Statistical analysis of areal quantities in the brain through permutation tests

DISSERTATION

to obtain the joint degree of Doctor at
Maastricht University and Université de Liège,
in Biomedical and Pharmaceutical Sciences.

on the authorities of the Rectores Magnifici,
Professor Rianne Letschert and Professor Albert Corhay.

in accordance with the decision of the Board of Deans,
to be defended in public on Monday the 10 July 2017 at 16h00.

by

Anderson M. Winkler

Supervisors:

- Prof. Dr. Paul M. Matthews, Maastricht University, The Netherlands
- Prof. Dr. Andre Luxen, Université de Liège, Belgium

Co-supervisor:

- Prof. Dr. Thomas E. Nichols, University of Warwick, United Kingdom

Assessment Committee:

- Ir. Dr. Christophe Phillips, Université de Liège, Belgium (*head*)
- Prof. Dr. Gerard J. P. van Breukelen, Maastricht University, The Netherlands
- Prof. Dr. Pierre Maquet, Université de Liège, Belgium
- Dr. Edouard Duchesnay, Commissariat à l'énergie atomique et aux énergies alternatives (CEA), France
- Prof. Dr. John Suckling, University of Cambridge, United Kingdom
- Dr. Giancarlo Valente, Maastricht University, The Netherlands

The work presented in this thesis was funded by the European Union within the PEOPLE Programme FP7: Marie Curie Initial Training Networks (FP7-PEOPLE-ITN-2008), Grant Agreement 238593 "Neurophysics", in which Universiteit Maastricht, Université de Liège, Forschungszentrum Jülich, and GlaxoSmithKline were network partners.

Abstract

In this thesis we demonstrate that direct measurement and comparison across subjects of the surface area of the cerebral cortex at a fine scale is possible using mass conservative interpolation methods. We present a framework for analyses of the cortical surface area, as well as for any other measurement distributed across the cortex that is areal by nature, including cortical gray matter volume. The method consists of the construction of a mesh representation of the cortex, registration to a common coordinate system and, crucially, interpolation using a pycnophylactic method. Statistical analysis of surface area is done with power-transformed data to address lognormality, and inference is done with permutation methods, which can provide exact control of false positives, making only weak assumptions about the data. We further report on results on approximate permutation methods that are more flexible with respect to the experimental design and nuisance variables, conducting detailed simulations to identify the best method for settings that are typical for imaging scenarios. We present a generic framework for permutation inference for complex general linear models (GLMs) when the errors are exchangeable and/or have a symmetric distribution, and show that, even in the presence of nuisance effects, these permutation inferences are powerful. We also demonstrate how the inference on GLM parameters, originally intended for independent data, can be used in certain special but useful cases in which independence is violated. Finally, we show how permutation methods can be applied to combination analyses such as those that include multiple imaging modalities, multiple data acquisitions of the same modality, or simply multiple hypotheses on the same data. For this, we use synchronised permutations, allowing flexibility to integrate imaging data with different spatial resolutions, surface and/or volume-based representations of the brain, including non-imaging data. For the problem of joint inference, we propose a modification of the Non-Parametric Combination (NPC) methodology, such that instead of a two-phase algorithm and large data storage requirements, the inference can be performed in a single phase, with more reasonable computational demands. We also evaluate various combining methods and identify those that provide the best control over error rate and power across. We show that one of these, the method of Tippett, provides a link between correction for the multiplicity of tests and their combination.

This page intentionally left blank.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in these, or any other Universities. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Anderson M. Winkler
January 2016

This page intentionally left blank.

Acknowledgements

À minha família.

To the whole network of supervisors and promoters involved in this multi-institutional project, especially Prof. Dr. Thomas E. Nichols and Prof. Dr. Stephen M. Smith, for their effective advisorship.

I am much thankful to the support of Prof. Peter de Weerd, Prof. Andre Luxen, Dr. Philip S. Murphy, and Prof. Paul M. Matthews. I also would like to thank the much helpful assistance of Mr. Ermo Daniëls, Ms. Christl van Veen and Ms. Jeannette Boschma.

I am extremely thankful to the funding provided by the Marie Curie — Initial Training Network (MC-ITN) "Methods in Neuroimaging", through its four core partners, Universiteit Maastricht, Université de Liège, Forschungszentrum Jülich and GlaxoSmithKline.

Some chapters benefited from strong, prolific, and enriching collaboration. The work on areal interpolation (Chapter 2) would have been impossible without the help of, first and foremost, David C. Glahn. I am also much thankful to Mert R. Sabuncu, B. T. Thomas Yeo, Bruce Fischl, Douglas N. Greve, Peter Kochunov, and John Blangero. The work on permutation for the general linear model (Chapter 3) greatly benefited from the participation of Gerard R. Ridgway and Matthew A. Webster. The work on combined inference (Chapter 4) was much improved thanks to the participation of Matthew A. Webster, Jonathan C. Brooks and Irene Tracey.

This page intentionally left blank.

Propositions

In complement of the dissertation:
Statistical analysis of areal quantities in the brain through permutation tests
by Anderson M. Winkler

Propositions 1–4 are related to the subject matter of the dissertation; Propositions 3–7 are related to the subject field of the doctoral candidate; Proposition 8 is not related to either.

1. Analysis of brain cortical surface area has received insufficient attention compared to thickness and volume, even though it provides a different kind of information about the cortex, particularly when compared to thickness.

2. Pycnophylactic interpolation is the most appropriate method to resample areal quantities to allow comparisons between individuals.

3. The G-statistic provides a simple generalisation over various well known statistics. Written in matrix form, it can be computed quickly for imaging data, and assessed through permutations, sign flippings, or permutations with sign flippings, either freely or with restrictions imposed by exchangeability blocks, depending on knowledge or assumptions about the data and residuals.

4. Non-parametric combination can be modified so as to run in a single phase, rendering its use feasible for imaging data, and offering in general higher power compared to classical multivariate tests.

5. Voxel-based morphometry (VBM) had its time, but should no longer be used for serious research of cortical anatomy, particularly given that other methods are readily available.

6. Cortical surface area at finer resolution can provide adequate traits that are closer to gene action and may be more successful for the identification of genes that influence brain structure and function.

7. Cortical surface area is heritable and has potential to be an endophenotype for psychiatric disorders.

8. Knowledge of genetic influences on brain structure and function should be used to fight disease and improve quality of life. Its influences on policy making, however, must be seen with caution, and receive due, wide consideration.

This page intentionally left blank.

Contents

1 Introduction **17**

1.1 Methods for areal quantities 20

1.2 Methods for permutation inference 21

1.3 Methods for joint permutation inference 22

2 Areal quantities in the cortex **25**

2.1 Introduction 25

2.2 Method 27

2.2.1 Area per face and other areal quantities 29

2.2.2 Computation of surface area 31

2.2.3 Volume as an areal quantity 31

2.2.4 Registration 32

2.2.5 Areal interpolation 33

2.2.6 Implementation 35

2.2.7 Geodesic spheres and areal inequalities 37

2.2.8 Smoothing 38

2.2.9 Conversion from facewise to vertexwise 39

2.2.10 Statistical analysis 40

2.2.11 Presentation of results 41

2.3 Evaluation 41

2.3.1 Registration 43

2.3.2 Distributional characterization 44

2.3.3 Comparison with expansion/contraction methods 54

2.3.4 Validation and stability 59

2.4 Discussion 61

2.4.1 Registration . . . 61
2.4.2 Areal interpolation . . . 62
2.4.3 Statistical analysis of areal quantities . . . 63
2.4.4 Box–Cox and log-normality . . . 64
2.4.5 Further developments and potential applications . . . 64
2.5 Chapter conclusion . . . 65

3 Permutation inference **67**
3.1 Introduction . . . 67
3.2 Theory . . . 70
3.2.1 Model and notation . . . 70
3.2.2 Model partitioning . . . 73
3.2.3 Permutations and exchangeability . . . 73
3.2.3.1 Unrestricted exchangeability . . . 77
3.2.3.2 Restricted exchangeability . . . 80
3.2.4 Number of permutations . . . 86
3.2.5 Multiple testing . . . 89
3.2.6 The randomise algorithm . . . 90
3.3 Worked examples . . . 93
3.4 Evaluation methods . . . 100
3.4.1 Choice of the statistic . . . 100
3.4.2 Permutation strategies . . . 102
3.5 Results . . . 103
3.5.1 Choice of the statistic . . . 103
3.5.2 Permutation strategies . . . 105
3.6 Discussion . . . 110
3.6.1 Permutation tests . . . 111
3.6.2 Pivotal statistics . . . 112
3.6.3 Permutation strategies . . . 113
3.7 Chapter conclusion . . . 114

4 Combined inference **115**
4.1 Introduction . . . 115

4.1.1 Multiple tests — but not the usual multiplicity 115
4.1.2 Combination of imaging modalities 118
4.1.3 Overview of the chapter 121
4.2 Theory 122
4.2.1 History 122
4.2.2 Notation and general aspects 124
4.2.3 Union–intersection and intersection–union tests 126
4.2.4 Closed testing 129
4.2.5 Non-parametric combination 130
4.2.6 Overview of combining functions 131
4.2.7 Transformation of the statistics 138
4.2.8 Directed, non-directed, and concordant hypotheses 141
4.2.9 Consistency of combined tests 143
4.2.10 Admissibility of combined tests 143
4.2.11 The method of Tippett 145
4.2.12 A unified procedure 147
4.2.13 Implementation 149
4.3 Evaluation methods 154
4.3.1 Validity of the modified NPC 154
4.3.2 Performance of combined tests 156
4.3.3 Example: Pain study 157
4.4 Results 158
4.4.1 Validity of the modified NPC 158
4.4.2 Performance of combined tests 161
4.4.3 Example: Pain study 163
4.5 Discussion 163
4.5.1 Validity of the modified NPC 163
4.5.2 Performance of combined tests 165
4.5.3 Interpretation of combined tests 166
4.5.4 Correction over contrasts and over modalities 167
4.5.5 Pain study 168
4.5.6 Relationship with meta-analysis 168
4.5.7 Applicability for cortical volumes 169

4.6 Chapter conclusion 169

A Valorisation **171**

A.1 Introduction 171

A.2 Thesis impact 172

A.2.1 Peer-reviewed publications 172

A.2.2 Presentations in conferences 173

A.2.3 Talks 173

A.2.4 Public engagement 174

A.2.5 Software 174

A.3 Further perspectives 175

B Supporting Information **177**

References **179**

List of Figures

1.1 Some possible analyses of cortical morphometric measurements using permutation tests. 20

2.1 Surface- and volume-based representations of the cortex. 26
2.2 Example demonstrating differences between area and point measurements. 28
2.3 Overview of areal analyses. 30
2.4 Overlapping areas used to weight areal quantities during interpolation. 34
2.5 Geodesic spheres. 38
2.6 Differences between presentation of facewise and vertexwise area. 42
2.7 Effect of registration method on areal analyses. 45
2.8 The distribution of surface area is lognormal. 47
2.9 Results of the Shapiro–Wilk normality test. 48
2.10 Maps of the skewness of the areal data. 49
2.11 Histograms of the skewness of the areal data. 50
2.12 Maps of the kurtosis of the areal data. 51
2.13 Histograms of the kurtosis of the areal data. 52
2.14 Spatial distribution of the parameter λ. 53
2.15 Example of a retessellated surface. 55
2.16 Comparison with expansion/contraction methods (I). 57
2.17 Comparison with expansion/contraction methods (II). 58
2.18 Comparison with expansion/contraction methods (III). 59

3.1 Examples of permutation and sign flipping matrix. 75

3.2 Example of permutation and sign flipping matrix for within-block exchangeability . . . 82
3.3 Example of permutation and sign flipping matrix for whole-block exchangeability . . . 84
3.4 Heatmaps for evaluation of pivotality for F and G statistics. . . . 104

4.1 Rejection regions for union–intersection and intersection–union tests. . . . 127
4.2 Overview of the original and modified NPC . . . 140
4.3 Rejection regions of two partial tests with four different combining functions. . . . 142
4.4 Examples of inconsistent combining functions. . . . 144
4.5 Examples of inadmissible combining functions. . . . 146
4.6 Overview of the main simulation parameters. . . . 156
4.7 Histograms of p-values for the simulations. . . . 159
4.8 Bland–Altman plots comparing original and modified NPC. . . . 160
4.9 Performance of the modified NPC using different combining functions, compared to Hotelling's T^2. . . . 162
4.10 Results of the example pain study. . . . 164

A.1 The proposed tests are available in PALM. . . . 175

List of Tables

2.1 Problems solved by the proposed method. 55
2.2 Stability of areal measurements. 60

3.1 Summary of assumptions of permutation methods. 82
3.2 Methods available to construct the null distribution in the presence of nuisance variables. 83
3.3 Some tests of which the statistic G is a generalisation. 86
3.4 Maximum number of unique permutations. 88
3.5 Coding for Example 1 94
3.6 Coding for Example 2 95
3.7 Coding for Example 3 96
3.8 Coding for Example 4 97
3.9 Coding for Example 5 98
3.10 Coding for Example 6 99
3.11 Summary of simulation scenarios. 101
3.12 False positive rate and power for the statistics F and G. 106
3.13 Summary of the amount of error type I and power for the different permutation strategies. 107
3.14 Amount of error type I for representative simulation scenarios. . 107

4.1 Summary of various combining functions 119
4.2 Joint hypotheses of UIT and IUT. 126

This page intentionally left blank.

Chapter 1

Introduction

It has been suggested that the processes that drive horizontal (tangential) and vertical (radial) development of the cerebral cortex are separate from each other (Rakic, 1988). Variations on these would result, respectively, in variations on the extent of cortical surface area and on the thickness of the cortical mantle. Through the use of genetically informative samples, it has been demonstrated these two processes are indeed uncorrelated genetically (Panizzon et al., 2009; Winkler et al., 2010) and are each influenced by regionally distinct genetic factors (Schmitt et al., 2008; Rimol et al., 2010b). Moreover, it is variation on surface area that explains most of the variation observed in the amount of gray matter assessed with methods that only measure volume, such as voxel-based morphometry (Winkler et al., 2010; Rimol et al., 2012).

These findings give prominence to the use of surface area alongside cortical thickness in studies of brain morphology and, and its interaction with brain function. However, cortical surface area been measured only over gross regions or approached indirectly via comparisons with a standard brain. Of studies using the latter, few that have used area measurements on every point of the cortex (vertex-wise) and have offered detailed insight on the exact procedures used for this assessment. Some studies described their methods in terms of "expansion/contraction", often using different definitions of what expansion or contraction would be. By 2011, various impromptu approaches had been considered, for example:

– Lyttelton et al. (2009): The authors describe that the asymmetry measurement is the logarithm of the ratio of the area per vertex of left and right

hemispheres. Expansion or contraction are in relation to the contralateral hemisphere.

- Joyner et al. (2009): After a brief description of the method of measurement, the authors state that "(...) this provides point-by-point estimates of the relative areal expansion or compression of each location in atlas space." Expansion/contraction are relative to the chosen template.

- Sun et al. (2009a): The authors state that "The distance between a center position of the brain (...) and each brain surface point was calculated (...). The difference of the above radial distances between the follow-up and baseline brain surfaces (...) was defined as brain surface contraction". Under this definition, not only contraction refers to an initial point in time, but it also refers not to a bidimensional feature, and instead to a linear distance between each point in the surface and a given central point in the brain.

- Sun et al. (2009b): The authors state that "The distance between two brain surfaces [*i.e. inner skull and pial*] was then measured at subvoxel resolution (...), the value in millimeters was assigned to the voxel as the intensity value and an image of the brain surface contraction was obtained". Under this definition, for a longitudinal study, the contraction is the difference between initial and final distances between inner skull and pial surfaces, assigned to a volumetric (voxel-based) space.

- Hill et al. (2010): The article discusses growth of the cortex from birth to adulthood and compares it with the cortex of the monkey. Here expansion can be interpreted as in relation to an initial, developmental and/or evolutionary stage, not to a given template or to the other hemisphere.

- Rimol et al. (2010a): Expansion and contraction are measured in relation to a template, as in Joyner et al. (2009).

- Palaniyappan et al. (2011): The authors state that "In line with Joyner et al. (2009), we use the term contraction to suggest group differences in the surface area in patients compared to controls, rather than a reduction from previously larger area." This in fact seems a new interpretation over the method

used by Joyner et al. (2009), as the authors here would then be using expansion/contraction to compare to the control group. Yet, reading through the article, it appears clear that expansion/contraction still refers to the chosen template.

– Chen et al. (2011, 2012): The authors use a method similar to Joyner et al. (2009) and Rimol et al. (2010a), and so, expansion/contraction refer to the template.

All these different operating definitions of what expansion/contraction would be create already difficulties in the interpretation of their meaning. However, even if only one of these existed, it would still be difficult to interpret, due to the dependence of all these methods on a reference brain or on the contra-lateral hemisphere, from which expansion or contraction is tentatively assessed.

In the present work, we propose a method that uses absolute quantities, as opposed to being relative to a reference brain. While initially addressing these concerns, we found yet others that required further investigation. The first is that we found that surface area is lognormally distributed, such that direct use of statistical methods based on the assumption of normality are likely to yield incorrect results. The second concerns use of data assigned to each face of a mesh representation of the brain, as opposed to each vertex, which cannot be analysed in software designed to handle vertexwise data, nor stored in vertexwise file formats, thus demanding the development of new tools for analysis and a file format. The third is that in neuroimaging thousands of tests are performed in an image representation of the brain. None of the parametric methods can be considered for control of the familywise error rate, given the lognormality and the spatial dependencies among the data assigned to each face of the mesh representation without appealing to many unrealistic assumptions, thus demanding the use of more flexible approaches.

Treating these problems eventually that led into a complete framework for the measurement and statistical analysis of areal quantities. It uses permutation tests in the general linear model, and yet allowing area and thickness to be studied jointly without appealing to cortical volume. Nonetheless, the method can also be used to study volume, either using the current approach of multiplying cortical area by cortical thickness, or else, using an improved method that we propose, in

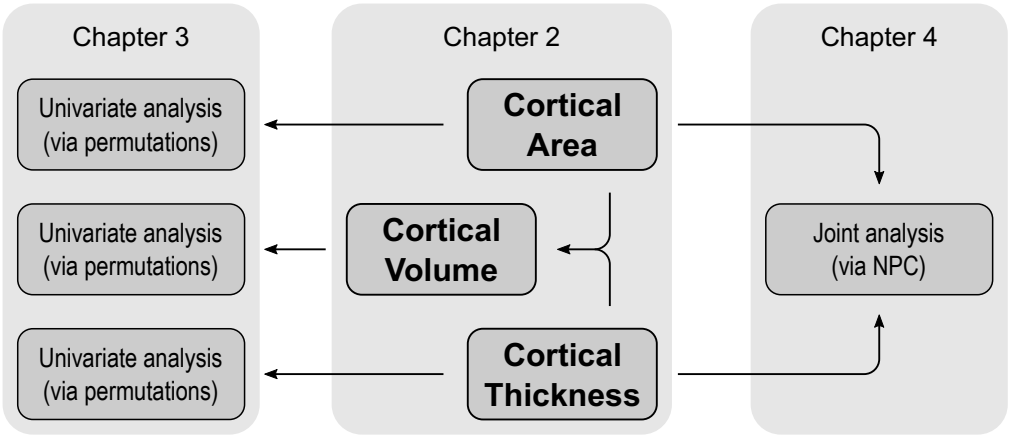


Figure 1.1: Some possible analyses of cortical morphometric measurements using permutation tests. Inter-subject comparisons of cortical area and other areal quantities, such as volume, that depends on both area and thickness, use the methods proposed in Chapter 2. Univariate statistical analysis of each of these separately use the strategy discussed in Chapter 3. Joint (combined) analysis of area and thickness, that bypass volumes altogether, would use the methods proposed in Chapter 4, in particular the non-parametric combination (NPC), but also classical multivariate tests.

which no pieces of the cortex are left over- or under-represented.

Although this work can be organised into three core topics that are relatively independent from each other, and that have each been published as separate papers (Winkler et al., 2012, 2014, 2016), the flow of information in a complete study of cortical morphology visits all three, as shown in Figure 1.1. The next sections outline these three main chapters. Each chapter offers a detailed introduction describing the problem that each aim to solve, along with review of the relevant literature, evaluation and implementation, including algorithms as needed, and a detailed discussion.

1.1 Methods for areal quantities

The general strategy for analyses of cortical measurements consists of the generation of a surface-representation of the brain and its subsequent transformation into a sphere. Vertices of this sphere are then shifted along its surface to allow alignment that matches some feature of interest, such as sulcal depth, myelin con-

tent, or functional markers. As the alignment is performed, quantities assigned to vertices or faces, such as thickness or area, are carried along these vertices and faces. Once registration is done, these quantities are interpolated to a common grid (mesh), where comparisons between subjects can be performed.

While methods to study thickness across subjects are available (Fischl and Dale, 2000), and use interpolation to a common reference grid using methods such as nearest neighbour or barycentric, such interpolation strategies cannot be used for either cortical area itself, nor to other areal quantities, such as cortical volume, as these are not mass-conservative (pycnophylactic). Chapter 2 clarifies the distinction between the nature of these measurements, and proposes the use of areal interpolation. This strategy permits quantities to be studied in absolute terms, as opposed to relative to some reference brain. The chapter proposes that areal quantities are analysed directly in the faces of the mesh from which they were computed, instead of resampled to vertices, which halves the resolution.

We demonstrate that areal data do not follow a normal distribution, being better characterised by a mixture of normal and lognormal distributions, in proportions that vary across the brain and possibly according to the scale of measurement. A power transformation can be considered to address lognormality, although a better alternative is to use permutation methods, that not only do not rely on distributional assumptions, but also allow correction for multiple testing and the use of non-standard statistics.

1.2 Methods for permutation inference

Permutation methods can provide exact control of false positives, making only weak assumptions about the data, and have been available in brain imaging for many particular cases (Holmes et al., 1996; Nichols and Holmes, 2002), although no implementation for surface-based methods, even less so for facewise data as we have developed, existed in the literature until this work. With the recent availability of fast and inexpensive computing, the main limitation of permutation tests would be a certain lack of flexibility with respect to arbitrary experimental designs, in particular with respect to nuisance variables in the model, as well as repeated measurements.

In Chapter 3 we report on results on approximate permutation strategies that are more flexible with respect to experimental designs that include such nuisances. We review the literature and conduct detailed simulations to identify the best method for settings that are typical for imaging research. A generic framework for permutation inference for complex general linear models (GLMs) when the errors are exchangeable and/or have a symmetric distribution, is presented. Even in the presence of nuisance effects, these permutation inferences are powerful and provide control of false positives in a wide range of common and relevant imaging research scenarios.

We also demonstrate how the inference on GLM parameters, originally intended for independent data, can be used in certain special but useful cases in which independence is violated, by means of using exchangeability blocks, that is, sets of observations with shared non-independence, and that can sometimes be treated as a single unit for permutation, i.e., shuffled as a whole, or sometimes serve as delimiters such that permutations happen only within block. The definition of exchangeability blocks allow for groups of observations with same variances, either known or assumed, thus requiring a statistic that preserves certain desirable properties for control of multiple testing even under such scenarios. We provide such a statistic, dubbed G-statistic, which is a generalisation of the F-statistic, as well as others.

1.3 Methods for joint permutation inference

While gray matter volume can be studied directly using the methods discussed in Chapter 2, it may be the case that true effects affecting thickness and area in opposite directions may cancel each other. Yet, analysing them separately using univariate methods as in Chapter 3 may not aggregate power from having effects acting simultaneously on both. Likewise, participants of an imaging study are often subjected to the acquisition of more than one imaging modality. These modalities are often analysed separately. However, a joint analysis has potential to answer more complex questions and to increment power. Moreover, even a single modality can sometimes be partitioned into subcomponents that disentangle different aspects of brain structure or function. Examples include independent component analysis, as

well as scalar measurements from diffusion-tensor imaging.

In Chapter 4 we show how permutation methods can be applied to combination analyses such as those that include multiple imaging modalities, multiple data acquisitions of the same modality, or simply multiple hypotheses on the same data. Using the well-known definition of union-intersection tests and closed testing procedures, we use synchronised permutations to correct for such multiplicity of tests, allowing flexibility to integrate imaging data with different spatial resolutions, surface and/or volume-based representations of the brain, including non-imaging data.

In particular for the problem of joint inference, we propose and evaluate a modification of the recently introduced Non-Parametric Combination (NPC) methodology (Pesarin and Salmaso, 2010a), such that instead of a two-phase algorithm and large data storage requirements, the inference can be performed in a single phase, with reasonable computational demands. We also evaluate, in the context of permutation tests, various combining methods that have been proposed in the past decades, and identify those that provide the best control over error rate and power across a range of situations. We show that one of these, the method of Tippett (1931), provides a link between correction for the multiplicity of tests and their combination.

Finally, we discuss how the correction can solve certain problems of multiple comparisons in common designs, and how the combination is distinguished from conjunctions, even though both can be assessed using permutation tests. We also provide a common algorithm that accommodates combination and correction.

This page intentionally left blank.

Chapter 2

Areal quantities in the cortex

2.1 Introduction

The surface area of the cerebral cortex greatly differs across species, whereas the cortical thickness has remained relatively constant during evolution (Mountcastle, 1998; Fish et al., 2008). At a microanatomic scale, regional morphology is closely related to functional specialization (Roland and Zilles, 1998; Zilles and Amunts, 2010), contrasting with the columnar organization of the cortex, in which cells from different layers respond to the same stimulus (Jones, 2000; Buxhoeveden and Casanova, 2002). In addition, Rakic (1988) proposed an ontogenetic model that explains the processes that lead to cortical arealization and differentiation of cortical layers according to related, yet independent mechanisms. Supporting evidence for this model has been found in studies with both rodent and primates, including humans (Chenn and Walsh, 2002; Rakic et al., 2009), as well as in pathological states (Rimol et al., 2010a; Bilgüvar et al., 2010).

At least some of the variability of the distinct genetic and developmental processes that seem to determine regional cortical area and thickness can be captured using polygon mesh (surface-based) representations of the cortex derived from T_1-weighted magnetic resonance imaging (MRI) (Panizzon et al., 2009; Winkler et al., 2010; Sanabria-Diaz et al., 2010). In contrast, volumetric (voxel-based) representations, also derived from MRI, were shown to be unable to readily disentangle these processes (Winkler et al., 2010). Figure 2.1 shows schematically the difference between these two representations.

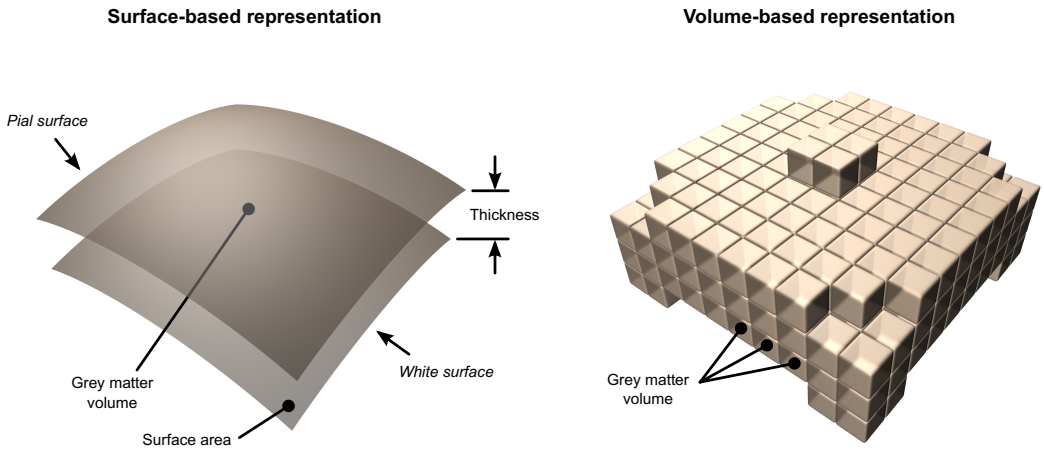


Figure 2.1: Geometrical relationship between cortical thickness, surface area and grey matter volume. In the surface-based representation, the grey matter volume is a quadratic function of distances in the surfaces and a linear function of the thickness. In the volume-based representation, only the volumes can be measured directly and require partial volume-effects to be considered (Winkler et al., 2010).

Mesh representations of the brain allow measurements of the cortical thickness at every point in the cortex, as well as estimation of the average thickness for pre-specified regions. However, to date, analyses of cortical surface area have been generally limited to two types of studies: (1) vertexwise comparisons with a standard brain, using some kind of expansion or contraction measurement, either of the surface itself (Joyner et al., 2009; Lyttelton et al., 2009; Hill et al., 2010; Rimol et al., 2010a; Palaniyappan et al., 2011), of linear distances between points in the brain (Sun et al., 2009a,b), or of geometric distortion (Wisco et al., 2007), or (2) analyses of the area of regions of interest (ROI) defined from postulated hypotheses or from macroscopic morphological landmarks (Dickerson et al., 2009; Nopoulos et al., 2010; Kähler et al., 2011; Durazzo et al., 2011; Schwarzkopf et al., 2011; Eyler et al., 2011; Chen et al., 2011, 2012). Analyses of expansion, however, do not deal with area directly, depending instead on non-linear functions associated with the warp to match the standard brain, such as the Jacobian of the transformation. Moreover, by not quantifying the amount of area, these analyses are only interpretable with respect to the brain used for the comparisons. ROI-based analyses, on the other hand, entail the assumption that each region is homogeneous with regard to the feature under study, and have maximum sensitivity only when the effect of interest is present throughout the ROI.

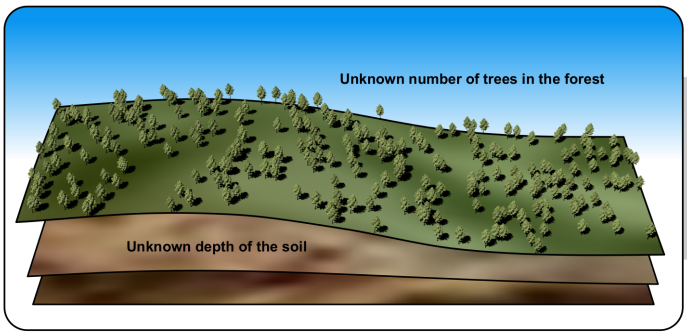
These difficulties can be obviated by analysing each point on the cortical surface of the mesh representation, a method already well established for cortical thickness (Fischl and Dale, 2000). Pointwise measurements, such as thickness, are generally taken at and assigned to each vertex of the mesh representation of the cortex. This kind of measurement can be transferred to a common grid and subjected to statistical analysis. Standard interpolation techniques, such as nearest neighbor, barycentric (Yiu, 2000), spline-based (De Boor, 1962) or distance-weighted (Shepard, 1968) can be used for this purpose. The resampled data can be further spatially smoothed to alleviate residual interpolation errors. However, this approach is not suitable for areal measurements, since area is not inherently a point feature. To illustrate this aspect, an example is given in Figure 2.2. Methods that can be used for interpolation of point features do not necessarily compensate for inclusion or removal of datapoints,[1] unduly increasing or reducing the global or regional sum of the quantities under study, precluding them for use with measurements that are, by nature, areal. The main contribution of this chapter is to address the technical difficulties in analysing the local brain surface area, *as well as any other cortical quantity that is areal by nature*. We propose a framework to analyse areal quantities and argue that a mass preserving interpolation method is a necessary step. We also study different processing strategies and characterize the distribution of *facewise* cortical surface area.

2.2 Method

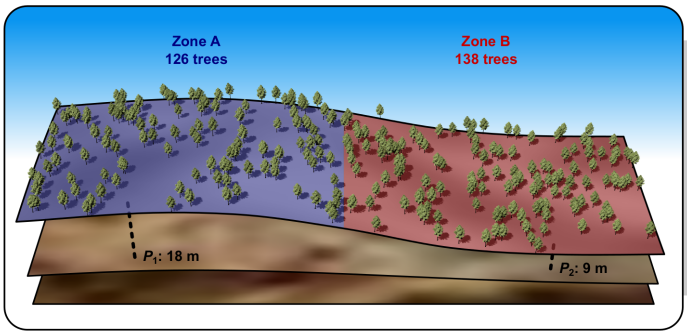
An overview of the method is presented in Figure 2.3. Comparisons of cortical area between subjects require a surface model for the cortex to be constructed. A number of approaches are available (Mangin et al., 1995; Dale et al., 1999; van Essen et al., 2001; Kim et al., 2005) and, in principle, any could be used. Here we adopt the method of Dale et al. (1999) and Fischl et al. (1999a), as implemented in the FreeSurfer software package (FS).[2] In this method, the T_1-weighted images

[1] A notable exception is the natural neighbor method (Sibson, 1981). However, the original method needs modification for use with areal analyses.

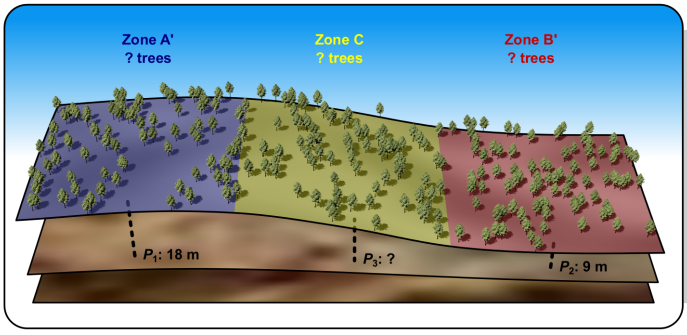
[2] Available at http://surfer.nmr.mgh.harvard.edu.



In a hypothetical, near future scenario, global warming has melted the polar icecaps and profoundly modified landscapes. A group of botanists and geologists decide to study a newly formed forest in Antarctica. Some of the variables to be collected include the number of trees of a certain slow growing species, and the depth of the soil until a specific rocky formation is found. They prepare a trip to the field for data collection.



The botanists divide the forest in zones A and B, and count the number of trees in each, finding 126 and 138, respectively, giving a total of 264. The geologists measure the depth at the points P_1 and P_2 as 18 and 9 m respectively. Back to the lab, they realise that the collected data is not sufficient to verify their hypotheses. Using maps, they redefine the zones A and B to be smaller, A' and B', and define a new one, called C. They also identify a point, P_3, at the mid-distance between P_1 and P_2, which depth needs to be known. Unable to return to the field before their grant is renewed, they decide to estimate the missing data using interpolation.



*Knowing that the depth of the soil there changes gradually, they choose linear interpolation for an estimate at P_3, obtaining (18+9)/2 = 13.5 m. To calculate the number of trees in zone C, the same interpolation method, if used, would give (126 +138)/2 = 132 trees, which is clearly incorrect. The botanists, however, observe that **the number of trees is a different kind of measurement, that is areal by nature and, therefore, needs different methods**. Using areal interpolation, they estimate the number of trees in zones A', B' and C as, respectively, 82, 88 and 94. Some months later, in a second visit to the field, they are pleased to confirm the accuracy of their estimates.*

Figure 2.2: An example demonstrating differences in the nature of measurements. In this analogy, the depth of the soil is similar to brain cortical thickness, whereas the number of trees is similar to areal quantities distributed across the cortex. These areal quantities can be the surface area itself (in this case, the area of the terrain), but can also be any other measurement that is areal by nature (such as the number of trees).

are initially corrected for magnetic field inhomogeneities and skull-stripped (Ségonne et al., 2004). The voxels belonging to the white matter (WM) are identified based on their locations, on their intensities, and on the intensities of the neighboring voxels. A mass of connected WM voxels is produced for each hemisphere, using a six-neighbors connectivity scheme, and a mesh of triangular faces is tightly built around this mass, using two triangles per exposed voxel face. The mesh is smoothed taking into account the local intensity in the original images (Dale and Sereno, 1993), at a subvoxel resolution. Topological defects are corrected (Fischl et al., 2001; Ségonne et al., 2007) ensuring that the surface has the same topological properties of a sphere. A second iteration of smoothing is applied, resulting in a realistic representation of the interface between gray and white matter (the *white surface*). The external cortical surface (the *pial surface*), which corresponds to the pia mater, is produced by nudging outwards the white surface towards a point where the tissue contrast is maximal, maintaining constraints on its smoothness and on the possibility of self-intersection (Fischl and Dale, 2000). The white surface is inflated in an area-preserving transformation and subsequently homeomorphically transformed to a sphere (Fischl et al., 1999b). After the spherical transformation, there is a one-to-one mapping between faces and vertices of the surfaces in the native geometry (white and pial) and the sphere. These surfaces are comprised exclusively of triangular faces.

2.2.1 Area per face and other areal quantities

The surface area for analysis is computed at the interface between gray and white matter, i.e. at the *white surface*. Another possible choice is to use the middle surface, i.e. a surface that runs at the mid-distance between white and pial. Although this surface is not guaranteed to match any specific cortical layer, it does not over or under-represent gyri or sulci (van Essen, 2005), which might be an useful property. The white surface, on the other hand, matches directly a morphological feature and also tends to be less sensitive to cortical thinning or thickening than the middle or pial surfaces. Whenever methods to produce surfaces that represent biologically meaningful cortical layers are available, these should be preferred.

In contrast to conventional approaches in which the area of all faces that meet

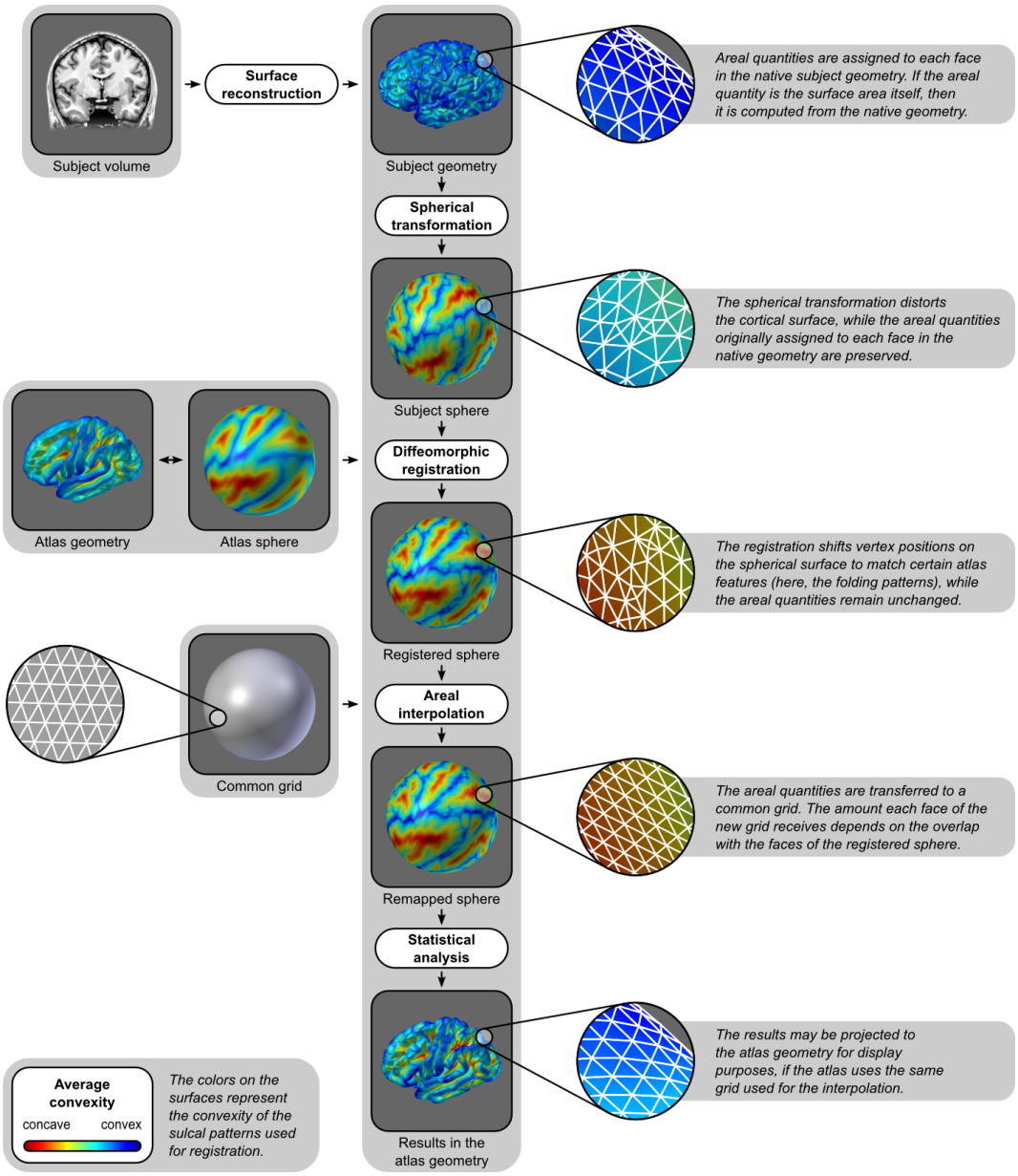


Figure 2.3: Diagram of the steps to analyse the cortical surface area. For clarity, the colors represent the convexity of the surface, as measured in the native geometry.

at a given vertex is summed and divided by three, producing a measure of the *area per vertex*, for facewise analysis it is the *area per face* that is measured and analysed. Since for each subject, each face in the native geometry has its corresponding face on the sphere, the value that represents area per face, as measured from the native geometry, can be mapped directly to the sphere, despite any areal distortion introduced by the spherical transformation.

Furthermore, since there is a direct mapping that is independent of the actual area in the native geometry, *any other quantity that is biologically areal can also be mapped to the spherical surface.* Perhaps the most prominent example is cortical volume (Section 2.2.3), although other cases of such quantities, that may potentially be better characterized as areal processes, are the extent of the neural activation as observed with functional MRI, the amount of amyloid deposited in Alzheimer's disease (Klunk et al., 2004; Clark et al., 2011), or simply the the number of cells counted from optic microscopy images reconstructed to a tri-dimensional space (Schormann and Zilles, 1998). Since areal interpolation (described below) conserves locally, regionally and globally the quantities under study, it allows accurate comparisons and analyses across subjects for measurements that are areal by nature, or that require mass conservation on the surface of the mesh representation.

2.2.2 Computation of surface area

The facewise areas in the mesh representation of the brain can be computed trivially: for a triangular face ABC with vertices $\mathbf{a} = [x_A\ y_A\ z_A]'$, $\mathbf{b} = [x_B\ y_B\ z_B]'$, and $\mathbf{c} = [x_C\ y_C\ z_C]'$, the area is $|\mathbf{u} \times \mathbf{v}|/2$, where $\mathbf{u} = \mathbf{a} - \mathbf{c}$, $\mathbf{v} = \mathbf{b} - \mathbf{c}$, $\times$ represents the cross product, and $|\bullet|$ represents the vector norm.

2.2.3 Volume as an areal quantity

Gray matter volume can be assessed using the partial volume effects of the gray matter in a per-voxel fashion using volume-based representations of the brain, such as in voxel-based morphometry (VBM Ashburner and Friston, 2000), or as the amount of tissue present between the gray and white surfaces in surface-based representations. Using the surface-based representation, software such as Free-

Surfer up to version 5.3.0 compute the volume by first calculating the area at each vertex as $^1/_3$ of the sum of the areas of all faces of the white surface that have that vertex in common, then multiplying that by the thickness at that vertex. Volume is also an areal quantity, that requires mass-conservative interpolation methods.

2.2.4 Registration

Registration to a common coordinate system is necessary to allow comparisons across subjects (Drury et al., 1996). The registration is performed by shifting vertex positions along the surface of the sphere until there is a good alignment between subject and template (target) spheres with respect to certain specific features, usually, but not necessarily, the cortical folding patterns. As the vertices move, the areal quantities assigned to the corresponding faces are also moved along the surface. The target for registration should be the less biased as possible in relation to the population under study (Thompson and Toga, 2002).

A registration method that produces a smooth, i.e. spatially differentiable, warp function enables the smooth transfer of areal quantities. A possible way to accomplish this is by using registration methods that are diffeomorphic. A diffeomorphism is an invertible transformation that has the elegant property that it and its inverse are both continuously differentiable (Christensen et al., 1996; Miller et al., 1997), minimising the risk of vagaries that would be introduced by the non-differentiability of the warp function.

Diffeomorphic methods are available for spherical meshes (Glaunès et al., 2004; Yeo et al., 2010a; Robinson et al., 2014), and here we adopt the Spherical Demons (SD) algorithm[3] (Yeo et al., 2010a). SD extends the Diffeomorphic Demons algorithm (Vercauteren et al., 2009) to spherical surfaces. The Diffeomorphic Demons algorithm is a diffeomorphic variant of the efficient, non-parametric Demons registration algorithm (Thirion, 1998). SD exploits spherical vector spline interpolation theory and efficiently approximates the regularization of the Demons objective function via spherical iterative smooting.

Methods that are not diffeomorphic by construction (Fischl et al., 1999b; Auzias

[3] Available at http://sites.google.com/site/yeoyeo02/software/sphericaldemonsrelease.

et al., 2013), but in practice produce invertible and smooth warps could, in principle, be used for registration for areal analyses. In the Evaluation section we study the performance of different registration strategies as well as the impact of the choice of the template.

2.2.5 Areal interpolation

After the registration, the correspondence between each face on the registered sphere and each face from the native geometry is maintained, and the surface area or other areal quantity under study can be transferred to a common grid, where statistical comparisons between subjects can be performed. The common grid is a mesh which vertices lie on the surface of a sphere. A geodesic sphere, which can be constructed by iterative subdivision of the faces of a regular icosahedron, has many advantages for this purpose, namely, ease of computation, edges of roughly similar sizes and, if the resolution is fine enough, edge lengths that are much smaller than the diameter of the sphere (see Section 2.2.7 for details). These two spheres, i.e. the registered, irregular spherical mesh (source), and the common grid (target), typically have different resolutions. The interpolation method must, nevertheless, *conserve the areal quantities*, globally, regionally and locally. In other words, the method has to be *pycnophylactic*[4] (Tobler, 1979). This is accomplished by assigning, to each face in the target sphere, the areal quantity of all overlapping faces from the source sphere, weighted by the fraction of overlap between them (Figure 2.4).

More specifically, let Q_i^S represent the areal quantity on the i-th face of the registered, source sphere S, $i = 1, 2, \ldots, I$. This areal quantity can be directly mapped back to the native geometry, and can be the area per face as measured in the native geometry, or any other quantity of interest that is areal by nature. Let the actual area of the same face on the source sphere be indicated by A_i^S. The quantities Q_i^S have to be transfered to a target sphere T, the common grid, which face areas are given by A_j^T for the j-th face, $j = 1, 2, \ldots, J$, $J \neq I$. Each target face j overlaps with K faces of the source sphere, being these overlapping faces

[4] From Greek *pyknos* = mass, density, and *phylaxis* = guard, protect, preserve, meaning that the method has to be mass conservative.

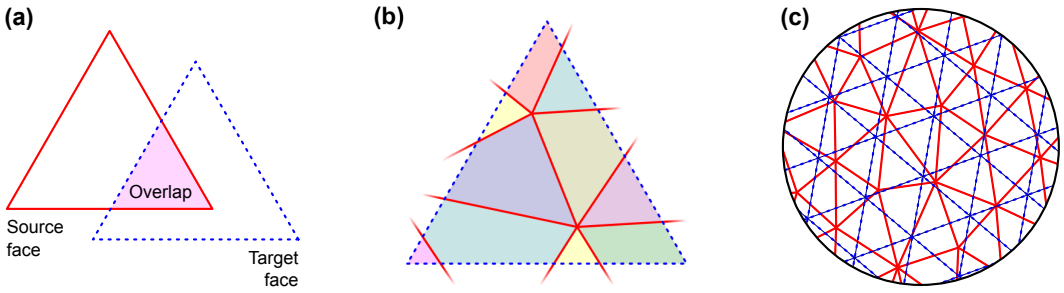


Figure 2.4: (*a*) Areal interpolation between a source and a target face uses the overlapping area as a weighting factor. (*b*) For a given target face, each overlapping source face contributes an amount of areal quantity. This amount is determined by the proportion between each overlapping area (represented in different colors) and the area of the respective source face. (*c*) The interpolation is performed at multiple faces of the target surface, so that the amount of areal quantity assigned to a given source face is conservatively redistributed across one or more target faces.

indicated by indices $k = 1, 2, \ldots, K$, and the area of each overlap indicated by A_k^O. The interpolated areal quantity to be assigned to the j-th target face is then given by:

$$Q_j^T = \sum_{k=1}^{K} \frac{A_k^O}{A_k^S} Q_k^S \tag{2.1}$$

Similar interpolation schemes have been devised to solve problems in geographic information systems (GIS) (Markoff and Shapiro, 1973; Goodchild and Lam, 1980; Flowerdew et al., 1991; Gregory et al., 2010). Surface models of the brain impose at least one additional challenge, which we address in the implementation (see Section 2.2.6). Differently than in other fields, where interpolation is performed over geographic territories that are small compared to Earth and, therefore, can be projected to a plane with acceptable areal distortion, here we have to interpolate across the whole sphere. Although other conservative interpolation methods exist for this purpose (Jones, 1999; Lauritzen and Nair, 2008; Ullrich et al., 2009), these methods either use regular latitude-longitude grids, cubed-spheres, or require a special treatment of points located above a certain latitude threshold to avoid singularities at the poles. These disadvantages may render these methods suboptimal for direct use in brain imaging.

2.2.6 Implementation

The areal interpolation for spheres is implemented in two parts. In the first, we compute inside of which source faces the target vertices are located, creating a lookup table to be used in the second part. This is the point-in-polygon problem found in vector graphics applications (Vince, 2005). Here we calculate the area of each source face, A_i^S, and the subsequent steps proceed iteratively for each face in the source. The barycentric coordinates of each candidate vertex in relation to the current face i is computed; if their sum equals to unity, the point is labelled as inside. However, to test if all vertices are inside every face would needlessly waste computational time. Moreover, since all points are on the surface of a sphere, the vertices in the target are never expected to be coplanar to the source triangular faces, so the test would always fail. The first problem is treated by testing only the vertices located within a bounding box defined, still in the 3D space, from the source face extreme coordinates. The second could naïvely be treated by converting the 3D Cartesian coordinates to 2D spherical coordinates, which allow a fast flattening of the sphere to the popular plate carrée cylindrical projection. However, latitude is ill-defined at the poles in cylindrical projections. Moreover, cylindrical projections introduce a specific type of deformation that is undesired here: straight lines on the surface (geodesic lines) are distorted. The solution we adopt is to rotate the Cartesian coordinate system so that the barycenter of the current source face lies at the point $[r\ 0\ 0]'$, where r is the radius of the source and target spheres. The barycenter is used for ease of calculation and for being always inside the triangle. After rotation, the current face and the nearby candidate target vertices are projected to a plane using the azimuthal gnomonic projection (Snyder, 1987), centered at the barycenter of the face. The point-in-polygon test can then be applied successfully. The key advantage of the gnomonic projection is that all geodesics project as straight lines, rather than loxodromic or other complex paths as with other projections, which would cause many target vertices to be incorrectly labelled. This projection can be obtained trivially after the rotation of the 3D Cartesian coordinate system as $\phi = y/x$ and $\theta = z/x$, where $[x\ y\ z]'$ are the 3D coordinates of the point being projected. A potential disadvantage of the gnomonic projection is the remarkable areal distortion for regions distant from

the center of the projection. Since in typical neuroimaging applications the source and target spheres are composed of a tessellation of approximately 3×10^6 faces, $A_i^S \ll 4\pi r^2$, and the distortion becomes negligible.

In the second part, the areal interpolation is performed, with the overlapping areas being calculated and used to weigh the areal quantity under study. The identification of intersections between two sets of polygons is also a well studied problem in vector graphics (Guibas and Seidel, 1987; Chazelle et al., 1994), which solution depends on optimally finding crossings between multiple line segments (Bentley and Ottmann, 1979; Chazelle and Edelsbrunner, 1992; Balaban, 1995). Most of the efficient available algorithms assume that the polygons are all coplanar; those that work in the surface of a sphere use coordinates expressed in latitude and longitude and require special treatment of the polar regions. The solution we adopt obviates these problems by first computing the area of each target face, A_j^T; the subsequent steps are performed iteratively for each face in the target sphere, using the azimuthal gnomonic projection, similarly as in the first part, but now centered at the barycenter of the current target face at every iteration. The areal quantities assigned to the faces in the target sphere are initialized as zero before the loop begins. If all three vertices of the current target face j lie inside the same source face k, as known from the lookup table produced in the first part, then to the current face the areal quantity given by $Q_j^T = Q_k^S A_j^T / A_k^S$ is assigned. Otherwise, the source faces that surround the target are examined to find overlaps. This is done by considering the edges of the current target face as vectors organised in counterclockwise orientation, and testing if the vertices of the candidate faces lie on the left, right or if they coincide with the edge. If all the three vertices of any candidate face are on the right of any edge, there is no overlap and the candidate face is removed from further consideration. If all the three vertices are on the left of all three edges, then the candidate source face is entirely inside the target, which has then its areal quantity incremented as $Q_j^T \leftarrow Q_j^T + Q_k^S$. The remaining faces are those that contain some vertices on the left and some on the right of the edges of the current, target face. The intersections between these source and target edges are computed and false intersections between edge extensions are ignored. A list containing the vertices for each candidate source face that are inside the target face (known for being on the left of the three target edges), the target vertices that are

inside each of the source faces (known from the lookup table) and the coordinates of the intersections between face edges, is used to compute the convex hull, using the Quickhull algorithm (Barber et al., 1996). The convex hull delimits the overlapping region between the current target face j and the candidate source face k, which area, A_k^O, is used to increment the areal quantity assigned to the target face as $Q_j^T \leftarrow Q_j^T + Q_k^S A_k^O / A_k^S$.

