Systematic misregistration and the statistical analysis of surface data

A. H. Gee
G. M. Treece

Systematic misregistration and the statistical analysis of surface data

A.H. Gee[a,*], G.M. Treece[a]

[a] *University of Cambridge Department of Engineering, Trumpington Street, Cambridge CB2 1PZ, UK*

Abstract

Spatial normalization is a key element of statistical parametric mapping and related techniques for analysing cohort statistics on voxel arrays and surfaces. The normalization process involves aligning each individual specimen to a template using some sort of registration algorithm. Any misregistration will result in data being mapped onto the template at the wrong location. At best, this will introduce spatial imprecision into the subsequent statistical analysis. At worst, when the misregistration varies systematically with a covariate of interest, it may lead to false statistical inference. Since misregistration generally depends on the specimen's shape, we investigate here the effect of allowing for shape as a confound in the statistical analysis, with shape represented by the dominant modes of variation observed in the cohort. In a series of experiments on synthetic surface data, we demonstrate how allowing for shape can reveal true effects that were previously masked by systematic misregistration, and also guard against misinterpreting systematic misregistration as a true effect. We introduce some heuristics for disentangling misregistration effects from true effects, and demonstrate the approach's practical utility in a case study of the cortical bone distribution in 268 human femurs.

Keywords: statistical parametric mapping, spatial normalization, misregistration

1. Introduction

A common procedure in medical image analysis is the calculation of cohort statistics expressed on voxel arrays or surfaces. Perhaps the most well-known exemplar is statistical parametric mapping (SPM) (Friston et al., 1994), which has become a standard tool for neuroimaging. In its voxel-based instantiation, SPM starts with an ensemble of voxel arrays containing, for example, fMRI activations measured across a number of subjects, where maybe the subjects are classified into two groups, the interest group and a set of controls. Since each subject's brain will be of a different size and shape, the fMRI data is then *spatially normalized*, a process which involves registering each voxel array to some standardized template. Now that the data is expressed on a common morphology, a general linear model (GLM) can be fitted, to explain the data at each voxel in terms of covariates of interest (e.g. group) and also confounding covariates (e.g. age, sex). Finally, F- or t-statistics are calculated at each voxel, to test whether the data depends significantly on the covariates, with random field theory furnishing the corresponding p-values, corrected for multiple comparisons to control the overall image-wise chance of false positives. The SPM paradigm can also be applied to data expressed on surfaces (Tucholka et al., 2012; Worsley et al., 2009) and is being increasingly adopted outside of its neuroimaging roots. For example, we have undertaken several studies analysing the thickness and mass of cortical bone in the proximal femur using a surface-based approach (Poole et al., 2011, 2012), while others have analysed both the cortical and trabecular compartments with a voxel-based approach (Carballido-Gamio et al., 2013; Li et al., 2009).

Statistical inference is rarely straightforward, and SPM p-maps need to be interpreted with caution. One possible source of error arises from the spatial normalization. There will always be a degree of misregistration. At best, this will just reduce the spatial precision of the p-maps. At worst, the nature of the misregistration will vary across the different study groups. Such *systematic* misregistration is dangerous, since it might lead to the p-maps showing effects that do not, in fact, correspond to different activations between groups, but instead to different registration errors between groups. This phenomenon is well understood and has been much discussed in the literature, most memorably in the context of voxel-based morphometry (VBM) (Ashburner and Friston, 2001; Bookstein, 2001), an SPM variant for analysing anatomical shape.

Perhaps not surprisingly, the standard approach to dealing with systematic misregistration is to employ a better registration algorithm[1]. However, despite much progress in medical image registration, particularly in human neuroimaging (Klein et al., 2009), SPM studies continue to cite systematic misregistration as a source of error (Acosta-

*Corresponding author, Tel./Fax. +44 1223 332750/332662
Email addresses: ahg13@cam.ac.uk (A.H. Gee), gmt11@cam.ac.uk (G.M. Treece)

[1] The ubiquitous SPM smoothing kernel helps when the misregistration is not systematic, increasing the likelihood of detecting effects at the expense of spatial resolution.

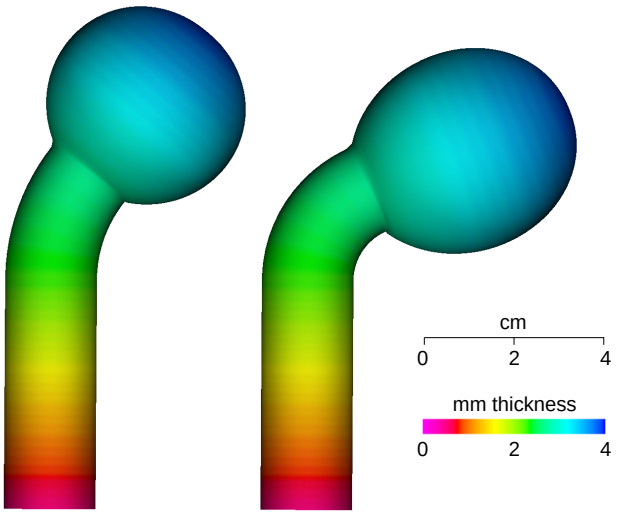


Figure 1: The synthetic lollipop data comprises 100 specimens all with the same thickness distribution. Two of the specimens are shown here.

Cabronero et al., 2010; Garrido et al., 2009; Jung and Haier, 2007; Mohammadi et al., 2012; Oakes et al., 2007; Vangberg et al., 2006). It is our contention that there will always be a degree of arbitrariness in the spatial normalization — we shall argue this point more strongly in Section 3 — and the arbitrariness may turn out to be systematic, affecting different groups in different ways.

In this brief paper, the focus is not on improved registration algorithms, but on methods for detecting and ameliorating false positive and negative inferences caused by systematic misregistration. In neuroimaging, it is not uncommon to allow for total intracranial and grey matter volumes in the GLM (Barnes et al., 2010; Peelle et al., 2012). Depending on the particular registration algorithm, global size measures such as these may correlate with misregistration, and allowing for them in the GLM may guard against false inference. Local misregistration may also be detected by VBM and allowed for in the GLM by way of voxelwise anatomical covariates (Casanova et al., 2007; Oakes et al., 2007). Beyond neuroimaging, we speculate that misregistration will generally depend on the individual specimen's *shape*, and present here a series of experiments designed to investigate the effects of allowing for global shape in the GLM. Our testbed in Section 2 is surface-based statistics on femur-like synthetic surfaces, followed by a case study with real femur data. In addition to allowing for shape, we also outline some well-motivated heuristics that attempt to disentangle misregistration effects from true effects. In Section 3 we discuss our findings before drawing some broad conclusions in Section 4.

