Accurate Non-Iterative Modelling and Inference of Longitudinal Neuroimaging Data

Dissertation

to obtain the joint degree of Doctor at
the University of Liège and Maastricht University

on the authority of the Rector Magnifici,
Prof. A. Corhay and Prof. Dr. L.L.G. Soete

in accordance with the decision of the Board of Deans,
to be defended in public
on Wednesday the 30th of September 2015 at 9:30 hours

by

Bryan Guillaume

Supervisors
Prof. Dr. P. Matthews, Maastricht University
Ir. Dr. C. Phillips, University of Liège

Co-supervisor
Prof. Dr. T.E. Nichols, Warwick University

Assessment Committee
Prof. Dr. K. Van Steen, University of Liège (Chair)
Prof. Dr. E. Formisano, Maastricht University
Dr. G. R. Ridgway, University of Oxford
Prof. Dr. S.A. Rombouts, Leiden University

ISBN: 978-94-6259-851-5

The work presented in this thesis was funded by the EU within the PEOPLE Programme (FP7): Initial Training Networks (FP7-PEOPLE-ITN-2008), Grant Agreement No. 238593 NEUROPHYSICS, in which Maastricht University, Université de Liège, Forschungszentrum Jülich GmbH, and GlaxoSmithKline Ltd were network partners. The PhD research was mainly conducted at GlaxoSmithKline Ltd.

I would like to dedicate this thesis to my family.

Declaration

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

Bryan Guillaume
August 2015

Acknowledgements

To start, I would like to thank all my supervisors for their support and help. They were always there when it was needed. I would like particularly to thank Prof. T.E. Nichols for all his help and guidance throughout my doctoral research.

I am hugely grateful to the Statistics Department of Warwick University for hosting me for such a long time (I would not be surprised if I was the longest visiting PhD student of Warwick University - a silly record, but a record nonetheless). In particular, from Warwick University, I would like to thank all the past and present members of the Neuroimaging Statistics Research group with who I have enjoyed the traditional brain tea on Mondays and the traditional reading group on Thursdays. This is without any doubt the coolest group in the department.

I would like also to thank a few people without who I would probably not have pursued a PhD. First, I would like to thanks Relja Jerkovic. Without our discussion about research and PhD on our way towards the Borçelik building site, I would probably not have thought about changing my industrial life for an academic life. I must also thanks Prof. Rodolphe Sepulchre and Dr. Eric Bullinger who were the first to help me make the transition to a PhD.

Finally, I would like to give a special thanks to my family, particularly my parents who always been there for me and Dragana Pavlovic who supported me throughout my PhD. Without her patience and encouragement at some low moments, I would probably have given up.

Abstract

Despite the growing importance of longitudinal data in neuroimaging, the standard analysis methods make restrictive or unrealistic assumptions. For example, the widely used SPM software package assumes spatially homogeneous longitudinal correlations while the FSL software package assumes Compound Symmetry, the state of all equal variances and equal correlations. While some new methods have been recently proposed to more accurately account for such data, these methods can be difficult to specify and are based on iterative algorithms that are generally slow and failure-prone. In this thesis, we propose and investigate the use of the Sandwich Estimator method which first estimates the parameters of interest with a (non-iterative) Ordinary Least Square model and, second, estimates variances/covariances with the "so-called" Sandwich Estimator (SwE) which accounts for the within-subject covariance structure existing in longitudinal data. We introduce the SwE method in its classic form, and review existing and propose new adjustments to improve its behaviour, specifically in small samples. We compare the SwE method to other popular methods, isolating the combination of SwE adjustments that provides valid and powerful inferences. While this result provides p-values at each voxel, it does not provide spatial inferences, e.g. voxel- or cluster-wise family-wise error-corrected p-values. For this, we investigate the use of the non-parametric inference approach called Wild Bootstrap. We again identify the set of procedures and adjustments that provide valid inferences. Finally, in the third and fourth projects, we investigate two ideas to improve the statistical power of the SwE method, by using a shrinkage estimator or a covariance spatial smoothing, respectively. For all the projects, in order to assess the methods, we use intensive Monte Carlo simulations in settings important for longitudinal neuroimaging studies and, for the first two projects, we also illustrate the methods by analysing a highly unbalanced longitudinal dataset obtained from the Alzheimer's Disease Neuroimaging Initiative.

Propositions

In complement of the dissertation

Accurate Non-Iterative Modelling and Inference of Longitudinal Neuroimaging Data

by

Bryan Guillaume

1. Analysis of longitudinal data must account for multiple sources of variation, including subject-specific temporal evolution and group- and subject-specific noise magnitude. Analysis of neuroimaging longitudinal data must further account for spatial variation in each of these aspects.

2. Many popular longitudinal neuroimaging analysis methods use restrictive assumptions (e.g., the assumption of Compound Symmetry—the state of all equal variances and all equal covariances—or the assumption of a common covariance structure for the whole brain) and may yield invalid inferences when these assumptions do not hold.

3. The Sandwich Estimator method is a very promising approach to analyse longitudinal neuroimaging data due, mainly, to its robustness against the misspecification of the working covariance matrix and to the fact that it is free of iterative algorithms (thus, fast and without convergence issues).

4. The specification of the design matrix of a regression model is generally much more complicated with longitudinal data than with cross-sectional data. For example, a time-varying covariate should, in general, be split into a pure cross-sectional covariate and a pure longitudinal covariate.

5. The best remedy against small sample issues is to increase the sample size.

6. There is a lack of diagnostic tools (e.g., for checking model assumptions or detecting influential data) in neuroimaging.

7. There is a crisis of reproducibility in neuroimaging, a crisis that can be addressed by freely distributing the data and code used to draw scientific conclusions.

8. The number of people living with dementia and the cost associated to it are currently estimated worldwide at 44 million and at US $604 billion a year, respectively (Prince et al., 2014a), explaining the growing number of initiatives put in place to collect longitudinal neuroimaging data about it.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Longitudinal data analysis is of increasing importance in neuroimaging, particularly in structural and functional MRI studies. This trend can be observed in Figure 1.1, where the yearly number of publications mentioning "longitudinal MRI" exhibits rapid growth, not only in volume, but also in percentage of the yearly number of "MRI" publications. Unfortunately, while the current versions of the two most widely used neuroimaging software packages (i.e. `SPM` and `FSL`) are computationally efficient, they make quite restrictive assumptions when the longitudinal data consists of anything other than two time points per subject. In particular, `SPM12` unrealistically assumes a common longitudinal covariance structure for the whole brain while `FSL v5.0` assumes Compound Symmetry (CS), a simple covariance structure where the variances and correlations of the repeated measures are constant over time. This motivates recent publications proposing methods to better model neuroimaging longitudinal data (Skup et al., 2012; Bernal-Rusiel et al., 2013a,b; Chen et al., 2013; Li et al., 2013). However, all of these methods entail iterative optimisation at each voxel and are not necessarily easy to specify in practice.

In neuroimaging, two of the most widely longitudinal approaches currently used are the Naïve Ordinary Least Squares (N-OLS) modelling and the Summary Statistics Ordinary Least Squares (SS-OLS) modelling. The N-OLS method tries to account for the within-subject correlations by including subject-specific indicator variables (i.e. an intercept per subject) in an OLS model. This approach is fast, but does not allow one to make valid inferences on pure between covariates (e.g., group intercept or gender) and is valid only under CS. The SS-OLS method proceeds by first extracting a summary statistic of interest for each subject (e.g., slope with time) and then uses a group OLS model to infer on the summary measures. This method is also fast and has the advantage of reducing the analysis of correlated data to an analysis of independent

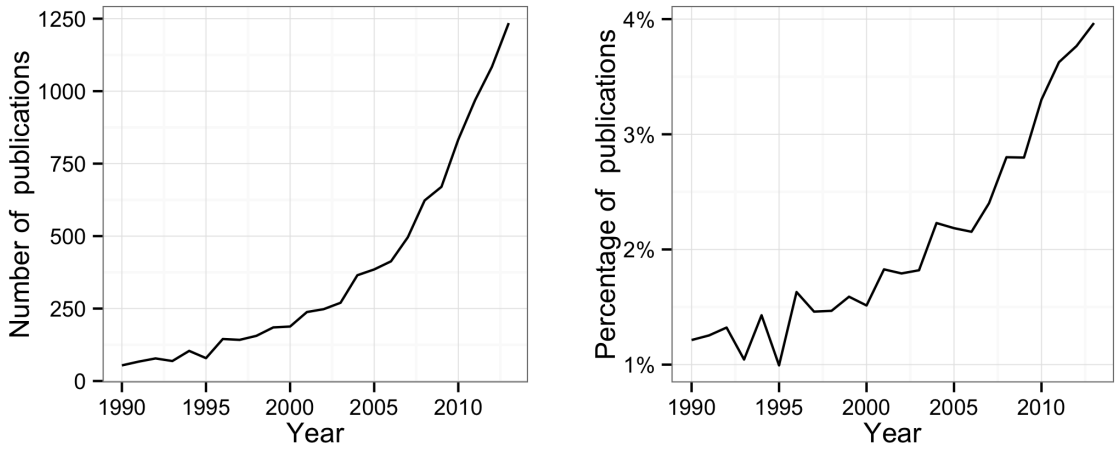

Fig. 1.1 Number of publications mentioning "longitudinal AND MRI" in all fields on a Pubmed search (left) and percentage of publications mentioning "longitudinal AND MRI" in all fields versus just "MRI" on Pubmed searches (right).

data, but this summary data may be highly variable as it is based on single-subject fits. In the context of one-sample t-tests, Mumford and Nichols (2009) showed that this approach is robust under heterogeneity, but warned that it is probably not the case for more general regression models.

In biostatistics, the analysis of longitudinal data is a long-standing problem and is generally done by using either Linear Mixed Effects (LME) models or marginal models. The LME models include random effects which account for the within-subject correlations existing in the data. Nevertheless, they require iterative algorithms which are generally slow and may fail to converge to the optimal solution. Another issue with LME models is the complexity of specifying and fitting the models. For example, the random effects and the covariance structure of the error terms need to be specified (e.g., only random intercepts? Also random slopes?) and, unfortunately, a misspecification of those may lead to invalid results. These are particularly serious problems in neuroimaging as model assessment is difficult and a single model must be used for the whole brain. As a consequence, the use of LME models in neuroimaging may be prohibitively slow, and may lead to statistical images with missing or invalid results for some voxels in the brain. To limit the convergence issues, one may be tempted to use a LME model with only a random intercept per subject. Unfortunately, like the N-OLS model, this model assumes CS which is probably not realistic, especially for long studies carried out over years and with many visits. In contrast, marginal models implicitly account for random effects, treat the intra-visit correlations as a

nuisance and focus the modelling only on population averages. They have appealing asymptotic properties, are robust against model misspecifications and, as there are no explicit random effects, are easier to specify than LME models. However, they only focus on population-averaged inferences or predictions, typically require iterative algorithms and assume large samples.

Recently, Bernal-Rusiel et al. (2013a) proposed the use of LME models to analyse longitudinal neuroimaging data, but only on a small number of regions of interest or biomarkers, thus, making the slow LME models practical. However, in the case where all voxels are analysed, overcoming all the LME model drawbacks seems daunting. Nevertheless, Chen et al. (2013) and Bernal-Rusiel et al. (2013b) extended the use of the LME models to mass-univariate settings. In particular, Bernal-Rusiel et al. (2013b) proposed the use of a spatiotemporal LME method, that first parcels the brain into homogeneous areas for which they separately model the full spatiotemporal covariance structure by notably assuming, for each area, a common temporal covariance structure across all the points and a simple covariance structure to model the spatial dependencies. Skup et al. (2012) and Li et al. (2013) proposed to use marginal models to analyse neuroimaging longitudinal data. Specifically, Skup et al. (2012) proposed a Multiscale Adaptive Generalised Method of Moments (MA-GMM) approach which combines a spatial regularisation method with a marginal model called Generalised Methods of Moments (GMM; Hansen, 1982; Lai and Small, 2007) and Li et al. (2013) proposed a Multiscale Adaptive Generalised Estimating Equations (MA-GEE) approach which also combines a spatial regularisation method, but with a marginal model called Generalised Estimating Equations (GEE; Liang and Zeger, 1986). Thanks to their appealing theoretical asymptotic properties, the two latter methods seem very promising for analysing longitudinal neuroimaging data. Nevertheless, like the LME models, they require iterative algorithms, which make them slow, and - due to the fact that they rely on asymptotic theoretical results - their use may be problematic in small samples.

In this thesis, we propose an alternative marginal approach. We use a simple OLS model for the marginal model (i.e. no subject indicator variables) to create estimates of the parameters of interest. For standard errors of these estimates, we use the so-called Sandwich Estimator (SwE; Eicker, 1963) to account for the repeated measures correlation. The main property of the SwE is that, under weak conditions, it is asymptotically robust against misspecification of the covariance model. In particular, this robustness allows us to combine the SwE with a simple OLS model which has no covariance model. Thus, this method is quite easy to specify and, with no need

for iterative computations, it is fast and has no convergence issues. Moreover, the proposed method can deal with unbalanced designs and heterogeneous variances across time and groups (or even subjects). In addition, the SwE method can also be used for cross-sectional designs where repeated measures exist, such as fMRI studies where multiple contrasts of interests are jointly modelled, or even for family designs where subjects from the same family cannot be assumed independent. Nevertheless, like the MA-GMM and MA-GEE methods, the SwE method typically relies on asymptotical theoretical results, guaranteeing accurate inferences only in large samples. Therefore, in this thesis, we also review and propose small sample adjustments that improve its behaviour in small samples.

The remainder of this thesis is organised as follows.

In Chapter 2, we give some background information about longitudinal neuroimaging data analysis, starting from the way this type of data can be acquired to the way statistical inferences can be made on them. In particular, we describe some longitudinal models which have been widely used or recently proposed in neuroimaging. We also introduce a real longitudinal neuroimaging dataset acquired as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI; Mueller et al., 2005).

In Chapter 3, we introduce the SwE method for the analysis of longitudinal data in its most standard form. Then, we review and propose many adjustments with the aim to improve its behaviour, mainly in small samples. In particular, we propose three novel statistical parametric tests that account for the potential small sample nature of the data. Using intensive Monte Carlo evaluations, we also assess the SwE method and all the adjustments presented, isolate the best combination of adjustments to be used in practice and compare the SwE method to popular alternative methods. Finally, we illustrate the SwE method on the real ADNI dataset introduced in Chapter 2.

In Chapter 4, we introduce the Wild Bootstrap, a method that can be used to make non-parametric inferences with the SwE method. We describe several versions of it and assess them using intensive Monte Carlo simulations. Finally, we apply the Wild Bootstrap to perform a non-parametric inference on the real ADNI dataset already analysed parametrically in Chapter 3.

In Chapter 5, we propose two new types of SwE, both based on the Ledoit-Wolf shrinkage estimator of covariance matrices (Ledoit and Wolf, 2003) with the goal of decreasing the estimation error of the SwE. We also assess these two new SwE versions using Monte Carlo simulations.

In Chapter 6, we investigate the possibility to enhance the power of detection of the SwE method by spatially smoothing the data covariance matrices. In particular, we

consider the use of several smoothing metrics and propose three smoothing procedures based on some forms of homogenisation of the data covariance matrices. Then, we use Monte Carlo simulations to study the proposed forms of smoothing.

In Chapter 7, we conclude by summarising and discussing the main findings of the thesis. We also discuss potential future work.

Chapter 2

Background

In this chapter, we first give a general overview of a typical neuroimaging longitudinal data analysis pipeline, starting from the acquisition of longitudinal brain images to their preprocessing, statistical modelling and inference. In particular, we describe several longitudinal models often used or recently proposed in neuroimaging. Finally, we introduce a real longitudinal dataset acquired as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI; Mueller et al., 2005) that will be used later to illustrate the methods developed in this thesis.

2.1 Neuroimaging modalities

Presently, there are many techniques that allow the acquisition of brain images in a non-invasive way. For example, Computed Tomography (CT), Positron Emission Tomography (PET), Electroencephalography (EEG), Magnetoencephalography (MEG), Magnetic Resonance Imaging (MRI) and Near-Infrared Spectroscopy (NIRS) are often used to achieve this goal and can potentially be used to produce longitudinal neuroimaging data. In this section, we briefly describe each of these modalities and, for the interested readers, provide further references.

Computed Tomography

Computed Tomography (CT) uses X-rays to construct 3D images of the brain (Hsieh, 2009). More precisely, this technique measures the X-ray attenuation which varies in accordance with the density of different tissues; the different density, mainly between bone, water and brain tissue, provides the contrast to visualise the brain in 3D. Although this modality is non-invasive, its main disadvantage is the subject's exposure

to ionising radiation which may pose some health hazards. Also, there is very poor contrast between the gray and white matter of the brain. Nevertheless, some longitudinal neuroimaging studies using CT have been carried out (e.g., Illowsky et al., 1988; Woods et al., 1990; Jaskiw et al., 1994; Davis et al., 1998).

Positron Emission Tomography

Positron Emission Tomography (PET) produces 3D images of the brain by detecting gamma rays emitted by a radioactive tracer injected into the subject. The tracer is part of a simple substance (oxygen or water) or biologically active molecules. In either case, the tracer makes its way to the brain and maps the *function* of the entire brain or the presence of pathological deposits such as amyloid plaques (see, e.g., Ossenkoppele et al., 2012) or tau-containing neurofibrillary tangles (see, e.g., Villemagne et al., 2014). The term *function* is used to refer to characterisation of physiology that changes on a short time scale (e.g., up to hours/days). It is in distinction to *structural* imaging, which reveals the anatomical detail of the brain. Note that, for the detection of pathological deposits, the time scale of change is longer (e.g., months/years). This modality is often combined with CT or MRI to produce a structural image of the brain allowing the anatomical localisation of the metabolic activities or pathologic deposits detected by the PET scan. Like CT, PET has the disadvantage of exposing the subjects to ionising radiation. While the latter may be a limitation for longitudinal studies, longitudinal PET studies have been conducted (e.g., Sturm et al., 2004; De Boissezon et al., 2005; Mueller et al., 2005; Ossenkoppele et al., 2012). For further information about this modality, a good technical review of it can be found in Holmes (1994, Chapter 1) and a recent overview of its practical applications in neuroimaging can be found in Nasrallah and Dubroff (2013).

Electroencephalography

Electroencephalography (EEG) measures the current on the scalp induced by electrical activity in the brain. This modality provides functional information with very good temporal resolution, but has relatively poor spatial resolution. Also, it has the advantages, compared to PET, that it does not use any ionising radiation. Some longitudinal studies using this modality can be found in the literature (e.g., Bangert and Altenmüller, 2003; Saggar et al., 2012; Seppänen et al., 2012). Further information about this modality can be found in Niedermeyer and Da Silva (2005).

Magnetoencephalography

Magnetoencephalography (MEG) measures the magnetic fields produced by the electrical activity of the brain. Like EEG, it is a functional modality that has very good temporal resolution, but relatively poor spatial resolution and does not use any ionising radiation. Longitudinal studies using MEG can also be found in the literature (e.g., Dubbelink et al., 2013, 2014; Van Dellen et al., 2014; Yoshimura et al., 2014). Further information about this modality can be found in Hansen et al. (2010).

Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a technique which uses the magnetic properties of atomic nuclei, generally those of hydrogen, to image the brain. An MRI scanner consists of an extremely strong static magnetic field to place the hydrogen atoms into an equilibrium state. Then, using a sequence of electromagnetic pulses, the scanner excites the hydrogen atoms and measures the signals they emit during their relaxation times. Spatial gradients applied to the magnetic field allow images to be formed, and careful manipulation of the timing of the electromagnetic pulse "sequences" creates images with different types of tissue contrast. Different kinds of sequences lead to different type of MRI sub-modalities such as, for example, structural MRI (sMRI) or functional MRI (fMRI). In particular, fMRI is based on how changes in the concentrations of oxygenated haemoglobin and deoxygenated haemoglobin tends to vary locally upon brain activity. Since oxygenated haemoglobin and the deoxygenated haemoglobin have different magnetic properties, the variation of their concentrations can be detected by the MRI scanner under the form of a signal, generally referred to as Blood-Oxygen-Level Dependent (BOLD) signal, reflecting the brain activity and allowing the acquisition of functional images (Ogawa et al., 1990).

Compared to EEG, MEG and NIRS (see below), fMRI has the advantage to have a better spatial resolution and a better ability to localise the signal source, but has a poorer temporal resolution. Compared to PET and CT, MRI has the advantage of not using any ionising radiation. Probably due to those advantages, MRI is the dominant modality which currently produces the largest quantity of longitudinal neuroimaging data (e.g., Draganski et al., 2004; Mueller et al., 2005; Meltzer et al., 2009; Kim et al., 2010; Andrews et al., 2013).

Near Infrared Spectroscopy

Near Infrared Spectroscopy (NIRS) uses light in the near infrared range to detect changes in the concentration of oxygenated haemoglobin and deoxygenated haemoglobin. As these changes are closely related to the brain activity, NIRS can be used to construct functional images of the brain. While NIRS has notably the advantage to have a better temporal resolution than fMRI, it has a poorer spatial resolution and, like EEG and MEG, the localisation of the signal sources is unfortunately not as straightforward as in fMRI. Some longitudinal studies using this technique can be found in the literature (e.g., Kono et al., 2007; Tsujii et al., 2009; Moriguchi and Hiraki, 2011). Additional information about NIRS can be found, for example, in Huppert et al. (2009).

2.2 Preprocessing

In general, the images obtained from any of the modalities mentioned in Section 2.1 cannot be immediately used for a statistical analysis and a series of preprocessing steps must be applied to the data. The main purpose of these preprocessing steps is to transform the data to a format suitable for a mass-univariate analysis. A mass-univariate model is a model fit at each voxel[1], independently of all other voxels. Crucially, a mass-univariate analysis assumes that the data at each voxel corresponds to the same region of the brain in every subject at each time point. Note that other types of models can be considered, e.g., multivariate models where multiple locations in space are jointly modelled. However, due to the large amount of data as well as the high spatial and temporal complexity of the data, these multivariate models are generally very challenging to specify and fit. Therefore, a mass-univariate approach is generally preferred in practice.

In the remainder of this section, we succinctly describe several important preprocessing steps used for the two modalities most widely used in longitudinal studies, fMRI and sMRI.

2.2.1 fMRI preprocessing

In fMRI, the data acquired consists, for each subject and session, of a time series of 3D images. To make them suitable for a mass-univariate modelling, several preprocessing

[1] A voxel is short for volume element. It is the 3D generalisation of a pixel, and is the smallest element of a 3D image

steps like slice timing, spatial realignment, spatial registration, spatial smoothing and intra-subject modelling are generally used. Here below, we briefly describe each of these steps.

Slice timing

In fMRI, a 3D image is not an instantaneous snapshot of the brain activity. Rather, the fMRI images are acquired slice by slice and, as it can take 2 seconds to acquire all the slices in the brain, significant differences between the acquisition times over slices exist. In practice, there are mainly three ways to correct for this: (i) an additional covariate able to account for the time differences is included in the General Linear Model (GLM) used to extract summary statistics from the fMRI time series (see below), (ii) a slice timing correction is applied, consisting of interpolating each slice in time to a common time point or (iii) slice-specific models with slice-specific time-shifted covariates are used (Henson et al., 1999).

Spatial realignment

The acquisition of an fMRI scan can last for several minutes. Even if some devices are usually used to maintain the subject's head still in the scanner, it is difficult to avoid subject's head motion. Due to this, the fMRI scans at different time points can be misaligned. To correct this misalignment, a rigid body realignment is typically used for each fMRI scan (Ashburner and Friston, 2007b). Note that more sophisticated approaches exist, such as the method proposed in Roche (2011) which simultaneously corrects for motion and slice timing.

Spatial coregistration

It is often desirable to map the functional information of the fMRI images onto the anatomical space of the subject. This can be helpful to localise where the brain activity occurs and can also ease the spatial normalisation step described next. To achieve this, an sMRI of the subject's head is acquired and registered with the fMRI images. This coregistration is generally performed by minimising a cross-modality loss function like the mutual information (Ashburner and Friston, 2007b).

Spatial normalisation

Inter-subject registration, or spatial normalisation, is required for a mass-univariate analysis, as the size and shape of each subject's brain is different. Spatial normalisation

consists of registering the images of each subject onto a common brain template. Several algorithms are available and generally use the information contained in the structural images to warp each subject's brain images onto a common brain template (Ashburner and Friston, 2007a).

Spatial smoothing

Even after spatial normalisation, it is expected that small misregistrations across subjects exist. In practice, this can be partially corrected by spatially smoothing the data using a Gaussian kernel. While this spatial smoothing decreases the spatial resolution of the images and may remove some high spatial frequency information, it has three other advantages that are important for the analysis of neuroimaging data. First, the spatial smoothing can also increase the signal-to-noise ratio, in particular for spatially extended signals. Second, by the central limit theorem, the spatial smoothing tends to make the data more Gaussian, allowing a better compatibility with the usual model assumptions. Third, the smoothing makes the data more compliant with the assumptions behind the Random Field Theory, a set of statistical inference procedures used to solve the issue of multiple comparisons (Nichols and Hayasaka, 2003).

Intra-subject Modelling

While ideally a model for fMRI would simultaneously consider all time series data from all subjects and all visits, the extreme size of the data makes this impracticable. Instead,"first level" intra-subject models are fit for each individual, and possibly for each visit separately. A General Linear Model (Kiebel and Holmes, 2007) is typically used to fit the time course data and produce a summary measure of interest. For example, while a verbal working memory task might have 0-, 1- and 2-word conditions, only the 1- vs 2-word comparison may be of interest. When a single visit/session is modelled, this "contrast" is the estimate that is submitted to the group "second level" longitudinal analysis. Alternatively, if all visits for a subject are modelled together, a single summary measure is generated for a so-called "summary statistic" analysis (described in detail in Section 2.3.3).

2.2.2 sMRI preprocessing

Longitudinal sMRI images are rarely used directly for a statistical analysis. Instead, some meaningful structural information is generally extracted from them and used as measure of interest. In this section, we describe two of such approaches referred to

as Tensor-Based Morphometry (TBM; Ashburner and Friston 2003) and Voxel-Based Morphometry (VBM; Ashburner and Friston 2007d).

Tensor-Based Morphometry

TBM generally consists of extracting, from each structural image, the Jacobian determinant image of the deformation field that is used to normalise each image onto a common template structural image. The Jacobian determinant values represents the relative volume expansion (value > 1) or contraction (value < 1) of the source image voxels compared to the template image voxels, allowing to study how the different brain areas tend to change across time and subjects. As the resulting Jacobian determinant images are already in the space of the template image, no additional normalisation is needed and they may be used without further preprocessing for a statistical modelling. Note that alternative approaches which separately perform a within-subject registration and a between-subject normalisation allow for a more precise estimation of the within-subject volume changes, but are subject to the difficulty of deciding how to spatially normalise the within-subject information (Ridgway, 2009; Ridgway et al., 2015).

Voxel-Based Morphometry

VBM generally consists of first extracting a segmentation image of grey matter from each structural image (Ashburner and Friston, 2007c). Then, before using them as data for a statistical analysis, each grey matter image is warped onto a common template image using a non-linear registration method (Ashburner and Friston, 2007a). As this normalisation step changes the volumes of brain regions, the actual amount of grey matter is unfortunately modified by the normalisation. To adjust for this, the normalised images are generally multiplied by the Jacobian determinant images, as they correspond to the relative volumes of each voxel before and after the normalisation. Next, these "modulated" images are smoothed for the same reasons mentioned previously for the preprocessing of fMRI data (see Section 2.2.1), and can then be used for a statistical modelling.

2.3 Models for longitudinal data

In this section, we first define what is the difference between cross-sectional and longitudinal data. Then, we review some methods currently used or recently proposed to

analyse longitudinal neuroimaging data.

2.3.1 Cross-sectional versus longitudinal data

Cross-sectional data is a type of data collected from several subjects at a single time point and is useful to test how the population of subjects may differ, but only at the time the data is acquired. We typically say that we are testing for cross-sectional effects in the population, which can be, for example, age, gender, IQ, etc. The data can generally be assumed to be independent across subjects, allowing the use of relatively simple models.

In contrast, longitudinal data is a type of data collected from several subjects repeatedly at several time points and is useful to test how the population of subjects may differ at a single time point (cross-sectional effects), but also how the population may change over time (longitudinal effects). While the assumption of independence across subjects is usually valid, the data from the same subject cannot generally be assumed independent. Therefore, the models used for cross-sectional data are, in most cases, not valid and more complicated models able to account for the within-subject dependence have to be used instead.

2.3.2 The Naïve-Ordinary Least Square method

A popular model used currently to fit repeated measures or longitudinal neuroimaging data is the Naïve-Ordinary Least Square model (N-OLS). The N-OLS model tries to account for the within-subject correlations by including subject indicator dummy variables (i.e. an intercept per subject) in an OLS model. This approach is fast, but does not allow one to make valid inferences on pure between-subject covariates (e.g., group intercept or gender) and is valid only under Compound Symmetry (CS), the state of all equal variances and all equal covariances. The latter is quite restrictive as we may expect the variance to increase over time and/or the correlation to decrease over time.

2.3.3 The Summary Statistics Ordinary Least Square model

The Summary Statistics Ordinary Least Square (SS-OLS) method is also very popular in neuroimaging and proceeds by first extracting a summary statistic of interest for each subject (e.g., slope with time) and then uses a group OLS model to infer on the summary measures. This method is fast and has the advantage of reducing the

analysis of correlated data to an analysis of independent data. In the context of one-sample t-tests, Mumford and Nichols (2009) showed that this approach is robust under heterogeneity, but warned that it is probably not the case for more general regression models. In particular, when there is imbalance in the number of visits per subject, the SS-OLS may yield inaccurate inferences (evaluated in depth in this work; see Section 3.3.2). That is, even if we assume that all the data point have the same variance, the variance of the summary measures will be different between subjects having a different number of data points. For example, the summary measure obtained from a subject with only two data points will clearly be more variable than the one obtained from a subject with six data points. Therefore, in the condition of an unbalanced design, heterogeneity is likely to exist between the subject summary measures, challenging the homogeneity assumption of the the group OLS model.

2.3.4 The SPM procedure

Another popular neuroimaging procedure currently used in `SPM`, probably the most widely used neuroimaging software package, relies on the estimation of a global covariance structure for the whole brain, using the most "promising" voxels in the brain as determined by an omnibus F-test from a preliminary OLS model fit. This estimated covariance structure is then assumed to be the true covariance structure for the whole brain and is used to "whiten" the data at every voxel, before using an OLS model on the "whitened" data. This method is fast and quite powerful for detecting effects, as the effective number of degrees of freedom of the global covariance structure is infinity (no uncertainty). Nevertheless, as it is unlikely that every voxel in the brain has the same covariance structure, this method can be perilous to use in practice without an additional procedure ensuring that the global covariance structure is valid everywhere in the brain or, at least, at the voxels showing effects. Indeed, for voxels with a true covariance structure different from the one assumed by the SPM procedure, the method will actually "colour" the data by altering the covariance structure to something other than the desired i.i.d. assumption, thus challenging the assumption made by the OLS model.

2.3.5 The Linear Mixed Effects model

In the biostatistics literature, the most popular model to analyse longitudinal data is likely the Linear Mixed Effects (LME) model and has been proposed recently to analyse longitudinal neuroimaging data (Bernal-Rusiel et al., 2013a,b; Chen et al.,

2013). Using the formulation of Laird and Ware (1982), the LME model for individual i is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{2.1}$$

where y_i is a vector of n_i observations for individual $i = 1, 2, \ldots, m$, β is a vector of p fixed effects which is linked to y_i by the $n_i \times p$ design matrix X_i, b_i is a vector of r individual random effects linked to y_i by the $n_i \times r$ design matrix Z_i, and ϵ_i is a vector of n_i individual error terms assumed to be normally distributed with mean 0 and covariance Σ_i. The individual random effects b_i are also assumed to be normally distributed, independently of ϵ_i, with mean 0 and covariance D. Typically, the p fixed effects might include an intercept per group, a linear effect of time per group, a quadratic effect of time per-group or per-visit measures effects like, in the case of Alzheimer's Disease, the MMSE (Mini-Mental State Examination) score. The r random effects usually include a "random intercept" for each subject (modelled by a constant in Z_i) and may also include a "random slope" for each subject.

In LME models, the randomness of the data is modelled by both the random effects b_i and the error terms ϵ_i. This makes LME models quite flexible in practice as we can use both the random effects b_i and the error terms ϵ_i to model the covariance structure existing in the data. Nevertheless, as pointed out in Hamer and Simpson (1999), this flexibility comes also with the risk of confusion and errors. Indeed, specifying an LME model comes with many questions such as "What random effects should I include in the model?", "Only a random intercept?", "Should I also add a random slope per subject?" or "Should I assume that the error terms are i.i.d. or should I assume a particular structure for Σ_i?". These questions are not easy to answer and, unfortunately, a misspecification of the model can easily lead to inaccurate inferences. In particular, the random effects b_i have an important impact on the covariance structure modelling and have to be chosen carefully. For example, an LME model with only a random-intercept per subject and i.i.d. error terms assumes by construction that the within-subject covariance structure is CS, just like the N-OLS model. This makes the LME models quite difficult to specify in practice, particularly in the context of neuroimaging where we need to fit thousands of models simultaneously. Indeed, as the covariance structure is likely to vary across the brain, a well specified model for some voxels may be invalid for other voxels in the brain. Nevertheless, an advantage of the LME models compared to other models is their ability to make inferences on random effects or to predict subject-specific profiles. In the context of neuroimaging, inferences on random effects have been studied in Lindquist et al. (2012).

For LME models, the estimate of the fixed effect parameters β and the estimate of the covariance matrix $\text{Cov}[\hat{\beta}]$ are given by

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i^\top \hat{V}_i X_i \right)^{-1} \sum_{i=1}^{m} X_i^\top \hat{V}_i y_i, \tag{2.2}$$

$$S = \left(\sum_{i=1}^{m} X_i^\top \hat{V}_i X_i \right)^{-1}, \tag{2.3}$$

respectively, where $\hat{V}_i$ is an estimate of $V_i = \Sigma_i + Z_i D Z_i^\top$. In practice, the elements of V_i are generally defined as functions of a set of covariance parameters θ, $V_i = V_i(\theta)$. These covariance parameters θ are then estimated by either Maximum Likelihood (ML) or Restricted Maximum Likelihood (ReML) and are used to construct an estimate of V_i (Harville, 1977), which can then be used to get an estimate of β and $\text{Cov}[\hat{\beta}]$.

2.3.6 The marginal model

Instead of posing a model that consists of subject-specific random components like in LME models, we can fit a model with only fixed components and let the random components induce structure on the random error terms. This is the so-called marginal model, that has, for subject i, the form

$$y_i = X_i \beta + \epsilon_i^* \tag{2.4}$$

where the individual marginal error terms ϵ_i^* have mean 0 and covariance V_i. Typically, the covariance is taken to be unstructured, but if data arise as per the LME model specified above, then we have $V_i = \Sigma_i + Z_i D Z_i^\top$.

In contrast to the LME models, in the marginal models, all the randomness is treated as a nuisance and is modelled by the marginal error terms ϵ_i^*. Therefore, the marginal models do not require the specification of random effects, making them easier to specify than LME models. Moreover, the marginal models are somehow less restrictive because only V_i is required to be positive semi-definite. In contrast, in the case of LME models, both Σ_i and D have to be positive semi-definite which is more restrictive (West et al., 2006; Verbeke and Molenberghs, 2009; Molenberghs and Verbeke, 2011). However, the marginal models are only focused on population-averaged inferences and predictions, and do not offer the possibility to make inferences on random effects or to predict subject-specific profiles like LME models can. Nevertheless, subject-specific inferences or predictions are not generally of interest in longitudinal

neuroimaging studies and, therefore, a marginal approach will be of great utility.

For marginal models, the estimate of the fixed effect parameters β and the estimate of the covariance matrix $\mathrm{Cov}[\hat{\beta}]$ are given by

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i^\top W_i X_i\right)^{-1} \sum_{i=1}^{m} X_i^\top W_i y_i \tag{2.5}$$

$$S = \underbrace{\left(\sum_{i=1}^{m} X_i^\top W_i X_i\right)^{-1}}_{Bread} \underbrace{\left(\sum_{i=1}^{m} X_i^\top W_i \hat{V}_i W_i X_i\right)}_{Meat} \underbrace{\left(\sum_{i=1}^{m} X_i^\top W_i X_i\right)^{-1}}_{Bread}, \tag{2.6}$$

where W_i is the so-called working covariance matrix of individual i and $\hat{V}_i$ is an estimate of the subject covariance matrix V_i (Liang and Zeger, 1986; Diggle et al., 1994). The central part of the covariance estimate S can be conceptualised as a piece of meat between two slices of bread, giving rise to the name of Sandwich Estimator (SwE). If $m^{-1}\sum_{i=1}^{m} X_i^\top W_i \hat{V}_i W_i X_i$ consistently[2] estimates $m^{-1}\sum_{i=1}^{m} X_i^\top W_i V_i W_i X_i$, the SwE converges asymptotically to the true covariance matrix $\mathrm{Cov}[\hat{\beta}]$, even if W_i is misspecified (Eicker, 1963, 1967; Huber, 1967; White, 1980; Diggle et al., 1994).

If $W_i = I$, the identity matrix, then the estimate of β becomes equivalent to the OLS estimate $\hat{\beta}_{\mathrm{OLS}}$, that assumes i.i.d. error terms, and we obtain the simplest form of SwE which was firstly introduced by Eicker (1963, 1967):

$$S = \underbrace{\left(\sum_{i=1}^{m} X_i^\top X_i\right)^{-1}}_{Bread} \underbrace{\left(\sum_{i=1}^{m} X_i^\top \hat{V}_i X_i\right)}_{Meat} \underbrace{\left(\sum_{i=1}^{m} X_i^\top X_i\right)^{-1}}_{Bread}. \tag{2.7}$$

Note that this is different from the OLS estimate of variance, $\hat{\sigma}^2 \left(\sum_{i=1}^{m} X_i^\top X_i\right)^{-1}$, where $\hat{\sigma}^2$ is the OLS estimate of the assumed common variance of the error terms. Note also that, in practice, other choices for W_i are possible, by assuming a non-identity structure for W_i and parametrising it with a vector of parameters, which then has to be estimated (Liang and Zeger, 1986; Diggle et al., 1994). Such alternative choices are motivated by the fact that, even if the use of $W_i = I$ yields consistent estimates and has been shown to be almost as efficient as the Generalised Least Squares estimator in some settings (Liang and Zeger, 1986; McDonald, 1993), it may lead to a non-negligible loss of efficiency[3] in some other settings and more complicated forms of W_i can be used to improve efficiency (Zhao et al., 1992; Fitzmaurice, 1995). In

[2] An estimator of a parameter is said to be consistent if it converges in probability to the true value of the parameter. Here, this is the case if $\mathrm{plim}_{m\to\infty} m^{-1}\sum_{i=1}^{m} X_i^\top W_i(\hat{V}_i - V_i)W_i X_i = 0$.

[3] The efficiency of a scalar estimator is the inverse of estimator variance.

particular, Fitzmaurice (1995) shows that, in the context of clustered binary data, an important loss of efficiency may arise for within-cluster covariates when the within-cluster correlation is high. Nevertheless, Pepe and Anderson (1994) showed that using a non-diagonal working covariance matrix may lead to inaccurate estimates of β and, further, using a non-identity covariance matrix generally requires the use of iterative algorithms to estimate β and $\text{Cov}[\hat{\beta}]$. Finally, as shown in Section 2.4, the loss of efficiency can be limited by an appropriate construction of the design matrix. For all these reasons, in this thesis, we only focus on the use of the identity for W_i and will refer to the use of the corresponding marginal model as the SwE method. See, however, Li et al. (2013) for the use of non-diagonal working covariance matrix within the framework of neuroimaging data, and Pepe and Anderson (1994) in order to check the validity of using such working covariance matrices.

2.4 Construction of the design matrix

In longitudinal data, the covariates have generally a between-subject component and a within-subject component. For example, in the ADNI study described later in Section 2.6, the *Age* covariate has a between-subject component which can be summarised by the subject mean $\overline{Age}_i$ and a within-subject component which can be summarised by the difference with the subject mean $Age - \overline{Age}_i$. Including only the *Age* covariate in the design matrix means that we implicitly assume that the effects on the response is the same for both components. Actually, the effects of each component can be very different and, as shown by Neuhaus and Kalbfleisch (1998), the assessment of the effect of such between/within-subject covariates on the response can be very misleading. Therefore, it seems essential to follow the recommendation of Neuhaus and Kalbfleisch (1998) and systematically split this kind of covariates into between- and within-subject components and include them both in the design matrix. Moreover, as shown in Table 2.1, this also helps to improve the efficiency of the SwE method when assuming an identity working covariance matrix. This result shows that splitting the *Age* covariate makes the SwE nearly as efficient as the Generalised Least Squares (GLS) estimator. It also demonstrates the (well-known) importance of centring covariates when inference is made on the intercepts, as this can be of interest in longitudinal neuroimaging studies. As the only reason to use a nontrivial working covariance matrix is to improve efficiency, we found that these covariate-splitting results were a compelling reason to only focus on the use of an identity working covariance matrix in this thesis.

Model	Covariate	Relative efficiency $\rho = 0$	$\rho = 0.5$	$\rho = 0.95$
1	*Intercept*	1	0.88	0.40
	Age	1	0.88	0.40
2	*Intercept*	1	0.94	0.89
	$Age - \overline{Age}$	1	0.88	0.40
3	*Intercept*	1	0.92	0.87
	$\overline{Age}_i$	1	0.92	0.87
	$Age - \overline{Age}_i$	1	1	1
4	*Intercept*	1	0.94	0.89
	$\overline{Age}_i - \overline{Age}$	1	0.92	0.87
	$Age - \overline{Age}_i$	1	1	1

Table 2.1 Impact of splitting covariates into separate within- and between-subject covariates. The ages of all 817 subjects of the ADNI dataset (see Section 2.6) were used to construct 4 models: (1) *Intercept* and *Age*, (2) *Intercept* and centred *Age*, (3) *Intercept*, mean age per subject $\overline{Age}_i$, and intra-subject-centred age $Age - \overline{Age}_i$, and (4) *Intercept*, centred mean age per subject $\overline{Age}_i - \overline{Age}$, and intra-subject-centred age $Age - \overline{Age}_i$. The relative efficiency is shown for each model for 3 possible values of ρ, the common intra-visit correlation. Here, we define relative efficiency as the ratio between the variance of the GLS estimate and the variance of the SwE estimate.

2.5 Inference

In this section, we first describe how inferences can be performed using one of the models described in Section 2.3 . Then, we discuss how inferences are generally carried out in neuroimaging using a voxel-wise inference approach corrected for the multiple testing issue.

2.5.1 Statistical tests for a longitudinal model

To perform inference on a combination of the parameters, $\mathcal{H}_0 : C\beta = b_0$, a Wald statistic (Wald, 1943) is generally used:

$$T = (C\hat{\beta} - b_0)^\top (CSC^\top)^{-1}(C\hat{\beta} - b_0)/q, \tag{2.8}$$

where $\hat{\beta}$ and S are the estimates of β and $\mathrm{Cov}[\hat{\beta}]$ obtained using one of the methods described in Section 2.3, C is a matrix (or a vector) defining the combination of the parameters (contrast) tested and q is the rank of C.

To make the inference, this Wald statistic is compared to the distribution that it would follow under the null hypothesis. Unfortunately, this null distribution is generally unknown and needs to be estimated. To achieve this, the first possibility is to assume a parametric null distribution which complies with the assumptions of the model used. For example, in the cases of the N-OLS and the SS-OLS methods, this would be an F-distribution with q and $n - p$ degrees of freedom, where n is the total number of data points and p is the total number of parameters used in their respective OLS models. For the LME models, this would be an F-distribution with q and ν degrees of freedom, where ν is generally estimated using the formula proposed in Pinheiro and Bates (2000) or the Kenward-Roger effective degrees of freedom formula proposed in Kenward and Roger (1997). For the marginal models, a χ^2-distribution is often assumed, but an F-distribution with q and ν degrees of freedom is also sometimes considered with ν usually estimated using a simple arbitrary quantity without strong justifications (Hardin, 2001). Note that, for the SwE method with identity working covariance matrix, a more advanced parametric test proposed in Pan and Wall (2002) is reviewed in Section 3.2.4 and, still in Section 3.2.4, three novel alternative parametric tests are proposed with the goal to improve the accuracy of the inferences, particularly in small samples.

A second possibility to estimate the null distribution of the Wald statistic is to use a non-parametric resampling approach. In Neuroimaging, this is typically done using a permutation test which is is based on a resampling scheme without replacement (Nichols and Holmes, 2002; Winkler et al., 2014). Unfortunately, permutation tests rely on the assumption that the data is exchangeable under the null hypothesis. While this assumption of exchangeability is valid in some cases, it is harder to validate in the context of longitudinal data where the data is correlated and where different sources of heterogeneity may exist. Nevertheless, some alternative resampling approaches based on resampling schemes with replacement, that are generally referred to as bootstrap methods, can also be considered instead. In Chapter 4, we investigate such a type of resampling method, called Wild Bootstrap, to make non-parametric inferences in the context of the SwE method.

2.5.2 Multiple testing corrections

For a test of a single voxel, the False Positive Rate (FPR) is well-defined, as the probability of rejecting the null hypothesis when it is true. In neuroimaging, however, we typically want to make inference on the whole image by performing a test at every in-mask voxel in the image. This means that thousands of tests are carried out

simultaneously and controlling for the usual FPR would be inappropriate. Indeed, if there were 200,000 in-mask voxels and we control the FPR at 5% at every voxel, then, we would expect on average $0.05 \times 200{,}000 = 10{,}000$ false positive voxels in the image. This inflation of false positives is referred to as the multiple testing problem, and a correction is generally used to overcome this issue. Typically, this is done by controlling either the Family-Wise Error Rate (FWER), the probability of having at least one false positive in the image, or the False Discovery Rate (FDR), the expected proportion of false positives among the rejected null hypotheses.

Several strategies can be used to control the FWER. The simplest is the Bonferroni correction which consists of dividing the significance level used for a FPR control by the number of tests. For our example with an FPR significance level of 5% and 200,000 in-mask voxels, the Bonferroni-corrected significance level would be 2.5×10^{-7}. While the Bonferroni correction is always valid, it becomes conservative when data are highly dependent. As image data exhibits strong spatial correlation, the Bonferroni correction does not perform well and is generally not used. Another strategy to control the FWER attempts to account for the spatial dependence using Random Field Theory (RFT; see, e.g., Worsley et al., 1996). Nevertheless, RFT relies on several assumptions (see, e.g., Petersson et al., 1999) that, to our knowledge, have not been validated in the context of longitudinal models. A third strategy is based on how the FWER actually corresponds to the probability that the maximum statistic in the image is detected as a false positive when the null hypothesis is true. Therefore, the control of the FWER can be achieved by comparing the statistics in the image to the maximum statistic null distribution of the image. While the latter can be difficult to estimate with a parametric approach, it can be relatively easily estimated using a non-parametric approach provided that its assumptions are justified. In neuroimaging, this is generally done using permutation. Nevertheless, as explained above in Section 2.5.1, in the context of longitudinal data, permutation is generally not feasible and hence we consider a bootstrap resampling method instead (see Chapter 4).

Instead of controlling for the FWER, another type of correction introduced to neuroimaging by Genovese et al. (2002) and generally less conservative, is based on the False Discovery Rate (FDR) control. In practice, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) is generally used to control FDR, and has been used widely in neuroimaging. This is due to in part how the Benjamini-Hochberg method is valid under positive dependence (Benjamini and Yekutieli, 2001), and doesn't show the same sort of ultra-conservativeness as the Bonferrroni correction with smooth image data.

2.6 The ADNI dataset

In this thesis, in order to demonstrate the proposed methods on a real longitudinal neuroimaging dataset, we use a dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

Scanning time	N	MCI	AD	Total
0 month	229	400	188	817
6 months	208	346	159	713
12 months	196	326	138	660
18 months	n/a	286	n/a	286
24 months	172	244	105	521
36 months	147	170	n/a	317

Table 2.2 Numbers of subjects scanned at baseline (0 month) and follow-up (6, 12, 18, 24 and 36 months) for the Normal controls (N), Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) subjects in the ADNI dataset.

The dataset considered in this thesis is a modified version of the dataset produced and detailed by Hua et al. (2013). In brief, the dataset in Hua et al. (2013) consisted on 3314 images obtained after applying Tensor Based Morphometry (TBM) on 3314 brain MRI scans from 229 healthy elderly Normal controls (age: 76.0 ± 5.0 years, 119 Male (M)/110 Female (F)), 400 individuals with amnestic MCI (age: 74.8 ± 7.4 years, 257 M/143 F), and 188 probable AD patients (age at screening: 75.4 ± 7.5 years, 99 M/89 F). As shown in Table 2.2, the subjects were scanned at screening and followed up at 6, 12, 18 (MCI only), 24, and 36 months (Normal and MCI only) with visit counts of 4.16 ± 1.21, 4.43 ± 1.61 and 3.14 ± 1.07 for the Normal, MCI and AD subjects, respectively. More precisely, 817 screening TBM images were produced by considering the 817 screening scans and a Minimal Deformation Target (MDT) image, obtained from the scans of 40 randomly selected Normal subjects, as baseline; 2497 longitudinal TBM images were produced by considering, for each subject, the follow-up scans and the corresponding screening scan as baseline. More details about this dataset can be found in Hua et al. (2013). The 2497 longitudinal TBM images corresponds to longitudinal changes and not absolute measurements. Therefore, we multiplied them with their corresponding TBM screening image in order to produce 2497 TBM images reflecting the brain volumes changes compared to a unique baseline, the MDT image. We considered these modified 2497 TBM images with the original 817 screening TBM images as the dataset to be analysed.

Chapter 3

The Sandwich Estimator method

3.1 Introduction

As mentioned in Section 2.3.6, the Sandwich Estimator (SwE) method, which uses the marginal model with the identity matrix as working covariance matrix, seems to be a very attractive tool to analyse longitudinal neuroimaging data. First, it is simpler to specify than alternative methods like Linear Mixed Effects (LME) models as there is no need to specify random effects or to make assumptions about the intra-visit covariance structure of the error terms. Second, it is free of any iterative algorithms, allowing the method to be fast and without the risk of convergence issues. Third, it is robust, at least asymptotically, against the misspecification of the working covariance matrix. Nevertheless, the main issue is that many longitudinal neuroimaging studies have a small number of subjects, breaking down the assumption of large samples, and, unfortunately, several studies have shown that the behaviour of the SwE method without any small sample considerations may lead to inaccurate results (MacKinnon and White, 1985; Chesher and Jewitt, 1987; Long and Ervin, 2000). That is why several authors proposed to use different adjustments to improve its behaviour, mainly in the case of small samples. However, there does not seem to exist a clear consensus about which of these adjustments should be used in practice. Moreover, the majority of the literature has been focused on cross-sectional data, meaning that the related conclusions may differ in the context of longitudinal data. For all these reasons, in this chapter, we review, extend and propose different adjustments aiming to improve the small sample behaviour of the SwE method. In order to isolate the best combinations of these adjustments, we evaluate them using intensive Monte Carlo simulations in settings important for longitudinal neuroimaging data analysis. Also, using Monte Carlo simulations, we compare the SwE method to popular alternative approaches. Finally,

we illustrate the SwE method by applying it to the real ADNI dataset introduced in Section 2.6. Note that a part of the content of this chapter has been published in Guillaume et al. (2014).

3.2 Methods

In this section, we first detail the SwE method in its most general form. Then, we review and propose several adjustments which can be used to improve its behaviour, particularly in small samples. We also detail how the SwE method was evaluated.

3.2.1 The classic SwE method

The SwE method considers the marginal linear regression model introduced in Section 2.3.6, such that, for each subject i, we have

$$y_i = X_i\beta + \epsilon_i^* \qquad \text{(2.4 revisited)}$$

where y_i is a vector of n_i observations for individual $i = 1, 2, \ldots, m$, β is a vector of p fixed effects which is linked to y_i by the $n_i \times p$ design matrix X_i and the individual marginal error terms ϵ_i^* have mean 0 and covariance V_i.

To estimate the parameters β, the SwE method assumes that the error terms are independent and identically distributed (i.i.d.) and simply uses the OLS estimator such that

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i^\top X_i\right)^{-1} \sum_{i=1}^{m} X_i^\top y_i. \qquad (3.1)$$

This may seem wrong to assume i.i.d. error terms in the estimation of β, but it was shown that $\hat{\beta}_{\text{OLS}}$ is a consistent estimator of β, even if the data is correlated (Liang and Zeger, 1986). The main issue of using OLS regressions in clustered data actually arises with the OLS estimator of the covariance matrix of the parameter $\text{Cov}(\hat{\beta})$ which is not consistent (Kauermann and Carroll, 2001). That is why, in the SwE method, we use instead the so-called SwE with the identity matrix as working covariance matrix (see Section 2.3.6) which was firstly introduced by Eicker (1963, 1967):

$$S = \underbrace{\left(\sum_{i=1}^{m} X_i^\top X_i\right)^{-1}}_{Bread} \underbrace{\left(\sum_{i=1}^{m} X_i^\top \hat{V}_i X_i\right)}_{Meat} \underbrace{\left(\sum_{i=1}^{m} X_i^\top X_i\right)^{-1}}_{Bread}. \qquad \text{(2.7 revisited)}$$

As mentioned in Section 2.3.6, its property is that it consistently estimates the true covariance matrix of the parameters $\text{Cov}(\hat{\beta})$ if $m^{-1}\sum_{i=1}^{m} X_i^\top \hat{V}_i X_i$ consistently estimates $m^{-1}\sum_{i=1}^{m} X_i^\top V_i X_i$. From this property, we learn that, in order to get an accurate estimate of $\text{Cov}(\hat{\beta})$, we need a large sample and a consistent estimate of $m^{-1}\sum_{i=1}^{m} X_i^\top V_i X_i$. In practice, obtaining a consistent estimate can be achieved by estimating the subject covariance matrices from the subject residuals $e_i = y_i - X_i\hat{\beta}$ by

$$\hat{V}_i = e_i e_i^\top \tag{3.2}$$

(Diggle et al., 1994). Using Equation (3.2) in Equation (2.7), we then obtain the most simple SwE version. In the literature, this SwE version is often referred to as HC0 (see, e.g., Long and Ervin, 2000) where "HC" stands for "Heteroscedasticity Consistent" and "0" stands for the fact that no small sample bias adjustment is used (see Section 3.2.2). In this thesis, however, we will refer to this SwE version as S_0^{Het} where the subscript "0" indicates that there is no bias adjustment, and the superscript "Het" stands for "Heterogeneous" subject covariance matrices and is used to contrast with another type of SwE introduced in Section 3.2.3.

To perform inference on a combination of the parameters, $\mathcal{H}_0 : C\beta = b_0$, a Wald statistic is generally used:

$$T = (C\hat{\beta} - b_0)^\top (CSC^\top)^{-1}(C\hat{\beta} - b_0)/q \tag{2.8 revisited}$$

where C is a matrix (or a vector) defining the combination of the parameters (contrast) tested and q is the rank of C. At large samples, this Wald statistic follows a χ^2_q distribution under the null hypothesis. In small samples, while the obvious choice would be an F-distribution with q and $n - p$ degrees of freedom, we show in Section 3.2.4 that this is not a good approximation of the true null distribution of T when the SwE method is used.

