Non-normality and transformations of random fields, with an application to voxel-based morphometry

Roberto Viviani*, Petra Beschoner*, Katja Ehrhard*[#],

Bernd Schmitz°, Jan Thöne*

Departments of *Psychiatry III and °Radiology,
[#]Transfer Centre for Neuroscience and Learning (ZNL),
University of Ulm, Germany

Corresponding author:
Roberto Viviani
University of Ulm
Department of Psychiatry III
Leimgrubenweg 12
89075 Ulm
Germany
Telephone: +49 731 50021486
Fax: +49 731 50026751
E-mail: roberto.viviani@uni-ulm.de

Non-normality and transformations of random fields, with an application to voxel-based morphometry

Abstract

Parametric tests of linear models for images modelled as random fields are based like ordinary univariate tests on distributional assumptions. It is here shown that the effect of departures from assumptions in random field tests is more pronounced than in the univariate condition. Simulations are presented investigating in detail the influence of smoothing, unbalancedness and leverages on empirical thresholds. In certain conditions, significance tests may become invalid. As a case study, the existence and effect of departures from normality of grey matter probability maps, commonly used in voxel-based morphometry, is investigated, as well as the effect of different transformation strategies involving estimating the degree of transformation from the data by maximum likelihood. The best results are achieved with a voxel-by-voxel transformation, suggesting heterogeneity of distributional form across the volume for this kind of data.

Non-normality and transformations of random fields, with an application to voxel-based morphometry

Introduction

The effect of non-normality on the performance of significance tests based on Student's *t* or Snedecor's *F* statistics were investigated early in the statistical literature (Pearson 1929, Pearson 1931, Pearson and Please 1975, Box 1953). The general conclusion from these and later studies is that, unless deviation from normality is severe, the impact on classic univariate parametric tests is limited, in many cases only reducing the power of the test (see summary in Miller 1986). In the first part of this study, simulations on the effect of non-normality on random fields were carried out to show their impact on the thresholds that determine achieved significance levels in procedures such as statistical parametric mapping (SPM, Friston et al. 1995). A specific study is justified by the fact that it is difficult to extrapolate the results from these early studies to the random field setting. On the one hand, random field theory tests (Worsley et al. 1992, 1996) are based on an approximation of the distribution of the extrema of the random field, and for this reason they may by more sensitive to deviations from normality than statistics of central tendency such as *t* in the univariate setting (Westfall and Young 1993, pp. 56-60). On the other hand, the common practice of enforcing a fixed spatial correlation structure by smoothing the data may reduce any original deviation from non-normality (Salmond et al. 2002).

The deviation from normality examined here only concern the distribution of the data viewed voxel-by-voxel (marginal normality). This means that we are not concerned with the full distributional specification of the random field model which would also require, for example, that the joint spatial distribution be normal, or that the extent of spatial correlation

be uniform across the volume (Hayasaka et al. 2004). In this respect, we also note that while the distributional requirements of random field theory tests do not coincide with voxel-by-voxel distributional assumptions, these latter may be important for other types of tests, or for multiple testing situations in general (Westfall and Young 1993). If the correct voxel-by-voxel significance threshold varies across the volume, applying a uniform threshold will have the effect of applying unequal requirements in weighting the evidence in individual voxels (Beran 1988).

This study is primarily motivated by the application of statistical tests on structural images, which often consist of probability values or coefficients. The issue is how much non-normality can be tolerated before the achieved Type I error rates are altered making the test invalid or overly conservative. This issue affects unbalanced comparisons, and especially single-subject studies (Colliot et al. 2005, Kassubek et al. 2002, Woermann et al. 1999a, b), since here the consequences of non-normality are most marked. In the first part of this study we will show with the help of Monte Carlo simulations that even after spatial smoothing departures from normality can indeed impact random field thresholds. It will be shown that this occurs by non-normality levels that would not raise concern in a univariate setting. However, the conservativeness of *t* random field tests generally ensures that the tests remain valid. Hence, one may be more afraid of loosing of power than committing a Type I error when using random field theory tests, but we will show that exceptions are likely to occur at extreme degrees of unbalancedness, such as in tests of individual images, and regressors with high leverages.

In the second part of this study, we will examine the behaviour of empirical thresholds computed from 114 gray matter probability maps estimated from MPRAGE images, a type of structural T1-weighted images used in voxel-based morphometry (Ashburner and Friston 2000). Voxel-based morphometry is a prominent methodology in the analysis of structural data. We will show here that smoothing is only partially effective in reducing the impact of

non-normality, and that in fact non-normality affects the exact thresholds following a spatial pattern. To overcome this problem, we investigate the application of data-driven maximum likelihood estimates of the required amount of data transformation.

This study extends existing work on voxel-based morphometry data (Salmond et al. 2002) in several respects. Firstly, we provide simulations exploring in detail the effect of smoothing, unbalancedness, and regressor leverages on the asymmetry of empirical thresholds in tests on random fields. Secondly, in the case study with gray matter probability maps we use a much larger sample of volumes from an adult population, while Salmond et al. (2002) focused on structural images of a small sample of children. The larger sample allows us to compute maps of the non-uniform distribution of areas where overthreshold voxels occur as a result of Type I errors. Thirdly, we investigate the impact of different transformation strategies on the empirical thresholds and the spatial distribution of overthreshold voxels. In particular, we study the shortcomings of uniform transformations at smoothing kernels of 4 mm or less, and compare their performance to that of transformations estimated from the data.

Materials and methods

Simulations

All code implementing the algorithms and the simulations presented here was developed on MATLAB 6.1 R12 (The Mathworks, Natick, MA) installed on a Pentium PC running Windows 2000 (Microsoft, Redmond, WA). Transformed artificial random fields of size $32 \times 32 \times 32$ voxels (as in the simulations of Nichols and Hayasaka 2003) were created by convolving a Gaussian kernel of full width half-maximum (FWHM) 4 voxels with a standard normal deviate x after applying the inverse Box-Cox transformation $f(x;\alpha,\beta)$:

$$f(x;\alpha,\beta) = \begin{cases} \exp(x+\alpha) & \text{when } \beta = 0 \\ \left[(x+\alpha)(\beta+1)\right]^{\frac{1}{\beta}} & \text{when } \beta \neq 0 \end{cases} \quad (1)$$

(see Box and Cox 1964). To investigate the effect of kurtosis, the following power transformation was used:

$$f(x;\vartheta) = |x|^{\vartheta} \cdot \mathrm{sgn}(x), \tag{2}$$

where sgn(x) is the sign of x: +1 or −1. In all transformations, data were centered and scaled to unit variance. To avoid edge effects, images were padded at the sides with random variates for 3 times the FWHM size of the kernel prior to smoothing. Random numbers were obtained from MATLAB's generator 5. In computing the test statistic in the unbalanced trials, the smaller group was subtracted from the larger group. Trials from artificial data and resampling trials were repeated 2000 times (the confidence interval of a frequency of 0.05 estimated from a series of 2000 Bernoulli trials has width less than 0.01). Variance, skewness, and kurtosis were computed using the algorithms described in Press et al. 1988, pp. 613-614. To obtain the theoretical random field theory thresholds, code from the SPM2 package was used (Wellcome Department of Cognitive Neurology, London; online at http://www.fil.ion.ucl.ac.uk). Significance of rates of overthreshold extrema was computed by comparing the rates of skewed or kurtotic random fields with the rates of normal fields. This was achieved by establishing the empirical threshold at which normal random fields gave overthreshold voxels at a rate of 0.05. This threshold was subsequently applied to non-normal fields, and the overthreshold voxel rates were computed. Pearson's χ^2 test was used (Collett 2003), rejecting the null hypothesis of equal rates in normal and non-normal fields at a significance level of 0.001.