2. Experiments and results

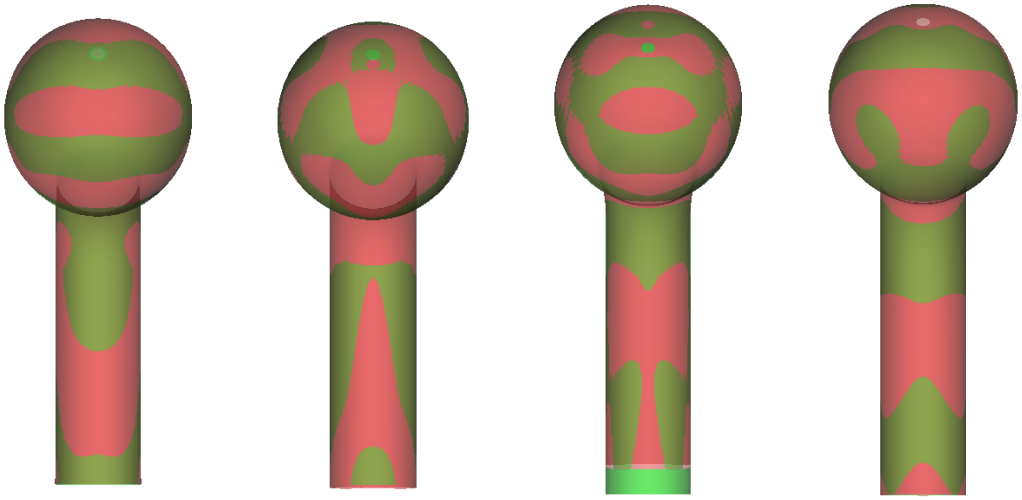
2.1. Synthetic experiments with fixed surface data

In order to explore how systematic misregistration affects surface statistics, we need a cohort of surfaces with known shape and surface data. To this end, we generated triangular meshes of 100 "lollipop" shapes, two of which can be seen in Figure 1. In a parody of our work analysing the cortex of the human femur, the lollipops had a "shaft" and a "head", and to every vertex we assigned a (cortical) "thickness". All 100 specimens had 6000 vertices and identical surface topology. The thickness at each vertex was the same for each specimen, increasing linearly from 0.5 mm at the bottom of the shaft to 4.0 mm at the apex of the head[2]. In contrast, the shapes varied randomly across two degrees of freedom. The head-shaft angle was uniformly distributed in the range 34°–71°, while the head length was uniformly distributed in the range 3.9 cm–4.8 cm. Full details of the procedures used to generate the synthetic data may be found in Appendix A.

We then proceeded to perform a classical, SPM-like analysis of the lollipops' thickness distributions. The first step was to map each individual distribution onto a common morphology, by registering a canonical lollipop (with average head-shaft angle and head length) to each individual. Registration was a two stage process. We first found the seven degree-of-freedom similarity transformation that best aligned the canonical lollipop to the individual. This was accomplished within an iterative framework. At each iteration, every vertex of the canonical mesh was matched with its nearest neighbour on the individual, and then the Levenberg-Marquardt algorithm (More, 1977) was used to find the similarity transformation that minimized the sum of the squared distances between the two point sets. This process was iterated until convergence, requiring typically 50–100 iterations. Following this rough, global alignment, we applied a B-spline free form deformation (FFD) to the canonical surface, with a $4 \times 4 \times 4$ grid of control points, again using iterative nearest neighbour vertex matching and Levenberg-Marquardt optimization to find the control point positions that best aligned the matched vertices. Finally, the individual's thickness distribution was projected onto the aligned canonical surface and smoothed with a 8 mm full-width-half-maximum filter, with all subsequent analysis taking place on the canonical morphology. This procedure is what we use in our femur work and is an unremarkable examplar of similar algorithms used for spatial alignment in medical imaging (Rueckert et al., 1999; Szeliski and Lavalle, 1996; Wang and Fei, 2013). We shall henceforth refer to it as B-spline-based point matching (BPM).

Figure 2 shows some illustrative registration results. Although the distance between the registered surfaces is

[2]In this respect the lollipops do not mimic real femurs, where the cortical thickness tends to decrease from the shaft to the head.



(a) BPM #38 (b) BPM #65 (c1) BPM #94 (c2) 1-to-1 #94

Figure 2: Alignments of the canonical lollipop to specimens 38 (a), 65 (b) and 94 (c). The canonical surface is shown in red with the individual specimens in green. Where the rendering appears red, the canonical surface is slightly in front of the individual specimen, and vice versa for green. (a), (b) and (c1) are BPM alignments. (c2) shows an alignment found not using BPM, but by enforcing the correct 1-to-1 mapping between red and green vertices, and then optimizing the B-spline FFD parameters to minimize the distances between corresponding vertices. Note that a small hole was left in the meshes at the apex of the head. This makes it easier to spot any misalignment since the holes should coincide, as they do in (a) and (c2).

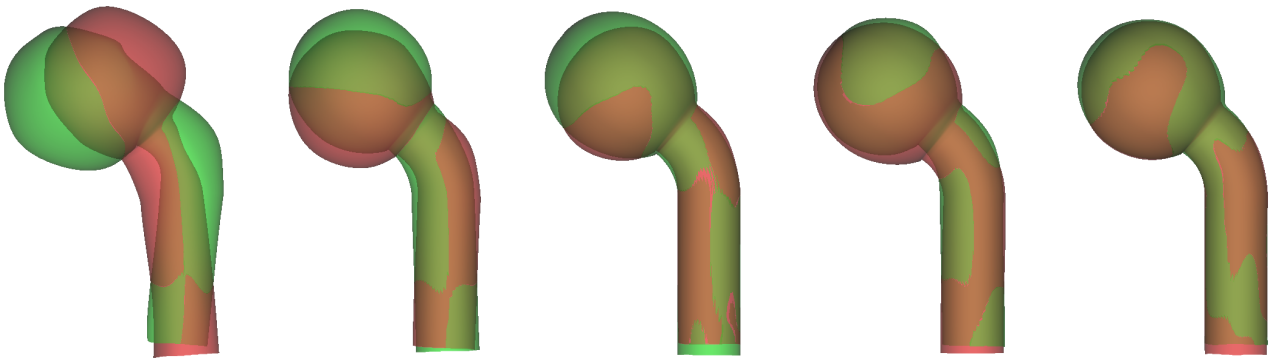
everywhere small, the BPM algorithm does not always find the correct 1-to-1 vertex correspondence, instead getting trapped in some local minimum that depends on the shape of the specimen. In Figure 2, specimen #38 is well registered in (a), but the canonical apex is too high in specimen #65 (b) and too low in specimen #94 (c1), where there is also significant misalignment at the bottom of the shaft. (c2) shows the FFD that best aligns 1-1 corresponding vertices, with a residual error of 0.119 mm per vertex, compared with 0.461 mm per vertex in (c1). Hence, most of the misregistration must be attributed to local minima, though a little (0.119 mm per vertex for specimen #94) arises from the inability of the FFD to capture perfectly the actual deformation.