3.2.2 Small sample bias adjustments

In small samples, it is well known that the use of the standard SwE S_0^{Het} may lead to incorrect inferences (MacKinnon and White, 1985; Chesher and Jewitt, 1987; Long and Ervin, 2000). One of the explanation for this effect is that, since S_0^{Het} uses the Maximum Likelihood Estimate for each V_i, it is generally biased downward and tends to make liberal inferences (i.e. inflated False Positive Rates). This has pushed several authors to propose different small sample bias adjustments to improve the behaviour

of the SwE in small samples. Nevertheless, before presenting these adjustments, it seems important to show how a bias appears in small samples. To do that, let us first write the relationship between the residuals of all the observations e and the data of all the observations y:

$$e = (I - H)y, \tag{3.3}$$

where $H = X(X^\top X)^{-1}X^\top$ is the so-called Hat matrix with X being the grand design matrix (i.e. the $n \times p$ stacked matrix of the m X_i's, where $n = \sum_i n_i$ is the total number of observations). Taking the covariance of both sides of Equation (3.3) and noting that H is symmetric, we get

$$\text{Cov}[e] = (I - H)\text{Cov}[y](I - H). \tag{3.4}$$

From Equation (3.4), we see that the covariance matrix of the residuals is not equal to the covariance matrix of the data. In particular, the expectation of the estimator $\hat{V}_i = e_i e_i^\top$ used in S_0^{Het} is given by

$$\begin{aligned}\mathbb{E}[e_i e_i^\top] =& \text{Cov}[e_i] \\ =& \sum_{j=1}^{m} (I - H)_{ij} V_j (I - H)_{ji},\end{aligned} \tag{3.5}$$

where $(I - H)_{ij}$ is the block matrix in $(I - H)$ whose rows correspond to subject i and columns to subject j. We can see that the expectation of $\hat{V}_i = e_i e_i^\top$ is typically not equal to V_i, indicating that it is a biased estimator of V_i. Nevertheless, noting that, in many practical situations, $(I - H)$ tends to become closer and closer of the identity matrix when the number of data points increases faster than the rank of X, the bias is generally expected to decrease when the number of data points increases, explaining why, in the literature, S_0^{Het} has been found accurate in large samples, but not in small samples. It seems therefore preferable to use an alternative estimator of V_i which accounts for the small sample bias. Below, we describe seven of these alternative estimators, leading to seven alternative SwE.

S_1^{Het}

S_1^{Het} (or HC1) was first proposed by Hinkley (1977) and consists of using the raw residuals e_i multiplied by $\sqrt{n/(n-p)}$ instead of the raw residuals e_i in the estimation of each V_i. This correction can be justified by first (wrongly) assuming that the error

terms are i.i.d. with variance σ^2. In this case, a good estimator for σ^2 would simply be the unbiased OLS estimator given by

$$\hat{\sigma}^2 = \frac{e^\top e}{n-p} \tag{3.6}$$

which is equivalent to use the raw residuals e_i multiplied by $\sqrt{n/(n-p)}$ instead of the raw residuals in the OLS Maximum Likelihood Estimator of variance. By analogy, we can expect each V_i to be estimated with less bias, making, in turn, S_1^{Het} less biased than S_0^{Het}. Note that, instead of adjusting the residuals, S_1^{Het} can simply be obtained by multiplying S_0^{Het} by $n/(n-p)$.

S_2^{Het}

S_2^{Het} (or HC2) was first proposed by Horn et al. (1975) and consists of using the adjusted residuals $e_{ik}/(1-h_{ik})^{1/2}$(where e_{ik} and h_{ik} are the raw residual and the diagonal element of H corresponding to the observation of subject i at visit k) instead of the raw residuals e_{ik}. To justify this correction, like with S_1^{Het}, it suffices to assume (wrongly) that the error terms are i.i.d. with variance σ^2. In this case, noting that the Hat matrix H is idempotent, Equation (3.4) becomes

$$\text{Cov}[e] = (I-H)\sigma^2 \tag{3.7}$$

and, from this, we get, for the residual of subject i at visit k,

$$\text{var}\left[\frac{e_{ik}}{(1-h_{ik})^{1/2}}\right] = \sigma^2. \tag{3.8}$$

This suggests that using the adjusted residual $e_{ik}/(1-h_{ik})^{1/2}$ instead of the raw residuals should yield less biased estimates of the within-subject covariance matrices and, consequently, should reduce the bias existing in S_0^{Het}.

S_3^{Het}

S_3^{Het} (or HC3) consists of using $e_{ik}/(1-h_{ik})$ instead of the raw residuals e_{ik}. It is actually a simplification of the jackknife estimator of $\text{Cov}[\hat{\beta}]$ proposed by MacKinnon and White (1985) in the context of independent data:

$$\frac{n-1}{n}(X^\top X)^{-1}\left(X^\top \Omega X - \frac{1}{n}(X^\top u u^\top X)\right)(X^\top X)^{-1} \tag{3.9}$$

where u is a vector of adjusted residuals with elements $e_{ik}/(1-h_{ik})$ and Ω is a diagonal matrix with diagonal elements $e_{ik}^2/(1-h_{ik})^2$. In practice, this jackknife estimator is generally simplified by dropping the multiplicative term $(n-1)/n$ and the term $1/n(X^\top uu^\top X)$, leading to a SwE version S_3^{Het} which uses $e_{ik}/(1-h_{ik})$ instead of the raw residuals. Note that, to our knowledge, the literature (see, e.g., Long and Ervin, 2000) does not give a clear justification about this simplification while the values of the two omitted terms could be influential in very small samples. However, we can observe that the first term is always inferior to one while the diagonal elements of the second term are always positive. Neglecting the fact that the off-diagonal elements of the second term may be negative, it seems clear that their omissions will have the tendency to inflate the estimate of $\text{Cov}[\hat{\beta}]$, which can be considered acceptable in practice to get valid, but not necessary optimal, inferences. Note also that, like S_1^{Het} and S_2^{Het}, S_3^{Het} is based on the assumption that the error terms are i.i.d. (Efron, 1982).

$S_{\text{C2}}^{\text{Het}}$ and $S_{\text{C3}}^{\text{Het}}$

All three SwE versions S_1^{Het}, S_2^{Het} and S_3^{Het} assume i.i.d. error terms and do not consider the clustered nature which may exist in the data. That is why, in the context of clustered data, some authors prefer to adjust the residuals by multiplying each subject residual e_i with $(I-H)_{ii}^{-1/2}$ (Kauermann and Carroll, 2001; Bell and McCaffrey, 2002) or $(I-H)_{ii}^{-1}$ (Mancl and DeRouen, 2001), leading to two alternative SwE versions which can be seen as clustered versions of S_2^{Het} and S_3^{Het}, respectively. Due to the latter, in this thesis, they are referred to as $S_{\text{C2}}^{\text{Het}}$ and $S_{\text{C3}}^{\text{Het}}$, respectively. Note that, considering the bias adjustment used in $S_{\text{C3}}^{\text{Het}}$ and using Equation (3.5), we get

$$\mathbb{E}[(I-H)_{ii}^{-1}e_ie_i^\top(I-H)_{ii}^{-1})] = V_i + (I-H)_{ii}^{-1}\left(\sum_{j\neq i}^{m}(I-H)_{ij}V_j(I-H)_{ji}\right)(I-H)_{ii}^{-1}, \tag{3.10}$$

indicating that $S_{\text{C3}}^{\text{Het}}$ is likely to be biased.

$S_{\text{U1}}^{\text{Het}}$ and $S_{\text{U2}}^{\text{Het}}$

All the bias adjustments reviewed previously are based on some assumptions (e.g., i.i.d. error terms) which may not be valid in practice. As a consequence, while they are expected to reduce the bias existing in the SwE, they may fail to remove it entirely. However, finding an unbiased estimator for each V_i and consequently for $\text{Cov}[\hat{\beta}]$ is not, in theory, impossible. Indeed, vectorising Equation (3.4) with the half-vectorisation

operator vech[1], using its relationship with the vectorisation operator vec[2] (Abadir and Magnus, 2005, Exercice 11.27) and using the relationship between the vec operator and the Kronecker product (Abadir and Magnus, 2005, Exercice 10.18), we get

$$\begin{aligned}\text{vech}[\text{Cov}[e]] =& \text{vech}[(I-H)\text{Cov}[y](I-H)] \\ =& D_n^+ \text{vec}[(I-H)\text{Cov}[y](I-H)] \\ =& D_n^+((I-H)\otimes(I-H))\text{vec}[\text{Cov}[y]] \\ =& D_n^+((I-H)\otimes(I-H))D_n\text{vech}[\text{Cov}[y]], \end{aligned} \tag{3.11}$$

where D_n is the so-called duplication matrix (Abadir and Magnus, 2005, Chapter 11) and $D_n^+ = (D_n^\top D_n)^{-1}D_n^\top$. Note that vech[Cov[$y$]] contains a lot of zeros since $\text{cov}[y_{ik}, y_{jk'}] = 0$ for all $i \neq j$. Therefore, we can define the column vector vecu[Cov[y]] as the vector obtained by removing all the elements of vech[Cov[y]] corresponding to the covariances involving two different subjects, and D_u as the matrix obtained by removing in D_n the columns corresponding to the elements removed in vech[Cov[y]] to obtain vecu[Cov[y]]. Then, Equation (3.11) becomes

$$\text{vech}[\text{Cov}[e]] = D_n^+((I-H)\otimes(I-H))D_u\text{vecu}[\text{Cov}[y]]. \tag{3.12}$$

From this equation, denoting the matrix $P = D_n^+((I-H)\otimes(I-H))D_u$ and multiplying both sides by $(P^\top P)^{-1}P^\top$, we get

$$(P^\top P)^{-1}P^\top\text{vech}[\text{Cov}[e]] = \text{vecu}[\text{Cov}[y]]. \tag{3.13}$$

We can then define an estimator for vecu[Cov[y]] as

$$(P^\top P)^{-1}P^\top\text{vech}[ee^\top]. \tag{3.14}$$

Taking its expectation, we directly see that it is unbiased. Therefore, if we use it to estimate each V_i in the SwE, it should produce an unbiased SwE that we will refer to as $S_{\text{U1}}^{\text{Het}}$ in this thesis.

As an alternative to $S_{\text{U1}}^{\text{Het}}$, we could only consider the within-subject covariance matrices Cov[e_i]'s and forget about the between-subject covariances. In this case, defining vecu[$ee^\top$] as the vector obtained by removing all the between-subject covariances in

[1]The vech operator is the operator which transforms a symmetric matrix A into a column vector by stacking all the columns of the lower triangular part of A on top of one another.

[2]The vec operator is the operator which transforms a matrix A into a column vector by stacking all the columns of A on top of one another.

vech$[ee^\top]$, and P_u as the matrix obtained by removing from P the rows corresponding to the between-subject covariances, we can define another estimator for vecu[Cov$[y]$] as

$$(P_u^\top P_u)^{-1} P_u^\top \text{vecu}[ee^\top], \tag{3.15}$$

which is also unbiased. In this thesis, we will refer to the resulting SwE as $S_{\text{U2}}^{\text{Het}}$.

Note that, while the estimators of vecu[Cov$[y]$] used in $S_{\text{U1}}^{\text{Het}}$ and $S_{\text{U2}}^{\text{Het}}$ are unbiased, they can yield very bad estimates for some V_i's. For example, there is no restriction imposing that the variances should be positive or that the correlations should be inferior or equal to 1. Therefore, in practice, we propose to systematically make a spectral decomposition of each $\hat{V}_i$ and, if at least one of the eigenvalue is negative, to replace it by a more stable estimate like, for example, the one used in S_3^{Het} or, in the case of very small negative eigenvalues, to set them to zero, before reconstructing the spectral decomposition. Finally, note that, to the best of our knowledge, we are the first to propose these two novel SwE versions.

3.2.3 The homogeneous SwE

The standard SwE versions S^{Het} estimates a separate V_i for each subject. In particular, the versions S_0^{Het}, S_1^{Het}, S_2^{Het}, S_3^{Het}, $S_{\text{C2}}^{\text{Het}}$ and $S_{\text{C3}}^{\text{Het}}$ estimate each V_i based only on the residuals of the i^{th} subject (see, for example, Equation 3.2). Nevertheless, as suggested in Pan (2001), if the studied population can be subdivided into n_{G} groups within which the subjects are sharing similar properties, we may assume that the variances and covariances over subjects within each group are actually homogeneous. For instance, in the ADNI study (see Section 2.6), the whole population can be divided into three groups: the Normal control (N), Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) groups in which the subjects can be assumed to share the same variances and covariances. We argue that this is a reasonable assumption as virtually the standard longitudinal neuroimaging analysis assumes homogeneous variance over all subjects. Therefore, we can define an alternative version of the SwE S_0^{Hom} which relies on the assumption of a common covariance matrix V_{0g} for all the individuals belonging to group $g = 1, \ldots, n_{\text{G}}$. To estimate V_{0g}, the observations have to be firstly classified into k_g visit categories (homogeneous groups) consistently defined between subjects in group g. For example, in the ADNI study, the MCI subjects were scanned at 0, 6, 12, 18, 24 and 36 months allowing us to divide the observations into $k_{\text{MCI}} = 6$ visit categories. Then, defining $m_{gkk'}$ as the number of subjects in group g

who have data at both visit k and k', e_{ik} as the residual of subject i at visit k and $\mathcal{I}(g,k,k')$ as the subset of subjects in group g who have data at both visit k and k', the k^{th}diagonal element of V_{0g} can then be estimated by

$$(\hat{V}_{0g})_{kk} = \frac{1}{m_{gkk}} \sum_{i \in \mathcal{I}(g,k,k)} e_{ik}^2. \tag{3.16}$$

The off-diagonal element of V_{0g} corresponding to the visits k and k' can be estimated by

$$(\hat{V}_{0g})_{kk'} = \hat{\rho}_{0gkk'} \sqrt{(\hat{V}_{0g})_{kk} (\hat{V}_{0g})_{k'k'}} \tag{3.17}$$

where $\hat{\rho}_{0gkk'}$ is an estimate of the correlation at visits k and k' in the group g and which can be computed by

$$\hat{\rho}_{0gkk'} = \frac{\sum\limits_{i \in \mathcal{I}(g,k,k')} e_{ik} e_{ik'}}{\sqrt{\left(\sum\limits_{i \in \mathcal{I}(g,k,k')} e_{ik}^2 \right) \left(\sum\limits_{i \in \mathcal{I}(g,k,k')} e_{ik'}^2 \right)}}. \tag{3.18}$$

Note that we could estimate the off-diagonal elements of V_{0g} directly by

$$(\hat{V}_{0g})_{kk'} = \frac{1}{m_{gkk'}} \sum_{i \in \mathcal{I}(g,k,k')} e_{ik} e_{ik'}. \tag{3.19}$$

Unfortunately, if there is missing data, as the subset of subjects could differ from the one used in the estimation of the variances $(V_{0g})_{kk}$ and $(V_{0g})_{k'k'}$, there will be no guarantee that the corresponding correlation estimates would be inferior or equal to 1. Using Equations (3.17) and (3.18), the correlations are ensured to be always inferior or equal to 1 and, therefore, we recommend their uses instead of Equation (3.19). Note also that, due to the possible presence of missing data, $\hat{V}_{0g}$ may not be positive semi-definite and, as a consequence, may lead to inaccurate results. Therefore, in presence of missing data, we can make a spectral decomposition of $\hat{V}_{0g}$ and check whether all the eigenvalues of $\hat{V}_{0g}$ are positive. If this is not the case, we set all the negative eigenvalues to zero and reconstruct $\hat{V}_{0g}$ with the new eigenvalues, ensuring that $\hat{V}_{0g}$ is positive semi-definite. Thus, in this SwE version that will be referred to as S_0^{Hom}, each $\hat{V}_i$ corresponds to a subset of the corresponding common covariance matrix $\hat{V}_{0g}$ depending on the visits measured for subject i. If the assumption of a common covariance matrix over subjects in a same group is valid, then each V_i should be more

efficiently estimated in comparison to the standard approach. Finally, note that this new SwE version depends on the way the population is subdivided and has two extreme cases, one assuming a single group and the other considering m homogeneous groups, equivalent to the standard SwE S_0^{Het}.

Small sample bias correction considerations

In Equations (3.16) and (3.18), we do not use any bias corrections as discussed in Subsection 3.2.2. That is for this reason that the resulting SwE version is referred to as S_0^{Hom}. However, like for the standard S^{Het} versions, we can apply similar bias corrections, leading to the bias corrected homogeneous SwE versions S_1^{Hom}, S_2^{Hom}, S_3^{Hom}, $S_{\text{C2}}^{\text{Hom}}$ and $S_{\text{C3}}^{\text{Hom}}$, $S_{\text{U2}}^{\text{Hom}}$ and $S_{\text{U2}}^{\text{Hom}}$. For the SwE versions S_1^{Hom}, S_2^{Hom}, S_3^{Hom}, $S_{\text{C2}}^{\text{Hom}}$ and $S_{\text{C3}}^{\text{Hom}}$, the corrections are the same as their corresponding heterogeneous versions and simply consist of replacing the raw residuals in Equations (3.16) and (3.18) by the adjusted residuals as described in Section 3.2.2. For the $S_{\text{U1}}^{\text{Hom}}$ and $S_{\text{U2}}^{\text{Hom}}$ versions, this is a little bit more complicated. For them, Equations (3.16) and (3.18) cannot be used anymore. Instead, we can modify Equation (3.12) as follows:

$$\begin{aligned}\text{vech}[\text{Cov}[e]] =& P\text{vecu}[\text{Cov}[y]] \\ =& P^{\text{Hom}}\text{vecu}[V_0],\end{aligned} \tag{3.20}$$

where $\text{vecu}[V_0]$ is a vector obtained by stacking together all the unique elements of each V_{0g} and P^{Hom} is a matrix obtained in such a way that the column corresponding to the element $(V_{0g})_{kk'}$ is the sum of all columns in P corresponding to the element $(V_{0g})_{kk'}$. We can then define the SwE version $S_{\text{U1}}^{\text{Hom}}$ for which the elements of $\text{vecu}[V_0]$ are estimated by

$$({P^{\text{Hom}}}^\top P^{\text{Hom}})^{-1}{P^{\text{Hom}}}^\top \text{vech}[ee^\top]. \tag{3.21}$$

Similarly, we can define the SwE version $S_{\text{U2}}^{\text{Hom}}$ for which the elements of $\text{vecu}[V_0]$ are estimated by

$$({P_u^{\text{Hom}}}^\top P_u^{\text{Hom}})^{-1}{P_u^{\text{Hom}}}^\top \text{vecu}[ee^\top]. \tag{3.22}$$

where the matrix P_u^{Hom} is obtained by removing from P^{Hom} the rows corresponding to the between-subject covariances in $\text{vech}[ee^\top]$. Like it is the case with their heterogeneous versions, there is no guarantee that each $\hat{V}_{0g}$ used in $S_{\text{U1}}^{\text{Hom}}$ and $S_{\text{U2}}^{\text{Hom}}$ will be positive semi-definite. Therefore, we propose to make a spectral decomposition of

them to check if all the eigenvalues are positive. If it is not the case and if the negative eigenvalue are small, then, we set them to 0 before reconstructing the spectral decomposition. If the negative eigenvalues are not negligible, we propose instead to replace each problematic $\hat{V}_{0g}$ by another bias corrected estimate of it like, for example, the one used in S_3^{Hom}.

3.2.4 Inferences in small samples

As mentioned at the end of Section 3.2.1, with the SwE method, inference on a combination of the parameters, $\mathcal{H}_0 : C\beta = b_0$, can be done using a Wald test:

$$T = (C\hat{\beta} - b_0)^\top (CSC^\top)^{-1}(C\hat{\beta} - b_0)/q, \qquad \text{(2.8 revisited)}$$

where C is a matrix (or a vector) defining the combination of the parameters (contrast) tested and q is the rank of C. As mentioned also in Section 3.2.1, under the null hypothesis and at large samples, this Wald test follows a χ^2_q distribution. While this null distribution is often considered in practice, it does not account for the variability of the SwE which can be non-negligible in small samples. That is the reason why some authors proposed to alter the null distribution of this Wald statistic by accounting for the small sample nature of the data (Lipsitz et al., 1999; Fay and Graubard, 2001; Hardin, 2001; Kauermann and Carroll, 2001; Mancl and DeRouen, 2001; Bell and McCaffrey, 2002; Pan and Wall, 2002; Waldorp, 2009). Most of the proposed adjustments consist of using an F-distribution with q and ν degrees of freedom instead of a χ^2_q distribution. The challenge is then to determine an appropriate value for ν. In this thesis, we focus our attention on the approximate F-test proposed in Pan and Wall (2002) (referred to as Pan test in this thesis) and develop three alternative F-tests (referred to, in this thesis, as Test I, Test II and Test III) which can be used with the SwE method.

The Pan statistical test

In the context of Generalised Estimating Equations, to account for the variability of the SwE in the Wald test, Pan and Wall (2002) proposed to assume that the contrasted SwE $CSC^\top$ follows a Wishart distribution $\mathcal{W}_q[\nu, C\text{Cov}[\hat{\beta}]C^\top/\nu]$. Noting that, under $\mathcal{H}_0$, $(C\hat{\beta} - b_0)/\sqrt{\nu} \sim \mathcal{N}[0, C\text{Cov}[\hat{\beta}]C^\top/\nu]$ and assuming that $C\hat{\beta}$ and

$CSC^\top$ are independent, we get (Härdle and Simar, 2012, Theorems 5.8 & 5.9)

$$\nu \left(\frac{C\hat{\beta} - b_0}{\sqrt{\nu}} \right)^\top (CSC^\top)^{-1} \left(\frac{C\hat{\beta} - b_0}{\sqrt{\nu}} \right) \sim \frac{\nu q}{\nu - q + 1} \mathcal{F}[q, \nu - q + 1], \tag{3.23}$$

which finally leads to the test statistic

$$\frac{\nu - q + 1}{\nu q} (C\hat{\beta} - b_0)^\top (CSC^\top)^{-1} (C\hat{\beta} - b_0) \sim \mathcal{F}[q, \nu - q + 1]. \tag{3.24}$$

The issue is then to use an appropriate value for the degrees of freedom ν. Pan and Wall (2002) proposed to choose ν such that an empirical estimate of the covariance matrix of $CSC^\top$ is close to the one predicted by the theory of Wishart distributions (Abadir and Magnus, 2005, Exercice 11.23), i.e.

$$\mathrm{Cov}[\mathrm{vec}[CSC^\top]] = \frac{2}{\nu} N_q \left((C\mathrm{Cov}[\hat{\beta}]C^\top) \otimes (C\mathrm{Cov}[\hat{\beta}]C^\top) \right), \tag{3.25}$$

where N_q is the so-called symmetrizer matrix (Abadir and Magnus, 2005, Chapter 11). Pan and Wall (2002) proposed to empirically estimate $\mathrm{Cov}[\mathrm{vec}[CSC^\top]]$ by

$$\widehat{\mathrm{Cov}}_{\mathrm{Pan}}[\mathrm{vec}[CSC^\top]] = \frac{m}{m-1} \sum_{i=1}^{m} (\mathrm{vec}[Q_i] - \frac{1}{m}\mathrm{vec}[CSC^\top])(\mathrm{vec}[Q_i] - \frac{1}{m}\mathrm{vec}[CSC^\top])^\top \tag{3.26}$$

where

$$Q_i = C \left(\sum_{j=1}^{m} X_j^\top X_j \right)^{-1} X_i^\top \hat{V}_i X_i \left(\sum_{j=1}^{m} X_j^\top X_j \right)^{-1} C^\top \tag{3.27}$$

is the contribution of subject i to the contrasted SwE $CSC^\top$. Finally, Pan and Wall (2002) suggested to estimate the degrees of freedom ν by minimising the sum of squared errors between $\nu\,\mathrm{vec}[\widehat{\mathrm{Cov}}_{\mathrm{Pan}}[\mathrm{vec}[CSC^\top]]]$ and $\nu\,\mathrm{vec}[\mathrm{Cov}[\mathrm{vec}[CSC^\top]]]$.

Note that the empirical estimator of $\mathrm{Cov}[\mathrm{vec}[CSC^\top]]$ given in Equation (3.26) was given in Pan and Wall (2002) without mentioning the conditions under which it is valid. Therefore, it is rather difficult for us to state with certainty the exact assumptions needed for its validity. Nevertheless, as a first guess, it seems that, in order to be valid, Equation (3.26) relies on the assumptions that the Q_i's are independent, have the same expectation and have the same covariance matrix. Unfortunately, as the residuals are typically correlated, the Q_i's will not be exactly independent, particularly in small

samples. Furthermore, for a general design, as the within-subject covariance matrices V_i's or the subject design matrices X_i's may differ across subject, it is also unlikely that the Q_i's will have the same expectation or the same covariance matrix. This indicates that the Pan test might be inaccurate in some designs. Also, Equation (3.26) will definitively not be valid if we are using a homogeneous version of the SwE. Indeed, in this case, the Q_i's related to the group g will be drawn from the same common covariance matrix estimate $\hat{V}_{0g}$ and will therefore be highly correlated, breaking down the assumption of independence. This means that the Pan test cannot be considered when a homogeneous version of the SwE is used.

Test I

Independently of Pan and Wall (2002), we have actually developed an alternative test which is partly similar to the Pan test and was published in Guillaume et al. (2014). It follows the same idea of assuming a Wishart distribution for the contrasted SwE $CSC^\top$ which then leads to the same test statistic as given in Equation (3.24). The difference with Pan and Wall (2002) is that we used a different strategy to estimate the degrees of freedom ν by first using an alternative empirical estimator for $\mathrm{Cov}[\mathrm{vec}[CSC^\top]]$ and, second, by equating the trace of it with the theoretical expression for $\mathrm{Cov}[\mathrm{vec}[CSC^\top]]$ given in Equation (3.25).

For this test, one of our goals was usability for both the heterogeneous and homogeneous versions of the SwE. Therefore, as the heterogeneous SwE can be seen as a particular case of the homogeneous version which would consider m homogeneous groups of one subject, we give only details for the homogeneous version.

To derive another estimator for $\mathrm{Cov}[\mathrm{vec}[CSC^\top]]$, let us first assume that there is no missing data. In such a case, assuming the use of a homogeneous version of the SwE with n_G groups, V_{0g} would be estimated by

$$\hat{V}_{0g} = \frac{1}{m_g} \sum_{i \in \mathcal{I}(g)} e_i^* e_i^{*\top} \tag{3.28}$$

where $\mathcal{I}(g)$ is the subset of subjects belonging to group g and e_i^* is an adjusted version of the residuals of subject i. If each e_i^* is correctly adjusted in such a way that each covariance matrix $\mathrm{Cov}[e_i^*]$ is equal to the covariance matrix of its corresponding true error term $\mathrm{Cov}[\epsilon_i^*]$, then they can be assumed to follow a Normal distribution with mean 0 and covariance matrix V_{0g} for all $i \in \mathcal{I}(g)$. Then, for all $i \in \mathcal{I}(g)$, we would

have

$$B_i = \frac{1}{m_g} e_i^* e_i^{*\top} \sim \mathcal{W}_{k_g}[1, V_{0g}/m_g] \tag{3.29}$$

by the definition of a Wishart distribution (Härdle and Simar, 2012, Section 5.2), where k_g is the size of e_i^*. If the subject residuals e_i^* were independent, we would have

$$\hat{V}_{0g} = \sum_{i \in \mathcal{I}(g)} B_i \sim \mathcal{W}_{k_g}[m_g, V_{0g}/m_g], \tag{3.30}$$

by the additive property of Wishart distributions. However, this is not the case due to covariates shared between subjects. To account for this dependence, let us first consider a $n \times p$ design matrix X that is separable into n_X sub-design matrices X_u of size $n_u \times p$ such that, defining A_u as the set of non-zero columns in X_u, the collection of sets $\{A_u : u = 1, \ldots, n_X\}$ is pairwise disjoint. Further, let X be composed of p_B pure between-subject covariates (e.g., group intercept, cross-sectional effect of age) and p_W pure within-subject (e.g., longitudinal effect of visit) as recommended in Section 2.4. In such a situation, the residuals e_i^* can be considered to be in a space of dimension $m_i - p_{Bi}$ where m_i is the number of subjects included in the sub-design matrix containing subject i and p_{Bi} is the number of pure between-subject covariates in this sub-design matrix that are not all-zero. Now, we treat the B_i's as independent random variables following a Wishart distribution $\mathcal{W}_{k_g}[\nu_i, V_{0g}/(m_g\nu_i)]$ with an effective number of degrees of freedom ν_i that is estimated by $1 - p_{Bi}/m_i$. This allows us to consider $\hat{V}_{0g}$ as a sum of independent Wishart distributions. From that, we approximate this sum of independent Wishart distributions by a Wishart distribution $\mathcal{W}_{k_g}[\nu_g, V_{0g}/\nu_g]$. Taking the expectations of this Wishart distribution and of the sum of independent Wishart distributions, we can easily verify that they are both equal (i.e. V_{0g}). Now, to get an estimate of ν_g, we equate the covariance matrices of the vectorised representation of the approximate Wishart distribution with the one of the sum of Wishart distributions and we get

$$\frac{2}{\nu_g} N_{k_g} (V_{0g} \otimes V_{0g}) = \sum_{i \in \mathcal{I}(g)} \frac{2}{\nu_i m_g^2} N_{k_g} (V_{0g} \otimes V_{0g}), \tag{3.31}$$

where we used, in both sides, the formula for the covariances of the vectorised representation of Wishart distribution (Abadir and Magnus, 2005, Exercice 11.23). From

Equation (3.31), we see that the degrees of freedom can therefore be estimated by

$$\nu_g = \frac{m_g^2}{\sum_{i \in \mathcal{I}(g)} \frac{1}{\nu_i}}. \tag{3.32}$$

Now, contrasting the SwE S with C, we have

$$\begin{aligned} CSC^\top &= C \left(\sum_{i=1}^{m} X_i^\top X_i \right)^{-1} \left(\sum_{g=1}^{n_\mathrm{G}} \sum_{i \in \mathcal{I}(g)} X_i^\top \hat{V}_{0g} X_i \right) \left(\sum_{i=1}^{m} X_i^\top X_i \right)^{-1} C^\top \\ &= \sum_{g=1}^{n_\mathrm{G}} \left(C \left(\sum_{i=1}^{m} X_i^\top X_i \right)^{-1} \left(\sum_{i \in \mathcal{I}(g)} X_i^\top \hat{V}_{0g} X_i \right) \left(\sum_{i=1}^{m} X_i^\top X_i \right)^{-1} C^\top \right) \\ &= \sum_{g=1}^{n_\mathrm{G}} (CSC^\top)_g \end{aligned} \tag{3.33}$$

where $(CSC^\top)_g$ is the contribution of group g to the contrasted SwE $CSC^\top$ and which can be rewritten as

$$(CSC^\top)_g = \sum_{i \in \mathcal{I}(g)} L_i \hat{V}_{0g} L_i^\top \tag{3.34}$$

where

$$L_i = C \left(\sum_{j=1}^{m} X_j^\top X_j \right)^{-1} X_i^\top. \tag{3.35}$$

Then, for all $i \in \mathcal{I}(g)$, we get

$$L_i \hat{V}_{0g} L_i^\top \sim \mathcal{W}_q[\nu_g, L_i V_{0g} L_i^\top / \nu_g], \tag{3.36}$$

where q is the rank of C (Härdle and Simar, 2012, Theorem 5.5). As each component $L_i \hat{V}_{0g} L_i^\top$ is obtained with the same estimate $\hat{V}_{0g}$, there is no contribution of additional degrees of freedom and, thus, we assume that

$$(CSC^\top)_g \sim \mathcal{W}_q[\nu_g, \sum_{i \in \mathcal{I}(g)} L_i V_{0g} L_i^\top / \nu_g]. \tag{3.37}$$

Assuming that the contributions $(CSC^\top)_g$ are independent and assuming that $CSC^\top$ follows a Wishart distribution $\mathcal{W}_q[\nu, C\mathrm{Cov}[\hat{\beta}]C^\top/\nu]$, we use an approximation originally proposed in Nel and Van der Merwe (1986) which consists of approximating a

sum of independent Wishart distribution by a Wishart distribution. This approximation consists of equating the trace of the expectation and the covariance matrix of the vectorised representation of the sum of Wishart distributions with those of the approximate Wishart distribution. In our case, noting that $\sum_{g=1}^{n_G}\sum_{i\in\mathcal{I}(g)} L_i V_{0g} L_i^\top = C\text{Cov}[\hat{\beta}]C^\top$, we directly see that the expectations are equal. Now, equating the trace of the covariance matrices, we get

$$\frac{2}{\nu}\text{tr}\left[N_q(C\text{Cov}[\hat{\beta}]C^\top)\otimes(C\text{Cov}[\hat{\beta}]C^\top)\right] = \sum_{g=1}^{n_G}\frac{2}{\nu_g}\text{tr}\left[N_q(C\text{Cov}[\hat{\beta}]C^\top)_g\otimes(C\text{Cov}[\hat{\beta}]C^\top)_g\right] \tag{3.38}$$

where $(C\text{Cov}[\hat{\beta}]C^\top)_g = \sum_{i\in\mathcal{I}(g)} L_i V_{0g} L_i^\top$.

Noting that, for any $q\times q$ matrix A, we have, using the relationship existing between the symmetrizer matrix N_q, the duplication matrix D_n and its pseudo inverse D_n^+ (Abadir and Magnus, 2005, Exercice 11.28), the cyclical property of the trace operator (Abadir and Magnus, 2005, Exercice 2.26) and the property of the matrix $D_n^+(A\otimes A)D_n$ (Abadir and Magnus, 2005, Exercice 11.33),

$$\begin{aligned}\text{tr}[N_q(A\otimes A)] =& \text{tr}[D_n D_n^+(A\otimes A)]\\ =& \text{tr}[D_n^+(A\otimes A)D_n]\\ =& \frac{1}{2}\text{tr}[A^2]+\frac{1}{2}(\text{tr}[A])^2,\end{aligned} \tag{3.39}$$

we can simplify and rearrange Equation (3.38) to get

$$\nu = \frac{\text{tr}[(C\text{Cov}[\hat{\beta}]C^\top)^2]+(\text{tr}[C\text{Cov}[\hat{\beta}]C^\top])^2}{\sum_{g=1}^{n_G}\frac{\text{tr}[(C\text{Cov}[\hat{\beta}]C^\top)_g^2]+(\text{tr}[(C\text{Cov}[\hat{\beta}]C^\top)_g])^2}{\nu_g}}. \tag{3.40}$$

In practice, $\text{Cov}(\hat{\beta})$ and V_{0g}'s are unknown, thus, their estimates S and $\hat{V}_{0g}$'s are used instead in (3.40) and we get

$$\nu = \frac{\text{tr}[(CSC^\top)^2]+(\text{tr}[CSC^\top])^2}{\sum_{g=1}^{n_G}\frac{\text{tr}[(CSC^\top)_g^2]+(\text{tr}[(CSC^\top)_g])^2}{\nu_g}}. \tag{3.41}$$

This number of degrees of freedom can then be used in Equation (3.24) to make inferences. While the test developed here, that we will refer to as Test I, is very similar to the Pan test (as they both used Equation (3.24)), the number of degrees of freedom ν obtained can be quite different and it relies on different assumptions,

making difficult a direct theoretical comparison between the two tests. However, some interesting remarks can be made. First, a relatively strong assumption of Test I is that it assumes no missing data. While this is not directly the case for the Pan test, the presence of missing data is likely to affect the mean and covariances of the Q_i's (see Equation (3.27)) and may break the (unstated) assumptions behind the Pan test. Therefore, the presence of missing data could be an issue for both tests. A second important remark is that, in Test I, the Q_i's are allowed to have different means and covariances, indicating that it could be more robust than the Pan test, at least in some designs. Third, while the Pan test does not account for the dependence existing between the residuals, in our test, this dependence has been partially taken into account by using an effective number of degrees of freedom (i.e. $1 - p_{Bi}/m_i$) for each B_i. Finally, while this type of practice seems to be common in the literature, as we will show it later (see Equation (3.48)), replacing the unknown $\text{Cov}[\hat{\beta}]$ and V_{0g}'s by their estimates in Equation (3.40) can be problematic as this typically leads to biased estimates of $\text{Cov}[\text{vec}[CSC^\top]]$. The latter issue and the fact that Test I assumes no missing data have led us to propose two alternative tests described next.

Test II

As mentioned above, Test I assumes no missing data and does not account for the potential bias introduced by replacing the unknown $\text{Cov}[\hat{\beta}]$ and each V_{0g} by their estimates. In this second test, we attempt to solve these two issues by defining an empirical estimator of $\text{Cov}[\text{vec}[CSC^\top]]$ which accounts for missing data and is unbiased. First, let us decompose $\text{Cov}[\text{vec}[CSC^\top]]$ in terms of each $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$. Using the linearity property of the vec operator (Abadir and Magnus, 2005, Exercice 10.16) and the relationship existing between the vec operator and the Kronecker product (Abadir and Magnus, 2005, Exercice 10.18), we have

$$\begin{aligned}\text{Cov}[\text{vec}[CSC^\top]] =& \text{Cov}\left[\sum_{g=1}^{n_\text{G}} \sum_{i \in \mathcal{I}(g)} \text{vec}[L_i \hat{V}_i L_i^\top]\right] \\ =& \text{Cov}\left[\sum_{g=1}^{n_\text{G}} \sum_{i \in \mathcal{I}(g)} (L_i \otimes L_i)\text{vec}[\hat{V}_i]\right], \end{aligned} \tag{3.42}$$

where L_i is given by Equation (3.35). As each $\hat{V}_i$ in group g is a sub-matrix of $\hat{V}_{0g}$, we can express the internal summation in Equation (3.42) as a function of $\text{vec}[\hat{V}_{0g}]$ such

that

$$\text{Cov}[\text{vec}[CSC^\top]] = \text{Cov}\left[\sum_{g=1}^{n_\text{G}} G_g \text{vec}[\hat{V}_{0g}]\right], \tag{3.43}$$

where G_g is a matrix easily constructed from the matrices $(L_i \otimes L_i)$ related to the subjects in group g. Now, assuming that $\hat{V}_{01}, \hat{V}_{02}, \ldots, \hat{V}_{0n_\text{G}}$ are mutually independent, we have

$$\text{Cov}[\text{vec}[CSC^\top]] = \sum_{g=1}^{n_\text{G}} G_g \text{Cov}[\text{vec}[\hat{V}_{0g}]] G_g^\top. \tag{3.44}$$

Note that the assumption of independence is slightly violated due to the covariates shared between subjects. Nevertheless, we will take this into account later by making an effective degrees of freedom correction.

From Equation (3.44), we see that, in order to estimate the covariance matrix of $\text{vec}[CSC^\top]$, we simply need to estimate the covariance matrix of each $\hat{V}_{0g}$. Unfortunately, getting an estimate of each $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ can be difficult under missing data and when the estimation of the off-diagonal elements of each V_{0g} is based on Equations (3.17) and (3.18), which are highly non-linear. Nevertheless, for the purpose of getting an estimate of $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$, it seems reasonable to, for now, assume that the estimation of the off-diagonal elements of V_{0g} is based instead on Equation (3.19) in which we use adjusted residuals such that

$$(\hat{V}_{0g})_{kk'} = \frac{1}{m_{gkk'}} \sum_{i \in \mathcal{I}(g,k,k')} e_{ik}^* e_{ik'}^*. \tag{3.45}$$

We now account for the dependence existing between the product $e_{ik}^* e_{ik'}^*$ in the same way as in Test I, that is, we treat the m outer products $e_i^* e_i^{*\top}$ as independent random variables, each following a Wishart distribution $\mathcal{W}_{n_i}[\nu_i, V_i/\nu_i]$ with an effective number of degrees of freedom ν_i that is estimated by $1 - p_{Bi}/m_i$. Using Equation (3.45) and using the property of Wishart distributions, the elements of $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ are then given by

$$\begin{aligned}
\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] =& \frac{1}{m_{gkk'} m_{gll'}} \sum_{i \in \mathcal{I}(g,k,k')} \sum_{j \in \mathcal{I}(g,l,l')} \text{cov}[e_{ik}^* e_{ik'}^*, e_{jl}^* e_{jl'}^*] \\
=& \frac{1}{m_{gkk'} m_{gll'}} \sum_{i \in \mathcal{I}(g,k,k') \cap \mathcal{I}(g,l,l')} \text{cov}[e_{ik}^* e_{ik'}^*, e_{il}^* e_{il'}^*]
\end{aligned}$$

$$= \frac{(V_{0g})_{kl}(V_{0g})_{k'l'} + (V_{0g})_{kl'}(V_{0g})_{k'l}}{m_{gkk'}m_{gll'}} \sum_{i \in \mathcal{I}(g,k,k') \cap \mathcal{I}(g,l,l')} \frac{1}{\nu_i}. \tag{3.46}$$

Note that, if there is no missing data, we would simply get

$$\text{Cov}[\text{vec}[\hat{V}_{0g}]] = \frac{2N_g(V_{0g} \otimes V_{0g})}{m_g^2} \sum_{i \in \mathcal{I}(g)} \frac{1}{\nu_i}. \tag{3.47}$$

Again, as, in Equation (3.46), each V_{0g} is unknown, we could replace them by each of their estimators $\hat{V}_{0g}$. Unfortunately, the resulting estimator $\widehat{\text{cov}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}]$ would be biased. Indeed, taking the expectation of the resulting estimator and using the fact that $\mathbb{E}[XY] = \text{cov}[X, Y] + \mathbb{E}[X]\mathbb{E}[Y]$, we get

$$\begin{aligned}
\mathbb{E}[\widehat{\text{cov}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}]] =& a_{gkk'll'} \left(\mathbb{E}[(\hat{V}_{0g})_{kl}(\hat{V}_{0g})_{k'l'}] + \mathbb{E}[(\hat{V}_{0g})_{kl'}(\hat{V}_{0g})_{k'l}] \right) \\
=& a_{gkk'll'} \Big(\text{cov}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'}] + (V_{0g})_{kl}(V_{0g})_{k'l'} \\
& + \text{cov}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l}] + (V_{0g})_{kl'}(V_{0g})_{k'l} \Big) \\
=& \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] + a_{gkk'll'} \Big(\text{cov}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'}] \\
& + \text{cov}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l}] \Big),
\end{aligned} \tag{3.48}$$

where

$$a_{gkk'll'} = \frac{\displaystyle\sum_{i \in \mathcal{I}(g,k,k') \cap \mathcal{I}(g,l,l')} \frac{1}{\nu_i}}{m_{gkk'}m_{gll'}}. \tag{3.49}$$

We can easily see that the bias is $a_{gkk'll'}(\text{cov}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'}] + \text{cov}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l}])$. Assuming a design without missing data, each covariances and the term $a_{gkk'll'}$ should be of the order m_{0g}^{-1}. As a consequence, the bias will tend to become more and more negligible when the number of subjects in each group g increases. From that, we can predict that the bias will be higher in the case of the heterogeneous SwE versions which have groups of one subject compared to the homogeneous SwE versions which typically have more than one subject per group. This indicates that Test I, which does not unfortunately account for the bias, should perform better for the homogeneous SwE than for the heterogeneous SwE.

Hopefully, it seems possible to correct for this bias. Indeed, let us consider the expectation of $\widehat{\text{cov}}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'})]$ and $\widehat{\text{cov}}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l})]$. They actually depend on the same three covariances as $\widehat{\text{cov}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'})]$. Therefore, we can express the

expectation of the three biased estimators as

$$\begin{pmatrix} \mathbb{E}[\widehat{\text{cov}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}]] \\ \mathbb{E}[\widehat{\text{cov}}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'}]] \\ \mathbb{E}[\widehat{\text{cov}}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l}]] \end{pmatrix} = \begin{pmatrix} 1 & a_{gkk'll'} & a_{gkk'll'} \\ a_{gklk'l'} & 1 & a_{gklk'l'} \\ a_{gkl'k'l} & a_{gkl'k'l} & 1 \end{pmatrix} \begin{pmatrix} \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'})] \\ \text{cov}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'})] \\ \text{cov}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l})] \end{pmatrix}. \tag{3.50}$$

Solving the system of equations, we obtain that

$$\begin{aligned} \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'})] =& \frac{1}{b_{gkk'll'}} \Big((1 - a_{gklk'l'}a_{gkl'k'l}) \mathbb{E}[\widehat{\text{cov}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}]] \\ &+ (a_{gkl'k'l} - 1) a_{gkk'll'} \mathbb{E}[\widehat{\text{cov}}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'}]] \\ &+ (a_{gklk'l'} - 1) a_{gkk'll'} \mathbb{E}[\widehat{\text{cov}}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l}]] \Big), \end{aligned} \tag{3.51}$$

where

$$b_{gkk'll} = 1 + 2a_{gkk'll'}a_{gklk'l'}a_{gkl'k'l} - a_{gkk'll'}a_{gklk'l'} - a_{gkk'll'}a_{gkl'k'l} - a_{gklk'l'}a_{gkl'k'l}. \tag{3.52}$$

This means that an unbiased estimator of $\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}]$ can be obtained by

$$\begin{aligned} \widehat{\text{cov}}_{\text{II}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] =& \frac{1}{b_{gkk'll'}} \Big((1 - a_{gklk'l'}a_{gkl'k'l}) \widehat{\text{cov}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] \\ &+ (a_{gkl'k'l} - 1) a_{gkk'll'} \widehat{\text{cov}}[(\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'}] \\ &+ (a_{gklk'l'} - 1) a_{gkk'll'} \widehat{\text{cov}}[(\hat{V}_{0g})_{kl'}, (\hat{V}_{0g})_{k'l}] \Big) \end{aligned} \tag{3.53}$$

or, replacing the biased estimators using Equation (3.46),

$$\begin{aligned} \widehat{\text{cov}}_{\text{II}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'})) =& \frac{a_{gkk'll'}}{b_{gkk'll'}} \Big((2a_{gklk'l'}a_{gkl'k'l} - a_{gklk'l'} - a_{gkl'k'l}) (\hat{V}_{0g})_{kk'} (\hat{V}_{0g})_{ll'} \\ &+ (1 - a_{gkl'k'l}) (\hat{V}_{0g})_{kl}, (\hat{V}_{0g})_{k'l'} \\ &+ (1 - a_{gklk'l'}) (\hat{V}_{0g})_{kl'} (\hat{V}_{0g})_{k'l} \Big). \end{aligned} \tag{3.54}$$

Equation (3.54) can then be used to construct an unbiased estimator $\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$ which can, in turn, be used to estimate $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ in Equation (3.44) and leads to an unbiased empirical estimator for $\text{Cov}[\text{vec}[CSC^\top]]$ given by

$$\widehat{\text{Cov}}_{\text{II}}[\text{vec}[CSC^\top]] = \sum_{g=1}^{n_{\text{G}}} G_g \widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]] G_g'. \tag{3.55}$$

Now, like the Pan test and Test I, we assume that $CSC^\top$ follows a Wishart distribution $\mathcal{W}_q[\nu, (C\text{Cov}[\hat{\beta}]C^\top)/\nu]$ which then leads to the same test statistic as given by Equation (3.24). We could therefore use Equation (3.25) for the theoretical expression of $\text{Cov}[\text{vec}[CSC^\top]]$ and, like the two other tests, replace $C\text{Cov}[\hat{\beta}]C^\top$ by its estimator $CSC^\top$. However, as for the empirical estimator, we would get a biased theoretical estimator. Therefore, instead, using similar considerations than the ones employed for $\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$, we can derive an unbiased theoretical estimator given by

$$\begin{aligned}\widehat{\text{Cov}}_{\text{II}}[(CSC^\top)_{kk'}, (CSC^\top)_{ll'})] =& \frac{\nu}{(1-\nu)(2+\nu)}\left(\frac{2}{\nu}(CSC^\top)_{kk'}(CSC^\top)_{ll'}\right. \\ &-(CSC^\top)_{kl}(CSC^\top)_{k'l'} \\ &\left.-(CSC^\top)_{kl'}(CSC^\top)_{k'l}\right). \end{aligned} \tag{3.56}$$

Now, to get an estimate of ν, we could equate the trace of both the theoretical and the empirical estimators like for Test I and solve the system of equations. However, if we equate instead the sum of the elements of both estimators, we actually get a simpler formula. Indeed, in this case, we get, for the theoretical estimator,

$$\sum_{k=1}^{q}\sum_{k'=1}^{q}\sum_{l=1}^{q}\sum_{l'=1}^{q}\widehat{\text{Cov}}_{\text{II}}[(CSC^\top)_{kk'}, (CSC^\top)_{ll'}] = \frac{2}{(2+\nu)}\left(\sum_{k=1}^{q}\sum_{k'=1}^{q}(CSC^\top)_{kk'}\right)^2, \tag{3.57}$$

and, after equating the sum of the elements of both estimators,

$$\nu = \frac{2\left(\sum_{k=1}^{q}\sum_{k'=1}^{q}(CSC^\top)_{kk'}\right)^2}{\sum_{i=1}^{q^2}\sum_{j=1}^{q^2}\left(\sum_{g=1}^{n_G} G_g\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]G_g'\right)_{ij}} - 2. \tag{3.58}$$

This new estimator for the degrees of freedom should be more accurate than the one proposed in Test I as it accounts for missing data and corrects for small sample biases. However, in this test, to simplify an already complicated problem, we have made use of Equation (3.45) instead of Equations (3.17) and (3.18) in which we would have used adjusted residuals such that

$$(\hat{V}_{0g})_{kk'} = (\hat{V}_{0gkk'})_{kk'}\sqrt{\frac{(\hat{V}_0)_{kk}(\hat{V}_0)_{k'k'}}{(\hat{V}_{0gkk'})_{kk}(\hat{V}_{0gkk'})_{k'k'}}}, \tag{3.59}$$

where , $(\hat{V}_{0g})_{kk} = \sum_{i\in\mathcal{I}(g,k,k)} e_{ik}^{*2}$, $(\hat{V}_{0g})_{k'k'} = \sum_{i\in\mathcal{I}(g,k',k')} e_{ik'}^{*2}$, $(\hat{V}_{0gkk'})_{kk} = \sum_{i\in\mathcal{I}(g,k,k')} e_{ik}^{*2}$,

$(\hat{V}_{0gkk'})_{k'k'} = \sum_{i\in\mathcal{I}(g,k,k')} e^{*2}_{ik'}$ and $(\hat{V}_{0gkk'})_{kk'} = \sum_{i\in\mathcal{I}(g,k,k')} e^{*}_{ik}e^{*}_{ik'}$. Unfortunately, as we will see next, if Equation (3.59) is used, another type of bias may arise and compromise the accuracy of the test.

Test III

As said previously, in practice, we recommend the use of Equation (3.59) instead of Equation (3.45). Thus, it seems important to attempt to develop another test assuming that Equation (3.59) was indeed used.

As Equation (3.59) is non-linear, it is difficult to derive the covariances involving $(\hat{V}_{0g})_{kk'}$. Therefore, we linearise it using a first order Taylor series around the true covariance matrix V_0 and get

$$(\hat{V}_{0g})_{kk'} \approx (\hat{V}_{0gkk'})_{kk'} + \frac{(V_{0g})_{kk'}}{2(V_{0g})_{kk}}((\hat{V}_{0g})_{kk} - (\hat{V}_{0gkk'})_{kk}) + \frac{(V_{0g})_{kk'}}{2(V_{0g})_{k'k'}}((\hat{V}_{0g})_{k'k'} - (\hat{V}_{0gkk'})_{kk}). \tag{3.60}$$

Using Equation (3.60), we can therefore get an approximative expression for the covariances given by

$$\begin{aligned}
\operatorname{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] \approx\ & \operatorname{cov}[(\hat{V}_{0gkk'})_{kk'}, (\hat{V}_{0gll'})_{ll'}] \\
& + \frac{(V_{0g})_{kk'}}{2} \sum_{i\in(k,k')} \frac{1}{(V_{0g})_{ii}} \Big(\operatorname{cov}[(\hat{V}_{0g})_{ii}, (\hat{V}_{0gll'})_{ll'}] \\
& - \operatorname{cov}[(\hat{V}_{0gkk'})_{ii}, (\hat{V}_{0gll'})_{ll'})] \Big) \\
& + \frac{(V_{0g})_{ll'}}{2} \sum_{i\in(l,l')} \frac{1}{(V_{0g})_{ii}} \Big(\operatorname{cov}[(\hat{V}_{0gkk'})_{kk'}, (\hat{V}_{0g})_{ii}] \\
& - \operatorname{cov}[(\hat{V}_{0gkk'})_{kk'}, (\hat{V}_{0gll'})_{ii}] \Big) \\
& + \frac{(V_{0g})_{ll'}(V_{0g})_{kk'}}{4} \sum_{i\in(k,k')} \sum_{j\in(l,l')} \frac{1}{(V_{0g})_{ii}(V_{0g})_{jj}} \\
& \Big(\operatorname{cov}[(\hat{V}_{0g})_{ii}, (\hat{V}_{0g})_{jj}] + \operatorname{cov}[(\hat{V}_{0gkk'})_{ii}, (\hat{V}_{0gll'})_{jj}] \\
& - \operatorname{cov}[(\hat{V}_{0g})_{ii}, (\hat{V}_{0gll'})_{jj}] - \operatorname{cov}[(\hat{V}_{0gkk'})_{ii}, (\hat{V}_{0g})_{jj}] \Big).
\end{aligned} \tag{3.61}$$

Similarly to Equation (3.46), we get that

$$\operatorname{cov}[(\hat{V}_{0gkk'})_{kk'}, (\hat{V}_{0gll'})_{ll'}] = a_{gkk'll'} \Big((V_{0g})_{kl}(V_{0g})_{k'l'} + (V_{0g})_{kl'}(V_{0g})_{k'l} \Big), \tag{3.62}$$

$$\operatorname{cov}[(\hat{V}_{0gkk'})_{kk'}, (\hat{V}_{0gll'})_{ll}] = 2a_{gkk'll'}(V_{0g})_{kl}(V_{0g})_{k'l}, \tag{3.63}$$

$$\begin{aligned}
\text{cov}[(\hat{V}_{0gkk'})_{kk'}, (\hat{V}_{0g})_{ll}] =& 2a_{gkk'll}(V_{0g})_{kl}(V_{0g})_{k'l}, && (3.64)\\
\text{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{ll}] =& 2a_{gkkll}(V_{0g})^2_{kl}, && (3.65)\\
\text{cov}[(\hat{V}_{0gkk'})_{kk}, (\hat{V}_{0gll'})_{ll}] =& 2a_{gkk'll'}(V_{0g})^2_{kl}, && (3.66)\\
\text{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0gll'})_{ll}] =& 2a_{gkkll'}(V_{0g})^2_{kl}. && (3.67)
\end{aligned}$$

Using these six equations in Equation (3.61), we get

$$\begin{aligned}
\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] \approx\ & a_{gkk'll'}\Big((V_{0g})_{kl}(V_{0g})_{k'l'} + (V_{0g})_{kl'}(V_{0g})_{k'l}\Big)\\
&+ (V_{0g})_{kk'} \sum_{i\in(k,k')} \frac{(V_{0g})_{il}(V_{0g})_{il'}}{(V_{0g})_{ii}}\Big(a_{giill'} - a_{gkk'll'}\Big)\\
&+ (V_{0g})_{ll'} \sum_{i\in(l,l')} \frac{(V_{0g})_{ki}(V_{0g})_{k'i}}{(V_{0g})_{ii}}\Big(a_{gkk'ii} - a_{gkk'll'}\Big)\\
&+ \frac{(V_{0g})_{ll'}(V_{0g})_{kk'}}{2} \sum_{i\in(k,k')}\sum_{j\in(l,l')} \frac{(V_{0g})^2_{ij}}{(V_{0g})_{ii}(V_{0g})_{jj}}\Big(\\
&a_{giijj} + a_{gkk'll'} - a_{giill'} - a_{gkk'jj}\Big). \qquad (3.68)
\end{aligned}$$

Comparing Equation (3.68) with Equation (3.46), we see that three additional terms appear. If there is no missing data, these three terms will vanish. Unfortunately, if it is not the case, they can be rather significant, particularly when the covariances are close to the variances and when the amount of missing data is appreciable. In particular, noting that $a_{gkk'll'}$ is likely to be greater or equal to $a_{gkkll'}$ or $a_{gkk'll}$ (due to a possible smaller number of subjects involved), the two first additional terms are likely to be negative (or equal to 0 if no missing data). For the third additional term, it is however more difficult to predict if it will be positive or negative, but it will vanish if there is no missing data. Finally, it seems important to note that they will not become negligible when the number of subjects increases, unless the degree of missing data is decreased. Therefore, these bias terms do not represent a small sample bias, but a missing data bias that seems important to account for when the amount of missing data is significant. We can therefore consider these three additional terms and replace the true covariances/variances by their estimators in Equation (3.68) and get a new estimator for $\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}]$ given by

$$\begin{aligned}
\widehat{\text{cov}}_{\text{III}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] =\ & a_{gkk'll'}\Big((\hat{V}_{0g})_{kl}(\hat{V}_{0g})_{k'l'} + (\hat{V}_{0g})_{kl'}(\hat{V}_{0g})_{k'l}\Big)\\
&+ (\hat{V}_{0g})_{kk'} \sum_{i\in(k,k')} \frac{(\hat{V}_{0g})_{il}(\hat{V}_{0g})_{il'}}{(\hat{V}_{0g})_{ii}}\Big(a_{giill'} - a_{gkk'll'}\Big)
\end{aligned}$$

$$+ (\hat{V}_{0g})_{ll'} \sum_{i\in(l,l')} \frac{(\hat{V}_{0g})_{ki}(\hat{V}_{0g})_{k'i}}{(\hat{V}_{0g})_{ii}} \Big(a_{gkk'ii} - a_{gkk'll'}\Big)$$
$$+ \frac{(\hat{V}_{0g})_{ll'}(\hat{V}_{0g})_{kk'}}{2} \sum_{i\in(k,k')} \sum_{j\in(l,l')} \frac{(\hat{V}_{0g})_{ij}^2}{(\hat{V}_{0g})_{ii}(\hat{V}_{0g})_{jj}} \Big(a_{giijj} + a_{gkk'll'} - a_{giill'} - a_{gkk'jj} \Big). \quad (3.69)$$

Note that, as mentioned for Test II, the estimator given in Equation (3.69) will be biased in small samples due to the variability of $\hat{V}_{0g}$. We could attempt to correct for it, but, due to the non-linearity of the three additional terms, it seems very challenging to do so and, therefore, here, we do not consider this kind of correction.

