Functional Data Analysis applied in Chemometrics

with focus on NMR Nutri-metabolomics

PhD thesis by

MARTHA MULLER

Department of Mathematical Sciences
University of Copenhagen
Denmark

PhD School of Science - Faculty of Science - University of Copenhagen

Martha Muller
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 København Ø
Denmark

muller.martie@gmail.com

PhD thesis submitted to the PhD School of Science, Faculty of Science, University of Copenhagen, Denmark, 31 October 2014.

Academic advisor:
Anders Tolver, Associate Professor
Department of Mathematical Sciences, University of Copenhagen, Denmark

Assessment Committee:
Bo Markussen (chair), Associate Professor
Department of Mathematical Sciences, University of Copenhagen, Denmark

Sara Sjöstedt de Luna, Professor
Department of Mathematics and Mathematical Statistics, Umeå University, Sweden

Ron Wehrens, Biometris / Biosciences Business Unit Manager
Plant Research International, Wageningen UR, The Netherlands

ISBN 978-87-7078-968-4

To my parents
Anton Muller and Anna Martha Muller
who are my first mathematics teachers and so much more,
to my love
Jerry Floyd Everett
"one letter belongs to you"
and to my Canadian parents
Floyd Emery Everett and Ina Ruth Everett
who always asked and prayed

Preface

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the Faculty of Science, University of Copenhagen, Denmark. The first two years, from August 2010 to September 2012, were conducted under the supervision of Christian Ritz at the then Department of Basic Science and Environment, Faculty of Life Sciences, University of Copenhagen. This faculty then merged with the Faculty of Science and what remained of the Biostatistics group moved to the Department of Mathematical Sciences. This PhD project was continued at the Department of Mathematical Sciences under the supervision of Anders Tolver until October 2014. It was financed by the Program of Excellence at the University of Copenhagen as well as the above mentioned departments.

The NMR nutri-metabolomics data were obtained from a pilot study of the DiOGenes trial. Lone Graasbøl Rasmussen, Francesco Savorani and Hanne Winning were instrumental in obtaining and explaining the data.

The juggling data were provided as part of the Mathematical Biosciences Institute (MBI) Current Topic Workshop (CTW) on Statistics of Time Warpings and Phase Variations, that took place from 13 to 16 November 2012 in Columbus, Ohio.

I am thankful to Ib Skovgaard and Christian Ritz who played fundamental roles in initiating this PhD and specific project. I am grateful to my supervisor, Anders Tolver, for many interesting discussions, problem-solving sessions and revisions - “Tusind tak”. To my last of many office mates, Nina Munkholdt - thank you for the fun and laughter. Sima Mashayekhi and Noura Mousavi, your friendliness, friendship and food tasting made lunch times fun.

My parents, family and friends were supportive and encouraging throughout this journey - thank you from my heart. Nina Vang, thank you for many heart-warming meals, walks and talks. To Thys, Jeroen, Nina, Julia and Dan - thank you for being there, listening and praying. Britta, Björn and May, your friendship, welcoming home and delicious food were always a light in the dark. Rosalind, thank you for listening, laughter and encouragement. To my parents - you taught me to count, to calculate and to do calculus with love and encouragement. Thank you for always being there and for visiting twice. Isabel, thank you for being a great sister and for listening with love. Ina and Floyd, thank you for always asking and for so many prayers. Jerry, your care, support, prayers and encouragement were an oasis in a sometimes dry landscape.

Summary

In this thesis we explore the use of functional data analysis as a method to analyse chemometric data, more specifically spectral data in metabolomics. Functional data analysis is a vibrant field in statistics. It has been rapidly expanding in both methodology and applications since it was made well known by Ramsay & Silverman's monograph in 1997. In functional data analysis, the data are curves instead of data points. Each curve is measured at discrete points along a continuum, for example, time or frequency. It is assumed that the underlying process generating the curves is smooth, but it is not assumed that the adjacent points measured along the continuum are independent. Standard chemometric methods originate from the field of multivariate analysis, where variables are often assumed to be independent. Typically these methods do not explore the rich functional nature of spectral data.

Metabolomics studies the 'unique chemical fingerprints' (Daviss, 2005) that cellular processes create in living systems. Metabolomics is used to study the influence of nutrition on the human metabolome. Nutritional metabolomics shows great potential for the discovery of novel biomarkers of food consumption, personal nutritional status and metabolic phenotype. We want to understand how metabolomic spectra can be analysed using functional data analysis to detect the influence of different factors on specific metabolites. These factors can include, for example, gender, diet culture or dietary intervention. In Paper I we apply wavelet-based functional mixed model methodology and use bootstrap-based inference on functions to find jointly significant differences in metabolites, or spectral regions. In more detail, wavelets are used to model sharp, localised peaks in the spectra. Wavelet shrinkage reduces the noise and provides a sparse representation of each spectrum. Subset selection of wavelet coefficients generates the input to mixed models. Mixed-model methodology enables us to take the study design into account while modelling covariates. Bootstrap-based inference preserves the correlation structure between curves and enables the estimation of functional confidence intervals for mean curves. We also discuss the many practical considerations in wavelet estimation and thresholding, and the important influence the choices can have on the resulting estimates.

On a conceptual level, the purpose of this thesis is to build a stronger connection between the worlds of statistics and chemometrics. We want to provide a glimpse of the essential and complex data pre-processing that is well known to chemometricians, but is generally unknown to statisticians. Pre-processing can potentially have a strong influence on the results of consequent data analysis. Our focus is on nuclear magnetic resonance data and

we discuss the inherent structure in this type of data. However, many of the methods covered in this thesis are also applicable to other spectral data, e.g. mass spectrometry or infrared.

In Paper II we give a brief overview of functional data analysis – a field that is known to statisticians, but often obscured from chemometricians. We illustrate the rich nature of functional derivatives in simulated nuclear magnetic peaks with characteristic Lorentzian line shape. Using phase-plane plots to explore the anatomy of NMR peaks, we introduce the novelty of heart plots for spectral data.

The important aspect of registration, also called warping or alignment, emerges from both the chemometric and statistical perspectives. In Paper III we apply functional registration in the context of biomechanics, specifically to data from a juggling experiment. The novelty of this work is that the registration is done towards an idealized biomechanical model. In this way, the warping is performed subject to biomechanical constraints.

The supplemental paper, Paper IV, demonstrates the application of classical mixed-model methodology in the context of targeted metabolomics. Dietary effects on biomarkers of bone turnover in children were investigated as part of the pan-European DiOGenes dietary intervention trial. The metabolomics data in paper I originated from a pilot study of the DiOGenes trial.

Overall this thesis gives an indication of the huge possibilities for functional data analysis in metabolomics and chemometrics. Spectral data are inherently functional in nature. Functional data analysis provides access to many functional equivalents of methods currently used in chemometrics, with the benefits of no strong assumptions regarding neighbouring observations. Functional data analysis also provides access to the data's derivatives and opens up the ability to analyse information that is otherwise locked away in the data. The use of functional data analysis in metabolomics can make a valuable contribution to the emerging technology in personalised medicine and health care, including personalised nutrition for prevention and treatment.

"Functional data analysis has a long historical shadow, extending at least back to the attempts of Gauss and Legendre to estimate a comet's trajectory (Gauss, 1809; Legendre, 1805)."

"Statistics shows its finest aspects when exciting data find existing statistical technology not entirely satisfactory. It is this process that ... ensures that unforeseen adventures in research awaits us all. "

Jim Ramsay & Bernard Silverman

Contents

List of Figures

List of Tables

Abbreviations

FDA	**F**unctional **D**aia **A**nalysis
FFT	**F**ast **F**ourier **T**ransform
FID	**F**ree **I**nduction **D**ecay
GAN	**G**roup **A**ggregating **N**ormalisation
GC-MS	**G**as-**C**hromatography **M**ass **S**pectrometry
LC-MS	**L**iquid-**C**hromatography **M**ass **S**pectrometry
mica	**m**ean **r**elative **c**hange in **a**rea
mcc	**m**ean **c**orrelation **c**oefficients
MS	**M**ass **S**pectrometry
NMR	**N**uclear **M**agnetic **R**esonance
PC	**P**rincipal **C**component
PCA	**P**rincipal **C**component **A**nalysis
PLS	**P**artial **L**east **S**quares
PLS-DA	**P**artial **L**east **S**quares - **D**iscriminant **A**nalysis
ppm	**p**arts **p**er **m**illion
PQN	**P**robabilistic **Q**uotient **N**ormalisation
STOCSY	**S**tatistical **TO**tal **C**orrelation **S**pectroscop**Y**
TSP	**T**rimethyl**s**ilyl **p**ropionate
TMS	**T**etra**m**ethyl**s**ilane
VSN	**V**ariance **S**tabilization **N**ormalisation

Terminology

The following terms/descriptions are generally used interchangeably in the literature:

- spectra, profiles
- metabolites, analytes, compounds, variables, chemical analytes
- concentration (of a metabolite), intensity, peak height
- metabolite concentrations, levels of individual peaks
- sample composition, chemical composition, amplitudes of metabolite peaks
- peak extraction, peak picking; resulting in peak lists

1 Introduction

Functional data analysis (FDA) is a well-established and fast-growing field in statistics. FDA is an especially exciting field of research based on its wide range of applications. Nevertheless, it is not well known in the field of chemometrics, which focuses on the statistical analysis of chemical data.

Metabolomics is a relatively new and expanding field where the chemical 'fingerprints' of metabolism are measured and analysed. In human nutrition, metabolomics shows great potential in dietary monitoring, discovery of biomarkers and development in personalised nutrition. The technology for chemical analysis of samples has rapidly expanded, but the technology for the analysis of the complex data gathered in metabolomics experiments is, in general, lagging behind. Close collaboration between metabolomics experts, chemometricians and statisticians would bring the complex and interesting data analysis problems from metabolomics to the attention of statisticians, thus fuelling new theoretical developments. Conversely, such a close collaboration will bring new statistical methods to the attention of chemometricians, and will enable novel statistical applications in metabolomics.

1.1 Objective of the Thesis

The purpose of this thesis is two fold. On a statistical level, we explore the use of functional data analysis as a chemometric tool in metabolomics. On a conceptual level, we aim to build a bridge between the worlds of chemometrics and statistics.

Considering these aims in more detail: Firstly, we want to understand how metabolomic spectra can be analysed using FDA to detect the influence of different factors on specific metabolites in the spectra. Secondly, we want to provide a glimpse of the essential and complex pre-processing of nuclear magnetic resonance data that are well known to chemometricians, but, generally, unknown to statisticians. Additionally, the important aspect of registration, also called warping or alignment, emerges from both the chemometric and statistical perspectives.

1.2 Thesis outline

Chapter 2 gives a general background on metabolomics and how it relates to human nutrition. This is relevant to Papers I, II and the supplementary Paper IV.

Chapter 3 provides an overview of the basic concepts in functional data analysis, as it pertains to Papers I, II and III. Chapter 4 presents the mathematical background for wavelets and the practical considerations relating to wavelet shrinkage, relevant to Paper I. In particular, we discuss the challenging parameter choices related to wavelet shrinkage of the data in Paper I.

In Chapter 5 we cover the structure of NMR data and the various steps involved in chemometric pre-processing. This is intended as background for statisticians and is specifically relevant to Papers I and II.

Papers I to III, focus on different aspects of functional data analysis: wavelet based functional mixed models (Paper I); the use of functional derivatives for phase-plane plots (Paper II); and, registration of functional data subject to constraints (Paper III). In Chapter 6 and Paper III we deviate from functional data analysis in the field of human nutrition metabolomics to functional data analysis in the field of human movement and biomechanics. The importance of registration (or alignment) is central in this chapter. In fact, alignment is a current topic in both functional data analysis and chemometrics.

Perspectives are discussed in Chapter 7.

As a supplemental paper, we include an application of mixed models in Paper IV. This work does not utilise functional data analysis, but the biomarker data originate from the same human nutrition metabolomics study that motivated the pilot study in Paper I.

The papers and are attached in the format of the journal where they were accepted for publication (Paper III) or published (Paper IV) and in manuscript form where they are in preparation for submission to be published (Papers I and II).

2 Metabolomics

2.1 Metabolomics

"Metabolomics is the systematic study of the unique chemical fingerprints that specific cellular processes leave behind" (Daviss, 2005).

This is only one of the numerous definitions of metabolomics. Van der Greef and Smilde (2005) defined metabolomics as "the comprehensive quantitative and qualitative analysis of all small molecules (in samples of cells, body fluids, tissues, etc.)".

The terms metabolomics and metabonomics are, in practice, used interchangeably and the distinction is largely philosophical: according to Nicholson and Lindon (2008) "*Metabolomics* seeks an analytical description of complex biological samples, and aims to characterize and quantify all the small molecules in such a sample" and "*Metabonomics* broadly aims to measure the global, dynamic metabolic response of living systems to biological stimuli or genetic manipulation. The focus is on understanding systemic change through time in complex multi cellular systems."

In humans, metabolic response can be due to lifestyle, diet (nutrition), disease, gut microflora, drugs, toxins, environment, and genetic modulations (Beckonert et al., 2007).

2.2 Analysis of urine - a short history

Metabonomic fingerprint data are generally sourced from cell extracts, tissue extracts, or biofluids. Biofluids include urine, serum, plasma, cerebrospinal fluid and saliva from animals or humans. (Beckonert et al., 2007)

Bouatra et al. (2013) described urine (in mammals) as surplus water, sugars, soluble wastes and other compounds extracted from the bloodstream by the kidneys. Metabolic breakdown products in urine can originate from a variety of sources: nutritional (solids or liquid), medication, by-products from bacteria, inner-waste metabolites and environmental contamination. High concentrations of certain compounds can be expected in urine: urea (generated by metabolism of amino acid), organic acids, creatinine, ammonia, coloured haemoglobin breakdown products, water-soluble toxins and inorganic salts (potassium, sodium and chloride) (Bouatra et al., 2013).

It has long been recognised that urine is more that just a waste product (Bouatra et al., 2013). Urinalysis has been around for over 6000 years (Echeverry et al., 2010) and medical texts from Babylon, Egypt, and the Far East refer to the use of urine to diagnose diseases (Eknoyan, 2007).

Around four centuries ago, Hindus practicing Ayurveda described insects being attracted to certain patients' urine. They also made the link between sweet-tasting urine and certain diseases. At about the same time Chinese traditional healers diagnosed diabetes, using ants to distinguish between high and low levels of glucose in urine. (Van der Greef and Smilde, 2005)

Hippocrates (around 400 BC) first described the use of urine to interpret human body functioning and for prognostic purposes, i.e. prediction of outcomes of illness. However, Theophilus (around 700 AD) described the systematic use of uroscopy for diagnosis of illnesses (Kouba et al., 2007).

