UMEÅ UNIVERSITY

Methods for Longitudinal Brain Imaging Studies with Dropout

Tetiana Gorbach

Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå 2019

Doctoral Thesis
Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå University
SE-901 87 Umeå

Statistical Studies No. 54
ISBN: 978-91-7855-011-1
ISSN: 1100-8989
Cover photo: Sunrise in the Carpathians, Ukraine, by Oksana Iakovenko
Electronic version available at http://umu.diva-portal.org/

Printed by: UmU Print Service, Umeå University
Umeå, Sweden 2019

Contents

List of papers

The thesis is based on the following papers:

I. Gorbach, T., Pudas, S., Lundquist, A., Orädd, G., Josefsson, M., Salami, A., de Luna, X. and Nyberg, L. (2017). Longitudinal association between hippocampus atrophy and episodic-memory decline. *Neurobiology of Aging*, 51: 167-176.

II. Gorbach, T. and de Luna, X. (2018). Inference for partial correlation when data are missing not at random. *Statistics & Probability Letters*, 141: 82-89.

III. Gorbach, T., Lundquist, A., de Luna, X., Nyberg, L. and Salami, A. (2019). A Hierarchical Bayesian Mixture Modeling Approach for Analysis of Resting-State Functional Brain Connectivity: An Alternative to Thresholding. *Manuscript*

IV. Gorbach, T., Lundquist, A., de Luna, X., Nyberg, L. and Salami, A. (2019). Bayesian mixture modeling for longitudinal fMRI connectivity studies with dropout. *Manuscript*.

Abstract

One of the challenges in aging research is to understand the brain mechanisms that underlie cognitive development in older adults. Such aging processes are investigated in longitudinal studies, where the within-individual changes over time are observed. However, several methodological issues exist in longitudinal analyses. One of them is loss of participants to follow-up, which occurs when individuals drop out from the study. Such dropout should be taken into account for valid conclusions from longitudinal investigations, and this is the focus of this thesis. The developed methods are used to explore brain aging and its relation to cognition within the Betula longitudinal study of aging.

Papers I and II consider the association between changes in brain structure and cognition. In the first paper, regression analysis is used to establish the statistical significance of brain-cognition associations while accounting for dropout. Paper II develops interval estimators directly for an association as measured by partial correlation, when some data are missing. The estimators of Paper II may be used in longitudinal as well as cross-sectional studies and are not limited to brain imaging.

Papers III and IV study functional brain connectivity, which is the statistical dependency between the functions of distinct brain regions. Typically, only brain regions with associations stronger than a predefined threshold are considered connected. However, the threshold is often arbitrarily set and does not reflect the individual differences in the overall connectivity patterns. Paper III proposes a mixture model for brain connectivity without explicit thresholding of associations and suggests an alternative connectivity measure. Paper IV extends the mixture modeling of Paper III to a longitudinal setting with dropout and investigates the impact of ignoring the dropout mechanism on the quality of the inferences made on longitudinal connectivity changes.

KEYWORDS: missing data, nonignorable dropout, sensitivity analysis, uncertainty intervals, pattern-mixture models, aging, cognition, MRI, brain structure, resting-state functional connectivity.

Sammanfattning

En stor utmaning inom forskning om åldrande är att förstå vilka mekanismer i hjärnan som ligger bakom kognitiv förändring. Longitudinella studier, där man följer deltagarna över tid och mäter deras individuella förändringar, är viktiga i sammanhanget. Tyvärr uppstår ofta problem relaterade till det longitudinella upplägget som kräver metodologisk eftertanke. Exempelvis drabbas de flesta longitudinella studier av bortfall, alltså att man inte kan följa alla individer under hela studietiden. Detta bör beaktas när slutsatser dras ifrån den aktuella studien, annars riskerar man att resultaten inte är representativa för den undersökta populationen. Avhandlingens huvudfokus är att utveckla och tillämpa metoder vars syfte är att undvika snedvridning av resultaten från longitudinella studier med bortfall. Tillämpning av metoderna sker på BETULA-studien, en longitudinell studie av åldrande.

Artikel I och II betraktar samband mellan förändringar i hjärnans struktur samt kognition, för olika kognitiva mått, med hänsyn tagen till bortfall. Artikel II behandlar teoretiska resultat för hantering av bortfall när partiell korrelation används. Speciellt betraktas intervallestimation av den partiella korrelationen. Estimatorerna är allmängiltiga och kan appliceras överallt där delar av data saknas, inte bara longitudinella studier.

Artikel III samt IV undersöker hjärnans funktionella konnektivitet - de statistiska sambanden mellan aktiviteten i olika par av hjärnområden. Normalt definieras ett tröskelvärde för sambandsmåttet, och alla regionpar vars värde på sambandsmåttet överstiger tröskelvärdet betraktas som sammanlänkade. Ett problem är att tröskelvärdet väljs godtyckligt utan hänsyn till individuell variation i såväl styrka som övergripande mönster i konnektivitet. Artikel III utvecklar en metod som inte kräver bestämning av tröskelvärden samt tar hänsyn till individuell variation. Metoden är applicerbar på tvärsnittsstudier. Artikel IV utökar metoderna från artikel III till att omfatta också longitudinella studier med bortfall, och utreder hur slutsatser om konnektivitet påverkas av att inte beakta bortfallet.

Preface

There are many fantastic people that I've been privileged to meet and work with during my PhD education. I deeply appreciate all your direct and indirect contributions to this thesis.

First and foremost, I would like to express my deepest gratitude to my supervisor Xavier de Luna. Thank you for generosity with your valuable time, your encouragement and openness to my ideas. Thank you for the opportunity to start this PhD! I also want to thank my co-supervisor Anders Lundquist for your constant engagement, helpful advice, and discussions outside of our weekly meetings. I appreciate them a lot! Xavier and Anders, you seem to have a recipe for making me much happier and more confident every single time we meet. Thank you for that!