The algorithm runs in $\mathcal{O}(n)$ for n faces, as opposed to $\mathcal{O}(n^2)$ that would be obtained by naïve search. Nevertheless, the current implementation, that runs in Octave (Eaton et al., 2015) or MATLAB (The MathWorks Inc., 2015), a dynamically typed, interpreted language, requires about 24 hours to run in a computer with 2.66 GHz Intel Xeon processors.

2.2.7 Geodesic spheres and areal inequalities

The only required feature for the common grid used for the areal interpolation is that all its vertices must lie on the surface of a sphere. The algorithm we present in Section 2.2.6 requires further that all faces of the sphere are triangular and that all edges of all faces are much smaller than the radius, so that areal distortion is minimised when projecting to a plane.

A common grid that meet these demands is a sufficiently fine geodesic sphere. There are different ways to construct such a sphere (Kenner, 1976). One method is to subdivide each face of a regular polyhedron with triangular faces, such as the icosahedron, into four new triangles. The new vertices are projected to the surface of the (virtual) circumscribed sphere along its radius and the process is repeated recursively a number of times (Lauchner et al., 1969). For the n-th iteration, the number of faces is given by $F = 4^n F_0$, the number of vertices by $V = 4^n(V_0 - 2) + 2$, and the number of edges by $E = 4^n E_0$, where F_0, V_0 and E_0 are, respectively, the number of faces, vertices and edges of the polyhedron with triangular faces used for the initial subdivision. For the icosahedron, $F_0 = 20$, $V_0 = 12$ and $E_0 = 30$ (Figure 2.5*a*). For the analyses in this manuscript, we used $n = 7$, producing geodesic spheres with 327680 faces and 163842 vertices.

These faces, however, do not have identical edge lengths and areas (Kenner, 1976), even though the initial icosahedron was perfectly regular. This is import-

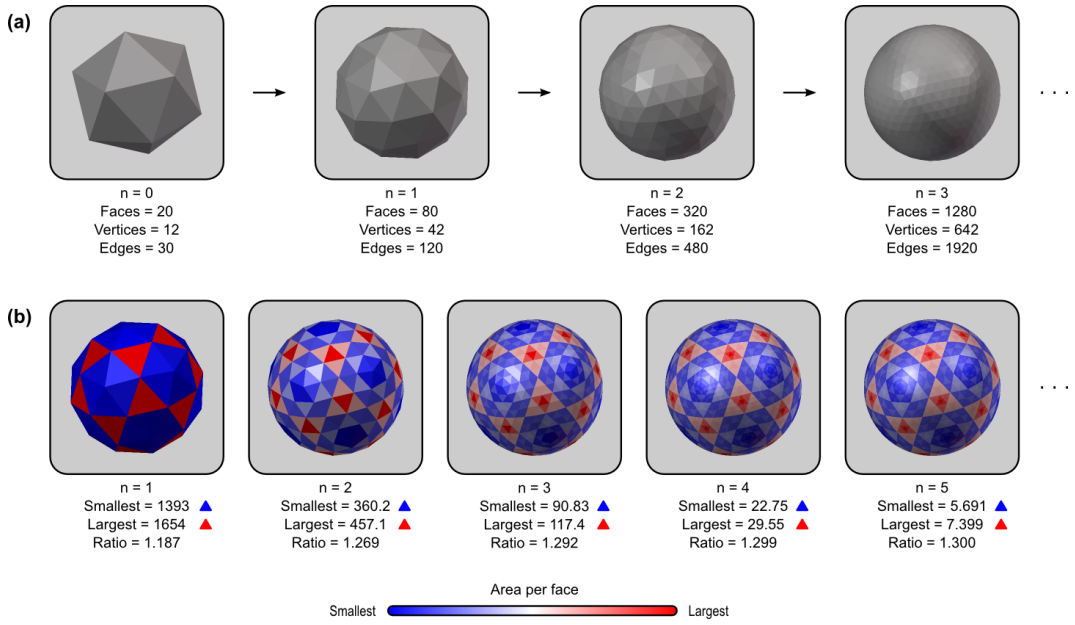


Figure 2.5: (*a*) The common grid can be a geodesic sphere produced from recursive subdivision of a regular icosahedron. At each iteration, the number of faces is quadruplied. (*b*) After the first iteration, however, the faces no longer have regular sizes, with the largest face being approximately 1.3 times larger than the smallest as n increases.

ant for areal interpolation, as larger faces on the target grid do overlap with more faces from the source surfaces, absorbing larger amounts of areal quantities, possibly causing confusion if one attempts to color-encode the interpolated image according to the actual areal quantities, in which case, geometric patterns such as in Figure 2.5*b* will become evident. Moreover, smoothing can cause quantities that are arbitrarily large or small due to face sizes to be blurred into the neighbors. Both potential problems can be addressed by multiplying the areal quantity at each face j, after interpolation, by a constant given by $4\pi r^2/(A_j^T F)$, where A_j^T is the area of the same face of the geodesic sphere, F is the number of faces, and r is the radius of the sphere.

2.2.8 Smoothing

Smoothing can be applied to alleviate residual discontinuities in the interpolated data due to unfavorable geometric configurations between faces of source and target spheres. For the purpose of smoothing, facewise data can be represented either by their barycenters, or converted to vertexwise (see Section 2.2.11 for a discussion

on how to convert), and should take into account differences on face sizes, as larger faces will tend to absorb more areal quantities (see Section 2.2.7). Smoothing can be applied using the moving weights method (Lombardi, 2002), defined as

$$\tilde{Q}_n^T = \frac{\sum_j Q_j^T G(g(\mathbf{x}_n, \mathbf{x}_j))}{\sum_j G(g(\mathbf{x}_n, \mathbf{x}_j))} \tag{2.2}$$

where $\tilde{Q}_n^T$ is the smoothed areal quantity at the n-th face, Q_j^T is the areal quantity assigned to each of the J faces of the same surface before smoothing, $g(\mathbf{x}_n, \mathbf{x}_j)$ is the scalar-valued distance along the surface between the barycenter $\mathbf{x}_n$ of the current face and the barycenter $\mathbf{x}_j$ of another face, and $G(g)$ is the Gaussian kernel.[5]

2.2.9 Conversion from facewise to vertexwise

Whenever it is necessary to perform analyses that include measurements taken at each vertex (such as some areal quantity versus cortical thickness) or when only software that can display vertexwise data is available (Section 2.2.11, it may be necessary to convert the areal quantities from facewise to vertexwise. The conversion can be done by redistributing the quantities at each face to their three constituent vertices. The areal values assigned to the faces that meet at a given vertex are summed, and divided by three, and reassigned to this vertex. Importantly, this procedure has to be done *after* the areal interpolation, since interpolation methods for vertexwise data are not appropriate for areal quantities, and *before* the statistical analysis, since the average of the results of the statistics of a test is not necessarily the same as the statistic for the average of the original data. It should also be observed that conversion from facewise to vertexwise data implies a loss of resolution to approximately half of the original and, therefore, should be performed only if resolution is not a concern and there is no other way to analyse, visualize, or present facewise data or results. The conversion does not change the underlying distribution, provided that the resolution of the initial mesh is sufficiently fine.

[5] As with other neuroimaging applications, smoothing after registration implies that the effective filter width is not spatially constant in native space, neither is the same across subjects. Smoothing on the sphere also contributes to different filter widths across space due to the deformation during spherical transformation.

2.2.10 Statistical analysis

After resampling to a common grid, the facewise data is ready for statistical analysis. The most straightforward method is to use the general linear model (GLM). The GLM is based on a number of assumptions, including that the observed values have a linear, additive structure, that the residuals of the model fit have the same variance and are normally distributed. When these assumptions are not met, a non-linear transformation can be applied, as long as the true, biological or physical meaning that underlies the observed data is preserved. In the Evaluation section, we show empirically that facewise cortical surface area is largely not normal. Instead, the distribution is skewed and can be better characterized as *lognormal*. A generic framework that can accommodate arbitrary areal quantities with skewed distributions is using a power transformation, such as the Box–Cox transformation (Box and Cox, 1964), which addresses possible violations of these specific assumptions, allied with permutation methods for inference (Holmes et al., 1996; Nichols and Hayasaka, 2003, see also Chapter 3) when the observations can be treated as independent, such as in most between-subject analysis.

The application of a statistical test at each face allows the creation of a statistical map and also introduces the multiple testing problem, which can also be addressed using permutation methods. These methods are known to allow exact significance values to be computed, even when distributional assumptions cannot be guaranteed, and also to facilitate strong control over family-wise error rate (FWER) if the distribution of the statistic under the null hypothesis is similar across tests. If not similar, the result is still valid, yet conservative. An alternative is to use a relatively assumption-free approach to address multiple testing, controlling instead the false discovery rate (FDR) (Benjamini and Hochberg, 1995; Genovese et al., 2002), which offers also weak control over FWER. Other approaches for inference, such as the Random Field Theory (RFT) for meshes (Worsley et al., 1999; Hagler et al., 2006) and the Threshold-Free Cluster Enhancement (TFCE) (Smith and Nichols, 2009) have potential to be used.

2.2.11 Presentation of results

To display results, facewise data can be projected from the common grid to the template geometry, which helps to visually identify anatomical landmarks and name structures. Projecting data from one surface to another is trivial as there is a one-to-one mapping between faces of the grid and the template geometry. The statistics and associated p-values can be encoded in colors, and a color scale can be shown along with the surface model.

However, the presentation of facewise data has conceptual differences in comparison with the presentation vertexwise data. For vertexwise data, each vertex cannot be directly colored, for being dimensionless. Instead, to display data per vertex, typically each face has its color interpolated according to the colors of its three defining vertices, forming a linear gradient that covers the whole face. For facewise data there is no need to perform such interpolation of colors, since the faces can be shown directly on the 3D space, each one in the uniform color that represents the underlying data. The difference is shown in Figure 2.6.

Interpolation of colors for vertexwise data should not be confused with the related, yet different concept of lightning and shading using interpolation. Both vertexwise and facewise data can be shaded to produce more realistic images. In Figure 2.6 we give an example of simple flat shading and shading based on linear interpolation of the lightning at each vertex (Gouraud, 1971).

Currently available software allow the presentation of color-encoded vertexwise data on the surface of meshes. However, only very few software applications can handle a large number of colors per 3D object, being one color per face. One example is Blender (Blender Foundation, Amsterdam, The Netherlands), which we used to produce the figures presented in this chapter. Another option, for instance, is to use low-level mesh commands in MATLAB (The MathWorks Inc., 2015), such as “patch”.

2.3 Evaluation

We illustrate the method using data from the Genetics of Brain Structure and Function Study, GOBS, a collaborative effort involving the Texas Biomedical Institute,

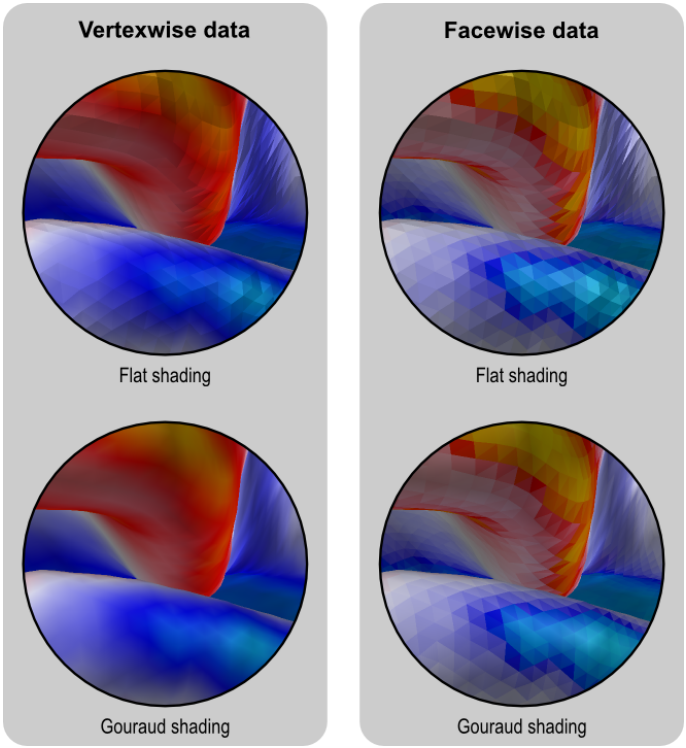


Figure 2.6: Differences between presentation of facewise and vertexwise data can be observed in this zoomed portion of the mesh representation of the cortex. Vertices are dimensionless and, to display vertexwise data, the faces have to be colored using linear interpolation. This is not necessary for facewise data, which can be shown directly in the uniform colors that represent the underlying data. In either case, the presentation can be improved by using a shading model, such as Gouraud in this example. Although the vertexwise presentation may be visually more appealing, it contains only half the resolution of the facewise image.

the University of Texas Health Science Center at San Antonio (UTHSCSA) and the Yale University School of Medicine. The participants are members of 42 families, and total sample size, at the time of the selection for this study, is 868 subjects. We randomly chose 84 subjects (9.2%), with the sparseness of the selection minimizing the possibility of drawing related individuals. The mean age of these subjects was 45.1 years, standard deviation 13.9, range 18.2–77.5, with 33 males and 51 females. All participants provided written informed consent on forms approved by each Institutional Review Board. The images were acquired using a Siemens MAGNETOM Trio 3 T system (Siemens AG, Erlangen, Germany) for 46 participants, or a Siemens MAGNETOM Trio/TIM 3 T system for 38 participants. We used a T_1-weighted, MPRAGE sequence with an adiabatic inversion contrast pulse with the following scan parameters: TE/TI/TR = 3.04/785/2100 ms, flip angle = 13°, voxel size (isotropic) = 0.8 mm. Each subject was scanned 7 (seven) times, consecutively, using the same protocol, and a single image was obtained by linearly coregistering these images and computing the average, allowing improvement over the signal-to-noise ratio, reduction of motion artifacts (Kochunov et al., 2006), and ensuring the generation of smooth, accurate meshes with no manual intervention. The image analysis followed the steps described in the Methods section, with some variation to test different registration strategies.

2.3.1 Registration

To isolate and evaluate the effect of registration, we computed the area per face after the spherical transformation[6] and registered each subject brain hemisphere to a common target using two different registration methods, the Spherical Demons (Yeo et al., 2010a) and the FreeSurfer registration algorithm (Fischl et al., 1999b)[7], each with and without a study-specific template as the target, resulting in four different variants. The study-specific targets for each of these methods were produced using the respective algorithms for registration, using all the 84 subjects

[6] Note that here the area was computed in the sphere with the aim of evaluating the registration method. For analyses of areal quantities, these quantities should be defined in the native geometry, as previously described.

[7] The software versions used were FS 5.0.0 and SD 1.5.1.

from the sample. The non-specific target was derived from an independent set of brain images of 40 subjects, the details of which have been described elsewhere (Desikan et al., 2006). Areal interpolation was used to resample the areal quantities to a common grid, a geodesic sphere produced by seven recursive subdivisions of a regular icosahedron.

The average area per face across subjects was computed after registration and interpolation to identify eventual systematic patterns of distortion caused by warping. This can be understood by observing that, as the vertices are shifted along the surface of the sphere, the faces that they define, and which carry areal quantities, are also shifted and distorted. The registration, therefore, causes displacement of areal quantities across the surface, which may accumulate on certain regions while other become depleted. Ideally, there should be no net accumulation when many subjects are considered and the target is unbiased with respect to the population under study. If pockets of accumulated or depleted areal quantities are present, this means that some regions are showing a tendency to systematically "receive" more areal quantities than others, which "donate" quantities. The average amount of area after the registration estimates this accumulation and, therefore, can be used as a measure of a specific kind of bias in the registration process, in which some regions consistently attract more vertices, resulting in these regions receiving more quantities. The result for this analysis is shown in Figure 2.7. Using default settings, SD caused less areal displacement across the surface, with less regional variation when compared to FS. The pattern was also more randomly distributed for SD, without spatial trends matching anatomical features, whereas FS showed a structure more influenced by brain morphology. Using a study specific template further helped to reduce areal shifts and biases. The subsequent analyses we present are based on the SD registration with a study-specific template.

2.3.2 Distributional characterization

To evaluate the normality for the cortical area at the white surface of the native geometry, we used the Shapiro–Wilk normality test (Shapiro and Wilk, 1965), implemented with the approximations for samples larger than 50 as described by Royston (1993). The test was applied after each hemisphere of the brain was re-

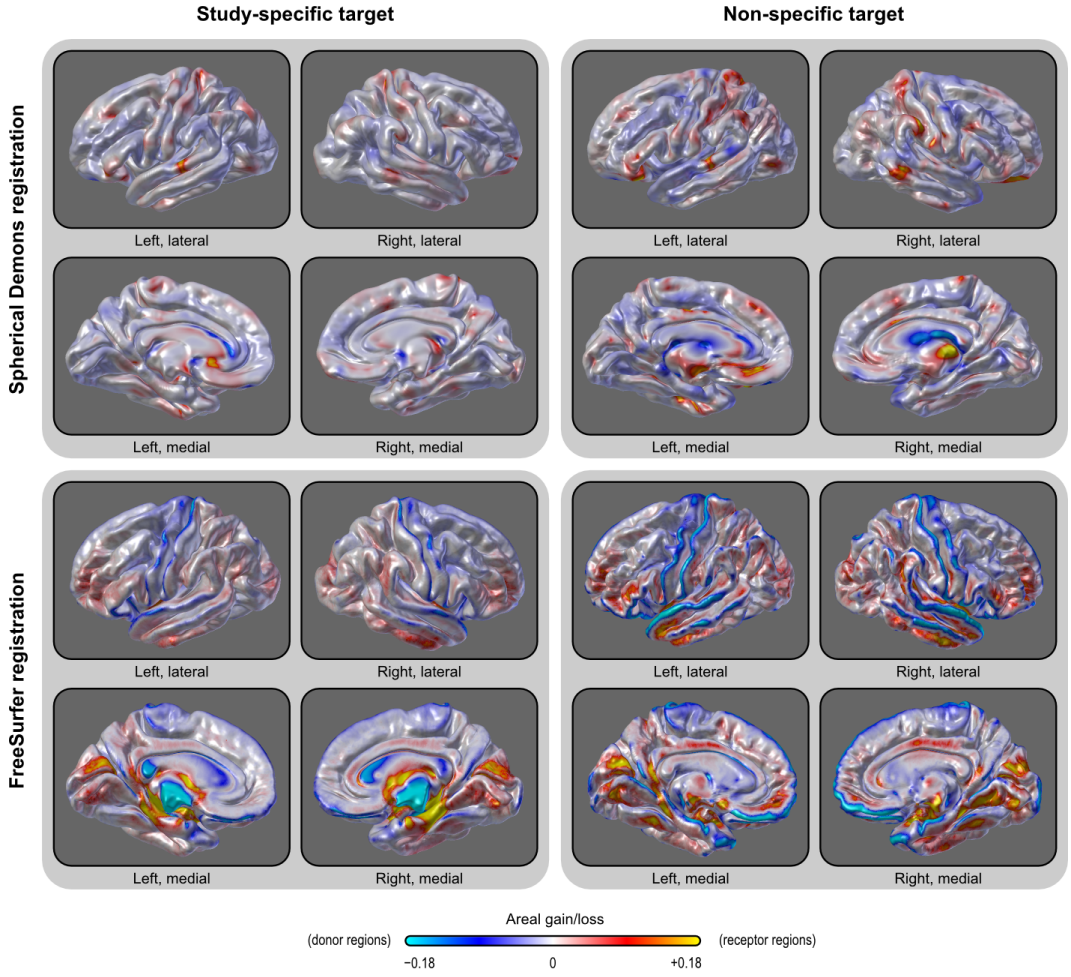


Figure 2.7: A study-specific template (target for the registration) caused less systematic accumulation of areal quantities across the brain when compared with a non-specific template. Using default parameters, areal accumulation wa less pronounced and unrelated to sulcal patterns using Spherical Demons in comparison with FreeSurfer registration. Gains and losses refer to the area per face that would be expected for areal quantities being redistributed with no bias, i.e. the zero corresponds to the average total surface area of all subjects, divided by the number of faces.

gistered to a study-specific template using the Spherical Demons and interpolated to the geodesic sphere using areal interpolation.

For the vast majority of the faces, the area of the white surface is *not* normally distributed (Figures 2.8–2.13). Instead, the lognormal distribution seems to be more appropriate to describe the data in most parts of the brain, with the test declaring a much larger number of faces as normally distributed after a simple logarithmic transformation. A log-transformation is a particular case of the Box–Cox transformation (Box and Cox, 1964). For a set of values $y = \{y_1, y_2, \ldots, y_n\}$, this transformation uses maximum-likelihood methods to seek a parameter λ that produces a transformed set $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n\}$ that approximately conforms to a normal distribution. The transformation is a piecewise function given by:

$$\tilde{y} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \ln y & (\lambda = 0) \end{cases} \tag{2.3}$$

Not surprisingly, the Box–Cox transformation rendered the data more normally distributed than a simple log-transformation. However, an interesting aspect of this transformation is that the parameter λ is allowed to vary continuously, and it approaches unity when the data is normally distributed, and zero if lognormally distributed, serving, therefore, as a summary metric of how normally or lognormally distributed the data is. Throughout most of the brain, λ is close to zero, although with a relatively wide variation (mode = -0.057, mean = -0.099, sd = 0.493 for the analysed dataset), indicating that, at the resolution used, the white surface cortical area can be better characterized across the surface as a gradient of skewed distributions, with the lognormal being the most common case. The same was observed for facewise data smoothed in the sphere after interpolation with FWHM = 10 mm (mode = -0.142, mean = -0.080, sd = 0.578).[8] Maps for the parameter λ are shown in Figure 2.14.

[8] For scale comparison, the sphere has radius fixed and set as 100 mm, such that the Gaussian filter has an HWHM (half width) = 1.59% of the geodesic distance between the barycenter of any face and its antipode.

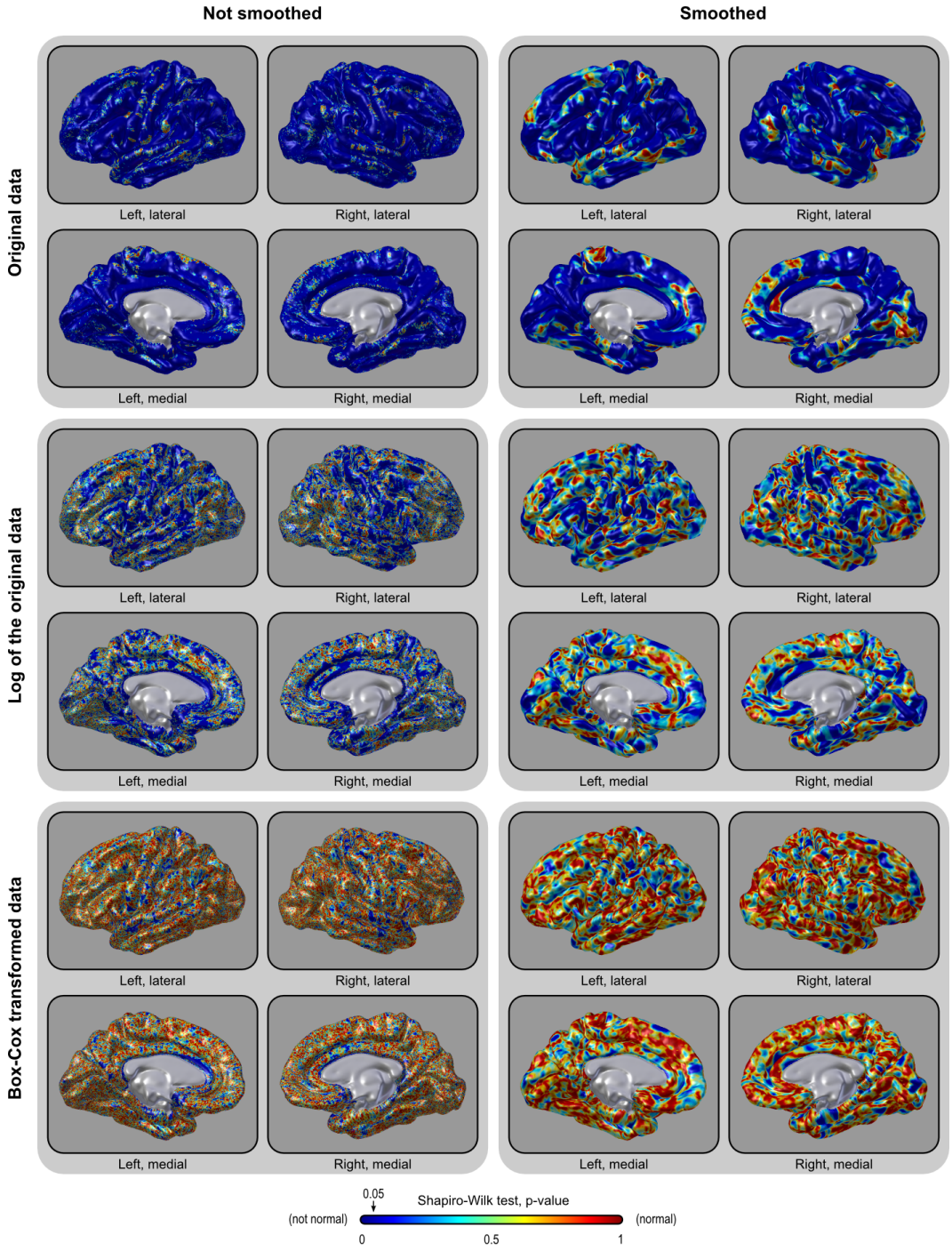


Figure 2.8: The area of the cortical surface is not normally distributed (*upper panels*). Instead, it is lognormally distributed throughout most of the brain (*middle panels*). A Box–Cox transformation can further improve normality (*lower panels*). The same pattern is present without (*left*) or with (*right*) smoothing (FWHM = 10 mm). Although normality is not an assumption for inference as proposed, it offers some advantages, as discussed in the text.

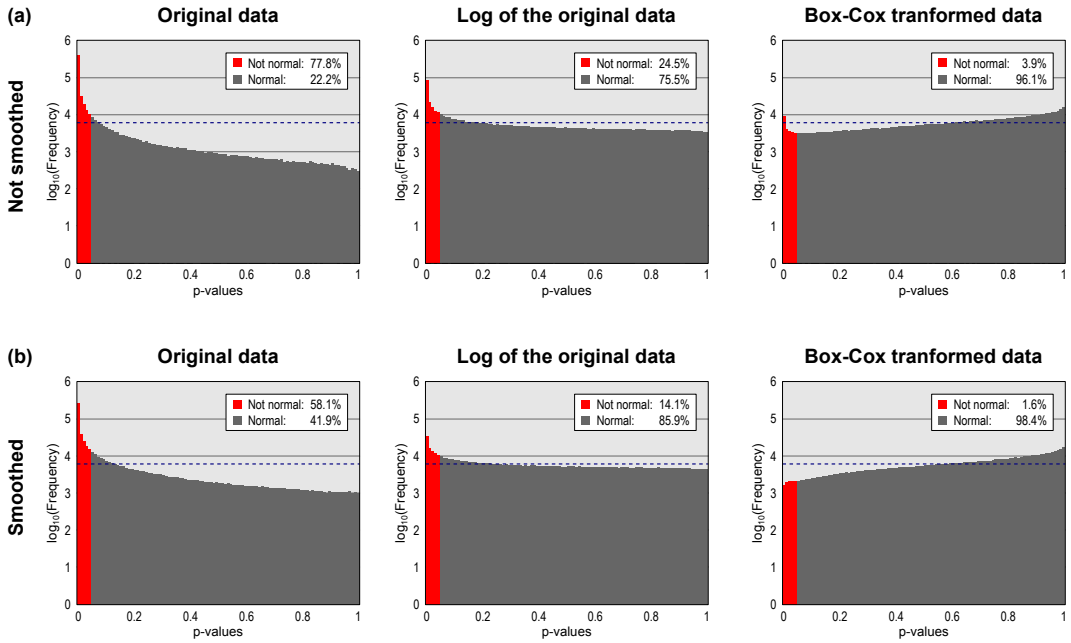


Figure 2.9: Distribution of the uncorrected p-values of the Shapiro–Wilk normality test. For normally distributed data, 5% of these tests are always expected to be declared as not normal with a significance level of $\alpha = 0.05$. Without transformation or smoothing, near 80% are found as not normal. Logarithmic and Box–Cox transformations render the data more normally distributed. Observe that the frequencies are shown in logscale. The dashed line (*blue*) is at the frequency that would be observed for uniformly distributed p-values.

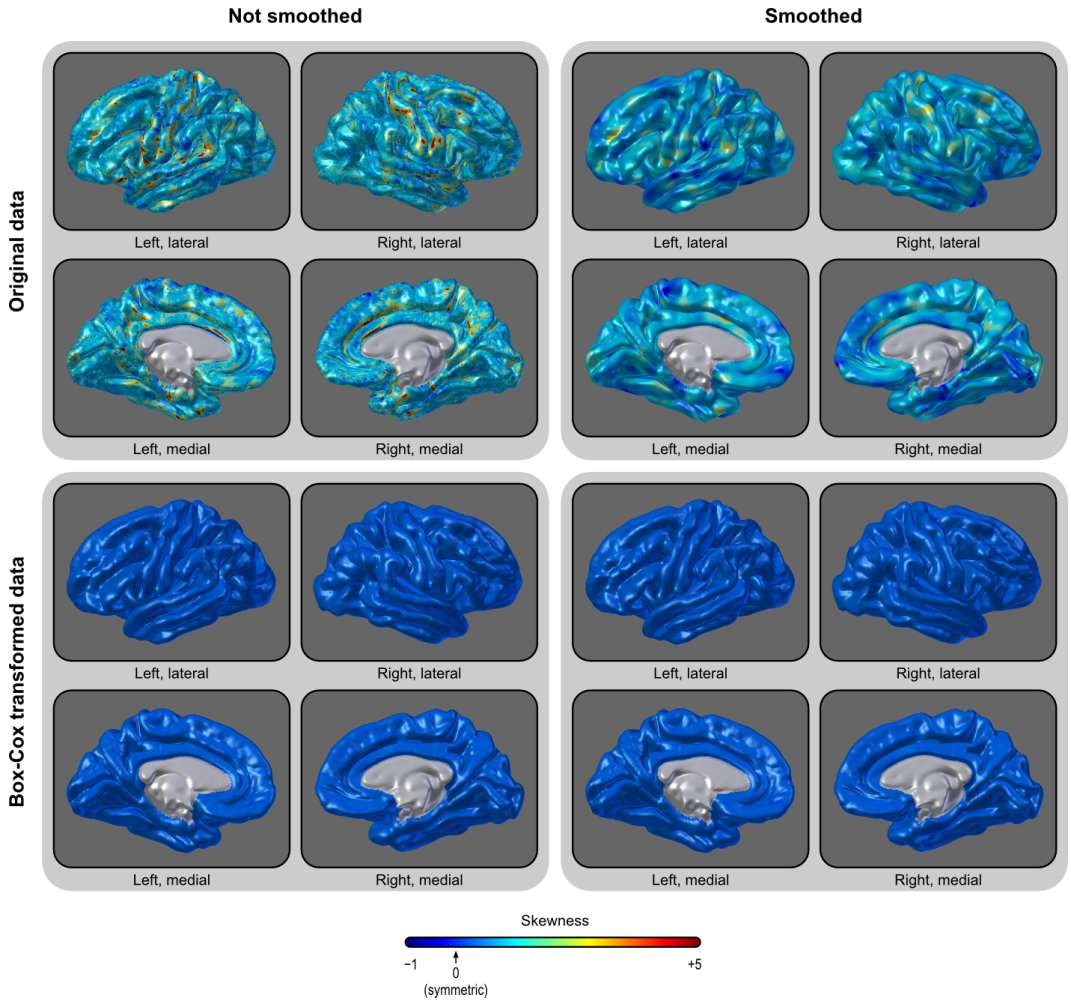


Figure 2.10: Maps of the skewness of the areal data, before and after the Box–Cox transformation, and with and without smoothing. The distribution is positively skewed (lognormal) throughout most of the brain, and the transformation successfully brings the data to symmetry (normality). The histograms are shown in Figure 2.11.

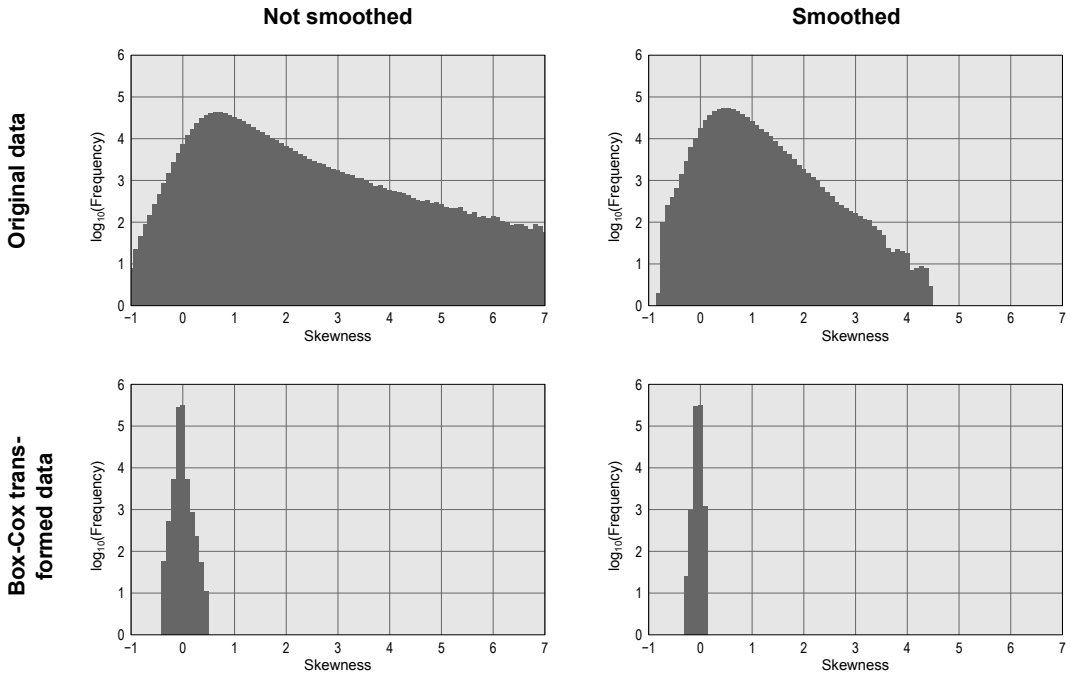


Figure 2.11: Histograms of the skewness of the areal data, before and after the Box–Cox transformation, and with and without smoothing. The distribution is positively skewed (lognormal) throughout most of the brain, and the transformation successfully brings the data to symmetry (normality). Note that the frequencies are shown in log scale. The corresponding maps are in Figure 2.10.

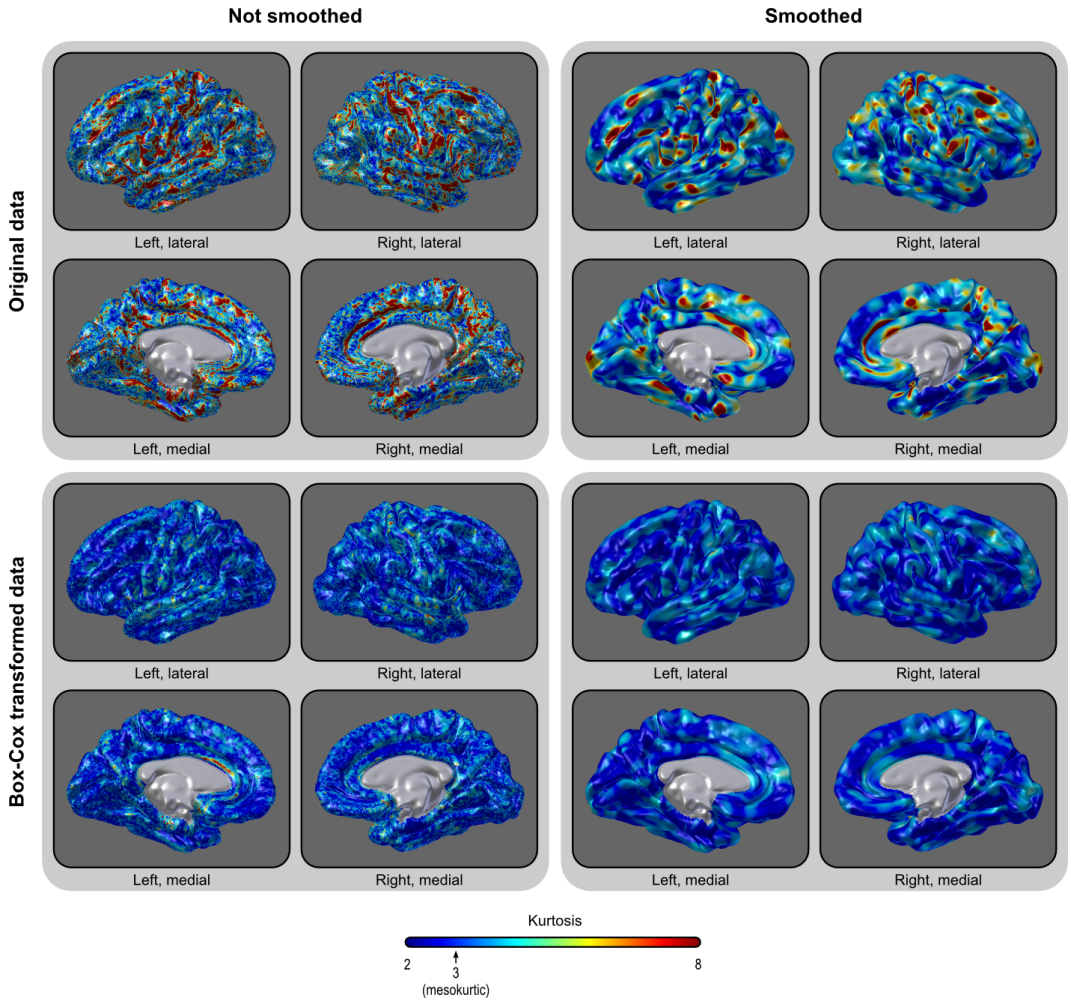


Figure 2.12: Maps of the kurtosis of the areal data, before and after the Box–Cox transformation, and with and without smoothing. The distribution is leptokurtic throughout most of the brain, and the transformation renders the kurtosis closer to the same value as for the normal distribution, i.e. closer to the value 3. The histograms are shown in Figure 2.13.

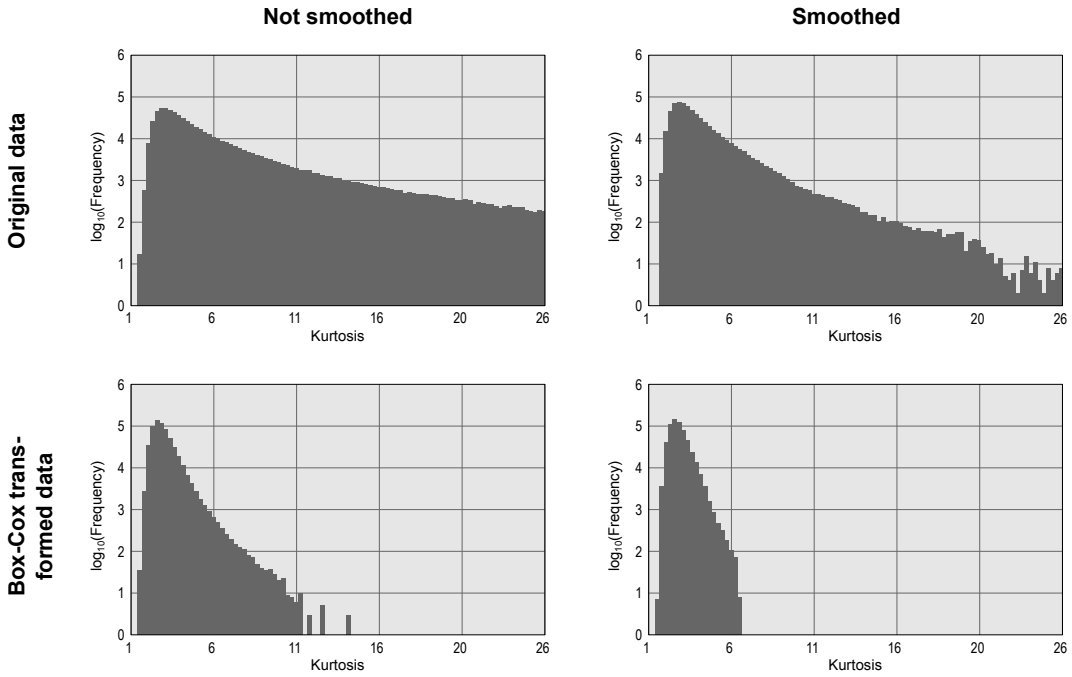


Figure 2.13: Histograms of the kurtosis of the areal data, before and after the Box–Cox transformation, and with and without smoothing. The distribution is leptokurtic throughout most of the brain, and the transformation renders the kurtosis closer to the same value as for the normal distribution, i.e. closer to the value 3. Note that the frequencies are shown in log scale. The corresponding maps are in Figure 2.12.

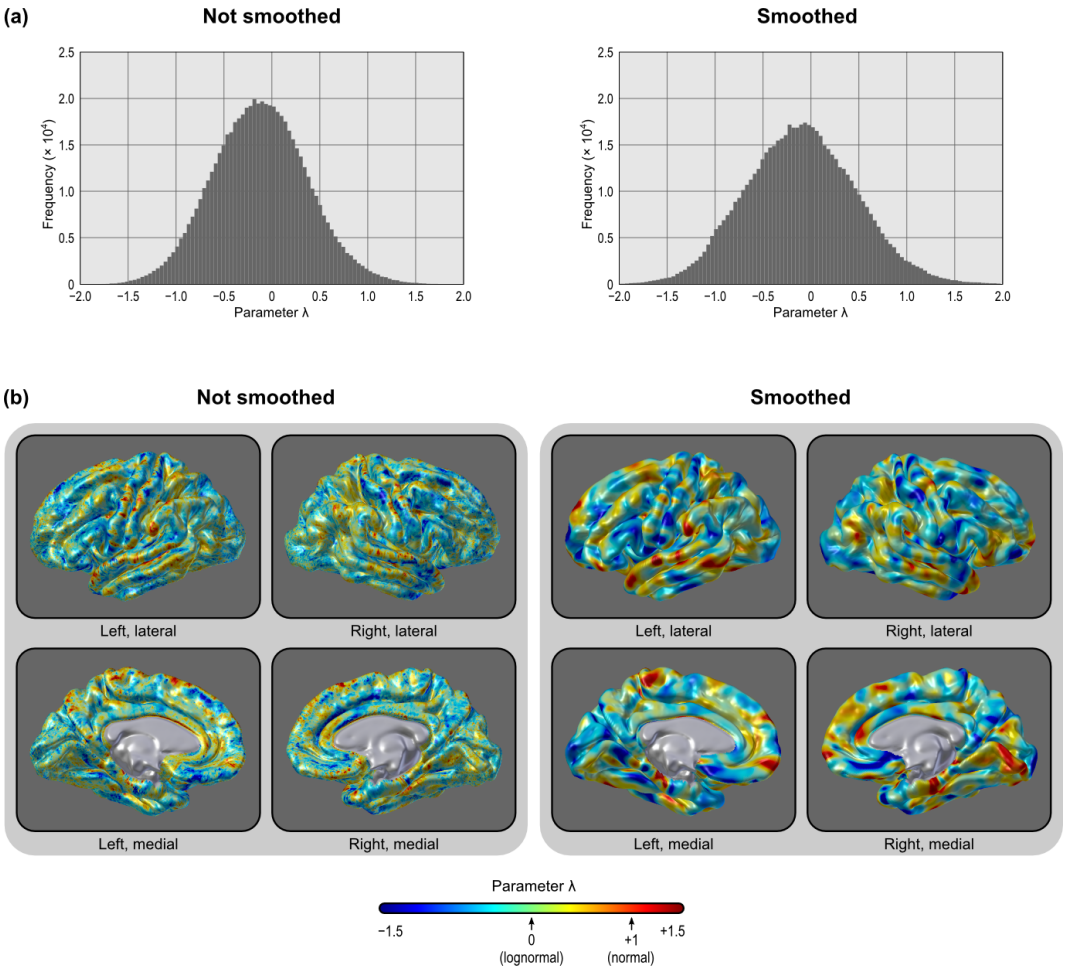


Figure 2.14: Distribution of the parameter λ of the Box–Cox transformation across the brain: (*a*) histogram and (*b*) spatial map. When λ approaches zero, the distribution of the underlying data is more lognormal.

2.3.3 Comparison with expansion/contraction methods

A number of studies have analysed what has been called expansion or contraction of the cortical surface when compared to a reference brain. Different studies adopted different operational definitions for what these terms would be [e.g. compare Joyner et al. (2009), Sun et al. (2009b), Hill et al. (2010)], and an unified approach has not been defined. Notwithstanding, the key difference between these methods and the proposed areal analysis is that, at the end of the processing pipeline, areal interpolation ensures the preservation of the amount (mass) of quantities, whereas these methods do not. Moreover, in the framework we present, a number of potential problems that may arise along the pipeline are explicitly addressed. These problems, along with the solutions we propose, are summarized in Table 2.1.

With a variety of expansion/contraction methods available, it is difficult to identify the best to which areal analysis could be compared. Here we retessellate the each subject brain in native space using the method described by Saad et al. (2004). The expansion/contraction method was implemented using the following steps: (1) From the native surface geometry, perform the spherical transformation; (2) Perform the spherical registration to a standard brain; (3) Treat the coordinates x, y and z of the vertices from the native geometry as three independent scalar fields over the registed sphere, and interpolate these values to the common spherical grid using barycentric interpolation[9]; (4) Use the interpolated coordinates, together with the same connectivity scheme between vertices as in the common grid, to construct a new model of the brain in a subject-specific geometry (Figure 2.15); (5) From this new model, compute the area per vertex and divide it by the area per vertex of the homologous point in the template. Call this measurement *expansion/contraction*; (6) Optionally, smooth this quantity.

[9] The three scalar fields can also be treated as a single vector field and the barycentric interpolation can be performed in a single step as

$$\begin{bmatrix} x_P \\ y_P \\ z_P \end{bmatrix} = \begin{bmatrix} x_A & x_B & x_C \\ y_A & y_B & y_C \\ z_A & z_B & z_C \end{bmatrix} \begin{bmatrix} \delta_A \\ \delta_B \\ \delta_C \end{bmatrix}$$

where x, y, z represent the coordinates of the triangular face ABC and of the interpolated point P, both in native geometry, and δ are the barycentric coordinates of P with respect to the same face after the spherical transformation.

Table 2.1: The proposed framework for areal analyses addresses a number of potential problems that may arise along the processing pipeline.

Processing step	Problem	Solution
Measurements assigned to vertices at the beginning of the analysis.	Vertices do not hold or convey the same spatial information as the original faces.	Analyse the faces directly.
Registration methods that not necessarily produce smooth and invertible warps.	Discontinuities on expansion or contraction that are not present in the actual brain.	Use diffeomorphic registration methods.
Interpolation based on points.	Areal quantities are not preserved at any scale (local, regional or global).	Use areal interpolation.
Use of a standard brain to compute the same measurement that is later analysed.	Results are interpretable only with respect to that same reference brain.	Measure and analyse absolute quantities, not relative to some reference.
Statistical analysis based on assumption of normality.	The local surface area follows a lognormal distribution.	Apply a data transformation. Use non-parametric methods.

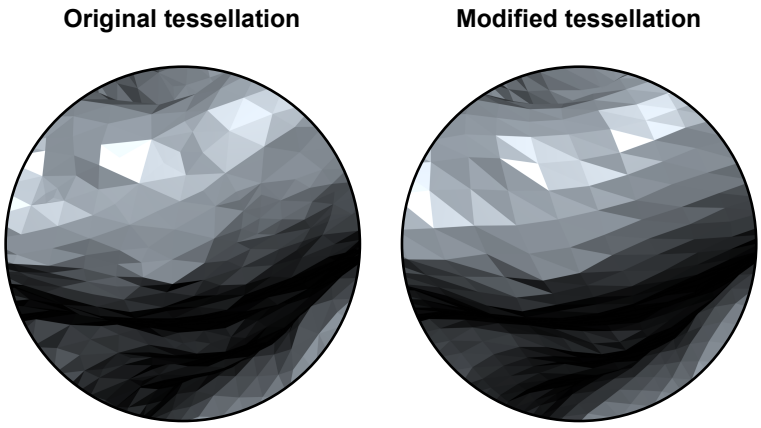


Figure 2.15: After barycentric interpolation of the coordinates in the surface of the sphere, a new, subject-specific retessellated model is constructed. Areas can be computed directly from the retessellated model and, once divided by the areas of the homologous vertices or faces of the reference brain, constitute the measurement of expansion/contraction.

For comparison with the expansion/contraction method, the original facewise area was converted to vertexwise, therefore halving the spatial resolution of the areal data (see Section 2.2.9). In this comparison, we addressed some of the problems presented in Table 2.1, namely, we registered using Spherical Demons, therefore ensuring smooth and invertible warps, and as target for registration, we used the study-specific template that produced the best results in Figure 2.7. Furthermore, the measurements were taken at the white surface, rather than the middle surface, as the last is more prone to be influenced by the cortical thickness. It is unclear if, when applicable, these aspects were taken care of in all the different studies that analysed some form of expansion/contraction.

After establishing an expansion/contraction procedure, there are still different ways to compare with areal analysis. The comparison can be made across subjects or across space, can be global or regional, and may or may not include smoothing. In Figure 2.16 we show that the average amount of area at each vertex did not produce a similar spatial map as the average expansion/contraction. Although the two methods follow remarkably different overall spatial patterns, when vertices across space were pooled together to produce a global measurement, they produced very similar results. Figure 2.17a shows the relationship between the global cortical surface area, computed from the sum of the area at each vertex, and a global measure of expansion computed by averaging the expansion/contraction at each vertex across space.[10] The correlation was very high and helps to validate both methods as a whole. Likewise, when each vertex was analysed separately, the correlation across subjects was also very high, as shown in Figure 2.18, with an R^2 above 0.9 throughout virtually the whole cortex. A spatial comparison of the average maps, on the other hand, showed a very poor relationship between both approaches, as shown in Figure 2.17b. When looking at each individual subject, rather than at the average, the correlation across space was still relatively low, albeit not as poor: for the 168 hemispheres analysed, we found an average linear $R^2 = 0.572$, sd $= 0.044$

[10] Note that an exact measurement of expansion/contraction relative to the template can be produced simply by dividing the global area in native geometry by the area of the template geometry. In this case, the points in Figure 2.17a would lie in a perfectly straight line, and nothing could be inferred about the relationship between regional variability on expansion estimates and global measurements.

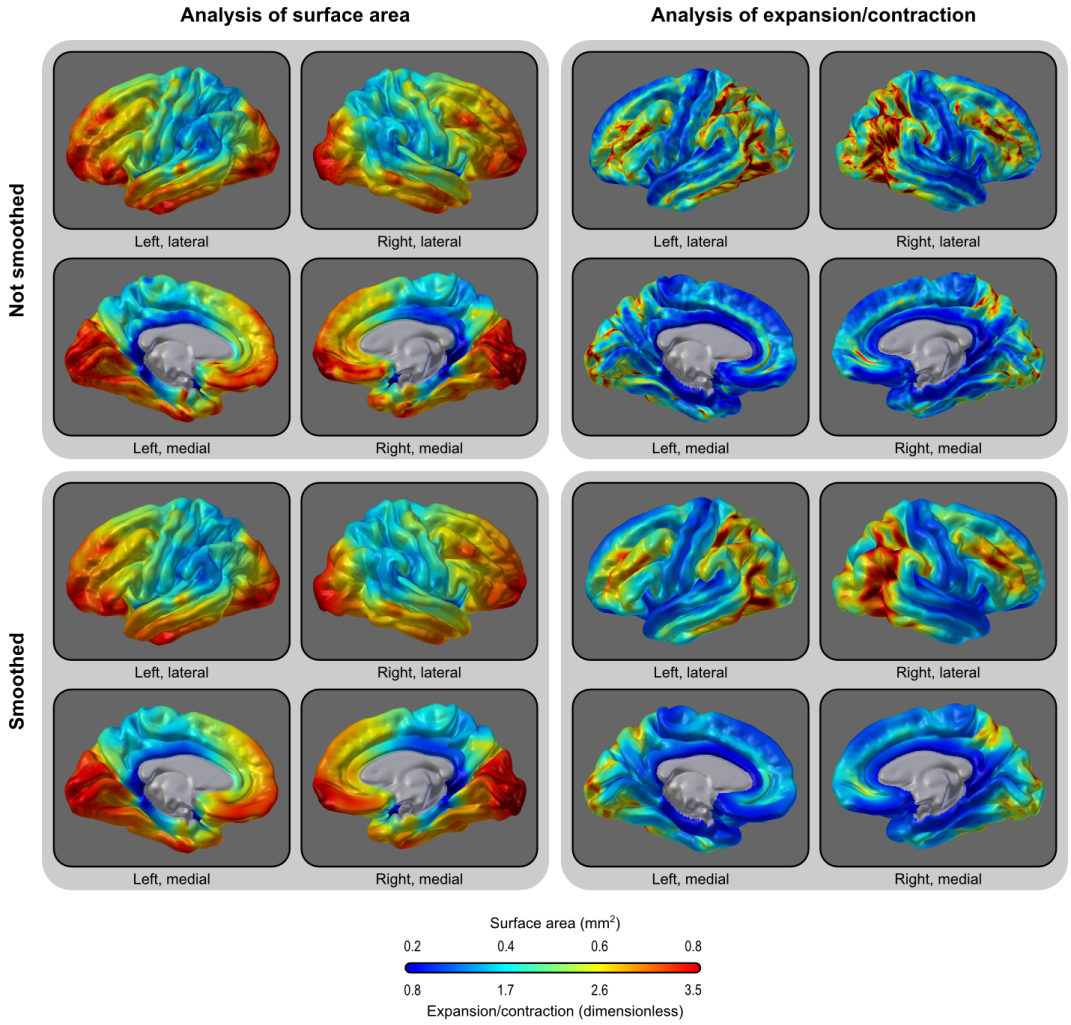


Figure 2.16: Average area (*left panels*) or expansion/contraction (*right panels*) per vertex, without (*upper panels*) and with smoothing (*lower panels*). Areal analyses and expansion/contraction differ across space. Smoothing has little global impact.

without smoothing, and $R^2 = 0.491$, sd $= 0.065$ after smoothing.