Gray matter maps

All magnetic resonance imaging (MRI) data were obtained with a 3-Tesla Magnetom Allegra (Siemens, Erlangen, Germany) MRI system equipped with a head volume coil. All 114 subjects (62 females, average age 25.6, min and max ages 18 and 55) were scanned over a predefined period at the Department of Psychiatry of the University of Ulm after obtaining informed consent. The study protocol was approved by the local ethical committee. Images were individually screened to exclude pathology. Images were obtained using a T1-weighted MPRAGE sequence. Image size was $256 \times 256 \times 19$ voxels, voxel size $1 \times 1 \times 1$ mm. The

images were acquired with TR 3300, TE 96, a flip angle of 90°, a bandwidth of 1220 Hz/Pixel, and a field of view of 240×240. We made use of the SPM5 package for realignment, stereotactic normalization, segmentation (after intensity modulation), and smoothing of volumes (Frackowiak et al. 1997). To obtain the segmentation maps, all volumes were realigned and registered to a T1 template (Montreal Neurological Institute) in one step using SPM5 (Wellcome Department of Cognitive Neurology, London; online at http://www.fil.ion.ucl.ac.uk), and resampled to obtain voxels of size $2 \times 2 \times 2$ mm. This procedure yields voxel-by-voxel maps of the probability of belonging to cerebrospinal fluid, gray or white matter compartments. Only the gray matter probability maps (most commonly examined in voxel-based morphometry) were considered in this study. The gray matter maps were smoothed using an isotropic Gaussian kernel of FWHM = 4, 8, and 12 mm (2, 4, and 8 voxels), and masked by excluding average values below 0.05 (the same mask as in Salmond et al. 2002).

Transformations of gray matter maps

Conformity to normality assumptions of probability values, which are constrained to lie between 0 and 1, is often achieved by means of the logit transformation of the raw signal value y (Atkinson 1985, Ashburner and Friston 2000):

$$\operatorname{logit}(y) = \log\left[y/(1-y)\right]. \tag{3}$$

To obtain intermediate transformation grades, the logit transformation may be replaced by a family of transformations having the logit as a particular case. For smoothing kernels of 4 mm, we applied the folded power transformation family (Atkinson 1985, pp. 138-139):

$$f_\lambda(y) = \begin{cases} y^\lambda - (1-y)^\lambda & 0 < \lambda \le 1, \\ \operatorname{logit}(y) & \lambda = 0 \end{cases} \tag{4}$$

Indexed by the parameter λ, this transformation yields the logit as λ approaches zero, and untransformed data (up to the addition of a constant) for $\lambda = 1$. An adequate value of λ can be chosen by maximizing the profile log-likelihood $L(\lambda)$, (Box and Cox 1964):

$$L(\lambda) = -(N/2)\log\hat{\sigma}_\lambda^2 + \log J_\lambda\,, \qquad J_\lambda = \prod_{i=1}^{N}\left|\frac{\partial f_\lambda(y_i)}{\partial y_i}\right|. \quad (5)$$

In this expression, $\hat{\sigma}_\lambda^2$ is the variance estimate of the errors obtained from the residuals of the maximum likelihood fit of the transformed data, and N the number of observations. J is the Jacobian of the transformed data, allowing for the change of scale of the response due to the transformation (details are available in standard textbooks such as Atkinson 1985 or Cook and Weisberg 1980).

The transformation may be estimated from the pooled data across the whole volume, or voxel by voxel. In this latter case it is imperative to use the normalized transformed data:

$$z_\lambda(y) = f_\lambda(y)\big/J_\lambda^{1/N}\,. \quad (6)$$

Normalization is required since the transformed values are not in the same scale when the transformation parameters differ in each voxel. This is a problem for enforcing smoothness by applying a Gaussian kernel, and for empirical smoothness estimation or any statistic computed on the residuals, since the residuals are in the scale of the data. Furthermore, transforming the data to the same scale makes it much easier to inspect the transformed data.

Results

Effect of skewness and kurtosis on distribution of extrema

In this first simulation, the inverse Box-Cox and power transformations were applied prior to smoothing artificial random fields, obtaining data with increasing skewness and kurtosis (see Materials and Methods for details). The plots of the empirical and the theoretical thresholds are displayed in Figure 1.

INSERT FIGURE 1 ABOUT HERE

One can see that the effect of skewness on the Monte Carlo thresholds can be substantial, and increases when the smoothing kernel is smaller (top half of Figure 1). The first trial, marked with ‘N’ on the abscissa, refers to untransformed data, and gives therefore a Monte Carlo estimate of the threshold for normal data at the chosen significance level (p = 0.05). In black dashed lines the theoretical thresholds computed according to random

field theory, which lie, as expected, between the Monte Carlo thresholds of the two extrema, and slightly above the Monte-Carlo estimate for normal data. Their position does not correspond exactly to the Monte Carlo thresholds, since the theoretical thresholds refer to continuous random fields and not to lattice approximations as in this simulation and in digital images (see Nichols and Hayasaka 2003, Worsley 2005). Realistic degrees of skewness prior to smoothing lead to residual skewness after smoothing that would be of little concern in a univariate distribution, but have an impact on random field thresholds.

Positive skewness, as induced by the inverse Box-Cox transformation in these simulations, moves the empirical thresholds for maxima upwards and lowers the empirical thresholds for minima. An equal amount of negative skewness would have the same effect, except that empirical thresholds for maxima would be lowered and those for minima would be raised (Miller 1986).

One may also note that the departure of the absolute empirical thresholds for maxima and minima from the normal case is not exactly symmetric. Rather, maxima thresholds appear to be slightly more affected by the inverse Box-Cox transformation than minima thresholds. This is due to the fact that the inverse Box-Cox transformation also induces a light kurtosis in the data. Co-occurrence of skewness and some degree of kurtosis is in fact a common occurrence in real data. When only kurtosis is present (bottom half of Figure 1), its effect is an increase of the Type I error rate on both sides of the distribution, and tests in both directions are affected identically, unlike in skewness. Hence, in the top half of Figure 1, where skewness and some kurtosis co-occur, the effect of kurtosis is to compound the effect of skewness for tests in one direction, and compensate it in the other.

The effect of kurtosis alone on the Monte Carlo thresholds can be substantial, but smoothing is here more effective in making the data more normal. Kurtosis is more quickly dampened as an effect of the central limit theorem than skewness (Miller 1986, p. 6).

Effect of skewness on error rates in t *tests in ANOVA models*

Two factors affect the impact of skewness in *t* tests in ANOVA models. First, the group averages are more normally distributed than the original data as an effect of the central limit theorem. Second, if the data are balanced and the skewness is the same in both groups, its effects will cancel out, and the test will not be affected (irrespective of the amount of smoothing). This is because, in the subtraction of the means of a *t* test, the long tail of the mean that is subtracted is flipped over to the other side, compensating the short tail of the first mean (for an analytical formulation and further details, see Miller 1986, p. 42). By contrast, if the data are unbalanced, the skewness of the smaller group will dominate the *t* statistic.