I am also grateful to my co-supervisor Lars Nyberg for seeing our research from a different perspective, asking sometimes tough questions that made me understand my work better. Thank you for your excellent expertise, without which this thesis would not have been written.

I would like to thank my co-authors Sara Pudas, Greger Orädd, Maria Josefsson, and Alireza Salami for our fruitful collaboration and your important contributions to our papers. It is a pleasure to work with you.

I am grateful for the opportunity to visit the University of Jyväskylä during my PhD studies. Many thanks to Juha Karvanen, Ilona Ruotsalainen, and Tiina Parviainen for our successful collaboration that maybe implicitly but definitely affected this thesis!

I want to thank my colleagues at the Department of Statistics. You all make the department an incredible place to work! I sincerely appreciate your smiles, hospitality, and a possibility to learn from you, both about statistics and life in general. Thanks to Katarina for helping me with my "newcomer" questions. Thank you, Jessica, for being a kind-hearted mentor in teaching. Thanks to Angel, Maria J., Maria K., Minna, and Kreske for taking the time to read earlier drafts of my papers. Gabriel and Anita, thank you for your support during the last months. Angel, thank you for bearing with me the rest of my PhD time:).

Thank you to all my school and university teachers that raised my interest in mathematics.

Special thanks to all my friends for your support. Thanks to Per and Natalya for being there for me when I needed advice and all "dysfunctional" family for laughter during our gatherings.

Thanks also to all other people, too many to mention here, for your help along the way.

Finally, I would like to thank my family for your understanding, care, and love. Thanks to Serhii for believing in me (when I don't), your support (not only the technical one :)), and your calmness (that was very much needed sometimes). Thank you for all the incredible journeys that we had and will have! Thank you!

Umeå, January 2019
Tetiana Gorbach

1 Introduction

Age-related changes in cognitive performance have been linked to brain changes (Cabeza et al., 2017, Geerligs et al., 2014, Salami et al., 2014, Salami et al., 2016). The strength of these links varies between different characteristics of the brain and cognitive domains (Damoiseaux, 2017). The conclusions also vary across studies (Kaup et al., 2011). Therefore, there is still much to be learned about cognitive and brain aging as well as their connection.

In aging research where within-individual change is of interest, longitudinal studies should be preferred over the cross-sectional ones (Diggle et al., 1994, Rönnlund et al., 2005, Schaie and Hofer, 2001). However, in nearly all longitudinal studies some participants drop out due to, for example, unwillingness to continue, relocation, health issues or death. Statistical analyses that ignore the reasons for missing data may cause severe bias in conclusions about aging (Josefsson, 2013). Therefore, dropout is one of the main methodological issues in longitudinal studies.

This thesis aims to provide methodology for studying brain aging when some data are missing due to dropout. The methods developed are applied to the longitudinal Betula project (Nilsson et al., 1997) that has unique information on aging trajectories of length up to 25 years.

Related to the diversity of brain-cognition relations, Paper I studies the association between various brain structural changes and decline of cognitive function. The dropout is taken into account by using interval estimation for regression coefficients (Genbäck et al., 2015). This approach allows statements about the significance of association between the brain and cognitive changes, but does not provide estimates of this association measured by partial correlation. Therefore, Paper II develops interval estimators for partial correlation when some data are missing.

Additionally to brain structure, aging may also be explored in terms of brain function and, in particular, statistical dependencies between the functions of distinct brain regions. When studying such connectivity, correlation is often chosen as a dependency measure. Pairs of regions with correlation stronger than some threshold are defined as connected, while other pairs are defined as non-connected (Rubinov and Sporns, 2010). Paper III adds to this field by proposing a Bayesian model for

functional brain connectivity that allows for inference without such explicit thresholding. The paper also suggests an alternative measure of brain connectivity.

Longitudinal studies of brain connectivity are still scarce (Damoiseaux, 2017). Paper IV contributes to this area of research by developing a model for longitudinal brain connectivity data with dropout. Simulation study investigates the impact of ignoring the dropout mechanism on the quality of inferences about connectivity changes during aging. To our knowledge, this is the first work which concerns nonignorable dropout in brain connectivity analyses.

This thesis is organized as follows. The Betula project is presented in Section 2. Section 3 outlines brain imaging techniques and brain connectivity analyses. Statistical methods for analysis of longitudinal data are discussed in Section 4. Section 5 presents methods for statistical analysis when some data are missing. A brief overview of Bayesian methods is given in Section 6. The papers are summarized in Section 7, followed by concluding remarks in Section 8.

2 The Betula project

The Betula project (Nilsson et al., 1997, Nilsson et al., 2004) is a longitudinal study of aging. The project aims to investigate memory and health changes during adulthood and old age as well as to explore dementia risk factors.

Participants are randomly selected from a population register in Umeå, Sweden. In total, approximately 4500 individuals have participated in at least some part of the study.

At present, six waves of data collection have been conducted since 1988 with approximately 5 years between the waves (Table 1). The subjects of Sample 1 entered the study at the first wave (T1) and have been followed since then. Sample 1 included 1000 participants with 100 subjects in each of 10 age groups (35, 40, 45, ..., 80 years old at T1). Five additional samples were recruited over time.

At each test wave, participants underwent an extensive health examination, filled in a questionnaire about socioeconomic factors, and underwent cognitive assessments. Additionally, a subsample of 376 individuals from Samples 1, 3, and 6 were scanned with structural and

Table 1: Testing occasions and subsamples of Betula study

	T1 1988-90	T2 1993-95	T3 1998-00	T4 2003-05	T5 2008-10	T6 2013-14
Sample 1	X	X	X	X	X	X
Sample 2		X	X			
Sample 3		X	X	X	X	X
Sample 4			X			
Sample 5				X		
Sample 6					X	X

Crosses represent data collection for a specific subsample. The letter T stands for the time point.

functional magnetic resonance imaging (MRI) at the fifth wave. Out of 376 initially scanned participants, 231 were also scanned at the sixth wave.