These results suggest that, if each vertex is analysed in isolation, analysis of surface area and analysis of expansion/contraction tend to produce similar results. This is the case, for instance, using mass univariate GLM-based approaches. However, for analysis that involve spatial information or that combine information across neighboring vertices, the results are expected to be rather dissimilar. The difference stems from the different units of measurement: areal analyses produce measurements in absolute units of area (e.g. mm^2), whereas expansion/contraction are relative to the a given reference. The result shown in Figure 2.18, left panel, also demonstrate, indirectly, that areas measured in the retessellated brain with

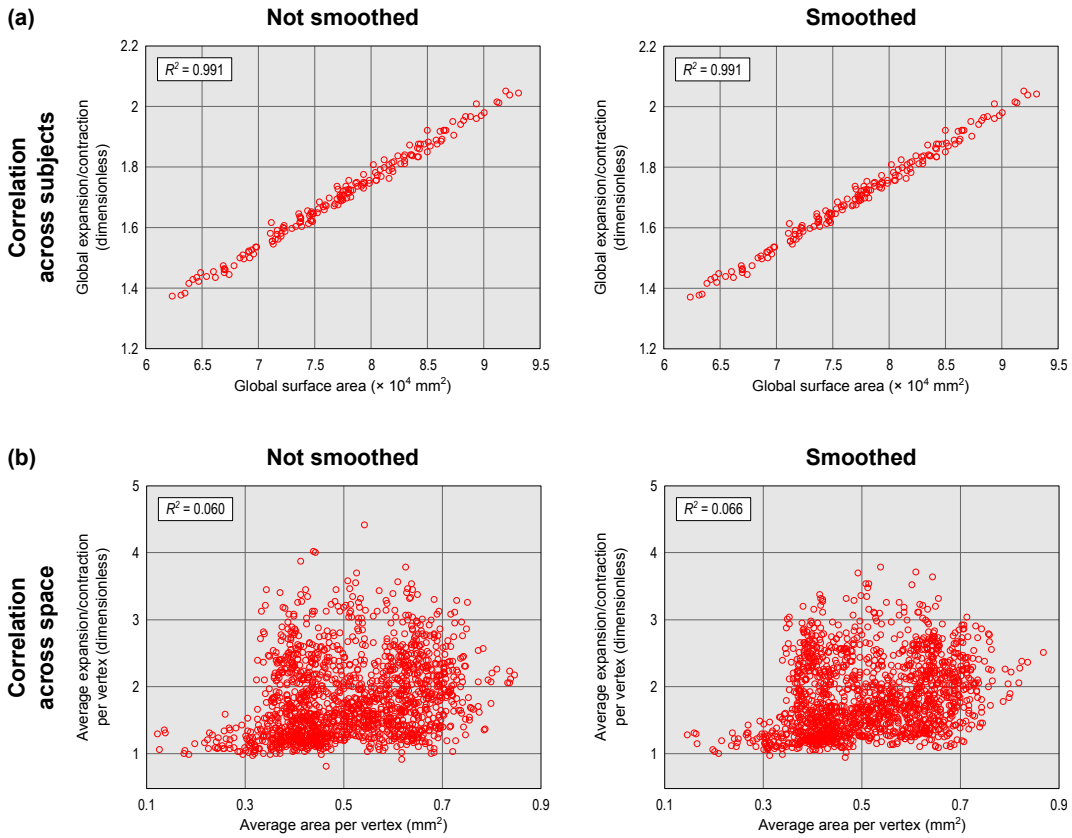


Figure 2.17: (*a*) The sum of the area per vertex correlates well with the average across space of the expansion/contraction at each vertex (i.e. equivalent to a weighed sum considering each vertex as having the same initial area) for the 168 hemispheres analysed. For the expansion/contraction, this is not the same as computing the ratio between the global surface area in native geometry and of the template, in which case, the result would be a perfectly straight line. The high correlation implies that the regional differences in general compensate each other to produce a similar global effect. (*b*) The correlation between average spatial maps across the 84 subjects, both hemispheres, is very poor between the methods. [Note that, for (*b*), attempts to simultaneously plot all the $>$ 300 thousand vertices would not produce meaningful plots in a small space; for this reason only 5% of the vertices were randomly selected for plotting. The R^2 were computed from all vertices and, for both (a) and (b), the value corresponds to the goodness of a linear fit.]

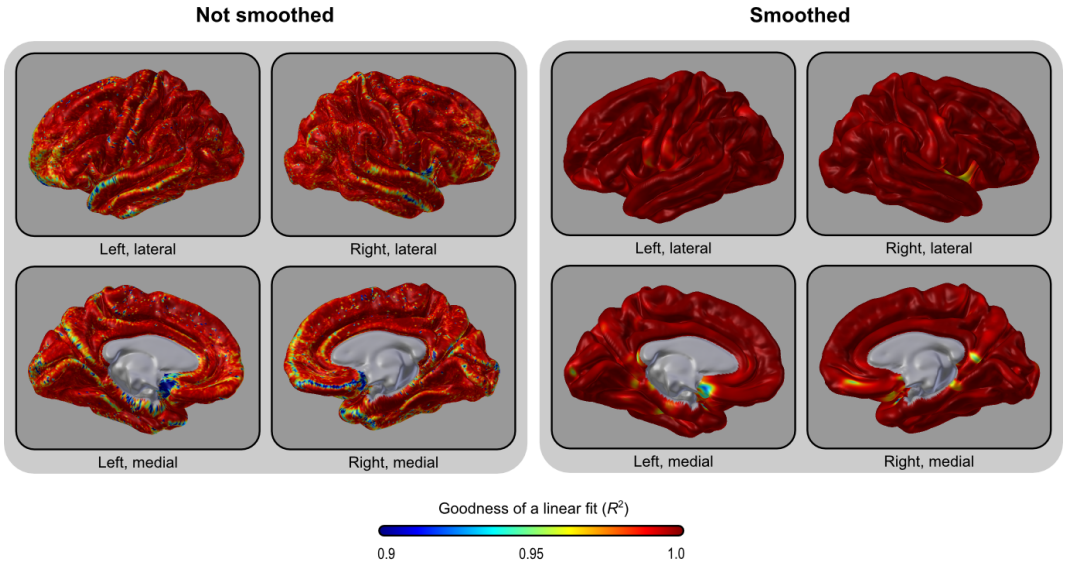


Figure 2.18: For each vertex, the linear relationship between areal analyses and expansion/contraction is very high across subjects, being above $R^2 = 0.90$ virtually across the whole cortex.

the resolution used correlate reasonably well with the areas obtained using areal interpolation, and so, have potential to be used as a fast approximation to areal interpolation (Section 2.2.6). Conversely, expansion/contraction measurements can be obtained after areal interpolation simply by dividing the area per face (or per vertex) by its homologous in the reference brain.

2.3.4 Validation and stability

Measurements of surface area are valid as long as the surface reconstruction from MR images produces accurate representations of the cortex. The suggested reconstruction method has been previously validated (Fischl and Dale, 2000), and is widely used for cortical thickness measurements. Comparison between subjects at the face level depends on good matching of homologies and the registration method we suggest has, likewise, been previously validated (Yeo et al., 2010a; Klein et al., 2010). As methods evolve, novel approaches for constructing surface representations of the cortex and for registration have potential to improve the overall quality of areal analyses. The validity of areal measurements other than surface area itself depend on each particular measurement technique.

To assess the stability across sessions and scanners, we compared MR images of

Table 2.2: Stability and robustness of measurements after registration and interpolation were assessed using three test images of the same subject. The measurements were similar across tests, with similar variability across space and high spatial correlation.

	Test A	**Test B**	**Test C**
Manufacturer and model	Siemens MAGNETOM Trio 3 T	Siemens MAGNETOM Trio/TIM 3 T	Siemens MAGNETOM Allegra 3 T
Sequence	MPRAGE	MPRAGE	MPRAGE
TE/TI/TR (ms)	3.04/785/2100	2.83/766/2200	2.74/900/2500
Flip angle	13°	13°	8°
Voxel size (mm)	$0.8 \times 0.8 \times 0.8$	$0.8 \times 0.8 \times 0.8$	$1.0 \times 1.0 \times 1.0$
Number of acquisitions	14	7	1
Scan date	March 2008	March 2008	April 2009
Cortical surface area (mm^2)	176,996	177,098	180,949
Not smoothed			
Average area per face (mm^2)	0.2937	0.2939	0.3003
Standard deviation	0.0938	0.0910	0.0962
Correlation with Test A	—	0.8218	0.7589
Correlation with Test B	0.8218	—	0.7863
Correlation with Test C	0.7589	0.7863	—
Smoothed (FWHM = 10 mm)			
Average area per face (mm^2)	0.2935	0.2936	0.3000
Standard deviation	0.0746	0.0712	0.0748
Correlation with Test A	—	0.9509	0.9074
Correlation with Test B	0.9509	—	0.9353
Correlation with Test C	0.9047	0.9353	—

the same subject acquired in three different sessions collected within a 1 year interval. The imaging protocol varied in terms of acquisition parameters, as well as the number of images used for averaging and improvements on signal and contrast-to-noise ratio. The details are summarized in Table 2.2. The estimated surface area produced by summing the facewise areas over the cortex after interpolation was very similar across tests, with the largest difference being 8.2% between Tests A and C (see Table 2.2), with or without smoothing. The mean and standard deviation for facewise areas were virtually identical across tests, again regardless of smoothing. The pairwise Pearson correlation between the tests for the facewise data after registration and interpolation was above 0.80 without smoothing, and above 0.90 after smoothing with FWHM = 10 mm, showing that the procedure is robust at the face level, even under different scanning conditions and degrees of smoothing.

2.4 Discussion

2.4.1 Registration

To be valid, facewise analyses rely on the assumption that microscopic structures can be localized using as reference the features that are identifiable with MRI and which drive the registration. Features with such localizing power are important because they help to ensure good overlap of homologous areas between subjects. Despite an implicit assumption already present in most imaging studies, only recently it has been demonstrated valid for some cytoarchitetonic areas when the references are the cortical folding patterns, even though for non-primary regions, the mismatch may still be substantial (Fischl et al., 2008, 2009; Hinds et al., 2008, 2009; Da Costa et al., 2011). Other features, some microscopic and detectable only under ultra-high field strengths (Augustinack et al., 2005; Bridge and Clare, 2006; Duyn et al., 2007; Kim et al., 2009), have the potential to be used as the reference as long as they are demonstrated to be markers of histologically or functionally defined areas, possibly replacing folding patterns altogether, or used to provide ancillary information. Myeloarchitectural features may be particularly useful for this application, for being resposible for most of the contrast observed with MRI (Geyer et al., 2011). Likewise, areal analyses can be conducted after registration based on features derived from functional MRI (Sabuncu et al., 2010).

Good matching of homologies, however, depends not only on the features used to guide the registration, but on the registration method itself. For facewise areal analyses, invertibility is necessary to prevent faces from being folded over others. In addition, methods that produce smooth warps are necessary to ensure that areal quantities are transferred smoothly, without abrupt variations. Such abrupt variations would only be acceptable if matching perfectly with areas where structure and/or function also changes abruptly. A spatial transformation that allows such perfect matching, however, cannot be obtained easily in practice, since these borders usually cannot be observed with current, conventional MRI methods, and importantly, since many of the differences between regions are subtle and the transitions are gradual. However, invertibility and smoothness, as guaranteed by diffeomorphic methods, albeit important, may not suffice. Our results show that

even methods that produce smooth varying warps can differ substantially with respect to how the areal quantities are shifted across the surface. It is possible that performance differences between these methods might be due to choices on regularization strategies and associated parameters (Fischl et al., 1999b; Yeo et al., 2010a), instigating further reseach on selections that may produce the most accurate results (Yeo et al., 2010b). Our experiments also demonstrate that the choice of the target used for registration affects the distortion in areal measurements.

2.4.2 Areal interpolation

Areal interpolation allows direct analysis of areal quantities in absolute values, including the surface area itself. This is because it is the areal quantity proper that is conservatively transfered between grids. Therefore, there is no need to apply corrections due to stretches or shrinkages, such as using the Jacobian of the transformation (Good et al., 2001), nor due to the choice of the parametrizable surface (Thompson and Toga, 1999). Moreover, the results are interpretable directly with regard to the actual amount of tissue or other measurement under study, rather than relative to concepts as expansion/contraction, which are always relative to a given reference, and can create difficulties in interpretation and comparison across studies, either due to different definitions adopted by different authors, or due to the the need of a reference brain. Notwithstanding, after areal interpolation, it continues to be possible to divide the areas by the areas of the homologous faces or vertices of a reference brain, and so, obtain an expansion/contraction measurement. Moreover, areal quantities that are not area itself can also be divided by the area of each face or vertex in native geometry, thus converting these quantities to densities if necessary.

It should be emphasized that, as with other interpolation strategies, areal interpolation is not perfectly reversible, i.e. once the cortical area of a subject is transferred to a different grid, remapping back to the subject surface will not produce locally identical results, although the global areal quantity is always conserved. This is because within each face, the areal quantity is implicitly assumed to be homogeneously distributed. This only becomes a problem if low resolution meshes are used and if several back-and-forth iterations are performed.

2.4.3 Statistical analysis of areal quantities

There are a number of reasons that go beyond purely methodological considerations to justify the transformation of the data before statistical analysis. Measurements related to biological morphology, such as lengths, areas, volumes or weights, are well known to follow non-normal distributions. If the diameter of a structure, for instance, is normally distributed, inevitably both its cross section and its surface area follow skewed distributions, whereas its volume follows an even more skewed (Kapteyn and van Uven, 1916; Gaddum, 1945). All these related measurements cannot be normally distributed simultaneously. The skewness is higher when the variability is relatively large in comparison to the measure of central tendency that best describes the data, such as the arithmetic or the geometric mean. If the non-normality is not considered, statistical models are likely to produce inaccurate results. In this scenario, a power transformation, such as the Box–Cox transformation, helps to identify subjacent, possibly causative, normally distributed effects.

The lognormal distribution, more specifically, is known to arise in a variety of biological processes. Of particular interest is the autocatalytic growth of tissue over time. The number of cells present on a tissue that grows in an unrestricted way can be given by the familiar formula $N = N_0 e^{ct}$, where N_0 is the initial number of cells, and t is the amount of time in which the cell grows under the circumstances represented by the constant c, a factor that incorporates a variety of influences, such as genetic and environmental. N will be lognormally distributed if either c or t are normally distributed (Koch, 1966; Limpert et al., 2001). The finding that the facewise cortical surface area follows mostly lognormal distributions may suggest that the method may capture these biological effects. Such interpretation can only subsist under the tenets of accurate and smooth registration.

From a statistical perspective, permutation methods do not rely on normality, rendering them appropriate in a variety of situations in which this assumption is not tenable. Nevertheless, the data should, still, undergo a transformation. As discussed above, the reason is not merely to conform to normality, although that comes as a bonus, but also to ensure that underlying biological effects, either multiplicative or proportionally dependent upon an initial value, can be treated as

additive in a linear model (Christensen, 2002). Areal quantities that are not the cortical surface area itself can, notwithstanding, be distributed differently, and the framework for statistical analysis outlined in the Methods section appears generic enough to accomodate a variety cases. The Box–Cox transformation has yet another advantage when used in combination with permutation methods under multiple testing conditions: the more stable variance after the transformation allows the distribution of the statistic under null hypothesis to become more similar across tests, allowing FWER to be controlled at a level closer to its nominal value using the distribution of the maximum statistic.

2.4.4 Box–Cox and log-normality

An interesting aspect related to the Box–Cox transformation is that here it was used as a metric to quantify how normal or log-normal the data are. This has clear applications in biology. Tissue growth that depends on cellular multiplication is exponential, with a constant factor that is often normally distributed, resulting in tissue size that is log-normally distributed (for a discussion, see McAlister, 1879; Koch, 1966, 1969). Measurements of the final tissue size in different individuals, however, does not perfectly conform to the log-normal due to external influences that may hinder tissue growth. Moreover, it is not always possible to measure the final amount of tissue that is the product of a single lineage of self-multiplicative cells. The combination of different cell lineages, each with their own growth rate, as well as external influences, tend to produce a distribution that is more normally (Gaussian) distributed. This appears to be the case of the cerebral cortex, with a *distribution gradient* between these two extremes of normality and log-normality. Estimating the parameter λ allows one to estimate also how closer to normal or log-normal certain measurements are. If $\lambda \approx 0$, the original data tend to be log-normally distributed, whereas if $\lambda \approx +1$, the data can be considered approximately more normally distributed.

2.4.5 Further developments and potential applications

Facewise analyses offer the possibility of studying surface area at a much finer scale than previously. This is a feature of interest in many research fields across

the neurosciences, as well as in medicine. Although the same applies to vertex-wise cortical thickness, thickness and area provide different and complementary insights into processes underlying the development of the brain and disorders (Voets et al., 2008; Winkler et al., 2010; Sanabria-Diaz et al., 2010).

Provided that the neurons in the cortex retain largely their same relative positions as the progenitor cells in the embryo (Rakic, 1988, 2009; Pierani and Wassef, 2009; Clowry et al., 2010), facewise comparison of surface area allows one to hypothesize about ontogenetic processes to the extent that they can be observed and localized with MRI, even long after the end of phases of massive tangential cellular proliferation. Until now, this kind of study could not be performed, either due to lack of methods to analyse cortical surface area without the constrains imposed by regions of interest, or due to inherent limitations of methods based on expansion or contraction.

The study of local cortical surface area offers, moreover, new possibilities for connectivity analyses, as the need for parcellations based on macroscopic anatomy is obviated. Indeed, the results of connectivity analyses are known to be influenced by the choice of the parcellation that define nodes of putative neuronal networks (Butts, 2009; Rubinov and Sporns, 2010). Notwithstanding, if a given set of regions is derived from a different method (Beckmann et al., 2009; Nelson et al., 2010), these can be directly associated with their corresponding surface-based areas or areal quantities by means of areal interpolation.

Another potential application is for genetic analyses. Given that cortical surface area and thickness are both heritable, yet genetically not correlated (Panizzon et al., 2009; Winkler et al., 2010), these traits, separately, can be used in a framework similar to voxelwise genome-wide association studies (VGWAS) (Stein et al., 2010). Identification of genes that influence surface area have potential to elucidate a myriad of developmental, neurologic and psychiatric disorders.

2.5 Chapter conclusion

We presented an interpolation method for between-subject analyses of cortical surface area. The method is also suitable for other quantities that are areal by nature and which require mass conservation (pycnophylactic property) during in-

terpolation and analysis. We demonstrated that, when the quantity under study is surface area itself, the distribution of the data does not follow a normal distribution, being instead better characterized as lognormal, and proposed a framework for statistical analysis and inference.

Chapter 3

Permutation inference

3.1 Introduction

The field of neuroimaging has continuously expanded to encompass an ever growing variety of experimental methods. From the early experiments using positron emission tomography (PET) and functional magnetic resonance imaging (FMRI), it is now often of interest to verify hypotheses using information obtained from, e.g. tensor-based morphometry (TBM), diffusion tensor imaging (DTI), cortical thickness and surface area, cerebral perfusion, as well as many others and variations and combinations of these. All these different modalities produce images that have different physical and biological properties, as well as different information content. Despite this variety, most of the strategies for statistical analysis constitute linear models, and can be formulated within the *general linear model* (GLM). The GLM is a simple, yet flexible framework of which many different types of analysis are particular cases (Scheffé, 1959; Searle, 1971; Christensen, 2002). The common strategy is to construct a plausible explanatory model for the observed data, estimate the parameters of this model, and compute a suitable statistic for testing of hypotheses on some or all of these parameters. The rejection or acceptance of a given hypothesis depends on the probablility of finding, due to chance alone, a statistic at least as high as the observed. Typically, but not necessarily, the hypothesis being tested is that one or more parameters are zero, being referred to as the *null hypothesis*.

If the parameters of the distribution of the statistic under the hypothesis being

tested, be it the null or not, is known, such probability can be ascertained using this distribution. In many particular cases, a mathematical expression describes the behaviour of the statistic as a function of these parameters, and this analytical representation of the distribution can be used for hypotheses testing as long as the data satisfies a certain set of requirements under which the distribution arises and can be recovered asymptotically. A conclusion based on these *parametric tests* will only be sound as long as the observed data possess these assumed stochastic properties, even if other methodological aspects are valid. Strategies that may be used when these assumptions are not met include, among others, the use of *non-parametric tests*.

Permutation tests are a class of non-parametric methods. They were pioneered by Fisher (1935a) and Pitman (1937a,b, 1938). Fisher demonstrated that the null hypothesis could be tested simply by observing, after permuting observations, how often the difference between means would exceed the difference found without permutation, and that for such test, no normality would be required. Pitman provided the first complete mathematical framework for permutation methods, although similar ideas, based on actually repeating an experiment many times with the experimental conditions being permuted, can be found even earlier (Peirce and Jastrow, 1884). Important theoretical and practical advances have been ongoing in the past decades (Pearson, 1937; Scheffé, 1943; Lehmann and Stein, 1949; Kempthorne, 1955; Edgington, 1995; Good, 2002, 2005; Westfall and Troendle, 2008; Pesarin and Salmaso, 2010a), and usage only became practical after the availability sufficient computing power (Efron, 1979).

In neuroimaging, permutation methods were first proposed by Blair and Karniski (1994) for electroencephalography, and later by Holmes et al. (1996) for positron-emission tomography, with the objective of allowing inferences while taking into account the multiplicity of tests. These early permutation approaches already accounted for the spatial smoothness of the image data. Arndt et al. (1996) proposed a permutation scheme for testing the omnibus hypothesis of whether two sets of images would differ. Structural magnetic resonance imaging (MRI) data were considered by Bullmore et al. (1999), who developed methods for omnibus, voxel and cluster-mass inference, controlling the expected number of false positives.

Single subject experiments from functional magnetic resonance imaging (FMRI)

presents a challenge to permutation methods, as serial autocorrelation in the time series violates the fundamental assumption needed for permutation, that of exchangeability (discussed below). Even though some early work did not fully account for autocorrelation (Belmonte and Yurgelun-Todd, 2001), other methods that accommodated the temporally correlated nature of the FMRI signal and noise were developed (Bullmore et al., 1996, 2001; Locascio et al., 1997; Brammer et al., 1997; Breakspear et al., 2004; Laird et al., 2004). Some of these methods use a single reference distribution constructed by pooling permutation statistics over space from a small number of random permutations, under the (untenable and often invalid) assumption of spatial homogeneity of distributions.

Nichols and Holmes (2002) provided a practical description of permutation methods for PET and multi-subject FMRI studies, but noted the challenges posed by nuisance variables. Permutation inference is grounded on *exchangeability* under the null hypothesis, that data can be permuted (exchanged) without affecting its joint distribution. However, if a nuisance effect is present in the model, the data cannot be considered exchangeable even under the null hypothesis. For example, if one wanted to test for sex differences while controlling for the linear effect of age, the null hypothesis is "male mean equals female mean", while allowing age differences; the problem is that, even when there is no sex effect, a possible age effect may be present, e.g., younger and older individuals being different, then the data are not directly exchangeable under this null hypothesis. Another case where this arises is in factorial experiments, where one factor is to be tested in the presence of another, or where their interaction is to be tested in the presence of main effects of either or both. Although permutation strategies for factorial experiments in neuroimaging were considered by Suckling and Bullmore (2004), a more complete general framework to account for nuisance variables is still missing.

In this chapter we review the statistical literature for the GLM with arbitrary designs and contrasts, emphasizing useful aspects, yet that have not been considered for neuroimaging, unify this diverse set of results into a single permutation strategy and a single generalised statistic, present implementation strategies for efficient computation and provide a complete algorithm, conduct detailed simulations and evaluations in various settings, and identify certain methods that generally outperforms others. We will not consider intrasubject (timeseries) FMRI data,

focusing instead on modelling data with independent observations or sets of repeated observations from independent subjects. We give examples of applications to common designs and discuss how these methods, originally intended for independent data, can in special cases be extended to repeated measurements and longitudinal designs.

3.2 Theory

3.2.1 Model and notation

At each spatial point (voxel, vertex or face) of an image representation of the brain, a general linear model (Searle, 1971) can be formulated and expressed as:

$$\mathbf{Y} = \mathbf{M}\boldsymbol{\psi} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\mathbf{Y}$ is the $N \times 1$ vector of observed data[1], $\mathbf{M}$ is the full-rank $N \times r$ design matrix that includes all effects of interest as well as all modelled nuisance effects, $\boldsymbol{\psi}$ is the $r \times 1$ vector of r regression coefficients, and $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of random errors. In permutation tests, the errors are not assumed to follow a normal distribution, although some distributional assumptions are needed, as detailed below. Estimates for the regression coefficients can be computed as $\hat{\boldsymbol{\psi}} = \mathbf{M}^{+}\mathbf{Y}$, where the superscript ($^{+}$) denotes the Moore–Penrose pseudo-inverse. Our interest is to test the null hypothesis that an arbitrary combination (contrast) of some or all of these parameters is equal to zero, i.e., $\mathcal{H}_0 : \mathbf{C}'\boldsymbol{\psi} = \mathbf{0}$, where $\mathbf{C}$ is a $r \times s$ full-rank matrix of s contrasts, $1 \leqslant s \leqslant r$.

For the discussion that follows, it is useful to consider a transformation of the model in Equation 3.1 into a partitioned one:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\mathbf{X}$ is the matrix with regressors of interest, $\mathbf{Z}$ is the matrix with nuisance re-

[1] While we focus on univariate data, the general principles presented can be applied to multivariate linear models.

gressors, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are respectively the vectors of regression coefficients. Even though such partitioning is not unique, it can be defined in terms of the contrast $\mathbf{C}$ in a way that inference on $\boldsymbol{\beta}$ is equivalent to inference on $\mathbf{C}'\boldsymbol{\psi}$, as described in Section 3.2.2.

As the models expressed in Equations 3.1 and 3.2 are equivalent, their residuals are the same and can be obtained as $\hat{\boldsymbol{\epsilon}} = \mathbf{R}_\mathbf{M}\mathbf{Y}$, where $\mathbf{R}_\mathbf{M} = \mathbf{I} - \mathbf{H}_\mathbf{M}$ is the residual-forming matrix, $\mathbf{H}_\mathbf{M} = \mathbf{M}\mathbf{M}^+$ is the projection ("hat") matrix, and $\mathbf{I}$ is the $N \times N$ identity matrix. The residuals due to the nuisance alone are $\hat{\boldsymbol{\epsilon}}_\mathbf{Z} = \mathbf{R}_\mathbf{Z}\mathbf{Y}$, where $\mathbf{R}_\mathbf{Z} = \mathbf{I} - \mathbf{H}_\mathbf{Z}$, and $\mathbf{H}_\mathbf{Z} = \mathbf{Z}\mathbf{Z}^+$. For permutation methods, an important detail of the linear model is the non-independence of residuals, even when errors $\boldsymbol{\epsilon}$ are independent and have constant variance, a fact that contributes to render these methods approximately exact. For example, in that setting $\mathrm{E}\left(\mathrm{Var}(\hat{\boldsymbol{\epsilon}}_\mathbf{Z})\right) = \mathbf{R}_\mathbf{Z} \neq \mathbf{I}$. The commonly used F statistic can be computed as (Christensen, 2002):

$$\begin{aligned} F &= \frac{\hat{\boldsymbol{\psi}}'\mathbf{C}\left(\mathbf{C}'(\mathbf{M}'\mathbf{M})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}'\hat{\boldsymbol{\psi}}}{\mathrm{rank}\left(\mathbf{C}\right)} \Bigg/ \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{N - \mathrm{rank}\left(\mathbf{M}\right)} \\ &= \frac{\hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}}{\mathrm{rank}\left(\mathbf{C}\right)} \Bigg/ \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n - \mathrm{rank}\left(\mathbf{X}\right) - \mathrm{rank}\left(\mathbf{Z}\right)} \end{aligned} \tag{3.3}$$

When $\mathrm{rank}(\mathbf{C}) = 1$, $\hat{\boldsymbol{\beta}}$ is a scalar and the Student's t statistic can be expressed as a function of F as $t = \mathrm{sign}(\hat{\boldsymbol{\beta}})\sqrt{F}$.

Choice of the statistic In non-parametric settings we are not constrained to the F or t statistics and, in principle, any statistic where large values reflect evidence against the null hypothesis could be used. This includes regression coefficients or descriptive statistics, such as differences between medians, trimmed means or ranks of observations (Ernst, 2004). However, the statistic should be chosen such that it does not depend on the scale of measurement or on any unknown parameter. The regression coefficients, for instance, whose variance depends both on the error variance and on the colinearity of that regressor with the others, are not in practice a good choice, as certain permutation schemes alter the colinearity among regressors (Kennedy and Cade, 1996). Specifically with respect to brain imaging, the correction for multiple testing (discussed later) requires that the statistic has a distribution that is spatially homogeneous, something that regression coefficients

cannot provide. In parametric settings, statistics that are independent of any unknown parameters are called *pivotal statistics*. Statistics that are pivotal or asymptotically pivotal are appropriate and facilitate the equivalence of the tests across the brain, and their advantages are well established for related non-parametric methods (Hall and Wilson, 1991; Westfall and Young, 1993). Examples of such pivotal statistics include the Student's t, the F ratio, as well as most other statistics used to construct confidence intervals and to compute p-values in parametric tests. We will return to the matter of pivotality when discussing exchangeability blocks, and the choice of an appropriate statistic for these cases.

p-values Regardless of the choice of the test statistic, p-values offer a common measure of evidence against the null hypothesis. For a certain test statistic T, which can be any of those discussed above, and a particular observed value T_0 of this statistic after the experiment has been conducted, the p-value is the probability of observing, by chance, a test statistic equal or larger than the one computed with the observed values, i.e., p-value $= P(T \geqslant T_0|\mathcal{H}_0)$. Although here we only consider one-sided tests, where evidence against $\mathcal{H}_0$ corresponds to larger values of T_0, two-sided or negative-valued tests and their p-values can be similarly defined. In parametric settings, under a number of assumptions, the p-values can be obtained by referring to the theoretical distribution of the chosen statistic (such as the F distribution), either through a known formula, or using tabulated values. In non-parametric settings, these assumptions are avoided. Instead, the data are randomly shuffled, many times, in a manner consistent with the null hypothesis. The model is fitted repeatedly once for every shuffle, and for each fit a new realisation of the statistic, T_j^*, is computed, being j a permutation index. An empirical distribution of T^* under the null hypothesis is constructed, and from this null distribution a p-value is computed as:

$$\frac{1}{J}\sum_{j=1}^{J} I(T_j^* \geqslant T_0) \tag{3.4}$$

where J is the number of shufflings performed, and $I(\cdot)$ is the indicator function. From this it can be seen that the non-parametric p-values are discrete, with each possible p-value being a multiple of $1/J$. The permutation distribution should

include the observed statistic without permutation (Edgington, 1969; Phipson and Smyth, 2010), and thus the smallest possible p-value is $1/J$, not zero. Even though such p-value is biased towards conservativeness (Pesarin and Salmaso, 2010a), it is the value that should be used in scientific research.

3.2.2 Model partitioning

The permutation methods discussed in this chapter require that the design matrix $\mathbf{M}$ is partitioned into effects of interest and nuisance effects. Such partitioning is not unique, and schemes can be as simple as separating apart the columns of $\mathbf{M}$ as $[\mathbf{X}\ \mathbf{Z}]$, with $\boldsymbol{\psi} = [\boldsymbol{\beta}'\ \boldsymbol{\gamma}']'$ (Guttman, 1982). More involved strategies can, however, be devised to obtain some practical benefits. One such partitioning is to define $\mathbf{X} = \mathbf{MDC}\left(\mathbf{C}'\mathbf{DC}\right)^{-1}$ and $\mathbf{Z} = \mathbf{MDC}_v\left(\mathbf{C}_v'\mathbf{DC}_v\right)^{-1}$, where $\mathbf{D} = (\mathbf{M}'\mathbf{M})^{-1}$, $\mathbf{C}_v = \mathbf{C}_u - \mathbf{C}(\mathbf{C}'\mathbf{DC})^{-1}\mathbf{C}'\mathbf{DC}_u$, and $\mathbf{C}_u$ has $r - \text{rank}\left(\mathbf{C}\right)$ columns that span the null space of $\mathbf{C}$, such that $[\mathbf{C}\ \mathbf{C}_u]$ is a $r \times r$ invertible, full-rank matrix (Beckmann et al., 2001; Smith et al., 2007). This partitioning has a number of features: $\hat{\boldsymbol{\beta}} = \mathbf{C}'\hat{\boldsymbol{\psi}}$, $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \mathbf{C}'\widehat{\text{Cov}}(\hat{\boldsymbol{\psi}})\mathbf{C}$, i.e., estimates and variances of $\boldsymbol{\beta}$ for inference on the partitioned model correspond exactly to the same inference on the original model, $\mathbf{X}$ is orthogonal to $\mathbf{Z}$, and $\text{span}(\mathbf{X}) \cup \text{span}(\mathbf{Z}) = \text{span}(\mathbf{M})$, i.e., the partitioned model spans the same space as the original. This is the partitioning strategy used in this chapter, and used in the randomise algorithm.

Another useful partitioning scheme, derived by Ridgway (2009), defines $\mathbf{X} = \mathbf{M}(\mathbf{C}^+)'$ and $\mathbf{Z} = \mathbf{M} - \mathbf{MCC}^+$. As with the previous strategy, the parameters of interest in the partitioned model are equal to the contrast of the original parameters. A full column rank nuisance partition can be obtained from the singular value decomposition (SVD) of $\mathbf{Z}$, which will also provide orthonormal columns for the nuisance partition. Orthogonality between regressors of interest and nuisance can be obtained by redefining the regressors of interest as $\mathbf{R}_\mathbf{Z}\mathbf{X}$.

3.2.3 Permutations and exchangeability

Perhaps the most important aspect of permutation tests is the manner in which data are shuffled under the null hypothesis. It is the null hypothesis, together with assumptions about exchangeability, that determines the permutation strategy. Let

the j-th permutation be expressed by $\mathbf{P}_j$, a $N \times N$ permutation matrix, a matrix that has all elements being either 0 or 1, each row and column having exactly one 1 (Figure 3.1a). Pre-multiplication of a matrix by $\mathbf{P}_j$ permutes its rows. We denote $\mathcal{P} = \{\mathbf{P}_j\}$ the set of all permutation matrices under consideration, indexed by the subscript j. We similarly define a sign flipping matrix $\mathbf{S}_j$, a $N \times N$ diagonal matrix whose non-zero elements consist only of $+1$ or -1 (Figure 3.1b). Pre-multiplication of a matrix by $\mathbf{S}_j$ implements a set of sign flips for each row. Likewise, $\mathcal{S} = \{\mathbf{S}_j\}$ denotes the set of all sign flipping matrices under consideration. We consider also both schemes together, where $\mathbf{B}_j = \mathbf{P}_{j'}\mathbf{S}_{j''}$ implements sign flips followed by permutation; the set of all possible such transformations is denoted as $\mathcal{B} = \{\mathbf{B}_j\}$. Throughout the chapter, we use generic terms as *shuffling* or *rearrangement* whenever the distinction between permutation, sign flipping or combined permutation with sign flipping is not pertinent. Finally, let $\hat{\boldsymbol{\beta}}_j^*$ and T_j^*, respectively, be the estimated regression coefficients and the computed statistic for the shuffling j.

The essential assumption of permutation methods is that, for a given set of variables, *their joint probability distribution does not change if the observations are rearranged.* This can be expressed in terms of exchangeable errors or independent and symmetric errors, each of these weakening different assumptions when compared to parametric methods.

Exchangeable errors (EE) is the traditional permutation requirement (Good, 2005). The formal statement is that, for any permutation $\mathbf{P}_j \in \mathcal{P}$,

$$\boldsymbol{\epsilon} \overset{\mathrm{d}}{=} \mathbf{P}_j\boldsymbol{\epsilon} \tag{3.5}$$

where the symbol $\overset{\mathrm{d}}{=}$ denotes equality of distributions. In other words, the errors are considered exchangeable if their joint distribution is invariant with respect to permutation. Exchangeability is similar to, yet more general than, independence, as exchangeable errors can have all-equal and homogeneous dependence. Relative to the common parametric assumptions of independent, normally and identically distributed (iid) errors, EE relaxes two aspects. First, normality is no longer assumed, although identical distributions are required. Second, the independence assumption is weakened slightly to allow exchangeability when the observations

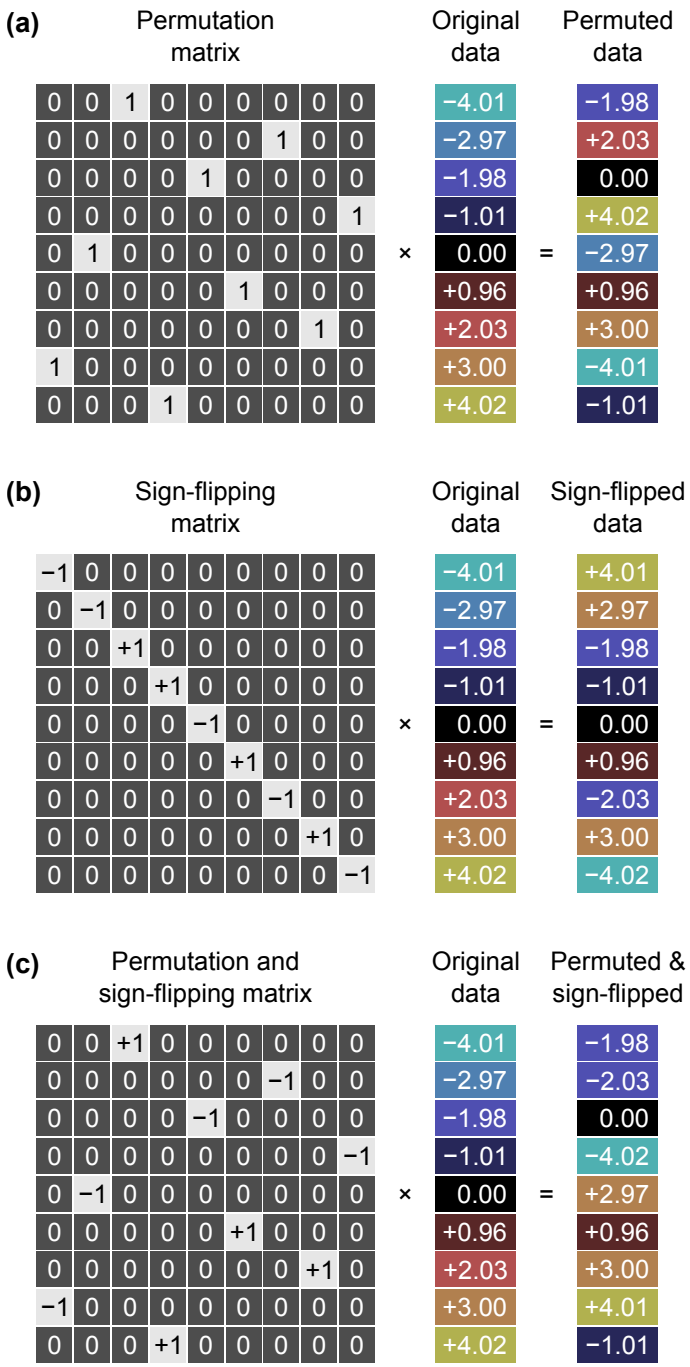


Figure 3.1: Examples of a permutation matrix (*a*), of a sign flipping matrix (*b*), and of a matrix that does permutation and sign flipping (*c*). Pre-multiplication by a permutation matrix shuffles the order of the data, whereas by a sign flipping matrix changes the sign of a random subset of datapoints.

are not independent, but their joint distribution is maintained after permutation. While exchangeability is a general condition that applies to any distribution, the multivariate normal distribution is indeed exchangeable if the marginal distributions are uncorrelated, or if the off-diagonal elements of the covariance matrix are identical to each other (not necessarily equal to zero).[2]

Independent and symmetric errors (ISE) can be considered for measurements that arise, for instance, from differences between two groups if the variances are not assumed to be the same. The formal statement for permutation under ISE is that for any sign flipping matrix $\mathbf{S}_j \in \mathcal{S}$,

$$\boldsymbol{\epsilon} \overset{\mathrm{d}}{=} \mathbf{S}_j \boldsymbol{\epsilon} \tag{3.6}$$

that is, the joint distribution of the error terms is invariant with respect to sign flipping. Relative to the parametric assumptions of independent, normally and identically distributed errors, ISE relaxes normality, although symmetry of distributions is required. Independence is also required to allow sign flipping of one observation without perturbing others.

Although the EE does not require symmetry for the distribution of the error terms, it requires that the variances and covariances of the error terms are all equal, or have a structure that is compatible with the definition of exchangebility blocks (discussed below). While the ISE assumption has yet more stringent requirements, if both EE and ISE are plausible and available for a given model, permutations and sign flippings can be performed together, increasing the number of possible rearrangements, a feature particularly useful for studies with small sample sizes. The formal statement for shuffling under both EE and ISE is that, as with the previous cases, for any matrix $\mathbf{B}_j \in \mathcal{B}$,

$$\boldsymbol{\epsilon} \overset{\mathrm{d}}{=} \mathbf{B}_j \boldsymbol{\epsilon} \tag{3.7}$$

that is, the joint distribution of the error terms remains unchanged under both permutation and sign flipping.

[2] In parametric settings, such dependence structure is often referred to as *compound symmetry*.

There are yet other important aspects related to exchangeability. The experimental design may dictate blocks of observations that are jointly exchangeable, allowing data to be permuted within block or, alternatively, that the blocks may themselves be exchangeable as a whole. This is the case, for instance, for designs that involve multiple observations from each subject. While permutation methods generally do not easily deal with non-independent data, the definition of these *exchangeability blocks* (EB) allows these special cases of well structured dependence to be accommodated. While EB determine how the data shufflings are performed, they should not be confused with *variance groups* (VG), i.e., groups of observations that are known or assumed to have similar variances, which can be pooled for estimation and computation of the statistic. Variance groups need to be compatible with, yet not necessarily identical to, the exchangeability blocks, as discussed in Section 3.2.3.2. A summary of the properties discussed this far and some benefits of permutation methods are shown in Table 3.1.

3.2.3.1 Unrestricted exchangeability

In the absence of nuisance variables, the model reduces to $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and under the null hypothesis $\mathcal{H}_0 : \boldsymbol{\beta} = \mathbf{0}$ the data are pure error, $\mathbf{Y} = \boldsymbol{\epsilon}$. Thus the EE or ISE assumptions on the *error* (presented above) justify freely permuting or sign flipping the *data* under $\mathcal{H}_0$. It is equivalent, however, to alter the design instead of the data. For example, for a nuisance-free design,

$$\mathbf{PY} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \iff \mathbf{Y} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\boldsymbol{\epsilon} \tag{3.8}$$

since permutation matrices $\mathbf{P}$ are orthogonal; the same holds for sign flipping matrices $\mathbf{S}$. This is an important computational consideration as altering the design is much less burdensome than altering the image data. Also note that the errors $\boldsymbol{\epsilon}$ are not observed and thus never directly altered; going forward we will suppress any notation indicating permutation or sign flipping of the errors.

In the presence of nuisance variables (Equation 3.2), however, the problem is more complex. If the nuisance coefficients $\boldsymbol{\gamma}$ were somehow known, an exact permutation test would be available:

$$\mathbf{Y} - \mathbf{Z}\boldsymbol{\gamma} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.9}$$

The perfectly adjusted data $\mathbf{Y} - \mathbf{Z}\boldsymbol{\gamma}$ are then pure error under $\mathcal{H}_0$ and inference could proceed as above. In practice, the nuisance cofficients have to be estimated and the adjusted data will not behave as $\boldsymbol{\epsilon}$. For example, the obvious solution is to use the nuisance-only residuals $\hat{\boldsymbol{\epsilon}}_Z$ as the adjusted data. However, as noted above, residuals induce dependence and any EE or ISE assumptions on $\boldsymbol{\epsilon}$ will not be conveyed to $\hat{\boldsymbol{\epsilon}}_Z$.

A number of approaches have been proposed to produce approximate p-values in these cases (Draper and Stoneman, 1966; Beaton, 1978; Still and White, 1981; Brown and Maritz, 1982; Levin and Robbins, 1983; Freedman and Lane, 1983; Oja, 1987; Gail et al., 1988; Welch, 1990; ter Braak, 1992; Kennedy, 1995; Edgington, 1995; Huh and Jhun, 2001; Jung et al., 2006; Manly, 2007; Kherad-Pajouh and Renaud, 2010). We present these methods in a common notation with detailed annotation in in Table 3.2. While a number of authors have made comparisons between some of these methods (Kennedy, 1995; Kennedy and Cade, 1996; Gonzalez and Manly, 1998; Anderson and Legendre, 1999; Anderson and Robinson, 2001; Anderson and ter Braak, 2003; O'Gorman, 2005; Dekker et al., 2007; Nichols et al., 2008; Ridgway, 2009), they often only approached particular cases, did not consider repeated measurements, did not use full matrix notation as more common in neuroimaging literature, and often did not consider implementation complexities due to the large size of imaging datasets. In this section we focus on the Freedman–Lane and the Dekker methods, which, as we show in Section 3.4.2, produce the best results in terms of control over error rates and power.

The *Freedman–Lane procedure* (Freedman and Lane, 1983) can be performed through the following steps:

1. Regress $\mathbf{Y}$ against the full model that contains both the effects of interest and the nuisance variables, i.e. $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. Use the estimated parameters $\hat{\boldsymbol{\beta}}$ to compute the statistic of interest, and call this statistic T_0.

2. Regress $\mathbf{Y}$ against a reduced model that contains only the nuisance effects, i.e. $\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_Z$, obtaining estimated parameters $\hat{\boldsymbol{\gamma}}$ and estimated residuals

$\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}}$.

3. Compute a set of permuted data $\mathbf{Y}_j^*$. This is done by pre-multiplying the residuals from the reduced model produced in the previous step, $\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}}$, by a permutation matrix, $\mathbf{P}_j$, then adding back the estimated nuisance effects, i.e. $\mathbf{Y}_j^* = \mathbf{P}_j\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}} + \mathbf{Z}\hat{\boldsymbol{\gamma}}$.

4. Regress the permuted data $\mathbf{Y}_j^*$ against the full model, i.e. $\mathbf{Y}_j^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, and use the estimated $\hat{\boldsymbol{\beta}}_j^*$ to compute the statistic of interest. Call this statistic T_j^*.

5. Repeat the Steps 2–4 many times to build the reference distribution of T^* under the null hypothesis.

6. Count how many times T_j^* was found to be equal or larger than T_0, and divide the count by the number of permutations; the result is the p-value.

For the Steps 2 and 3, it is not necessary to actually fit the reduced model at each point in the image. The permuted dataset can equivalently be obtained as $\mathbf{Y}_j^* = (\mathbf{P}_j\mathbf{R}_{\mathbf{Z}} + \mathbf{H}_{\mathbf{Z}})\,\mathbf{Y}$, which is particularly efficient for neuroimaging applications in the typical case of a single design matrix for all image points, as the term $\mathbf{P}_j\mathbf{R}_{\mathbf{Z}} + \mathbf{H}_{\mathbf{Z}}$ is then constant throughout the image and so, needs to be computed just once. Moreover, add the nuisance variables back in Step 3 is not strictly necessary, and the model can be expressed simply as $\mathbf{P}_j\mathbf{R}_{\mathbf{Z}}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, implying that the permutations can actually be performed just by permuting the rows of the residual-forming matrix $\mathbf{R}_{\mathbf{Z}}$. The Freedman–Lane strategy is the one used in the randomise algorithm, discussed in Section 3.2.6.

The rationale for this permutation method is that, if the null hypothesis is true, then $\boldsymbol{\beta} = \mathbf{0}$, and so the residuals from the reduced model with only nuisance variables, $\boldsymbol{\epsilon}_{\mathbf{Z}}$, should not be different than the residuals from the full model, $\boldsymbol{\epsilon}$, and can, therefore, be used to create the reference distribution from which p-values can be obtained.

The *Dekker procedure* consists of orthogonalising the regressors of interest with respect to the nuisance variables. This is done by pre-multiplication of $\mathbf{X}$ by the

residual forming matrix due to $\mathbf{Z}$, i.e., $\mathbf{R}_{\mathbf{Z}}$, then permuting this orthogonalised version of the regressors of interest. The nuisance regressors remain in the model.[3]

For both the Freedman–Lane and the Dekker procedures, if the errors are independent and symmetric (ISE), the permutation matrices $\mathbf{P}_j$ can be replaced for sign flipping matrices $\mathbf{S}_j$. If both EE and ISE are considered appropriate, then permutation and sign flipping can be used concomitantly.

3.2.3.2 Restricted exchangeability

Some experimental designs involve multiple observations from each subject, or the subjects may come from groups that may possess characteristics that may render their distributions not perfectly comparable. Both situations violate exchangeability. However, when the dependence between observations has a block structure, this structure can be taken into account when permuting the model, restricting the set of all otherwise possible permutations to only those that respect the relationship between observations (Pesarin, 2001).[4] The EE and ISE assumptions are then asserted at the level of these exchangeability blocks, rather than for each observation individually. The experimental hypothesis and the study design determine how the EBs should be formed and how the permutation or sign flipping matrices should be constructed. Except Huh–Jhun, the other methods can be applied at the block level as in the unrestricted case.

Within-block exchangeability Observations that share the same dependence structure, either assumed or known in advance, can be used to define EBs such that EE

[3] In Winkler et al. (2014) we named this method as "Smith" because, although orthogonalisation is a well known procedure, it did not seem to have been proposed by anyone to address the issues with permutation methods with the GLM until Smith and others presented it in a conference poster (Nichols et al., 2008). We also tried to keep it consistent with Ridgway (2009), with the convention of calling the methods by the earliest author that we could identify as the proponent for each method, even though this method seems to have been proposed by an anonymous referee of O'Gorman (2005). After publication, however, it was brought to our knowledge that the same method had in fact been proposed earlier, by Dekker, Krackhard and Snijders, in a conference in 2003 (Dekker et al., 2003, 2007). To give the proper credit to the original authors, we henceforth call this method simply as "Dekker".

[4] Observations that are exchangeable only in some subsets of all possible permutations are said *weakly exchangeable* (Good, 2002).

are asserted with respect to these blocks only, and the empirical distribution is constructed by permuting exclusively within block, as shown in Figure 3.2. Once the blocks have been defined, the regression of nuisance variables and the construction of the reference distribution can follow strategies as Freedman–Lane or Dekker, as above. The ISE, when applicable, is transparent to this kind of block structure, so that the sign flips occur as under unrestricted exchangeability. For within-block exchangeability, in general each EB corresponds to a VG for the computation of the test statistic. See Section 3.3 for examples.

Whole-block exchangeability Certain experimental hypotheses may require the comparison of sets of observations to be treated as a whole, being not exchangeable within set. Exchangeability blocks can be constructed such that each include, in a consistent order, all the observations pertaining to a given set and, differently than in within-block exchangeability, here each block is exchanged with the others on their entirety, while maintaining the order of observations within block unaltered. For ISE, the signs are flipped for all observations within block at once. Variance groups are not constructed one per block; instead, each VG encompasses one or more observations per block, all in the same order, e.g., one VG with the first observation of each block, another with the second of each block and so on. Consequently, all blocks must be of the same size, and all with their observations ordered consistently, either for EE or for ISE. Examples of permutation and sign flipping matrices for whole block permutation are shown in Figure 3.3. See Section 3.3 for examples.

Variance groups mismatching exchangeability blocks While variance groups can be defined implicity, as above, according to whether within- or whole-block permutation is to be perfomed, this is not compulsory. In some cases the EBs are defined based on the non-independence between observations, even if the variances across all observations can still be assumed to be identical. See Section 3.3 for an example using a paired t-test.

Choice of the statistic with exchangebility blocks The statistics F and t, described in Section 3.2.1, are pivotal and follow known distributions when, among

Table 3.1: Compared with parametric methods, permutation tests relax a number of assumptions and can be used in a wider variety of situations. Some of these assumptions can be further relaxed with the definition of exchangeability blocks.

Assumptions	EE	ISE	Parametric
With respect to the dependence structure between error terms:			
Independent	✓	✓	✓
Non-independent, exchangeable	✓	✗	✗
Non-independent, non-exchangeable	✗	✗	✗
With respect to the distributions of the error terms:			
Normal, identical	✓	✓	✓
Symmetrical, identical	✓	✓	✗
Symmetrical, non-identical	✗	✓	✗
Skewed, identical	✓	✗	✗
Skewed, non-identical	✗	✗	✗

✓ Can be used directly if the assumptions regarding dependence structure and distribution of the error terms are both met.

✗ Cannot be used directly, or can be used in particular cases.