The new estimator given in Equation (3.69) can then be used to get an alternative empirical estimate of $\text{cov}[\text{vec}[CSC^\top]]$ which can, in turn, be used to estimate the number of degrees of freedom ν. As no small sample bias correction is considered for the empirical estimator, we propose to use the same procedure as in Test I, i.e. we equate the trace of the empirical estimator with the one of the theoretical estimator (obtained without small sample bias correction) to finally get

$$\nu = \frac{\text{tr}[(CSC^\top)^2] + (\text{tr}[CSC^\top])^2}{\text{tr}\left[\sum_{g=1}^{n_G} G_g \widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]] G_g'\right]}. \quad (3.70)$$

3.2.5 Monte Carlo evaluations

As discussed previously in this chapter, many versions of the SwE method can be used in practice, notably depending on the choices between:

1. the heterogeneous and the homogeneous SwE (S^{Het} or S^{Hom}),
2. the small sample bias adjustments (S_0, S_1, S_2, S_3, S_{C2}, S_{C3}, S_{U1} or S_{U2}),
3. the statistical test (asymptotic χ^2-test, Pan test, Test I, Test II or Test III).

Note that we could have added the possibility of using another type of SwE where the null hypothesis model is considered explicitly (referred to as Restricted SwE in this thesis) or non-parametric statistical tests, but this is separately covered in Chapter 4.

While several authors have evaluated some of these versions of the SwE, they have only focused their attention on a subset of them and in settings which cannot be considered as representative of neuroimaging longitudinal data. For instance, many evaluations were made only on cross-sectional data (MacKinnon and White, 1985;

Chesher and Jewitt, 1987; Lipsitz et al., 1999; Long and Ervin, 2000), some only considered the asymptotic χ^2-test (Pan, 2001), or some, while considering clustered data, have not considered the possibility of missing data and have used within-cluster covariance structures which are not necessary representative of neuroimaging longitudinal studies (Kauermann and Carroll, 2001; Fay and Graubard, 2001; Mancl and DeRouen, 2001; Pan and Wall, 2002; Bell and McCaffrey, 2002). Moreover, some small sample bias adjustments (S_{U1} and S_{U2}) and statistical tests (Test I, Test II and Test III) seem, at our knowledge, to have been introduced for the first time in this thesis. Therefore, from the literature, while we can already conclude that some SwE method versions do not work well (e.g., S_0^{Het} with a χ^2-test), it is harder to draw conclusions about some other versions which use combinations of adjustments which have not been investigated before (e.g., S_3^{Hom} with Test II).

For this thesis, we carried out two large sets of Monte Carlo simulations. The first set focused on the comparison between the different SwE versions presented in this chapter while the second set compared the best SwE method versions versus popular alternative methods used in neuroimaging (i.e. the N-OLS, SS-OLS and LME models). These two sets of simulations are described with more details below.

Simulations I

As a first set of simulations, we considered a selection of balanced and unbalanced designs. We used balanced designs consisting of longitudinal data generated for sample sizes of $m = 12, 25, 50, 100$ or 200 subjects with $3, 5$ or 8 visits for each subject (a total of $5 \times 3 = 15$ distinct sample sizes). The subjects were divided into two groups A and B of equal sizes (except for $m = 25$ where the group A and B had 13 and 12 subjects, respectively) and we considered models consisting of, for each group, an intercept, a linear effect of visit and a quadratic effect of visit using orthogonal polynomials. In addition to these 15 balanced designs, we also considered the unbalanced design corresponding to the real ADNI dataset described in Section 2.6. In order to also assess the methods in an unbalanced design but with a smaller number of subjects, we also considered five subsets of the full ADNI dataset obtained by iteratively removing half of the subjects at random in each group, leading to smaller and smaller sample sizes ($m_{\mathrm{N}} = 229, 114, 57, 29, 14$ and 7; $m_{\mathrm{MCI}} = 400, 200, 100, 50, 25$ and 12; $m_{\mathrm{AD}} = 188, 94, 47, 24, 12$ and 6). For these real unbalanced data designs, we considered models consisting of, for each group, an intercept, the centred mean age per subject $\overline{Age}_i - \overline{Age}$ (referred to as cross-sectional "age" effect), the intra-subject centred age $Age - \overline{Age}_i$ (referred to as longitudinal "visit" effect) and their interaction

(referred to as "acceleration").

For each realised dataset, each observation was first generated independently from a standard Normal distribution $\mathcal{N}[0,1]$. Then, the data for each subject $y_i = (y_{i1}, \ldots, y_{ik}, \ldots, y_{in_i})^\top$ was correlated according to one of six different types of within-subject covariance structure by premultiplying y_i by a square-root factor of the desired covariance matrix. The six covariance structures were generated according to the two following equations:

$$\text{var}(y_{ik}) = \alpha_g(1 + \gamma t_k), \tag{3.71}$$

$$\text{corr}(y_{ik}, y_{ik'}) = \rho(1 - \psi|t_k - t_{k'}|), \tag{3.72}$$

where α_g allowed for different variances in each group, γ allowed the variance to vary with visit, t_k ($t_{k'}$, respectively) was the time of measurement at visit k (visit k'), ρ controlled the constant correlation over time and $\psi > 0$ allowed for a linear decrease of the correlation over time. Table 3.1 summarises the parameter values used for the six covariance structures in the simulations for both the balanced and unbalanced ADNI designs.

		Covariance parameters							
Design	Covariance structure	α_A	α_B	α_N	α_{MCI}	α_{AD}	γ	ρ	ψ
Balanced	CS	1	1	-	-	-	0	0.95	0
	Toeplitz	1	1	-	-	-	0	1	0.1
	Group heterogeneity	1	2	-	-	-	0	0	0
	Visit heterogeneity	1	1	-	-	-	1	0	0
	CS corr. & vis. var. het.	1	1	-	-	-	1	0.95	0
	Toepl. corr. & vis. var. het.	1	1	-	-	-	1	1	0.1
ADNI	CS	-	-	1	1	1	0	0.95	0
	Toeplitz	-	-	1	1	1	0	1	0.2
	Group heterogeneity	-	-	1	2	3	0	0	0
	Visit heterogeneity	-	-	1	1	1	2	0	0
	CS corr. & vis. var. het.	-	-	1	1	1	2	0.95	0
	Toepl. corr. & vis. var. het.	-	-	1	1	1	2	1	0.2

Table 3.1 Covariance parameter values used in Simulations I and II of Chapter 3; γ and ψ are expressed as "per visit" for the balanced designs and "per year" for the ADNI designs.

For null simulations, the data was used immediately after being correlated. For non-null simulations, a signal was added according to the (per-subject centred) effect

of visit. For each scenario, we used 10,000 realisations.

We used custom `R` functions to analyse each realised dataset with the SwE method. In total, 16 versions of the SwE were used: S_0^{Het}, S_1^{Het}, S_2^{Het}, S_3^{Het}, $S_{\text{C2}}^{\text{Het}}$, $S_{\text{C3}}^{\text{Het}}$, two versions of $S_{\text{U2}}^{\text{Het}}$, S_0^{Hom}, S_1^{Hom}, $S_2^{\text{Hom}}$and S_3^{Hom}, $S_{\text{C2}}^{\text{Hom}}$, $S_{\text{C3}}^{\text{Hom}}$ and two versions of $S_{\text{U2}}^{\text{Hom}}$, where the homogeneous groups were defined as groups A and B for the balanced designs and Normal, MCI and AD groups for the unbalanced ADNI designs (see Sections 3.2.2 and 3.2.3 for descriptions about these SwE versions). The first versions of $S_{\text{U2}}^{\text{Het}}$ and $S_{\text{U2}}^{\text{Hom}}$, that we will refer to as $S_{\text{U2-0}}^{\text{Het}}$ and $S_{\text{U2-0}}^{\text{Hom}}$, corresponded to versions where we replaced the sample covariance matrices which were not positive semi-definite by those obtained by zeroing the negative eigenvalues; the second versions of $S_{\text{U2}}^{\text{Het}}$ and $S_{\text{U2}}^{\text{Hom}}$, that we will refer to as $S_{\text{U2-S3}}^{\text{Het}}$ and $S_{\text{U2-S3}}^{\text{Hom}}$, corresponded to versions where we replaced the sample covariance matrices which were not positive semi-definite by those obtained in S_3^{Hom} (see Section 3.2.2 for more details on this). Note that the S_{U1} versions were not investigated due to the sizes of the matrices P and P^{Hom} which can be quite large in the scenarios with large samples (e.g., for the full ADNI dataset, the size of P is 5,492,955 $\times$ 9,279). The design matrices included all the effects described at the beginning of this section. For inference in the balanced designs, we considered 9 contrasts consisting of testing the 6 parameters alone (e.g., linear effect of visits in Group A) and the three differences between groups (e.g., difference of the linear effect of visits in Group B vs. Group A). For inference in the ADNI designs, we considered 24 contrasts consisting of testing the 12 parameters alone and the 12 differences between pair of groups. For each realisation and contrast, we used the asymptotic χ^2-test, the Pan test, Test I, Test II and Test III introduced in Section 3.2.4 considering a 5% level of significance.

To assess the different SwE approaches, we first used, for each contrast, the relative Bias defined as

$$\text{rel. Bias} = \frac{\mathbb{E}[CSC^\top]}{\text{var}(C\hat{\beta})} - 1, \tag{3.73}$$

where $\mathbb{E}[CSC^\top]$ was estimated by the Monte Carlo mean of $CSC^\top$ and $\text{var}(C\hat{\beta})$ by the Monte Carlo variance of $C\hat{\beta}$. Then, for null data, each significant realisation was counted as a False Positive detection and was used to compute the observed False Positive Rates (FPRs). The FPR of a valid test does not exceed the nominal level, while an invalid or liberal test has an FPR in excess of the nominal level. Using a Normal approximation to binomial counts over 10,000 realisations, an exact test (FPR = 5%) should have a FPR between (4.57%, 5.43%) with 95% probability. Finally, non-

null simulations allowed us to estimate power with the True Positive Rates (TPRs).

Simulations II

As a second set of simulations, we compared the best versions of the SwE method isolated from Simulations I to the N-OLS (see Section 2.3.2), SS-OLS (see Section 2.3.3) and LME (see Section 2.3.5) methods. The scenarios considered were the same as the ones used in Simulations I.

The N-OLS model included per-subject dummy variables, and thus precluded the use of the group intercepts and the age effect (as they are a linear combination of the dummy variables). Nevertheless, the N-OLS design matrices included all the visit effects like in the SwE design matrices. The SS-OLS approach used per-subject models, with a design matrix extracted from the appropriate rows and columns of the SwE design matrices, and contrasts that extracted quantities equivalent to the contrasts of interest used with the other models; the final model used with the SS-OLS approach was always a one-measure-per-subject OLS model allowing to test group effects equivalent to the one tested with the other methods. For both the N-OLS and SS-OLS methods, the function `lm` of the `stats R` package was used to estimate the model parameters, their variances/covariances and the degrees of freedom used in the Wald tests (i.e. the number of observations minus the number of parameters present in the considered model).

For the LME method, we considered three different models that we will refer to as LME I, LME II and LME III in this thesis. All three models used the SwE design matrices for the fixed effects, but differed in terms of the random effects. LME I used only a random intercept, LME II used a random intercept and a random linear effect of time, and LME III used a random intercept, a random linear effect of time and a random quadratic effect of time. For LME II and LME III, non-null covariances between random effects were allowed. The functions `lme` from the `R` package `nlme` (Pinheiro et al., 2013) and `lmer` from the `R` package `lme4` (Bates et al., 2012) were used to estimate the LME model parameters, their variances/covariances and the number of degrees of freedom used in the Wald tests. Note that, as the `lme4` package did not propose any estimation for the degrees of freedom, we used the ones estimated by the `nlme` package (Pinheiro and Bates, 2000) for all the `nlme` and `lme4` Wald tests. In addition, for all the LME models fitted with the function `lmer`, we used the `vcovAdj` and `get_ddf_Lb` functions of the `pbkrtest R` package (Halekoh and Højsgaard, 2013) to compute the Kenward-Roger-adjusted covariance matrices and the Kenward-Roger effective degrees of freedom, which were used to conduct Kenward-Roger-adjusted

F-tests (Kenward and Roger, 1997). The results obtained using these two Kenward-Roger adjustments will be referred to as LME-KR I, LME-KR II and LME-KR III to contrast them with those obtained without adjustments that will simply be referred to as LME I, LME II and LME III.

To assess and compare the methods, we used the same metrics of assessment as in Simulations I, i.e. the rel. Bias, the FPR and the TPR (or power).

3.2.6 Real data analysis

The real ADNI dataset described in Section 2.6 was analysed by using the N-OLS, SS-OLS and SwE methods with the same design matrices as used in the simulations (see Section 3.2.5). `SPM8` was used for the N-OLS and SS-OLS methods and a home-made `SPM8` plug-in, which has been made freely available at http://warwick.ac.uk/tenichols/SwE, was used for the SwE method.

In addition, in order to check the validity of the N-OLS approach, we conducted a Box' s test of Compound Symmetry (Box, 1950) with a reduced dataset of 483 subjects who were all scanned at screening and followed up at 6, 12 and 24 months (i.e. no missing data).

3.3 Results

In this section, we summarise the results obtained from the Monte Carlo simulations described in Section 3.2.5 and from the real data analysis described in Section 3.2.6.

3.3.1 Comparison between the SwE versions

Here, we summarise the results of Simulations I, first in terms of relative bias, then in terms of FPR control.

Relative bias

Figure 3.1 shows several boxplots of the relative bias of the 16 SwE versions assessed in Simulations I over several scenarios in the balanced designs. From this figure, we can see that, in the balanced designs, the best SwE versions were $S^{\text{Hom}}_{\text{C2}}$, $S^{\text{Het}}_{\text{C2}}$, $S^{\text{Hom}}_{\text{U2-0}}$ and $S^{\text{Hom}}_{\text{U2-S3}}$. They seemed unbiased, even in the challenging settings with only 12 subjects. Regarding the other versions, while they appeared unbiased in large samples, they tended to be more and more biased when the sample size decreased. In particular, the

versions S_0 tended to underestimate the true $\text{Cov}[\hat{\beta}]$. The S_1 and S_2 versions were less biased than S_0, but still tended to underestimate $\text{Cov}[\hat{\beta}]$. In contrast, the S_3, S_{C3}, $S^{\text{Het}}_{\text{U2-0}}$ and $S^{\text{Het}}_{\text{U2-S3}}$ versions tended to overestimate $\text{Cov}[\hat{\beta}]$. The reason why $S^{\text{Het}}_{\text{U2-0}}$ and $S^{\text{Het}}_{\text{U2-S3}}$ behaved like this in the simulations is due to the fact that they very often yielded covariance matrices which were not positive semi-definite, leading to a replacement of them by those obtained by zeroing the negative eigenvalues or by those obtained in S^{Het}_3, respectively. Those replacements clearly induced the bias observed. On this remark, we see that, it would have been better to substitute the misbehaved covariance matrix estimates by those obtained in $S^{\text{Het}}_{\text{C2}}$ as they seemed unbiased. Finally, note that, in the balanced designs, the heterogeneous and homogeneous SwE versions were equivalent for S_0, S_1, S_2, S_3, S_{C2} and S_{C3}, but not for S_{U2}.

Regarding the much more challenging ADNI designs, as shown in Figure 3.2, while the results seems less accurate than for the balanced designs, the same global trends were observed with the S_{C2} versions performing best and all the versions improving when the sample size increases. Typically, the S_0, S_1 and S_2 versions tended to underestimate the true variances with the S_1 versions performing better than the S_0 versions, but worse than the S_2 versions. The S_3 versions were generally performing better than those three versions, but not as well as the S_{C2} versions which seemed unbiased in almost all the scenarios. The S_{C3} and $S^{\text{Het}}_{\text{U2-0}}$ versions tended to overestimate quite strongly the true variances. The $S^{\text{Hom}}_{\text{U2}}$ seemed to be accurate in some scenarios, but unstable in others. Note that, particularly in small samples, all the homogeneous versions seemed to struggle in the scenarios under CS and when a within-subject effect was tested (see outliers with a positive bias in Figure 3.2). This effect can be observed with more details for $S^{\text{Hom}}_{\text{C2}}$ in Figure 3.3. We clearly see that $S^{\text{Hom}}_{\text{C2}}$ performed quite well in all scenarios, except when a within-subject effect was tested under CS, in which case, $S^{\text{Hom}}_{\text{C2}}$ overestimated the true variances. Note also that this effect tended to vanish when the sample size increased. A possible cause for this misbehaviour resides in the fact that the true covariance matrices V_{0g}'s used in the CS scenarios were close to the boundary of the positive definite-matrix space. Indeed, in this case, it is not surprising that the sample estimates $\hat{V}_{0g}$'s used in the S^{Hom} versions had the tendency to be non-positive semi-definite, particularly in small samples. As, in such cases, we replaced the negative eigenvalues by zero, we probably introduced the bias which is observed in the results. An interesting remark is that we do not observe this for the case with CS correlations and heterogeneous visit variances. This may surprise as the degree of correlation was the same as in the CS case with homogeneous variances (i.e. $\rho = 0.95$). Finally, we can also observe in Figure 3.3 that, in some scenarios with 25 subjects, the

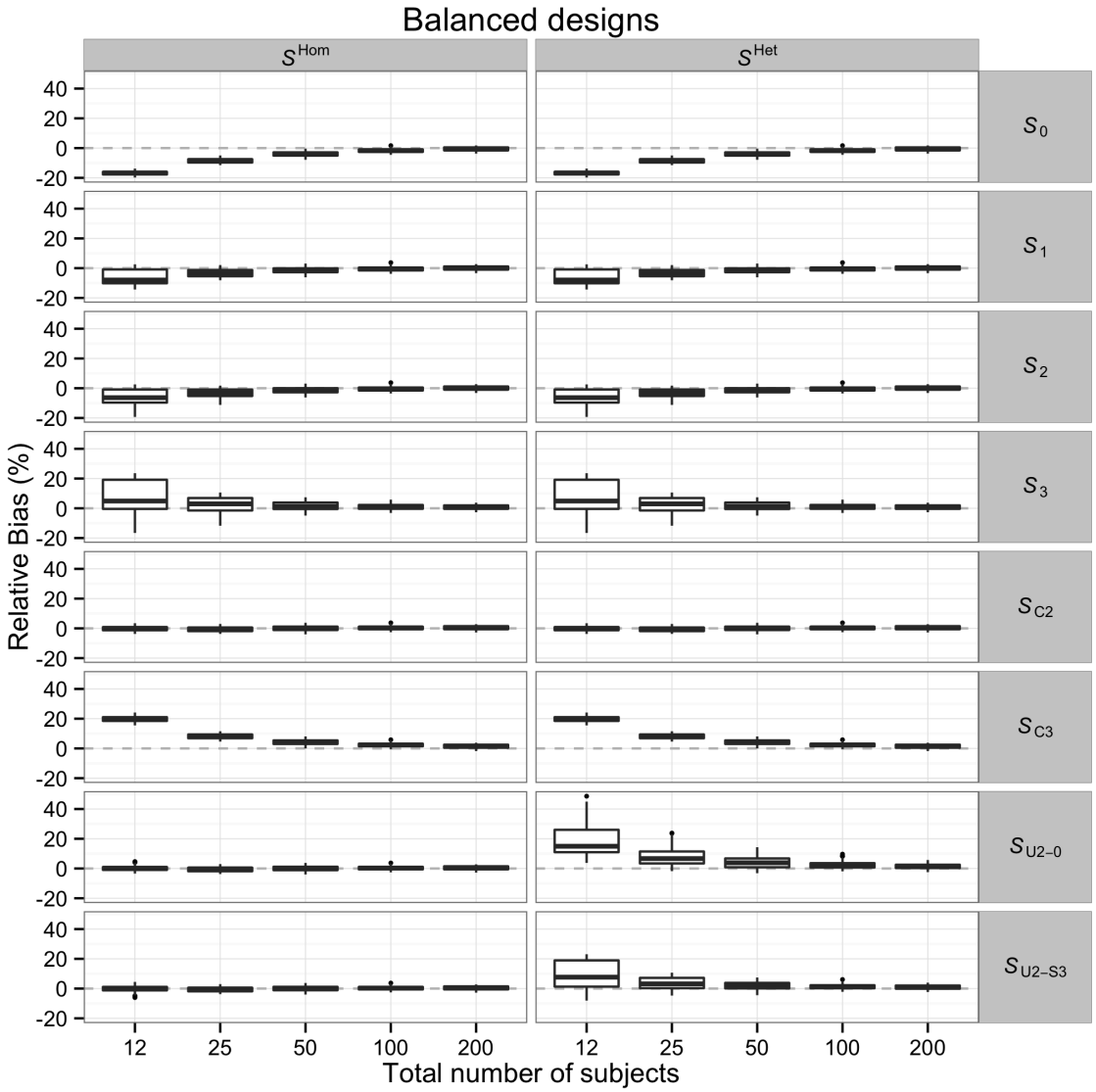

Fig. 3.1 Boxplots showing the Monte Carlo relative bias of 16 SwE versions as a function of the total number of subjects in the balanced designs over 162 scenarios (consisting of the 9 contrasts tested, the 6 within-subject covariance structures and the 3 numbers of visits per subject considered in Simulation I).

S_{C2} versions, particularly S_{C2}^{Het}, tended to slightly underestimate the true variances. Nevertheless, it is worth pointing out that these designs were rather challenging due to the presence of missing data and due to the effective number of subjects involved in each covariance matrix estimation ($m_N = 7$, $m_{MCI} = 12$ and $m_{AD} = 6$).

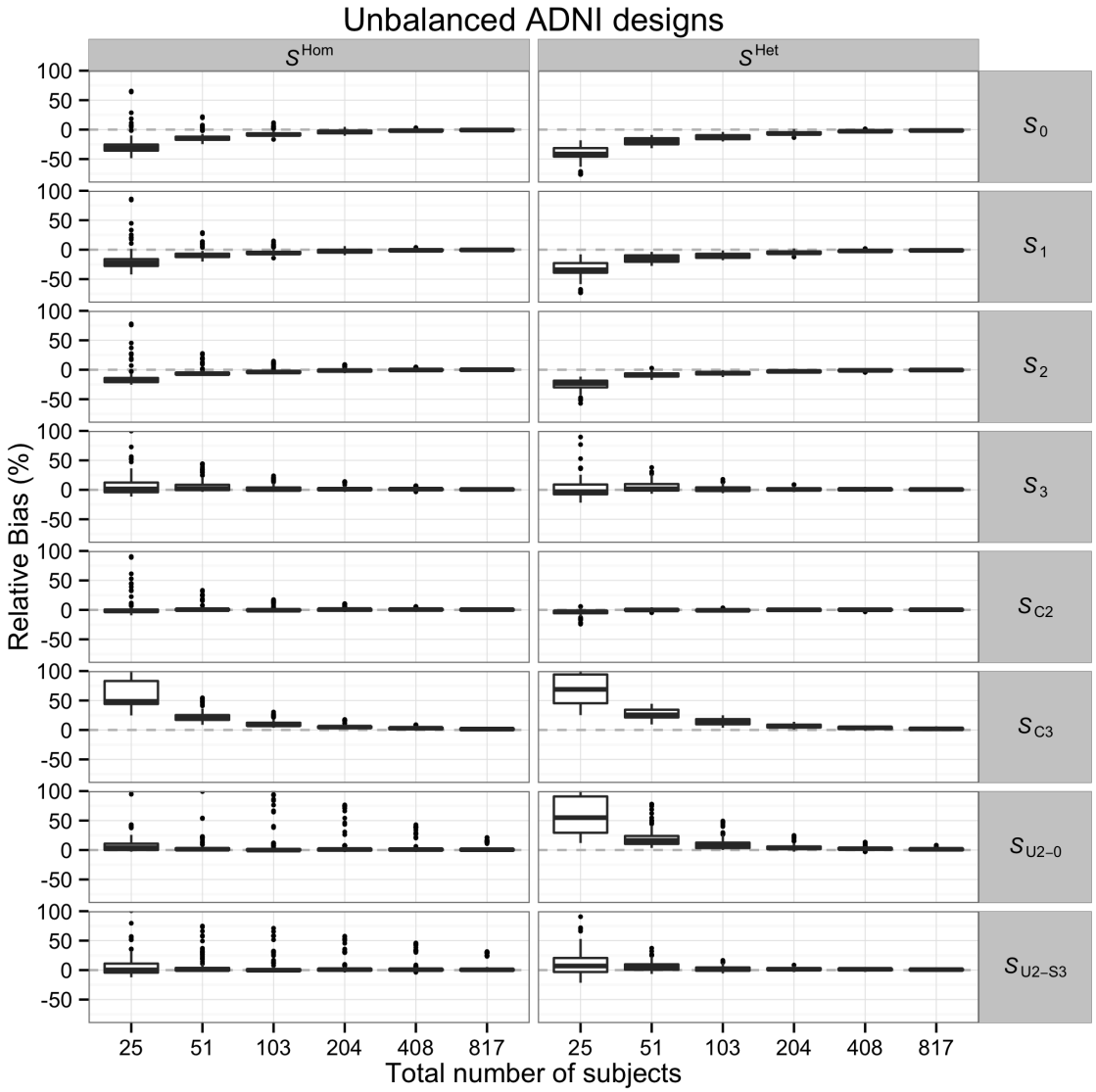

Fig. 3.2 Boxplots showing the Monte Carlo relative bias of 16 SwE versions as a function of the total number of subjects in the unbalanced ADNI designs over 144 scenarios (consisting of the 24 contrasts tested and the 6 within-subject covariance structures considered in Simulations I). For clarity, only the points in the interval $[-90\%, 100\%]$ are shown. This affects only the S_3, S_{C3}, $S^{Het}_{U2\text{-}0}$ & $S_{U2\text{-}S3}$ versions in the designs with a total of 25 subjects and $S^{Hom}_{U2\text{-}0}$ in the designs with a total of 25, 51 & 103 subjects, for which some relative bias were superior to 100%. More detailed results about the S_{C2} versions are given in Figures 3.3.

FPR control

Figures 3.4 shows the results obtained for the five statistical tests when the S_{C2} versions were used in the balanced designs. It appears clearly that the χ^2-test was the worst

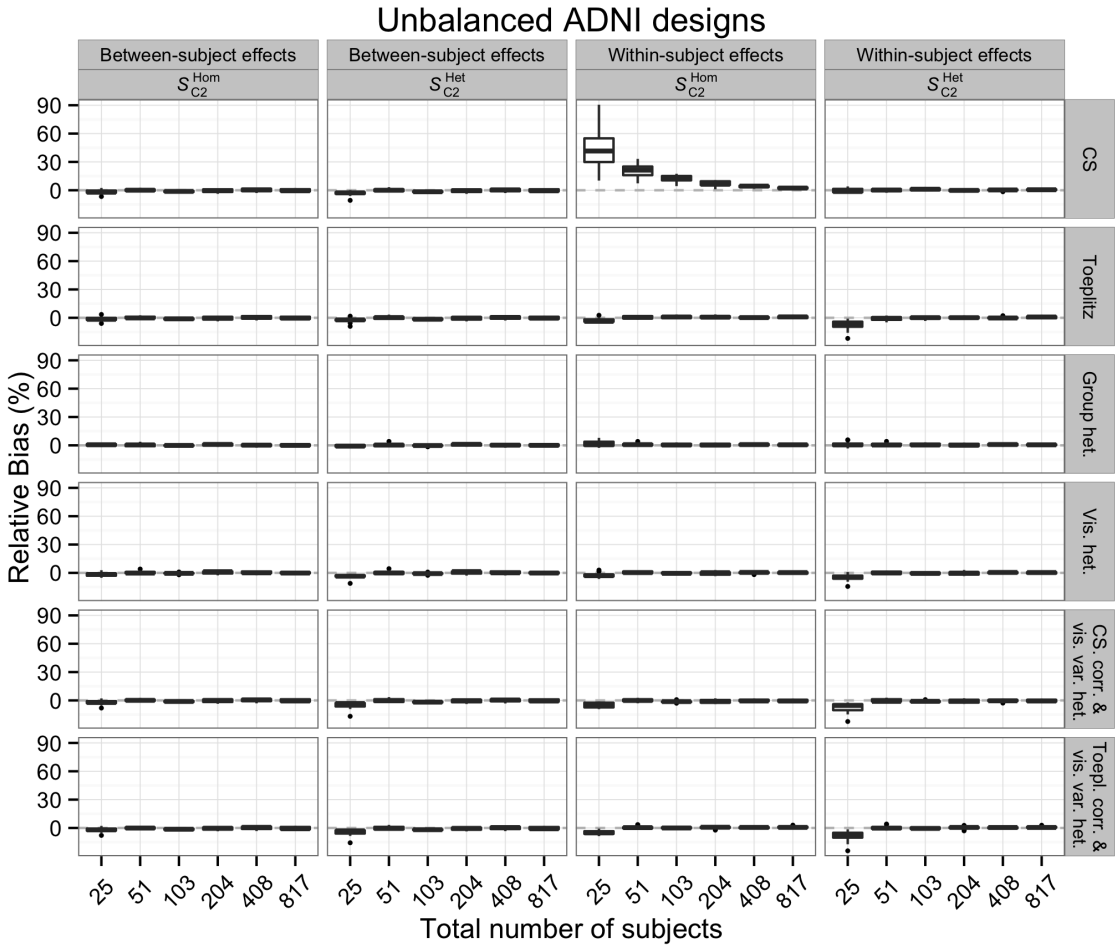

Fig. 3.3 Boxplots showing the Monte Carlo relative bias of the S_{C2} versions as a function of the total number of subjects in the unbalanced ADNI designs over 144 scenarios (consisting of the 24 contrasts and the 6 within-subject covariance structures considered in Simulations I). The results are split in terms of the within-subject covariance structures in the rows, and in terms of the two S_{C2} versions and the type of effects (between-subject or within-subject effects) in the columns. The between-subject effects corresponded to the 12 contrasts involving the intercepts or the cross-sectional effects of age while the within-subject effects corresponded to the 12 contrasts involving the longitudinal effects of age or the acceleration effects.

test, yielding systematically liberal inferences in small samples, but improving when the sample size increased. The best results were obtained for S_{C2}^{Hom} with Test II, for which the inferences were always accurate, even in the designs with a total of 12 subjects. The second best results were obtained for S_{C2}^{Hom} with Test I or Test III (which were equivalent in the balanced designs), for which the inferences were almost

as good as with Test II, but slightly conservative in the designs with 12 subjects. This effect can be explained by the fact that those two tests do not correct for the small sample bias appearing in Equation (3.48) while Test II does. The inferences obtained with $S_{\text{C2}}^{\text{Het}}$ were systematically less good than those obtained with $S_{\text{C2}}^{\text{Hom}}$. In particular, with $S_{\text{C2}}^{\text{Het}}$, the Pan test and Test II appeared slightly liberal in small samples, but seemed to improve quickly when the number of subjects were increased. Comparing these two tests, Test II seemed slightly more accurate than the Pan test, particularly in the designs with 12 subjects. Finally, with $S_{\text{C2}}^{\text{Het}}$, Test I and Test III (which are always equivalent for a heterogeneous SwE version) yielded conservative inferences. Also, while the inferences were improving when the number of subjects were increasing, the improvement seemed slower than the one observed for Test II or the Pan test. This conservativeness can be explained by the small sample bias observed in Equation (3.48). Indeed, as the term $a_{gkk'll'}$ appearing in the bias is approximatively inversely proportional to the number of subjects per homogeneous group, it will always be significant when a heterogeneous SwE is used as there is only one subject per group, even when the sample size increases. In contrast, it will be smaller when a homogeneous SwE is used as there are more subjects per group and will typically decrease when the number of subjects per group increases, explaining the strong differences observed between the results for $S_{\text{C2}}^{\text{Het}}$ and $S_{\text{C2}}^{\text{Hom}}$ with Test I and Test III.

Figure 3.5 shows the results in the unbalanced ADNI designs. Like in the balanced designs, the worst test was clearly the χ^2-test which tended to be liberal, particularly in small samples. The results related to $S_{\text{C2}}^{\text{Het}}$ were similar to those obtained in the balanced designs, i.e. Test I and Test III were conservative while the Pan test and Test II were liberal in small samples. In particular, the Pan test and Test II yielded similar results. Nevertheless, while the differences between these two tests did not seem significant, it is worth noting that, in our simulations, the results obtained with the Pan test were "helped" by the fact that the designs had several groups. Indeed, taking the example in which the contrast tested only involved the MCI subjects, approximatively 50% of the subject contributions to the SwE were zero. This typically tended to inflate the empirical estimates of $\text{Cov}[\text{vec}[CSC^\top]]$ (see Equation (3.26)) proposed in Pan and Wall (2002), making the Pan test typically less liberal than if we had a design with only the MCI subjects included. This kind of artefact were not present in Test II which would give the same results in both cases. This seems to indicate that the difference between the Pan test and Test II may be larger (in favour of Test II) in other designs than those investigated in this thesis. Under the use of $S_{\text{C2}}^{\text{Hom}}$, Test I, Test II and Test

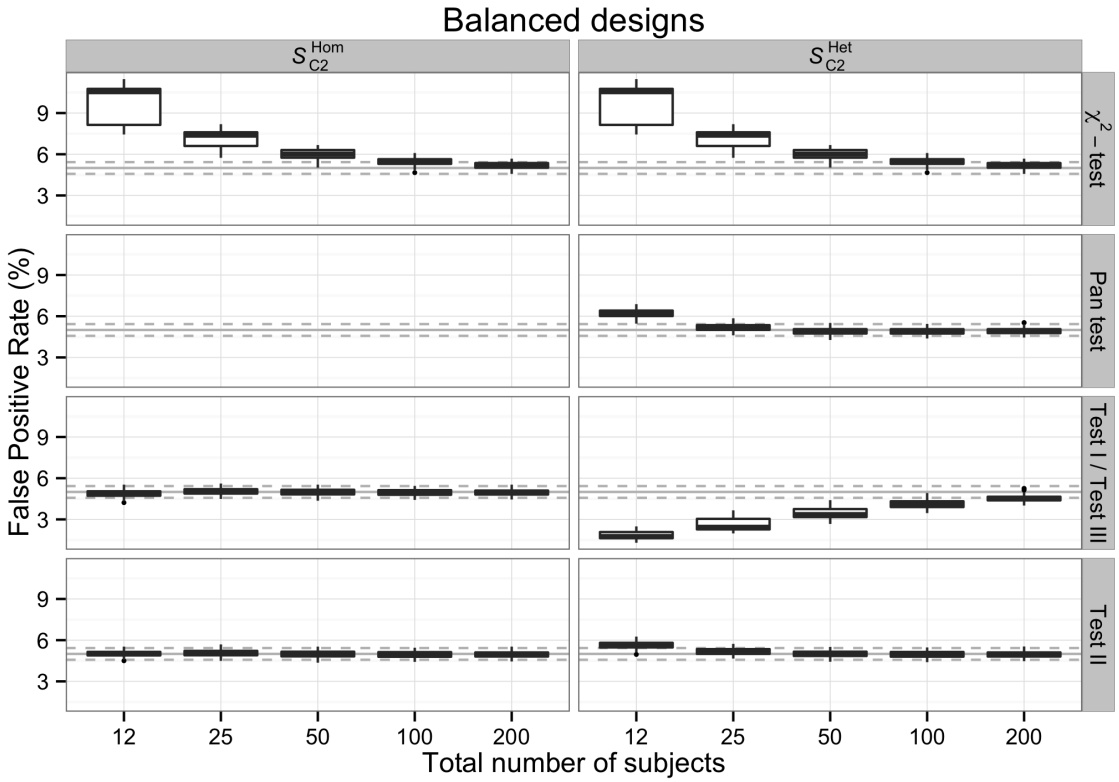

Fig. 3.4 Boxplots showing the Monte Carlo FPR of the two S_{C2} SwE versions as a function of the total number of subjects in the balanced designs over 162 scenarios (consisting of the 9 contrasts tested, the 6 within-subject covariance structures and the 3 numbers of visits per subject considered in Simulation I). The results are split in terms of the statistical tests in the rows, and in terms of the two S_{C2} versions in the columns. Note that Test I and Test III are identical in the balanced designs and the Pan test is invalid with S_{C2}^{Hom}.

III appeared accurate in many scenarios, but struggled in some others. This can be observed in more details in Figure 3.3. All three tests seemed to be relatively accurate when a between-subject effect was tested, struggling slightly in the design with a total of 25 subjects. For the within-subject effects, we see that Test I tended to be liberal, except in the CS designs where it tended to be conservative. The liberality of Test I can simply be explained by the fact that it does not account at all for the presence of missing data. Except in the CS designs, Test II typically performed better than Test I and had the tendency to be conservative in some scenarios. Except in the CS designs, Test III was outperforming the two other tests, being almost accurate in all the scenarios. The only scenarios where it was struggling was in the CS designs, for which it was highly conservative in small samples, but not as much as Test II which

yielded a FPR of 0% in almost all the scenarios, even in large samples. Note that the conservativeness observed for Test III under CS is not necessarily inherent to the test itself, but rather to the bias observed in the SwE S_{C2}^{Hom}.

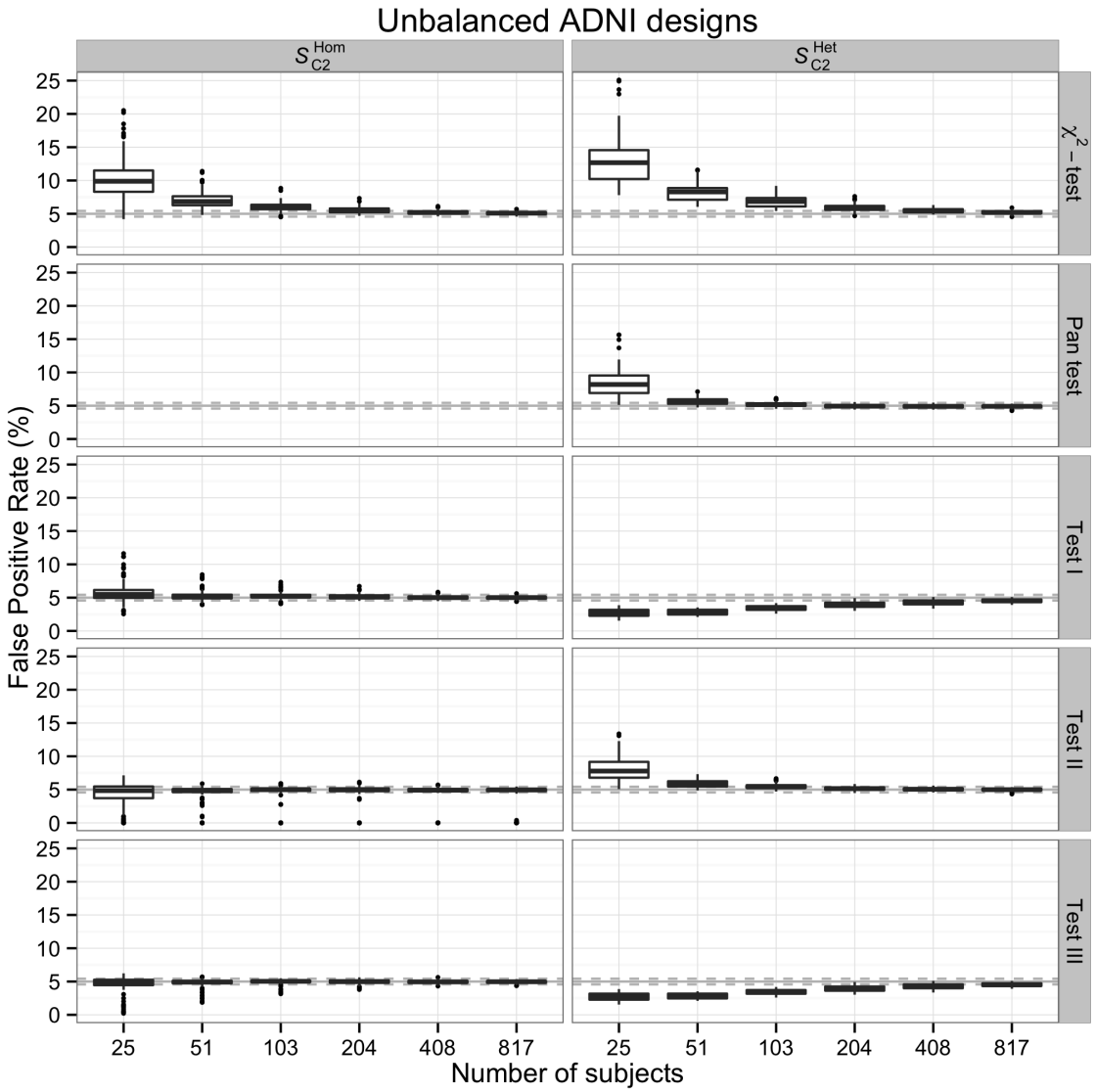

Fig. 3.5 Boxplot showing the Monte Carlo FPR of the S_{C2} SwE versions as a function of the total number of subjects in the unbalanced ADNI designs over 144 scenarios (consisting of the 24 contrasts tested and the 6 within-subject covariance structures considered in Simulations I). The results are split in terms of the statistical tests in the rows, and in terms of the two S_{C2} versions in the columns. Note that Test I and Test III are identical for S_{C2}^{Het} and the Pan test is invalid with S_{C2}^{Hom}.

To summarise the results obtained from Simulations I, it seems that the best SwE

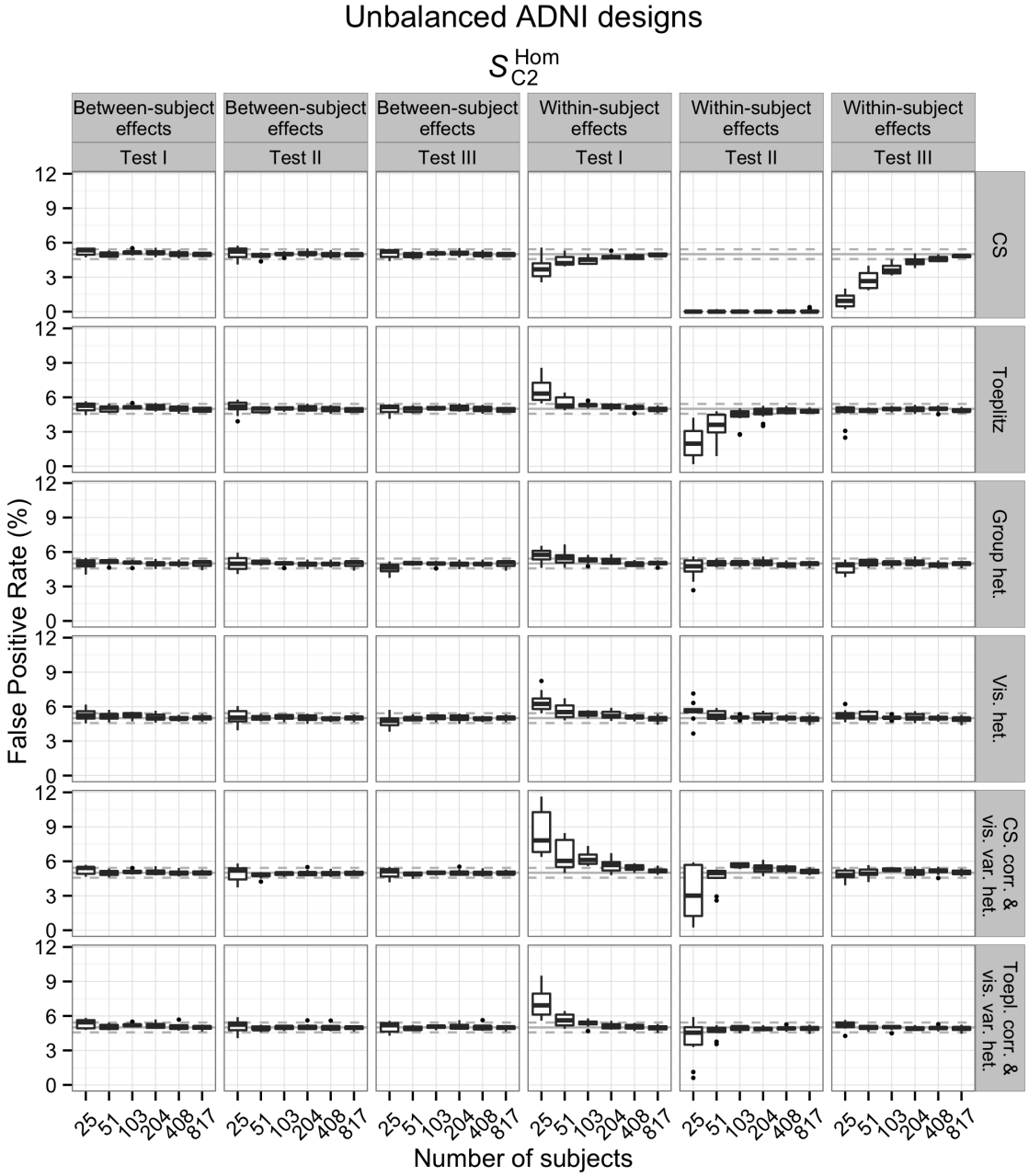

Fig. 3.6 Boxplot showing the Monte Carlo FPR of $S^{\text{Hom}}_{\text{C2}}$ as a function of the total number of subjects in the unbalanced ADNI designs over 144 scenarios (consisting of the 24 contrasts tested and the 6 within-subject covariance structures considered in Simulations I). The results are split in terms of the covariance structures in the rows, and in terms of the statistical tests (Test I, II or III) and the type of effects (between-subject or within-subject effects) in the columns.

versions were S_{C2}^{Het} and S_{C2}^{Hom}. Regarding the statistical tests, Test III seemed overall the best when S_{C2}^{Hom} was used, but not necessary when S_{C2}^{Het} was used as it seemed quite conservative in small samples. The best alternative when S_{C2}^{Het} was used seemed to be Test II, which unfortunately appeared liberal in small samples, but was yielding more quickly accurate inferences than Test III when the sample size increased.

3.3.2 Comparison with alternative methods

In this section, we summarise the results obtained from Simulations II, first in terms of relative bias, then in terms of FPR control and finally in terms of power.

Before presenting the results, it seems important to note that, when we used the function `lme` of the `R` package `nlme` to fit the LME models, convergence failures occurred frequently. In such cases, the function simply returned an error message without any solutions. Therefore, we abandoned the use of the function `lme` and used the function `lmer` of the `R` package `lme4` for all the LME models. In our simulations, this function always returned a solution without error or warning messages. Nevertheless, at the time of our simulations, we used the function `lmer` of the `R` package `lme4` in version 1.0.5 (released on 24/10/2013) which appeared to lack some features to check if the solutions converge. During the writing of this thesis, we have however noticed that more recent versions of the `R` package `lme4` (e.g., version 1.1.7 released on 19/07/2014) have such features. Using the function `lmer` of the `R` package `lme4` in version 1.1.7 in the same settings as in Simulations II, we have found that some solutions did not converge. This seems to indicate that some convergence issues may have occurred when we fitted the LME models during Simulations II. However, as, at that time, we did not have the means to check for this, we are not able here to give any information about this.

Note also that the model LME III could not be fitted in the balanced designs with three visits as there were not enough data per subject to fit a model with three random effects.

Finally, as the function `get_ddf_Lb` which was used to compute the Kenward-Roger degrees of freedom appeared prohibitively slow, particularly in the designs with a large sample size, we only computed the Kenward-Roger degrees of freedom for the balanced designs with 12, 25 and 50 subjects and for the unbalanced ADNI designs with 25, 51 and 103 subjects.

Relative bias

Figures 3.7 shows the relative biases obtained in Simulations II for the balanced designs. The N-OLS method was only accurate under CS or under visit heterogeneity. In the other scenarios, it was either underestimating or overestimating the true variances. The SS-OLS method worked well, except under group variance heterogeneity, where it was either underestimating or overestimating the true variances. Regarding the LME models, LME I seemed accurate only under CS or under visit heterogeneity when the sample size was not too small. LME II seemed to struggle in all designs, even if it seemed to perform better than LME I under some covariance structures. In particular, under the two designs with a Toeplitz correlation structure, it underestimated quite strongly the variances related to the quadratic effect of visits (see outliers in row 4, columns 2 and 6 of Figure 3.7). Under those two scenarios, only LME III seemed actually able to yield accurate estimates of variances. Nevertheless, LME III still failed strongly under group variance heterogeneity and seemed to struggle in all the other scenarios, generally overestimating the true variances. Still regarding the LME models, it seemed that the covariance matrices were unchanged after using the Kenward-Roger covariance correction in the balanced designs, meaning that the results of LME-KR I, II and III were identical to those of LME I, II and III, respectively. Finally, we see that the only method yielding accurate estimates in all the scenarios was the SwE method using either S_{C2}^{Hom} or S_{C2}^{Het}.

Figure 3.8 shows the relative biases obtained in Simulations II for the unbalanced ADNI designs. The N-OLS method was only accurate under CS. In the other scenarios, except under group variance heterogeneity, it seemed to always underestimate the true variances. The SS-OLS method, which worked relatively well in the balanced designs, failed to be accurate under all the covariance structures investigated, either underestimating or overestimating the true variances. Like the N-OLS method, LME I was accurate only under CS. LME II and LME III were clearly better than LME I, except under CS, where they seemed to struggle in small samples. Also, LME II appeared slightly better than LME III which seemed to struggle more in small samples. The Kenward-Roger covariance matrix correction seemed to affect LME II and LME III, particularly in small samples. While the correction did not seem to change significantly the results for LME II, it seemed to affect significantly those of LME III in small samples. The observed changes appeared to inflate the estimates of variances, either making them more accurate (e.g., under Toeplitz covariance structure) or increasing the overestimation (e.g., under CS covariance structure). Finally, the SwE method seemed overall the best method, yielding identical results to those observed in

Simulations I, where S_{C2}^{Hom} or S_{C2}^{Het} were almost always accurate, except in the designs with a total of 25 subjects where they seemed to slightly underestimate the true variances and, under CS, where S_{C2}^{Hom} had the tendency to overestimate the true variances in small samples.

FPR control

Figure 3.9, 3.10, 3.11 and 3.12 summarise the results obtained in Simulations II in terms of FPR. Broadly speaking, the results are consistent with the results obtained in terms of relative bias. Typically, in a scenario where a method underestimated the true variance, the inference tended to be liberal. Conversely, when there were an overestimation, the inference tended to be conservative. The N-OLS, LME I and LME-KR I (which seemed identical to LME I) methods were typically accurate only under CS. The SS-OLS method seemed to be valid only under a balanced design without group variance heterogeneity. In the other scenarios, it seemed to yield either conservative or liberal inferences. LME II and LME III gave relatively poor inferences in all the scenarios. Nevertheless, when the Kenward-Roger corrections were used (see LME-KR II and LM-KR III in the figures), the inferences improved in some scenarios. For example, in the designs with Toeplitz correlations (see columns 2 and 6 in the figures), LME-KR II and LM-KR III seemed more accurate than LME II or LME III. Nevertheless, in other scenarios like those under CS or visit variance heterogeneity, the inferences seemed better without the Kenward-Roger corrections. Finally, the SwE S_{C2}^{Hom} combined with Test III seemed overall the most accurate method, struggling almost only in the unbalanced designs under CS.

Power

An analysis of power is only valid for methods which are able to validly control the FPR, i.e. only when the control is accurate or conservative (FPR $\leq 5\%$). Unfortunately, as shown previously (see Figures 3.9, 3.10, 3.11 and 3.12), in our simulations, this happened only in a few scenarios which were generally different across methods, making very difficult a fair comparison between methods. As it seemed that the scenarios under CS were the ones where the largest number of methods were valid, we decided to make the power comparison only under this covariance structure. Nevertheless, it is important to note that this choice is quite unfair for some of the methods which were more conservative than the other methods in these scenarios, but more accurate in some others. In particular, it was almost the only scenarios where the SwE method with S_{C2}^{Hom} under Test III was not accurate, but conservative.