The term uroscopy comes from the Greek words for urine and visual examination. It refers to the macroscopic examination of urine and it informed diagnosis and treatment (Pardalidis et al., 2008). In uroscopy, four criteria were applied: the consistency, odour, non-soluble constituents and, most importantly, the colour of urine (Wittern-Sterzel, 1999).

The Byzantines adopted uroscopy and closely associated urine with food intake, digestive disorders and the liver's bile production that influenced the heart and the body at large (Pardalidis et al., 2008). Papers from Byzantine (Theophilos, 7^{th} century) and Egypt (Judãus, 10^{th} century) were translated into Latin and influenced medieval Western medicine.

Throughout the Byzantine era and even past the Middle Ages (Voswinckel, 2000), urine colour wheels were widely used as diagnostic tools (Nicholson and Lindon, 2008). The oldest known colour wheel dates to 1400 (Wittern-Sterzel, 1999). These diagrams associated the colours, odours and tastes of urine with different diseases. Obviously, these characteristics are of metabolic origin. Although metabonomics/metabolomics relies on state-of-the-art analytical chemistry, the fundamental concept remains unaltered: linking chemical patterns to biology. (Nicholson and Lindon, 2008)

2.3 The Human Metabolome

The metabolome can be defined as "the entire complement of all the small molecular weight molecules (metabolites in cells, body fluids, tissues etc.)" (Van der Greef and Smilde, 2005). Metabolomics is a more recent development in the 'omics' sciences, following genomics, transcriptomics and proteomics. Unlike other 'omes' the metabolome is not even close to near-complete coverage. (Bouatra et al., 2013).

The Human Metabolome Database (HMDB) (www.hmdb.ca), first published in 2007, provides the most recent and complete coverage of the human metabolome. The latest HMDB (version 3.5) contains more than 41 519 metabolite entries. These metabolites comprise

water-soluble as well as lipid soluble metabolites. Furthermore, both rare (< 1 nM) and abundant (> 1 uM) metabolites are included. (Wishart et al., 2013)

The HMDB contains detected and expected metabolites in blood, urine, cerebrospinal fluid (CSF), saliva, other biofluids and tissue. The origin of these metabolites can be generated by human cells or endogenous gut microflora, a toxin/pollutant, drug derived, microbial and/or food derived. Apart from spectroscopic information about human metabolites, the HMDB also contains their associated enzymes, their abundance and their relation to diseases. The HMDB is freely available at www.hmdb.ca (Wishart et al., 2013)

2.3.1 The human urine metabolome

Consequently, a metabolome-wide characterisation of human urine was conducted. The Urine Metabolome Database (UMDB: www.urinemetabolome.ca) describes metabolites that are detectable with today's technology, as well as their concentrations and known associated diseases. They identified 445 unique urine metabolites or metabolite species through experiments. Literature mining produced identification of an additional 2206 compounds found in urine. The UMDB contains all 2651 small-molecule metabolites found in human urine, their concentrations and known related diseases. (Bouatra et al., 2013)

2.3.2 NMR in human urine metabolomics

In metabolomics, nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) are widely used to measure a biological system's metabolic state (Liland, 2011).

For the characterisation of urine, NMR is currently the most complete quantitative method. A typical ^{1}H-NMR spectra from human urine is displayed in Figure 2.1. Compared to other analytical techniques, NMR identified and quantified the greatest number of metabolites in urine (209, of which 108 are unique compared to other analytical methods) and also produced the largest chemical diversity. Additionally, NMR requires minimal sample preparation and is non-destructive. For untargeted (global) metabolomics urine analysis, NMR spectroscopy emerged as the method of choice. Nevertheless, NMR is only able to measure approximately 8% (209/2561) of the known human urine metabolome. (Bouatra et al., 2013)

Compared to GC-MS and LC-MS spectroscopy, NMR spectroscopy has the advantages of being non-destructive (the sample is recoverable), fast (2–3 min per sample vs. 20–30 min), requires no separation, and allows for identification of novel compounds. The latter is difficult in GC-MS and LC-MS. The disadvantages of NMR spectroscopy are the requirement for larger samples (0.5 ml), a large instrument footprint, less sensitivity and the inability to detect inorganic molecules, salts or non-protonated compounds. Similar to GC-MS, NMR spectroscopy is a robust and mature technology, and there are various databases and software for identification of metabolites. (Wishart, 2009)

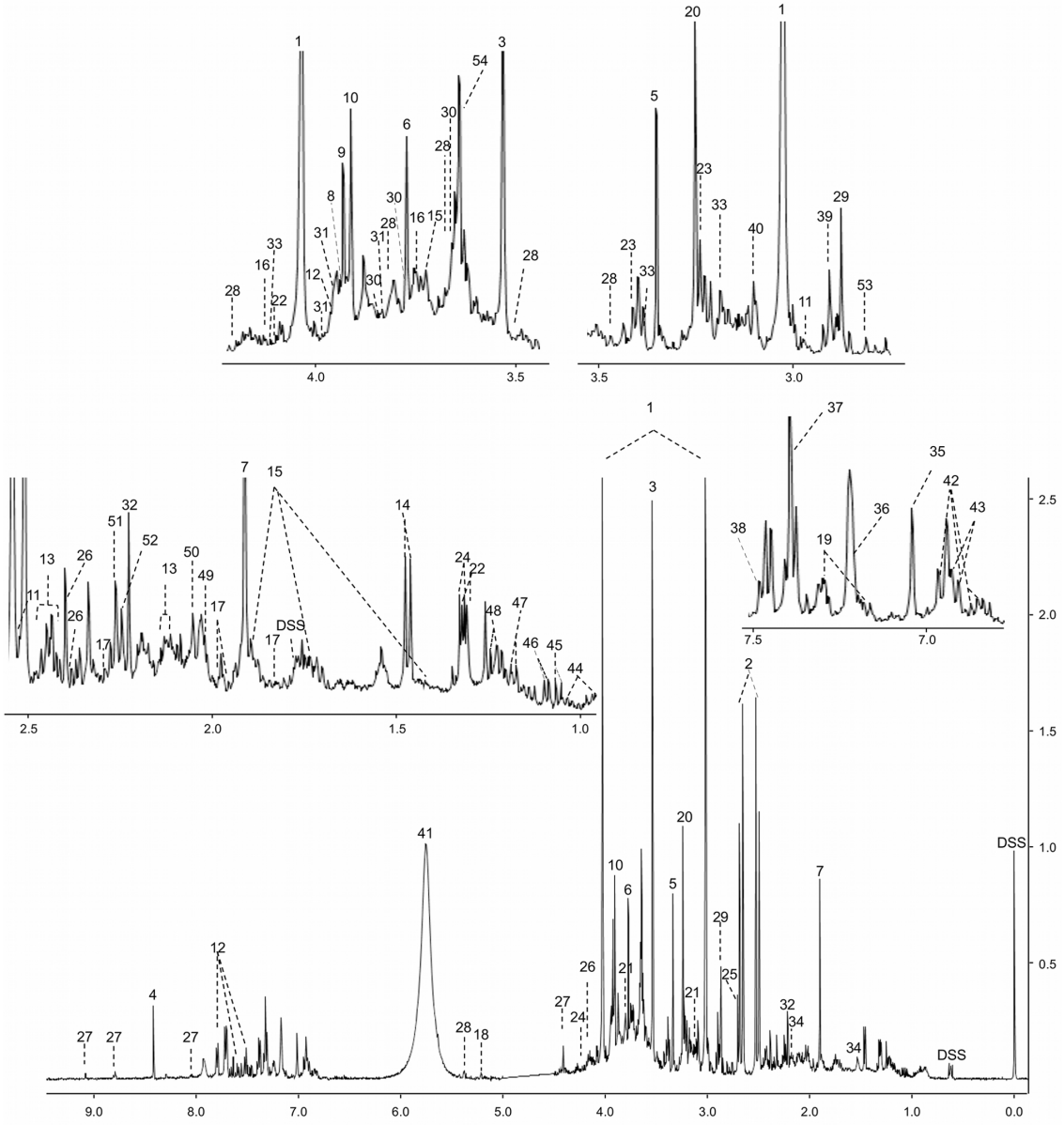

Figure 2.1: Typical 500 MHz ^{1}H-NMR spectra from human urine. Source: Bouatra et al. (2013). Numbers indicates the following metabolites: 1: creatinine; 2: citric acid; 3: glycine; 4: formic acid; 5: methanol; 6: guanidoacetic acid; 7: acetic acid; 8: L-cysteine; 9: glycolic acid; 10: creatine; 11: isocitric acid; 12: hippuric acid; 13: L-glutamine; 14: L-alanine; 15: L-lysine; 16: gluconic acid; 17: 2-hydroxyglutaric acid; 18: D-glucose; 19: indoxyl sulfate; 20: trimethyl-N-oxide; 21: ethanolamine; 22: L-lactic acid; 23: taurine; 24: L-threonine; 25: dimethylamine; 26: pyroglutamic acid; 27: trigonelline; 28: sucrose; 29: trimethylamine; 30: mannitol; 31: L-serine; 32: acetone; 33: L-cystine; 34: adipic acid; 35: L-histidine; 36: L-tyrosine; 37: imidazole; 38: mandelic acid; 39: dimethylglycine; 40: Cis-aconitic acid; 41: urea; 42: 3-(3-hydroxyphenyl)-3-hydroxypropanoic acid (HPHPA); 43: phenol; 45: isobutyric acid; 46: methylsuccinic acid; 47: 3-aminoisobutyric acid; 48: L-fucose; 49: N-acetylaspartic acid; 50: N-acetylneuraminic acid; 51: acetoacetic acid; 52: Alpha-aminoadipic acid; 53: methylguanidine; 54: phenylacetylglutamine

For a detailed discussion of the strengths and weaknesses of different technologies in metabolomics, see Lenz and Wilson (2006). More technical details on NMR data are provided in section 5.1 in Chapter 5.

2.4 Nutritional metabolomics

Savorani et al. (2013) describes nutritional metabolomics as "metabolomics applied to the study of the human (or animal) metabolome as a function of nutritional status or as a function of a nutritional challenge".

It is widely known that nutrition plays a role in both the development as well as the prevention of disease and the promotion of health. Nevertheless, the relationship between a person's nutrition and explicit health/disease results is largely unknown. For example, for two persons with the same diet, what are the reasons that one develops diabetes type 2 and the other remains healthy? (McNiven et al., 2011)

Many lifestyle diseases, including obesity, cardiovascular disease and type-II diabetes, are metabolic disorders which imply a mismatch between what is ingested (i.e. diet) and the needs of the (human) organism (O'Sullivan et al., 2011; Savorani et al., 2013). The aim of nutritional metabolomics is to understand how diet, and adjustment in diet, influences the metabolome (McNiven et al., 2011). For example, Martin et al. (2009) showed that dark chocolate, consumed daily for two weeks, is sufficient to modify the metabolism of healthy human subjects. This is evident in the decreased levels of stress-associated hormones and normalisation of the systemic stress metabolic signatures.

There are two complementary approaches to metabolomics:
Targeted profiling is based on the analysis of a pre-specified group of metabolites associated with a specific metabolic pathway. Certain information regarding the complete metabolic network and its links to the physiological processes underlying health/disease may be lost. (Llorach et al., 2012; Dettmer and Hammock, 2004)

Targeted profiling is often driven by a hypothesis. The selection of metabolites for analysis is based on the questions asked. In nutrition, targeted profiling is used to determine bioavailability, concentration, turnover, or metabolism of nutritional compounds. (Astarita and Langridge, 2013)

In Paper IV Dalskov et al. (2014) we used targeted profiling driven by the hypothesis that high protein intake compromises bone mineralisation in children. Plasma osteocalcin and urinary N-terminal telopeptide of collagen type I were used as biomarkers of bone turnover in children.

Targeted profiling is also called *targeted metabolomics* (Astarita and Langridge, 2013), *metabolic profiling* (Dettmer and Hammock, 2004), *chemometric metabolomics* or *non-quantative metabolomics* (Wishart, 2009).

Non-targeted fingerprinting is a global approach that aims to get an extensive portrait of a whole metabolome. This includes metabolites that are not well characterised or that

are unknown. (Llorach et al., 2012)

The intention of fingerprinting is not to identify all observed metabolites, instead it intends to compare patterns or fingerprints of metabolites that differ in response to an exposure, e.g. diet (Dettmer and Hammock, 2004). Fingerprinting is often hypothesis-generating as opposed to a targeted profiling approach which is hypothesis-driven (Llorach et al., 2012). In nutrition, fingerprinting is used to define individuals' metabolic phenotypes, study metabolite patterns in response to dietary interventions and scan food to determine molecular composition (Astarita and Langridge, 2013).

The benefit of this untargeted approach is that there are no assumptions about candidate metabolites. These candidate metabolites are often unforeseen based on the inadequacy of prevailing knowledge (Primrose et al., 2011).

The wavelet-based functional mixed model approach to chemometrics, as applied in our Paper I (Muller and Tolver, 2014) corresponds to a fingerprinting approach in metabolomics. We study metabolite fingerprint patterns in response to a dietary intervention in the presence of gender and dietary culture differences.

Fingerprinting is also called *untargeted metabolomics* (Astarita and Langridge, 2013), *metabolic fingerprinting* (Dettmer and Hammock, 2004), *metabonomics*, *global metabolic profiling* (Lenz and Wilson, 2006) or *quantative metabolomics* (Wishart, 2009).

2.4.1 Metabotypes and variation

Metabolites provide snapshots of metabolic processes and metabolomics enables the characterisation of individual metabolic phenotypes or metabotypes (Rezzi et al., 2007; Rubio-Aliaga et al., 2012).

Gavaghan et al. (2000) defined the metabotype as a "probabalistic multi-parametric description of an organism in a given physiological state based on analysis of its cell types, biofluids or tissues". Metabotype is n-dimensional and makes it possible to statistically compare the influence of interventions or disease progression on metabolism (Gavaghan et al., 2000).

Variation in the human metabolome, and thus in metabolic fingerprints, can be attributed to a number of factors: age, gender, body composition, body mass index (BMI), cultural differences, dietary factors (e.g. nutrient intake, nutrient-nutrient interactions), diurnal variation, physiological and lifestyle factors (e.g. exercise, smoking, stress), menstrual cycle, gut microflora, genetic variability and host-microbial interactions (Brennan, 2008; Heinzmann et al., 2011; Jenab et al., 2009).

In metabolomics studies the differences among individuals are often greater than the treatment effect (Scalbert et al., 2009). A standardised diet can, to some extent, control the inter-individual differences in a study of urine samples (Walsh et al., 2006). In Paper I, Muller and Tolver (2014), we use a functional data approach to analyse data from a diet standardisation study with participants of both genders and from different cultures.