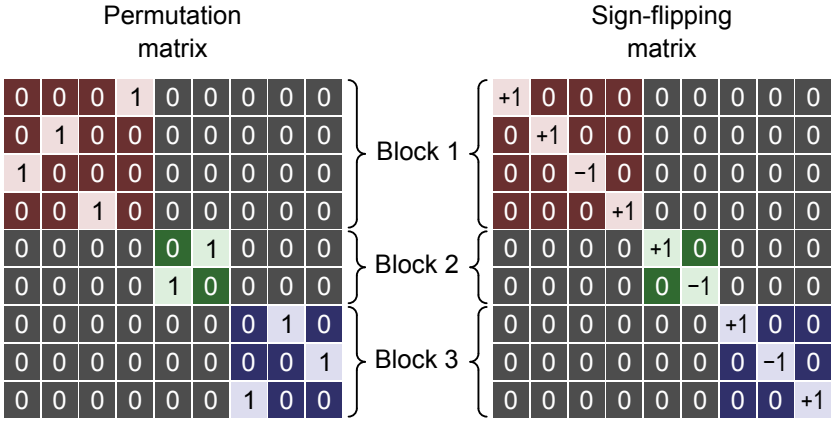


Figure 3.2: Left: Example of a permutation matrix that shuffles data within block only. The blocks are not required to be of the same size. The elements outside the diagonal blocks are always equal to zero, such that data cannot be swapped across blocks. *Right:* Example of a sign flipping matrix. Differently than within-block permutation matrices, here sign flipping matrices are transparent to the definitions of the blocks, such that the block definitions do not need to be taken into account, albeit their corresponding variance groups are considered when computing the statistic.

Table 3.2: A number of methods are available to obtain parameter estimates and construct a reference distribution in the presence of nuisance variables.

Method	Model
Draper–Stoneman$^{(a)}$	$\mathbf{Y} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
Still–White$^{(b)}$	$\mathbf{PR_Z Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
Freedman–Lane$^{(c)}$	$(\mathbf{PR_Z} + \mathbf{H_Z})\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
ter Braak$^{(d)}$	$(\mathbf{PR_M} + \mathbf{H_M})\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
Kennedy$^{(e)}$	$\mathbf{PR_Z Y} = \mathbf{R_Z X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
Manly$^{(f)}$	$\mathbf{PY} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
Huh–Jhun$^{(g)}$	$\mathbf{PQ'R_Z Y} = \mathbf{Q'R_Z X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
Dekker$^{(h)}$	$\mathbf{Y} = \mathbf{PR_Z X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
Parametric$^{(i)}$	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$

(a) Draper and Stoneman (1966). This method was called "Shuffle Z" by Kennedy (1995), and using the same notation adopted here, it would be called "Shuffle X". (b) Still and White (1981); Levin and Robbins (1983); Gail et al. (1988). (c) Freedman and Lane (1983). (d) ter Braak (1992). The null distribution for this method considers $\hat{\boldsymbol{\beta}}_j^* = \hat{\boldsymbol{\beta}}$, i.e., the permutation happens under the alternative hypothesis, rather than the null. (e) Kennedy (1995); Kennedy and Cade (1996). This method was referred to as "Residualize both Y and Z" in the original publication, and using the same notation adopted here, it would be called "Residualize both Y and X". (f) Manly (2007). (g) Huh and Jhun (2001); Jung et al. (2006); Kherad-Pajouh and Renaud (2010). $\mathbf{Q}$ is a $N \times N'$ matrix, where N' is the rank of $\mathbf{R_Z}$. $\mathbf{Q}$ is computed through Schur decomposition of $\mathbf{R_Z}$, such that $\mathbf{R_Z} = \mathbf{QQ'}$ and $\mathbf{I}_{N' \times N'} = \mathbf{Q'Q}$. For this method, $\mathbf{P}$ is $N' \times N'$. From the methods in the table, this is the only that cannot be used directly under restricted exchangeability, as the block structure is not preserved. (h) The Dekker method consists of orthogonalization of $\mathbf{X}$ with respect to $\mathbf{Z}$. In the permutation and multiple regression literature, this method was proposed by Dekker et al. (2003, 2007), then later by an anonymous referee of O'Gorman (2005), by Nichols et al. (2008) and discussed by Ridgway (2009). (i) The parametric method does not use permutations, being instead based on distributional assumptions. □ For all the methods, the left side of the equations contains the data (regressand), the right side the regressors and error terms. The unpermuted models can be obtained by replacing $\mathbf{P}$ for $\mathbf{I}$. Even for the unpermuted models, and even if $\mathbf{X}$ and $\mathbf{Z}$ are orthogonal, not all these methods produce the same error terms $\boldsymbol{\epsilon}$. This is the case, for instance, of the Kennedy and Huh–Jhun methods. Under orthogonality between $\mathbf{X}$ and $\mathbf{Z}$, some regression methods are equivalent to each other.

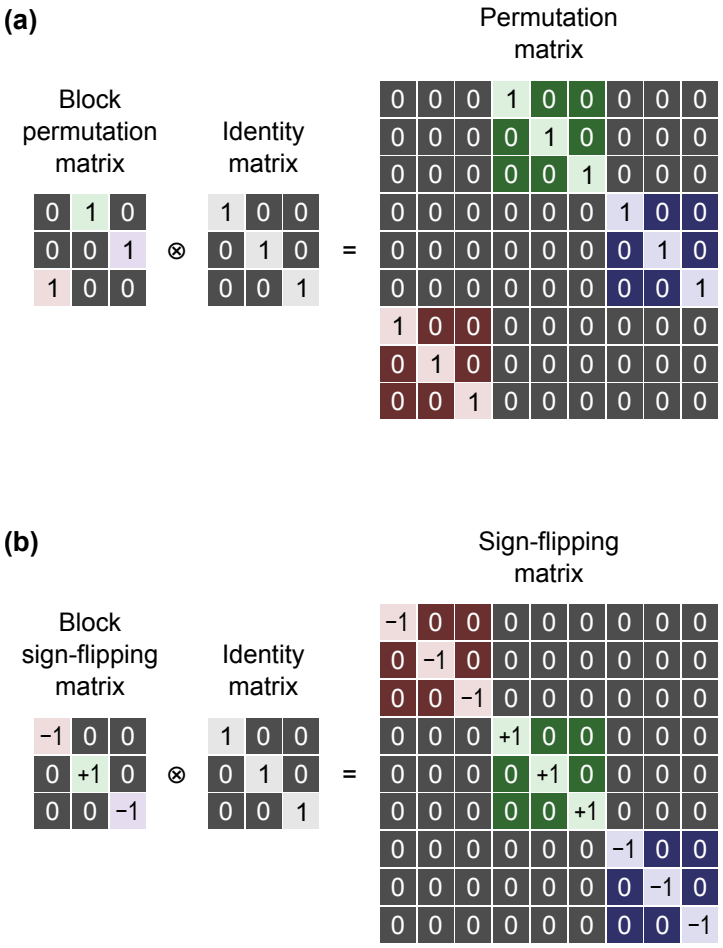


Figure 3.3: (*a*) Example of a permutation matrix that shuffles whole blocks of data. The blocks need to be of the same size. (*b*) Example of a sign flipping matrix that changes the signs of the blocks as a whole. Both matrices can be constructed by the Kronecker product (represented by the symbol $\otimes$) of a permutation or a sign flipping matrix (with size determined by the number of blocks) and an identity matrix (with size determined by the number of observations per block).

other assumptions, the error terms for all observations are identically distributed. Under these assumptions, all the errors terms can be pooled to compute the residual sum of squares (the term $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$ in the Equation 3.3) and so, the variance of the parameter estimates. This forms the basis for parametric inference, and is also useful for non-parametric tests. However, the presence of EBs is incompatible with the equality of distributions across all observations, with the undesired consequence that pivotality is lost, as shown in Sections 3.4.1 and 3.5.1. Although these statistics can still be used with permutation methods in general, the lack of pivotality for imaging applications can cause problems for correction for multiple testing. When exchangebility blocks are present, a suitable statistic can be computed as:

$$G = \frac{\hat{\boldsymbol{\psi}}'\mathbf{C}\left(\mathbf{C}'(\mathbf{M}'\mathbf{W}\mathbf{M})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}'\hat{\boldsymbol{\psi}}}{\Lambda \cdot \text{rank}\left(\mathbf{C}\right)} \tag{3.10}$$

where $\mathbf{W}$ is a $N \times N$ diagonal weighting matrix that has elements

$$W_{nn} = \frac{\sum_{n' \in g_n} R_{n'n'}}{\hat{\boldsymbol{\epsilon}}'_{g_n}\hat{\boldsymbol{\epsilon}}_{g_n}} \tag{3.11}$$

where g_n represents the variance group to which the n-th observation belongs, $R_{n'n'}$ is the n'-th diagonal element of the residual forming matrix, and $\hat{\boldsymbol{\epsilon}}_{g_n}$ is the vector of residuals associated with the same VG.[5] In other words, each diagonal element of $\mathbf{W}$ is the reciprocal of the estimated variance for their corresponding group. This variance estimator is equivalent to the one proposed by Horn et al. (1975), and corresponds to the "HC2" of (MacKinnon and White, 1985); see also Guillaume et al. (2014) for further discussion. The remaining term in Equation 3.10 is given by (Welch, 1951):

$$\Lambda = 1 + \frac{2(s-1)}{s(s+2)}\sum_{g}\frac{1}{\sum_{n\in g} R_{nn}}\left(1 - \frac{\sum_{n\in g} W_{nn}}{\text{trace}\left(\mathbf{W}\right)}\right)^2 \tag{3.12}$$

where $s = \text{rank}\left(\mathbf{C}\right)$ as before. The statistic G provides a generalisation of a num-

[5] Note that, for clarity, G is defined in Equation 3.10 as a function of $\mathbf{M}$, $\boldsymbol{\psi}$ and $\mathbf{C}$ in the unpartitioned model. With the partitioning described in Section 3.2.2, each of these variables is replaced by their equivalents in the partitioned, full model, i.e., $[\mathbf{X}\ \mathbf{Z}]$, $[\boldsymbol{\beta}'\ \boldsymbol{\gamma}']'$ and $[\mathbf{I}_{s\times s}\ \mathbf{0}_{s\times(r-s)}]'$ respectively.

Table 3.3: The statistic G provides a generalisation for a number of well known statistical tests.

	$\text{rank}\,(\mathbf{C}) = 1$	$\text{rank}\,(\mathbf{C}) > 1$
Homoscedastic errors, unrestricted exchangeability ($\Lambda = 1$)	Square of Student's t	F-ratio
Homoscedastic within VG, restricted exchangeability ($\Lambda \neq 1$)	Square of Aspin–Welch v	Welch's v^2

ber of well known statistical tests, some of them summarised in Table 3.3. When there is only one VG, variance estimates can be pooled across all observations, resulting in $\Lambda = 1$ and so, $G = F$. If $\mathbf{W} = \mathbf{V}^{-1}$, the inverse of the true covariance matrix, G is the statistic for an F-test in a weighted least squares model (WLS) (Christensen, 2002). If there are multiple variance groups, G is equivalent to the v^2 statistic for the problem of testing the means for these groups under no homoscedasticity assumption, i.e., when the variances cannot be assumed to be all equal (Welch, 1951).[6] If, despite heteroscedasticity, Λ is replaced by 1, G is equivalent to the James' statistic for the same problem (James, 1951). When $\text{rank}\,(\mathbf{C}) = 1$, and if there are more than one VG, $\text{sign}(\hat{\boldsymbol{\beta}})\sqrt{G}$ is the well-known v statistic for the Behrens–Fisher problem (Fisher, 1935b; Aspin and Welch, 1949); with only one VG present, the same expression produces the Student's t statistic, as shown earlier. If the definition of the blocks is respected, all these particular cases produce pivotal statistics, and the generalisation provided by G allows straightforward implementation.

3.2.4 Number of permutations

For a study with N observations, the maximum number of possible permutations is $N!$, and the maximum number of possible sign flips is 2^N. However, in the presence of B exchangebility blocks that are exchangeable as a whole, the maximum number

[6] If the errors are independent and normally distributed, yet not necessarily with equal variances (i.e., $\Lambda \neq 1$), parametric p-values for G can be approximated by referring to the F-distribution with degrees of freedom $\nu_1 = s$ and $\nu_2 = 2(s-1)/3/(\Lambda - 1)$.

of possible permutations drops to no more than $B!$, and the maximum number of sign flips to 2^B. For designs where data is only exchangeable within-block, the maximum number of possible permutations is $\prod_{b=1}^{B} N_b!$, where N_b is the number of observations for the b-th block, and the maximum number of sign flips continues to be 2^N.

However, the actual number of possible rearrangements may be smaller depending on the null hypothesis, the permutation strategy, or other aspects of the study design. If there are discrete covariates, or if there are ties among continuous regressors, many permutations may not alter the model at all. The maximum number of permutations can be calculated generically from the design matrix observing the number of repeated rows in $\mathbf{X}$ for the Freedman–Lane and most other methods, or in $\mathbf{M}$ for the ter Braak and Manly methods. The maximum number of possible permutations or sign flips, for different restrictions on exchangeability, is shown in Table 3.4.

Even considering the restrictions dictacted by the study design, the number of possible shufflings tends to be very large, even for samples of moderate size, and grows very rapidly as observations are included. When the number of possible rearrangements is large, not all of them need to be performed for the test to be valid (Dwass, 1957; Chung and Fraser, 1958), and the resulting procedure will be approximately exact (Edgington, 1969). The number can be chosen according to the availability of computational resources and considerations about power and precision. The smallest p-value that can be obtained is given by $1/J$, where J is the number of permutations performed. The precision of permutation p-values may be determined considering the confidence interval around the significance level.

To efficiently avoid permutations that do not change the design matrix, the Algorithm "L" (Knuth, 2005) can be used. This algorithm is simple and has the benefit of generating only permutations that are unique, i.e., in the presence of repeated elements, it correctly avoids synonymous permutations. This is appropriate when enumerating all possible permutations. However, the algorithm produces sequentially permutations that are in lexicographic order. Although this can be advantageous in other settings, here this behaviour can be problematic when running only a subset of $\mathcal{P}$, and has potential to bias the results. For imaging ap-

Table 3.4: Maximum number of unique permutations considering exchangeability blocks.

Exchangeability	EE	ISE
Unrestricted	$N!$	2^N
Unrestricted, repeated rows	$N! \prod_{m=1}^{M} \frac{1}{N_m!}$	2^N
Within-block	$\prod_{b=1}^{B} N_b!$	2^N
Within-block, repeated rows	$\prod_{b=1}^{B} N_b! \prod_{m=1}^{M\|b} \frac{1}{N_{m\|b}!}$	2^N
Whole-block	$B!$	2^B
Whole-block, repeated blocks	$B! \prod_{\tilde{m}=1}^{\tilde{M}} \frac{1}{N_{\tilde{m}}!}$	2^B

B Number of exchangebility blocks (EB).

M Number of distinct rows in $\mathbf{X}$.

$M|b$ Number of distinct rows in $\mathbf{X}$ within the b-th block.

$\tilde{M}$ Number of distinct blocks of rows in $\mathbf{X}$.

N Number of observations.

N_b Number of observations in the b-th block.

N_m Number of times each of the M distinct rows occurs in $\mathbf{X}$.

$N_{m|b}$ Number of times each of the m-th unique row occurs within the b-th block.

$N_{\tilde{m}}$ Number of times each of the $\tilde{M}$ distinct blocks occurs in $\mathbf{X}$.

plications, where there are many points (voxels, vertices, faces) being analysed, it is in general computationally less expensive to shuffle many times a sequence of values and store these permuted sequences, than actually fit the permuted model for all points. As a consequence, the problem with lexicographically ordered permutations can be solved by generating all the possible permutations, and randomly drawing J elements from $\mathcal{P}$ to do the actual shufflings of the model, or generating random permutations and checking for duplicates. Alternatively, the procedure can be conducted without attention to repeated permutations using simple shuffling of the data. This strategy is known as *conditional Monte Carlo* (CMC) (Trotter and Tukey, 1956; Pesarin and Salmaso, 2010a), as each of the random realisations is conditional on the available observed data.

Sign flipping matrices, on the other hand, can be listed using a numeral system with radix 2, and the sign flipped models can be performed without the need to enumerate all possible flips or to appeal to CMC. The simplest strategy is to use the digits 0 and 1 of the binary numeral system, treating 0 as -1 when assembling the matrix. In a binary system, each sign flipping matrix is also its own numerical identifier, such that avoiding repeated sign flippings is trivial. The binary representation can be converted to and from radix 10 if needed, e.g., to allow easier human readability.

For within-block exchangeability, permutation matrices are constructed within-block, then concatenated along their diagonal to assemble $\mathbf{P}_j$, which also has a block structure. The elements outside the blocks are filled with zeros as needed (Figure 3.2). The block definitions can be ignored for sign flipping matrices for designs where ISE is asserted within-block. For whole-block exchangeability, permutation and sign flipping matrices are generated by treating each block as an element, and the final $\mathbf{P}_j$ or $\mathbf{S}_j$ are assembled via Kronecker multiplication by an identity matrix of the same size as the blocks (Figure 3.3).

3.2.5 Multiple testing

Differently than with parametric methods, correction for multiple testing using permutation does not require the introduction of more assumptions. For familywise error rate correction (FWER), the method was described by Holmes et al.

(1996). As the statistics T_j^* are calculated for each shuffling to build the reference distribution at each point, the maximum value of T_j^* across the image, $T_j^{\max}$, is also recorded for each rearrangement, and its empirical distribution is obtained. For each test in the image, an FWER-corrected p-value can then be obtained by computing the proportion of $T_j^{\max}$ that is above T_0 for each test. A single FWER threshold can also be applied to the statistical map of T_0 values using the distribution of $T_j^{\max}$. The same strategy can be used for statistics that combine spatial extent of signals, such as cluster extent or mass (Bullmore et al., 1999), threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009) and others (Marroquin et al., 2011). For these spatial statistics, the effect of lack of pivotality can be mitigated by non-stationarity correction (Hayasaka et al., 2004; Salimi-Khorshidi et al., 2011).

The p-values under the null hypothesis are uniformly distributed in the interval $[0, 1]$. As a consequence, the p-values themselves are pivotal quantities and, in principle, could be used for multiple testing correction as above. The distribution of minimum p-value, $p_j^{\min}$, instead of $T_j^{\max}$, can be used. Due to the discreteness of the p-values, this approach, however, entails some computational difficulties that may cause considerable loss of power (Pantazis et al., 2005). Correction based on false-discovery rate (FDR) can be used once the uncorrected p-values have been obtained for each point in the image. Either a single FDR threshold can be applied to the map of uncorrected p-values (Benjamini and Hochberg, 1995; Genovese et al., 2002) or an FDR-adjusted p-value can be calculated at each point (Yekutieli and Benjamini, 1999).

3.2.6 The randomise algorithm

Algorithm 1 describes a procedure for permutation inference on contrasts of the GLM parameter estimates using the Freedman–Lane method. Modifications for other methods are trivial. For this algorithm, consider $\mathbf{Y}$ as a four-dimensional array, being the first three dimensions for space and the last for an observation index. A variable $\mathbf{v} = [x, y, z]$ is used to specify the point position in space, so that the vector of n different observations per point is represented as $\mathbf{Y}[\mathbf{v}]$. A set $\mathcal{C}$ of contrasts is specified, as well as the unpartitioned design matrix $\mathbf{M}$. Indicator

variables are used to specify whether the errors should be treated as exchangeable (EE = TRUE), independent and symmetric (ISE = TRUE), or both, which allows for permutations to happen together with sign flipping. A positive integer J is specified as the number permutations to be performed. Optionally, a $n \times 1$ vector $\mathbf{b}$ is provided to indicate the B exchangebility blocks that group the observations, along with an indicator variable PB that informs whether blocks should be permuted as a whole (PB = TRUE), or if permutations should happen within block only (PB = FALSE). The specification of $\mathbf{b}$ and PB obviate the need to specify the variance groups, as these can be defined implicitly for within or whole-block permutation when the pivotal statistic is computed.

Algorithm 1: The randomise algorithm.

Require: $\mathbf{Y}, \mathbf{M}, \mathcal{C}$, EE, ISE, J. **Optional**: $\mathbf{b}$, PB. ▷ Input variables.
if ¬ exist(PB) **then** ▷ If PB was not provided.
  PB ← FALSE ▷ Permutations happen within block.
end if
if ¬ exist($\mathbf{b}$) **then** ▷ If $\mathbf{b}$ was not provided.
  $\mathbf{b} \leftarrow \mathbf{1}_{n\times 1}$ ▷ A vector of ones is used for $\mathbf{b}$.
  PB ← FALSE ▷ Permutations happen within the single block.
end if
for all $\mathbf{C} \in \mathcal{C}$ **do** ▷ For each contrast.
  $\mathbf{X}, \mathbf{Z} \leftarrow \text{partition}(\mathbf{M}, \mathbf{C})$ ▷ Partition the model.
  $\mathbf{M} \leftarrow [\mathbf{X}\ \mathbf{Z}]$ ▷ For simplicity, replace $\mathbf{M}$.
  $J^{\max} \leftarrow \text{calc\_maxshuf}(\mathbf{X}, \mathbf{b}, \text{PB}, \text{EE}, \text{ISE})$ ▷ Maximum possible shufflings.
  if EE **then** ▷ If errors are exchangeable.
    if $J \geqslant J^{\max}$ **then** ▷ Exhaustive or too many permutations requested.
      $\mathcal{P} \leftarrow \text{algorithm\_L}(\mathbf{X}, \mathbf{b}, \text{PB})$ ▷ List all possible permutations.
    else
      $\mathcal{P} \leftarrow \text{permute\_randomly}(\mathbf{X}, \mathbf{b}, \text{PB}, J-1)$ ▷ Ignore repeated $\mathbf{P}_j$.
      $\mathcal{P} \leftarrow \{\mathcal{P}, \mathbf{I}\}$ ▷ Ensure inclusion of the unpermuted model.
    end if
  end if
  if ISE **then** ▷ If errors are independent and symmetric.
    if $J \geqslant J^{\max}$ **then** ▷ Exhaustive or too many sign flips requested.
      $\mathcal{S} \leftarrow \text{list\_signflips}(\mathbf{b}, \text{PB})$ ▷ List all possible sign flippings.
    else
      $\mathcal{S} \leftarrow \text{signflip\_randomly}(n, \mathbf{b}, \text{PB}, J-1)$ ▷ Ignore repeated $\mathbf{S}_j$.
      $\mathcal{S} \leftarrow \{\mathcal{S}, \mathbf{I}\}$ ▷ Ensure inclusion of the non-sign flipped model.
    end if
  end if

if EE $\wedge$ ISE **then** ▷ Errors independent, symmetric and exchangeable.
  $\mathcal{B} \leftarrow$ draw_products$(\mathcal{P}, \mathcal{S}, J)$ ▷ Draw J random products $\mathbf{P}_{j'}\mathbf{S}_{j''}$.
  if $\mathbf{I} \notin \mathcal{B}$ **then** ▷ If non-shuffled model is absent from $\mathcal{B}$.
    $\mathcal{B} \leftarrow \{\mathbf{B}_1, \ldots, \mathbf{B}_{J-1}, \mathbf{I}\}$ ▷ Ensure non-shuffled model is included.
  end if
  $\mathcal{B} \leftarrow \mathcal{P}$ ▷ Treat $\mathcal{B}$ as $\mathcal{P}$ for simplicity.
else if ISE $\wedge \neg$ EE **then** ▷ If errors are only independent and symmetric.
  $\mathcal{P} \leftarrow \mathcal{S}$ ▷ Treat $\mathcal{S}$ as $\mathcal{P}$ for simplicity.
end if
for all **v** **do** ▷ For each image point.
  $\mathbf{U}[\mathbf{v}] \leftarrow 0$ ▷ Initialise counter for uncorrected p-value.
  $\mathbf{F}[\mathbf{v}] \leftarrow 0$ ▷ Initialise counter for FWER-corrected p-value.
  $\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}}[\mathbf{v}] \leftarrow (\mathbf{I} - \mathbf{Z}\mathbf{Z}^{+})\mathbf{Y}[\mathbf{v}]$ ▷ Remove the nuisance effects.
  $\hat{\boldsymbol{\psi}}[\mathbf{v}] \leftarrow \mathbf{M}^{+}\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}}[\mathbf{v}]$ ▷ Estimate regression coefficients.
  $\hat{\boldsymbol{\epsilon}}[\mathbf{v}] \leftarrow (\mathbf{I} - \mathbf{M}\mathbf{M}^{+})\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}}[\mathbf{v}]$ ▷ Estimate the residuals.
  $\mathbf{T}_0[\mathbf{v}] \leftarrow \text{pivotal}(\hat{\boldsymbol{\psi}}[\mathbf{v}], \hat{\boldsymbol{\epsilon}}[\mathbf{v}], \mathbf{M}, \mathbf{b}, \text{PB})$ ▷ Compute a pivotal statistic.
end for
for $\mathbf{P}_j \in \mathcal{P}$ **do** ▷ For each shuffling (permutation and/or sign flipping).
  $\mathbf{M}^{*}_j \leftarrow \mathbf{P}_j\mathbf{M}$ ▷ Shuffle the model.
  for all **v** **do** ▷ For each image point.
    $\hat{\boldsymbol{\psi}}^{*}_j[\mathbf{v}] \leftarrow (\mathbf{M}^{*}_j)^{+}\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}}[\mathbf{v}]$ ▷ Fit permuted model.
    $\hat{\boldsymbol{\epsilon}}^{*}_j[\mathbf{v}] \leftarrow (\mathbf{I} - \mathbf{M}^{*}_j(\mathbf{M}^{*}_j)^{+})\hat{\boldsymbol{\epsilon}}_{\mathbf{Z}}[\mathbf{v}]$ ▷ Residuals.
    $\mathbf{T}^{*}_j[\mathbf{v}] \leftarrow \text{pivotal}(\hat{\boldsymbol{\psi}}^{*}_j[\mathbf{v}], \hat{\boldsymbol{\epsilon}}^{*}_j[\mathbf{v}], \mathbf{M}^{*}_j, \mathbf{b}, \text{PB})$ ▷ Shuffled statistic.
    if $\mathbf{T}^{*}_j[\mathbf{v}] \geqslant \mathbf{T}_0[\mathbf{v}]$ **then** ▷ If shuffled statistic is larger.
      $\mathbf{U}[\mathbf{v}] \leftarrow \mathbf{U}[\mathbf{v}] + 1$ ▷ Increment counter for uncorrected.
    end if
  end for
  $T^{\max}_j \leftarrow \max(\mathbf{T}^{*}_j)$ ▷ Find the largest T^{*}_j across space.
  for all **v** **do** ▷ For each image point.
    if $T^{\max}_j \geqslant \mathbf{T}_0[\mathbf{v}]$ **then** ▷ If $T^{\max}_j$ is larger.
      $\mathbf{F}[\mathbf{v}] \leftarrow \mathbf{F}[\mathbf{v}] + 1$ ▷ Increment counter for FWER-corrected.
    end if
  end for
end for
p-value $\leftarrow \mathbf{U}/J$ ▷ Significance map for this **C**, uncorrected.
p_{FWER}-value $\leftarrow \mathbf{F}/J$ ▷ Significance map for this **C**, FWER-corrected.
return p-value, p_{FWER}-value. ▷ Save significance images to disk.
end for

In the algorithm, the statistics T for each point (voxel, vertex, face) are stored in the array **T**, whereas the counters are stored in the arrays **U** and **F**. The design matrix as well as the contrasts can be specific for each image point (voxelwise, vertexwise, facewise), and there is no challenge other than implementation. It is

possible to omit the for-loop between lines 56 and 60, and instead store the distribution of the largest statistic as a vector of size J, which is then used to assess significance. The code runs faster, but it would be slightly less clear to present. In programming languages that offer good matrix manipulation capabilities, e.g. Octave, MATLAB or R, the for-loops that iterate for each point $\mathbf{v}$ can be replaced by matrix operations that are executed all in a single step. In the FMRIB Software Library (FSL)[7], a fast implementation, in C++, of the randomise algorithm is available.

3.3 Worked examples

The examples below serve to illustrate the permutation aspects discussed in the chapter, all with tiny samples, $N = 12$ only, so that the design matrices can be shown in their full extent. While permutation tests in general remain valid even with such small samples, these examples are by no means to be understood as a recommendation for sample sizes. There are many reasons why larger samples are more appropriate (see Button et al. (2013) for a recent review), and in what concerns permutations methods, larger samples allow smaller p-values, improve the variance estimates for each VG (which are embodied in the weighting matrix under restricted exchangeability), and allow finer control over the familywise error rate. For each example, the relevant contrasts are also shown.

Example 1: Mean effect Consider a multi-subject FMRI study to investigate the BOLD response associated with a novel experimental task. After the first-level analysis (within subject), maps of contrasts of parameter estimates for each subject are used in a second level analysis. The design matrix for the mean effect is simply a column of ones, and permutations of the data or of the design matrix do not change the model with respect to the regressor of interest. However, by treating the errors as symmetric, instead of permutation, the signs of the ones in the design matrix, or of each datapoint, can be flipped randomly to create the empirical distribution from which inference can be performed. In the presence of nuisance variables, such as handedness, the procedure is performed as in either the Freedman–Lane

[7] Available for download at `http://www.fmrib.ox.ac.uk/fsl`.

or Dekker methods, replacing the permutation matrix for a sign flipping matrix (Table 3.5).

Table 3.5: Coding of the design matrix, exchangeability blocks and variance groups for **Example 1**. Under unrestricted exchangeability, all subjects are assigned to a single block, and with identical variances, all to a single variance group. The regressor $\mathbf{m}_1$ codes for the overall mean, whereas $\mathbf{m}_2$ codes for handedness.

Coded data (**Y**)	EB	VG	Model (**M**)	
			$\mathbf{m}_1$	$\mathbf{m}_2$
Subject 1	1	1	1	h_1
Subject 2	1	1	1	h_2
Subject 3	1	1	1	h_3
Subject 4	1	1	1	h_4
Subject 5	1	1	1	h_5
Subject 6	1	1	1	h_6
Subject 7	1	1	1	h_7
Subject 8	1	1	1	h_8
Subject 9	1	1	1	h_9
Subject 10	1	1	1	h_{10}
Subject 11	1	1	1	h_{11}
Subject 12	1	1	1	h_{12}
Contrast 1 ($\mathbf{C}_1'$)			$+1$	0
Contrast 2 ($\mathbf{C}_2'$)			-1	0

Example 2: Multiple regression Consider the analysis of a study that compares patients and controls with respect to brain cortical thickness, and that recruiting process ensured that all selected subjects are exchangeable. Elder subjects may, however, have thinner cortices, regardless of the diagnosis so, without considering the possibility of interaction. To control for the confounding effect of age, it is included in the design as a nuisance regressor. Sex is also included. The permutation strategy follows the Freedman–Lane or Dekker methods, with the residuals of the reduced model being permuted under unrestricted exchangeability (Table 3.6).

Example 3: Paired t-test Consider a study to investigate the effect of the use of a certain analgesic in the magnitude of the BOLD response associated with painful stimulation. In this example, the response after the treatment is compared with

Table 3.6: Coding for **Example 2**. Under unrestricted exchangeability, all subjects are assigned to a single block. The regressors $\mathbf{m}_1$ and $\mathbf{m}_2$ code for the experimental groups, $\mathbf{m}_3$ and $\mathbf{m}_4$ for age and sex.

Coded data (**Y**)	EB	VG	Model (**M**)			
			$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$	$\mathbf{m}_4$
Subject 1	1	1	1	0	a_1	s_1
Subject 2	1	1	1	0	a_2	s_2
Subject 3	1	1	1	0	a_3	s_3
Subject 4	1	1	1	0	a_4	s_4
Subject 5	1	1	1	0	a_5	s_5
Subject 6	1	1	1	0	a_6	s_6
Subject 7	1	1	0	1	a_7	s_7
Subject 8	1	1	0	1	a_8	s_8
Subject 9	1	1	0	1	a_9	s_9
Subject 10	1	1	0	1	a_{10}	s_{10}
Subject 11	1	1	0	1	a_{11}	s_{11}
Subject 12	1	1	0	1	a_{12}	s_{12}
Contrast 1 ($\mathbf{C}_1'$)			$+1$	-1	0	0
Contrast 2 ($\mathbf{C}_2'$)			-1	$+1$	0	0

the response before the treatment, i.e., each subject is their own control. The experimental design is the "paired t-test". One EB is defined per subject, as the observatios are not exchangeable across subjects, and as the variance can be assumed to be homogeneous across all observations, only one VG is defined encompassing all observations (Table 3.7).

Example 4: Unequal group variances Consider a study using FMRI to compare whether the BOLD response associated with a certain cognitive task would differ among subjects with autistic spectrum disorder (ASD) and control subjects, while taking into account differences in age and sex. In this hypothetical example, the cognitive task is known to produce more erratic signal changes in the patient group than in controls. Therefore, variances cannot be assumed to be homogeneous with respect to the group assignment of subjects. This is an example of the classical Behrens–Fisher problem. To accommodate heteroscedasticity, two permutation blocks are defined according to the group of subjects. Under the assumption of

Table 3.7: Coding of the design matrix exchangebility blocks and variance groups for **Example 3**. Observations are exchangeable only within subject, and variance can be estimated considering all observations as a single group. The regressor $\mathbf{m}_1$ codes for treatment, whereas $\mathbf{m}_2$ to $\mathbf{m}_7$ code for subject-specific mean.

Coded data (**Y**)	EB	VG	Model (**M**)						
			$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$	$\mathbf{m}_4$	$\mathbf{m}_5$	$\mathbf{m}_6$	$\mathbf{m}_7$
Subj. 1, obs. 1	1	1	$+1$	1	0	0	0	0	0
Subj. 2, obs. 1	2	1	$+1$	0	1	0	0	0	0
Subj. 3, obs. 1	3	1	$+1$	0	0	1	0	0	0
Subj. 4, obs. 1	4	1	$+1$	0	0	0	1	0	0
Subj. 5, obs. 1	5	1	$+1$	0	0	0	0	1	0
Subj. 6, obs. 1	6	1	$+1$	0	0	0	0	0	1
Subj. 1, obs. 2	1	1	-1	1	0	0	0	0	0
Subj. 2, obs. 2	2	1	-1	0	1	0	0	0	0
Subj. 3, obs. 2	3	1	-1	0	0	1	0	0	0
Subj. 4, obs. 2	4	1	-1	0	0	0	1	0	0
Subj. 5, obs. 2	5	1	-1	0	0	0	0	1	0
Subj. 6, obs. 2	6	1	-1	0	0	0	0	0	1
Contrast 1 ($\mathbf{C}_1'$)			$+1$	0	0	0	0	0	0
Contrast 2 ($\mathbf{C}_2'$)			-1	0	0	0	0	0	0

independent and symmetric errors, the problem is solved by means of random sign-flipping (Pesarin, 1995), using the well known Welch's v statistic, a particular case of the statistic G shown in Equation 3.10 (Table 3.8).

Table 3.8: Coding of the design matrix and exchangebility blocks for **Example 4**. As the group variances cannot be assumed to be the same, each group constitutes an EB and VG; sign flippings happen within block. The regressors $\mathbf{m}_1$ and $\mathbf{m}_2$ code for the experimental groups, $\mathbf{m}_3$ and $\mathbf{m}_4$ for age and sex.

Coded data (**Y**)	EB	VG	Model (**M**)			
			$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$	$\mathbf{m}_4$
Subject 1	1	1	1	0	a_1	s_1
Subject 2	1	1	1	0	a_2	s_2
Subject 3	1	1	1	0	a_3	s_3
Subject 4	1	1	1	0	a_4	s_4
Subject 5	1	1	1	0	a_5	s_5
Subject 6	1	1	1	0	a_6	s_6
Subject 7	2	2	0	1	a_7	s_7
Subject 8	2	2	0	1	a_8	s_8
Subject 9	2	2	0	1	a_9	s_9
Subject 10	2	2	0	1	a_{10}	s_{10}
Subject 11	2	2	0	1	a_{11}	s_{11}
Subject 12	2	2	0	1	a_{12}	s_{12}
Contrast 1 ($\mathbf{C}_1'$)			$+1$	-1	0	0
Contrast 2 ($\mathbf{C}_2'$)			-1	$+1$	0	0

Example 5: Variance as a confound Consider a study using FMRI to compare whether a given medication would modify the BOLD response associated with a certain attention task. The subjects are allocated in two groups, one receiving the drug, the other not. In this hypothetical example, the task is known to produce very robust and, on average, similar responses for male and female subjects, although it is also known that males tend to display more erratic signal changes, either very strong or very weak. Therefore, variances cannot be assumed to be homogeneous with respect to the sex of the subjects. To accommodate heteroscedasticity, two permutation blocks are defined according to sex, and each permutation matrix is constructed such that permutations only happen within each of these blocks (Table 3.9).

Table 3.9: Coding for **Example 5**. The different variances restrict exchangeability for within same sex only, and two exchangebility blocks are defined, for shuffling within block. The regressors $\mathbf{m}_1$ and $\mathbf{m}_2$ code for group (patients and controls), whereas $\mathbf{m}_3$ codes for sex.

Coded data (**Y**)	EB	VG	Model (**M**)		
			$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$
Subject 1	1	1	1	0	1
Subject 2	1	1	1	0	1
Subject 3	1	1	1	0	1
Subject 4	2	2	1	0	−1
Subject 5	2	2	1	0	−1
Subject 6	2	2	1	0	−1
Subject 7	1	1	0	1	1
Subject 8	1	1	0	1	1
Subject 9	1	1	0	1	1
Subject 10	2	2	0	1	−1
Subject 11	2	2	0	1	−1
Subject 12	2	2	0	1	−1
Contrast 1 ($\mathbf{C}_1'$)			1	−1	0
Contrast 2 ($\mathbf{C}_2'$)			−1	1	0

Example 6: Longitudinal study Consider a study to evaluate whether fractional anisoptropy (FA) would mature differently between boys and girls during middle childhood. Each child recruited to the study is examined three times, at the ages of 9, 10 and 11 years, and none of them are related in any known way. Permutation of observations within child cannot be considered, as the null hypothesis is not the one that FA itself would be zero, but instead, that there would be no changes in the value of FA along the three yearly observations. The permutations must, therefore, always keep in the same order the three observations. Blocks are defined as one per subject, each encompassing all the three observations, and permutation of each block as a whole is performed. If the variances cannot be assumed to be equal along time, one variance group can be defined per time point, otherwise all are assigned to the same VG. If there are nuisance variables to be considered, these can be included in the model and the procedure is performed using the same Freedman–Lane or Dekker strategies (Table 3.10).

Table 3.10: Coding of the design matrix, exchangebility blocks and variance groups for **Example 6**. Shufflings happen for the blocks as a whole, and variances are not assumed to be the same across all timepoints.

Coded data ($\mathbf{Y}$)	EB	VG	Model ($\mathbf{M}$)					
			$\mathbf{m}_1$	$\mathbf{m}_2$	$\mathbf{m}_3$	$\mathbf{m}_4$	$\mathbf{m}_5$	$\mathbf{m}_6$
Subject 1, Timepoint 1	1	1	a_{11}	0	1	0	0	0
Subject 1, Timepoint 2	1	2	a_{12}	0	1	0	0	0
Subject 1, Timepoint 3	1	3	a_{13}	0	1	0	0	0
Subject 2, Timepoint 1	2	1	a_{21}	0	0	1	0	0
Subject 2, Timepoint 2	2	2	a_{22}	0	0	1	0	0
Subject 2, Timepoint 3	2	3	a_{23}	0	0	1	0	0
Subject 3, Timepoint 1	3	1	0	a_{31}	0	0	1	0
Subject 3, Timepoint 2	3	2	0	a_{32}	0	0	1	0
Subject 3, Timepoint 3	3	3	0	a_{33}	0	0	1	0
Subject 4, Timepoint 1	4	1	0	a_{41}	0	0	0	1
Subject 4, Timepoint 2	4	2	0	a_{42}	0	0	0	1
Subject 4, Timepoint 3	4	3	0	a_{43}	0	0	0	1
Contrast 1 ($\mathbf{C}_1'$)			1	-1	0	0	0	0
Contrast 2 ($\mathbf{C}_2'$)			-1	1	0	0	0	0

3.4 Evaluation methods

3.4.1 Choice of the statistic

We conducted extensive simulations to study the behaviour of the common F statistic (Equation 3.3) as well as of the generalised G statistic (Equation 3.10), proposed here for use in neuroimaging, in various scenarios of balanced and unbalanced designs and variances for the variance groups. Some of the most representative of these scenarios are shown in Table 3.11. The main objective of the simulations was to assess whether these statistics would retain their distributions when the variances are not equal for each sample. Within each scenario, 3 or 5 different configurations of simulated variances were tested, pairwise, for the equality of distributions using the two-sample Kolmogorov–Smirnov test (KS) (Press et al., 1992), with a signficance level $\alpha = 0.05$, corrected for multiple testing within each scenario using the Bonferroni correction, as these tests are independent.

For each variance configuration, 1000 voxels containing normally distributed random noise, with zero expected mean, were simulated and tested for the null hypothesis of no difference between the means of the groups. The empirical distribution of the statistic for each configuration was obtained by pooling the results for the simulated voxels, then compared with the KS test. The process was repeated 1000 times, and the number of times in which the distributions were found to be significantly different from the others in the scame scenario was recorded. Confidence intervals (95%) were computed using the Wilson method (Wilson, 1927).

By comparing the distributions of the same statistic obtained in different variance settings, this evaluation strategy mimics what is observed when the variances for each voxel varies across space in the same imaging experiment. The statistic must be robust to these differences and retain its distributional properties, even if assessed non-parametrically, otherwise FWER using the distribution of the maximum statistic is compromised. The same applies for multiple testing that combines more than one imaging modality.

In addition, the same scenarios and variance configurations were used to assess the proportion of error type I and the power of the F and G statistics. To assess power, a simulated signal was added to each of the groups; for the scenarios with

Table 3.11: The eight different simulation scenarios, each with its own same sample sizes and different variances. The distributions of the statistic (F or G) for each pair of variance configuration within scenario were compared using the KS test. The letters in the last column (marked with a star, $\star$) indicate the variance configurations represented in the pairwise comparisons shown in Figure 3.4 and results shown in Table 3.12.

Simulation scenario	Sample sizes for each VG	Variances for each VG	$\star$
1	8, 4	5, 1	(a)
		1.2, 1	(b)
		1, 1	(c)
		1, 1.2	(d)
		1, 5	(e)
2	20, 5	5, 1	(a)
		1.2, 1	(b)
		1, 1	(c)
		1, 1.2	(d)
		1, 5	(e)
3	80, 30	5, 1	(a)
		1.2, 1	(b)
		1, 1	(c)
		1, 1.2	(d)
		1, 5	(e)
4	40, 30, 20, 10	15, 10, 5, 1	(a)
		3.6, 2.4, 1.2, 1	(b)
		1, 1, 1, 1	(c)
		1, 1.2, 2.4, 3.6	(d)
		1, 5, 10, 15	(e)
5	4, 4	1, 1	(a)
		1, 1.2	(b)
		1, 5	(c)
6	20, 20	1, 1	(a)
		1, 1.2	(b)
		1, 5	(c)
7	4, 4, 4, 4	1, 1, 1, 1	(a)
		1, 1.2, 2.4, 3.6	(b)
		1, 5, 10, 15	(c)
8	20, 20, 20, 20	1, 1, 1, 1	(a)
		1, 1.2, 2.4, 3.6	(b)
		1, 5, 10, 15	(c)

two groups, the true ψ was defined as $[0\ -1]'$, whereas for the scenarios with four groups, it was defined as $[0\ -0.33\ -0.67\ -1]'$. In either case, the null hypothesis was that the group means were all equal. Significance values were computed using 1000 permutations, with $\alpha = 0.05$, and 95% confidence intervals were calculated using the Wilson method.

3.4.2 Permutation strategies

We compared the 10 methods described in Table 3.2 simulating different regression scenarios. The design considered one regressor of interest, $\mathbf{x}_1$, and two regressors of no interest, $\mathbf{z}_1$ and $\mathbf{z}_2$, $\mathbf{z}_2$ being a column-vector of just ones (intercept). The simulation scenarios considered different sample sizes, $n = \{12, 24, 48, 96\}$; different combinations for continuous and categorical $\mathbf{x}_1$ and $\mathbf{z}_1$; different degrees of correlation between $\mathbf{x}_1$ and $\mathbf{z}_1$, $\rho = \{0, 0.8\}$; different sizes for the regressor of interest, $\beta_1 = \{0, 0.5\}$; and different distributions for the error terms, $\boldsymbol{\epsilon}$, as normal ($\mu = 0$, $\sigma^2 = 1$), uniform ($\left[-\sqrt{3},\ +\sqrt{3}\ \right]$), exponential ($\lambda = 1$) and Weibull ($\lambda = 1$, $k = 1/3$). The coefficients for the first regressor of no interest and for the intercept were kept constant as $\gamma_1 = 0.5$ and $\gamma_2 = 1$ respectively, and the distributions of the errors were shifted or scaled as needed to have expected zero mean and expected unit variance.

The continuous regressors were constructed as a linear trend ranging from -1 to $+1$ for $\mathbf{x}_1$, and the square of this trend, mean-centered, for $\mathbf{z}_1$. For this symmetric range around zero for $\mathbf{x}_1$, this procedure causes $\mathbf{x}_1$ and $\mathbf{z}_1$ to be orthogonal and uncorrelated. For the discrete regressors, a vector of $n/2$ ones and $n/2$ negative ones was used, the first $n/2$ values being only $+1$ and the remaining -1 for $\mathbf{x}_1$, whereas for $\mathbf{z}_1$, the first and last $n/4$ were -1 and the $n/2$ middle values were $+1$. This procedure also causes $\mathbf{x}_1$ and $\mathbf{z}_1$ to be orthogonal and uncorrelated. For each different configuration, 1000 simulated vectors $\mathbf{Y}$ were constructed as $\mathbf{Y} = [\mathbf{x}_1\ \mathbf{z}_1\ \mathbf{z}_2][\beta_1\ \gamma_1\ \gamma_2]' + \boldsymbol{\epsilon}$.

Correlation was introduced in the regression models through Cholesky decomposition of the desired correlation matrix $\mathbf{K}$, such that $\mathbf{K} = \mathbf{L}'\mathbf{L}$, then defining the regressors by multiplication by $\mathbf{L}$, i.e., $[\mathbf{x}_1^\rho\ \mathbf{z}_1^\rho] = [\mathbf{x}_1\ \mathbf{z}_1]\mathbf{L}$. The unpartitioned design matrix was constructed as $\mathbf{M} = [\mathbf{x}_1^\rho\ \mathbf{z}_1^\rho\ \mathbf{z}_2]$. A contrast $\mathbf{C} = [1\ 0\ 0]'$ was defined

to test the null hypothesis $\mathcal{H}_0 : \mathbf{C}'\boldsymbol{\psi} = \beta_1 = 0$. This contrast tests only the first column of the design matrix, so partitioning $\mathbf{M} = [\mathbf{X}\ \mathbf{Z}]$ using the scheme shown in Section 3.2.2 might seem unnecessary. However, we wanted to test also the effect of non-orthogonality between columns of the design matrix for the different permutation methods, with and without the more involved partitioning scheme shown in the Appendix. Permutation, sign flipping, and permutation with sign flipping were tested. Up to 1000 permutations and/or sign flippings were performed using CMC, being less when the maximum possible number of shufflings was not large enough. In these cases, all the permutations and/or sign flippings were performed exhaustively.

Error type I was computed using $\alpha = 0.05$ for configurations that used $\beta_1 = 0$. The other configurations were used to examine power. As previously, confidence intervals (95%) were estimated using the Wilson method.

3.5 Results

3.5.1 Choice of the statistic

Figure 3.4 shows heatmaps with the results of the pairwise comparisons between variance configurations, within each of the simulation scenarios presented in Table 3.11, using either F or G statistic. For unbalanced scenarios with only two samples (simulation scenarios 1 to 3), and with modest variance differences between groups (configurations b to d), the F statistic often retained its distributional properties, albeit less often than the G statistic. For large variance differences, however, this relative stability was lost for F, but not for G (a and e). Moreover, the inclusion of more groups (scenario 4), with unequal sample sizes, caused the distribution of the F statistic to be much more sensitive to heteroscedasticity, such that almost always the KS test identified different distributions across different variance configurations. The G statistic, on the other hand, remained robust to heteroscedasticity even in these cases.

In balanced designs, either with two (simulation scenarios 5 and 6) or more (scenarios 7 and 8) groups, the F statistic had a better behaviour than in unbalanced cases. For two samples of the same size, there is no difference between F

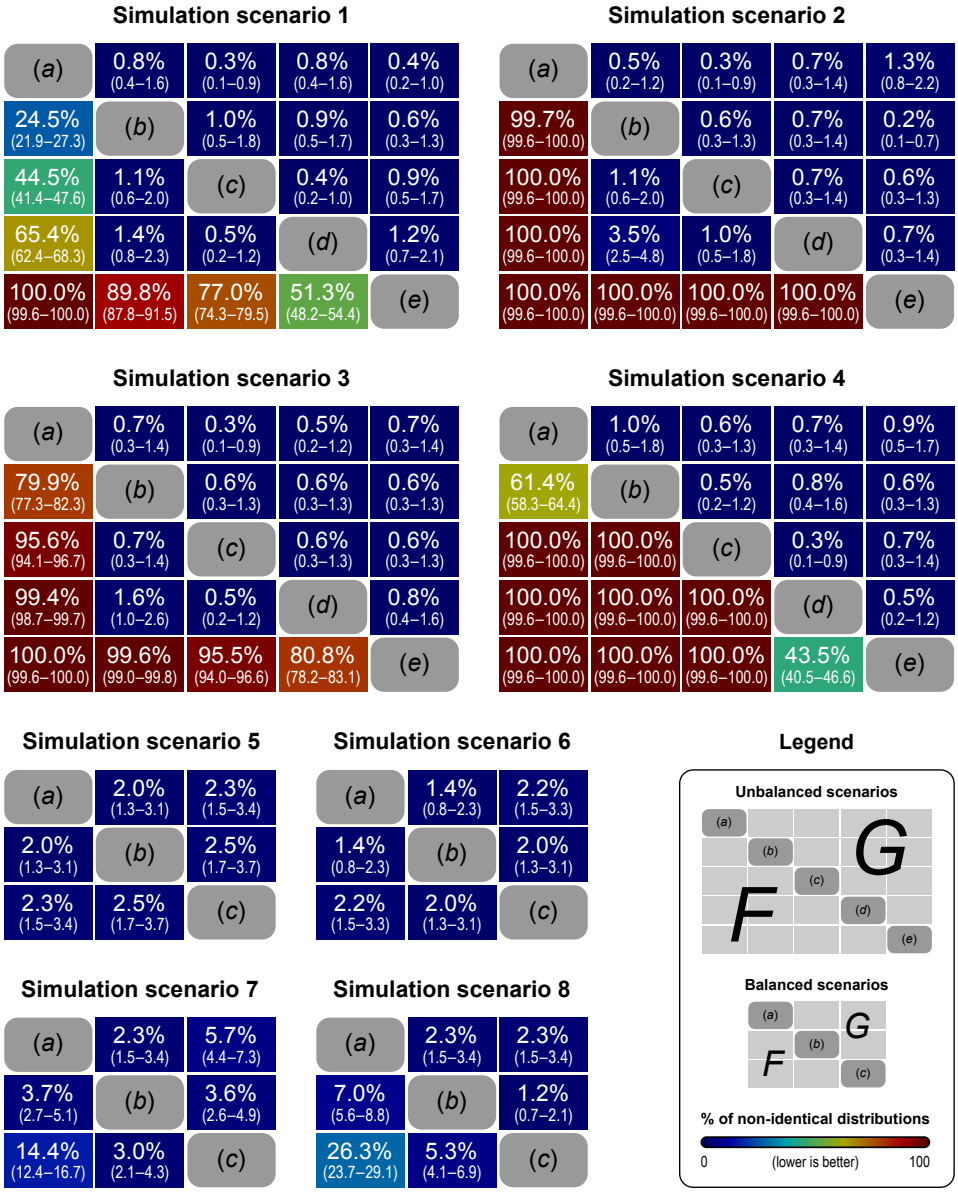


Figure 3.4: Heatmaps for the comparison of the distributions obtained under different variance settings for identical sample sizes. In each map, the cells below the main diagonal contain the results for the pairwise F statistic, and above, for the G statistic. The percentages refer to the fraction of the 1000 tests in which the distribution of the statistic for one variance setting was found different than for another in the same simulation scenario. Each variance setting is indicated by letters (a–e), corresponding to the same letters in Table 3.11. Smaller percentages indicate robustness of the statistic to heteroscedasticity. Confidence intervals (95%) are shown in parenthesis.

and G. For more than two groups, the G statistic behaved consistently better than F, particularly for large variance differences.

These results suggest that the G statistic is more appropriate under heteroscedasticity, with balanced or unbalanced designs, as it preserves its distributional properties, indicating more adequacy for use with neuroimaging. The F statistic, on the other hand, does not preserve pivotality and can, nonetheless, be used under heteroscedasticity when the groups have the same size.

With respect to error type I, both F and G resulted in similar amount of false positives when assessed non-parametrically. The G yielded generally higher power than F, particularly in the presence of heteroscedasticity and with unequal sample sizes. These results are presented in Table 3.12.

3.5.2 Permutation strategies

The different simulation parameters allowed 1536 different regression scenarios, being 768 without signal and 768 with signal; a summary is shown in Table 3.13, and some of the most representative in Table 3.14. In "well behaved" scenarios, i.e., large number of observations, orthogonal regressors and normally distributed errors, all methods tended to behave generally well, with adequate control over type I error and fairly similar power. However, performance differences between the permutation strategies shown in Table 3.2 became more noticeable as the sample sizes were decreased and skewed errors were introduced.

Some of the methods are identical to each other in certain circumstances. If $\mathbf{X}$ and $\mathbf{Z}$ are orthogonal, Draper–Stoneman and Dekker are equivalent. Likewise under orthogonality, Still–White produces identical regression coefficients as Freedman–Lane, although the statistic will only be the same if the loss in degrees of freedom due to $\mathbf{Z}$ is taken into account, something not always possible when the data has already been residualised and no information about the original nuisance variables is available. Nonetheless, the two methods remain asymptotically equivalent as the number of observations diverges from the number of nuisance regressors.