The Betula study was approved by the Regional Ethical Vetting Board at Umeå University and written consent was obtained from all participants.

3 Brain imaging

3.1 Imaging techniques

Techniques for imaging of brain structure and function include computed tomography, positron emission tomography, electroencephalography, magnetoencephalography, MRI, and functional MRI (fMRI), among others. This thesis uses MRI data to study brain structure (Papers I and II) and fMRI data to investigate brain function (Papers III and IV).

MRI utilizes differences in magnetic properties between gray matter, white matter, and cerebrospinal fluid to construct 3D images of the brain anatomy. FMRI uses magnetic properties of blood to measure brain function (Lazar, 2008). When a brain area is activated by stimuli blood flow to this area increases in order to meet metabolic demands. This results in change of blood oxygenation, which is captured by a blood oxygenation-level-dependent signal (BOLD, Ogawa et al., 1992) measured by fMRI. Brain function may be studied in response to external stimuli, such as some cognitive tasks, or during the resting state, when a subject is at rest.

3.2 Brain connectivity

Brain function is typically investigated in terms of an activation of specific regions or in terms of relations between activity of distinct brain regions, which is called brain connectivity (Bowman, 2014).

The analysis of brain connectivity usually starts with the identification of a set of brain regions of interest. This is followed by the construction of the set of connections between these regions in terms of association (functional connectivity) or causal relation between the regions' activities (effective connectivity, Friston, 2011, Smith, 2012). Properties of the constructed network may then be investigated using, for example, graph analysis (Bullmore and Sporns, 2009, Rubinov and Sporns, 2010).

Regions may be defined as voxels (the smallest volume elements in fMRI data), predefined nodes from a known parcellation of the brain or spatial areas of interest obtained with data-driven clustering (e.g., independent component analysis). Each region is then assigned with its time course using, for example, the mean of a BOLD signal from voxels within the region.

The connections in functional connectivity are determined by quantifying measures of statistical dependency between the nodes' signal, such as correlation, partial correlation, mutual information, etc. Effective connectivity, in turn, is analyzed in the literature using, among others, Granger causality and dynamic causal models (Friston, 2011).

Weak and negative correlations of fMRI signals may represent spurious connections due to subject motion (Power et al., 2012) or may be artifacts of data preprocessing (Murphy and Fox, 2017). These links are often discarded based on the absolute or proportional threshold (van den Heuvel et al., 2017). As an alternative, mixture models for brain connectivity have been proposed (Chen et al., 2016, Bielczyk et al., 2018). The mixture components represent the distribution of a connectivity measure for connected and non-connected brain regions. This idea was developed further in this thesis by imposing a mixed-effect structure on the distribution of connected component that allows simultaneous inferences on the population and subject level without explicit thresholding. We also suggest that the posterior probability of connection calculated from the fitted model might be used as an alternative measure of connectivity.

4 Statistical analyses of longitudinal data

Age-related changes are investigated using cross-sectional or longitudinal studies. In the cross-sectional approach, characteristics of interest are measured for subjects of different age. Changes in the characteristics are then inferred from the observed differences between age cohorts. On the other hand, in longitudinal studies, individuals are followed over time which enables direct analysis of within-individual change. This is the fundamental advantage of longitudinal approach, since it allows to separate aging effects from cohort effects. Such separation is impossible in cross-sectional studies. As a result, the conclusions from the longitudinal and cross-sectional approaches may differ (Diggle et al., 1994, Section 1; and Nyberg et al., 2010, Rönnlund et al., 2005 for examples from the Betula project) and, when a change over time is of interest, longitudinal studies should be preferred. Indeed, this thesis provides an example of differences between the two approaches: brain connectivity strength is found to increase with age cross-sectionally (Paper III) while longitudinal effects are negative (Paper IV).

Statistical methods for longitudinal data include two-stage analysis and modeling of individual responses (Diggle et al., 1994). In the two-stage method, the repeated measures for each subject are first reduced into a summary, which is then used in the second stage analyses. We apply such an approach in Paper I, where the relation between changes in brain structure and cognition is of interest. We define measures of change for brain structure and cognition and proceed with the correlation and regression analysis of these measures.

The individual responses are typically modeled using marginal, transition or random-effects models. All of these methods take into account the correlation between the repeated measures. In the marginal analysis, the mean and the covariance structure are modeled marginally as in a cross-sectional approach (Fitzmaurice et al., 2012). Transition models specify the covariance structure through the dependence of the current response on past observations. Random-effects modeling assumes that association between the repeated measures occurs due to variation of regression coefficients across individuals. We apply the latter approach in Paper IV.

5 Missing data in longitudinal studies

Missing data occur, e.g., when scheduled measurements are not available and is an inevitable feature of all longitudinal studies. This may happen when participants do not come to a particular measurement occasion, do not answer a questionnaire entirely, have unusable data quality or cannot be contacted.

The data may also be missing when individuals drop out, that is, stop participating in a study. In studies of aging in particular, subjects might drop out due to health issues. When health is also related to the outcome of interest, available data represents a biased sample from the population. Analyses which are based only on the available data and which ignore the dependency between the dropout and the outcome may lead to erroneous inferences about the entire population.

For example, the individuals that remain longer in longitudinal aging studies are believed to be healthier than dropouts (Chatfield et al., 2005, Josefsson, 2013, Pudas et al., 2013). Complete cases analyses, which exclude subjects with incomplete data, provide overoptimistic conclusions on the population's health. Therefore, it is crucial for the analysis to take into account the mechanisms that led to the missing data.

