A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Widening the applicability of permutation inference

Anderson M. Winkler
St. Edmund Hall

FMRIB Analysis Group
Nuffield Department of Clinical Neurosciences
University of Oxford

Trinity 2016

This page intentionally left blank.

Abstract

This thesis is divided into three main parts. In the FIRST, we discuss that, although permutation tests can provide exact control of false positives under the reasonable assumption of exchangeability, there are common examples in which global exchangeability does not hold, such as in experiments with repeated measurements or tests in which subjects are related to each other. To allow permutation inference in such cases, we propose an extension of the well known concept of exchangeability blocks, allowing these to be nested in a hierarchical, multi-level definition. This definition allows permutations that retain the original joint distribution unaltered, thus preserving exchangeability. The null hypothesis is tested using only a subset of all otherwise possible permutations. We do not need to explicitly model the degree of dependence between observations; rather the use of such permutation scheme leaves any dependence intact. The strategy is compatible with heteroscedasticity and can be used with permutations, sign flippings, or both combined. In the SECOND part, we exploit properties of test statistics to obtain accelerations irrespective of generic software or hardware improvements. We compare six different approaches using synthetic and real data, assessing the methods in terms of their error rates, power, agreement with a reference result, and the risk of taking a different decision regarding the rejection of the null hypotheses (known as the resampling risk). In the THIRD part, we investigate and compare the different methods for assessment of cortical volume and area from magnetic resonance images using surface-based methods. Using data from young adults born with very low birth weight and coetaneous controls, we show that instead of volume, the permutation-based non-parametric combination (NPC) of thickness and area is a more sensitive option for studying joint effects on these two quantities, giving equal weight to variation in both, and allowing a better characterisation of biological processes that can affect brain morphology.

This page intentionally left blank.

Acknowledgements

À minha família.

To my friends, for their infinite patience and support.

To my doctoral advisors, Prof. Stephen M. Smith and Prof. Thomas E. Nichols, endless sources of knowledge and inspiration.

I am very much indebted to the National Research Council of Brazil (Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq), without which this work would have been impossible.

A strong, prolific, and enriching collaboration have helped and improved all the chapters, and I am very much thankful for these fruitful interactions. Chapter 2 received helpful feedback from Matthew A. Webster, Diego Vidaurre, and Gergő Nemes. Likewise, Chapter 3 received much welcome input from Gerard R. Ridgway and Gwenaëlle Douaud. Chapter 4 would have been impossible without the close collaboration with Lars M. Rimol, Jon Skranes, Knut Jørgen Bjuland, Douglas N. Greve, Asta K. Håberg, and Mert R. Sabuncu.

My time in Oxford was made even better thanks to the people in the colleges to which I was at one time or another affiliated: St. Cross and St. Edmund Hall. Under the risk unfair omissions, some I would like to thank nominally Prof. Keith Gull, Dr. Richard Willden. Prof. Heidi Johansen–Berg, Prof. Paul M. Matthews, Dr. Charlotte Stagg, Sir Mark Jones, Prof. Petros Ligoxygakis, and Dr. Katie Warnaby, all of whom, through their remarkable work and friendliness, have made these years extremely enjoyable.

At FMRIB, the list is long and, again, in the hope of not unfairly omitting anyone, I am much thankful to the members of the Analysis Group listed here in no particular order: Moisés Hernández–Fernández, Fidel Alfaro–Almagro, Samuel J. Harrison, Jonathan Hadida, Giles L. Colclough, Zobair Arya, Emmanuel Vallee, Riham Satti, Gwenaëlle Douaud, Eugene P. Duff, Gerard R. Ridgway, Diego Vidaurre, Matteo Bastiani, Evangelos Roussos, Michiel Cottaar, Eelke Visser, Stamatios Sotiropoulos, Mark Jenkinson, Jesper L. Andersson, Saad Jbabdi, Mark Woolrich, Christian F. Beckmann, Janine Bijsterbosch, Oiwi Parker Jones, Matthew A. Webster, Sean Fitzgibbon, Paul McCarthy, Ludovica Griffanti, Emma C. Robinson, Jelena Božek Mouthuy, David Flitney, Duncan Mortimer, Thomas E. Nichols and Stephen M. Smith. Additionally, I would like to thank Susan Field and Marilyn Goulding for being tirelessly solicitous since even before the time I arrived, and Marion Greenleaves for all her support.

It would have been far more difficult to do this work without the use of free software. I very much thank every person who contributes, directly or indirectly, to the development of these tools.

Finally, I would like to thank the Sir Edward Penley Abraham Fund and St. Edmund Hall for the honour of being twice the recipient of the "Cephalosporin Award", providing the much needed further support during this period, and the Nuffield Department of Clinical Neurosciences and St. Edmund Hall for the funding that facilitated attending and presenting work at conferences.

Contents

List of Figures

List of Tables

Abbreviations

2-D/3-D	Two-dimensional/tri-dimensional.
AFNI	Analysis of Functional Neuroimages.
ASL	Arterial spin labelling.
BMI	Body mass index.
CCA	Canonical correlation analysis.
CMV	Classical multivariate test.
CTP	Closed testing procedure.
DOI	Digital object identifier.
DZ	Dizygotic twin.
EB	Exchangeability block.
EE	Exchangeable errors.
FDR	False discovery rate.
FILM	FMRIB's Improved Linear Model
FLAME	FMRIB's Local Analysis of Mixed Effects.
FMRIB	Oxford Centre for Functional MRI of the Brain.
FMRI	Functional magnetic resonance imaging.

FSL	FMRIB Software Library.
FS	FreeSurfer.
FWER	Familywise error rate.
FWHM	Full-width at half-maximum.
GEV	Generalised extreme value distribution.
GLM	General linear model.
GPD	Generalised Pareto distribution.
GPL	General Public License.
GPU	Graphics processing unit.
HCP	Human Connectome Project.
IC3/IC5/IC7	Icosahedron recursively subdivided 3, 5 or 7 times.
ISE	Independent and symmetric errors.
IUT	Intersection-union test.
JNH	Joint null hypothesis.
MANOVA/MANCOVA	Multivariate analysis of variance/covariance.
MNI	Montreal Neurological Institute.
MPRAGE	Magnetization-prepared rapid gradient-echo.
MRI	Magnetic resonance imaging.
MSM	Multi-modal surface matching.
MZ	Monozygotic twin.
NPC	Non-parametric combination.

NS	Non-sibling.
ORA	Oxford Research Archive.
PALM	Permutation Analysis of Linear Models.
PET	Positron-emission tomography.
RFT	Random field theory.
ROI	Region of interest.
SD	Spherical Demons.
SE	Standard error.
SOLAR	Sequential Oligogenic Linkage Analysis Routines.
SVD	Singular value decomposition.
TFCE	Threshold-free cluster enhancement.
UIT	Union-intersection test.
URL	Uniform resource locator.
VBM	Voxel-based morphometry.
VG	Variance group.
VLBW	Very-low birth weight.
WM	White matter.

This page intentionally left blank.

Chapter 1

Introduction

1.1 The rise of permutation tests

Provided with the task of investigating an association between two variables from a number of samples, how to quantify the likelihood that a measurement of such association is not merely due to chance? If no true association exists, any pairing of values is incidental, and for a given measurement of one variable, the other could have taken any value. We could thus replace one variable by a set of random values, recompute the association, and keep repeating this procedure multiple times to have a sense of how the association varies when there is no actual effect. *Any* value, as suggested above, however, is not realistic: possible values and their frequencies lie within certain sets or intervals, and occur with frequencies that are not always known. The observed data, however, gives an indication of the possible values and their frequencies. Instead of replacing one variable by sets of random values, the observations for that variable can be *permuted* randomly, the association quantified and recorded, and the process repeated many times. How often a stronger association than the one observed without any permutation, is the answer being sought.

Permutation tests, in their essence, do exactly as above, and despite this simplicity, they entail an obvious practical difficulty: the need to repeat the shuff-

ling many times. Before computers became available, recalculating a measurement of association between variables thousands of times was beyond the bounds of feasibility, rendering them "little more than curiosities" (Bradley, 1968). The fact that Fisher, their most prominent early proponent (Fisher, 1935) became silent about these tests in his later years (Basu, 1980) certainly did not help much to increase their popularity. In practice, only the simplest cases could be performed, although some shortcuts, mostly based on ranks, were eventually developed (Wilcoxon, 1945; Mann and Whitney, 1947; Box and Andersen, 1955). With the possibility of automated execution using computers (Efron, 1979), the landscape changed completely, and these otherwise extremely laborious tests became feasible.

Although the idea of repeating an experiment many times while randomising experimental conditions can be traced to the 19th century (Peirce and Jastrow, 1884), it was not until the decade of 1930 that strategies started to be studied in depth (Fisher, 1935; Pitman, 1937a,b, 1938), notably about the then astonishing idea of shuffling the data that had already been obtained, as opposed to repeating the experiment multiple times. Theoretical and practical advances have been ongoing ever since (Pearson, 1937; Scheffé, 1943; Lehmann and Stein, 1949; Kempthorne, 1955; Freedman and Lane, 1983; Westfall and Young, 1993; Edgington, 1995; Good, 2002, 2005; Westfall and Troendle, 2008; Pesarin and Salmaso, 2010); a detailed, extensive historical account is provided by Berry et al. (2014).

Quite often, permutation tests are presented as an alternative, or in contradistinction, to parametric approaches. This need not be the case: permutation tests do not depend on parametric methods for their existence, and were not created as a direct response or as an alternative to limitations of parametric ones. On the contrary: parametric methods and the assumptions they entail were introduced to solve practical problems at a time in which the ability to perform large number of computations was beyond even the most epic efforts. One could speculate that if computers already existed before the 20th century, parametric methods such as the t-test or the F-test would have never been devised as known today; perhaps not

even inverse probability, from the 17th century and recently in vogue again, would have been developed had computers been available at the time. However, permutation tests did benefit from the parametric literature in that the latter provides test statistics that have some interesting, desirable properties, such as pivotality (Hall and Wilson, 1991).

In fairness, some comparisons could still be made. Differently from parametric tests, permutation methods do not depend on underlying theoretical distributions, do not suffer from not quite as stringent assumptions (normality, independence, homogeneous variances), allow the use of non-random samples even of small sizes, and permit a wider variety of test statistics such as those that have poorly known distributions. They are also relatively resilient to outliers (and robust statistics can be used, even without a limiting distribution). All the information necessary for inference is contained in the data itself, not on some idealised population or frequency distribution. An assumption is however needed: the one of exchangeability, that is, that the data after shuffling remain just as likely as the original.

In the brain imaging field, parametric inference using the general linear model (GLM; Scheffé, 1959; Searle, 1971; Friston et al., 1994) has, for various reasons, been the stock method: the same framework is treated as applicable to diverse imaging modalities, can be very fast, is well understood, was the first to be available in software packages and, among other benefits, is relatively robust to certain departures from its assumptions. These assumptions refer to (I) having observations that are independent, (II) that have the same variance, and (III) that are normally distributed. In one way or another, these can be violated: repeated measurements, subjects that are recruited based on familial relationship with other subjects in the same sample, effects or confounds that affect the variance, lack of normality, non-linear effects, different physical properties and information content of different imaging modalities, among others.

On top of this, there is the multiple testing problem: the parametric solution builds on properties of a random field that must satisfy a myriad of other supposi-

tions: the probability distribution must be the same across space, the field must be a sufficiently "good" lattice representation of an underlying continuous process, which on its turn implies that the field must be smooth, and that this smoothness is larger than the voxel size. Plus, the smoothness must be known or estimable at each point, and (ideally), be the same for all positions. Moreover, it is also required that the spatial autocorrelation function is twice differentiable at the origin, that the field is thresholded at high levels, and that supra-threshold regions do not touch the borders of the segment of the field that is being inspected. Furthermore, direct results are currently available only for fields that are either Gaussian, or that can be directly related to a Gaussian distribution, as t, F, χ^2 and T^2 (Worsley et al., 1996; Cao and Worsley, 1999).

Permutation tests obviate all such issues, that are mostly intrinsic to parametric approaches. In the imaging literature, they were proposed by Blair and Karniski (1994) and Holmes et al. (1996); subsequent, relevant work has been done for various cases and applications (Arndt et al., 1996; Locascio et al., 1997; Brammer et al., 1997; Belmonte and Yurgelun-Todd, 2001; Bullmore et al., 1996, 1999, 2001; Nichols and Holmes, 2002; Breakspear et al., 2004; Laird et al., 2004; Suckling and Bullmore, 2004; Hayasaka et al., 2004; Mëriaux et al., 2006; Eklund et al., 2012; Ge et al., 2012; Winkler et al., 2014, 2016c; McFarquhar et al., 2016). For all its benefits, and now feasible computing, it is of no surprise that interest in this type of test is expected to increase in the upcoming future.

That future is not quite distant. Two recent developments have further propelled the interest on permutation tests. The first is the ongoing reproducibility crisis, that while not specific to brain imaging and affecting essentially all research areas (Baker, 2016), relates to it through applications in the fields of psychology and social sciences mainly, but also medical sciences, physics and engineering; it also relates simply for suffering from similar vices and practices (Carp, 2012; Gorgolewski and Poldrack, 2016). While the causes for the crisis are multiple (Eicken, 2013; Begley and Ioannidis, 2015) and outside the scope of this thesis, there is little

doubt that poor statistical practices and too liberal test levels account for some of the ease with which non-existing effects can be labelled as significant (Ioannidis et al., 2011). Studies that lead to p-values too close to the test level, and that rely on too many assumptions are those most vulnerable to errors if such assumptions are not met.

This brings to the second recent development: the finding that at least one of the prevailing parametric strategies used for inference in brain imaging, the one that computes p-values for the size of clusters of voxels that survive a predefined threshold using the aforementioned random field theory, yields unacceptably high false positive rates (Eklund et al., 2016), precisely because some of its suppositions are not met. Permutation tests offer a remedy to issues such as these, provided that their own assumptions, which are fewer and more easily attainable, are satisfied; in some cases, this requires a change of the test statistic.

1.2 Overview of the thesis

Statistics occupy a prominent place in the assessment of brain imaging data. As an example, consider the core analysis tools currently available in the FMRIB Software Library (FSL), which are depicted as a flow chart in Figure 1.1. Direct statistical modelling and inference are present in the tools FILM, FLAME, `randomise` and PALM, with the first two using parametric methods; statistics also play a key role in all other tools. A single statistical framework, if used for various imaging modalities, is more likely to be successful if based on a minimal set of assumptions, for which permutation methods are strong candidates.

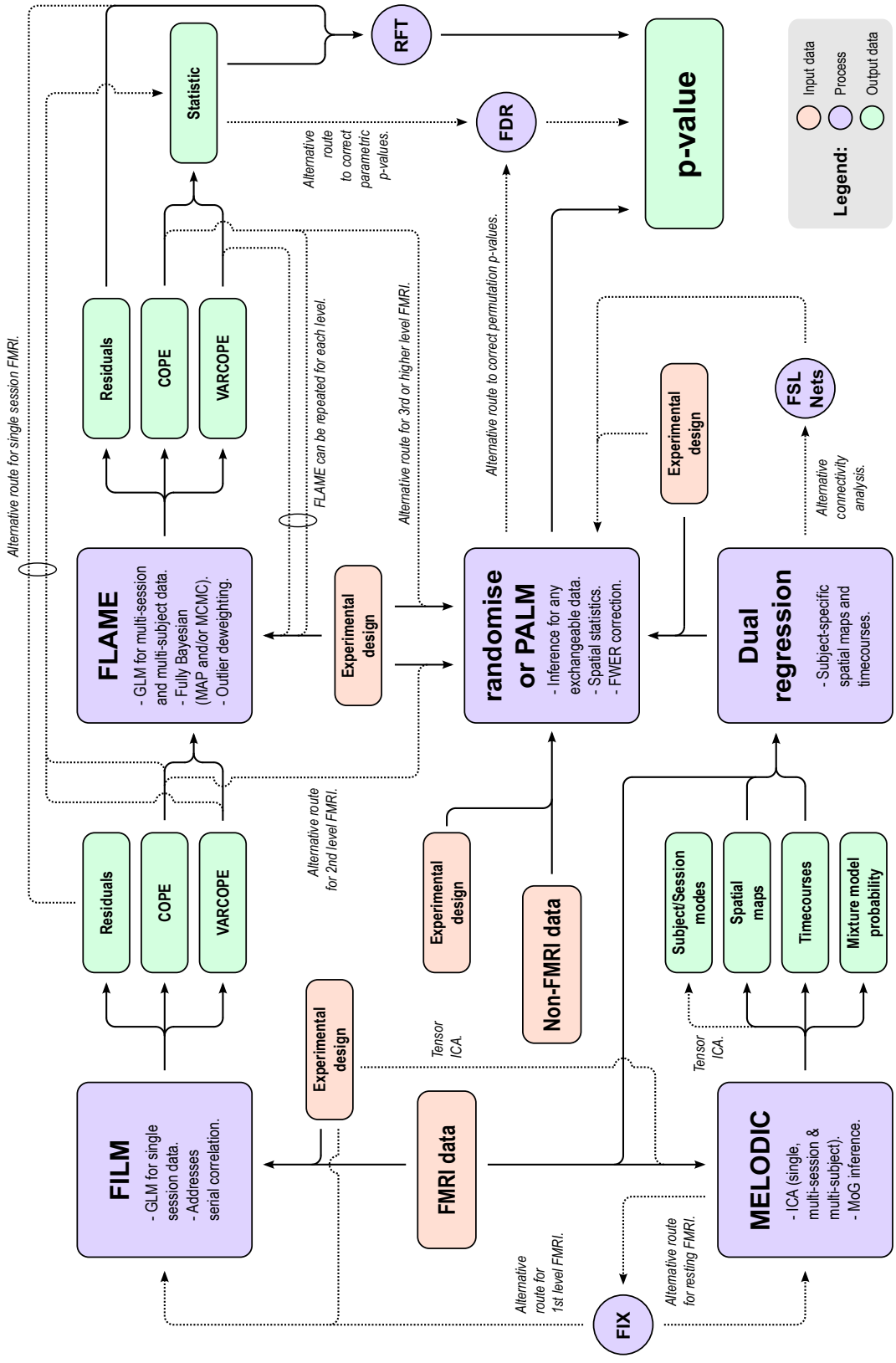
In the example of FSL, although `randomise` already covers the most common types of experimental design, currently it does not address various other useful possibilities that include, for instance, (I) statistics that are robust to heteroscedasticity, (II) multivariate methods, (III) flexibility to accommodate certain designs with repeated measurements, (IV) correction for multiple testing based on minimal assumptions, not only across voxels, but also across multiple contrasts

and multiple designs, (v) inference with missing data, (vi) multi-level inference on complete generality, for fixed-, random-, and mixed-effects, and (vii) all these performed in a reasonable amount of time even for large datasets. The absence of these techniques is not merely a lack of implementation in FSL or other software: some these topics have only been superficially covered by the literature, with even less work having addressed aspects that are specific to neuroimaging.

In *this* doctoral thesis, methods for the items (iii) and (vii) are introduced, thus widening the applicability of permutation tests to more cases than otherwise possible; the respective papers have been published (Winkler et al., 2015, 2016b). Item (i) was discussed in Winkler et al. (2014), whereas items (ii) and (iv) were presented in Winkler et al. (2016c); these two last papers are the product of *another* thesis by the same author, and care was taken to avoid overlapping original material. Items (v) and (vi) are left as future work. This thesis also explores an application of a permutation-based multivariate method, item (ii) above, using the theory presented in Winkler et al. (2016c) as the foundation that allows the replacement of analyses of the volume of the cerebral cortex for a joint analysis cortical thickness and cortical surface area. The respective paper is, at the time of this writing, under review, and a pre-print is publicly available in the *bioRxiv* server (Winkler et al., 2016a).

Each chapter is a self-contained piece of work, under the common motif that is

Figure 1.1: *(page 23)* Flow chart of the main analysis methods available in FSL. Statistical methods occupy a central role, with the tools FILM, FLAME, `randomise`/PALM, RFT and FDR. In particular, permutation methods occupy a privileged, central position: with the exception of single-subject FMRI, all strategies for analysis can pass through `randomise` or PALM. Abbreviations in this figure: COPE: contrast of parameter estimates; FDR: false discovery rate; FILM: FMRIB's Improved Linear Model; FIX: FMRIB's ICA-based Xnoiseifier; FLAME: FMRIB's Local Analysis of Mixed Effects; FMRI: functional magnetic resonance images; FSL: FMRIB Software Library; FWER: Familywise error rate; GLM: general linear model; ICA: independent component analysis; MAP: maximum a posteriori probability estimation; MCMC: Markov-chain Monte Carlo; MOG: Mixture of Gaussians; PALM: Permutation Analysis of Linear Models; RFT: Random field theory; VARCOPE: variance of the contrast of parameter estimates.


Legend:
Input data
Process
Output data
FMRI data
FILM
- GLM for single session data.
- Addresses serial correlation.
MELODIC
- ICA (single, multi-session & multi-subject).
- MoG inference.
FIX
FLAME
- GLM for multi-session and multi-subject data.
- Fully Bayesian (MAP and/or MCMC).
- Outlier deweighting.
randomise or PALM
- Inference for any exchangeable data.
- Spatial statistics.
- FWER correction.
Dual regression
- Subject-specific spatial maps and timecourses.
Residuals
COPE
VARCOPE
Statistic
RFT
FDR
p-value
FSL Nets
Experimental design
Non-FMRI data
Subject/Session modes
Spatial maps
Timecourses
Mixture model probability
Tensor ICA.
Alternative route for single session FMRI.
Alternative route for 2nd level FMRI.
Alternative route for 3rd or higher level FMRI.
FLAME can be repeated for each level.
Alternative route to correct permutation p-values.
Alternative route to correct parametric p-values.
Alternative connectivity analysis.
Alternative route for 1st level FMRI.
Alternative route for resting FMRI.

the extension of possible uses of permutation tests. Each individual chapter can be read together with the others or in isolation, although there is some intertextual elements, particularly on what concerns the non-parametric combination (NPC), that appears in Chapters 3 and 4. All three chapters build on previous research from the author, in particular Winkler et al. (2014, 2016c), but also, to some extent, Winkler et al. (2010, 2012). However, none of these readings are strictly necessary, and each chapter provides an introduction with a review of the relevant literature that should help to update the reader who is already familiar with the topic; readers who are not familiar are referred to the cited background literature.

1.2.1 Permutation under structured dependence

In Chapter 2 we propose a solution to a crucial problem that arises when using permutation tests for the analysis of experimental data using repeated measurements, which, depending on the hypothesis, violates exchangeability: the various measurements obtained from a given subject are not independent from each other, thus not exchangeable. Another case is for data such as those from the Human Connectome Project (HCP; Van Essen et al., 2012, 2013): as each subject does not constitute an independent observation — many are twins, with additional siblings also participating of the study — conventional permutation strategies, despite their superior properties in general, if used, would dangerously expose the researcher to an increased risk of false positives.

We address this problem by constructing the permutations in a way that respects the data structure, without the need to explicitly model such structure, by imposing restrictions on exchangeability. Observations are organised in blocks, which are nested within other blocks in a multi-level fashion; these blocks can be shuffled a whole, and inside them, sub-blocks are further allowed to be shuffled, in a recursive process. The method is flexible enough to accommodate ordinary permutations of the data, random sign flipings of the data (sometimes also called *wild*

bootstrap; Guillaume et al., 2014), and permutations together with sign flippings.[1] In particular, this permutation scheme allows the data such as those from the HCP to be analysed via permutations: subjects are allowed to be shuffled with their siblings while keeping the joint distribution intra-sibship maintained. Then each sibship is allowed to be shuffled with others of the same type, ultimately ensuring that the false positive rate is controlled at the level of the test.

The results from this chapter show that the error type I is controlled at the nominal level, and the power is just marginally smaller than that would be obtained by permuting freely if that were allowed. The more complex the block structure, the larger the reductions in power, although with large sample sizes, the difference reduces. Importantly, simply ignoring family structure in designs as these causes the error rates not to be controlled, with excess false positives, and invalid results. We show examples of false positives that can arise, even after correction for multiple testing, when testing associations between cortical thickness, cortical area, and measures of body size, such as height, weight, and body-mass index, all of which are known to be highly heritable. Such false positives can be avoided with permutation tests that respect the family structure.

1.2.2 Accelerating permutation tests

For small sample sizes, low resolutions, or small regions of interest, permutation tests can run very quickly (in a matter of minutes with current personal computers). For larger data, however, they become computationally intensive, and accelerations become useful. The natural strategies to increase the speed consist of the use of faster computers, parallel processing, graphical processing units (GPUs),

[1] In conventional permutations, the data is randomly reshuffled multiple times, the model re-fit at each time, the test statistic recomputed, and the empirical null distribution constructed. In sign flippings, instead of reshuffling the data, these are multiplied randomly by either $+1$ or -1, with the remaining of the procedure similar. In permutations together with sign flippings, both things are done simultaneously. For a detailed discussion on the assumptions and the merits of each, see (Winkler et al., 2014).

or use of optimised code. These are general methods that can be considered to various applications.

Instead, Chapter 3 does not build on such generic software or hardware improvements, but on properties of certain test statistics and other statistical devices. We evaluate six different strategies for acceleration — some already existing in the literature, others introduced, improved, or adapted for brain imaging. These strategies allow accelerations larger than two orders of magnitude, yielding nearly identical results as in the non-accelerated case. Some, such as tail approximation, are generic enough to be used in nearly all the most common scenarios, including univariate and multivariate tests, spatial statistics, and for correction for multiple testing.

In addition to accelerating the tests, some of these methods allow continuous p-values to be obtained, and refine them far into the tail of the distribution of the test statistic, thus avoiding the usual discreteness of p-values in permutation methods, which can be a problem in some applications if too few permutations are done. Based on the results of extensive simulations and on the reanalysis of previously published real (Douaud et al., 2007), we provide a set of recommendations for cases in which each method should be preferred.

1.2.3 Application to cortical morphometry

A joint, combined analysis of multiple variables can be done using the *non-parametric combination* (NPC). While NPC is only relatively new (Pesarin and Salmaso, 2010), its use in neuroimaging was impossible for two reasons: the need to store very large amounts of data, and low power caused by the fact that, in imaging, correction for multiple testing is mandatory. In a work unrelated to this thesis, Winkler et al. (2016c) have made modifications to the original NPC so as to render it feasible, valid, and powerful for brain imaging, addressing specific details that include multiple testing and the treatment of spatial statistics.

In Chapter 4, the application of this modified NPC to structural imaging reveals

something quite remarkable: that there is a viable way to assess cortical volume that, differently than the wildly popular voxel-based morphometry (VBM), does not mirror the cortical surface area. More: the method is able to reveal patterns of cortical pathology that may affect thickness and area in opposing directions — something that VBM cannot do — and even jointly remains a directional test, i.e., in which the direction of the effect (positive or negative) can be ascertained, thus in contrast with classical tests such as MANCOVA, which do not possess a sign and thus, are unable disambiguate directional effects.

However, given that thickness and surface area have quite recently been shown to reflect distinct biological processes, cortical volume could still find use for the study of disorders that affect both cortical area and thickness simultaneously. Here we demonstrate that NPC is a better suited solution. This superiority remains even when compared to an improved, analytic method to measure volume, that is not affected by the local configuration of the cortical morphology, and that is also novel and introduced in this thesis.

In the same chapter we also provide, obiter dictum, a comparison of the four existing methods for areal resampling, answering a well known question that had so far not been approached in the literature: can the simple nearest neighbour method for areal interpolation be trusted? The answer is a confident yes, provided that there is smoothing. This translates into substantial gains for those interested, as the alternative, exact method, is computationally too intensive. The chapter equips researchers focused in the in vivo study of cortical anatomy with the information necessary to make the best use of the state-of-the-art morphometry methods.

This page intentionally left blank.

Chapter 2

Multi-level block permutation

2.1 Introduction

In the context of hypothesis testing using the general linear model (GLM) (Scheffé, 1959; Searle, 1971), permutation tests can provide exact or approximately exact control of false positives, and allow the use of various non-standard statistics, all under weak and reasonable assumptions, mainly that the data are *exchangeable* under the null hypothesis, that is, that the joint distribution of the error terms remains unaltered after permutation. Permutation tests that compare, for instance, groups of subjects, are of great value for neuroimaging (Holmes et al., 1996; Nichols and Holmes, 2002), and in Winkler et al. (2014), extensions were presented to more broadly allow tests in the form of a GLM, and also to account for certain types of well structured non-independence between observations, which ordinarily would preclude the use of permutation methods. This was accomplished by redefining the basic exchangeable unit from each individual datum to blocks of data, i.e., rather than asserting exchangeability across all observations of a given experiment, blocks of exchangeable units are defined; these *exchangeability blocks* (EBS) can be rearranged as a whole (*whole-block exchangeability*), or the observations within block can be shuffled among themselves (*within-block exchangeability*), using either permutations, sign flippings, or permutations combined with sign

flippings.

In the same work, the G-statistic, a generalisation over various commonly used statistics, including the F-statistic, was proposed. G is robust to known heteroscedasticity (i.e., the situation in which the variances are known to be not equal across all observations, which can be then classified into variance groups) and can be used with the GLM, ensuring that pivotality is preserved, a crucial requisite for exact control over familywise error rate (FWER) using the distribution of the most extreme statistic (Westfall and Young, 1993), as needed in many neuroimaging studies. A *pivotal* statistic has a sampling distribution that does not depend on unknown parameters. Indeed, the use of EBS allows for variances to be heterogeneous, provided that the groups of observations sharing the same variance (i.e., *variance groups*, VGS) (Woolrich et al., 2004) are compatible with the EBS; specifically, for within-block exchangeability the VGS must coincide with the blocks, and for whole-block exchangeability they must include one or more observations from each block in a consistent order.

This arrangement, using a statistic that is robust to heteroscedasticity, the use of variance groups, and the imposition of restrictions on exchangeability through the use of EBS, allows inference on various designs that, otherwise, would be much more difficult to do non-parametrically. These designs include paired tests, longitudinal designs, and other common tests that involve repeated measurements. However, certain study designs, despite exhibiting well-structured dependence between observations, still cannot be accommodated in the above framework. This occurs when the overall covariance structure is known, but its exact magnitude is not. An example occurs when multiple measurements per subject are performed in more than one session, with more than one measurement per session: the measurements within session may be exchangeable, but not across sessions. Another example is for studies using siblings, such as designs using discordant sib-pairs (in which only one sibling is affected by a given disorder), or using twins: permutations that disrupt the constitution of any sibship cannot be performed, as this

would violate exchangeability.

Studies such as these are relatively common, notably those that involve siblings. However, whereas in classical twin designs the central objective is to quantify the fraction of the variation in a measurement (trait) that can be explained by the familial relationship between subjects after potential confounds have been taken into account, a quantity known as *heritability*, here the concern is with a general linear model, and the objective is to test the influence of explanatory variables on the observed data. In other words, the interest lies on the relationship between the covariates and the main trait, while the non-independence between observations, which is a feature of interest in a heritability study, is here a form of nuisance that imposes restrictions on exchangeability for permutation inference for the GLM.

Rather than inadvertently breaking these restrictions, here we propose to test the null hypothesis using a subset of all otherwise possible permutations, only allowing the rearrangements that respect exchangeability, thus retaining original joint distribution unaltered. Exchangeability with respect to a subset of all possible permutations is termed *weak exchangeability* (Good, 2005). For conciseness, we will use the solitary term "exchangeability", while making clear the subsets of permutations for which this is valid. As in our previous work, we treat observations or entire blocks of data as weakly exchangeable, but here we further extend the definition of EBs to allow more complex designs to be addressed. This is accomplished through the use of *multi-level exchangeability blocks*, in which levels consist of *nested* blocks; for each such block the state of within- or whole-block exchangeability can be specified. The blocks are defined hierarchically, based on information about the dependence within data, but not requiring the modelling of the actual dependency. Even though the possibility of using nested blocks was anticipated in Winkler et al. (2014) ("Whole-block and within-block can be mixed with each other in various levels of increasing complexity", page 386), nothing further was studied or presented at the time. Here we provide a comprehensive description of the approach, investigate its performance, its power, and present

an applied example using the data structure of the ongoing Human Connectome Project (HCP). In the Section 2.3, we present an implementation strategy.

2.2 Theory

2.2.1 Terminology

When contrasting the method described in this chapter with simple data rearrangement, various terms could be adopted: *single-level* vs. *multi-level* block shuffling, emphasising the levels of relationship between observations; *unrestricted* vs. *restricted*, emphasising the imposition of restrictions on how the data are allowed to be rearranged at each shuffling; *free* vs. *tree* shuffling, emphasising the tree-like structure of the relationships between observations that allow shuffling. All these terms have equivalent meaning in the context of this chapter, and are used interchangeably throughout. The generic terms *shuffling* and *rearrangement* are used when the distinction between permutations, sign flippings or permutations with sign flippings is not relevant.

2.2.2 Notation

We consider a GLM that can be expressed as $\mathbf{Y} = \mathbf{M}\boldsymbol{\psi} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is the $N \times 1$ vector of observed data, $\mathbf{M}$ is the full-rank $N \times r$ design matrix that includes explanatory variables (i.e., effects of interest and possibly 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. Estimates for the $\boldsymbol{\psi}$ can be computed by ordinary least squares, i.e., $\hat{\boldsymbol{\psi}} = \mathbf{M}^{+}\mathbf{Y}$, where the superscript $(^{+})$ denotes a pseudo-inverse. One generally wants to test the null hypothesis that a given combination (contrast) of the elements in $\boldsymbol{\psi}$ equals to zero, that is, $\mathcal{H}_0 : \mathbf{C}'\boldsymbol{\psi} = \mathbf{0}$, where $\mathbf{C}$ is a $r \times s$ full-rank matrix of s contrasts, $1 \leqslant s \leqslant r$. The commonly used F statistic can be computed as usual and used to test the null hypothesis. When $s = 1$, the Student's t statistic can be computed as $t = \text{sign}(\hat{\psi})\sqrt{F}$. A p-value for the statistic is calculated by means

of shuffling the data, the model, the residuals, or variants of these (Winkler et al., 2014, Table 2). In any of these cases, to allow rearrangements of the data, 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. These rearrangements are represented by permutation and/or sign flipping matrices **P**, and the set of all such matrices allowed for a given design is denoted as $\mathcal{P}$.

At its simplest, the EBs for within- or whole-block exchangeability can be identified or represented by a set of indices $\{1, 2, ..., B\}$, one for each of the B blocks. A vector of size $N \times 1$, can be used to indicate to which EB each observation from **Y** belongs (Figure 2.1, *left*); an extra flag is passed to the shuffling algorithm (such as the `randomise` algorithm) to indicate whether the rearrangements of the data should happen as within- or as whole-block. While this notation probably covers the majority of the most common study designs, it allows only within- *or* whole-block, but not *both* simultaneously; in other words, if in a study the observations can be permuted within block, and the blocks as a whole can also be permuted, such notation does not convey all possibilities for reorganising the data while preserving their joint distribution unaltered, and algorithms would perform fewer shufflings than those that are effectively allowed.

This can be addressed by extending the notation from a single column to a multi-column array, allowing nested EBs to be defined, such that blocks can contain sub-blocks, in a hierarchical fashion, and where each column represents a level; we use the leftward columns to indicate higher, and rightward to indicate lower levels. More columns alone, however, are not sufficient, because at each level, shufflings of observations or of sub-blocks can be allowed within-block, or the blocks at that level can be shuffled as a whole. Hence to discriminate between one type or the other, we use negative indices to indicate that the exchangeable units at the level immediately below should not be permuted, and positive indices

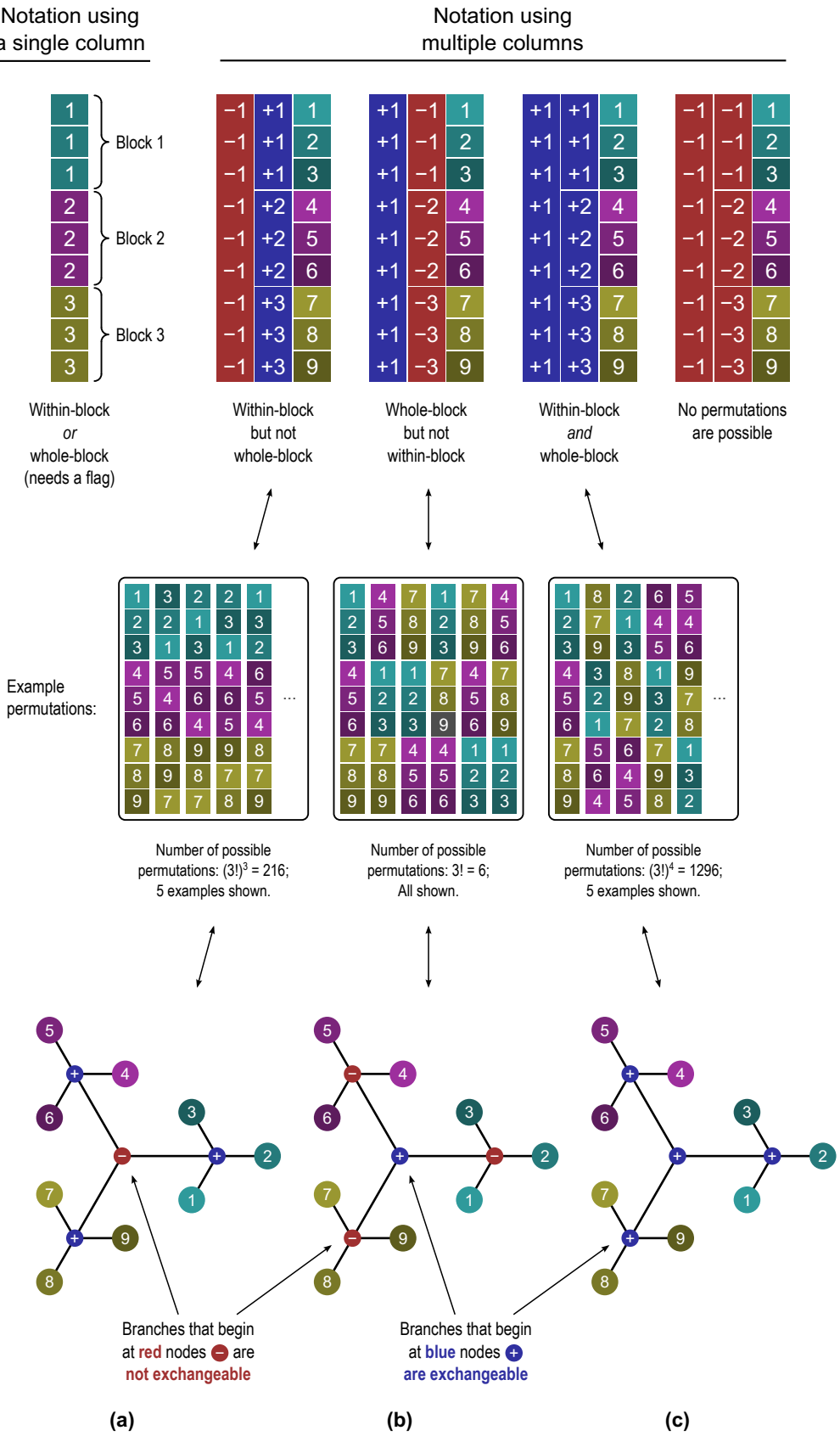
indicate that shuffling of these units is allowed as usual (Figure 2.1, *right*). The exchangeable units can be sub-blocks, which can contain yet other sub-blocks, or observations if the next level immediately below is the last.

These two notations, i.e., using single- or multi-column indices, do not represent mathematical entities, and are not meant to be used for algebraic manipulation; rather, these notations are shorthand methods to represent structured relationships between observations. The covariance structure prevents unrestricted shuffling from being considered, but it often permits shufflings to happen in a certain orderly manner that preserves the joint distribution of the data. These notations are to be used by the algorithm that performs the test to construct the permutation and/or sign flipping matrices, which then can be used to effectively disarrange the model to construct the distribution of the statistic under the null hypothesis.

2.2.3 Visual representation

The notation using multiple columns encapsulates all the information necessary not only for the rearrangements to be constructed, but also to depict the relation-

Figure 2.1: *(page 35)* Different notations for the specification of exchangeability blocks; in this example, 3 blocks of 3 observations each. *Top-left:* In a single-column notation, each block has its index (here 1, 2, and 3, shown in different, random colours for clarity), and either within- or whole-block exchangeability are possible, but not both simultaneously. The specification of which kind of shuffling is to be done requires extra information, as a flag passed to the algorithm that permutes the data. *Top-right:* In a multiple-column notation, that information is encoded by virtue of the indices having a sign indicating whether the exchangeable units of a block at a given level should be shuffled as a whole ($+$) or kept fixed ($-$); these are shown respectively in blue and red. The signs define whether it is possible to perform rearrangements within-block, or of the blocks as a whole, or both. The rightmost example serves only to illustrate the notation, and is not useful in practice as all the observations would need to remain still. *Middle:* Example permutations are shown, with the observation indices coloured for clarity. *Bottom:* Visual representations using a tree diagram. The levels can be depicted as branching from a central (top/central) node, akin to a tree in which the most peripheral elements (leaves) represent the observations. The nodes from which the branches depart can be labelled as allowing permutations ($+$) or not ($-$), shown respectively in blue and red colours. The letters (a) through (c) refer to the variance groups in Figure 2.3.


Notation using a single column
Notation using multiple columns
Block 1
Block 2
Block 3
Within-block or whole-block (needs a flag)
Within-block but not whole-block
Whole-block but not within-block
Within-block and whole-block
No permutations are possible
Example permutations:
Number of possible permutations: (3!)^3 = 216; 5 examples shown.
Number of possible permutations: 3! = 6; All shown.
Number of possible permutations: (3!)^4 = 1296; 5 examples shown.
Branches that begin at red nodes ⊖ are not exchangeable
Branches that begin at blue nodes ⊕ are exchangeable
(a)
(b)
(c)

ships between the observations in a tree-like diagram, highlighting their hierarchy, as shown in the lower panel of Figure 2.1. Branches can only be shuffled with each other if their size and internal covariance structure are perfectly identical; this information is contained in the signs and indices used to represent each block: positive indices (shown in blue) allow them to be permuted; negative (in red) prevent permutation and keeps the branches in their fixed positions. The permutation of branches at lower levels (when these exist) is controlled by the nodes at these lower levels, independently from those at higher levels or within the same level.

Using the tree diagram, it becomes clear that the terms "within-block" and "whole-block", that have been used so far to describe exchangeability and permutation strategies, become no longer necessary, as either the branches can be shuffled, or they cannot. It is also helpful in emphasising that more complicated designs can be considered using multi-level blocks, in which even the distinction between within- and whole-block is softened, as each level in the multi-column notation is not restricted to contain purely positive or negative indices restricting (or not) the shuffling of their constituent sub-blocks (branches). These can be present alongside each other if immediately below a level in which shuffling is not allowed, such that some branches may be allowed to be shuffled, whereas others are not. It may also be the case that some levels need to be included in the notation only so that the number of levels remains the same across all branches of the tree, from the top node to the most distal (leaves), without affecting the construction of $\mathcal{P}$, but ensuring that the notation can be stored, without gaps, in a two-dimensional array; in the visual representation these are shown as small, sign-less nodes. Figure 2.2 (*left* and *centre*) exemplifies these cases. Although the multi-column notation and the corresponding tree can become very complex, the simple, unrestricted exchangeability can also be accommodated, as shown in Figure 2.2 (*right*).