In the following simulation, distributions of extrema of 2000 *t* random fields were obtained by carrying out *t* tests on 30 random fields with smoothing kernel of 2 voxels divided into two groups of varying sizes, from perfectly balanced (15/15) to extremely unbalanced (1/29) (top row of Figure 2).

INSERT FIGURE 2 ABOUT HERE

The simulation shows that, as an effect of averaging, the impact of skewness is less than in the previous simulation, except in the single-image test. For all group sizes except the single image case, random field theory tests remain valid, also as a consequence of the increased conservativeness of *t* relative to Gaussian fields (Nichols and Hayasaka 2003). However, differences in the Monte Carlo thresholds between minima and maxima are noticeable.

Effect of skewness and kurtosis on error rates in t *tests in regression models*

Skewness is a common concern in linear regression contexts (Atkinson 1985). To characterize the possible susceptibility of a covariate coefficient estimate to be influenced by skewness, we turned to the *leverage*, a measure of remoteness in regression space of the values of the predictors x_i for a single observation i relative to the other N observations modeled by the design matrix $\mathbf{X}$:

$$h_i = x_i' \left(\mathbf{X}'\mathbf{X} \right)^{-1} x_i , \qquad i = 1, 2, \ldots N, \quad (7)$$

The leverage h_i of the observation i is the ith diagonal of the hat matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'$ and ranges from $1/N$ to 1. The larger the leverage of an observation, the more determined the fitted value $\hat{y}_i$ from the observed y_i alone. When $h_i = 1$, $\hat{y}_i$ and y_i are identical and the residual is zero (Atkinson 1985).

In the classic texts on linear model diagnostics (such as Cook and Weisberg 1980), the importance of high leverage observations derives from their possible large influence on the fit that occurs when they correspond to outliers in the data. In the present context, the 'outliers' are samples from the tails of a non-normal distribution. Since the same model with high leverage observations is repeatedly fitted thousands of times, it is quite likely that a high leverage observation co-occurs with a sample from the tail of the distribution somewhere in the volume. The test statistic will have a large value here, impacting on the distribution of the extrema. This explains the effects on the rates of overthreshold voxels shown in Figure 3.

INSERT FIGURE 3 ABOUT HERE

One can see that here the effect of high leverages on the overthreshold rates is qualitatively similar to that of unbalancedness in ANOVA (Figure 2). Hence, high leverages indicate that statistical inference may be affected by skewness in regression.

Departure from assumptions in gray maps from segmented MPRAGE images

In this section the distribution of extrema in mean comparisons of gray matter probability maps from MPRAGE images of 114 normal subjects will be examined at degrees of kernel smoothing from 4 to 12 mm FWHM (see the Materials and Methods section for details on the acquisition of these images). Being probability values, these maps are expected to depart from normality to the extent to which they approach 0 or 1. Parametric maps of the initial mean, variance, skewness, and kurtosis of these images are shown in Figure 4. It can bee seen that variance is especially high at the border zone between gray and white matter, while skewness affects the distribution at the edge of the brain and in the white matter. The pattern of kurtosis

is more complex. There are isolated small patches of kurtosis values larger than 10, located especially at the border of the gray matter mask. Most voxels have more moderate kurtosis values which show a tendency to form a spatial pattern following the convolutions of the cortical mantle.

INSERT FIGURE 4 ABOUT HERE

To explore the impact of non-normality on significance thresholds, 30 volumes were sampled at random without replacement at each of 2000 trials, and allocated randomly to two groups. Differences between group means were computed voxel by voxel and subjected to a *t* test as in the Monte Carlo studies of the previous sections. Unbalanced (10/20, 5/25) as well as balanced (15/15) reference comparisons were computed. In Figure 5, the thresholds for tests of increased gray matter (corresponding to maxima thresholds) are higher than those for decreases (corresponding to minima thresholds), a finding compatible with a preponderance of data with a long right tail coming from low probability values in the map. In the 8 mm smoothing group and the unbalanced comparison 10/20, thresholds are about a half *t* value apart; in groups of sizes 5/25, the difference is almost 1.5 *t* values. It is also worth noting that with decreasing smoothing kernels, the overthreshold rates not only diverge, but also shift upwards (Figure 5, right). This indicates that there are sources of non-normality other than skewness whose influence becomes perceptible at narrow smoothing kernels.

The dashed black lines of Figure 5 also show random fields theory thresholds computed from an estimate of the smoothness of the centered gray matter probability maps (the residuals from a fit to the intercept). This choice is intended to provide a rough equivalent of the SPM implementation, where smoothness is estimated from the residuals to the fit. [We gave much thought to how estimating the smoothness in the present context, since the residuals are not distributed like the errors. It turned out that it was computationally prohibitive to compute the smoothness estimates on the residuals of the 2000 trials, exactly mimicking SPM. Note that the gist of our argument concerns the asymmetry of the empirical

thresholds and the spatial distribution of the overthreshold voxels, and these aren't affected by the way smoothness is estimated for the parametric test]. Because of the conservativeness of the random fields test, their thresholds are in most cases higher than the empirical thresholds. This makes the overall test valid at the nominal significance level. In the 12 mm smoothing group, the achieved alpha rates of random field theory tests of increased gray matter are always within the nominal alpha rates (confirming the findings of Salmond et al. 2002), but the empirical thresholds are about one *t* value higher than in tests for reduced gray matter.

INSERT FIGURE 5 ABOUT HERE

In Figure 6 maps of the occurrence of overthreshold voxels are shown (uncorrected thresholds at $p = 0.001$ were used, instead of the corrected thresholds on the previous simulations, as it may well be unwise to estimate the distribution of extrema that far in its tail from only 114 MPRAGE images). Because of the sparseness of overthreshold clusters, even at this relatively liberal threshold, good spatial rendering was obtained using 8000 trials (instead of the 2000 of previous simulations). In this and all the following maps, the colour range is adapted to the range of the data to achieve a good contrast and help to recognize anatomical features, when they exist. To appreciate the differences in uniformity of the occurrence of overthreshold voxels, attention should be paid to the range of the colour scales, displayed on the right of the maps: the smaller the range, the more uniform is the occurrence of overthreshold voxels.

INSERT FIGURE 6 ABOUT HERE

The maps of Figure 6 show that there are much more overthreshold than underthreshold voxels, and that they follow a spatial pattern instead of being randomly distributed across the masked image. In the map with the smaller smoothing kernel of 4 mm, it is possible to recognize the preponderant influence of the voxel-by-voxel skewness of the gray matter probability maps (Figure 4). Thus, even when nominal alpha rates are respected, the spatial distribution of overthreshold voxels is biased. The maps also show that, while guaranteeing

correct tests, smoothing kernel sizes of 12 mm obliterate much anatomical detail. At these levels of smoothing, validity of statistical tests is achieved at the expense of the fine-grained structure of the cortical mantle. Furthermore, the asymmetry of the occurrence of overthreshold and underthreshold voxels is still present.

Transformations of segmented MPRAGE images

To reduce the impact of non-normality in voxel-based morphometry, it has been proposed to transform the data with the logit transformation (equation 3, Ashburner and Friston 2000). When applied to these data, however, the logit transformation brought about a reversal of the overthreshold rates pattern visible at small smoothing kernels (Figure 7). At intermediate and large smoothing kernels, by contrast, the logit transformation appears adequate.