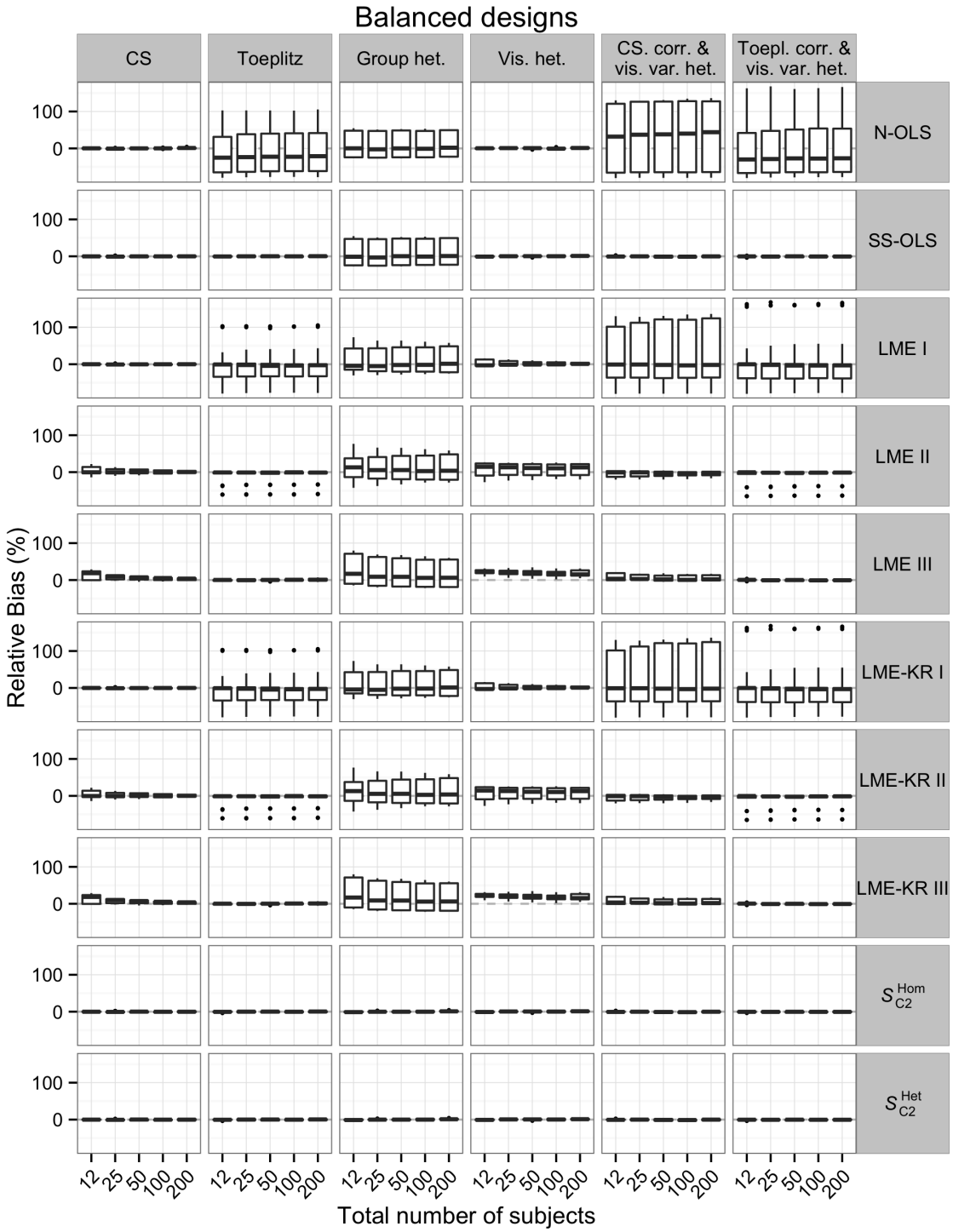

Fig. 3.7 Boxplots showing the Monte Carlo relative bias of several methods as a function of the total number of subjects in the balanced designs over 162 scenarios (consisting of the 9 contrasts tested, the 6 within-subject covariance structures and the 3 numbers of visits per subject considered in Simulation II). Note that no results were obtained for LME III and LME-KR III with the designs consisting of 3 visits per subject as models with 3 random effects cannot be fitted.

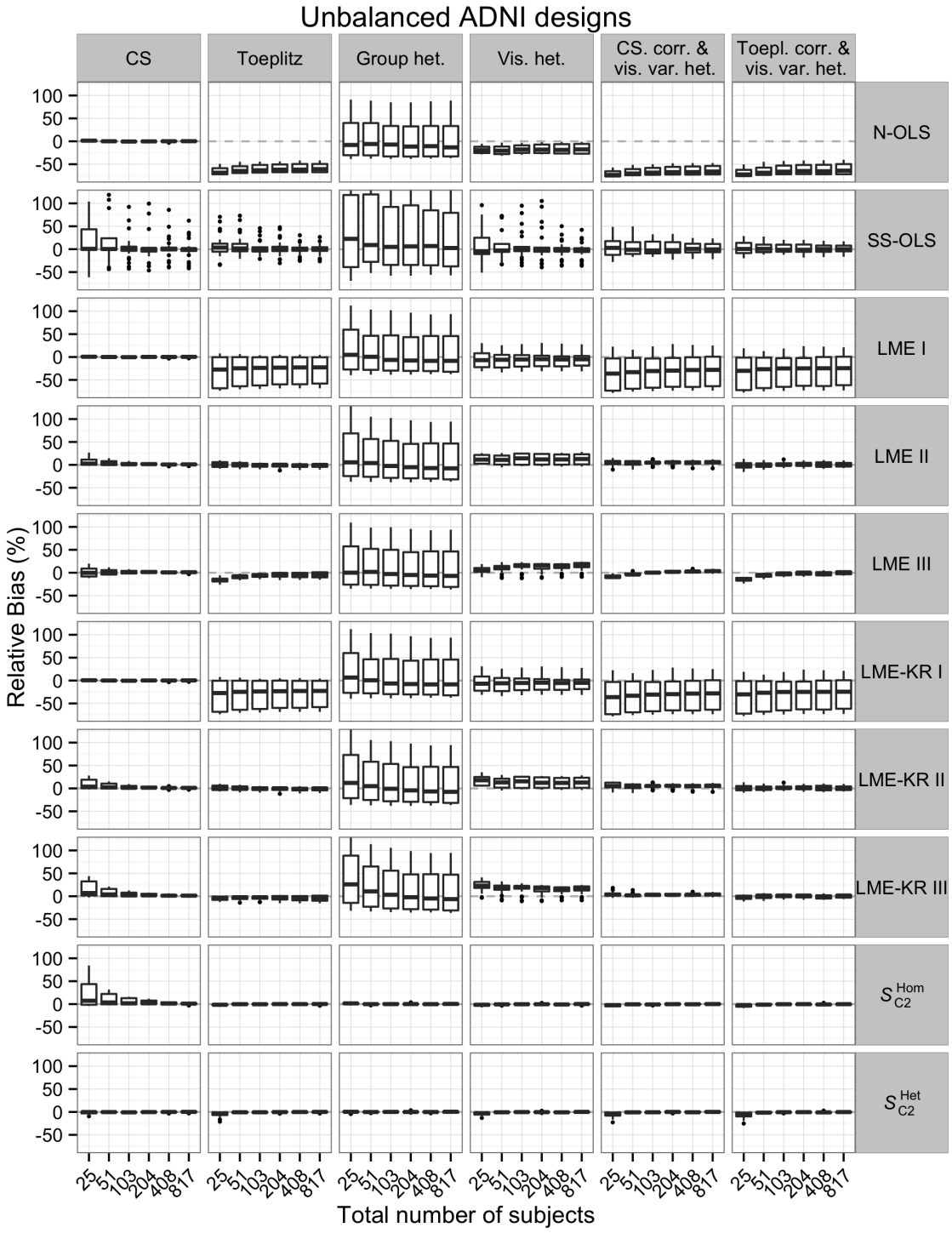

Fig. 3.8 Boxplots showing the Monte Carlo relative bias of several methods as a function of the total number of subjects in the unbalanced ADNI designs over 144 scenarios (consisting of the 24 contrasts tested and the 6 within-subject covariance structures considered in Simulations I).

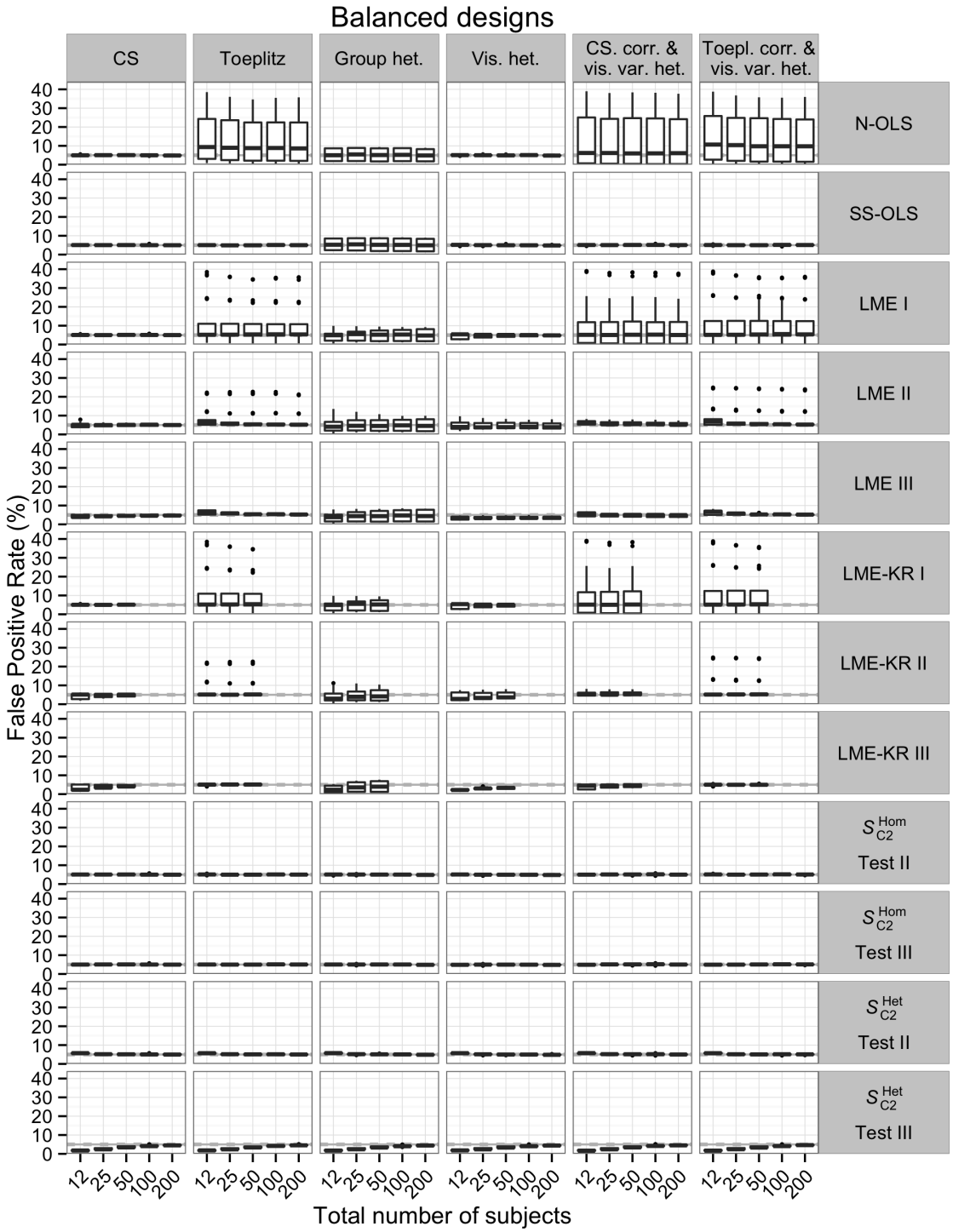

Fig. 3.9 Boxplots showing the Monte Carlo FPR of several methods as a function of the total number of subjects in the balanced designs over 162 scenarios (consisting of the 9 contrasts tested, the 6 within-subject covariance structures and the 3 numbers of visits per subject considered in Simulation II). Results for the LME-KR models in the designs with 100 or 200 subjects were not computed due to the prohibitive computation time of the function `get_ddf_Lb` used to compute the Kenward-Roger degrees of freedom.

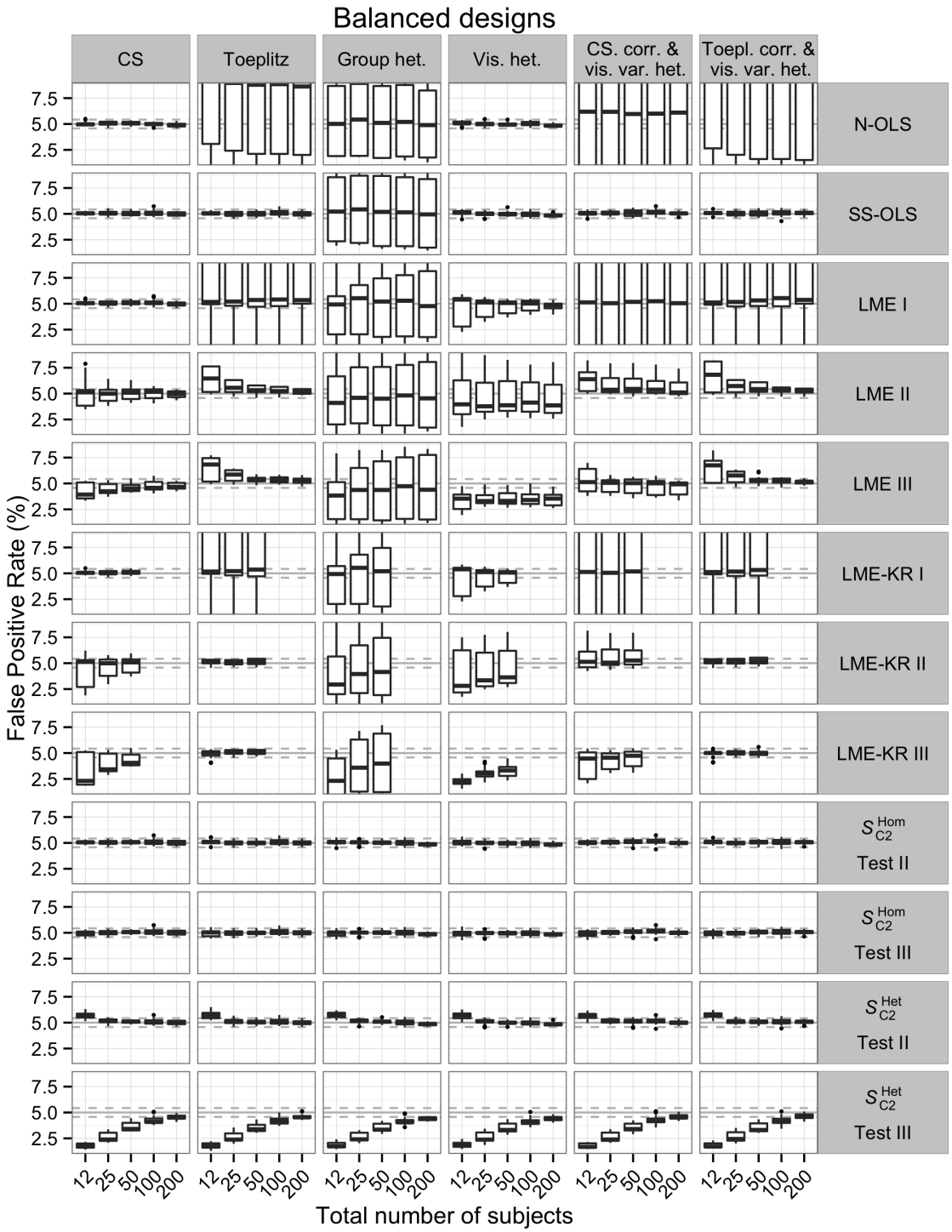

Fig. 3.10 Zoomed version of Figure 3.9 where only the FPRs between 1% and 9% are shown.

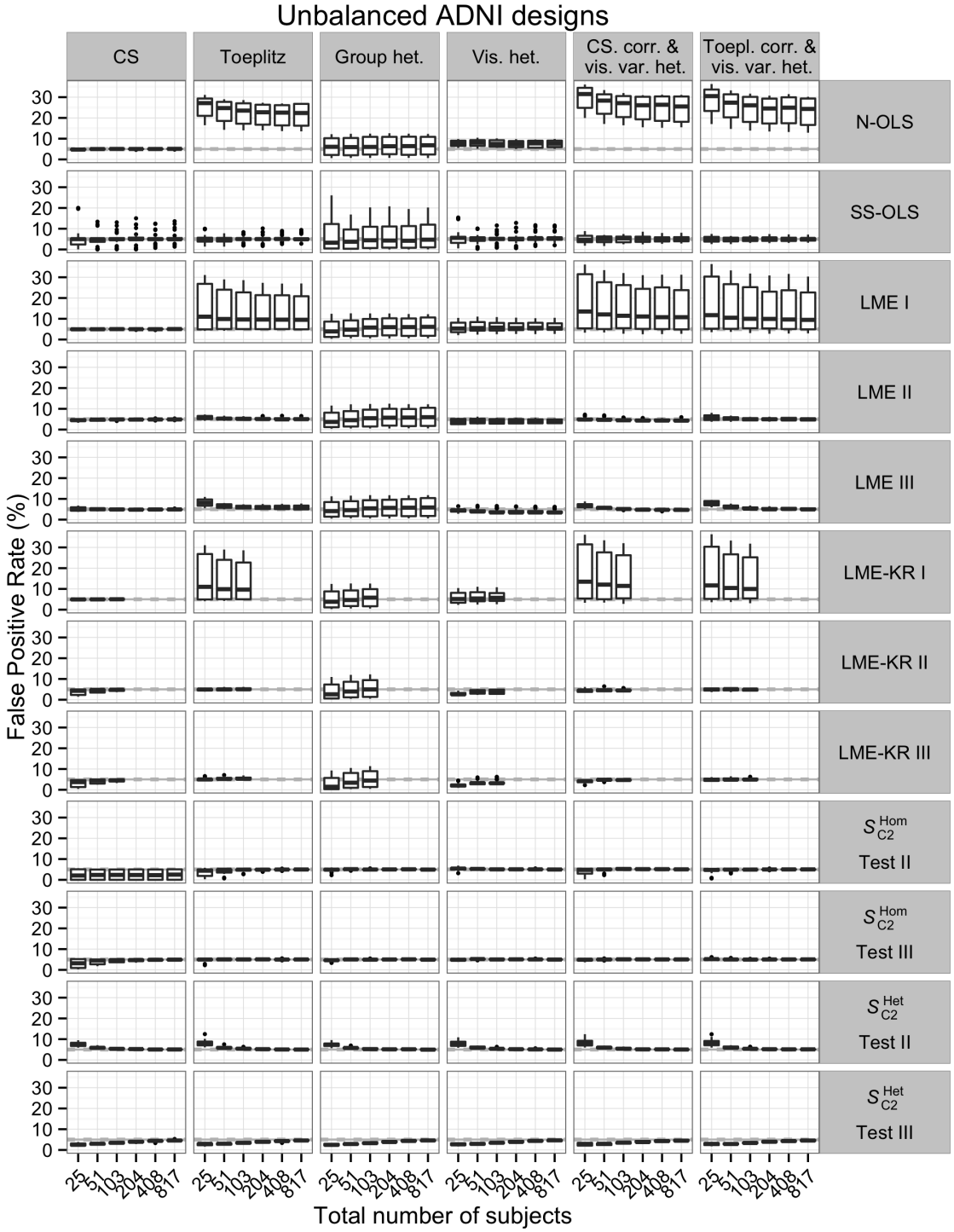

Fig. 3.11 Boxplots showing the Monte Carlo FPR of several methods as a function of the total number of subjects in the unbalanced ADNI designs over 144 scenarios (consisting of the 24 contrasts tested and the 6 within-subject covariance structures considered in Simulations II). Results for the LME-KR models in the designs with 204, 408 or 817 subjects were not computed due to the prohibitive computation time of the function `get_ddf_Lb` used to compute the Kenward-Roger degrees of freedom.

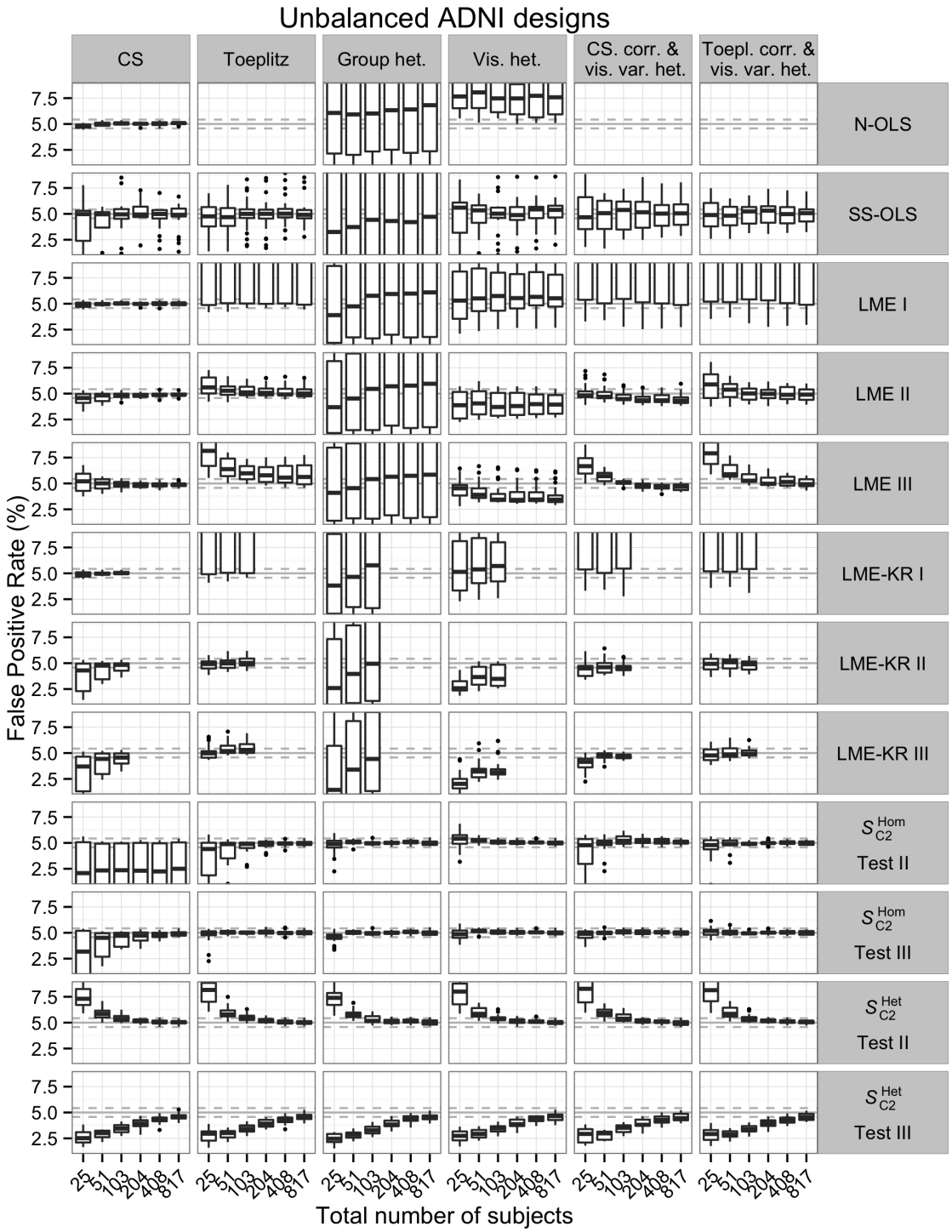

Fig. 3.12 Zoomed version of Figure 3.11 where only the FPRs between 1% and 9% are shown.

Figure 3.13 shows the FPR (first column in the figure) and the power obtained for two effect sizes (second and third columns in the figure) of the difference between Group B and Group A in terms of the linear effect of visits in the balanced designs with 5 visits and under CS. First, in these scenarios, we see that all the methods were valid, but LME II, LME III, LME-KR II, LME-KR III and the SwE versions under Test III tended to be conservative, particularly in small samples. We observe that the most powerful methods were the N-OLS, LME I and LME-KR I methods. The SS-OLS method, $S^{\text{Hom}}_{\text{C2}}$ with Test II, $S^{\text{Hom}}_{\text{C2}}$ with Test III and $S^{\text{Het}}_{\text{C2}}$ with Test II were less powerful but not by much. Finally, mainly due to their strong conservativeness in the control of the FPR, all the other methods seemed the least powerful methods. In particular, we observed that, in these scenarios, the use of the Kenward-Roger corrections tended to make the LME methods more conservative and consequently less powerful. Nevertheless, it is worth pointing out that, in other scenarios, these corrections were really useful to improve the uncorrected LME methods which were liberal and therefore invalid (e.g., under Toeplitz for LME II and III; see Figure 3.10). Finally, from these results, we also observe that the differences of power between the methods decreased quickly when the sample size increased and did not seem significant in the largest samples.

Figure 3.14 shows the FPR (first column in the figure) and the power obtained for two effect sizes (second and third columns in the figure) of the difference between AD and MCI in terms of the longitudinal effect of age in the ADNI designs under CS. First, in these scenarios, we observe that the SS-OLS method was liberal for almost all the sample sizes, the SwE method with $S^{\text{Het}}_{\text{C2}}$ and Test II was liberal in the design with 25 subjects and the three other SwE procedure were relatively conservative, mainly in small samples. We observe, like in the balanced designs, that the N-OLS, LME I and LME-KR I methods were the most powerful methods. The use of more complicated LME models, particularly those using the Kenward-Roger corrections were less powerful. The SwE method using $S^{\text{Het}}_{\text{C2}}$ and Test II, except for the case with 25 subjects where it was not valid, seemed to have a power in the range of the more complicated LME approaches. When the sample size was not too large, it seemed even more powerful than LME-KR II and LME-KR III which appeared to be slightly conservative in small samples. Nevertheless, in some other designs, the SwE method appeared slightly less powerful than all the LME approaches. This can be explained by two facts. First, all the LME models imposed by construction a structure to the within-subject covariance matrices while the SwE method does not. This implies that the variability of the estimates of the within-subject covariance matrices in the LME

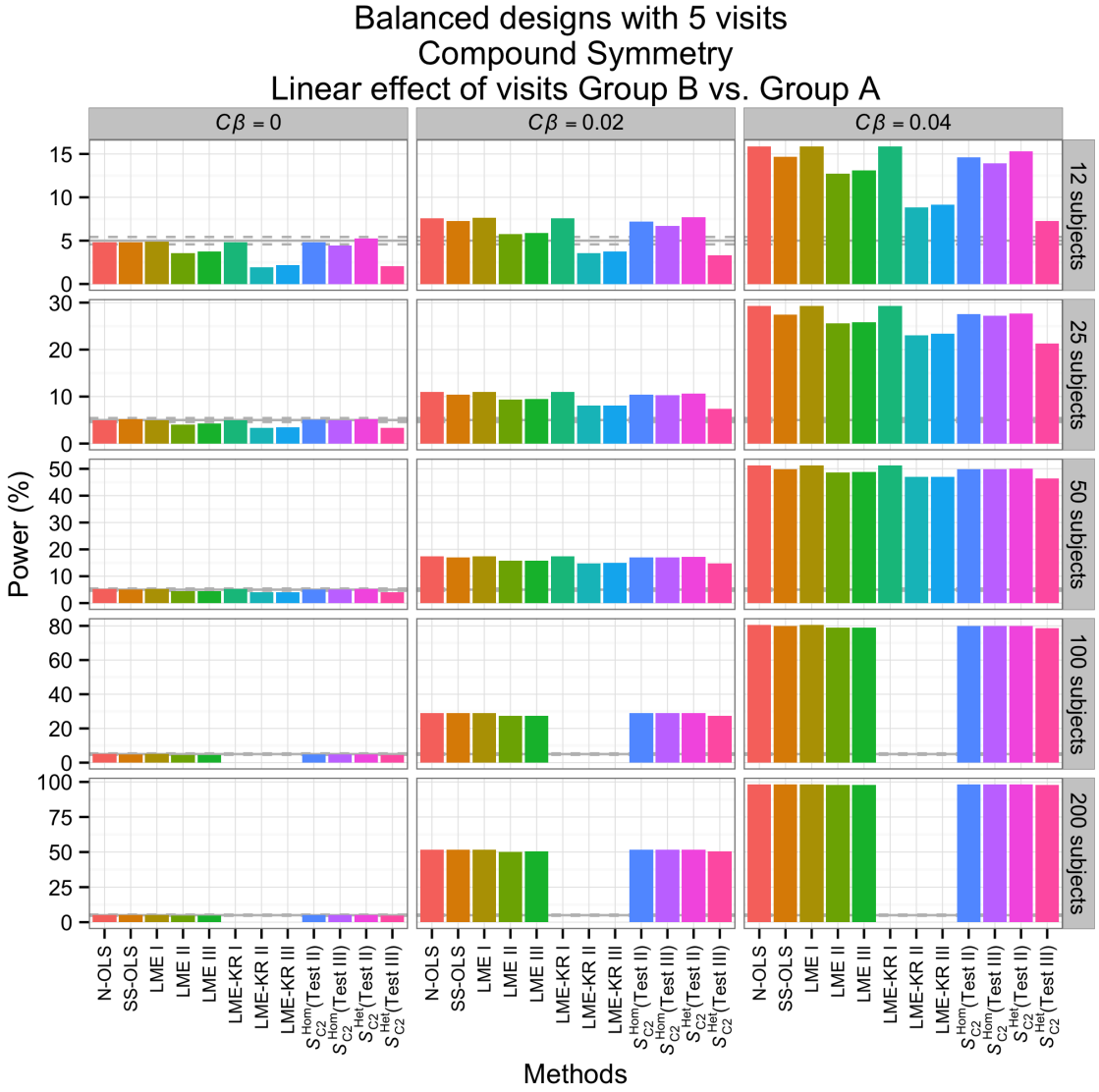

Fig. 3.13 Barplots showing the Monte Carlo FPR and Power of several methods for the balanced designs with 5 visits and under CS obtained from Simulation II. For computational reasons, the results for LME-KR I, II and III in the designs with 100 or 200 subjects were not computed and therefore are not shown. Note that, for clarity, the scales for the FPR and power are different over sample sizes.

models is expected to be lower than in the SwE method, allowing more powerful inferences. Second, in the SwE method, as we do not assume the same covariance structure across subjects or groups of subjects, when we used a test involving only a subset of the subjects like, for example, the AD and MCI subjects, we do not use any information from the other subjects (in the example, the Normal subjects), decreasing

the effective sample size. This is not the case for all the other methods which assumes a common within-subject covariance matrices for all the subjects and, therefore, would also use the information from the Normal subjects in a test involving only the AD and MCI subjects. Note that the latter can be dangerous, as if there is heterogeneity across groups of subjects, the inference can be inaccurate as it was observed for the control of the FPR under group heterogeneity in Figures 3.9, 3.10, 3.11 and 3.12. Regarding the use of $S_{\text{C2}}^{\text{Hom}}$ in the particular scenario of Figure 3.14 , we see that, it suffered from conservativeness, penalising it in terms of power. Nonetheless, we see that, under Test III and with enough subjects, its conservative nature tended to disappear allowing it to be almost as powerful as the LME methods in large samples. It would be however unfair to state that the SwE method using $S_{\text{C2}}^{\text{Hom}}$ under Test III is strongly less powerful than the other methods in small samples as these scenarios with CS were almost the only ones where it was conservative. Indeed, in the other scenarios, it seemed almost always accurate while all the other methods were frequently struggling to control the FPR (see, for example, Figure 3.12). Finally, while the SS-OLS method was generally liberal, it clearly tended to be less powerful than the other approaches. This effect can be explained by the fact that the true variance of the parameters were larger for the SS-OLS method than for the other methods, making the SS-OLS method less efficient to detect effects.

3.3.3 Real data analysis

Figure 3.15 shows the Box's test F-score image (centred at the anterior commissure) thresholded after controlling for a False Discovery Rate (FDR) of 5% (on the left) and after using a Bonferroni correction at 5% level of significance (on the right). 97% of the in-mask voxels survived the FDR thresholding while 56% of them survived the Bonferroni thresholding, indicating a strong evidence of non-Compound Symmetry in the brain and challenging the validity of the N-OLS method.

Figure 3.16 compares the t-score images obtained by the N-OLS, SwE ($S_{\text{C2}}^{\text{Hom}}$, Test III) and SS-OLS methods with the real images for contrasts on the difference between groups in terms of visit effect on the brain atrophy. For comparison, we used a threshold of 5 for positive effects (i.e. greater atrophy rate) and -5 for negative effects (i.e. greater expansion rate). The N-OLS method had larger t-values and more supra-threshold voxels than the SwE method. While this could be attributed to power differences, with 817 subjects, we expect negligible differences in power. Hence a more likely explanation is the presence of a complex (non-CS) longitudinal covariance structure that results in inflated significance (see Figures 3.11 and 3.12, first row). The

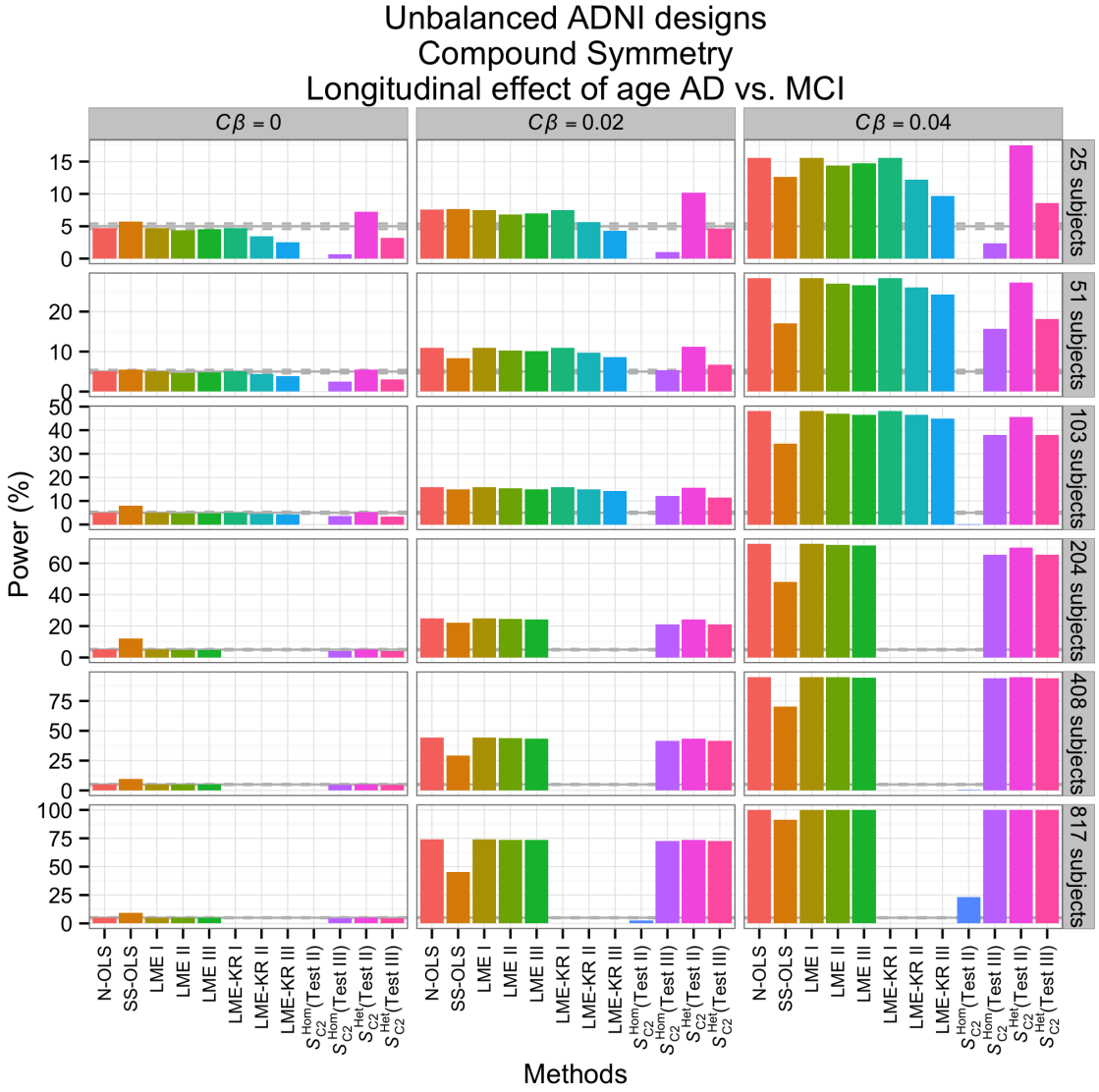

Fig. 3.14 Barplots showing the Monte Carlo FPR and Power of several methods for the unbalanced ADNI designs under CS obtained from Simulation II. For computational reasons, the results for LME-KR I, II and III in the designs with 204, 408 or 817 subjects were not computed and therefore are not shown. Note that, for clarity, the scales for the FPR and power are different over sample sizes.

SS-OLS had smaller t-values and fewer supra-threshold voxels than the SwE method, likely attributable to conservativeness (see Figure 3.12, second row) and/or reduced power (Figure 3.14, sixth row).

Figures 3.17 , 3.18 and 3.19 shows the regression fits for three particular voxels situated in different areas of the brain. Note that these voxels were not selected based

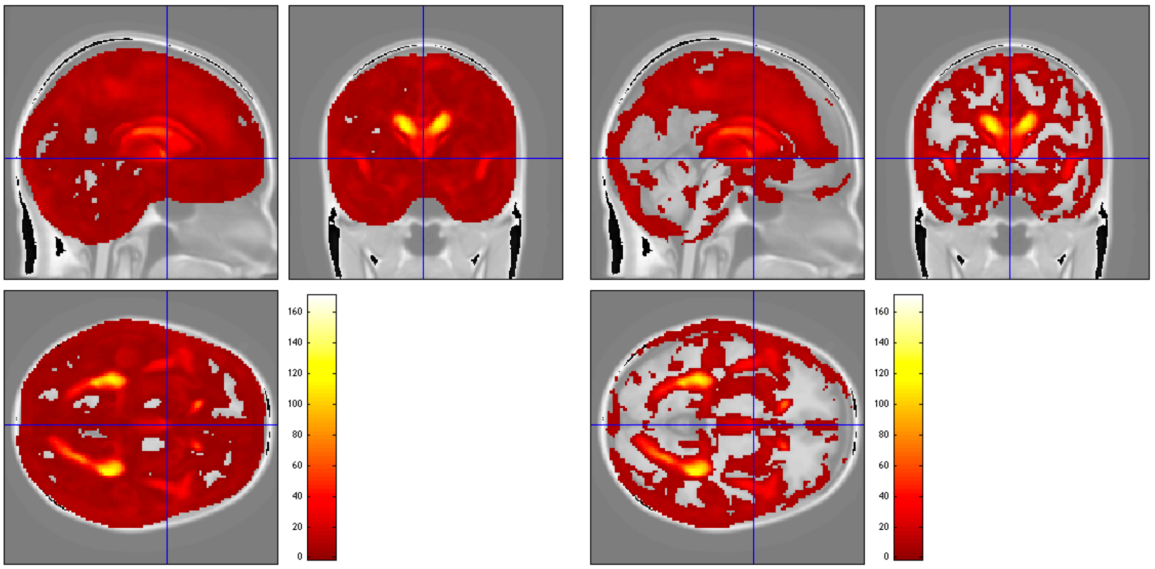

Fig. 3.15 Box's test of Compound Symmetry F-score image on the ADNI dataset thresholded at 5% after using an FDR correction (left) and a Bonferroni correction (right). 97% of the in-mask voxels survived the FDR thresholding while 56% of the in-mask voxels survived the Bonferroni thresholding, indicating extensive regions incompatible with the CS assumption.

on maximal difference between the SwE and N-OLS (or SS-OLS) methods, but rather based on relatively high significance in term of age, visit or acceleration effects in all of the methods (qualitatively, the statistic maps for the three methods are similar). As a reminder from Section 2.6, all the scans represent the relative difference in brain volume from the MDT reference image, as such, a value of 10% in the plots indicates that the brain volume is 10% bigger than in the MDT image. Figure 3.17 shows results for a voxel in the right anterior cingulate where there is strong evidence of brain atrophy with age and also with the visit effect. The rate of brain atrophy seems similar for each group and is similar for both the age and the visit effect, indicating consistent cross-sectional and longitudinal volume changes. Figure 3.18 shows a voxel in the right ventricle where there is strong evidence of an expansion in volume. As expected, this is greater in AD subjects than in MCI or Normal subjects. Figure 3.19 shows a voxel in the right posterior cingulate where we observe strong brain atrophy for the AD subjects compared to the Normal subjects. In Figures 3.17, 3.18 and 3.19, the Normal subjects have similar intra- and inter-subject effects of time (visit and age effects, respectively), and we generally observe this throughout the brain. In contrast, in the AD and MCI groups, there are inconsistent longitudinal and cross-sectional effects of time. Specifically, there is evidence of a "deceleration", where the

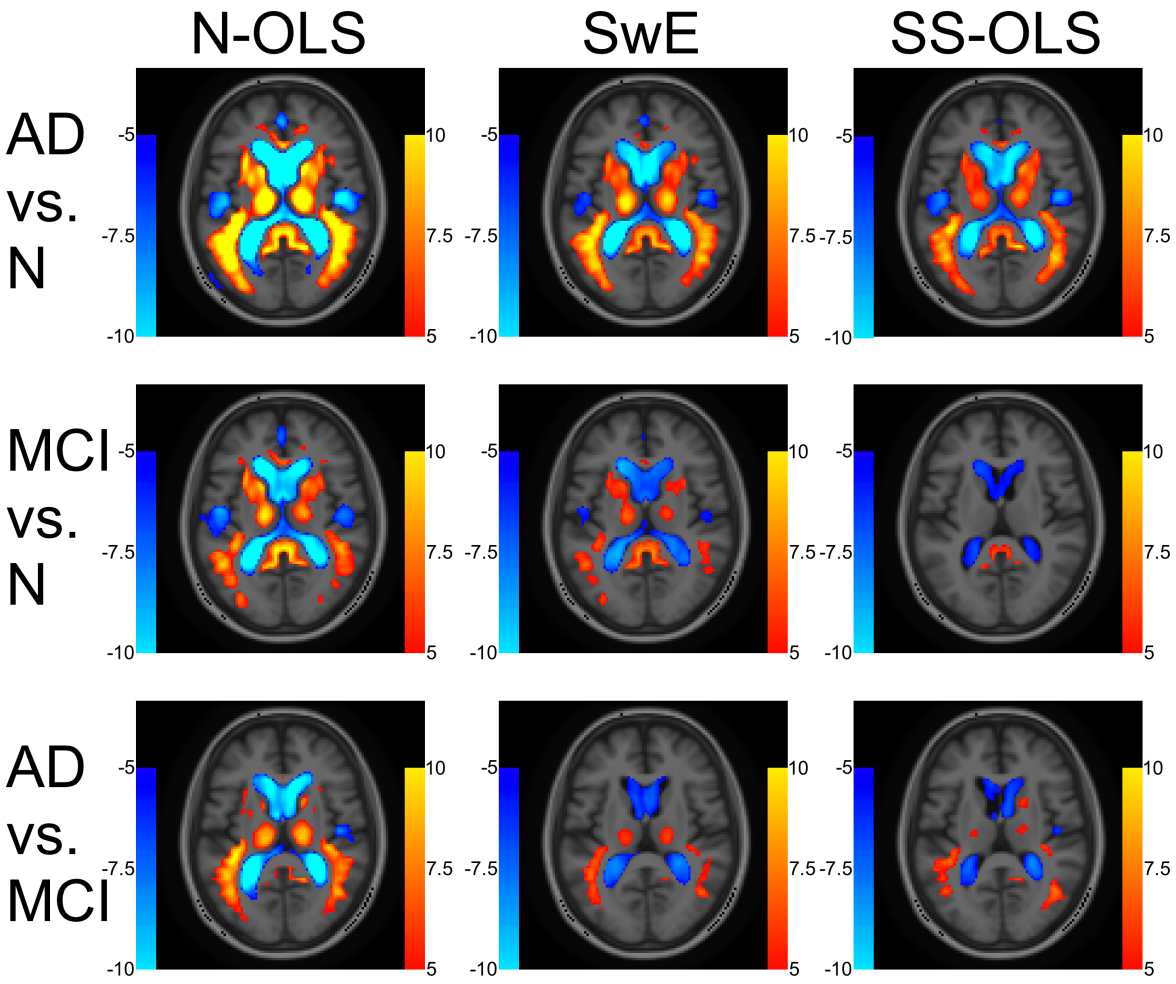

Fig. 3.16 Thresholded t-score images (axial section at $z = 14$ mm superior of the anterior commissure) for the differential visit effect, greater decline in volume in AD relative to N, MCI relative to N and AD relative to MCI, for the N-OLS, SwE (S_{C2}^{Hom}, Test III) and SS-OLS methods. For all the methods, a threshold of 5 for the positive effects (i.e. greater atrophy rate) and a thereshold of -5 for the negative effects (i.e. greater expansion rate) was used. Apparent superior sensitivity of the N-OLS method (left) is likely due to inflated significance and poor FPR control; see text and Figures 3.11 and 3.12.

oldest patients exhibit reduced rates of expansion (or contraction) relative to younger patients. One interpretation is a "saturation" effect, where, with advancing disease progress, there is less gray matter left to atrophy and less space in the cranial vault for the ventricles to expand. However, as the ADNI only follows subjects for at most three years, an alternative interpretation must be considered. Specifically, instead of this deceleration reflecting an aspect of the disease process, it rather reflects age-dependent heterogeneity in the ADNI cohort. For example, MCI subjects in their 80's

are likely to have systematic differences from the MCI subjects in their 60's, as the former group have survived to their 8th decade free of severe dementia, while some of the latter group will convert to AD in the next 20 years. This kind of explanation has already been reported in Thompson et al. (2011).

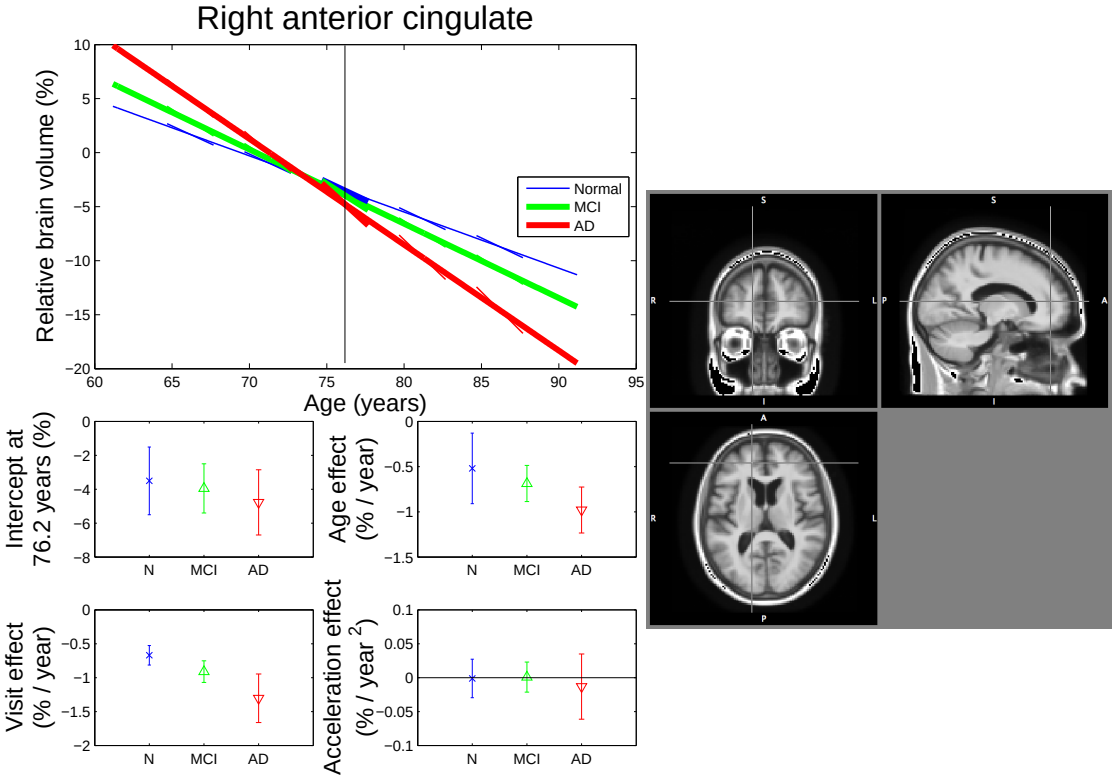

Fig. 3.17 Model fit in the right anterior cingulate cortex. Top plot: linear regression fit obtained with the SwE method ($S_{\mathrm{C2}}^{\mathrm{Hom}}$) at voxel $(x, y, z) = (16, 45, 14)$ mm; the vertical line at 76.2 years marks the average age of the study participants; the thickness of the lines reflects the strength of the t-scores obtained for the age effect (the three main lines), the visit effect (the three secondary lines centred at 76.2 years) and the acceleration effect (the secondary lines centred at 66.2, 71.2, 81.2 and 86.2 years). Bottom plots: 95% confidence intervals for all the parameters of the linear regression. Right image: location of the selected voxel. The confidence intervals suggest that the rate of brain atrophy is similar for each group and for both the age and the visit effect, indicating consistent cross-sectional and longitudinal volume changes.

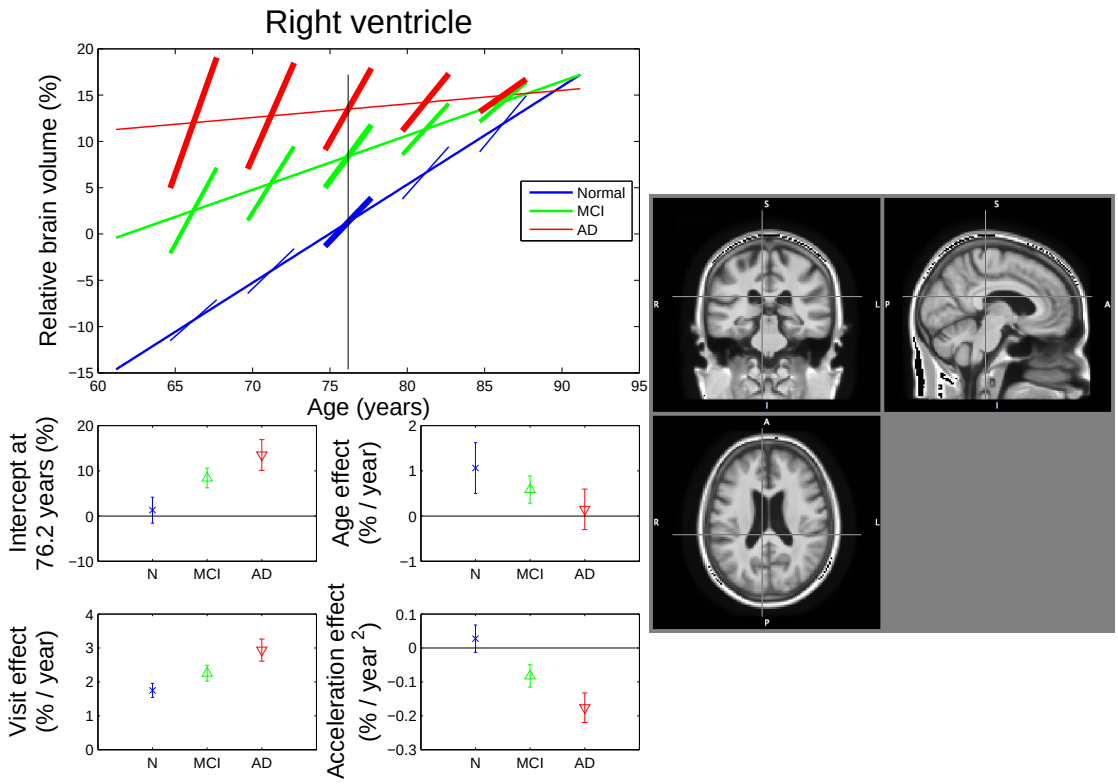

Fig. 3.18 Model fit in the right ventricle. Top plot: linear regression fit obtained with the SwE method ($S^{\text{Hom}}_{\text{C2}}$) for voxel $(x, y, z) = (8, \text{-}35, 24)$ mm. (See Figure 3.17 caption for a description of the different figure components). In the AD and MCI groups a mismatch is observed between cross-sectional and longitudinal effects of time, with a reduced rate of change with increasing age; see body text for more discussion.