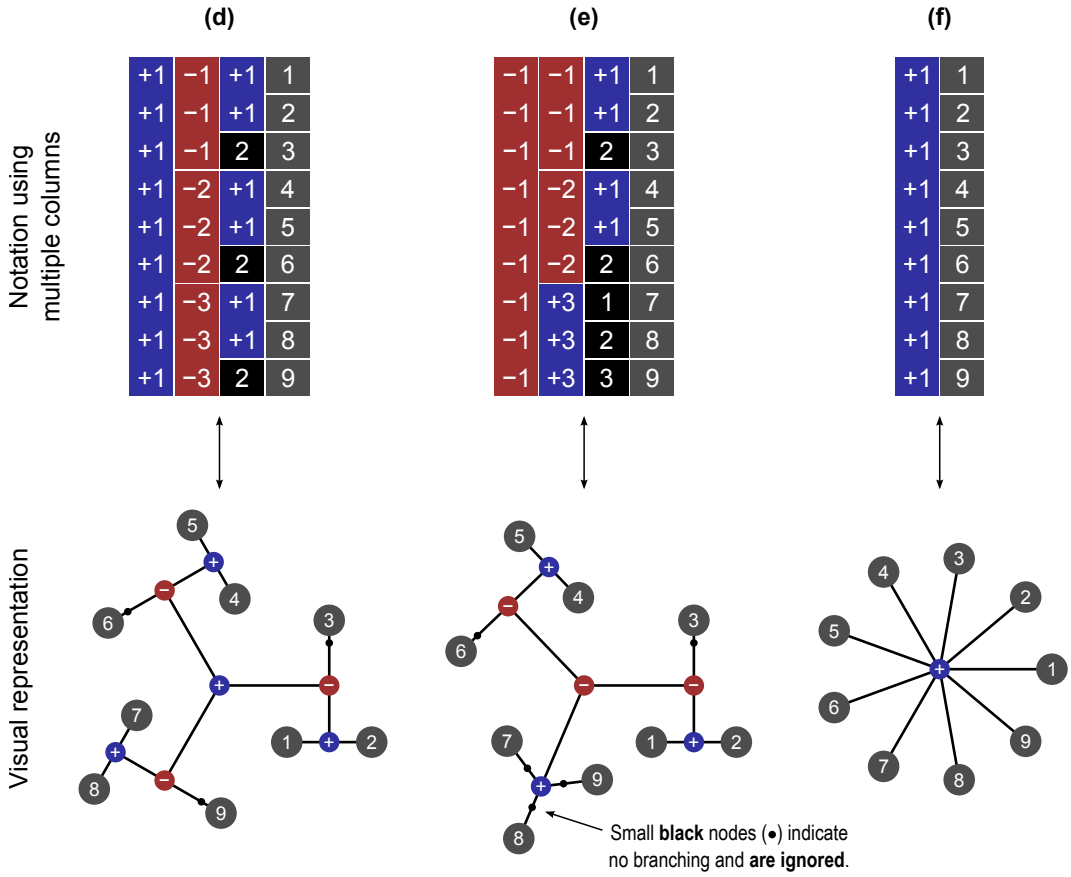


Figure 2.2: The multi-level definition of blocks allows more complex relationships between observations. *Left:* Three blocks of identical structure (2nd column) can be shuffled as a whole (as indicated by the positive indices in the 1st column); within each (3rd column), only two out of their three constituting observations can be swapped (1 and 2, 4 and 5, and 7 and 8), whereas the third on each (3, 6 and 9) cannot; levels for these last branches are completed with blocks for which the sign has no meaning (in black), as they remain unaltered towards the next level (4th column), and represent no actual branching. In the visual representation, these black blocks are shown as small black dots on continuous branches. This example could represent 3 sets of siblings, each composed of a pair of monozygotic twins and a third non-twin. *Centre:* An example showing that it is possible to mix types of blocks in the same level (2nd column). As shown, the first two blocks in the 2nd column cannot be swapped despite similar coding, and neither of these can be permuted with the third, which has a different structure consisting of three observations (7, 8 and 9) that can be shuffled freely. This example could represent 3 sets of siblings, the first a pair of monozygotic twins and a non-twin, the second a pair of dizygotic twins and a non-twin (if certain environmental effects are considered), and the third a set of three non-twin siblings. *Right:* The same notation can also accommodate simple designs. Here all 9 observations can be permuted without restrictions on exchangeability.

2.2.4 Variance groups and the G-statistic

When the variances can be assumed to be the same throughout the sample, the classical F and the Student's t statistics can be used; these statistics have sampling distributions that do not depend on any unknown population parameters, but solely on the degrees of freedom, i.e., these are pivotal statistics. However, if homoscedasticity cannot be assumed, although F and t can still be used with permutation tests in general, they cannot be used to correct for multiple testing using the distribution of the most extreme statistic. The reason is that under heteroscedasticity, these statistics cease to be pivotal, and follow instead distributions that depend on the heterogeneous variances for the different groups of observations, causing them to be no longer adequate for FWER correction. Instead, a statistic that is robust to heteroscedasticity is necessary.

The G-statistic (Winkler et al., 2014) was proposed to address this concern; this statistic is a generalisation of various other well established statistics, including F and t, as well as the v-statistic used for the classical Behrens–Fisher problem. The definition of the variance groups used to calculate G is based on knowledge about the data, and such groups need to be constructed together with the definition of the blocks. However, VGs and EBs represent different concepts; although they may coincide for simple designs, they do not need to. The EBs are used to indicate sets of observations that must remain together in every permutation due to having a non-diagonal *covariance* structure, and are used by the permutation algorithm to rearrange the data many times to build the empirical distribution. The VGs, however, are used to indicate sets of observations that possess the same *variance*, and are used to estimate the sample variance(s) when computing the statistic. Despite the distinction, any pair of observations that have the possibility of being swapped according to the EB structure must be in the same VG; observations in different variance groups cannot be permuted as that would modify the joint distribution, thus violating exchangeability.

For simple within-block permutation, the most restrictive configuration for the

variance groups, that is, the configuration in which fewer observations need to be assumed to share the same variance, is such that each block corresponds to its own VG. For simple whole-block permutation, on the other hand, the first observation from each block, together, constitute a VG, the second observation from each block, together, another VG, and so forth. The minimum set of variance groups for more complicated designs can be derived similarly from the configuration of the exchangeability blocks; examples are shown in Figure 2.3. The stringency of this definition lies in that, depending on the configuration of the EBS, each VG can contain only the smallest possible number of observations that can be assumed to have the same variance given the covariance structure imposed by the blocks. Such definition can, however, be relaxed by merging these minimum groups whenever homoscedasticity across more than one VG can be considered, while retaining the EBS unaltered. Whether merger, or any other definition, for the VGS should be sought for a given design may depend on information about the data or on the design itself. For a simple paired t-test, for instance, although each pair could in principle constitute a VG on its own, homogeneous variances can fairly well be assumed, with the benefit of much better variance estimates than would be obtained with groups of two sole observations.

Regardless of which strategy is used to define the variance groups, and irrespective of the indices used to represent each of them, the column vector containing these indices must be invariant with respect to the permutations that are allowed for a given design. In other words, let $\mathbf{v}$ be the column vector of length N containing the indices that represent each variance group, such as those in Figure 2.3. For any permutation matrix $\mathbf{P} \in \mathcal{P}$, $\mathbf{Pv} = \mathbf{v}$, that is, $\mathbf{v}$ is a common eigenvector for all permutation matrices in $\mathcal{P}$. Any permutation that breaks this equality must not be used to test the null hypothesis, as this would mix observations that belong to different VGS, thus violating exchangeability. Likewise, a definition of groups that does not meet this criterion must not be used.

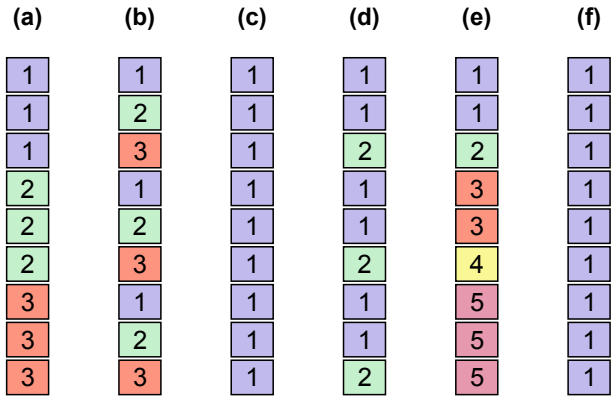


Figure 2.3: Variance groups defined from the exchangeability blocks (a)–(c) shown in Figure 2.1, and (d)–(f) in Figure 2.2. These are the most restrictive configurations for the VGs that are possible given the structure imposed by the EBs. If, however, despite the covariance structure between observations, their variances are known to be or can be assumed to be homogeneous, some or all of these groups can be merged, with the additional benefit of improving the variance estimates. Alternatively, the groups can be entirely replaced by a different definition if additional information from the variance of the data is available. In (e), note two groups with only one observation each; see the main text for details.

2.2.5 Number of permutations

With the multi-level block permutation strategy, the rules to calculate the number of permutations are similar, yet more general than in the case of a single level that could be represented with a single-column notation. The number still depends on the number of repeated rows in the design matrix for methods as Manly and ter Braak (Manly, 1986; ter Braak, 1992) or, for methods as Draper–Stoneman and Freedman–Lane (Draper and Stoneman, 1966; Freedman and Lane, 1983), on the number of repeated rows across only the columns that are tested in the contrast after the model has been partitioned into effects of interest and nuisance effects.

Once the tree has been constructed, for the EE assumption, and in the absence of repetitions in the design as described above, the number of permutations can be calculated separately for each node in which shuffling is allowed as $B!$, with B denoting the number of branches that begin at that node. If however, there are branches with identical structure and containing the repetitions in the design mat-

rix, the number of possible permutations for that node is then $B! \Big/ \prod_{m=1}^{M} B_m!$, where M is the number of unique branches beginning at that node, and B_m the number of times each of the M unique branches begins at that node. The number of permutations for nodes that cannot be permuted is simply 1, that is, no permutation. With the number of permutations at each node calculated, the overall number of possible permutations for the whole design (whole tree) is the product of the number of possible permutations for all the nodes.

For ISE, the number of sign flippings at the highest node in which shuffling is allowed is 2^B, and 1 for all other nodes that lie below (distal) in the hierarchy. For the nodes in which shuffling is not allowed, the number of possible flips is 1, that is, no sign flippings are allowed, but it can still be higher than 1 for the nodes that lie below in the hierarchy. Unlike with permutations, the eventual presence of repeated elements in the design matrix does not affect the number of possible sign flippings. The number of possible sign flippings for the whole design is the product of the number of sign flippings for all the nodes.

When both EE and ISE assumptions are valid for a given design, permutations can happen with sign flipping, and the total number of possible rearrangements is simply the product of the number of permutations with the number of sign flippings. Regardless of the kind of shuffling strategy adopted, the number of possible rearrangements can be extremely large, even for sample sizes of relatively moderate size. Not all of them need to be performed for the test to be valid and sufficiently exact; a random subset of all possible rearrangements can be sufficient for accurate hypotheses testing.

2.2.6 Power

The set of all rearrangements that can be performed while respecting the structure of the data is termed the *permutation space* (Pesarin and Salmaso, 2010). The restrictions imposed by the EBs cause this space to be reduced, sometimes considerably, as none of the rearrangements that would violate exchangeability are

performed. If the restrictions are such that the permutation space is not a representative, uniform sample from what the space would be without such restrictions, power may be reduced. In the Section 2.4 we assess various configurations for the multi-level EBs and their impact on the ability to detect true effects.

2.2.7 Outliers

Permutation tests strategies in general tend to be robust to outliers (Anderson and Legendre, 1999; Good, 2005): although large outliers can bias the test statistic, the bias is present for all permutations; if a given outlier can be shuffled with *any* other observation, their effect on the distribution of the test statistic tends to spread uniformly throughout its whole support. However, the dependence structure and the multi-level blocks may amplify the effect of outliers, since a given extreme value may no longer be shuffled with *any* other observation, such that their their effect on the distribution of the test statistic may concentrate more strongly in some intervals than in others. The ultimate results then become difficult to predict, either in terms of conservativeness or anticonservativeness. Such problems may be minimised by providing some treatment of these extreme values; some possible remedies include censoring, trimming, replacement for ranks or quantiles, conversion of quantiles to a normal distribution, and robust regression.

2.3 Implementation

2.3.1 Permutation of the tree branches

A naïve way to select only the permutations that respect the data structure could be to create (randomly or lexicographically) permutation matrices as if the data could be shuffled freely, and then test whether these matrices would respect the configuration of the EBs at the various levels. If yes, the permutation matrix is used, otherwise it is discarded. The problem with this approach is not only that the process of testing can be slow, but also the restrictions imposed by the blocks

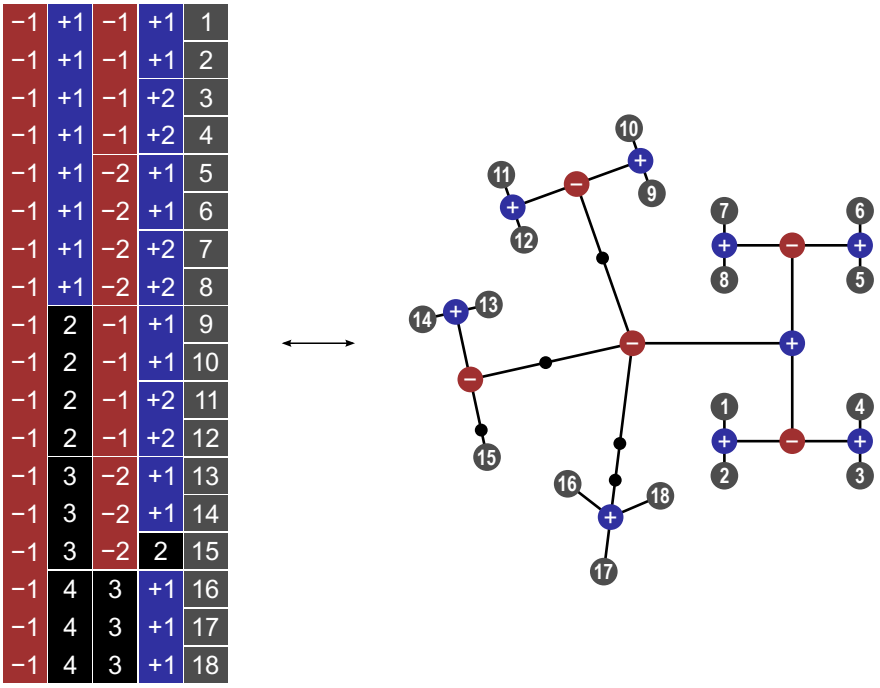


Figure 2.4: An example multi-level block structure, to be compared with the example that illustrate the implementation; compare with Figures 2.5 and 2.6, following the text description of the implementation strategy. (This is also the structure D shown in Figure 2.8.)

can reduce the number of possible permutations by many orders of magnitude, implying that for some designs, for each valid permutation that is found, enormous numbers of permutations would need to be discarded, in a very inefficient process.

However, the diagram using a tree to represent the blocks and sub-blocks, and their hierarchical relationship, does not constitute merely a visual resource. The same tree structure can be used to efficiently implement a permutation algorithm that shuffles the branches, even in programming languages that do not offer natively a tree type, such as C or Octave/MATLAB, as trees can be constructed with pointers, or with generic types, such as cells. Figure 2.4 shows an example, further expanded in Figures 2.5 and 2.6. When constructing the tree representation, each node must store three pieces of information, which are all updated as the permutations and/or sign flips are performed. These three pieces are labelled as "pieces" in Figures 2.5 and 2.6 (note the recursion), and described below:

1. A three-column array, to be used under the EE assumption, and modified at each permutation, with as many rows as branches beginning at the node. The first column contains a sequence of integers that represents each of the branches. Branches that have identical structure and contain identical leaves (i.e., repeated rows in the relevant part of the design matrix necessary to test the null hypothesis) receive the same index and so this column can contain repeated values. This sequence is used by the Algorithm "L" (Knuth, 2005), the algorithm that performs lexicographic permutations without repetitions; as the algorithm runs, the rows are permuted as a whole. The second column contains the indices that represent the current permutation in relation to the original sequence, that is, the indices that rearrange the original sequence of branches into the current state after permuting. Reverse-indexing this sequence can reset the permuted branches back to the original, unpermuted state. The third column contains the indices that rearrange the previous state into the current state and, just before running the Algorithm "L" for the next permutation, this column is regenerated as a sequence $\{1, 2, ..., B\}$, with B here denoting the number of branches (sub-blocks) in the current level; these are the indices that effectively permute the branches. If branches that begin at the node cannot be permuted (negative indices in the multi-level block notation), the whole array is replaced by some distinct marker that can be tested quickly, such as a not-a-number (NaN) or simply a zero (0).

2. A vector of counters, to be used under the ISE assumption. The counter can be in a numeral system of radix 2 (e.g., using -1 and $+1$ as symbols in lieu of the conventional binary 0 and 1), with as many digits as branches starting at the node, and representing in a direct manner the current state of the sign flippings for each branch that begin at that node. As in the case for permutations, if the signs in branches that begin at the node cannot be flipped, the vector of counters is replaced by a distinct marker, such as a not-a-number (NaN). For convenience, a the counter can be programmed using

radix 10, then converted to radix 2 when the permutations are generated (in Figures 2.5 and 2.6, the vectors with the counters are shown in base 10).

3. The branches that begin at the node. Each branch is constituted of another tree structure, in a pattern that replicates itself recursively from the top level to the most distal branches.

Once the tree has been constructed, shufflings can be performed exhaustively or, if the number of possible rearrangements is too large, only a subset needs to be performed. To generate any single permutation in lexicographic order, the tree is swept from the top node, proceeding recursively down to the next lower level before moving into the next branch in the same level, skipping the nodes in which the last lexicographic permutation has been reached (that is, the Algorithm "L" is unable to do any further shuffling), and stopping when a single pairwise permutation of branches can be performed. The respective branches are then swapped, all previous nodes already visited are reset back to their original, unpermuted state. A permutation vector is then constructed by concatenating the (now permuted) indices of the observations (leaves, at the far end of the tree), from which the permutation matrix is generated.

For sign flippings, the process is similar: the tree is swept in a similar order, but instead of computing the next permutation, the counter for a particular branch is incremented. Nodes that have reached the last possible sign flipping (i.e., with all signs reversed) are skipped; when a node can have its counter incremented, the sweeping stops, the counter is incremented by one, and all previous nodes that had been skipped are reset back to their original state. As the counter uses a numeral system with base 2, the counter itself constitutes a vector of sign flips that can be applied to the branches that begin at that node, and generating the sign flipping matrix is then trivial.

As described, the tree representation allows computing exhaustively and lexicographically all possible rearrangements. However, the tree can also be perturbed randomly, with the branches that begin in all its nodes being permuted

and/or sign flipped, a feature useful when performing only a small subset of all possible shufflings.

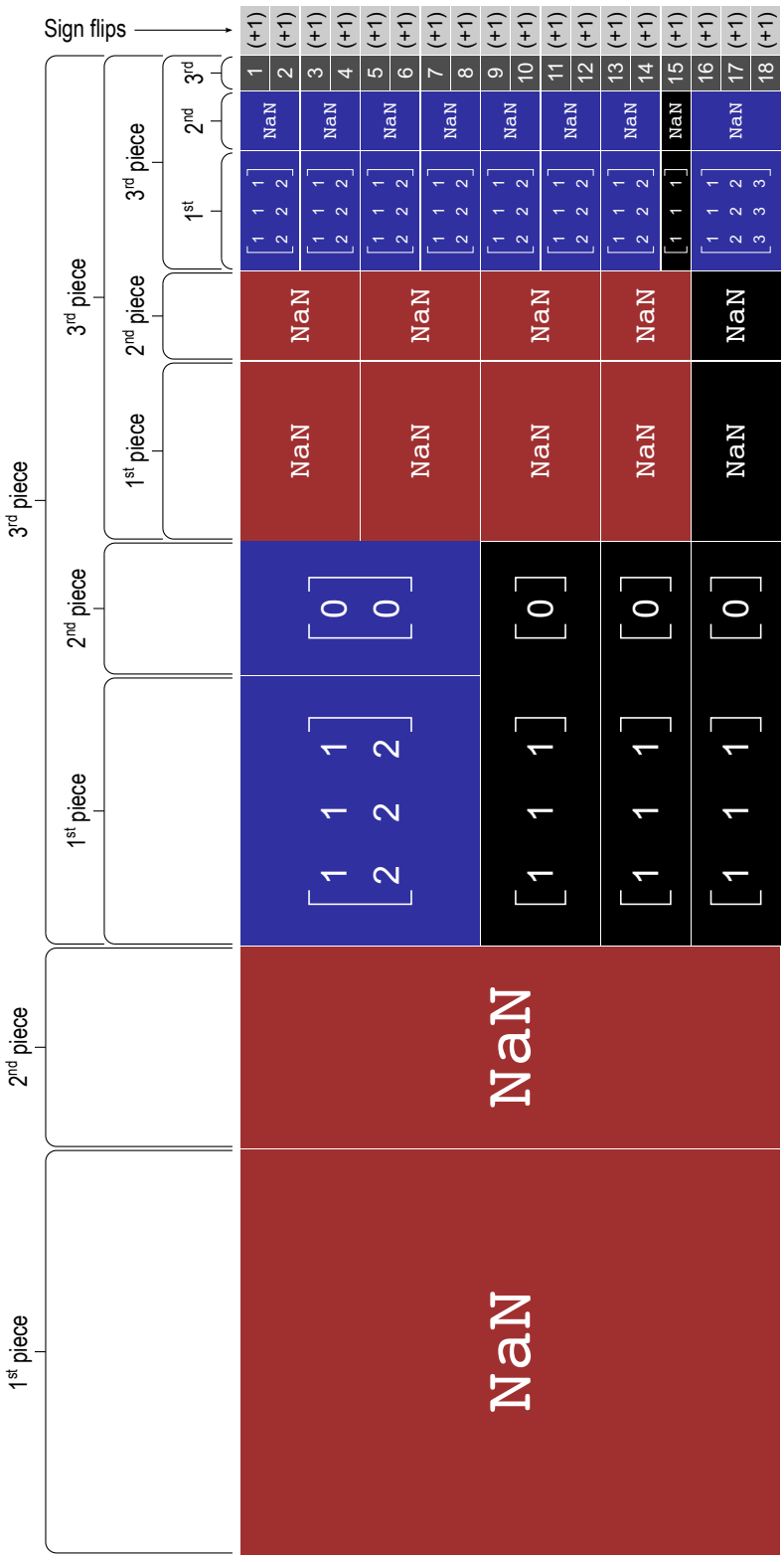
2.3.2 Variance groups

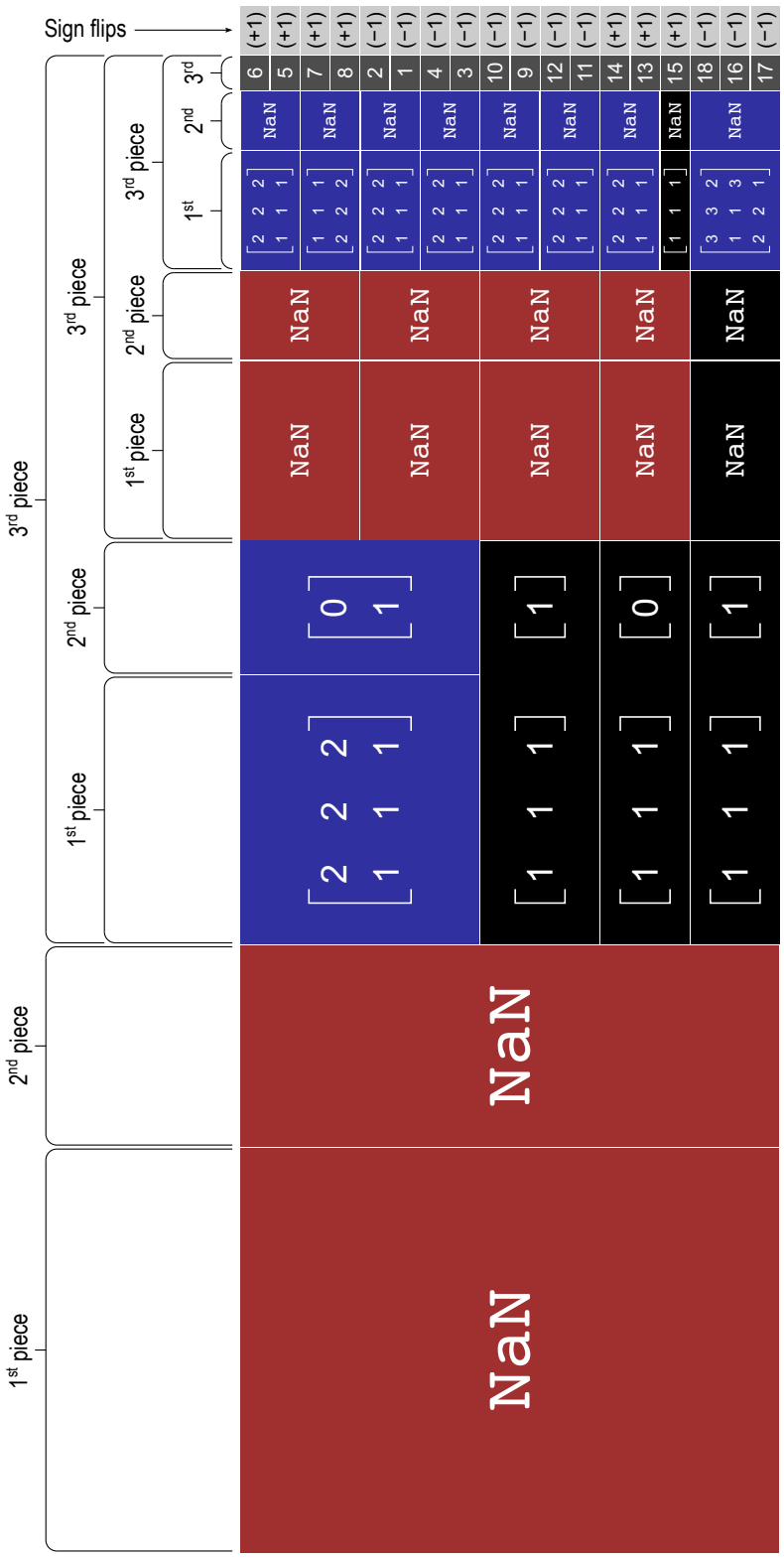
The most restrictive set of VGs can also be defined from the same tree structure. The nodes of the tree are swept from the first branch in the top node, proceeding recursively down to the next lower level before moving into the next branch in the same level. At the lowest level, observations (leaves) that can be permuted are assigned to the same variance group; those that cannot are each assigned to a distinct group. At the intermediate nodes, between the top node and the terminal leaves, those in which permutation of the branches is allowed have their corresponding VGs defined for the first branch, then replicated for all remaining branches at that level without the need to visit their lower levels. For the nodes in which permutation of branches is not allowed, each branch has its own set of VGs defined. A counter is passed down and up the levels as each node is visited, being incremented to the next integer every time a new VG is created.

Actual code for the above implementation, including the assembly of the tree, its shuffling using permutations and/or sign flippings, with complete enumeration in lexicographic order using the Algorithm "L", as well as generation of the variance groups, all written in MATLAB, is available in the tool PALM – Permutation Analysis of Linear Models; consult Section 5.2 for details.

Figure 2.5: *(page 47)* An example of a multi-level tree structure, showing the elements used in the implementation, at the first permutation, in which all the terminal leaves (observation indices) are in their original position, and no sign flippings have been performed. At each level, three "pieces" of information are present, the third one being, recursively, another such level. Compare with Figure 2.4 and the text for a complete description.

Figure 2.6: *(page 48)* The same example multi-level structure shown in Figure 2.6, after random permutations of branches and sign flippings have been performed.


Sign flips
1st piece
2nd piece
3rd piece
1st piece
2nd piece
3rd piece
1st piece
2nd piece
3rd piece
NaN
(+1)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18


Sign flips
1st piece
2nd piece
3rd piece
1st piece
2nd piece
3rd piece
1st piece
2nd piece
3rd piece
1st
2nd
3rd
NaN
(+1)
(−1)

2.4 Evaluation method

2.4.1 Error rates and power

The error rate is calculated as the proportion of tests that are significant when there is no effect. Power is calculated as the proportion of tests that are significant when a true effect is present. In either case, it is necessary to know whether true effects are present or not. Moreover, multiple realisations are necessary for the proportions to be calculated. In the simulations below, two dependence structures, named datasets A and B, were created to evaluate the permutation strategy. Both use mixtures of levels that can or that cannot be shuffled. For the dataset A, $N = 36$ observations were simulated, grouped into nine exchangeability blocks of four observations each, and each of these further divided into two blocks of two. Not all levels were allowed to be shuffled freely, and the structure is shown in Figure 2.7 (*left*). For dataset B, $N = 27$ observations were divided into nine EBs of three observations each; and each of these further divided into two blocks, one with two, and one with one observation, as shown in Figure 2.7 (*right*). Although these may appear somewhat artificial for practical use, we wanted examples that would restrict the number of possible shufflings, to test the multi-level strategy in relatively difficult scenarios. The structure in dataset A precisely emulates a twin study with nine sets of siblings, each comprised of a pair of monozygotic twins and a pair of non-twins (or of dizygotic twins). Dataset B uses a similar scheme, but further restricts the possibilities for shuffling by having just one non-twin in each set of siblings.

Using the same notation as in Section 2.2.2, 500 response variables (data vectors $\mathbf{Y}$) were simulated for each dataset, using the model $\mathbf{Y} = \mathbf{M}\boldsymbol{\psi} + \boldsymbol{\epsilon}$; each variable might represent, for instance, a voxel or vertex in a brain image. The residuals, $\boldsymbol{\epsilon}$, were simulated following either a Gaussian distribution (with zero mean and unit variance), 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),

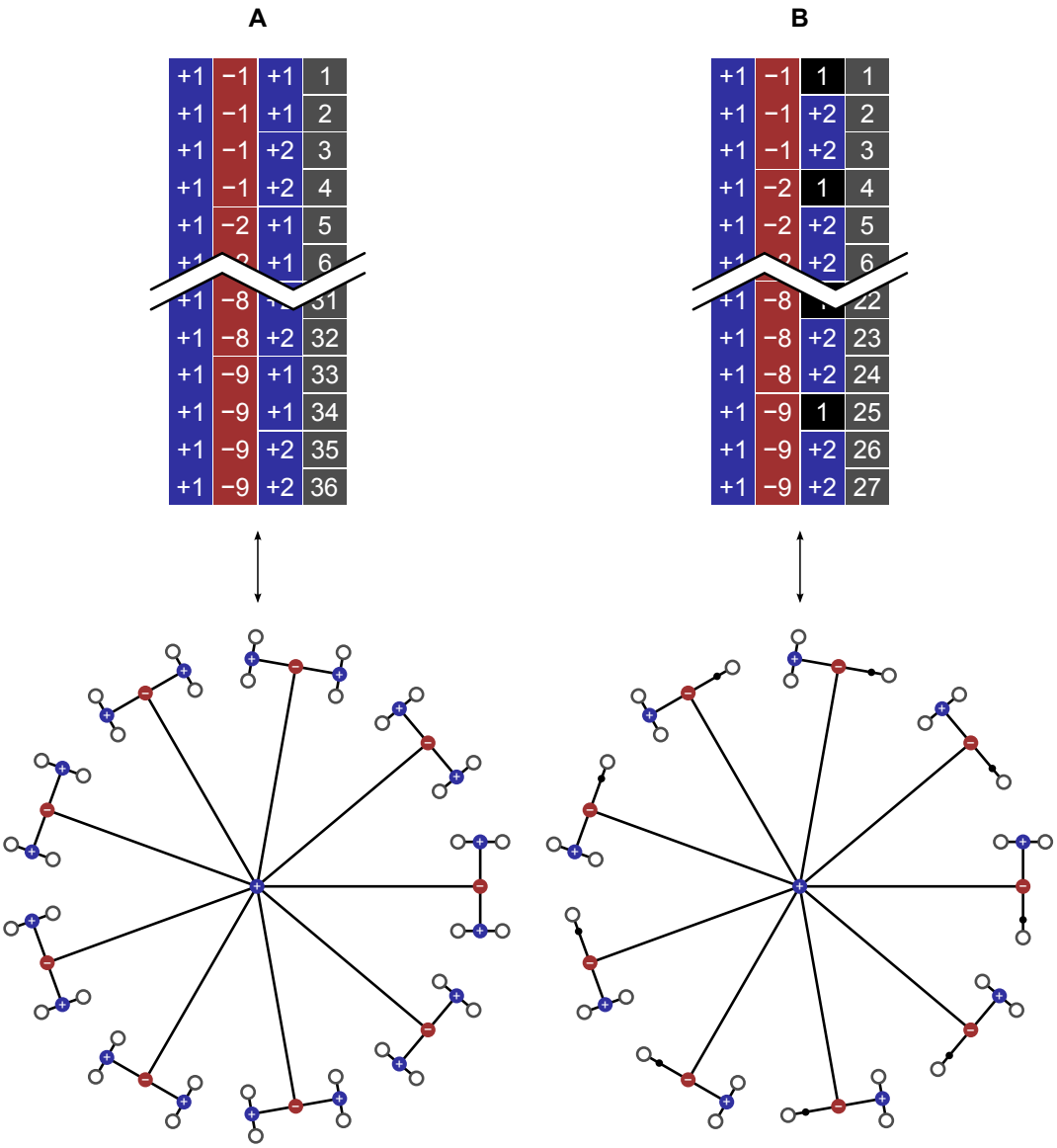


Figure 2.7: The two dependence structures, A and B, used to assess error rates and power. *Top:* Multi-level block definition. *Bottom:* Visualisation as a tree diagram.

or a Laplace distribution (kurtotic, with zero mean and unit variance)[1]. In order to introduce dependence between the residuals, for simplicity and without loss of generality to any study in which there is dependence among the data, including repeated measurements, each observation was treated as if from a participant in a twin study design, as described above, and an $N \times N$ correlation matrix $\mathbf{\Omega}$, was created using the coefficient of kinship, $2\phi_{ij}$, between subjects i and j (Jacquard, 1970), such that $\mathbf{\Omega} = 2\mathbf{\Phi}h^2_{\boldsymbol{\epsilon}} + \mathbf{I}(1 - h^2_{\boldsymbol{\epsilon}})$, where $\mathbf{\Phi}$ is the matrix with the coefficients ϕ_{ij}, and $\mathbf{I}$ is the identity matrix. The benefit of constructing the simulations in this way is that the strength of the dependence structure can vary linearly in the interval 0 to 1 using a single parameter, here denoted as $h^2_{\boldsymbol{\epsilon}}$, which coincides, in quantitative genetics and under certain assumptions, with the heritability of the measurement after explanatory or nuisance variables have been considered. The coefficient of kinship ($2\phi_{ij}$) is set to 1 for monozygotic twins, 0.5 full siblings that are not monozygotic twins, 0.25 for half siblings, and 0 for unrelated subjects. For these simulations, we used different values for the heritability of the residuals as $h^2_{\boldsymbol{\epsilon}} = \{0, 0.4, 0.8\}$. To introduce the desired correlation structure, $\mathbf{\Omega}$ was subjected to a Cholesky decomposition such that $\mathbf{\Omega} = \mathbf{L}'\mathbf{L}$, then redefining the residuals as $\mathbf{L}'\boldsymbol{\epsilon}$.

The dependent data, $\mathbf{Y}$, were generated by adding the simulated effects, $\mathbf{M}\boldsymbol{\psi}$, to the residuals, $\boldsymbol{\epsilon}$, with $\boldsymbol{\psi} = [\psi_1\ 0]'$, ψ_1 being either 0 or:

$$t^{-1}_{\text{cdf}}\left(1 - \alpha; N - \text{rank}\left(\mathbf{M}\right)\right) \Big/ \sqrt{N}$$

where $\alpha = 0.05$ is the significance level of the permutation test to be performed at a later stage, ensuring a calibrated signal strength sufficient to yield an approximate power of 50% with Gaussian errors, irrespective of the sample size. The actual

[1] The actual skewness and kurtosis for these two distributions are either fixed or functions of their parameters, and therefore were held constant throughout the simulations. For Weibull, the skewness is $\left(\Gamma\left(1 + 3/k\right)\lambda^3 - 3\mu\sigma^2 - \mu^3\right)/\sigma^3 \approx 19.58$, where k is the shape and λ the scale parameter. For Laplace, the excess kurtosis is 3.

effect was coded in the first regressor only, here denoted **m**, the second regressor modelling an intercept. This regressor was constructed as a set of random values following a Gaussian distribution with zero mean and unit variance. As in real experiments, such effects of interest may be (as with the residuals) not independent across observations, three different values for the strength of this dependence were simulated, using $h^2_{\mathbf{m}} = \{0, 0.4, 0.8\}$. These values are equivalent to the heritability of **m** in the context of genetics, yet without loss of generality to studies in which there is dependence between the data that constitute any individual independent variable, including certain designs involving repeated measurements.

Permutations, sign flippings, and permutations with sign flippings were performed, either freely or respecting the dependence structure. In each case, 500 shufflings were performed for each of the 500 variables, 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). Significance levels 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.

2.4.2 Power

The evaluations above were used to assess error rates and power according to the degree of non-independence between observations and distribution of the errors. To further investigate how the restrictions imposed by the exchangeability blocks could affect power, other dependence structures were considered to shuffle the data, in addition to the datasets A and B above; these were named C through I (Figures 2.8, 2.9, and 2.10). The configuration C corresponds to freely shuffling 11 observations; D corresponds to a small set of 5 sibships with a total of 18 subjects, mixing whole-block and within-block at different levels; E is formed by 15 observations, organised in 5 blocks of 3 observations each, with shufflings being allowed

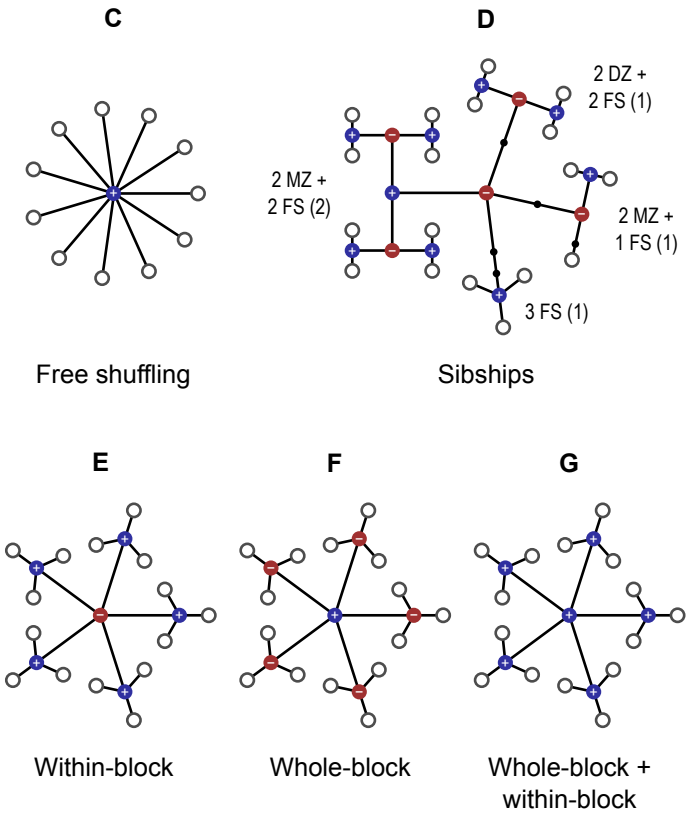


Figure 2.8: Tree diagrams C–G, used to assess power, in addition to A, B, textsch and I (shown in Figures 2.7, 2.9 and 2.10). In C, observations can be shuffled without restrictions. In D, which represent a set of five sibships, MZ refers to each subject of a pair of monozygotic twins, DZ to dizygotic twins, and FS to full siblings (non-twin and not half siblings); the numbers in parentheses indicate the number of each type of sibship in the tree (see also Figure 2.9). In E, observations can be shuffled only within-block; in F the blocks as a whole can be shuffled, and in G, shufflings are allowed within-block, and the blocks as a whole can also be shuffled.

within-block only; F is similar, but with whole-block rearrangements only, and G also similar, but allowing both whole-block and within-block simultaneously; configurations H and I use the family structure of the Human Connectome Project at the time of the HCP-S500 release (more details below): in H, dizygotic twins are treated as a category on its own, thus accounting for the possibility of shared, non-genetic effects within twin pair, whereas in I, dizygotic twins are treated as ordinary, non-twin full siblings. The number of possible permutations and sign flippings for each of these structures is shown in Table 2.1.

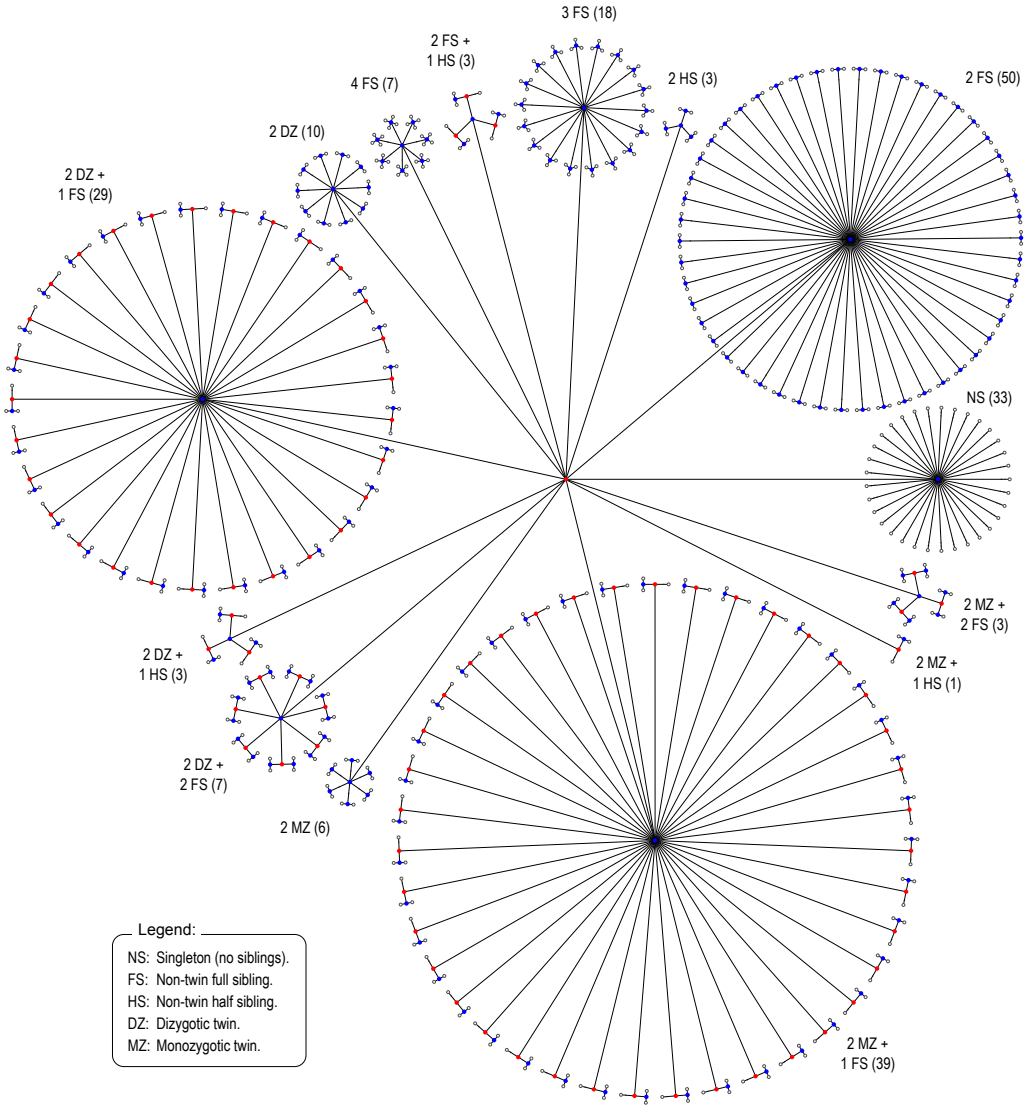


Figure 2.9: Tree diagram depicting the structure present among the subjects of the Human Connectome Project HCP, at the time of the release HCP-S500, with 518 subjects. The numbers in parentheses indicate how many of each type of sibship set are present.