INSERT FIGURE 7 ABOUT HERE

One simple strategy to limit the impact of non-normality is to raise the threshold of the gray matter values over which the parametric map is computed, hoping to strike the right trade-off between resolution requirements and spatial coverage. This would exclude the areas where the probability of the voxel belonging to gray matter is low, and hence the areas where the distribution most departs from normality. Here, an alternative strategy based on intermediate transformation grades is investigated to gain insight on the characteristics of the signal, and explore an automatic procedure based on statistical methodology. To obtain intermediate transformation grades, the appropriate parameter λ of the folded power transformation (equation 4) was estimated by maximum likelihood (equation 5). Under the assumption that the same transformation is appropriate for the whole volume, λ can be estimated by pooling the data from all voxels, obtaining $\lambda = 0.4$. When applied to the data at the narrow smoothing kernel of 4 mm, the transformation redressed the unbalance of overthreshold voxel frequency observed with the logit transformation (Figure 8, left). Note, however, that the tendency of the thresholds to rise with unbalanced comparisons is still present.

INSERT FIGURE 8 ABOUT HERE

Unfortunately, when the maps of overthreshold voxels in the 5/25 comparison are inspected (Figure 8, right), one can see that overthreshold signals still occur in a specific spatial pattern involving white matter, indicating that no adequate normalization of the data can be achieved using a single transformation across the whole volume. The consequence of this is that any test procedure that uses a simultaneous threshold over the whole volume will be more liberal with some voxels, and more conservative with others.

To correct for inhomogeneous skewness, separate λ parameters were estimated voxel by voxel, normalizing the transformed data (equation 6) to obtain meaningful smoothness estimates. The resulting composite transformation brought about small unfavourable changes to the empirical thresholds (Figure 9, left), but improved the spatial distribution of overthreshold voxels considerably (note the narrower range of overthreshold voxels in the colour scale of Figure 9, right). However, a spatial pattern constituted by a rim parallel to the surface of the brain, already present but less evident in the overthreshold maps of Figures 6 and 8, is visible in the tests for overthreshold maxima in the comparison 5/25. Interestingly, this rim is also visible in the kurtosis maps of the untransformed images (Figure 4).

INSERT FIGURE 9 ABOUT HERE

More insight into the workings of the voxel-by-voxel transformation and its role in reversing the impact of skewness can be gained by inspecting Figure 10, where the parameter λ is shown in the images on the left. On the right, the skewness values of the gray matter probability maps have been redrawn, showing that the transformation is most marked (values of lambda around zero) when the skewness is large. This is especially the case around the ventricles.

INSERT FIGURE 10 ABOUT HERE

Discussion

The Monte Carlo studies presented here demonstrate that the effect of sampling extrema dominates over that of spatial smoothing (Figure 1). Even at relatively large smoothing kernels the effect of skewness on effective Type I error rates is noticeable. In *t* tests, the effect of the central limit theorem is that of further limiting the impact of skewness. In these circumstances, the effect on effective Type I error rates is complex: any amount of unbalancedness in the data leads to noticeable effects in the distribution of extrema, especially at smaller smoothing kernels. However, *t* random field thresholds are more conservative at narrower smoothing kernels, compensating or compounding the effect of skewness depending on the direction of the test. Empirical thresholds are also sensitive to kurtosis, but in this case both spatial smoothing and the effect of averaging are more effective in enforcing normality onto the data (Figure 2).

Skewness in the data can affect empirical thresholds in the presence of non-uniform leverages in the predictors (Figure 3). This is of some practical importance: leverage plots developed for linear regression diagnostics are applicable to statistical parametric mapping without modifications, since they do not involve residuals or fitted values, which are different in each voxel. As in standard linear regression, however, it should be remembered that high leverages do not automatically translate into high-influence observations; for this to happen, they need to co-occur with a sample located in the tail of its distribution.

For the purpose of inducing normality, averaging is at least as important as smoothing. In functional neuroimaging it is common to acquire hundreds of volumes per experiment; statistics based on so many observations are not likely to be severely affected due to the effect of the central limit theorem on their distribution, unless produced by a fit on a predictor giving large leverage to individual observations. The situation may be different, however, in studies of structural images, especially those of single subjects (Figure 5). In our sample of gray matter probability maps, skewness failed to be completely removed by spatial smoothing,

even at large smoothing kernels, resulting in a spatial pattern of overthreshold extrema (Figure 6).

Our results show that it might not be an entirely trivial task to correct for the non-normal distribution of gray matter probability maps, since the best results might be obtained with voxel-by-voxel computations of the required transformation (Figures 8 and 9). This suggests that the deviations from normality do not follow a uniform pattern across the volume. Even with this technique, however, small asymmetries in the overthreshold rates and a residual spatial pattern in the occurrence of overthreshold voxels (possibly caused by kurtosis) were present.

It is important to mention that our primary objective here was not to provide a specific technique to address the problem in the context of a given model, but to characterize the distribution of this type of data more generally. In particular, we opted here for using all images at once in computing the transformation parameter λ, instead of using the residuals of the model in each trial. This would not only have increased the cost of our computations prohibitively, but is also justified by our aim to characterize the data in general terms. When a specific model is given, the transformation parameter will be computed from the residuals of the model. For studies with small samples, the applicability of voxel-by-voxel transformations remains an open issue, so that the importance of balanced comparisons must be stressed. In any case, if a uniform transformation is used, then a folded power transformation with $\lambda = 0.4$ appears to be preferable to the logit (Figures 7 and 8).

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry - The methods. NeuroImage 11, 805-821.

Atkinson, A.C., 1985. Plots, Transformations, and Regression. An Introduction to Graphical Methods of Diagnostic Regression Analysis. Oxford University Press, Oxford.

Beran, R., 1988. Balanced simultaneous confidence sets. J. Amer. Statist. Assoc. 83, 679-686.

Box, G.E.P., 1953. Non-normality and tests on variances. Biometrika 40, 318-335.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. B 17, 1-26.

Collett, D., 2003. Modelling Binary Data. Chapman & Hall, London.

Colliot, O., Bernasconi, N., Khalili, N., Antel, S.B., Naessens, V., Bernasconi, A., 2005. Individual voxel-based analysis of gray matter in focal cortical dysplasia. NeuroImage 29, 162-171.

Cook, R.D., Weisberg, S., 1980. Residuals and Influence in Regression. Chapman and Hall, London.

Frackowiak, R.S.J., Friston, K.J., Frith, C.D., Dolan, R.J., Mazziotta, J.C., 1997. Human Brain Function. Academic Press, London.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: A general linear approach. Human Br. Mapping 2, 189-210.

Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E., 2004. Nonstationary cluster-size inference with random field and permutation methods. NeuroImage 22, 676-687.

Kassubek, J., Hupperz, H.J., Spreer, J., Schulze-Bonhage, A., 2002. Detection and localization of focal cortical dysplasia by voxel-based 3-D MRI analysis. Epilepsia 43, 596-602.

Miller, R.G., 1986. Beyond ANOVA, Basics of Applied Statistics. John Wiley & Sons, New York.

Nichols, T.E., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: A comparative review. Stat. Meth. Med. Res. 12, 419-446.

Pearson, E.S., 1929. The distribution of frequency constants in small samples from non-normal symmetrical and dkew populations. Biometrika 21, 259-286.

Pearson, E.S., 1931. The analysis of variance in cases of non-normal variation. Biometrika 23, 114-133.