Variation can also be caused by non-compliance to dietary interventions and differences in the time at which samples were taken related to meals and fluid intake (Scalbert et al., 2009). Variation caused by sampling and analytical methods should be controlled, e.g. sample collection and treatment, storage conditions and analytical instrument performance (Jenab et al., 2009). Unwanted variation in metabolomics studies should be controlled, as far as possible, and should be taken into account in the interpretation of results (Scalbert et al., 2009).

2.4.2 Biomarkers in nutritional metabolomics

Nutritional metabolomics often focuses on the definition of normal physiological variation and differences in metabolomic profiles due to specific dietary interventions. The potential for identification of dietary biomarkers has emerged more recently. (O'Sullivan et al., 2011)

A nutritional (or dietary) biomarker is an indicator of nutritional status, dietary intake, nutrient metabolism, or biological results of dietary patterns or intake. Biomarkers can be clinical, biochemical or functional in nature. (Potischman, 2003)

Nutritional biomarkers from metabolomics can potentially be used as markers of (Llorach et al., 2012; Potischman, 2003)

- dietary intakes in observational studies (nutritional/dietary exposure, food consumption);
- biological effects of a nutritional intervention (nutritional impacts);
- biological effects of dietary habits (personal nutritional status);
- dietary compliance in controlled trials; and
- metabolic mechanisms in a particular metabolic phenotype, in response to a diet.

Dietary biomarkers can more accurately assess nutritional intake compared to self-reported methods (Potischman, 2003). Food-specific biomarkers in urine have been associated with dietary intake of red meat, cooked meats, fish, vegetables, citrus fruits, coffee, green and black tea (O'Gorman et al., 2013; Astarita and Langridge, 2013).

Metabolomics in human nutrition is a growing field and may encourage novel biomarker discovery for specific food consumption and, as a result, for health status (Zivkovic and German (2009) cited in Hedrick et al. (2012)). Nutritional deficiencies in population cohorts could, in the future, be routinely assessed, once rapid assays for biomarkers of food intake have been developed (Primrose et al., 2011). Research encouraging compliance to national nutrition recommendations could benefit greatly from biomarkers that estimate the intake of specific dietary components and foods (Hedrick et al., 2012). Screening of metabolites could soon be used to monitor food consumption in epidemiological or dietary intervention studies, together with self-reported methods for dietary intake (Llorach et al., 2012; O'Sullivan et al., 2011).

Combining metabolic responders with non-responders in an intervention study is an important source of variation and could be the reason for studies with different conclusions. Biomarkers of response could, in future, be used to stratify subjects and reduce variance in data, enhancing identification of biologically significant effects. (Zeisel et al., 2013)

2.4.3 Personalised health and nutrition

Nutrition scientists used to assume that humans are metabolically alike, but evidence is increasingly pointing to considerable metabolic individuality. This is sparking interest in personalised nutrition and lifestyle recommendations. (Zeisel et al., 2013; Heinzmann et al., 2011; Brennan, 2008)

Specific human metabolic phenotypes display variations in dietary requirements. This suggests the potential role of metabolomics in personalised nutrition, where diet is attuned to the nutritional needs of the individual (Astarita and Langridge, 2013; Brennan, 2008).

Heinzmann et al. (2011) demonstrated the importance of individuals' metabotype identification as a starting point for lifestyle intervention. Future stratified medicine programmes and personalised health care are likely to rest on this new paradigm of metabotype stratification of individuals for implementation of dietary and drug interventions (Heinzmann et al., 2011).

By comparing metabotypes between healthy versus diseased groups, or between treatment versus control groups, patterns of variation can be determined. Furthermore, positive treatment outcomes of pharmacological or dietary interventions can be monitored by transition of patients from the cluster of diseased metabolic phenotype to the cluster of healthy metabolic phenotype. (Nicholson et al., 2012)

2.5 Chemometric methods in Metabolomics

Spectral data in metabolomics, e.g. NMR spectra, are often analysed using chemometric methods. According to Wold (1995) "The art of extracting chemically relevant information from data produced in chemical experiments is given the name of chemometrics in analogy with biometrics, econometrics, etc.". Wishart (2007) defined chemometrics as "the application of mathematical, statistical, graphical or symbolic methods to maximize the information which can be extracted from chemical or spectral data".

Chemometric methods are typically among methods that are called *multivariate analysis* (MVA) in statistics. In metabonomics, the most widely used chemometric method is principal component analysis (PCA) (Trygg et al., 2006). Projections to latent structures (PLS), also called partial least squares (PLS) regression, and orthogonal PLS (O-PLS) are also popular methods (Barding et al., 2012). When the response variable is categorical, partial least squares discriminant analysis (PLS-DA) and orthogonal PLS-DA (O-PLS-DA)

can be used. Statistical total correlation spectroscopy (STOCSY) generates a pseudo-two-dimensional spectrum from a set of spectra and visualises the correlation among peak heights across all spectra Cloarec et al. (2005). Other data analysis techniques include soft independent modelling of class analogy (SIMCA), analysis of variance (ANOVA), multivariate ANOVA (MANOVA), ANOVA-simultaneous component analysis (ASCA), k-means clustering, hierarchical clustering, artificial neural networks and support vector machines (Wishart, 2009).

Standard chemometric methods typically do not explore the rich functional nature of metabolomics data.

3 Functional Data Analysis

In this chapter we aim to give a concise conceptual overview of the basic steps involved in functional data analysis (FDA) while keeping mathematical details to a minimum. We also mention some of the many possible statistical methods that can be applied to functional data. This chapter is aimed at the reader not familiar with FDA.

In conventional data analysis the data consist of a set of measurements or observations. In functional data analysis the data consist of a set of functions or curves. Each function is measured at discrete points along a continuum. The continuum is often time, but can be any continuous domain.

The assumption in FDA is that the underlying process generating the data is smooth, although the data are still observed at discrete time points and subject to measurement error, i.e. noise. The underlying process may typically be measured on as few as 20 or up to tens of thousands of discrete points on the continuum. Additionally the process may also be measured repeatedly, either multiple samples of a single process (within subjects), or samples from the process measured in multiple subjects (across subjects). In many data sets a given observation is dependent on adjacent observations, i.e. correlated. This situation violates the independence assumption in traditional multivariate analysis. In FDA we do not assume that adjacent observations are independent.

FDA operates on functions instead of single data values. For example, in conventional data analysis we calculate the mean of a sample of single data values, but in functional data analysis we calculate the mean of a sample of functions.

Ramsay and Silverman (2005) made FDA widely known through the first edition of their monograph in 1997.

3.1 Smoothing

The first step in FDA is to smooth discretely observed data points to obtain a single functional datum or object. The original discrete data points are then 'discarded' and only the functions, and possibly their derivatives, are used in the analyses that follow.

A variety of smoothing methods are available. Often basis expansions are used and smoothness is imposed by either restricting the basis or by explicitly specifying a roughness

penalty. Fourier bases, polynomial spline bases and B-spline bases are popular. Alternatively, free-knot splines and wavelets provide data-adaptive basis systems. Wavelets are especially useful for data with sharp peaks. Splines are well suited to cases where derivatives of functions are required.

We used wavelets to smooth discretely observed nuclear magnetic spectra (Chapter 5 and (Muller and Tolver, 2014)).

3.2 Registration or Feature Alignment

There are two sources of variability present in smoothed curves that form the functional data. Amplitude variation is displayed in the different size of features between curves: the height of curve peaks and the depth of curve valleys. Phase variation can be seen in the difference in the timing, or location on the continuum, of specific features between curves. Phase variation is often referred to as misalignment of curves. The aim of registration in functional data is to separate amplitude and phase variation by aligning curves. Functional registration is also called warping, time warping or alignment.

The most well known registration methods in FDA are landmark registration and continuous registration. Landmark registration uses well-defined features in the data and warps the curves in such a way that these features appear at the same time (or at the same position on the horizontal axis) for all curves. Continuous registration uses a measure of closeness to quantify the similarity between curves. The method aligns curves by warping their time (or horizontal axis) parameters by selecting the optimal warping function from a class of warping functions in order to maximise the similarity between curves. In practice alignment is done to a reference curve. However, in the absence of a reference curve an iterative process is used: estimating a reference curve, for example, the mean curve, and aligning to this estimated reference curve and repeating this process. Note that the functional registration is always performed on curves and not on data points.

It is essential to perform registration of the smooth functions before further analyses, since misalignment can have a serious effect on results.

3.3 Derivatives and Phase-plane plots

Smoothness of a curve usually implies that a number of derivatives can be calculated from the data. The first derivative, 'velocity' indicates the rate of change. The second derivative 'acceleration' indicates the curvature of the function. Analysis of these derivatives is an important aspect of FDA: phase-plane plots display velocity versus acceleration and differential equations are applied as models to describe dynamic processes. Classical multivariate statistical methods typically either do not have access to or do not take advantage of the derivatives of the underlying functions.

We used the first and second derivates of nuclear magnetic resonance peaks to create phase-plane plots in Paper II. We call these phase-plane plots of Lorentzian curves 'heart plots' (Muller and Ramsay, 2014).

3.4 Analysis

Many statistical methods have counterparts in functional data analysis: functional ANOVA, (Zhang, 2013), functional linear models, generalised functional linear models, functional principal component analysis, functional clustering and functional classification to name a few. Functional regression can use a functional response and/or functional covariates: scalar-on-function, function-on-scalar and function-on-function regression. Functional data analysis methods respect the structure found in complex data and can accommodate data measured within and across subjects.

In Paper I we used wavelet-based functional mixed models on NMR data from a nutri-metabolomics study (Muller and Tolver, 2014).

3.5 Functional data analysis in chemometrics

Alsberg (1993) introduced the idea to represent spectra by continuous functions to the chemometric community. More than a decade later Saeys et al. (2008) mentioned that the potential of functional data analysis was still not well known to most chemometricians and suggested a functional data approach to spectrometric data. Nevertheless, standard and widely used methods in chemometrics (and metabolomics) rely heavily on multivariate statistical methods (Section 2.5) and rarely utilise FDA. A bi-annual review of the field of chemometrics summarised the development of new methods in chemometrics and novel or important applications of these methods over the past two decades (Lavine and Workman, 2013, 2010, 2008, 2006, 2004, 2002; Lavine, 2000, 1998). Functional data analysis has never featured in these reviews. Nevertheless, there have been a number of publications that considered FDA applications in chemometrics.

Published applications of FDA in chemometrics include the following: functional principal component regression and functional partial least squares (Reiss and Ogden, 2007), functional linear regression with a scalar response, functional ANOVA to analyse spectroscopic data from designed experiments (Saeys et al., 2008) linear regression with functional predictors and scalar responses (Zhao et al., 2012) All of the above applications were in near-infrared (NIR) spectroscopy where the data are measured as functions of wavelengths.

Berk et al. (2011) described a smoothing splines mixed effects (SME) model for metabolomic time course data. They treated longitudinal measurements (within each spectral bin) as a smooth function of time and performed a functional t-test to detect between-group differences. Statistical significance was assessed using non-parametric bootstrapping. Our approach is similar in terms of using an FDA approach, more specifically a functional

mixed model, to NMR metabolomic data with the aim of detecting biomarker differences between groups. However, we consider each NMR profile (consisting of many metabolites) as a function over chemical shift (ppm) and model differences between groups as a fixed effect. These groups can be defined by a covariate, a treatment or even discrete times. We can alternatively include time as a continuous covariate in our model. We use wavelets for estimating the spiky NMR profiles, before applying the mixed effect model. On the contrary, Berk et al. (2011) used smoothing splines to estimated differences in groups for individual metabolite (more technically individual spectral bin) functions over time (not chemical shift) to be modelled using a mixed effect model. Berk et al. (2011) determined p-values for each spectral bin from a nonparametric bootstrap procedure and corrected for multiple testing using the false discovery rate.

3.6 Further reading

For further reading Levitin et al. (2007) provides a conceptual introduction to FDA in the context of psychology and behavioural science. Sørensen et al. (2013) gives an introduction to FDA with medical applications. The monograph on functional data analysis Ramsay and Silverman (2005) and accompanying book on case studies Ramsay and Silverman (2002) provide a comprehensive theoretical basis and a wide variety of applications. For an introduction to FDA with R and MATLAB, see Ramsay et al. (2009).

4 Wavelets and wavelet shrinkage

In functional data analysis, a variety of smoothing methods are used (see section 3.1). We chose wavelets to convert discretely observed noisy spectral data to smooth functions. Wavelets are families of basis functions and are widely used in signal processing. Wavelets exhibit time and frequency localisation, i.e the ability to accommodate smooth as well as spiky functions efficiently (Hastie et al., 2009). This is an important feature for the smoothing of spectral data, since a spectral signal often contains relatively smooth features on a larger scale as well as characteristic spiky (or bumpy) features on a smaller, more local scale.

We provide a mathematical introduction to wavelets in Section 4.1. This includes multiresolution analysis, families of orthonormal wavelet bases and the discrete wavelet transform. For a detailed explanation of wavelet theory the reader is referred to the original text of Daubechies (1992). Percival and Walden (2006) covers wavelets for time series analysis and Ogden (1997a) covers the use of wavelets in statistics. In Section 4.2 we cover wavelet shrinkage and thresholding. A number of practical issues relating to the choice of parameters in wavelet transformation and shrinkage of spectral data are discussed in Section 4.3.

In essence, wavelet coefficients describe features of a function at different times and frequencies. In this way, the wavelet decomposition gives a *time and frequency localisation*, also called a *location and scale decomposition* of the underlying function. Wavelet decomposition provides a sparse representation of a function and is fast to compute.

4.1 A mathematical introduction to wavelets

Functional data analysis deals with discrete observations $(y_{i1}, \ldots, y_{in_i})$ of functional variables $f_i : I \to \mathbb{R}$, $i = 1, \ldots, N$. Typically, we will assume that f_i belongs to the Hilbert space $L^2(\mathbb{R})$ of square integrable functions and we may be interested in expanding $f_1, \ldots, f_N$ in a suitable finite basis.

Obviously, there is no unified basis capable of representing (or approximating) any finite sample of functions $f_1, \ldots, f_N$ using only a limited number of non-zero coefficients. However, it is possible to construct families of different basis systems allowing a reasonably sparse representation of many functional data set in terms of a least one of the basis

systems. The purpose of this section is to introduce families of basis systems based on orthogonal wavelets and discuss their ability to efficiently yield a sparse approximation to a discrete sample of a function.