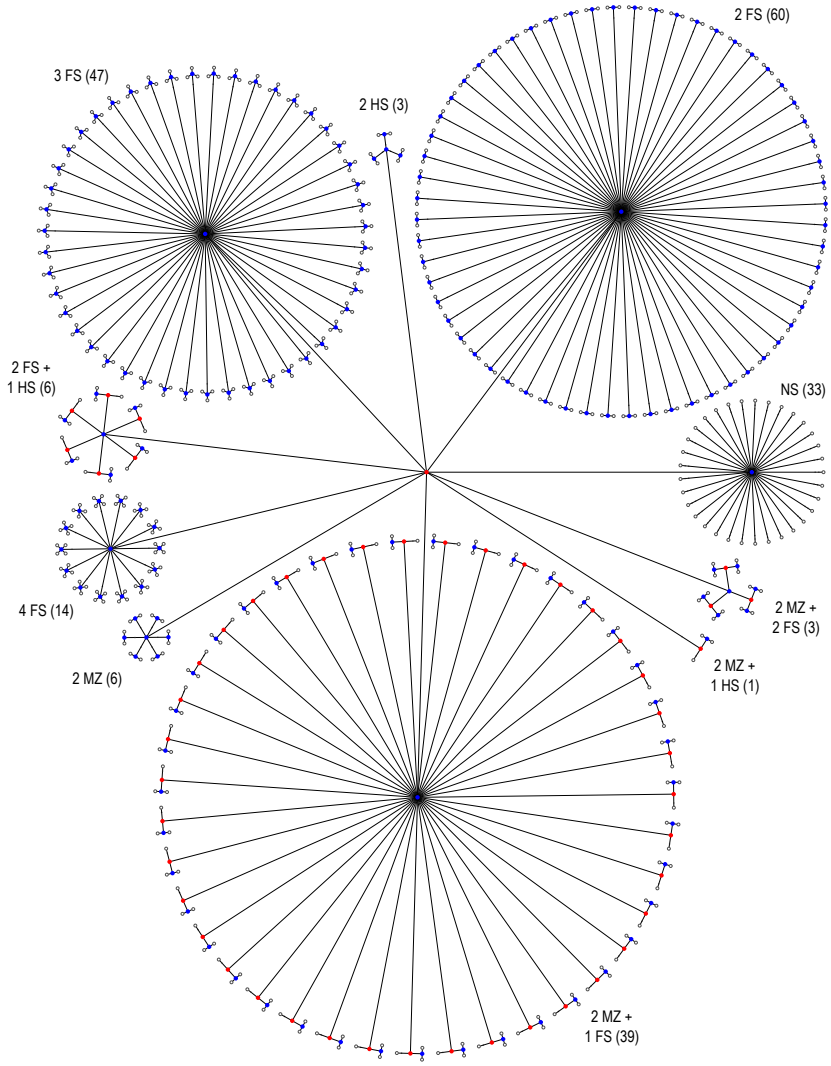


Figure 2.10: Tree diagram representing the structure among the same 518 subjects of the HCP-S500 release, shown in Figure 2.9, but treating dizygotic twins as ordinary siblings, therefore not accounting for the possibility of shared common non-genetic effects within dizygotic twin pair.

Table 2.1: Number of permutations (EE) and sign flippings (ISE) for the 9 dependence structures simulated to examine power. If there were ties in the data, the number of possible permutations would be smaller. When both EE and ISE can be used, such that the data can be permuted and sign flipped, the number of possible rearrangements is simply the product of the number of permutations with the number of sign flippings. The footnote shows in detail how these values were calculated for the more complex configurations.

	Unrestricted shuffling		Restricted shuffling	
Set	EE	ISE	EE	ISE
A	$36! \approx 3.7 \cdot 10^{41}$	$2^{36} \approx 6.9 \cdot 10^{10}$	$4^9 \cdot 9! \approx 9.5 \cdot 10^{10}$	$2^9 = 512$
B	$27! \approx 1.1 \cdot 10^{28}$	$2^{27} \approx 1.3 \cdot 10^{8}$	$2^9 \cdot 9! \approx 1.9 \cdot 10^{8}$	$2^9 = 512$
C	$11! \approx 4.0 \cdot 10^{7}$	$2^{11} = 2048$	$11! \approx 4.0 \cdot 10^{7}$	$2^{11} = 2048$
D	$18! \approx 6.4 \cdot 10^{15}$	$2^{18} = 262144$	$2^8 \cdot 3! = 1536$	$2^5 = 32$
E	$15! \approx 1.3 \cdot 10^{12}$	$2^{15} = 32768$	$(3!)^5 = 7776$	$2^{15} = 32768$
F	$15! \approx 1.3 \cdot 10^{12}$	$2^{15} = 32768$	$5! = 120$	$2^5 = 32$
G	$15! \approx 1.3 \cdot 10^{12}$	$2^{15} = 32768$	$(3!)^5 \cdot 5! = 933120$	$2^5 = 32$
H	$518! \approx 6.5 \cdot 10^{1182}$	$2^{518} \approx 8.6 \cdot 10^{155}$	$a \approx 2.9 \cdot 10^{287}$	$c \approx 6.6 \cdot 10^{63}$
I	$518! \approx 6.5 \cdot 10^{1182}$	$2^{518} \approx 8.6 \cdot 10^{155}$	$b \approx 1.3 \cdot 10^{335}$	$d \approx 6.6 \cdot 10^{63}$

$a = [33!] \cdot [2^{50} \cdot 50!] \cdot [2^3 \cdot 3!] \cdot [(3!)^{18} \cdot 18!] \cdot [2^3 \cdot 3!] \cdot [(4!)^7 \cdot 7!] \cdot [2^{10} \cdot 10!] \cdot [2^{29} \cdot 29!] \cdot [2^3 \cdot 3!] \cdot [(2^2)^7 \cdot 7!] \cdot [2^6 \cdot 6!] \cdot [2^{39} \cdot 39!] \cdot [2] \cdot [(2^2)^3 \cdot 3!]$.

$b = [33!] \cdot [2^{60} \cdot 60!] \cdot [2^3 \cdot 3!] \cdot [(3!)^{47} \cdot 47!] \cdot [2^6 \cdot 6!] \cdot [(4!)^{14} \cdot 14!] \cdot [2^6 \cdot 6!] \cdot [2^{39} \cdot 39!] \cdot [2] \cdot [(2^2)^3 \cdot 3!]$.

$c = 2^{33} \cdot 2^{50} \cdot 2^3 \cdot 2^{18} \cdot 2^3 \cdot 2^7 \cdot 2^{10} \cdot 2^{29} \cdot 2^3 \cdot 2^7 \cdot 2^6 \cdot 2^{39} \cdot 2 \cdot 2^3 = 2^{212}$.

$d = 2^{33} \cdot 2^{60} \cdot 2^3 \cdot 2^{47} \cdot 2^6 \cdot 2^{14} \cdot 2^6 \cdot 2^{39} \cdot 2 \cdot 2^3 = 2^{212}$.

Compare products a, b, c and d with Figures 2.9 and 2.10, that depict respectively the HCP structures H and I; the factors are shown starting from the singletons (labelled as NS in the figures) and running counter-clockwise around the central node.

For each of these nine datasets, an artificial effect (signal) was introduced, in the same way as described in the previous section, but here exclusively using independent Gaussian errors and preserving this independence throughout the simulations while still using multi-level exchangeability blocks for shuffling, as if dependencies among the data existed. Power was then compared with what would be observed if the same data were shuffled without the restrictions imposed by the EBS. For each configuration, 100 repetitions were performed, each simulating 1000 variables (as before, each could represent a voxel or vertex in an image). Up to 512 shufflings were used, either permutations, sign flippings, or permutations with sign flippings. Each repetition used a different set of random observations and a different set of shufflings when the maximum number of possible rearrangements was larger than the number of shufflings performed. The significance level was set as $\alpha = {}^{1}/_{16} = 0.0625$. Both the number of permutations and the significance level were chosen so as to allow compatible resolutions of the p-values among runs, allowing a more direct comparison between each case.

Power changes were assessed in relation to what would be observed if the data were shuffled freely, and compared to a measure of the amount of shuffling applied to the data, given the restrictions imposed by the permutation tree. For this purpose, Hamming distance (Hamming, 1950) was used; this distance counts the number of observations that change their position at each permutation (EE) or that change their sign at each sign flip (ISE), or both when permutations are performed together with sign flippings. While the Hamming distance cannot be interpreted as a direct quantification of perturbation on the data, it is appropriate to quantify the effect of the shufflings proper, which do not depend on actual data values.

2.4.3 Real data

The ongoing Human Connectome Project (HCP) involves the assessment of about four hundred sibships, in many cases with up to four subjects, and with at least one pair of monozygotic (MZ) or dizygotic (DZ) twins (Van Essen et al., 2012, 2013).

The inclusion of additional siblings to the classic twin design is known to improve the power to detect sources of variation in the observable traits (Posthuma and Boomsma, 2000; Keller et al., 2010). The objective is to have a genetically informative sample as in a classical twin design, enriched with the inclusion of relatives from the same nuclear family. The coefficient of kinship between MZ twins is the same for all such pairs, and so are their expected covariance. Likewise, the covariance is the same for all pairs of DZ twins. While kinship can be modelled, such modelling is contingent upon various assumptions that may not always be valid, or that can be hardly checked for all the imaging modalities and exploratory analyses that the HCP entails. Instead, such dependence structure can be represented as a tree that indicates which pieces of data can be shuffled for inference, rendering the permutation methods described this far directly applicable to the HCP data, and without the need to explicitly model the exact degree of dependence present in the data. Depending on whether there is interest in considering or not common effects in dizygotic twins, these can be treated as a category on their own, that cannot be shuffled with ordinary, non-twin siblings, or be allowed to be shuffled with them (Figures 2.9 and 2.10).

Virtually all data being collected in the HCP are to be publicly released,[2] and for this analysis, we used the set named HCP-S500, which includes various imaging and non-imaging measurements for approximately five hundred subjects. Here, measurements of height, weight, and body mass index (BMI) (Barch et al., 2013) were investigated for positive and negative associations with the cortical area and thickness as measured at each point of the cortex. These traits are well known to be highly heritable, with most studies reporting h^2 estimates that are well above 0.70, so that measurements on subjects from the same family cannot be considered independent. [For the heritabilities of height, weight and BMI, see Farooqi and O'Rahilly (2005); Visscher et al. (2006); Walley et al. (2006); Silventoinen and Kaprio

[2] Detailed information can be found at www.humanconnectome.org.

(2009); Silventoinen et al. (2012); Min et al. (2013); for cortical thickness and area, see Panizzon et al. (2009); Winkler et al. (2010); Joshi et al. (2011); Eyler et al. (2011, 2012); Kremen et al. (2013); McKay et al. (2014), among others.] To confirm the heritability of these traits specifically in the HCP sample, the variance of these traits was decomposed into genetic and environmental components using the maximum-likelihood methods described in Almasy and Blangero (1998), and as implemented in the package Sequential Oligogenic Linkage Analysis Routines – SOLAR (Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA). The released HCP data do not include an index that could directly categorise subjects according to a common environment or household. Nonetheless, ignoring these possible effects has the potential to overestimate heritabilities. To minimise this possibility, two models were tested: one in which a common environment term (c^2) was not included, and another in which a rather conservative surrogate for household effects was included; such proxy was defined by assigning all subjects sharing the same mother to a common environment. The reasoning is twofold: to account for potential maternal effects, which could affect half-siblings sharing the same mother, but not those sharing the same father, and also considering that, most commonly, children of divorced couples tend to stay or dwell with their mothers for most of the time. To ensure normality, the traits were subjected to a rank-based inverse-normal transformation before estimation.[3] The nuisance variables included in the model were age, age-squared, race and ethnicity, the interactions of these with sex, as well as sex itself. The test statistic, for either h^2 and c^2, is twice the difference between the log-likelihood of a model in which the parameter being tested is constrained to zero and the log-likelihood of a model in which that parameter is allowed to vary; this statistic is distributed as 50:50 mixture of a point mass and a χ^2 distribution with one degree of freedom (Self and Liang, 1987); here we represent this statistic (known as deviance) as $2D_{\mathrm{LL}}$. For

[3] Normality is not required for the permutation test. However, it is an assumption of the maximum likelihood method used to estimate heritability.

this analysis, 502 subjects with complete data for all these variables were selected (mean age: 29.22, standard deviation: 3.47, range 22–36 years; 296 females; 49 MZ pairs, 356 non-MZ sibling pairs, 16 half-sibling pairs).

The imaging protocol used for the structural magnetic resonance scans, as well as the steps necessary to construct the surface representation of the cortical mantle, have been described extensively in Glasser et al. (2013) (see also the references therein); FreeSurfer (Martinos Center for Biomedical Imaging, Massachussetts General Hospital, Boston, MA, USA) was used to generate the surfaces and to obtain cortical thickness measurements (Dale et al., 1999; Fischl et al., 1999a; Fischl and Dale, 2000); image registration was performed using the Multimodal Surface Matching framework (MSM) (Robinson et al., 2014). The surface area was processed using the methods described in Winkler et al. (2012, see also Section 4.2.7): the area was measured at each face of the white surface, then interpolated using a pycnophylactic method to a common grid (an icosahedron recursively subdivided five times, therefore with 10242 vertices and 20480 faces), and finally converted from facewise to vertexwise. Cortical thickness was also resampled to the same resolution, using barycentric interpolation. Both thickness and area were smoothed on the surface of a homeomorphic sphere with 100 mm radius using a Gaussian kernel with full width at half maximum (FWHM) of 20 mm. For these analyses, 5000 permutations were used, and DZ twins were considered as constituting a category on their own, and therefore not allowed to be permuted with non-twin siblings in the same family. Nuisance variables were the same used for the heritability analyses described above, plus global cortical surface area and average thickness. Visualisation of imaging results used Blender (The Blender Foundation, Amsterdam, The Netherlands). Sample statistics for the analysed traits are shown in Table 2.2.

Table 2.2: Descriptive statistics for the indices of body size and for global cortical surface area and global average thickness on the sample of subjects from the HCP.

Trait	Mean ± SD	Range
Height (m)	1.708 ± 0.096	1.473 — 1.956
Weight (kg)	77.712 ± 17.342	44.906 — 128.820
BMI (kg/m^2)	26.581 ± 5.252	16.788 — 45.171
Area (cm^2)	1666.80 ± 169.79	1292.14 — 2112.00
Thickness (mm)	2.620 ± 0.087	2.239 — 2.824

2.5 Results

2.5.1 Error rates and power

Despite the differences in the relationship between the observations that constituted datasets A and B, the results were very similar. With errors that were independent and symmetric, i.e., either normally distributed (Gaussian) or kurtotic (Laplacian), the false positive rates (error type I) were controlled at the nominal level ($\alpha = 0.05$) using unrestricted permutations, sign flippings, or permutations with sign flippings, whenever there were no true dependence between observations or elements of the regressor of interest, that is, when either $h^2_{\boldsymbol{\epsilon}}$ or $h^2_{\mathbf{m}}$ were equal to zero. With both $h^2_{\boldsymbol{\epsilon}}$ and $h^2_{\mathbf{m}}$ higher than zero, however, the conventional test in which the data are shuffled freely became, as expected, invalid. Using instead the shuffling strategy that we propose, that respects the covariance structure present or assumed to exist in the data, the false positive rates were controlled at α, even when the dependence was at high levels. These results are shown in Table 2.3 (Gaussian) and in the Table 2.4 (Laplacian).

With skewed (Weibullian) errors, sign flippings were generally conservative when $h^2_{\boldsymbol{\epsilon}}$ or $h^2_{\mathbf{m}}$ were equal to zero and the data were shuffled freely. With $h^2_{\boldsymbol{\epsilon}}$ and $h^2_{\mathbf{m}}$ higher than zero, the test not only reversed its conservativeness, but became invalid if flippings ignored the data structure. If, however, the shufflings were performed respecting the restrictions imposed by the relationships among the data-

points, the test was valid in all cases, with its conservativeness maintained. These results are shown in the Table 2.5.

These Tables also show the power of each shuffling strategy when there is true signal present. For the cases in which the false positive rate is not controlled, the test is invalid, and as a consequence, considerations of power are irrelevant; in these cases, the values that would represent power are shown crossed by a line. When the data are truly independent, hence unrestricted shuffling could be performed, the proposed restricted permutations caused a slight, yet consistent, loss of power for the datasets A and B. This is revisited in the next section, with the other synthetic datasets. Histograms of p-values using permutations, sign flippings, and permutations with sign flippings, for cases of normal and skewed distributions, and using both unrestricted and restricted shuffling, for dataset A, are shown in Figure 2.11 (the pattern is similar for dataset B). These extend the results shown in Tables 2.3, 2.4 and 2.5, with an overview of the frequencies of p-values throughout the whole $[0\ 1]$ interval. Except for the use of the ISE with skewed errors (assumptions violated), in general the use of restricted shuffling ensured that the histograms were all flat, as desirable, with no excesses of p-values at any range. Even for ISE with skewed errors, a test that otherwise would be invalid, became valid on average and therefore useful in practice, although conservative. Power changes, though slight, are visible. Bland–Altman plots shown in the Figure 2.12 reveal that, even in the absence of dependence for data and for design, and without any simulated effect, the p-values for each test are not identical for unrestricted versus restricted shuffling, with the differences falling well outside the confidence interval for a much larger fraction of tests than the 5% that would be expected by chance.

2.5.2 Power

For the nine synthetic datasets, a slight, yet consistent loss of power was observed when using the proposed restricted shuffling strategy, compared to the results us-

Table 2.3: Proportion of error type I and power (%) for the simulated sets A and B, with **Gaussian errors**, at the level $\alpha = 0.05$, using different degrees of dependence for the error terms ($h^2_{\boldsymbol{\epsilon}}$) and for the regressor of interest ($h^2_{\mathbf{m}}$), using permutations (EE), sign flippings (ISE), or permutations with sign flippings (EE+ISE). Confidence intervals (95%) are shown between parentheses. The values that appear ~~striked out~~ are not valid, as they refer to power observed when the corresponding error rates are not controlled (i.e., the lower bound of the confidence interval is above the nominal level α when there is no actual effect).

		Unrestricted shuffling						Restricted shuffling					
Set	$h^2_{\boldsymbol{\epsilon}}$	Without effect (error rate)			With effect (power)			Without effect (error rate)			With effect (power)		
		$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$
Permutations only:													
A	0.0	5.0 (3.4–7.3)	4.9 (3.3–7.2)	5.1 (3.5–7.3)	49.1 (44.7–53.5)	47.4 (43.1–51.8)	46.5 (42.2–50.9)	5.0 (3.4–7.3)	4.9 (3.3–7.2)	5.1 (3.5–7.4)	47.6 (43.3–52.0)	46.1 (41.7–50.5)	44.3 (40.0–48.7)
	0.4	5.0 (3.4–7.3)	6.4 (4.5–8.9)	7.8 (5.7–10.5)	49.8 (45.4–54.2)	49.8 (45.4–54.1)	~~48.5 (44.2–52.9)~~	5.1 (3.5–7.3)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	48.5 (44.1–52.9)	44.1 (39.8–48.5)	38.6 (34.4–42.9)
	0.8	4.9 (3.4–7.2)	7.8 (5.8–10.5)	10.4 (8.0–13.3)	51.5 (47.1–55.8)	~~50.5 (46.2–54.9)~~	~~49.7 (45.3–54.1)~~	4.9 (3.3–7.1)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	50.6 (46.2–54.9)	41.9 (37.6–46.2)	34.3 (30.2–38.5)
B	0.0	5.0 (3.4–7.2)	5.0 (3.4–7.2)	4.9 (3.4–7.2)	48.4 (44.1–52.8)	47.4 (43.0–51.7)	46.5 (42.1–50.8)	4.9 (3.3–7.2)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	46.7 (42.4–51.1)	45.4 (41.1–49.8)	43.8 (39.5–48.2)
	0.4	5.0 (3.4–7.3)	6.2 (4.4–8.7)	7.6 (5.6–10.2)	49.6 (45.2–53.9)	49.2 (44.8–53.6)	~~48.2 (43.9–52.6)~~	5.0 (3.4–7.2)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	48.3 (43.9–52.7)	43.5 (39.2–47.8)	38.1 (34.0–42.5)
	0.8	5.0 (3.4–7.3)	7.4 (5.4–10.0)	10.0 (7.7–12.9)	50.4 (46.0–54.8)	~~50.3 (45.9–54.6)~~	~~49.2 (44.8–53.6)~~	5.0 (3.4–7.3)	4.9 (3.3–7.2)	5.0 (3.4–7.3)	50.0 (45.6–54.4)	41.7 (37.5–46.1)	33.7 (29.7–37.9)
Sign flippings only:													
A	0.0	5.1 (3.5–7.4)	5.0 (3.4–7.3)	4.9 (3.3–7.2)	45.6 (41.3–50.0)	45.6 (41.3–50.0)	45.1 (40.8–49.5)	5.0 (3.4–7.3)	5.1 (3.5–7.4)	5.2 (3.6–7.5)	41.5 (37.2–45.9)	41.7 (37.5–46.1)	40.9 (36.7–45.3)
	0.4	5.0 (3.4–7.2)	6.2 (4.4–8.6)	7.7 (5.7–10.4)	47.3 (42.9–51.6)	47.1 (42.7–51.5)	~~46.0 (41.7–50.4)~~	4.9 (3.3–7.1)	5.0 (3.4–7.3)	5.2 (3.6–7.5)	43.0 (38.8–47.4)	39.2 (35.0–43.6)	34.8 (30.8–39.1)
	0.8	5.1 (3.5–7.4)	7.6 (5.6–10.3)	10.7 (8.3–13.7)	48.5 (44.1–52.9)	~~48.3 (43.9–52.7)~~	~~48.6 (44.2–52.9)~~	4.9 (3.3–7.1)	5.0 (3.4–7.3)	5.2 (3.6–7.5)	45.2 (40.8–49.5)	37.6 (33.5–41.9)	31.6 (27.7–35.9)
B	0.0	5.0 (3.4–7.2)	4.9 (3.4–7.2)	5.0 (3.4–7.2)	45.1 (40.8–49.5)	44.3 (40.0–48.7)	43.8 (39.6–48.2)	4.9 (3.4–7.2)	5.1 (3.5–7.4)	5.3 (3.6–7.6)	41.5 (37.3–45.9)	40.9 (36.7–45.3)	39.9 (35.7–44.2)
	0.4	5.0 (3.4–7.3)	6.3 (4.4–8.7)	7.4 (5.4–10.0)	46.3 (42.0–50.7)	45.3 (41.0–49.7)	~~46.2 (41.9–50.6)~~	4.9 (3.3–7.2)	5.1 (3.5–7.4)	5.1 (3.5–7.4)	42.4 (38.2–46.8)	38.3 (34.2–42.7)	35.2 (31.2–39.5)
	0.8	5.1 (3.5–7.3)	7.6 (5.6–10.3)	10.1 (7.8–13.1)	47.5 (43.2–51.9)	~~47.2 (42.8–51.5)~~	~~47.1 (42.8–51.5)~~	4.8 (3.2–7.0)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	44.2 (39.9–48.6)	37.3 (33.2–41.6)	30.8 (26.9–34.9)
Permutations + sign flippings:													
A	0.0	5.1 (3.5–7.4)	5.0 (3.4–7.2)	5.0 (3.4–7.2)	48.6 (44.3–53.0)	48.6 (44.3–53.0)	46.8 (42.5–51.2)	5.1 (3.5–7.4)	5.1 (3.5–7.4)	5.3 (3.6–7.6)	48.4 (44.0–52.7)	48.8 (44.4–53.1)	46.9 (42.6–51.3)
	0.4	5.0 (3.4–7.3)	6.4 (4.6–8.9)	7.7 (5.7–10.4)	49.7 (45.3–54.0)	48.6 (44.2–52.9)	~~48.0 (43.6–52.4)~~	5.0 (3.4–7.2)	5.1 (3.5–7.4)	5.4 (3.7–7.7)	49.7 (45.4–54.1)	44.5 (40.2–48.9)	40.2 (36.0–44.6)
	0.8	5.0 (3.4–7.2)	7.7 (5.7–10.4)	10.5 (8.1–13.5)	51.3 (46.9–55.6)	~~50.3 (45.9–54.6)~~	~~50.2 (45.8–54.5)~~	4.9 (3.3–7.2)	5.1 (3.5–7.4)	5.4 (3.7–7.7)	52.1 (47.7–56.5)	43.2 (38.9–47.6)	36.3 (32.2–40.6)
B	0.0	5.1 (3.5–7.4)	5.1 (3.5–7.3)	5.0 (3.4–7.2)	48.8 (44.4–53.1)	48.2 (43.8–52.5)	47.1 (42.8–51.5)	5.1 (3.5–7.4)	5.2 (3.6–7.5)	5.3 (3.7–7.7)	48.3 (44.0–52.7)	48.1 (43.7–52.4)	47.0 (42.7–51.4)
	0.4	5.1 (3.5–7.4)	6.2 (4.4–8.7)	7.5 (5.5–10.2)	49.2 (44.8–53.5)	49.5 (45.1–53.8)	~~48.2 (43.9–52.6)~~	5.0 (3.4–7.3)	5.2 (3.5–7.5)	5.3 (3.7–7.7)	49.0 (44.6–53.4)	45.8 (41.5–50.2)	40.6 (36.4–44.9)
	0.8	5.0 (3.4–7.2)	7.6 (5.6–10.3)	10.2 (7.8–13.2)	50.3 (45.9–54.7)	~~50.3 (46.0–54.7)~~	~~50.0 (45.6–54.4)~~	5.0 (3.4–7.2)	5.2 (3.5–7.5)	5.2 (3.6–7.5)	51.2 (46.8–55.5)	43.6 (39.3–47.9)	36.6 (32.5–40.9)

Table 2.4: Proportion of error type I and power (%) for the simulated sets A and B, with **Laplacian (kurtotic) errors**, at the level $\alpha = 0.05$, using different degrees of dependence for the error terms ($h^2_{\boldsymbol{\epsilon}}$) and for the regressor of interest ($h^2_{\mathbf{m}}$), using permutations (EE), sign flippings (ISE), or permutations with sign flippings (EE+ISE). Confidence intervals (95%) are shown between parentheses. The values that appear ~~striked out~~ are not valid, as they refer to power observed when the corresponding error rates are not controlled (i.e., the lower bound of the confidence interval is above the nominal level α when there is no actual effect).

		Unrestricted shuffling						Restricted shuffling					
		Without effect (error rate)			With effect (power)			Without effect (error rate)			With effect (power)		
Set	$h^2_{\boldsymbol{\epsilon}}$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$
Permutations only:													
A	0.0	5.0 (3.4–7.3)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	50.5 (46.1–54.8)	50.0 (45.6–54.3)	49.0 (44.6–53.3)	5.0 (3.4–7.3)	5.0 (3.4–7.2)	5.0 (3.4–7.3)	49.3 (44.9–53.7)	48.6 (44.2–53.0)	46.8 (42.4–51.1)
	0.4	4.9 (3.4–7.2)	6.3 (4.5–8.8)	7.7 (5.7–10.4)	51.7 (47.3–56.1)	50.7 (46.4–55.1)	~~49.8 (45.4–54.1)~~	5.0 (3.4–7.2)	4.9 (3.4–7.2)	5.0 (3.4–7.2)	50.7 (46.3–55.1)	45.3 (40.9–49.6)	39.9 (35.7–44.3)
	0.8	5.0 (3.4–7.2)	7.6 (5.6–10.3)	10.3 (8.0–13.3)	52.9 (48.5–57.3)	~~51.9 (47.5–56.2)~~	~~51.3 (46.9–55.7)~~	5.0 (3.4–7.3)	5.0 (3.4–7.3)	4.9 (3.4–7.2)	52.8 (48.5–57.2)	43.8 (39.5–48.2)	36.3 (32.2–40.6)
B	0.0	5.0 (3.4–7.2)	5.0 (3.4–7.3)	5.0 (3.4–7.2)	50.4 (46.0–54.7)	49.4 (45.1–53.8)	48.1 (43.7–52.4)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	5.1 (3.5–7.4)	48.5 (44.2–52.9)	47.5 (43.2–51.9)	45.3 (41.0–49.7)
	0.4	4.9 (3.3–7.2)	6.2 (4.4–8.7)	7.5 (5.5–10.1)	51.9 (47.6–56.3)	50.8 (46.4–55.2)	~~50.0 (45.6–54.3)~~	5.0 (3.4–7.3)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	50.4 (46.0–54.7)	45.4 (41.1–49.7)	40.0 (35.8–44.3)
	0.8	4.9 (3.3–7.1)	7.3 (5.4–10.0)	9.8 (7.5–12.8)	53.9 (49.5–58.2)	~~52.4 (48.0–56.7)~~	~~51.8 (47.4–56.1)~~	5.0 (3.4–7.3)	5.0 (3.4–7.3)	4.9 (3.3–7.1)	53.1 (48.8–57.5)	44.6 (40.3–48.9)	36.9 (32.8–41.2)
Sign flippings only:													
A	0.0	5.0 (3.4–7.2)	5.0 (3.4–7.3)	5.0 (3.4–7.2)	49.7 (45.4–54.1)	49.0 (44.6–53.4)	47.8 (43.5–52.2)	5.0 (3.4–7.3)	5.1 (3.5–7.4)	5.2 (3.6–7.5)	45.8 (41.5–50.2)	45.2 (40.8–49.5)	43.4 (39.1–47.8)
	0.4	4.9 (3.4–7.2)	6.4 (4.6–8.9)	7.8 (5.8–10.5)	50.6 (46.2–55.0)	50.5 (46.2–54.9)	~~49.2 (44.9–53.6)~~	4.9 (3.3–7.2)	5.0 (3.4–7.2)	5.1 (3.5–7.4)	46.7 (42.4–51.1)	43.1 (38.8–47.4)	38.0 (33.9–42.4)
	0.8	4.9 (3.3–7.2)	7.6 (5.6–10.3)	10.5 (8.1–13.5)	52.1 (47.7–56.4)	~~51.4 (47.1–55.8)~~	~~50.6 (46.2–55.0)~~	4.8 (3.3–7.1)	5.0 (3.4–7.3)	5.1 (3.5–7.4)	49.3 (44.9–53.7)	42.1 (37.8–46.5)	34.9 (30.9–39.2)
B	0.0	4.9 (3.3–7.1)	5.0 (3.4–7.2)	4.9 (3.3–7.2)	48.9 (44.5–53.2)	48.8 (44.4–53.1)	47.9 (43.6–52.3)	4.9 (3.3–7.2)	5.1 (3.5–7.4)	5.2 (3.6–7.5)	45.2 (40.9–49.6)	45.2 (40.9–49.6)	43.8 (39.5–48.1)
	0.4	5.0 (3.4–7.3)	6.2 (4.4–8.7)	7.6 (5.6–10.3)	49.4 (45.1–53.8)	49.9 (45.5–54.3)	~~49.8 (45.4–54.1)~~	4.9 (3.3–7.1)	5.0 (3.4–7.3)	5.2 (3.5–7.5)	45.8 (41.5–50.2)	43.4 (39.2–47.8)	38.7 (34.6–43.1)
	0.8	5.0 (3.4–7.3)	7.2 (5.2–9.8)	10.0 (7.7–13.0)	50.8 (46.4–55.2)	~~51.5 (47.1–55.8)~~	~~50.6 (46.3–55.0)~~	4.7 (3.2–7.0)	4.9 (3.3–7.1)	4.9 (3.4–7.2)	48.1 (43.7–52.5)	42.2 (38.0–46.6)	35.4 (31.4–39.7)
Permutations + sign flippings:													
A	0.0	4.9 (3.4–7.2)	5.0 (3.4–7.2)	5.1 (3.5–7.3)	50.6 (46.2–55.0)	49.3 (44.9–53.7)	48.5 (44.2–52.9)	4.9 (3.3–7.2)	5.1 (3.5–7.4)	5.3 (3.7–7.7)	50.3 (46.0–54.7)	49.4 (45.0–53.8)	48.6 (44.2–53.0)
	0.4	5.0 (3.4–7.3)	6.3 (4.5–8.8)	7.6 (5.6–10.3)	51.6 (47.2–55.9)	50.3 (46.0–54.7)	~~49.5 (45.2–53.9)~~	4.9 (3.4–7.2)	5.2 (3.6–7.5)	5.2 (3.6–7.6)	51.5 (47.1–55.8)	46.5 (42.1–50.8)	41.9 (37.6–46.2)
	0.8	5.0 (3.4–7.2)	7.9 (5.8–10.6)	10.2 (7.9–13.2)	52.8 (48.4–57.1)	~~52.9 (48.5–57.2)~~	~~51.9 (47.5–56.2)~~	5.0 (3.4–7.3)	5.2 (3.5–7.5)	5.3 (3.7–7.6)	53.2 (48.8–57.5)	45.7 (41.4–50.1)	38.6 (34.5–43.0)
B	0.0	5.1 (3.5–7.4)	5.0 (3.4–7.2)	4.9 (3.4–7.2)	51.1 (46.7–55.4)	50.3 (46.0–54.7)	48.4 (44.1–52.8)	5.0 (3.4–7.3)	5.2 (3.5–7.5)	5.3 (3.6–7.6)	50.8 (46.4–55.2)	50.4 (46.1–54.8)	48.4 (44.0–52.7)
	0.4	4.9 (3.3–7.2)	6.2 (4.4–8.6)	7.5 (5.5–10.2)	51.7 (47.3–56.0)	51.5 (47.1–55.8)	~~49.9 (45.5–54.3)~~	4.9 (3.3–7.1)	5.2 (3.6–7.5)	5.3 (3.7–7.6)	51.5 (47.1–55.9)	48.0 (43.7–52.4)	42.4 (38.2–46.8)
	0.8	4.9 (3.3–7.1)	7.4 (5.4–10.1)	10.1 (7.8–13.1)	53.5 (49.2–57.9)	~~52.1 (47.7–56.4)~~	~~51.4 (47.1–55.8)~~	5.0 (3.4–7.3)	5.2 (3.5–7.5)	5.4 (3.7–7.7)	53.8 (49.5–58.2)	45.9 (41.6–50.3)	38.6 (34.5–43.0)

Table 2.5: Proportion of error type I and power (%) for the simulated sets A and B, with **Weibullian (skewed) errors**, at the level $\alpha = 0.05$, using different degrees of dependence for the error terms ($h^2_{\boldsymbol{\epsilon}}$) and for the regressor of interest ($h^2_{\mathbf{m}}$), using permutations (EE), sign flippings (ISE), or permutations with sign flippings (EE+ISE). Confidence intervals (95%) are shown between parentheses. The values that appear ~~striked out~~ are not valid, as they refer to power observed when the corresponding error rates are not controlled (i.e., the lower bound of the confidence interval is above the nominal level α when there is no actual effect).

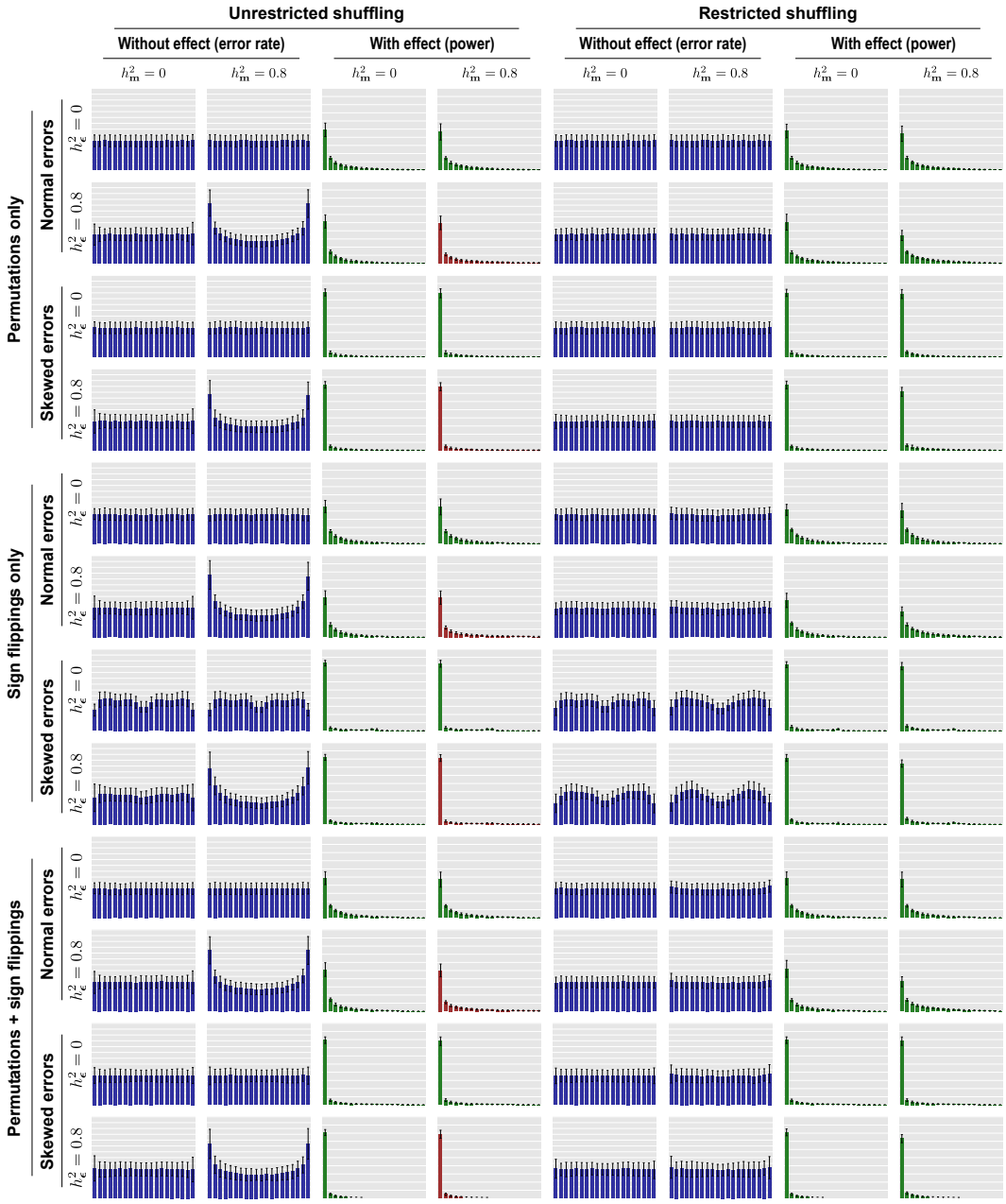
		Unrestricted shuffling						Restricted shuffling					
		Without effect (error rate)			With effect (power)			Without effect (error rate)			With effect (power)		
Set	$h^2_{\boldsymbol{\epsilon}}$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$	$h^2_{\mathbf{m}}=0$	$h^2_{\mathbf{m}}=0.4$	$h^2_{\mathbf{m}}=0.8$
Permutations only:													
A	0.0	5.0 (3.4–7.3)	5.0 (3.4–7.3)	4.9 (3.4–7.2)	79.1 (75.3–82.4)	78.3 (74.5–81.7)	78.1 (74.2–81.5)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	5.0 (3.4–7.2)	78.5 (74.7–81.9)	77.5 (73.7–81.0)	77.1 (73.3–80.6)
	0.4	5.1 (3.5–7.3)	6.3 (4.5–8.8)	7.7 (5.7–10.4)	79.4 (75.6–82.7)	78.6 (74.8–82.0)	~~77.7 (73.9–81.1)~~	4.9 (3.4–7.2)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	78.9 (75.1–82.3)	76.3 (72.4–79.8)	73.3 (69.3–77.0)
	0.8	5.0 (3.4–7.3)	7.2 (5.2–9.8)	9.6 (7.3–12.5)	80.4 (76.7–83.7)	~~79.8 (76.0–83.1)~~	~~78.1 (74.3–81.5)~~	5.0 (3.4–7.3)	5.1 (3.5–7.4)	5.0 (3.4–7.3)	80.3 (76.6–83.5)	76.7 (72.8–80.2)	71.9 (67.8–75.6)
B	0.0	4.9 (3.3–7.2)	5.0 (3.4–7.2)	4.9 (3.3–7.1)	80.6 (76.9–83.8)	80.1 (76.3–83.3)	79.3 (75.5–82.6)	5.0 (3.4–7.3)	4.9 (3.4–7.2)	4.9 (3.4–7.2)	80.1 (76.4–83.4)	79.3 (75.5–82.6)	78.3 (74.5–81.7)
	0.4	5.0 (3.4–7.2)	6.2 (4.4–8.7)	7.5 (5.5–10.2)	81.2 (77.5–84.4)	80.7 (77.0–83.9)	~~79.5 (75.7–82.8)~~	4.9 (3.3–7.2)	5.0 (3.4–7.2)	5.0 (3.4–7.3)	80.6 (76.9–83.8)	78.7 (74.9–82.1)	75.5 (71.6–79.1)
	0.8	5.1 (3.5–7.4)	7.2 (5.2–9.8)	9.1 (6.9–12.0)	81.9 (78.3–85.1)	~~81.4 (77.7–84.5)~~	~~80.6 (76.9–83.8)~~	5.0 (3.4–7.3)	5.0 (3.4–7.3)	4.9 (3.4–7.2)	81.7 (78.1–84.8)	78.7 (74.9–82.1)	75.2 (71.2–78.8)
Sign flippings only:													
A	0.0	3.5 (2.2–5.6)	3.6 (2.3–5.6)	3.6 (2.3–5.6)	82.9 (79.3–85.9)	82.7 (79.2–85.8)	82.2 (78.6–85.3)	3.8 (2.5–5.9)	4.0 (2.6–6.0)	4.0 (2.6–6.1)	81.0 (77.4–84.2)	80.6 (76.9–83.8)	79.2 (75.4–82.5)
	0.4	4.2 (2.7–6.3)	5.9 (4.2–8.3)	7.8 (5.8–10.5)	82.3 (78.7–85.4)	82.1 (78.5–85.2)	~~81.3 (77.7–84.5)~~	3.7 (2.3–5.7)	3.7 (2.4–5.8)	3.8 (2.4–5.8)	81.0 (77.3–84.2)	78.9 (75.1–82.2)	75.7 (71.7–79.2)
	0.8	4.5 (3.0–6.7)	7.1 (5.1–9.6)	9.6 (7.3–12.5)	82.0 (78.4–85.2)	~~81.7 (78.0–84.8)~~	~~81.2 (77.6–84.4)~~	3.6 (2.3–5.6)	3.7 (2.4–5.7)	3.7 (2.3–5.7)	81.4 (77.7–84.5)	78.4 (74.6–81.8)	74.5 (70.5–78.1)
B	0.0	3.2 (2.0–5.1)	3.1 (1.9–5.1)	3.2 (1.9–5.1)	83.7 (80.2–86.7)	83.3 (79.8–86.3)	83.0 (79.5–86.1)	3.5 (2.2–5.5)	3.7 (2.3–5.7)	3.6 (2.3–5.6)	82.1 (78.5–85.3)	81.5 (77.8–84.6)	80.2 (76.5–83.4)
	0.4	3.7 (2.4–5.8)	5.3 (3.7–7.6)	6.6 (4.7–9.1)	83.2 (79.7–86.2)	82.7 (79.2–85.8)	82.2 (78.6–85.3)	3.4 (2.2–5.4)	3.5 (2.2–5.4)	3.5 (2.2–5.4)	82.0 (78.4–85.2)	80.0 (76.3–83.3)	77.0 (73.1–80.5)
	0.8	4.0 (2.6–6.1)	5.9 (4.2–8.4)	8.3 (6.2–11.0)	83.3 (79.8–86.3)	83.0 (79.5–86.0)	~~82.4 (78.9–85.5)~~	3.3 (2.1–5.3)	3.2 (2.0–5.1)	3.3 (2.1–5.3)	82.7 (79.2–85.8)	79.9 (76.2–83.2)	76.1 (72.1–79.6)
Permutations + sign flippings:													
A	0.0	5.0 (3.4–7.3)	5.0 (3.4–7.3)	5.0 (3.4–7.2)	79.3 (75.5–82.6)	78.5 (74.7–81.9)	78.0 (74.2–81.4)	5.0 (3.4–7.3)	5.2 (3.6–7.5)	5.3 (3.7–7.6)	79.3 (75.5–82.6)	78.7 (74.9–82.1)	78.1 (74.3–81.5)
	0.4	4.8 (3.3–7.1)	6.1 (4.4–8.6)	7.7 (5.7–10.3)	79.3 (75.6–82.6)	78.7 (74.9–82.0)	~~77.9 (74.1–81.4)~~	4.9 (3.3–7.1)	5.0 (3.4–7.3)	5.3 (3.7–7.7)	79.4 (75.7–82.7)	77.1 (73.2–80.5)	74.7 (70.7–78.3)
	0.8	5.1 (3.5–7.5)	7.3 (5.3–9.9)	9.3 (7.1–12.2)	80.7 (77.0–83.9)	~~79.9 (76.1–83.1)~~	~~78.6 (74.8–82.0)~~	5.1 (3.5–7.4)	5.3 (3.6–7.6)	5.3 (3.7–7.7)	80.9 (77.3–84.1)	77.2 (73.3–80.6)	73.2 (69.2–76.9)
B	0.0	4.9 (3.4–7.2)	5.0 (3.4–7.3)	5.0 (3.4–7.3)	80.5 (76.8–83.8)	80.3 (76.6–83.6)	79.5 (75.8–82.8)	4.9 (3.4–7.2)	5.2 (3.6–7.5)	5.4 (3.7–7.8)	80.5 (76.8–83.7)	80.3 (76.6–83.6)	79.4 (75.7–82.7)
	0.4	5.0 (3.4–7.3)	6.2 (4.4–8.6)	7.4 (5.4–10.0)	81.1 (77.4–84.3)	80.5 (76.8–83.7)	~~79.6 (75.8–82.9)~~	5.0 (3.4–7.2)	5.0 (3.4–7.3)	5.3 (3.7–7.7)	81.2 (77.5–84.4)	79.2 (75.4–82.5)	76.6 (72.6–80.1)
	0.8	5.1 (3.5–7.4)	7.1 (5.2–9.7)	9.0 (6.8–11.9)	82.0 (78.3–85.1)	~~81.5 (77.9–84.7)~~	~~80.4 (76.7–83.7)~~	4.9 (3.4–7.2)	5.1 (3.5–7.4)	5.3 (3.6–7.6)	82.1 (78.5–85.2)	79.3 (75.6–82.7)	75.7 (71.8–79.3)

ing unrestricted shuffling when the last was, in fact, possible. These results are shown in Table 2.6. The loss appears to be larger for the datasets with more involved dependence structures (e.g., dataset D), or when restrictions on permutations are imposed at higher levels (e.g., dataset E), or on sign flippings at lower levels (e.g., datasets F and G). Even so, this is not quite as conspicuous with samples that are just modestly larger (e.g., A and B), or much larger (e.g., datasets H and I, that use the data structure from the HCP).