Pearson, E.S., Please, N.W., 1975. Relation between the shape of population distribution and the robustness of four simple test statistics. Biometrika 62, 223-241.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1988. Numerical Recipes in C. Cambridge University Press, Cambridge, 2nd edition.

Salmond, C.H., Ashburner, J., Vargha-Khadem, F., Connelly, A., Gadian, D.G., Friston, K.J., 2002. Distributional assumptions in voxel-based morphometry. NeuroImage 17, 1027-1030.

Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing. Examples and Methods for *p*-Value Adjustement. Wiley, New York.

Woermann, F.G., Free, S.L., Koepp, M.J., Ashburner, J., Duncan, J.S., 1999a. Voxel-by-voxel comparison of automatically segmented cerebral grey 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., Duncan, J.S., 1999b. Abnormal cerebral structure in juvenile myoclonic epilepsy demonstrated with voxel-based analysis of MRI. Brain 122, 1202-1208.

Worsley, K.J., Marrett, S., Neelin, P., Evans, A.C., 1992. A three-dimensional statistical analysis for CBF activations studies in human brain. J. Cerebr. Bl. Flow Metab. 12, 900-918.

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Human Br. Mapping 4, 58-73.

Worsley, K.J., 2005. An improved theoretical *P*-value for SPMs based on discrete local maxima. NeuroImage 28, 1056-1062.

Figure legends

Figure 1. Monte-Carlo estimate of the thresholds at significance level 0.05 for data departing from normality, obtained from 2000 trials. Columns from left to right: decreasing degrees of smoothing. Top row: random fields at different degrees of skewness (Box-Cox transformation with $\alpha = 10$). On the abscissa, decreasing β values indicate increasing skewness; the letter 'N' denotes data without transformation (normally distributed). Bottom row: random fields at different degrees of kurtosis (power transformation, exponent on the abscissa). At the top of each plot, marginal skewness or kurtosis measured from the artificial data before and after smoothing. The plots of asterisks and circles refer to absolute empirical thresholds for maxima and minima, respectively, while the random field theory thresholds are shown by black dashed horizontal lines. Overthreshold rates that significantly differ from rates of untransformed normal fields are marked by '×' ($p < 0.001$).

Figure 2. Monte-Carlo estimate of the thresholds at significance level 0.05 from 2000 *t* maps with 28 degrees of freedom for increasingly unbalanced comparisons (columns from left to right). Top row: inverse Box-Cox transformation (α was set to 10 as in the previous simulation). Bottom row: power transformation (same parameters as in the previous simulation). At the top of each plot, marginal skewness or kurtosis measured from the artificial data before smoothing. The skewness of the smoothed *t* maps or the kurtosis of the residuals (*t* maps are kurtotic even if the data are normal) are also shown**.** Other conventions as in Figure 1.

Figure 3. Monte-Carlo estimate of the thresholds at significance level 0.05 from 2000 *t* maps (28 degrees of freedom) from skewed data at increasing leverages (columns from left to right) at smoothing kernel FWHM = 2. As in the previous simulations, skewness was obtained by applying the inverse Box-Cox transformation ($\alpha = 10$). Increasing leverages were obtained by creating a regressor from a normal random variate, and adding a fixed amount (1, 3, 5, 7 standard deviations) to the first predictor. Top row: empirical and parametric thresholds (same conventions as in Figure 1). Bottom row: histogram of the leverages obtained by pooling the leverages from all 2000 trials. In this simulation, leverages above 0.4 (for the bulk of the leverages lying below 0.2) lead to substantial asymmetries in the thresholds for maxima and minima.

Figure 4. From left to right, mean, variance, skewness, and kurtosis (computed so that zero means no kurtosis) of gray matter probability maps before smoothing, masked at average gray matter probability values larger than 0.05. Note that the kurtosis map has been drawn with a colour scale with most of its dynamic range at the lower values; this was done to increase the contrast at kurtosis values between 0 and 10, which are those that involve the large majority of voxels and follow recognizable spatial patterns.

Figure 5. Empirical thresholds at significance level 0.05 from 2000 *t* tests of resampled gray matter probability maps at different degrees of smoothing (columns from left to right) and unbalancedness (abscissa). Values for maxima and minima are drawn as asterisks and circles, respectively. The approximate theoretical random fields theory thresholds are drawn as dashed horizontal lines. Overthreshold rates that significantly differ from rates of balanced comparisons are marked by '×' ($p < 0.001$).

Figure 6. Transversal slices of colour-coded number of voxels over the upper and lower thresholds for $p = 0.001$, uncorrected, in the comparison 5/25, at smoothing kernels of 4 (left) and 12 mm (right) computed from 8000 *t* tests between randomly resampled volumes.

Figure 7. Empirical thresholds at significance level 0.05 from 2000 *t* tests of resampled logit-transformed gray matter probability maps at different degrees of smoothing (columns from left to right) and unbalancedness (abscissa). Conventions as in Figure 5.

Figure 8. Left: Empirical thresholds at significance level 0.05 from *t* tests of resampled gray matter probability maps, after applying a smoothing kernel of 4 mm and the folded power transformation with $\lambda = 0.4$ at different degrees of unbalancedness (abscissa). Conventions as in Figure 5. Right: Transversal slices of colour-coded number of voxels over the upper and lower thresholds for $p = 0.001$, uncorrected, computed from 8000 trials of *t* tests between resampled volumes in the comparison 5/25.

Figure 9. Left: Empirical thresholds at significance level 0.05 from *t* tests of resampled gray matter probability maps, after applying a smoothing kernel of 4 mm and the folded power transformation estimated separately at each voxel at different degrees of unbalancedness (abscissa). Conventions as in Figure 5. Right: Transversal slices of colour-coded number of voxels over the upper and lower thresholds for $p = 0.001$, uncorrected, computed from 8000 trials of *t* tests between resampled volumes in the comparison 5/25.

Figure 10. Left: Lambda of the folded power transformation, estimated voxel-by-voxel. Right: Skewness of gray matter probability values, redrawn here for convenience from Figure 4.