Following registration, principal component analysis was used to build a point-based, statistical shape model from the 100 sets of canonical vertex coordinates obtained by applying the 100 FFDs. Let $\mathbf{X}_i$ be the 18000-element vector formed by concatenating the canonical vertex coordinates following registration with specimen i, and let $\hat{\mathbf{X}} = \frac{1}{100}\sum_{i=1}^{100}\mathbf{X}_i$. Then the principal modes of shape variation are the eigenvectors $\mathbf{m}_i$ of the sample covariance matrix $\frac{1}{99}\sum_{i=1}^{100}(\mathbf{X}_i - \hat{\mathbf{X}})(\mathbf{X}_i - \hat{\mathbf{X}})^T$. The first five shape modes are shown in Figure 3. Even though there were only two degrees of freedom in the synthetic lollipop data, there are 99 in the shape model, since the registrations are imperfect and the actual deformations are not additive. Shape models of this nature are the standard way to obtain compact shape descriptors of individual specimens, which may be represented according to $\mathbf{X}_i \approx \hat{\mathbf{X}} + \sum_{i=1}^{n} S_i\mathbf{m}_i$. For example, setting $n = 5$ would produce a 5-element shape vector $[S_1 \dots S_5]$ accounting for 90% of the shape variation observed in the population of 100. We shall refer to S_i as the *shape coefficients*.

The next stage in the SPM analysis is to fit a GLM to the thickness distributions (now all expressed on the canonical morphology), to investigate how the thickness depends on explanatory and confounding variables of interest. We used the SurfStat package (Worsley et al., 2009) to fit the GLM and subsequently perform statistical tests on the resulting coefficients. In our first example study, we fitted the GLM $1 + \sum_{i=1}^{10} S_i$ and then performed F-tests on the individual shape coefficients, in order to test whether thickness depends on shape[3] . Figure 4 shows the GLM coefficients and resulting p-maps for S_1, S_3 and S_5. Even though all lollipops had exactly the same thickness distribution, the SPM analysis reveals large areas where there appears to be a statistically significant relationship between thickness and shape[4]. This is because the misregistration depends *systematically* on each specimen's shape: lollipops with a large head-shaft angle tend to misregister

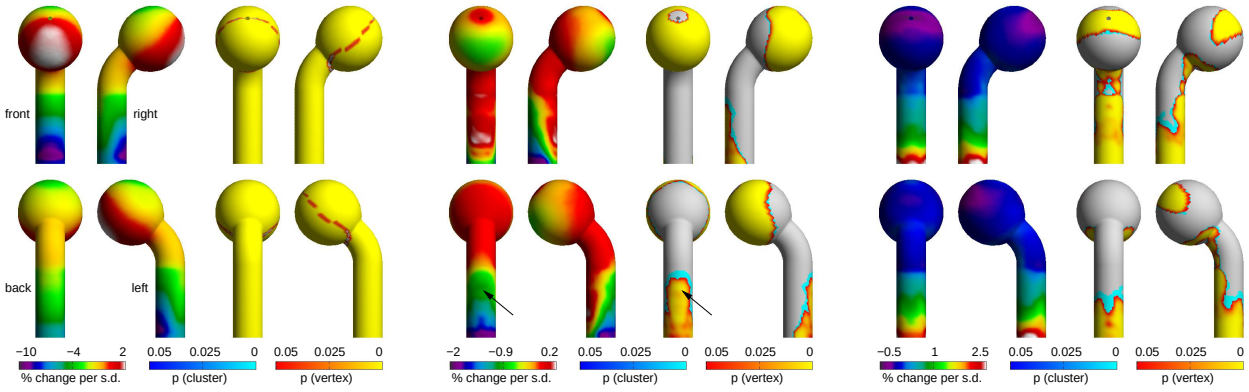
[3]For concision, and in common with many statistics packages, we use the model formula to specify the independent variables in the GLM. A model formula of the form $1 + \sum_{i=1}^{10} S_i$ implies the GLM $y_j = \beta_{0,j} + \sum_{i=1}^{10}\beta_{i,j}S_i + \epsilon_j$, where y_j is the dependent data (in this case, thickness) at vertex j, $\beta_{i,j}$ are the model coefficients and ϵ_j is the residual error.

[4]The p-maps have been corrected for multiple comparisons over vertices, but not for multiple comparisons over different shape modes. The latter multiple comparison problem is far less severe than the former, and is generally ignored in neuroimaging SPM. Regardless, in Figure 4 the various connected clusters are all significant with $p < 0.0002$ and would therefore survive a simple Bonferroni correction.



(a) Mode 1 ±3 s.d. (b) Mode 2 ±3 s.d. (c) Mode 3 ±3 s.d. (d) Mode 4 ±3 s.d. (e) Mode 5 ±3 s.d.

Figure 3: The first five modes of the statistical shape model, accounting for 90% of the population variance.



(a) Mode 1 (b) Mode 3 (c) Mode 5

Figure 4: SPM analysis of the relationship between lollipop thickness and shape. The GLM fitted was $1 + \sum_{i=1}^{10} S_i$. The percentage change maps are for the coefficients of S_1, S_3 and S_5 in the GLM. The corresponding p-maps are for F-tests on S_1, S_3 and S_5. The p-maps are based on the magnitudes of vertex peaks (yellow-orange colour map, sensitive to focal effects) and on the extent of connected clusters exceeding an uncorrected p-value threshold of 0.001 (cyan-blue colour map, sensitive to distributed effects). The arrows in (b) are for comparison with Figure 7(b).

one way, those with a small angle another way, and so on. Consequently, each specimen's thickness distribution "slips" around the canonical surface in a manner that depends on shape, and the resulting, false thickness variation is incorrectly interpreted as a true effect.

The danger of making a false inference is not limited to studies that are overtly concerned with shape. There is also the possibility that some other covariate of interest might depend on shape. Figure 5 shows an example where a covariate C_1 happens to be correlated with S_3. A conventional SPM analysis with GLM $1 + C_1$ (a) reveals a thickness effect where there should not be one. Suspecting systematic misregistration, we might allow for shape in the GLM to avoid such false positive results, effectively treating the shape coefficients as confounding, as opposed to explanatory, variables. However, changing the model to $1 + C_1 + \sum_{i=1}^{2} S_i$ (b) actually strengthens the false signal, since S_1 and S_2 explain much of the variance that was previously interpreted as noise. Only by allowing for more shape modes with the GLM $1 + C_1 + \sum_{i=1}^{10} S_i$ (c) do we remove the false signal, since C_1 does not explain much variance that cannot be explained just as well by S_3. Careful judgement is required to determine how many shape modes to allow for: too many will result in an elaborate model, leaving few degrees of freedom for the statistical analysis and hence compromising statistical power. There is also the matter of how to interpret any residual patches of significance after allowing for n modes: such regions might be caused by systematic misregistration associated with mode $n+1$. We shall return to this question in Section 2.3.

