a sufficiently similar sample, processed by the same confounds of misregistration, may well yield the same incorrectly delimited or detected finding a second time—indeed, will likely do so if the study is of a disease that entails some local neuroanatomical abnormality, whether familiar or not. Nor is it enough for VBM to appear to confirm an ROI analysis, as they both omit the same information (about deformation)—their flaws are very similar.

In summary, regardless of the technicalities of statistical inference that concerned AF, the method of VBM reviewed there is mathematically vitiated by the unfortunate confound between its spatial normalization step and all subsequent computations. This interaction is not noted in the list of conventional assumptions of which the VBM user is presumed aware. Put forward as a method to be used in the absence of local registration accuracy, in fact VBM is capable of providing reliable, sensible answers only in the presence of state-of-the-art registrations such as those of Thompson *et al.* (2000) or Joshi *et al.* (1995), not the low-parameter superpositions with which AF recommends one begin. Failure to register correctly on all pertinent image gradients confound the resulting voxelwise tests to a great distance from the gradient, in fact, everywhere that the image has not plateaued to stationarity. There may be no voxel anywhere that is far enough from the nearest gradient for any of the voxelwise statistics to be trusted. In the vicinity of any strong gradient, VBM findings are seriously confounded by the imperfections of registration in a manner that cannot be stabilized, within VBM, against improvements in registration. In neither case, whether registration is at large scale or at small, can VBM findings that purport to visualize group differences be considered empirically reliable in the absence of verifications so strenuous as to render the methodology impractical in most of the applications proposed by its developers.

ACKNOWLEDGMENTS

This work was supported in part by USPHS Grant GM-37251 to the author. I thank the anonymous reviewer who pointed out the exact expression for EΦ and two other reviewers who expressed support for the main message here.

REFERENCES

Andreasen, N. C., Arndt, S., Swayze, V., 2nd, Cizadlo, T., Flaum, M., O'Leary, D., Ehrhardt, J. C., and Yuh, W. T. 1994. Thalamic

abnormalities in schizophrenia visualized through magnetic resonance image averaging. *Science* **266:** 294–298.

Ashburner, J., Neelin, P., Collins, D. L., Evans, A. C., and Friston, K. J. 1997. Incorporating prior knowledge into image registration. *NeuroImage* **6:** 344–352.

Ashburner, J., and Friston, K. J. 1999. Nonlinear spatial normalization using basis functions. *Hum. Brain Mapp.* **7:** 254–266.

Ashburner, J., and Friston, K. J. 2000. Voxel-based morphometry—The methods. *NeuroImage* **11:** 805–821.

Bookstein, F. L. 1999. Linear methods for nonlinear maps: Procrustes fits, thin-plate splines, and the biometric analysis of shape variability. In *Brain Warping* (A. W. Toga, Ed.), pp. 157–181. Academic Press, San Diego.

Bullmore, E. T., Suckling, O., Rabe-Hesketh, S., Taylor, E., and Brammer, M. J. 1999. Global, voxel, and cluster tests, by theory and permutation for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imag.* **18:** 32–42.

Dryden, I. L., and Mardia, K. V. 1998. *Statistical Shape Analysis.* Wiley.

Joshi, S., Miller, M. I., Christensen, G. E., Banerjee, A., Coogan, T., and Grenander, U. 1995. Hierarchical brain mapping via a generalized Dirichlet solution for mapping brain manifolds. In *Vision Geometry IV* (R. Melter, A. Wu, F. Bookstein, and W. Green, Eds.), S.P.I.E. Proceedings, Vol. 2573, pp. 278–289.

Paus, T., Zijdenbos, A., Worsley, K. J., Collins, D. L., Blumenthal, J., Giedd, J. N., Rapoport, J. L., and Evans, A. C. 1999. Structural maturation of neural pathways in children and adolescents: An *in vivo* study. *Science* **283:** 1908–1911.

Stuart, A., and Ord, K. 1994. *Kendall's Advanced Theory of Statistics*, 6th ed., Vol. 1: *Distribution Theory.* Halsted Press, New York.

Thompson, P. M., Giedd, J. N., Woods, R. P., MacDonald, D., Evans, A. C., and Toga, A. W. 2000. Growth patterns in the developing human brain detected using continuum-mechanical tensor mapping. *Nature* **404:** 190–193.

Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D., and Evans, A. C. 1999. Detecting changes in nonisotropic images. *Hum. Brain Mapp.* **8:** 98–101.