5.1 Missing data mechanisms

Let $\boldsymbol{Y} = (Y_{ij})$ denote a matrix of full data scheduled to be observed, where Y_{ij} represents variable j for subject i. Let $\boldsymbol{M} = (M_{ij})$, where $M_{ij} = 1$ if intended measurement Y_{ij} is missing and 0 otherwise. A missing data mechanism is the conditional distribution of missingness indicators $\boldsymbol{M}$ given $\boldsymbol{Y} = \boldsymbol{y}$ and indexed by a parameter $\boldsymbol{\phi}$ (Little and Rubin, 2002).

Let $\tilde{\boldsymbol{m}}$ represent the observed value the missing data pattern, which is the value of the missing indicator matrix $\boldsymbol{M}$. The data $\boldsymbol{Y}$ may then be partitioned based on $\tilde{\boldsymbol{m}}$ into the observed part $\boldsymbol{Y}_{(1)} = (Y_{ij} : \tilde{m}_{ij} = 1)$ and missing data $\boldsymbol{Y}_{(0)} = (Y_{ij} : \tilde{m}_{ij} = 0)$. Let $\tilde{\boldsymbol{y}}_{(1)}$ denote an observed value of $\boldsymbol{Y}_{(1)}$, and $\boldsymbol{y}_{(0)}$ and $\boldsymbol{y}'_{(0)}$ denote realizations of the random variable $\boldsymbol{Y}_{(0)}$.

In this thesis, we are mostly interested in distinction between situations where missingness depends on the value of the missing data (missing not at random data mechanisms) or not (missing at random).

The missing data mechanism is called missing at random (MAR) if the conditional probability of the observed pattern of missing data, given the missing and observed data, is the same for all possible values of the missing data and the parameter $\boldsymbol{\phi}$ (Rubin, 1976, Mealli and Rubin, 2015):

$$\begin{aligned} &P(\boldsymbol{M} = \tilde{\boldsymbol{m}} | \boldsymbol{Y}_{(1)} = \tilde{\boldsymbol{y}}_{(1)}, \boldsymbol{Y}_{(0)} = \boldsymbol{y}_{(0)}, \boldsymbol{\phi}) = \\ &P(\boldsymbol{M} = \tilde{\boldsymbol{m}} | \boldsymbol{Y}_{(1)} = \tilde{\boldsymbol{y}}_{(1)}, \boldsymbol{Y}_{(0)} = \boldsymbol{y}'_{(0)}, \boldsymbol{\phi}) \\ &\text{for all } \boldsymbol{\phi}, \boldsymbol{y}_{(0)} \text{ and } \boldsymbol{y}'_{(0)}. \end{aligned}$$

The missing data mechanism is called missing not at random (MNAR) when it depends on missing data for some value of the parameter $\boldsymbol{\phi}$:

$$\begin{aligned} &P(\boldsymbol{M} = \tilde{\boldsymbol{m}} | \boldsymbol{Y}_{(1)} = \tilde{\boldsymbol{y}}_{(1)}, \boldsymbol{Y}_{(0)} = \boldsymbol{y}_{(0)}, \boldsymbol{\phi}) \neq \\ &P(\boldsymbol{M} = \tilde{\boldsymbol{m}} | \boldsymbol{Y}_{(1)} = \tilde{\boldsymbol{y}}_{(1)}, \boldsymbol{Y}_{(0)} = \boldsymbol{y}'_{(0)}, \boldsymbol{\phi}) \\ &\text{for some } \boldsymbol{\phi}, \text{ and some } \boldsymbol{y}_{(0)} \neq \boldsymbol{y}'_{(0)}. \end{aligned}$$

When the missing data mechanism does not depend on the data at all, it is called missing completely at random (MCAR):

$$P(\boldsymbol{M} = \tilde{\boldsymbol{m}} | \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\phi}) = P(\boldsymbol{M} = \tilde{\boldsymbol{m}} | \boldsymbol{Y} = \boldsymbol{y}', \boldsymbol{\phi}) \text{ for all } \boldsymbol{\phi}, \boldsymbol{y} \text{ and } \boldsymbol{y}'.$$

In case of dropout, the terminology can be simplified by using, for example, one missing data indicator $M_i = j$ if subject i drops out between time points $j-1$ and j.

For illustration, let us consider a case of two scheduled measurements. The first measurement Y_1 is observed, while Y_2 may be missing due to dropout. Such missingness is the focus of Paper IV. Here, only one indicator variable is sufficient to represent missingness: M is equal to one if the second observation is missing and zero otherwise. For example, if the observed missing data pattern is $M = 1$, then $\boldsymbol{Y}_{(1)} = Y_1$, $\boldsymbol{Y}_{(0)} = Y_2$ and $\tilde{\boldsymbol{y}}_{(1)} = \tilde{y}_1$, the value of the observed data Y_1. The dropout mechanism is MAR when

$$\begin{aligned} &P(M = 1 | Y_1 = \tilde{y}_1, Y_2 = y_2, \boldsymbol{\phi}) = P(M = 1 | Y_1 = \tilde{y}_1, Y_2 = y'_2, \boldsymbol{\phi}) \\ &\text{for all values of the parameter } \boldsymbol{\phi} \text{ and missing data } y_2 \text{ and } y'_2. \end{aligned}$$

The missing data mechanism is MNAR if

$$\begin{aligned} &P(M = 1 | Y_1 = \tilde{y}_1, Y_2 = y_2, \boldsymbol{\phi}) \neq P(M = 1 | Y_1 = \tilde{y}_1, Y_2 = y'_2, \boldsymbol{\phi}) \\ &\text{for some } \boldsymbol{\phi}, \text{ and some } y_2 \neq y'_2. \end{aligned}$$

Since their introduction by Rubin (1976), the definitions of missing data mechanisms have not been used consistently. For example, some authors separate response data and fully-observed covariates (Daniels and Hogan, 2008, Fitzmaurice et al., 2012), while others consider response and covariates together in one data vector (Little and Rubin, 2002, Diggle et al., 1994). Additionally, the definitions of MAR were interpreted differently by researchers. Seaman et al. (2013) discusses this issue and the validity of inferences under different missing data mechanisms. Following the publication of Seaman et al. (2013), these definitions were clarified in Mealli and Rubin (2015).