Figures

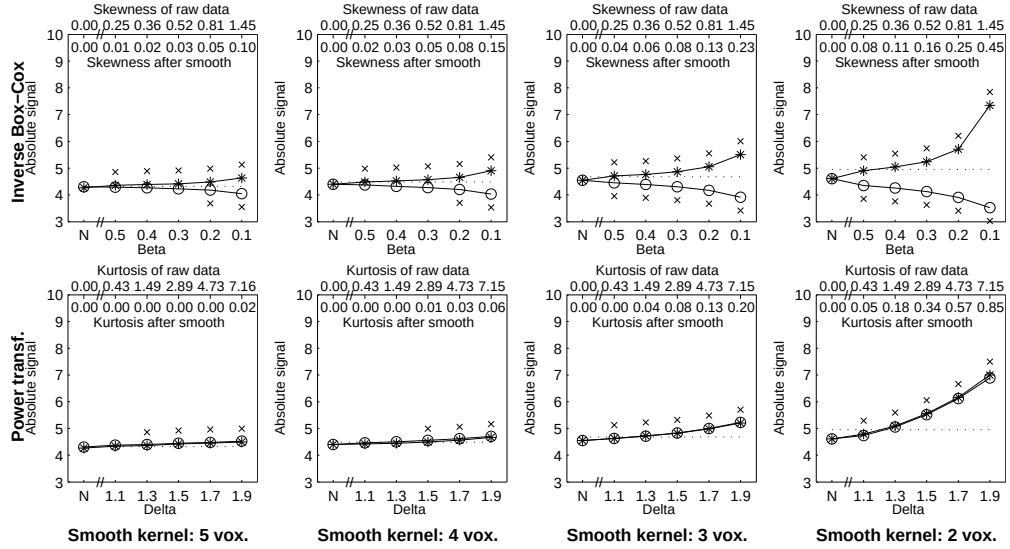


Figure 1

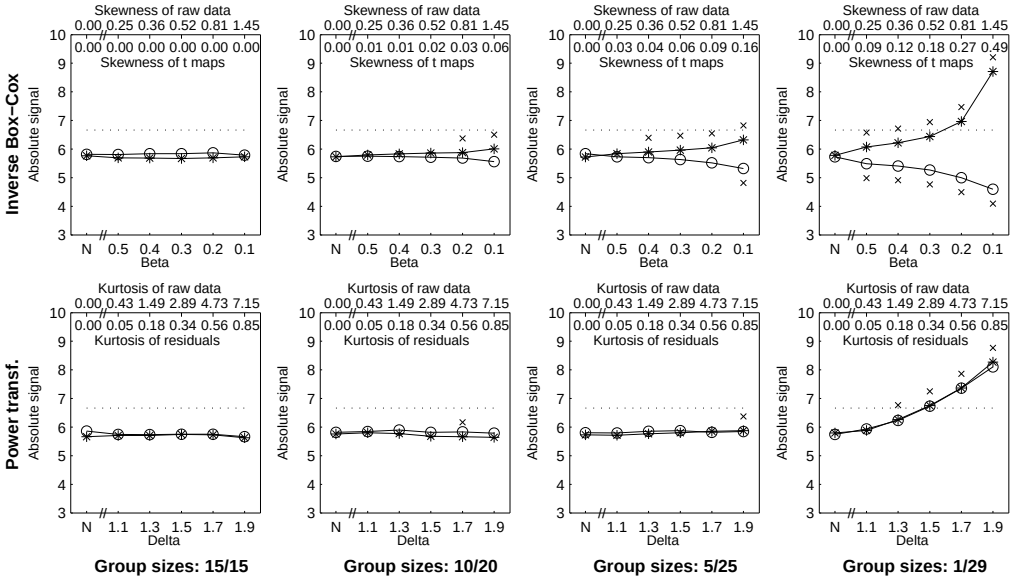


Figure 2

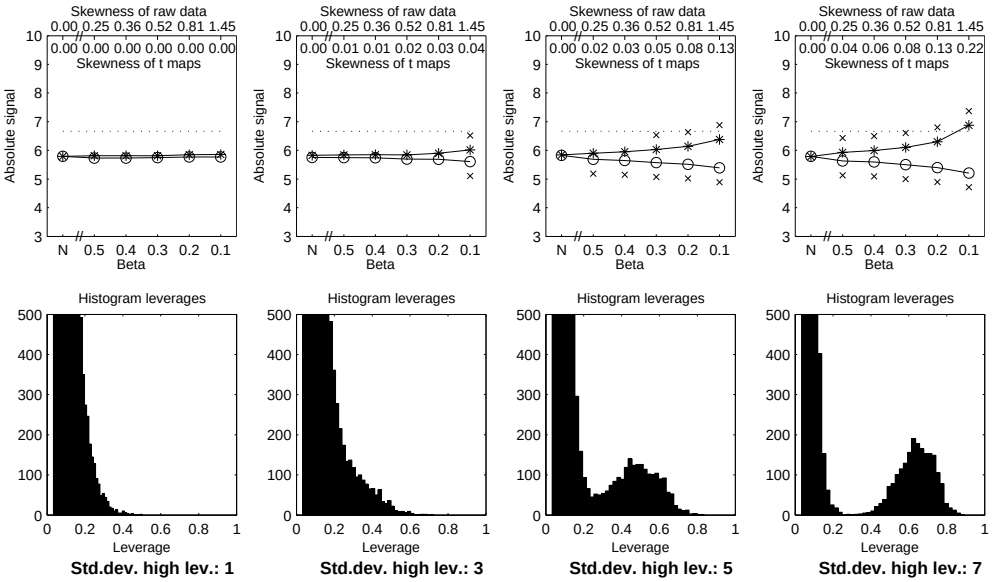


Figure 3

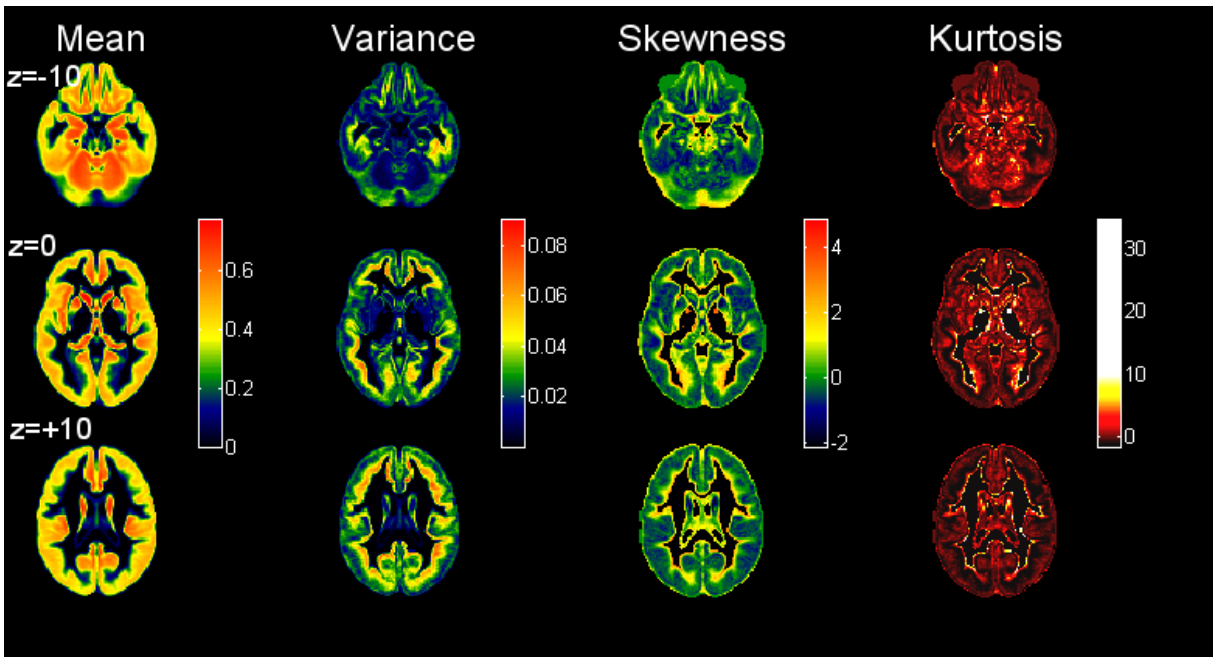


Figure 4

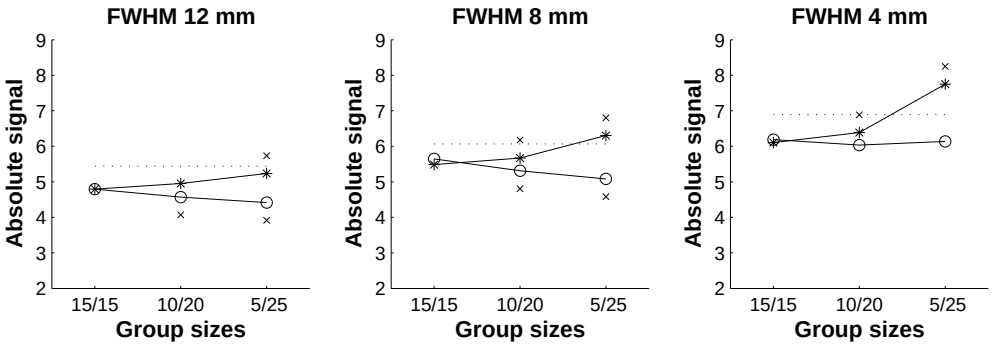