2.2. Synthetic experiments with varying surface data

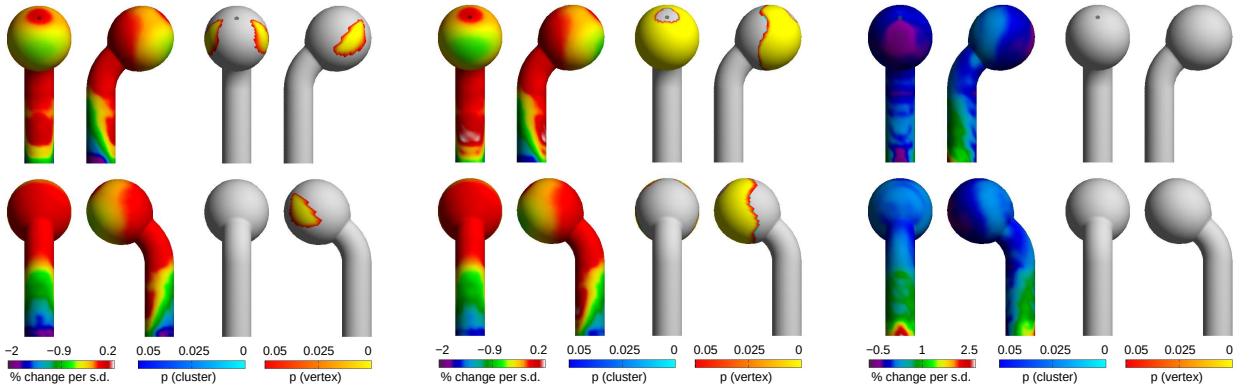
In Section 2.1, the synthetic lollipop data had fixed thickness. In contrast, here we adjust the synthetic data so that there actually *is* a dependence between a covariate C_2 and thickness in a narrow band around the lollipop shaft. Unlike C_1, C_2 was not contrived to correlate with shape. In Figure 6(a), the band of increased thickness is barely significant under a conventional SPM analysis with GLM $1 + C_2$, since the effect is weak compared with the misregistration effects, which are unaccounted for. It is necessary to allow for shape as a confounding variable to see the C_2 effect, using models $1 + C_2 + \sum_{i=1}^{2} S_i$ (b), $1 + C_2 + \sum_{i=1}^{10} S_i$ (c) or even $1 + C_2 + \sum_{i=1}^{30} S_i$ (d). Allowing for more shape modes (e–f) becomes problematic, since the statistical power is gradually reduced until the C_2 effect is no longer significant at the 5% level with 73 modes (not shown). As we remarked in Section 2.1, however many modes we allow for, we need to be aware that any apparent effect might actually be caused by systematic misregistration associated with an unallowed-for shape mode.

2.3. Synthetic experiments with varying surface data that depends on shape

For our final series of synthetic experiments we again adjust the data, this time to induce a dependence between shape, specifically S_3, and thickness in a narrow band around the lollipop shaft. Figure 7 shows the results of two different experiments with the GLM $1 + \sum_{i=1}^{10} S_i$, both looking for any relationship between thickness and S_3. In (a), the data was contrived to show a strong positive correlation between thickness and S_3 all around the band, while in (b) the effect was much weaker. The SPM analysis conflates the true thickness effect with the misregistration effect. Note from Figure 4(b, arrows) that the misregistration results in a negative correlation between thickness and S_3 at the back of the shaft, which cancels with the true, weak effect in Figure 7(b, arrows).

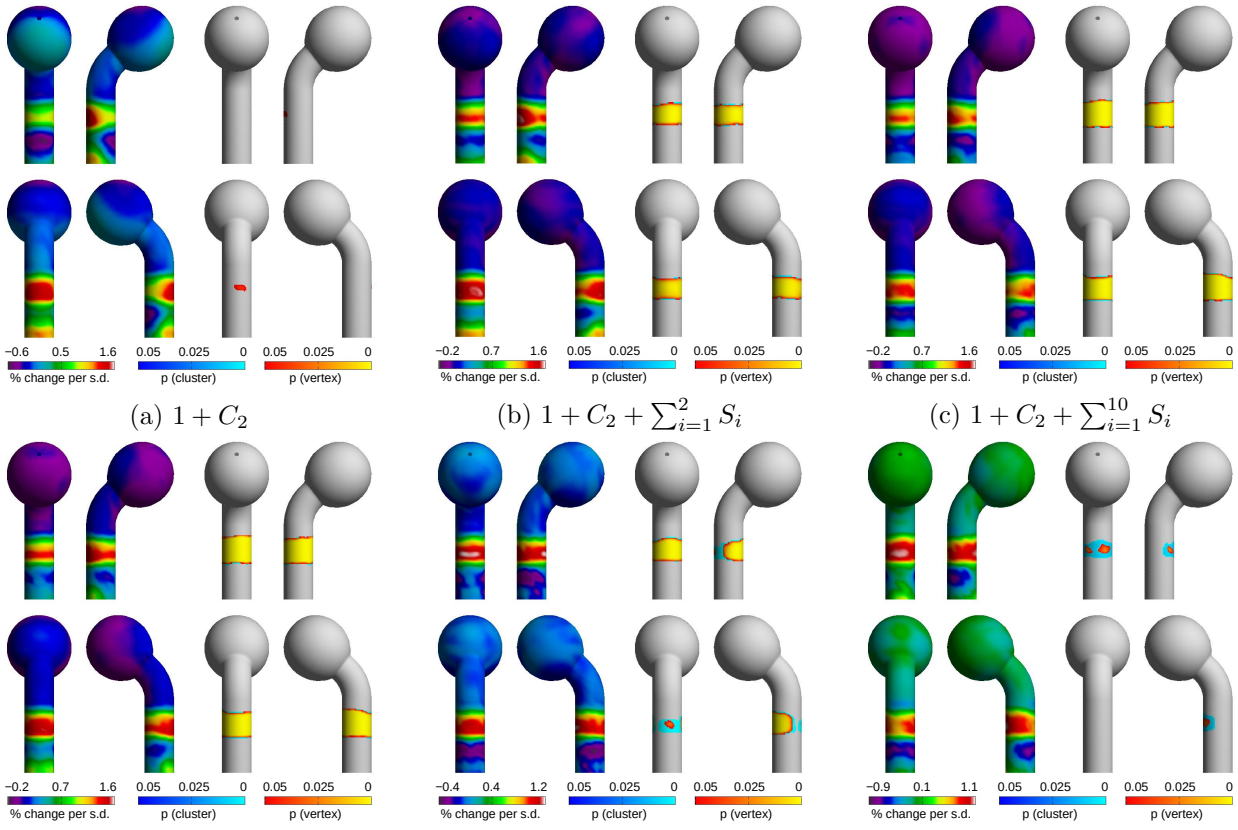
We shall use these two experiments to explore the question that arose at the ends of Sections 2.1 and 2.2: can we disambiguate a true effect from a misregistration effect? One approach would be to compare the *scale* of the apparent effect with the gradient of the mean thickness distribution. In Figure 8(a), we see the mean thickness of the 100 lollipops alongside a very crude estimate of the peak directional derivative computed on the surface. In Figures 8(b) and (c), we divide the S_3 GLM coefficients (left) by the peak gradient, to obtain an estimate (right) of the amount of misregistration, in mm, that would account for the change in thickness induced by one standard deviation of S_3. Note that we are computing an *independent* misregistration at each vertex: we are disregarding the fact that in any physically plausible, smooth misregistration, neighbouring vertices are displaced in roughly the same direction. Looking at Figure 8(b), to explain the thickness effect at the shaft would require a local misregistration in excess of 5 mm per standard deviation, whereas everywhere else only a 1 mm misregistration is necessary. This strong discontinuity suggests that there is a true thickness effect at the shaft. In contrast, in Figure 8(c) all the thickness effects are consistent with a misregistration of less than 1 mm. Note, though, that this might not correspond to a physically plausible, smooth misregistration. So negative results of this nature are inconclusive.