With exchangeable errors, in which only permutations are performed, power reductions were more noticeable for some datasets and related well to how the data could be disarranged at each permutation, as quantified by the average Hamming

Figure 2.11: *(page 67)* Pictorial table showing the histograms of p-values in different settings for the dataset A. In each histogram, the horizontal axis contains p-values in the range 0 to 1, split into 20 bins, while the vertical axis contains the relative frequencies. For the error rates, the vertical axis in the range 0 to 14%, in steps of 1%, and for power, in the range 0 to 100%, in steps of 10%. The error bars indicate one standard deviation on the height of each bar after 500 repetitions. To facilitate viewing, the bars for the error rates are shown in blue; for power, in green when the respective error rate was controlled at the nominal level of the test, or in red when the test became invalid. In general, when there is no actual dependence structure among either the data or model, the false positive rate was controlled; however, the tests became invalid when observations in both were not independent. This can be observed easily by noting that when $h^2_{\mathbf{m}}$ and $h^2_{\boldsymbol{\epsilon}}$ are both larger than zero, and in the absence of signal, there are excesses of very low and very high p-values. See the main text for details; for FWER-corrected, and for results using dataset B, see Appendix A.2.1.

Figure 2.12: *(page 68)* Pictorial table showing the Bland–Altman plots comparing the unrestricted with the restricted permutations for the dataset A. In each scatter plot, the horizontal axis contains average of the restricted and unrestricted p-values, in the range 0 to 1 and in steps of 0.1 in the marked grid, while the vertical axis contains the difference between the unrestricted and the restricted p-values, such that negative differences correspond to larger (less significant) restricted p-values. For the plots without effect, the vertical axis is in the range -0.3 to 0.3 and in steps of 0.1, and for those with effect, in the range -0.8 to 0.8 and in steps of 0.2. Because there would be too many p-values to be shown ($2.5 \cdot 10^5$, that is, 500 repetitions of 500 tests for each configuration), here only 1000 dots are shown on each of the 96 panels, two from each set at every repetition, selected randomly. To facilitate viewing, the dots for the error rates are shown in blue; for power, in green. The ellipsoids, shown in red, indicate the 95% confidence intervals. For FWER-corrected, and for results using dataset B, see Appendix A.2.1.


Unrestricted shuffling
Restricted shuffling
Without effect (error rate)
With effect (power)
$h^2_m = 0$
$h^2_m = 0.8$
Permutations only
Sign flippings only
Permutations + sign flippings
Normal errors
Skewed errors
$h^2_\epsilon = 0$
$h^2_\epsilon = 0.8$

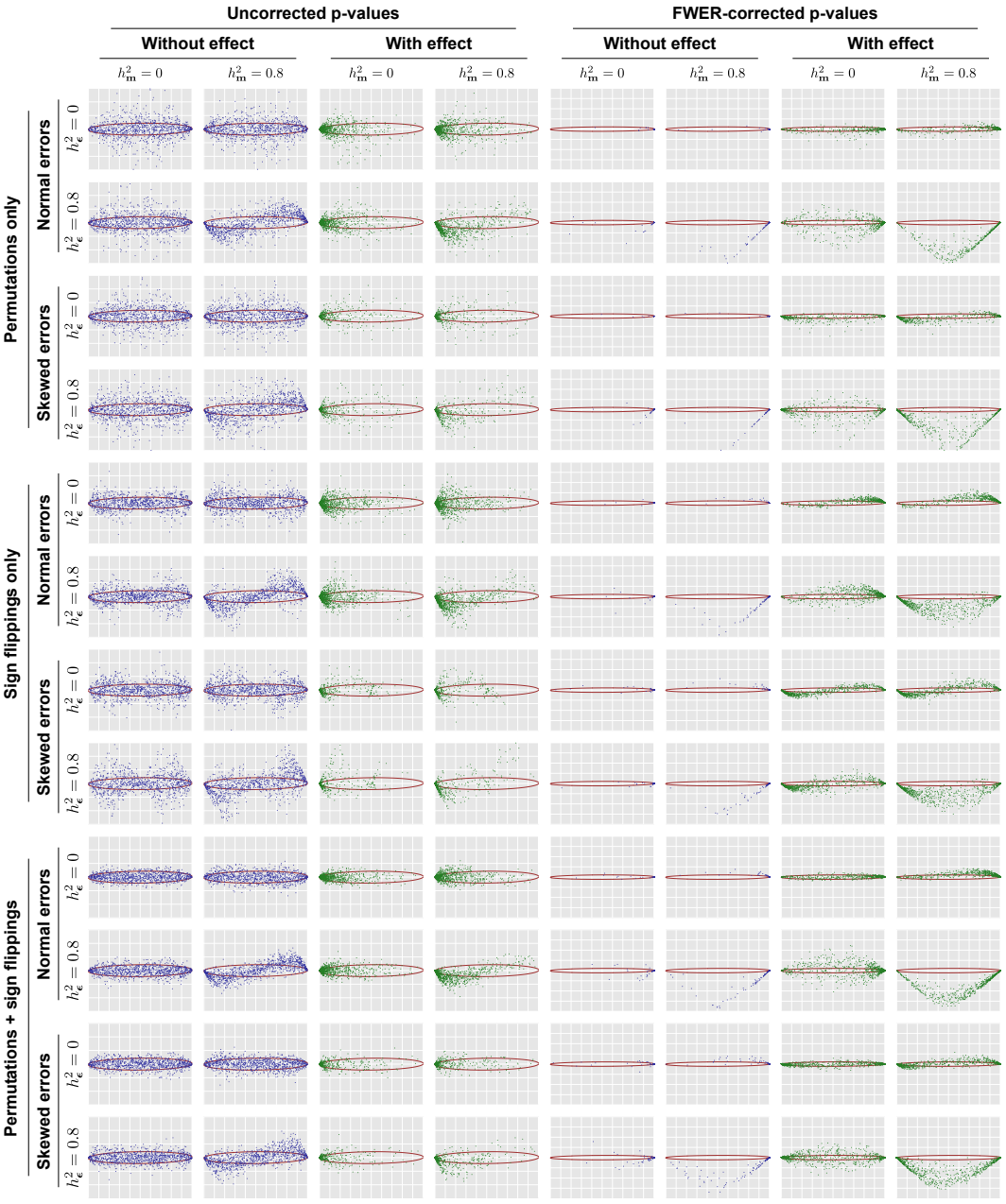

Uncorrected p-values
FWER-corrected p-values
Without effect
With effect
$h^2_{\mathrm{m}} = 0$
$h^2_{\mathrm{m}} = 0.8$
Permutations only
Sign flippings only
Permutations + sign flippings
Normal errors
Skewed errors
$h^2_{\epsilon} = 0$
$h^2_{\epsilon} = 0.8$

Table 2.6: Relationship between the average Hamming distance across shufflings and the observed power (± standard deviation). In general, larger reductions in the Hamming distance when using restricted permutations (EE) caused more noticeable losses in power (see also Figure 2.13). The loss did not correlate with the Hamming distance when using sign flippings only (ISE) or permutations with sign flippings (EE+ISE). In these cases, the power changes were generally minimal.

	Unrestricted shuffling		Restricted shuffling	
Set	Hamming distance	Power (%)	Hamming distance	Power (%)
Permutations only:				
A	34.929 ± 0.051	49.17 ± 7.18	33.956 ± 0.123	47.97 ± 7.22
B	25.945 ± 0.052	48.45 ± 7.77	24.956 ± 0.104	46.68 ± 7.57
C	9.980 ± 0.041	46.52 ± 10.92	9.981 ± 0.044	46.57 ± 10.73
D	16.965 ± 0.044	48.01 ± 10.93	11.003 ± 0.122	32.48 ± 8.48
E	13.973 ± 0.048	47.57 ± 10.23	9.991 ± 0.106	34.58 ± 8.80
F	13.867 ± 0.084	45.16 ± 10.11	12.000 ± 0.000	38.14 ± 8.69
G	13.972 ± 0.047	47.46 ± 10.18	13.975 ± 0.066	46.93 ± 10.21
H	515.996 ± 0.042	49.59 ± 2.37	496.000 ± 0.307	48.41 ± 2.16
I	515.994 ± 0.043	49.56 ± 2.25	496.063 ± 0.326	48.45 ± 2.23
Sign flippings only:				
A	17.959 ± 0.131	48.13 ± 6.81	18.000 ± 0.000	44.74 ± 6.28
B	13.476 ± 0.109	47.23 ± 7.66	13.500 ± 0.000	44.06 ± 7.17
C	5.495 ± 0.062	44.16 ± 10.27	5.489 ± 0.069	44.20 ± 10.07
D	8.715 ± 0.375	42.83 ± 10.97	9.000 ± 0.000	38.59 ± 8.75
E	7.492 ± 0.076	45.04 ± 9.05	7.479 ± 0.074	45.00 ± 9.29
F	7.215 ± 0.332	41.81 ± 10.35	7.500 ± 0.000	38.45 ± 7.99
G	7.292 ± 0.366	42.06 ± 10.22	7.500 ± 0.000	38.45 ± 7.99
H	258.528 ± 0.478	49.48 ± 2.23	258.440 ± 0.900	49.36 ± 2.34
I	258.542 ± 0.505	49.59 ± 2.33	258.525 ± 0.761	49.41 ± 2.33
Permutations with sign flippings:				
A	35.432 ± 0.042	50.07 ± 7.19	34.913 ± 0.095	50.00 ± 7.19
B	26.449 ± 0.040	49.72 ± 7.74	25.914 ± 0.092	49.24 ± 7.56
C	10.483 ± 0.041	50.30 ± 10.97	10.471 ± 0.033	50.32 ± 10.91
D	17.465 ± 0.041	49.94 ± 11.11	14.450 ± 0.137	46.29 ± 10.55
E	14.471 ± 0.039	49.87 ± 10.19	12.452 ± 0.088	48.81 ± 10.11
F	14.472 ± 0.039	49.93 ± 10.18	13.470 ± 0.085	47.79 ± 9.85
G	14.474 ± 0.035	49.95 ± 10.15	14.456 ± 0.055	49.05 ± 10.16
H	516.480 ± 0.036	49.68 ± 2.26	505.953 ± 0.569	49.65 ± 2.31
I	516.481 ± 0.038	49.61 ± 2.32	506.072 ± 0.570	49.69 ± 2.25

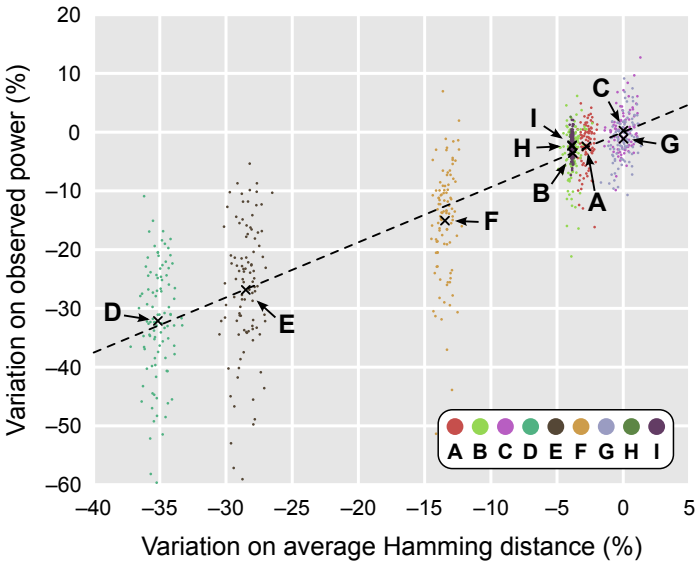


Figure 2.13: Changes in power related well to the average Hamming distance across permutations for the nine simulated datasets A–I (see also Table 2.6). When all dots are considered, $R^2 = 0.7557$ for a linear fit (dashed line); when only the centres of mass for each dataset (marked with "×" and indicated with arrows) are considered, $R^2 = 0.9902$.

distance across the permutations that were performed. This is shown in Table 2.6, and also visually in Figure 2.13. With independent and symmetric errors, in which only sign flippings are performed, the power losses were considerably smaller, and unrelated to the Hamming distance. In the same manner, permutations combined with sign flippings showed power changes that were minimal, and unrelated to the Hamming distance. Moreover, in these cases the resulting power was, for all datasets, higher than for just permutations or just sign flippings.

Table 2.6 and Figure 2.13 show a considerable dispersion of the observed power around the average. In the simulations, this dispersion can be reduced by one order of magnitude approximately just by using the same data and design for all repetitions, varying only the set of shufflings that are performed. Although this reduced dispersion would reflect more accurately the actual variation that different shufflings would cause on a given real experiment, the average power would be dependent on the exact, but random values used for the simulations, and would not be appropriate for the investigation performed here. The magnitude of variations on power as shown does not translate to actual experiments and should not be

Table 2.7: Heritabilities (h^2) for the indices of body size and for global cortical surface area and global average thickness on the sample of HCP subjects when a surrogate for common enviroment effects (c^2) is included in the model. The standard errors (SE), the test statistic ($2D_{\text{LL}}$), and the p-values are also shown. Only for height the common environment effect was estimated to be different than zero. All traits being highly heritable implies that permutation for analysis of their relationship must take the dependence structure into account.

	Additive genetic				Common environment			
Trait	h^2	SE	$2D_{\text{LL}}$	p-value	c^2	SE	$2D_{\text{LL}}$	p-value
Height	0.7346	0.1035	35.4	$1.3 \cdot 10^{-9}$	0.1409	0.0990	1.9	$8.7 \cdot 10^{-2}$
Weight	0.7248	0.0580	61.1	$2.8 \cdot 10^{-15}$	0.0000	–	–	–
BMI	0.7390	0.0572	62.1	$1.6 \cdot 10^{-15}$	0.0000	–	–	–
Area	0.8697	0.0274	125.3	$2.2 \cdot 10^{-29}$	0.0000	–	–	–
Thickness	0.8961	0.0232	125.5	$1.9 \cdot 10^{-29}$	0.0000	–	–	–

interpreted as such.

2.5.3 Real data

Summary statistics for height, weight, BMI, global cortical surface area, and global average thickness for the analysed HCP sample are shown in Table 2.7. The same Table also shows that, consistently with the literature, all these quantities are highly and significantly heritable, even when a conservative surrogate for common environment is included in the model. In fact, the estimated common environment fraction of the variance (c^2) was zero for all traits except for height. When the shared environment term is removed from the model, the estimated heritability for height increases to 0.8771 (standard error: 0.0244, $2D_{\text{LL}}$ = 146.9, p-value: $4.1 \cdot 10^{-34}$).

Permuting the data freely to test the hypotheses of correlation between thickness or area and indices of body size, therefore not respecting the structure of the sibships, allowed the identification of a few seemingly significant associations, even after FWER correction across the whole brain, and considering that both positive and negative tests were being performed. These regions, shown in Figure 2.14,

are (1) left anterior cingulate for a positive correlation between height and cortical surface area, (2) right orbitofrontal medial cortex for a positive correlation between thickness and BMI, (3), right temporal pole, at the confluence of the inferior temporal gyrus, for a negative correlation between thickness and body weight, and (4) right inferior temporal gyrus for a negative correlation between thickness and height. All these regions are very small, two of them comprising just one vertex at the resolution of the surfaces. However, using the proposed multi-level permutation strategy, in which shufflings only happen within siblings of the same type, and in which families with identical structure are allowed to be permuted as a whole, therefore respecting the kinship structure, all these findings became no longer significant. Table 2.8 shows the minimum (most significant) p-value throughout the brain for both unrestricted and restricted permutation.

2.6 Discussion

2.6.1 Error rates and power

The proposed multi-level shuffling strategy controlled the false positive rate at the nominal level in all configurations evaluated. With the only exception of sign flippings in the presence of skewed errors, which clearly violates assumptions, the empirical distribution of p-values was uniform, as desired, whenever shufflings respected the dependence structure present in the data; ignoring the dependence resulted in inflated error rates, rendering the test invalid. Dependencies in both data and model are present, for example, whenever an association between two heritable variables are investigated in a genetically informative sample, such as sibships (or even general pedigrees) as in the HCP case shown.

The guaranteed validity and exactness of p-values came, however, at the price of a small, yet noticeable and consistent reduction in power, that related to the complexity of the dependence structure and the ensuing restrictions on exchangeability. This can be understood by noting that the restricted permutation strategy

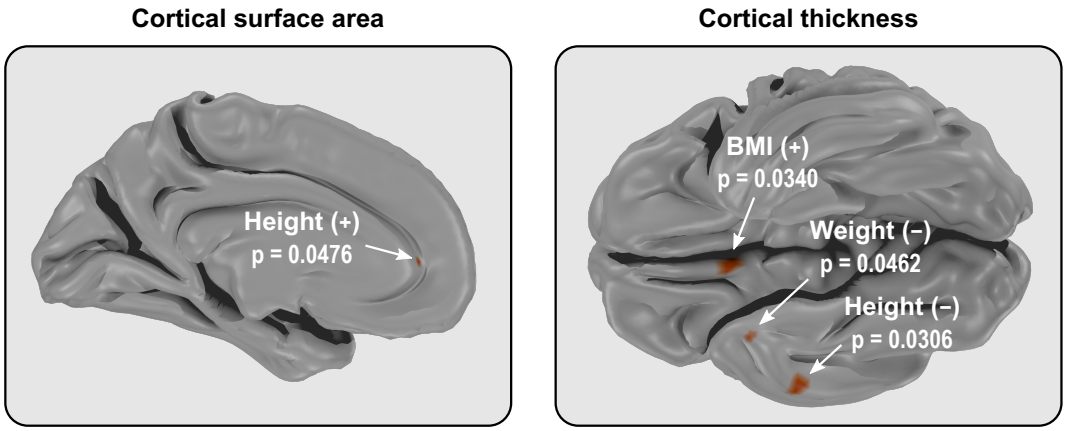


Figure 2.14: Maps showing the locations of the peaks of significance, for positive (+) and negative (−) correlations of height, weight, and BMI with cortical surface area and thickness. For conciseness, and given their lack of overlap, the original maps for thickness were thresholded at 0.05 and added together, allowing the regions to be displayed in the same figure. Even after using FWER-correction across the brain and contrasts, the unrestricted shuffling identified seemingly significant regions; these regions were not found significant using the restricted permutations that respect the family structure in the HCP sample. Provided that these traits are highly non-independent between subjects (i.e., heritable) this suggests that these results, produced with simple, unrestricted permutation, are in fact **false positives** (the peaks of significance for both restricted and unrestricted are listed in Table 2.8).

Table 2.8: Peak significance levels for each of the correlations between height, weight and BMI, and cortical thickness and local cortical surface area. The p-values below $\alpha = 0.05$ are marked in **bold**; all values are corrected controlling the familywise-error rate (FWER) across the whole brain and across both positive and negative correlations. Permuting the data freely, therefore violating exchangeability, identified seemingly significant associations. However, using the proposed permutation strategy that respects the data structure, these findings disappear, suggesting that these significant results are, in fact, **false positives**. See also Figure 2.14 in the main text.

Trait	Cortical surface area				Cortical thickness			
	Unrestricted		Restricted		Unrestricted		Restricted	
	Left	Right	Left	Right	Left	Right	Left	Right
Positive correlation:								
Height	**0.0476**	0.2178	0.1530	0.4398	0.4168	0.0734	0.6438	0.1694
Weight	0.9524	0.7448	0.9724	0.8346	0.2348	0.1706	0.3118	0.2376
BMI	1.0000	0.9368	1.0000	0.9586	0.0950	**0.0340**	0.1376	0.0522
Negative correlation:								
Height	0.9674	0.9936	0.9944	0.9996	0.2206	**0.0306**	0.4204	0.0788
Weight	0.9960	0.9908	0.9992	0.9978	0.1246	**0.0462**	0.1860	0.0810
BMI	0.8002	0.9848	0.8644	0.9926	0.4492	0.2674	0.5620	0.3588

does not disarrange the data as much as the unrestricted shuffling, with the consequence that the statistics computed after permuting the data may not be as distant from the statistic computed from the unpermuted data. With sign flippings, the power losses were smaller, and unrelated to the Hamming distance, presumably because even changes seemingly small, such as a single sign swap, can cause large perturbations on the shuffled data, sufficient to minimise reductions on sensitivity. Permutations combined with sign flippings showed minimal power changes that were also unrelated to the average Hamming distance, and with losses that were smaller than for just permutations or just sign flippings, suggesting that when both EE and ISE are valid for a given model, permutations with sign flippings can allow maximum efficiency.[4]

Although the diminished sensitivity could suggest that the multi-level permutation strategy would be "conservative" this is not the case, as can be attested by the exact control over error rates. This apparent incongruity can be understood through the Bland–Altman plots in Figure 2.12, that show that the differences in uncorrected p-values between both strategies is largely outside the margins of the confidence interval in both directions, suggesting that, under the null, variations in p-values can go in any direction when the strategies are compared. Nonetheless, in the presence of signal, or when the p-values are corrected for multiple testing using the distribution of the largest statistic across variables (such as voxels), the p-values for the restricted strategy tend to be stochastically larger than those for the free shuffling.

The restrictions imposed on the possible rearrangements that can be performed, with the consequent reduction in the number of possible permutations, as well as the lessened sensitivity, could be seen as negative, but in fact, such restrictions establish a set of rules under which permutation inference can be performed

[4] It should be noted, however, that explicitly selecting permutations that maximise the amount of disarrangement applied to the data, i.e., performing only those with highest Hamming distance, cause the error rates not to be controlled; conversely, a selection of only those that cause less shuffling cause the test to become conservative (results not shown).

in settings where otherwise it would not possible without appealing to often untenable assumptions, or that would not be possible at all. Simple permutation, if performed, would create data that could be impossible to be observed in practice, and thus, that should not be used to construct the reference distribution to test the hypotheses of interest. Moreover, the stronger the dependency is between observations, the fewer genuinely independent pieces of information are available to test a given hypothesis; in this scenario, smaller power does not appear unexpected.

2.6.2 Body size and cortical morphology

Height, weight, and BMI are known to be highly heritable in general, and were so for the HCP sample. Likewise, the heritability for global cortical surface area and average thickness are known to be heritable, and were found as such in the sample analysed. All these traits remained highly heritable even when a potential confound — a surrogate for household and maternal effects — was included. Even if estimated heritability were reduced, common effects would still cause the observations not to be independent. The fact that for all these traits there is a strong dependence between the observations implies that a permutation test that ignores the relationship between observations would not be valid, by violating the exchangeability assumption.

Indeed, the test that shuffled the data freely identified a few positive and negative localised significant associations between indices of body size and cortical area and thickness, even after FWER correction considering all tests in both hemispheres and the fact that positive and negative hypotheses were being tested. None of these areas were found to be significant if the test used only permutations that respected the structure present in the data, in the multi-level fashion, suggesting that these findings are likely false positives. None of the regions implicated were reported in previous studies that investigated relationships between indices of body size and cortical morphology (Pannacciulli et al., 2006; Raji et al., 2010; Ho et al., 2010,

2011; Smucny et al., 2012; Curran et al., 2013; Marqués-Iturria et al., 2013; Melka et al., 2013; Bond et al., 2014) that we could identify. It should be conceded, however, that not all these studies used the same methods, with some having analysed gray matter volumes in voxel-based representations of the brain, and some, despite using surface-based methods, performed analyses in macroscopic regions, as opposed to at each point in the cortical mesh. Still, as the simulations demonstrated, the violation of the exchangeability assumption makes the free permutation prone to inflated amounts of error type I if the observations are not independent, and the absence of similar findings from the literature supports the likelihood that these seemingly significant findings are not genuine, being instead false positives.[5]

Another aspect is that, although FWER-correction was applied considering the multiplicity of vertices in the mesh representation of the cortex and the two contrasts (positive and negative), no correction was applied considering that overall six tests were performed (three independent variables versus two dependent variables); FWER-controlling procedures that would take into account the non-independence between these tests are currently not available. Using Bonferroni correction, the results using the free permutation, which are likely false-positives as discussed above, disappear. Since most studies — and in fact, most of those referenced in the previous paragraph — investigated only the relationship between one independent versus one dependent variable, for which no such correction is necessary, the results shown emulate well the risk of false positives in similar, real studies.

2.6.3 Applications and other considerations

In addition to the above examples, and most clearly, with a direct application for the HCP data, the multi-level permutation strategy can be considered for repeated

[5] We have run a side analysis (not shown) in which we treated DZ pairs as ordinary siblings, i.e., using the tree shown in Figure 2.10 instead of the one in Figure 2.9, and observed results that are very similar to those reported, that is, nothing significant.

measurements designs in which within- and between-subject factors are tested in the same model or contrast, such as for a continuous regressor. A direct comparison of the power observed for datasets E, F and G, using permutations only, shows that even with the same number of subjects, the combination of within-block with whole-block permutation can be more powerful than each of these used in isolation. Moreover, the strategy can also be considered when not all measurements for a given subject are available, as long as compound symmetry within subject remains valid, without the need to exclude entirely the data for subjects as would be the case for whole-block permutation.

As experiments are planned considering strategies for subsequent analyses, the use of permutation tests can be included among the tools available, especially given its simplicity and, as demonstrated here and in a large body of literature, flexibility to accommodate designs and data that can be quite complex. Adequate planning includes ensuring that assumptions for permutation tests are met from the beginning, such as that the random errors of the instrument are stable along time, and do not vary with the values they measure, that the observations if not independent, possess a dependence structure that can be accommodated in a framework as the one shown here, and that observations, if not homogeneous, can be broadly qualified into a few number of variance groups.

Indeed, regarding VGs, the compatibility of these with the blocks ensures the feasibility of permutation tests, but it also allows that the necessary assumptions are reduced to a minimum: instead of requiring that all observations are homoscedastic (strong), the maximum possible amount of heterogeneity of variances that could still permit the shuffling as indicated by the blocks (weaker assumption) can be allowed. In this case, homogeneity would still be there, although not across all and every observation, just the minimal amount necessary so that the experiment can still be analysed. These considerations may not be relevant if the recruiting process, experimental conditions or data collection can guarantee that the same variance is homogeneous, but may be necessary when the data collec-

tion or samples are not under direct control of the researcher (e.g., reanalysis of past experiments, or census data).

This page intentionally left blank.

Chapter 3

Faster permutation inference

3.1 Introduction

Permutation tests allow exact control of error rates, with minimal assumptions. However, permutation tests are computationally intensive. For small, non-imaging datasets, recomputing a model thousands of times is seldom a problem, but for imaging applications, that involve testing at thousands of spatial points (voxels, vertices, faces, edges), large models that involve many subjects, multiple measurements, pointwise (voxelwise) regressors, spatial statistics, as well as other sources of complexity, even with the availability of inexpensive computing power, the same procedure can be prohibitively slow. Strategies to accelerate the process include the use of efficient or optimised code, the use of parallel, multi-threaded, or distributed computing, and the use of graphics processing units (GPUs) (for example applications of the latter, see Eklund et al., 2012, 2013; Hernández et al., 2013). While these methods are attractive for increases in speed, none reduce the amount of tasks that effectively need to be executed, and the improvements in speed happen through more efficient use of resources available, or through the introduction of yet more resources. At a time in which Moore's law (Moore, 1965) approaches physical limits (Waldrop, 2016), alternative methods to expedite computation are expected to gain prominence.

Here we exploit properties of the statistics themselves and their distributions, which could be used to accelerate the evaluation of the test in order to accept or reject the null hypothesis in a fraction of the time that otherwise would be needed with a large number of permutations. The main tenet of these approaches is to obtain a reduction of the number of actual computations that need to be performed, such that acceleration can be obtained in addition to, or irrespective of, generic improvements of software or hardware. In particular, we discuss the following approaches: (i) performing a small number of shufflings (with no other change from the usual case of permutation tests); (ii) estimation of the p-value as a parameter of a negative binomial distribution; (iii) fitting of a generalised Pareto distribution to the tail of the empirical permutation distribution; (iv) computing the p-values based on the expected moments of the empirical distribution, approximated from a gamma distribution; (v) direct fitting of a gamma distribution to the empirical distribution; and (vi) shuffling of a reduced number of points (e.g., voxels), with completion of the remainder using low rank matrix theory. Details of each are provided in Section 3.2.

Very few of such acceleration strategies have been investigated or used in brain imaging. The tail approximation was considered by Ge et al. (2012) for an imaging genetics application in which, due to the sheer volume of data, conventional permutation tests were not considered feasible. A variant of many possible algorithms for low rank matrix completion was proposed by Hinrichs et al. (2013). The fitting of a gamma distribution without the need for permutations was proposed recently for a range of statistics by Minas and Montana (2014). Here we aim to study, evaluate, and in some cases propose, solutions that can accelerate permutation tests for the general linear model (GLM), considering aspects that are specially relevant to imaging, such as the multiplicity of tests and the use of spatial statistics. In particular, we make the following main contributions: (I) show how a connection between Pillai's trace and the popular univariate t statistic allows the direct computation of the p-values from the permutation distribution, even without performing ac-

tual permutations, (II) use the moments of the empirical permutation distribution for the fit of a gamma distribution, and (III) propose a novel low rank matrix completion algorithm, writing the test statistic as the product of two matrices that can be sampled sparsely, and allowing exact recovery of what otherwise would be an approximation.

In Section 3.2 we begin by briefly reviewing the uni- and multivariate GLM, their assessment using permutation tests, and introduce the notation used throughout the chapter. The six different acceleration strategies are then presented in sequence, followed by certain aspects related to spatial statistics and multiple testing correction in the context of these methods. In the Sections 3.3 and 3.4 we assess the performance of these different methods on both synthetic and real data. In Section 3.5 we discuss these findings and provide recommendations for general circumstances. A summary of the acceleration strategies is provided in Table 3.1. Figure 3.1 illustrates four of them.

Table 3.1: *(page 84)* Overview of various strategies that can be considered to accelerate permutation tests. CMV: Classical multivariate test (such as MANCOVA); NPC: Non-Parametric Combination; see Winkler et al. (2016c) for details. Although the tail and gamma approximations can be considered for essentially any permutation distribution (the latter particularly for unimodal distributions), the Section 3.4 showed that the fit performs better for the distribution of the extremum statistic, as used for familywise error rate (FWER) correction. The negative binomial can be used for NPC, although unlikely with any acceleration benefit. For low rank matrix completion, many algorithmic variants can be considered, and the complexity needed for CMV and NPC may offset speed benefits; for this method, spatial statistics can be computed from the completed non-spatial (pointwise) statistics, although a direct computation, in a similar way as for the pointwise, would require a different algorithm with results that would likely not be exact. See main text for details on this and on all other methods.

Method	Brief description	Univariate				CMV				NPC			
		Pointwise		Spatial		Pointwise		Spatial		Pointwise		Spatial	
		unc.	corr.	unc.	corr.	unc.	corr.	unc.	corr.	unc.	corr.	unc.	corr.
Few permutations	Compute the p-values using just a few permutations, e.g., less than a thousand.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Negative binomial	Run for each voxel as many permutations as needed until a predefined number of exceedances is found. Then divide this number of by the number of permutations.	✓	•	✗	•	✓	•	✗	•	•	•	✗	•
Tail approximation	Run a small number of permutations and, for the p-values below a certain threshold (e.g., 0.10), fit a generalised Pareto distribution, modelling the tail of the permutation distribution.	•	✓	•	✓	•	✓	•	✓	•	✓	•	✓
No permutation	For statistics that can be written as $\text{trace}(\mathbf{AW})$, where $\mathbf{A} = \mathbf{X}\mathbf{X}^{+}$, $\mathbf{W} = \mathbf{U}\mathbf{U}'$, and $\mathbf{U}\mathbf{S}\mathbf{V}' = \text{svd}(\mathbf{R}_Z\mathbf{Y})$, compute analytically the moments of the permutation distribution, then fit a gamma distribution.	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
Gamma approximation	Run a small number of permutations, compute empirically the moments of the permutation distribution, then fit a gamma distribution.	•	✓	•	✓	•	✓	•	✓	•	✓	•	✓
Low rank matrix completion	Run a certain number of permutations, define orthonormal bases for matrices that are linear functions of the data and from which the statistic can be obtained; continue permuting a random subset of tests, filling the missing ones via projection to these bases.	✓	✓	•	•	•	•	•	•	•	•	•	•

✓ Can be used. • Can be used, although there are particularities (see main text). ✗ Cannot be used.

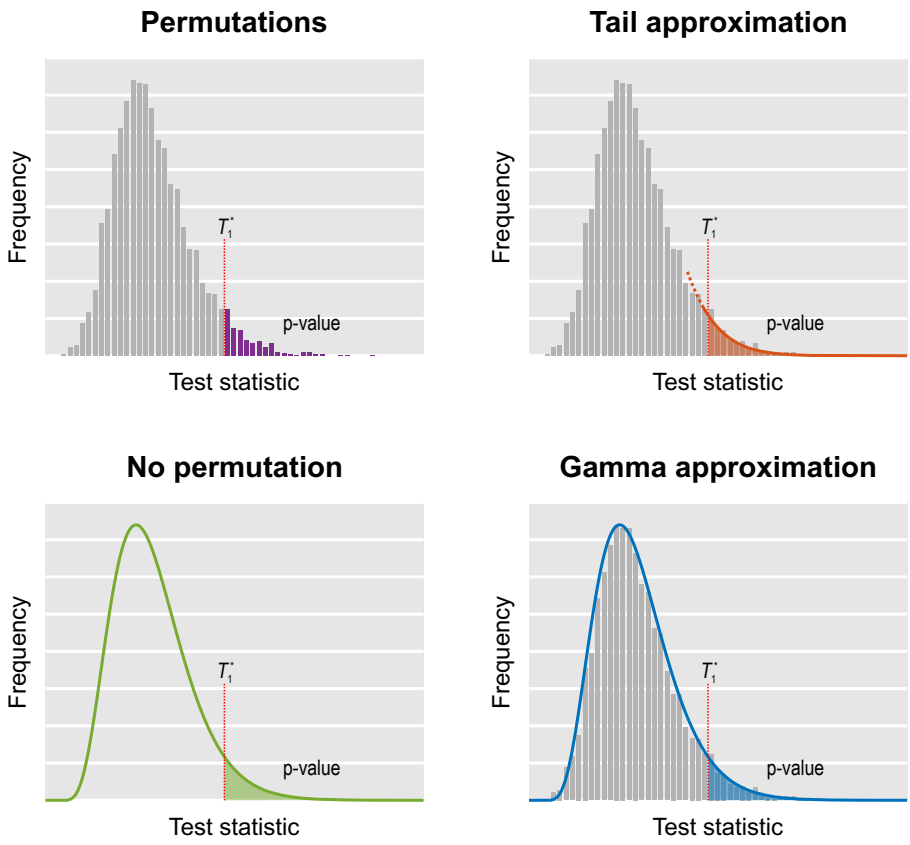


Figure 3.1: With permutations (i.e., any number of rearrangements, the use of the negative binomial distribution, or the low rank matrix completion), the p-value is the fraction of the test statistics obtained after permuting that are higher than in the unpermuted $T \equiv T_1^*$. In the tail approximation, the tail of the permutation distribution is subjected to the fit of a generalised Pareto distribution (GPD), from which the p-values are computed. In the method in which no permutations are performed, the first three moments of the permutation distribution are computed from data and model, and to these which a gamma distribution (Pearson type III) can be fitted, and from which the p-values are computed. In the gamma approximation, the moments of the empirical permutation distribution are used to the fit of the gamma distribution. The figure is merely illustrative: the actual fit uses the cumulative distribution function, such that histograms are not constructed in practice, hence the fit does not depend on binning.

3.2 Theory

3.2.1 Notation and general aspects

At each spatial point of an image representation of the brain, consider a general linear model (GLM) (Scheffé, 1959; Searle, 1971) expressed as:

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

where $\mathbf{Y}$ is the $N \times K$ matrix of observed data, with N observations of K distinct (possibly non-independent) variables, $\mathbf{M}$ is the full-rank $N \times R$ design matrix of explanatory variables (i.e., effects of interest and possibly nuisance effects), $\boldsymbol{\psi}$ is the $R \times K$ matrix of regression coefficients, and $\boldsymbol{\epsilon}$ is the $N \times K$ matrix of random errors. Estimates for the regression coefficients can be computed as $\hat{\boldsymbol{\psi}} = \mathbf{M}^{+}\mathbf{Y}$, where the superscript $(^{+})$ denotes a generalised inverse. One is generally interested in testing the null hypothesis that a contrast of regression coefficients is equal to zero, i.e., $\mathcal{H}_0 : \mathbf{C}'\boldsymbol{\psi}\mathbf{D} = \mathbf{0}$, where $\mathbf{C}$ is an $R \times S$ full-rank matrix of S contrasts of coefficients on the regressors encoded in $\mathbf{M}$, $1 \leqslant S \leqslant R$ and $\mathbf{D}$ is a $K \times Q$ full-rank matrix of Q contrasts of coefficients on the dependent, response variables in $\mathbf{Y}$, $1 \leqslant Q \leqslant K$; if $K = 1$ or $Q = 1$, the model is univariate. Once the hypothesis has been established, $\mathbf{Y}$ can be equivalently redefined as $\mathbf{YD}$, such that the contrast $\mathbf{D}$ can be omitted for simplicity, and the null hypothesis stated as $\mathcal{H}_0 : \mathbf{C}'\boldsymbol{\psi} = \mathbf{0}$. Another useful simplification is to consider a transformation of the model 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 regressors, 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 $\mathbf{C}$ such that the columns of $\mathbf{X}$ and $\mathbf{Z}$ are orthogonal to each other, and inference

on $\boldsymbol{\beta}$ is equivalent to inference on $\mathbf{C}'\boldsymbol{\psi}$ (Beckmann et al., 2001; Smith et al., 2007; Winkler et al., 2014). A suitable, pivotal test statistic, here generically termed T, is computed and its significance assessed through permutations and/or sign flippings of the data, the model, the residuals, or variants of these. We sometimes use the terms *rearrangement* or *shuffling* when the distinction between permutations or sign flippings is not pertinent. The p-value is computed as:

$$p = \frac{1}{J} \sum_{j=1}^{J} I\left(T_j^* \geqslant T\right) \tag{3.3}$$

where $I(\cdot)$ is the indicator function, T_j^* is the test statistic observed at the j-th shuffling of the data, J is the number of rearrangements performed, of which the first (i.e., $j = 1$) is the unpermuted case. We denote the significance level of the test as α. In typical cases, J is much smaller than the number of unique possible rearrangements allowed by the design and data, $J^{\max}$. The same procedure can be used with classical multivariate tests (CMV), such as MANOVA/MANCOVA or canonical correlation analysis (CCA), as well as with Non-Parametric Combination (NPC); details for both the univariate and multivariate GLM in the context of imaging are discussed in Winkler et al. (2014, 2016c).

Resampling risk Two methods may have similar error rates and power, yet fail to agree on which tests should have their null hypotheses rejected or retained. The *resampling risk* is a quantity that represents the probability of taking a different decision regarding the rejection or acceptance of the null hypothesis if the procedure is repeated using the same input data, but different methods (Jöckel, 1984). Compared to confidence intervals, which can be calculated for p-values derived from permutations through a binomial approximation (see the Section 3.2.2.1), the resampling risk is a more generic quantity in that it provides information on the chance of reaching a different decision regarding the null hypothesis that is computable for all the different methods, including, for instance, the one in which no permutations are used.

3.2.2 Acceleration methods

Nearly all of the acceleration strategies below can be applied to univariate, uncorrected pointwise tests, as shown in Table 3.1 ("pointwise" as an umbrella term encompassing voxelwise, vertexwise, facewise, as well as nodewise and edgewise graph theoretical measurements, or any other relevant imaging test). If $Q > 1$ or $K > 1$, the model is multivariate, and CMV or NPC can be considered (Winkler et al., 2016c). Some of the methods can also be used with spatial test statistics, and for inferences corrected for the familywise error rate (FWER) using the distribution of the extremum statistic (see below).