4.1.1 Motivation

Suppose $(y_1, \ldots, y_n)$ are discrete observations of a function f on [0,1]. We consider the approximation of f by the piecewise constant function

$$\tilde{f}(t) = \sum_{k=1}^{n} y_k \mathbb{I}\{\tfrac{k-1}{n} \leq t < \tfrac{k}{n}\}$$

where $\mathbb{I}$ is the indicator function. The sum can be considered as a representation of $\tilde{f}$ in terms of the orthonormal basis in $L_2(\mathbb{R})$ given by

$$\sqrt{n}\mathbb{I}\{\tfrac{k-1}{n} \leq t < \tfrac{k}{n}\}, \quad k = 1, \ldots, n. \tag{4.1}$$

If $n = 2^J$ denote by V_J the subspace of $L_2(\mathbb{R})$ spanned by (4.1) and denote

$$\phi_{J,k}(t) = 2^{J/2}\mathbb{I}\{\tfrac{k-1}{2^J} \leq t < \tfrac{k}{2^J}\}, \quad k = 1, \ldots, 2^J.$$

Since $n = 2^J$ is even $V_{J-1} \subset V_J$ and the following relation exists between the basis vectors

$$\begin{aligned}
\phi_{J-1,k}(t) &= 2^{(J-1)/2}\mathbb{I}\{\tfrac{k-1}{2^{J-1}} \leq t < \tfrac{k}{2^{J-1}}\} \\
&= 2^{(J-1)/2}\mathbb{I}\{\tfrac{2k-2}{2^J} \leq t < \tfrac{2k-1}{2^J}\} + 2^{(J-1)/2}\mathbb{I}\{\tfrac{2k-1}{2^J} \leq t < \tfrac{2k}{2^J}\} \\
&= \frac{1}{\sqrt{2}}\phi_{J,2k-1}(t) + \frac{1}{\sqrt{2}}\phi_{J,2k}(t).
\end{aligned}$$

To formulate an orthonormal basis for the orthogonal complement of V_{J-1} within V_J let

$$W_{J-1} = V_J \cap V_{J-1}^{\perp}$$

and

$$\psi_{J-1,k}(t) = 2^{(J-1)/2}\mathbb{I}\{\tfrac{2k-2}{2^J} \leq t < \tfrac{2k-1}{2^J}\} - 2^{(J-1)/2}\mathbb{I}\{\tfrac{2k-1}{2^J} \leq t < \tfrac{2k}{2^J}\}, \quad k = 1, \ldots, 2^{J-1}.$$

It follows that $\{\psi_{J-1,k}\}_k$ are mutually orthogonal as well as orthogonal to any basis vector $\phi_{J-1,k}$ of V_{J-1}. Note that

$$\phi_{J-1,k}(t) + \psi_{J-1,k}(t) = 2 \cdot 2^{(J-1)/2}\mathbb{I}\{\tfrac{2k-2}{2^J} \leq t < \tfrac{2k-1}{2^J}\} = \sqrt{2}\phi_{J,2k-1}(t)$$

and

$$\phi_{J-1,k}(t) - \psi_{J-1,k}(t) = 2 \cdot 2^{(J-1)/2}\mathbb{I}\{\tfrac{2k-1}{2^J} \leq t < \tfrac{2k}{2^J}\} = \sqrt{2}\phi_{J,2k}(t).$$

Thus $\psi_{J-1,k}(t)$ spans W_{J-1}. Thus, we have two different orthonormal bases for V_J: $\{\phi_{J,k}\}_{k=1}^{2^J}$ or the union of $\{\phi_{J-1,k}\}_{k=1}^{2^{J-1}}$ and $\{\psi_{J-1,k}\}_{k=1}^{2^{J-1}}$. Changing from coordinates in the former

basis (y) to the latter basis (say d) is a linear mapping and can be represented by an $n \times n$ matrix W where

$$d = Wy. \tag{4.2}$$

In general, the complexity of computing the matrix product Wy is of $O(n^2)$. However, for the particular bases considered here, the computation can be performed in only $O(n)$ operations, using the fast pyramid algorithm (Mallat, 1989). This is quite remarkable. This 'fast' algorithm enables us to change rapidly between the two orthonormal basis systems used to represent functions in V_J.

By iterating the construction above we obtain orthogonal decompositions

$$V_J = V_{J_0} \oplus W_{J_0} \oplus \ldots \oplus W_{J-1}$$

for any $J_0 = 0, \ldots, J-1$, where

$$\phi_{j,k}(t) = 2^{j/2}\mathbb{I}\{\tfrac{k-1}{2^J} \leq t < \tfrac{k}{2^J}\} \quad , k = 1, \ldots, 2^j$$

is an orthonormal basis for V_j and

$$\psi_{j,k}(t) = 2^{j/2}\mathbb{I}\{\tfrac{2k-2}{2^{j+1}} \leq t < \tfrac{2k-1}{2^{j+1}}\} - 2^{j/2}\mathbb{I}\{\tfrac{2k-1}{2^{j+1}} \leq t < \tfrac{2k}{2^{j+1}}\} \quad , k = 1, \ldots, 2^{j-1}$$

is an orthonormal basis for W_j. Computationally fast algorithms exist to move between coordinate representations corresponding to different choices of J_0.

The motivation for changing between the basis systems is that for many functions there exists a choice of J_0 such that the corresponding basis representation is sparse. Moreover, we have computationally fast algorithms to find a basis representation allowing us to disregard many of the coefficients that are close to zero and yet still have a good approximation to the original function.

The discontinuity of the basis functions considered above transfers to the functions in the vector spaces V_j and W_j. This is unsatisfactory if the observation $(y_1, \ldots, y_n)$ is a discrete sample of a continuous function. Therefore we seek (in the next section) to generalise the construction to obtain other systems of orthonormal basis functions based on smoother functions. We still focus on the ability to efficiently move from coordinate representations with respect to the different basis systems.

4.1.2 Multiresolution analysis and wavelets

Any increasing family

$$\ldots \subset V_{j-1} \subset V_j \subset V_{j+1} \subset \ldots, \quad j \in \mathbb{Z},$$

of closed subspaces of $L_2(\mathbb{R})$ allows us to consider the orthogonal complement $W_j = V_{j+1} \cap V_j^{\perp}$ and decompositions of the form

$$V_j = V_{j_0} \oplus W_{j_0} \oplus \ldots \oplus W_{j-1}$$

for any $j_0 \leq j$. The question arises: when do we have efficient algorithms to switch between representations of the same function $f \in V_j$ corresponding to different $j_0 \leq j$?

Similar to the motivating example above, we impose more structure on the spaces V_j. We require the spaces V_j to be scaled versions of each other

$$t \to f(t) \in V_j \Leftrightarrow t \to f(2^j t) \in V_0$$

and that V_0 is invariant to translations

$$t \to f(t) \in V_0 \Leftrightarrow t \to f(t-k) \in V_0,$$

for any $k \in \mathbb{Z}$. If further $\overline{\cup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R})$ and $\cap_{j\in\mathbb{Z}} V_j = \emptyset$ we say that $\{V_j\}_{j\in\mathbb{Z}}$ constitutes a *multiresolution analysis* in $L_2(\mathbb{R})$.

The definition of a multiresolution analysis given above does not involve any basis system for V_j. Consequently, the concept is too general to guarantee efficient orthogonal decompositions as in Section 4.1.1. It is therefore common to consider multiresolution analyses where each V_j is given as the closed vector space spanned by

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k), \quad j,k \in \mathbb{Z},$$

where the generating function ϕ is referred to as the *father wavelet* of the family $\{V_j\}_{j\in\mathbb{Z}}$. We first note that the requirement that $\{V_j\}_{j\in\mathbb{Z}}$ be increasing puts heavy restrictions on which father wavelets may be used as generator for a multiresolution analysis. Further, observe that if $\{\phi_{0,k}\}_{k\in\mathbb{Z}}$ happens to be an orthonormal basis for V_0 then $\{\phi_{j,k}\}_{k\in\mathbb{Z}}$ will be a orthonormal basis for V_j for any $j \in \mathbb{Z}$.

Example 1. *The vector spaces V_j generated by the father wavelet*

$$\phi(t) = \begin{cases} 1 & , t \in [0,1) \\ 0 & , otherwise \end{cases}$$

constitute a multiresolution analysis. This follows from the fact that the father wavelet may be written as

$$\phi(t) = \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t) + \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t-1)$$

where the righthand side is a linear combination of functions in V_1. Since $\phi_{0,k}$ lives on disjoint intervals, we further have that $\{\phi_{j,k}\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_j for any $j \in \mathbb{Z}$.

It follows from Section 4.1.1 that an orthonormal basis for $W_j = V_{j+1} \cap V_j^{\perp}$ for this multiresolution analysis is given by

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \quad k \in \mathbb{Z},$$

where

$$\psi(t) = \begin{cases} 1 & , t \in [0,1/2) \\ -1 & , t \in [1/2,1) \\ 0 & , otherwise \end{cases}.$$

The generating function ψ is referred to as the mother wavelet. □

The functions $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ of the previous example constitute a, so-called, orthonormal *wavelet basis* for $L_2(\mathbb{R})$. In this example, the wavelet basis is associated with a multiresolution analysis and we can write down a father wavelet ϕ generating the multiresolution analysis. The fundamental identity behind the multiresolution analysis is that

$$\phi(t) = \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t) + \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t-1)$$

and the identity linking the father and mother wavelets is

$$\psi(t) = \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t) - \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t-1).$$

This leads to the following definition.

Definition 1 (Wavelet analysis)**.** *A* wavelet analysis *is an orthonormal wavelet basis for* $L_2(\mathbb{R})$ *consisting of functions*

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \quad j,k \in \mathbb{Z}, \tag{4.3}$$

generated by a mother wavelet ψ*, with* j *the dilation index and* k *the translation index. Furthermore, the wavelet basis should be associated to a multiresolution analysis* $\{V_j\}_{j\in\mathbb{Z}}$ *generated by some father wavelet* ϕ *so that*

- $W_j = V_{j+1} \cap V_j^{\perp} = \overline{Span\{\psi_{j,k}|k\in\mathbb{Z}\}}$
- $\{\phi_{0,k}\}_{k\in\mathbb{Z}}$ *is an orthonormal basis for* V_0
- $V_j = \overline{Span\{\phi_{j,k}|k\in\mathbb{Z}\}}$

where we have defined $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$, $j,k\in\mathbb{Z}$. □

For a given wavelet basis constituting a wavelet analysis, any function $f \in V_J$ has infinitely many coordinate representations (one for each $j_0 \leq J$):

$$f(t) = \sum_{k\in\mathbb{Z}} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t).$$

For a wavelet analysis we refer to V_j as the *scale space* at level j generated by the *scale function* ϕ.

4.1.3 Families of orthonormal wavelet bases

In this section we discuss the construction of families of orthonormal wavelet bases. Any of these families of orthonormal wavelet bases can form the basis for a wavelet analysis. Though it may seem difficult to meet all requirements in Definition 1 a key result by Daubechies (1992) allows the construction of a wavelet analysis from a multiresolution analysis.

Start with a father wavelet ϕ. Let V_0 denote the vector space spanned by the functions

$$\phi_{0,k}(t) = \phi(t-k), \quad k \in \mathbb{Z}.$$

For $\{\phi_{0,k}(t)\}_{k\in\mathbb{Z}}$ to be an orthonormal basis of V_0 we need $||\phi|| = 1$ and

$$\forall k \in \mathbb{Z} : \int \phi(t)\phi(t-k)dt = 0.$$

For any $j \in \mathbb{Z}$ then $\{\phi_{j,k}\}_{k\in\mathbb{Z}}$ defined as

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$$

will also be an orthonormal basis for a vector space V_j. For V_0 to be a subspace of V_1 we must have

$$\phi(t) = \sum_{k\in\mathbb{Z}} h_k \phi_{1,k}(t) \tag{4.4}$$

for appropriate *filter coefficients* h_k. The main result by Daubechies (1992) states that if $\{V_j\}_{j\in\mathbb{Z}}$ forms a multiresolution analysis then there is a wavelet basis $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ for the spaces $W_j = V_{j+1} \cap V_j^{\perp}$ and the mother wavelet *can* be chosen as

$$\psi(t) = \sum_{k\in\mathbb{Z}} \underbrace{(-1)^{k-1}\overline{h_{-k-1}}}_{:=g_k} \phi_{1,k}(t) \tag{4.5}$$

Note that the choice of mother wavelet is not unique!

The above result enables us to construct orthonormal wavelet bases for $L_2(\mathbb{R})$: Look for generating functions (ϕ) such that

1. $t \to \phi(t-k)$ are orthogonal for $k \in \mathbb{Z}$
2. there exists filter coefficients $\{h_k\}_{k\in\mathbb{Z}}$ such that

$$\phi(t) = \sum_{k\in\mathbb{Z}} h_k 2^{1/2}\phi(2t-k).$$

3. the sequence of scaling spaces V_j spans $L_2(\mathbb{R})$

For the Haar wavelets discussed in Section 4.1.1 and Example 1 we had $\phi(t) = \mathbb{I}_{[0,1)}(t)$ and $h_0 = h_1 = \frac{1}{\sqrt{2}}$. This implies the following mother wavelet

$$\begin{aligned}\psi(t) &= -\overline{h_0}\phi_{1,-1}(t) + \overline{h_1}\phi_{1,-2}(t) \\ &= -\frac{1}{\sqrt{2}}\sqrt{2}\phi(2t-(-1)) + \frac{1}{\sqrt{2}}\sqrt{2}\phi(2t-(-2)) \\ &= \mathbb{I}_{[-1,-1/2)}(t) - \mathbb{I}_{[-1/2,0)}(t).\end{aligned}$$

Note that the choice of mother wavelet is not unique and usually the preferred choice is

$$\psi(t) = \mathbb{I}_{[0,1/2)}(t) - \mathbb{I}_{[1/2,1)}(t).$$

Example 2 (Compactly supported wavelets)**.** *To simplify the problem of constructing a mother wavelet for an orthonormal wavelet basis, it is convenient to consider families of orthonormal wavelet bases generated by a compactly supported father wavelet ϕ. Assume that the support for some ϕ with $||\phi|| = 1$ is contained in $[0, k_0]$ for some $k_0 \in \mathbb{N}$ then $\{\phi_{0,k}\}_{k\in\mathbb{Z}}$ is an orthonormal system if only*

$$\int \phi(t)\phi(t-k)dt = 0$$

for $0 < |k| < k_0$. Then we only need to consider the finite number of equations for $|k| < k_0$. Consider the vector space V_1 spanned by the orthonormal functions

$$\phi_{1,k}(t) = 2^{1/2}\phi(2t-k), \quad k \in \mathbb{Z}.$$

If we can show that $V_0 \subset V_1$ then it follows more generally that the vector spaces V_j spanned by

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k), \quad k \in \mathbb{Z}$$

constitute a multiresolution analysis provided that $\cap_j V_j = \emptyset$ and $\overline{\cup_j V_j} = L_2(\mathbb{R})$. Assume the filter coefficients h_k are known for

$$\phi(t) = \sum_{k\in\mathbb{Z}} h_k \phi_{1,k}(t).$$

We then have a method for constructing the mother wavelet of an orthonormal wavelet basis. Several examples constructed along this line can be found in Daubechies (1992). □

4.1.4 The discrete wavelet transform

The *discrete wavelet transform* acts on a vector $(y_1, \ldots, y_n)$ of length $n = 2^J$. The first step is to represent the discrete sample by a square integrable function $f \in L_2(\mathbb{R})$. Formally, this is done by approximating $(y_1, \ldots, y_n)$ with $f \in V_J$ where f is of the form

$$f(t) = \sum_{k\in\mathbb{Z}} c_{J,k}\phi_{J,k}(t).$$

We think of $(y_1, \ldots, y_n)$ as a discrete sample of a function $f \in V_J$, obtained over a uniform grid of length $n = 2^J$ on $[0, 1]$. From a theoretical point of view the k-th *scaling coefficient* at level J (i.e. $c_{J,k}$) should be chosen to approximate the integral

$$\int f(t)\phi_{J,k}(t)dt.$$

For the Haar basis we just get $c_{J,k} = y_k$, $k = 1, \ldots, n$ ($c_{J,k} = 0$ otherwise) but better approximations exist for other wavelet bases.