To disambiguate the weak S_3 effect from the misregistration artefacts, we must resort to a more sophisticated test that establishes whether the apparent effects are consistent with a *smooth* misregistration. To some extent, we could attempt this by eye. For example, looking at Figure 6(c), it is fairly obvious that there is no smooth warp of the average lollipop thickness distribution that would produce this effect. To produce the band of positive thickness change, we would need to stretch the shaft downwards, but this would cause collateral damage above and below the band. For less trivial examples, we might seek the assistance of an automatic registration algorithm. Consider again the weak S_3 effect that the simple gradient test failed to disambiguate from a misregistration in Figure 8(c). In Figure 9(a), we see (right) the mean thickness of all 100 specimens, mapped onto the canonical lollipop and (left) the mean thickness plus the apparent thickening explained by S_3. Figure 9(b) shows the difference between the two thickness distributions: this correlates with the percentage



(a) $1 + C_1$ (b) $1 + C_1 + \sum_{i=1}^{2} S_i$ (c) $1 + C_1 + \sum_{i=1}^{10} S_i$

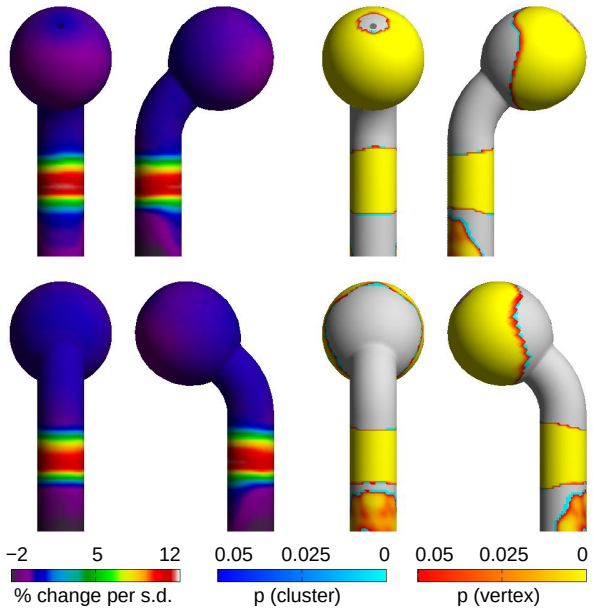
Figure 5: SPM analysis of the relationship between thickness and a covariate C_1 using three different GLMs. In each case, the percentage change maps are for the coefficient of C_1 in the GLM. The corresponding p-maps are for F-tests on C_1.



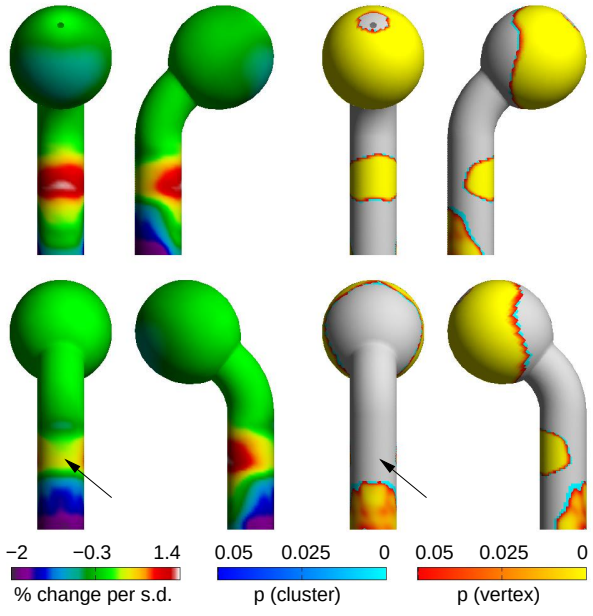
(a) $1 + C_2$ (b) $1 + C_2 + \sum_{i=1}^{2} S_i$ (c) $1 + C_2 + \sum_{i=1}^{10} S_i$

(d) $1 + C_2 + \sum_{i=1}^{30} S_i$ (e) $1 + C_2 + \sum_{i=1}^{50} S_i$ (f) $1 + C_2 + \sum_{i=1}^{70} S_i$

Figure 6: SPM analysis of the relationship between thickness and a covariate C_2 using six different GLMs. In each case, the percentage change maps are for the coefficient of C_2 in the GLM. The corresponding p-maps are for F-tests on C_2.



(a) Strong S_3 effect



(b) Weak S_3 effect

Figure 7: SPM analysis of the relationship between lollipop thickness and S_3. The GLM fitted was $1 + \sum_{i=1}^{10} S_i$. The percentage change maps are for the coefficient of S_3 in the GLM. The corresponding p-maps are for F-tests on S_3. In (a), the synthetic data featured a strong dependence between thickness and S_3, while in (b) the effect was weaker. The arrows in (b) are for comparison with Figure 4(b).

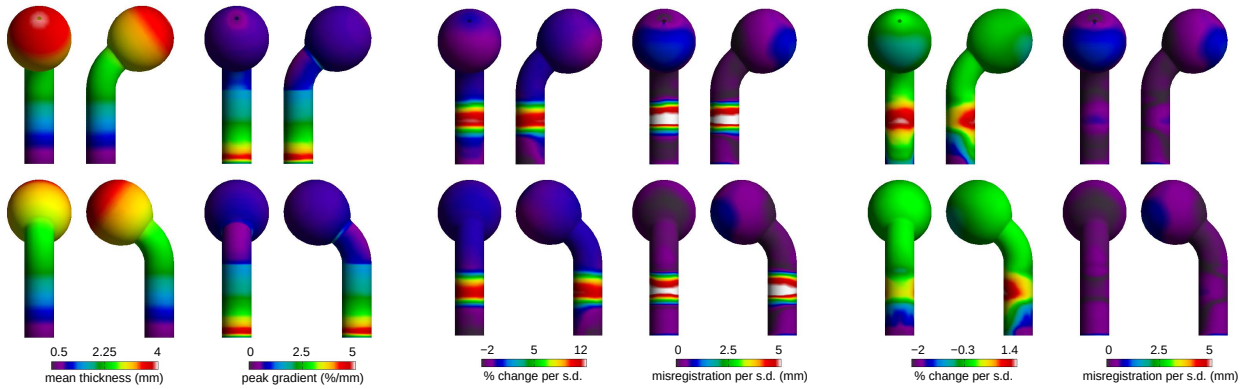
change map in Figure 7(b). We now use a B-spline FFD to deform the right hand canonical lollipop so as best to align the two thickness distributions, producing the result in Figure 9(c). The post-registration thickness difference in Figure 9(d) demonstrates that most of the S_3 effect can be explained by warping the mean thickness distribution (i.e. a misregistration), except in a band around the shaft where there appears to be a true thickness difference. For this proof of concept, we aligned the two thickness distributions using a trivial variation of the BPM algorithm. Instead of matching nearest neighbours, we matched according to thickness, with the additional constraint that matched vertices had to be reasonably proximate. While this approach did succeed in producing the one-off result in Figure 9, in other experiments it was less successful, converging to local minima that were clearly suboptimal. Further research is required to identify, develop and evaluate more robust registration algorithms capable of reliably registering some parts of the distributions (those caused by misregistration) while ignoring others (those caused by true thickness effects). Whether this is even possible, given the inevitable difficulties with registration that are a central tenet of this paper, is a moot point.