3.2.2.1 Few permutations

Conditional on the observed data, if all possible rearrangements are performed, a permutation test is exact in that it yields results that are not based on distributional assumptions or asymptotic approximations, but rather represent the exact probability of rejecting the null hypothesis when it is true. If fewer than all possible rearrangements are performed, the p-value obtained is an estimate of the true and unknown p-value; the test continues to be exact in that the probability of obtaining an estimate $\hat{p}$ less than or equal to the significance level α, is α itself, i.e., $P(\hat{p} \leqslant \alpha) = \alpha$, provided that the level is sensibly chosen considering the discreteness of the permutation p-values. Thus, a simple strategy for acceleration consists in running only a small number of permutations. As indicated above, this results in an unbiased (i.e., correct on average) estimate of the p-value, but with higher variance (variability around the true value) than when using a large number of permutations. Confidence intervals around $\hat{p}$ can be computed using one of the various methods for Bernoulli trials, such as those proposed by Wilson (1927), Clopper and Pearson (1934) or Agresti and Coull (1998) (for a comparative review, see Brown et al., 2001). Whichever is used, fewer permutations imply wider intervals (Table 3.2), such that the resampling risk can be expected to increase; in the Section 3.3 we assess this risk for the case of a few permutations, as well as for the

Table 3.2: Confidence intervals (95%), computed using the Wilson method, for a p-value $P = 0.05$ as a function of the number of permutations (J). More permutations narrow the confidence interval.

Number of permutations	Confidence interval
40	0.0138–0.1650
60	0.0171–0.1370
100	0.0215–0.1118
200	0.0274–0.0896
300	0.0305–0.0808
500	0.0341–0.0728
1000	0.0381–0.0653
2000	0.0413–0.0604
5000	0.0443–0.0564
10000	0.0459–0.0544
50000	0.0481–0.0519

other acceleration methods.

3.2.2.2 Negative binomial

If the permutations are performed randomly (as opposed to in some order, such as lexicographic), after a few permutations there may already be sufficient information on whether the null should be rejected, and continuation of the process narrows the confidence interval around $\hat{p}$, although with little chance of changing a decision about the rejection of the null hypothesis if the estimated p-value lies far from the test level α. The process can therefore be interrupted after some criterion has been reached. Various such criteria have been proposed (Andrews and Buchinsky, 2000; Davidson and MacKinnon, 2000; Fay and Follmann, 2002; Fay et al., 2007; Gandy, 2009; Kim, 2010; Sandve et al., 2011; Gandy and Rubin-Delanchy, 2013; Ruxton and Neuhäuser, 2013), and of particular interest is the interruption after a predefined number n of exceedances $T_j^* \geqslant T$ has been found. Weaker effects will quickly be exceeded after a few random shufflings, whereas stronger effects require insistence in doing more shufflings until exceedances are found.

The ensuing p-value is the estimated parameter of a negative binomial distribution (Haldane, 1945) as $\hat{p} = (n-1)/(j-1)$, where j is the permutation in which n was reached; this does not include the unpermuted case, and once that is considered, the permutation p-value becomes $\hat{p} = n/j$. This method was proposed by Besag and Clifford (1991), and compared to other approaches, it is attractive for its negligible computational overhead, and for bypassing the need that α or any other parameter is defined beforehand. If n has not been reached after a sufficiently large predefined number J of permutations, the process can be interrupted regardless, and the p-value computed as in Equation 3.3.

3.2.2.3 Tail approximation

The limiting distribution of the maximum of a set of identically distributed random variables converges to one of three well known families of distributions, under a form given by the generalised extreme value distribution (GEV) [Gnedenko (1943); for reviews, see Leadbetter et al. (1983); Davison and Huser (2015)]. More broadly, however, the tail of the distribution of an arbitrary random variable can be approximated using a generalised Pareto distribution (GPD) (Picklands III, 1975). For a threshold $u \to \infty$, the limiting distribution of the quantity $y = T - u$, for $T > u$, is $F(y) = 1 - (1 - \xi y/\sigma)^{1/\xi}$, defined for $y > 0$ and $\xi y/\sigma < 1$, with parameters ξ (shape) and σ (scale).[1] The effect of these two parameters on the density distribution is shown in Figure 3.2. Methods to estimate the two parameters of the GPD from the observed permutation statistics include maximum likelihood, the method of moments, or the method of probability-weighted moments; all three have similar estimation efficiency for $-1/2 < \xi < 1/2$, as typical in real world applications (Hosking and Wallis, 1987; Knijnenburg et al., 2009). Using the method of moments, the estimators of the scale and shape parameters are $\hat{\sigma} = \bar{y}(\bar{y}^2/s^2 + 1)/2$

[1] The shape parameter ξ of the GPD corresponds to the shape parameter of the generalised extreme value distribution, whereas the scale parameter σ relates to the GEV scale s as $\sigma = s - \xi(u - \mu)$, where μ is the GEV location parameter.

and $\hat{\xi} = (\bar{y}^2/s^2 - 1)/2$, where $\bar{y}$ and s^2 are respectively the sample mean and variance of the values y (Hosking and Wallis, 1987). Goodness of fit can be assessed with the Anderson–Darling test (Anderson and Darling, 1952; Choulakian and Stephens, 2001; Knijnenburg et al., 2009).

The algorithm proceeds as follows: a small number of permutations is initially performed, the set of test statistics T_j^* is recorded for each image point, and initial p-values computed as in Equation 3.3. The voxels with p-values above a loose, liberal significance level (such as twice the chosen α) remain unchanged; the others have the tail of their permutation distribution used to estimate the GPD parameters. For these, a reasonable, initial threshold u is the T_j^* that defines the upper quartile of their respective permutation distribution. This threshold is iteratively increased until a good fit of the GPD is found; if a good fit is not found when the permutation distribution has been exhausted, no approximation is made, and the initial p-value is not modified; otherwise, a new p-value is computed using the tail of the GPD fitted for that voxel. For the initial permutation distribution, the unpermuted statistic (T_1^*) may or may not be included in the process of tail fitting, and the impact of its inclusion depends on the number of permutations used for the initial distribution, as we show in the Section 3.4.

3.2.2.4 No permutation

Pillai's trace (Pillai, 1955) is a suitable statistic that can be considered to test $\mathcal{H}_0$. With the partitioned model, it can be computed as $T = \text{trace}(\tilde{\mathbf{Y}}'\mathbf{H}_\mathbf{X}\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1})$, where $\mathbf{H}_\mathbf{X} = \mathbf{X}\mathbf{X}^+$, $\tilde{\mathbf{Y}} = \mathbf{R}_\mathbf{Z}\mathbf{Y}$, $\mathbf{R}_\mathbf{Z} = \mathbf{I} - \mathbf{Z}\mathbf{Z}^+$, and $\mathbf{I}$ is the $N \times N$ identity matrix. Alternatively, it can be computed as $T = \text{trace}(\mathbf{H}_\mathbf{X}\mathbf{U}\mathbf{U}')$, where $\mathbf{U}$ is a $N \times K$ matrix containing the K left singular vectors of $\tilde{\mathbf{Y}}$ that have non-zero singular values. To see this, let $\mathbf{H} = (\mathbf{C}'\hat{\boldsymbol{\psi}}\mathbf{D})' \left(\mathbf{C}'(\mathbf{M}'\mathbf{M})^{-1}\mathbf{C}\right)^{-1} (\mathbf{C}'\hat{\boldsymbol{\psi}}\mathbf{D})$ and $\mathbf{E} = (\hat{\boldsymbol{\epsilon}}\mathbf{D})'(\hat{\boldsymbol{\epsilon}}\mathbf{D})$ be, respectively, the sums of products explained by the model (hypothesis) and the sums of the products of the residuals, i.e., that remain unexplained. Pillai's statistic is $T = \text{trace}\left(\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\right)$. With the model simplification and partitioning,

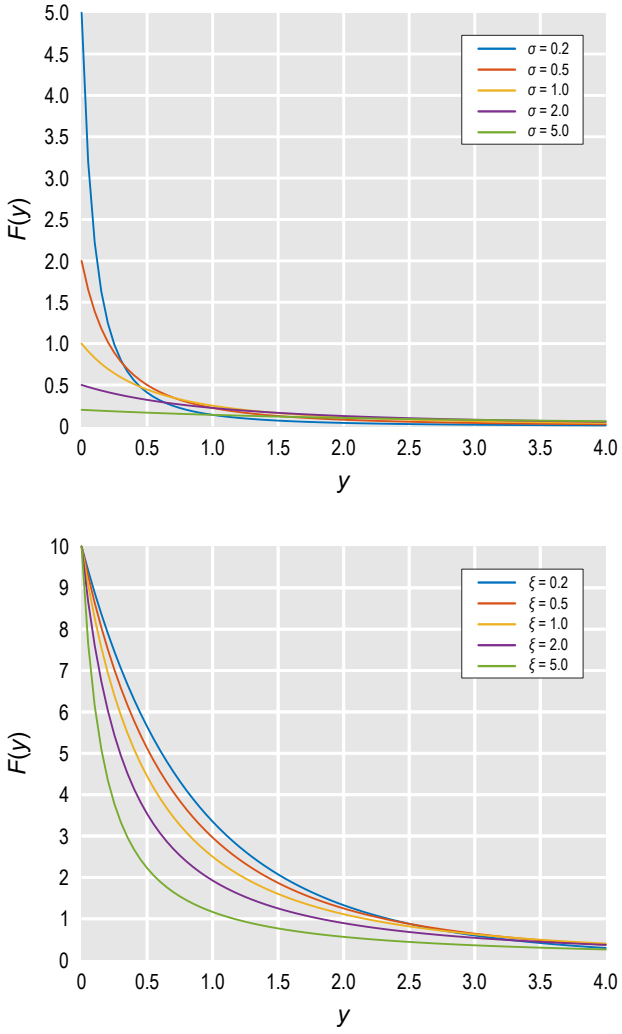


Figure 3.2: Probability density function (pdf) of the generalised Pareto distribution (GPD). The fit to the tail of the permutation distribution is accomplished by finding the optimal scale (σ) and shape (ξ) parameters that lead to the best approximation to the empirical distribution obtained from a reduced set of permutations.

$\mathbf{H} = (\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{H}_\mathbf{X}\tilde{\mathbf{Y}})'(\mathbf{H}_\mathbf{X}\tilde{\mathbf{Y}}) = \tilde{\mathbf{Y}}'\mathbf{H}_\mathbf{X}\tilde{\mathbf{Y}}$ and $\mathbf{E} = (\mathbf{R}_\mathbf{X}\tilde{\mathbf{Y}})'(\mathbf{R}_\mathbf{X}\tilde{\mathbf{Y}}) = \tilde{\mathbf{Y}}'\mathbf{R}_\mathbf{X}\tilde{\mathbf{Y}}$, where $\mathbf{R}_\mathbf{X} = \mathbf{I} - \mathbf{H}_\mathbf{X}$. Thus, $\mathbf{H} + \mathbf{E} = \tilde{\mathbf{Y}}'(\mathbf{H}_\mathbf{X} + \mathbf{R}_\mathbf{X})\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$. The trace of a product is invariant to a circular permutation of the factors, such that $T = \text{trace}(\tilde{\mathbf{Y}}'\mathbf{H}_\mathbf{X}\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}) = \text{trace}(\mathbf{H}_\mathbf{X}\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}') = \text{trace}(\mathbf{H}_\mathbf{X}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{+})$. Using the factors of a singular value decomposition, $\tilde{\mathbf{Y}} = \mathbf{USV}'$ and $\tilde{\mathbf{Y}}^{+} = \mathbf{VS}^{+}\mathbf{U}'$, where $\mathbf{U}$ contains only the K columns that correspond to non-zero singular values, the statistic becomes $T = \text{trace}(\mathbf{H}_\mathbf{X}\mathbf{USV}'\mathbf{VS}^{+}\mathbf{U}') = \text{trace}(\mathbf{H}_\mathbf{X}\mathbf{UU}')$. The matrices $\mathbf{H}_\mathbf{X}$, $\mathbf{UU}'$, and $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{+}$ are $N \times N$.

Let $\mathbf{A} \equiv \mathbf{H}_\mathbf{X}$ and $\mathbf{W} \equiv \mathbf{UU}'$, such that $T = \text{trace}(\mathbf{AW})$. For statistics that can be written in this form, with $\mathbf{A}$ and $\mathbf{W}$ being $N \times N$ symmetric matrices with mean-centered columns, the first three moments of the permutation distribution of the $N!$ possible values for T can be computed analytically under the assumption of symmetry of the error terms (Box and Watson, 1962; Mardia, 1971; Kazi-Aoual et al., 1995). With the moments known, a gamma distribution can be fitted, from which p-values can be obtained without permutations. The gamma distribution is the Type III distribution in the Pearson system (Pearson, 1895); references to the classical name often appear when the distribution is parameterised with respect to its moments, although here the current name is used to keep in pace with modern terminology.

The requirement of mean-centered columns for $\mathbf{A}$ and $\mathbf{W}$ implies that the model intercept is entirely represented in $\mathbf{Z}$, and that all columns of $\mathbf{X}$ have zero mean. This imposes a restriction on the set of designs for which this method can be considered. Simple group comparisons and correlations between continuous variables, for instance, are easily accommodated, whereas the means of individual groups are not.

When $\text{rank}(\mathbf{C}) = 1$ and $K = 1$ (or $Q = 1$), which is by far the most commonly encountered situation, the contrast has a direction (positive or negative), but Pillai's trace is two-tailed, which in principle would seem to diminish its usefulness, and limit the uses of the above relationship to just a few situations. This

is not a problem in practice: if T is Pillai's trace, then $\text{sign}(\beta)\sqrt{T}$ is the partial correlation coefficient, which has a monotonic relationship with, and therefore is permutationally equivalent to, the t statistic. Assuming that the (unknown) distribution of t is symmetric around zero, a p-value for the directional test can be computed by halving the p-value obtained from the gamma fit to the distribution of T, then subtracting the result from unity if the sign of the regression coefficient in the partitioned model (β) is negative. Thus, these relationships allow p-values that are based on the moments of the permutation distribution for Student's t-tests to be obtained, without doing any actual permutation.

3.2.2.5 Gamma approximation

Even for statistics that cannot be written in the form $T = \text{trace}(\mathbf{AW})$, the fit of a gamma distribution through moment matching has potential to yield valid, useful approximations (Solomon and Stephens, 1978; Minas and Montana, 2014). This category includes the distributions of spatial statistics, as well as the distribution of the extremum statistic, which is used to control error rates for the multiplicity of tests (both are discussed below). For such statistics, a small number of permutations is performed, the first three moments (mean, variance, and skewness) are estimated from the permutation distribution, and a gamma distribution with the corresponding moments fitted, from which the p-values are computed. As with the tail approximation, the unpermuted statistic (T_1^*) may or may not be included in the initial permutation distribution (we evaluated both ways, and return to this aspect below). The gamma distribution does not have infinite support in both directions, but some test statistics do have, and sometimes the unpermuted test statistic may fall outside the support of the fitted curve. To address this issue, depending on the direction of the skewness, the respective p-value is replaced by either 1 or $1/J$, i.e., the smallest attainable if no approximation had been done.

3.2.2.6 Low rank matrix completion

The statistics computed for each permutation can be organised in a matrix $\mathbf{T}$ of size $J \times V$, where J is the number of permutations and V is the number of image points (voxels, vertices, etc). Assuming that $\mathbf{T}$ has a low rank, only a small, random subset of its entries needs to be sampled; the missing ones can instead be recovered approximately using results from low rank matrix completion theory (Candès and Recht, 2009; Candès and Tao, 2010), with appreciable acceleration. However, despite the fact that $\mathbf{T}$ tends to have a dominant low rank component, with many small values in the eigenspectrum, it is still of full rank for statistics that are non-linear functions of the data, which is the case for nearly all the useful ones. Ignoring the end of the spectrum causes loss of information. While the rank can be recovered through the introduction of random noise with similar moments (Hinrichs et al., 2013), there is no guarantee that it will possess the same spatial structure that would preserve the distribution of spatial statistics used in imaging. There is also no guarantee that the residual noise can be characterised by the parameters of a particular distribution, which is at odds with a usable recovery of this matrix. This is the case even considering that some of the acceleration methods discussed in this chapter explicitly make this assumption in different contexts.

Here we follow a different strategy: we factorise $\mathbf{T}$ into a pair of matrices that can be assembled from linear functions of the data, thus allowing $\mathbf{T}$ to be recovered *exactly*. We begin by recalling that, using the partitioned model, when $\text{rank}(\mathbf{C}) = 1$ and $Q = 1$, a suitable statistic is the t statistic, such that each element of $\mathbf{T}$ is computed as $T_{jv} = \hat{\beta}_{jv}(\mathbf{X}'\mathbf{X})^{1/2}/\hat{\sigma}_{jv}$, where $\hat{\beta}_{jv}$ are the estimated regression coefficients for the j-th permutation and v-th voxel, and $\hat{\sigma}_{jv}$ is the standard deviation of the respective residuals, $\hat{\sigma}^2_{jv} = \hat{\boldsymbol{\epsilon}}'_{jv}\hat{\boldsymbol{\epsilon}}_{jv}/(N - \text{rank}(\mathbf{M}))$. Thus, $\mathbf{T} = \kappa\mathbf{B} \odot \boldsymbol{\Sigma}^{[-1/2]}$, where $\mathbf{B}$ is a $J \times V$ matrix that has entries $\hat{\beta}_{jv}$, $\boldsymbol{\Sigma}$ is a similarly sized matrix whose entries are the sums of squares of the residuals, $\varsigma_{jv} = \hat{\boldsymbol{\epsilon}}'_{jv}\hat{\boldsymbol{\epsilon}}_{jv}$, $\kappa = (\mathbf{X}'\mathbf{X}(N - \text{rank}(\mathbf{M})))^{1/2}$ is a scalar constant, $\odot$ is the Hadamard (elementwise) product, and the bracketed exponent in $\boldsymbol{\Sigma}$ indicates elementwise power. In this for-

mulation, it is $\mathbf{B}$ and $\mathbf{\Sigma}$ that are subjected to sparse sampling and low rank matrix completion, instead of $\mathbf{T}$ directly; the results of completion are used to compute $\mathbf{T}$ exactly, rather than approximately, provided that certain conditions are met.

Such exact matrix recovery is not possible unless at least as many entries as the degrees of freedom of the matrix, ν, are observed, a quantity that depends on the size and rank of the matrix to be recovered (Candès and Tao, 2010), and that should not to be confused with the degrees of freedom associated with the GLM. For a $J \times V$ matrix, $\nu = r(J + V) - r^2$, where r is the matrix rank. For full rank matrices, this implies observing all their entries, and doing so would not bring any speed improvement. However, provided that the matrix to be completed has rank $r < \min(J, V)$, then $\nu < J \cdot V$, so that not all its entries need to be seen or sampled. Moreover, if an orthonormal basis spanning the range of the matrix is known, such as its left singular vectors, complete recovery of the missing entries on any row or column can be performed using ordinary least squares regression (Troyanskaya et al., 2001), provided that, respectively, at least r observations are available on each row or column. If fewer are available, approximate recovery may still be possible.

Our objective is to sample some of the entries of $\mathbf{B}$ and $\mathbf{\Sigma}$, fill the missing ones, and compute $\mathbf{T}$. Although $\mathbf{B}$ and $\mathbf{\Sigma}$ do not need to have a matching set of entries sampled, it is convenient to do the sampling simultaneously, as both are produced from the same regression of the GLM. The number of entries that needs to be sampled depends then on which of these two matrices has the highest rank. To determine that, note that $\mathbf{B}$ can be computed as a product of a $J \times N$ and an $N \times V$ matrix. The rows and columns of each of these are determined, respectively, by the permutation and regression strategy, as shown in Table 3.3. With any of these strategies, the matrix product makes it clear that the upper bound on the rank of $\mathbf{B}$ is N. Likewise, $\mathbf{\Sigma}$ depends on the permutation and regression strategy, and its rank cannot be larger than the number of possible distinct pairs of N observations, which imposes an upper bound on the rank of $\mathbf{\Sigma}$ at $N(N + 1)/2$.

Table 3.3: A number of methods are available to obtain parameter estimates and construct the permutation distribution in the presence of nuisance variables. Comparative details and references for each of these approaches are in Winkler et al. (2014, Table 2); see also Anderson and Legendre (1999); Anderson and Robinson (2001). For the method of low rank matrix completion, $\mathbf{B}$ can be written as a product $\tilde{\mathbf{X}}\tilde{\mathbf{Y}}$, where $\tilde{\mathbf{X}}$ is a $J \times N$ matrix that contains the pseudo-inverse of the model on each row, and $\tilde{\mathbf{Y}}$ is an $N \times V$ matrix that contains the data. The j-th row of $\tilde{\mathbf{X}}$ is shown as $\tilde{\mathbf{x}}_j$, whereas the v-th column of $\tilde{\mathbf{Y}}$ is shown as $\tilde{\mathbf{y}}_v$. The rank$(\mathbf{B})$ is at most N, and can be smaller for most methods, even when $V > N$ and $J > N$, given the projection to subspaces due to $\mathbf{R_Z}$ and $\mathbf{R_M}$. The matrix $\boldsymbol{\Sigma}$ has rows $\boldsymbol{\varsigma}_j = \mathrm{diag}(\tilde{\mathbf{Y}}'\mathbf{R}\tilde{\mathbf{Y}})$, and its rank is, at most, $N(N+1)/2$. This determines the number J_0 of initial permutations to identify an orthonormal basis, and the number v_0 of tests that need to be done to allow exact recovery. See the text for details.

Method	Model	$\tilde{\mathbf{x}}_j$	$\tilde{\mathbf{y}}_v$	$\mathbf{R}$
Draper–Stoneman	$\mathbf{Y} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$	$\tilde{\mathbf{C}}'[\mathbf{P}_j\mathbf{X}, \mathbf{Z}]^+$	$\mathbf{Y}$	$\mathbf{I} - [\mathbf{P}_j\mathbf{X}, \mathbf{Z}][\mathbf{P}_j\mathbf{X}, \mathbf{Z}]^+$
Still–White	$\mathbf{PR_Z Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\mathbf{X}^+\mathbf{P}_j$	$\mathbf{R_Z Y}$	$\mathbf{I} - \mathbf{P}_j'\mathbf{X}\mathbf{X}^+\mathbf{P}_j$
Freedman–Lane	$(\mathbf{PR_Z} + \mathbf{H_Z})\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$	$\tilde{\mathbf{C}}'[\mathbf{X}, \mathbf{Z}]^+\mathbf{P}_j$	$\mathbf{R_Z Y}$	$\mathbf{I} - \mathbf{P}_j'[\mathbf{X}, \mathbf{Z}][\mathbf{X}, \mathbf{Z}]^+\mathbf{P}_j$
Manly	$\mathbf{PY} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$	$\tilde{\mathbf{C}}'[\mathbf{X}, \mathbf{Z}]^+\mathbf{P}_j$	$\mathbf{Y}$	$\mathbf{I} - \mathbf{P}_j'[\mathbf{X}, \mathbf{Z}][\mathbf{X}, \mathbf{Z}]^+\mathbf{P}_j$
ter Braak	$(\mathbf{PR_M} + \mathbf{H_M})\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$	$\tilde{\mathbf{C}}'[\mathbf{X}, \mathbf{Z}]^+\mathbf{P}_j$	$\mathbf{R_M Y}$	$\mathbf{I} - \mathbf{P}_j'[\mathbf{X}, \mathbf{Z}][\mathbf{X}, \mathbf{Z}]^+\mathbf{P}_j$
Kennedy	$\mathbf{PR_Z Y} = \mathbf{R_Z X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\mathbf{X}^+\mathbf{R_Z}\mathbf{P}_j$	$\mathbf{R_Z Y}$	$\mathbf{I} - \mathbf{P}_j'\mathbf{R_Z}\mathbf{X}\mathbf{X}^+\mathbf{R_Z}\mathbf{P}_j$
Huh–Jhun	$\mathbf{PQ}'\mathbf{R_Z Y} = \mathbf{Q}'\mathbf{R_Z X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\mathbf{X}^+\mathbf{R_Z}\mathbf{Q}'^+\mathbf{P}_j$	$\mathbf{Q}'\mathbf{R_Z Y}$	$\mathbf{I} - \mathbf{P}_j'\mathbf{Q}'\mathbf{R_Z}\mathbf{X}\mathbf{X}^+\mathbf{R_Z}\mathbf{Q}'^+\mathbf{P}_j$
Dekker	$\mathbf{Y} = \mathbf{PR_Z X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$	$\tilde{\mathbf{C}}'[\mathbf{P}_j\mathbf{R_Z}\mathbf{X}', \mathbf{Z}]^+$	$\mathbf{Y}$	$\mathbf{I} - [\mathbf{P}_j\mathbf{R_Z}\mathbf{X}', \mathbf{Z}][\mathbf{P}_j\mathbf{R_Z}\mathbf{X}', \mathbf{Z}]^+$

While the models as shown can be used for any general linear model (uni or multivariate), here the focus is on the univariate case ($K = 1$ or $Q = 1$) and in which rank$(\mathbf{C}) = 1$, such that $\mathbf{Y}$ and $\mathbf{X}$ are $N \times 1$ matrices (column vectors). After the partitioning, the effective contrast, $\tilde{\mathbf{C}}$, is a column vector of length R, full of zeroes except for the first element, that is equal to one. $\mathbf{Q}$ is an $N \times N'$ 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 $\mathbf{I}_{N' \times N'} = \mathbf{Q}'\mathbf{Q}$ (for this method only, $\mathbf{P}$ is $N' \times N'$; otherwise it is $N \times N$). $\mathbf{R_M} = \mathbf{I}_{N \times N} - \mathbf{MM}^+$. All other variables are described in the text. (It has been brought to our attention that the Smith method cited in Winkler et al. (2014) had been proposed previously by Dekker et al. (2007), hence it is here renamed.)

Thus we have the conditions in which not all samples are needed, that allow exact recovery of $\mathbf{T}$, and from which an algorithm arises naturally: (I) $\min(J, V) > N(N+1)/2$, (II) orthonormal bases spanning the range of $\mathbf{\Sigma}$ are known, and (III) for each permutation j, at least as many tests (e.g., voxels) as the rank of $\mathbf{\Sigma}$ are observed. For condition (I), the number N of subjects should ideally not be chosen based on speed considerations, but rather on statistical power and costs associated with data collection, and can be considered fixed for an experiment. The number V of points in an image is typically very large, such that this condition is trivially satisfied. The number J of permutations, however, can be varied, and should be chosen so as to satisfy (I). For condition (III), at least as many voxels than the rank of $\mathbf{\Sigma}$ are randomly sampled. For condition (II), orthonormal bases can be identified by first running a number $J_0 = N(N+1)/2$ of permutations using all V tests, and assembling initial fully sampled $\mathbf{B}_0$ and $\mathbf{\Sigma}_0$ matrices, which are subjected to SVD. With the two bases known, subsequent permutations $j = \{J_0 + 1, \ldots, J\}$ are done using a much smaller set of voxels; the results for these are projected to the respective orthonormal bases, recovering the complete j-th row of $\mathbf{B}$ and $\mathbf{\Sigma}$ for that permutation, and hence the corresponding row of $\mathbf{T}$. This proceeds as follows: consider the singular value decomposition $\mathbf{USV}' = \mathbf{B}_0$, where $\mathbf{U}$ is an $r \times V$ orthonormal basis, $r = \text{rank}(\mathbf{B}_0)$, $r < V$. In a given permutation j, a (possibly random) number v, $r \leqslant v < V$ of entries of the row $\boldsymbol{\beta}_j$ of $\mathbf{B}$ is observed; call this $1 \times v$ row $\tilde{\boldsymbol{\beta}}_j$. The complete row can be recovered as $\boldsymbol{\beta}_j = \tilde{\boldsymbol{\beta}}_j \tilde{\mathbf{U}}^{+} \mathbf{U}$, where $\tilde{\mathbf{U}}$ contains the respective v columns of $\mathbf{U}$ that match the observed row entries. The same procedure can be applied to the rows $\boldsymbol{\varsigma}_j$ of $\mathbf{\Sigma}$, using the basis derived from $\mathbf{\Sigma}_0$. $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_0$ have only positive entries, and to minimise the effects of sign ambiguity on the recovered data (for a description of the problem, see Bro et al., 2007), the mean can be subtracted before SVD, and added back after recovery.

The full matrix $\mathbf{T}$ is never actually needed. Instead, at each permutation, its j-th row is computed using completion as above, and discarded after counters have been incremented (Equation 3.3). To ensure that all permutations are

treated equally, particularly if the number of sampled voxels is smaller than $v_0 = N(N+1)/2$, the permutations $j = \{1, \ldots, J_0\}$ can be revisited and recomputed through low rank matrix completion once the orthonormal bases for $\mathbf{B}_0$ and $\mathbf{\Sigma}_0$ have been obtained.

A similar strategy can be considered for cases in which $\text{rank}(\mathbf{C}) > 1$ or $Q > 1$, for statistics other than t. However, to accommodate more regression coefficients for the F-statistic, or the various off-diagonal sums of products in the multivariate case for statistics as Wilks' λ or Pillai's trace, more than just two matrices would need to be sampled and filled, causing further computational costs that have potential to nullify, or even reverse, acceleration improvements. Finally, the dependence of the completion on a common design for all V tests does not allow for pointwise (voxelwise) regressors in the design matrix; all other acceleration methods discussed in this chapter, however, allow for this possibility.

3.2.3 Inference for spatial statistics

The distribution of spatial statistics, such as cluster extent (Friston et al., 1994), cluster mass (Poline et al., 1997; Bullmore et al., 1999) and threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009), can be computed using few permutations, from which p-values can be assessed. These can be further refined, at the tails, with a generalised Pareto distribution, or using the fit of a gamma distribution. The performance of these approaches for spatial statistics are assessed below. The negative binomial approximation cannot be used, because the permutations at each voxel are interrupted after a different number of permutations, preventing spatial statistics from being computed correctly (except for FWER, see below). Moreover, these statistics cannot be trivially written as $\text{trace}(\mathbf{AW})$, such that the method with no permutations cannot be used either. Finally, with low rank matrix completion, while it is possible to compute these statistics after missing voxels have been filled, it is unlikely that useful improvements on speed can be obtained, as most of the time spent on spatial statistics rests on the computation

of neighbourhood information.

3.2.4 Multiple testing correction

3.2.4.1 Overview

In brain imaging, each voxel (or vertex, or face, or edge) constitutes a single statistical test. Because thousands such voxels are present in an image, a single experiment results in thousands of statistical tests being performed. With so many tests, the chance of obtaining a spuriously significant result in at least one increases. This is known as the *multiple testing problem* (for a review in the context of brain imaging, see Nichols and Hayasaka, 2003). To take the multiple testing problem into account, either the test level or the p-values can be adjusted, such that instead of controlling the error rate at each individual test, the error rate is controlled for the whole set (family) of tests. Controlling such *family-wise error rate* (FWER) ensures that the chance of finding a significant result *anywhere* in the image is expected to be within a certain defined level. Either the adjustment of the test level, or the adjustment of the p-values for each test uses the distribution of the most extreme statistic across tests. The "most extreme" statistic is the one towards the direction in which the presence of an eventual effect would be manifest. Typically, the extremum is the maximum, although for some tests (e.g., Wilks' λ, for which smaller values are the ones more significant) the most extremum is the minimum.

The use of the distribution of the maximum (extremum) statistic can be understood in three ways: First, in the context of multiple testing, controlling the family-wise error rate amounts to testing a joint null hypothesis (JNH) that there is no effect *anywhere* across the image; if there is an effect *somewhere*, even in a single voxel, JNH can be rejected. If the tests are independent, the probability of retaining the JNH is the product of the probabilities of retaining the null hypotheses at each of the voxels; the probability of rejecting the JNH is its complement. This forms the basis of the well known Šidák procedure. Consider now the cumulative distribution function of the maximum of a set of independent random variables,

which itself is a random variable. Finding an instance of such variable that is smaller or equal than some arbitrary threshold t is only possible if all variables within the set are also smaller or equal than t. Thus, the probability that the maximum is smaller than or equal than t can be computed as the product of the probabilities that all variables within the set are also below t. Thus, for both cases (the JNH and the maximum of a set of random variables), the reference cumulative distribution function for obtaining probabilities is the same.

The second way in which the use of the distribution of the extremum can be understood is through a closed testing procedure (CTP). Marcus et al. (1976) showed that, in the context of multiple testing, the null hypothesis for a particular test can be rejected provided that all possible sub-JNH within the set that include that particular test are also significant, and doing so controls the FWER. Such joint test can be quite much any valid test, including CMV or NPC, all of which are based on recomputing the test statistic from the original data, or others, based on the test statistics or p-values of each of the separate (partial) tests, as in a meta-analysis. A CTP incurs a substantial computational overhead, though: the number of sub-JNHs that needs to be tested grows exponentially with the number of tests present in the family, which in imaging applications renders them unfeasible. However, there is one particular joint test that provides a direct algorithmic shortcut: using the maximum as the test statistic. The maximum across all tests is also the maximum for any subset of tests, such that all the sub-JNHs can be skipped altogether. This gives a vastly efficient algorithmic shortcut to a CTP, as shown by (Westfall and Young, 1993). Additional information on CTP in this context can be found in Winkler et al. (2016c).

The third way is merely intuitive: if the distribution of some test statistic that is not the distribution of the maximum within an image were used as the reference to compute the (FWER-adjusted) p-values at a given voxel v, then the probability of finding a voxel with a test statistic larger than t_v anywhere could not be determined: there could always be some other voxel v', with an even larger statistic (i.e.,

$t_{v'} > t_v$), but the probability of such happening would not be captured by the distribution of a non-maximum. Hence the chance of finding a significant voxel anywhere in the image under the null hypothesis (the very definition of FWER) would not be controlled. Using the absolute maximum eliminates this logical leakage.

The family-wise error rate is not the only measure of uncertainty that can be controlled in the context of multiple tests. Of particular interest is the control of the proportion of false positives among the tests. This quantity, that came to be know as *false discovery rate* (FDR) can be controlled, on average, using a simple multi-step procedure that operates on the ranked p-values (?). Both strategies (FWER and FDR) are popular in imaging data analysis.

3.2.4.2 Correction under acceleration

As the control of the FWER requires the distribution of the extremum statistic, the method in which no permutations are done cannot be used, as the extremum cannot be written as trace($\mathbf{AW}$). The negative binomial, as proposed, if operating individually at each test (voxel) cannot be used either: later rearrangements include fewer voxels than the initial ones, thus changing the skewness of the distribution of the extremum as the shufflings are performed. A possible workaround for the negative binomial is to interrupt the shufflings once the extremum across tests in a given permutation exceeds (a number n of times) the extremum in the unpermuted case; the empirical distribution of the maximum statistic obtained at this point is used for the adjustment of the p-values. This permits also the use of spatial statistics. A potential problem for this approach is that all voxels in an image would depend entirely on the result found for the single, most extreme test in the unpermuted case: an incidental incorrect result at that single voxel would affect the results across the whole image.

Other methods can be used directly for FWER-correction: few permutations, tail and gamma approximations, and low rank matrix completion can all be used. For the tail and gamma, the GPD and the gamma distribution are, respectively, fit-

ted to the distribution of the extremum after a fixed, possibly small number of permutations have been performed. For the low rank matrix completion, the distribution is obtained by taking the maximum across the V columns of $\mathbf{T}$, thus producing a vector of length J containing the extrema, from which p-values can be computed for all voxels in the image.

Such correction is not limited to the points within an image: under the same principles the extremum statistic can be used to correct across multiple imaging modalities, multiple contrasts (i.e., multiple hypotheses using the same data), as well as a mixture of imaging and non-imaging data (Winkler et al., 2016c), provided that the test statistic is pivotal, that is, that its asymptotic sampling distribution does not depend on unknown parameters (Winkler et al., 2014).

Controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995; Genovese et al., 2002) requires that, under the null, the distribution of the p-values is uniform on the interval $[0\ 1]$. This condition can be relaxed by accepting p-values that are valid for any significance level smaller than or equal to the proportion of false discoveries that the researcher is willing to tolerate, i.e., $\alpha \leqslant q_{\text{FDR}}$, which not only encompasses the original definition, but also accommodates the cases (e.g., with TFCE) in which the uniformity of the distribution of p-values is lost only for high p-values, which are typically of no interest. It should be noted, however, that from its own definition, FDR is expected to be conservative with discrete p-values if too few permutations are performed, which can be predicted from the original formulation, and as it has been described in the literature (Gilbert, 2005). This can be the case if some tests are found significant (the true proportion of false discoveries may be smaller than the level q_{FDR}, due to ties), or if none is found significant (the true familywise error rate, usually weakly controlled by FDR, may be below q_{FDR} or even equal to zero, as the lower bound on the p-values, dictated by the number of permutations, may not be sufficiently small to allow any rejection).

3.2.5 Algorithmic complexity

The actual time needed to perform each method depends on choices made at implementation, including programming strategies, resources offered by the programming language and the compiler, as well as the available hardware. Asymptotic bounds and memory requirements are more realistic as means to provide a fairer comparison, and a summary is shown in Table 3.4. Compared to an ideal method in which a very large, potentially exhaustive ($J^{\max}$), number of shufflings is performed, and that would have asymptotic computational complexity $\Theta(NVJ^{\max})$, each method uses a different strategy to increase speed. Few permutations, tail and gamma approximations use small J. Speed is increased in the negative binomial case by means of reducing the number of shufflings based on the number n of exceedances needed, thus having a stochastic runtime. The no permutation case bypasses the need for permutations altogether. Compared to the others, the low rank matrix completion has lower asymptotic run time when N is small in relation to V and J.

As the acceleration in each of the methods is due to different mechanisms, the stage at which the increments in speed happen varies. For few permutations, as well as for tail and gamma approximations, the increases in speed happen through the use of fewer shufflings; the latter two, however, need additional time to allow the fit of a GPD or gamma distribution respectively, to the initial, permutation distribution. For FWER-corrected results, such fitting is quick, as it needs to be performed for only one distribution (of the extremum statistic); for uncorrected results, however, this process takes considerably longer, as each voxel needs its own curve fitting. The negative binomial benefits from fewer permutations, and further, benefits from a reduction in the number of tests (voxels) that need to be assessed, although there is a computational overhead due to the selection of tests that did not reach the number of exceedances and need to continue to undergo permutations. The low rank matrix completion benefits from a dramatic reduction in the number of tests that need to be done, a quantity that depends only on the

Table 3.4: Computational complexity and memory requirements for the different methods.

Method	Computational complexity	Specific storage
Few permutations	$\Theta(NVJ)$	$2V$
Negative binomial	$\Theta(nN\log(V))$	$2V$
Tail approximation	$\Theta(V(NJ+1))$	$V(J+1)$
No permutation	$\Theta(NV)$	V
Gamma approximation	$\Theta(V(NJ+1))$	$V(J+1)$
Low rank matrix completion	$\Theta(N^3(V+J))$	$2V(2J_0+1)$

N is the sample size, V the number of tests in an image (such as voxels or vertices), n the number of exceedances, and J the number of permutations, and J_0 the number of fully sample permutations in the low rank matrix completion method. The computational complexity refers to the acceleration, and does not include steps that are common to all methods, such as the model partitioning, computation of the test statistic and other procedures. Likewise, the specific storage refers to the amount of memory needed to store the bulk of the intermediate data that are particular for each method, and ignores storage needs that are common to all methods, such as for the data itself, the design matrix, the set of permutations, etc.; it also ignores small transitory variables that occupy insignificant amounts of memory. Tail and gamma as indicated consider the fitting for uncorrected p-values, that need one fit per test (voxel); if only FWER is required, the cost of a single fit is negligible, and these can be considered $\Theta(NVJ)$.

number of subjects and not on the size of the actual images. The method in which no permutations are performed benefits from the analytical solution and, as the name suggests, the waiver of the need to permute anything.

The memory requirements also vary. For the few permutations and negative binomial, only the array of V elements containing the test statistic, and another of the same size for the counters to produce p-values are needed. For the tail and gamma approximations, the test statistics for all J permutations need to be stored, from which the moment matching is performed. The no permutation does not require counters. The low rank matrix completion needs two arrays of size $V \times J_0$ to store the values of $\mathbf{B}_0$ and $\mathbf{\Sigma}_0$, and two further arrays of the same size to store the orthonormal bases (at which point $\mathbf{B}_0$ and $\mathbf{\Sigma}_0$ are no longer needed).

3.3 Evaluation methods

In an initial phase, we explored all methods using synthetic univariate and multivariate data and a wide variety of parameters. We assessed their performance in terms of agreement of the p-values with those obtained from a reference set constructed from a relatively large number of permutations, which provide information on error rates and power. In a second phase, using a more parsimonious set of parameters, univariate data, and a hundred repetitions, we assessed the resampling risk and speed. Real data was used as an illustration in which speed and resampling risk were also evaluated.

3.3.1 Synthetic data: Phase 1

The dataset consisted of $N = 20$ synthetic images of size $12 \times 12 \times 12$ voxels, containing random variables 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[2]). The use of these two distributions is to cover a large set of real world problems, with a well-behaved (Gaussian) and a skewed (Weibull) distribution. While the methods are not limited to imaging data, the use of images is helpful for permitting the assessment of the methods using spatial statistics.

To these images, and following the notation from the Section 3.2.1, simulated effects were added as $\mathbf{M}\boldsymbol{\psi}$, with $\boldsymbol{\psi} = [\psi_1 \;\; 0]'$, ψ_1 being either 0 (no effect) or $t_{\text{cdf}}^{-1}\left(1-\alpha; N - \text{rank}\left(\mathbf{M}\right)\right) \big/ (\mathbf{C}'(\mathbf{M}'\mathbf{M})^{+}\mathbf{C})^{1/2}$, where $\mathbf{C} = [1 \; 0]'$ is the contrast and $\alpha = 0.05$ is the significance level of the permutation test to be performed at a later stage, thus ensuring a calibrated signal strength sufficient to yield an approximate power of 50% with Gaussian errors, irrespective of the sample size; for the Weibull distribution, the signal was further weakened by a factor $5/8$, also ensuring power

[2] Thus with actual skewness $\left(\Gamma\left(1+3/k\right)\lambda^3 - 3\mu\sigma^2 - \mu^3\right)/\sigma^3 \approx 19.58$, where here μ and σ^2 represent the mean and variance of this distribution, and k and λ the shape and scale parameters.

of approximately 50%. Signal was added to all voxels, thus avoiding the usual problems of signal bleeding, due to smoothing, to areas of otherwise pure noise. The actual effect was coded in the first regressor only, with the second regressor modelling an intercept. The first regressor was constructed as a set of random values following a Gaussian distribution with zero mean and unit variance. Smoothing was applied with a Gaussian kernel of full width at half maximum (FWHM) of 4 voxels in all three directions, implemented as multiplication in the frequency domain, without zero padding, such that positive dependencies among voxels was introduced as desired, and without producing edge artefacts.

It should be noted that the performance of the voxelwise tests does not depend on the size of the dataset, such that although the synthetic data is smaller than in typical real world problems, the results still generalise. For spatial statistics, that depend on the size of the dataset, the size chosen (12 voxels in each direction) was verified empirically to be sufficient, considering also the size of the smoothing kernel, and the fact that smoothing was done in the frequency domain without padding. A smaller synthetic dataset could risk jeopardising the analysis of spatial statistics, whereas a substantially larger dataset would make the whole analysis computationally prohibitive (even considering the accelerations provided by the methods).

Tests were performed using just one such simulated image (univariate) or three (multivariate data). For the latter, both CMV and NPC test statistics were considered, using Wilks' λ, and Pillai's trace for CMV, and the combining functions of Tippett and Fisher for NPC (Winkler et al., 2016c). These cover the most common cases. For all these statistics, permutations (for exchangeable errors, EE), sign flippings (for independent and symmetric errors, ISE), and permutations with sign flippings (EE and ISE) were performed. To assess how the parameters needed for each acceleration could impact results, these were varied:

- Few permutations: $J = \{40,\ 60,\ 100,\ 200,\ 300,\ 500,\ 1000,\ 2000,\ 5000\}$, where J is the number of permutations.