3.4 Conclusions

In this chapter, we have reviewed several versions of the SwE and, using intensive Monte Carlo simulations in a range of settings important for longitudinal neuroimaging data, we have isolated the best two versions as $S^{\text{Het}}_{\text{C2}}$ and $S^{\text{Hom}}_{\text{C2}}$. They were almost always unbiased in our simulations (see Figures 3.1, 3.2 and 3.3), except in the unbalanced ADNI designs under CS for which $S^{\text{Hom}}_{\text{C2}}$ had the tendency to overestimate the true variances, particularly in small samples (see Figure 3.3). A possible explanation for this misbehaviour is that, in these scenarios with missing data and true covariance matrices close to the boundary of the set of positive semi-definite matrices, the estimator we used to estimate the common within-subject covariance matrices might have yielded non positive semi-definite matrices for some realisations. As, in such cases,

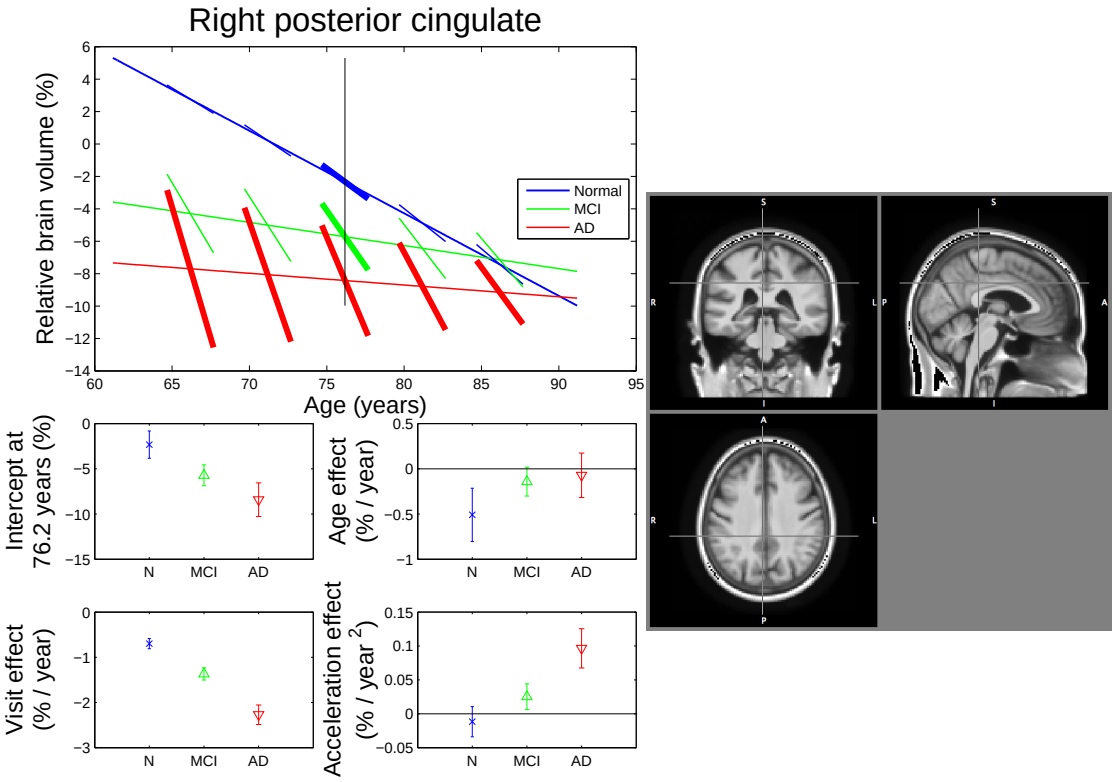

Fig. 3.19 Model fit in the right posterior cingulate. Top plot: linear regression fit obtained with the SwE method ($S^{\text{Hom}}_{\text{C2}}$) for voxel (x, y, z) = (4, -39, 38) mm. (See Figure 3.17 caption for a description of the different figure components). In the AD and MCI groups, there is a mismatch between cross-sectional and longitudinal effects of time, with a reduced rate of change with increasing age; see body text for more discussion.

the problematic estimates were adjusted by zeroing their negative eigenvalues, it is likely that the latter procedure induced a positive bias in the estimation, explaining the results observed. Also, the bias reduction observed when the number of subjects increased may simply be attributed to the reduction of variability of the estimator. Indeed, with a reduced variability, the estimator probably yielded estimates closer to the true covariances, reducing the frequency of having an estimate outside the boundary, but also reducing the average distance of the non positive semi-definite estimates from the boundary, explaining the bias reduction. However, further research is needed to confirm this explanation and also to find a way to adjust for this misbehaviour.

We have proposed three new statistical parametric tests (Test I, Test II and Test III) to make inference with the SwE and compared them to the widely used χ^2-test

and the Pan test (Pan and Wall, 2002). When S_{C2}^{Hom} was used, Test III, which accounts for the presence of missing data, seemed to be the best test, only struggling under CS in the unbalanced ADNI designs where it tended to be conservative, particularly in small samples. Nevertheless, this misbehaviour is likely due to the bias observed in the SwE in these scenarios and not specifically to Test III itself, indicating a very good behaviour of the test. Also, while Test II, which accounts only partially for the presence of missing data, did not perform so well in the unbalanced ADNI designs, it worked extremely well in the balanced designs and seemed even to work slightly better than Test III in the scenarios with only 12 subjects for which Test III was slightly conservative. The likely explanation for the latter resides in the fact that Test II accounts for the presence of a small sample bias while Test III does not. When S_{C2}^{Het} was used, no statistical test was accurate in the smallest samples. Test I and Test III, which were identical with S^{Het}, were conservative, while the Pan test and Test II were liberal. The conservativeness of Test I (or Test III) can simply be explained by the fact they do not account for the small sample bias. The liberality of the Pan test is difficult to explain as it is unclear what all the assumptions are behind the test. A possible explanation for the liberality of Test II, which accounts for the small sample bias, is that it makes the assumption that the dependence between the subject residuals is only due to the presence of the pure between-subject covariates and not the presence of the within-subject covariates as well. While this seems to be a reasonable assumption as we can expect that the influence of the within-subject covariates would be in general negligible, this may not be the case in very small samples, explaining that the degree of dependence is higher than assumed. Further research on this is needed to confirm this explanation and to eventually find a way to account for the influence of the within-subject covariates. Nevertheless, test II seemed to become relatively quickly accurate when the sample size increased. Therefore, in small samples and if it is possible to classify the observations into consistently defined visit categories, it seems that S_{C2}^{Hom} should be preferred to S_{C2}^{Het} as it yields more accurate inferences. However, when the sample size is large enough, S_{C2}^{Het} may be considered as well, and has the advantages of being free from the need to classify the observations into visit categories and allowing for heterogeneity across all subjects while S_{C2}^{Hom} allows this only across groups of subjects. Here, it is worth also mentioning that, for the development of Test II and Test III, we have proposed two estimators for $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$, both accounting, at least partially, for the presence of missing data. In addition, the one developed for Test II accounts for a small sample bias while the one developed for Test III accounts for a missing data bias. To the best of our knowledge, we are the first to propose

such kind of estimators. They might be useful in other applications than the SwE method where an estimate of the covariance matrix of the elements of a covariance matrix estimator would be needed. In particular, any application using an estimator following a Wishart distribution might use the estimator developed for Test II as it should be unbiased in such cases.

We have shown that the SwE method is a flexible computationally efficient (no iterative algorithms) alternative to the N-OLS, SS-OLS and LME methods. When the simplest covariance structure, CS, cannot be assumed, the SwE (S_{C2}^{Hom}) method was the only method that consistently controlled the FPR. In particular, the SS-OLS method was not able to control the FPR in the ADNI designs. This effect can be explained by the fact that an inhomogeneity in the distribution of the summary statistics is likely to occur when the subjects do not have the same number of observations, leading to a lack of control of the FPR as observed in our simulations. We have also shown that the N-OLS, SS-OLS and LME methods may be inaccurate when there exists heterogeneity in group variance. Nevertheless, it is worth noting that all of these methods can be adapted to accommodate such a heterogeneity by, for example, specifying different variances for each group in their model. In the SwE method, the use of a marginal model simplifies the specification of the predictors and the interpretation of parameters. In particular, both within- and between-subject covariates can be used (that is not possible with the N-OLS method), and we have illustrated the ease with which cross-sectional and longitudinal time effects can be used. In particular, testing the interaction of these two time effects revealed a "deceleration" effect in the MCI and AD patient groups that was missing from the healthy controls. We have noted, however, the importance of replacing an arbitrary covariate with two, one purely within-subject and one purely between-subject.

The principal limitation of the SwE method regards power. It is generally less powerful than other methods like the random-intercept LME and N-OLS methods. However, the difference observed in our simulations was generally small and when CS did not hold or when there was variance heterogeneity, the N-OLS, SS-OLS and random-intercept LME generally failed to control the FPR and were unusable. In our simulations, more complicated LME models were more accurate than the random-intercept LME model, but still seemed to struggle to give accurate inferences, being either liberal, accurate or conservative. So, even if they have the potential to give more powerful inferences than the SwE method, they can be invalid or penalised by their conservative nature. Thus, the potential lack of power of the SwE method seems like a reasonable price to pay for validity.

If more power is needed, one can use some form of spatial regularisation or more complicated models like in Skup et al. (2012), Bernal-Rusiel et al. (2013b) or Li et al. (2013). Nevertheless, while those methods are expected to be more powerful, they require iterative algorithms, which makes them slower than the SwE method. Moreover, there is no evidence that, at least in some settings, they will do this with a good control of the FPR. Notably, Zhang (2008) showed that using a spatial regularisation will tend to decrease the variance of the estimates (which will tend to increase the power), but also increase their bias (which will tend to alter the accuracy). On this, a spatial regularisation of the data covariance matrix estimates used in the SwE is investigated in Chapter 6, but have not shown promising results so far.

Note that we have not investigated one-sample t-tests on subject summary statistics. While one-sample t-tests have been shown to be robust under heterogeneity (Mumford and Nichols, 2009), these methods are however less flexible than other regression methods which allow for the inclusion of covariates. Another approach not investigated in this chapter is the SPM procedure introduced in Section 2.3.4. The main reason for this is that this method is difficult to assess in simulations or in a real data analysis. In particular, it seems important to check for the validity of the assumption of a common covariance structure across the brain made by the SPM procedure and a formal statistical test for this would then be needed.

Regarding the real data analysis, we have found ample evidence, through the use of the Box's test of CS, that the ADNI data's within-subject covariance matrices are inconsistent with CS, challenging the validity of the N-OLS and the random-intercept LME methods with this dataset. The N-OLS, SS-OLS and SwE methods showed clearly different results with the SwE method finding fewer significant voxels than the N-OLS method, but more than the SS-OLS method. This seems to be in accordance with our non-CS simulations (see Section 3.2.5) in which the N-OLS method poorly controls the FPR (and thus has inflated significance; see Figure 3.11) and the SS-OLS method which is less powerful than the SwE method (see Figure 3.14). In the simulations, except for the CS scenarios in small samples where it had the tendency to be conservative in the unbalanced ADNI designs, the SwE was accurate for all the different types of covariance structures tested and this seems to make the SwE one of the most trustworthy methods for the analysis of the ADNI data.

It would be desirable to use permutation methods (see, e.g., Nichols and Holmes, 2002) in combination with the SwE to produce non-parametric inferences. However, permutation tests assume that the scans are exchangeable under the null hypothesis, incompatible with longitudinal or repeated measures data. Bootstrap methods (see,

e.g., Efron and Tibshirani, 1994), in contrast, do not require the exchangeability assumption and may be applicable. The use of a particular type of bootstrap method to use with the SwE, called Wild Bootstrap, is investigated in Chapter 4.

As another future direction, we also intend to check the validity of the Random Field Theory (see, e.g., Worsley et al., 1996) with the SwE method. It is indeed not guaranteed that the assumptions required by the Random Field Theory hold when the SwE method is used. As such, at present, we can only recommend the use of a False Discovery Rate control in order to deal with the multiple comparison problem.

Finally, note that an `SPM` extension implementing some versions of the SwE method presented in this chapter has been made freely available for use at http://warwick.ac.uk/tenichols/SwE.

Chapter 4

Non-parametric inference with the Sandwich Estimator

4.1 Introduction

In Chapter 3, we assumed that the test statistics (Wald scores) obtained with the SwE method follow a parametric distribution under the null hypothesis, by notably assuming Normal error terms and that the contrasted SwE $CSC^\top$ follows a Wishart distribution. In biostatistics, it is however often desirable to relax these type of assumptions by using a resampling method instead to estimate the null distribution of the test statistics and use this estimated distribution to make inferences. As mentioned in section 2.5.1, resampling methods can be divided into two main categories: (a) resampling methods with replacement and (b) resampling methods without replacement. In neuroimaging, the most popular resampling method is the permutation test (Nichols and Holmes, 2002; Winkler et al., 2014) which is based on resampling schemes without replacement. While very useful in many cases, it relies on the assumption that the data we want to permute is exchangeable under the null hypothesis. In the case of longitudinal data, this assumption of exchangeability is in general not supported as the data is correlated and heterogeneous over subjects (e.g., variable number of visits per subjects, or heterogeneous within-subject covariance matrices). Therefore, the possibility of using permutation tests to analyse longitudinal neuroimaging data is limited to particular cases. For that reason, it seems important to find another resampling method which would be valid in a large range of cases.

As an alternative resampling method for non-parametric inference in longitudinal neuroimaging studies, we consider a bootstrap method (see, e.g., Efron and Tibshirani, 1994). In the context of hypothesis testing, bootstrap methods attempt to approxi-

mate the null distribution of the statistic of interest by randomly resampling the data with replacement. However, in order to be accurate, the Data Generating Process (DGP) used to resample the data should be as close as possible to the DGP which would generate data observed under the null hypothesis. Unfortunately, in the context of longitudinal data, the within-subject covariance matrices are generally unknown and may vary across subjects or groups of subjects. Therefore, it seems difficult to imitate directly the true DGP and the use of typical bootstrap procedures does not seem appropriate in our context. Nevertheless, a particular type of bootstrapping called the Wild Bootstrap (WB) addresses some of these issues. This bootstrap method was initially proposed in the context of linear regression models with heteroskedastic independent errors by Liu (1988), following the work made by Wu (1986) and its extension to the case of clustered data seems to first appear in the work of Brownstone and Valletta (2001). In the literature, several versions of the WB can be found and have mainly been studied in the case of cross-sectional data (Davidson and Flachaire, 2001; Flachaire, 2005; Davidson and Flachaire, 2008). While a few studies exist in the case of clustered data (Cameron et al., 2008; Webb, 2013), to our knowledge, none seems to be specifically related to longitudinal data.

In neuroimaging, the WB has mainly been used in the context of Diffusion Tensor Imaging (DTI; see, e.g., Whitcher et al., 2005; Chung et al., 2006; Whitcher et al., 2008; Zhu et al., 2008). However, with DTI, the model uses independent errors and has the objective of estimating the sampling distribution of measures like the Fractional Anisotropy (FA). The application is therefore relatively different from ours, where we have correlated errors and where the objective is not the distribution of the model parameters, but the null distribution of the test statistics related to the combination of the parameters we want to test. Nevertheless, while still in the context of models with independent errors, a more similar application to ours can be found in Zhu et al. (2007) where the authors used the WB to make inference on associations which may exist between some brain morphology measures and some covariates of interest.

In this chapter, we introduce the WB and describe several versions of it. Then, we compare these different WB versions using intensive Monte Carlo simulations in a large range of scenarios important for longitudinal neuroimaging data analysis and isolate the best versions. Finally, we illustrate the WB by using it to analyse the real ADNI dataset described in Section 2.6 and notably show that it can be used to control the Family-Wise Error Rate (FWER).

4.2 Methods

4.2.1 The Unrestricted Wild Bootstrap

Let us consider the marginal model defined in Section 2.3.6, where, for each subject i, we have

$$y_i = X_i\beta + \epsilon_i^*, \tag{2.4 revisited}$$

and that the SwE method is used to estimate β and the covariance matrix $\text{Cov}[\hat{\beta}]$ (see Chapter 3). To make inference on a combination of the parameters $\mathcal{H}_0 : C\beta = b_0$ where C is a matrix (or a vector) of rank q defining the combination of the parameters (contrast) tested, the Unrestricted WB (U-WB) considers, like in a standard parametric test, the Wald statistic as defined in Section 2.5.1,

$$T = (C\hat{\beta} - b_0)^\top (CSC^\top)^{-1}(C\hat{\beta} - b_0)/q, \tag{2.8 revisited}$$

that we will refer to as T_0 in this chapter. To estimate the null distribution of this statistic, the U-WB first resamples the data, subject by subject, using the DGP defined as

$$y_i^* = X_i\hat{\beta} + e_i^* f_i, \tag{4.1}$$

where each e_i^* is a vector of small sample bias adjusted subject residuals of subject i and the f_i's are i.i.d. scalar random variables, independent of the original data and respecting, at least, the two following conditions: $\mathbb{E}[f_i] = 0$ and $\mathbb{E}[f_i^2] = 1$. From Equation (4.1), we see that the resampling scheme is relatively simple and consists of resampling independently the adjusted residuals across subjects, but, in such a way that the adjusted residuals of each subject are resampled using the same multiplicative value. The latter is relatively convenient as it allows to preserve, for each resampling, the within-subject covariance structure of each subject without making any assumption on it.

Using Equation (4.1), n_B bootstrap samples are generated. For each bootstrap sample, the SwE method is used to fit the resampled data with the original model in order to get the WB estimates $C\hat{\beta}_b$ and $CS_bC^\top$, which are then used to compute the

U-WB Wald statistic

$$T_b = (C\hat{\beta}_b - C\hat{\beta})^\top (CS_bC^\top)^{-1}(C\hat{\beta}_b - C\hat{\beta})/q. \tag{4.2}$$

Note that the U-WB Wald statistics are centred around the original fitted value $C\hat{\beta}$ instead of b_0 in Equation (4.2). This is justified by the fact that the U-WB DGP as defined by Equation (4.1) resamples the data assuming that $\beta = \hat{\beta}$ and, consequently, the U-WB Wald statistics need to be centred around $C\hat{\beta}$ and not b_0.

The n_B bootstraps and the original Wald statistics are then used to estimate the null distribution of the original statistic T_0, which can be used to make inference on it. Specifically, we can compute the U-WB p-value as the proportion of bootstrap Wald statistics (the original Wald statistic T_0 included) which are superior or equal to the original Wald statistic T_0, i.e.

$$\frac{1}{(n_\mathrm{B}+1)} \sum_{b=0}^{n_\mathrm{B}} \mathcal{I}[T_b \geq T_0], \tag{4.3}$$

where $\mathcal{I}$ is the indicator function.

4.2.2 The Restricted Wild Bootstrap

As mentioned in Section 4.2.1, the U-WB DGP as defined in Equation (4.1) resamples the data assuming that $\beta = \hat{\beta}$ and, consequently, the resampled data cannot be considered as representative of the null hypothesis. Therefore, instead of using the U-WB DGP, we can use the following DGP which assumes that the null hypothesis is true:

$$y_i^* = X_i\tilde{\beta} + \tilde{e}_i^* f_i \tag{4.4}$$

where $\tilde{\beta}$ and the $\tilde{e}_i^*$'s are the restricted OLS parameter estimates and some restricted small sample adjusted residuals obtained after imposing the null hypothesis. As we resample the data using a restricted model, we will refer this WB procedure as the Restricted WB (R-WB) to contrast with the U-WB which uses an unrestricted model. The restricted OLS parameter estimates $\tilde{\beta}$ can be obtained by the formula given in Zhu et al. (2007):

$$\tilde{\beta} = \hat{\beta} - (X^\top X)^{-1}C^\top (C(X^\top X)^{-1}C^\top)^{-1}(C\hat{\beta} - b_0). \tag{4.5}$$

The restricted raw residuals are simply given by $y_i - X_i\tilde{\beta}$ and can be small sample bias corrected as the unrestricted residuals, but by accounting also for the restriction imposed to the model. Typically, the latter can be done by replacing the Hat matrix H by the matrix $H - X(X^\top X)^{-1}C^\top(C(X^\top X)^{-1}C^\top)^{-1}C(X^\top X)^{-1}X^\top$ in the small sample bias adjustments defined in Section 3.2.2.

The R-WB procedure to estimate the null distribution of the original statistic and to get a R-WB p-value is the same as the U-WB procedure except that the R-WB Wald statistics are now centred around b_0 and not $C\hat{\beta}$ such that

$$T_b = (C\hat{\beta}_b - b_0)^\top(CS_bC^\top)^{-1}(C\hat{\beta}_b - b_0)/q. \tag{4.6}$$

Note that, while it makes somehow more sense to use the R-WB instead of the U-WB as the resampled data respects the null hypothesis and due to the fact that it has been shown to perform better than the U-WB in some cross-sectional designs (see, e.g., Davidson and Flachaire, 2008), the R-WB has the disadvantage to be contrast-specific. The latter can be annoying in cases where many contrasts need to be tested as the R-WB procedure would need to be repeated for each contrast. This, however, would not be the case for the U-WB which can be used to test several contrasts with the same set of resampled data.

4.2.3 The Restricted SwE vs. the Unrestricted SwE

In Chapter 3, we have always assumed that the SwE was computed using unrestricted residuals. However, Davidson et al. (1985) proposed to use restricted residuals instead, obtained by fitting a restricted OLS model which imposes the null hypothesis as it is done in the R-WB procedure (see Section 4.2.2). This yields a new type of SwE that we will refer to as the Restricted SwE (R-SwE) to contrast with the SwE which uses unrestricted residuals that we will refer to as the Unrestricted SwE (U-SwE) in this chapter. In the context of asymptotic parametric tests in cross-sectional designs, Davidson et al. (1985) showed that the R-SwE was performing better than the U-SwE. Nevertheless, it seems that no literature exists about the use of the R-SwE for parametric inferences in longitudinal settings or when an F- or a t-distribution is used instead of a χ^2- or a Normal distribution, respectively. This might explain why its use does not seem popular with a parametric test. However, the R-SwE seems to be often considered when a WB procedure is used. This can be explained by the fact that some studies (Flachaire, 2005; Davidson and Flachaire, 2008) made in the context of cross-sectional designs seemed to indicate that combining the R-SwE with the R-WB yields

more accurate inferences than any of the three other possible combinations (U-WB & U-SwE, U-WB & R-SwE and R-WB & U-SwE).

In the computation of the R-SwE, the restricted residuals can also be adjusted like this can be done in the R-WB procedure (see Section 4.2.2). Moreover, while, in the WB literature, the U-SwE and the R-SwE are always considered in their heterogeneous form S^{Het}, we can also consider them in one of their homogeneous form S^{Hom} (see Section 3.2.3).

4.2.4 The WB resampling distribution

For both the U-WB and the R-WB, the resampling is done through the random variables $f_1, f_2, \ldots, f_m$ which, at least, need to respect the two conditions: $\mathbb{E}[f_i] = 0$, $\mathbb{E}[f_i^2] = 1$. Using formal Edgeworth expansions, Davidson and Flachaire (2001) gave some indications that, in order to get accurate inferences in small samples, the ideal resampling distribution of the f_i's should also respect the two additional conditions: $\mathbb{E}[f_i^3] = 1$ and $\mathbb{E}[f_i^4] = 1$. Unfortunately, due to the inequality $\mathbb{E}[f_i^4] \geq 1 + (\mathbb{E}[f_i^3])^2$, no distribution can fulfil these two additional conditions (Davidson et al., 2007; Webb, 2013). Nonetheless, several distributions have been proposed in the literature and the two most popular distributions seems to be the Rademacher and the Mammen distributions. The Rademacher distribution was first mentioned in Liu (1988) and is defined as

$$F_{\text{Rademacher}} : f_i = \begin{cases} -1 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2. \end{cases} \tag{4.7}$$

As shown in Table 4.1, this distribution respects the first, the second and the fourth moments of the ideal distribution, but not the third moment. The Mammen distribution was suggested in Mammen (1993) and is defined as

$$F_{\text{Mammen}} : f_i = \begin{cases} (1+\sqrt{5})/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}) \\ (1-\sqrt{5})/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}). \end{cases} \tag{4.8}$$

As shown in Table 4.1, this distribution respects the first, the second and the third moments of the ideal distribution, but not the fourth.

We see that both $F_{\text{Rademacher}}$ and F_{Mammen} fail to respect the four first moments of the ideal distribution, indicating that we can generally expect some inaccuracies for both in the control of the FPR in small samples. However, the works of Davidson and

Distribution	$\mathbb{E}(f_i)$	$\mathbb{E}(f_i^2)$	$\mathbb{E}(f_i^3)$	$\mathbb{E}(f_i^4)$
F_{Ideal}	0	1	1	1
$F_{\text{Rademacher}}$	0	1	0	1
F_{Mammen}	0	1	1	2
F_{Webb4}	0	1	0	5/4
F_{Webb6}	0	1	0	7/6
$F_{\text{Normal(0,1)}}$	0	1	0	3

Table 4.1 The four first moments of the ideal and candidate resampling distributions.

Flachaire (2001), Flachaire (2005) and Davidson and Flachaire (2008) seem to indicate that $F_{\text{Rademacher}}$ generally yields more accurate results than F_{Mammen} and explains why $F_{\text{Rademacher}}$ seems to be the preferred choice in recent works (see, e.g., Zhu et al., 2007).

While popular, the use of $F_{\text{Rademacher}}$ or F_{Mammen} can be problematic when the number of subjects m is very small. Indeed, they are both two-point distributions, meaning that, when m is small, the maximum number of unique bootstraps cannot be superior to 2^m, which could be rather small. For example, in the case of a group with six subjects, only a maximum of 64 unique WB samples are possible. This is definitively not enough to make accurate inferences. That is the reason why Webb (2013) proposed two alternative resampling distributions with more points in their respective distribution: a four-point and a six-point distributions. The four-point distribution, that we will refer to as F_{Webb4} in this thesis, is

$$F_{\text{Webb4}} : f_i = \begin{cases} -\sqrt{3/2} & \text{with probability } 1/4 \\ -\sqrt{1/2} & \text{with probability } 1/4 \\ \sqrt{1/2} & \text{with probability } 1/4 \\ \sqrt{3/2} & \text{with probability } 1/4. \end{cases} \tag{4.9}$$

The six-point distribution, that we will refer to as F_{Webb6} in this thesis, is

$$F_{\text{Webb6}} : f_i = \begin{cases} -\sqrt{3/2} & \text{with probability } 1/6 \\ -1 & \text{with probability } 1/6 \\ -\sqrt{1/2} & \text{with probability } 1/6 \\ \sqrt{1/2} & \text{with probability } 1/6 \\ 1 & \text{with probability } 1/6 \\ \sqrt{3/2} & \text{with probability } 1/6. \end{cases} \tag{4.10}$$

We directly see that using one of these two distributions will yield a higher number of unique bootstraps. For example, in the case of a group with six subjects, F_{Webb4} should yield a maximum of 4,096 unique bootstraps while F_{Webb6} should yield a maximum of 46,656 unique bootstraps, allowing a less discrete estimation of the null distribution of the original Wald statistic T_0. This can be observed in Figure 4.1 where we considered a balanced design with 2 groups (A and B) of 6 subjects having 5 visits each. We simulated a dataset with a Toeplitz correlation structure with a correlation decrease of 0.1 per visit and used the R-WB combined with the R-SwE S_{C2}^{Hom} to test for a linear effect of visits in Group A alone and for a different linear effect of visits in Group B versus Group A. We can observe that, for the first contrast which involved only 6 subjects (first row in Figure 4.1), the use of $F_{\text{Rademacher}}$ yielded a rather discrete estimate of the Wald statistic null distribution while this was not the case for F_{Webb4} or F_{Webb6}, indicating that one of the these two distributions should be preferred to $F_{\text{Rademacher}}$ in this case. However, for the second contrast which involved 12 subjects (second row in Figure 4.1), the Wald statistic null distribution estimated under $F_{\text{Rademacher}}$ was far less discrete than with the first contrast and seemed relatively similar to the ones estimated under F_{Webb4} or F_{Webb6}, indicating that the discretisation issue of $F_{\text{Rademacher}}$ appears only in very small samples. Note also that, as shown in Table 4.1, F_{Webb4} and F_{Webb6} respect only the first and the second moments of the ideal distribution, but try to mimic $F_{\text{Rademacher}}$ by making the fourth moment close to one. This seems to indicate that, if the sample size is large enough to avoid the issue of having a small number of unique bootstraps, $F_{\text{Rademacher}}$ might yield better results than F_{Webb4} or F_{Webb6}.

Finally, note that we could imagine to use other distributions in practice. For example, we could consider the Normal distribution with mean 0 and variance 1, $F_{\text{Normal}(0,1)}$, which satisfies the first and the second moments of the ideal distribution. However, as shown in Table 4.1, its third and fourth moments are both quite different

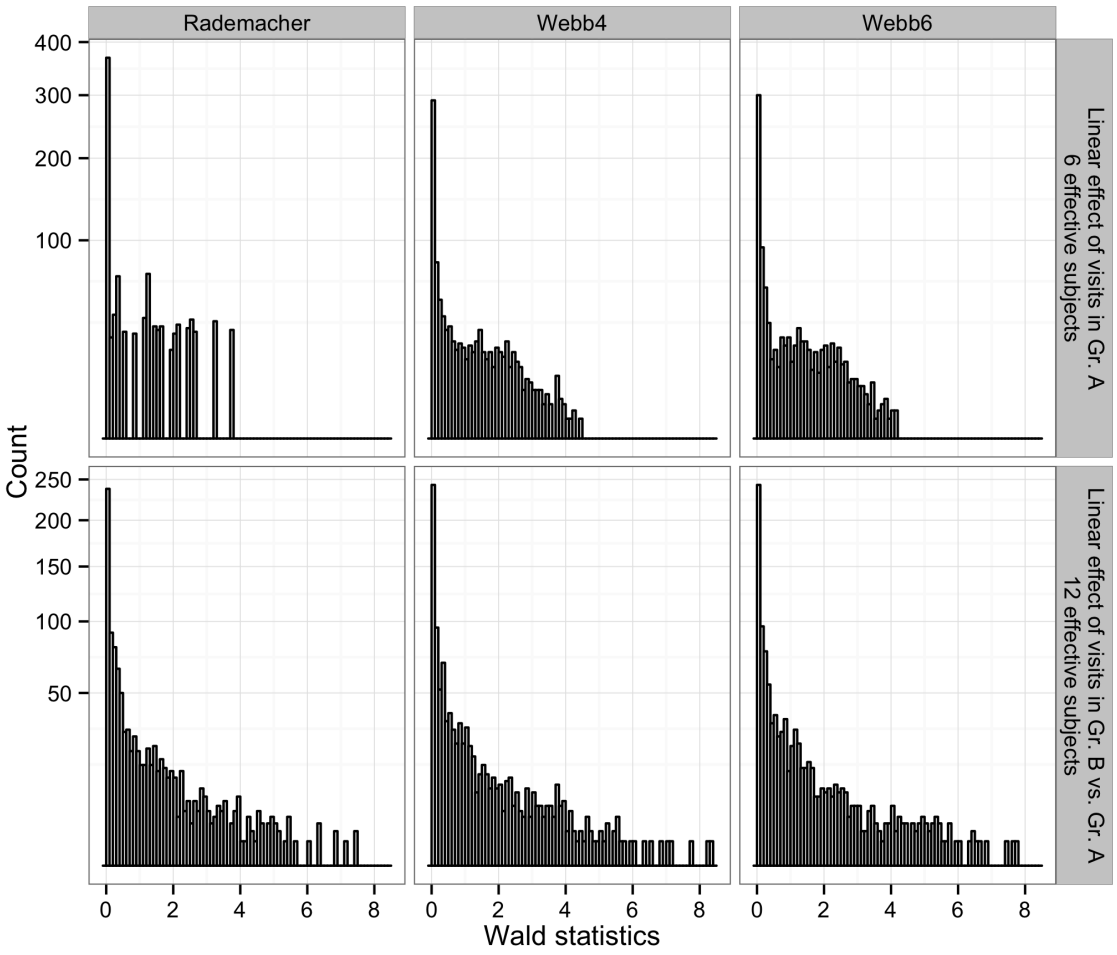

Fig. 4.1 Histograms of the WB Wald statistics obtained with a simulated dataset having a Toeplitz within-subject correlation structure (with a correlation decrease of 0.1 per visit) in a balanced design with 2 groups (A and B) of 6 subjects having 5 visits each. In rows are two different contrasts, while in columns are three different resampling distributions. In each case, the R-WB ($n_{\mathrm{B}} = 999$ bootstraps) combined with the R-SwE $S_{\mathrm{C2}}^{\mathrm{Hom}}$ was used.

from the ones of the ideal distribution, indicating a probable poorer behaviour than with the use of one of the four other distributions mentioned previously.

4.2.5 Multiple testing correction with the WB

To correct for the multiple testing issue mentionned in Section 2.5.2, we could compute a WB p-value at each voxel using Equation (4.3) and, like it was recommended for the parametric tests proposed in Chapter 3, use an FDR correction to make infer-

ence on the p-value image. However, as is generally done with permutation tests in neuroimaging, we can take advantage of the WB procedure to control for the FWER instead of the FDR. To achieve that, it suffices to record, for each bootstrap b, the maximum WB Wald statistic in the score image, $T_b^{\max}$, and use them to estimate the WB maximum statistic null distribution of the score image. As the FWER corresponds to the probability that the maximum statistic in the image is superior or equal to a particular threshold value under the null hypothesis, we can estimate this FWER-corrected threshold or compute the WB FWER-corrected p-values at each voxel using the estimated WB maximum statistic null distribution such that, at each voxel v, we have

$$\frac{1}{(n_{\mathrm{B}}+1)}\sum_{b=0}^{n_{\mathrm{B}}}\mathcal{I}[T_b^{\max}\geq T_0[v]], \tag{4.11}$$

where $T_0[v]$ is the original Wald statistic at voxel v.

Nevertheless, it is worth noting that, within the framework of the SwE, we typically expect that the variability of the Wald statistics under the null hypothesis may vary across the brain. Also, for voxels where the null hypothesis is false, we also may expect large values for the original and some bootstrap Wald statistics. Therefore, the maximum statistic distribution is likely to be driven by such voxels (i.e. with high variability or high departure from the null hypothesis). While this does not affect the validity of the inference, this will typically penalise, in term of power, the other voxels in the brain. This kind of issue has already been reported in the context of permutation tests in Nichols and Holmes (2002). To overcome this issue, Nichols and Holmes (2002) suggested the use of a multi-step approach where we iteratively identify significant voxels, remove them for the next step and reanalyse the remaining voxels through the procedure. This multi-step procedure can also be used in the context of the WB, but, due to its iterative nature, it can be computationally heavy. A more computationally efficient solution to overcome this issue was proposed, also in the context of permutation tests, by Belmonte and Yurgelun-Todd (2001) and discussed in Ridgway (2009). It consists of recording, for each sample, the scores and locations of a small number of voxels with the highest scores. When a voxel is declared significant, we can then replace, for each sample, its eventual contribution to the maximum distribution by the next highest recorded value. A potential issue is that, for some samples, all the recorded voxels could be declared significant, in which case we do not have the next highest score recorded. In this case, an option, implemented in AFNI by Belmonte and discussed in Ridgway (2009), consists of dropping these samples,

assuming that the number of samples left is large enough (e.g., thousands). This second alternative can be seen as an interesting trade-off between the single-step and the multi-step approaches, potentially more powerful than the single-step approach (but, not necessary as powerful as the multi-step approach) while being much more computationally efficient than the multi-step approach.

Furthermore, instead of considering voxel-wise FWER inferences, the WB can also be used, like permutation tests, to make cluster-wise FWER inferences. The procedure is identical to the one used for permutation tests and consists of first thresholding the original and all the bootstrap score images by a primary threshold and, for each thresholded image, recording the size of the largest cluster surviving the thresholding. These maximum cluster sizes can then be used to estimate the maximum cluster size null distribution which can, in turn, be used to make a cluster-wise FWER inference on the original score image. One of the downside of this method is that it is depending of the primary threshold. Also, due to the nature of the SwE method, the Wald statistics are typically expected to have different variability across the brain, meaning that using a common threshold for the whole brain might not be an appropriate way to form the clusters. A solution to this issue would be to homogenise first the Wald statistic images before thresholding them by, for example, using one of the parametric tests proposed in Chapter 3 in order to transform the Wald statistic images into p-value images. The use of a common primary threshold on the p-value images should therefore be more adequate, improving the quality of the inference. Another alternative would be to obtain the p-value images using the WB procedure at every voxel as proposed by Pantazis et al. (2005) in the context of permutation tests. However, the latter strategy could be prohibitively time consuming as it would require two layers of WB, which are by nature already time consuming, or require a large amount of disk space to record the entire set of score images. Another issue that may arise for cluster-wise inferences is the presence of spatial non-stationarity in the smoothness. This non-stationarity is likely to alter the quality of cluster-wise inferences as bigger clusters are expected in smoother areas. A solution to this problem would be to adjust the cluster sizes with a local estimate of smoothness obtained using a first pass of WB as it was proposed by Salimi-Khorshidi et al. (2011) in the context of permutation tests.

4.2.6 Monte-Carlo evaluations

As discussed previously, several WB procedures can be used in practice depending on the choice between the U-WB and the R-WB (see Sections 4.2.1 and 4.2.2), the choice between the U-SwE and the R-SwE (see Section 4.2.3), the choice between S^{Het}

and S^{Hom}, the resampling distribution (see Section 4.2.4), and the small sample bias adjustment used on the residuals (for the resampling and the SwE computation). In the literature, several Monte Carlo evaluations assessing the WB can be found (e.g., Flachaire, 2005; Davidson and Flachaire, 2008; Cameron et al., 2008; Webb, 2013). Nevertheless, they were generally focused on cross-sectional designs or assessed only a small number of WB procedures. In particular, Flachaire (2005) and Davidson and Flachaire (2008) assessed the WB only in the context of cross-sectional designs, Cameron et al. (2008) seemed to only consider the R-WB combined with the U-SwE and $F_{\text{Rademacher}}$ while Webb (2013) seemed to consider only procedures combining the R-WB and the U-SwE. Moreover, all these evaluations considered only the heterogeneous SwE S^{Het} and never the homogeneous SwE S^{Hom}. Therefore, it seems difficult to infer from these studies what would be the best WB procedure to use in the context of longitudinal neuroimaging data and further investigations seems needed.

For this thesis, we conducted several Monte Carlo simulations with the goal to assess and compare 80 WB procedures in some of the scenarios that were investigated in Simulation I of Chapter 3 (see Section 3.2.5). More precisely, the scenarios considered were the same as in Section 3.2.5, except the ones with 200 subjects in the balanced designs and with more than 103 subjects in the unbalanced ADNI designs. The 80 WB procedures differed by the choices between the U-WB & the R-WB, between the U-SwE & the R-SwE, between S^{Hom} & S^{Het}, between S_0 & S_{C2} (used both for the resampling and the SwE), and between the five WB resampling distributions described in Section 4.2.4 (i.e. the Rademacher, Mammen, Webb4, Webb6 and Normal distributions). In addition to these 80 WB procedures, we also used 24 versions of the parametric tests which differed by the choices between the U-SwE & the R-SwE, between S^{Hom} & S^{Het}, between S_0 & S_{C2}, and between Test II, Test III & the asymptotic χ^2- test (see Section 3.2.4 for more details about the parametric tests).

Null and non-null data were generated exactly in the same way as for the simulations of Section 3.2.5 and therefore, we do not repeat the explanation here. For each WB procedure, we use $n_{\text{B}} = 399$ bootstraps. Note that we recommend to use a larger number of bootstraps in a real data analysis (e.g., $n_{\text{B}} = 999$) in order to get a more accurate estimation of the Wald statistic null distribution. Nevertheless, in the case of Monte Carlo simulations, it is not important to have such a large number of bootstraps due to the averaging effect occurring across the 10,000 Monte Carlo realisations that we considered. This would increase unnecessarily the computational time which is already large with $n_{\text{B}} = 399$. Like in Section 3.2.5, we consider 9 contrasts for the balanced designs and 24 contrasts for the unbalanced designs. For each contrast, we

used the 80 WB procedures and the 24 parametric tests to test for significance at 5%. For null data, each significant test was counted as a False Positive and was used to compute the observed FPR of each procedure. For non-null data, each significant test was counted as a True Positive and was used to compute the observed power of each procedure.

4.2.7 Real data analysis

Using an `SPM` toolbox developed for this work, we applied the WB method on the ADNI dataset described in Section 2.6 to test for stronger atrophy visit effects in the AD cohort vs. the Normal cohort. We used one of the best procedure isolated from the Monte Carlo simulation, i.e. the R-WB with the R-SwE $S_{\text{C2}}^{\text{Hom}}$ and $F_{\text{Rademacher}}$. We used 999 WB samples to control for a voxel-wise FWER at 5%.

In order to contrast the obtained results, we also analyse the ADNI dataset using the U-SwE $S_{\text{C2}}^{\text{Hom}}$, the parametric test Test III and the Benjamini-Hochberg procedure (see Section 2.5.2) to control for a voxel-wise FDR at 5%. Finally, also for comparison and using the software package `SPM12`, we analysed the ADNI dataset with the N-OLS and SS-OLS methods using Random Field theory (see Section 2.5.2) to control for a voxel-wise FWER at 5% and using the Benjamini-Hochberg procedure to control of a voxel-wise FDR at 5%.

4.3 Results

In this section, we present the results obtained from the Monte Carlo simulations described in Section 4.2.6 and from the real data analysis described in 4.2.7.

4.3.1 Monte Carlo simulations

The results of the Monte Carlo simulations have been summarised in Figures 4.2, 4.3, 4.4 and 4.5. Due to the large amount of procedure combinations investigated in the simulations (80 for the WB and 24 for the parametric tests), the results are commented below by comparing the different choices separately. Note that, when some procedures were found accurate in terms of FPR control, they did not seem to exhibit any significant differences in terms of power, explaining why, below, we do not show any results regarding the power.

Comparison between the U-WB and R-WB

In the unbalanced ADNI designs (see Figure 4.3), the R-WB clearly outperformed the U-WB which tended to be liberal, particularly in small samples. However, in the balanced designs (see Figure 4.2), while the U-WB combined with the R-SwE performed poorly, the U-WB combined with the U-SwE seemed to perform as well as the R-WB and seemed even to control slightly better the FPR in the scenarios with 12 subjects. Note that, when both the U-WB and the R-WB were accurate, no significant differences in terms of power were observed in the results.

Comparison between the resampling distributions

From Figures 4.2 and 4.3, we can see that the best distributions were clearly the Rademacher, Webb4 and Webb6 distributions. The Mammen distribution yielded liberal inferences when the U-SwE was used and conservative inferences when the R-SwE was used. The Normal distribution yielded liberal inferences in almost all the scenarios. The results did not seem to show strong differences between the Rademacher, Webb4 and Webb6 distributions. Nevertheless, in the smaller samples, some small differences could be observed. The Rademacher distributions seemed to yield slightly more accurate inferences than the Webb6 distribution that, in turn, seemed to yield slightly better inferences than the Webb4 distribution. This can simply be explained by the fact that, among these three resampling distributions, the Rademacher distribution is the closest from the ideal distribution while the Webb4 distribution is the farthest (see Table 4.1). Nevertheless, the results obtained from the Monte Carlo simulations can be misleading when the contrasts tested involved less than 12 subjects (i.e. for the 6 contrasts involving only one group in the balanced designs with a total of 12 subjects and for the 8 contrasts involving only the Normal or the AD subjects in the ADNI designs with a total of 25 subjects). Indeed, in such cases, the Rademacher distribution systematically yielded a very discrete estimate of the statistic null distribution (see, e.g., Figure 4.1), leading to probable bad inferences, but which appeared better in the results due to the averaging effect of the Monte Carlo simulations. Note that, for those scenarios, while the Webb4 and Webb6 distributions had the advantage to yield less discrete estimates of the statistic null distribution (see, e.g., Figure 4.1), they did not seem to yield accurate inferences.

Finally, when the resampling distributions were accurate, no significant differences were observed between them in terms of power.

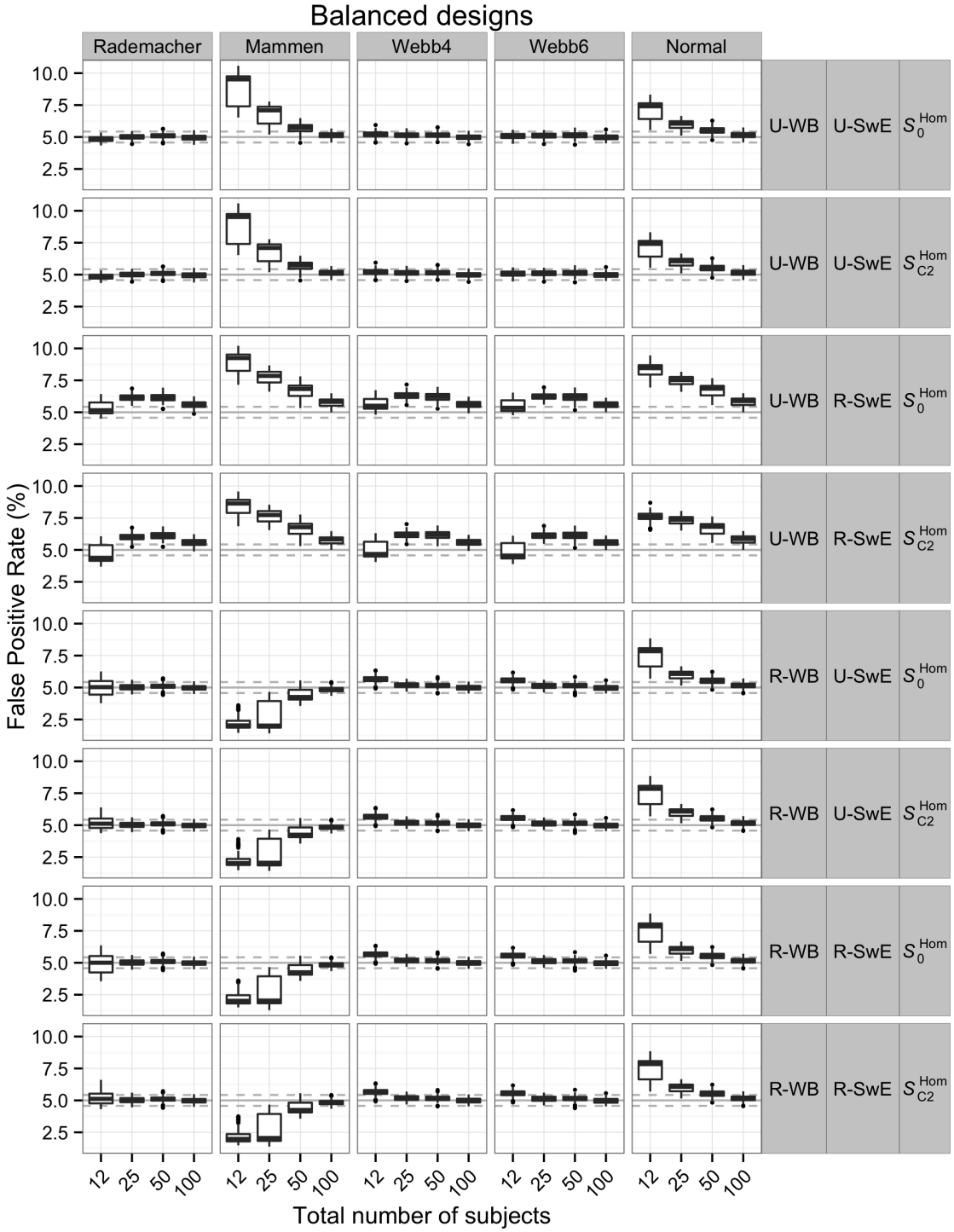

Fig. 4.2 Boxplots showing the FPR control of 40 WB procedures as a function of the total number of subjects in the balanced designs over 162 scenarios (consisting of the 9 contrasts tested, the 6 within-subject covariance structures and the 3 numbers of visits per subject considered in the Monte Carlo simulations). Note that, in these scenarios, the results obtained with the heterogeneous SwE S^{Het} were identical to the ones obtained with the homogeneous SwE and are therefore not shown.

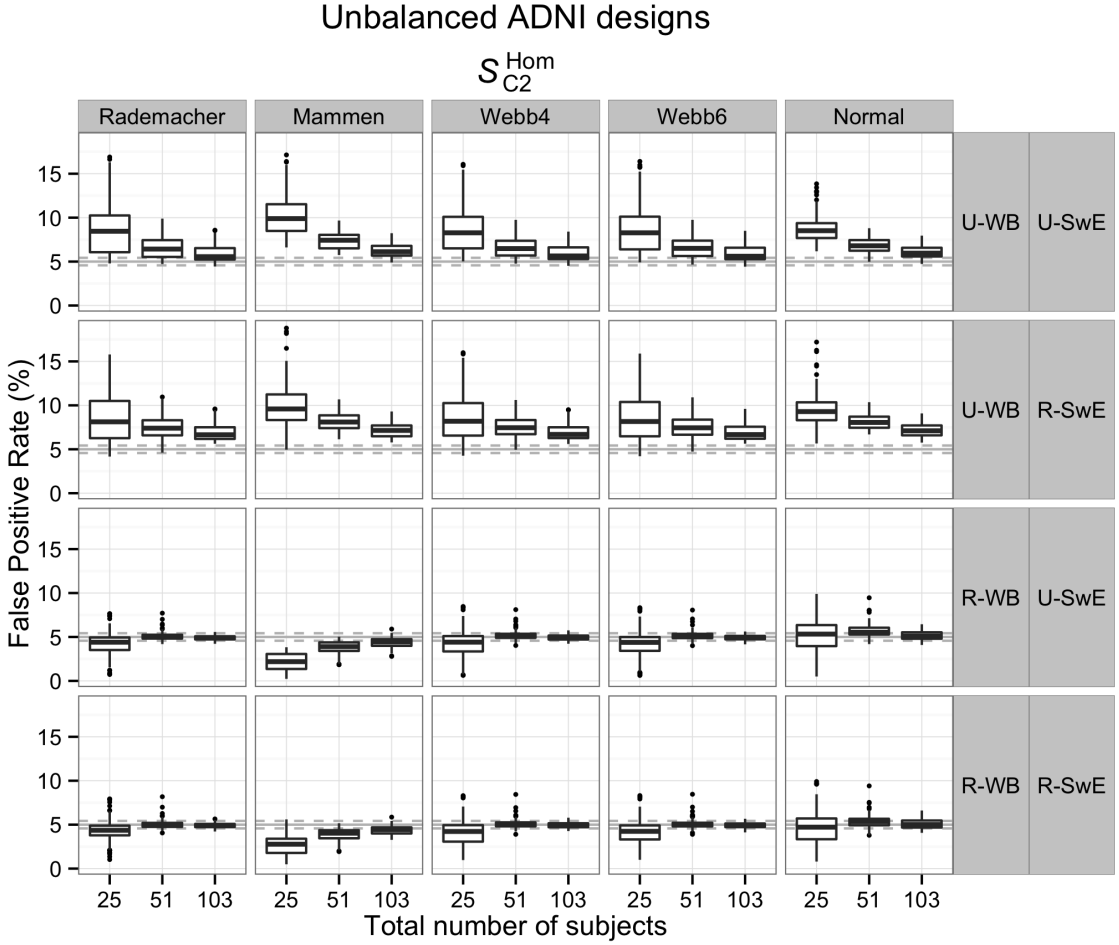

Fig. 4.3 Boxplots showing the FPR control of 20 WB procedures (all using S_{C2}^{Hom}) as a function of the total number of subjects in the unbalanced ADNI designs over 144 scenarios (consisting of the 24 contrasts tested and the 6 within-subject covariance structures considered in the Monte Carlo simulations).

Comparison between the SwE versions used in the WB

In general, when the R-WB was combined with with the Rademacher, Webb4 or Webb6 distributions, no significant differences was observed in the results between the U-SwE and the R-SwE. The only strong differences seemed to occur when the U-WB was used in the balanced designs. In such cases, the U-SwE was clearly better than the R-SwE.

Regarding the choice between the SwE versions S_0^{Het}, S_{C2}^{Het}, S_0^{Hom} and S_{C2}^{Hom}, no compelling differences could be observed in the balanced designs. Nevertheless, in the unbalanced ADNI designs, while, in general, the differences did not appear very strong,

it seemed that S_{C2}^{Hom} was slightly better than the other versions (see, e.g., second and fourth columns in Figure 4.5), explaining also why only this SwE version is shown in Figure 4.3.

Comparison between the WB and the parametric tests

The best WB procedures are compared to some parametric tests in Figures 4.4 and 4.5. We can see that the parametric tests were very sensitive to the choice of SwE and seemed to require the use of the U-SwE S_{C2}^{Hom} or S_{C2}^{Het} in order to be accurate. This was not the case for the WB procedures which seemed almost insensitive to the choice of SwE, indicating a strong robustness of the WB against potential biases existing in the SwE. Nevertheless, the inferences obtained with the parametric test Test III combined with the U-SwE S_{C2}^{Hom} seemed to be as good as those obtained with the best WB procedures or even better in the majority of the small sample scenarios. The only exceptions seemed to be in the scenarios with CS covariance structures in the unbalanced ADNI designs where the WB procedures did not seem affected by the conservativeness observed for the parametric test Test III combined with the U-SwE S_{C2}^{Hom} and yielded more powerful inferences.

Also, as it can be observed in Figure 4.4, the use of the R-SwE in the parametric tests typically yielded conservative inferences which seemed the most accurate when a χ^2-test was used. The explanation for this is that the R-SwE systematically overestimated the true variance of the parameters, making the inferences generally conservative and compensating for the liberal nature of the χ^2-test.

4.3.2 Real data analysis

Figure 4.6 shows the estimated WB null distribution of the maximum statistic and indicates strong evidence that the observed maximum statistic does not occur by chance.