It is frequently said that likelihood methods that ignore the missing data mechanism provide valid inferences when the data are MAR. However, the validity of any inferences, not only maximum likelihood, also depends on the research question at hand. If in the illustration above, for example, inferences about the population mean of Y_1 are of interest, then the sample mean $\overline{Y}_1$ is an unbiased estimator even if the data are MNAR for Y_2. In contrast, when Y_1 and Y_2 are dependent, the sample mean $\overline{Y}_2$ calculated over the observed values of Y_2 may be a biased estimator for the population mean of Y_2 even if data are MAR.

5.2 Statistical methods for analysis with missing data

Strategies for handling incomplete data include complete cases analysis, weighting methods, available cases analyses, imputation, and model-based procedures (Little and Rubin, 2002).

Complete cases analysis (listwise deletion) uses only the data from the subjects that have all data that is intended to be observed available on all variables in the model of interest. It is a popular choice due to its simplicity. It is also a default choice in standard statistical software.

When the missing mechanism is MCAR, the complete cases are just a random sample from the original sample (Figure 1). As a result, unbiased estimators based on the original sample remain unbiased when applied to the reduced sample. Complete cases analysis may sometimes be valid even in MAR or MNAR cases. For example, in a correctly specified regression model where missingness only occurs on the predictors and does not depend on the value of the response (Allison, 2001, p. 7); and in case of MNAR data on the response in a logistic regression when missingness does not depend on the predictors (Prentice and

Pyke, 1979).

However, complete cases analyses generally provide biased estimates when the MCAR assumption is not fulfilled. For example, when less healthy people do not report their health, an average health based on complete cases will overestimate the health of the population. See as an illustration Figure 1, where sample mean of Y_2 based on complete cases is a biased estimator of the population mean of Y_2 in the MAR setting, and where the dependency between Y_1 and Y_2 for the full data and complete cases are of opposite sign in the MNAR. Additionally, estimates based on the complete cases have high variance when a large amount of incomplete data is discarded.

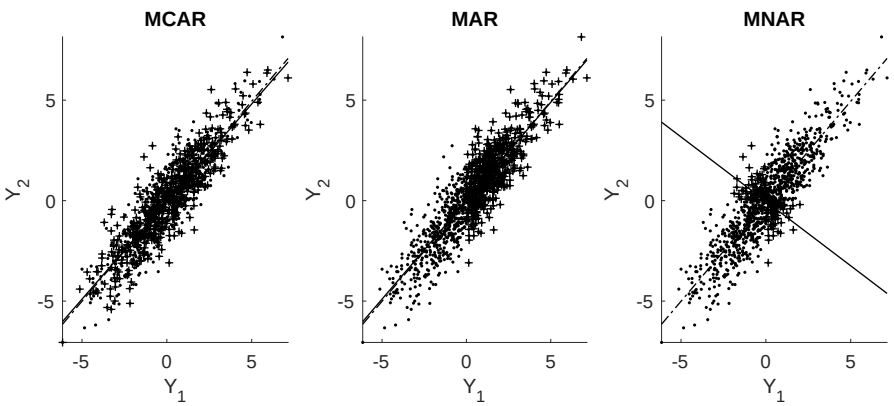

Figure 1: Examples of data with two scheduled measurements Y_1 and Y_2, missingness in Y_2, and MCAR ($M_i = 0$ if the observation Y_{2i} is observed and 1 otherwise, with $P(M_i = 0) = 0.5$ regardless of Y_1 and Y_2), MAR ($M_i = 0$ if $Y_{1i} > 0$ regardless of Y_2), and MNAR ($M_i = 0$ if $Y_{2i}Y_{1i} < 0$) dropout mechanisms. Crosses represent the complete cases, dots represent the cases with missing Y_2. Missing data mechanism is specified in the title of each plot. The lines represent least-squares fit to the full data (dash-dotted line) and complete cases (solid line).

One solution to the inefficiency of the complete cases analysis is to consider available cases instead. This method uses all data available for the calculation of a particular statistics of interest. For example, suppose there are n_1 observations for Y_1 and n_2 for Y_2. The estimate of the population mean of Y_1 is then based on n_1 observations while the mean of Y_2 is estimated from n_2 data points. When MCAR is true, available cases estimators might be less variable than the ones

from complete cases analyses. The disadvantage of this method is that estimates using different variables are calculated on samples of different size which may result in final estimates outside of the acceptable range (Little and Rubin, 2002).

Weighting techniques extend the survey sampling approach to estimation (Cochran, 1977) by adjusting sampling weights to handle nonresponse in the data.

Imputation-based procedures are used to avoid inefficiency from discarding the observed data. Here, the missing values are filled in once or possibly several times (Rubin, 2004, van Buuren, 2012). The resulting complete datasets are further analyzed. The values to impute may be observations from other subjects in the sample (hot-deck imputation), available observations from the subject with missing data (last observation carried forward, baseline observation carried forward), sample means (mean imputation) or estimated via regression techniques from the observed data (regression imputation or stochastic imputation).