2.4. Case study: the cortical mass of the human femur

Finally, we illustrate how systematic misregistration can affect real studies of surface data. In this case, the data is the cortical mass (expressed in mg/cm^2) of the human proximal femur, measured from CT scans of 268 females[5] using the technique described in Treece et al. (2012). Figure 10 shows an SPM analysis examining dependence between cortical mass and femur size. In Figure 10(a), the GLM included femur size as well as other obvious covariates (subject age, weight etc.) but not shape. In Figure 10(b), the GLM was augmented to allow for the first ten shape modes. Note the increased signal strength when allowing for shape, most evident in the sizes of the various significant regions (the p-values are also lower).

Figure 11 shows a second SPM analysis examining dependence between cortical mass and shape modes 1 and 2. These modes correspond, approximately and respectively, to femoral neck length and neck-shaft angle. In (c1) and (d1), we see significant areas where cortical mass appears to depend on shape. While this is perfectly plausible from a physiological perspective — a lifetime of walking on different shaped femurs, with different mechanical stress distributions, is likely to stimulate bone remodelling in different ways — there is also the possibility that the effects arise from systematic misregistration. It is highly suspicious that many of the significant regions for mode 1 coincide with areas of high gradient on the average mass distribution: see Figure 11(b). To investigate further, we

[5]Data from the *FEMCO* study, courtesy of Dr. Ken Poole, School of Clinical Medicine, University of Cambridge, and the *Surgical Treatment of the Hip Joint in Trauma* study, courtesy of Dr. Jan Štěpán, Charles University and Institute of Rheumatology, Prague.



(a) Mean thickness and gradient (b) Strong S_3 effect (c) Weak S_3 effect

Figure 8: A scale-comparison heuristic for disentangling true effects from misregistration artefacts. To obtain the peak gradient estimate in (a), we calculated the thickness gradient along each edge of the mesh, and then labelled the gradient at each vertex with the maximum absolute gradient found amongst the edges incident at the vertex.

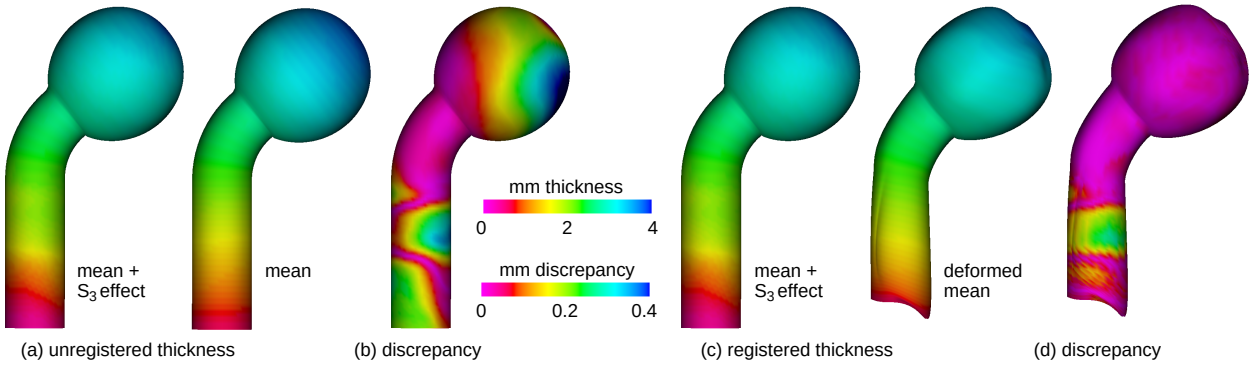
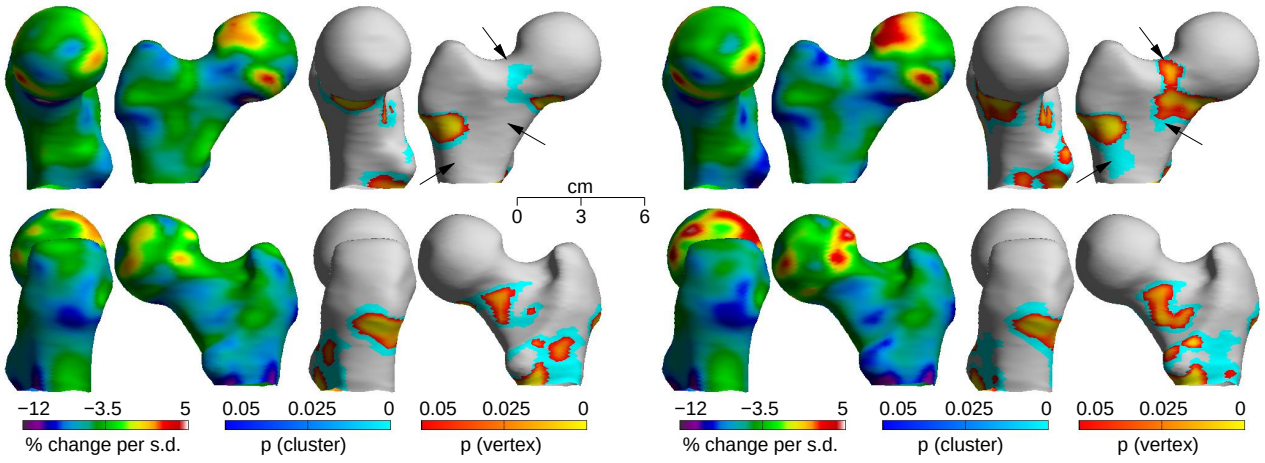
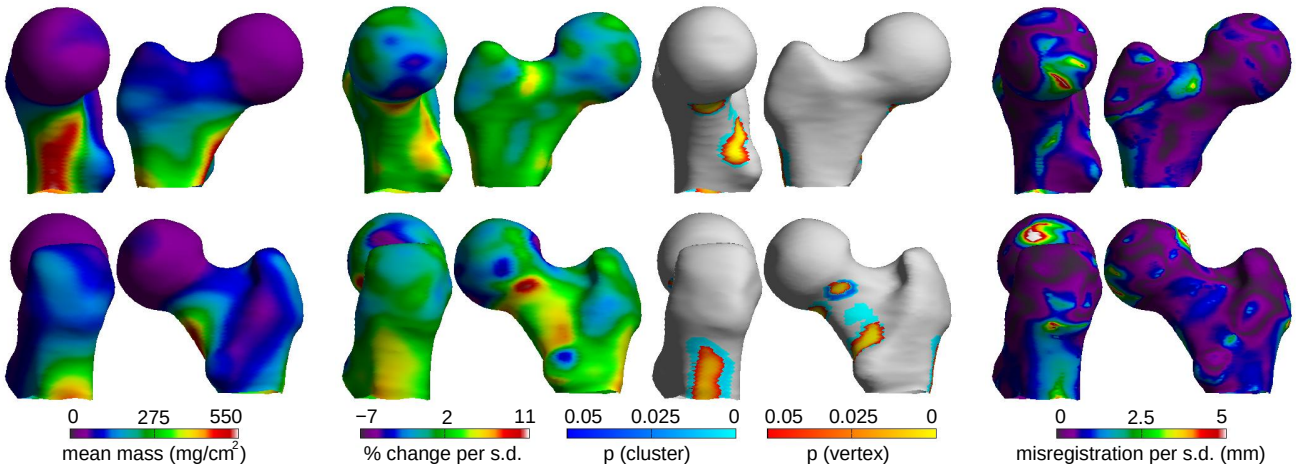


Figure 9: A registration heuristic for disentangling true effects from misregistration artefacts. This example is for the weak S_3 effect for which the scale-comparison heuristic in Figure 8(c) was inconclusive. The thickness distribution in (a, left) and (c, left) was generated by adding $k\times$ the GLM coefficient of S_3 to the mean thickness. We tuned k to coax the best performance out of the modified BPM registration algorithm.