Figure 4.7 shows the score images of the N-OLS, SS-OLS and SwE methods, thresholded after correcting for the FWER and the FDR as described in Section 4.2.7. Note that the score images in Figure 4.7 are not equivalent across methods. In particular, they are all t-score images, except for the SwE method controlling the FDR for which the image is an equivalent Z-score image (needed to homogenise the threshold which is spatially varying for a t-score image when Test III is used). As the score images are not equivalent, they cannot be compared in terms of score values, but they can be in terms of number of voxels surviving the thresholding. These numbers are re-

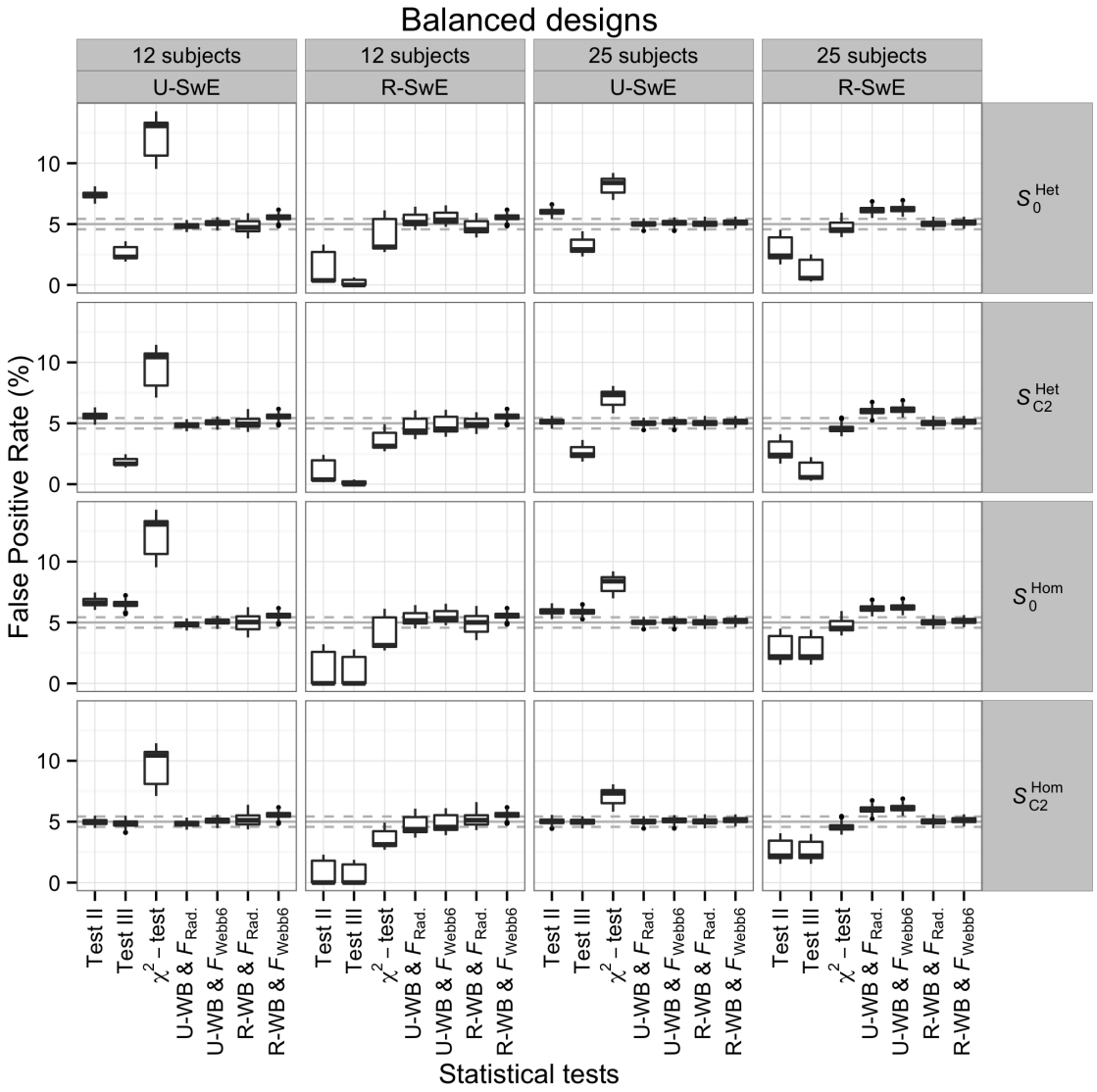

Fig. 4.4 Boxplots comparing the FPR control of some WB procedures and some parametric tests in the balanced designs over 162 scenarios (consisting of the 9 contrasts tested, the 6 within-subject covariance structures and the 3 numbers of visits per subject considered in the Monte Carlo simulations); $F_{\text{Rad.}}$ stands for $F_{\text{Rademacher}}$.

ported in Table 4.2. As is typically expected in neuroimaging, we clearly see that, for all the methods, the use of an FWER-corrected threshold yielded less significant voxels than an FDR-corrected threshold. Like already observed in Section 3.3.3 for an uncorrected threshold, the N-OLS method had more supra-threshold voxels than the SwE method, likely attributable to the presence of a complex (non-CS) longitudinal covariance structure that results in inflated significance (see Figures 3.11 and 3.12,

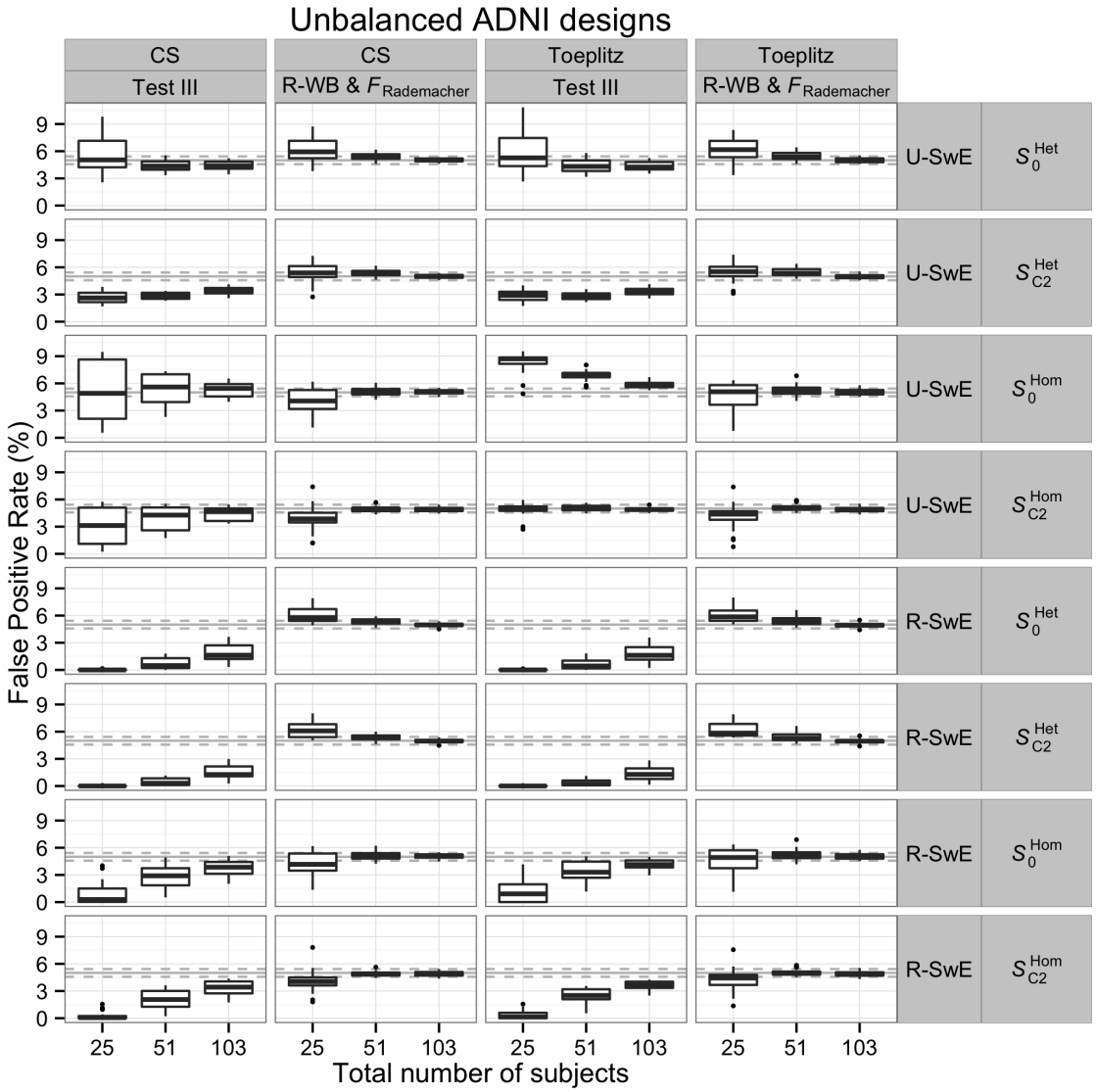

Fig. 4.5 Boxplots comparing the FPR control of the R-WB using the Rademacher resampling distribution and the parametric test Test III in the unbalanced ADNI designs under CS and Toeplitz covariance structures over 24 scenarios (consisting of the 24 contrasts tested in the Monte Carlo simulations).

first row). The SS-OLS had fewer supra-threshold voxels than the SwE method, likely attributable to conservativeness (see Figure 3.12, second row) and/or reduced power (Figure 3.14, sixth row).

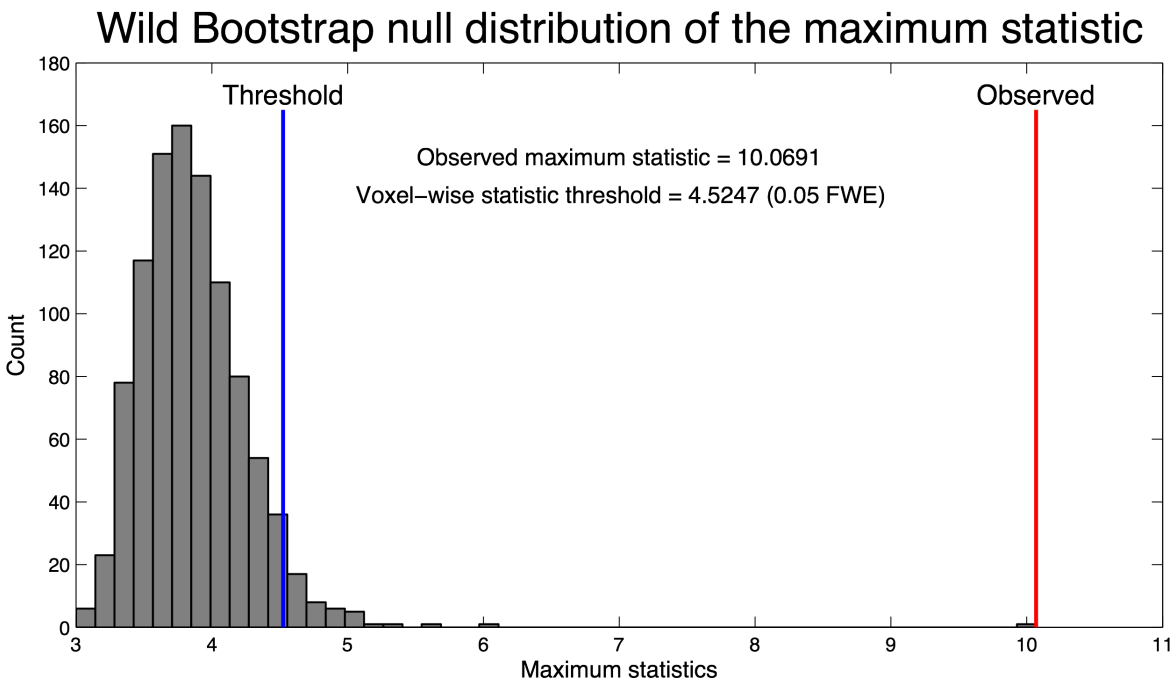

Fig. 4.6 Histogram of the WB null distribution of the maximum statistic obtained using the R-WB combined with the R-SwE $S_{\text{C2}}^{\text{Hom}}$ and the Rademacher distribution ($n_{\text{B}} = 999$ bootstrap samples) on the longitudinal atrophy effect difference (AD vs. N) in the real ADNI dataset.

	N-OLS	SwE	SS-OLS
FDR	110,196	99,951	85,728
FWER	57,515	44,783	32,669

Table 4.2 Number of voxels surviving the FDR and FWER thresholding at 5% significance level after using the parametric N-OLS and SS-OLS methods (both using Random Field Theory for the FWER control) and the SwE method (using the R-WB combined with the R-SwE $S_{\text{C2}}^{\text{Hom}}$, the Rademacher distribution and 999 bootstrap samples to control the FWER, and $S_{\text{C2}}^{\text{Hom}}$ under Test III to control for the FDR). Note that the total number of in-mask voxels was 336,331 voxels for all the methods.

4.4 Conclusion

In this chapter, we have introduced the WB as a resampling method able to make non-parametric inference with the SwE method for the analysis of longitudinal neuroimaging data. We have demonstrated that it can be used as a valid alternative to permutation tests and to the parametric tests developed in Chapter 3.

Using Monte Carlo simulations, we have compared 80 possible variants of the WB procedures and have isolated some of them as the best to use in practice. More pre-

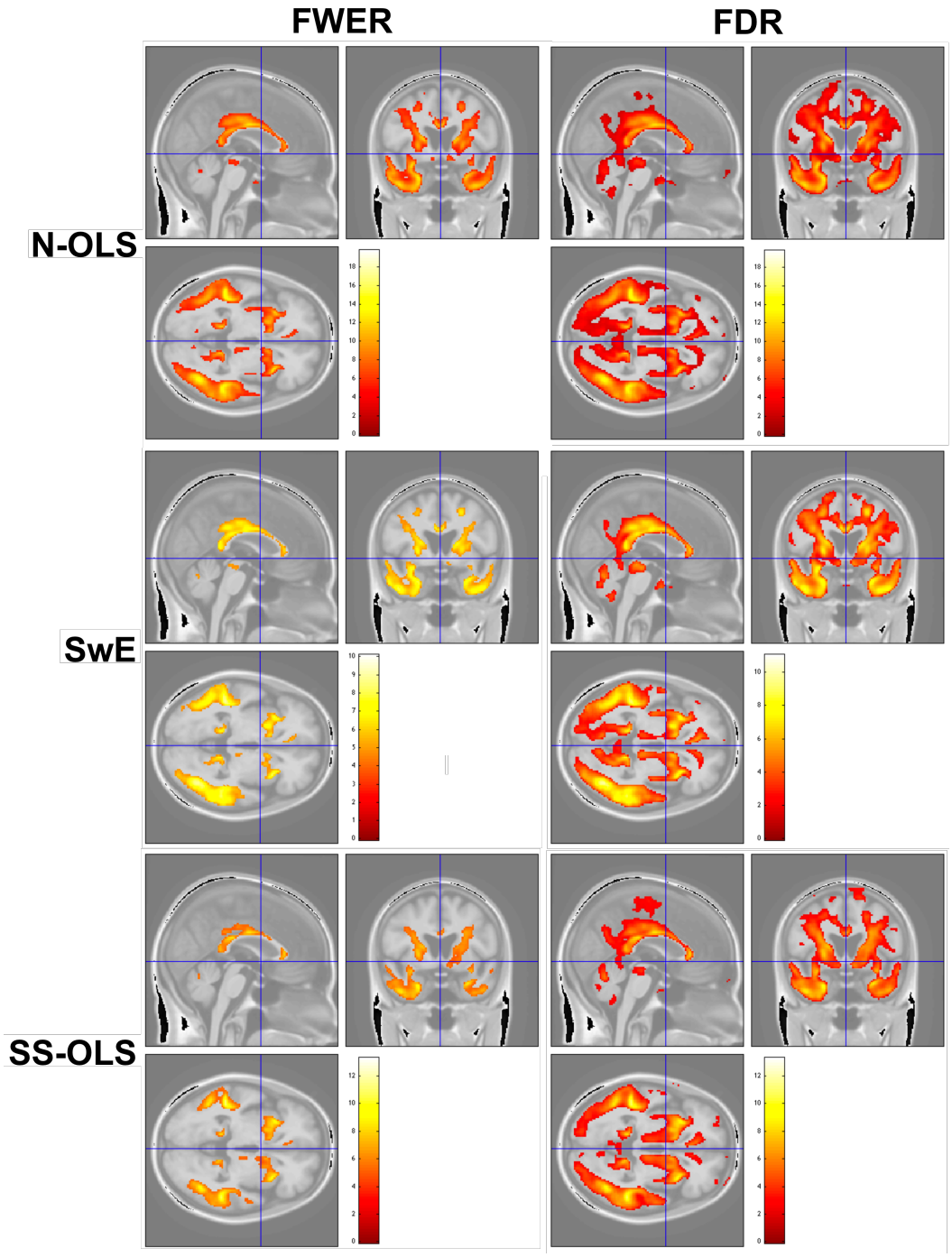

Fig. 4.7 FWER-corrected and FDR-corrected thresholded score images at 5% significance level (centred at the anterior commissure) on the longitudinal atrophy effect difference (AD vs. N) obtained with the parametric N-OLS and SS-OLS methods (both using Random Field Theory for the FWER control), and the SwE method (using the R-WB combined with the R-SwE S_{C2}^{Hom}, the Rademacher distribution and 999 bootstrap samples to control the FWER, and S_{C2}^{Hom} under Test III to control for the FDR). Note that the score images are not equivalent across methods. In particular, they are all t-score images, except for the FDR-thresholded SwE method image which is an equivalent Z-score image.

cisely, it seems that, in general, the R-WB combined with the Rademacher distribution and the SwE S_{C2}^{Hom} should be preferred in practice. Note, however, that we have not found large differences when the other SwE versions were used and have not observed any significant differences between the use of the U-SwE or the R-SWE, indicating that the R-WB was relatively robust against the presence of bias in the SwE. This seems to be an advantage compared to the best parametric tests developed in Chapter 3 which appeared very sensitive to the presence of bias in the SwE. On this, we can expect that the R-WB procedure could also be more robust than the best parametric tests when the error terms cannot be assumed multivariate Normal. In particular, the latter could be interesting for Voxel-Based Morphometry data that is known to be skewed (Viviani et al., 2007; Scarpazza et al., 2013). Nevertheless, further investigation is needed to check this.

We have also used the WB to control for the FWER in the context of a real longitudinal neuroimaging dataset. This is clearly an advantage compared to the parametric tests developed in Chapter 3 which cannot so far be used to control the FWER. A possible way to envision the use of FWER inferences with a parametric test would be to use Random Field Theory. However, further work needs to be done to check if this would be valid in the context of the SwE method and we leave this as a future work.

Nevertheless, some limitations of the WB need to be noted. First, in very small samples (e.g., 6 subjects), the WB procedure using the Rademacher distribution can be inaccurate due to the maximum of number of unique bootstraps which can be rather small and yield a very discrete estimate of the null distribution of the statistics. While we investigated the use of more promising distributions for such cases (i.e. the Webb4 and Webb6 distributions), which yields less discrete estimate of the null distribution, they did not seem so accurate in such small samples, indicating that the WB procedure should probably not be used for contrasts involving less than 12 subjects.

Another issue that can occur with the use of the WB to control for a voxel-wise or a cluster-wise FWER resides in the fact that a spatial heterogeneity of the null distribution of the statistics of interests is likely to exist. For a voxel-wise FWER control, while this should not break the validity of the inference, this might penalise it in terms of power. A solution would be to use the multi-step approach proposed by Nichols and Holmes (2002) or the procedure proposed by Belmonte and Yurgelun-Todd (2001) as described in Section 4.2.5. For a cluster-wise FWER control, this spatial heterogeneity could be more problematic and may challenge the validity of the primary thresholding step, compromising, in turn, the validity of the inference. As suggested in

Section 4.2.5, a solution would be to transform the Wald statistics into p-values (e.g., by using one of the parametric tests developed in Chapter 3) before applying a primary threshold. Another issue for cluster-wise FWER inferences is the spatial heterogeneity of the smoothness that can influence the size of clusters. As discussed in Section 4.2.5, this issue can be addressed using the procedure proposed by Salimi-Khorshidi et al. (2011). Further work would be needed to check the influence of spatial heterogeneity on the WB procedures. However, this seems to be a very challenging task due to the variety of spatial heterogeneities that may exist in longitudinal neuroimaging data.

Chapter 5

The Shrinkage Sandwich Estimator

5.1 Introduction

In the homogeneous SwE S^{Hom} discussed in Section 3.2.3, no structure is assumed for the n_{G} common within-subject covariance matrix V_{0g}'s. Provided that the assumption of a common within-subject covariance matrix within groups is true and if the residuals are appropriately bias corrected, the estimator $\hat{V}_{0g}$ introduced in Section 3.2.3 (see Equations (3.16), (3.17) and (3.18)) has the advantage to be unbiased (i.e. $\mathbb{E}[\hat{V}_{0g}] = V_{0g}$). Thanks to this, the SwE method works well in practice to control the FPR for any type of within-subject covariance structure existing in the data. Nevertheless, in small samples, $\hat{V}_{0g}$ has the disadvantage to carry a lot of estimation error due to its high variability (i.e. $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ is very large). This is somehow not an important issue for the FPR control as the variability of $\hat{V}_{0g}$ is taken into account in the statistical test (see Section 3.2.4). However, this variability may affect quite strongly the power to detect effects, particularly in very small samples. Therefore, instead of using $\hat{V}_{0g}$, we could consider an alternative estimator $\hat{\Theta}_{0g}$ with a lot of structure (e.g., with an imposed compound symmetric structure) which has less variability and consequently carries less estimation error due the variability of the estimator. Unfortunately, if the imposed structure does not hold, this type of estimators can be strongly biased and can consequently exhibit a lot of estimation error due to the bias which, in turn, is likely to affect badly the FPR control of the SwE method. For that reason, in this chapter, instead of using the unstructured estimator $\hat{V}_{0g}$ or a highly structured estimator $\hat{\Theta}_{0g}$, we propose to make a trade-off between both of them and use a shrinkage estimator $\hat{R}_{0g}$ which simply consists of a convex linear combination of both estimators such that

$$\hat{R}_{0g} = \lambda_g \hat{\Theta}_{0g} + (1 - \lambda_g)\hat{V}_{0g}, \tag{5.1}$$

where the structured $\hat{\Theta}_{0g}$ is generally referred to as the target estimator and the shrinkage intensity λ_g can take any values between 0 and 1. In other words, using the typical jargon of the shrinkage estimator literature, we propose to shrink the unstructured estimator $\hat{V}_{0g}$ towards the target $\hat{\Theta}_{0g}$.

This idea of using a shrinkage estimator for the estimation of a covariance matrix has already been proposed in many fields where a better estimate of the covariance matrix is needed. Nevertheless, in the context of the SwE, as far as we are aware of, the use of a shrinkage estimator can only be found in Warton (2011). More precisely, Warton (2011) proposed, in the context of Generalised Estimating Equations (GEE) applied to multivariate abundance data in ecology, to shrink the common within-cluster covariance matrix (only one group considered) towards the within-cluster working covariance matrix. The author also proposed to select an optimal shrinkage intensity using a cross-validation procedure and to make inference using a permutation test procedure. Finally, he demonstrated a substantial gain of power of its methodology compared to the use of a SwE without shrinkage. Unfortunately, the methodology he proposed suffers of several drawbacks that are not really appropriate for the analysis of longitudinal neuroimaging data with the SwE method described in Chapter 3. First, his proposal of using the working covariance matrix as target means that, in our case, we would shrink each $\hat{V}_{0g}$ towards an identity matrix that is likely to be a strongly biased target. Second, the proposed cross-validation procedure to select an optimal shrinkage intensity is computationally intensive and would make the SwE method prohibitively slow in the context of neuroimaging data. Third, the proposed permutation test for inference is also computationally intensive and limited by the assumption of exchangeability under the null hypothesis which is difficult to validate in the general context of longitudinal data. Finally, the evaluation made in Warton (2011) was only focused on the power, not on the essential FPR control, and was performed only in the case of abundance data which is different from longitudinal neuroimaging data.

In this chapter, we propose to modify the methodology proposed by Warton (2011) in order to make it more suitable for longitudinal neuroimaging data. First, we propose the use of other targets, different from identity matrices and more representative of longitudinal data within-subject covariance matrices. Second, instead of using a cross-validation procedure for the selection of an optimal shrinkage intensity, we propose the use of the much faster Ledoit-Wolf procedure (Ledoit and Wolf, 2003) which is non-iterative, but unfortunately more complicated to implement. More precisely, we propose two new SwE versions based on the the Ledoit-Wolf procedure that we will refer to as the Ordinary Ledoit-Wolf Shrinkage SwE (OLWS-SwE) and the Gener-

alised Ledoit-Wolf Shrinkage SwE (GLWS-SwE). Third, for inference, instead of using permutation tests, we propose to extend the much faster parametric tests developed in Section 3.2.4 to account for the shrinkage. Finally, we evaluate the modified methodology using intensive Monte Carlo simulations in settings important for longitudinal neuroimaging studies.

5.2 Methods

In this section, assuming the use of a general target, we first describe the two new SwE versions based on the Ledoit-Wolf procedure to select an optimal shrinkage estimator. Then, for seven specific targets, we develop the specific equations needed to compute the optimal shrinkage intensities. In a third part, we show how the parametric test developed in Section 3.2.4 can be modified to account for the shrinkage. Finally, we describe in details how the evaluations were conducted to assess the proposed methodologies.

5.2.1 The Ordinary Ledoit-Wolf Shrinkage SwE

To find an optimal shrinkage intensity in Equation (5.1), Ledoit and Wolf (2003) proposed to minimise the expectation of a loss function $L_g^{\text{OLW}}[\lambda_g]$ consisting of the square of the Frobenius norm $||\cdot||_{\text{F}}$ of the difference between the shrinkage estimator and the true covariance matrix such that, in our case,

$$L_g^{\text{OLW}}[\lambda_g] = ||\lambda_g \hat{\Theta}_{0g} + (1-\lambda_g)\hat{V}_{0g} - V_{0g}||_{\text{F}}^2. \tag{5.2}$$

Note that we use the superscript "OLW", which stands for "Ordinary Ledoit-Wolf", to contrast with another loss function that is proposed in Section 5.2.2. After minimisation of the expectation of the loss function $L_g^{\text{OLW}}[\lambda_g]$ (which actually corresponds to the Mean Squared Error of the shrinkage estimator in the Frobenius norm sense), Ledoit and Wolf (2003) showed that the optimal shrinkage intensity is, using our notation, obtained by

$$\lambda_g^{\text{OLW}} = \frac{\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \text{var}[(\hat{V}_{0g})_{kk'}] - \text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{V}_{0g})_{kk'}]}{\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \text{var}[(\hat{\Theta}_{0g})_{kk'} - (\hat{V}_{0g})_{kk'}] + \Big(\mathbb{E}[(\hat{\Theta}_{0g})_{kk'}] - (V_{0g})_{kk'}\Big)^2}, \tag{5.3}$$

where n_{0g} is the number of visits in group g. Note that, using the fact that $\mathbb{E}[\cdot^2] = \text{var}[\cdot] + (\mathbb{E}[\cdot])^2$, the terms in the denominator of Equation (5.3) can be replaced by $\mathbb{E}[((\hat{\Theta}_{0g})_{kk'} - (\hat{V}_{0g})_{kk'})^2]$.

As, in practice, the expectations, variances, covariances and the true within-subject covariance matrix used in Equation (5.3) are unknown, we typically replace them by sample estimates of them, leading to an estimator $\hat{\lambda}_g^{\text{OLW}}$. While this is usually not mentioned in the literature, it is worth noting that this implies that $\hat{\lambda}_g^{\text{OLW}}$ cannot be considered as a constant, but as a random variable. This is likely to introduce another bias than the one from the target itself. Indeed, taking the expectation of the shrinkage estimator $\hat{R}_{0g}^{\text{OLW}}$, we obtain

$$\begin{aligned}\mathbb{E}[\hat{R}_{0g}^{\text{OLW}}] &= V_{0g} + \mathbb{E}[\hat{\lambda}_g^{\text{OLW}}(\hat{\Theta}_{0g} - \hat{V}_{0g})] \\ &= V_{0g} + \mathbb{E}[\hat{\lambda}_g^{\text{OLW}}](\mathbb{E}[\hat{\Theta}_{0g}] - V_{0g}) + \text{cov}[\hat{\lambda}_g^{\text{OLW}}, \hat{\Theta}_{0g} - \hat{V}_{0g}], \end{aligned} \tag{5.4}$$

and we can see that two bias terms appear. The first term $\mathbb{E}[\hat{\lambda}_g^{\text{OLW}}](\mathbb{E}[\hat{\Theta}_{0g}] - V_{0g})$ corresponds to the bias introduced by the target and will exist even if $\hat{\lambda}_g^{\text{OLW}}$ is not random. The second term $\text{cov}[\hat{\lambda}_g^{\text{OLW}}, \hat{\Theta}_{0g} - \hat{V}_{0g}]$ is due to the randomness of $\hat{\lambda}_g^{\text{OLW}}$ and would typically disappear only if $\hat{\lambda}_g^{\text{OLW}}$ is not random or if it is independent of $\hat{\Theta}_{0g} - \hat{V}_{0g}$. For a general target, the latter is unlikely to happen, even in the case of an unbiased target. Therefore, it is important to note that, even if we select an unbiased target, the shrinkage estimator may be biased. Finally, another important remark is that $\hat{\lambda}_g^{\text{OLW}}$ is likely to be a biased estimator of λ_g^{OLW}. Unfortunately, due to the non-linear form of $\hat{\lambda}_g^{\text{OLW}}$, it seems difficult to give an exact expression for this bias.

Replacing each $\hat{V}_{0g}$ by their corresponding OLW shrinkage estimators $\hat{R}_{0g}^{\text{OLW}}$ in the computation of the homogeneous SwE S^{Hom} (see Section 3.2.3) yields a new type of SwE that we will refer to as the Ordinary Ledoit-Wolf Shrinkage SwE (OLWS-SwE).

5.2.2 The Generalised Ledoit-Wolf Shrinkage SwE

As described in Section 5.2.1, the Ordinary Ledoit-Wolf procedure aims to reduce the Mean Squared Error (MSE) of the covariance matrix estimators used in the SwE. However, even if the MSE of the covariance matrix estimators is reduced, it does not mean that the one of the SwE S or the one of a contrasted version of it, $CSC^\top$, will be reduced as well. Therefore, instead of attempting to select shrinkage intensities which minimise the MSE of each $\hat{R}_{0g}$, it seems more adequate to select shrinkage intensities which minimise the MSE of S or $CSC^\top$. To achieve this, we can define, for each group

of subjects with a common covariance, an alternative loss function $L_g^{\text{GLW}}[\lambda_g]$ consisting of the square of the Frobenius norm of the difference between the group contribution to the contrasted shrinkage SwE and the one of the true contrasted covariance matrix of the parameters such that

$$L_g^{\text{GLW}}[\lambda_g] = ||\lambda_g C S_{\Theta g} C^\top + (1-\lambda_g) C S_g C^\top - C\text{Cov}(\hat{\beta})_g C^\top||_{\text{F}}^2, \tag{5.5}$$

where S_g, $S_{\Theta g}$ and $\text{Cov}(\hat{\beta})_g$ are the contributions of group g to the SwE obtained using $\hat{V}_{0g}$, $\hat{\Theta}_{0g}$ and V_{0g}, respectively. Note that, if we were interested to only minimise the MSE of the SwE S, it suffices to take the identity matrix as contrast matrix C. Minimising the expectation of this loss function, we obtain an optimal shrinkage intensity given by

$$\lambda_g^{\text{GLW}} = \frac{\sum_{k=1}^{q}\sum_{k'=1}^{q} \text{var}[(CS_gC^\top)_{kk'}] - \text{cov}[(CS_{\Theta g}C^\top)_{kk'}, (CS_gC^\top)_{kk'}]}{\sum_{k=1}^{q}\sum_{k'=1}^{q} \text{var}[(CS_{\Theta g}C^\top)_{kk'} - (CS_gC^\top)_{kk'}] + (\mathbb{E}[(CS_{\Theta g}C^\top)_{kk'}] - (C\text{Cov}(\hat{\beta})_g C^\top)_{kk'})^2}. \tag{5.6}$$

Using the fact that $\text{vec}[CS_gC^\top] = G_g\text{vec}[\hat{V}_{0g}]$ (see Equation (3.43)), we can rewrite Equation (5.6) such that

$$\lambda_g^{\text{GLW}} = \frac{\text{tr}[G_g\text{Cov}[\text{vec}[\hat{V}_{0g}], \text{vec}[\hat{V}_{0g} - \hat{\Theta}_{0g}]]G_g^\top]}{\text{tr}[G_g\text{Cov}[\text{vec}[\hat{V}_{0g} - \hat{\Theta}_{0g}]]G_g^\top] + ||G_g\text{vec}[\mathbb{E}[\hat{\Theta}_{0g}] - V_{0g}]||_{\text{E}}^2}, \tag{5.7}$$

where $||\cdot||_{\text{E}}$ is the Euclidean norm.

Note that, if we replace G_g by the identity matrix in Equation (5.7), we retrieve the OLW optimal shrinkage intensity as given in Equation (5.3). Therefore, this alternative optimal shrinkage intensity can be seen as a generalisation of the OLW optimal shrinkage intensity, explaining why, in this thesis, we refer to it as the Generalised Ledoit-Wolf (GLW) shrinkage intensity.

Similarly to the OLWS-SwE, in practice, we simply replace the covariances and expectations in Equation (5.7) by sample estimates of them, leading to the estimator $\hat{\lambda}_g^{\text{GLW}}$, which can be used in Equation (5.1) to compute the shrinkage estimator $\hat{R}_{0g}^{\text{GLW}}$ which, in turn, can be used to obtain a new version of the SwE that we will refer to as the Generalised Ledoit-Wolf Shrinkage SwE (GLWS-SwE).

5.2.3 Choice of the target matrix

Ideally, we should select a target with low variability (to decrease the estimation error due to the variability), but also with low bias (to minimise the error due to the bias). Unfortunately, as these two properties are generally antagonistic, the choice of a target may be difficult in practice and may depend on the data we want to analyse. Here, inspired by the work in Schäfer and Strimmer (2005), we review seven popular target choices (see Table 5.1) and give, for each of them, the necessary formulas to ease the computation of the corresponding OLWS-SwE and the GLWS-SwE. For six of these targets, this work has already been done for $\hat{\lambda}_g^{\text{OLW}}$ in Schäfer and Strimmer (2005). However, as mentioned in Kwan (2011), in the denominator of Equation (5.3), Schäfer and Strimmer (2005) generally assumed that the term $\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \text{var}[(\hat{\Theta}_{0g})_{kk'} - (\hat{V}_{0g})_{kk'}] = 0$, which is unlikely to be true in practice. Moreover, for some targets, they made the additional assumption that the term at the numerator $\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{V}_{0g})_{kk'}] = 0$ and, finally, for one of the targets, consisting of heterogeneous variances and homogeneous correlations, they assumed that the homogeneous correlations are not random, which is also not true. Therefore, in this section, in addition to give the formulas for $\hat{\lambda}_g^{\text{OLW}}$ derived in Schäfer and Strimmer (2005), we also rederive them without making the same simplifications. For the computation of the shrinkage intensities $\hat{\lambda}_g^{\text{GLW}}$, we do not provide the exact formulas as they can be very complicated to write down. However, we provide the formulas expressing the elements of $\text{Cov}[\text{vec}[\hat{\Theta}_{0g}]]$ and $\text{Cov}[\text{vec}[\hat{V}_{0g}], \text{vec}[\hat{\Theta}_{0g}]]$ as a function of the elements of V_{0g} and $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$. These formulas can then be used to express the two terms $\text{Cov}[\text{vec}[\hat{V}_{0g} - \hat{\Theta}_{0g}]]$ and $\text{Cov}[\text{vec}[\hat{V}_{0g}], \text{vec}[\hat{V}_{0g} - \hat{\Theta}_{0g}]]$ present in Equation (5.7) as a function of the elements of V_{0g} and $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$.

Note that the equations provided in this section are given as a function of V_{0g} and $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ or sample estimates of them without specifying how these can actually be obtained. In practice, V_{0g} may be estimated as described in Section 3.2.3 using Equations (3.16), (3.17) and (3.18) while $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ can be estimated using either the estimator developed for Test II or the one developed for Test III (see Section 3.2.4, Equations (3.54) and (3.69), respectively).

		$(\hat{\Theta}_{0g})_{kk'}$	
Target name	Description	$k = k'$	$k \neq k'$
Target A	Identity matrix	1	0
Target B	Hom. var. & no corr.	$\frac{1}{n_{0g}} \sum_{k=1}^{n_{0g}} (\hat{V}_{0g})_{kk}$	0
Target C	Hom. var. & hom. corr.	$\frac{1}{n_{0g}} \sum_{k=1}^{n_{0g}} (\hat{V}_{0g})_{kk}$	$\frac{2}{n_{0g}(n_{0g}-1)} \sum_{k=1}^{n_{0g}} \sum_{k'>k}^{n_{0g}} (\hat{V}_{0g})_{kk'}$
Target D	Het. var. & no corr.	$(\hat{V}_{0g})_{kk}$	0
Target E	Het. var. & perfect positive corr.	$(\hat{V}_{0g})_{kk}$	$\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}$
Target F	Het. var. & hom. corr.	$(\hat{V}_{0g})_{kk}$	$\hat{\rho}\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}$
Target G	Hom. var. & perfect positive corr.	$\frac{1}{n_{0g}} \sum_{k=1}^{n_{0g}} (\hat{V}_{0g})_{kk}$	$\frac{1}{n_{0g}} \sum_{k=1}^{n_{0g}} (\hat{V}_{0g})_{kk}$

Table 5.1 Popular targets for covariance matrices. The labelling of the targets corresponds to the one used in Schäfer and Strimmer (2005), except for Target G which was not investigated therein. "Het.", "hom.", "var." and "corr." stand for "heterogeneous", "homogeneous", "variances' and "correlations", respectively. The expression for $\hat{\rho}$ is given by Equation (5.37).

Target A: the identity matrix

The first target considered in Schäfer and Strimmer (2005) was very simple and consisted of the identity matrix, i.e. with

$$(\hat{\Theta}_{0g})_{kk'} = \begin{cases} 1 & \text{if } k = k', \\ 0 & \text{if } k \neq k', \end{cases} \tag{5.8}$$

and, in this case, the optimal Schäfer-Strimmer OLW shrinkage intensity was given by

$$\hat{\lambda}_g^{\text{OLW-SS}} = \frac{\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{V}_{0g})_{kk'}^2 + \sum_{k=1}^{n_{0g}} (1 - (\hat{V}_{0g})_{kk})^2}. \tag{5.9}$$

Accounting for the variance of each term $(\hat{\Theta}_{0g})_{kk'} - (\hat{V}_{0g})_{kk'}$, the "correct" OLW shrinkage intensity is given by

$$\hat{\lambda}_g^{\text{OLW-C}} = \frac{\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}] + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{V}_{0g})_{kk'}^2 + \sum_{k=1}^{n_{0g}} (1 - (\hat{V}_{0g})_{kk})^2}. \tag{5.10}$$

Comparing Equation (5.9) with Equation (5.10), we directly see that, as the term $\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]$ will be always superior to 0, the Schäfer-Strimmer OLW shrinkage intensity will tend to overestimate the correct OLW shrinkage intensity as defined by Ledoit and Wolf (2003).

Finally, as the target is not random, we simply get

$$\text{Cov}[\text{vec}[\hat{\Theta}_{0g}]] = \text{Cov}[\text{vec}[\hat{V}_{0g}], \text{vec}[\hat{\Theta}_{0g}]] = 0. \tag{5.11}$$

Target B: homogeneous variances and no correlation

The second target considered in Schäfer and Strimmer (2005) consisted of the diagonal matrix with a common variance, i.e. with

$$(\hat{\Theta}_{0g})_{kk'} = \begin{cases} \hat{v} = \dfrac{1}{n_{0g}} \displaystyle\sum_{k=1}^{n_{0g}} (\hat{V}_{0g})_{kk} & \text{if } k = k', \\ 0 & \text{if } k \neq k', \end{cases} \tag{5.12}$$

and, in this case, the optimal Schäfer-Strimmer OLW shrinkage intensity was given by

$$\hat{\lambda}_g^{\text{OLW-SS}} = \frac{\displaystyle\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{\displaystyle\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{V}_{0g})_{kk'}^2 + \sum_{k=1}^{n_{0g}} (\hat{v} - (\hat{V}_{0g})_{kk})^2}. \tag{5.13}$$

For this target, in addition to assuming that $\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \text{var}[(\hat{\Theta}_{0g})_{kk'} - (\hat{V}_{0g})_{kk'}] = 0$ in the denominator of Equation (5.3), Schäfer and Strimmer (2005) also assumed that $\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{V}_{0g})_{kk'}] = 0$ in the numerator. Avoiding these two simplifications, we get instead

$$\hat{\lambda}_g^{\text{OLW-C}} = \frac{-\hat{a}_1 + \displaystyle\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{-\hat{a}_1 + \displaystyle\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}] + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{V}_{0g})_{kk'}^2 + \sum_{k=1}^{n_{0g}} (\hat{v} - (\hat{V}_{0g})_{kk})^2} \tag{5.14}$$

where

$$\hat{a}_1 = \frac{1}{n_{0g}} \sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\text{cov}}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{k'k'}]. \tag{5.15}$$

In this case, only the diagonal elements of the target are random and are all the same. Thus, for all $k, k', l, l' = 1, \ldots, n_{0g}$, we have

$$\operatorname{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] = \begin{cases} 0 & \text{if } k \neq k' \text{ or } l \neq l' \\ \dfrac{1}{n_{0g}^2} \displaystyle\sum_{i=1}^{n_{0g}} \sum_{j=1}^{n_{0g}} \operatorname{cov}[(\hat{V}_{0g})_{ii}, (\hat{V}_{0g})_{jj}] & \text{if } k = k' \text{ and } l = l', \end{cases} \tag{5.16}$$

$$\operatorname{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] = \begin{cases} 0 & \text{if } l \neq l', \\ \dfrac{1}{n_{0g}} \displaystyle\sum_{i=1}^{n_{0g}} \operatorname{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ii}] & \text{if } l = l'. \end{cases} \tag{5.17}$$

Target C: homogeneous variances and homogeneous covariances

The third target considered in Schäfer and Strimmer (2005) consisted of a matrix with common variances and common covariances, i.e. with

$$(\hat{\Theta}_{0g})_{kk'} = \begin{cases} \hat{v} = \dfrac{1}{n_{0g}} \displaystyle\sum_{k=1}^{n_{0g}} (\hat{V}_{0g})_{kk} & \text{if } k = k', \\ \hat{c} = \dfrac{2}{n_{0g}(n_{0g}-1)} \displaystyle\sum_{k=1}^{n_{0g}} \sum_{k'>k}^{n_{0g}} (\hat{V}_{0g})_{kk'} & \text{if } k \neq k', \end{cases} \tag{5.18}$$

and, in this case, the optimal Schäfer-Strimmer OLW shrinkage intensity was given by

$$\hat{\lambda}_g^{\text{OLW-SS}} = \frac{\displaystyle\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}]}{\displaystyle\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{c} - (\hat{V}_{0g})_{kk'})^2 + \sum_{k=1}^{n_{0g}} (\hat{v} - (\hat{V}_{0g})_{kk})^2}. \tag{5.19}$$

Avoiding the simplifications made in the numerator and the denominator in Schäfer and Strimmer (2005), we get instead

$$\hat{\lambda}_g^{\text{OLW-C}} = \frac{-\hat{a}_1 - \hat{a}_2 + \displaystyle\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}]}{-\hat{a}_1 - \hat{a}_2 + \displaystyle\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}] + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{c} - (\hat{V}_{0g})_{kk'})^2 + \sum_{k=1}^{n_{0g}} (\hat{v} - (\hat{V}_{0g})_{kk})^2} \tag{5.20}$$

where

$$\hat{a}_2 = \frac{2}{n_{0g}(n_{0g}-1)} \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \sum_{l=1}^{n_{0g}} \sum_{l'>l}^{n_{0g}} \widehat{\text{cov}}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}]. \tag{5.21}$$

For all $k, k', l, l' = 1, \ldots, n_{0g}$, we have

$$\text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] = \begin{cases} \frac{4}{n_{0g}^2(n_{0g}-1)^2} \sum_{i=1}^{n_{0g}} \sum_{i'>i}^{n_{0g}} \sum_{j=1}^{n_{0g}} \sum_{j'>j}^{n_{0g}} \text{cov}[(\hat{V}_{0g})_{ii'}, (\hat{V}_{0g})_{jj'}] & \text{if } k \neq k' \text{ and } l \neq l', \\ \frac{1}{n_{0g}^2} \sum_{i=1}^{n_{0g}} \sum_{j=1}^{n_{0g}} \text{cov}[(\hat{V}_{0g})_{ii}, (\hat{V}_{0g})_{jj}] & \text{if } k = k' \text{ and } l = l', \\ \frac{2}{n_{0g}^2(n_{0g}-1)} \sum_{i=1}^{n_{0g}} \sum_{i'>i}^{n_{0g}} \sum_{j=1}^{n_{0g}} \text{cov}[(\hat{V}_{0g})_{ii'}, (\hat{V}_{0g})_{jj}] & \text{otherwise,} \end{cases} \tag{5.22}$$

$$\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] = \begin{cases} \frac{2}{n_{0g}(n_{0g}-1)} \sum_{i=1}^{n_{0g}} \sum_{i'>i}^{n_{0g}} \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ii'}] & \text{if } l \neq l', \\ \frac{1}{n_{0g}} \sum_{i=1}^{n_{0g}} \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ii}] & \text{if } l = l'. \end{cases} \tag{5.23}$$

Target D: heterogeneous variances and no correlation

The fourth target considered in Schäfer and Strimmer (2005) consisted of a diagonal matrix with heterogeneous variances such that

$$(\hat{\Theta}_{0g})_{kk'} = \begin{cases} (\hat{V}_{0g})_{kk'} & \text{if } k = k', \\ 0 & \text{if } k \neq k', \end{cases} \tag{5.24}$$

and, in this case, the optimal Schäfer-Strimmer OLW shrinkage intensity was given by

$$\hat{\lambda}_g^{\text{OLW-SS}} = \frac{\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{V}_{0g})_{kk'}^2}. \tag{5.25}$$

Avoiding the simplification made in the denominator in Schäfer and Strimmer

(2005), we get instead

$$\hat{\lambda}_g^{\text{OLW-C}} = \frac{\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}] + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{V}_{0g})_{kk'}^2}. \tag{5.26}$$

For all $k, k', l, l' = 1, \ldots, n_{0g}$, we have

$$\text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'})] = \begin{cases} 0 & \text{if } k \neq k' \text{ or } l \neq l', \\ \text{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{ll}] & \text{if } k = k' \text{ and } l = l', \end{cases} \tag{5.27}$$

$$\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] = \begin{cases} 0 & \text{if } l \neq l', \\ \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll}] & \text{if } l = l'. \end{cases} \tag{5.28}$$

Target E: heterogeneous variances and perfect positive correlation

The fifth target considered in Schäfer and Strimmer (2005) consisted of a matrix with heterogeneous variances and correlations equal to one such that

$$(\hat{\Theta}_{0g})_{kk'} = \begin{cases} (\hat{V}_{0g})_{kk'} & \text{if } k = k', \\ \sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}} & \text{if } k \neq k', \end{cases} \tag{5.29}$$

and, in this case, the optimal Schäfer-Strimmer OLW shrinkage intensity was given by

$$\hat{\lambda}_g^{\text{OLW-SS}} = \frac{-\hat{a}_3 + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{\sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \left(\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}} - (\hat{V}_{0g})_{kk'} \right)^2} \tag{5.30}$$

where

$$\hat{a}_3 = \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \sqrt{\frac{(\hat{V}_{0g})_{k'k'}}{(\hat{V}_{0g})_{kk}}} \widehat{\text{cov}}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{kk'}]. \tag{5.31}$$

Note that, in order to estimate the covariances $\text{cov}\left[\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}, (\hat{V}_{0g})_{kk'}\right]$, Schäfer and Strimmer (2005) used the delta method which is based on the approximation of $\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}$ by a first-order Taylor series around the point estimates

of $(\hat{V}_{0g})_{kk}$ and $(\hat{V}_{0g})_{k'k'}$.

Avoiding the simplification made in the denominator in Schäfer and Strimmer (2005) and using the delta method to estimate the variances $\operatorname{var}\left[\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}\right]$, we get instead

$$\hat{\lambda}_g^{\text{OLW-C}} = \frac{-\hat{a}_3 + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}]}{\hat{a}_4 - 2\hat{a}_3 + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}] + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \left(\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}} - (\hat{V}_{0g})_{kk'}\right)^2}, \tag{5.32}$$

where

$$\hat{a}_4 = \frac{1}{2} \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \left(\frac{(\hat{V}_{0g})_{k'k'}}{(\hat{V}_{0g})_{kk}} \widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk}] + \widehat{\operatorname{cov}}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{k'k'}] \right). \tag{5.33}$$

Using the delta method, we get, for all $k, k', l, l' = 1, \ldots, n_{0g}$,

$$\operatorname{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] \approx \begin{cases} \frac{1}{4} \left(\sqrt{\frac{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{ll}}} \operatorname{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{ll}] \right. \\ + \sqrt{\frac{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{ll}}} \operatorname{cov}[(\hat{V}_{0g})_{k'k'}, (\hat{V}_{0g})_{ll}] \\ + \sqrt{\frac{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{l'l'}}} \operatorname{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{l'l'}] \\ \left. + \sqrt{\frac{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{l'l'}}} \operatorname{cov}[(\hat{V}_{0g})_{k'k'}, (\hat{V}_{0g})_{l'l'}] \right) & \text{if } k \neq k' \text{ and } l \neq l', \\ \frac{1}{2} \left(\sqrt{\frac{(\hat{V}_{0g})_{k'k'}}{(\hat{V}_{0g})_{kk}}} \operatorname{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{ll}] \right. \\ \left. + \sqrt{\frac{(\hat{V}_{0g})_{kk}}{(\hat{V}_{0g})_{k'k'}}} \operatorname{cov}[(\hat{V}_{0g})_{k'k'}, (\hat{V}_{0g})_{ll}] \right) & \text{if } k \neq k' \text{ and } l = l' \\ \frac{1}{2} \left(\sqrt{\frac{(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{ll}}} \operatorname{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{ll}] \right. \\ \left. + \sqrt{\frac{(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{l'l'}}} \operatorname{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{l'l'}] \right) & \text{if } k = k' \text{ and } l \neq l' \\ \operatorname{cov}[(\hat{V}_{0g})_{kk}, (\hat{V}_{0g})_{ll}] & \text{if } k = k' \text{ and } l = l', \end{cases} \tag{5.34}$$

$$\operatorname{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] \approx \begin{cases} \frac{1}{2}\left(\sqrt{\frac{(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{ll}}}\operatorname{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll}] \right. \\ \left. +\sqrt{\frac{(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{l'l'}}}\operatorname{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{l'l'}]\right) & \text{if } l \neq l', \\ \operatorname{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll}] & \text{if } l = l'. \end{cases} \tag{5.35}$$

Target F: heterogeneous variances and homogeneous correlations

The last target considered in Schäfer and Strimmer (2005) consisted of a matrix with heterogeneous variances and homogeneous correlations such that

$$(\hat{\Theta}_{0g})_{kk'} = \begin{cases} (\hat{V}_{0g})_{kk'} & \text{if } k = k' \\ \hat{\rho}\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}} & \text{if } k \neq k', \end{cases} \tag{5.36}$$

where

$$\hat{\rho} = \frac{2}{n_{0g}(n_{0g}-1)} \sum_{k=1}^{n_{0g}} \sum_{k'>k}^{n_{0g}} \left(\frac{(\hat{V}_{0g})_{kk'}}{\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}} \right). \tag{5.37}$$

For this target, the optimal Schäfer-Strimmer OLW shrinkage intensity was given by

$$\hat{\lambda}_g^{\text{OLW-SS}} = \frac{-\hat{\rho}\hat{a}_3 + \sum_{k=1}^{n_{0g}} \sum_{k'\neq k}^{n_{0g}} \widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}]}{\sum_{k=1}^{n_{0g}} \sum_{k'\neq k}^{n_{0g}} \left(\hat{\rho}\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}} - (\hat{V}_{0g})_{kk'}\right)^2}. \tag{5.38}$$

For this target, Schäfer and Strimmer (2005) assumed again that the denominator term $\sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \operatorname{var}[(\hat{\Theta}_{0g})_{kk'} - (\hat{V}_{0g})_{kk'}] = 0$ and used the delta method to estimate the covariances $\operatorname{cov}\left[\hat{\rho}\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}, (\hat{V}_{0g})_{kk'}\right]$, but assumed that $\hat{\rho}$ is not random. Note that Kwan (2008) modified the Equation (5.38) by accounting for the randomness of $\hat{\rho}$ in the delta method, but did not correct for the simplification at the denominator of Equation (5.3). Here, using the correction proposed by Kwan (2008) and avoiding

the simplification made at the denominator, we get

$$\hat{\lambda}_g^{\text{OLW-C}} = \frac{-\hat{\rho}\hat{a}_3 - \hat{a}_5 + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \widehat{\text{var}}[(\hat{V}_{0g})_{kk'}]}{\hat{a}_6 + \hat{\rho}^2 \hat{a}_4 - 2\hat{\rho}\hat{a}_3 - 2\hat{a}_5 + \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \left(\widehat{\text{var}}[(\hat{V}_{0g})_{kk'}] + \left(\hat{\rho} \sqrt{(\hat{V}_{0g})_{kk} (\hat{V}_{0g})_{k'k'}} - (\hat{V}_{0g})_{kk'} \right)^2 \right)} \tag{5.39}$$

where

$$\begin{aligned}
\hat{a}_5 =& \frac{1}{n_{0g}(n_{0g}-1)} \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \sum_{l=1}^{n_{0g}} \sum_{l' \neq l}^{n_{0g}} \sqrt{\frac{(\hat{V}_{0g})_{kk} (\hat{V}_{0g})_{k'k'}}{(\hat{V}_{0g})_{ll} (\hat{V}_{0g})_{l'l'}}} \Big(\widehat{\text{cov}}[(\hat{V}_{0g})_{ll'}, (\hat{V}_{0g})_{kk'}] \\
& - \frac{(\hat{V}_{0g})_{ll'}}{(\hat{V}_{0g})_{ll}} \widehat{\text{cov}}[(\hat{V}_{0g})_{ll}, (\hat{V}_{0g})_{kk'}] \Big), \qquad (5.40) \\
\hat{a}_6 =& \widehat{\text{var}}[\hat{\rho}] \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} (\hat{V}_{0g})_{kk} (\hat{V}_{0g})_{k'k'} \\
& + \frac{2\hat{\rho}}{n_{0g}(n_{0g}-1)} \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \sum_{l=1}^{n_{0g}} \sum_{l' \neq l}^{n_{0g}} \frac{(\hat{V}_{0g})_{k'k'}}{\sqrt{(\hat{V}_{0g})_{ll} (\hat{V}_{0g})_{l'l'}}} \Big(\widehat{\text{cov}}[(\hat{V}_{0g})_{ll'}, (\hat{V}_{0g})_{kk}] \\
& - \frac{(\hat{V}_{0g})_{ll'}}{(\hat{V}_{0g})_{ll}} \widehat{\text{cov}}[(\hat{V}_{0g})_{ll}, (\hat{V}_{0g})_{kk}) \Big), \qquad (5.41) \\
\widehat{\text{var}}[\hat{\rho}] =& \frac{1}{n_{0g}^2 (n_{0g}-1)^2} \sum_{k=1}^{n_{0g}} \sum_{k' \neq k}^{n_{0g}} \sum_{l=1}^{n_{0g}} \sum_{l' \neq l}^{n_{0g}} \frac{1}{\sqrt{(\hat{V}_{0g})_{kk} (\hat{V}_{0g})_{k'k'} (\hat{V}_{0g})_{ll} (\hat{V}_{0g})_{l'l'}}} \\
& \times \Big(\widehat{\text{cov}}[(\hat{V}_{0g})_{ll'}, (\hat{V}_{0g})_{kk'}] - 2 \frac{(\hat{V}_{0g})_{ll'}}{(\hat{V}_{0g})_{ll}} \widehat{\text{cov}}[(\hat{V}_{0g})_{ll}, (\hat{V}_{0g})_{kk'}] \\
& + \frac{(\hat{V}_{0g})_{ll'} (\hat{V}_{0g})_{kk'}}{(\hat{V}_{0g})_{ll} (\hat{V}_{0g})_{kk}} \widehat{\text{cov}}[(\hat{V}_{0g})_{ll}, (\hat{V}_{0g})_{kk}] \Big). \qquad (5.42)
\end{aligned}$$