Model-based procedures specify a model of full data, indexed by some parameter $\boldsymbol{\omega}$, $P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{M} = \boldsymbol{m}|\boldsymbol{\omega})$. When missing mechanism is MAR and the parameter $\boldsymbol{\theta}$ of the data distribution $P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta})$ is different from the parameter $\boldsymbol{\phi}$ of the missing data mechanism (a priori independent in Bayesian analysis), the missing data mechanism is called ignorable. In such case, likelihood inferences for $\boldsymbol{\theta}$ based on the observed data distribution $P(\boldsymbol{Y}_{(1)} = \tilde{\boldsymbol{y}}_{(1)}|\boldsymbol{\theta})$ are the same as those using the full data model $P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{M} = \boldsymbol{m}|\boldsymbol{\omega})$ (Little and Rubin, 2002). MAR is necessary (but not sufficient) for ignorability of the missing data mechanism. Therefore, MNAR missing data mechanisms are sometimes called nonignorable. Note that a mechanism may be nonignorable even if it is MAR.

Under a MNAR data mechanism, all above-mentioned techniques for analysis with missing data typically provide biased results since they ignore the dependency of the missing data mechanism on unobserved data. Two common ways to model MNAR data are selection and pattern-mixture models (Little and Rubin, 2002, Diggle et al., 1994, Daniels and Hogan, 2008). Selection models factor the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{M}$ as

$$P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{M} = \boldsymbol{m}|\boldsymbol{\omega}) = P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{M} = \boldsymbol{m}|\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\phi}),$$

while pattern-mixture models use the factorization:

$$P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{M} = \boldsymbol{m}|\boldsymbol{\omega}) = P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{M} = \boldsymbol{m}, \boldsymbol{\theta}')P(\boldsymbol{M} = \boldsymbol{m}|\boldsymbol{\phi}').$$

A famous example of a selection model is the Heckman selection model (Heckman, 1979). It is utilized with some changes in Papers I and II in this thesis. Paper IV uses the pattern-mixture approach.

The missing data mechanism may be known, when, for example, incomplete data is due to censoring at some known censoring point. When the missing mechanism is unknown, the analysis relies on the assumptions about the missing data mechanism. Association between $\boldsymbol{M}$ and the observed data $\boldsymbol{Y}_{(1)}$ may be used to provide evidence against MCAR assumption (Little, 1988). However, the assumptions about MAR and MNAR missing data mechanisms are untestable from the observed data since their definitions include unavailable data $\boldsymbol{Y}_{(0)}$. Therefore, sensitivity of inferences to the specification of the full data model $P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{M} = \boldsymbol{m}|\boldsymbol{\omega})$ should be investigated.

5.3 Sensitivity analysis using uncertainty regions

The full data model $P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{M} = \boldsymbol{m}|\boldsymbol{\omega})$ may involve some parameter $\boldsymbol{\gamma}$ that is unidentified from the observed data. As a result, several full data distributions, indexed by different values of $\boldsymbol{\gamma}$, may correspond to the same observed data distribution. The parameter $\boldsymbol{\gamma}$ is sometimes called a sensitivity parameter (Daniels and Hogan, 2008).

Sensitivity analysis explores the variation of inferences about the parameters of interest $\boldsymbol{\theta}$ depending on the assumption that $\boldsymbol{\gamma}$ takes a specific value. This can be done by evaluating estimators of $\boldsymbol{\theta}$ for a range of plausible values of $\boldsymbol{\gamma}$. The range of the obtained estimates may be used as a final estimate that represents a sensitivity of results to missing data. When interval estimators are considered, the sensitivity analysis may be performed by constructing uncertainty regions as a union of the confidence intervals over plausible values of the sensitivity parameter (Figure 2). The coverage of such uncertainty intervals is at least as good as the coverage of the confidence interval constructed under the true unknown value of the sensitivity parameter $\boldsymbol{\gamma}$.

Vansteelandt et al. (2006) proposed to construct uncertainty regions for an unidentified parameter by adding confidence limits to estimated

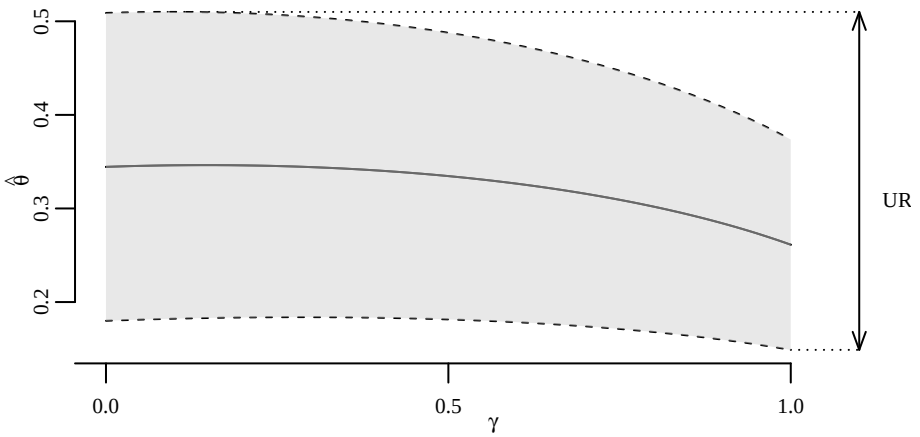

Figure 2: Let γ index the full data distributions that correspond to the observed data distribution and let $[0, 1]$ be reasonable values of γ. Point estimates, $\hat{\theta}$, of the parameter θ for $\gamma \in [0, 1]$ are represented as a solid line. The union of 95% confidence intervals (bounds represented as dashed lines) for θ over $\gamma \in [0, 1]$ is an uncertainty region (UR).

bounds of a range of parameter values that correspond to different full data distributions compatible with the observed data law.

Genbäck et al. (2015) used uncertainty regions for estimation of regression parameters. This approach is also used in Paper I in the thesis. Lindmark et al. (2018) proposed the uncertainty regions for estimation of direct and indirect effects in the presence of unobserved confounding.

Paper II of the thesis develops a theory for construction of the uncertainty regions for partial correlation when data are missing not at random.

6 Bayesian inference

Bayesian inference uses the model for the data and prior distributions for the parameters in this model to provide posterior inferences on the parameters of the model. These priors reflect knowledge about the model available before the data are observed, for example, from previous studies. The model is then used to update prior information and to get a posterior distribution of the parameters.