The second step of the discrete wavelet transform is to establish the filter equations to obtain the coefficients corresponding to wavelet decompositions of the form

$$f(t) = \sum_{k\in\mathbb{Z}} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J-1}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t). \tag{4.6}$$

for $j_0 \leq J$. We know that the *wavelet coefficients* ($d_{j,k}$) and the *scaling coefficients* ($c_{j,k}$) are given by inner products

$$c_{j,k} = \int f(t)\phi_{j,k}(t)dt \quad \text{and} \quad d_{j,k} = \int f(t)\psi_{j,k}(t)dt. \tag{4.7}$$

Assume the wavelet basis is given by filter equations of the form (4.4) and (4.5) then, in general,

$$\begin{aligned}
\phi_{j-1,l}(t) &= 2^{(j-1)/2}\phi(2^{j-1}t - l) = 2^{(j-1)/2}\sum_{k\in\mathbb{Z}} h_k\phi_{1,k}(2^{j-1}t - l) \\
&= 2^{(j-1)/2}\sum_{k\in\mathbb{Z}} h_k 2^{1/2}\phi(2(2^{j-1}t - l) - k) \\
&= 2^{(j-1)/2}\sum_{k\in\mathbb{Z}} h_k 2^{1/2}\phi(2^j - (2l + k)) \\
&= \sum_{k\in\mathbb{Z}} h_k\phi_{j,2l+k}(t)
\end{aligned}$$

and

$$\psi_{j-1,l}(t) = \sum_{k\in\mathbb{Z}} g_k\phi_{j,2l+k}(t).$$

The above formulations allow us to recursively compute scaling and wavelet coefficients at level $j - 1$ from scaling coefficients at level j.

Due to orthogonality of the discrete wavelet transform (4.2), the inverse discrete wavelet transform is given by

$$\boldsymbol{Y} = W^T\boldsymbol{d}. \tag{4.8}$$

A wavelet is characterised by a number of vanishing (or zero) moments for a given support (point where the function is not zero). A function $\psi \in L_2(\mathbb{R})$ have v vanishing moments if

$$\int x^m\psi(x)dx = 0$$

for $m = 0, \ldots, v-1$ (unders specific technical conditions). For a wavelet with v vanishing moments all wavelet coefficients of any v-degree polynomial or polynomials of lesser degree will equal zero. When v increases, the wavelet ϕ is smoother, and so is the scaling function ϕ.

4.1.5 Energy preservation and data compression

The *energy* of a signal is the sum of the squared values of the function

$$\|y\|^2 = y_1^2 + y_2^2 + \cdots + y_n^2$$

The orthonormal wavelet transform *preserves energy*: the energy of the wavelet coefficients of a function equals the energy of the function:

$$\|d\|^2 = d^T d = (Wy)^T Wy = y^T W^T Wy = y^T y = \|y\|^2$$

since W is an orthogonal matrix.

On a relatively smooth function the wavelet coefficients will be very close to or equal to zero for the smooth parts where the function behaves as a polynomial of order v or less. Thus, the wavelet transform of a relatively smooth function will be sparse: many wavelet coefficients will have zero values and can be disregarded. On the other hand, discontinuities and noise in a function will be represented by non-zero coefficients. Thus we say the wavelet transform *compacts energy*: a function is 'compressed' into a small set of, typically large, wavelet coefficients with the remaining wavelet coefficients equal to or close to zero. However, the wavelet transform does not compress noise: an orthogonal wavelet transform will transform *iid* Gaussian noise to a set of *iid* Gaussian wavelet coefficients. (Nason, 2008; Walker, 2008)

4.2 Wavelet shrinkage

In the literature wavelet methods are often used as a form of nonparametric regression (Nason (2008), p.83) and occur under various names, including wavelet shrinkage, curve estimation and wavelet regression.

In the process of wavelet shrinkage we observe a function contaminated with additive noise, (1) take a wavelet transform (DWT), (2) modify or shrink the noisy function's wavelet coefficients, and (3) take the inverse wavelet transform to estimate the function (IDWT) (Nason, 2008). In this three-step procedure the estimates of the functions are regularised such that local features, like sharp peaks, are kept but noise is removed (Morris and Carroll, 2006). This is called *adaptive regularisation*. The modification or shrinkage of wavelet coefficients in step 2 can be done by thresholding methods described in sections 4.2.1 and 4.3.3.

In statistical terms, we think of observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ arising from the model

$$y_k = f(t_k) + \epsilon_k, \text{ for } k = 1, \ldots, n \tag{4.9}$$

where $t_k = k/n$. The objective is to estimate the unknown function $f(\boldsymbol{t})$, for $\boldsymbol{t} \in [0, 1]$, using the noisy observations $\boldsymbol{y}$. We assume that $\epsilon_k \sim N(0, \sigma^2)$ are independent, i.e. white noise.

Donoho and Johnstone (1994) introduced the concept of wavelet shrinkage to the statistical literature (Donoho and Johnstone, 1994, 1995; Donoho et al., 1995). Their general idea is that the discrete wavelet transform is applied to (4.9) as described below.

Let W denote the discrete wavelet transform that we choose and let y denote the vector of observations, f the true unknown function and ϵ the noise. Since the discrete wavelet transform is linear, the wavelet transformed model can be written as

$$\begin{aligned} Wy &= Wf + W\epsilon & (4.10)\\ d^* &= d + e & (4.11) \end{aligned}$$

where W is the $n \times n$ orthogonal wavelet transform matrix associated with the orthonormal periodic wavelet basis chosen. d^* is the $n \times 1$ vector of empirical wavelet coefficients.

Donoho and Johnstone (1994) proposed the following wavelet shrinkage technique for estimation of $g(x)$, as described in Nason (2008):

Large values of the empirical wavelet coefficients, d^*, probably consist of true signal (and noise); in contrast, small coefficients probably consist of only noise. Thus, to estimate d, the *thresholding* idea creates and estimates, $\hat{d}$, by removing coefficients in d^* that are smaller than some threshold, and thereby keeps the coefficients that are larger.

4.2.1 Thresholding

The hard and soft thesholding functions are defined by (Donoho and Johnstone, 1994)

$$\begin{aligned} \hat{d} &= \eta_H(d^*, \lambda) = d^* \mathbb{I}\{|d^*| > \lambda\} & (4.12)\\ \hat{d} &= \eta_S(d^*, \lambda) = sgn(d^*)(|d^*| - \lambda)\mathbb{I}\{|d^*| > \lambda\} & (4.13) \end{aligned}$$

where $\mathbb{I}$ is the indicator function, d^* is the empirical coefficient to be thresholded, and λ is the *threshold*. Hard thresholding takes a 'keep' or 'kill' approach in the sense that wavelet coefficients, in absolute value, greater than the threshold λ are kept and those smaller than or equal to λ are set to zero. Soft thresholding also sets wavelet coefficients with absolute value smaller than or equal to λ to zero, but shrinks the remaining coefficients to zero by an amount λ. The choice of λ is crucial and different thresholding methods are discussed in section 4.3.3.

Soft thresholding shrinks large coefficients uniformly towards 0 by λ and thus results in larger bias than hard thresholding (Vidakovic, 1999). On the other hand, the hard

thresholding rule is discontinuous and thus results in larger variance (Vidakovic, 1999). Marron et al. (1998) showed that hard shrinkage has smaller bias but larger variance than soft shrinkage, and that significantly smaller thresholds should be used for soft shrinkage.

There are numerous thresholding methods that can be used in combination with either a hard or soft threshold. Donoho and Johnstone (1994) introduced the universal threshold

$$\lambda = \sigma\sqrt{2\log n} \tag{4.14}$$

where n is the number of observations in the signal and $\hat{\sigma}$, the estimate of the noise level is calculated from some measure of the common standard deviation of the noise ϵ_i. See Section 4.3.3 for more on thresholding methods, specifically *SureShrink*.

In certain applications it may be desirable to retain relatively large-scale components in g. Thresholding is then limited to higher levels, say $j > j_0$ of the empirical wavelet coefficients d^*. For lower levels $j \leq j_0$ $\hat{d}_{j,k} = d^*_{j,k}$. In Section 4.3.4 we discuss the choice of primary resolution j_0.

Using the IDWT with the thresholded wavelet coefficients $\hat{d}$,

$$\hat{f} = W^T\hat{d}. \tag{4.15}$$

we obtain an estimate $\hat{f}$ of the true underlying function f in (4.9) and a smoothed, i.e. denoised version of the noisy observations y.

Measure of error

To judge the successful estimation of f an error measure is defined. The most commonly used error is the l_2 or integrated square error (ISE) which is given by

$$\hat{M} = n^{-1}\sum_{k=1}^{n}\{\hat{f}(t_k) - f(t_k)\}^2. \tag{4.16}$$

This error depends on $\hat{f}$ which depends on the specific error sequence $\{e_k\}$. The mean ISE (MISE), or risk, is defined by $M = \mathbb{E}(\hat{M})$ where M may depend on the estimator, the true function, the number of observations, and the properties of the sequence $\{e_k\}$. M depends not only on the chosen 'smoothing parameters' of the estimator, but also on the underlying wavelet family selected to perform the smoothing. (Nason, 2008)

4.3 Parameter choices and practical considerations

Throughout this section we refer to the *Bumps* test function, defined by Donoho and Johnstone (1994) and displayed in Figure 4.1. Together with other test functions, it has been used in a number of wavelet-related articles. The *Bumps* test function resembles a typical spectrum with a flat baseline and several sharp peaks and can be used as a template for a simplified NMR spectrum.

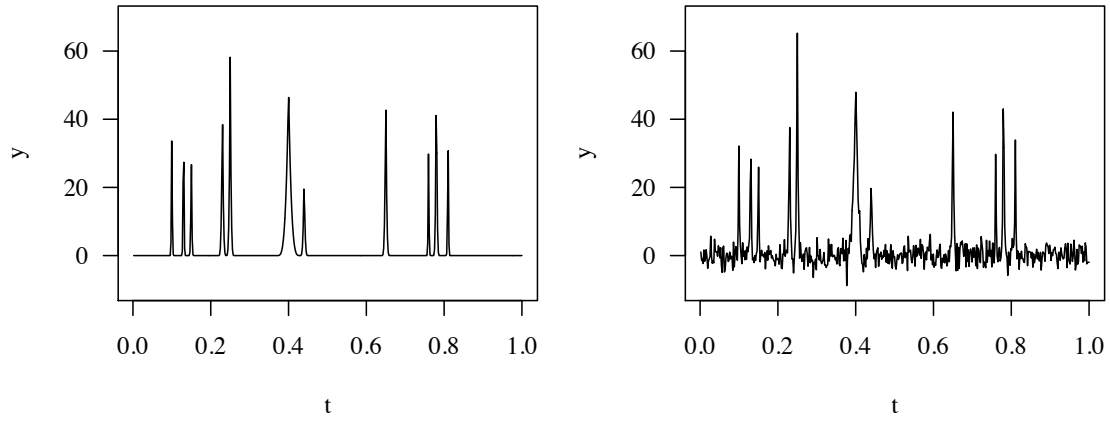

Figure 4.1: The Bumps test function (left) (Donoho and Johnstone, 1994) and with added Gaussian noise of root signal to noise ratio of 3 (right) as used by Antoniadis et al. (2001). Axes are scaled to be similar to Donoho and Johnstone (1994).

4.3.1 Boundary conditions

The discrete wavelet transform takes the vector of scaling coefficients $c_{J,k}$ at level J as its starting point. Depending on the support of the father wavelet ϕ then computation of the coefficients needed to restore the function f on $[0, 1]$ requires that we extend f just outside the interval $[0, 1]$. The two common approaches implemented in the `waveThresh` package (Nason, 2013) is to either recycle or reflect the values of $(y_1, \ldots, y_n)$ near the boundaries.

These two simple solutions are respectively called periodic and symmetric boundary handling. For periodic boundaries, the wavelet and scaling functions are basically 'wrapped around' by pasting the function together at the start and end of the interval. It assumes $f(-x) = f(1 - x)$ where $x \in (-1, 0)$ or $x \in (0, 1)$. Symmetric boundaries assume $f(-x) = f(x)$ and $f(1 + x) = f(1 - x)$ where $x \in (0, 1)$. (Nason, 2008)

The disadvantage of periodic boundaries is the possibility of large wavelet coefficients, without data-related interpretation, for wavelets centred near the boundaries (Ogden, 1997a). Nevertheless, the advantage of independent empirical wavelet coefficients with identical variances for orthogonal wavelet families and Gaussian noise has made periodic boundary handling a popular method (Ogden, 1997a). Symmetric boundary handling preserves continuity of the function, while periodic boundary handling does not, but results in more than 2^j wavelet coefficients at level j and introduces dependencies by having more coefficients than data points (Ogden, 1997a).

Abramovich and Benjamini (1996) and Zhao et al. (2012) used periodic boundary conditions for a modified version of the Bumps test function (Figure 4.1). We also chose to use periodic boundary handling conditions for our analysis of the NMR data from the diet standardisation study.

4.3.2 Sample sizes that are not a power of two

To perform a discrete wavelet transform, the data are required to have a sample size of $n = 2^J$ where J is a positive integer.

The original data set can be pre-conditioned to be of length 2^J where J is a positive integer. Ogden (1997b) described two computationally simple approaches and assumed periodic boundary handling. Firstly, the data set can be extended to the next larger power of two by 'padding' with zeros, or 'padding' with a data value like the last value in the data set. Secondly, interpolation of data values can be done to create a new data set with length of 2^J. Ogden (1997a) (section 6.4) pointed out that 'padding' with zeros will result in zeros being averaged into computation of wavelet coefficients and as such will 'dilute' the signal towards the end of the original data set. In the case of NMR data, where large areas at the ends of each signal are typically regarded as not containing any meaningful peaks, a reasonable alternative is to 'cut' the data carefully at the ends, not removing meaningful peaks, but reducing the data to length 2^J. The residual water peak and urea peak are typically removed by 'cutting' a part from the centre of the NMR signal (section 5.4). It may arguably be better to leave this area intact and thus avoid a jump in the signal where the otherwise remaining parts of the signal would be 'joined'.