Using the delta method, we get, for all $k, k', l, l' = 1, \ldots, n_{0g}$,

$$\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] \approx \begin{cases} \sqrt{(\hat{V}_{0g})_{ll} (\hat{V}_{0g})_{l'l'}} \text{cov}[\hat{\rho}, (\hat{V}_{0g})_{kk'}] \\ + \dfrac{\hat{\rho}}{2} \left(\sqrt{\dfrac{(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{ll}}} \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll}] \right. \\ \left. + \sqrt{\dfrac{(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{l'l'}}} \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{l'l'}] \right) & \text{if } l \neq l', \\ \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll}] & \text{if } l = l', \end{cases} \tag{5.43}$$

$$\operatorname{cov}[(\hat{\Theta}_{0g})_{kk'},(\hat{\Theta}_{0g})_{ll'}] \approx \begin{cases} \sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{ll}(\hat{V}_{0g})_{l'l'}}\operatorname{var}[\hat{\rho}] \\ +\frac{\hat{\rho}}{2}\left(\sqrt{\frac{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{ll}}}\operatorname{cov}[\hat{\rho},(\hat{V}_{0g})_{ll}]\right. \\ +\sqrt{\frac{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{l'l'}}}\operatorname{cov}[\hat{\rho},(\hat{V}_{0g})_{l'l'}] \\ +\sqrt{\frac{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{l}(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{kk}}}\operatorname{cov}[\hat{\rho},(\hat{V}_{0g})_{kk}] \\ \left.+\sqrt{\frac{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{ll}(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{k'k'}}}\operatorname{cov}[\hat{\rho},(\hat{V}_{0g})_{k'k'}]\right) \\ +\frac{\hat{\rho}^2}{4}\left(\sqrt{\frac{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{ll}}}\operatorname{cov}[(\hat{V}_{0g})_{kk},(\hat{V}_{0g})_{ll}]\right. \\ +\sqrt{\frac{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{ll}}}\operatorname{cov}[(\hat{V}_{0g})_{k'k'},(\hat{V}_{0g})_{ll}] \\ +\sqrt{\frac{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{l'l'}}}\operatorname{cov}[(\hat{V}_{0g})_{kk},(\hat{V}_{0g})_{l'l'}] \\ \left.+\sqrt{\frac{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{l'l'}}}\operatorname{cov}[(\hat{V}_{0g})_{k'k'},(\hat{V}_{0g})_{l'l'}]\right) & \text{if } k\neq k' \text{ and } l\neq l', \\ \sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}\operatorname{cov}[\hat{\rho},(\hat{V}_{0g})_{ll}] \\ +\frac{\hat{\rho}}{2}\left(\sqrt{\frac{(\hat{V}_{0g})_{k'k'}}{(\hat{V}_{0g})_{kk}}}\operatorname{cov}[(\hat{V}_{0g})_{kk},(\hat{V}_{0g})_{ll}]\right. \\ \left.+\sqrt{\frac{(\hat{V}_{0g})_{kk}}{(\hat{V}_{0g})_{k'k'}}}\operatorname{cov}[(\hat{V}_{0g})_{k'k'},(\hat{V}_{0g})_{ll}]\right) & \text{if } k\neq k' \text{ and } l=l', \\ \sqrt{(\hat{V}_{0g})_{ll}(\hat{V}_{0g})_{l'l'}}\operatorname{cov}[\hat{\rho},(\hat{V}_{0g})_{kk}] \\ +\frac{\hat{\rho}}{2}\left(\sqrt{\frac{(\hat{V}_{0g})_{l'l'}}{(\hat{V}_{0g})_{ll}}}\operatorname{cov}[(\hat{V}_{0g})_{kk},(\hat{V}_{0g})_{ll}]\right. \\ \left.+\sqrt{\frac{(\hat{V}_{0g})_{ll}}{(\hat{V}_{0g})_{l'l'}}}\operatorname{cov}[(\hat{V}_{0g})_{kk},(\hat{V}_{0g})_{l'l'}]\right) & \text{if } k=k' \text{ and } l\neq l', \\ \operatorname{cov}[(\hat{V}_{0g})_{kk},(\hat{V}_{0g})_{ll}] & \text{if } k=k' \text{ and } l=l', \end{cases} \tag{5.44}$$

where

$$\operatorname{cov}[\hat{\rho},(\hat{V}_{0g})_{ll'}] \approx \frac{1}{n_{0g}(n_{0g}-1)}\sum_{k=1}^{n_{0g}}\sum_{k'\neq k}^{n_{0g}}\frac{1}{\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}}} \times\left(\operatorname{cov}[(\hat{V}_{0g})_{kk'},(\hat{V}_{0g})_{ll'}]-\frac{(\hat{V}_{0g})_{kk'}}{(\hat{V}_{0g})_{kk}}\operatorname{cov}[(\hat{V}_{0g})_{kk},(\hat{V}_{0g})_{ll'}]\right), \tag{5.45}$$

$$\begin{aligned}\operatorname{var}[\hat{\rho}] \approx & \frac{1}{n_{0g}^2(n_{0g}-1)^2}\sum_{k=1}^{n_{0g}}\sum_{k'\neq k}^{n_{0g}}\sum_{l=1}^{n_{0g}}\sum_{l'\neq l}^{n_{0g}}\frac{1}{\sqrt{(\hat{V}_{0g})_{kk}(\hat{V}_{0g})_{k'k'}(\hat{V}_{0g})_{ll}(\hat{V}_{0g})_{l'l'}}} \\ & \times\Bigg(\operatorname{cov}[(\hat{V}_{0g})_{ll'},(\hat{V}_{0g})_{kk'}]-2\frac{(\hat{V}_{0g})_{ll'}}{(\hat{V}_{0g})_{ll}}\operatorname{cov}[(\hat{V}_{0g})_{ll},(\hat{V}_{0g})_{kk'}] \\ & +\frac{(\hat{V}_{0g})_{ll'}(\hat{V}_{0g})_{kk'}}{(\hat{V}_{0g})_{ll}(\hat{V}_{0g})_{kk}}\operatorname{cov}[(\hat{V}_{0g})_{ll},(\hat{V}_{0g})_{kk}]\Bigg). \end{aligned} \tag{5.46}$$

Target G: homogeneous variances and perfect positive correlations

In this thesis, we also consider a seventh target, similar to Target E, but with homogeneous variances such that

$$(\hat{\Theta}_{0g})_{kk'} = \begin{cases} \hat{v} = \dfrac{1}{n_{0g}}\displaystyle\sum_{k=1}^{n_{0g}}(\hat{V}_{0g})_{kk} & \text{if } k = k', \\ \hat{v} & \text{if } k \neq k'. \end{cases} \tag{5.47}$$

This target was not investigated in Schäfer and Strimmer (2005), but, we can make the same simplifications as performed for Targets B and C, and the optimal Schäfer-Strimmer OLW shrinkage intensity would be given by

$$\hat{\lambda}_g^{\text{OLW-SS}} = \frac{\displaystyle\sum_{k=1}^{n_{0g}}\sum_{k'=1}^{n_{0g}}\widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}]}{\displaystyle\sum_{k=1}^{n_{0g}}\sum_{k'=1}^{n_{0g}}(\hat{v}-(\hat{V}_{0g})_{kk})^2}. \tag{5.48}$$

Avoiding the simplifications, we get instead

$$\hat{\lambda}_g^{\text{OLW-C}} = \frac{-\hat{a}_7+\displaystyle\sum_{k=1}^{n_{0g}}\sum_{k'=1}^{n_{0g}}\widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}]}{n_{0g}\hat{a}_1-2\hat{a}_7+\displaystyle\sum_{k=1}^{n_{0g}}\sum_{k'=1}^{n_{0g}}\widehat{\operatorname{var}}[(\hat{V}_{0g})_{kk'}]+\sum_{k=1}^{n_{0g}}\sum_{k'=1}^{n_{0g}}(\hat{v}-(\hat{V}_{0g})_{kk})^2}, \tag{5.49}$$

where

$$\hat{a}_7 = \frac{1}{n_{0g}}\sum_{k=1}^{n_{0g}}\sum_{k'=1}^{n_{0g}}\sum_{l=1}^{n_{0g}}\widehat{\operatorname{cov}}[(\hat{V}_{0g})_{ll},(\hat{V}_{0g})_{kk'}]. \tag{5.50}$$

Finally, for all $k, k', l, l' = 1, \ldots, n_{0g}$, we have

$$\text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] = \frac{1}{n_{0g}^2} \sum_{i=1}^{n} \sum_{j=1}^{n_{0g}} \text{cov}[(\hat{V}_{0g})_{ii}, (\hat{V}_{0g})_{jj}], \tag{5.51}$$

$$\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] = \frac{1}{n_{0g}} \sum_{i=1}^{n_{0g}} \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ii}]. \tag{5.52}$$

5.2.4 Parametric inferences

If the shrinkage intensity $\hat{\lambda}_g$ is greater than 0, the variability of the shrinkage estimator $\hat{R}_{0g}$ will typically be different from the one of the unstructured estimator $\hat{V}_{0g}$. Therefore, the parametric tests developed in Section 3.2.4 have to be modified to account for the change of variability implied by the shrinkage. Here, we propose to modify Test II and Test III by replacing the estimator of each $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ by an estimator of $\text{Cov}[\text{vec}[\hat{R}_{0g}]]$. Nevertheless, to achieve this, we need to get an estimator of $\text{Cov}[\text{vec}[\hat{R}_{0g}]]$. Unfortunately, due to the non-linearity of $\hat{\lambda}_g$, this seems to be a very challenging task if we do not assume that $\hat{\lambda}_g$ is not a random variable. Therefore, here, we first make the assumption that $\hat{\lambda}_g$ is not random and we then get

$$\begin{aligned}\text{cov}[(\hat{R}_{0g})_{kk'}, (\hat{R}_{0g})_{ll'}] =& \hat{\lambda}_g^2 \text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}] + (1-\hat{\lambda}_g)^2 \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] \\ &+ \hat{\lambda}_g(1-\hat{\lambda}_g)(\text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{V}_{0g})_{ll'}] + \text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'}]).\end{aligned} \tag{5.53}$$

We can see that the elements of $\text{Cov}[\text{vec}[\hat{R}_{0g}]]$ depends on the target choice. For each of the seven targets introduced in Section 5.2.3, the formulas provided in that section for $\text{cov}[(\hat{\Theta}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'})]$ and $\text{cov}[(\hat{V}_{0g})_{kk'}, (\hat{\Theta}_{0g})_{ll'})]$ can be used to express $\text{cov}[(\hat{R}_{0g})_{kk'}, (\hat{R}_{0g})_{ll'}]$ as a function of the elements of V_{0g} and $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$. Then, we propose to replace V_{0g} and $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$ by sample estimates of them. Specifically, for V_{0g}, we propose to use the unstructured estimator $\hat{V}_{0g}$ and, for $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$, we propose to use the estimator proposed for Test II (see Section 3.2.4, Equation (3.54)) or the one proposed for Test III (see Section 3.2.4, Equation (3.69)), leading to two estimators for $\text{Cov}[\text{vec}[\hat{R}_{0g}]]$ that we will refer to as $\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{R}_{0g}]]$ and $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{R}_{0g}]]$.

Finally, in order to modify Test II, we propose to simply replace $\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$ by $\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{R}_{0g}]]$ in Equation (3.58). Similarly, we propose to modify Test III by replacing $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$ by $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{R}_{0g}]]$ in Equation (3.70).

5.2.5 Monte Carlo evaluations

To evaluate the shrinkage method proposed in this chapter, we performed several Monte Carlo simulations in the same scenarios as in the Monte Carlo Simulations performed in Chapter 3 (see Section 3.2.5). Each simulated dataset was analysed using 56 versions of the shrinkage SwE, differing by the use of one of the eight shrinkage SwE versions mentioned in Table 5.2 combined with one of the seven targets (A-G) presented in Section 5.2.3.

Shrinkage SwE name	Estimator to shrink	Formula type	Covariance estimator type
OLWS-SwE-SS II	$\hat{V}_{0g}$	"Schäfer-Strimmer"	$\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$
OLWS-SwE-SS III	$\hat{V}_{0g}$	"Schäfer-Strimmer"	$\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$
OLWS-SwE-C II	$\hat{V}_{0g}$	"Correct"	$\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$
OLWS-SwE-C III	$\hat{V}_{0g}$	"Correct"	$\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$
GLWS-SwE-S II	S	-	$\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$
GLWS-SwE-S III	S	-	$\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$
GLWS-SwE-CSC II	$CSC^\top$	-	$\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$
GLWS-SwE-CSC III	$CSC^\top$	-	$\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$

Table 5.2 Shrinkage SwE versions investigated in the Monte Carlo simulations. The first column gives the name of the shrinkage SwE versions. The second column indicates which estimator is targeted by the loss functions $L_g[\lambda_g]$ for MSE reduction. The third column indicates, when applicable, which formula is used to compute the optimal shrinkage intensity (i.e. $\hat{\lambda}_g^{\text{OLW-SS}}$ or $\hat{\lambda}_g^{\text{OLW-C}}$ given in Section 5.2.3). Finally, the fourth columns indicates which estimator is used to estimate $\text{Cov}[\text{vec}[\hat{V}_{0g}]]$; note that $\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$ is given by Equation (3.54) while $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$ is given by Equation (3.69).

All the shrinkage SwE versions were computed using the small sample adjustment used in S_{C2}^{Hom} as it was found to be the best in the simulations performed in Chapter 3. As baseline for comparison, we also used the SwE S_{C2}^{Hom} without any shrinkage. For inference, we considered the same contrasts as in the simulations of Chapter 3, and used Test II and Test III, modified as described in Section 5.2.4.

Assessment metrics

As a first set of assessment metrics, we used the Mean Squared Error (MSE_{F}), the Variance (VAR_{F}) and the Squared Bias (SBIAS_{F}) of each shrinkage estimator $\hat{R}_{0g}$

defined in the sense of the Frobenius norm, i.e. given by

$$\mathrm{MSE}_{\mathrm{F}} = \sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \mathbb{E}[((\hat{R}_{0g})_{kk'} - (V_{0g})_{kk'})^2], \tag{5.54}$$

$$\mathrm{VAR}_{\mathrm{F}} = \sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} \mathbb{E}[((\hat{R}_{0g})_{kk'} - \mathbb{E}[(R_{0g})_{kk'}])^2], \tag{5.55}$$

$$\mathrm{SBIAS}_{\mathrm{F}} = \sum_{k=1}^{n_{0g}} \sum_{k'=1}^{n_{0g}} (\mathbb{E}[(\hat{R}_{0g})_{kk'}] - (V_{0g})_{kk'})^2. \tag{5.56}$$

Note that it can be easily verified that $\mathrm{MSE}_{\mathrm{F}} = \mathrm{VAR}_{\mathrm{F}} + \mathrm{SBIAS}_{\mathrm{F}}$.

As a second set of assessment metrics, we used the Mean Squared Error (MSE), the Variance (VAR) and the Bias (BIAS) of each contrasted SwE $CSC^\top$, i.e. given by

$$\mathrm{MSE} = \mathbb{E}[(CSC^\top - \mathrm{var}[C\hat{\beta}])^2], \tag{5.57}$$

$$\mathrm{VAR} = \mathbb{E}[(CSC^\top - \mathbb{E}[CSC^\top])^2], \tag{5.58}$$

$$\mathrm{BIAS} = \mathbb{E}[CSC^\top] - \mathrm{var}[C\hat{\beta}], \tag{5.59}$$

which are linked by the fact that $\mathrm{MSE} = \mathrm{VAR} + \mathrm{BIAS}^2$.

Finally, as a third set of assessment metrics, we used the FPR and power in order to assess the quality of the inferences.

5.3 Results

In this section, we summarise the results of the Monte Carlo simulations described in Section 5.2.5. Note that we only show results for the smallest sample sizes (i.e. for the balanced designs with 12 subjects and the unbalanced ADNI designs with 25 subjects) as these corresponded to the designs where the effect of shrinkage was the most important.

5.3.1 Estimation error of $\hat{R}_{0g}$

Figure 5.1 and 5.2 compare the OLWS-SwE versions in terms of the estimations errors of $\hat{R}_{0g}$ for two specific groups of subjects (the group B in the balanced designs with 12 subjects and the MCI group in the unbalanced ADNI designs with 25 subjects) across several scenarios. Overall, it seems that the best results were achieved when the Schäfer-Strimmer simplifications were not used, particularly in the four scenarios with correlated data (see rows 2-5 in the figures). When the Schäfer-Strimmer simpli-

fications were used, the shrinkage estimator had the tendency to overshrink, increasing quite strongly the bias and even leading to an increase of the MSE_F in some scenarios.

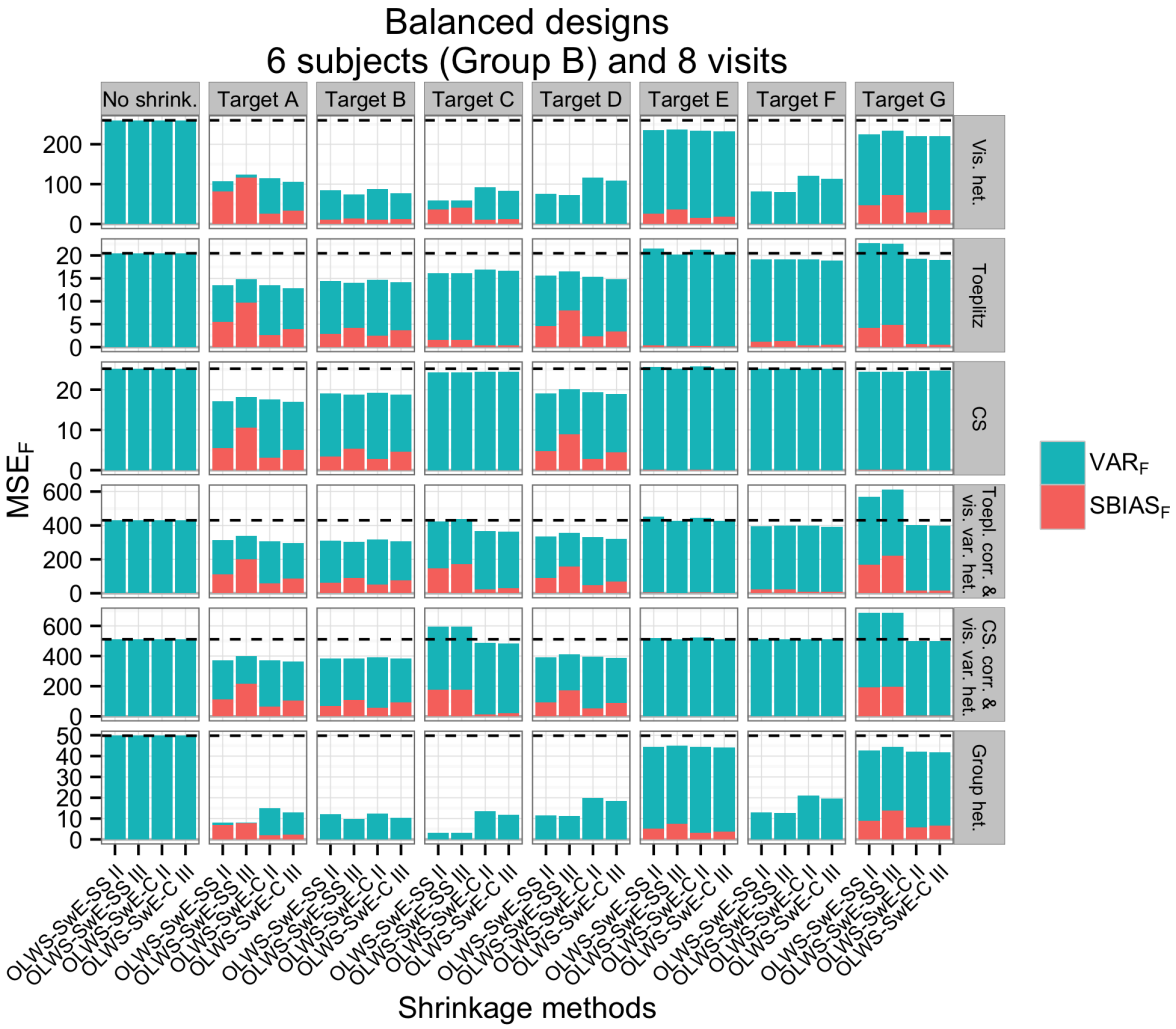

Fig. 5.1 Barplots comparing the MSE_F, VAR_F and SBIAS_F of the shrinkage estimator $\hat{R}_\text{B}$ obtained using several versions of the OLWS-SwE in the balanced designs with 12 subjects in total. A description of the target can be found in Table 5.1 while a description of the shrinkage methods can be found in Table 5.2.

Regarding the target choice, Targets E and G, which both assume perfect positive correlations, yielded the smallest reductions of MSE_F while Targets A, B and D, which all assume no correlation, seemed to yield overall the largest reduction of MSE_F. Targets C and F, which both assume homogeneous correlations, seemed to have an intermediary behaviour with Target C working better than Target F. While Target C was less performant that Target A, B and D in some scenarios (see, e.g., rows 2-5

in Figure 5.1), it seemed to perform as well or even better than them in some other scenarios (see, e.g., Figure 5.2) and had the tendency to carry a lot of less estimation error due to the bias, making it an interesting candidate.

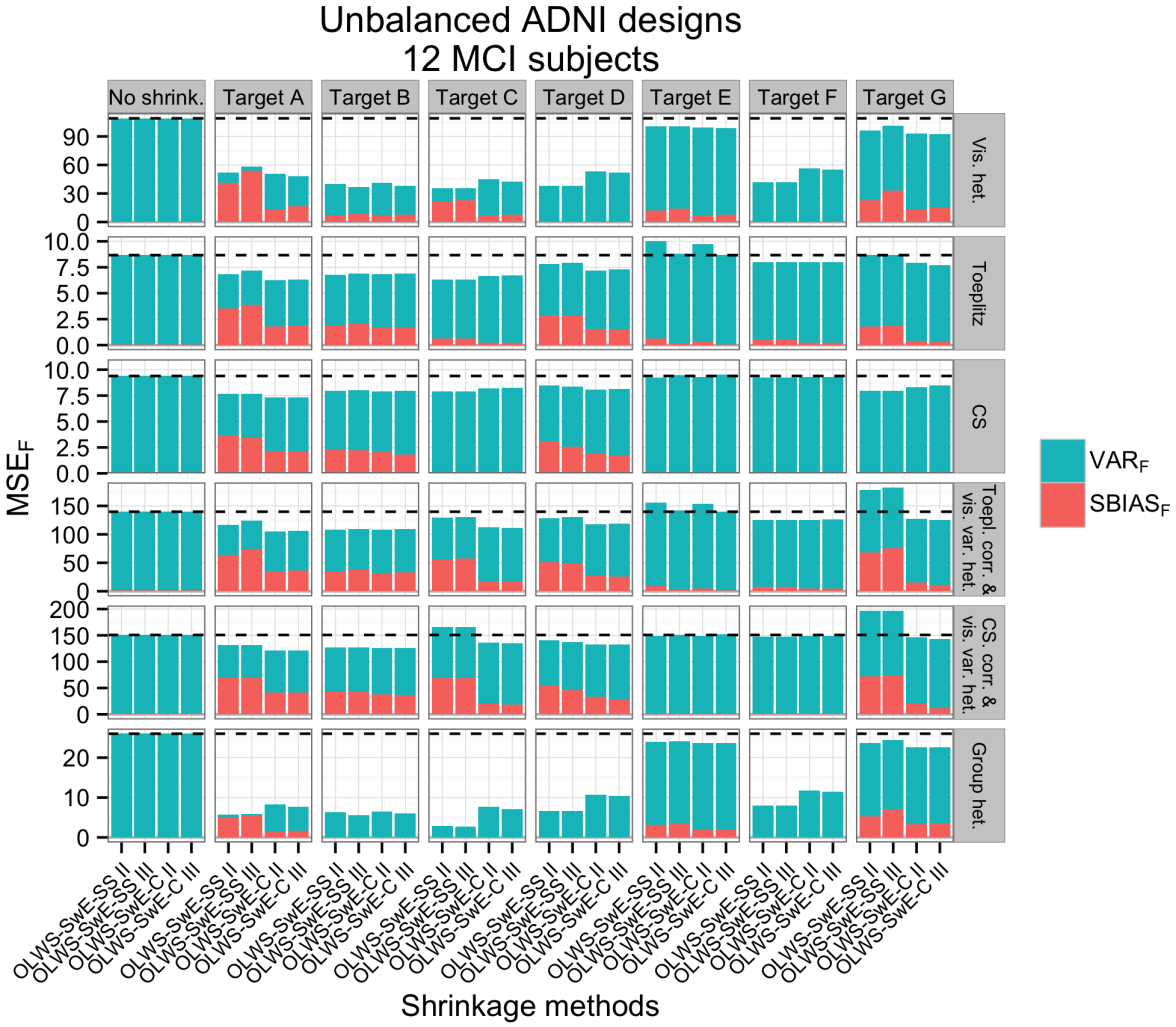

Fig. 5.2 Barplots comparing the MSE_F, VAR_F and SBIAS_F of the shrinkage estimator $\hat{R}_{\text{MCI}}$ obtained using several versions of the OLWS-SwE in the unbalanced ADNI designs with 25 subjects in total. A description of the target can be found in Table 5.1 while a description of the shrinkage methods can be found in Table 5.2.

Finally, for the OLWS-SwE using the Schäfer-Strimmer simplifications, the results obtained with $\widehat{\text{Cov}}_{\text{II}}[\text{vec}[\hat{V}_{0g}]]$ seemed better than those obtained with $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$. However, for the OLWS-SwE avoiding the Schäfer-Strimmer simplifications, the results seemed to indicate the converse, i.e. better performances with $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$.

5.3.2 Estimation error of the contrasted SwE $CSC^\top$

Figures 5.3 and 5.4 compares the relative MSE (defined as the ratio between the MSE obtained with shrinkage and the one obtained without shrinkage) of several contrasted shrinkage SwE. We clearly see that al the OLWS-SWE versions yielded poor results exhibiting, in some scenarios, an increase of MSE, mainly when the Schäfer-Strimmer simplifications were used. The GLWS-SwE versions targeting the reduction of the MSE_F of the whole SwE (GLWS-SwE-S) seemed to work poorly as well, exhibiting, in some scenarios, an increase of MSE. Only the GLWS-SwE versions focusing directly on the minimisation of the MSE of the contrasted SwE (GLWS-SwE-CSC) seemed to work appropriately, particularly when the estimator of covariances $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$ was used (GLWS-SwE-CSC III).

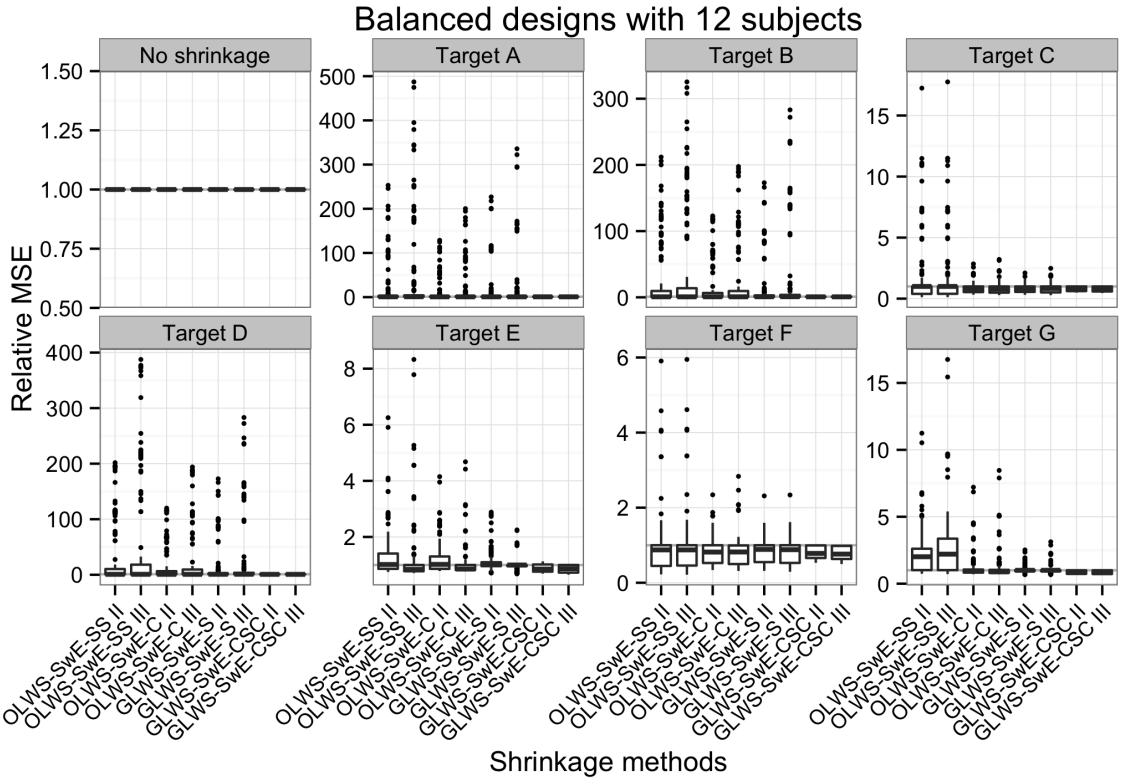

Fig. 5.3 Boxplots showing the relative MSE (defined as the ratio between the MSE obtained with shrinkage and the one obtained without shrinkage) of several contrasted shrinkage SwE over 9 contrasts, 3 numbers of visits and 6 covariance matrix structures. Note that, for clarity, the scales are different over targets.

Figures 5.5 shows the results in more details for GLWS-SwE-CSC III in the unbalanced ADNI designs with 25 subjects. It seemed that the best target choice depends

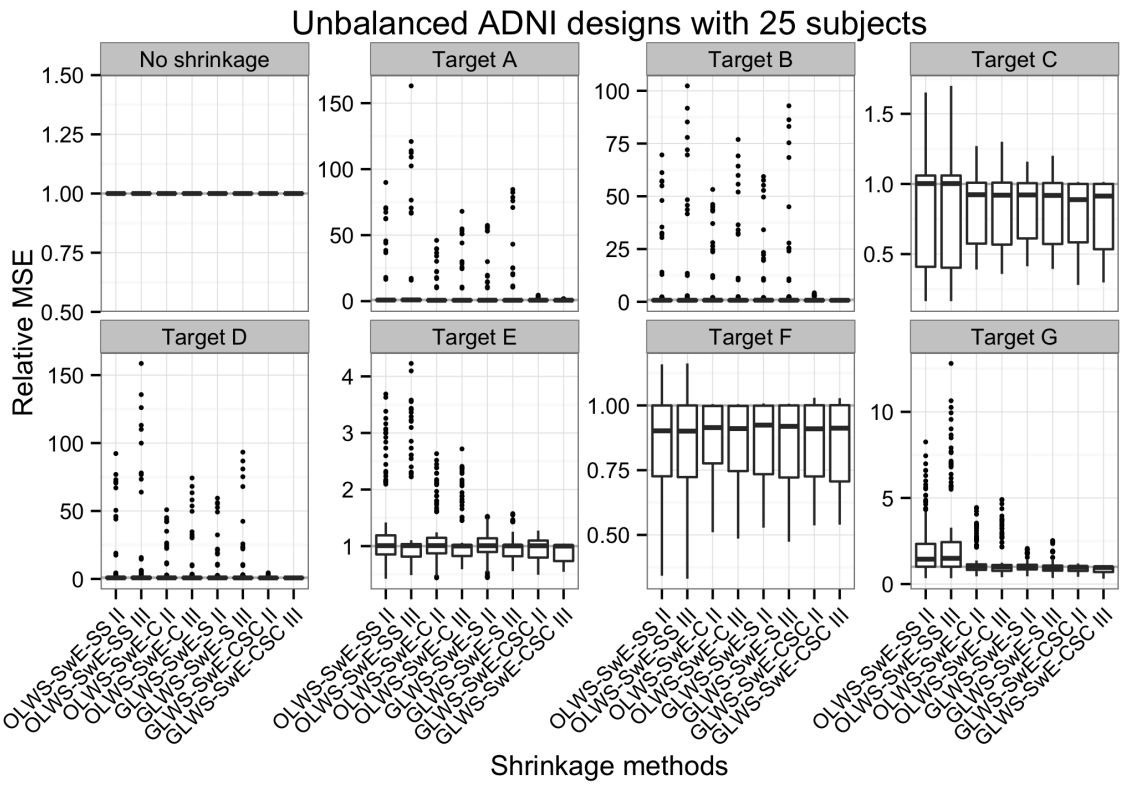

Fig. 5.4 Boxplots showing the relative MSE (defined as the ratio between the MSE obtained with shrinkage and the one obtained without shrinkage) of several contrasted shrinkage SwE over 24 contrasts and 6 covariance matrix structures. Note that, for clarity, the scales are different over targets.

strongly on the type of effects. Indeed, for the cross-sectional effects considered, Targets A, B and D, all assuming no correlation, seemed to outperform the other targets while this was not the case for the longitudinal effects. For those effects, Target C, which has a compound symmetric structure, seemed overall the best performing choice.

5.3.3 Parametric inference

Figure 5.6 (bottom) shows some typical results about the control of the FPR obtained with Test III using a shrinkage estimator. We clearly see that the use of a shrinkage SwE systematically yielded a liberal control of the FPR. This can be explained by the fact that the shrinkage tended to introduce a negative bias (see top of Figure 5.6) into the SwE which, in turn, tended to inflate the Wald scores. This liberal control of the FPR were observed across the majority of the scenarios investigated, indicating that no valid parametric inference was typically obtained when a shrinkage estimator was

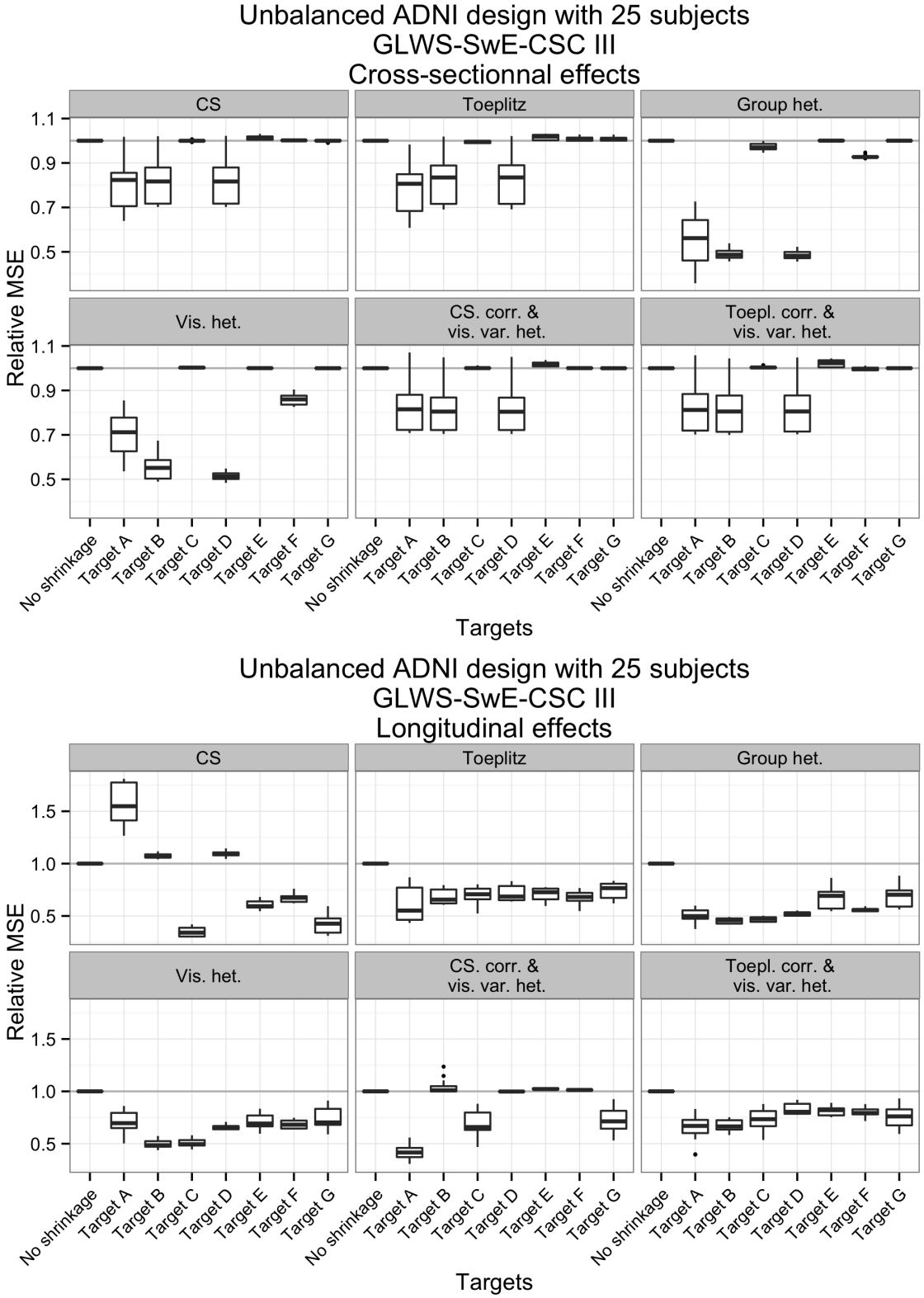

Fig. 5.5 Boxplots showing the relative MSE of several contrasted shrinkage SwE after using the GLWS-SWE-CSC III over 12 cross-sectional contrasts (top) and 12 longitudinal contrasts (bottom) in the unbalanced ADNI designs with 25 subjects in total.

used.

As the observed inferences were typically invalid, we do not show any results regarding the power. Note, though, that the use of shrinkage seemed to improve the power. Nevertheless, this can be attributed to both the reduction of variability and the negative bias observed in the shrinkage SwE, which is not acceptable.

5.4 Conclusion

In this chapter, inspired by the work of Warton (2011), we have investigated the use of shrinkage in the computation of the SwE. More precisely, we have proposed two main version of shrinkage SwE, the OLWS-SwE and the GLWS-SwE, both based on the use of the Ledoit-Wolf procedure (Ledoit and Wolf, 2003) to estimate the shrinkage intensity. These have the advantage, compared to cross-validation procedures, to be relatively fast to compute. For seven potential targets, we have also provided the necessary equations allowing an easy computation of these shrinkage SwE versions.

Using Monte Carlo simulations, we have showed that the shrinkage estimator $\hat{R}_{0g}$ used in the OLWS-SwE versions may effectively be used to reduce the MSE_F of the covariance matrix estimator, but that it is preferable to avoid the simplifications made in Schäfer and Strimmer (2005) as they had the tendency to overshrink. However, the Monte Carlo simulations also showed that the OLWS-SwE is not a reliable way to reduce the MSE of a contrasted SwE $CSC^\top$ and that only the GLWS-SwE version targeting specifically the reduction of MSE of $CSC^\top$ and using the estimator $\widehat{\text{Cov}}_{\text{III}}[\text{vec}[\hat{V}_{0g}]]$ was able to achieve valid reductions of MSE.

Regarding the target choice, the Monte Carlo simulations seemed to indicate that the best choice may depend on the type of effects we are interesting in. For cross-sectional effects, it seemed that the targets assuming no correlation (Targets A, B and D) were the most appropriate while Target C, which assumes a compound symmetric structure, seemed the best choice for longitudinal effects. Note that alternative targets could be considered. For example, we could use a target obtained by pooling information from several voxels, such as one of the smooth estimators of covariance matrix proposed in Chapter 6 or, as suggested by Gerard Ridgway, a target based on the global covariance structure estimated by the SPM procedure described in Section 2.3.4. However, further work is needed to investigate this idea.

Unfortunately, while the GLWS-SwE was able to somehow improve the quality of the estimation of $\text{var}[C\hat{\beta}]$, at least in terms of MSE, it seemed clear from the Monte Carlo simulations that it was unsuitable to make parametric inferences due to the bias

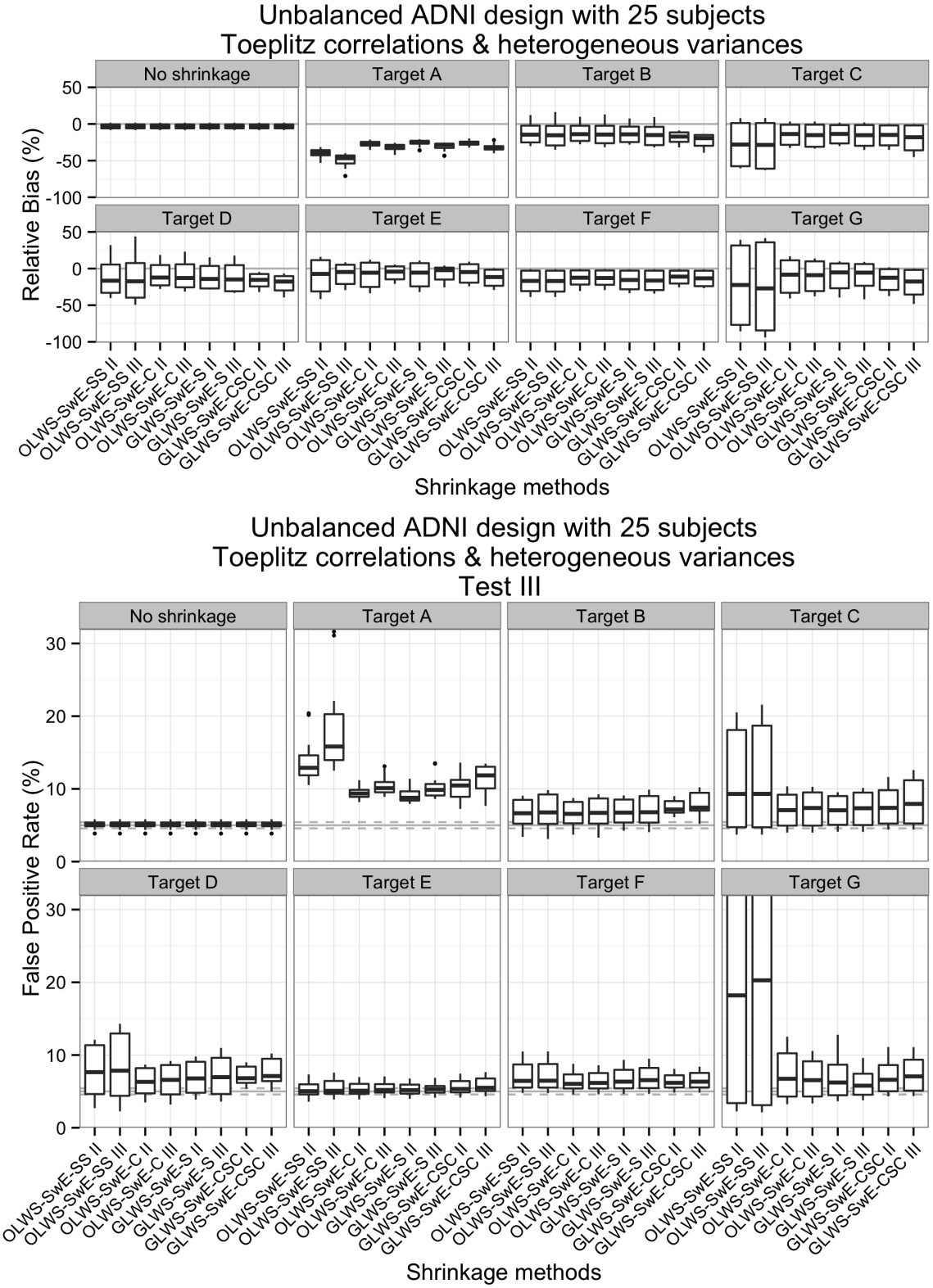

Fig. 5.6 Boxplots showing the relative bias (defined as the ratio between the bias and var$[C\hat{\beta}]$) and the FPR obtained after using Test III for several shrinkage SwE versions over 24 contrasts in the unbalanced ADNI design with 25 subjects under Toeplitz correlations and heterogeneous variances. Note that, for clarity, only the results between 0% and 32% of FPR are shown, affecting only the results related to Target G.

induced by the shrinkage. Nevertheless, as a future project, it would be interesting to investigate the use of the Wild Bootstrap investigated in Chapter 4 in order to make non-parametric inferences with the shrinkage SwE. Indeed, in Section 4.3.1, the Wild Bootstrap appeared to be relatively robust against the presence of bias in the SwE, indicating that it might also be robust against the bias typically observed in the shrinkage SwE.

Finally, note that we did not apply the shrinkage SwE approach to the real ADNI dataset as the Monte Carlo simulations showed that the parametric tests developed in Section 5.2.4 are invalid. Nevertheless, even if we have not found any direct application of the shrinkage SwE in our context, as our main goal is to make statistical inference, the work made in this chapter could still be useful in other contexts, for example, when an estimate of variance (or covariance matrix) with less estimation error is needed.

Chapter 6

Covariance matrix smoothing in the Sandwich Estimator

6.1 Introduction

When the number of subjects is small, each covariance matrix estimator $\hat{V}_{0g}$ used in the homogeneous SwE S^{Hom} (see Section 3.2.3) is expected to be highly variable, potentially limiting the power to detect effects. In Chapter 5, we attempted to reduce the variability of the covariance matrix estimators by shrinking them towards less variable estimators. While the latter allowed a reduction of variability, we showed that it also typically induced a bias in the SwE, which unfortunately prevented to make valid parametric inference. Nevertheless, instead of attempting to reduce the variability using only the information at the voxel level, we may try to use the information from several voxels. Indeed, if each covariance matrice V_{0g} is homogeneous across the brain, at least locally, we may want to pool the information across voxels by, for example, spatially smoothing the covariance matrix images.

This idea has already been proposed in neuroimaging, but, to our knowledge, only in the context of a scalar variance (Nichols and Holmes, 2002; Worsley et al., 2002). In particular, Nichols and Holmes (2002) proposed this in the context of permutation tests while Worsley et al. (2002) proposed this in the context of parametric tests where the scalar variance could be decomposed into two components, a fixed effect variance with low variability and a random effect variance with high variability. In this context, Worsley et al. (2002) argued that the global variance image is in general not homogeneous and proposed instead to smooth the ratio between the random effect variance and the fixed effect variance, arguing that it is a more homogeneous image.

In our case, however, we do not have scalar variances, but covariance matrices that

are far more challenging to handle. First, a covariance matrix belongs to the space of positive semi-definite matrices and, as such, the use of a simple metric like the Euclidean metric may not be appropriate for the smoothing. Indeed, in the context of Diffusion Tensor Imaging (DTI), Dryden et al. (2009) reported that other metrics such as the Root-Euclidean or the Log-Euclidean metric should be preferred when one wants to average or interpolate covariance matrices. However, the sample covariance matrices we encounter in longitudinal neuroimaging data and, more particularly, in the context of the SwE method may differ from the ones we encounter in DTI. Therefore, it is relatively unclear which metric should be used in our context. Another challenge resides in the fact that, in general, the covariance matrices are not homogeneous across the brain, even locally. Therefore, it would be desirable to first spatially homogenise the covariance matrices before smoothing them, like it was proposed in Worsley et al. (2002). Nevertheless, the homogenisation proposed in Worsley et al. (2002) concerned scalar variances which can be decomposed into fixed and random effect terms and, therefore, it cannot be reproduced for our purpose. In our case, we need to find a proper way to homogenise covariance matrices, which is definitively more challenging. Finally, in the context of parametric inference, even if we find a good smoothing metric and a good homogenisation, we still need to quantify how the effective number of degrees of freedom is increased after the spatial smoothing. This might be also challenging as this depends on several factors such as the smoothing metric, the homogenisation and the degree of spatial smoothness initially present in the images.

In this chapter, inspired by the work in Dryden et al. (2009), we first review several possible smoothing metrics. Then, we propose several homogenisations of the covariance matrices and discuss how we could estimate the effective number of degrees of freedom after spatial smoothing. Finally, we use Monte Carlo simulations to assess the smoothing metrics and the proposed homogenisation in settings important for longitudinal neuroimaging data.

6.2 Methods

6.2.1 Smoothing metrics

In Nichols and Holmes (2002) and in Worsley et al. (2002), the smoothing was made on scalar values using the Euclidean metric. In our case, we can also consider the Euclidean metric, for which, the smoothed covariance matrix at voxel v is given by

$$\mathrm{sm}_{\mathrm{E}}[\hat{V}_{0g}[v]] = \sum_{v'} K[v, v']\hat{V}_{0g}[v'], \tag{6.1}$$

where $\hat{V}_{0g}[v']$ is the original sample covariance matrix estimate at voxel v' and $K[v, v']$ is a Gaussian kernel.

In order to define other metrics, let us first consider the spectral decomposition $\hat{V}_{0g}[v] = U[v]\Lambda[v]U^{\top}[v]$, where $U[v]$ is a matrix containing the eigenvectors of $\hat{V}_{0g}[v]$ and $\Lambda[v]$ is a diagonal matrix containing the eigenvalues of $\hat{V}_{0g}[v]$ in its diagonal. Based on this spectral decomposition, we can define the matrix logarithm, the matrix exponential and the matrix square root of $\hat{V}_{0g}[v]$ as

$$\log[\hat{V}_{0g}[v]] = U[v]\log[\Lambda[v]]U^{\top}[v], \tag{6.2}$$

$$\exp[\hat{V}_{0g}[v]] = U[v]\exp[\Lambda[v]]U^{\top}[v], \tag{6.3}$$

$$\hat{V}_{0g}^{1/2}[v] = U[v]\Lambda^{1/2}[v]U^{\top}[v], \tag{6.4}$$

respectively, where $\log[\Lambda[v]]$, $\exp[\Lambda[v]]$ and $\Lambda^{1/2}[v]$ are the diagonal matrices with diagonal elements obtained by taking the logarithms, the exponentials and the square roots of the eigenvalues of $\hat{V}_{0g}[v]$, respectively.

Using these definitions, we can define the Log-Euclidean metric (Arsigny et al., 2007), for which, the smoothed covariance matrix at voxel v is given by

$$\mathrm{sm}_{\mathrm{L}}[\hat{V}_{0g}[v]] = \exp\Big[\sum_{v'} K[v, v']\log[\hat{V}_{0g}[v']]\Big]. \tag{6.5}$$

Also, we can define the Square-Root-Euclidean metric (Dryden et al., 2009), for which, the smoothed covariance matrix at voxel v is given by

$$\mathrm{sm}_{\mathrm{SR}}[\hat{V}_{0g}[v]] = \Big(\sum_{v'} K[v, v']\hat{V}_{0g}^{1/2}[v']\Big)\Big(\sum_{v'} K[v, v']\hat{V}_{0g}^{1/2}[v']\Big). \tag{6.6}$$

Finally, using the Cholesky decomposition $\hat{V}_{0g}[v] = \mathrm{chol}[\hat{V}_{0g}[v]]\mathrm{chol}[\hat{V}_{0g}[v]]^{\top}$, we can define the Cholesky metric (Wang et al., 2004), for which, the smoothed covariance matrix at voxel v is given by

$$\mathrm{sm}_{\mathrm{C}}[\hat{V}_{0g}[v]] = \Big(\sum_{v'} K[v, v']\mathrm{chol}[\hat{V}_{0g}[v]]\Big)\Big(\sum_{v'} K[v, v']\mathrm{chol}[\hat{V}_{0g}[v]]\Big)^{\top}. \tag{6.7}$$

Other metrics may also be considered such as, for example, the Procrustes size-and-

shape metric (Dryden et al., 2009). However, these type of metric seems generally more complicated to handle and, therefore, in this thesis, we only consider the Euclidean, the Log-Euclidean, the Square-Root-Euclidean and the Cholesky metrics as defined above.

6.2.2 Spatial homogenisations

Using Equations (6.1), (6.5), (6.6) or (6.7), the sample covariance matrix images are smoothed directly without any transformation. To be valid, the covariance matrices have to be homogeneous across the brain or at least locally homogeneous. Unfortunately, this is unlikely to be true in practice. Therefore, we can attempt to first homogenise the sample covariance matrices, before smoothing them. This type of strategy has already been proposed in Worsley et al. (2002), but only on scalar values and, to our knowledge, no extension to covariance matrices exists in the literature. Therefore, here, we propose several smoothing strategies relying on some form of homogenisation of the sample covariance matrices as described below.

The first idea is to homogenise the sample covariance matrices by dividing them by their respective traces. All the transformed covariance matrices should then have a trace equal to one and lead to a smoothing for which the smoothed covariance matrix at voxel v is given by

$$\widetilde{V}_{0g}^{\mathrm{tr}}[v] = \mathrm{sm}\left[\frac{\hat{V}_{0g}[v]}{\mathrm{tr}[\hat{V}_{0g}[v]]}\right]\mathrm{tr}[\hat{V}_{0g}[v]], \tag{6.8}$$

where the smoothing operator sm can be either sm_{E}, sm_{L}, $\mathrm{sm}_{\mathrm{SR}}$ or sm_{C} as defined in Section 6.2.1. Note that $\hat{V}_{0g}[v]$ and $\mathrm{tr}[\hat{V}_{0g}[v]]$ are not independent random variables. This could be problematic as, even if the true covariance matrices are perfectly homogeneous (i.e all equal to a common covariance matrix), the expectation of $\widetilde{V}_{0g}^{\mathrm{tr}}[v]$ could be different from the true common covariance matrix due to the dependence of the sample covariance matrix with its trace. Also, with this homogenisation, even if the extent of smoothing is very large, we will not be able to decrease $\mathrm{Cov}[\mathrm{vec}[\widetilde{V}_{0g}^{\mathrm{tr}}[v]]]$ towards zero, but towards the value

$$\mathrm{var}[\mathrm{tr}[\hat{V}_{0g}[v]]]\mathrm{vec}\left[\mathrm{sm}\left[\frac{\hat{V}_{0g}[v]}{\mathrm{tr}[\hat{V}_{0g}[v]]}\right]\right]\mathrm{vec}\left[\mathrm{sm}\left[\frac{\hat{V}_{0g}[v]}{\mathrm{tr}[\hat{V}_{0g}[v]]}\right]\right]^{\top}, \tag{6.9}$$

meaning that the possible reduction of variability could be rather limited in practice.

To develop another type of homogenisation, let us assume for a moment that

the true covariance matrix at each voxel is known. In such a circumstance, one of the most straightforward homogenisations would be to use the square root or the Cholesky decomposition of the true covariance matrix such that the smoothed covariance matrix at voxel v would be given by

$$\widetilde{V}_{0g}^{\mathrm{SR}}[v] = V_{0g}^{1/2}[v]\mathrm{sm}\Big[V_{0g}^{-1/2}[v]\hat{V}_{0g}[v]V_{0g}^{-1/2}[v]\Big]V_{0g}^{1/2}[v] \text{ or} \tag{6.10}$$

$$\widetilde{V}_{0g}^{\mathrm{C}}[v] = \mathrm{chol}[V_{0g}[v]]\mathrm{sm}\Big[\mathrm{chol}[V_{0g}[v]]^{-1}\hat{V}_{0g}[v]\mathrm{chol}[V_{0g}[v]]^{-\top}\Big]\mathrm{chol}[V_{0g}[v]]^{\top}, \tag{6.11}$$

respectively.

Unfortunately, in practice, the true covariance matrices used in Equations (6.10) and (6.11) are unknown and, therefore, we need to replace them by sample estimates of them which are less variable than $\hat{V}_{0g}$, but also with as little bias as possible. One of such candidates could be one of the targets $\hat{\Theta}_{0g}$ investigated in Section 5.2.3, yielding smoothed estimates given, at voxel v, by

$$\widetilde{V}_{0g}^{\mathrm{SR}}[v] = \hat{\Theta}_{0g}^{1/2}[v]\mathrm{sm}\Big[\hat{\Theta}_{0g}^{-1/2}[v]\hat{V}_{0g}[v]\hat{\Theta}_{0g}^{-1/2}[v]\Big]\hat{\Theta}_{0g}^{1/2}[v] \text{ or} \tag{6.12}$$

$$\widetilde{V}_{0g}^{\mathrm{C}}[v] = \mathrm{chol}[\hat{\Theta}_{0g}[v]]\mathrm{sm}\Big[\mathrm{chol}[\hat{\Theta}_{0g}[v]]^{-1}\hat{V}_{0g}[v]\mathrm{chol}[\hat{\Theta}_{0g}[v]]^{-\top}\Big]\mathrm{chol}[\hat{\Theta}_{0g}[v]]^{\top}. \tag{6.13}$$

Unfortunately, while we can expect a decrease of variability, we can also expect some bias appearing and it is relatively unclear what would be the effect of this on the smoothing. Moreover, similarly to the trace homogenisation, it will not be possible to decrease $\mathrm{Cov}[\mathrm{vec}[\widetilde{V}_{0g}^{\mathrm{SR}}[v]]]$ or $\mathrm{Cov}[\mathrm{vec}[\widetilde{V}_{0g}^{\mathrm{C}}[v]]]$ towards zero, but only towards a matrix of finite values closely related to the variability of the target $\hat{\Theta}_{0g}[v]$ used for the homogenisation. This means again that the reduction of variability could be relatively limited in practice.