Given a prior distribution $p(\boldsymbol{\theta})$, the posterior distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)}) = \frac{L(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)})p(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)})p(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

where the likelihood $L(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)})$ for a parameter $\boldsymbol{\theta}$ is the joint probability density (or mass for discrete random variables) function of the observed data when regarded as a function of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)}) = p(\tilde{\boldsymbol{y}}_{(1)}|\boldsymbol{\theta}).$$

Typically, one is interested in evaluation of some function of the posterior distribution, e.g., the expectation of some function f of the parameter $\boldsymbol{\theta}$:

$$Ef(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)}) = \int f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)})d\boldsymbol{\theta}.$$

Often, due to intractability of the high-dimensional integral in the denominator in $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)})$, a sample from the posterior distribution is obtained without explicit evaluation of the denominator using techniques such as Markov chain Monte Carlo (MCMC, Metropolis et al., 1953).

Markov Chain Monte Carlo evaluates the integral $Ef(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)})$ using the convergence:

$$\frac{\sum_{i=1}^{n} f(\boldsymbol{\theta}_i)}{n} \rightarrow_{a.s.} Ef(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)}),$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ is generated as a Markov chain with stationary distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}_{(1)})$. The convergence is ensured by the ergodic theorem (Nummelin, 1984). Metropolis-Hastings, Metropolis, and Gibbs sampler (Gilks et al., 1996) algorithms are commonly used to construct Markov chains that satisfy the conditions of the ergodic theorem.

The rate of MCMC convergence is known to depend on the parameterization of the model and the choice of parameters' batches for the updates (Gilks et al., 1996). Thus, we found that in mixed-effects modeling in Paper III, updating all fixed and random effects together provided the quickest convergence compared to the update of each parameter separately, the update of fixed effects together but separately from the random effects, which, in turn, were updated separately for each individual or together.

When interval estimation is of interest, credible intervals, which are analogues to confidence sets in frequentist statistics, may be constructed using the quantiles of posterior distribution.

7 Summary of papers

7.1 Paper I

The main purpose of Paper I, *Longitudinal association between hippocampus atrophy and episodic-memory decline*, is to explore the association between aging-related cognitive decline and brain changes in gray matter volume, white matter integrity, and white matter hyperintensities volume. The cognition was measured longitudinally over 15 years while the brain changes were measured at two time points approximately four years apart.

A two-stage analysis of the longitudinal data is used in the paper. The paper examines change-change brain-cognition relation using partial correlation while controlling for potential confounding factors of age and hypertension. Complete cases analysis is followed by a sensitivity analysis of results for data missing not at random. The sensitivity analysis uses the relation between partial correlation and a regression coefficient, and uncertainty intervals for a regression parameter proposed by Genbäck et al. (2015). As predicted, atrophy in the hippocampus was related to episodic-memory decline, and this association remained significant when data missing not at random were accounted for. The remaining brain-cognition change-change relations were weak and nonsignificant.

7.2 Paper II

In Paper I, significance testing for partial correlation is performed using the relationship between partial correlation and a regression coefficient. However, the method does not provide estimates for the partial correlation. In Paper II, entitled *Inference for partial correlation when data are missing not at random,* we define interval estimators, called uncertainty regions, for the partial correlations when data are missing not at random. The performance of these estimators is illustrated via a simulation study. For illustration, we construct uncertainty intervals for the partial correlation between longitudinal changes in gray matter volume of the hippocampus and episodic memory decline analyzed in Paper I. The theoretical results of Paper II are general, not limited to brain imaging studies, and may be used in longitudinal as well as cross-sectional studies.

7.3 Paper III

The third paper, *A Hierarchical Bayesian Mixture Modeling Approach for Analysis of Resting-State Functional Brain Connectivity: An Alternative to Thresholding*, develops a mixture model to study functional resting-state brain connectivity. The mixture components represent connected and non-connected brain regions. Methodologically, this paper adds to the existing literature by imposing a mixed-effect structure on the distribution of the connected component. This allows simultaneous inferences on the population and subject levels.

For each pair of regions and subject, posterior probabilities of being in the connected component can be computed. The paper suggests the posterior probability as an alternative connectivity measure since the posterior probabilities reflect connectivity of a brain region pair in relation to overall connectivity pattern of an individual, which is neglected in a traditional correlation analysis.

This paper proposes an approach for analysis of connectivity without explicit thresholding. However, if the connectivity matrix has to be thresholded for subsequent analysis, we show that absolute thresholding based on posterior probabilities may be superior to the one based on the correlation.

The introduced method is applied to study the relation of brain connectivity to age and cognition using data from the Betula project. This cross-sectional study indicates that older individuals might have stronger connections on average and that the proportion of connections varies considerably between individuals. The cognition was not strongly related to the considered resting-state connectivity.

7.4 Paper IV

Paper IV, *Bayesian mixture modeling for longitudinal fMRI connectivity studies with dropout*, extends the cross-sectional model of Paper III to a longitudinal setting in the presence of nonignorable dropout. The proposed method allows to take into account the uncertainty in the dropout mechanism. The paper discovers differences in longitudinal and cross-sectional estimates of brain changes based on complete cases analyses. A simulation study in Paper IV shows that inferences might be highly biased when dropout is ignored in the estimation procedure.

8 Final remarks and further research

A subsample of Betula participants underwent MRI and cognitive assessment approximately four years after the final Betula wave. This may allow, additionally to the analysis of associations in Paper I, to study causal relationships between brain structural and cognitive changes using, for example, graphical models (Pearl, 2009). It would be of interest to investigate which brain structural changes cause cognitive changes and if cognitive changes may also affect subsequent changes in brain structure.