- Negative binomial: $n = \{2, 5, 10, 15, 20, 50, 100\}$ and $J = \{50000\}$, where n is the number of exceedances before interrupting the process.

- Tail approximation: $J = \{40, 60, 100, 200, 300, 500, 1000, 2000, 5000\}$, using $p = 0.10$ as the threshold below which the p-values are refined, and including or not the first permutation test statistic, $T_1^* \equiv T$ in the initial null distribution to which tail the GPD is fit.

- No permutation: No parameters to be varied for this method.

- Gamma approximation: $J = \{40, 60, 100, 200, 300, 500, 1000, 2000, 5000\}$, and including or not the first permutation test statistic in the initial null distribution, to which the gamma is fit.

- Low rank matrix completion: $v = \{42, 105, 210, 864\}$ and $J = \{210, 300, 500, 1000, 2000, 5000, 50000\}$, where v is the number of voxels randomly selected to infer the values of all others. The value $v = 210$ corresponds to $v_0 = N(N + 1)/2$. We expected that v equal to or larger than this critical value would allow perfect reconstruction of the test statistic, but wanted to assess whether smaller values (one half or one fifth of this value) would still be acceptable as approximations; the $v = 864$ corresponds to oversampling. For the univariate case only, a further run using $J = 50000$ and the exact same permutations as the reference set was used to verify their equality.

The 81 possible configurations above generated 709 sets of results considering the univariate, the two CMV, and the two NPC, and the univariate non-spatial statistics (uncorrected and FWER-corrected), TFCE (uncorrected and FWER-corrected) and cluster extent and mass (corrected). Further, the 12 combinations of signal, noise and shuffling strategy required a total of 8508 scenarios to be considered. Each of the six acceleration methods were compared to a reference set produced with $J = 50000$ permutations, which were assessed using PP and QQ plots, constructed in logarithmic scale [henceforth log(PP) and log(QQ)] so as to emphasise

the smaller, more interesting p-values, and Bland–Altman plots (Bland and Altman, 1986), all with 95% confidence intervals estimated from an approximation to the binomial distribution using the Wilson method (Wilson, 1927). Error rates and power were computed using respectively the simulations without and with signal.

3.3.2 Synthetic data: Phase II

In addition, for the univariate, Gaussian errors, with and without signal, and exchangeable errors (permutations only), 100 realisations were performed using all the various methods and respective parameters, except low rank matrix completion (Phase I demonstrated it produces identical results as using ordinary permutations; see the Section 3.4). This allowed empirical standard deviations, as opposed to estimated confidence intervals, to be computed and included in the log(PP) and Bland–Altman plots. Histograms of p-values, with the variability on the heights of the bars, could also be computed. Estimates of error rates, power, and resampling risk were obtained, as well as elapsed times. Error rate and power were computed as the proportions of the tests that were declared significant in the absence or presence of a true effect respectively. The resampling risk was calculated as the proportion of tests in which a decision whether accepting or rejecting the null hypothesis changed when using a given method compared to the reference set of 50000 permutations. These simulations also allowed log(QQ) plots for the extremum statistic, based on the 100 repetitions, as opposed to plots for the corrected FWER p-values as in Phase I.

3.3.3 Real data

We conducted a re-analysis of the data of the voxel-based morphometry (VBM) study by Douaud et al. (2007). In brief, T_1-weighted magnetic resonance images of 25 subjects diagnosed with schizophrenia and 25 controls matched for sex and age

were obtained. These images were analysed with FSL-VBM[3] (Douaud et al., 2007), an optimised VBM protocol (Good et al., 2001) carried out with the FMRIB Software Library (FSL; Smith et al., 2004). In short, the grey matter was segmented from the T_1-weighted image, non-linearly registered to a common space, modulated and smoothed, and the two groups of subjects compared using a design corresponding to a two-sample t-test. This is the same dataset used in the original evaluation of TFCE (Smith and Nichols, 2009) and for the present re-analysis, we considered the same two levels of smoothing, i.e., with $\sigma = 3$, that correspond to FWHM of approximately 7 mm. The overall number of voxels included in this analysis was $V = 231,259$.

The parameters used for the acceleration strategies are the same used for Phase I of the simulations, except that for low rank matrix completion, and considering the $N = 50$, the parameters were held fixed at $v_0 = N(N+1)/2 = 1275$ and $J = 5000$. The reason is that using a smaller v would cause the method to fail to recover, even approximately, the non-sampled statistics not even approximately as the simulations in Phases I and II demonstrated (see the Section 3.4), and varying J, once v has been fixed, is equivalent to the few permutations method.

3.4 Results

Phase I allowed a comparison between p-values obtained from the reference set with those obtained by the various acceleration methods and uncorrected error rates, whereas Phase II allowed an estimation of the familywise error rate after multiple repetitions. The VBM example permitted inspection of the results of a practical example of an imaging modality that offers various statistical challenges, particularly with respect to non-stationarity (Hayasaka et al., 2004; Salimi-Khorshidi et al., 2011) and skewness (Salmond et al., 2002; Viviani et al., 2007). The multiplicity of scenarios resulted in the construction of more than 25 thou-

[3] Available at fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLVBM

sand plots and maps, which do not fit the thesis format; these were organised in a browsable set of pages that can be viewed online, or downloaded and browsed locally; details are in Appendix A.

Despite the large and multidimensional nature of the simulations and analysis of the real data, both of which considered many possible parameters, and the fact that each method may have strengths under different evaluation metrics, the overall results are generally simple to describe, and are summarised below.

Error rate Nearly all methods, when used according to their respective theory, yielded, on average, exact error rates. Evidence for this assertion comes from the log(QQ) plots produced in Phase I, that show p-values running along the identity line, or not deviating more than by their respective 95% confidence interval, and the log(PP) and histograms produced from the hundred repetitions performed in Phase II, as shown in the Appendix A.3.1. A notable exception occurred, for the uncorrected case, if the unpermuted statistic T_1^* was not included in the null distribution for the gamma and tail approximations, and if less than 500 or 1000 permutations respectively were performed, in which case the error rate was on average above the nominal level. For the corrected, error rates were controlled regardless, and the difference between inclusion or not was negligible. Another exception was, for low rank matrix completion, the use of fewer than the prescribed v_0 tests, which led to error rates being not well controlled; using at least this quantity not only allowed the method to remain exact, but produced results in complete agreement (that is, perfectly identical) to using the same number of permutations and full sampling (that is, without completion).

Power Conditional on the error rate being controlled, all methods yielded generally similar power, as evidenced by the histograms in produced in Phase II, shown in the Appendix A.3.2. It should be noted, however, that although more permutations did not intrinsically increase power, they allowed smaller p-values to be found, thus being beneficial for methods that use permutation (few permuta-

tions, tail approximation, gamma approximation, and low rank matrix completion) if the significance level were smaller than $\alpha = 0.05$, and certainly for the use of FDR.

Agreement with the reference set The smaller p-values (e.g., smaller than 0.10), were generally similar across methods, agreeing well with the reference set of results produced with 50000 simple permutations, without considerable variations that would result in entirely different results, both in the presence and absence of signal, although for p-values in the middle of the distributions, results often varied widely. In the Appendix A, this can observed in the log(PP) and Bland–Altman plots. The two important exceptions were: (I) for low rank matrix completion using fewer tests (voxels) than v_0, that led to widespread disagreement with the reference set and often nonsensical results, and (II) for the no permutation method if the resampling used only sign flipping, or if the errors were skewed. Moreover, for p-values away from the tail, the disagreement of the no permutation method with the reference set was substantial, even with symmetric errors and permutations only.

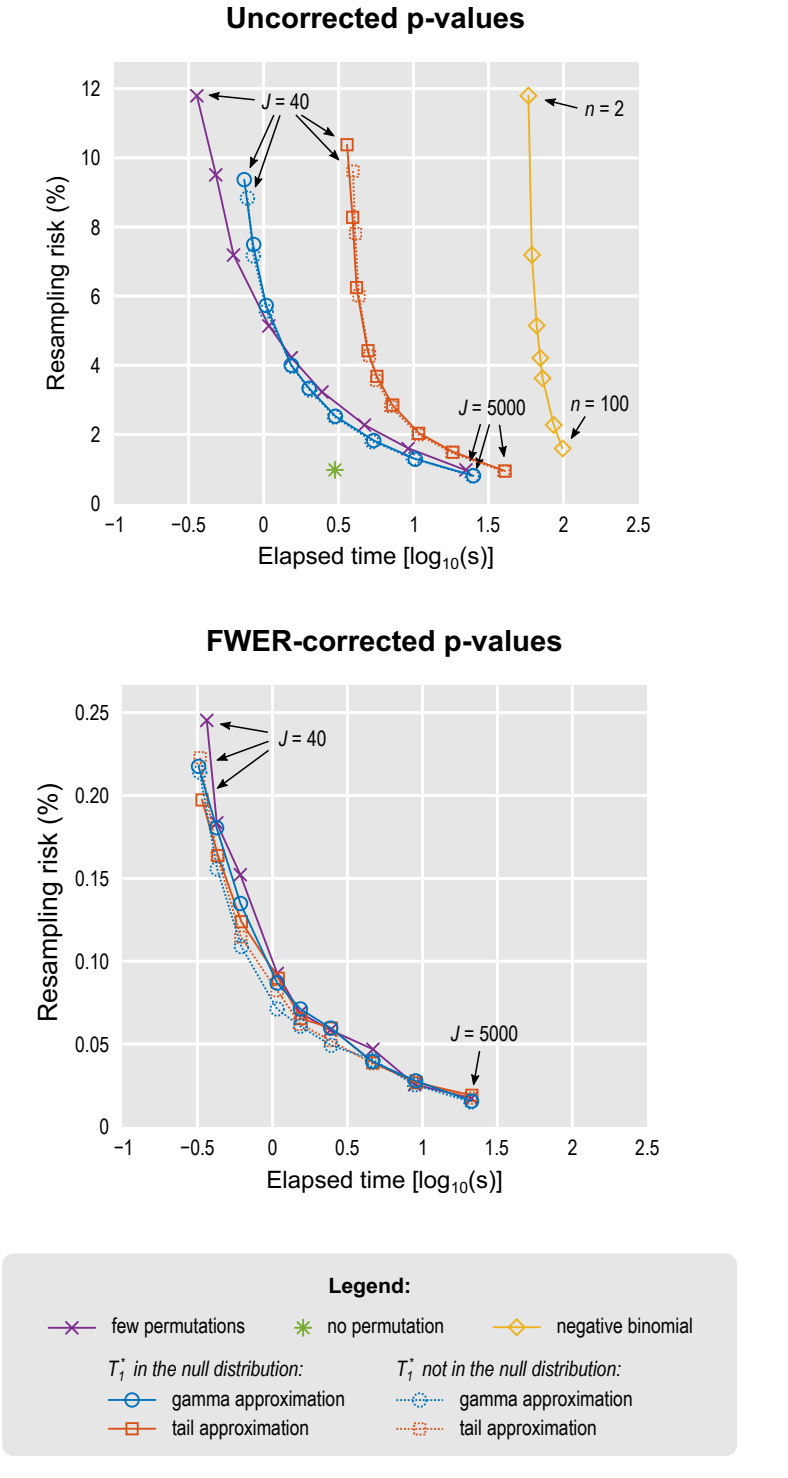
Resampling risk The risk of altering decisions about the rejection of null hypotheses was higher when fewer rearrangements were used for methods where J was varied. This could be observed in both uncorrected and corrected p-values. Removal of T_1^* in the methods that fit a distribution reduced marginally the resampling risk compared with keeping the unpermuted statistic in the distribution, although making the test invalid; in either case, the resampling risk was always smaller than for using only few permutations, with either uncorrected or FWER-corrected p-values. For the negative binomial, resampling risk was higher with fewer exceedances. The method with no permutations yielded the lowest resampling risk overall for the settings assessed. In any case, the resampling risk can be said to have been generally small, and well below 1% for corrected p-values in the simulations. Figure 3.3 shows the trade-off between speed and resampling

risk for the more conservative case in which T_1^* is included in the permutation distribution.

Speed For comparable resampling risks, the method in which no permutations are performed was the absolute fastest. Few permutations, gamma and tail approximations were generally quick, with tail being slower than gamma for the same number of permutations, and gamma slightly slower than few permutations. This considers a voxelwise fit, for uncorrected p-values; if only corrected p-values are needed, the time needed for the single fit of the GPD or gamma for the distribution of extremum statistic is negligible. The negative binomial and, specially, low rank matrix completion were the slowest. Low rank, however, is expected to perform better in settings where there are more tests to be performed (more voxels) than those used in the simulations and real data, and with a relatively smaller sample size (Table 3.4).

Noise distribution and shuffling strategy The performance of the various methods was similar in terms of error rates, power, resampling risk, and speed, regardless of the errors being Gaussian or Weibull (skewed). However, as expected given its assumptions, the method in which no permutations are used did not produce correct results that could be compared with the reference set if the reference set used sign-flippings (for either error distribution), or if the errors were skewed (re-

Figure 3.3: *(page 114)* Balance between resampling risk when compared to a reference set of $J = 50000$ permutations and the respective running time, with the data simulated for Phase II (hence, 100 repetitions, Gaussian noise). Some methods have parameters that could be varied: few permutations, tail approximation and gamma approximation use a certain number of permutations that varied in the simulations as $J = \{40, 60, 100, 200, 300, 500, 1000, 2000, 5000\}$. The negative binomial distribution uses a fixed upper limit on the number of permutations (set as $J = 50000$) and a number of exceedances that varied as $n = \{2, 5, 10, 15, 20, 50, 100\}$. The no permutation method has no parameter to be varied. The low rank matrix completion has the same resampling risk as the few permutations, but the running time is too dependent on the size of the data, hence is not shown. More permutations reduce the resampling risk, but take longer to run.


Uncorrected p-values
Resampling risk (%)
Elapsed time [log10(s)]
J = 40
n = 2
J = 5000
n = 100
FWER-corrected p-values
Resampling risk (%)
Elapsed time [log10(s)]
J = 40
J = 5000
Legend:
few permutations
no permutation
negative binomial
T1* in the null distribution:
gamma approximation
tail approximation
T1* not in the null distribution:
gamma approximation
tail approximation

gardless of the shuffling strategy, i.e., permutations, sign-flippings, or permutations with sign-flippings).

Spatial statistics For all acceleration methods, the behaviour for spatial statistics followed the same trends as for the voxelwise, non-spatial statistics, in terms of error rates, power, agreement with the reference set, and resampling risk.

Multivariate statistics and non-parametric combination Likewise, the results for CMV and for NPC followed similar trends as above, with error rates controlled exactly, and yielding similar power as the reference set, as evidenced by the results of Phase I shown in the Appendix A.3.1.

Real data All methods yielded visually similar maps for the real data, with smaller p-values observable with more permutations for the methods that use permutations, or more exceedances for the negative binomial. In the TFCE, FWER-corrected maps, stronger effects of interest could be revealed by tail and gamma methods for equivalent J of few permutations. These results are remarkably similar to the results seen in the reference set, even using about a hundred times fewer permutations, with proportional increases in speed, as summarised in Figures 3.4 and 3.5, and shown in greater detail in the Appendix A.3.3. The timings refer to the implementation available in PALM, as described in Section 5.2. The acceleration methods worked similarly, and yielded similar increases in speed, for the two levels of smoothing considered.

Figure 3.4: *(page 117)* VBM results, showing **uncorrected** p-value maps (axial slices $z = 10$ and $z = 48$ mm, MNI space), and the overall amount of time taken by each method. The tail and gamma methods generally have higher power compared to few permutations with the same J, even with these not including the unpermuted statistic in the null distribution; see the Appendix A.3.3 for other maps.

3.5 Discussion

3.5.1 Assumptions

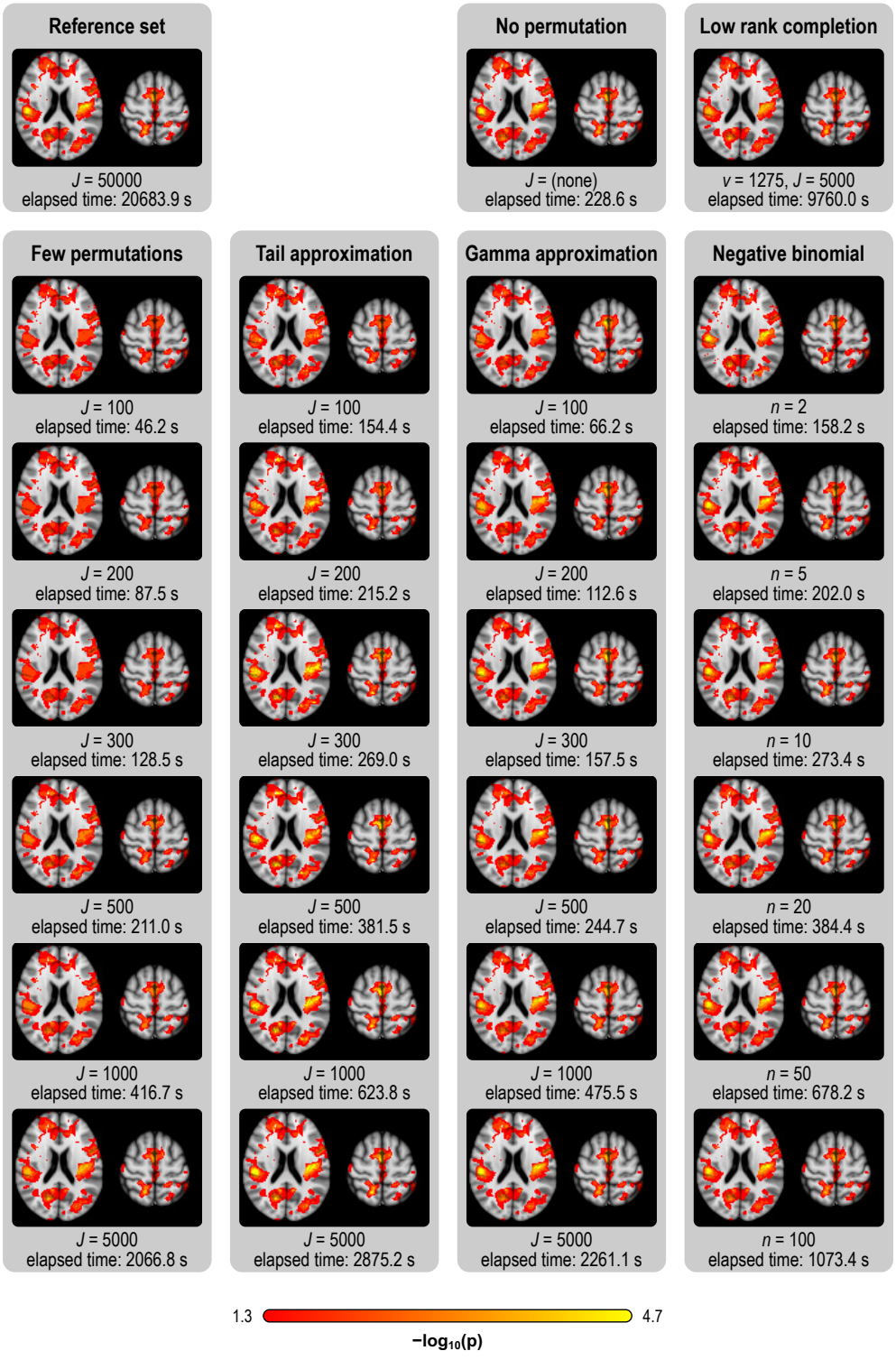
All six methods presented are non-parametric in the sense that they do not depend on the distribution of the test statistic. Some of the methods can still be said to be parametric in that certain parameters need to be estimated, such as for the gamma or for the generalised Pareto distribution, although they remain non-parametric in that the distribution from which these parameters are estimated is based on permutations (or at least conceptually, as in the case of the no permutation method). Some methods nevertheless require certain assumptions: for the gamma approximation, a fit can only be adequate if the distribution of the test statistic is unimodal; for the method in which no permutations are performed, the results are an approximation only to permutations proper, not to sign-flippings, and only if the distribution of the errors is symmetric.

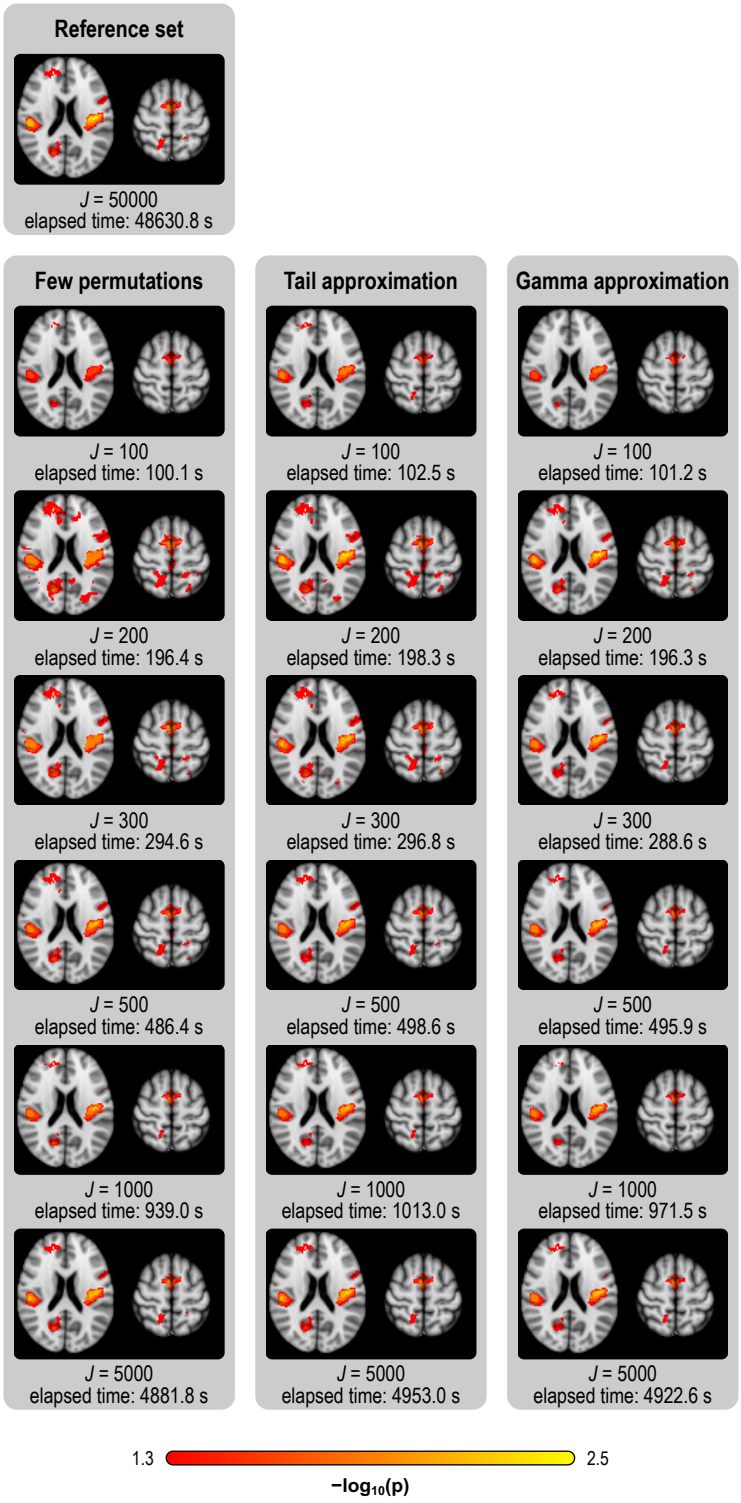
3.5.2 Resampling risk and number of permutations

Although the p-values can vary considerably between the methods, as evidenced by the Bland–Altman plots, at the tails they are remarkably similar, thus allowing similar inferences to be drawn, and presenting an overall low resampling risk for the corrected maps. This means that for most methods, the overall result upon rejection or not of the null hypothesis is expected to remain broadly the same.

The results permit relaxing the usual common sense that more permutations are better. Although more permutations do reduce the resampling risk, the high computational cost may not bring additional information upon acceptance or rejection of the null hypotheses, even considering the large number of tests usually

Figure 3.5: *(page 118)* VBM results, showing **corrected** (FWER) p-value TFCE maps (axial slices $z = 10$ and $z = 48$ mm, MNI space), and the overall amount of time taken by each method. As with the uncorrected, the methods generally have higher power compared to few permutations with the same J, and approximate better the reference set.


Reference set
J = 50000
elapsed time: 20683.9 s
No permutation
J = (none)
elapsed time: 228.6 s
Low rank completion
v = 1275, J = 5000
elapsed time: 9760.0 s
Few permutations
J = 100
elapsed time: 46.2 s
J = 200
elapsed time: 87.5 s
J = 300
elapsed time: 128.5 s
J = 500
elapsed time: 211.0 s
J = 1000
elapsed time: 416.7 s
J = 5000
elapsed time: 2066.8 s
Tail approximation
J = 100
elapsed time: 154.4 s
J = 200
elapsed time: 215.2 s
J = 300
elapsed time: 269.0 s
J = 500
elapsed time: 381.5 s
J = 1000
elapsed time: 623.8 s
J = 5000
elapsed time: 2875.2 s
Gamma approximation
J = 100
elapsed time: 66.2 s
J = 200
elapsed time: 112.6 s
J = 300
elapsed time: 157.5 s
J = 500
elapsed time: 244.7 s
J = 1000
elapsed time: 475.5 s
J = 5000
elapsed time: 2261.1 s
Negative binomial
n = 2
elapsed time: 158.2 s
n = 5
elapsed time: 202.0 s
n = 10
elapsed time: 273.4 s
n = 20
elapsed time: 384.4 s
n = 50
elapsed time: 678.2 s
n = 100
elapsed time: 1073.4 s
1.3
4.7
−log10(p)


Reference set
J = 50000
elapsed time: 48630.8 s
Few permutations
J = 100
elapsed time: 100.1 s
J = 200
elapsed time: 196.4 s
J = 300
elapsed time: 294.6 s
J = 500
elapsed time: 486.4 s
J = 1000
elapsed time: 939.0 s
J = 5000
elapsed time: 4881.8 s
Tail approximation
J = 100
elapsed time: 102.5 s
J = 200
elapsed time: 198.3 s
J = 300
elapsed time: 296.8 s
J = 500
elapsed time: 498.6 s
J = 1000
elapsed time: 1013.0 s
J = 5000
elapsed time: 4953.0 s
Gamma approximation
J = 100
elapsed time: 101.2 s
J = 200
elapsed time: 196.3 s
J = 300
elapsed time: 288.6 s
J = 500
elapsed time: 495.9 s
J = 1000
elapsed time: 971.5 s
J = 5000
elapsed time: 4922.6 s
1.3
2.5
−log10(p)

performed in brain imaging. This is particularly the case for FWER corrected results, for which the resampling risk, even for moderate to small number of permutations, was quite small.

It should be noted that, although more permutations do not intrinsically increase power, they allow smaller p-values to be found (Equation 3.3). Even though p-values much smaller than needed to reach a decision on the null hypotheses may be not needed, such as for FWER correction, methods that use uncorrected p-values as a starting point for further computations, such as for subsequent FDR correction, stand to benefit more from the greater resolution and potentially greater significance of p-values derived with a larger number of permutations. This compounds with more accurate fitting of a distribution, such as the GPD (tail) and gamma, enabled by the larger number of points available in the empirical distribution.

3.5.3 Tail, gamma, and no permutation

For tail and gamma approximations, a small number of permutations is initially performed, from which a low resolution null distribution is built and used for the GPD (tail) or gamma (full distribution) fit. The results show that inclusion or not of the unpermuted test statistic (T_1^*) in this null distribution makes a substantial difference in the uncorrected case if too few permutations are performed, with p-values that, at the tail, are either conservative (if included) or invalid (if not included). Thus, if interest lies solely on uncorrected p-values, such as in the absence of multiple testing, or for subsequent use of FDR, other acceleration methods that do not suffer from either conservativeness or invalidity at the tails are advisable. For FWER-corrected p-values, as the number of tests (voxels) increase, the difference between including or not the unpermuted statistic in the null distribution becomes negligible.

This is not an unexpected finding, particularly for test statistics that happen to be at the tail, such as when there is a true, strong effect of interest: by being at the tail, T_1^* is among the rarest values found with the permutations, hence a single

extra observation of the statistic is considerably influential if too few permutations are done; for test statistics lying towards the mode of the distribution, where most of the other values are located, a single extra observation has little noticeable effect.

These two methods allow p-values to extend further into the tail of the null distribution than otherwise is possible when only few permutations are used, and are particularly useful for the FWER case, offering a complement for the no permutation method, that is available to produce uncorrected p-values. The latter, however, requires both symmetric error terms and that the intercept is entirely contained in **Z**. Tail and gamma approximation can also be used even if the number of permutations is reasonably large (such as 5000), yielding corrected results that are remarkably similar to what would be obtained with far more shufflings.

3.5.4 Low rank matrix completion

Various methods can be considered that could make use of low rank matrix completion. The method proposed here performs completion of two matrices, using the data from potentially far fewer tests (voxels) than those present in an image. While completing two matrices, instead of only one, may seem an undesirable computational cost, by restricting the completion to only matrices that can be constructed through linear operations on the data and model, exact recovery is possible. Therefore, problems with unrecoverable residuals due to imperfect reconstruction of the matrix that stores the statistic itself are eschewed, and no assumptions need to be introduced, such as for *ad hoc* attempts for the recovery of the residuals themselves, or for the characterisation of its parameters. The conditions for completion are easily attainable in brain imaging, and the method produces identical results to those obtained with the conventional permutation test.

The method is expected to perform faster with large images and with small samples, although performance gains also need a fast implementation. The simulations were too expensive to use a sufficiently large image, hence potential advantages of low rank completion could not be illustrated. Yet, the method remains

an option as a potential replacement for simple permutations, and as the initial step for tail and gamma approximations. It has also the benefit that, from the recovered statistics, spatial statistics can further be calculated, although direct recovery of such spatial statistics, that are not linear functions of the data, would require approximated results.

3.5.5 Applicability

Most of the assessed methods are generic and can accommodate many cases of potential interest. In particular, the tail and gamma approximations, as well as few permutations, can be applied in a variety of situations that include univariate and multivariate tests (both CMV and NPC), spatial statistics, and for the correction using the distribution of the extremum statistic (minimum or maximum). The low rank matrix completion, by producing identical result to few permutations, can likewise be considered a generic solution, although its computational benefits only arise for large images and with relatively smaller sample sizes, and even so, only for univariate statistics.

Except for the method in which no permutations are performed, all others can be considered for experiments that use non-independent data, as long as dependencies between observations have been taken into account by means of exchangeability blocks, including multiple levels of exchangeability (Winkler et al., 2015), with the consequence that these acceleration methods can be used for experiments that used repeated measurements, heterogeneous variances, or other types of structured dependencies.

3.5.6 Real data

Using a VBM dataset was especially useful as this imaging method is known to suffer from non-normality, particularly skewness, and spatial non-stationarity, which could pose difficulties. Yet, the acceleration methods performed generally well, and the results of the reanalysis are in line with those of the original study (Douaud

et al., 2007). Of note, at $J = 500$, the tail approximation seemed to produce spatial results closer to the reference set than the gamma approximation, with fewer false positives and importantly fewer false negatives in relation to that set, especially in the left Broca's area and the inferior temporal gyri. Using of any of the acceleration methods that can produce FWER-corrected p-values resulted in the same conclusions about rejection of the null, only with considerable increases in speed. Even though the method in which no permutations are done worked reasonably well with the real and presumably skewed VBM data, it should be noted that assumptions were violated, and this method should not in general be recommended in the presence of skewness.

3.5.7 Recommendations

As a general rule, given its generalisability, its lack of dependence on symmetry or on unimodality of the permutation distribution, the need to consider the multiplicity of tests in brain imaging, its availability not only for univariate tests, but also CMV and NPC, as well as spatial statistics, and in the absence of any reasonable information about the data, the tail approximation can be in general recommended. The gamma approximation can be recommended for the same circumstances, and it tends to be slightly faster than the tail approximation, although it requires that the whole permutation distribution is well behaved, and the assumption that its entirety can be approximated by a gamma distribution.

For uncorrected p-values, and without spatial statistics, if symmetry of the error terms can be assumed, the method in which no permutations are performed can be recommended, given its speed. If symmetry cannot be assumed, negative binomial distribution and tail approximation can be used; for the latter, the unpermuted statistic may be excluded from the null distribution if the number of permutations is large given the significance level (such as about a thousand for an $\alpha = 0.05$, as considered in the Section 3.3), or if the approximation is used for FWER corrected p-values. The low rank matrix completion can be considered when

the number of tests (voxels) is much larger than the number of subjects, as a replacement to the few permutations or to build the initial null distribution before tail or gamma approximations.

As for the number of shufflings to be used, the choice depends on how small the p-value needs to be for a given significance level while maintaining a reasonably small resampling risk. The results seem to indicate that, even without tail or gamma approximations, using about 500 permutations can give stable results for FWER corrected inference, although whenever computational resources are available, more should be considered. The fitting of a GPD or gamma distributions can help with the discreteness that can render FDR conservative. A flow chart summarising these recommendations is shown in Figure 3.6.

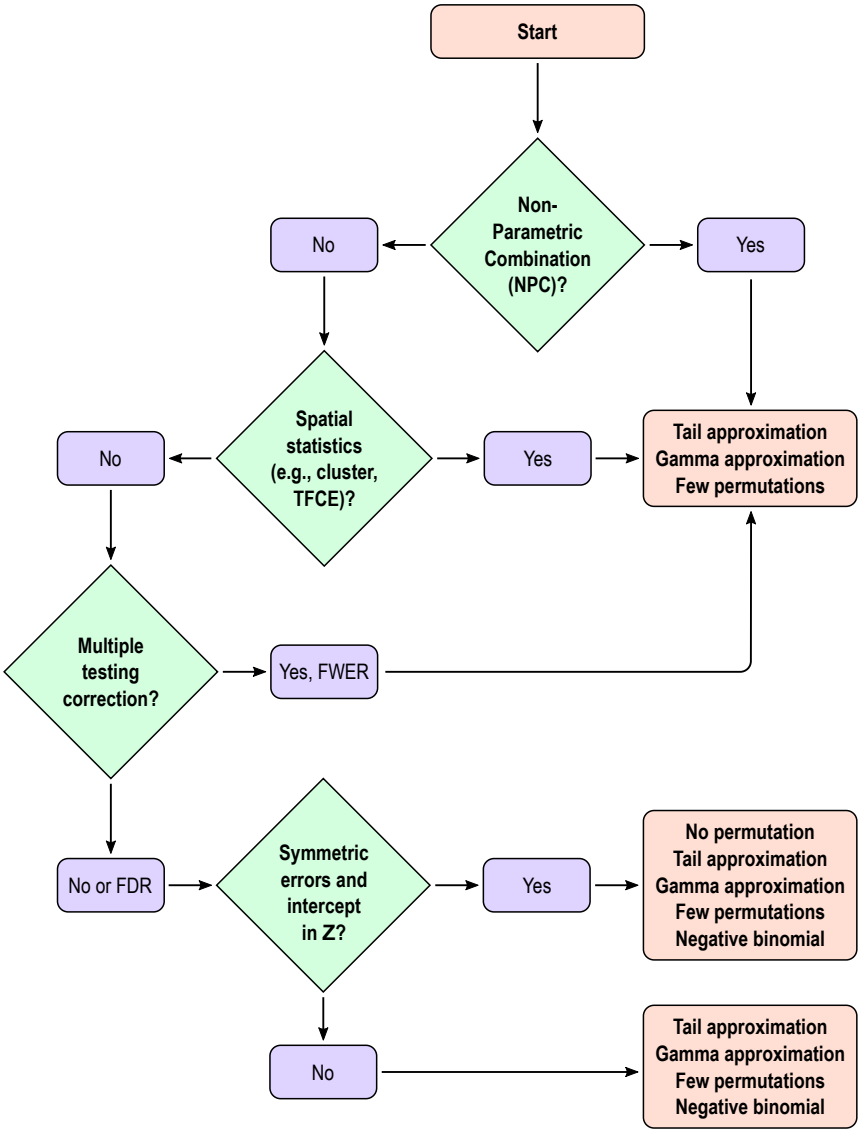


Figure 3.6: Decision tree regarding the various acceleration methods. Each of the terminal boxes show, in order, the preferred methods. For NPC, spatial statistics, or for FWER-corrected p-values, tail and gamma approximations, and few permutations are in general recommendead; gamma is faster than tail fitting, but the latter is more generic. For uncorrected p-values, without spatial statistics, and if the errors can be assumed symmetric, the no permutation method is preferred; if symmetry cannot be assumed, the negative binomial is favoured. The low rank matrix completion (not shown) can be used if $N \ll V$, as a replacement to the few permutations or to build the initial null distribution before tail or gamma approximations.

Chapter 4

Permutation tests for cortical morphometry

4.1 Introduction

It has been suggested that biological processes that drive horizontal (tangential) and vertical (radial) development of the cerebral cortex are separate from each other (Rakic, 1988; Geschwind and Rakic, 2013), influencing cortical area and thickness independently. These two indices of cerebral morphology are uncorrelated genetically (Panizzon et al., 2009; Winkler et al., 2010), are each influenced by regionally distinct genetic factors (Schmitt et al., 2008; Rimol et al., 2010b; Chen et al., 2012, 2015), follow different trajectories over the lifespan (O'Leary et al., 2007; Hogstrom et al., 2013; Fjell et al., 2015), and are differentially associated with cognitive abilities and disorders (Schnack et al., 2015; Noble et al., 2015; Lee et al., 2016; Vuoksimaa et al., 2016). Moreover, it is cortical area, not thickness, that differs substantially across species (Rakic, 1995). These findings give prominence to the use of surface area alongside thickness in studies of cortical morphology and its relationship to function. The literature contains a variety of approaches and terminologies for its assessment and cortical volume, which commingles thickness and area, continues to be a popular metric, thanks largely to the wide availability of

voxel-based morphometry (VBM) (Ashburner and Friston, 2000; Good et al., 2001; Douaud et al., 2007). While analysing cortical thickness and cortical area separately improves specificity (Rimol et al., 2012), it may still be of interest to combine these two measurements so as to increase power to investigate non-specific challenges that could affect thickness and area simultaneously, such as auto-immune (Ceccarelli et al., 2008; Zhang et al., 2016) and infectious neurological disorders (Gitelman et al., 2001; Küper et al., 2011), or in relation to inflammatory markers (Marsland et al., 2008; Zhang et al., 2015). The contributions of this chapter are threefold: we (I) expose certain aspects of the various methods for cortical area analysis, in particular the interpolation between surfaces at different resolutions; (II) propose an improved, analytic measurement of volume; (III) show that a joint analysis using the non-parametric combination NPC (Pesarin and Salmaso, 2010; Winkler et al., 2016c) of thickness and area provides a sensible solution to the investigation of factors that may affect cortical morphology, which can replace the analysis of cortical volume altogether.

4.1.1 Cortical surface area

Using continuous cortical maps to compare surface area across subjects offers advantages over a region-of-interest (ROI) approach, since it does not require that the effects map onto a previously defined ROI scheme. For instance, a group difference present in a subregion of a given ROI may be cancelled out by the absence of a difference in the rest of the ROI and thus remain undetected. Nevertheless, surface area analyses still depend on registration of the cortical surface and interpolation to a common resolution. Such resampling must preserve the amount of area at local, regional and global scales, i.e., it must be mass-conservative. A well known interpolation method is the *nearest-neighbour*, which can be enhanced by correction for stretches and shrinkages of the surface during the registration, as available

Table 4.1: Overview of the four different methods to interpolate surface area and areal quantities. A detailed description is in Section 4.2.

Method	Description
Nearest-neighbour	Nearest-neighbour interpolation of areal quantities on the sphere, followed by Jacobian correction.
Retessellation	Barycentric interpolation on the sphere of the native vertex coordinates.
Redistributive	Vertexwise redistribution of areal quantities based on barycentric coordinates of the source in relation to the target.
Pycnophylactic	Mass-conservative facewise interpolation method that uses the overlapping areas between faces of source and target.

in the function `mris_preproc`, part of the FreeSurfer (FS) software package.[1] Another approach is the *retessellation* of the mesh of each subject to the geometry of a common grid, as proposed by Saad et al. (2004) as a way to produce meshes with similar geometry across subjects, and as available in the package AFNI/SUMA. [2] Even though the method has been mostly used to compute areal expansion, it can be used for surface area itself, as well as for other areal quantities. A third approach is the use of the barycentric coordinates of each vertex with reference to the vertices of the common grid to *redistribute* the areal quantities, in an approximately mass preserving process. Lastly, a strategy for analysis of areal quantities was presented in Winkler et al. (2012) using a *pycnophylactic* (mass-preserving) interpolation method (Table 4.1).

4.1.2 Measuring volume and other areal quantities

The volume of cortical grey matter is also an areal quantity, which therefore requires mass-conservative interpolation methods. Volume can be estimated

[1] Available at freesurfer.net

[2] Available at afni.nimh.nih.gov

through the use of voxelwise partial volume effects using volume-based representations of the brain, such as in VBM, or from a surface representation, in which it can be measured as the amount of tissue present between the surface placed at the site of the pia-mater, and the surface at the interface between gray and white matter. If the area of either of these surfaces is known, or if the area of a mid-surface, i.e., the surface running half-distance between pial and white surfaces (Van Essen, 2005) is known, an estimate of the volume can be obtained by multiplying, at each vertex, area by thickness. This procedure, while providing a reasonable approximation that improves over voxel-based measurements, since it is less susceptible to various artefacts (for a discussion of artefacts in VBM, see Ashburner, 2009), is still problematic as it underestimates the volume of tissue that is external to the convexity of the surface, and overestimates volume that is internal to it; both cases are undesirable, and cannot be solved by merely resorting to using an intermediate surface as the mid-surface (Figure 4.1a). Here a different approach is proposed: each face of the white surface and its matching face in the pial surface are used to define an oblique truncated pyramid, the volume of which is computed analytically, without introducing additional error other than what is intrinsic to the placement of these surfaces (Figure 4.1b for a 2-D schema and Figure 4.2 for 3-D).

Quantitative measurements, such as from positron emission tomography (PET), cerebral blood flow, cerebral blood volume, the mass, or number of molecules of a given compound (Leahy and Qi, 2000; van den Hoff, 2005), are all areal quantities whenever these are expressed in absolute quantities. Likewise, cerebral blood flow and volume obtained using methods based on magnetic resonance imaging (MRI), such as arterial spin labelling (ASL), as well as other forms of quantitative MRI, as those involving contrast enhancement (Parker and Padhani, 2003), quantitative magnetisation transfer (Levesque et al., 2010; Harrison et al., 2015), or quantitative assessment of myelination, are also areal quantities that require mass conservation when measured in absolute terms. The methods used for statistical analysis surface area can be applied for these areal quantities as well.