Table 3.12: Proportion of error type I and power (%) for the statistics F and G in the various simulation scenarios and variance configurations shown in Table 3.11. Confidence intervals (95%) are shown in parenthesis.

Simulation scenario	$\star$	Proportion of error type I		Power	
		F	G	F	G
1	(a)	5.9 (4.6–7.5)	6.1 (4.8–7.8)	20.1 (17.7–22.7)	23.8 (21.3–26.5)
	(b)	4.9 (3.7–6.4)	5.3 (4.1–6.9)	28.3 (25.6–31.2)	31.9 (29.1–34.9)
	(c)	4.7 (3.6–6.2)	4.5 (3.4–6.0)	29.3 (26.6–32.2)	32.6 (29.8–35.6)
	(d)	4.9 (3.7–6.4)	4.6 (3.5–6.1)	29.9 (27.1–32.8)	32.0 (29.2–35.0)
	(e)	3.9 (2.9–5.3)	4.1 (3.0–5.5)	14.0 (12.0–16.3)	14.1 (12.1–16.4)
2	(a)	6.7 (5.3–8.4)	6.6 (5.2–8.3)	29.1 (26.4–32.0)	38.3 (35.3–41.4)
	(b)	5.0 (3.8–6.5)	4.6 (3.5–6.1)	42.4 (39.4–45.5)	48.8 (45.7–51.9)
	(c)	5.0 (3.8–6.5)	5.8 (4.5–7.4)	44.6 (41.6–47.7)	48.9 (45.8–52.0)
	(d)	6.1 (4.8–7.8)	6.2 (4.9–7.9)	42.3 (39.3–45.4)	46.7 (43.6–49.8)
	(e)	5.9 (4.6–7.5)	6.2 (4.9–7.9)	19.5 (17.2–22.1)	19.0 (16.7–21.6)
3	(a)	5.2 (4.0–6.8)	5.0 (3.8–6.5)	90.4 (88.4–92.1)	92.3 (90.5–93.8)
	(b)	4.9 (3.7–6.4)	5.1 (3.9–6.6)	99.7 (99.1–99.9)	99.8 (99.3–100)
	(c)	6.3 (5.0–8.0)	6.2 (4.9–7.9)	99.8 (99.3–100)	99.8 (99.3–100)
	(d)	4.4 (3.3–5.9)	4.4 (3.3–5.9)	99.6 (99.0–99.8)	99.6 (99.0–99.8)
	(e)	4.4 (3.3–5.9)	4.4 (3.3–5.9)	72.9 (70.1–75.6)	72.9 (70.1–75.6)
4	(a)	6.4 (5.0–8.1)	5.7 (4.4–7.3)	10.2 (8.5–12.2)	19.4 (17.1–22.0)
	(b)	5.3 (4.1–6.9)	5.6 (4.3–7.2)	37.8 (34.9–40.9)	45.6 (42.5–48.7)
	(c)	5.7 (4.4–7.3)	4.9 (3.7–6.4)	72.2 (69.3–74.9)	74.9 (72.1–77.5)
	(d)	3.1 (2.2–4.4)	3.7 (2.7–5.1)	34.6 (31.7–37.6)	44.6 (41.6–47.7)
	(e)	4.5 (3.4–6.0)	4.2 (3.1–5.6)	9.7 (8.0–11.7)	15.7 (13.6–18.1)
5	(a)	4.3 (3.2–5.7)	4.3 (3.2–5.7)	29.9 (27.1–32.8)	29.9 (27.1–32.8)
	(b)	4.3 (3.2–5.7)	4.3 (3.2–5.7)	30.6 (27.8–33.5)	30.6 (27.8–33.5)
	(c)	6.9 (5.5–8.6)	6.9 (5.5–8.6)	14.5 (12.5–16.8)	14.5 (12.5–16.8)
6	(a)	3.3 (2.4–4.6)	3.3 (2.4–4.6)	92.6 (90.8–94.1)	92.6 (90.8–94.1)
	(b)	4.4 (3.3–5.9)	4.4 (3.3–5.9)	90.5 (88.5–92.2)	90.5 (88.5–92.2)
	(c)	4.4 (3.3–5.9)	4.4 (3.3–5.9)	53.7 (50.6–56.8)	53.7 (50.6–56.8)
7	(a)	5.6 (4.3–7.2)	5.5 (4.3–7.1)	11.0 (9.2–13.1)	8.8 (7.2–10.7)
	(b)	5.2 (4.0–6.8)	4.4 (3.3–5.9)	6.5 (5.1–8.2)	7.8 (6.3–9.6)
	(c)	5.7 (4.4–7.3)	4.8 (3.6–6.3)	5.8 (4.5–7.4)	6.9 (5.5–8.6)
8	(a)	4.6 (3.5–6.1)	4.5 (3.4–6.0)	78.7 (76.1–81.1)	78.1 (75.4–80.6)
	(b)	4.6 (3.5–6.1)	5.6 (4.3–7.2)	40.7 (37.7–43.8)	45.5 (42.4–48.6)
	(c)	4.7 (3.6–6.2)	4.8 (3.6–6.3)	11.6 (9.8–13.7)	19.3 (17.0–21.9)

Table 3.13: A summary of the results for the 1536 simulations with different parameters. The amount of error type I is calculated for the 768 simulations without signal (β_1 = 0), whereas the power was calculated for the remaining 768 simulations with signal (β_1 = 0.5). Confidence intervals (CI) are at 95%.

Method	Proportion of error type I			Average power
	Within CI	Below CI	Above CI	
Draper–Stoneman	86.33%	8.20%	5.47%	72.96%
Still–White	67.84%	14.58%	17.58%	71.82%
Freedman–Lane	88.67%	8.46%	2.86%	73.09%
ter Braak	83.59%	11.07%	5.34%	73.38%
Kennedy	77.60%	1.04%	21.35%	74.81%
Manly	73.31%	15.89%	10.81%	73.38%
Dekker	89.32%	7.81%	2.86%	72.90%
Huh–Jhun	85.81%	9.24%	4.95%	71.62%
Parametric	77.47%	14.84%	7.68%	72.73%

When the amount of errors is below the nominal level (here, $\alpha = 0.05$), the test is said to be *conservative*. If above, it is *invalid*.

Sample size Increasing the sample size had the effect of approaching the error rate to closer to the nominal level $\alpha = 0.05$ for all methods in virtually all parameter configurations. For small samples, most methods were slightly conservative, whereas Still–White and Kennedy were anticonservative and often invalid, particularly if the distributions of the errors were skewed.

Continuous or categorical regressors of interest For all methods, using continuous or categorical regressors of interest did not produce remarkable differences

Table 3.14: *(Page 108)* Proportion of error type I (for α = 0.05), for some representative of the 768 simulation scenarios that did not have signal, using the different permutation methods, and with G as the statistic in the absence of EB (so, equivalent to the F statistic). Confidence intervals (95%) are shown in parenthesis.

N: number of observations; $\mathbf{x}_1$ and $\mathbf{z}_1$: regressors of interest and of no interest, respectively, being either continuous (C) or discrete (D). ρ: correlation between $\mathbf{x}_1$ and $\mathbf{z}_1$; $\nsim$: model partitioned or not (using the Beckmann et al. (2001) scheme, shown in Section 3.2.2); ϵ: distribution of the simulated errors, which can be normal ($\mathcal{N}$), uniform ($\mathcal{U}$), exponential ($\mathcal{E}$) or Weibull ($\mathcal{W}$); EE: errors treated as exchangeable; ISE: errors treated as independent and symmetric. The methods are the same shown in Table 3.2: Draper–Stoneman (D–S), Still–White (S–W), Freedman–Lane (F–L), ter Braak (tB), Kennedy (K), Manly (M), Huh–Jhun (H–J), Dekker (D) and parametric (P), the last not using permutations.

Simulation parameters								Proportion of error type I (%)								
N	$\mathbf{x}_1$	$\mathbf{z}_1$	ρ	✂	ϵ	EE	ISE	D–S	S–W	F–L	tB	K	M	S	H–J	P
12	C	C	0	✗	$\mathcal{N}$	✓	✗	4.9 (3.7–6.4)	5.3 (4.1–6.9)	5.1 (3.9–6.6)	5.3 (4.1–6.9)	5.3 (4.1–6.9)	5.0 (3.8–6.5)	4.9 (3.7–6.4)	4.7 (3.6–6.2)	4.4 (3.3–5.9)
12	C	C	0	✗	$\mathcal{U}$	✓	✓	5.3 (4.1–6.9)	6.9 (5.5–8.6)	5.1 (3.9–6.6)	5.2 (4.0–6.8)	6.9 (5.5–8.6)	5.8 (4.5–7.4)	5.3 (4.1–6.9)	5.2 (4.0–6.8)	4.6 (3.5–6.1)
12	C	C	0	✗	$\mathcal{W}$	✓	✗	5.9 (4.6–7.5)	6.5 (5.1–8.2)	5.2 (4.0–6.8)	5.4 (4.2–7.0)	6.5 (5.1–8.2)	5.0 (3.8–6.5)	5.9 (4.6–7.5)	5.4 (4.2–7.0)	8.3 (6.7–10.2)
12	C	C	0	✗	$\mathcal{E}$	✓	✓	5.3 (4.1–6.9)	6.9 (5.5–8.6)	5.1 (3.9–6.6)	4.7 (3.6–6.2)	6.9 (5.5–8.6)	5.0 (3.8–6.5)	5.3 (4.1–6.9)	4.8 (3.6–6.3)	5.7 (4.4–7.3)
12	C	C	0.8	✗	$\mathcal{N}$	✓	✗	4.4 (3.3–5.9)	3.6 (2.6–4.9)	5.1 (3.9–6.6)	5.2 (4.0–6.8)	5.8 (4.5–7.4)	4.8 (3.6–6.3)	5.1 (3.9–6.6)	4.4 (3.3–5.9)	4.4 (3.3–5.9)
12	C	C	0.8	✗	$\mathcal{W}$	✓	✗	1.5 (0.9–2.5)	1.2 (0.7–2.1)	4.8 (3.6–6.3)	5.2 (4.0–6.8)	6.5 (5.1–8.2)	4.9 (3.7–6.4)	5.8 (4.5–7.4)	5.8 (4.5–7.4)	8.5 (6.9–10.4)
12	C	C	0.8	✗	$\mathcal{N}$	✓	✓	5.5 (4.2–7.1)	5.4 (4.2–7.0)	4.9 (3.7–6.4)	5.4 (4.2–7.0)	7.5 (6.0–9.3)	4.8 (3.6–6.3)	4.8 (3.6–6.3)	5.8 (4.5–7.4)	4.6 (3.5–6.1)
12	C	C	0.8	✓	$\mathcal{N}$	✓	✓	5.1 (3.9–6.6)	7.2 (5.8–9.0)	5.4 (4.2–7.0)	4.3 (3.2–5.7)	7.2 (5.8–9.0)	5.2 (4.0–6.8)	5.1 (3.9–6.6)	4.6 (3.5–6.1)	4.6 (3.5–6.1)
12	C	D	0	✗	$\mathcal{W}$	✓	✗	5.6 (4.3–7.2)	6.8 (5.4–8.5)	5.4 (4.2–7.0)	4.7 (3.6–6.2)	6.8 (5.4–8.5)	4.0 (3.0–5.4)	5.6 (4.3–7.2)	3.7 (2.7–5.1)	8.9 (7.3–10.8)
12	C	D	0	✗	$\mathcal{N}$	✓	✗	3.9 (2.9–5.3)	4.9 (3.7–6.4)	3.9 (2.9–5.3)	4.0 (3.0–5.4)	4.9 (3.7–6.4)	4.3 (3.2–5.7)	3.9 (2.9–5.3)	4.2 (3.1–5.6)	3.7 (2.7–5.1)
12	C	D	0	✗	$\mathcal{W}$	✗	✓	2.9 (2.0–4.1)	4.3 (3.2–5.7)	2.6 (1.8–3.8)	2.8 (1.9–4.0)	4.3 (3.2–5.7)	14.1 (12.1–16.4)	2.9 (2.0–4.1)	16.4 (14.2–18.8)	9.0 (7.4–10.9)
12	D	D	0	✗	$\mathcal{W}$	✓	✗	3.2 (2.3–4.5)	4.6 (3.5–6.1)	2.2 (1.5–3.3)	2.0 (1.3–3.1)	4.6 (3.5–6.1)	3.8 (2.8–5.2)	3.2 (2.3–4.5)	2.6 (1.8–3.8)	0.5 (0.2–1.2)
24	C	C	0.8	✗	$\mathcal{N}$	✓	✗	4.4 (3.3–5.9)	3.5 (2.5–4.8)	4.3 (3.2–5.7)	4.4 (3.3–5.9)	4.9 (3.7–6.4)	4.4 (3.3–5.9)	4.3 (3.2–5.7)	4.5 (3.4–6.0)	4.4 (3.3–5.9)
24	D	D	0	✗	$\mathcal{N}$	✓	✗	5.0 (3.8–6.5)	5.4 (4.2–7.0)	5.1 (3.9–6.6)	5.1 (3.9–6.6)	5.4 (4.2–7.0)	4.9 (3.7–6.4)	5.0 (3.8–6.5)	4.5 (3.4–6.0)	5.0 (3.8–6.5)
24	D	D	0	✗	$\mathcal{U}$	✓	✗	6.2 (4.9–7.9)	6.6 (5.2–8.3)	6.3 (5.0–8.0)	5.9 (4.6–7.5)	6.6 (5.2–8.3)	5.5 (4.2–7.1)	6.2 (4.9–7.9)	5.9 (4.6–7.5)	5.8 (4.5–7.4)
24	D	D	0.8	✗	$\mathcal{U}$	✓	✗	4.9 (3.7–6.4)	1.8 (1.1–2.8)	5.1 (3.9–6.6)	4.8 (3.6–6.3)	5.4 (4.2–7.0)	5.1 (3.9–6.6)	5.2 (4.0–6.8)	5.7 (4.4–7.3)	5.4 (4.2–7.0)
48	C	C	0	✗	$\mathcal{N}$	✗	✓	4.9 (3.7–6.4)	5.4 (4.2–7.0)	5.0 (3.8–6.5)	5.6 (4.3–7.2)	5.4 (4.2–7.0)	3.8 (2.8–5.2)	4.9 (3.7–6.4)	6.0 (4.7–7.6)	5.0 (3.8–6.5)
48	C	C	0.8	✓	$\mathcal{U}$	✓	✗	5.1 (3.9–6.6)	5.4 (4.2–7.0)	5.0 (3.8–6.5)	5.7 (4.4–7.3)	5.4 (4.2–7.0)	5.2 (4.0–6.8)	5.1 (3.9–6.6)	5.6 (4.3–7.2)	5.6 (4.3–7.2)
48	C	C	0.8	✓	$\mathcal{N}$	✓	✗	4.6 (3.5–6.1)	4.8 (3.6–6.3)	4.7 (3.6–6.2)	4.7 (3.6–6.2)	4.8 (3.6–6.3)	4.6 (3.5–6.1)	4.6 (3.5–6.1)	4.4 (3.3–5.9)	4.5 (3.4–6.0)
48	C	D	0	✗	$\mathcal{E}$	✗	✓	5.4 (4.2–7.0)	5.7 (4.4–7.3)	5.1 (3.9–6.6)	5.5 (4.2–7.1)	5.7 (4.4–7.3)	9.2 (7.6–11.2)	5.4 (4.2–7.0)	4.3 (3.2–5.7)	5.1 (3.9–6.6)
48	C	D	0.8	✗	$\mathcal{E}$	✓	✗	5.5 (4.2–7.1)	0.3 (0.1–0.9)	5.0 (3.8–6.5)	5.0 (3.8–6.5)	5.0 (3.8–6.5)	4.9 (3.7–6.4)	5.0 (3.8–6.5)	5.0 (3.8–6.5)	4.9 (3.7–6.4)
96	C	C	0	✗	$\mathcal{N}$	✓	✓	5.1 (3.9–6.6)	5.3 (4.1–6.9)	5.1 (3.9–6.6)	4.9 (3.7–6.4)	5.3 (4.1–6.9)	4.6 (3.5–6.1)	5.1 (3.9–6.6)	5.3 (4.1–6.9)	4.9 (3.7–6.4)
96	C	C	0.8	✗	$\mathcal{N}$	✗	✓	5.0 (3.8–6.5)	3.6 (2.6–4.9)	5.0 (3.8–6.5)	4.8 (3.6–6.3)	5.2 (4.0–6.8)	4.4 (3.3–5.9)	5.1 (3.9–6.6)	5.2 (4.0–6.8)	4.9 (3.7–6.4)
96	D	C	0	✗	$\mathcal{W}$	✓	✗	4.9 (3.7–6.4)	5.2 (4.0–6.8)	4.7 (3.6–6.2)	4.8 (3.6–6.3)	5.2 (4.0–6.8)	4.5 (3.4–6.0)	4.9 (3.7–6.4)	3.9 (2.9–5.3)	3.6 (2.6–4.9)

See caption in page 107.

in the observed proportions of type I error, except if the distribution of the errors was skewed and sign flipping was used (in violation of assumptions), in which case Manly and Huh–Jhun methods showed erratic control over the amount of errors.

Continuous or categorical nuisance regressors The presence of continuous or categorical nuisance variables did not substantially interfere with either control over error type I or power, for any of the methods, except in the presence of correlated regressors.

Degree of non-orthogonality and partitioning All methods provided relatively adequate control over error type I in the presence of a correlated nuisance regressor, except Still–White (conservative) and Kennedy (inflated rates). The partitioning scheme mitigated the conservativeness of the former, and the anticonservativeness of the latter.

Distribution of the errors Different distributions did not substantially improve or worsen error rates when using permutation alone. Still–White and Kennedy tended to fail control over error type I in virtually all situations. Sign flipping alone, when used with asymmetric distributions (in violation of assumptions), required larger samples to allow approximately exact control over the amount of error type I. In these cases, and with small samples, the methods Draper–Stoneman, Manly and Huh–Jhun tended to display erratic behaviour, with extremes of conservativeness and anticonservativeness depending on the other simulation parameters. The same happened with the parametric method. Freedman–Lane and Dekker methods, on the other hand, tended to have a relatively constant and somewhat conservative behaviour in these situations. Permutation combined with sign-flipping generally alleviated these issues where they were observed.

From all the methods, the Freedman–Lane and Dekker were those that performed better in most cases, and with their 95% confidence interval covering the desired error level of 0.05 more often than any of the other methods. The Still–White and Kennedy methods did not generally control the error type I for most of the simulation parameters, particularly for smaller sample sizes. On the other hand, with a few exceptions, the Freedman–Lane and the Dekker methods effect-

ively controlled the error rates in most cases, even with skewed errors and sign-flipping, being, at worst, conservative or only slightly above the nominal level. All methods were, overall, similarly powerful, with only marginal differences among those that were on average valid.

3.6 Discussion

Ideally, criteria to accept or reject a given hypothesis should be sensitive to changes in the parameters of interest (powerful), and insensitive to changes in nuisance factors (robust). As pointed out long ago by Box and Andersen (1955), the assumptions on which parametric tests are built are such that the first criterion is generally satisfied, albeit not necessarily the second. Many non-parametric tests, on the other hand, are constructed such that most or all these assumptions are not demanded, satisfying the second criterion, but not necessarily the first. For current applications in neuroimaging, however, this compromise between robustness and power gains new contours and a different balance. First, in neuroimaging it is necessary to address the multiple testing problem, in which one or more tests are applied to each of thousands of points (commonly voxels, vertices or faces) of the image representation of the brain. Parametric methods require the introduction of an even larger set of assumptions to deal with multiple testing. Second, different imaging modalities not necessarily follow the same set of assumptions regarding distributions under the null at each test, neither for the covariance between tests across the brain, so that those that might be acceptably used with one method, may cause others to be invalid. Third, under non-random sampling, as common in case-control studies, the very presence of the features under investigation (such as a disorder) may compromise the assumptions on which parametric tests depend. For all these reasons, parametric methods, despite common use, are more likely to fail as candidates to provide a general statistical framework for the current variety of imaging modalities for research applications, where not only the assumptions may not be met, but also where robustness may be seen as a key factor. Permutation methods are a viable alternative, flexible enough to accommodate several experimental needs. Further to all this, our simulations showed similar and sometimes higher power compared to the parametric approach.

3.6.1 Permutation tests

Permutation tests require very few assumptions about the data and, therefore, can be applied in a wider variety of situations than parametric tests. None of the most common parametric assumptions need to hold for non-parametric tests to be valid. The assumptions that are eschewed include, for instance, the need of normality for the error terms, the need of homoscedasticity and the need of random sampling. With a very basic knowledge of sample properties or of the study design, errors can be treated as exchangeable (EE) and/or independent and symmetric (ISE) and inferences that otherwise would not be possible with parametric methods become feasible. Furthermore, permutation tests permit the use of the very same regression and hypothesis testing framework, even with disparate imaging modalities, without the need to verify the validity of parametric assumptions for each of them. The ISE can be an alternative to EE when the errors themselves can be considered exchangeable, but the design is not affected by permutations, as for one-sample tests. And if the assumptions for EE and ISE are both met, permutation and sign flipping can both be performed to construct the empirical distribution.

The justification for permutation tests has, moreover, more solid foundations than their parametric counterparts. While the validity of parametric tests rely on random sampling, permutation tests rely on the idea of random allocation of experimental units, with no reference to any underlying population. This aspect has a key importance in biomedical research — including neuroimaging — where only a small minority of studies effectively use random population sampling. Most experimental studies need to use the subjects that are available in a given area, and who accept to participate (e.g. patients of a hospital or students of an university near where the MRI equipment is installed). True random sampling is rarely achieved in real applications because, often and for different reasons, selection criteria are not truly unbiased (Ludbrook and Dudley, 1998; Pesarin and Salmaso, 2010a). Non-parametric methods allow valid inferences to be performed in these scenarios.

3.6.2 Pivotal statistics

In addition, permutation methods have the remarkable feature of allowing the use of non-standard statistics, or for which closed mathematical forms have not been derived, even asymptotically. Statistics that can be used include, for instance, those based on ranks of observations (Brunner and Munzel, 2000; Rorden et al., 2007), derived from regression methods other than least squares (Cade and Richards, 1996) or that are robust to outliers (Theil, 1950; Sen, 1968). For imaging applications, statistics that can be considered include the pseudo-t statistic after variance smoothing (Holmes et al., 1996), the mass of connected voxels (Bullmore et al., 1999), threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009), as well as cases in which the distribution of the statistic may lie in a gradient between distributions, each of them with known analytical forms, such as the distribution of surface area, as demonstrated in Chapter 2 (published as Winkler et al., 2012). The only requirement, in the context of neuroimaging, is that these statistics retain their distributional properties irrespective to the (unknown) population parameters.

Indeed, a large part of the voluminous literature on statistical tests when the errors cannot be assumed to be homoscedastic is concerned with the identification of the asymptotic distribution of the statistics, its analytical form, and the consequences of experimental scenarios that include unbalancedness and/or small samples. This is true even considering that in parametric settings, the statistics are invariably chosen such that their sampling distribution is independent of underlying and unknown population parameters. Permutation tests render all these issues irrelevant, as the asymptotic properties of the distributions do not need to be ascertained. For imaging, all that is needed is that the distribution remains invariant to unknown population parameters, i.e., the statistic needs to be pivotal. Parameters of the distribution proper do not need to be known, nor the distribution needs to be characterised analytically. The proposed statistic G, being a generalisation over various tests that have their niche applications in parametric settings, is appropriate for use with the general linear model and with a permutation framework, for being pivotal and easily implementable using simple matrix operations. Moreover, as the simulations showed, this statistic is not less powerful than the commonly

used F statistic.

3.6.3 Permutation strategies

From the different permutation strategies presented in Table 3.2, the Freedman–Lane and the Dekker methods provided the most adequate control of type I error across the various simulation scenarios. This is in line with the study by Anderson and Legendre (1999), who found that the Freedman–Lane method is the most accurate and powerful in various different models. The Dekker method was a somewhat positive surprise, not only for the overall very good performance in our simulations, but also because this method had not been extensively evaluated in previous literature, is computationally simple, and has an intuitive appeal.

Welch (1990) commented that the Freedman–Lane procedure would violate the ancillarity principle, as the permutation procedure would destroy the relationship between $\mathbf{X}$ and $\mathbf{Z}$, even if these are orthogonal. Notwithstanding, even with ancillarity violated, this and other methods perform satisfactorily well.

Freedman and Lane (1983) described their method as having a "non-stochastic" interpretation, and so, that the computed p-value would be a descriptive statistic. On the contrary, we share the same view expressed by Anderson and Legendre (1999), that the rationale for the test and the procedure effectively produces a p-value that can be interpreted as a true probability for the underlying model.

Regarding differences between the methods, and even though for this study we did not evaluate the effect of extremely strong signals or of outliers, it is worth commenting that previous research have shown that the Freedman–Lane method is relatively robust to the presence of extreme outliers, whereas the ter Braak tends to become more conservative in these cases (Anderson and Legendre, 1999). The ter Braak method, however, was shown to be more robust to extremely strong signals in the data, situations in which signal may "leak" into the permutation distribution with the Freedman–Lane method (Salimi-Khorshidi et al., 2011).

It should be noted that the Still–White method, as implemented for these simulations, used for the regression the model containing only the regressors of interest when computing the statistic. as shown in Table 3.2. It is done in this way to emulate what probably is its more common use, i.e., rearrange the data that has

already been residualised from nuisance, and when the nuisance regressors are no longer available. Had the full model been used when computing the statistic, it is possible that this method might have performed somewhat similarly as Freedman–Lane, specially for larger samples. Moreover, neither the original publication (Still and White, 1981), nor a related method published shortly after (Levin and Robbins, 1983), specify how the degrees of freedom should be treated when computing the statistic in a generic formulation as we present here.

Finally, although non-parametric methods are generally considered less powerful than their parametric counterparts, we found in the simulations performed that most of the permutation methods are not substantially less powerful than the parametric method, and sometimes are even more powerful, even when the assumptions of the latter are met. With the availability of computing power and reliable software implementation, there is almost no reason for not using these permutation methods.

3.7 Chapter conclusion

We presented a generic framework that allows permutation inference using the general linear model with experimental designs of arbitrary complexity, and which depends only on the weak requirements of exchangeable or independent and symmetric errors, which define permutations, sign flippings, or both. Structured dependence between observations is addressed through the definition of exchangeability blocks. We also proposed a statistic that is robust to heteroscedasticity, can be used for multiple-testing correction, and can be implemented easily with matrix operations. Based on evaluations, we recommend the Freedman–Lane and the Dekker methods to construct the empirical distribution, and use Freedman–Lane in the randomise algorithm (3.2.6).

Chapter 4

Combined inference

4.1 Introduction

In this chapter we show that permutation tests can provide a common solution to seemingly disparate problems that arise when dealing with multiple imaging measurements. These problems refer to the multiplicity of tests, and to the combination of information across multiple modalities for joint inference. We begin by describing each of these problems separately, then show how they are related, and offer a complete and generic solution that can accommodate a myriad of designs that can mix imaging and non-imaging data. We also present an algorithm that has with amenable computational demands for treating these problems.

4.1.1 Multiple tests — but not the usual multiplicity

Because in neuroimaging one statistical test is typically performed at each of many thousands of imaging units (e.g., voxels or vertices), the problems related to such multiplicity of tests were recognised almost as early as these techniques were developed (for pioneering examples, see Fox et al., 1988; Friston et al., 1991). There is now a comprehensive body of literature on multiple testing correction methods that include those based on the random field theory, on permutation tests, as well as on other strategies that control the familywise error rate (FWER) or the false discovery rate (FDR) (for reviews, see Nichols and Hayasaka, 2003; Nichols, 2012). However, the multiplicity of tests in neuroimaging can appear in other ways that

are less explicit, and most importantly, that have not been fully appreciated or made available in software packages. In the context of the general linear model (GLM, Scheffé, 1959), these *other* multiple tests include:

A. *Multiple hypotheses in the same model:* Testing more than one hypothesis regarding a set of explanatory variables. An example is testing the effects of multiple variables, such as presence of a disease along with its duration, some clinical score, age and/or sex of the subjects, on a given imaging measurement, such as maps from functional magnetic resonance imaging (FMRI) experiments.

B. *Multiple pairwise group comparisons:* Often an initial global (omnibus) test is performed, such as an F-test in the context of analysis of variance (ANOVA), and if this test is significant, subsequent (post hoc) tests are performed to verify which pairwise difference(s) drove the global result, thus introducing a multiple comparisons problem.

C. *Multiple models:* Testing more than one set of explanatory variables on one given dataset, that is, assembling and testing more than one design matrix, each with its own set of regressors, which may differ across designs, and each with its own set of contrasts. An example is interrogating the effect of distinct seeds, one at a time, in a resting-state FMRI experiment; another is in an imaging genetics experiment, testing multiple candidate polymorphisms.

D. *Multiple modalities:* Testing separately, in the same study, more than one imaging modality as the response variable, such as FMRI and positron-emission tomography (PET), or different metrics from the same modality, such as various measurements from diffusion tensor imaging (DTI), as fractional anisotropy (FA), mean diffusivity (MD), or radial diffusivity (RD), or the effect of various networks identified using independent component analysis (ICA).

E. *Imaging and non-imaging:* Testing separately, in the same study, imaging and non-imaging measurements as response variables. An example is studying group effects on FMRI and on behavioural or cognitive scores, such as IQ, or disease severity scores, among countless other non-imaging measurements.

f. *Multiple processing pipelines:* Testing the same imaging modality multiple times, each time after a different processing pipeline, such as using filters with different widths for smoothing, or using different strategies for registration to a common space.

g. *Multiple multivariate analyses:* Testing more than one multivariate hypothesis with the GLM in repeated measurements designs, such as in profile analyses, in which the same data allows various different hypotheses about the relationships between explanatory and response variables.

In all these cases, the multiple tests cannot be assumed to be independent, so that the simple FWER correction using the conventional Bonferroni method risks a considerable loss in power. Modelling the degree of dependence between these tests can be a daunting task, and be suboptimal by invariably requiring the introduction of assumptions about the data, which, if at all valid, may not be sufficient. By contrast, robust, generic, multi-step procedures, which do not depend as much on assumptions, or on independence among tests, such as the Benjamini–Hochberg procedure that controls the false discovery rate (FDR) (Benjamini and Hochberg, 1995; Genovese et al., 2002), do not guarantee that the spatial relationship between voxels or vertices within test is preserved when applied across *these* multiple tests, therefore being not as useful as in other settings. More specifically, the difficulty relates to correcting across various distinct imaging tests, while maintaining control across space within any given test, as opposed to controlling just within a single imaging test as commonly done. For the same reason, various multiple testing approaches that are applicable to many particular cases, can hardly be used for the problems we discuss here; extensive details on these tests can be found in Hochberg and Tamhane (1987) and in Hsu (1996).

We call the multiple tests that arise in situations as those listed above "multiple testing problem II" (MTP-II), to allow a distinction from the usual multiple testing problem due to the many voxels/vertices/faces that constitute an image, which we denote "multiple testing problem I" (MTP-I). Methods that can be used in neuroimaging for the MTP-I not always can be considered for the MTP-II, a problem that has remained largely without treatment; for two rare counter examples in which the MTP-II *was* considered, we point to the studies by Licata et al. (2013)

and Abou Elseoud et al. (2014).

4.1.2 Combination of imaging modalities

Acquisition of multiple imaging modalities on the same subjects can allow the examination of more complex hypotheses about physiological processes, and has potential to increase power to detect group differences. Such combination of modalities can refer strictly to data acquired from different instruments (e.g., MRI, PET, EEG), or more broadly, to data acquired from the same instrument using different acquisition parameters (e.g., different MRI sequences, different PET ligands); for an overview, see Uludağ and Roebroeck (2014); Zhu et al. (2014); Calhoun and Sui (in press), and for example applications, see Hayasaka et al. (2006); Thomas et al. (in press). Irrespective of which the modalities are, the options in the context of the GLM rest in testing for a single multivariate hypothesis, or in testing for a combination of multiple univariate hypotheses. Single multivariate tests encompass various classical tests, known in particular cases as multivariate analysis of variance (MANOVA), multivariate analysis of covariance (MANCOVA), or canonical correlation/variates analysis (CCA/CVA); these tests will be referred here as *classical multivariate tests*, or CMV.

The combination of multiple univariate hypotheses requires that each is analysed separately, and that these results are grouped together to test, at each voxel (or vertex, or face) a *joint null hypothesis* (JNH); in this context, the separate tests are termed *partial tests*. Different criteria to decide upon rejection of the JNH give rise to three broad categories of combined tests: (I) reject if *any* partial test is significant; (II) reject if *all* partial tests are significant; and (III) reject if some *aggregate measure* from the partial tests is significant. The first of these can be traced back to Tippett (1931), and in current terminology, could be defined as rejecting the joint null hypothesis if any partial test is rejected at the FWER level using the Šidák correction (Šidák, 1967); it also corresponds to a *union–intersection test* (UIT, Roy, 1953). The second is the *intersection–union test* (IUT, Berger, 1982), that in neuroimaging came to be known as *conjunction test* (Nichols et al., 2005). The third offers a trade-off between the two other approaches, and gives rise to a large number of possible tests, each with a different rejection region, and therefore

with different sensitivity and specificity profiles; some of these tests are popular in meta-analyses, with the method of Fisher (Fisher, 1932) being one of the most popular, and new approaches are continually being developed. A summary is shown in Table 4.1, and a brief overview of these and yet other tests, along with bibliographic information, is in Section 4.2.6.

Both cases — a single multivariate test or the combination of multiple univariate tests — can be assessed parametrically when the asymptotic distribution of the test statistic is known, which may sometimes be the case if various assumptions about the data are met. These generally refer to the the independence between observations and between tests, to the distribution of the error terms, and for brain imaging, to yet other assumptions regarding the relationship, across space, between the tests. However, if the observations are *exchangeable*, that is, if their joint distribution remains unchanged after shuffling, then all such assumptions can be eschewed at once, and instead, permutation tests can be performed. The p-values can then be computed for either the classical multivariate tests, or for the combination of univariate tests; when used in the last case, the strategy corresponds to Pesarin's method of *non-parametric combination* (NPC, Pesarin, 1990, 2001), discussed below. Exchangeability is assumed only for the observations within each partial test (or for the errors terms of the respective models,

Table 4.1: *(page 120)* Various functions are available for joint inference on multiple tests. For each method, both its statistic (T) and associated p-value, P are shown. These p-values are only valid if, for each method, certain assumptions are met, particularly with respect to the independence between tests, but sometimes also with respect to underlying distributions. Under exchangeability, the p-values can be computed using permutation tests, and the formulæ in the last column are no longer necessary. The tests are shown in chronological order; see Section 4.2.6 for details and bibliographic information.

T is the statistic for each method and P its asymptotic p-value. All methods are shown as function of the p-values for the partial tests. For certain methods, however, the test statistic for the partial tests, if available, can be used directly. K is the number of tests being combined, p_k, $k = \{1, 2, \ldots, K\}$ are the partial p-values, w_k are positive weights assigned to the respective p_k, $p_{(r)}$ are the p_k with rank r in ascending order (most significant first), α is the significance level for the partial tests, $I(\cdot)$ is an indicator function that evaluates as 1 if the condition is satisfied, 0 otherwise, $\lfloor \cdot \rfloor$ represents the floor function, χ^2_ν is the cumulative distribution function (cdf) for a χ^2 distribution, with the ν degrees of freedom, t_{cdf} is the cdf of the Student's t distribution with degrees of freedom ν, and t^{-1}_{cdf} its inverse, Φ is the cdf of the normal distribution with mean μ and variance σ^2, and Φ^{-1} its inverse, and F and G are the cdf of arbitrary, yet well chosen distributions. For the two Dudbridge–Koeleman methods, $A(T, a, b) = I\left(T > a^b\right) a^b + I\left(T \leqslant a^b\right) T \sum_{j=0}^{b-1} \left(b \ln a - \ln T\right)^j / j!$.

Method	Test statistic (T)	p-value (P)
Tippett	$\min\left(p_k\right)$	$1-\left(1-T\right)^K$
Fisher	$-2\sum_{k=1}^K \ln\left(p_k\right)$	$1-\chi^2\left(T;\ \nu=2K\right)$
Stouffer	$\frac{1}{\sqrt{K}}\sum_{k=1}^K \Phi^{-1}\left(1-p_k\right)$	$1-\Phi\left(T;\ \mu=0,\ \sigma^2=1\right)$
Wilkinson	$\sum_{k=1}^K I\left(p_k \leqslant \alpha\right)$	$\sum_{k=T}^K \binom{K}{k}\alpha^k(1-\alpha)^{K-k}$
Good	$\prod_{k=1}^K p_k^{w_k}$	$\sum_{k=1}^K w_k^{K-1}T^{1/w_k}\left(\prod_{i=1}^{k-1}\left(w_k-w_i\right)^{-1}\right)\left(\prod_{i=k+1}^{K}\left(w_k-w_i\right)^{-1}\right)$
Lancaster	$\sum_{k=1}^K w_k F_k^{-1}\left(1-p_k\right)$	$1-G\left(T\right)$
Winer	$\sum_{k=1}^K t_{\text{cdf}}^{-1}\left(1-p_k;\ \nu_k\right) \Big/ \sqrt{\sum_{k=1}^K \frac{\nu_k}{\nu_k-2}}$	$1-\Phi\left(T;\ \mu=0,\ \sigma^2=1\right)$
Edgington	$\sum_{k=1}^K p_k$	$\sum_{j=0}^{\lfloor T\rfloor}(-1)^j\binom{K}{j}\frac{(T-j)^K}{K!}$
Mudholkar–George	$\frac{1}{\pi}\sqrt{\frac{3(5K+4)}{K(5K+2)}}\sum_{k=1}^K \ln\left(\frac{1-p_k}{p_k}\right)$	$1-t_{\text{cdf}}(T;\ \nu=5K+4)$
Darlington–Hayes	$\frac{1}{r}\sum_{k=1}^r \Phi^{-1}\left(1-p_{(k)}\right)$	Computed through Monte Carlo methods. Tables are available.
Zaykin et al. (TPM)	$\prod_{k=1}^K p_k^{I(p_k\leqslant\alpha)}$	$\sum_{k=1}^K \binom{K}{k}\left(1-\alpha\right)^{K-k}\left(I\left(T>\alpha^k\right)\alpha^k + I\left(T\leqslant\alpha^k\right)T\sum_{j=0}^{k-1}\frac{(k\ln\alpha-\ln T)^j}{j!}\right)$
Dudbridge–Koeleman (RTP)	$\prod_{k=1}^r p_{(k)}$	$\binom{K}{r+1}\left(r+1\right)\int_0^1\left(1-t\right)^{K-r-1}A\left(T,t,K\right)\mathrm{d}t$
Dudbridge–Koeleman (DTP)	$\max\left(\prod_{k=1}^r p_{(k)}, \prod_{k=1}^K p_k^{I(p_k\leqslant\alpha)}\right)$	$\sum_{k=1}^r \binom{K}{k}\left(1-\alpha\right)^{K-k}A\left(T,\alpha,k\right) +$ $I\left(r<K\right)\binom{K}{r+1}\left(r+1\right)\int_0^\alpha\left(1-t\right)^{K-r-1}A\left(T,t,K\right)\mathrm{d}t$
Taylor–Tibshirani (TS)	$\frac{1}{K}\sum_{k=1}^K\left(1-p_{(k)}\frac{K+1}{k}\right)$	$1-\Phi\left(T;\ \mu=0,\ \sigma^2\approx\frac{1}{K}\right)$
Jiang et al. (TTS)	$\frac{1}{K}\sum_{k=1}^K I\left(p_{(k)}\leqslant\alpha\right)\left(1-p_{(k)}\frac{K+1}{k}\right)$	Computed through Monte Carlo methods.

See caption in page 119.

see below); exchangeability is not assumed between the partial tests for either CMV or NPC. Moreover, non-independence does not need to be explicitly modelled, either between observations, between partial tests, or across space for imaging data, thus making such tests applicable to a wide variety of situations.

4.1.3 Overview of the chapter

We show that a single, elegant permutation solution is available for all the situations described above, addressing the *comparisons* of response variables when these can be put in comparable scale, the *correction* of p-values, via adjustment to allow exact control over FWER in the various multiple testing scenarios described above, and the *combination* of multiple imaging modalities to allow for joint inference. The *conjunction* of multiple tests is a special case in which the null hypothesis differs from that of a combination, even though it can be approached in a similar fashion; because the distinction is quite an important one, it is also discussed.

In the next section we outline the notation used throughout the chapter. We then use the definition of union-intersection tests, closed testing procedures, and synchronised permutations to correct for multiple hypotheses, allowing flexibility to mix in the same framework imaging data with different spatial resolutions, surface and/or volume-based representations of the brain, and even non-imaging data. For the problem of joint inference, we propose and evaluate a modification of the NPC, such that instead of two phases and large data storage requirements, the permutation inference can be performed in a single phase, without prohibitive memory needs. We also evaluate, in the context of permutation tests, various combining methods that have been proposed in the past decades, and identify those that provide the best control over error rate and power across a range of situations. We also exemplify the potential gains in power with the reanalysis of the data from a pain study. In the Appendix, we provide a brief historical review of various combining functions, discuss criteria of consistency and admissibility, and provide an algorithm that allows combination and correction in a unified framework.

4.2 Theory

4.2.1 History

Historically, as in the case of the permutation tests discussed in Chapter 3, Ronald A. Fisher was among the first to propose such joint analysis of various tests. In the fourth edition of his now classical book *Statistical Methods for Research Workers* (Fisher, 1932), his approach was described rather succinctly:

> *When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance. It is sometimes desired, taking account only of these probabilities, and not of the detailed composition of the data from which they are derived, which may be of very different kinds, to obtain a single test of the significance of the aggregate, based on the product of the probabilities individually observed.*
>
> *The circumstance that the sum of a number of values of χ^2 is itself distributed in the χ^2 distribution with the appropriate number of degrees of freedom, may be made the basis of such a test. For in the particular case when $n = 2$, the natural logarithm of the probability is equal to $\frac{1}{2}\chi^2$. If therefore we take the natural logarithm of a probability, change its sign and double it, we have the equivalent value of χ^2 for 2 degrees of freedom. Any number of such values may be added together, to give a composite test, using the Table of χ^2 to examine the significance of the result.* — *Fisher (1932)*

The logic of this test is based on the fact that the probability of rejecting the global null hypothesis is related to intersection of the probabilities of each individual test, i.e., $\prod_i P_i$. However, $\prod_i P_i$ is not uniformly distributed, even if the null is true for all partial tests, and cannot be used itself as the joint significance level for the global test. To remediate this fact, some interesting properties and relationships among distributions of random variables were exploited by Fisher and embodied in the succinct excerpt above:

The logarithm of uniform is exponential The cumulative distribution function (cdf) of an exponential distribution is $F(x) = 1 - e^{-\lambda x}$ where λ is the rate parameter, the only parameter of this distribution. The inverse cdf is, therefore, given by $x = -\frac{1}{\lambda}\ln(1 - F(x))$. $F(x) = P$ is a random variable uniformly distributed in the interval $[0, 1]$, and so is $1 - P$, and it is immaterial to differ between them. As a consequence, the same can be written as $x = -\frac{1}{\lambda}\ln(P)$, where $P \sim \mathcal{U}(0, 1)$, which highlights the fact that the negative of the natural logarithm of a random variable distributed uniformly between 0 and 1 follows an exponential distribution with rate parameter $\lambda = 1$.

An exponential with rate 1/2 is χ^2 distributed The cdf of a chi-squared distribution with ν degrees of freedom, i.e. χ^2_ν, is given by $F(x; \nu) = \frac{\int_0^{x/2} t^{\frac{\nu}{2}-1} e^{-t} \mathrm{d}t}{\left(\frac{\nu}{2}-1\right)!}$. If $\nu = 2$, the expression simplifies to $F(x; \nu = 2) = 1 - e^{-x/2}$. In other words, a χ^2 distribution with $\nu = 2$ is equivalent to an exponential distribution with rate parameter $\lambda = 1/2$.

The sum of chi-squared is also chi-squared The moment-generating function (mgf) of a sum of independent variables is the product of the mgfs of the respective variables. The mgf of a χ^2_ν is $M(t) = (1 - 2t)^{-\nu/2}$. The mgf of the sum of K independent variables that follow each a χ^2_2 distribution is then given by $M_{\text{sum}}(t) = \prod_{i=1}^{K}(1-2t)^{-2/2} = (1-2t)^{-K}$, which also defines a χ^2 distribution, however with degrees of freedom $\nu = 2K$.

With these facts in mind, the product $\prod_i P_i$ can be transformed into a p-value that is uniformly distributed when the global null is true. The product can be converted into a sum by taking the logarithm. And as shown above, the logarithm of uniformly distributed variables follows an exponential distribution with rate parameter $\lambda = 1$. Multiplication of each $\ln(P_i)$ by 2 changes the rate parameter to $\lambda = 1/2$ and makes this distribution equivalent to a χ^2 distribution with degrees of freedom $\nu = 2$. The sum of k of these logarithms also follow a χ^2 distribution, now with $\nu = 2K$ degrees of freedom, i.e., χ^2_{2K}. Thus, the statistic for the Fisher method is given by $T_{\text{Fisher}} = -2\sum_{i=1}^{K}\ln(P_i)$, with T_{Fisher} following a χ^2_{2K} distribution, from which a p-value for the global hypothesis can be easily obtained.

The elegance of the combination strategy devised by Fisher resides in that it depends solely on the uniformity of the distribution of the p-values for each of the separate K tests under the null hypothesis, something that, by definition, is always attained whenever a statistical test is exact. This also renders the test, in a certain sense, non-parametric, although arguably, the number of tests can be considered a parameter upon which the resulting combined statistic depends.

4.2.2 Notation and general aspects

For a given voxel (or vertex, or face), consider a multivariate GLM:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.1}$$

where $\mathbf{Y}$ is the $N \times K$ matrix of observed data, with N observations of K distinct (possibly non-independent) variables, $\mathbf{X}$ is the full-rank $N \times R$ design matrix that includes explanatory variables (i.e., effects of interest and possibly nuisance effects), $\boldsymbol{\beta}$ is the $R \times K$ matrix of R regression coefficients for each of the K variables, and $\boldsymbol{\epsilon}$ is the $N \times K$ array of random errors. Estimates for $\boldsymbol{\beta}$ can be computed by ordinary least squares, i.e., $\hat{\boldsymbol{\beta}} = \mathbf{X}^{+}\mathbf{Y}$, where the superscript $(^{+})$ denotes a pseudo-inverse. One generally wants to test the null hypothesis that a given combination (contrast) of the elements in $\boldsymbol{\beta}$ equals to zero, that is, $\mathcal{H}^0 : \mathbf{C}'\boldsymbol{\beta}\mathbf{D} = \mathbf{0}$, where $\mathbf{C}$ is a $R \times S$ full-rank matrix of S contrasts of coefficients on the regressors encoded in $\mathbf{X}$, $1 \leqslant S \leqslant R$ and $\mathbf{D}$ is a $K \times Q$ full-rank matrix of Q contrasts of coefficients on the dependent, response variables in $\mathbf{Y}$, $1 \leqslant Q \leqslant K$. Often more than one such standard multivariate hypothesis is tested, each regarding different aspects of the same data, and each using a different pair of contrasts $\mathbf{C}$ and $\mathbf{D}$. Not uncommonly, even different sets of explanatory variables are considered, sometimes arranged in entirely different designs. We denote the set of such design matrices as $\mathcal{X} = \{\mathbf{X}\}$, the set of pairs of contrasts for each hypothesis related to that design as $\mathcal{C}_{\mathbf{X}} = \{(\mathbf{C}, \mathbf{D})\}$, and the set of sets of such contrasts as $\{\mathcal{C}_{\mathbf{X}}\}$.

Depending on the values of K, Q, and S, $\mathcal{H}^0$ can be tested using various common statistics. If $K = 1$, or if $K > 1$ and $Q = 1$, the problem reduces to the univariate case, in which a t statistic can be used if $S = 1$, or an F-statistic if $S \geqslant 1$. If $K > 1$ and $Q > 1$, the problem is a multivariate proper and can be

approached via CMV when respective multivariate Gaussian assumptions are satisfied; in these cases, if $S = 1$, the Hotelling's T^2 statistic can be used (Hotelling, 1931), whereas if $S > 1$, various other statistics are available, such as the Wilks' λ (Wilks, 1932), the Lawley–Hotelling's trace (Lawley, 1938; Hotelling, 1951), the Roy's largest root(s) (Roy, 1953; Kuhfeld, 1986), and the Pillai's trace (Pillai, 1955); the merits of each in the parametric case are discussed in various textbooks (e.g., Christensen, 2001; Timm, 2002; Anderson and ter Braak, 2003; Johnson and Wichern, 2007), and such tests have been applied to neuroimaging applications (Chen et al., 2014).

The model in Equation 4.1 can be rewritten as $\tilde{\mathbf{Y}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\epsilon}}$, where $\tilde{\mathbf{Y}} = \mathbf{YD}$, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\mathbf{D}$ and $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}\mathbf{D}$. If $Q = 1$, this is a univariate model, otherwise it remains multivariate, with $\tilde{\mathbf{Y}}$ having $\tilde{K} = Q$ columns, and the null hypothesis simplified as $\mathcal{H}^0 : \mathbf{C}'\tilde{\boldsymbol{\beta}} = \mathbf{0}$. This null is equivalent to the original, and can be split into multiple partial hypotheses $\mathcal{H}^0_{\tilde{k}} : \mathbf{C}'\tilde{\boldsymbol{\beta}}_{\tilde{k}} = \mathbf{0}$, where $\tilde{\boldsymbol{\beta}}_{\tilde{k}}$ is the $\tilde{k}$-th column of $\tilde{\boldsymbol{\beta}}$, $\tilde{k} = 1, \ldots, \tilde{K}$. This transformation is useful as it defines a set of separate, even if not independent, partial hypotheses, that can be tested and interpreted separately. We drop heretofore the "$\sim$" symbol, with the modified model always implied.

Non-parametric inference for these tests can be obtained via permutations, by means of shuffling the data, the model, the residuals, or variants of these, in a direct extension from the univariate case (Winkler et al., 2014, Table 3.2, also published in). To allow such rearrangements, some assumptions need to be made: either of *exchangeable errors* (EE) or of *independent and symmetric errors* (ISE). The first allows permutations, the second sign flippings; if both are available for a given model, permutations and sign flippings can be performed together. We use generically the terms *rearrangement* or *shuffling* when the distinction between permutations or sign flippings is not pertinent. These are represented by permutation and/or sign flipping matrices $\mathbf{P}_j$, $j = 1, \ldots, J$, where J is the number of such rearrangements.

Another aspect that concerns permutation tests refers to the use of statistics that are *pivotal*, i.e., that have sampling distributions that do not depend on unknown parameters. Most statistics used with parametric tests (and all the uni- and multivariate examples from the previous paragraph) are pivotal if certain assumptions are met, especially homoscedasticity. Their benefits in non-parametric tests

Table 4.2: Joint hypotheses tested with union–intersection and intersection–union of K partial tests. In the UIT, the null is also called *global null hypothesis*, whereas in the IUT, the null is also called *conjunction null hypothesis*.

	UIT	IUT
Null hypothesis ($\mathcal{H}^0$)	$\bigcap_{k=1}^{K} \mathcal{H}_k^0$	$\bigcup_{k=1}^{K} \mathcal{H}_k^0$
Alternative hypothesis ($\mathcal{H}^1$)	$\bigcup_{k=1}^{K} \mathcal{H}_k^1$	$\bigcap_{k=1}^{K} \mathcal{H}_k^1$

are well known (Hall and Wilson, 1991), and for neuroimaging, pivotal statistics are useful to allow exact correction for the MTP-I.

4.2.3 Union–intersection and intersection–union tests

Consider the set of p-values $\{p_k\}$ for testing the respective set of partial null hypotheses $\{\mathcal{H}_k^0\}$. A union–intersection test (UIT, Roy, 1953) considers the JNH corresponding to a *global null hypothesis* that *all* $\mathcal{H}_k^0$ are true; if any such partial null is rejected, the global null hypothesis is also rejected. An intersection–union test (IUT, Berger, 1982) considers the JNH corresponding to a *conjunction null hypothesis* (also termed *disjunction of null hypotheses*) that *any* $\mathcal{H}_k^0$ is true; if all partial nulls are rejected, the conjunction null hypothesis is also rejected. In the UIT, the null is the intersection of the null hypotheses for all partial tests; the alternative is the union of the alternatives. In the IUT, the null is the union of the null hypotheses for all partial tests; the alternative is the intersection of the alternatives. A UIT is significant if the smallest p_k is significant, whereas an IUT is significant if the largest p_k is significant. Figure 4.1 illustrates the rejection regions for UIT and IUT cases based on two independent t-tests, in which the statistic larger than a certain critical level is considered significant. Table 4.2 shows the null and alternative hypotheses for each case.