Ogden (1997b) warned that wavelet coefficients resulting from 'padding' or interpolation should never be blindly thresholded by a procedure that depends on independent wavelet coefficients with equal variance. Apart from choosing a pre-conditioning method that serves the application of interest, regarding the importance of estimating the correct mean, correct variance or having minimal correlation, the variance of wavelet coefficients should be variance corrected before thresholding Ogden (1997b).

In our analysis of the NMR data from the diet standardisation study, we reduced the number of values per spectrum from 19 930 (after pre-processing) to the largest power of 2, smaller than 19 930, i.e. 2^{14} =16 384. We did this by cutting the data carefully at the ends of each of the two sections, before joining the two sections.

4.3.3 Thresholding methods

Antoniadis et al. (2001) conducted an extensive simulation study to compare a wide variety of wavelet thresholding and wavelet shrinkage estimators, to denoise signals containing additive Gaussian noise. Of the 12 test functions they used, the *Bumps* signal (Figure 4.1) is the most similar to a typical NMR signal in structure (See section 4.3). Among many other simulation results, they evaluate the performance of 34 chosen wavelet denoising procedures using 100 simulations for the *Bumps* signal with a high noise level (root signal-to-noise ratio of 3) at larger sample size ($n = 512$) and using a symmlet-8 wavelet filter (Antoniadis et al. (2001) Figure 8.8). This scenario (from among other scenarios described) for the *Bumps* signal is the one most closely resembling a typical NMR signal, although the sample size would be into the tens of thousands.

A discussion of the wide variety of available thresholding methods is outside the scope of this thesis, but we briefly mention the relevant results from the graphical output in the abovementioned simulation study (Antoniadis et al. (2001) Figure 8.8) Considering the Bayesian denoising methods (both term-by-term and block thresholding and shrinkage methods) 15 of the 16 methods performed well in terms of RMSE, root mean squared bias (RMSB) and maximum deviation (MXDV). Among the non-Bayesian denoising methods, the only level-dependent thresholding methods that performed equally well were the *SureShrink* and hybrid *SureShrink*, both using soft thresholds. Among the non-Bayesian methods, a number of global thresholding methods performed equally well, but only when using a hard threshold: VisuShrink, Minimax, False Discovery Rate and Translation- Invariant. Cross-Validation also performed equally well with both hard and soft thresholds. In terms of CPU time, non-Bayesian methods were superior to Bayesian methods (Antoniadis et al., 2001). The authors concluded that "no wavelet-based denoising procedure uniformly dominates in all aspects".

For our analysis of NMR data from the diet standardisation study, we chose the hybrid *SureShrink* procedure. In the next section we discuss this procedure that adapts to unknown smoothness in more detail.

SureShrink thresholding

Donoho and Johnstone (1995) introduced *SureShrink*, an automatically smoothness adaptive thresholding of empirical wavelet coefficients to suppress noise. They considered the *Bumps* signal (Figure 4.1) as a signal that mimics a simple NMR or other spectrum. In a small simulation study they compared the RMSE for *SureShrink* (using the Daubechies D4, Coiflet-3 and Symmlet-8 wavelet filters), RiskShrink (i.e. Minimax) (using Coiflet-3 and Symmlet-8) and VisuShrink (using Symmlet-8) for sample sizes of 2^7 to 2^{14} with 20 replications for each sample size (except for smaller number of replications for the last two sample sizes) (Donoho and Johnstone (1995), Table 2). For the *Bumps* signal RiskShrink performed better than VisuShrink, but *SureShrink* performed the overall best. The performance of all methods improved with sample size. Results for *SureShrink* were very similar regardless of which wavelet filter was used. Hastie et al. (2009) used *SureShrink* for adaptive wavelet filtering of an NMR signal, which is similar in structure to the NMR data that we consider.

SureShrink assigns a threshold level to each resolution level by minimising the Stein (1981) unbiased risk estimator (SURE) for threshold estimates. Let $\mu = (\mu_1, \ldots, \mu_n)$ and let $x_i \sim N(\mu_i, 1)$ be multivariate normal observations and $x = (x_1, \ldots, x_n)$. Let $\hat{\mu}(x)$ be a specific 'nearly arbitrary, nonlinear biased' estimator for μ. Stein demonstrated that the loss $\| \hat{\mu} - \mu \|^2$ can be estimated unbiasedly. Let $\hat{\mu}(x) = x + g(x)$, where $g = (g_1, \ldots, g_n)$ and $g : \mathbb{R}^n \to \mathbb{R}^n$. Stein (1981) proved that, for $g(x)$ weakly differentiable

$$E_\mu \| \hat{\mu} - \mu \|^2 = n + \mathbb{E}_\mu\{\| g(x) \|^2 + 2\nabla \cdot g(x)\} \tag{4.17}$$

where

$$\nabla \cdot g = \sum_i \frac{\partial}{\partial x_i} g_i.$$

Referring to d and d^* in the wavelet transform (4.11) Donoho and Johnstone (1995) wrote the mean vector d as $\mu = (\mu_1, \ldots, \mu_n)$ and the elements of d^* as independent $x_i \sim N(\mu_i, 1)$ and applied Stein's result to the soft threshold estimator (4.13) as an estimate of μ. We can write $\hat{\mu}_i^{(\lambda)}(x) = \eta_S(x_i, \lambda)$. Then it can be shown that, where $y \wedge z = min(y, z)$,

$$SURE(\lambda; x) = n - 2 \cdot \#\{i : |x_i| \leq \lambda\} + \sum_{i=1}^{n} (|x_i| \wedge \lambda)^2 \tag{4.18}$$

is an unbiased estimate of risk, i.e. $E_\mu SURE(\lambda, x) = E_\mu \parallel \hat{\mu} - \mu \parallel^2$. The SURE risk estimator is then used to select a threshold, by finding $0 \leq \lambda \leq \sqrt{2 \log n}$ that minimises (4.18).

When the true signal wavelet coefficients are extremely sparse, the SURE principle has disadvantages and a hybrid method is implemented in *SureShrink* (Donoho and Johnstone, 1995). It uses the universal threshold (4.14) where the signal is sparse and the SURE threshold otherwise. In more detail, let $s_n^2 = n^{-1} \sum_i (x_i^2 - 1)$ and γ_n be a critical value, typically taken as $log_2^{3/2} n / \sqrt{n}$, then the estimator $\hat{\mu}^*$ is defined (Donoho and Johnstone, 1995) as

$$\begin{aligned} \hat{\mu}_i^{*(\lambda)} &= \sqrt{2 \log n} \qquad s_n^2 \leq \gamma_n \\ &= \eta_S(x_i, \lambda) \qquad s_n^2 > \gamma_n \end{aligned} \tag{4.19}$$

In practice, the noise level σ is not assumed as known, but estimated from the data. In the `threshold.wd` function in the R package WaveThresh (Nason, 2013), the *SureShrink* procedure by default uses the median absolute deviation function to compute the noise level across all levels to be thresholded and adjusts s_n accordingly:

$$\begin{aligned} \hat{\sigma} &= median(|x_i - median(x)|) \\ s_n^2 &= n^{-1} \sum_i \left(\left(\frac{x_i}{\hat{\sigma}} \right)^2 - 1 \right) \end{aligned} \tag{4.20}$$

and in (4.19) x_i is replaced by $x_i / \hat{\sigma}$.

The performance of a thresholding method does not only depend on the choice of thresholding method, but also on the type of wavelet underlying the transform. According to Nason (2008) there appears to be "very little systematic work" done on the choice of wavelet. He expressed his disappointment with this state of affairs, since the type of wavelet can have a "potentially dramatic effect on concrete performance" of thresholding methods. We return to this issue in Section 4.3.5.

4.3.4 Primary resolution

In wavelet shrinkage the *primary resolution*, j_0, is the coarsest level at which thresholding is applied. Some authors refer to the primary resolution level as the level where thresholding begins (i.e. the first coarse levels up to j_0 are not thresholded) or the lowest level of decomposition. (Abramovich and Benjamini (1996); Nason (2008); Nason's discussion of Donoho and Johnstone (1995)).

The primary resolution can take on values from 0 to $log_2(n) - 1$. If the primary resolution level is set very low, over-smoothing typically appears in the sections where the underlying curve is smooth (Hall and Penev, 2001). Also, if n increases but j_0 is kept fixed, over-smoothing will result, specifically when the underlying curve is smooth or piecewise smooth (Hall and Panil's comment on Donoho and Johnstone (1995)).

Zhao et al. (2012) experienced that the choice of j_0 impacted the resulting wavelet estimate and therefore recommends that the choice of j_0 should be carefully considered. In practice, the choice of j_0 has a large influence on the accuracy of the estimate and can determine the success of the thresholding method: the choice of j_0 influences the accuracy of the estimate nearly to the same extent that the choice of threshold does (Nason (2008); Nason's discussion of Donoho and Johnstone (1995)).

The primary resolution level is a smoothing parameter. One option to avoid large bias in peaks and valleys is to choose j_0 empirically (Hall and Penev, 2001). The choice of an appropriate value for j_0 should intuitively depend on the noise level as well as on the smoothness of the estimated function and j_0 should arguably be smaller for smooth functions and larger for oscillating functions (Abramovich and Benjamini, 1996).

The coefficients on the lower coarse levels characterise 'low-frequency' terms. These 'low-frequency' terms often contain vital components of the underlying function and should ideally be kept intact (i.e. not thresholded) (Antoniadis et al., 2001). The `threshold.wd()` function in the R package WaveThresh uses a default value of 3 for the lowest level of decomposition, but this parameter can be changed by the user (Nason, 2013).

Concerning the choice of primary resolution in the literature, Hastie et al. (2009) used $j_0 = 4$ for $n = 2^{10}$ for an NMR signal. Donoho and Johnstone (1995) used $j_0 = 6$ for samples sizes $n = 2^J$, for $J = 7 \dots 14$ for a simulation study that included the Bumps signal.

Antoniadis et al. (2001) chose the primary resolution level to be $j_0(n) = log_2(log(n)) + 1$ in a simulation study using $n = 128, 256, 512, 1024$. For our analysis of the NMR data from the diet standardisation study $n = 16\ 384$ and we chose to select a primary resolution of $j_0 = 11$ (see Section 4.4.1).

4.3.5 Type of wavelet and number of vanishing moments

The most simple mother wavelet is the Haar wavelet, but Daubechies extremal phase wavelets, Daubechies least asymmetric (LA) wavelets and coiflets, among others, are com-

monly used. Daubechies least asymmetric wavelets are also known as symmlets. The symmlets are more symmetric than the Daubechies extremal phase wavelets. A good choice of an appropriate wavelet filter depends on the application (Percival and Walden, 2006).

The number of vanishing moments, v, provides an index for the specific member of a wavelet family, e.g. Daubechies least asymmetric 8. Some references, however, use v to denote the length of the filter coefficients, which is twice the number of vanishing moments (Nason, 2008). The Daubechies least asymmetric wavelet family has vanishing moments starting from 4 (i.e. 8 filter coefficients). See Section 4.1.4 for more on vanishing moments.

Regarding the choice of wavelet (specifically number of vanishing moments) Percival and Walden (2006) recommends a strategy directed by balancing two aspects: on the one hand, avoiding possible artifacts introduced by wavelets of very short widths (i.e. 2, 4 or 6) and, on the other hand, using wavelets with larger widths to better correspond to features in a signal. Wavelets with very short widths may typically introduce artifacts of triangular, block-like or 'shark fin'-like shapes in the results. However, wavelets with large width have drawbacks in terms of more computational effort, boundary conditions negatively affecting more wavelet coefficients and a lower degree of localisation of the discrete wavelet coefficients. A reasonable strategy would be to use the smallest wavelet width that produces a reasonable wavelet analysis (Percival and Walden, 2006). By increasing wavelet width and comparing the wavelet analyses results, the wavelet width should be chosen to correspond with the first wavelet analysis that does not produce artifacts that are due to the wavelet only (Percival and Walden, 2006).

Considering the application of wavelets to spectral data, Antoniadis et al. (2001) used the symmlet-8 and coiflet-3 wavelet filters to model the *Bumps* function (Figure 4.1) and reported similar results for the two wavelet filters. Abramovich and Benjamini (1996) used the Daubechies-4 wavelet transform. Donoho and Johnstone (1994) also used symmlet-8 wavelets. Zhao et al. (2012) used Daubechies least asymmetric wavelets (i.e. symmlets) with eight vanishing moments for a modified version of the Bumps test function.

Specifically for NMR spectra Kim et al. (2008) used the *symmlet-16* wavelet transform. Astle et al. (2012) used *symmlet-6* to model NMR peaks. They motivate the choice of wavelets by the similarity in shape between these wavelets and the Lorentzian peaks obtained from NMR signals. The Lorentzian shape (Cauchy distribution shape) of these peaks are specified by the physics of NMR (Hore and Compton, 1995). Astle et al. (2012) reports that other wavelet bases gave very similar results to the *symmlet-6* transform in terms of reconstructing spectra. Morris et al. (2008) used Daubechies wavelets with four vanishing moments for two different mass spectrometry proteomics (similar in structure to NMR data) examples. They also report that other wavelet bases gave similar results. Hastie et al. (2009) utilised the *symmlet-8* basis on an NMR signal.

We chose to use the Daubechies least asymmetric wavelet with four vanishing moments (Figure 4.2) for our analysis of the NMR data from the diet standardisation study (see Section 4.4.2). Some authors would call it the *symmlet-4* basis but other authors would call it a *symmlet-8* basis.

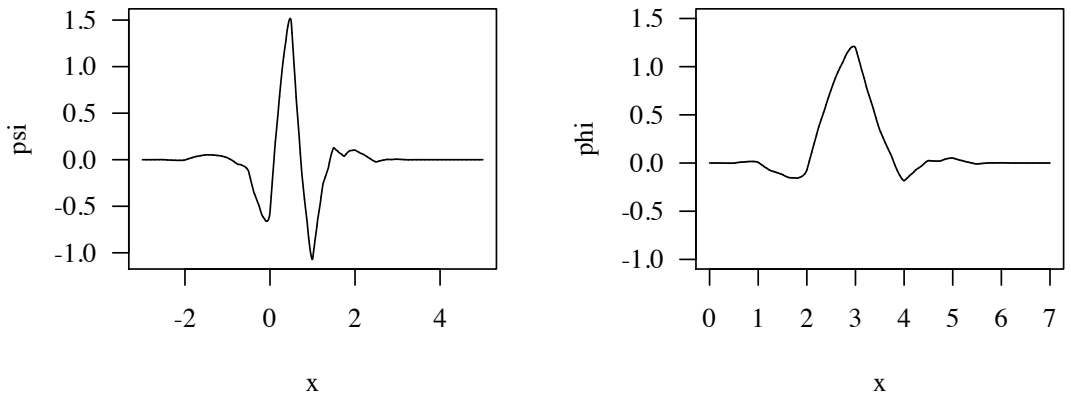

Figure 4.2: The Daubechies least asymmetric wavelet (symmlet) with four vanishing moments and the corresponding scaling function.