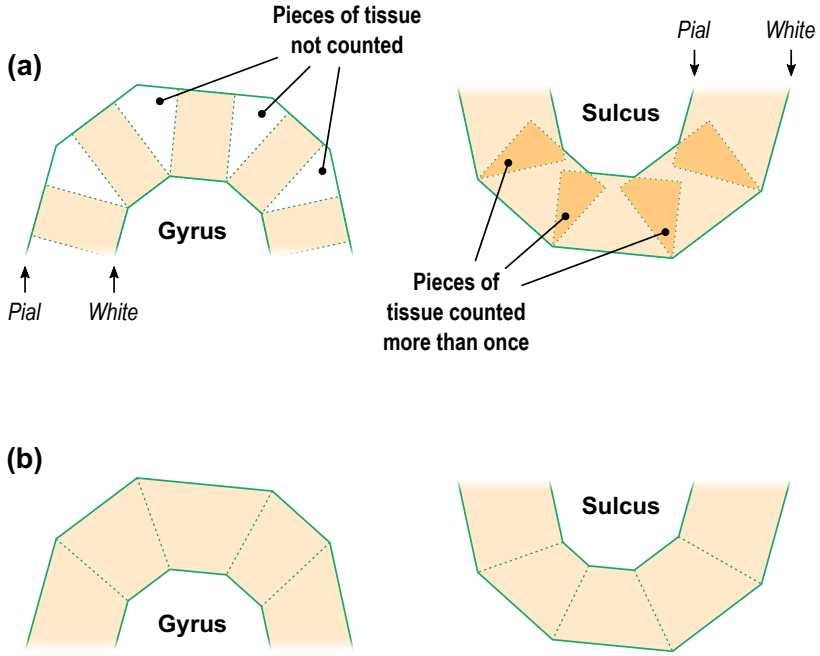


Figure 4.1: A diagram in two dimensions of the problem of measuring the cortical volume. (*a*) If volume is computed using multiplication of thickness by area, considerable amount of tissue is left unmeasured in the gyri, or measured repeatedly in sulci. The problem is minimised, but not solved, with the use of the mid-surface. (*b*) Instead, vertex coordinates can be used to compute analytically the volume of tissue between matching faces of white and pial surfaces, leaving no tissue under- or over-represented.

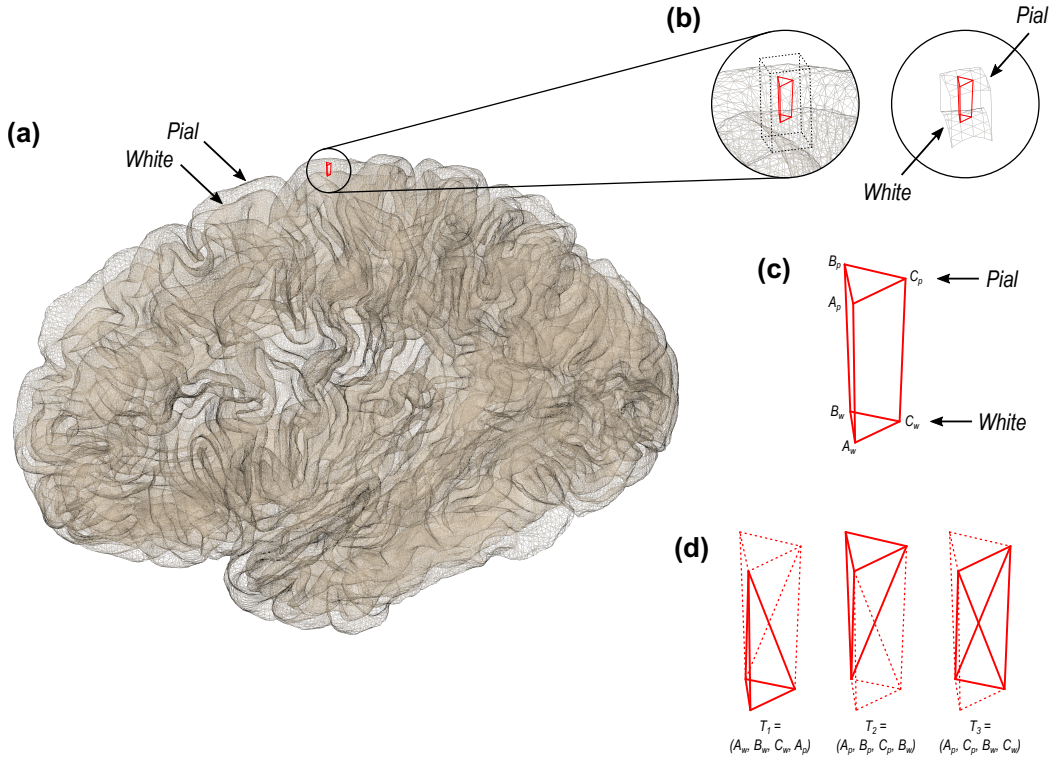


Figure 4.2: A 3-D diagram with the proposed solution to measure the cortical volume. (*a*) In the surface representation, the cortex is limited internally by the white and externally by the pial surface. (*b*) and (*c*) These two surfaces have matching vertices that can be used to delineate an oblique truncated triangular pyramid. (*d*) The six vertices of this pyramid can be used to define three tetrahedra, the volumes of which are computed analytically.

4.1.3 Non-parametric combination (NPC)

Instead of volume, we suggest that a joint approach can be used for the analysis of thickness and area. Classical multivariate tests such as MANCOVA, however, are not informative about the direction of any observed effects, and are based on assumptions that are known not to hold for surface area, such as normality. Rather, the permutation-based non-parametric combination (NPC) (Pesarin and Salmaso, 2010; Winkler et al., 2016c) provides a test for directional as well as two-tailed hypotheses, and is based on minimal assumptions, mainly that of exchangeability, that is, swapping one datum for another keeps the data just as likely. The NPC consists of, in a first phase, testing separately hypotheses on each available metric using permutations that are performed in synchrony; these tests are termed *partial tests*. The resulting statistics for each and every permutation are recorded, allowing an estimate of the complete empirical cumulative distribution function (cdf) to be constructed for each one. In a second phase, the empirical p-values for each test are combined, for each permutation, into a *joint statistic*. As the joint statistic is produced from the previous permutations, all of which have been recorded, an estimate of its empirical cdf function is immediately known, and so is its corresponding p-value (Pesarin and Salmaso, 2010).

As originally proposed, and as described above, NPC is not practicable in brain imaging: as the statistics for all partial tests for all permutations need to be recorded, an enormous amount of space for data storage is necessary. However, even if storage space were not a problem, the discreteness of the p-values for the partial tests is problematic when correcting for multiple testing, because with thousands of vertices in a surface, ties occur frequently, further causing ties among the combined statistics. If too many tests across an image share the same most extreme statistic, correction for the multiplicity, while still valid, is 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 stor-

Table 4.2: Some common combining functions that can be used with NPC. For a more comprehensive list, see Winkler et al. (2016c).

Function	Test statistic
Tippett (1931)	$\min(p_k)$
Fisher (1932)	$-2\sum_{k=1}^{K}\ln(p_k)$
Stouffer et al. (1949)	$\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\Phi^{-1}(1-p_k)$
Mudholkar and George (1979)	$\frac{1}{\pi}\sqrt{\frac{3(5K+4)}{K(5K+2)}}\sum_{k=1}^{K}\ln\left(\frac{1-p_k}{p_k}\right)$

The combining functions are shown as function of the p-values for the partial tests, but for certain methods, 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, Φ^{-1} is the probit function

age. The solution to this problem is loosely based on the direct combination of the test statistics, by converting the statistics of the partial tests to values that behave as p-values using the asymptotic distribution of the statistics, and using these for the combination (Winkler et al., 2016c).

Combining functions The null hypothesis of the NPC is that the null hypotheses for all partial tests are true, and the alternative is that any test is false, which is the same that a union-intersection test (UIT) (Roy, 1953). The rejection region depends on how the combined statistic is produced. Various combining functions can be considered, particularly those used in meta-analyses, such as those listed in Table 4.2. These and most other combining functions, related statistics and their distributions were originally derived under the assumption of independence among the partial tests, which is not always valid, particularly under the tenable hypothesis of shared environmental effects affecting both area and thickness. Such lack of independence is not a problem for NPC: the synchronised permutations implicitly capture the dependencies among the tests that would cause a parametric combination to be invalid, even if using the same combining functions.

4.2 Methods

The general workflow for surface-based morphometry consists of the generation of a surface-representation of the cortex and its subsequent homeomorphic transformation into a sphere. Vertices of this sphere are shifted tangentially along its surface to allow alignment matching a particular feature of interest of a reference brain (i.e., an atlas), such as sulcal depth, myelin content, or functional markers. Once registration has been done, interpolation to a common grid is performed; it is at the resolution of this grid that analyses across subjects are performed. While the order of these processing stages remains generally fixed, the stage in which areal quantities are calculated or obtained varies according to the method: for the nearest neighbour, redistributive, and pycnophylactic methods, these are computed in the native space, using native geometry. With the retessellation method, area is computed in native space, with a new geometry produced after interpolation of the surface coordinates to the common grid. An overview of the whole process is in Figure 4.3.

We evaluate (I) if and where the four different interpolation methods (nearest neighbour, retessellation, redistributive and pycnophylactic) differ; (II) if and where these methods vary according to the resolution of the common grid used as target; (III) if and where the two ways of measuring volumes (the product method and the analytic method) differ from each other; (IV) how the NPC results would relate to the separate analyses of thickness, area and volume.

4.2.1 Subjects

In the period 1986–88, 121 preterm newborns with very low birth weight (VLBW; $\leqslant$ 1500g) were admitted to the Neonatal Intensive Care Unit at the St. Olav University Hospital in Trondheim, Norway. At age 20, a total of 41 VLBW subjects consented to participate and had usable MRI data. The term-born controls were born at the same hospital in the same period. A random sample of women with parities 1 or 2 was selected for follow-up during pregnancy. At birth, 122 chil-

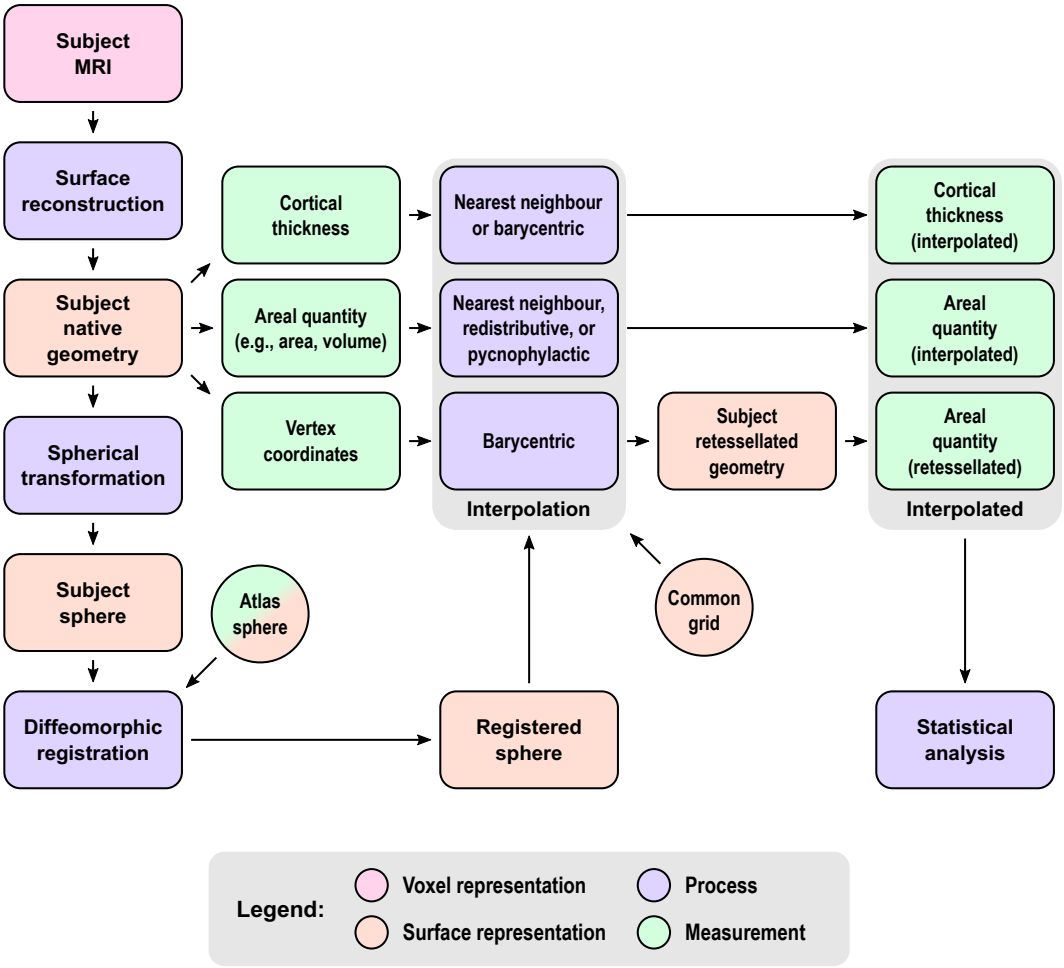


Figure 4.3: An overview of the steps for the analysis of surface area and thickness using different resampling methods. The subject magnetic resonance images are used to reconstruct a pair of surfaces (pial and white) representing the cortex, which initially are in the subject space and individual geometry. From this pair of surfaces, cortical thickness can be measured. From the same surfaces, area and volume can be measured. Finally, the coordinates of the vertices can be stored for subsequent use. The subject native surfaces are homeomorphically transformed to a sphere, registered to a spherical atlas, and used for the interpolation, which for thickness can be either nearest neighbour or barycentric, for area can be nearest neighbour, redistributive or pycnophylactic, and for the vertex coordinates can be barycentric. In the latter, the interpolation of coordinates allows the construction of a new retessellated surface in subject space, from which area can alternatively be measured. The interpolated quantities are then ready to undergo statistical analyses. See references in the main text.

Table 4.3: Characteristics of the sample of participants of the study; additional details can be found in Martinussen et al. (2005); Skranes et al. (2007).

Clinical characteristics	VLBW	Controls
Birth weight, in grams	1232.1 $\pm$ 221.9	3693.3 $\pm$ 504.5
Gestational age, in weeks	29.3 $\pm$ 2.51	39.8 $\pm$ 1.28
Age at scanning, in years	20.1 $\pm$ 0.77	20.3 $\pm$ 0.52
Sex (M/F)	23/18	35/16

dren with birth weight above the tenth percentile for gestational age from this sample were included as controls. At age 20, a total of 59 control subjects consented to participate and had usable MRI data. Details about the sample are provided in Table 4.3, additional details can be found in Martinussen et al. (2005); Skranes et al. (2007).

4.2.2 Data acquisition

Magnetic resonance imaging was performed on a 1.5 T Siemens MAGNETOM Symphony scanner equipped with a quadrature head coil. In each scanning session, two sagittal T_1-weighted magnetization prepared rapid gradient echo (MPRAGE) scans/sequences were acquired (echo time = 3.45 ms, repetition time = 2730 ms, inversion time = 1000 ms, flip angle = 7°; field of view = 256 mm, voxel size = 1 $\times$ 1 $\times$ 1.33 mm, acquisition matrix 256 $\times$ 192 $\times$ 128).

4.2.3 Reconstruction of the cortical surface

We used the method implemented in the FreeSurfer software package (version 5.3.0; Dale et al., 1999; Fischl et al., 1999b): T_1-weighted images are first corrected for magnetic field inhomogeneities and then skull-stripped (Ségonne et al., 2004). Voxels belonging to the white matter (WM) are identified based on their locations, on their intensities, and on the intensities of the neighbouring voxels. A mass of connected WM voxels is produced for each hemisphere, using a six-neighbours con-

nectivity scheme, and a mesh of triangular faces is tightly built around this mass, using two triangles for each externally facing voxel side. The mesh is smoothed taking into account the local intensity in the original images (Dale and Sereno, 1993), at a subvoxel resolution. Defects are corrected (Fischl et al., 2001; Ségonne et al., 2007) to ensure 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, between the gray matter and the cerebrospinal fluid, maintaining constraints on its smoothness while preventing self-intersection. Cortical thickness is measured as the distance between the matching vertices of these two surfaces (Fischl and Dale, 2000).

4.2.4 Measurement of areal quantities

Areal quantities are measured in native space, i.e., before spherical transformation and registration. For the retessellation method, the measurement is made in native space after the surface has been reconstructed to a particular resolution; for nearest neighbour, redistributive, and pycnophylactic, measurement uses native space, with the original, subject-specific mesh geometry.

Cortical area For a triangular face ABC of the surface representation, with vertex coordinates $\mathbf{a} = [x_A \;\; y_A \;\; z_A]'$, $\mathbf{b} = [x_B \;\; y_B \;\; z_B]'$, 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 the bars $|\ |$ represent the vector norm. Although the area per face (i.e., the *facewise* area) can be used in subsequent steps, it remains the case that most software packages can only deal with values assigned to each vertex of the mesh (i.e., *vertexwise*). Conversion from facewise to vertexwise is achieved by assigning to each vertex one-third of the sum of the areas of all faces that have that vertex in

common (Winkler et al., 2012).

Cortical volume The conventional method for computing surface-based volume consists of computing the area at each vertex as above, then multiplying this value by the thickness at that vertex, in a procedure that leaves tissue under- or over-represented in gyri and sulci (Figure 4.1). Instead, volumes can be computed using the three vertices that define a face in the white surface and the three matching vertices in the pial surface, defining an *oblique truncated triangular pyramid*, which in turn is perfectly subdivided into three tetrahedra. The volumes of these are computed analytically, summed, and assigned to each face of the surface representation, viz.:

1. For a given face $A_w B_w C_w$ in the white surface, and its corresponding face $A_p B_p C_p$ in the pial surface, define an oblique truncated triangular pyramid.

2. Split this truncated pyramid into three tetrahedra, defined as:

$$\begin{aligned} T_1 &= (\ A_w,\ B_w,\ C_w,\ A_p\ ) \\ T_2 &= (\ A_p,\ B_p,\ C_p,\ B_w\ ) \\ T_3 &= (\ A_p,\ C_p,\ C_w,\ B_w\ ) \end{aligned}$$

   This division leaves no volume under- or over-represented.

3. For each such tetrahedra, let **a**, **b**, **c** and **d** represent its four vertices in terms of coordinates $[x\ y\ z]'$. Compute the volume as $|\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})|/6$, where $\mathbf{u} = \mathbf{a} - \mathbf{d}$, $\mathbf{v} = \mathbf{b} - \mathbf{d}$, $\mathbf{w} = \mathbf{c} - \mathbf{d}$, the symbol $\times$ represents the cross product, $\cdot$ represents the dot product, and the bars $|\ |$ represent the vector norm.

Computation can be accelerated by setting $\mathbf{d} = A_p$, the common vertex for the three tetrahedra, such that the vector subtractions can happen only once. Conversion from facewise volume to vertexwise is possible, and done in the same manner as for facewise area. The above method will be the default in the next FreeSurfer release.

4.2.5 Spherical transformation

The white surface is homeomorphically transformed to a sphere (Fischl et al., 1999c), thus keeping a one-to-one mapping between faces and vertices of the native geometry (white and pial) and the sphere. All these surfaces comprise triangular faces exclusively. Measurements of interest obtained from native geometry or in native space, such as area and thickness, are stored separately and are not affected by the transformation, nor by registration (next step; see diagram in Figure 4.3).

4.2.6 Registration

Various strategies are available to place all surfaces in register and allow inter-subject comparisons, including the ones used by FreeSurfer (Fischl et al., 1999c), Spherical Demons (SD) (Yeo et al., 2010), Multimodal Surface Matching (MSM) (Robinson et al., 2014), among others. Methods that are diffeomorphic (i.e., smooth and invertible) should be favoured. Methods that are not diffeomorphic by design but in practice produce invertible and smooth warps can, in principle, be used in registration for areal analyses. In the present analyses, FreeSurfer was used; a side comparison with SD is in Appendix A.4.1).

4.2.7 Interpolation methods

Statistical comparisons require meshes with a common resolution where each point (vertex, face) represents homologous locations across individuals. One type of mesh that can act as a common grid is a geodesic sphere constructed by iterative subdivision of the faces of a regular (Platonic) icosahedron. A geodesic sphere has many advantages as the target for interpolation: 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 (Kenner, 1976). We compared four different interpolation methods each at three different mesh resolutions: IC3 (lowest resolution, with 642 vertices and 1280 faces), IC5 (intermediate resolution, with

10242 vertices and 20480 faces), and IC7 (163842 vertices and 327680 faces).

Nearest neighbour interpolation The well known nearest neighbour interpolation does not guarantee preservation of areal quantities, although modifications can be introduced to render it approximately mass conservative: for each vertex in the target, the closest vertex is found in the source sphere, and the area from the source vertex is assigned to the target vertex; if a given source vertex maps to multiple target vertices, its area is divided between them so as to preserve the total area. If there are any source vertices that have not been represented in the target, for each one of these, the closest target vertex is located and the corresponding area from the source surface is incremented to any area already stored on it. This method ensures that total area remains unchanged after mapping onto the group surface. This process is a surface equivalent of Jacobian correction[3] used in volume-based methods in that it accounts for stretches and shrinkages while preserving the overall amount of areal quantities. Nearest neighbour interpolation is currently the default method in FreeSurfer.

Retessellation of the native geometry This method appeared in Saad et al. (2004). It consists of generating a new mesh by interpolating not the area assigned to vertices but the coordinates of the corresponding vertices in the native geometry. The set of three coordinates is used, together with the connectivity scheme between vertices from the common grid, to construct a new mesh that has similar overall shape as the original brain, but with the geometry of the common grid. The area for each face (or vertex) can be computed from this new mesh and used for statistical comparison across subjects. Equivalently, the coordinates of each vertex can be treated as a single vector and the barycentric interpolation can be

[3] Not to be confused with the computation of the Jacobian itself, that is defined, for the i-th vertex, as $J_i = \frac{A_i^S}{A_i^w} \frac{\sum_i A_i^w}{\sum_i A_i^S}$, where A_i^S is the area of the vertex in the source (registered) sphere, A_i^w is the area of the same vertex in the white surface (native space and native geometry), and the sums are over the entire surface, i.e., all vertices.

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. From the four methods considered in this chapter, this is the only one that does not directly interpolate either area or areal quantities, but the mesh in native space.

Redistribution of areas This method works by splitting the areal quantity present at each vertex in the source sphere using the proportion given by the barycentric coordinates of that vertex in relation to the face in the target sphere (common grid) on which it lies, redistributing these quantities to the three vertices that constitute that face in the target. If some quantity was already present in the target vertex (e.g., from other source vertices lying on the same target face), that quantity is incremented. The method is represented by:

$$Q_i^T = \sum_{f=1}^{F} \sum_{v=1}^{V_f} Q_{vf}^S \delta_{ivf}$$

where Q_{vf}^S is the areal quantity in the source vertex v, $v \in \{1, \ldots, V_f\}$ lying on the target face f, $f \in \{1, \ldots, F\}$, F being the number of faces that meet at the target vertex i, and δ_{ivf} is the barycentric coordinate of v, lying on face f, and in relation to the target vertex i. This method has similarities with the conventional barycentric interpolation (as used for the interpolation of coordinates in the retesselation method). The key difference is that in the barycentric interpolation, it is the barycentric coordinates of the target vertex in relation to their containing source face that are used to weight the quantities, in a process that therefore is not mass conservative. Here it is the barycentric coordinates of the source ver-

tex in relation to their containing target face that are used; the quantities are split proportionately, and redistributed across target vertices.

Pycnophylactic interpolation The ideal interpolation method should conserve the areal quantities globally, regionally and locally. In other words, the method has to be *pycnophylactic*. 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 (Markoff and Shapiro, 1973; Winkler et al., 2012). The pycnophylactic method operates on the faces directly, not on vertices and the area (or any other areal quantity) is transferred from source to target surface via weighting by the overlapping area between any pairs of faces. The interpolated areal quantity, Q_i^T, of a face i in the target surface, that overlaps with F faces from the source surface, is given by:

$$Q_i^T = \sum_{f=1}^{F} \frac{A_f^O}{A_f^S} Q_f^S$$

where A_f^S is the area of the f-th overlapping face from the source sphere, which contains a quantity Q_f^S of some areal measurement (such as the surface area measured in the native space), and A_f^O is the overlapping area with the face i.

Correction for unequal face sizes and smoothing Regardless of the interpolation method used, larger faces in the common grid inherit larger amounts of areal quantities. If the analysis will compare regions that are topographically distinct, or if the data are to be smoothed, a correction for different face sizes is needed (Winkler et al., 2012). For facewise data, such a correction consists of weighting the the areal quantity at each face or vertex, after interpolation, by a constant that depends on the respective area in the common grid. Smoothing was considered at two levels for the comparison of areal interpolation and volume methods: no smoothing, and smoothing with a Gaussian kernel with full width at half maximum (FWHM) of 10 mm, small so as to preserve the effect of different resolutions

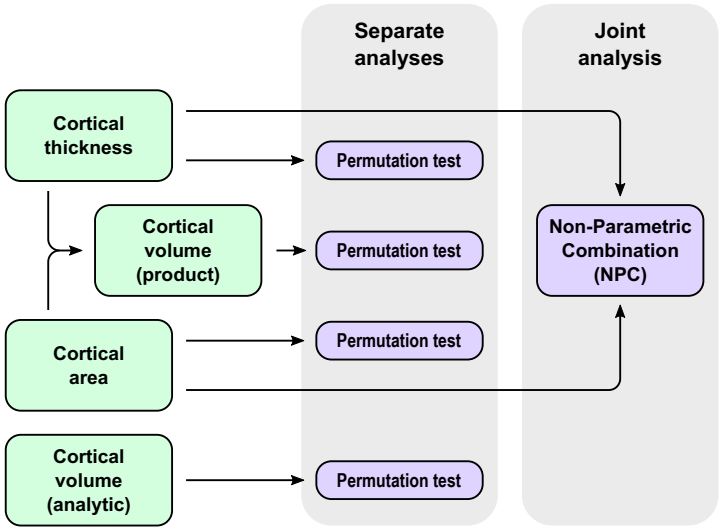


Figure 4.4: Overview of the separate and joint analyses of thickness, area and volume.

being investigated. For the comparison between VLBW and controls, 30 mm, as in Skranes et al. (2013).

4.2.8 Statistical analysis

The statistical analysis was performed using the tool Permutation Analysis of Linear Models (PALM; Winkler et al., 2014, 2016c, see also Section 5.2). The number of permutations was set to 1000, followed by approximation of the tail of the distribution by a generalised Pareto distribution (GPD; Winkler et al., 2016b, see also Chapter 3), and familywise error rate correction (FWER) was applied considering both hemispheres and both test directions for the null hypothesis of no difference between the two groups. Analyses were performed separately for cortical thickness, area, and volume (both methods), and also using NPC for the joint analysis of thickness and area; see diagram in Figure 4.4 .

4.2.9 Presentation of results

The large number of scenarios evaluated, that involved two different registration and four different interpolation methods, three grid resolutions, two different

smoothing levels, four different indices of cortical morphology, plus NPC, resulted in more than 16 thousand maps and Bland-Altman plots (Bland and Altman, 1986). These have been organised in a set of browsable pages accessible online; see details in Appendix A.

4.3 Results

4.3.1 Preservation of areal quantities

All methods preserve generally well the global amount of surface area, and therefore, of other areal quantities, at the highest resolution of the common grid (IC7). At lower resolutions, massive amounts of area are lost with the retessellation method: about 40% on average for IC3 (lowest resolution, with 642 vertices and 1280 faces) and 9% for IC5 (intermediate resolution, with 10242 vertices and 20480 faces), although only 1% for IC7 (163842 vertices and 327680 faces). Areal losses, when existing, tend to be uniformly distributed across the cortex (Figure 4.5, upper panels), with no trends affecting particular regions and, except for retessellation, can be substantially alleviated by smoothing (Appendix A.4.1).

4.3.2 Differences between interpolation methods

While there are no spatial trends in terms of areal gains or losses, the inexactness of the non-pycnophylactic interpolation methods introduces noise that substantially reduces their correlation when assessed between subjects (Figure 4.5, lower panels). The only exception is between the retessellation and the pycnophylactic methods, which have near perfect correlation even without any smoothing. Smoothing increases the correlation between all methods to near unity throughout the cortex (Appendix A.4.1*a*). At the subject level, the spatial correlation between the nearest neighbour and the pycnophylactic methods is only about 0.60, although approaching unity when the subjects are averaged (Appendix A.4.1*b*). Smoothing leads to a dramatic improvement on agreement, causing nearest neighbour

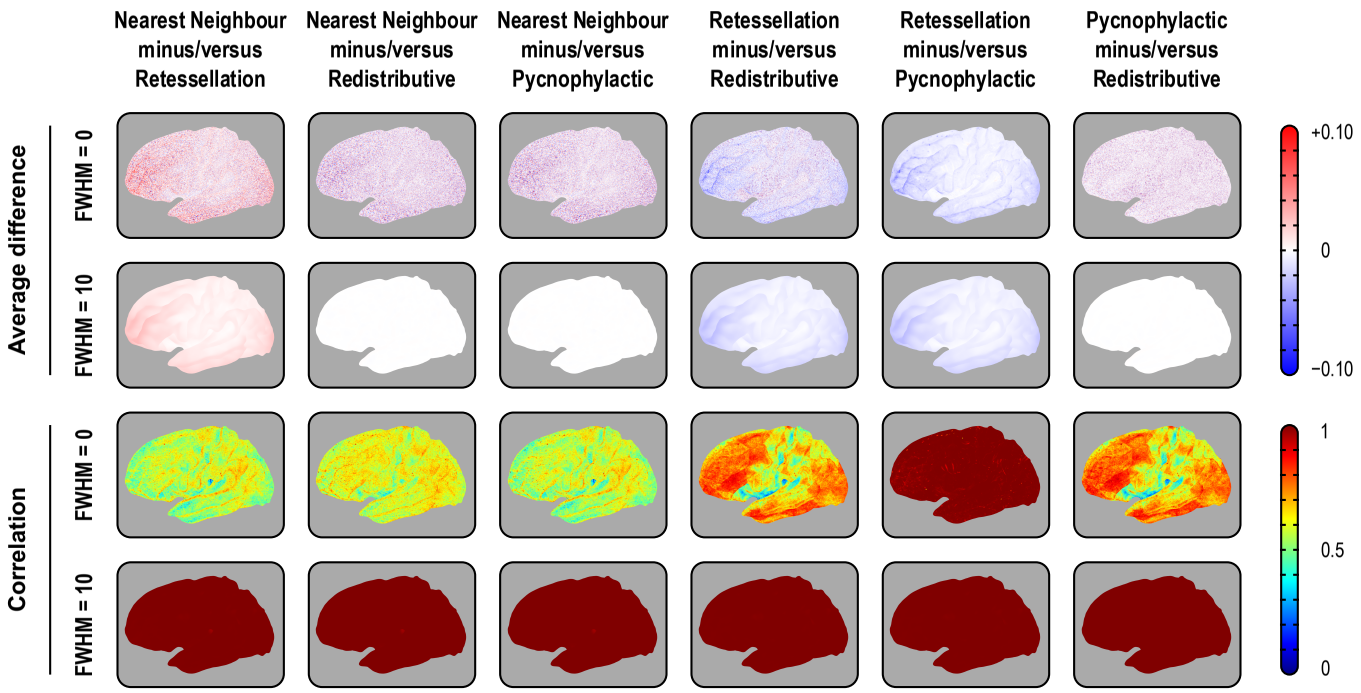


Figure 4.5: Pairwise average differences (in mm^2) and correlations between the four resampling methods, using the IC7 as target, with or without smoothing with a Gaussian kernel of FWHM = 10 mm, projected to the average white surface. Although the four methods differ, with some leading to substantial, undesirable losses and gains in surface area, and the introduction of noise manifested by lower correlations, the average variation was zero for nearest neighbour, redistributive and pycnophylactic. The retessellation method led to substantial losses of area that could not be recovered or compensated by blurring. Although this method showed excellent correlation with pycnophylactic, quantitative results after interpolation are biased downwards. For the medial views, for the right hemisphere, for IC3 and IC5, and for projections to the pial and inflated surfaces, consult Appendix A.4.

to be nearly indistinguishable from the pycnophylactic method. The redistributive method performed in a similar manner, although with a higher correlation without smoothing, i.e., about 0.75 (Appendix A.4.1*b*).

4.3.3 Cortical volume measurements

At the local scale, differences between the product and the analytic methods of volume estimation are as high as 20% in some regions (Appendix A.4.2), an amount that could not be alleviated by smoothing or by changes in resolution. As predicted by Figure 4.1, differences were larger in the crowns of gyri and depths of sulci, in either case with the reverse polarity (Figure 4.6, upper panels). The vertexwise correlation between the methods across subjects, however, was in general very high, approaching unity throughout the whole cortex, with or without smoothing, and at different resolutions. In regions of higher sulcal variability, however, the correlations were not as high, sometimes as low as 0.80, such as in the insular cortex and at the confluence of parieto-occipital and calcarine sulci, between the lingual and the isthmus of the cingulate gyrus (Figure 4.6, lower panels). At least in the case of the insula, this effect may be partly attributed to a misplacement of the white surface in the region lateral to the claustrum (Glasser et al., 2016).

4.3.4 Global measurements and their variability

Average global cortical area, thickness, and volume (using both methods) across subjects in the sample are shown in Table 4.4. Cortical volumes assessed with the multiplicative method are significantly higher ($p < 0.0001$) than using the analytic method. Variability for area is higher than for thickness, and even higher for volume: the average coefficient of variation across subjects ($100 \cdot \sigma/\mu$) was, respectively, 9.9%, 3.2% and 10.5%, after adjusting for the variables group, age, and sex (that is, computing the global mean in a general linear model that included these variables in the design, and using the standard deviation of the residuals). The parietal region (bilateral) was the most variable for all measurements; cor-

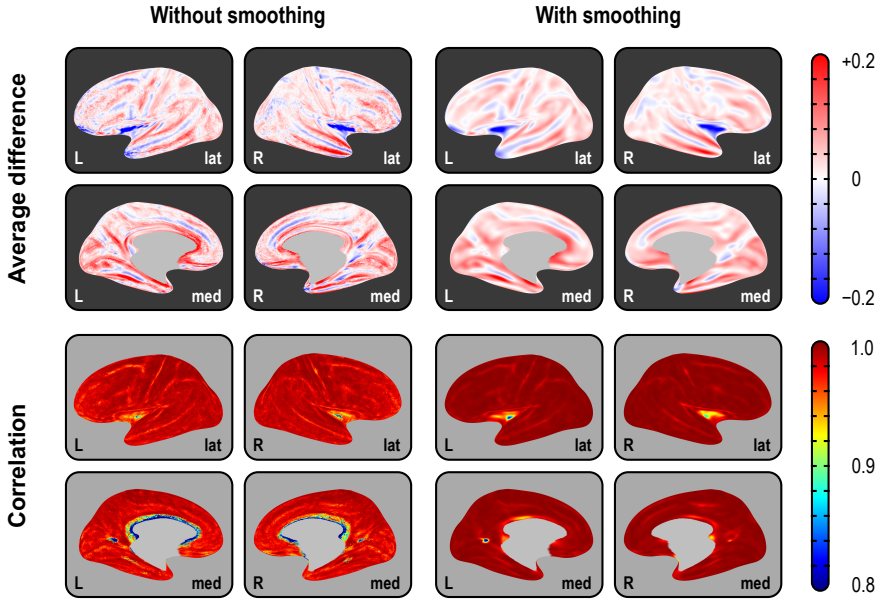


Figure 4.6: Average difference (in mm^3) between the two methods of assessing volume and their correlation (across subjects), using the highest resolution (IC7) as the interpolation target, projected to the average inflated surface. As predicated from Fig. 4.1, differences are larger in the crowns of gyri and in the depths of sulci, with gains/losses in volume in these locations following opposite patterns. Although the correlations tend to be generally high, and increase with smoothing, they are lower in regions of higher inter-individual morphological variability, such as at the anterior end of the cuneus, and in the insular cortex. For IC3 and IC5, and for projections to the white and pial surfaces, consult Appendix A.4.

Table 4.4: Average $\pm$ standard deviation of area (in mm^2), thickness (in mm) and volume (in mm^3) across subjects. Volumes are shown assessed using the multiplicative (m) and analytic (a) methods, as well as their difference.

Measure	Left hemisphere	Right hemisphere	Both hemispheres
Area	97104.3 ± 9594.8	97767.7 ± 9684.4	194872.0 ± 19247.6
Thickness	2.5357 ± 0.0951	2.5273 ± 0.0914	2.5314 ± 0.0921
Volume$^{(m)}$	246268.9 ± 26416.7	247131.0 ± 26529.9	493399.9 ± 52855.7
Volume$^{(a)}$	242580.3 ± 26141.4	243688.3 ± 26214.0	486268.6 ± 52266.3
Difference$^{(m-a)}$	3688.6 ± 569.2	3442.7 ± 605.8	7131.4 ± 1087.2

responding spatial maps are shown in Figure 4.7. Correlation and Bland–Altman plots are shown in Appendix A.4.3.

4.3.5 Differences between VLBW and controls

Analysing cortical thickness and area separately, the comparisons between the VLBW subjects and the controls suggest a distinct pattern of significant differences ($p \leqslant 0.05$, FWER-corrected). Surface area maps show a significant bilateral reduction in the middle temporal gyrus, the superior banks of the lateral sulcus, and the occipito-temporal lateral (fusiform) gyrus, as well as a diffuse bilateral pattern of areal losses affecting the superior frontal gyrus, posterior parietal cortex and, in the right hemisphere, the subgenual area of the cingulate cortex. Cortical thickness maps suggest a diffuse bilateral thinning in the parietal lobes, left middle temporal gyrus, right superior temporal sulcus, and bilateral thickening of the medial orbito-frontal cortex of the VLBW subjects compared to controls (Figure 4.8, upper panels, light blue background). Maps of cortical volume differences largely mimic the surface area results, although with a few differences: diffuse signs of volume reduction in the parietal lobes, ascribable to cortical thinning and, contrary to the analysis of area and thickness, no effects found in the medial-orbitofrontal or in the subgenual region of the cingulate gyrus (Figure 4.8, middle panels, light red background).

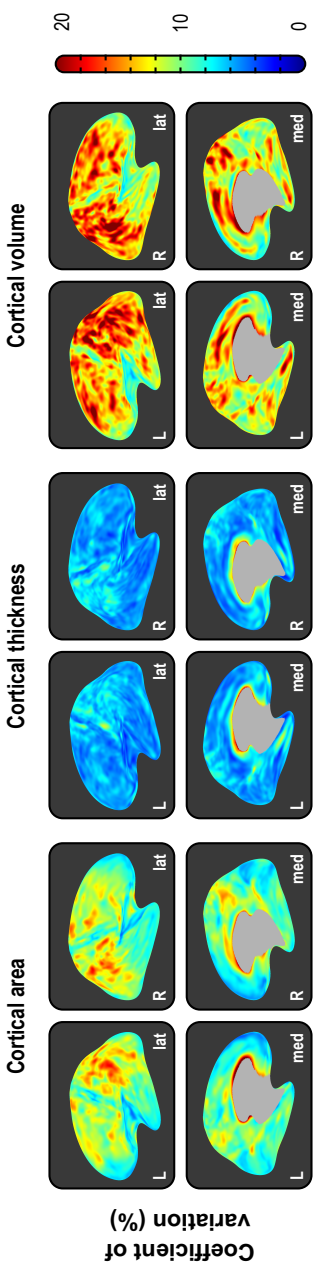
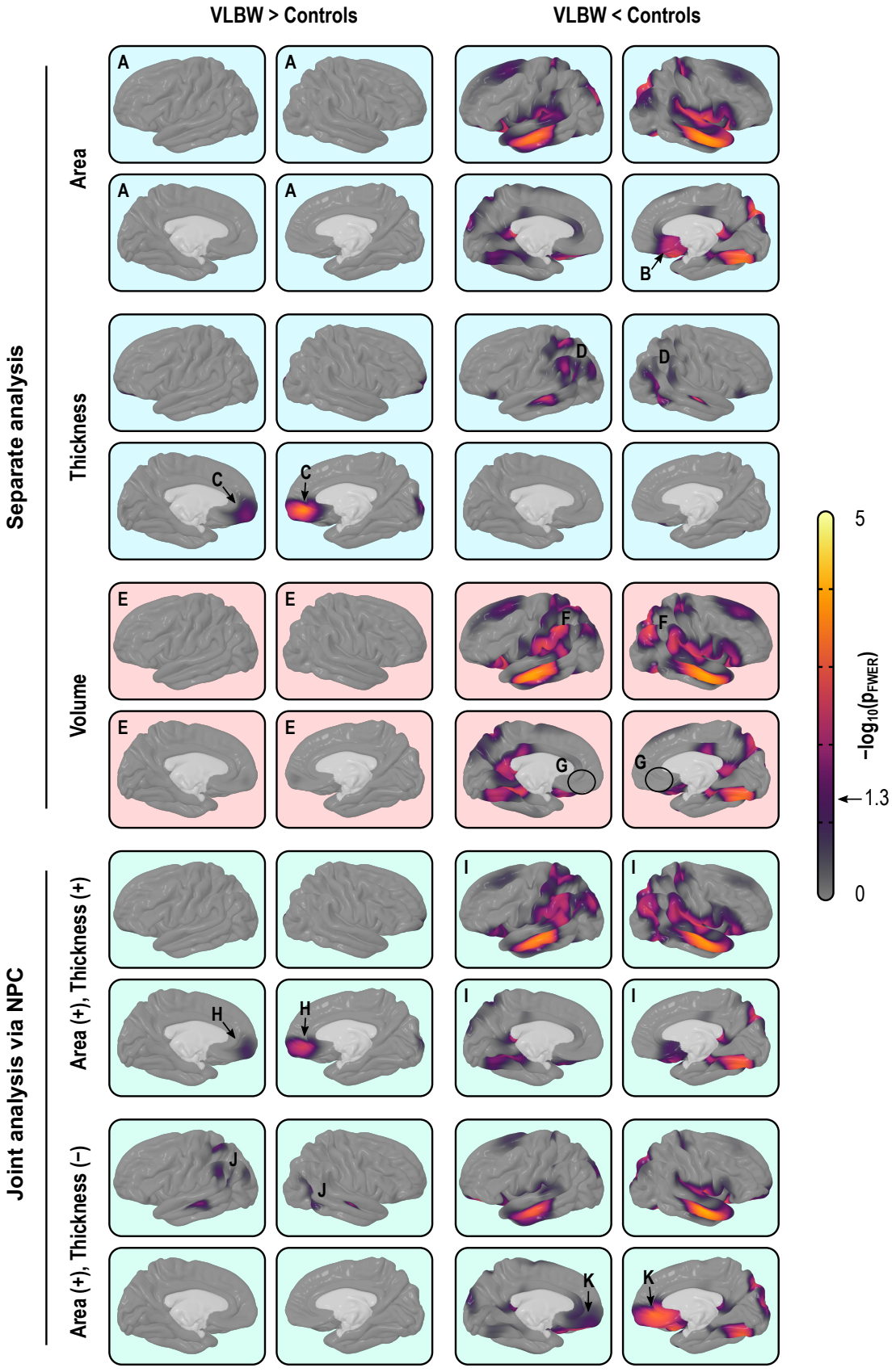


Figure 4.7: Coefficient of variation (σ/μ) after regressing the variability due to age, sex, and group. The variability across subjects is higher for area than for thickness, and even higher for volume. In all cases, the parietal cortex (parietal) is the region with the highest variability. For projections to the white and pial surfaces, consult the Appendix A.4.

4.3.6 Joint analysis via NPC

Non-parametric combination of thickness and area provides information about patterns of group differences not visible in cortical volume analyses, or that appear split or not visible in separate maps of area and thickness (Figure 4.8, lower panels, light green background). In the present data, the joint analysis suggests a decrease in the amount of tissue in VLBW subjects in the medial orbito-frontal cortex, and a bilateral decrease throughout most of the parietal cortex, as well as in the middle temporal and fusiform gyri. Furthermore, it suggests weaker bilateral increase in the amount of tissue in the parietal region, that alternates in space with that of tissue loss. Finally, NPC shows simultaneous bilateral decrease in surface area and increase in thickness in the medial orbito-frontal gyrus, none of which was observed using simple volume measurements (for additional maps, see Appendix A.4.4).