Enlarging the number of tests affects UITs and IUTs differently. For the UIT with

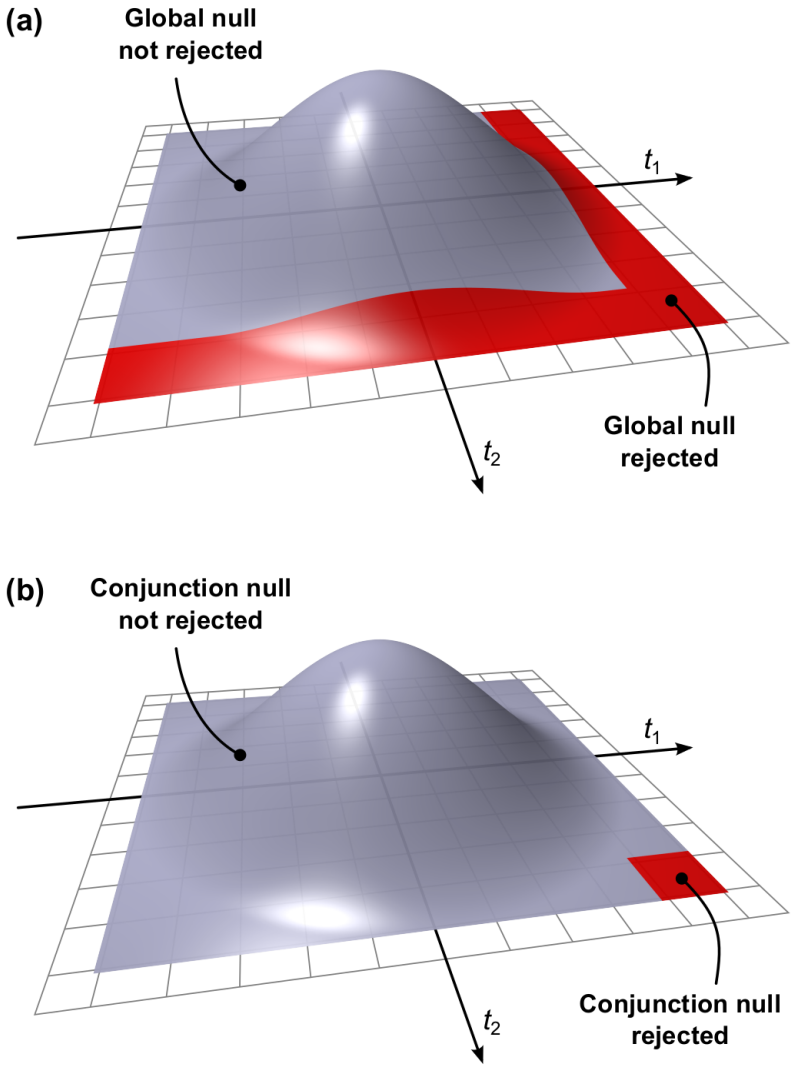


Figure 4.1: *(a)* Rejection region of a union–intersection test (UIT) based on two independent t-tests. The null is rejected if either of the partial tests has a statistic that is large enough to be qualified as significant. *(b)* Rejection region of an intersection–union test (IUT) based the same tests. The null is rejected if both the partial tests have a statistic is large enough to be qualified as significant.

a given statistic threshold, more tests increase the chances of false positives, and correction for this multiplicity needs to be applied. In fact, it can be shown that a UIT at a significance level α is equivalent to controlling the FWER at α for the same tests. In other words, a union-intersection procedure is an FWER procedure. For an IUT, in contrast, the procedure does not change with more tests. The conjunction null hypothesis is *composite*, consisting of different parameter settings. For the extreme case that exactly one partial null is true and $K - 1$ effects are real, an IUT is exact for any K; if two or more more partial nulls are true, an IUT becomes increasingly conservative with larger K.

The null hypothesis of the UIT can be rejected if the smallest p_k is significant or, equivalently, its corresponding statistic, that is, the extremum statistic. For tests in which larger statistics provide evidence against the null hypothesis, the relevant extremum is the maximum. Conversely, for tests in which smaller statistics provide evidence against the null, the extremum is the minimum. Clearly, if the most extreme statistic is significant, at least one partial hypothesis is rejected, therefore the global null hypothesis can be rejected without the need to continue testing the other $K - 1$ partial hypotheses. The null hypothesis of the IUT can be rejected if the largest p_k is significant or, equivalently, its corresponding least extreme statistic. Clearly, if the least extreme statistic is significant, all partial hypotheses can be rejected, therefore the conjunction hypothesis can be rejected without the need to continue testing all other $K - 1$ partial hypotheses.

In brain imaging, the term *conjunction* refers to a test performed when one wants to localise regions where there is signal in all partial tests, that is, a logical AND of all alternative hypotheses (Nichols et al., 2005), and is synonymous with the IUT. In noting the lack of power of such a proper conjunction test, Friston et al. (2005) suggested a *partial conjunction*, in which fewer than all alternatives need to intersect. Using the same notation of Table 4.1, both approaches have the same statistic, $T = \max(p_k)$, but the p-value of the latter can be computed as T^{K-v+1}, so that the test is a conjunction of at least v alternative hypotheses; if $v = K$, it is an IUT, and if $v = 1$ the null is equivalent to that of a UIT (such a test, however, is inconsistent for a UIT; see Section 4.2.9). Benjamini and Heller (2008) further generalised the procedure by allowing the combination of the largest p-values using any of various possible combining functions, such as those we present

in Table 4.1 and in Section 4.2.6.

4.2.4 Closed testing

In a *closed testing procedure* (CTP), each $\mathcal{H}_k^0$ is rejected if, and only if, it is significant in its own right at a certain level α, and if all possible sub-JNHs that include the same $\mathcal{H}_k^0$ and comprise some or all of the partial hypotheses (that is, subsets of the global JNH formed by some of the partial tests) are also rejected at α using a suitable test. Various such tests can be considered, including CMVs and NPC (next section).

A CTP guarantees strong control over FWER (Marcus et al., 1976). To produce adjusted p-values, the original method requires that all $2^K - 1$ sub-JNHs are tested[1], a requirement that is computationally onerous, even for a moderate number of tests, a problem aggravated by the large number of tests that are considered in an imaging experiment. There exists, however, a particular test for the sub-JNHs that obviates the need for such a gargantuan computational venture: the union–intersection test. In a UIT using the extremum statistic, the most extreme of the global JNH that comprises all the K partial tests is also the most extreme of any other sub-JNH that includes that particular partial hypothesis, such that the other joint subtests can be bypassed altogether. As a UIT is also an FWER-controlling procedure, this raises various possibilities for correction of both MTP-I and MTP-II. While such a shortcut can be considered for both parametric (Holm, 1979) and non-parametric cases (Westfall and Young, 1993), for the non-parametric methods using permutation, one additional feature is needed: that the joint sampling distribution of the statistic used to test each of the sub-JNH is the same regardless whether the null is true for all the K partial tests, or just some of them. This property is called *subset pivotality* (Westfall and Young, 1993; Westfall and Troendle, 2008), and it constitutes the multivariate counterpart to the univariate pivotality.

[1] From the Pascal triangle: $\sum_{i=1}^{K} \binom{K}{i} = 2^K - 1$.

4.2.5 Non-parametric combination

The NPC consists of testing each of the $\mathcal{H}_k^0$ using shufflings that are performed synchronously for all K partial tests. The resulting statistics for each permutation are recorded, allowing an estimate of the complete empirical null distribution to be constructed for each partial test. In a second stage, the empirical p-values for each statistic are combined, for each permutation, into a *joint statistic*. As such a combined joint statistic is produced from the previous permutations, an estimate of its empirical distribution function is immediately known, and so the p-value of the unpermuted statistic, hence of the joint test, can be assessed. The method was proposed by Pesarin (1990, 1992), and independently, though less generically, by Blair and Karniski (1993); Blair et al. (1994); a thorough description is available in Pesarin (2001) and Pesarin and Salmaso (2010a). An early application to brain imaging can be found in Hayasaka et al. (2006), its use to combine different statistics within the same modality in Hayasaka and Nichols (2004), and a summary description and practical examples are presented in Brombin et al. (2013). The JNH of the combined test is that all partial null hypotheses are true, and the alternative that any is false, which is the same null of a UIT, although the rejection region may differ widely from the example in Figure 4.1*a*, depending on the combining function.

The only two requirements for the validity of the NPC are that the partial test statistics have the same direction suggesting the rejection of the null hypothesis, and that they are consistent (see Section 4.2.9). For the combining function, it is desirable that (I) it is non-decreasing with respect to all its arguments (which are the p-values p_k, or $1 - p_k$, depending on the combining function), (II) that it approaches its maximum (or minimum, depending on the function) when at least one of the partial tests approaches maximum significance (that is, when at least one p-value approaches zero), and (III) that for a test level $\alpha > 0$, the critical significance threshold is smaller than the function maximum value. These requirements are easily satisfied by almost all functions shown in Table 4.1, which therefore can be used as combining functions in the framework of NPC (see Section 4.2.9 for a discussion on the few exceptions).

One of the most remarkable features of NPC is that the synchronised permuta-

tions implicitly account for the dependence structure among the partial tests. This means that even combining methods originally derived under an assumption of independence, such as Tippett or Fisher, can be used even when independence is untenable. In fact, modifications to these procedures to account for non-independence (e.g., Brown, 1975; Kost and McDermott, 2002, for the Fisher method) are made redundant. As the p-values are assessed via permutations, distributional restrictions are likewise not necessary, rendering the NPC free of most assumptions that thwart parametric methods in general. This is why NPC methods are an alternative to CMV tests, as each of the response variables in a MANOVA or MANCOVA analysis can be seen as an univariate partial test in the context of the combination.

4.2.6 Overview of combining functions

Below are a few details and references for the methods shown in Table 4.1, plus a few others, presented in chronological order. A number of studies comparing some of these functions in various scenarios have been published (Birnbaum, 1954; van Zwet and Oosterhoff, 1967; Oosterhoff, 1969; Rosenthal, 1978; Berk and Cohen, 1979; Westberg, 1985; Lazar et al., 2002; Loughin, 2004; Whitlock, 2005; Wu, 2006; Won et al., 2009; Bhandary and Zhang, 2011; Chen, 2011; Zaykin, 2011; Chang et al., 2013). Some of these are permutationally equivalent to each other, that is, their rejection region under permutation is the same, and it becomes immaterial which is chosen.

Tippett This is the oldest and probably the simplest of the combination methods, having appeared in Tippett (1931). The combined test statistic is simply the minimum p-value across all partial tests, i.e. $T_{\text{Tippett}} = \min_k (p_k)$. The probability is computed as $P_{\text{Tippett}} = 1 - \left(1 - T_{\text{Tippett}}\right)^K$.

Fisher This is certainly the most well known of the combination strategies. It appeared in Fisher (1932) and follows from the idea of treating the joint probability as the intersection of all partial tests, which is given by their product $\prod_k p_k$. A statistic for the global hypothesis can be constructed as $T_{\text{Fisher}} = -2 \sum_k \ln (p_k)$, as shown earlier in this chapter, which follows a χ^2 distribution with $2k$ degrees of

freedom, and from which an uniformly distributed significance level, P_{Fisher}, can be obtained.

Pearson–David The same product suggested by Fisher, $\prod_k p_k$, was used by Pearson (1933) to test equality of distributions. David (1934) discussed that a similar test could be used with $\prod_k (1 - p_k)$ and suggested using the most extreme of these two products as the statistic, a view later shared by Pearson himself (Pearson, 1934). The test statistic is, therefore, given by $T_{\text{Pearson–David}} = -2 \min \left(\sum_k \ln (p_k) , \sum_k \ln (1 - p_k) \right)$, which, as in the Fisher method, follows a χ^2 distribution with $2k$ degrees of freedom, and from which the significance $P_{\text{Pearson–David}}$ can be computed.[2]

Stouffer This method appeared as a footnote in the report of the sociological study conducted among veterans of the World War II by Stouffer et al. (1949). The idea is to convert the p-values to normally-distributed z-scores, sum these scores, and compute a new p-value. The conversion to a normal distribution is irrespective to the distributions from which the partial p-values, p_k, may have arisen. The test statistic is given by $T_{\text{Stouffer}} = \frac{1}{\sqrt{K}} \sum_k \Phi^{-1} (1 - p_k)$, where Φ^{-1} is the inverse cumulative distribution function (cdf) of the normal distribution (i.e. the probit function). The statistic T_{Stouffer} follows a normal distribution with zero mean and unit variance, from which a probability P_{Stouffer} can be obtained.

Wilkinson The probability of observing r significant p-values at level α out of the K tests performed can be computed using a binomial expansion as proposed by Wilkinson (1951). The statistic $T_{\text{Wilkinson}}$ is simply r, and the probabilty of finding no more or less than r by chance is given by $P_{\text{Wilkinson}} = \sum_{k=r}^{K} \binom{K}{k} \alpha^k (1 - \alpha)^{K-k}$. If the partial p-values are sorted in ascending order, $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant, p_{(K)}$, and if the significance level is defined as $\alpha = p_{(1)}$, the approach is equivalent to the Tippett method. Note that the probability does not depend on the actual probabilities for the partial tests, but only on r and α.

[2] Historical details regarding this method are recounted in Owen (2009). The authors also comment that the significance level could be doubled to account for the fact that two tests are being performed, although this is not in the original publications.

Good A generalisation of the Fisher method, and which assigns arbitrary, unequal positive weights w_k for each of the p-values of the partial tests, was suggested by Good (1955). Each partial test can be weighted according to some criteria, for instance, the sample size for each of the partial test, the number of degrees of freedom, or some other desirable feature, such as ecological or internal validity (Rosenthal, 1978). The statistic is given by $T_{\text{Good}} = \prod_k p_k^{w_k}$, and its significance can be assessed as $P_{\text{Good}} = \sum_k W_k T_{\text{Good}}^{1/w_k}$, where $W_k = w_k^{K-1} \left(\prod_{i=1}^{k-1} (w_k - w_i)^{-1}\right) \left(\prod_{i=k+1}^{K} (w_k - w_i)^{-1}\right)$.

Lipták Another generalised combined statistic can be produced using the inverse cdf, F^{-1}, of the p_k, summing the values of the statistics, and computing a new p-value for the global null using the cdf G of the sum of the statistics, a method proposed by Lipták (1958). Each summand can be arbitrarily weighted, as in the Good method. In principle, any continuously increasing function with support in the interval $[0;\ 1]$ can be used for F, albeit a more obvious choice is the cdf of the normal distribution, which can be used as both F and G, and which makes the approach virtually identical to the Stouffer method if all weights are 1 (van Zwet and Oosterhoff, 1967). In this case, the statistic for the method is given by $T_{\text{Lipták}} = \sum_k w_k \Phi^{-1}(1 - p_k)$, which follows a normal distribution with zero mean and variance K. F can also be a χ^2_ν distribution, in which case, and also when all $w_k = 1$, G is a $\chi^2_{K\nu}$ distribution. If $\nu = 2$, the approach is equivalent to the Fisher method.

Lancaster While Lipták method generalises combining strategies such as Fisher and Stouffer, the Lancaster method (Lancaster, 1961) further generalises the Lipták approach by allowing different F_k^{-1} for each partial test. Choices for F_k^{-1} include, for instance, the cdf of the gamma distribution with scale parameter $\theta = 2$, possibly with different shape parameters taking the place of the weights w_k for each partial test. If the weights are all positive integers, the significances can be assessed from the cdf of a χ^2 distribution, with degrees of freedom $\nu = 2\sum_k w_k$ (Berk and Cohen, 1979).

Winer A combination strategy that resembles the Stouffer method, but uses Student's t statistics, rather than z-scores was proposed by Winer (1962). The idea is to sum the t statistics for all the K partial tests, and normalising the sum so that the resulting statistic follows a standard normal distribution. The normalisation is based on the fact that the variance of the t distribution can be determined from its the degrees of freedom ν as $\nu/(\nu-2)$. The statistic for this method is given by $T_{\text{Winer}} = \sum_k t_k \Big/ \sqrt{\sum_k \frac{\nu_k}{\nu_k - 2}}$. The Winer method cannot be applied if $\nu_k \leqslant 2$ for any of the partial tests. Moreover, ν_k should not be too small for the normal approximation to be reasonably valid (e.g., $\nu_k \geqslant 10$). The Winer method is a particular case of the Lancaster method.

Edgington The probability of observing, due to chance, a value equal or smaller than the sum of the partial p-values, $T_{\text{Edgington}} = \sum_k p_k$, was proposed by Edgington (1972) as a more powerful alternative to the Fisher method. This probability can be calculated as $P_{\text{Edgington}} = \frac{T^K}{K!}$ when $T \leqslant 1$, where T is the $T_{\text{Edgington}}$ statistic. More generally, or if $T > 1$ the probability can be computed as $P_{\text{Edgington}} = \sum_{j=0}^{\lfloor T \rfloor} (-1)^j \binom{K}{j} \frac{(T-j)^K}{K!}$, where $\lfloor \cdot \rfloor$ is the floor function.

Mudholkar–George It is possible to use a simple logit transformation to compute a statistic that approximates a scaled version of the Student's t distribution, as shown by Mudholkar and George (1979). The scaling can be applied to the result of the logit transformation itself, such that the statistic is computed as $T_{\text{Mudholkar–George}} = \frac{1}{\pi}\sqrt{\frac{3(5K+4)}{K(5K+2)}} \sum_k \ln\left(\frac{1-p_k}{p_k}\right)$, which follows a t distribution with $5K + 4$ degrees of freedom.

Friston (global null) Friston et al. (1999) proposed the use of the minimum statistic, or equivalently, the maximum p_k, across the K tests as a way to test the null hypothesis of no effect for all the tests. The fact that it had originally been called a "conjunction" caused some confusion in the literature, because the eventual rejection of the global null cannot be used to infer that the null for each of the partial tests are all rejected, as it would be in a logical conjunction (Nichols et al., 2005). The statistic for this method can be expressed in terms of the p-values for the partial tests as $T_{\text{Friston}} = \max_k (p_k)$, and its significance can be assessed as

$P_{\text{Friston-GN}} = T^K_{\text{Friston}}$. The Friston method is equivalent to the Wilkinson method if $\alpha = p_{(K)}$ and so, $r = K$.

Darlington–Hayes In a discussion about pooling p-values for meta-analysis, Darlington and Hayes (2000) raised a number of limitations of these methods, and proposed a modification over the Stouffer method that would address some of these concerns. The modified method, called *Stouffer-max*, uses as test statistic the mean of the r highest z-scores, i.e. $T_{\text{Darlington–Hayes}} = \frac{1}{r}\sum_{k=1}^{r}\Phi^{-1}\left(1-p_{(k)}\right)$, rather than the normalised sum all the z-scores as in the Stouffer method. When $r = 1$, it is equivalent to the Tippett method, whereas when $r = K$, equivalent to the original Stouffer. Significances can be computed for intermediate values of r through Monte Carlo simulation, and the authors provided tables with critical values.

Zaykin This method, called *truncated product method* (TPM) was proposed by Zaykin et al. (2002) as a way to combine features of the Fisher and Wilkinson methods. The statistic is given by $T_{\text{Zaykin}} = \prod_{k=1}^{K} p_k^{I(p_k \leqslant \alpha)}$, where $I(\cdot)$ is an indicator function that evaluates as 1 if the given condition is satisfied, and 0 otherwise. In other words, the statistic is the product of only the partial p-values that are significant at the level α, whereas in the Fisher method, all p-values are used. The significance for the combination is given by $P_{\text{Zaykin}} = \sum_{k=1}^{K}\binom{K}{k}(1-\alpha)^{K-k}\left(I\left(T > \alpha^k\right)\alpha^k + I\left(T \leqslant \alpha^k\right)T\sum_{j=0}^{k-1}\frac{(k\ln\alpha - \ln T)^j}{j!}\right)$, where T is T_{Zaykin}. If $\alpha = \min_k(p_k)$, then the approach is equivalent to the Tippett method. If $\max_k(p_k) \leqslant \alpha \leqslant 1$, the approach is equivalent to the Fisher method. Although exact, computationally the expression for P_{Zaykin} is prone to over/underflows for certain combinations of large K and α, and because of this, when a global significance cannot be obtained analytically, Monte Carlo methods can be used.

Dudbridge–Koeleman While the Zaykin method combines only the partial tests that are significant at the level α, it is also possible to create a statistic that combines only the most r significant tests, where r is specified in advance. This method was proposed by Dudbridge and Koeleman (2003) and called *rank truncated product* (RTP). The main benefit of this strategy is that it depends only on a predetermined number of partial tests to be rejected, rather than on their significances, which

are random quantities. The statistic is computed as $T_{\text{Dudbridge–Koeleman}} = \prod_{k=1}^{r} p_{(k)}$, where $p_{(k)}$ is the p-value for the k-th most significant partial test. The significance can be assessed as $P_{\text{Dudbridge–Koeleman}} = \binom{K}{r+1} (r+1) \times \int_0^1 (1-t)^{K-r-1} \left(I(T > t^r)\, t^r + I(T \leqslant t^r)\, T \sum_{j=0}^{r-1} \frac{(r \ln t - \ln T)^j}{j!} \right) \mathrm{d}t$, where $T = T_{\text{Dudbridge–Koeleman}}$. As with the Zaykin method, for certain combinations of r and large K, the significances need to be computed through Monte Carlo methods.[3]

Nichols Addressing logical issues regarding the original Friston method[4] when used for conjunctions, Nichols et al. (2005) observed that the same minimum statistic (or, equivalently, the maximum p-value) could still be used for true conjunction inference. The idea is that, if the least significant test, i.e. the largest p_k, is significant at α, then all the partial tests are also significant at that level, and so, the *conjunction null hypothesis*[5], i.e. the hypothesis that there is no effect for all or for some of the tests, can be rejected. This was the first conjunction test proposed in the neuroimaging literature[6] and it does not assume independence between the partial tests.

Friston (conjunction null) To address the issues that emerged about the misuse of the original test to reject the global null as a "conjunction", Friston et al. (2005) suggested another test, which uses the same statistic, but with the significance being computed as $P_{\text{Friston-CN}} = T_{\text{Friston}}^{K-u+1}$, where u is the minimum number of partial tests that need to be rejected so that the test is a true conjunction of at least u tests. When $u = K$, the approach is equivalent to the Nichols method, and when $u = 1$, it is equivalent to the original Friston method. For other values of u, the test can

[3] A combination of the TPM and RTP has been also proposed and named *rank-and-threshold truncated product* or *dual truncated product* (DTP). The statistic is $\max\left(T_{\text{Zaykin}}, T_{\text{Dudbridge–Koeleman}}\right)$ and its significance can be computed analytically or via Monte Carlo methods. See the Appendix of Dudbridge and Koeleman (2003) for details.

[4] By original we mean the method in Friston et al. (1999). Another conjunction method had previously been proposed (Price and Friston, 1997), which suffered from different issues (Caplan and Moo, 2004).

[5] Also called *disjunction of null hypotheses* (Benjamini and Heller, 2008).

[6] The authors had presented a the test in a poster at the X Annual Meeting of the Organization for Human Brain Mapping (OHBM), in 2004 in Budapest, Hungary (Brett et al., 2004). A similar test, with the null and alternative hypotheses reversed, had been proposed by Berger (1982).

be termed a *partial conjunction test.*

Taylor–Tibshirani If the p-values are sorted in ascending order, $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant, p_{(K)}$, these ranked significances can be compared to their expectations under the global null hypothesis. Large deviations from the expected values suggest the presence of the effect among the tests. Taylor and Tibshirani (2006) suggested that a measurement of this deviation could be used to infer the overall significance of the tests. This measurement, termed *tail strength* (TS), is defined as $T_{\text{Taylor–Tibshirani}} = \frac{1}{K} \sum_{k=1}^{K} \left(1 - p_{(k)} \frac{K+1}{k}\right)$. Under the assumptions that the global null is true and the tests are independent, this statistic follows a normal distribution with zero mean and a variance that can be approximated as $\sigma^2 = \frac{1}{K}$ when $K \to \infty$, from which significance can be assessed. When these assumptions are not met, bootstrap inference can be used.

Benjamini–Heller Recognising that sometimes a compromise between the global null and the conjunction null may be necessary, as in the Friston (conjunction null) method, Benjamini and Heller (2008) proposed a generic approach in which a probability for rejecting the conjunction null in at least u out of the K tests is computed. In this method, the p-values are sorted in ascending order, and only those larger than $p_{(u)}$ are combined. The combination can use any of the methods that reject the global null discussed above, or others, including methods that take non-independence into account.

Jiang The statistic of the Taylor–Tibshirani method has a variance that depends asymptotically only on the number of tests K. However, the value of the statistic can be small when effect is truly present in only a few partial tests, therefore reducing the power of the method. In an analogy with the Zaykin method, Jiang et al. (2011) proposed to compute the tail strength using only partial tests with p-values smaller than a certain level α. The method is called *truncated tail strength* (TTS), and the statistic is computed as $T_{\text{Jiang}} = \frac{1}{K} \sum_{k=1}^{K} I\left(p_{(k)} \leqslant \alpha\right) \left(1 - p_{(k)} \frac{K+1}{k}\right)$. This statistic has no known analytical distribution and the authors propose computing their significance using Monte Carlo or permutation methods.

4.2.7 Transformation of the statistics

While NPC offers flexibility in a simple and uncomplicated formulation, its implementation for brain imaging applications poses certain challenges. Because the statistics for all partial tests for all permutations need to be recorded, enormous amounts of data storage space may be necessary, a problem further aggravated when more recent, high resolution imaging methods are considered. Even if storage space were not a problem, however, the discreteness of the p-values for the partial tests becomes problematic when correcting for multiple testing, because with thousands of tests in an image, ties are very likely to occur among the p-values, further causing ties among the combined statistics. If too many tests across an image share the same most extreme statistic, correction for the MTP-I, while still valid, becomes less powerful (Westfall and Young, 1993; Pantazis et al., 2005). The most obvious workaround — run an ever larger number of permutations to break the ties — may not be possible for small sample sizes, or when possible, requires correspondingly larger data storage.

However, another possible approach can be considered after examining the two requirements for the partial tests, and also the desirable properties (I)–(III) of the combining functions, all listed earlier. These requirements and properties are quite mild, and if the sample size is reasonably large and the test statistics homogeneous, i.e., they share the same asymptotic permutation distribution, a *direct combination* based not on the p-values, but on the statistics themselves, such as their sum, can be considered (Pesarin and Salmaso, 2010a, page 131). Sums of statistics are indeed present in combining functions such as of Stouffer, Lancaster, Winer, and Darlington–Hayes, but not others listed in Table 4.1 and Section 4.2.6. In order to use these other combining functions, most of them based on p-values for the partial tests, and under the same premises, the statistics need to be transformed to quantities that behave as p-values. In the parametric case, these would be the parametric p-values, computed from the parametric cumulative distribution function (cdf) of the test statistic. If the parametric assumptions are all met for the partial tests, their respective parametric p-values are all valid and exact; if the assumptions are not met, these values are no longer appropriate for inference on the partial tests, but may still be valid for NPC, for satisfying all requirements and

desirable properties of the combining functions. As they are not guaranteed to be appropriate for inference on the partial tests, to avoid confusion, we call these parametric p-values "u-values".

Another reason for not treating u-values as valid p-values is that they do not necessarily need to be obtained via an assumed, parametric cumulative distribution function for the statistics of the partial tests. If appropriate, other transformations applied to the statistics for the partial tests can be considered; whichever is more accurate to yield values in the interval $[0; 1]$ can be used. The interpretation of a u-value should not be that of a probability, but merely of a monotonic, deterministic transformation of the statistic of a partial test, so that it conforms to the needs of the combining functions.

Transformation of the statistic to produce quantities that can be used in place of the non-parametric p-values effectively simplifies the NPC algorithm, greatly reducing the data storage requirements and computational overhead, and avoiding the losses in power induced by the discreteness of p-values. This simplification is shown in Figure 4.2, alongside the original NPC algorithm.

Regardless of the above transformation, the distribution of the combined statistic, T, may vary greatly depending on the combining function, and it is always assessed non-parametrically, via permutations. Different distributions for different combining functions can, however, pose practical difficulties when computing spatial statistics such as cluster extent, cluster mass, and even threshold-free cluster enhancement (TFCE, Smith and Nichols, 2009). Consider for instance the threshold used to define clusters: prescribed values such as 2.3 or 3.1 (Woo et al., 2014) relate to the normal distribution and are not necessarily sensible choices for combining functions such as Tippett or Fisher. Moreover, for some combining functions, such as Tippett and Edgington, smaller values for the statistic are evidence towards the rejection of the null, as opposed to larger as with most of the others. To address these practical issues, a monotonic transformation can be applied to the combined statistic, so that its behaviour becomes more similar to, for instance, the z-statistic (Efron, 2004). This can be done again by resorting to the asymptotic behaviour of the tests: the combined statistic is converted to a parametric p-value (the formulas are summarised in Table 4.1), which, although not valid for inference unless certain assumptions are met, particularly with respect to the independence among the

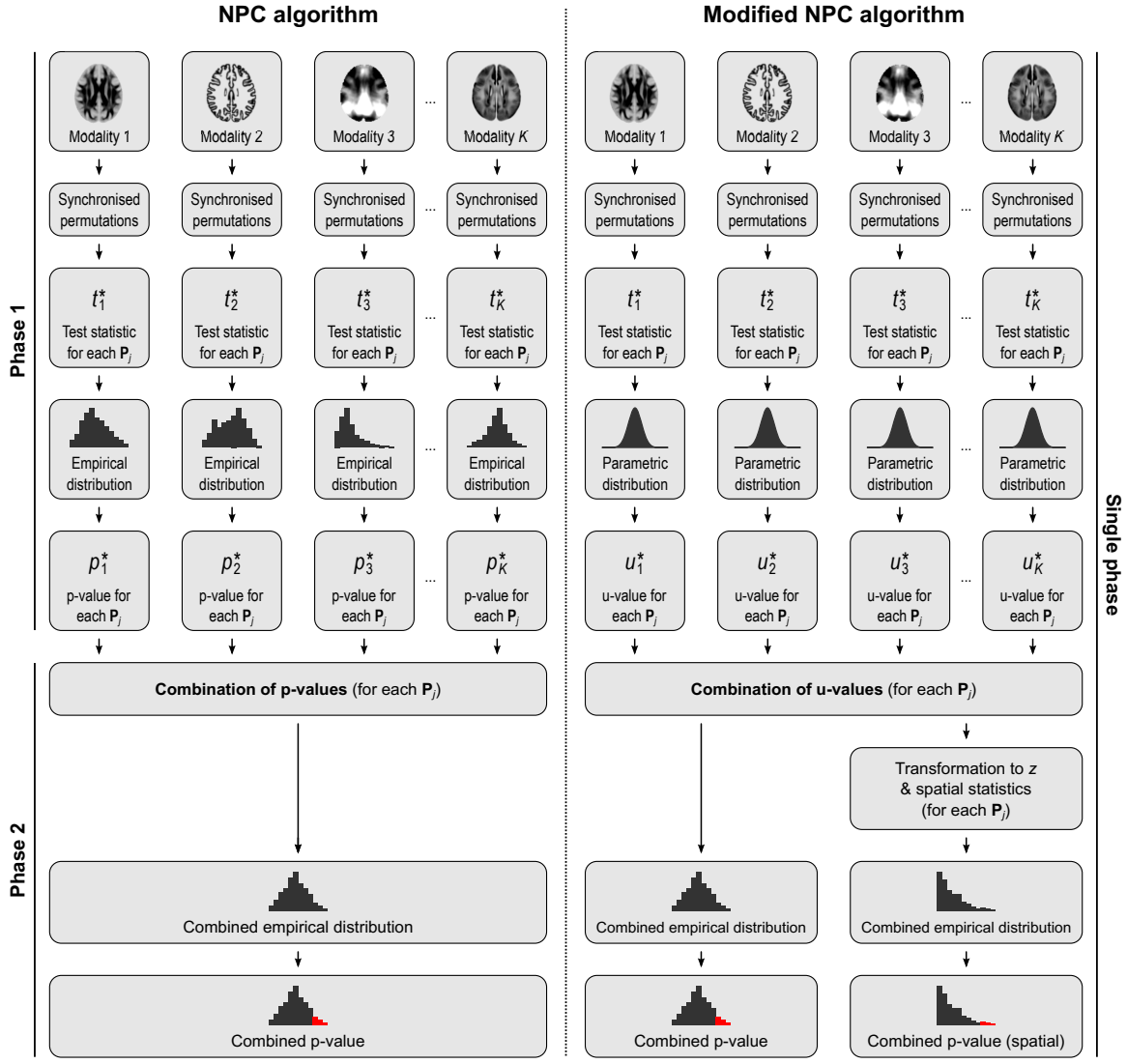


Figure 4.2: The original NPC algorithm combines non-parametric p-values and, for imaging applications, requires substantial amount of data storage space. Two modifications simplify the procedures: (I) the statistic t_k for each partial test k is transformed into a related quantity u_k that has a behaviour similar to the p-values, and (II) the combined statistic is transformed to a variable that follows approximately a normal distribution, so that spatial statistics (such as cluster extent, cluster mass, and TFCE) can be computed as usual. The first simplification allows the procedure to run in a single phase, without the need to retrieve data for the empirical distribution of the partial tests.

partial tests, are useful to transform, at each permutation, the combined statistic to the z-statistic, which can then be used for inference using cluster extent, mass, or TFCE.

4.2.8 Directed, non-directed, and concordant hypotheses

When the partial hypotheses are one-sided, i.e., $\mathcal{H}_k^0 : \mathbf{C}'\boldsymbol{\beta}_k > 0$ or $\mathcal{H}_k^0 : \mathbf{C}'\boldsymbol{\beta}_k < 0$, and all have the same direction (either), the methods presented thus far can be used as described. If not all have the same direction, a subset of the tests can be scaled by -1 to ensure a common direction for all.

If the direction is not relevant, but the concordance of signs towards one of them (either) is, a new combining test can be constructed using one-sided p-values, p_k, and another using $1-p_k$, then taking the best of these two results after correcting for the fact that two tests were performed. For example, for the Fisher method, we would have:

$$T = \max\left(-2\sum_{k=1}^{K}\ln\left(p_k\right), -2\sum_{k=1}^{K}\ln\left(1-p_k\right)\right) \tag{4.2}$$

where T is the combined test statistic, with its p-value, P, assessed via permutations.

If direction or concordance of the signs are not relevant, two-sided (non-directed) tests and p-values can be used before combining, that is, ignoring the sign of the test statistic for the partial tests, or using a statistic that is non-directional (e.g., with F-tests for the partial hypotheses). It worth mentioning, however, that it is not appropriate to simultaneously ignore directions of the partial tests *and* use a combination that favours concordant signs. Such a test would lack meaning and would be inadmissible, with examples shown in Section 4.2.10.

Rejection regions for these three cases, for four different combining functions, are shown in Figure 4.3, as functions of the partial p-values, for $K = 2$ partial tests.

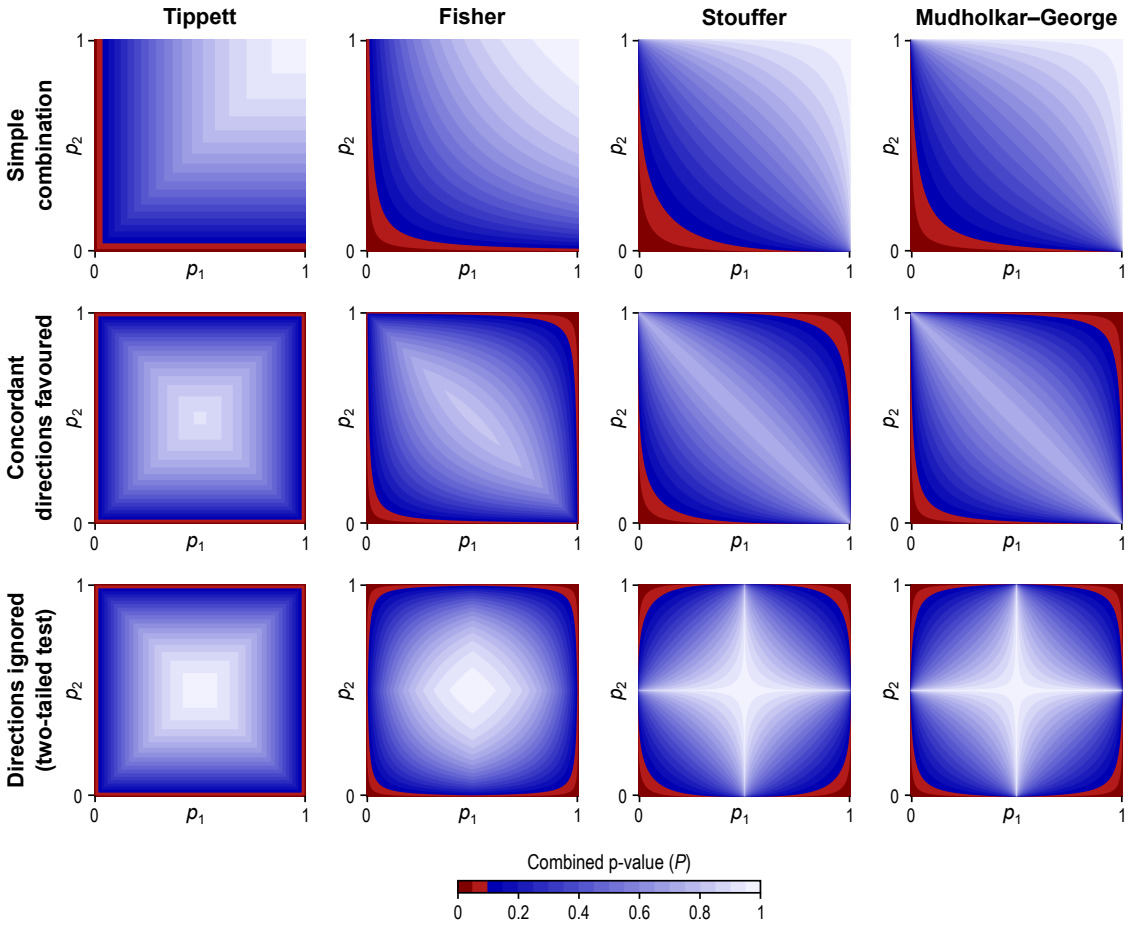


Figure 4.3: *Upper row:* Rejection regions for the combination of two partial tests using four different combining functions, and with the p-values assessed parametrically (Table 4.1). The regions are shown as function of the p-values of the partial tests (p_k). *Middle row:* Rejection regions for the same functions with the modification to favour alternative hypotheses with concordant directions. *Lower row:* Rejection regions for the same functions with the modification to ignore the direction altogether, that is, for two-tailed partial tests.

4.2.9 Consistency of combined tests

A hypothesis test is said to be *consistent* if, for a fixed test level, its power goes to unity as the sample size increases to infinity. The use of a non-consistent combining function to form an NPC test is problematic, as the rejection region may not be reached even if the p-value for one or more of the partial tests approach zero, thus violating the second of the three desirable properties of the combining functions, presented in Section 4.2.5.

Among the functions shown in Table 4.1, the notable non-consistent combining functions are the Edgington and Wilkinson (see Section 4.2.6). Also, it should be noted that functions that define conjunctions (IUT), such as those based on $\max(p_k)$, are likewise not consistent in the context of NPC, as the latter serves to test the global null hypothesis. Figure 4.4 shows rejection regions for some inconsistent combining functions, and variants, similarly as for the (consistent) shown in Figure 4.3.

4.2.10 Admissibility of combined tests

A combined hypothesis test is said to be *admissible* if there exists no other test that, at the same significance level, without being less powerful to all possible alternative hypotheses, is more powerful to at least one alternative (Lehmann and Romano, 2005). This can be stated in terms of either of two sufficient conditions for admissibility: (I) that rejection of the null for a given p-value implies the rejection of the null for all other p-values smaller or equal than that, or (II) that the rejection region is convex in the space of the test statistic.

Combinations that favour tests with concordant directions (Section 4.2.8), if used with of non-directional partial tests, create tests that are inadmissible, that is, tests that are not optimal in the sense that there exist other tests that, without being less powerful to some true alternative hypotheses, are more powerful to at least one true alternative. Inadmissibility implies that the test cannot be used, as certain combinations of partial tests lead to nonsensical results, such as rejecting the JNH for some partial p-values, and failing to reject for some p-values that are even smaller. Figure 4.5 shows rejection regions of inadmissible versions of the combining functions considered in Figures 4.3 and 4.4; clearly none of the two

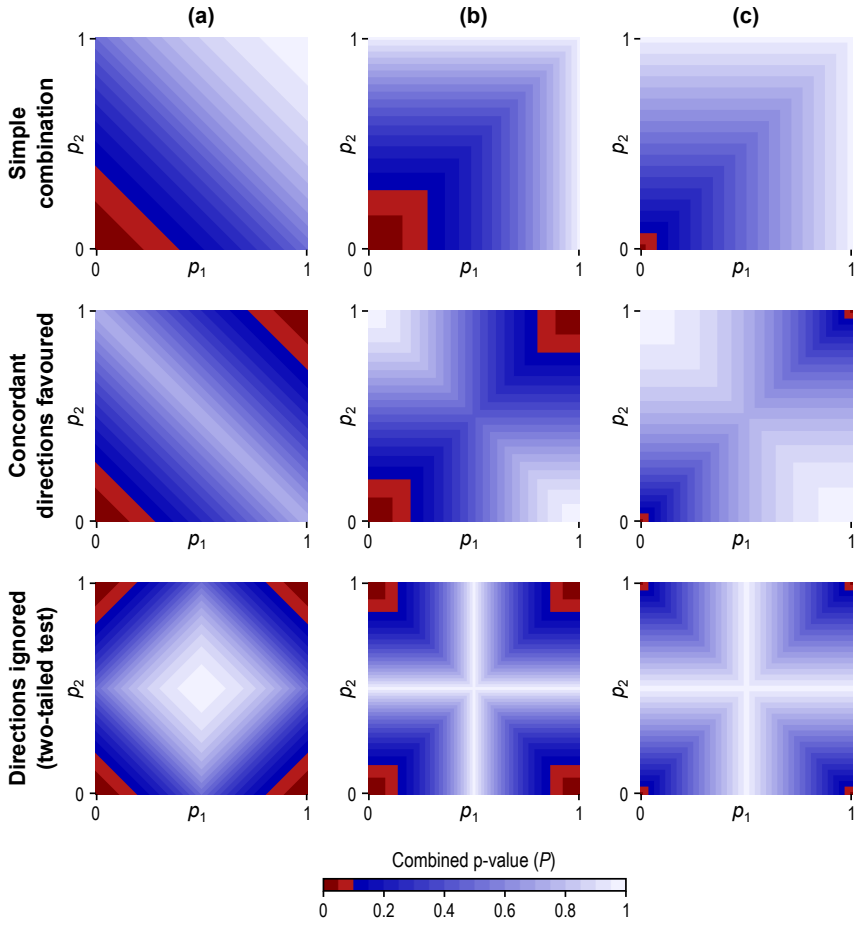


Figure 4.4: Examples of inconsistent combining functions for testing the global null hypothesis: *(a)* Addition of p-values for the partial tests (Edgington, 1972); *(b)* Maximum of p-values for the partial tests, with the p-value computed as T^K (Friston et al., 1999, 2005); *(c)* Maximum of p-values for the partial tests, but with the p-value computed as T (Nichols et al., 2005). While the last is not appropriate for testing the global null, it is appropriate for the conjunction null.

conditions above are satisfied. The particular combining function shown in Equation 4.2 was suggested by Pearson (1933) and used by David (1934), but after an influential paper by Birnbaum (1954), it was for decades thought to be inadmissible. However, it is in fact admissible (Owen, 2009).

Admissibility is important in that it allows, for more than just two partial tests, combined tests that favour alternative hypotheses with the same direction. Other possibilities favouring alternatives with common direction, such as multiplying together the partial test statistics to produce a combined statistic, work for two partial tests only (Hayasaka et al., 2006).

4.2.11 The method of Tippett

From the various combining functions listed in Table 4.1, consider the combining function of Tippett (1931), that has statistic $T = \min p_k$ and, when all partial tests are independent, a p-value $P = 1 - (1 - T)^K$. This test has interesting properties that render it particularly attractive for imaging:

- It defines a UIT test: If the minimum p-value remains significant when all tests are considered, clearly the global null hypothesis can be rejected.

- It controls the FWER: Controlling the error rate of a UIT is equivalent to an FWER-controlling procedure over the partial tests.

- If the partial tests are independent, it defines an exact FWER threshold: The function is closely related to Šidák (1967) correction: set $P = \alpha^{\text{FWER}}$, then $T^{\text{FWER}} = 1 - (1 - \alpha^{\text{FWER}})^{\frac{1}{K}}$; one can retain only the partial p-values that satisfy $p_k \leqslant T^{\text{FWER}}$. Adjusted p-values can be obtained similarly through the Šidák procedure, that is $p_k^{\text{FWER}} = 1 - (1 - p_k)^{\frac{1}{K}}$.

- If the partial tests are not independent, it still defines an FWER threshold and adjusted p-values: As a UIT, the Tippett function can be used in a closed testing procedure. Further, it is *the* function that makes CTP with large K feasible in practice; adjusted p-values are obtained with the distribution of the minimum p-value (or of the extremum statistic).

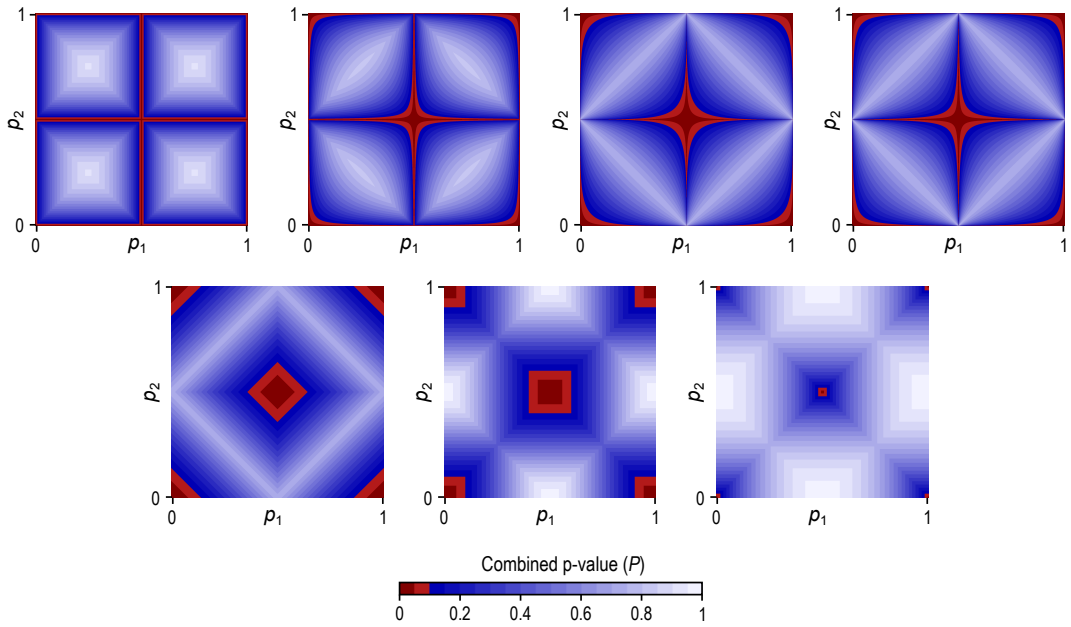


Figure 4.5: *Upper row:* Inadmissible versions of the four consistent combining functions shown in Figure 4.3 (in the same order). *Lower row:* Inadmissible versions of the three inconsistent combining functions shown in Figure 4.4 (in the same order). These inadmissible functions arise if one attempts to favour alternatives with the same sign while performing two-tailed partial tests.

- Because it subsumes correction using the extremum statistic that is already in use in imaging to account for MTP-I, the correction for the MTP-II can be done by pooling the maximum statistics across both space and the set of partial tests. This allows algorithmic advantages that we exploit in the proposed implementation shown in Section 4.2.13.

- It can be used as the combining function with NPC, thus providing a common procedure for correction and for combination of p-values.

- It is fast to compute: Taking the extremum statistic or minimum p-value is trivial compared to other functions that require cumulative sums or products, multiple parameters, integrations, or that depend on Monte Carlo simulations.

While the Tippett function is advantageous for all these reasons, note that, even when other combining functions are used for NPC, the extremal statistic (equivalent to the Tippett combining function) is also used for the MTP-I to control FWER over space.

4.2.12 A unified procedure

Armed with these concepts, and with the modifications to the original NPC algorithm, we are positioned to tackle the various problems identified in the Introduction:

Combination of multiple modalities With K modalities, all in register and with the same spatial resolution, each is tested separately, using synchronised permutations, and their statistics converted to u-values for each shuffling. These are are combined using a suitable combining function, such as one from those shown in Table 4.1. The p-values for the combined statistic are produced using the same set of permutations used to assess each test separately. This is the modified NPC algorithm that we propose, shown in Figure 4.2.

Correction for multiple modalities With K modalities, which are not necessarily in register, nor with the same resolution, nor of the same type (e.g., some from

volumetric, some from surface representations of the brain), or which may not necessarily be all related to imaging (e.g., some imaging and some non-imaging data), each is tested separately using a suitable test statistic. The permutation distribution of the extremum statistic across *all* tests is produced and used to compute FWER-adjusted p-values that simultaneously address the MTP-I and MTP-II.

Correction for multiple designs and contrasts Each pair of contrasts defined by $(\mathbf{C}, \mathbf{D})$ allows the corresponding design matrix to be partitioned into effects of interest and nuisance effects (Section 3.2.2), and also the redefinition of the response variables (Section 4.2.2). Thus, multiple designs and their respective contrasts can be tested separately. Differently than for the correction for multiple modalities, however, with different contrasts, their respective statistics may possess different asymptotic behaviour (due to, e.g., the contrasts having different ranks, or the designs having different degrees of freedom), thus precluding the use of the distribution of the extremum statistic. When known, the asymptotic behaviour can be used to convert these statistics — univariate or multivariate — to a z-statistic. The distribution of the maximum across the results of the various designs and contrasts can then be computed and used for correction.

Correction for multiple modalities, designs and contrasts Following the same principles, it is also possible to account for the multiplicity of input modalities, each tested with their respective design and set of contrasts, or each tested versus all designs and contrasts. Each test is applied separately, statistics converted to a z-statistic based on their asymptotic behaviour, and the distribution of the extremum used to obtain adjusted p-values for all in a CTP using a UIT. It is not necessary that all are in register, neither that all use the same kind of image representation of the brain (i.e., volume or surface), nor that they are even all (or any) imaging-related, and can therefore include clinical or behavioural, biomarkers, and other types of data.

Conjunctions An IUT can be assessed through permutations simply by computing $\max(p_k)$, which is, in its own right, the p-value of the IUT, such that there is no need for transformation into u-values for the assessment of the combined statistic.

In the context of imaging, such conjunctions can be used with statistics at every voxel (or vertex or face), thus allowing also certain spatial statistics such as TFCE.

Since combinations and conjunctions are performed at each individual image point, it is necessary that all images have been registered to the same common space and possess similar spatial resolution (Lazar et al., 2002). This can be accomplished through intra-subject and inter-subject registration and resampling. By contrast, correction for the multiplicity of tests uses the maximum statistic across such tests, thus not requiring that the tests match on space, or even that they are all related to imaging. However, they explicitly require pivotal statistics, as discussed in Section 3.2.1, so that the extreme is taken from statistics that share the same sampling distribution. The statistics used with CMV and NPC are all pivotal and therefore can be used. Spatial statistics, however, lack this property and require similar search volumes and resolutions, even for correction. Moreover, by including information from neighbouring voxels, such as using spatial smoothing or spatial statistics like TFCE (Smith and Nichols, 2009), subset pivotality is lost, meaning that strong control of FWER cannot be guaranteed. In practice, though, the power gained by pooling information over space is essential. In the Section 4.2.13 we provide an algorithm that generically implements the combination and correction methods presented.

4.2.13 Implementation

A unified algorithm for combination and correction that is amenable for use with imaging applications is shown below. It has many similarities with the `randomise` algorithm (Section 3.2.6; Winkler et al., 2014), with various modifications to accommodate combination and correction. The p-values adjusted for the multiplicity of tests are computed using the distribution of the extremum statistic, which can be collapsed across modalities and/or designs and contrasts for each case, rendering the algorithm simpler. The notation below is slightly different than that used throughout the chapter. The inputs are:

- $\mathbf{Y}$: The input data for each of the K modalities and image points. Each column vector of N observations for the k-th modality is accessed as $\mathbf{Y}[k, \mathbf{v}]$,

where $\mathbf{v} = [x, y, z]$ is used to specify the point position in space; this is so without loss of generality for non-imaging data.

- $\mathcal{X}$: The set of design matrices $\mathbf{X}$.
- $\{\mathcal{C}_\mathrm{X}\}$: The set of sets of contrasts for each design matrix $\mathbf{X}$. Each element of each subset is a pair of multivariate contrasts $(\mathbf{C}, \mathbf{D})$. This definition allows each design to be tested with multiple such pairs of contrasts, and allows various designs to be tested with the same input data.
- $\mathbf{B}$: Definition of exchangeability blocks, used to define valid shufflings that respect the data structure; these can be multi-level (Winkler et al., 2015).
- $\mathbf{V}$: Definition of the variance groups, useful to compute statistics that are robust to heteroscedasticity.
- EE, ISE: Boolean flags (true/false) indicating whether errors can be treated as exchangeable (EE), allowing permutations, independent and symmetric (ISE), allowing sign-flippings, or both.
- J: Number of permutations to be performed.
- NPCMOD, NPCCON: Boolean indicating whether combination should be performed respectively across modalities, across designs and contrasts, or both.
- FWEMOD, FWECON: Boolean indicating whether familywise error rate correction should be performed respectively across modalities, across designs and contrasts, or both.

The output of interest is the p-value. For simplicity, as shown, the output is always FWER-adjusted across the image points indexed by $\mathbf{v}$, and for the non-combined, further adjusted based on the contrasts and modalities; these are shown in the algorithm topped by a tilde, that is, as "p̃-value", as opposed to simply "p-value". Also for simplicity, p-values for combination of modalities are not shown adjusted for multiple contrasts, nor vice-versa. These can also be obtained following the same logic used for the FWER-adjustment of the non-combined statistics. Uncorrected p-values, useful for correction using false discovery rate (FDR, Benjamini and Hochberg, 1995) can be obtained with trivial modifications.