4.3.6 Subset selection of wavelet coefficients across multiple signals

In the case of a single signal $y = (y_1, \ldots, y_n)$, a set of non-zero wavelet coefficients, say D^0, where

$$D^0 = \{(j,k)\mathbb{I}(\hat{d}_{j,k} \neq 0) : j,k \in \mathbb{Z}, j_0 \leq j \leq J-1\} \tag{4.21}$$

$$\tag{4.22}$$

is obtained after wavelet decomposition and thresholding. When there are multiple signals $y_i = (y_{i1}, \ldots, y_{in}), i = 1, \ldots, N$ of the same length n, wavelet decomposition and thresholding will result in a set of retained (non-zero) wavelet coefficients, say $\hat{d}^0_{(i)}$ for each signal i. However, it is possible that the sets of retained coefficients differ, i.e.

$$D^0_{(i)} \neq D^0_{(i')} \qquad \text{where } i, i' = 1, \ldots, N \text{ and } i \neq i'.$$

For our purpose of a wavelet-based functional mixed model, we want to model each wavelet coefficient $d_{j,k}$ using a mixed model. The model assumptions require that the random effects as well as the errors should be normally distributed. This assumption will be violated when a wavelet coefficient has zero values $d_{(i),j,k}$ for a number of i's in the N signals in the data set. The challenge is to select a subset of wavelet coefficients from the various sets of non-zero coefficients $D^0_{(1)} \cup D^0_{(2)} \cup \cdots \cup D^0_{(N)}$. We could select only those wavelet coefficients that have non-zero values for all 48 signals

$$D^0_{(1)} \cap D^0_{(2)} \cap \cdots \cap D^0_{(N)}$$

and will thus definitely satisfy the model assumptions. In this way we risk excluding wavelet coefficients where a small number of signals have zero values for the specific coefficient. At the other extreme we could select all wavelet coefficients with at least one non-zero value across the 48 signals

$$D^0_{(1)} \cup D^0_{(2)} \cup \cdots \cup D^0_{(N)}.$$

In this way we will not exclude any wavelet coefficients from the subset, but for model assumptions may be violated for some wavelet coefficient models.

For our analysis of the NMR data from the diet standardisation study we chose to select only those wavelet coefficients that have non-zero values for all 48 signals (see Section 4.4.3).

4.4 Parameter choices: NMR diet standardisation data

As mentioned in Section 4.3 we made the following choices for our analysis of the NMR data from the diet standardisation study:

- Periodic boundary handling;
- Reduction of the number of values per spectrum to the largest power of 2, smaller than n;
- Hybrid *SureShrink* thresholding by level and using a soft threshold;
- Daubechies least asymmetric wavelet with four vanishing moments; and
- Primary resolution 11 combined with subset selection of wavelet coefficients that are non-zero (after thresholding) for all 48 spectra.

We discuss the motivation for our choices regarding type of wavelet, primary resolution and subset selection in the following sections.

4.4.1 Primary resolution

For the spectra in the NMR diet standardisation study the primary resolution j_0 can take on values from 0 to 13 ($log_2(16384) - 1$) (See Section 4.3.4). Here we carefully and empirically consider the choice of j_0 and its impact on the resulting wavelet estimates.

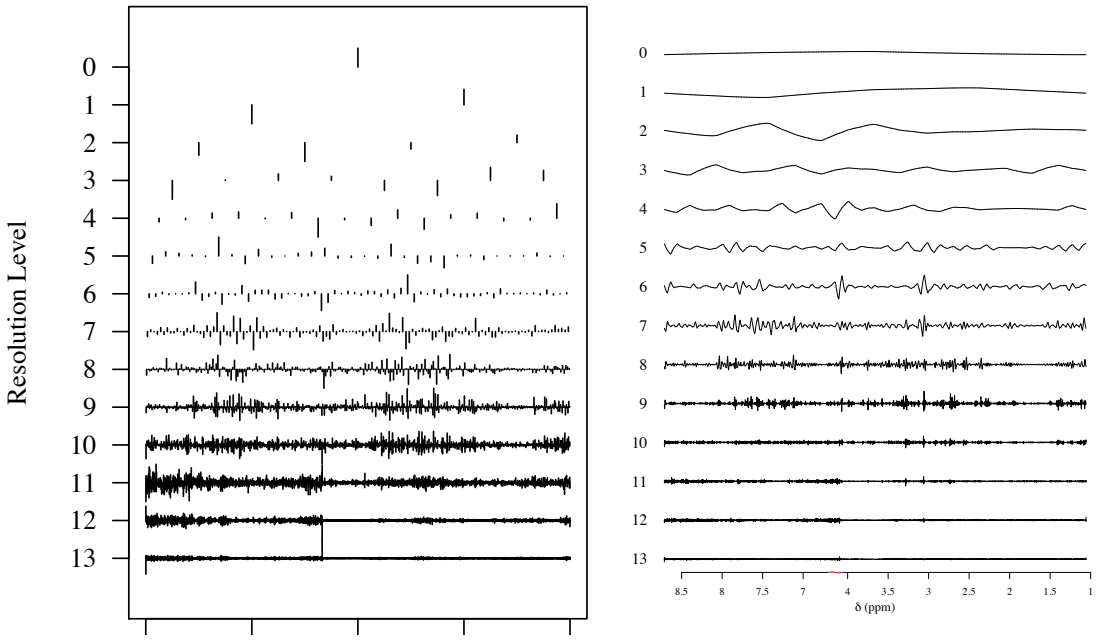

Figure 4.3: Wavelet decomposition coefficients for a single spectrum from the diet standardisation study (left) using Daubechies Least Asymmetric wavelets with four vanishing moments. Coefficients are scaled for each resolution level separately and depends on the largest absolute value of coefficients in that level. For the wavelet coefficients (left), the inverse wavelet transform for each level separately is shown on the right.

Figure 4.3 shows the wavelet decomposition for one of the 48 original spectra. Each wavelet coefficient is represented by a vertical bar and the (positive or negative) height of the bar indicates the size of the coefficient. The corresponding inverse wavelet transform on the right gives an indication of the component in the signal represented by the specific level of wavelet coefficients. The first number of levels contain very broad 'low-frequency' terms that span the entire width of the spectrum. The higher the level (the further down in the figure), the more localised the effects are, representing 'higher-frequency' terms. The highest level, level 13, can contain mostly noise.

Figure 4.4 shows the same wavelet decomposition as in Figure 4.3 (left), with the difference that SureShrink thresholding was applied with primary resolution 3 and 11 respectively. Most strikingly, all coefficients on level 13 ($n = 8\ 192$) were shrunk to 0 (and are not displayed) in both thresholding procedures. On level 12 only 2 of the 4096 coefficients

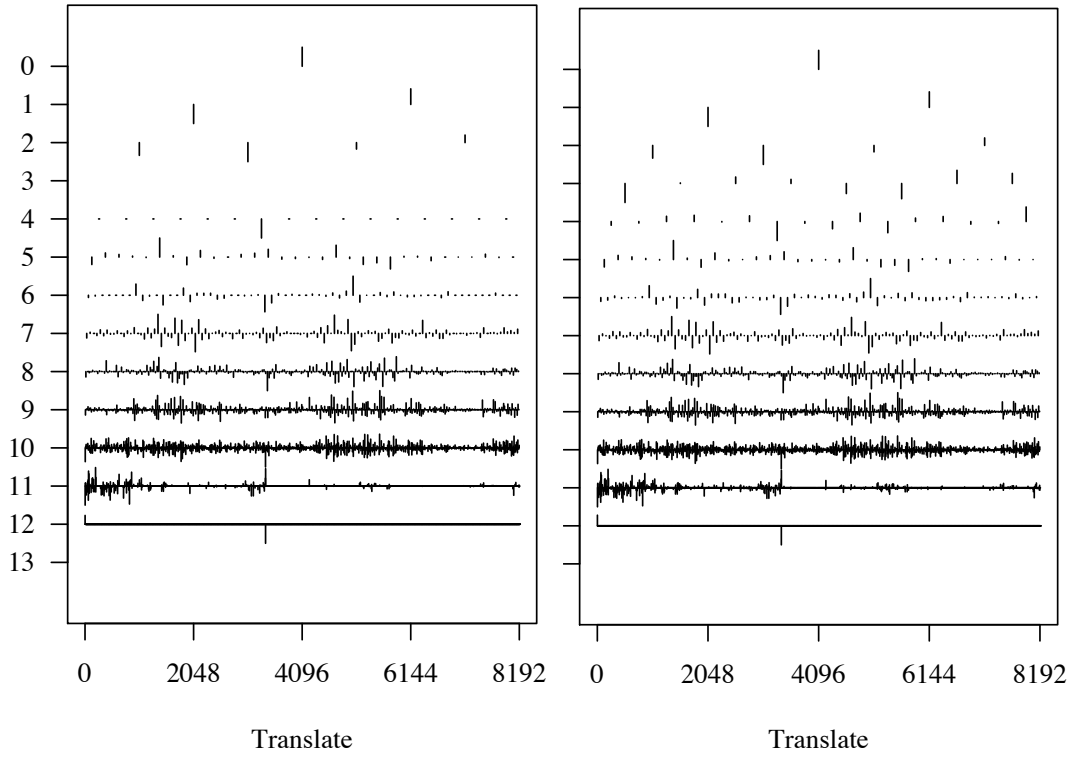

Figure 4.4: Wavelet decomposition coefficients for a single spectrum from the diet standardisation study: *SureShrink* with primary resolution 3 (left) and 11 (right). Compare with Figure 4.3

were not shrunk to zero in both graphs. One of these coefficients is related to a boundary effect and the other to the location where two parts of the spectrum were joined. On level 11, for both cases, a substantial number of the 2048 coefficients were shrunk to zero, but a number of coefficients are retained. From level 10 and lower (upwards in the figure), the coefficients in the graph on the right remain unchanged (compared to Figure 4.3 (left)) since thresholding only started from level 11. For the graph on the left, it is difficult to see changes at levels 10 to 7 and level 5. The shrinkage in coefficients may just not be visible at this scale. In the graph of the left, for levels 6 and 4, smaller coefficients were clearly shrunk to zero. At level 3 no non-zero coefficients remained after thresholding.

Considering only a small section of the same spectrum, Figure 4.5 demonstrates the smoothing effect (top row) of using SureShrink with primary resolution of 11. The corresponding wavelet coefficients (bottom row) indicate that, in this specific section, all coefficients on levels 12 and 13 were regarded as noise and shrunk to zero. On level 11 a few large coefficients were retained. We only show coefficients from level 5, since lower levels (0 to 4) contain very wide 'low-frequency' terms.

Figures 4.6 demonstrates the smoothing effect of different primary resolutions for a single

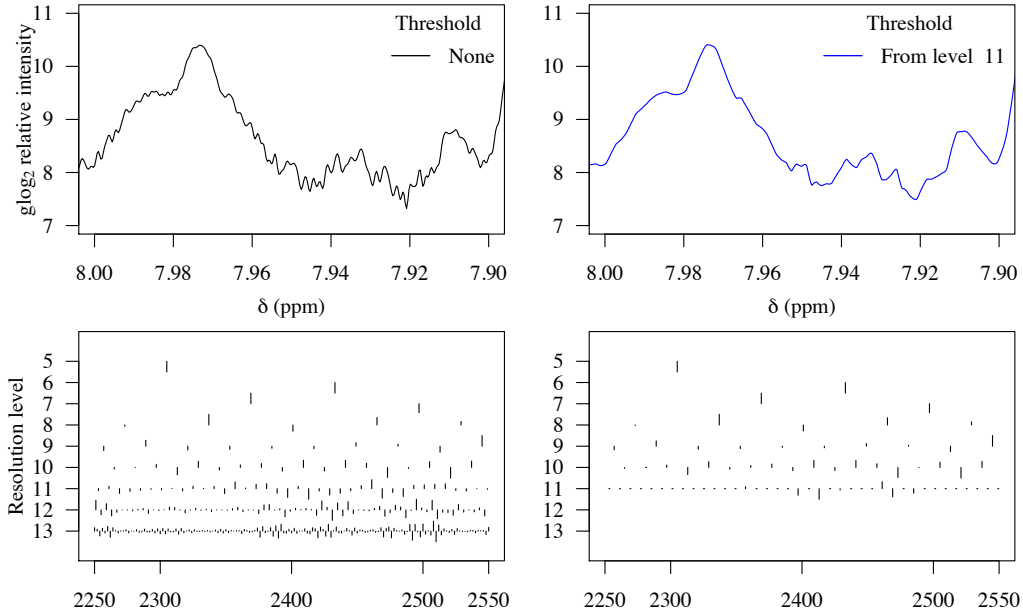

Figure 4.5: An enlarged section (8.0 to 7.9 ppm) of a single spectrum from the diet standardisation study. The effect of *SureShrink* with primary resolution level 11 (top right, blue) compared to the inverse wavelet transform of the data (top left). Corresponding wavelet coefficients (from level 5) are shown in the bottom row.

spectrum from the diet standardisation study: thresholding from level 0 (to 13) over-smoothes some of the 'low-frequency terms' and notably overestimates some sections (8.0 to 7.7 ppm) and underestimates other sections (7.6 to 7.5 ppm) in the spectrum; thresholding from level 3 still over-smoothes some of the 'low-frequency terms' and thus overestimates and underestimates some sections of the spectrum (most notably 7.8 to 7.7 ppm and 8.0 to 7.95 ppm), but the errors are not necessarily in the same direction as for thresholding from level 0 (note 8.0 to 7.9 ppm); thresholding from level 6 seem to better preserve most of the 'low-frequency terms' and the broad areas of overestimation and underestimation disappear, although some over-smoothing remains (smaller areas of overestimation or underestimation around 7.8 and 7.99 ppm).

Figures 4.7 demonstrates the effect of thresholding from even higher levels (compare with 8.0 to 7.9 ppm in Figure 4.6): there is no over-smoothing of 'low-frequency terms' and all three estimates follow the shape of the spectrum; thresholding from level 7 possibly still results in over-smoothing some smaller areas and underestimates the peak at 7.97 ppm and overestimates the valley around 7.92 ppm; there are very little visible differences between the estimates that threshold from levels 9 and 11.