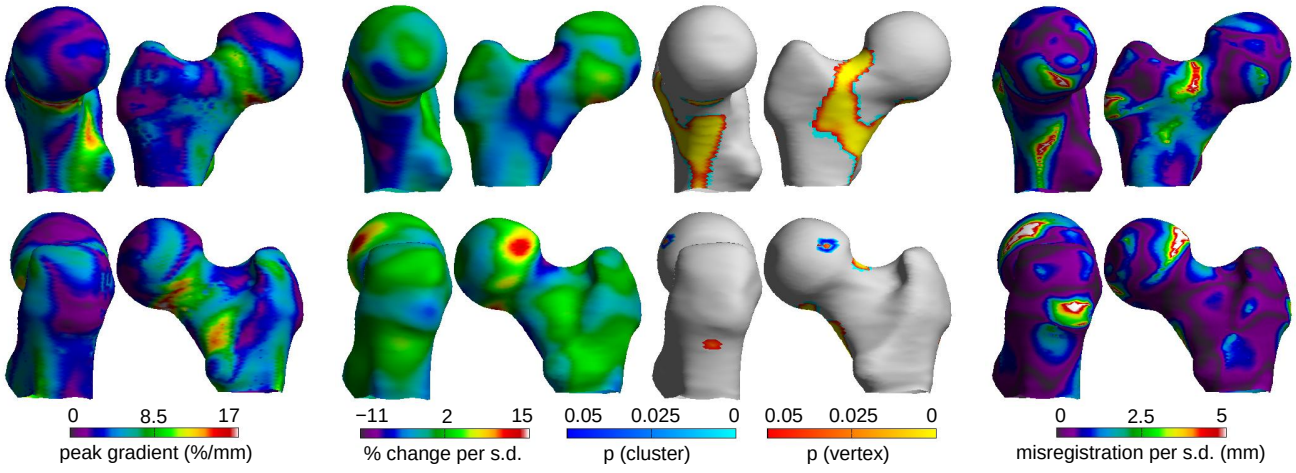


(a) Not allowing for shape (b) Allowing for shape

Figure 10: (a) A conventional SPM analysis of the dependence of cortical mass on femur size, allowing for obvious covariates (subject age, weight etc.) but not shape. (b) The same analysis, but this time allowing for the first ten shape modes. In each case, the percentage change maps are for the coefficient of femur size in the GLM. The corresponding p-maps are for F-tests on femur size. In (a), the significant connected clusters span 1011 vertices, whereas in (b) they span 1329 vertices. The arrows indicate areas where the clusters are enlarged in (b).



(a) Mean mass (c1) Mass dependence on mode 1 (c2) Misregistration



(b) Peak mass gradient (d1) Mass dependence on mode 2 (d2) Misregistration

Figure 11: The scale-comparison heuristic for disentangling true effects from misregistration artefacts, applied to the study of 268 proximal femurs. (a) and (b) show, respectively, the mean cortical mass and the peak mass gradient of the 268 specimens. (c1) and (d1) show an apparent relationship between cortical mass and shape modes 1 and 2. (c2) and (d2) show the amount of misregistration that would explain the apparent shape effects.

performed the simple scale-comparison test of Figure 8, resulting in the misregistration maps in Figures 11(c2) and (d2). True mass effects are likely where the statistically significant regions in (c1) and (d1) coincide with elevated, discontinuous regions in the misregistration maps. It appears that most of the mode 2 effects are associated with true mass changes. The mode 1 effects are ambiguous, they could be caused by small misregistrations, though not necessarily smooth ones.

3. Discussion

One criticism of this work might be that the systematic misregistration arises through the use of an unsophisticated registration algorithm acting on relatively featureless surfaces. While this might indeed affect the scale and extent of the artefact, systematic misregistration remains an issue, even when applying sophisticated registration algorithms to feature-rich surfaces like the human brain (Mohammadi et al., 2012). When registering real biological shapes, the actual modes of shape variation are numerous and unknown, and no parametric registration algorithm will replicate them faithfully, let alone find the global minimum of the objective function. Nonparametric registration algorithms, which allow arbitrary correspondences between surface points, are no panacea. In the intervals between distinguished points, there is an inevitable arbitrariness to the alignment, bridged by way of some reasonable, though arbitrary, smoothness criterion (see Section 3.1 of Ashburner and Ridgway (2013) for a succinct illustration of this point). Hence, quite apart from the simple failings of unsophisticated registration algorithms, we must acknowledge the fact that it is generally impossible to define a uniquely "correct" registration[6]. And all three sources of misregistration — under-parameterization, local minima, arbitrariness — tend to depend systematically on the surface's shape.

4. Conclusions

The main contribution of this paper has been to investigate the benefits of allowing for global shape in the GLM when performing SPM-like analyses. In the presence of systematic misregistration, this simple step can improve the signal strength and also guard against making false inferences. Nevertheless, there remains the risk that systematic misregistration, caused by unaccounted-for shape modes, lies behind an apparently significant effect, and it is for this reason that we have suggested two heuristics that might help disambiguate true effects from misregistration artefacts. These heuristics are neither definitive nor exhaustive, they are simply two further ideas for best practice verification of SPM results. Other due diligence checks one might carry out include: investigating correlations between covariates and shape modes; comparing the sizes of significant regions with the sizes of any effects associated with the first unallowed-for shape mode; and checking that the results are reasonably invariant to different registration algorithms, and different templates, and different random initializations of the registration algorithm.

[6]This must be borne in mind even when there is no systematic misregistration: the spatial localization of any significant effects is limited by the arbitrariness of the registration. In practice, however, this imprecision is most likely small compared with the greater amount of smoothing (8 mm full-width half-maximum in this paper) that is applied to the surface data, in order to ensure compatibility with the Gaussian random field theory that underpins SPM.

Acknowledgments

The authors would like to thank Dr. Ged Ridgway, Wellcome Trust Centre for Neuroimaging, University College London, and Dr. Tristan Whitmarsh, Department of Engineering, University of Cambridge, for their most helpful comments on drafts of this paper. Thanks are also due to the anonymous referees, whose constructive comments helped to improve the original version of the manuscript.