Figure 5

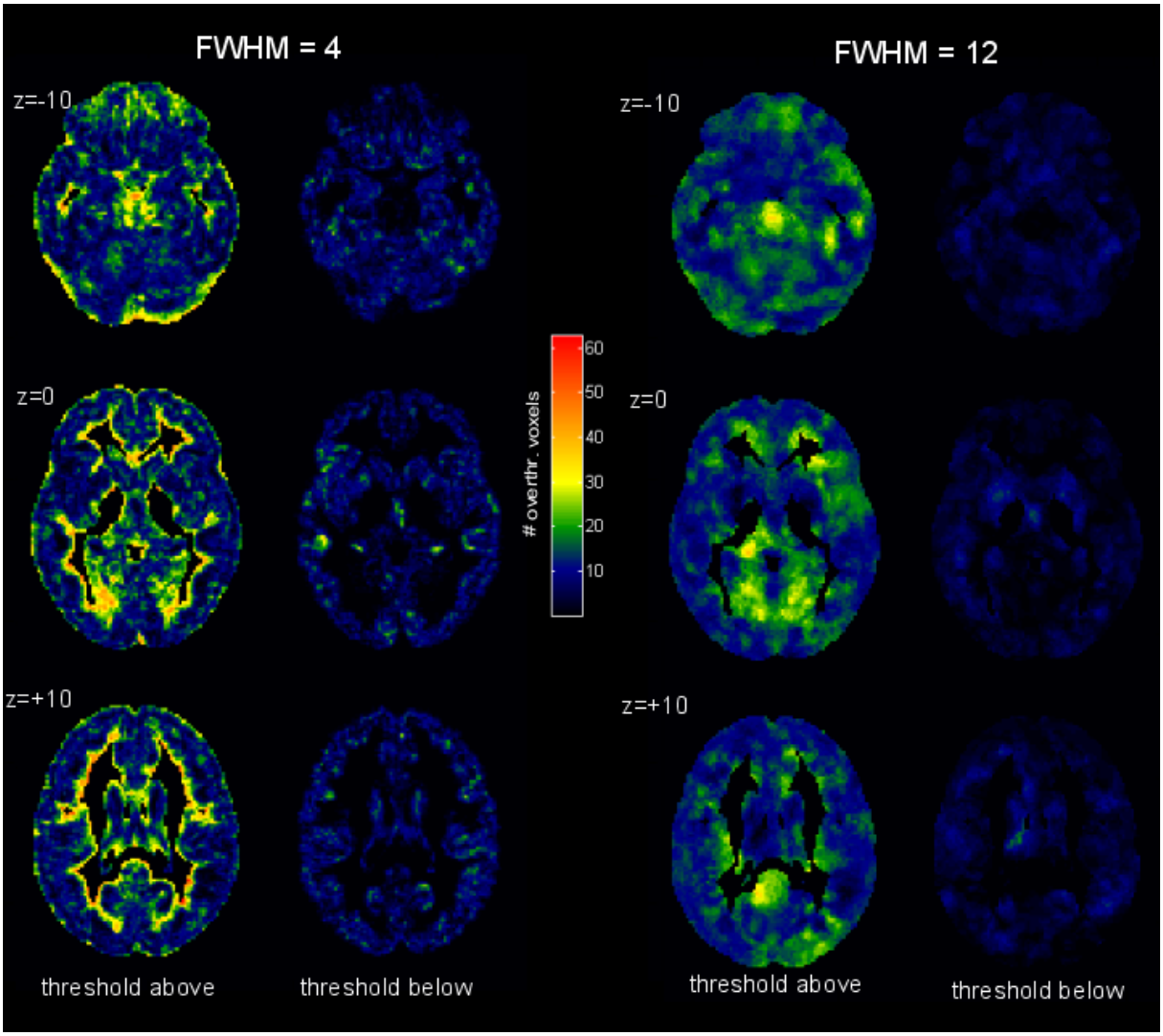


Figure 6

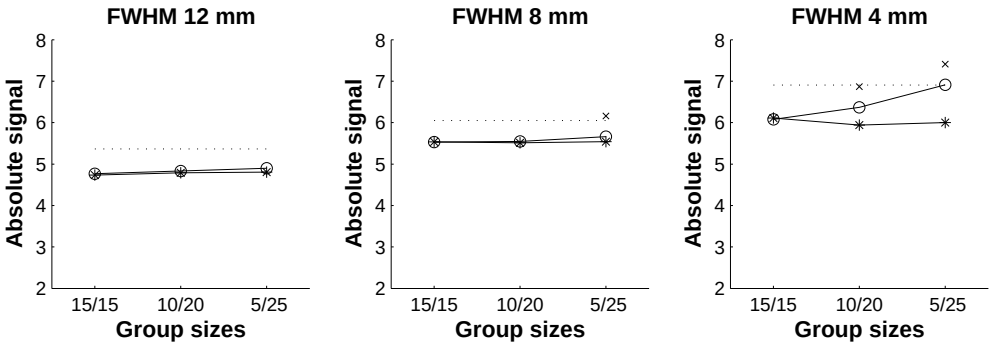


Figure 7

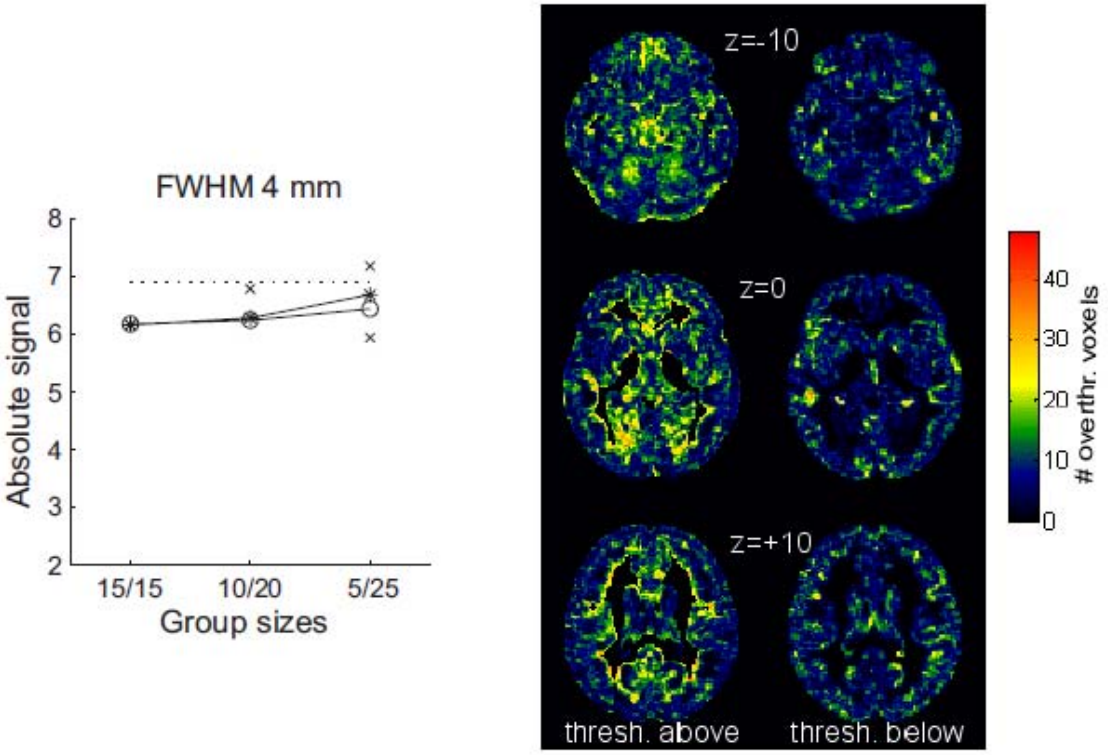
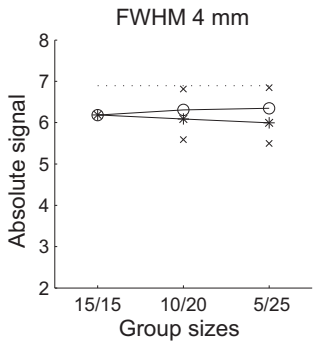
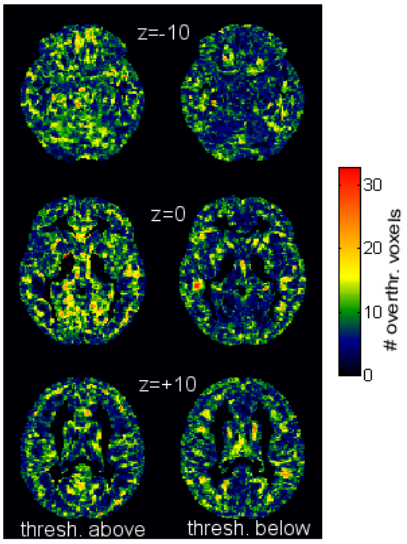