Algorithm 2: Unified algorithm. See the main text for details.

Require: $\mathbf{Y}, \mathcal{X}, \{\mathcal{C}_\mathbf{X}\}, \mathbf{B}, \mathbf{V}$, EE, ISE, J, NPCMOD, NPCCON, FWEMOD, FWECON.

$\mathcal{P} \leftarrow$ sync_perms$(\mathcal{X}, \{\mathcal{C}_\mathbf{X}\}, \mathbf{B}$, EE, ISE, $J-1)$ ▷ Define the permutation set.
$\mathcal{P} \leftarrow \{\mathbf{I}, \mathcal{P}\}$ ▷ Ensure first permutation is no permutation.
for $j = 1, \ldots, J$ **do** ▷ For each shuffling.
  $c \leftarrow 1$ ▷ Counter for the number of designs and contrasts.
  for all $\mathbf{X} \in \mathcal{X}$ **do** ▷ For each design matrix.
    for all $(\mathbf{C}, \mathbf{D}) \in \mathcal{C}_\mathbf{X}$ **do** ▷ For each pair of contrasts.
      $\mathbf{Y} \leftarrow \mathbf{YD}$ ▷ Redefine the data, discard $\mathbf{D}$.
      $\mathbf{X}^* \leftarrow \mathbf{P}_j\mathbf{X}$ ▷ Shuffle the model.
      for all $k \in \{1, \ldots, K\}$ **do** ▷ For each partial test.
        for all v do ▷ For each image point.
          $\hat{\boldsymbol{\beta}} \leftarrow (\mathbf{X}^*)^+\mathbf{Y}[k, \mathbf{v}]$ ▷ Estimated regression coefficients.
          $\hat{\boldsymbol{\epsilon}} \leftarrow \mathbf{Y}[k, \mathbf{v}] - \mathbf{X}^*\hat{\boldsymbol{\beta}}$ ▷ Estimation residuals.
          $\mathbf{G} \leftarrow$ pivotal$(\mathbf{X}^*, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}, \mathbf{C}, \mathbf{V})$ ▷ Test statistic.
          $\mathbf{U}[j, k, c, \mathbf{v}] \leftarrow$ transform$(\mathbf{G})$ ▷ Transform to u-value.
          if $j = 1$ **then** ▷ In the first permutation (no permutation).
            $\mathbf{U}_0[k, c, \mathbf{v}] \leftarrow \mathbf{U}[1, k, c, \mathbf{v}]$ ▷ Keep the unpermuted u-value.
          end if
        end for
        $\mathbf{U}_\mathrm{e}[j, k, c] \leftarrow$ extremum$(\mathbf{U}[j, k, c, \cdot])$ ▷ Extremum across space.
      end for
      $c \leftarrow c + 1$ ▷ Increment counter for the number of designs and contrasts.
    end for
  end for
  $C \leftarrow c$ ▷ Keep the total number of designs and contrasts for later use.
  if NPCMOD $\wedge \neg$ NPCCON **then** ▷ Combine modalities only.
    for all $c \in \{1, \ldots, C\}$ **do** ▷ For each design/contrast.
      for all v do ▷ For each image point.
        $\mathbf{T}[c, \mathbf{v}] \leftarrow$ combine$(\mathbf{U}[j, \cdot, c, \mathbf{v}])$ ▷ Combined statistic.
      end for
      $\mathbf{T}_\mathrm{e}[j, c] \leftarrow$ extremum$(\mathbf{T}[c, \cdot])$ ▷ Distribution of the extrema across tests.
    end for
  else if NPCCON $\wedge \neg$ NPCMOD **then** ▷ Combine designs/contrasts only.
    for all $k \in \{1, \ldots, K\}$ **do** ▷ For each design/contrast.
      for all v do ▷ For each image point.
        $\mathbf{T}[k, \mathbf{v}] \leftarrow$ combine$(\mathbf{U}[j, k, \cdot, \mathbf{v}])$ ▷ Combined statistic.
      end for
      $\mathbf{T}_\mathrm{e}[j, k] \leftarrow$ extremum$(\mathbf{T}[k, \cdot])$ ▷ Distribution of the extrema across tests.
    end for
  else if NPCMOD $\wedge$ NPCCON **then** ▷ Combine modalities & designs/contrasts.
    for all v do ▷ For each image point.
      $\mathbf{T}[\mathbf{v}] \leftarrow$ combine$(\mathbf{U}[j, \cdot, \cdot, \mathbf{v}])$ ▷ Combined statistic.
    end for
    $\mathbf{T}_\mathrm{e}[j] \leftarrow$ extremum$(\mathbf{T}[\cdot])$ ▷ Distribution of the extrema across tests.

end if
if $j = 1$ **then** ▷ In the first permutation (no permutation).
$\mathbf{T}_0 \leftarrow \mathbf{T}$ ▷ Keep the unpermuted combined statistic.
end if
end for
if NPCMOD $\wedge$ $\neg$ NPCCON **then** ▷ Combine modalities only.
for all $c \in \{1, \ldots, C\}$ **do** ▷ For each design/contrast.
for all $\mathbf{v}$ **do** ▷ For each image point.
p-value$[c, \mathbf{v}] \leftarrow$ data_pval$(\mathbf{T}_0[c, \mathbf{v}], \mathbf{T}_\mathrm{e}[\cdot, c])$ ▷ Combined p-value.
end for
end for
else if NPCCON $\wedge$ $\neg$ NPCMOD **then** ▷ Combine designs/contrasts only.
for all $k \in \{1, \ldots, K\}$ **do** ▷ For each design/contrast.
for all $\mathbf{v}$ **do** ▷ For each image point.
p-value$[k, \mathbf{v}] \leftarrow$ data_pval$(\mathbf{T}_0[k, \mathbf{v}], \mathbf{T}_\mathrm{e}[\cdot, k])$ ▷ Combined p-value.
end for
end for
else if NPCMOD $\wedge$ NPCCON **then** ▷ Combine modalities & designs/contrasts.
for all $\mathbf{v}$ **do** ▷ For each image point.
p-value$[\mathbf{v}] \leftarrow$ data_pval$(\mathbf{T}_0[\mathbf{v}], \mathbf{T}_\mathrm{e}[\cdot])$ ▷ Combined p-value.
end for
end if
if FWEMOD $\wedge$ $\neg$ FWECON **then** ▷ Correct over modalities only.
for all $c \in \{1, \ldots, C\}$ **do** ▷ For each design/contrast.
for all $j \in \{1, \ldots, J\}$ **do** ▷ For each shuffling.
$\mathbf{U}'_\mathrm{e}[j, c] \leftarrow$ extremum$(\mathbf{U}_\mathrm{e}[j, \cdot, c])$ ▷ Distribution of the extrema.
end for
for all $k \in \{1, \ldots, K\}$ **do** ▷ For each modality.
for all $\mathbf{v}$ **do** ▷ For each image point.
$\tilde{\text{p}}$-value$[k, c, \mathbf{v}] \leftarrow$ data_pval$(\mathbf{U}_0[k, c, \mathbf{v}], \mathbf{U}'_\mathrm{e}[\cdot, c])$ ▷ Adjusted p-value.
end for
end for
end for
else if FWECON $\wedge$ $\neg$ FWEMOD **then** ▷ Correct over designs/contrasts only.
for all $k \in \{1, \ldots, K\}$ **do** ▷ For each modality.
for all $j \in \{1, \ldots, J\}$ **do** ▷ For each shuffling.
$\mathbf{U}'_\mathrm{e}[j, k] \leftarrow$ extremum$(\mathbf{U}_\mathrm{e}[j, k, \cdot])$ ▷ Distribution of the extrema.
end for
for all $c \in \{1, \ldots, C\}$ **do** ▷ For each design/contrast.
for all $\mathbf{v}$ **do** ▷ For each image point.
$\tilde{\text{p}}$-value$[k, c, \mathbf{v}] \leftarrow$ data_pval$(\mathbf{U}_0[k, c, \mathbf{v}], \mathbf{U}'_\mathrm{e}[\cdot, k])$ ▷ Adjusted p-value.
end for
end for
end for
else if FWEMOD $\wedge$ FWECON **then** ▷ Correct over modalities & des./contr.
for all $j \in \{1, \ldots, J\}$ **do** ▷ For each shuffling.
$\mathbf{U}'_\mathrm{e}[j] \leftarrow$ extremum$(\mathbf{U}_\mathrm{e}[j, \cdot, \cdot])$ ▷ Distribution of the extrema.

```
    end for
    for all k ∈ {1, ..., K} do                          ▷ For each modality.
      for all c ∈ {1, ..., C} do                        ▷ For each design/contrast.
        for all v do                                    ▷ For each image point.
          p̃-value[k, c, v] ← data_pval(U_0[k, c, v], U'_e[·])   ▷ Adjusted p-value.
        end for
      end for
    end for
  end if
```

Within the algorithm, the functions are:

- *sync_perms*: This function produces a set $\mathcal{P}$ of permutation and/or sign flipping matrices that can be performed sychronously to test a joint null hypotheses about the input data. The synchronisation is always necessary to allow combination/correction over modalities, and it may also be necessary across multiple designs and/or contrasts if these are to be combined/corrected as well. If synchronisation is not necessary for designs and/or contrasts, the algorithm can be modified so that $\mathcal{P}$ can be defined inside the for-loops that iterate over designs and contrasts.

- *transform*: This converts the test statistic into a u-value, thus rendering the NPC method feasible for imaging applications. If no combination is to be performed, the algorithm can be modified to skip this step and work directly with the test statistic.

- *extremum*: For statistics in which larger values are evidence against the null hypothesis, this function takes the maximum. For statistics in which smaller values are indication against the null, this takes the minimum. In either case, it is always the most extreme towards evidence favouring the alternative. This function effectively implements a CTP using an IUT.

- *combine*: This combines the inputs (p- or u-values) into a new, combined statistic. Any of the combining functions from Table 4.1 can be considered. For the method of Tippett, *combine* and *extremum* are the same.

– *data_pval*: This function produces a p-value based on a set of empirical values for the test statistic after shuffling. This works by computing the fraction of the test statistics after shuffling that is larger or equal than the unpermuted test statistic, while taking care of ties.

The algorithm has four major parts: the first consists of the loop that begins in line 5 of the pseudocode above, and which consists of a simplified version of the `randomise` algorithm. The second begins with the conditional structure in line 27, which performs the combination and computes distribution of the extremum statistics for each case of NPC, thus also treating the MTP-I. These initial two parts are repeated for each shufflong, in the loop that begins in line 3. The third part begins with the conditional in line 51, that is, once all rearrangements have been performed; in this part, the distributions are used to compute the combined p-values. Finally, the fourth part begins with the conditional in line 68, in which the MTP-II is addressed.

As shown, the algorithm is simplified so as to emphasise the most important aspects of combination and correction. However, various modifications and improvements can be applied for particular circumstances, and for speed, including the partitioning discussed in the Section 3.2.2. An open-source working implementation, that can be executed in MATLAB (The MathWorks Inc., 2015) or Octave (Eaton et al., 2015), is available in the tool *Permutation Analysis of Linear Models* (PALM), available for download at `www.fmrib.ox.ac.uk/fsl`.

4.3 Evaluation methods

4.3.1 Validity of the modified NPC

To assess the validity of the proposed modification to the NPC, we consider one of the simplest scenarios that would have potential to invalidate the method and reduce power: this is the case of having a small number of partial tests, small sample size, and with each partial test possessing substantially different distributions for the error terms. We investigated such a scenario with $K = 2$, varying sample sizes $N = \{8, 12, 20, 30, 40, 50, 60, 70, 80, 120, 200\}$, and different error distributions. Using the notation defined in Section 4.2.2, response variables were generated for

each simulation using the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbf{Y}$ sized $N \times K$. Each modality was simulated as having 500 points, these representing, for instance, voxels or vertices of an image representation of the brain. The errors, $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2]$, were simulated following either a Gaussian distribution with zero mean and unit variance, or a Weibull distribution (skewed), with scale parameter 1 and shape parameter 1/3, shifted and scaled so as to have expected zero mean and unit variance. Different combinations of error distributions were used: Gaussian for both partial tests, Weibull for both partial tests, or Gaussian for the first, and Weibull for the second partial test.

The response data, $\mathbf{Y}$, were constructed by adding the simulated effects, $\mathbf{X}\boldsymbol{\beta}$, to the simulated errors, where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$, with $\boldsymbol{\beta}_k = [\beta_1, 0]'$, β_1 being either 0 (no signal) or $t_{\text{cdf}}^{-1}\left(1 - \alpha; N - \text{rank}\left(\mathbf{X}\right)\right) \Big/ \sqrt{N}$ (with signal), where $\alpha = 0.05$ is the significance level of the permutation test to be performed. This procedure ensures a calibrated signal strength sufficient to yield an approximate power of 50% for each partial test, with Gaussian errors, irrespective of the sample size; for non-Gaussian errors this procedure does not guarantee power at the same level. The actual effect was coded in the first regressor of $\mathbf{X}$, constructed as a vector of random values following a Gaussian distribution with zero mean and unit variance; the second regressor was modelled an intercept. All four possible combinations of presence/absence of effect among the $K = 2$ partial tests were simulated, that is, (1) with no signal in any of the two partial tests, (2) with signal in the first partial test only, (3) with signal in the second partial test only, and (4) with signal in both partial tests.

The simulated data was tested using the Tippett and Fisher methods. The case with complete absence of signal was used to assess error rates, and the others to assess power. The p-values were computed with 500 permutations, and the whole process was repeated 500 times, allowing histograms of p-values to be constructed, as well as to estimate the variability around the heights of the histogram bars. Confidence intervals (95%) were computed for the empirical error rates and power using the Wilson method (Wilson, 1927). The p-values were also compared using Bland–Altman plots (Bland and Altman, 1986), modified so as to include the confidence intervals around the means of the methods.

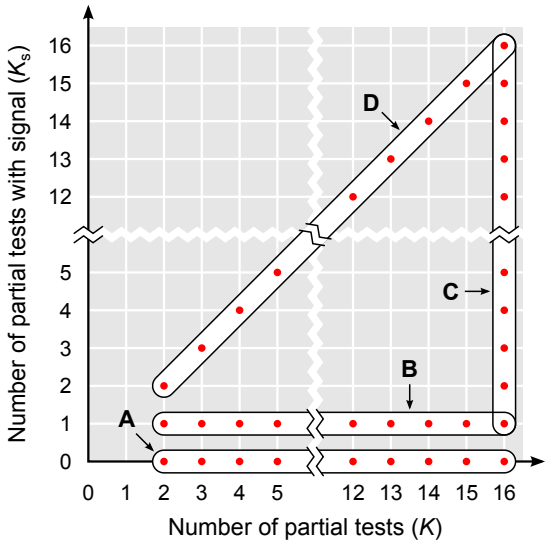


Figure 4.6: The simulations A–D. Each was constructed with a set of K partial tests, a number of which (K_s) had synthetic signal added.

4.3.2 Performance of combined tests

We also took the opportunity to compare the combining functions shown in Table 4.1. While other comparisons have been made in the past (for a list of references, see Section 4.2.6), none included *all* these functions, nor explored their performance under permutation or NPC, and therefore, did not consider the modifications that we introduce to the procedure to render it feasible for imaging applications. In addition, we investigate the performance of two classical multivariate tests, the Hotelling's T^2, and the Wilks' λ, both assessed through permutations.

Four different simulation sets were conducted, named A–D; in all, the number of partial tests being combined could vary in the range $K = 2, \ldots, 16$, and the number of partial tests containing true, synthetic signal could vary in the range $K_s = 0, \ldots, K$. In simulation A, K varied, while K_s was held fixed at 0, that is, no synthetic signal was added. In simulation B, K varied, while K_s was held fixed at 1, that is, just one partial test had signal added. In simulation C, K was held fixed at 16, while K_s varied. Finally, in simulation D, K varied, and K_s was set as equal to K, that is, all partial tests had synthetic signal added. Figure 4.6 shows graphically how K and K_s varied in each simulation.

The response variables $\mathbf{Y}$ had size $N \times K$, $N = 20$, that is, simulating measure-

ments for 20 subjects, each with K image modalities (partial tests). Each modality was simulated as having 500 points, these representing, for instance, voxels or vertices. The errors were simulated following either a Gaussian distribution with zero mean and unit variance, or a Weibull distribution, with scale parameter 1 and shape parameter $1/3$, shifted and scaled so as to have expected zero mean and unit variance. The response data were constructed by adding to the errors the simulated effects — either no signal, or a signal with strength callibrated to yield an approximate power of 50% with Gaussian errors, irrespective of the sample size, as described above for the simulations that tested the validity of the modified NPC; for the Weibull errors, the signal was further decreased, in all these four simulations, by a factor $5/8$, thus minimising saturation at maximum power in simulation D. The actual effect was coded in the first regressor only, which was constructed as a set of random values following a Gaussian distribution with zero mean and unit variance; the second regressor was modelled as an intercept.

The simulated data was tested using 500 shufflings (permutations, sign-flippings, and permutations with sign-flippings). For all the simulations, the whole process was repeated 100 times, allowing histograms of p-values to be constructed, as well as to estimate the variability around the heights of the histogram bars. Confidence intervals (95%) were computed for the empirical error rates and power using the Wilson method.

4.3.3 Example: Pain study

While the proposed correction for the MTP-II has a predictable consequence, that is, controlling the familywise error rate at the nominal level, the combination of modalities, designs, and contrasts may not be quite as obvious. In this section we show a re-analysis of the data of the pain study by Brooks et al. (2005). In brief, subjects received, in separate tests, painful, hot stimuli in the right side of the face (just below the lower lip), dorsum of the right hand, and dorsum of the right foot. The objective was to investigate somatotopic organisation of the pain response in the insular cortex using FMRI, and the complete experimental details, stimulation and imaging acquisition protocols, analysis and conclusions can be found in the original publication. Here we sought to identify, at the group level, in

standard space, areas within the insula that jointly respond to hot painful stimuli across the three topologically distinct body regions. We used the modified NPC, comparing the combining functions of Tippett, Fisher, Stouffer and Mudholkar–George, as well as the Hotelling's T^2 statistic, and an IUT (conjunction). At the group level, the design is a one-sample t-test, for which only sign flippings can be used to test the null hypothesis. We used twelve of the original subjects, and performed exhaustively all the 4096 sign flippings possible.

4.4 Results

A large number of plots and tables were produced and are shown in the Appendix B. The Figures below contain only the most representative results, that are sufficient to highlight the major points.

4.4.1 Validity of the modified NPC

Both the original and the modified NPC methods controlled the error rates at exactly the level of the test. Such validity was not limited to $\alpha = 0.05$, and the histograms of uncorrected p-values under complete absence of signal were flat throughout the whole $[0, 1]$ interval for both the original and modified NPC methods, using either the Tippett or the Fisher combining functions. A representative subset of the results, for the Fisher method only, and for sample sizes $N = \{8, 12, 20, 40\}$, is shown in Figure 4.7.

When considering the uncorrected p-values, the modified NPC yielded a mostly negligible increase in power when compared to the original NPC, with the difference always within the 95% confidence interval. Although this slight gain can be hardly observed in the histograms and Bland–Altman plots for the uncorrected p-values, they are clearly visible in the Bland–Altman plots for the p-values corrected across the 500 tests. In these plots, the predominance of smaller (towards more significant) p-values can be seen as a positive difference between the original and modified NPC p-values. A representative subset of the results is shown in Figure 4.8.

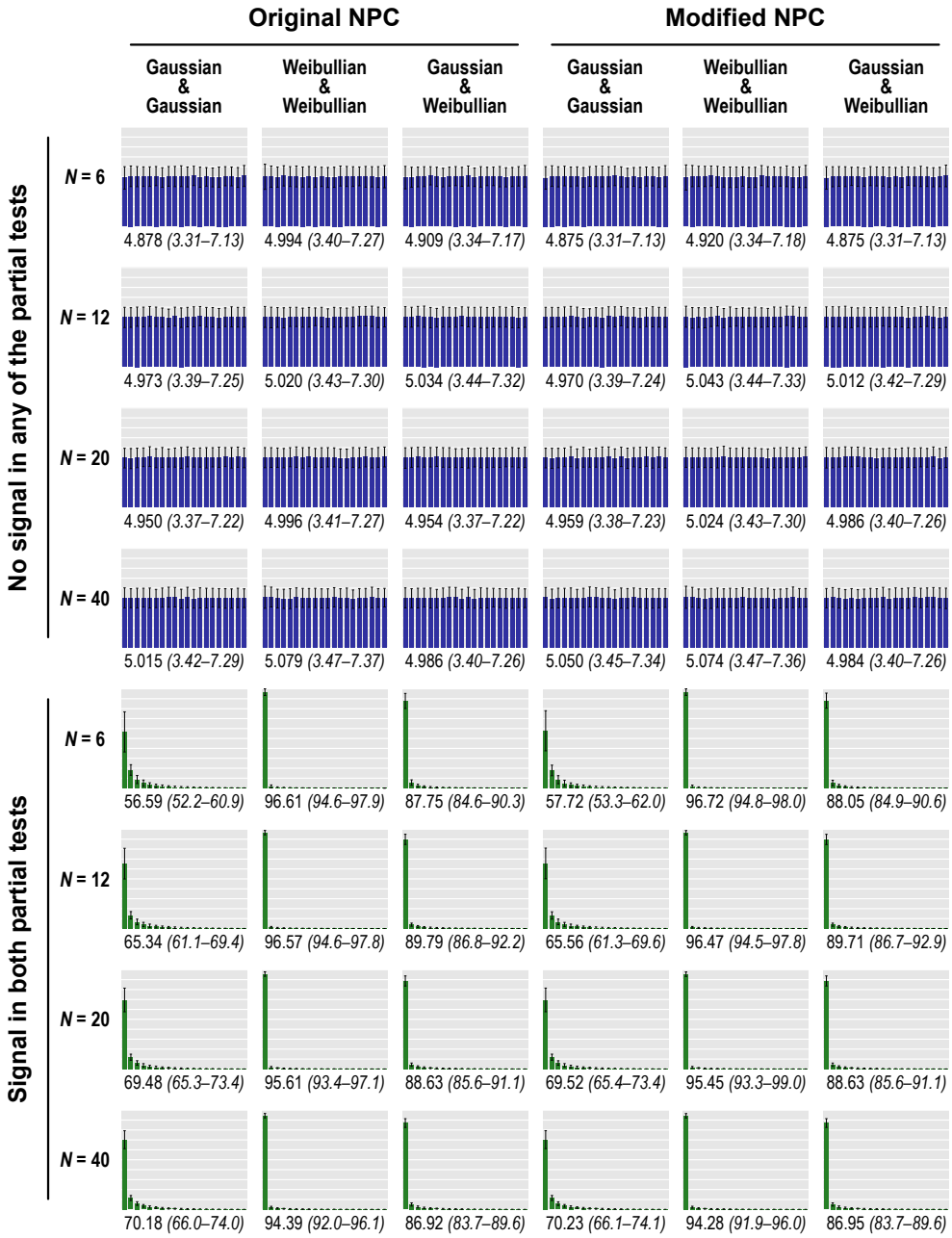


Figure 4.7: Histograms of p-values for the simulation without signal in either of the two partial tests (upper panel, blue bars) or with signal in both (lower panel, green bars). The values below each plot indicate the height (in percentage) of the first bar, which corresponds to p-values smaller than or equal to 0.05, along with the confidence interval (95%, italic). Both original and modified NPC methods controlled the error rates at the nominal level, and produced flat histograms in the absence of signal. The histograms suggest similar power for both approaches. See also Appendix B.

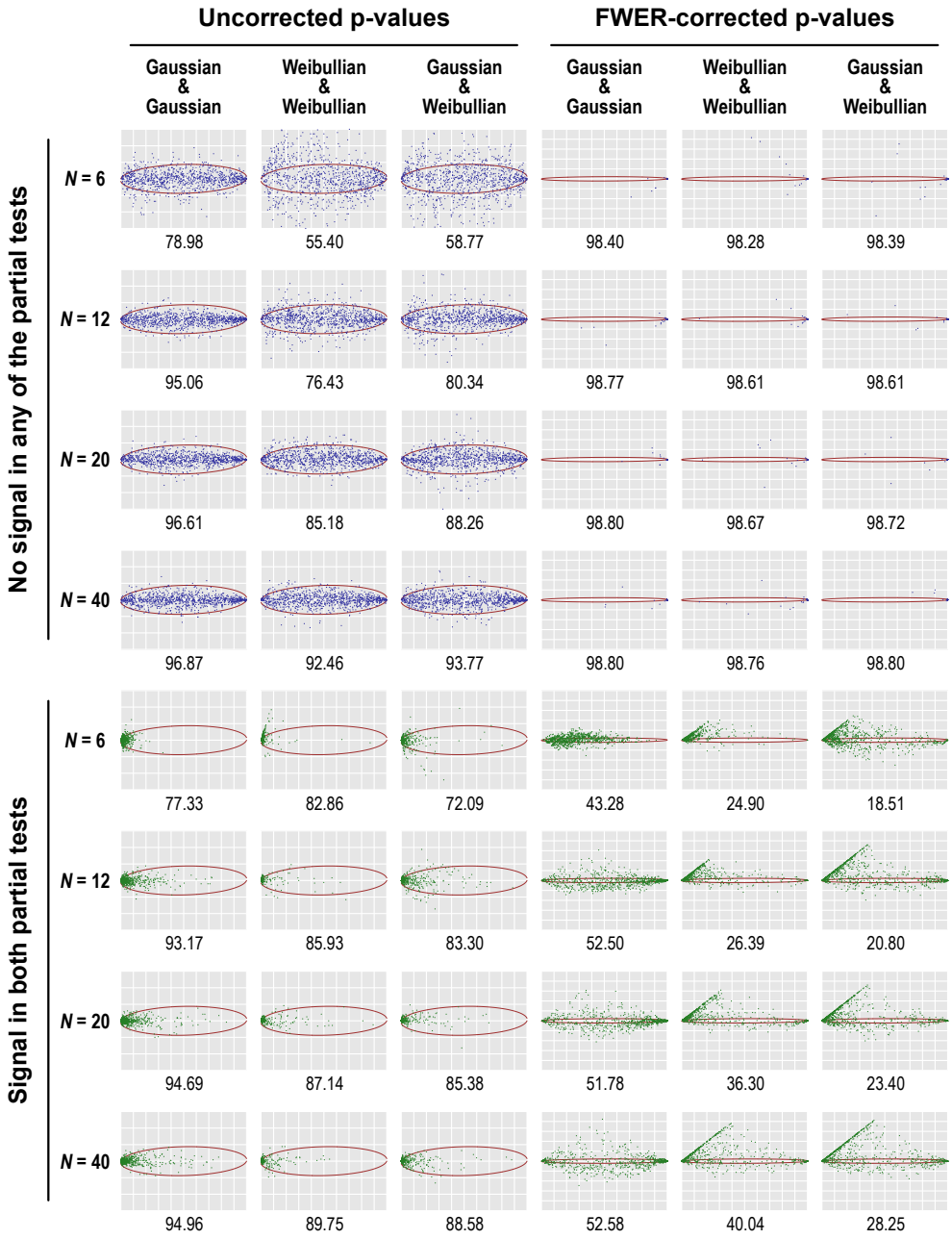


Figure 4.8: Bland–Altman plots comparing the original and modified NPC, for both uncorrected and corrected p-values, without signal in either of the two partial tests (upper panel, blue dots) or with signal in both (lower panel, green dots). The values below each plot indicate the percentage of points within the 95% confidence interval ellipsoid. For smaller sample sizes and non-Gaussian error distributions, the methods differ, but the differences become negligible as the sample size increases. In the presence of signal, the modification caused increases in power, particularly for the corrected p-values, with dots outside and above the ellipsoid. See the Supplementary Material for zoomed in plots, in which axes tick labels are visible.

4.4.2 Performance of combined tests

Representative results demonstrating the performance of the methods of Tippett, Fisher, Stouffer, Mudholkar–George, as well as Hotelling's T^2, is shown in Figure 4.9. The remaining results are browsable in Appendix B. In the absence of signal (simulation A), all combining functions controlled the error rate at the level of the test or below it, never above, thus confirming their validity. With normally distributed (Gaussian) errors, most functions yielded uniformly distributed p-values, although some functions seemed to converge towards uniformity only as the number of partial tests is increased; this was the case for the methods of Wilkinson, Zaykin, Dudbridge–Koeleman (DTP) and Jiang. With skewed (Weibullian) errors, the error rate was controlled at the test level with the use of permutations; with sign-flippings or permutations with sign-flippings, the combined results tended to be conservative, and more so for the Hotelling's T^2 statistics (and likewise the Wilks' λ).

With signal added to just one of the partial tests (simulation B), the method of Tippett was generally the most powerful, followed by the methods of Fisher and Dudbridge–Koeleman (both RTP and DTP variants). As the number of tests was increased, predictably, the power was reduced for all tests. The method of Stouffer did not in general have good performance with skewed errors, presumably because the dependence on z-statistics strengthens the dependence on the assumption of normality of the statistics for the partial tests in the modified NPC. The CMV did not deliver a good performance either, being generally among the least powerful.

With the number of partial tests held fixed, as the number of tests with signal was increased (simulation C), the power of the method of Fisher increased more quickly than of the other methods, although when most of the partial tests had signal, most of the combining functions reached similar power, all close to 100% for both normal or skewed errors. Hotelling's T^2 test was the considerably less powerful than any of the combining functions used with the modified NPC.

As the total number of partial tests and the number of partial tests with signal were both increased (simulation D), almost all combined tests had similar power, and reached saturation (100% power) quickly, particularly for the Weibullian errors, in which the calibration, even after reduction with the $5/8$ factor, yielded power

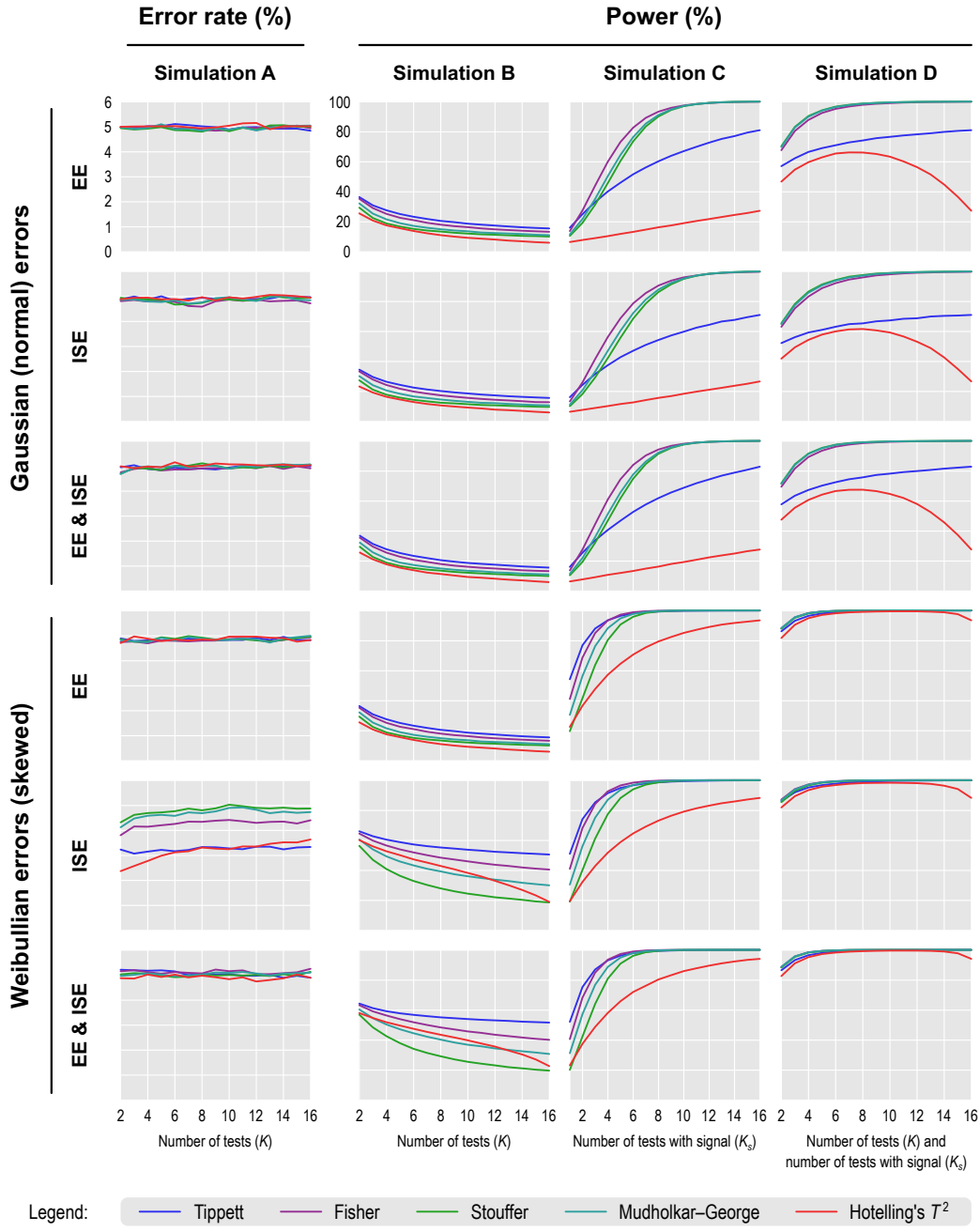


Figure 4.9: Performance of the modified NPC with four representative combining functions (Tippett, Fisher, Stouffer, and Mudholkar–George) and of one CMV (Hotelling's T^2), using normal or skewed errors, and using permutations (EE), sign flippings (ISE), or both. All resulted in error rates controlled at or below the level of the test. The Tippett and Fisher were generally the most powerful, with Tippett outperforming others with signal present in a small fraction of the tests, and with Fisher having the best power in the other settings.

above 50% for each partial test. With Gaussian errors, in which calibration ensured average 50% power, two tests had considerably lower sensitivity: Tippett's and Hotelling's T^2, the last with the remarkable result that power reached a peak, then began to fall as the number of tests kept increasing.

4.4.3 Example: Pain study

Using a conventional, mass univariate voxelwise tests, assessed through sign flippings, and after correction for multiple testing (MTP-I), only a few, sparse voxels could be identified at the group level for face, hand, and foot stimulation separately, in all cases with multiple distinct foci of activity observed bilaterally in the anterior and posterior insula. However, the joint analysis using the modified NPC with Fisher, Stouffer and Mudholkar–George evidenced robust activity in the anterior insula bilaterally, posterior insula, secondary somatosensory cortex (SII), and a small focus of activity in the midbrain, in the periaqueductal gray area. The combining function of Tippett, however, did not identify these regions, presumably because this method is less sensitive than the others when signal is present in more than a single partial test, as suggested by the findings in the previous section.

The Hotelling's T^2 was not able to identify these regions, with almost negligible, sparse, single-voxel findings in the anterior insula, bilaterally. The conjunction test, that has a different JNH, and searches for areas where all partial tests are significant, identified a single, barely visible, isolated voxel in the right anterior insula.

The above results are shown in Figure 4.10. Cluster-level maps that can directly be compared to the original findings of Brooks et al. (2005) are shown in Appendix B.

4.5 Discussion

4.5.1 Validity of the modified NPC

The modified NPC combines u-values, which are simply parametric p-values here renamed to avoid confusion. The renaming, however, emphasises the fact that the conversion to u-values via a parametric approximation should only be seen as

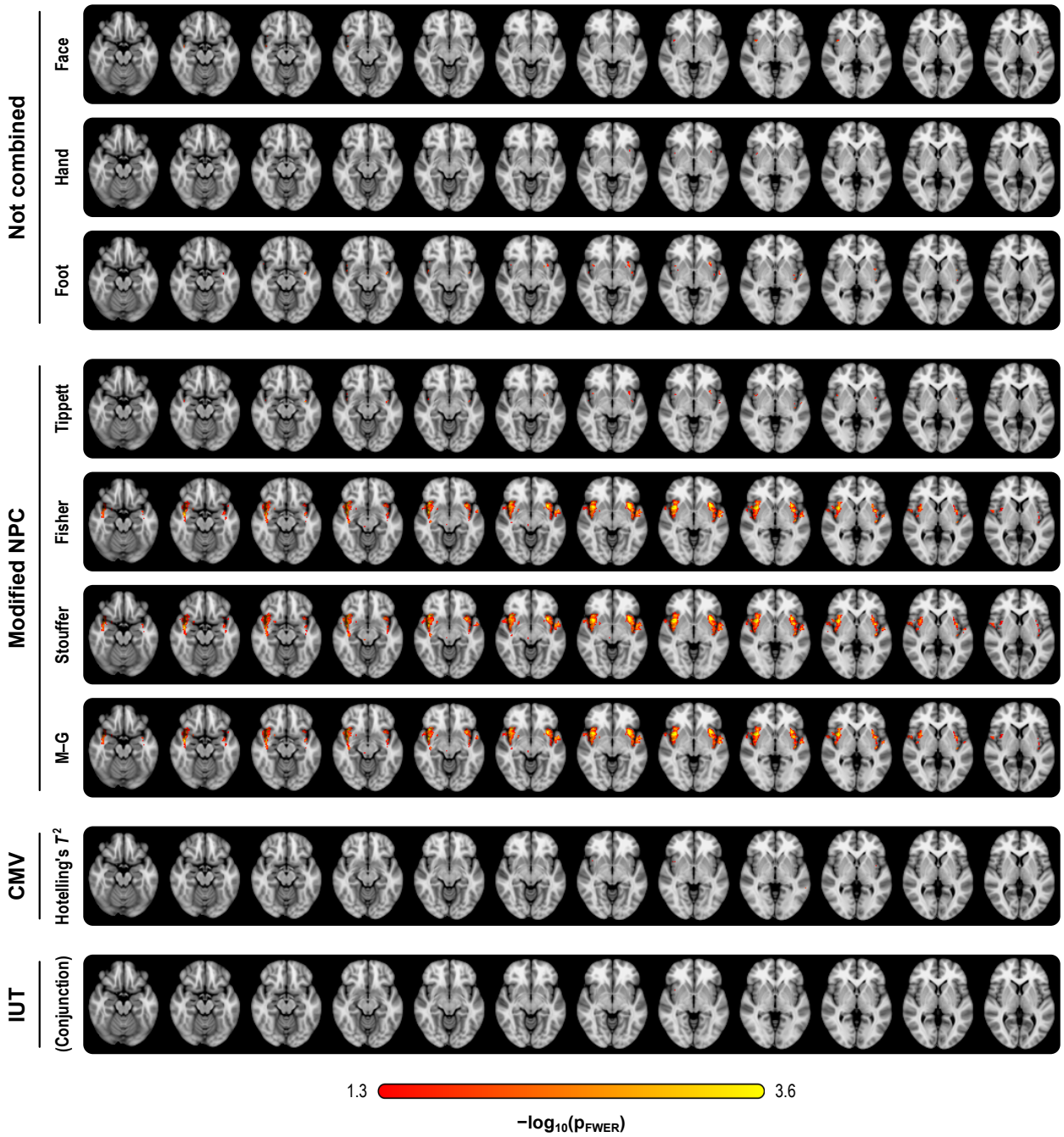


Figure 4.10: Without combination, and with correction across voxels (MTP-I), no significant results were observed at the group level for any of the three tests. Combination using the methods of Fisher, Stouffer and Mudholkar–George (M–G), however, evidenced bilateral activity in the insula in response to hot, painful stimulation. A classical multivariate test, Hotelling's T^2, as well as the Tippett method, failed to identify these areas. An intersection-union test (conjunction) could not locate significant results; such a test has a different null hypothesis that distinguishes it from the others. Images are in radiological orientation. For cluster-level results, comparable to Brooks et al. (2005), see Appendix B.

a data transformation, in which the interpretation as a p-value is not preserved due to untenable assumptions. The combination method continues to be non-parametric as the combined statistic is assessed non-parametrically. More importantly, irrespective of the validity of parametric assumptions, any dependence between the tests is accounted for, implicitly, by the combination procedure, without the need of any modelling that could, at best, introduce complex and perhaps untenable assumptions, and at worst, be completely intractable.

The results suggest that, even in the cases in which the modified NPC could have failed, i.e., with small sample sizes and different distributions, the combined statistic controlled the error rate at the level of the test. This control, maintained even in such difficult scenarios, suggests that the modified NPC controls the error rates in general. The results also suggest that the modification increases power, even if such increase is minute in some scenarios. The Bland–Altman plots indicate that gains in sensitivity are more pronounced in the results corrected for the MTP-I, suggesting that the modified method is appropriate not merely due to its expediency for imaging applications, but also for having increased sensitivity compared to the original NPC.

4.5.2 Performance of combined tests

The results also demonstrate that the NPC method is more powerful than the Hotelling's T^2. The superiority of combined permutation tests when compared to classical multivariate tests has been observed in the literature (Blair et al., 1994), and the fact that power increases as the number of partial tests with signal increases is one of its most remarkable features. While CMV depends on the positive-definiteness of the covariance matrix of the vectors of residuals, such limitation does not apply to NPC (Pesarin and Salmaso, 2010b). As a consequence, although in the comparisons only the Hotelling's T^2 and the Wilks' λ statistics were used (in the simulations, rank $(\mathbf{C}) = 1$), and had their p-values assessed through permutations, similar behaviour can be expected when using other CMVs, such as Pillai's trace (and with rank $(\mathbf{C}) > 1$). With effect, NPC can be used even when the number of variables equals or even greatly exceeds the number of observations, that is, when $K \geqslant N$. In the results shown in Figure 4.9, this can be noted as a reduction

in power that can be seen with the Hotelling's T^2, particularly for simulation D, and this is the case even considering that the test is assessed through permutations.

Regarding the different combining functions, the simulations show that the method of Tippett is the most powerful when signal is present in only a small fraction of the partial tests. For other cases, other combining functions, particularly that of Fisher, tend to be considerably more powerful.

The results also indicate that the use of sign flipping when the errors are not symmetric (a violation of assumptions) tends to produce a conservative test, with error rates below the nominal level, even if the power eventually remained unaltered when compared with permutations. While permutations together with sign flippings did alleviate conservativeness, at least for the Tippett method, the error rate remained below the nominal level. In general, if the errors are known to be skewed, only permutations should be used; if sign flippings are used, the error rate can be expected to be below the nominal level.

4.5.3 Interpretation of combined tests

The key aspect of the NPC is that these tests seek to identify, *on the aggregate* of the partial tests, a measure of evidence against the JNH, even if only some or none of them can be considered significant when seen in isolation, just as originally pointed out by Fisher (1932) (Section 4.2.1). This is the logic and interpretation of all of these combining statistics, with the exception of the conjunction inference. Combination is known to be able to answer questions that could otherwise not be answered be at all, or be answered less accurately if each information source were considered separately (Draper et al., 1992). Here the simulations and the pain study exemplify these aspects, and the improved sensitivity compared to each partial test when seen in separate.

As they depend on fewer assumptions than classical multivariate tests, NPC can be considered whenever the validity of the former cannot be guaranteed. Even when parametric CMV assumptions hold, note that the NPC can have superior power when sample size is small and prevents precise estimation of a covariance.

It should be noted that the aggregation of information follows a different principle than using different measurements separately to interrogate particular as-

pects of the brain (or of any other experiment or physiological phenomenon). Used judiciously, NPC provides a complete framework that can be used for both the aggregate and for the correction of tests separately, with the valuable feature of being based on minimal assumptions.

4.5.4 Correction over contrasts and over modalities

Correction over contrasts using synchronised permutations provides a novel solution to the multiple comparisons problem for certain common experimental designs, in particular, for the popular one-way ANOVA layout, that is, when the means of multiple groups are compared. The classical Fisher's protected least significant difference (LSD), that consists of performing an omnibus F-test and only proceeding to the group-wise post hoc tests if this initial test is significant, is known to fail to control the error rate if there are more than just three groups (Hayter, 1986; Hsu, 1996; Meier, 2006), and the failure can be by a wide margin, that grows as the number of groups being compared increases. Even though the same may not happen with other correction methods (e.g., Tukey's range test, Tukey, 1949), the correction done non-parametrically also renders these older, parametric methods, redundant.

The correction over contrasts further obviates methods that are based on what has been termed "logical constraints" among hypotheses (Shaffer, 1986; Hochberg and Tamhane, 1987), as the dependencies among the tests are implicitly taken into account by the correction using the distribution of the extremum across contrasts, with or without concomitant combination or correction across multiple K variables. In fact, the use of an omnibus F-test as a way to guard against multiple testing becomes quite unnecessary.

In the same manner, while combination across multiple modalities is a powerful substitute for classical multivariate tests as shown earlier, the correction across such modalities can replace the post hoc tests that are usually performed after significant results are found with CMVS.

4.5.5 Pain study

Joint significance is an important consideration when trying to interpret data such as these, that are distinct in some aspects (here, the topography of the stimulation), but similar in others (here, the type of stimulation, hot and painful), strengthening the case for distinct representations in some brain regions, but not in others. In terms of identifying areas with significant joint activity, the results suggest involvement of large portions of the anterior insula and secondary somatosensory cortex. The Fisher, Stouffer and Mudholkar–George combining functions were particularly successful in recovering a small area of activity in the midbrain and periaqueductal gray area that would be expected from previous studies on pain (Reynolds, 1969; Petrovic et al., 2002; Tracey et al., 2002; Roy et al., 2014), but that could not be located from the original, non-combined data.

4.5.6 Relationship with meta-analysis

Most of the combining functions shown in Table 4.1 were originally defined based on p-values, and some of them are popular in meta-analyses, such as those of Fisher and Stouffer (Borenstein et al., 2009). Although there are commonalities between these meta-analytical methods and NPC, it is worth emphasising that the two constitute distinct approaches to entirely different problems. In the NPC, the objective is to interrogate joint significance across the multiple observed variables (or multiple designs and contrasts if these are instead combined) when the data for each individual observation is readily available to the researcher. Meta-analyses methods based on p-values, while sometimes using the same combining functions, attempt to identify a joint effect across multiple studies that not have necessarily been performed on the same experimental units, and when the data for the individual observations are not available. Moreover, the p-value of the combined statistic in the NPC is produced through permutations, a procedure that is not available for ordinary meta-analytical methods.

The fact that NPC and meta-analysis form different approaches to separate problems also imply that certain criticisms levelled at the use of certain combined functions in the context of meta-analysis do not extend trivially to NPC. As the simulations show, various of the combining functions more recently developed did

not in general outperform older combining methods, such as Fisher and Stouffer, even though these were developed precisely for that purpose, in the context of meta-analyses, or for problems framed as such.

4.5.7 Applicability for cortical volumes

In the scope of this doctoral thesis, it is convenient to ask about the differences between investigating gray matter volume (an areal quantity) directly, as an univariate measurement computed as proposed in Section 2.2.3, or as a multivariate measurement, assessed through thickness and area separately, then combined through the modified NPC as proposed in this chapter. These two types of test would provide information about gray matter in two different ways. In the univariate case, a test could be significant in the presence of an effect affecting volume as a whole, but not necessarily either thickness or area in significant ways. Conversely, if either thickness or area is significant, due to the consistency of the combining functions used in NPC, it is expected that the joint analysis would assist in rejecting the null hypothesis of no effects in neither of the two. This can be the case even if effects acting on thickness and area follow opposite trends, such that volumes would remain generally unaltered (Brown and Jernigan, 2012).

4.6 Chapter conclusion

We proposed and evaluated a modified version of Non-Parametric Combination that is feasible and useful for imaging applications, and serves as a more powerful alternative to classical multivariate tests. We presented and discussed aspects related multiple testing problems in brain imaging, and proposed a single framework that addresses all these concerns at once. We showed that combination and correction of multiple imaging modalities, designs, and contrasts, are related to each other in the logic of their implementation, and also through the use of the simplest and the oldest of the combining functions, attributed to Tippett.

This page intentionally left blank.

Appendix A

Valorisation

According to the regulation governing the attainment of doctoral degrees at Universiteit Maastricht, an addendum about valorisation must be added to each doctoral thesis. This is the purpose of this section. Note that each of the previous chapters has its own conclusion.

A.1 Introduction

Before this work, there was a considerable confusion on what would be the proper way of studying the surface area of the cerebral cortex at a finer scale, across subjects, with many improvised approaches having appeared. None of these, however had sufficiently considered the deleterious impact that non-pycnophylactic methods would have on measurements. The work presented in Chapter 2 provided a much missed ground truth to which other, possibly faster even if approximate, strategies can be compared.

While the study of surface area is definitely not the only one to benefit from permutation inference, it is one where such non-parametric techniques can be applied in their whole potential, for providing robust inference even when assumptions about normality do not hold, and allowing correction for multiplicity of tests in spite of the complex dependence structure and irregular lattice of facewise areal data. The investigation of various regression and permutation strategies provided in Chapter 3, allied with the use of exchangeability blocks, variance groups, permutations together with sign-flippings, and a statistic that is robust to heterosce-

dasticity, the G-statistic, allows researchers to use permutation tests confidently even with complex designs.

For multimodal data, and for unexplored, yet pervasive types of multiple testing, the non-parametric combination (NPC) framework provided in Chapter 4 compares favourably against classical multivariate tests, such as MANCOVA, with power that grows intuitively and consistently with the inclusion of data that truly contains the effect being sought. The NPC allows the analysis of cortical gray matter volume from surface-based methods, even when its separate constituent parts have possibly cancelling effects on each other.

A.2 Thesis impact

The most obvious impact of this thesis is that it provided:

- A method to study the surface area of the cortex, of populations of subjects, at a fine resolution.
- A thorough investigation of permutation tests in the presence of nuisance variables.
- Assessment of permutations with sign flippings.
- Whole-block and within-block permutation.
- A heteroscedasticity-robust test statistic.
- A version of NPC that is feasible for neuroimaging applications.
- A demonstration of the superior power of NPC when compared to classical tests.

A.2.1 Peer-reviewed publications

- Winkler AM, Sabuncu MR, Yeo BT, Fischl B, Greve DN, Kochunov P, Nichols TE, Blangero J, Glahn DC. Measuring and comparing brain cortical surface area and other areal quantities. *NeuroImage*. 2012 Mar 15;61(4):1428-43.

– Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. *NeuroImage*. 2014 May 15;92:381-97.

– Winkler AM, Webster MA, Brooks JC, Tracey I, Smith SM and Nichols TE. Non-parametric combination and related permutation tests for neuroimaging. *Human Brain Mapping*. 2016 (*in press*).

A.2.2 Presentations in conferences

– Winkler AM, Sabuncu MR, Yeo BT, Fischl B, Greve DN, Kochunov P, Nichols TE, Blangero J, Glahn DC. *Measuring and comparing brain cortical surface area and other areal quantities.* 18th Human Brain Mapping, 10-14 June 2012, Beijing, China.

– Winkler AM, Smith SM, Nichols TE. *Non-parametric combination for analyses of multi-modal imaging.* 19th Human Brain Mapping, 16-20 June 2013, Seattle, WA, USA; also presented at: Neuroimaging Data Analysis. Statistical and Applied Mathematical Sciences Institute (SAMSI), 04-14 June 2013, Research Triangle Park, North Carolina, USA.

– Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. *Permutation inference for the general linear model and the G-statistic.* 20th Human Brain Mapping, 8-12 June 2014, Hamburg, Germany.

A.2.3 Talks

– *Areal analysis.* Talk at the FMRIB Analysis Group meeting, 13 February 2013, University of Oxford, UK.

– *Permutation for the general linear model.* Talk at at the Department of Statistics, University of Warwick, UK, on 27 March 2014.

– *Permutation for the general linear model.* This was a series of three talks delivered in 3, 10 and 17 September 2015 at FMRIB, University of Oxford, UK. The talks covered permutation for the general linear model, block permutation, multivariate methods (classical MANOVA, etc), non-parametric combin-

ation for brain imaging data (NPC), and multi-level block permutation, therefore including topics beyond those covered in this dissertation.

– *Areal and volumetric analyses.* Talk at the Department of Laboratory Medicine, Children's and Women's Health, 14 April 2015, Norwegian University of Science and Technology, Trondheim, Norway.

A.2.4 Public engagement

Various small pieces of information that were studied during the development of this dissertation have been published in the blog of the author: brainder.org. This includes entries about normality tests, the Box–Cox transformation and log-normality, quality inspection of results of FreeSurfer surface reconstruction, vertexwise and facewise file formats, confidence intervals for Bernoulli trials, biases on permutation p-values, the logic of the method of Fisher to combine p-values, the classic "lady tasting tea" experiment (a form of permutation test), among others. Scripts for areal analyses, for smoothing quickly vertexwise or facewise data after interpolation to a regular mesh, to generate a regular spherical mesh, and for visualisation of areal quantities (facewise) are also provided in the blog.

A.2.5 Software

Scripts for areal interpolation, smoothing, conversion from facewise to vertexwise, and facewise data visualisation, as presented in Chapter 2, have been made available, and can be obtained from brainder.org. The permutation tests, as discussed in Chapters 3 and 4, have been made available in the tool *Permutation Analysis of Linear Models* (PALM), a text-based application that can be invoked from scripts, and that can be downloaded from fsl.fmrib.ox.ac.uk/fsl/fslwiki/PALM (Figure A.1). It should be noted, however, that PALM includes various other features that are not covered by this dissertation. Both PALM and the scripts for areal analysis are licensed under the General Public Licence (GPL), thus can be distributed freely, and run in either MATLAB or Octave.