Figure 4.8: *(page 150)* Separate (*light blue background*) and joint (*green*) analysis of cortical area and thickness, as well as volume (*red*), using the IC7 resolution and smoothing with FWHM = 30 mm. Analysis of area indicates no reductions in the control group anywhere in the cortex (A), and among other regions, the subgenual region of the cingulate cortex (B). Analysis of thickness indicates that VLBW subjects have thicker cortex in the medial orbito-frontal cortex (C) and diffuse bilateral thinning mostly in the parietal and middle temporal regions (D). Analysis of volume alone broadly mimics analysis of area, with no evidence for increased volume in VLBW subjects (E), although some maps there seems to be a partial superimposition, with signs of bilateral decreased volume in the parietal lobe (F), but differently than for the analysis of area, no signs for reductions in the subgenual cortex (G). Jointly analysing area and thickness gives equal weight to both measurements, and allows directional effects to be inferred. Differently than in the case for volume, it is possible to know that there is an increase in the amount of cortical tissue in VLBW subjects in the medial orbito-frontal cortex (H) when compared to controls, and a bilateral decrease throughout most of the parietal cortex, more strongly in the middle temporal and fusiform gyri, in both hemispheres (I). Moreover, the joint analysis allows search for effects that can negate each other, such as in this case weaker effects in the parietal region (J), that partially overlap in space with those shown in (I). Finally, strong effects in the middle orbito-frontal, that were missed with simple volumes (G) become clearly visible (K).


VLBW > Controls
VLBW < Controls
Separate analysis
Area
Thickness
Volume
Joint analysis via NPC
Area (+), Thickness (+)
Area (+), Thickness (−)
A
B
C
D
E
F
G
H
I
J
K
5
−log10(pFWER)
1.3
0

4.4 Discussion

4.4.1 Interpolation of areal quantities

The different resampling methods do not perform similarly in all settings. Nearest neighbour and redistributive require smoothing in order to become comparable to, and interchangeable with, the pycnophylactic method. Since data is usually smoothed in neuroimaging studies in order to improve the matching of homologies and to improve the signal-to-noise ratio, this is not a limitation. Retessellation, particularly at lower resolutions, leads to substantial areal losses that cannot be recovered even with smoothing. Moreover, the vertices of the retessellated surfaces are not guaranteed to lie at the tissue boundaries they represent, introducing uncertainties to the obtained measurements. Regarding speed, although the various implementations run on linear time, $\Theta(n)$, the pycnophylactic method has to perform a larger number of computations that may not pay off when compared with nearest neighbour, provided that smoothing is used.

4.4.2 Areal expansion and absolute area

Few studies of cortical surface area have offered insight into the procedures adopted. Sometimes the methods were described in terms of areal expansion/contraction, as opposed to surface area itself. Furthermore, different definitions of areal expansion/contraction have been used, e.g., relative to the contra-lateral hemisphere (Lyttelton et al., 2009), to some earlier point in time (Hill et al., 2010), to a control group (Palaniyappan et al., 2011), or in relation to a standard brain, possibly the default brain (average or atlas) used in the respective software package (Joyner et al., 2009; Rimol et al., 2010a, 2012; Chen et al., 2011, 2012; Vuoksimaa et al., 2016); other studies considered linear distances as proxies for expansion/contraction (Sun et al., 2009a,b). Some of the studies that used a default brain as reference did use nearest neighbour interpolation followed by smoothing (Joyner et al., 2009; Rimol et al., 2010a, 2012), which as we have shown, assesses cortical area itself; never-

theless, the measurements were described in terms of areal expansion/contraction. These multiple definitions make the interpretation and comparison between studies challenging. Notwithstanding, measurements of areal expansion/contraction in relation to a given reference can also be obtained once interpolation has been performed using these methods. It suffices to divide the area per face (or per vertex) by the area of the corresponding face (or vertex) in the reference brain, which can be an atlas, the contralateral hemisphere, the same brain at an earlier point in time, or a brain from a different species. Surface area and areal expansion/contraction are related to each other by a factor that varies spatially.

4.4.3 Volumes improved, yet problematic

The large absolute difference between the product and the analytic method for cortical volume indicates that if interest lies in the actual values (for instance, for predictive models), the analytic method is to be preferred. The high correlation across subjects, however, suggests that, for group comparisons and similar analyses, both methods generally lead to similar results, except in a few regions of higher morphological inter-individual variability. However, even in these cases, cortical volume is a poor choice of trait of interest, since it is largely insensitive to changes in cortical thickness. While volume encapsulates information on both area and thickness, research has suggested that the proportion in which the variability of these two measurements coalesces varies spatially across the cortical mantle (Winkler et al., 2010), and moreover, that most of the variability of cortical volume, including that measured using VBM, can be explained by the variability of surface area alone (Voets et al., 2008; Lenroot et al., 2009; Winkler et al., 2010; Rimol et al., 2012), rendering volume a largely redundant metric. The continuous cortical maps in Figure 4.8 provide evidence that the results for cortical volume, even without using the simple product of thickness by area, mostly mirror the results for cortical surface area. However, volume and area are not interchangeable and since cortical thickness has some influence on volume, even in the absence

of volume effects it cannot be excluded that reductions in area have been compensated by increases in thickness, or vice versa, to yield no net volume effect.

4.4.4 Joint analyses via NPC

Such problems with cortical volume can be eschewed through the use of a joint statistical analysis of area and thickness. The NPC methodology gives equal (or otherwise predefined) weights for thickness and area, which therefore no longer have their variability mixed in unknown and variable proportions across the cortical mantle. Various combining functions can be considered, and the well known Fisher method of combination of p-values (Fisher, 1932) is a simple and computationally efficient choice. By using two distinct metrics in a single test, power is increased (Pesarin and Salmaso, 2010; Winkler et al., 2016c), allowing detection of effects that otherwise may remain unseen when analysing volume, or when thickness and area are used separately. NPC can be particularly useful for the investigation of processes affecting cortical area and thickness simultaneously, and can effectively replace volume as the measurement of interest in these cases, with various beneficial effects, and essentially none of the shortcomings. It constitutes a general method that can be applied to any number of partial tests, each relating to hypotheses on data that may be of a different nature, obtained using different measurement units, and related to each other arbitrarily.

Moreover, NPC allows testing directional hypotheses (by reversing the signs of partial tests), hypotheses with concordant directions (taking the extremum of both after multiple testing correction), and two-tailed hypotheses (with two-tailed partial tests). Power increases consistently with the introduction of more partial tests when there is a true effect, while strictly controlling the error rate. This is in stark contrast with classical multivariate tests based on regression, such as MANOVA or MANCOVA, that do not provide information on directionality of the effects, and lose power as the number of partial tests increase past a certain optimal point.

A joint test using NPC has similarities with, yet it is distinct from, the test

known as *conjunction* or *intersection-union test* (IUT) (Nichols et al., 2005). The NPC tests a joint null hypothesis that all partial tests have no effect; if the null is rejected in any partial test at a suitable level, the joint null is rejected. The conjunction tests a null hypothesis that at least one partial test has no effect; the alternative is that all partial tests have an effect. While a conjunction seeks an effect across all tests, NPC seeks an effect in any, or in an aggregate of the partial tests. Usage of NPC is not constrained to the replacement of cortical volume, and the method can be considered for analyses involving other cortical indices, including myelination (Glasser and Van Essen, 2011; Sereno et al., 2013) and folding and gyrification metrics (Mangin et al., 2004; Schaer et al., 2008; Toro et al., 2008), among others.

4.4.5 Permutation inference

Permutation tests provide exact inference based on minimal assumptions, while allowing multiple testing correction with strong control over the error rate. Even though permutation tests still have certain requirements, such that the data are exchangeable, certain types of structured dependency can be accommodated by means of restricted permutation strategies. Being based on permutations in each of the partial tests, NPC does not preclude the analysis of thickness and area (or of other partial tests) separately, and through the synchronised shuffling, correction for multiplicity of tests while taking into account their non-independence is trivial. This includes correction for multiple tests that may be used using various combinations of positive and negative directions for the partial tests. Permutation tests do not depend on distributional assumptions, which favours the analysis of surface area, which at the local level shows positive skewness, and is better characterised as log-normal (Winkler et al., 2012).

4.4.6 Area and thickness of VLBW subjects

The reduced cortical surface area observed in VLBW subjects compared to controls replicates previous findings from the same cohort at 20 years of age (Skranes

et al., 2013), and is consistent with findings from a younger cohort of VLBW subjects (Sølsnes et al., 2015) and teenagers born with extremely low birth weight ($\leqslant$ 1000g) (Grunewaldt et al., 2014). The combined evidence from these studies suggests that surface area reductions in the preterm brain are present from early childhood and remain reasonably constant from childhood until adulthood (Rimol et al., 2016). Proposed mechanisms for gray matter injury in preterm birth include hypoxia-ischemia and inflammation arising from intrauterine infections or from postnatal sepsis (Volpe, 2009, 2011), which may adversely affect critical phases of brain maturation before and after birth and cause diffuse white matter damage, including hypomyelination and primary or secondary gray matter dysmaturation (Hagberg et al., 2015). Cortical area reductions may not be explained by primary white matter damage alone, especially since area reductions are also observed in younger cohorts of preterms with less perinatal morbidity and less pathology in white matter microstructure, evaluated with diffusion tensor imaging (Eikenes et al., 2011; Rimol et al., 2016). Reduced neuropil is a possible explanation for cortical thinning in the lateral parietal and temporal cortex in VLBW subjects, but the thickening of the medial orbito-frontal cortex must be due to different mechanisms (Marín-Padilla, 1997; Bjuland et al., 2013; Grunewaldt et al., 2014). The combination of thickening and reduced area in medial orbito-frontal cortex has been observed in multiple cohorts and, more generally, these changes in both thickness and area could be related to prenatal factors, such as foetal growth restriction, or to postnatal exposure to extra-uterine environmental stressors (Sølsnes et al., 2015; Rimol et al., 2016). Regardless of underlying pathological aspects, the morphological indices appear to be robust markers of perinatal brain injury and maldevelopment (Raznahan et al., 2011; Skranes et al., 2013; Rimol et al., 2016).

This page intentionally left blank.

Chapter 5

Conclusion

In the thesis we deal with various improvements that make permutation tests more widely applicable, and show an example of such use for the analysis of cortical anatomy of the cerebrum.

Chapter 2 shows that the multi-level block permutation effectively controls the false positive rate, even in the presence of strong dependence between observations, and can be used as a general inference tool when the dependence structure can be organised in blocks, hierarchically if necessary. There is an unavoidable loss of power due to the reduced scope of shuffling, although in large datasets, even if with relatively complex dependence structure, such as the HCP, this loss is expected to be quite small, especially if permutations can be combined with sign flippings.

Chapter 3 shows that a number of statistical devices can be considered to accelerate permutation tests in addition to, or irrespective of, generic improvements to accelerations that depend on software implementation or on hardware. The methods considered yielded generally similar results, and as the different scenarios of error terms and shuffling strategy varied, the methods performed only marginally better or worse than each other in terms of conservativeness, agreement with the reference set, and resampling risk, and were in general substantially faster than the alternative of running a large number of permutations.

Chapter 4 shows that a joint analysis of cortical thickness and cortical surface area can reveal patterns of cortical pathology that would remain unseen with the usual methods that measure cortical volume. It also introduces a geometrically exact method to measure such volume, and provides a much needed comparison between the four extant methods to resample cortical area and related measurements that require mass conservation.

The hope is that these methods will help to accelerate the pace in which discoveries are made, while maintaining the hallmarks of permutation tests: robustness, exactness, and strong control over error rates.

5.1 Applications and future work

The thesis opens new avenues for research. In particular, permutation tests for large datasets, that could not be considered due to the elevated computational load, should become feasible with the acceleration strategies proposed in Chapter 3, which can be further combined with general software and hardware improvements. Moreover, the fact that some of the acceleration methods, such as tail approximation, gamma approximation, and the no permutation, allow continuous p-values to be found, tends to increase the power of methods for corretion for multiple testing, such as false discovery rate (FDR; Genovese et al., 2002), and can be considered in tandem with the NPC, so as to produce continuous p-values for partial tests before their combination. Investigation into these topics can be considered for future work.

The multi-level block permutation discussed in Chapter 2 is currently the only method that can yield valid inferences for data from the Human Connectome Project, and has been used, for instance to identify a mode of population covariation that links various imaging and non-imaging variables (Smith et al., 2015).

The joint analysis introduced in Chapter 4 is expected to provide a more informative option over the analysis of cortical volume. The method can be used to study pathological conditions affecting the cortex, but also, to provide localising

power for such effects. Moreover, in being mass-conservative, it has potential to help, for instance, to elucidate general questions, such as those related to the recent debate on whether regional proportions within the neocortex would be relatively constant across primates (Schoenemann et al., 2005; Barton and Venditti, 2013; Gabi et al., 2016).

In addition to the above, future work that can be considered immediately after the thesis consists of using both the multi-level block permutation and the acceleration strategies to allow faster, exact, multi-level inference for FMRI data (Woolrich et al., 2004), in a frequentist manner, yet free from assumptions related to frequency distributions. Such inference could replace current mixed-effects strategies that are based on parametric tests, or purely fixed or random-effects currently possible with permutation, yet without permuting FMRI time series. The acceleration strategies can likewise be considered for other cases of non-parametric combination that can be computationally onerous, such as for the statistical treatment of designs with missing data, especially when such missingness is not at random, which requires joint modelling of a potentially large number of designs.

5.2 Availability

A working implementation for Octave/MATLAB of all the permutation methods presented in this thesis are available under the General Public Licence (GPL) in the tool *Permutation Analysis of Linear Models* (PALM), which can be downloaded freely from www.fmrib.ox.ac.uk/fsl. Additionally, PALM includes various other features that were not discussed in the thesis, and which were presented elsewhere.

This page intentionally left blank.

Appendix A

Supplementary Material

The multiplicity of scenarios evaluated in Chapters 2, 3 and 4 resulted in the construction of more than 52 thousand plots and maps, which do not fit the thesis format. Although the most interesting and relevant results are already shown in the main text, a further selection of a few results would unduly overemphasise certain aspects at the expense of others. Instead, to facilitate presentation, these plots were organised in a browsable set of pages and packaged them into a single, 9.8 GB file that can be downloaded and browsed locally. This file is deposited for long term preservation and public access at the Research Archive of the Bodleian Libraries (ORA-Data), and it constitutes the Appendix that accompanies this thesis, and is accessible under its own Digital Object Identifier (DOI): **10.5287/bodleian:AePBKpxV8**. The results shown in the body of the text make ample reference to this material, and its inspection is encouraged.

This page intentionally left blank.

Bibliography

Agresti, A., and Coull, B. A. Approximate is better than exact for interval estimation of binomial proportions. *American Statistician*, 52(2):119–126, 1998.

Almasy, L., and Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62(5):1198–211, 1998.

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.

Anderson, T., and Darling, D. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23 (2):193–212, 1952.

Andrews, D. W. K., and Buchinsky, M. A three-step method for choosing the number of bootstrap repetitions. *Econometrica*, 68(1):23–51, 2000.

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. Computational anatomy with the SPM software. *Magnetic Resonance Imaging*, 27(8):1163–74, 2009.

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

Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., and Van Essen, D. C. Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage*, 80: 169–89, 2013.

Barton, R. A., and Venditti, C. Human frontal lobes are not relatively large. *Proceedings of the National Academy of Sciences of the United States of America*, 110(22):9001–6, 2013.

Basu, D. Randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association*, 75(371):575–582, 1980.

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.

Begley, C. G., and Ioannidis, J. P. A. Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, 116(1):116–126, 2015.

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 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.

Berry, K. J., Johnston, J. E., , and Jr., P. W. M. *A Chronicle of Permutation Statistical Methods: 1920–2000, and Beyond*. Springer, Heidelberg, 2014.

Besag, J., and Clifford, P. Sequential Monte Carlo p-values. *Biometrika*, 78(2): 301–304, 1991.

Bjuland, K. J., Løhaugen, G. C. C., Martinussen, M., and Skranes, J. Cortical thickness and cognition in very-low-birth-weight late teenagers. *Early Human Development*, 89(6):371–380, 2013.

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.

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

Bond, D. J., Ha, T. H., Lang, D. J., Su, W., Torres, I. J., Honer, W. G., Lam, R. W., and Yatham, L. N. Body mass index-related regional gray and white matter volume reductions in first-episode mania patients. *Biological Psychiatry*, 76(2): 138–45, 2014.

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.

Box, G. E. P., and Watson, G. S. Robustness to non-normality of regression tests. *Biometrika*, 49(1-2):93–106, 1962.

Bradley, J. V. *Distribution-Free Statistical Stests*. Prentice Hall, Eaglewood Cliffs, NJ, USA, 2 edition, 1968.

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.

Bro, R., Acar, E., and Kolda, T. Resolving the sign ambiguity in the singular value decomposition. Technical Report 2007-6422, Sandia National Laboratories, Albuquerque, NM, 2007.

Brown, L. D., Cai, T. T., and DasGupta, A. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001.

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(1):32–42, 1999.

Candès, E., and Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Candès, E. J., and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Cao, J., and Worsley, K. The detection of local shape changes via the geometry of Hotelling's t^2 fields. *The Annals of Statistics*, 27(3):925–942, 1999.

Carp, J. The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, 63(1):289–300, 2012.

Ceccarelli, A., Rocca, M. A., Pagani, E., Colombo, B., Martinelli, V., Comi, G., and Filippi, M. A voxel-based morphometry study of grey matter loss in MS patients with different clinical phenotypes. *NeuroImage*, 42(1):315–322, 2008.

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, C.-H., Peng, Q., Schork, A. J., Lo, M.-T., Fan, C.-C., Wang, Y., Desikan, R. S., Bettella, F., Hagler, D. J., Westlye, L. T., Kremen, W. S., Jernigan, T. L., Le Hellard, S., Steen, V. M., Espeseth, T., Huentelman, M., Håberg, A. K., Agartz, I., Djurovic, S., Andreassen, O. A., Schork, N., and Dale, A. M. Large-scale genomics unveil polygenic architecture of human cortical surface area. *Nature Communications*, 6(May):7549, 2015.

Choulakian, V., and Stephens, M. A. Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4):478–484, 2001.

Clopper, C. J., and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Curran, J. E., McKay, D. R., Winkler, A. M., Olvera, R. L., Carless, M. a., Dyer, T. D., Kent Jr, J. W., Kochunov, P., Sprooten, E., Knowles, E. E., Comuzzie, A. G., Fox, P. T., Almasy, L., Duggirala, R., Blangero, J., and Glahn, D. C. Identification of pleiotropic genetic effects on obesity and brain anatomy. *Human Heredity*, 75(2-4):136–143, 2013.

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*, 9:179–94, 1999.

Davidson, R., and MacKinnon, J. G. Bootstrap tests: how many bootstraps? *Econometric Reviews*, 19(1):55–68, 2000.

Davison, A., and Huser, R. Statistics of extremes. *Annual Review of Statistics and Its Application*, 2(1):203–235, 2015.

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

Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., James, S., Voets, N., Watkins, K., Matthews, P. M., and James, A. Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*, 130(9):2375–2386, 2007.

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.

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.

Eicken, H. Six red flags for suspect work. *Nature*, 497:433–434, 2013.

Eikenes, L., Løhaugen, G. C., Brubakk, A. M., Skranes, J., and Håberg, A. K. Young adults born preterm with very low birth weight demonstrate widespread white matter alterations on brain DTI. *NeuroImage*, 54(3):1774–1785, 2011.

Eklund, A., Andersson, M., and Knutsson, H. fMRI analysis on the GPU-possibilities and challenges. *Computer Methods and Programs in Biomedicine*, 105(2):145–61, 2012.

Eklund, A., Dufort, P., Forsberg, D., and LaConte, S. M. Medical image processing on the GPU: Past, present and future. *Medical Image Analysis*, 17(8): 1073–1094, 2013.

Eklund, A., Nichols, T. E., and Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):7900–5, 2016.

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.

Eyler, L. T., Chen, C.-H., Panizzon, M. S., Fennema-Notestine, C., Neale, M. C., Jak, A., Jernigan, T. L., Fischl, B., Franz, C. E., Lyons, M. J., Grant, M., Prom-Wormley, E., Seidman, L. J., Tsuang, M. T., Fiecas, M. J. A., Dale, A. M., and Kremen, W. S. A comparison of heritability maps of cortical surface area and

thickness and the influence of adjustment for whole brain measures: A magnetic resonance imaging twin study. *Twin Research and Human Genetics*, 15(03):304–314, 2012.

Farooqi, I. S., and O'Rahilly, S. New advances in the genetics of early onset obesity. *International Journal of Obesity*, 29(10):1149–52, 2005.

Fay, M. P., and Follmann, D. A. Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests. *The American Statistician*, 56(1):63–70, 2002.

Fay, M. P., Kim, H.-J., and Hachey, M. On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests. *Journal of Computational and Graphical Statistics*, 16(4):946–967, 2007.

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 the United States of America*, 97:11050–5, 2000.

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

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, 1999b.

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, 1999c.

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.

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, 1935.

Fjell, A. M., Grydeland, H., Krogsrud, S. K., Amlien, I., Rohani, D. A., Ferschmann, L., Storsve, A. B., Tamnes, C. K., Sala-Llonch, R., Due-Tønnessen, P., Bjørnerud, A., Sølsnes, A. E., Håberg, A. K., Skranes, J., Bartsch, H., Chen, C.-H., Thompson, W. K., Panizzon, M. S., Kremen, W. S., Dale, A. M., and Walhovd, K. B. Development and aging of cortical thickness correspond to genetic organization patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 112(50):15462–15467, dec 2015.

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., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994.

Gabi, M., Neves, K., Masseron, C., Ribeiro, P. F. M., Ventura-Antunes, L., Torres, L., Mota, B., Kaas, J. H., and Herculano-Houzel, S. No relative expansion of the number of prefrontal neurons in primate and human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 113(34):9617–22, 2016.

Gandy, A. Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104 (488):1505–1511, 2009.

Gandy, A., and Rubin-Delanchy, P. An algorithm to compute the power of Monte Carlo tests with guaranteed precision. *The Annals of Statistics*, 41(1): 125–142, 2013.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *NeuroImage*, 63 (2):858–73, 2012.

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.

Geschwind, D. H., and Rakic, P. Cortical evolution: judge the brain by its cover. *Neuron*, 80(3):633–47, 2013.

Gilbert, P. B. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 54(1):143–158, 2005.

Gitelman, D. R., Ashburner, J., Friston, K. J., Tyler, L. K., and Price, C. J. Voxel-based morphometry of herpes simplex encephalitis. *NeuroImage*, 13(4):623–31, 2001.

Glasser, M. F., and Van Essen, D. C. Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *Journal of Neuroscience*, 31(32):11597–616, 2011.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen,

D. C., and Jenkinson, M. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–24, 2013.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–8, 2016.

Gnedenko, B. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):423–453, 1943.

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, 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.

Gorgolewski, K. J., and Poldrack, R. A. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS Biology*, 14(7):1–13, 2016.

Grunewaldt, K. H., Fjørtoft, T., Bjuland, K. J., Brubakk, A.-M., Eikenes, L., Håberg, A. K., Løhaugen, G. C. C., and Skranes, J. Follow-up at age 10 years in ELBW children - functional outcome, brain morphology and results from motor assessments in infancy. *Early Human Development*, 90(10):571–8, 2014.

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.

Hagberg, H., Mallard, C., Ferriero, D. M., Vannucci, S. J., Levison, S. W., Vexler, Z. S., and Gressens, P. The role of inflammation in perinatal brain injury. *Nature Reviews Neurology*, 11(4):192–208, 2015.

Haldane, J. On a method of estimating frequencies. *Biometrika*, 33(3):222–225, 1945.

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

Hamming, R. W. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.

Harrison, N. A., Cooper, E., Dowell, N. G., Keramida, G., Voon, V., Critchley, H. D., and Cercignani, M. Quantitative magnetization transfer imaging as a biomarker for effects of systemic inflammation on the brain. *Biological Psychiatry*, 78(1):49–57, 2015.

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.

Hernández, M., Guerrero, G. D., Cecilia, J. M., García, J. M., Inuggi, A., Jbabdi, S., Behrens, T. E. J., and Sotiropoulos, S. N. Accelerating fibre orientation estimation from diffusion weighted magnetic resonance imaging using GPUs. *PLoS One*, 8(4):e61892, 2013.

Hill, J., Inder, T., Neil, J., Dierker, D., Harwell, J., and Essen, D. V. 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.

Hinrichs, C., Ithapu, V., Sun, Q., Johnson, S., and Singh, V. Speeding up permutation testing in neuroimaging. *Advances in Neural Information Processing Systems*, pages 890–898, 2013.

Ho, A. J., Stein, J. L., Hua, X., Lee, S., Hibar, D. P., Leow, A. D., Dinov, I. 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., Stephan, D. A., DeCarli, C. S., DeChairo, B. M., Potkin, S. G., Jack, C. R., Weiner, M. W., Raji, C. A., Lopez, O. L., Becker, J. T., Carmichael, O. T., and Thompson, P. M. A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly. *Proceedings of the National Academy of Sciences of the United States of America*, 107(18):8404–9, 2010.

Ho, A. J., Raji, C. A., Saharan, P., DeGiorgio, A., Madsen, S. K., Hibar, D. P., Stein, J. L., Becker, J. T., Lopez, O. L., Toga, A. W., and Thompson, P. M. Hippocampal volume is related to body mass index in Alzheimer's disease. *NeuroReport*, 22(1):10–14, 2011.

Hogstrom, L. J., Westlye, L. T., Walhovd, K. B., and Fjell, A. M. The structure of the cerebral cortex across adult life: age-related patterns of surface area, thickness, and gyrification. *Cerebral Cortex*, 23(11):2521–30, 2013.

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.

Hosking, J. R. M., and Wallis, J. R. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29:339–349, 1987.

Ioannidis, J. P. A., Tarone, R., and McLaughlin, J. K. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, 22(4):450–6, 2011.

Jacquard, A. *Structures génétiques des populations*. Masson, Paris, France, 1970.

Joshi, A. A., Leporé, N., Joshi, S. H., Lee, A. D., Barysheva, M., Stein, J. L., McMahon, K. L., Johnson, K., de Zubicaray, G. I., Martin, N. G., Wright, M. J., Toga, A. W., and Thompson, P. M. The contribution of genes to cortical thickness and volume. *Neuroreport*, 22(3):101–105, 2011.

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.

Jöckel, K.-H. Computational aspects of Monte Carlo tests. In Havránek, T., Šidák, Z., and Novák, M., editors, *Compstat 1984*, pages 183–188. Physica-Verlag HD, 1984. ISBN 978-3-7051-0007-7.

Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J.-D. Refined approximations to permutation tests for multivariate inference. *Computational Statistics & Data Analysis*, 20(94):643–656, 1995.

Keller, M. C., Medland, S. E., and Duncan, L. E. Are extended twin family designs worth the trouble? A comparison of the bias, precision, and accuracy of parameters estimated in four twin family models. *Behavior Genetics*, 40(3): 377–93, 2010.

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

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

Kim, H. J. Bounding the resampling risk for sequential Monte Carlo implementation of hypothesis tests. *Journal of Statistical Planning and Inference*, 140(7): 1834–1843, 2010.

Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T., and Shmulevich, I. Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):i161–8, 2009.

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

Kremen, W. S., Fennema-Notestine, C., Eyler, L. T., Panizzon, M. S., Chen, C.-H., Franz, C. E., Lyons, M. J., Thompson, W. K., and Dale, A. M. Genetics of brain structure: contributions from the Vietnam Era Twin Study of Aging. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics*, 162B(7): 751–61, 2013.

Küper, M., Rabe, K., Esser, S., Gizewski, E. R., Husstedt, I. W., Maschke, M., and Obermann, M. Structural gray and white matter changes in patients with HIV. *Journal of Neurology*, 258(6):1066–1075, 2011.

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.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. *Extremes and related properties of random sequences and processes.* Springer-Verlag, New York, 1983.

Leahy, R. M., and Qi, J. Statistical approaches in quantitative positron emission tomography. *Statistics and Computing*, 10(2):147–165, 2000.

Lee, N. R., Adeyemi, E. I., Lin, A., Clasen, L. S., Lalonde, F. M., Condon, E., Driver, D. I., Shaw, P., Gogtay, N., Raznahan, A., and Giedd, J. N. Dissociations in cortical morphometry in youth with down syndrome: Evidence for reduced surface area but increased thickness. *Cerebral Cortex*, 26(7):2982–2990, 2016.

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

Lenroot, R. K., Schmitt, J. E., Ordaz, S. J., Wallace, G. L., Neale, M. C., Lerch, J. P., Kendler, K. S., Evans, A. C., and Giedd, J. N. Differences in genetic and environmental influences on the human cerebral cortex associated with development during childhood and adolescence. *Human Brain Mapping*, 30(1):163–74, 2009.

Levesque, I. R., Giacomini, P. S., Narayanan, S., Ribeiro, L. T., Sled, J. G., Arnold, D. L., and Pike, G. B. Quantitative magnetization transfer and myelin water imaging of the evolution of acute multiple sclerosis lesions. *Magnetic Resonance in Medicine*, 63(3):633–640, 2010.

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.

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.

Mangin, J.-F., Rivière, D., Cachia, A., Duchesnay, E., Cointepas, Y., Papadopoulos-Orfanos, D., Scifo, P., Ochiai, T., Brunelle, F., and Régis, J. A framework to study the cortical folding patterns. *NeuroImage*, 23 Suppl 1: S129–38, 2004.

Manly, B. F. J. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. *Researches on Population Ecology*, 28(2):201–218, 1986.

Mann, H., and Whitney, D. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1): 50–60, 1947.

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.

Mardia, K. V. The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, 58(1):105–121, 1971.

Marín-Padilla, M. Developmental neuropathology and impact of perinatal brain damage. II: white matter lesions of the neocortex. *Journal of Neuropathology & Experimental Neurology*, 56(3):219–35, 1997.

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

Marqués-Iturria, I., Pueyo, R., Garolera, M., Segura, B., Junqué, C., García-García, I., José Sender-Palacios, M., Vernet-Vernet, M., Narberhaus, A., Ariza, M., and Jurado, M. A. Frontal cortical thinning and subcortical volume reductions in early adulthood obesity. *Psychiatry Research*, 214(2):109–15, 2013.

Marsland, A. L., Gianaros, P. J., Abramowitch, S. M., Manuck, S. B., and Hariri, A. R. Interleukin-6 covaries inversely with hippocampal grey matter volume in middle-aged adults. *Biological Psychiatry*, 64(6):484–490, 2008.

Martinussen, M., Fischl, B., Larsson, H. B., Skranes, J., Kulseng, S., Vangberg, T. R., Vik, T., Brubakk, A. M., Haraldseth, O., and Dale, A. M. Cerebral cortex thickness in 15-year-old adolescents with low birth weight measured by an automated MRI-based method. *Brain*, 128(11):2588–2596, 2005.

McFarquhar, M., McKie, S., Emsley, R., Suckling, J., Elliott, R., and Williams, S. Multivariate and repeated measures (MRM): A new toolbox for dependent and multimodal group-level neuroimaging data. *NeuroImage*, 132:373–389, 2016.

McKay, D. R., Knowles, E. E. M., Winkler, A. A. M., Sprooten, E., Kochunov, P., Olvera, R. L., Curran, J. E., Kent, J. W., Carless, M. A., Göring, H. H. H., Dyer, T. D., Duggirala, R., Almasy, L., Fox, P. T., Blangero, J., and Glahn,

D. C. Influence of age, sex and genetic factors on the human brain. *Brain Imaging and Behavior*, 8(2):143–52, 2014.

Melka, M. G., Gillis, J., Bernard, M., Abrahamowicz, M., Chakravarty, M. M., Leonard, G. T., Perron, M., Richer, L., Veillette, S., Banaschewski, T., Barker, G. J., Büchel, C., Conrod, P., Flor, H., Heinz, A., Garavan, H., Brühl, R., Mann, K., Artiges, E., Lourdusamy, A., Lathrop, M., Loth, E., Schwartz, Y., Frouin, V., Rietschel, M., Smolka, M. N., Ströhle, A., Gallinat, J., Struve, M., Lattka, E., Waldenberger, M., Schumann, G., Pavlidis, P., Gaudet, D., Paus, T., and Pausova, Z. FTO, obesity and the adolescent brain. *Human Molecular Genetics*, 22(5):1050–8, 2013.

Mëriaux, S., Roche, A., Dehaene-Lambertz, G., Thirion, B., and Poline, J. B. Combined permutation test and mixed-effect model for group average analysis in fMRI. *Human Brain Mapping*, 27:402–410, 2006.

Min, J., Chiu, D. T., and Wang, Y. Variation in the heritability of body mass index based on diverse twin studies: a systematic review. *Obesity Reviews*, 14 (11):871–82, 2013.

Minas, C., and Montana, G. Distance-based analysis of variance: Approximate inference. *Statistical Analysis and Data Mining*, 7(6):450–470, 2014.

Moore, G. E. Cramming more components onto integrated circuits. *Electronics*, pages 114–117, 1965.

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.

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.

Noble, K. G., Houston, S. M., Brito, N. H., Bartsch, H., Kan, E., Kuperman, J. M., Akshoomoff, N., Amaral, D. G., Bloss, C. S., Libiger, O., Schork, N. J., Murray, S. S., Casey, B. J., Chang, L., Ernst, T. M., Frazier, J. A., Gruen, J. R., Kennedy, D. N., Van Zijl, P., Mostofsky, S., Kaufmann, W. E., Kenet, T., Dale, A. M., Jernigan, T. L., and Sowell, E. R. Family income, parental education and brain structure in children and adolescents. *Nature Neuroscience*, 18(5):773–778, 2015.

O'Leary, D. D. M., Chou, S.-J., and Sahara, S. Area patterning of the mammalian cortex. *Neuron*, 56(2):252–69, 2007.

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.

Pannacciulli, N., Del Parigi, A., Chen, K., Le, D. S. N. T., Reiman, E. M., and Tataranni, P. A. Brain abnormalities in human obesity: a voxel-based morphometric study. *NeuroImage*, 31(4):1419–25, 2006.

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.

Parker, G. J. M., and Padhani, A. R. T_1-w DCE-MRI: T_1-weighted Dynamic Contrast-Enhanced MRI. In Tofts, P., editor, *Quantitative MRI of the Brain: Measuring Changes Caused by Disease*, pages 341–364. John Wiley & Sons, 2003.

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

Pearson, K. Contributions to the mathematical theory of evolution. II. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. Series A*, 186(1895):343– 414, 1895.

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

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

Picklands iii, J. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

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.

Poline, J. B., Worsley, K. J., Evans, A. C., and Friston, K. J. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, 5(2):83–96, 1997.

Posthuma, D., and Boomsma, D. I. A note on the statistical power in extended twin designs. *Behavior Genetics*, 30(2):147–58, 2000.

Raji, C. A., Ho, A. J., Parikshak, N. N., Becker, J. T., Lopez, O. L., Kuller, L. H., Hua, X., Leow, A. D., Toga, A. W., and Thompson, P. M. Brain structure and obesity. *Human Brain Mapping*, 31(3):353–64, 2010.

Rakic, P. Specification of cerebral cortical areas. *Science*, 241(4862):170–6, 1988.

Rakic, P. A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. *Trends in Neurosciences*, 18(9):383–8, sep 1995.

Raznahan, A., Shaw, P., Lalonde, F., Stockman, M., Wallace, G. L., Greenstein, D., Clasen, L., Gogtay, N., and Giedd, J. N. How does your cortex grow? *Journal of Neuroscience*, 31(19):7174–7, 2011.

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.

Rimol, L. M., Bjuland, K. J., Løhaugen, G. C. C., Martinussen, M., Evensen, K. A. I., Indredavik, M. S., Brubakk, A.-M., Eikenes, L., Håberg, A. K., and Skranes, J. Cortical trajectories during adolescence in preterm born teenagers with very low birthweight. *Cortex*, 75:120–31, 2016.

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.

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.

Ruxton, G. D., and Neuhäuser, M. Improving the reporting of p-values generated by randomization methods. *Methods in Ecology and Evolution*, 4(11):1033–1036, 2013.

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.

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.

Salmond, C., Ashburner, J., Vargha-Khadem, F., Connelly, A., Gadian, D., and Friston, K. Distributional assumptions in voxel-based morphometry. *NeuroImage*, 17(2):1027–1030, 2002.

Sandve, G. K., Ferkingstad, E., and Nygård, S. Sequential Monte Carlo multiple testing. *Bioinformatics*, 27(23):3235–3241, 2011.

Schaer, M., Cuadra, M. B., Tamarit, L., Lazeyras, F., Eliez, S., and Thiran, J.-P. A surface-based approach to quantify local cortical gyrification. *IEEE Transactions on Medical Imaging*, 27(2):161–70, 2008.

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.

Schnack, H. G., Van Haren, N. E. M., Brouwer, R. M., Evans, A., Durston, S., Boomsma, D. I., Kahn, R. S., and Hulshoff Pol, H. E. Changes in thickness and surface area of the human cortex and their relationship with intelligence. *Cerebral Cortex*, 25(6):1608–1617, 2015.

Schoenemann, P. T., Sheehan, M. J., and Glotzer, L. D. Prefrontal white matter volume is disproportionately larger in humans than in other primates. *Nature Neuroscience*, 8(2):242–52, 2005.

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.

Self, S. G., and Liang, K.-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–10, 1987.

Sereno, M. I., Lutti, A., Weiskopf, N., and Dick, F. Mapping the human cortical surface by combining quantitative T1 with retinotopy. *Cerebral Cortex*, 23(9): 2261–8, 2013.

Silventoinen, K., and Kaprio, J. Genetics of tracking of body mass index from birth to late middle age: evidence from twin and family studies. *Obesity Facts*, 2 (3):196–202, 2009.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., de Lange, M., Harris, J. R., Hjelmborg, J. V., Luciano, M., Martin, N. G., Mortensen, J., Nisticò, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G., and Kaprio, J. Heritability of adult body height: A comparative study of twin cohorts in eight countries. *Twin Research*, 6(05):399–408, 2012.

Skranes, J., Vangberg, T. R., Kulseng, S., Indredavik, M. S., Evensen, K. A. I., Martinussen, M., Dale, A. M., Haraldseth, O., and Brubakk, A. M. Clinical findings and white matter abnormalities seen on diffusion tensor imaging in adolescents with very low birth weight. *Brain*, 130(3):654–666, 2007.

Skranes, J., Løhaugen, G. C. C., Martinussen, M., Håberg, A., Brubakk, A. M., and Dale, A. M. Cortical surface area and IQ in very-low-birth-weight (VLBW) young adults. *Cortex*, 49(8):2264–2271, 2013.

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.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(Suppl. 1):208–219, 2004.

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E. J., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., and Miller, K. L. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565–1567, 2015.

Smucny, J., Cornier, M.-A., Eichman, L. C., Thomas, E. A., Bechtell, J. L., and Tregellas, J. R. Brain structure predicts risk for obesity. *Appetite*, 59(3):859–65, 2012.

Solomon, H., and Stephens, M. Approximations to density functions using pearson curves. *Journal of the American Statistical Association*, 73(361):153–160, 1978.

Sølsnes, A. E., Grunewaldt, K. H., Bjuland, K. J., Stavnes, E. M., Bastholm, I. a., Aanes, S., Østgård, H. F., Håberg, A., Løhaugen, G. C. C., Skranes, J., and Rimol, L. M. Cortical morphometry and IQ in VLBW children without cerebral palsy born in 2003-2007. *NeuroImage. Clinical*, 8:193–201, 2015.

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.

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.

Tippett, L. H. C. *The methods of statistics*. Williams and Northgate, London, 1931.

Toro, R., Perron, M., Pike, B., Richer, L., Veillette, S., Pausova, Z., and Paus, T. Brain size and folding of the human cerebral cortex. *Cerebral Cortex*, 18(10): 2352–7, 2008.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

van den Hoff, J. Principles of quantitative positron emission tomography. *Amino acids*, 29(4):341–53, 2005.

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., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., Chang, a., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S. E., Prior, F., Schlaggar, B. L., Smith, S. M., Snyder, A. Z., Xu, J., and Yacoub, E. The Human Connectome Project: a data acquisition perspective. *NeuroImage*, 62(4):2222–31, Oct. 2012.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80:62–79, 2013.

Visscher, P. M., Medland, S. E., Ferreira, M. a. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., and Martin, N. G. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, 2(3):e41, 2006.

Viviani, R., Beschoner, P., Ehrhard, K., Schmitz, B., and Thöne, J. Non-normality and transformations of random fields, with an application to voxel-based morphometry. *NeuroImage*, 35(1):121–30, 2007.

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.

Volpe, J. J. Brain injury in premature infants: a complex amalgam of destructive and developmental disturbances. *Lancet Neurol*, 8(1):110–124, 2009.

Volpe, J. J. Systemic inflammation, oligodendroglial maturation, and the encephalopathy of prematurity. *Annals of Neurology*, 70(4):525–529, 2011.

Vuoksimaa, E., Panizzon, M. S., Chen, C.-H., Fiecas, M., Eyler, L. T., Fennema-Notestine, C., Hagler, D. J., Franz, C. E., Jak, A. J., Lyons, M. J., Neale, M. C., Rinker, D. A., Thompson, W. K., Tsuang, M. T., Dale, A. M., and Kremen, W. S.

Is bigger always better? the importance of cortical configuration with respect to cognitive ability. *NeuroImage*, 129:356–366, 2016.

Waldrop, M. M. The chips are down for Moore's law. *Nature*, 530(7589):144–147, feb 2016.

Walley, A. J., Blakemore, A. I. F., and Froguel, P. Genetics of obesity and the prediction of risk for health. *Human Molecular Genetics*, 15 Spec No 2(2):R124–30, 2006.

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*. John Wiley and Sons, New York, 1993.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1 (6):80, 1945.

Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

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*, 53(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):1428–1443, 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., Greve, D. N., Bjuland, K. J., Nichols, T. E., Sabuncu, M. R., Håberg, A. K., Skranes, J., and Rimol, L. M. Joint analysis of area and thickness of the cerebral cortex replaces cortical volume measurements. *bioRxiv*, 2016a. doi: 10.1101/074666.

Winkler, A. M., Ridgway, G. R., Douaud, G., Nichols, T. E., and Smith, S. M. Faster permutation inference in brain imaging. *NeuroImage*, 141:502–516, 2016b.

Winkler, A. M., Webster, M. A., Brooks, J. C., Tracey, I., Smith, S. M., and Nichols, T. E. Non-Parametric Combination and related permutation tests for neuroimaging. *Human Brain Mapping*, 37(4):1486–511, 2016c.

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., and Smith, S. M. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4):1732–47, 2004.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1):58–73, 1996.

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, 2010.

Zhang, H., Sachdev, P. S., Wen, W., Crawford, J. D., Brodaty, H., Baune, B. T., Kochan, N. A., Slavin, M. J., Reppermund, S., Kang, K., and Trollor, J. N. The relationship between inflammatory markers and voxel-based gray matter volumes in nondemented older adults. *Neurobiology of Aging*, 37:138–146, 2015.

Zhang, Z., Wang, Y., Shen, Z., Yang, Z., Li, L., Chen, D., Yan, G., Cheng, X., Shen, Y., Tang, X., Hu, W., and Wu, R. The neurochemical and microstructural changes in the brain of systemic lupus erythematosus patients: A multimodal MRI study. *Scientific Reports*, 6:19026, 2016.

This page intentionally left blank.

St. Edmund Hall