Appendix A. Generation of synthetic data

The lollipop meshes comprise 100 circular contours with 60 vertices per contour. The first 50 contours have radius 10 mm and are stacked at regular intervals of 1 mm to form a cylindrical shaft. The next 12 contours form the lollipop neck. They also have radius 10 mm, but their centres are placed at 2 mm intervals around a circular arc which subtends a random angle uniformly distributed in the range 34°–71°. The remaining 38 contours form the lollipop head. Their radii follow a sine curve, starting at 10 mm, peaking at 22 mm and finishing at 1.6 mm. Their centres are regularly distributed along a straight line, with a random (but uniform) separation such that the accumulated head length is uniformly distributed in the range 3.9 cm–4.8 cm. For the experiments in Section 2.1, the thickness at each vertex was set to $0.5 + 3.5c/99$, where $c \in \{0 \dots 99\}$ is the contour number.

The standard deviation of S_3 was 37.1. Covariate C_1 was generated by adding Gaussian noise to S_3 according to $C_1 = 0.1(S_3 + N(0, 13))$. This gives a correlation coefficient between C_1 and S_3 of 0.94. A high correlation was necessary to produce an effect in Figure 5(a): had C_1 been correlated instead with S_1 or S_2, a much lower correlation would have sufficed. We chose S_3 to illustrate how allowing for shape modes can, in certain circumstances, strengthen false signals. Covariate C_2 was pure Gaussian noise according to $C_2 = N(0, 4)$.

For the experiments in Figure 6, the thickness at vertices 1500–2500 (roughly contours 26–42) was incremented by $C_2/200$. For Figure 7(a), the thickness increment was $S_3/200$ while in Figure 7(b) it was $S_3/1800$.

Acosta-Cabronero, J., Williams, G.B., Pengas, G., Nestor, P.J., 2010. Absolute diffusivities define the landscape of white matter degeneration in Alzheimer's disease. Brain 133, 529–539.

Ashburner, J., Friston, K.J., 2001. Why voxel-based morphometry should be used. Neuroimage 14, 1238–1243.

Ashburner, J., Ridgway, G.R., 2013. Symmetric diffeomorphic modelling of longitudinal structural MRI. Front Neurosci 6.

Barnes, J., Ridgway, G.R., Bartlett, J., Henley, S.M., Lehmann, M., Hobbs, N., Clarkson, M.J., MacManus, D.G., Ourselin, S., Fox, N.C., 2010. Head size, age and gender adjustment in MRI studies: a necessary nuisance? Neuroimage 53, 1244–1255.

Bookstein, F.L., 2001. Voxel-based morphometry should not be used with imperfectly registered images. Neuroimage 14, 1454–1462.

Carballido-Gamio, J., Harnish, R., Saeed, I., Streeper, T., Sigurdsson, S., Amin, S., Atkinson, E.J., Therneau, T.M., Siggeirsdottir, K., Cheng, X., III, L.J.M., Keyak, J., Gudnason, V., Khosla, S., Harris, T.B., Lang, T.F., 2013. Proximal femoral density distribution and structure in relation to age and hip fracture risk in women. J Bone Miner Res 28, 537–546.

Casanova, R., Srikanth, R., Baer, A., Laurienti, P.J., Burdette, J.H., Hayasaka, S., Flowers, L., Wood, F., Maldjian, J.A., 2007. Biological parametric mapping: A statistical toolbox for multimodality brain image analysis. Neuroimage 34, 137–143.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: A general linear approach. Hum Brain Mapp 2, 189–210.

Garrido, L., Furl, N., Draganski, B., Weiskopf, N., Stevens, J., Tan, G.C.Y., Driver, J., Dolan, R.J., Duchaine, B., 2009. Voxel-based morphometry reveals reduced grey matter volume in the temporal cortex of developmental prosopagnosics. Brain 132, 3443–3455.

Jung, R.E., Haier, R.J., 2007. The parieto-frontal integration theory (p-fit) of intelligence: Converging neuroimaging evidence. Behav Brain Sci 30, 135–154.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46, 786–802.

Li, W., Kornak, J., Harris, T., Keyak, J., Li, C., Lu, Y., Cheng, X., Lang, T., 2009. Identify fracture-critical regions inside the proximal femur using statistical parametric mapping. Bone 44, 596–602.

Mohammadi, S., Keller, S.S., Glauche, V., Kugel, H., Jansen, A., Hutton, C., Flel, A., Deppe, M., 2012. The influence of spatial registration on detection of cerebral asymmetries using voxel-based statistics of fractional anisotropy images and TBSS. PLOS ONE 7, e36851.

More, J.J., 1977. The Levenberg-Marquardt algorithm: Implementation and theory, in: Watson, A. (Ed.), Numerical Analysis. Lecture Notes in Mathematics 630, Springer-Verlag, pp. 105–116.

Oakes, T.R., Fox, A.S., Johnstone, T., Chung, M.K., Kalin, N., Davidson, R.J., 2007. Integrating VBM into the general linear model with voxelwise anatomical covariates. Neuroimage 34, 500–508.

Peelle, J.E., Cusack, R., Henson, R.N., 2012. Adjusting for global effects in voxel-based morphometry: Gray matter decline in normal aging. Neuroimage 60, 1503–1516.

Poole, K.E.S., Treece, G.M., Mayhew, P.M., Vaculik, J., Dungl, P., Horák, M., Štěpán, J.J., 2012. Cortical thickness mapping to identify focal osteoporosis in patients with hip fracture. PLOS ONE 7, e38466.

Poole, K.E.S., Treece, G.M., Ridgway, G.R., Mayhew, P.M., Borggrefe, J., Gee, A.H., 2011. Targeted regeneration of bone in the osteoporotic human femur. PLOS ONE 6, e16190.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE T Med Imaging 18, 712–721.

Szeliski, R., Lavalle, S., 1996. Matching 3-D anatomical surfaces with non-rigid deformations using octree-splines. Int J Comput Vision 18, 171–186.

Treece, G.M., Poole, K.E.S., Gee, A.H., 2012. Imaging the femoral cortex: thickness, density and mass from clinical CT. Med Image Anal 16, 952–965.

Tucholka, A., Fritsch, V., Poline, J.B., Thirion, B., 2012. An empirical comparison of surface-based and volume-based group studies in neuroimaging. Neuroimage 63, 1443–1453.

Vangberg, T.R., Skranes, J., Dale, A.M., Martinussen, M., Brubakk, A.M., Haraldseth, O., 2006. Changes in white matter diffusion anisotropy in adolescents born prematurely. Neuroimage 32, 1538–1548.

Wang, H., Fei, B., 2013. Nonrigid point registration for 2D curves and 3D surfaces and its various applications. Phys Med Biol 58, 4315–4330.

Worsley, K., Taylor, J., Carbonell, F., Chung, M., Duerden, E., Bernhardt, B., Lyttelton, O., Boucher, M., Evans, A., 2009. Surfstat: A Matlab toolbox for the statistical analysis of univariate and multivariate surface and volumetric data using linear mixed effects models and random field theory. Neuroimage 47, S102–S102. Organization for Human Brain Mapping, 2009 Annual Meeting.