Figure 8


FWHM 4 mm
Absolute signal
8
7
6
5
4
3
2
15/15
10/20
5/25
Group sizes


z=-10
z=0
z=+10
thresh. above
thresh. below
30
20
10
0
overthr. voxels

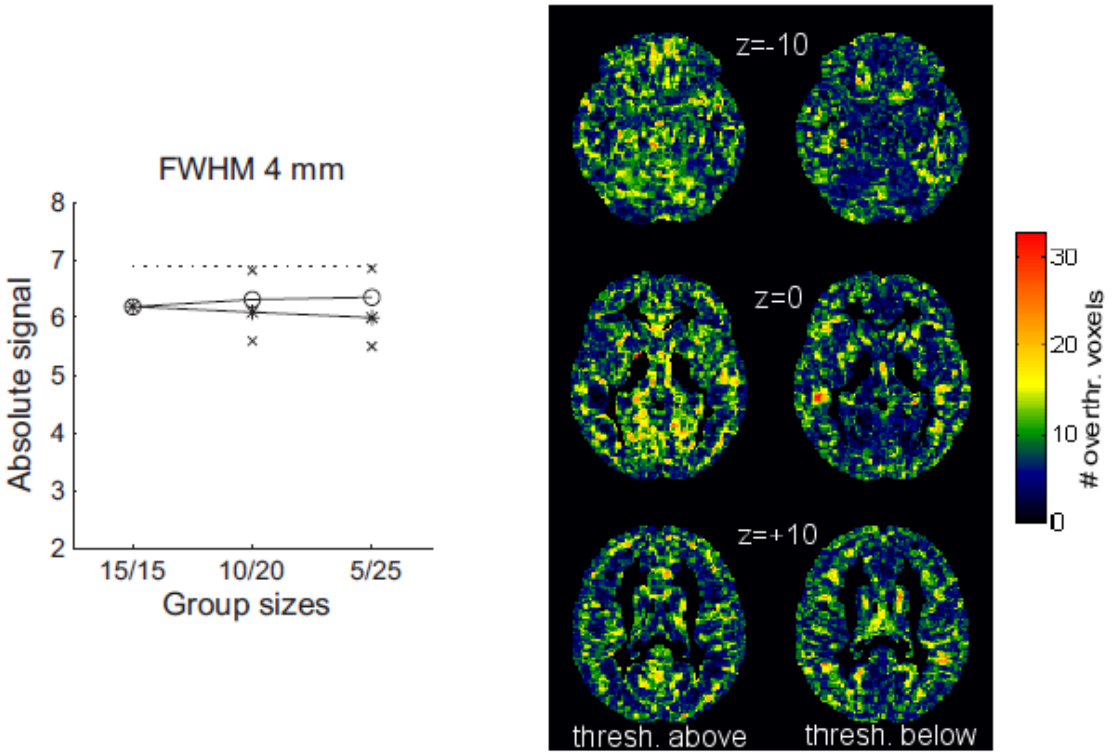


Figure 9

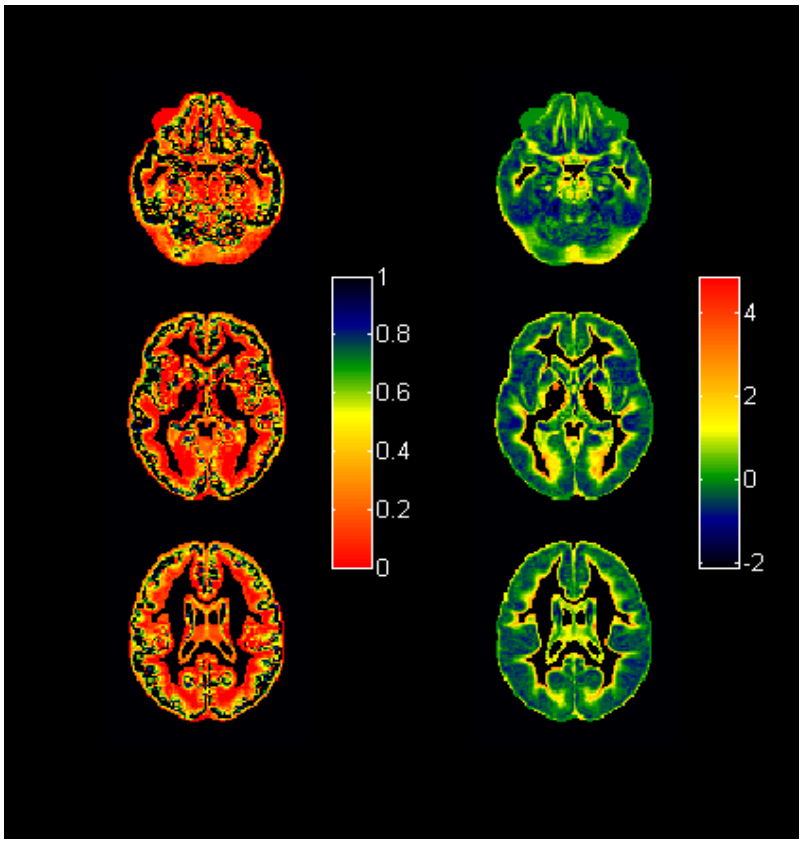


Figure 10