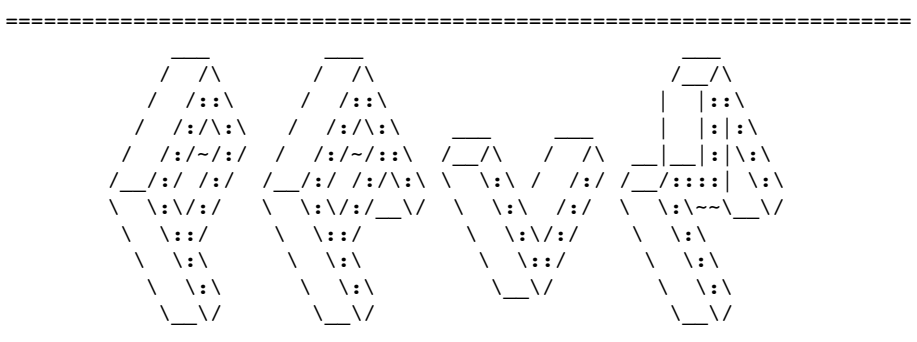


Figure A.1: The permutation tests discussed and proposed in Chapters 3 and 4 have been made available in the tool *Permutation Analysis of Linear Models* (PALM), a text-based application that can be invoked from scripts.

A.3 Further perspectives

The work in this thesis has already unfolded into rich consequences. The whole-block and within-block exchangeability can be nested into each other, so as to allow multi-level exchangeability blocks, a work that the author has already developed as a side-project, and that is already published (Winkler et al., 2015).

While the pycnophylactic interpolation is clearly the most appropriate way of assessing area, it is computationally demanding given the high resolution of the cortical meshes typically used, and an assessment of other, approximate methods, is necessary so as to reduce computational costs. In the same line, permutation tests are much slower than their parametric counterparts. Strategies for acceleration should to be considered if these methods are to be used for routine analysis. These are left as future work. Likewise, the promising use of NPC for joint analysis of cortical surface area and thickness, as a replacement for analysis of cortical volume, is left as future work.

Finally, the demonstration that permutation tests can be used in complex experimental designs as those considered opens up many possibilities, particularly for the analysis of multi-level FMRI data, as well as for cases in which not all observations are present. This includes cases that currently can only be treated with linear mixed effects models, which has all the disadvantages inherent to iterative

methods.

Appendix B

Supporting Information

Chapter 4 is complemented by supporting materials: a browsable set of pages with the results of the various simulations mentioned in that chapter. This material accessible at the URL http://bit.ly/2dOmt3M. The same material is available as the Supporting Information of the paper that includes most of the content of this chapter, Winkler et al. (2016).

This page intentionally left blank.

Bibliography

Abou Elseoud, A., Nissilä, J., Liettu, A., Remes, J., Jokelainen, J., Takala, T., Aunio, A., Starck, T., Nikkinen, J., Koponen, H., Zang, Y. F., Tervonen, O., Timonen, M., and Kiviniemi, V. Altered resting-state activity in seasonal affective disorder. *Human Brain Mapping*, 35(1):161–172, 2014.

Anderson, M., and ter Braak, C. J. F. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, 73(2):85–113, 2003.

Anderson, M. J., and Legendre, P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62(3):271–303, 1999.

Anderson, M. J., and Robinson, J. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2001.

Arndt, S., Cizadlo, T., Andreasen, N. C., Heckel, D., Gold, S., and O'Leary, D. S. Tests for comparing images based on randomization and permutation methods. *Journal of Cerebral Blood Flow and Metabolism*, 16(6):1271–9, 1996.

Ashburner, J., and Friston, K. J. Voxel-based morphometry - the methods. *NeuroImage*, 11:805–21, 2000.

Aspin, A. A., and Welch, B. L. Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika*, 36(3):290–296, 1949.

Augustinack, J. C., van der Kouwe, A. J., Blackwell, M. L., Salat, D. H., Wiggins, C. J., Frosch, M. P., Wiggins, C. C., Potthast, A., Wald, L. L., and Fischl, B. R. Detection of entorhinal layer II using 7 Tesla magnetic resonance imaging. *Annals of Neurology*, 57(4):489–94, 2005.

Auzias, G., Lefevre, J., Le Troter, a., Fischer, C., Perrot, M., Regis, J., and Coulon, O. Model-driven harmonic parameterization of the cortical surface: HIP-HOP. *IEEE Transactions on Medical Imaging*, 32(5):873–87, 2013.

Balaban, I. J. An optimal algorithm for finding segments intersections. *Proceedings of the 11th Annual Symposium on Computational Geometry*, pages 211–219, 1995.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. T. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–83, 1996.

Beaton, A. E. Salvaging experiments: Interpreting least squares in non-random samples. In Hogben, D., and Fife, D., editors, *Computer Science and Statistics: Tenth Annual Symposium of the Interface*, pages 137–45, Gaithersburg, Maryland, 1978. United States Department of Commerce.

Beckmann, C. F., Jenkinson, M., and Smith, S. M. General multi-level linear modelling for group analysis in FMRI. Technical report, University of Oxford, Oxford, 2001.

Beckmann, M., Johansen-Berg, H., and Rushworth, M. F. S. Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. *Journal of Neuroscience*, 29(4):1175–90, 2009.

Belmonte, M., and Yurgelun-Todd, D. Permutation testing made practical for functional magnetic resonance image analysis. *IEEE Transactions on Medical Imaging*, 20(3):243–8, 2001.

Benjamini, Y., and Heller, R. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–22, 2008.

Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Bentley, J. L., and Ottmann, T. A. Algorithms for reporting and counting geometric intersections. *IEEE Transactions on Computers*, C-28(9):643–7, 1979.

Berger, R. L. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982.

Berk, R. H., and Cohen, A. Asymptotically optimal methods of combining tests. *Journal of the American Statistical Association*, 74(368):812–814, 1979.

Bhandary, M., and Zhang, X. Comparison of several tests for combining several independent tests. *Journal of Modern Applied Statistical Methods*, 10(2):436–446, 2011.

Bilgüvar, K., Oztürk, A. K., Louvi, A., Kwan, K. Y., Choi, M., Tatli, B., Yalnizoğlu, D., Tüysüz, B., Cağlayan, A. O., Gökben, S., Kaymakćalan, H., Barak, T., Bakircioğlu, M., Yasuno, K., Ho, W., Sanders, S., Zhu, Y., Yilmaz, S., Dinçer, A., Johnson, M. H., Bronen, R. A., Koçer, N., Per, H., Mane, S., Pamir, M. N., Yalçinkaya, C., S Kumanda s., Topçu, M., Ozmen, M., Sestan, N., Lifton, R. P., State, M. W., and Günel, M. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*, 467(7312):207–10, 2010.

Birnbaum, A. Combining independent tests of significance. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

Blair, R. C., and Karniski, W. An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, 30(5):518–24, 1993.

Blair, R. C., and Karniski, W. Distribution-free statistical analyses of surface and volumetric maps. In Thatcher, R. W., Hallett, M., Zeffiro, T., John, E. R., and Huerta, M., editors, *Functional Neuroimaging: Technical Foundations*, pages 19–28. Academic Press, San Diego, 1994.

Blair, R. C., Higgins, J. J., Karniski, W., and Kromrey, J. D. A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances. *Multivariate Behavioral Research*, 29(2):141–163, 1994.

Bland, J. M., and Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327(8476):307–10, 1986.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. *Introduction to Meta-Analysis*. Wiley, West Sussex, UK, 2009.

Box, G., and Cox, D. An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, 26(2):211–252, 1964.

Box, G. E. P., and Andersen, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society. Series B*, 17(1):1–34, 1955.

Brammer, M. J., Bullmore, E. T., Simmons, A., Williams, S. C., Grasby, P. M., Howard, R. J., Woodruff, P. W., and Rabe-Hesketh, S. Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. *Magnetic Resonance Imaging*, 15(7):763–70, 1997.

Breakspear, M., Brammer, M. J., Bullmore, E. T., Das, P., and Williams, L. M. Spatiotemporal wavelet resampling for functional neuroimaging data. *Human Brain Mapping*, 23(1):1–25, 2004.

Brett, M., Nichols, T., Andersson, J., Wager, T., and Poline, J.-B. When a conjunction is not a conjunction. Poster presented at the X Annual Meeting of the Organization for Human Brain Mapping, Budapest, Hungary, 2004.

Bridge, H., and Clare, S. High-resolution MRI: in vivo histology? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1465):137–46, 2006.

Brombin, C., Midena, E., and Salmaso, L. Robust non-parametric tests for complex-repeated measures problems in ophthalmology. *Statistical Methods in Medical Research*, 22(6):643–660, 2013.

Brooks, J. C. W., Zambreanu, L., Godinez, A., Craig, A. D. B., and Tracey, I. Somatotopic organisation of the human insula to painful heat studied with high resolution functional imaging. *NeuroImage*, 27(1):201–9, 2005.

Brown, B. M., and Maritz, J. S. Distribution-free methods in regression. *Australian Journal of Statistics*, 24(3):318–331, 1982.

Brown, M. B. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31(4):987–992, 1975.

Brown, T. T., and Jernigan, T. L. Brain development during the preschool years. *Neuropsychology Review*, 22(4):313–333, 2012.

Brunner, E., and Munzel, U. The nonparametric Behrens–Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42 (1):17–25, 2000.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35 (2):261–77, 1996.

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T. A., and Brammer, M. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Human Brain Mapping*, 12(2):61–78, 2001.

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

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5):365–76, 2013.

Butts, C. T. Revisiting the foundations of network analysis. *Science*, 325(5939): 414–6, 2009.

Buxhoeveden, D. P., and Casanova, M. F. The minicolumn hypothesis in neuroscience. *Brain*, 125(5):935–51, 2002.

Cade, B. S., and Richards, J. D. Permutation tests for least absolute deviation regression. *Biometrics*, 52(3):886–902, 1996.

Calhoun, V. D., and Sui, J. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, in press.

Caplan, D., and Moo, L. Cognitive conjunction and cognitive functions. *NeuroImage*, 21(2):751–6, 2004.

Chang, L.-C., Lin, H.-M., Sibille, E., and Tseng, G. C. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, 14(vi):368, 2013.

Chazelle, B., and Edelsbrunner, H. An optimal algorithm for intersecting line segments in the plane. *Journal of the ACM*, 39(1):1–54, 1992.

Chazelle, B., Edelsbrunner, H., Guibas, L. J., and Sharir, M. Algorithms for bichromatic line-segment problems and polyhedral terrains. *Algorithmica*, 11 (2):116–32, 1994.

Chen, C.-H., Panizzon, M. S., Eyler, L. T., Jernigan, T. L., Thompson, W., Fennema-Notestine, C., Jak, A. J., Neale, M. C., Franz, C. E., Hamza, S., Lyons, M. J., Grant, M. D., Fischl, B., Seidman, L. J., Tsuang, M. T., Kremen, W. S., and Dale, A. M. Genetic influences on cortical regionalization in the human brain. *Neuron*, 72(4):537–544, 2011.

Chen, C.-H., Gutierrez, E. D., Thompson, W., Panizzon, M. S., Jernigan, T. L., Eyler, L. T., Fennema-Notestine, C., Jak, A. J., Neale, M. C., Franz, C. E., Lyons, M. J., Grant, M. D., Fischl, B., Seidman, L. J., Tsuang, M. T., Kremen, W. S., and Dale, A. M. Hierarchical genetic organization of human cortical surface area. *Science*, 335(6076):1634–6, 2012.

Chen, G., Adleman, N. E., Saad, Z. S., Leibenluft, E., and Cox, R. W. Applications of multivariate modeling to neuroimaging group analysis: a comprehensive alternative to univariate general linear model. *NeuroImage*, 99:571–88, 2014.

Chen, Z. Is the weighted z-test the best method for combining probabilities from independent tests? *Journal of Evolutionary Biology*, 24(4):926–30, 2011.

Chenn, A., and Walsh, C. A. Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science*, 297(5580):365–9, 2002.

Christensen, G. E., Rabbitt, R. D., and Miller, M. I. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5 (10):1435–47, 1996.

Christensen, R. *Advanced Linear Modelling*. Springer, New York, USA, 2 edition, 2001.

Christensen, R. *Plane answers to complex questions: the theory of linear models*. Springer, New York, 3 edition, 2002.

Chung, J. H., and Fraser, D. A. S. Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association*, 53(283):729–735, 1958.

Clark, C. M., Schneider, J. A., Bedell, B. J., Beach, T. G., Bilker, W. B., Mintun, M. A., Pontecorvo, M. J., Hefti, F., Carpenter, A. P., Flitter, M. L., Krautkramer, M. J., Kung, H. F., Coleman, R. E., Doraiswamy, P. M., Fleisher, A. S., Sabbagh, M. N., Sadowsky, C. H., Reiman, E. P., Reiman, P. E. M., Zehntner, S. P., and Skovronsky, D. M. Use of florbetapir-PET for imaging beta-amyloid pathology. *Journal of the American Medical Association*, 305 (3):275–83, 2011.

Clowry, G., Molnár, Z., and Rakic, P. Renewed focus on the developing human neocortex. *Journal of Anatomy*, 217(4):276–88, 2010.

Da Costa, S., van der Zwaag, W., Marques, J. P., Frackowiak, R. S. J., Clarke, S., and Saenz, M. Human primary auditory cortex follows the shape of heschl's gyrus. *Journal of Neuroscience*, 31(40):14067–14075, 2011.

Dale, A. M., and Sereno, M. I. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction. *Journal of Cognitive Neuroscience*, 5(2):162–176, 1993.

Dale, A. M., Fischl, B., and Sereno, M. I. Cortical surface-based analysis I: Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–94, 1999.

Darlington, R. B., and Hayes, A. F. Combining independent p values: extensions of the stouffer and binomial methods. *Psychological Methods*, 5(4):496–515, 2000.

David, F. N. On the P_{λ_n} test for randomness: Remarks, further illustration, and table of P_{λ_n} for given values of $-\log_{10}\lambda_n$. *Biometrika*, 26(1-2):1–11, 1934.

De Boor, C. Bicubic spline interpolation. *Journal of Mathematics and Physics*, 41 (3):212–218, 1962.

Dekker, D., Krackhardt, D., and Snijders, T. Multicollinearity robust QAP for multipleregression. In *1st Annual Conference of the North American Association for Computational Social and Organizational Science (June 22-25), Pittsburg, PA, USA*, 2003.

Dekker, D., Krackhardt, D., and Snijders, T. A. B. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika*, 72(4):563–581, 2007.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31:968–80, 2006.

Dickerson, B. C., Feczko, E., Augustinack, J. C., Pacheco, J., Morris, J. C., Fischl, B., and Buckner, R. L. Differential effects of aging and Alzheimer's

disease on medial temporal lobe cortical thickness and surface area. *Neurobiology of Aging*, 30(3):432–40, 2009.

Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., and Waternaux, C. *Combining information: statistical issues and opportunities for research.* National Academy Press Washington, DC, USA, 1992.

Draper, N. R., and Stoneman, D. M. Testing for the inclusion of variables in linear regression by a randomisation technique. *Technometrics*, 8(4):695–699, 1966.

Drury, H. A., Van Essen, D. C., Anderson, C. H., Lee, C. W., Coogan, T. A., and Lewis, J. W. Computerized mappings of the cerebral cortex: a multiresolution flattening method and a surface-based coordinate system. *Journal of Cognitive Neuroscience*, 8(1):1–28, 1996.

Dudbridge, F., and Koeleman, B. P. C. Rank truncated product of P-values, with application to genomewide association scans. *Genetic Epidemiology*, 25(4):360–6, 2003.

Durazzo, T. C., Tosun, D., Buckley, S., Gazdzinski, S., Mon, A., Fryer, S. L., and Meyerhoff, D. J. Cortical thickness, surface area, and volume of the brain reward system in alcohol dependence: relationships to relapse and extended abstinence. *Alcoholism: Clinical & Experimental Research*, 35(6):1–14, 2011.

Duyn, J. H., van Gelderen, P., Li, T. Q., de Zwart, J. A., Koretsky, A. P., and Fukunaga, M. High-field MRI of brain cortical substructure based on signal phase. *Proceedings of the National Academy of Sciences of U. S. A.*, 104(28):11796–801, 2007.

Dwass, M. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187, 1957.

Eaton, J. W., Bateman, D., Hauberg, S., and Wehbring, R. *GNU Octave: A high-level interactive language for numerical computations.* Samurai Media Limited, 2015. URL `http://www.gnu.org/software/octave`.

Edgington, E. S. Approximate randomization tests. *The Journal of Psychology*, 72 (2):143–149, 1969.

Edgington, E. S. An additive method for combining probability values from independent experiments. *The Journal of Psychology*, 80(2):351–363, 1972.

Edgington, E. S. *Randomization Tests.* Marcel Dekker, New York, 1995.

Efron, B. Computers and the theory of statistics: thinking the unthinkable. *SIAM Review*, 21(4):460–480, 1979.

Efron, B. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104, mar 2004.

Ernst, M. D. Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):676–685, 2004.

Eyler, L. T., Prom-Wormley, E., Panizzon, M. S., Kaup, A. R., Fennema-Notestine, C., Neale, M. C., Jernigan, T. L., Fischl, B., Franz, C. E., Lyons, M. J., Grant, M., Stevens, A., Pacheco, J., Perry, M. E., Schmitt, J. E., Seidman, L. J., Thermenos, H. W., Tsuang, M. T., Chen, C. H., Thompson, W. K., Jak, A., Dale, A. M., and Kremen, W. S. Genetic and environmental contributions to regional cortical surface area in humans: a magnetic resonance imaging twin study. *Cerebral Cortex*, 21(10):2313–21, 2011.

Fischl, B., and Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of U. S. A.*, 97(20):11050–5, 2000.

Fischl, B., Sereno, M. I., and Dale, A. M. Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2): 195–207, 1999a.

Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–84, 1999b.

Fischl, B., Liu, A., and Dale, A. M. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, 20(1):70–80, 2001.

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T., Mohlberg, H., Amunts, K., and Zilles, K. Cortical folding patterns and predicting cytoarchitecture. *Cerebral Cortex*, 18(8):1973–80, 2008.

Fischl, B., Stevens, A. A., Rajendran, N., Yeo, B. T., Greve, D. N., Leemput, K. V., Polimeni, J. R., Kakunoori, S., Buckner, R. L., Pacheco, J., Salat, D. H., Melcher, J., Frosch, M. P., Hyman, B. T., Grant, P. E., Rosen, B. R., van der Kouwe, A. J., Wiggins, G. C., Wald, L. L., and Augustinack, J. C. Predicting the location of entorhinal cortex from MRI. *NeuroImage*, 47(1):8–17, 2009.

Fish, J. L., Dehay, C., Kennedy, H., and Huttner, W. B. Making bigger brains - the evolution of neural-progenitor-cell division. *Journal of Cell Science*, 121: 2783–93, 2008.

Fisher, R. A. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 4 edition, 1932.

Fisher, R. A. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935a.

Fisher, R. A. The fiducial argument in statistical inference. *Annals of Eugenics*, 6 (4):391–398, 1935b.

Flowerdew, R., Green, M., and Kehris, E. Using areal interpolation methods in geographic information systems. *Papers in Regional Science*, 70(3):303–15, 1991.

Fox, P. T., Mintun, M. A., Reiman, E. M., and Raichle, M. E. Enhanced detection of focal brain responses using intersubject averaging and change-distribution analysis of subtracted PET images. *Journal of Cerebral Blood Flow and Metabolism*, 8(5):642–53, 1988.

Freedman, D., and Lane, D. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–8, 1983.

Friston, K. J., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. Comparing functional (PET) images: the assessment of significant change. *Journal of Cerebral Blood Flow and Metabolism*, 11(4):690–9, 1991.

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., and Worsley, K. J. Multisubject fMRI studies and conjunction analyses. *NeuroImage*, 10(4):385–96, 1999.

Friston, K. J., Penny, W. D., and Glaser, D. E. Conjunction revisited. *NeuroImage*, 25(3):661–7, 2005.

Gaddum, J. H. Lognormal distributions. *Nature*, 156(3964):463–6, 1945.

Gail, M. H., Tan, W. Y., and Piantadosi, S. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75(1):57–64, 1988.

Genovese, C. R., Lazar, N. A., and Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4): 870–8, 2002.

Geyer, S., Weiss, M., Reimann, K., Lohmann, G., and Turner, R. Microstructural parcellation of the human cerebral cortex - from brodmann's post-mortem map to in vivo mapping with high-field magnetic resonance imaging. *Frontiers Human Neuroscience*, 5:19, 2011.

Glaunès, J., Vaillant, M., and Miller, M. I. Landmark matching via large deformation diffeomorphisms on the sphere. *Journal of Mathematical Imaging and Vision*, 1-2:179–200, 2004.

Gonzalez, L., and Manly, B. F. J. Analysis of variance by randomization with small data sets. *Environmetrics*, 9(1):53–65, 1998.

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., and Frackowiak, R. S. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14(1 Pt 1):21–36, 2001.

Good, I. J. On the weighted combination of significance tests. *Journal of the Royal Statistical Society. Series B*, 17(2):264–265, 1955.

Good, P. Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, 1(2):243–247, 2002.

Good, P. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, New York, 2005.

Goodchild, M. F., and Lam, N. S.-N. Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing*, 1:297–312, 1980.

Gouraud, H. Continuous shading of curved surfaces. *IEEE Transactions on Computers*, C-20(6):623–629, 1971.

Gregory, I. N., Marti-Henneberg, J., and Tapiador, F. J. Modelling long-term pan-european population change from 1870 to 2000 by using geographical information systems. *Journal of the Royal Statistical Society. Series A*, 173(1):31–50, 2010.

Guibas, L. J., and Seidel, R. Computing convolutions by reciprocal search. *Discrete & Computational Geometry*, 2(1):175–93, 1987.

Guillaume, B., Hua, X., Thompson, P. M., Waldorp, L., and Nichols, T. E. Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. *NeuroImage*, 94:287–302, 2014.

Guttman, I. *Linear Models: An Introduction*. Wiley, New York, 1982.

Hagler, D. J., Saygin, A. P., and Sereno, M. I. Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage*, 33(4):1093–103, 2006.

Hall, P., and Wilson, S. R. Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47(2):757–762, 1991.

Hayasaka, S., and Nichols, T. E. Combining voxel intensity and cluster extent with permutation test framework. *NeuroImage*, 23(1):54–63, 2004.

Hayasaka, S., Phan, K. L., Liberzon, I., Worsley, K. J., and Nichols, T. E. Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*, 22(2):676–87, 2004.

Hayasaka, S., Du, A.-T., Duarte, A., Kornak, J., Jahng, G.-H., Weiner, M. W., and Schuff, N. A non-parametric approach for co-analysis of multi-modal brain imaging data: application to alzheimer's disease. *NeuroImage*, 30(3):768–79, 2006.

Hayter, A. A. J. The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association*, 81(396):1000–1004, 1986.

Hill, J., Inder, T., Neil, J., Dierker, D., Harwell, J., and van Essen, D. Similar patterns of cortical expansion during human development and evolution. *Proceedings of the National Academy of Sciences of U. S. A.*, 107(29):13135–40, 2010.

Hinds, O., Polimeni, J. R., Rajendran, N., Balasubramanian, M., Amunts, K., Zilles, K., Schwartz, E. L., Fischl, B., and Triantafyllou, C. Locating the functional and anatomical boundaries of human primary visual cortex. *NeuroImage*, 46(4):915–22, 2009.

Hinds, O. P., Rajendran, N., Polimeni, J. R., Augustinack, J. C., Wiggins, G., Wald, L. L., Rosas, H. D., Potthast, A., Schwartz, E. L., and Fischl, B. Accurate prediction of V1 location from cortical folds in a surface coordinate system. *NeuroImage*, 39(4):1585–99, 2008.

Hochberg, Y., and Tamhane, A. C. *Multiple Comparison Procedures*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Holmes, A. P., Blair, R. C., Watson, J. D., and Ford, I. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism*, 16(1):7–22, 1996.

Horn, S. D., Horn, R. A., and Duncan, D. B. Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70(350):380–385, 1975.

Hotelling, H. The generalization of Student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.

Hotelling, H. A generalized T test and measure of multivariate dispersion. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, number 042, pages 23–41, Berkeley, 1951. University of California Press.

Hsu, J. C. *Multiple Comparison: Theory and Methods*. Chapman & Hall/CRC, Boca Raton, FL, USA, 1996.

Huh, M. H., and Jhun, M. Random permutation testing in multiple linear regression. *Communications in Statistics – Theory and Methods*, 30(10):2023–2032, 2001.

James, G. The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38(3):324–329, 1951.

Jiang, B., Zhang, X., Zuo, Y., and Kang, G. A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *Journal of Theoretical Biology*, 277(1):67–73, 2011.

Johnson, R. A., and Wichern, D. W. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Sadle River, NJ USA, 6 edition, 2007.

Jones, E. G. Microcolumns in the cerebral cortex. *Proceedings of the National Academy of Sciences of U. S. A.*, 97(10):5019–21, 2000.

Jones, P. W. First- and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review*, 127(9):2204–10, 1999.

Joyner, A. H., Roddey, J. C., Bloss, C. S., Bakken, T. E., Rimol, L. M., Melle, I., Agartz, I., Djurovic, S., Topol, E. J., Schork, N. J., Andreassen, O. A., and Dale, A. M. A common MECP2 haplotype associates with reduced cortical surface area in humans in two independent populations. *Proceedings of the National Academy of Sciences of U. S. A.*, 106(36):15483–8, 2009.

Jung, B. C., Jhun, M., and Song, S. H. A new random permutation test in ANOVA models. *Statistical Papers*, 48(1):47–62, 2006.

Kähler, A. K., Djurovic, S., Rimol, L. M., Brown, A. A., Athanasiu, L., Jönsson, E. G., Hansen, T., Gústafsson, O., Hall, H., Giegling, I., Muglia, P., Cichon, S., Rietschel, M., Pietiläinen, O. P., Peltonen, L., Bramon, E., Collier, D., Clair, D. S., Sigurdsson, E., Petursson, H., Rujescu, D., Melle, I., Werge, T., Steen, V. M., Dale, A. M., Matthews, R. T., Agartz, I., and Andreassen, O. A. Candidate gene analysis of the human natural killer-1 carbohydrate pathway and perineuronal nets in schizophrenia: B3GAT2 is associated with disease risk and cortical surface area. *Biological Psychiatry*, 69(1):90–6, 2011.

Kapteyn, J. C., and van Uven, M. J. *Skew frequency curves in biology and statistics*. Hoitsema Brothers, Groningen, The Netherlands, 1916.

Kempthorne, O. The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967, 1955.

Kennedy, P. E. Randomization tests in econometrics. *Journal of Business & Economic Statistics*, 13(1):85–94, 1995.

Kennedy, P. E., and Cade, B. S. Randomization tests for multiple regression. *Communications in Statistics-Simulation*, 25:923–936, 1996.

Kenner, H. *Geodesic math and how to use it*. University of California Press, Los Angeles, CA, USA, 1976.

Kherad-Pajouh, S., and Renaud, O. An exact permutation method for testing any effect in balanced and unbalanced fixed effect ANOVA. *Computational Statistics & Data Analysis*, 54(7):1881–1893, 2010.

Kim, E. Y., Kim, D. H., Chang, J. H., Yoo, E., Lee, J. W., and Park, H. J. Triple-layer appearance of Brodmann area 4 at thin-section double inversion-recovery MR imaging. *Radiology*, 250(2):515–22, 2009.

Kim, J. S., Singh, V., Lee, J. K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J. M., Kim, S. I., and Evans, A. C. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage*, 27(1):210–21, 2005.

Klein, A., Ghosh, S. S., Avants, B., Yeo, B. T. T., Fischl, B., Ardekani, B., Gee, J. C., Mann, J. J., and Parsey, R. V. Evaluation of volume-based and surface-based brain image registration methods. *NeuroImage*, 51(1):214–20, 2010.

Klunk, W. E., Engler, H., Nordberg, A., Wang, Y., Blomqvist, G., Holt, D. P., Bergström, M., Savitcheva, I., Huang, G.-F., Estrada, S., Ausén, B., Debnath, M. L., Barletta, J., Price, J. C., Sandell, J., Lopresti, B. J., Wall, A., Koivisto, P., Antoni, G., Mathis, C. A., and Långström, B. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Annals of Neurology*, 55(3):306–19, 2004.

Knuth, D. E. *The Art of Computer Programming. Volume 4, Fascicle 2.* Addison-Wesley, 2005.

Koch, A. L. The logarithm in biology. 1. mechanisms generating the log-normal distribution exactly. *Journal of Theoretical Biology*, 12(2):276–90, 1966.

Koch, A. L. The logarithm in biology. II. distributions simulating the log-normal. *Journal of Theoretical Biology*, 23(2):251–68, 1969.

Kochunov, P., Lancaster, J. L., Glahn, D. C., Purdy, D., Laird, A. R., Gao, F., and Fox, P. T. Retrospective motion correction protocol for high-resolution anatomical MRI. *Human Brain Mappping*, 27(12):957–62, 2006.

Kost, J. T., and McDermott, M. P. Combining dependent p-values. *Statistics & Probability Letters*, 60(2):183–190, 2002.

Kuhfeld, W. F. A note on Roy's largest root. *Psychometrika*, 51(3):479–481, 1986.

Laird, A. R., Rogers, B. P., and Meyerand, M. E. Comparison of Fourier and wavelet resampling methods. *Magnetic Resonance in Medicine*, 51(2):418–22, 2004.

Lancaster, H. O. The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, 3(1):20–33, 1961.

Lauchner, J. H., Buckminster Fuller, R., Clinton, J. D., Mabee, M. B., Moeller, R. M., and Flood, R. Structural design concepts for future space missions. NASA contract NGR-14-008-002. Technical report, Southern Illinois University, 1969.

Lauritzen, P. H., and Nair, R. D. Monotone and conservative cascade remapping between spherical grids (CaRS): Regular latitude-longitude and cubed-sphere grids. *Monthly Weather Review*, 136(4):1416–32, 2008.

Lawley, D. N. A generalization of Fisher's z test. *Biometrika*, 30(1):180–187, 1938.

Lazar, N. A., Luna, B., Sweeney, J. A., and Eddy, W. F. Combining brains: a survey of methods for statistical pooling of information. *NeuroImage*, 16(2): 538–50, June 2002.

Lehmann, E., and Stein, C. On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics*, 20(1):28–45, 1949.

Lehmann, E. L., and Romano, J. P. *Testing Statistical Hypotheses*. Springer, New York, NY, USA, 3 edition, 2005.

Levin, B., and Robbins, H. Urn models for regression analysis, with applications to employment discrimination studies. *Law and Contemporary Problems*, 46(4): 247–267, 1983.

Licata, S. C., Nickerson, L. D., Lowen, S. B., Trksak, G. H., MacLean, R. R., and Lukas, S. E. The hypnotic zolpidem increases the synchrony of BOLD signal fluctuations in widespread brain networks during a resting paradigm. *NeuroImage*, 70:211–222, 2013.

Limpert, E., Stahel, W. a., and Abbt, M. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352, 2001.

Lipták, T. On the combination of independent tests. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közlémenyei*, 3:171–197, 1958.

Locascio, J. J., Jennings, P. J., Moore, C. I., and Corkin, S. Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping*, 5(3):168–93, 1997.

Lombardi, M. Interpolation and smoothing. *Astronomy & Astrophysics*, 395(2): 733–45, 2002.

Loughin, T. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, 47(3):467–485, 2004.

Ludbrook, J., and Dudley, H. Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, 52(2):127–132, 1998.

Lyttelton, O. C., Karama, S., Ad-Dab'bagh, Y., Zatorre, R. J., Carbonell, F., Worsley, K., and Evans, A. C. Positional and surface area asymmetry of the human cerebral cortex. *NeuroImage*, 46(4):895–903, 2009.

MacKinnon, J. G., and White, H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985.

Mangin, J.-F., Frouin, V., Bloch, I., Regis, J., and Lopes-Krabe, J. From 3D MR images to structural representations of the cortex topography using topology preserving deformations. *Journal of Mathematical Imaging and Vision*, 5:297–318, 1995.

Manly, B. F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London, 3rd edition, 2007.

Marcus, R., Peritz, E., and Gabriel, K. R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655, 1976.

Markoff, J., and Shapiro, G. The linkage of data describing overlapping geographical units. *Historical Methods Newsletter*, 7(1):34–46, 1973.

Marroquin, J. L., Biscay, R. J., Ruiz-Correa, S., Alba, A., Ramirez, R., and Armony, J. L. Morphology-based hypothesis testing in discrete random fields: a non-parametric method to address the multiple-comparison problem in neuroimaging. *NeuroImage*, 56(4):1954–67, 2011.

McAlister, D. The law of the geometric mean. *Proceedings of the Royal Society of London*, 29:367–76, 1879.

Meier, U. A note on the power of Fisher's least significant difference procedure. *Pharmaceutical Statistics*, 5(4):253–263, 2006.

Miller, M., Banerjeea, A., Christensen, G., Joshi, S., Khaneja, N., Grenander, U., and Matejic, L. Statistical methods in computational anatomy. *Statistical methods in Medical Research*, 6(3):267–99, 1997.

Mountcastle, V. P. *Perceptual Neuroscience: The Cerebral Cortex*. Harvard University Press, Cambridge, MA, USA, 1998.

Mudholkar, G. S., and George, E. O. The logit statistic for combining probabilities. In Rustagi, J., editor, *Symposium on Optimizing Methods in Statistics*, pages 345–366. Academic Press, New York, 1979.

Nelson, S. M., Cohen, A. L., Power, J. D., Wig, G. S., Miezin, F. M., Wheeler, M. E., Velanova, K., Donaldson, D. I., Phillips, J. S., Schlaggar, B. L., and Petersen, S. E. A parcellation scheme for human left lateral parietal cortex. *Neuron*, 67(1):156–70, 2010.

Nichols, T. Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, 62(2):811–815, 2012.

Nichols, T., and Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12 (5):419–46, 2003.

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.-B. Valid conjunction inference with the minimum statistic. *NeuroImage*, 25(3):653–60, 2005.

Nichols, T. E., and Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002.

Nichols, T. E., Ridgway, G. R., Webster, M. G., and Smith, S. M. GLM permutation: Nonparametric inference for arbitrary general linear models. *NeuroImage*, 41(S1):S72, 2008.

Nopoulos, P. C., Aylward, E. H., Ross, C. A., Johnson, H. J., Magnotta, V. A., Juhl, A. R., Pierson, R. K., Mills, J., Langbehn, D. R., Paulsen, J. S., and PREDICT-HD Investigators Coordinators of Huntington Study Group (HSG). Cerebral cortex structure in prodromal Huntington disease. *Neurobiology of Disease*, 40(3):544–54, 2010.

O'Gorman, T. W. The performance of randomization tests that use permutations of independent variables. *Communications in Statistics – Simulation and Computation*, 34(4):895–908, 2005.

Oja, H. On permutation tests in multiple regression and analysis of covariance problems. *Australian Journal of Statistics*, 29(1):91–100, 1987.

Oosterhoff, J. *Combination of one-sided statistical tests*. Mathematisch Centrum, Amsterdam, The Netherlands, 1969.

Owen, A. B. Karl Pearson's meta-analysis revisited. *The Annals of Statistics*, 37 (6B):3867–3892, 2009.

Palaniyappan, L., Mallikarjun, P., Joseph, V., White, T. P., and Liddle, P. F. Regional contraction of brain surface area involves three large-scale networks in schizophrenia. *Schizophrenia Research*, 129(2-3):163–8, 2011.

Panizzon, M. S., Fennema-Notestine, C., Eyler, L. T., Jernigan, T. L., Prom-Wormley, E., Neale, M., Jacobson, K., Lyons, M. J., Grant, M. D., Franz, C. E., Xian, H., Tsuang, M., Fischl, B., Seidman, L., Dale, A., and Kremen, W. S. Distinct genetic influences on cortical surface area and cortical thickness. *Cerebral Cortex*, 19(11):2728–35, 2009.

Pantazis, D., Nichols, T. E., Baillet, S., and Leahy, R. M. A comparison of random field theory and permutation methods for the statistical analysis of MEG data. *NeuroImage*, 25(2):383–94, 2005.

Pearson, E. S. Some aspects of the problem of randomization. *Biometrika*, 29(1/2): 53–64, 1937.

Pearson, K. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, 25(3/4):379–410, 1933.

Pearson, K. On a new method of determining "goodness of fit". *Biometrika*, 26(4): 425–442, 1934.

Peirce, C. S., and Jastrow, J. On small differences of sensation. *Memoirs of the National Academy of Sciences*, 3:75–83, 1884.

Pesarin, F. On a nonparametric combination method for dependent permutation tests with applications. *Psychotherapy and Psychosomatics*, 54(2–3):172–179, 1990.

Pesarin, F. A resampling procedure for nonparametric combination of several dependent tests. *Journal of the Italian Statistical Society*, 1(1):87–101, 1992.

Pesarin, F. A new solution for the generalized Behrens–Fisher problem. *Statistica*, 55(2):131–146, 1995.

Pesarin, F. *Multivariate Permutation Tests: With Applications in Biostatistics*. John Wiley and Sons, West Sussex, England, UK, 2001.

Pesarin, F., and Salmaso, L. *Permutation Tests for Complex Data: Theory, Applications and Software*. John Wiley and Sons, West Sussex, England, UK, 2010a.

Pesarin, F., and Salmaso, L. Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparametric Statistics*, 22(5):669–684, 2010b.

Petrovic, P., Kalso, E., Petersson, K. M., and Ingvar, M. Placebo and opioid analgesia– imaging a shared neuronal network. *Science*, 295(5560):1737–1740, 2002.

Phipson, B., and Smyth, G. K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1):Article39, 2010.

Pierani, A., and Wassef, M. Cerebral cortex development: from progenitors to patterning to neocortical size during evolution. *Development, Growth & Differentiation*, 51:325–42, 2009.

Pillai, K. C. S. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, 26(1):117–121, 1955.

Pitman, E. J. G. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937a.

Pitman, E. J. G. Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4(2):225–232, 1937b.

Pitman, E. J. G. Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, 29(3/4):322–335, 1938.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical recipes in C*. Cambridge University Press, Cambridge, UK, 1992.

Price, C. J., and Friston, K. J. Cognitive conjunction: a new approach to brain activation experiments. *NeuroImage*, 5(4 Pt 1):261–70, 1997.

Rakic, P. Specification of cerebral cortical areas. *Science*, 241(4862):170–6, 1988.

Rakic, P. Evolution of the neocortex: a perspective from developmental biology. *Nature Reviews Neuroscience*, 10(10):724–35, 2009.

Rakic, P., Ayoub, A. E., Breunig, J. J., and Dominguez, M. H. Decision by division: making cortical maps. *Trends in Neurosciences*, 32(5):291–301, 2009.

Reynolds, D. V. Surgery in the rat during electrical analgesia induced by focal brain stimulation. *Science*, 164(878):444–445, 1969.

Ridgway, G. R. *Statistical analysis for longitudinal MR imaging of dementia*. PhD Thesis, University College London, 2009.

Rimol, L. M., Agartz, I., Djurovic, S., Brown, A. A., Roddey, J. C., Kähler, A. K., Mattingsdal, M., Athanasiu, L., Joyner, A. H., Schork, N. J., Halgren, E., Sundet, K., Melle, I., Dale, A. M., Andreassen, O. A., and Alzheimer's Disease Neuroimaging Initiative. Sex-dependent association of common variants of microcephaly genes with brain structure. *Proceedings of the National Academy of Sciences of U. S. A.*, 107(1):384–8, 2010a.

Rimol, L. M., Panizzon, M. S., Fennema-Notestine, C., Eyler, L. T., Fischl, B., Franz, C. E., Hagler, D. J., Lyons, M. J., Neale, M. C., Pacheco, J., Perry, M. E., Schmitt, J. E., Grant, M. D., Seidman, L. J., Thermenos, H. W., Tsuang, M. T., Eisen, S. A., Kremen, W. S., and Dale, A. M. Cortical thickness is influenced by regionally specific genetic factors. *Biological Psychiatry*, 67(5):493–9, 2010b.

Rimol, L. M., Nesvå g, R., Hagler, D. J., Bergmann, O., Fennema-Notestine, C., Hartberg, C. B., Haukvik, U. K., Lange, E., Pung, C. J., Server, A., Melle, I., Andreassen, O. A., Agartz, I., and Dale, A. M. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biological Psychiatry*, 71(6): 552–60, 2012.

Robinson, E. C., Jbabdi, S., Glasser, M. F., Andersson, J., Burgess, G. C., Harms, M. P., Smith, S. M., Van Essen, D. C., and Jenkinson, M. MSM: A new flexible framework for multimodal surface matching. *NeuroImage*, 100:414–26, 2014.

Roland, P. E., and Zilles, K. Structural divisions and functional fields in the human cerebral cortex. *Brain Research Reviews*, 26(2-3):87–105, 1998.

Rorden, C., Bonilha, L., and Nichols, T. E. Rank-order versus mean based statistics for neuroimaging. *NeuroImage*, 35(4):1531–7, 2007.

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin*, 85(1):185–193, 1978.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., and Wager, T. D. Representation of aversive prediction errors in the human periaqueductal gray. *Nature Neuroscience*, 17(11):1607–1612, 2014.

Roy, S. N. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2):220–238, June 1953.

Royston, P. A toolkit for testing for non-normality in complete and censored samples. *The Statistician*, 42(1):37, 1993.

Rubinov, M., and Sporns, O. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–69, 2010.

Saad, Z. S., Reynolds, R. C., Argall, B., Japee, S., and Cox, R. W. SUMA: An interface for surface-based intra- and inter-subject analysis with AFNI. *IEEE International Symposium on Biomedical Imaging*, pages 1510–1513, 2004.

Sabuncu, M. R., Singer, B. D., Conroy, B., Bryan, R. E., Ramadge, P. J., and Haxby, J. V. Function-based intersubject alignment of human cortical anatomy. *Cerebral Cortex*, 20(1):130–40, 2010.

Salimi-Khorshidi, G., Smith, S. M., and Nichols, T. E. Adjusting the effect of nonstationarity in cluster-based and TFCE inference. *NeuroImage*, 54(3):2006–2019, 2011.

Sanabria-Diaz, G., Melie-García, L., Iturria-Medina, Y., Alemán-Gómez, Y., Hernández-González, G., Valdés-Urrutia, L., Galán, L., and Valdés-Sosa, P. Surface area and cortical thickness descriptors reveal different attributes of the structural human brain networks. *NeuroImage*, 50(4):1497–510, 2010.

Scheffé, H. Statistical inference in the non-parametric case. *The Annals of Mathematical Statistics*, 14(4):305–332, 1943.

Scheffé, H. *The Analysis of Variance*. John Wiley and Sons, New York, 1959.

Schmitt, J. E., Lenroot, R. K., Wallace, G. L., Ordaz, S., Taylor, K. N., Kabani, N. J., Greenstein, D., Lerch, J. P., Kendler, K. S., Neale, M. C., and Giedd, J. N. Identification of genetically mediated cortical networks: a multivariate study of pediatric twins and siblings. *Cerebral Cortex*, 18(8):1737–47, 2008.

Schormann, T., and Zilles, K. Three-dimensional linear and nonlinear transformations: an integration of light microscopical and MRI data. *Human Brain Mapping*, 6(5-6):339–47, 1998.

Schwarzkopf, D. S., Song, C., and Rees, G. The surface area of human V1 predicts the subjective experience of object size. *Nature Neuroscience*, 14(1):28–30, 2011.

Searle, S. R. *Linear Models.* John Wiley and Sons, New York, 1971.

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., and Fischl, B. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3):1060–75, 2004.

Ségonne, F., Pacheco, J., and Fischl, B. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Transactions on Medical Imaging*, 26(4):518–29, 2007.

Sen, P. K. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968.

Shaffer, J. P. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831, 1986.

Shapiro, S. S., and Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

Shepard, D. A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 23rd National Conference of the ACM*, pages 517–524, 1968.

Sibson, R. A brief description of natural neighbour interpolation. In Barnett, V., editor, *Interpreting multivariate data*, pages 21–36, New York, 1981. John Wiley & Sons.

Smith, S., Jenkinson, M., Beckmann, C., Miller, K., and Woolrich, M. Meaningful design and contrast estimability in FMRI. *NeuroImage*, 34(1):127–36, 2007.

Smith, S. M., and Nichols, T. E. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98, 2009.

Snyder, J. P. *Map Projections: A Working Manual - U.S. Geological Survey Professional Paper 1395.* United States Government Printing Office, Washington, DC, USA, 1987.

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., Corneveaux, J. J., Dechairo, B. M., Potkin, S. G., Weiner, M. W., and Thompson, P. M. Voxelwise genome-wide association study (vGWAS). *NeuroImage*, 53(3):1160–1174, 2010.

Still, A. W., and White, A. P. The approximate randomization test as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology*, 34(2):243–252, 1981.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Jr., R. M. W. *The American Soldier: Adjustment During Army Life (Volume 1)*. Princeton University Press, Princeton, New Jersey, USA, 1949.

Suckling, J., and Bullmore, E. Permutation tests for factorially designed neuroimaging experiments. *Human Brain Mapping*, 22(3):193–205, 2004.

Sun, D., Phillips, L., Velakoulis, D., Yung, A., McGorry, P. D., Wood, S. J., van Erp, T. G. M., Thompson, P. M., Toga, A. W., Cannon, T. D., and Pantelis, C. Progressive brain structural changes mapped as psychosis develops in 'at risk' individuals. *Schizophrenia Research*, 108(1-3):85–92, 2009a.

Sun, D., Stuart, G. W., Jenkinson, M., Wood, S. J., McGorry, P. D., Velakoulis, D., van Erp, T. G. M., Thompson, P. M., Toga, a. W., Smith, D. J., Cannon, T. D., and Pantelis, C. Brain surface contraction mapped in first-episode schizophrenia: a longitudinal magnetic resonance imaging study. *Molecular Psychiatry*, 14(10):976–86, 2009b.

Taylor, J., and Tibshirani, R. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*, 7(2):167–81, 2006.

ter Braak, C. J. F. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In Jöckel, K.-H., Rothe, G., and Sendler, W., editors, *Bootstrapping and related techniques*, number 1989, pages 79–86. Springer-Verlag, Berlin, 1992.

The MathWorks Inc. *MATLAB Version 8.5 (R2015a)*. Natick, Massachusetts, USA, 2015.

Theil, H. A rank-invariant method for linear and polynomial regression. I. II. III. *Proceedings of the Section of Sciences, Koninklijke Akademie van Wetenschappen te Amsterdam*, 53:386–392, 521–525, 1397–1412, 1950.

Thirion, J. P. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243–60, 1998.

Thomas, A. G., Dennis, A., Rawlings, N. B., Stagg, C. J., Matthews, L., Morris, M., Kolind, S. H., Foxley, S., Jenkinson, M., Nichols, T. E., Dawes, H., Bandettini, P. A., and Johansen-Berg, H. Multi-modal characterization of rapid anterior hippocampal volume increase associated with aerobic exercise. *NeuroImage*, in press. doi: 10.1016/j.neuroimage.2015.10.090.

Thompson, P. M., and Toga, A. W. Anatomically driven strategies for high-dimensional brain image warping and pathology detection. In Toga, A. W., editor, *Brain Warping*, pages 311–36. Academic Press, 1999.

Thompson, P. M., and Toga, A. W. A framework for computational anatomy. *Computing and Visualization in Science*, 5:13–34, 2002.

Timm, N. H. *Applied multivariate analysis.* Springer, New York, 2002.

Tippett, L. H. C. *The methods of statistics.* Williams and Northgate, London, 1931.

Tobler, W. R. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–30, 1979.

Tracey, I., Ploghaus, A., Gati, J. S., Clare, S., Smith, S., Menon, R. S., and Matthews, P. M. Imaging attentional modulation of pain in the periaqueductal gray in humans. *The Journal of Neuroscience*, 22(7):2748–2752, 2002.

Trotter, H. F., and Tukey, J. W. Conditional Monte Carlo techniques in a complex problem about normal samples. In Meyer, H. A., editor, *Symposium on Monte Carlo methods*, pages 64–79, New York, 1956. Wiley.

Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114, 1949.

Ullrich, P. A., Lauritzen, P. H., and Jablonowski, C. Geometrically exact conservative remapping (GECoRe): Regular latitude-longitude and cubed-sphere grids. *Monthly Weather Reviews*, 137(6):1721–41, 2009.

Uludağ, K., and Roebroeck, A. General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage*, 102:3–10, 2014.

van Essen, D. C. A Population-Average, Landmark- and Surface-based (PALS) atlas of human cerebral cortex. *NeuroImage*, 28(3):635–62, 2005.

van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., and Anderson, C. H. An integrated software suite for surface-based analyses of cerebral cortex. *Journal of American Medical Informatics Association*, 8(5):443–59, 2001.

van Zwet, W., and Oosterhoff, J. On the combination of independent test statistics. *The Annals of Mathematical Statistics*, 38(3):659–680, 1967.

Vercauteren, T., Pennec, X., Perchant, A., and Ayache, N. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage*, 45(1 Suppl): S61–72, 2009.

Vince, J. *Mathematics for Computer Graphics.* Springer, London, UK, 2005.

Voets, N. L., Hough, M. G., Douaud, G., Matthews, P. M., James, A., Winmill, L., Webster, P., and Smith, S. Evidence for abnormalities of cortical development in adolescent-onset schizophrenia. *NeuroImage*, 43(4):665–75, 2008.

Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

Welch, B. L. On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3):330–336, 1951.

Welch, W. J. Construction of permutation tests. *Journal of the American Statistical Association*, 85(411):693–698, 1990.

Westberg, M. Combining independent statistical tests. *The Statistician*, 34(3): 287–296, 1985.

Westfall, P. H., and Troendle, J. F. Multiple testing with minimal assumptions. *Biometrical Journal*, 50(5):745–55, 2008.

Westfall, P. H., and Young, S. S. *Resampling-Based Multiple Testing: Examples And Methods for p-Value Adjustment*. John Wiley and Sons, New York, 1993.

Whitlock, M. C. Combining probability from independent tests: the weighted z-method is superior to Fisher's approach. *Journal of Evolutionary biology*, 18 (5):1368–73, 2005.

Wilkinson, B. A statistical consideration in psychological research. *Psychological Bulletin*, 48(3):156–8, 1951.

Wilks, S. S. Certain generalizations in the analysis of variance. *Biometrika*, 24(3): 471–494, 1932.

Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Winer, B. J. *Statistical Principles in Experimental Design*. McGraw-Hill, New York, 1962.

Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., Duggirala, R., and Glahn, D. C. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage*, 15(3):1135–46, 2010.

Winkler, A. M., Sabuncu, M. R., Yeo, B. T. T., Fischl, B., Greve, D. N., Kochunov, P., Nichols, T. E., Blangero, J., and Glahn, D. C. Measuring and comparing brain cortical surface area and other areal quantities. *NeuroImage*, 61(4):1428–43, 2012.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. Permutation inference for the general linear model. *NeuroImage*, 92:381–97, 2014.

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., and Smith, S. M. Multi-level block permutation. *NeuroImage*, 123:253–68, 2015.

Winkler, A. M., Webster, M. A., Brooks, J. C., Tracey, I., Nichols, T. E., and Smith, S. M. Non-parametric combination and related permutation tests for neuroimaging. *Human Brain Mapping*, 37(4):1486–511, 2016.

Wisco, J. J., Kuperberg, G., Manoach, D., Quinn, B. T., Busa, E., Fischl, B., Heckers, S., and Sorensen, A. G. Abnormal cortical folding patterns within Broca's area in schizophrenia: evidence from structural MRI. *Schizophrenia Research*, 94 (1-3):317–27, 2007.

Won, S., Morris, N., Lu, Q., and Elston, R. C. Choosing an optimal method to combine p-values. *Statistics in Medicine*, 28(11):1537–53, 2009.

Woo, C.-W., Krishnan, A., and Wager, T. D. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, 91:412–9, 2014.

Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D., and Evans, A. C. Detecting changes in nonisotropic images. *Human Brain Mapping*, 8(2-3):98–101, 1999.

Wu, S. S. Combining univariate tests for multivariate location problem. *Communications in Statistics: Theory and Methods*, 35(8):1483–1494, 2006.

Yekutieli, D., and Benjamini, Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2):171–196, 1999.

Yeo, B. T., Sabuncu, M. R., Vercauteren, T., Ayache, N., Fischl, B., and Golland, P. Spherical demons: fast diffeomorphic landmark-free surface registration. *IEEE Transactions on Medical Imaging*, 29(3):650–68, 2010a.

Yeo, B. T. T., Sabuncu, M. R., Vercauteren, T., Holt, D. J., Amunts, K., Zilles, K., Golland, P., and Fischl, B. Learning task-optimal registration cost functions for localizing cytoarchitecture and function in the cerebral cortex. *IEEE Transactions on Medical Imaging*, 29(7):1424–41, 2010b.

Yiu, P. The uses of homogeneous barycentric coordinates in plane Euclidean geometry. *International Journal of Mathematical Education in Science and Technology*, 31(4):569–578, 2000.

Zaykin, D. V. Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, 24(8):1836–1841, 2011.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–85, 2002.

Zhu, D., Zhang, T., Jiang, X., Hu, X., Chen, H., Yang, N., Lv, J., Han, J., Guo, L., and Liu, T. Fusing DTI and fMRI data: A survey of methods and applications. *NeuroImage*, 102:184–191, 2014.

Zilles, K., and Amunts, K. Centenary of Brodmann's map-conception and fate. *Nature Reviews Neuroscience*, 11(2):139–45, 2010.

This page intentionally left blank.

Curriculum Vitæ

Anderson M. Winkler studied Electronics for two years before joining the Medical School at the Universidade Federal do Paraná, Curitiba, Brazil, from which he graduated in early 2005. He worked as a physician at the Centro Municipal de Urgências Médicas Boa Vista at the same time in which he studied for a Masters in Biomedical Engineering at the Universidade Tecnológica Federal do Paraná, Curitiba, Brazil. Upon completion, he worked for one year as Postdoctoral Associate at the Department of Psychiatry of University of Texas Health Science at San Antonio, Texas, United States, then for three years at the Department of Psychiatry of the Yale University School of Medicine. During this period he worked with brain phenotypes for imaging genetics. In 2011, he joined the Marie-Curie Initial Training Network "Neurophysics" doctoral program which, with this thesis, has now reached completion.

This page intentionally left blank.