To make the interval estimators for partial correlation proposed in Paper II more accessible, we are planning to implement them as part of the publicly available R package, called `ui`, for sensitivity analysis using uncertainty regions (Stat4Reg, 2019).

The simulation study of Paper IV investigates the effect of ignoring the missing data mechanism on the inferences about longitudinal functional brain connectivity. In order to provide simulation scenarios, complete cases analysis of data from the Betula project was performed. In the nearest future we plan to carry out an analysis of the Betula data assuming a MNAR dropout mechanism.

References

Allison, P. D. (2001). *Missing data*. Sage publications, Thousand Oaks, CA.

Bielczyk, N. Z., Walocha, F., Ebel, P. W., Haak, K. V., Llera, A., Buitelaar, J. K., Glennon, J. C., and Beckmann, C. F. (2018). Thresholding functional connectomes by means of mixture modeling. *NeuroImage*, 171:402–414.

Bowman, F. D. (2014). Brain imaging analysis. *Annual Review of Statistics and Its Application*, 1:61–85.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10:186–198.

Cabeza, R., Nyberg, L., and Park, D. C. (2017). *Cognitive Neuroscience*

of Aging: Linking Cognitive and Cerebral Aging. Oxford University Press, New York.

Chatfield, M. D., Brayne, C. E., and Matthews, F. E. (2005). A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies. *Journal of Clinical Epidemiology*, 58(1):13–19.

Chen, S., Bowman, F. D., and Mayberg, H. S. (2016). A Bayesian hierarchical framework for modeling brain connectivity for neuroimaging data. *Biometrics*, 72(2):596–605.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, 3 edition.

Damoiseaux, J. S. (2017). Effects of aging on functional and structural brain connectivity. *NeuroImage*, 160:32–40.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall/CRC, Boca Raton.

Diggle, P., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*. John Wiley & Sons Inc., Hoboken, NJ, 2 edition.

Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain connectivity*, 1(1):13–36.

Geerligs, L., Renken, R. J., Saliasi, E., Maurits, N. M., and Lorist, M. M. (2014). A brain-wide study of age-related changes in functional connectivity. *Cerebral Cortex*, 25(7):1987–1999.

Genbäck, M., Stanghellini, E., and de Luna, X. (2015). Uncertainty intervals for regression parameters with non-ignorable missingness in the outcome. *Statistical Papers*, 56(3):829–847.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, Boca Raton.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Josefsson, M. (2013). *Attrition in Studies of Cognitive Aging*. PhD thesis, Umeå universitet, Umeå.

Kaup, A. R., Mirzakhanian, H., Jeste, D. V., and Eyler, L. T. (2011). A review of the brain structure correlates of successful cognitive aging. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 23(1):6–15.

Lazar, N. (2008). *The Statistical Analysis of Functional MRI Data*. Springer, New York.

Lindmark, A., de Luna, X., and Eriksson, M. (2018). Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals. *Statistics in Medicine*, 37(10):1744–1762.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken, NJ, 2 edition.

Mealli, F. and Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102(4):995–1000.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Murphy, K. and Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *NeuroImage*, 154:169–173.

Nilsson, L.-G., Adolfsson, R., Bäckman, L., de Frias, C. M., Molander, B., and Nyberg, L. (2004). Betula: A prospective cohort study on memory, health and aging. *Aging, Neuropsychology, and Cognition*, 11(2-3):134–148.

Nilsson, L.-G., Bäckman, L., Erngrund, K., Nyberg, L., Adolfsson, R., Bucht, G., Karlsson, S., Widing, M., and Winblad, B. (1997). The Betula prospective cohort study: Memory, health, and aging. *Aging, Neuropsychology, and Cognition*, 4(1):1–32.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, Cambridge.

Nyberg, L., Salami, A., Andersson, M., Eriksson, J., Kalpouzos, G., Kauppi, K., Lind, J., Pudas, S., Persson, J., and Nilsson, L.-G. (2010). Longitudinal evidence for diminished frontal cortex function in aging. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22682–22686.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):5951–5955.

Pearl, J. (2009). *Causality*. Cambridge University Press, New York, 2 edition.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

Pudas, S., Persson, J., Josefsson, M., de Luna, X., Nilsson, L.-G., and Nyberg, L. (2013). Brain characteristics of individuals resisting age-related cognitive decline over two decades. *Journal of Neuroscience*, 33(20):8668–8677.

Rönnlund, M., Nyberg, L., Bäckman, L., and Nilsson, L.-G. (2005). Stability, growth, and decline in adult life span development of declarative memory: cross-sectional and longitudinal data from a population-based study. *Psychology and Aging*, 20(1):3.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken, NJ.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52:1059–1069.

Salami, A., Pudas, S., and Nyberg, L. (2014). Elevated hippocampal resting-state connectivity underlies deficient neurocognitive function in aging. *Proceedings of the National Academy of Sciences of the United States of America*, 111(49):17654–17659.

Salami, A., Wåhlin, A., Kaboodvand, N., Lundquist, A., and Nyberg, L. (2016). Longitudinal evidence for dissociation of anterior and posterior MTL resting-state connectivity in aging: Links to perfusion and memory. *Cerebral Cortex*, 26:3953–3963.

Schaie, K. W. and Hofer, S. M. (2001). Longitudinal studies in aging research. In Birren, J. E. and Schaie, K. W., editors, *Handbook of The Psychology of Aging*, pages 53–77. Academic Press, 5 edition.

Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "missing at random"? *Statistical Science*, 28(2):257–268.

Smith, S. M. (2012). The future of FMRI connectivity. *NeuroImage*, 62(2):1257–1266.

Stat4Reg (2019). ui R-package. Online, accessed: 2019-01-10, https://www.stat4reg.se/software/uipackage.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, New York, 2 edition.

van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T. T., and Schmidt, R. (2017). Proportional thresholding in resting-state fMRI functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations. *NeuroImage*, 152:437–449.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16:953–979.