6.2.3 The smooth SwE

After smoothing the sample covariance estimate images using one of the metrics and one of the homogenisations (or none) described in the two previous section, we simply use, at each voxel v, the resulting smooth estimate $\widetilde{V}_{0g}[v]$ to estimate each subject covariance matrice V_i in the SwE (see Equation (2.7)) to obtain a new version of the SwE that we refer to as the smooth SwE in this thesis.

6.2.4 Parametric inferences

To make inference using smoothed estimates of the covariance matrices, we could adapt one of the parametric tests introduced in Section 3.2.4. However, this would require the estimation of $\text{Cov}[\text{vec}[\widetilde{V}_{0g}]]$ which does not seem straightforward, particularly when a non-Euclidean metric and/or a homogenisation are used. Nevertheless, in some strict circumstances, we can provide some ways to estimate $\text{Cov}[\text{vec}[\widetilde{V}_{0g}]]$ and consequently the number of degrees of freedom used for the parametric testing.

The first circumstance would be when we consider the Euclidean metric, no homogenisation and the assumptions that the true covariance matrices and the data smoothness are spatially homogeneous. With the additional assumption that we have only one measure per subject (i.e. V_{0g} would be a scalar variance) and an infinite number of voxels, Zhang (2008, Section B.1.1) showed that

$$\text{var}[\tilde{V}_{0g}] = \text{var}[\hat{V}_{0g}] \left(1 + 2\left(\frac{\text{FWHM}_\text{V}}{\text{FWHM}_\text{D}}\right)^2\right)^{-3/2}, \tag{6.14}$$

where FWHM_V is the Full Width at Half Maximum of the Gaussian smoothing kernel and FWHM_D is the Full Width at Half Maximum of the data smoothness (assuming a Gaussian kernel). Here, we simply propose to extend Equation (6.14) to a longitudinal setting in which case we get

$$\text{Cov}[\text{vec}[\widetilde{V}_{0g}]] = \text{Cov}[\text{vec}[\hat{V}_{0g}]] \left(1 + 2\left(\frac{\text{FWHM}_V}{\text{FWHM}_D}\right)^2\right)^{-3/2}. \tag{6.15}$$

Note that, from Equation (6.15), it is interesting to note that, for a smoothing with $\text{FWHM}_\text{V} = \text{FWHM}_\text{D}$, we can expect to reduce the variability by a factor superior to five.

For all the other cases, due to the non-linearity of the metrics or of the homogenisations, it seems harder to estimate the variability of $\widetilde{V}_{0g}$. Nevertheless, if we use the trace homogenisation (see Equation (6.8)) and if the extent of smoothing is very large, we can assume that the smoothed part of Equation (6.8) is not random and use the value given in Equation (6.9) to estimate $\text{Cov}[\text{vec}[\widetilde{V}^{\text{tr}}_{0g}]]$.

Similarly, the same kind of considerations could be made for $\text{Cov}[\text{vec}[\widetilde{V}^{\text{SR}}_{0g}]]$ and $\text{Cov}[\text{vec}[\widetilde{V}^{\text{C}}_{0g}]]$ if we assume that the extent of smoothing is very large and that the smoothed parts of Equations (6.12) and (6.13) asymptotically yield the identity matrix. In such a circumstance, $\text{Cov}[\text{vec}[\widetilde{V}^{\text{SR}}_{0g}]]$ and $\text{Cov}[\text{vec}[\widetilde{V}^{\text{C}}_{0g}]]$ can then be estimated by

$\text{Cov}[\text{vec}[\hat{\Theta}_{0g}]]$, which can, in turn, be estimated using the relevant equations given in Section 5.2.3.

6.2.5 Monte Carlo evaluations

To assess the smoothing metrics defined in Section 6.2.1 and the homogenisation procedures proposed in Section 6.2.2, we used Monte Carlo simulations with 10,000 realisations. To simplify the simulations, we only simulated spatial longitudinal data in two dimensions (21×21 pixels) without spatial correlation. The data at each pixel was generated as in the Monte Carlo simulations of Chapter 3 (see Section 3.2.5), but only for the balanced design scenarios. For each scenario, the data was generated with the same covariance matrix (one of the 6 covariance matrix described in Section 3.2.5) at every pixel, i.e. without spatial heterogeneity of the covariance matrix. We then estimated each covariance matrix $\hat{V}_{0g}$ at every pixel as described in Section 3.2.3, but using the small-sample bias adjustment used for $S_{\text{C2}}^{\text{Hom}}$, as it was shown to be the best in the simulations of Chapter 3 (see Section 3.3.1). We then applied 16 smoothing procedures on the central pixel differing by the use of one of the four metrics described in Section 6.2.1, and the use of one of the three homogenisations proposed in Section 6.2.2 or without any homogenisation. For the homogenisation based on the Cholesky decomposition and the one based on the square root decomposition, we used the target with a compound symmetric structure (Target C in Chapter 5). For each smoothing procedure, a discrete Gaussian smoothing with four different FWHM of 1, 2, 3 and 4 pixels was used. Finally, using the resulting smooth estimates, we computed the smooth SwE at the central pixel. Also, for comparison, we considered the SwE $S_{\text{C2}}^{\text{Hom}}$ at the central pixel obtained without any smoothing.

To assess the smoothing procedures, we used the relative bias (defined as the ratio between the bias and the true value) and the variance ratio (defined as the ratio between the variance after and before smoothing) of nine contrasted SwE $CSC^{\top}$ (see Section 3.2.5 for details about the contrasts used) extracted from the central pixel after using in turn every smoothing procedure.

6.3 Results

Figure 6.1 compare the smoothing procedures in terms of the relative bias of several contrasted SwE in the balanced designs with 12 subjects. In all the scenarios, the smoothing using the Cholesky, Root-Euclidean or Log-Euclidean metrics seemed to

introduce an important negative bias in the estimation of the SwE, indicating a strong underestimation of the covariance matrix of the parameters. The Euclidean metric seemed to work better than the three other metrics and seemed even to yield unbiased estimates when no homogenisation was used. However, when a homogenisation was used, the smoothing had the tendency to introduce a bias, but clearly smaller in absolute value than those observed for the three other metrics. As it can be more clearly observed in Figure 6.2, the bias was in general positive, but also appeared to be negative in a few scenarios. The homogenisation by the trace seemed to yield the worst results while the homogenisation based on the square-root decomposition seemed to yield the best results, but still inducing a non-negligible bias in some scenarios.

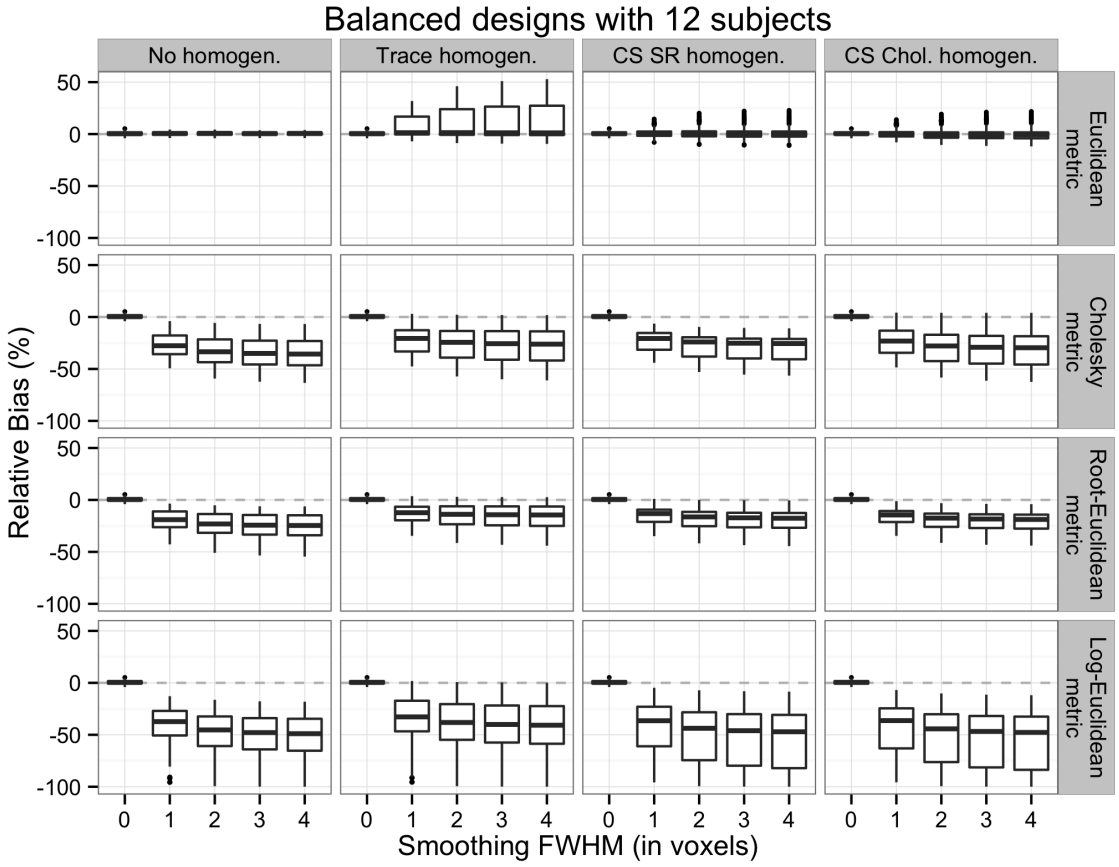

Fig. 6.1 Boxplots showing the effect of smoothing in terms of the relative bias of several contrasted SwE in the balanced designs with 12 subjects over 162 scenarios (consisting of the 9 contrasts tested, the 6 within-subject covariance structures and the 3 numbers of visits per subject considered in the Monte Carlo simulations).

Figure 6.3 shows how the variance of the contrasted SwE was modified after

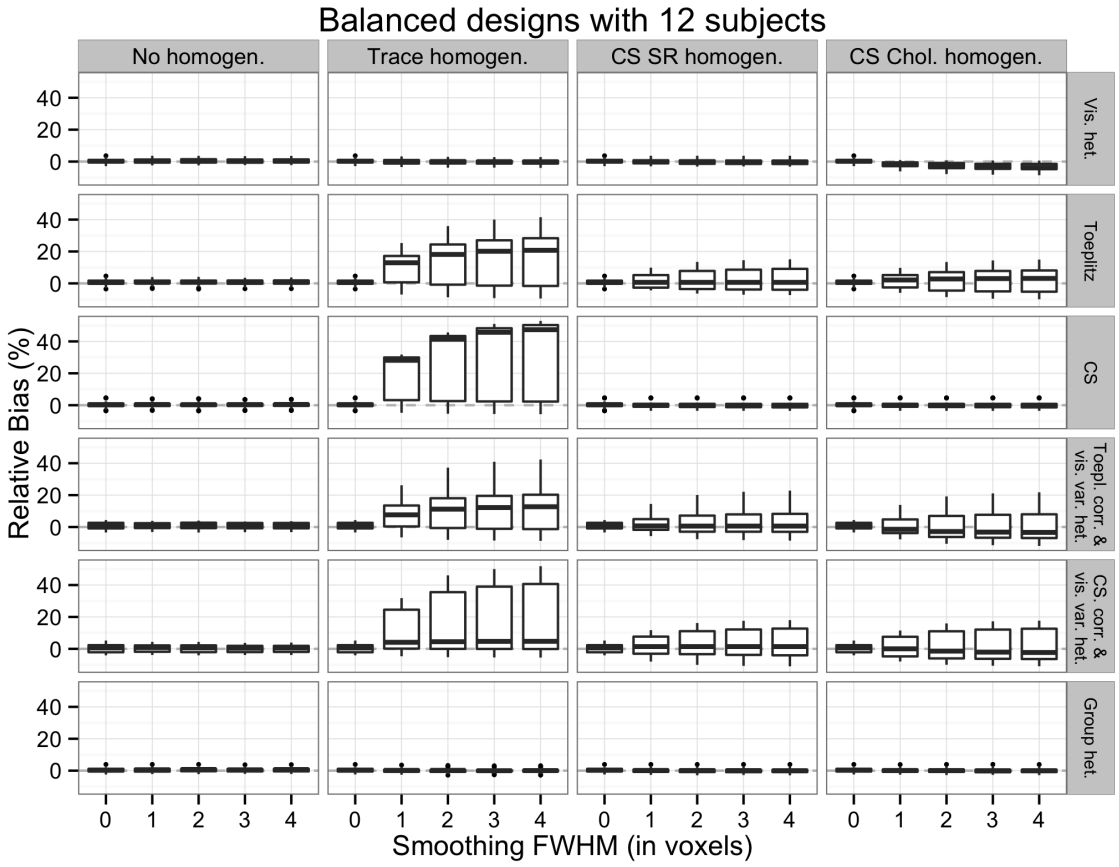

Fig. 6.2 Boxplots showing the effect of smoothing for the Euclidean metric in terms of the relative bias of several contrasted SwE in the balanced designs with 12 subjects over 27 scenarios (consisting of the 9 contrasts tested and the 3 numbers of visits per subject considered in the Monte Carlo simulations).

smoothing using the Euclidean metric. As expected, the strongest decreases were obtained for the cases without homogenisation. The homogenisation by the trace seemed to be able to reduce the variance for cross-sectional effects, but struggled with the longitudinal effects, even yielding an increase of variance in many scenarios that are likely attributable to the presence of a large positive bias. The homogenisations based on the square-root and Cholesky decompositions of the compound symmetric estimator were able to decrease the variance of longitudinal effects with approximatively the same extent, but far less strongly than the smoothing without homogenisation. Clearly, the latter is due to the fact that both of these homogenisations are asymptotically limited by the variability of the compound symmetric estimator. While the homogenisations based on the square-root and Cholesky decomposition had similar performance for

the longitudinal effects, their behaviours differed for the cross-sectional effects. For those effects, the homogenisation based on the Cholesky decomposition seemed to decrease the variance of the cross-sectional effects while the one based on the square-root decomposition could not. This difference can simply be explained by the fact that the homogenisation based on the square-root decomposition seemed unbiased for those effects while the one based on the Cholesky decomposition had the tendency to introduce a negative bias, which naturally tended to decrease the variances.

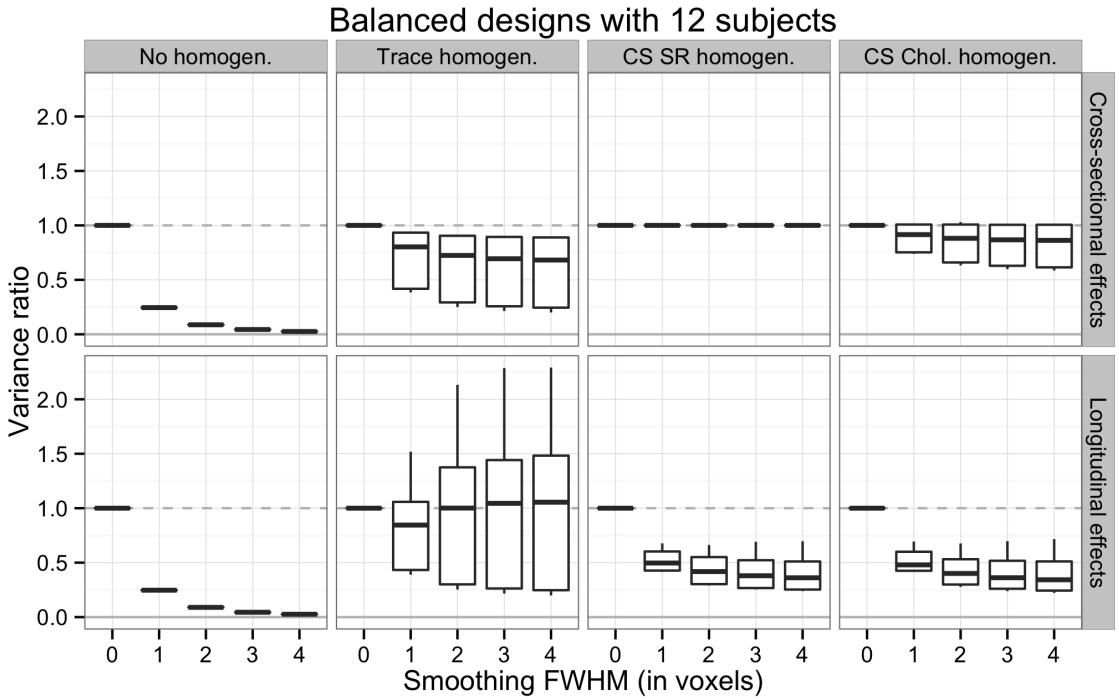

Fig. 6.3 Boxplots showing the effect of smoothing for the Euclidean metric in terms of the variance ratio (defined as the ratio between the variance of $CSC^\top$ after smoothing and the one before smoothing) of several contrasted SwE in the balanced designs with 12 subjects over 27 scenarios (consisting of the 9 contrasts tested and the 3 numbers of visits per subject considered in the Monte Carlo simulations).

6.4 Conclusion

In this chapter, we have studied the possibility to spatially smooth the common covariance matrices $\hat{V}_{0g}$ used in the SwE in order to decrease its variability. More precisely, inspired by the work of Dryden et al. (2009), we have investigated the use of four smoothing metrics. Moreover, we have also proposed three forms of smoothing based

on three type of homogenisation of the covariance matrices and assessed them in combination of the four smoothing metrics using Monte Carlo simulations in simple spatial scenarios.

From the Monte Carlo simulations, it appeared clearly that the best metric was the Euclidean metric. This is in total contradiction with the results found in Dryden et al. (2009) where the authors found that the three other metrics were outperforming the Euclidean metric in their simulations. Nevertheless, it seems that the sample covariance matrices simulated in Dryden et al. (2009) were not representative of those we encounter in the context of the SwE method. Indeed, taking the expectation of the models they used to produce the sample covariance matrices, a positive bias term seems to systematically appear for the diagonal elements, meaning that the sample covariance matrices they generated were positively biased estimates of the population covariance matrix. Now, for the sake of understanding why the results differ, we can imagine to add a positive bias in the results of Figure 6.1. Consequentely, the results obtained with the Cholesky, Root-Euclidean and Log-Euclidean metrics would appear better than they are while the results obtained with the Euclidean metric would appear worse than they are. This may simply explains the contradiction existing between the results of this thesis with those in Dryden et al. (2009). Nevertheless, a more rigorous verification might be needed to confirm this explanation. On this, it is important to note that, in the context of the SwE method, we try hard to get sample covariance matrices as unbiased as possible and, as such, the conclusion of our simulations seems to be more trustworthy than those in Dryden et al. (2009) in our context. Nevertheless, this comment might not be valid in other context like the one of DTI investigated in Dryden et al. (2009) where the sample covariance matrices can be different from those encountered in this thesis. Moreover, the results obtained in this thesis are only valid for the averaging of unbiased sample covariance matrices coming from the same distribution. In other applications like, for example, the interpolation of covariance matrices, it would be surprising to find the Euclidean metric to be a good metric.

Even if the Euclidean metric was the best metric in our simulations, it seemed to perform well only when no homogenisation was used. As, for a real longitudinal neuroimaging dataset, it is unlikely that the covariance matrices would be spatially homogeneous, it seems essential to use some form of homogenisation before smoothing. Unfortunately, some non-negligible bias was observed in the smooth SwE when one of the three proposed homogenisations was used. This is likely to be an issue if a parametric test is used to make inference. Moreover, in order to use a parametric test, it is important to estimate accurately the effective number of degrees of freedom

after smoothing. While we have suggested some ways to achieve this in Section 6.2.4, this seems to be a very challenging and error prone task as the effective number of degrees of freedom after smoothing typically depends of several factors such as the homogenisation, which is generally non-linear, and the degree of smoothness initially present in the data. Therefore, a more promising strategy to make inference would be to use the Wild Bootstrap introduced in Chapter 4 which appeared to be relatively robust against the presence of bias in the SwE and does not need any estimation of degrees of freedom due to its non-parametric nature. Nevertheless, further investigations are required to check this in the simple spatial scenarios investigated here, but also in more complicated scenarios with more realistic spatial correlations and spatially heterogeneous covariance matrices. We leave this as a future work.

Finally, it seems evident that other metrics or homogenisations than those investigated in this chapter could be considered. For example, as suggested by Gerard Ridgway, one could homogenise the covariance matrices by dividing them by their respective determinants. This would be particularly relevant for the log-euclidean metric which is known to smooth the determinant (Arsigny et al., 2007).

Chapter 7

Discussion

Inspired by the growing importance of longitudinal neuroimaging studies and the need for more appropriate tools to analyse the data obtained from such studies, the initial goal of this thesis was to improve the analysis of longitudinal neuroimaging data. Our main proposition to achieve this has been the use of the SwE method due to its appealing simplicity, its robustness against misspecifications and the fact that it is free of iterative algorithms (thus, fast and without convergence failure).

In Chapter 3, we have reviewed and proposed many adjustments of the SwE method to improve its behaviour, specifically in small samples. In particular, we have proposed three novel parametric statistical tests that showed promising results. Using Monte Carlo simulations, we have isolated the best variants of the SwE method and have shown their strengths and weaknesses compared to other popular approaches such as the N-OLS, SS-OLS and LME methods.

While the SwE method exhibits many advantages, we have however noticed some limitations. First, in scenarios with missing data and where the covariances were close to the variances, the homogeneous SwE had the tendency to yield conservative inference (see Section 3.3.1), likely due to a bias induced by the correction (consisting of zeroing the negative eigenvalues) made on non-positive semi-definite covariance matrix estimates. While this does not break the validity of the method, it would be desirable to find some solutions to adjust for the observed conservativeness. One solution would be to use the heterogeneous SwE instead, as it did not appear affected by the bias observed for the homogeneous version. Nevertheless, the heterogeneous SwE seemed to struggle to control the FPR when the number of subjects was small, indicating that it could be a good alternative only for moderate or large sample sizes. Another solution would be to decrease the effective number of missing data per homogeneous group by splitting the problematic homogeneous groups into sub-homogeneous groups

with less missing data per group. While the latter strategy will increase the number of homogeneous group and yield groups with fewer subjects, it should make the estimator $\hat{V}_{0g}$ (see Section 3.2.3) be less prone to non-positive semi-definite covariance matrix estimation. Further research might be useful to check if better solutions exists to solve this issue.

A second limitation regards the power of the SwE method, which may be lower than alternative methods (see Section 3.3.2). This is somehow the reasonable price to pay for all the advantages that the SwE method has compared to other alternative methods (e.g., no iterative algorithms, accurate inferences in scenarios where other methods are inaccurate). However, this is probably the main disadvantage of the method and some propositions to improve this have been investigated in Chapter 5 and 6. Unfortunately, the shrinkage SwE (Chapter 5) and the smooth SwE (Chapter 6), while showing some promising results for the reduction of estimation error, were typically characterised by the introduction of bias which unfortunately prevents the use of valid parametric inference. Nevertheless, some hope exists that the Wild Bootstrap investigated in Chapter 4 might be combined with one of these two new SwE variants and yields accurate and more powerful inferences than with a more traditional SwE. Indeed, the results obtained in Chapter 4 seems to indicate that the the Wild Bootstrap procedure is robust against the presence of bias (in the resampling or in the SwE), indicating that it might be robust against the bias observed in the shrinkage SwE or in the smooth SwE. Nevertheless, further research is needed to investigate this in more details.

Another way which could potentially improve the power of the SwE method would be to use, for each subject, a non-identity working covariance matrix W_i that is to be estimated alongside the other parameters using Generalised Estimating Equations (Liang and Zeger, 1986) as suggested by Li et al. (2013) in the context of neuroimaging. However, this procedure typically requires the use of an iterative algorithm and, due to the random nature of the working covariance matrices, the parametric tests developed in Section 3.2.4 might not be valid. Alternatively, we could define a "poor man's Generalised Estimating Equations" procedure by estimating the working covariance matrices in a first pass and then use them as non-random working covariance matrices, in which case no iterative algorithm would be needed and the parametric tests developed in Section 3.2.4 might be valid. One possible way to achieve this would be to estimate first the subject covariance matrices separately at every voxel as it is performed for the heterogeneous or homogeneous SwE, apply some form of smoothing to the resulting covariance matrix image (see Chapter 6 for some smoothing procedure examples) to reduce the variability of the estimates enough to treat

them as non-random, and finally used the resulting smooth estimates as working covariance matrices. As suggested by Gerard Ridgway, a second option would be to use the global covariance matrix structure estimate obtained by the SPM procedure (see Section 2.3.4) and use it to define the working covariance matrices, which would then be identical across voxels. Further research is however needed to validate this "poor man's Generalised Estimating Equations" approach.

A third limitation is the current impossibility with the SwE method to control for a FWER with a parametric test. A possible solution would be to validate the use of Random Field Theory (see, e.g., Worsley et al., 1996) to achieve this like it is done with more standard neuroimaging models. However, further research is needed for this validation. Thus, so far, the alternative is to parametrically control the FDR or to use the Wild Bootstrap to make a non-parametric control of the FWER.

Another possible improvement of the SwE method concerns the use of the heterogeneous SwE in small samples. In such scenarios, the inferences obtained through Test I and Test III (which are equivalent for a heterogeneous SwE) appeared to be conservative. This misbehaviour can simply be explained by the use of biased estimators for $\text{Cov}[\text{vec}[\hat{V}_i]]$ and $\text{Cov}[\text{vec}[CSC']]$. While we have attempted to solve this issue in Test II, the inferences had the tendency to be liberal in very small samples. A possible explanation for this is the assumption that the pure within-subject covariates do not affect the effective number of degrees of freedom. While this seems to be a reasonable assumption for many sample sizes (even moderately small), in very small samples, it seems that the their influence might become non-negligible, explaining the observed liberality. Nevertheless, further work is needed to check this explanation and find an appropriate solution to correct for this liberality in very small samples. Note, however, that this liberality was not present when a homogeneous SwE was used, indicating that such kind of SwE might be preferred in practice in very small samples. As a reminder, this accuracy of the homogeneous SwE can be explained by the fact that the bias terms that need to be adjusted are inherently smaller for a homogeneous SwE than for a heterogeneous SwE as they are approximatively inversely proportional to the number of subjects per homogeneous group.

An aspect of longitudinal data, which has not be studied in this thesis, is about the process behind the missing data and its effect on the SwE method. We have always assumed that the data is, using the terminology of Little and Rubin (2002), missing completely at random (i.e. the probability of missingness does not depend on observed or unobserved data). If the process of missingness is more complicated such as, using the terminology of Little and Rubin (2002), missing at random (i.e., given

the observed data, the probability of missingness does not depend on the unobserved data) or missing not at random (i.e., even given the observed data, the probability of missingness depends on the unobserved data), the SwE method might yield inaccurate results. Nevertheless, there is a possibility to modify the SwE method to handle data which is missing at random. Indeed, Robins et al. (1995) proposed a modification of Generalised Estimating Equations, generally referred to as Weighted Generalised Estimating Equations, which can handle data that is missing at random by weighting the data according to their probability of missingness. As the SwE method can be seen as a particular case of Generalised Estimating Equations, it should therefore be relatively easy to modify the methods developed in this thesis to handle data that is missing at random. Nonetheless, further research would be needed to check this.

While we implemented the SwE method into an `SPM` toolbox and made it freely available at http://warwick.ac.uk/tenichols/SwE, the last release was based on an early work that was published in Guillaume et al. (2014). Since then, additional research has been conducted and has notably allowed the discovery of a more accurate bias-adjustment (i.e. S_{C2}) and more accurate parametric statistical tests (i.e. Test II and Test III). Also, the last release did not implement the use of the WB for non-parametric inference. However, for the purpose of analysing the ADNI dataset, these features have actually been implemented in a non-released version of the toolbox, but unfortunately in a non-user-friendly format. Therefore, we intend to modify the implementation of these features in a more user-friendly format before the next release of the toolbox.

An important aspect of statistical analyses which was missing in this thesis and, more generally, often neglected in neuroimaging is the use of model diagnostics that check the assumption of the model used (e.g., the assumption of multivariate normality of the error terms) or detect potential outliers. Diagnostic tools specifically developed for the SwE method seems unfortunately missing in the literature, but, it does not seem too complicated to adapt existing diagnostic tools developed for other models like the Generalised Linear Model. Such tools could, in turn, be implemented in the SwE toolbox that we have released and be part of the whole analysis process.

Finally, while, in this thesis, we have focused our attention on the use of the SwE method to analyse specifically longitudinal neuroimaging data, it is worth noting that it could also be used for many other types of data, in neuroimaging or in other fields. For example, in the neuroimaging context, it could also be used to analyse repeated-measures data, family data (where subjects from the same family cannot be assumed independent) or even cross-sectional data where the assumption of homogeneous error

terms may not hold. In particular, it could be an alternative option to multivariate linear regression models that are sometimes used in neuroimaging (see, e.g. Naylor et al., 2014). It would be, however, interesting to compare them, particularly in term of the quality of inference, to see what would be their respective advantages and disadvantages.

References

Abadir, K. M. and Magnus, J. R. (2005). *Matrix algebra*, volume 1. Cambridge University Press.

Andrews, E., Frigau, L., Voyvodic-Casabo, C., Voyvodic, J., and Wright, J. (2013). Multilingualism and fmri: Longitudinal study of second language acquisition. *Brain Sciences*, 3(2):849–876.

Arsigny, V., Fillard, P., Pennec, X., Ayache, N., et al. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347.

Ashburner, J. and Friston, K. J. (2003). Morphometry. *Human brain function*, pages 707–722.

Ashburner, J. and Friston, K. J. (2007a). Non-linear registration. *Statistical parametric mapping: The analysis of functional brain images*, pages 63–80.

Ashburner, J. and Friston, K. J. (2007b). Rigid body registration. *Statistical parametric mapping: The analysis of functional brain images*, pages 49–62.

Ashburner, J. and Friston, K. J. (2007c). Segmentation. *Statistical parametric mapping: The analysis of functional brain images*, pages 81–91.

Ashburner, J. and Friston, K. J. (2007d). Voxel-based morphometry. *Statistical parametric mapping: The analysis of functional brain images*, pages 92–98.

Bangert, M. and Altenmüller, E. O. (2003). Mapping perception to action in piano practice: a longitudinal dc-eeg study. *BMC neuroscience*, 4(1):26.

Bates, D., Maechler, M., and Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-0.

Bell, R. M. and McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182.

Belmonte, M. and Yurgelun-Todd, D. (2001). Permutation testing made practical for functional magnetic resonance image analysis. *Medical Imaging, IEEE Transactions on*, 20(3):243–248.

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

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Bernal-Rusiel, J. L., Greve, D. N., Reuter, M., Fischl, B., and Sabuncu, M. R. (2013a). Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage*, 66:249–260.

Bernal-Rusiel, J. L., Reuter, M., Greve, D. N., Fischl, B., and Sabuncu, M. R. (2013b). Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. *NeuroImage*, 81(0):358 – 370.

Box, G. E. (1950). Problems in the analysis of growth and wear curves. *Biometrics*, 6(4):362–389.

Brownstone, D. and Valletta, R. (2001). The bootstrap and multiple imputations: harnessing increased computing power for improved statistical tests. *Journal of Economic Perspectives*, pages 129–141.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., and Cox, R. W. (2013). Linear mixed-effects modeling approach to fmri group analysis. *NeuroImage*.

Chesher, A. and Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica: Journal of the Econometric Society*, pages 1217–1222.

Chung, S., Lu, Y., and Henry, R. G. (2006). Comparison of bootstrap approaches for estimation of uncertainties of dti parameters. *NeuroImage*, 33(2):531–541.

Davidson, J., Monticini, A., and Peel, D. (2007). Implementing the wild bootstrap using a two-point distribution. *Economics Letters*, 96(3):309–315.

Davidson, R. and Flachaire, E. (2001). The wild bootstrap, tamed at last. Technical report.

Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169.

Davidson, R., MacKinnon, J. G., and Davidson, R. (1985). Heteroskedasticity-robust tests in regressions directions. In *Annales de l'INSEE*, pages 183–218. JSTOR.

Davis, K. L., Buchsbaum, M. S., Shihabuddin, L., Spiegel-Cohen, J., Metzger, M., Frecska, E., Keefe, R. S., and Powchik, P. (1998). Ventricular enlargement in poor-outcome schizophrenia. *Biological psychiatry*, 43(11):783–793.

De Boissezon, X., Démonet, J.-F., Puel, M., Marie, N., Raboyeau, G., Albucher, J.-F., Chollet, F., and Cardebat, D. (2005). Subcortical aphasia a longitudinal pet study. *Stroke*, 36(7):1467–1473.

Diggle, P., Liang, K., and Zeger, S. (1994). Analysis of longitudinal data oxford statistical science series 13.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., and May, A. (2004). Neuroplasticity: changes in grey matter induced by training. *Nature*, 427(6972):311–312.

Dryden, I., Koloydenko, A., and Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123.

Dubbelink, K. T. O., Hillebrand, A., Stoffers, D., Deijen, J. B., Twisk, J. W., Stam, C. J., and Berendse, H. W. (2014). Disrupted brain network topology in parkinson's disease: a longitudinal magnetoencephalography study. *Brain*, 137(1):197–207.

Dubbelink, K. T. O., Stoffers, D., Deijen, J. B., Twisk, J. W., Stam, C. J., Hillebrand, A., and Berendse, H. W. (2013). Resting-state functional connectivity as a marker of disease progression in parkinson's disease: A longitudinal meg study. *NeuroImage: Clinical*, 2(Complete):612–619.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. In *CBMS-NSF regional conference series in applied mathematics (*, number 38. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. and Tibshirani, R. J. (1994). An introduction to the bootstrap (chapman & hall/crc monographs on statistics & applied probability).

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, 34(2):447–456.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82. University of California Press Berkeley.

Fay, M. and Graubard, B. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*, 57(4):1198–1206.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, pages 309–317.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, 49(2):361–376.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

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

Halekoh, U. and Højsgaard, S. (2013). *pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison*. R package version 0.3-8.

Hamer, R. and Simpson, P. (1999). Mixed-up models: Things that look like they should work but don't, and things that look like they shouldn't work but do. In *SAS User's Group International conference*.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hansen, P., Kringelbach, M., and Salmelin, R. (2010). *MEG: An introduction to methods*. Oxford university press.

Hardin, J. (2001). Small sample adjustments to the sandwich estimate of variance. http://www.stata.com/support/faqs/stat/sandwich.html.

Härdle, W. K. and Simar, L. (2012). *Applied multivariate statistical analysis*. Springer.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, pages 320–338.

Henson, R., Buechel, C., Josephs, O., and Friston, K. (1999). The slice-timing problem in event-related fmri. *NeuroImage*, 9:125.

Hinkley, D. (1977). Jackknifing in unbalanced situations. *Technometrics*, pages 285–292.

Holmes, A. (1994). *Statistical issues in functional brain mapping.* PhD thesis, University of Glasgow.

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

Hsieh, J. (2009). Computed tomography: principles, design, artifacts, and recent advances. SPIE Bellingham, WA.

Hua, X., Hibar, D. P., Ching, C. R., Boyle, C. P., Rajagopalan, P., Gutman, B. A., Leow, A. D., Toga, A. W., Jr., C. R. J., Harvey, D., Weiner, M. W., and Thompson, P. M. (2013). Unbiased tensor-based morphometry: Improved robustness and sample size estimates for alzheimer's disease clinical trials. *NeuroImage*, 66(0):648 – 661.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–33.

Huppert, T. J., Diamond, S. G., Franceschini, M. A., and Boas, D. A. (2009). Homer: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied optics*, 48(10):D280–D298.

Illowsky, B. P., Juliano, D. M., Bigelow, L. B., and Weinberger, D. R. (1988). Stability of ct scan findings in schizophrenia: results of an 8 year follow-up study. *Journal of Neurology, Neurosurgery & Psychiatry*, 51(2):209–213.

Jaskiw, G. E., Juliano, D. M., Goldberg, T. E., Hertzman, M., Urow-Hamell, E., and Weinberger, D. R. (1994). Cerebral ventricular enlargement in schizophreniform disorder does not progress a seven year follow-up study. *Schizophrenia research*, 14(1):23–28.

Kauermann, G. and Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, pages 983–997.

Kiebel, S. and Holmes, A. (2007). The general linear model. *Statistical parametric mapping: The analysis of functional brain images*, pages 101–125.

Kim, J.-H., Lee, J.-M., Kang, E., Kim, J. S., Song, I. C., and Chung, C. K. (2010). Functional reorganization associated with semantic language processing in temporal lobe epilepsy patients after anterior temporal lobectomy: a longitudinal functional magnetic resonance image study. *Journal of Korean Neurosurgical Society*, 47(1):17–25.

Kono, T., Matsuo, K., Tsunashima, K., Kasai, K., Takizawa, R., Rogers, M. A., Yamasue, H., Yano, T., Taketani, Y., and Kato, N. (2007). Multiple-time replicability of near-infrared spectroscopy recording during prefrontal activation task in healthy men. *Neuroscience research*, 57(4):504–512.

Kwan, C. C. (2008). Estimation error in the average correlation of security returns and shrinkage estimation of covariance and correlation matrices. *Finance Research Letters*, 5(4):236–244.

Kwan, C. C. (2011). An introduction to shrinkage estimation of the covariance matrix: a pedagogic illustration. *Spreadsheets in Education (eJSiE)*, 4(3):6.

Lai, T. L. and Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):79–99.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621.

Li, Y., Gilmore, J. H., Shen, D., Styner, M., Lin, W., and Zhu, H. (2013). Multiscale adaptive generalized estimating equations for longitudinal neuroimaging data. *NeuroImage*.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lindquist, M., Spicer, J., Asllani, I., and Wager, T. (2012). Estimating and testing variance components in a multi-level glm. *Neuroimage*, 59(1):490–501.

Lipsitz, S., Ibrahim, J., and Parzen, M. (1999). A degrees-of-freedom approximation for a t-statistic with heterogeneous variance. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 48(4):495–506.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

Liu, R. Y. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708.

Long, J. and Ervin, L. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, pages 217–224.

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

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285.

Mancl, L. and DeRouen, T. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1):126–134.

McDonald, B. W. (1993). Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 391–397.

Meltzer, J. A., Postman-Caucheteux, W. A., McArdle, J. J., and Braun, A. R. (2009). Strategies for longitudinal neuroimaging studies of overt language production. *Neuroimage*, 47(2):745–755.

Molenberghs, G. and Verbeke, G. (2011). A note on a hierarchical interpretation for negative variance components. *Statistical Modelling*, 11(5):389–408.

Moriguchi, Y. and Hiraki, K. (2011). Longitudinal development of prefrontal function during early childhood. *Developmental cognitive neuroscience*, 1(2):153–162.

Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J., Toga, A., and Beckett, L. (2005). The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869.

Mumford, J. A. and Nichols, T. (2009). Simple group fmri modeling and inference. *Neuroimage*, 47(4):1469–1475.

Nasrallah, I. and Dubroff, J. (2013). An overview of pet neuroimaging. In *Seminars in nuclear medicine*, volume 43, pages 449–461. Elsevier.

Naylor, M. G., Cardenas, V. A., Tosun, D., Schuff, N., Weiner, M., and Schwartzman, A. (2014). Voxelwise multivariate analysis of multimodality magnetic resonance imaging. *Human brain mapping*, 35(3):831–846.

Nel, D. and Van der Merwe, C. (1986). A solution to the multivariate behrens-fisher problem. *Communications in Statistics-Theory and Methods*, 15(12):3719–3735.

Neuhaus, J. and Kalbfleisch, J. (1998). Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, pages 638–645.

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446.

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

Niedermeyer, E. and Da Silva, F. L. (2005). *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.

Ogawa, S., Lee, T., Kay, A., and Tank, D. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872.

Ossenkoppele, R., Tolboom, N., Foster-Dingley, J. C., Adriaanse, S. F., Boellaard, R., Yaqub, M., Windhorst, A. D., Barkhof, F., Lammertsma, A. A., Scheltens, P., et al. (2012). Longitudinal imaging of alzheimer pathology using [11c] pib,[18f] fddnp and [18f] fdg pet. *European journal of nuclear medicine and molecular imaging*, 39(6):990–1000.

Pan, W. (2001). On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3):901–906.

Pan, W. and Wall, M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in medicine*, 21(10):1429–1441.

Pantazis, D., Nichols, T. E., Baillet, S., and Leahy, R. M. (2005). A comparison of random field theory and permutation methods for the statistical analysis of meg data. *Neuroimage*, 25(2):383–394.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation*, 23(4):939–951.

Petersson, K. M., Nichols, T. E., Poline, J.-B., and Holmes, A. P. (1999). Statistical limitations in functional neuroimaging ii. signal detection and statistical inference. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 354(1387):1261–1281.

Pinheiro, J. and Bates, D. (2000). Mixed-effects models in s and s-plus. *Statistics and Computing. Springer-Verlag, Berlin, D.*

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-113.

Prince, M., Albanese, E., Guerchet, M., and Prina, M. (2014a). World alzheimer report 2014: Dementia and risk reduction – an analysis of protective and modifiable factors. *Londres: Alzheimers Disease International*.

Prince, M., Knapp, M., Guerchet, M., McCrone, P., Prina, M., Comas-Herrera, A., Wittenberg, R., Adelaja, B., Hu, B., King, D., et al. (2014b). Dementia uk: Second edition – overview.

Ridgway, G., Leung, K., and Ashburner, J. (2015). Computing brain change over time. In Toga, A. W., editor, *Brain Mapping*, pages 417 – 428. Academic Press, Waltham.

Ridgway, G. R. (2009). *Statistical analysis for longitudinal MR imaging of dementia*. PhD thesis, UCL (University College London).

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Roche, A. (2011). A four-dimensional registration algorithm with application to joint correction of motion and slice timing in fmri. *Medical Imaging, IEEE Transactions on*, 30(8):1546–1554.

Saggar, M., King, B. G., Zanesco, A. P., MacLean, K. A., Aichele, S. R., Jacobs, T. L., Bridwell, D. A., Shaver, P. R., Rosenberg, E. L., Sahdra, B. K., et al. (2012). Intensive training induces longitudinal changes in meditation state-related eeg oscillatory activity. *Frontiers in human neuroscience*, 6.

Salimi-Khorshidi, G., Smith, S. M., and Nichols, T. E. (2011). Adjusting the effect of nonstationarity in cluster-based and tfce inference. *Neuroimage*, 54(3):2006–2019.

Scarpazza, C., Sartori, G., De Simone, M., and Mechelli, A. (2013). When the single matters more than the group: very high false positive rates in single case voxel based morphometry. *Neuroimage*, 70:175–188.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

Seppänen, M., Pesonen, A.-K., and Tervaniemi, M. (2012). Music training enhances the rapid plasticity of p3a/p3b event-related brain potentials for unattended and attended target sounds. *Attention, Perception, & Psychophysics*, 74(3):600–612.

Skup, M., Zhu, H., and Zhang, H. (2012). Multiscale adaptive marginal analysis of longitudinal neuroimaging data with time-varying covariates. *Biometrics*.

Sturm, W., Longoni, F., Weis, S., Specht, K., Herzog, H., Vohn, R., Thimm, M., and Willmes, K. (2004). Functional reorganisation in patients with right hemisphere stroke after training of alertness: a longitudinal pet and fmri study in eight cases. *Neuropsychologia*, 42(4):434–450.

Thompson, W. K., Hallmayer, J., and O'Hara, R. (2011). Design considerations for characterizing psychiatric trajectories across the lifespan: Application to effects of apoe-e4 on cerebral cortical thickness in alzheimer's disease. *American Journal of Psychiatry*, 168(9):894–903.

Tsujii, T., Yamamoto, E., Masuda, S., and Watanabe, S. (2009). Longitudinal study of spatial working memory development in young children. *Neuroreport*, 20(8):759–763.

Van Dellen, E., Douw, L., Hillebrand, A., De Witt Hamer, P. C., Baayen, J. C., Heimans, J. J., Reijneveld, J. C., and Stam, C. J. (2014). Epilepsy surgery outcome and functional network alterations in longitudinal meg: A minimum spanning tree analysis. *NeuroImage*, 86:354–363.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer.

Villemagne, V. L., Furumoto, S., Fodero-Tavoletti, M. T., Mulligan, R. S., Hodges, J., Harada, R., Yates, P., Piguet, O., Pejoska, S., Doré, V., et al. (2014). In vivo evaluation of a novel tau imaging tracer for alzheimer's disease. *European journal of nuclear medicine and molecular imaging*, 41(5):816–826.

Viviani, R., Beschoner, P., Ehrhard, K., Schmitz, B., and Thöne, J. (2007). Non-normality and transformations of random fields, with an application to voxel-based morphometry. *NeuroImage*, 35(1):121–130.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Waldorp, L. (2009). Robust and unbiased variance of glm coefficients for misspecified autocorrelation and hemodynamic response models in fmri. *Journal of Biomedical Imaging*, 2009:15.

Wang, Z., Vemuri, B., Chen, Y., and Mareci, T. (2004). A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from complex dwi. *Medical Imaging, IEEE Transactions on*, 23(8):930–939.

Warton, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics*, 67(1):116–123.

Webb, M. D. (2013). Reworking wild bootstrap based inference for clustered errors. Technical report, Queen's Economics Department Working Paper.

West, B., Welch, K. B., and Galecki, A. T. (2006). *Linear mixed models: a practical guide using statistical software*. CRC Press.

Whitcher, B., Tuch, D., and Wang, L. (2005). The wild bootstrap to quantify variability in diffusion tensor mri. In *Proceedings of ISMRM Annual Meeting*.

Whitcher, B., Tuch, D. S., Wisco, J. J., Sorensen, A. G., and Wang, L. (2008). Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging. *Human brain mapping*, 29(3):346–362.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92:381–397.

Woods, B. T., Yurgelun-Todd, D., Benes, F. M., Frankenburg, F. R., Pope Jr, H. G., and McSparren, J. (1990). Progressive ventricular enlargement in schizophrenia: comparison to bipolar affective disorder and correlation with clinical course. *Biological psychiatry*, 27(3):341–352.

Worsley, K. J., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fmri data. *Neuroimage*, 15(1):1–15.

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

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, pages 1261–1295.

Yoshimura, Y., Kikuchi, M., Ueno, S., Shitamichi, K., Remijn, G. B., Hiraishi, H., Hasegawa, C., Furutani, N., Oi, M., Munesue, T., et al. (2014). A longitudinal study of auditory evoked field and language development in young children. *NeuroImage*, 101:440–447.

Zhang, H. (2008). *Advances in Modeling and Inference of Neuroimaging Data*. PhD thesis, The University of Michigan.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 805–811.

Zhu, H., Ibrahim, J. G., Tang, N., Rowe, D. B., Hao, X., Bansal, R., and Peterson, B. S. (2007). A statistical analysis of brain morphology using wild bootstrapping. *Medical Imaging, IEEE Transactions on*, 26(7):954–966.

Zhu, T., Liu, X., Connelly, P. R., and Zhong, J. (2008). An optimized wild bootstrap method for evaluation of measurement uncertainties of dti-derived parameters in human brain. *Neuroimage*, 40(3):1144–1156.

Appendix A

Valorisation

According to the regulation governing the attainment of doctoral degrees at Maastricht University, an addendum about valorisation must be added to each doctoral thesis. This is the purpose of this appendix.

A.1 Introduction

As described in Chapter 1, the number of longitudinal neuroimaging studies has been increasing in recent years. This is not surprising as this kind of studies can help to study longitudinal changes occurring in the brain while this is not possible with cross-sectional studies. In particular, for the past few years, this type of studies has been increasingly used to study dementia, which affects an increasing number of people that is currently estimated at 44 million worldwide (Prince et al., 2014a) and around 850,000 in the United Kingdom (Prince et al., 2014b), and has an important cost on society which is currently estimated at US $604 billion a year worldwide (Prince et al., 2014a) and at £26 billion a year in the United Kingdom (Prince et al., 2014b). Unless some actions are taken, all these alarming figures are even set to rise in the next years due to the population ageing (Prince et al., 2014a,b). That is why, in recent years, several initiatives have been conducted with the goal to lessen the impact of dementia on individuals and society. One of such initiatives is the Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005) which has collected a large amount of longitudinal neuroimaging data (see, e.g., the dataset described in Section 2.6). In order to be effective, it is however essential to analyse the data obtained from these initiatives as accurately as possible. This was notably pointed out by the "World Alzheimer Report 2014" (Prince et al., 2014b, pages 93 & 94) that highlighted the importance of enhancing the quality and relevance of evidence from

observational studies. This thesis completely enters into this philosophy as its main goal is to improve the quality of the analysis of longitudinal neuroimaging data. In the remainder of this appendix, we summarise the actual impacts of this thesis to achieve this goal and discuss the potential impacts that this thesis may have on other type of studies as well as the importance of further research on this topic.

A.2 Thesis impact

The impact of this thesis on the enhancement of the quality of the analysis of longitudinal neuroimaging data can be divided into four main points that are described below.

A.2.1 Raising awareness about the limitations of current popular analysis methods

The first impact of this thesis has been to raise awareness about the limitations of popular analysis methods that are currently used to analyse longitudinal neuroimaging data. This has been achieved in Chapters 2 and 3 by discussing and evaluating them. This is important because many users of these methods are unfortunately unaware about their limitations and may use them when it is not appropriate. The latter can be very problematic as this may lead to very misleading conclusions that can, in turn, yield negative socio-economic impacts.

A.2.2 Proposition of alternative methods

The second impact of this thesis has been the proposition and evaluation of promising alternative methods. In particular, the Sandwich Estimator method investigated in Chapter 3 appeared to be accurate in much more scenarios than alternative methods, indicating that it might be a better choice to analyse longitudinal neuroimaging data. Also, the non-parametric Wild Bootstrap methodology investigated in Chapter 4 seemed to be a promising way to non-parametrically control for the Family-Wise Error Rate with longitudinal neuroimaging data.

A.2.3 Dissemination

A third impact of this thesis work regards the dissemination of the research results. Indeed, throughout this doctoral work, an important amount of time and effort has

been dedicated to communicate the results obtained for this thesis through the use of many poster presentations (e.g., at OHBM 2012, 2013 and 2014), many oral presentations (e.g., at OHBM 2012 and at the Reading Emotions Workshop 2014: Capturing Brain Changes Across the Lifespan) as well as a publication in NeuroImage (Guillaume et al., 2014).

A.2.4 Software

Finally, an important contribution of this doctoral work has been the implementation of the Sandwich Estimator method (see Chapter 3) into an `SPM` toolbox that has been made freely available at http://warwick.ac.uk/tenichols/SwE (see Figure A.1 for an overview of the toolbox user interface). An important effort has been made to make it easy to use by mimicking as much as possible a typical analysis made with the `SPM` software package. Additional features for the toolbox are currently under implementation and are expected to be available soon (e.g., the Wild Bootstrap methodology proposed and studied in Chapter 4).

A.3 Further perspectives

While, in this thesis, we have specifically focused on the analysis of longitudinal neuroimaging data, the results obtained could also be useful for the analysis of other type of neuroimaging data like repeated-measures data, family data or cross-sectional data where the subject variances cannot be assumed homogeneous. Furthermore, the results could also be useful for the analysis of such kind of data, but not obtained from neuroimaging. This means that the potential socio-economic impact of this thesis can be wider than it could appear at first thought.

To finish this valorisation appendix, it seems important to note that much more research could be done to improve further the quality of the analysis of longitudinal neuroimaging data. For example, there is an obvious lack of diagnostic tools that are able to check the validity of the model used to analyse the data. While this kind of tools exists for other type of data and their use are considered as a mandatory step in the analysis process, this seems lacking in neuroimaging, particularly for longitudinal data. We can therefore imagine the importance that future research on this thesis topic may have to enhance further the quality of the analysis of longitudinal neuroimaging data.

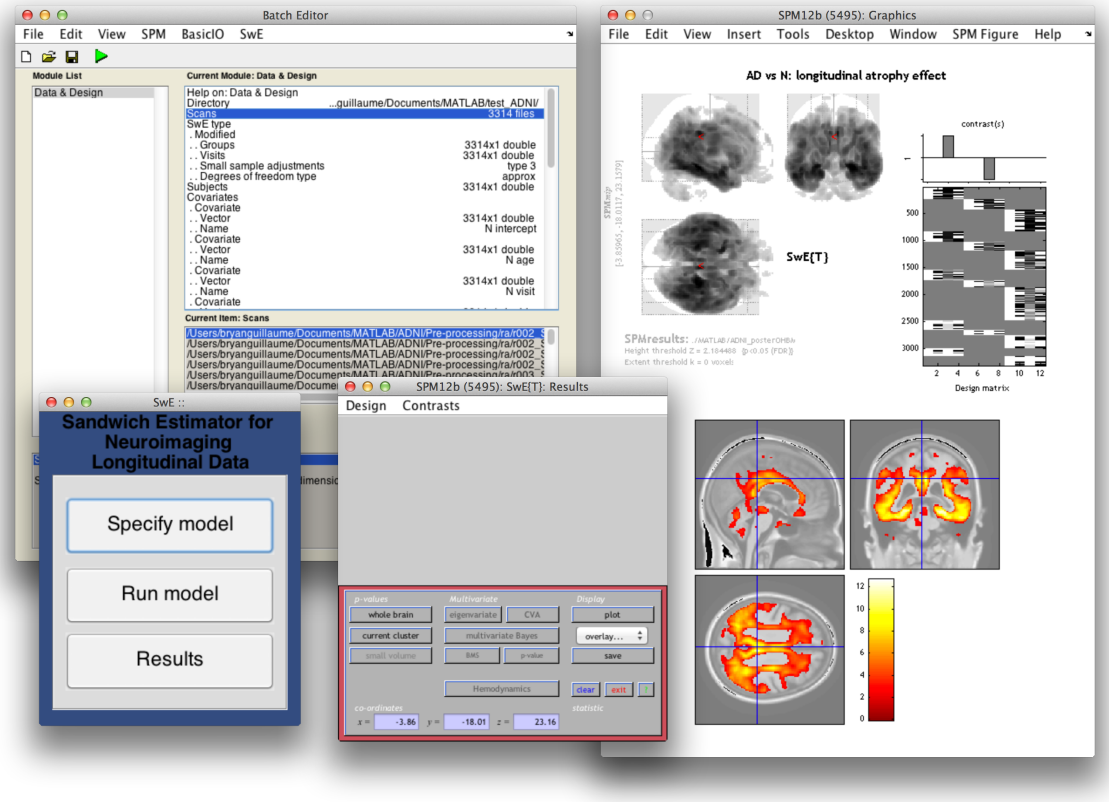

Fig. A.1 User interface of the SwE toolbox. Bottom right: the main interface window, top right: the batch system used to specify the model, middle and left: interface windows for the analysis of results.

Curriculum Vitae

Bryan Guillaume was born on September 15th, 1983 in Seraing, Belgium. In 2003, he graduated with a Bachelor's degree in engineering from Liège University (Belgium). In 2005, he graduated with a Master's degree in general engineering from the Ecole Centrale Paris (France). In 2007, he graduated with a Master's degree in mechatronics engineering from Liège University (Belgium). From 2007 to 2010, he worked as a project engineer at the company CMI (Belgium) in the context of thermal processes. In 2011, he moved to UK to begin his doctoral program as part of the EU FP7 Marie Curie Initial Training Network "Neurophysics" at GlaxoSmithKline and was based at Warwick University.