The challenge is to find a good primary resolution. Starting thresholding too low (e.g. level 3 or 6) will over-smooth 'low-frequency' terms and starting too high (e.g. level 13) will not remove enough noise. In Figure 4.8 the contributions from different levels of the

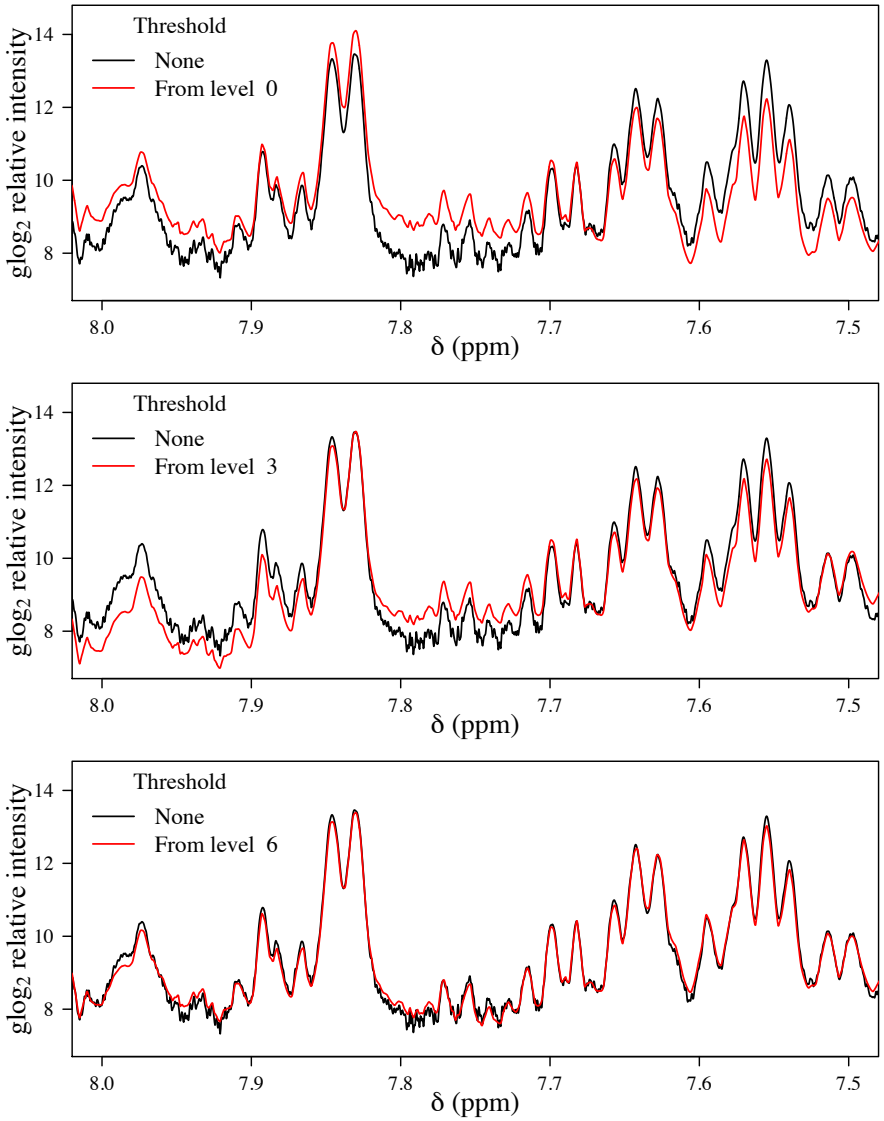

Figure 4.6: A section (8.0 to 7.5 ppm) of a single spectrum from the diet standardisation study. The effect of *SureShrink* thresholding for primary resolution levels: 0 (top), 3 (middle) and 6 (bottom). The curve with no thresholding (black) is the inverse wavelet transform of the data.

wavelet decomposition to the spectrum is illustrated. Roughly speaking: levels 0 to 4 seem to constitute broad 'low-frequency' effects that reach over the entire range of the spectrum (and cannot be seen in this small section of the spectrum), levels 5 to 7 seem to construct the basic shape of large peaks, levels 8 to 11 seem to construct smaller peaks and add the detail to construct larger peaks, and levels 12 and 13 seem to contain mostly noise. Ideally, we want to select the primary resolution such that levels containing peaks and noise are thresholded, while levels containing the broad underlying structure of the spectrum are

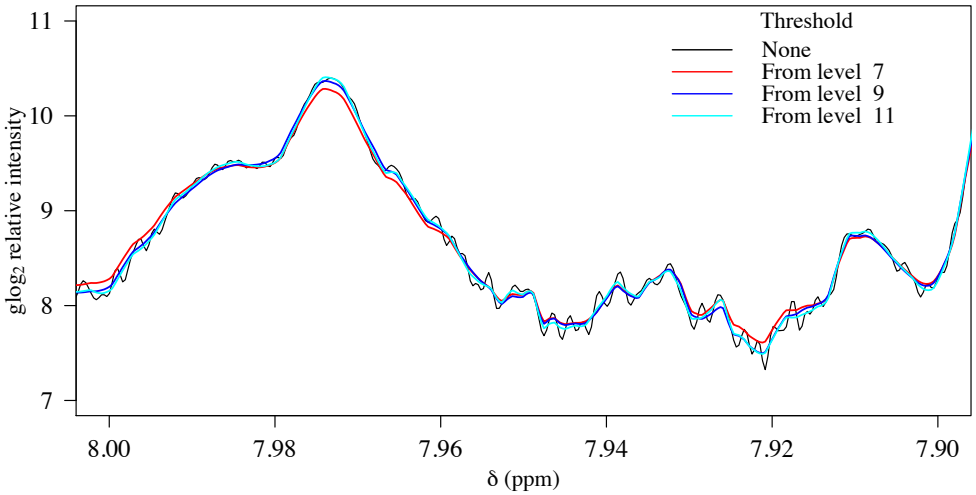

Figure 4.7: An enlarged section (from the region in Figure 4.6) of a single spectrum from the diet standardisation study. The effect of *SureShrink* thresholding for primary resolution levels: 7 (red), 9 (blue) and 11 (light blue). The curve with no thresholding (black) is the inverse wavelet transform of the data.

left unaltered.

For the diet standardisation with $n = 16\ 384$ for each of 48 spectra. By using $j_0 = log_2(logN)) + 1$ (Antoniadis et al., 2001) we obtain a primary resolution level of $j_0 \approx 4$. Visually (Figure 4.6) it seems that we still remove low-frequency terms, even with a primary resolution of 6. These visual results, however, are only shown for a small section of one of the 48 spectra in the NMR data set. To investigate the effect of different primary resolutions on our entire data set, we calculated the MISE (4.16) for each spectrum. We present a box plot per primary resolution level (Figure 4.9). For an individual spectrum it is to be expected that the MISE will decrease as the primary resolution increase. Note the decreasing, yet large, range in MISE per primary resolution level, up to approximately level 6. The median and range of the MISE across the 48 spectra seem to stabilise at level 8. Here we used Daubechies Least Assymetric wavelets with four vanishing moments. In Section 4.3.5 we discuss the choice of the type of wavelet.

Knowledge of the shape of NMR peaks may further guide our choice of primary resolution level (Section 5.1). NMR peaks have a Lorentzian shape and the peak width is related to the peak height. Consequently we consider the highest peak in the data as having the largest possible peak width present in the data. This peak occurs at 3.048 ppm in all spectra. Empirically, the width of this peak points to a primary resolution of at least 7 (Figure 4.10. Obviously it will also depend on the location of the peak with respect to the location of wavelets on level 7, whether a wavelet from this level could represent this peak. If not, wavelets from higher levels would automatically be used to model this peak. If we chose level 7 as the primary resolution, we should arguably include all peaks (and noise) in the thresholding and exclude most 'low-frequency terms' from thresholding.

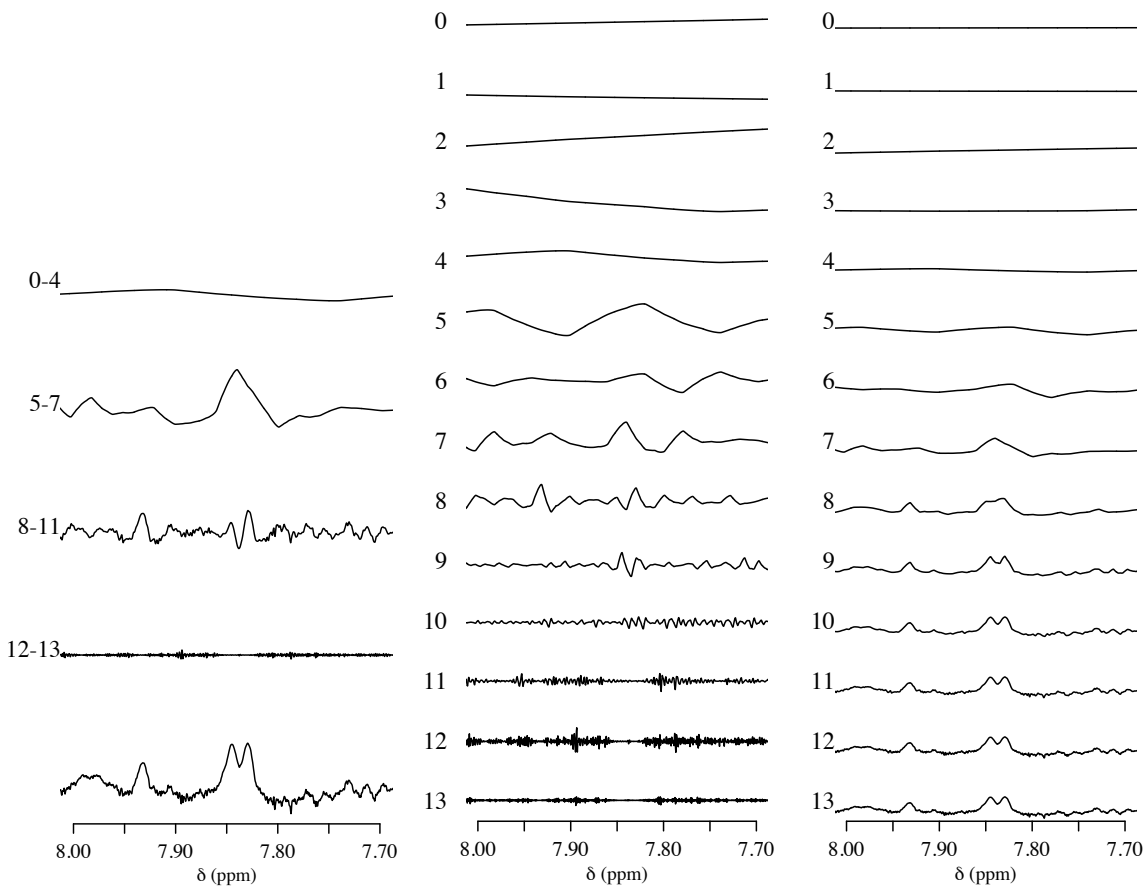

Figure 4.8: An enlarged section (8.0 to 7.9 ppm) of a single spectrum from the diet standardisation study. The inverse wavelet transform: by level (middle), combined for levels 0 – 4, 5 – 7, 8 – 11 and 12 –13 (left) and cumulative by level (right).

We further consider the type of wavelet, subject to the choice of primary resolution level, in Section 4.4.3. In our application, the subset selection of thresholded wavelet coefficients across spectra, for the purpose of modelling, further complicates the choice of primary resolution. We discuss this issue in Section 4.4.3.

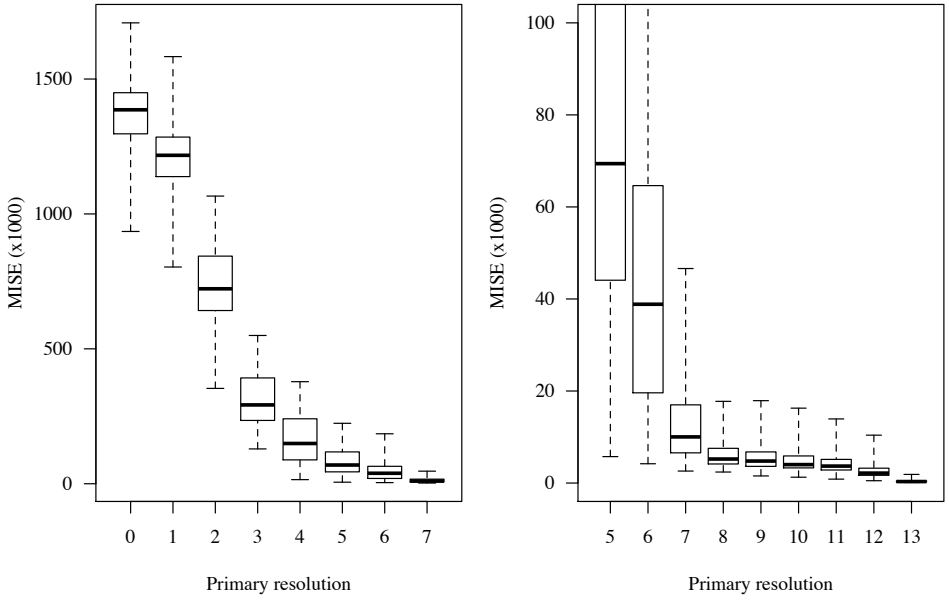

Figure 4.9: Box plots of mean integrated square error (MISE) (×1000) for 48 signals from the diet standardisation study, by primary resolution 0 to 7 (left) and 5 to 13 (right) (5 to 7 are include in both graphs for comparison purposes). We used *SureShrink* and Daubechies Least Asymmetric wavelets with four vanishing moments.

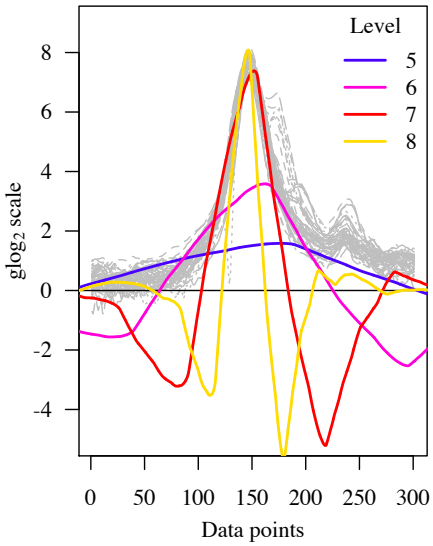

Figure 4.10: The largest peak in relative height and width from the diet standardisation study. Possible 'fitting' Daubechies Least Asymmetric wavelets with four vanishing moments and dilation corresponding to levels 5 to 8.

4.4.2 Type of wavelet and number of vanishing moments

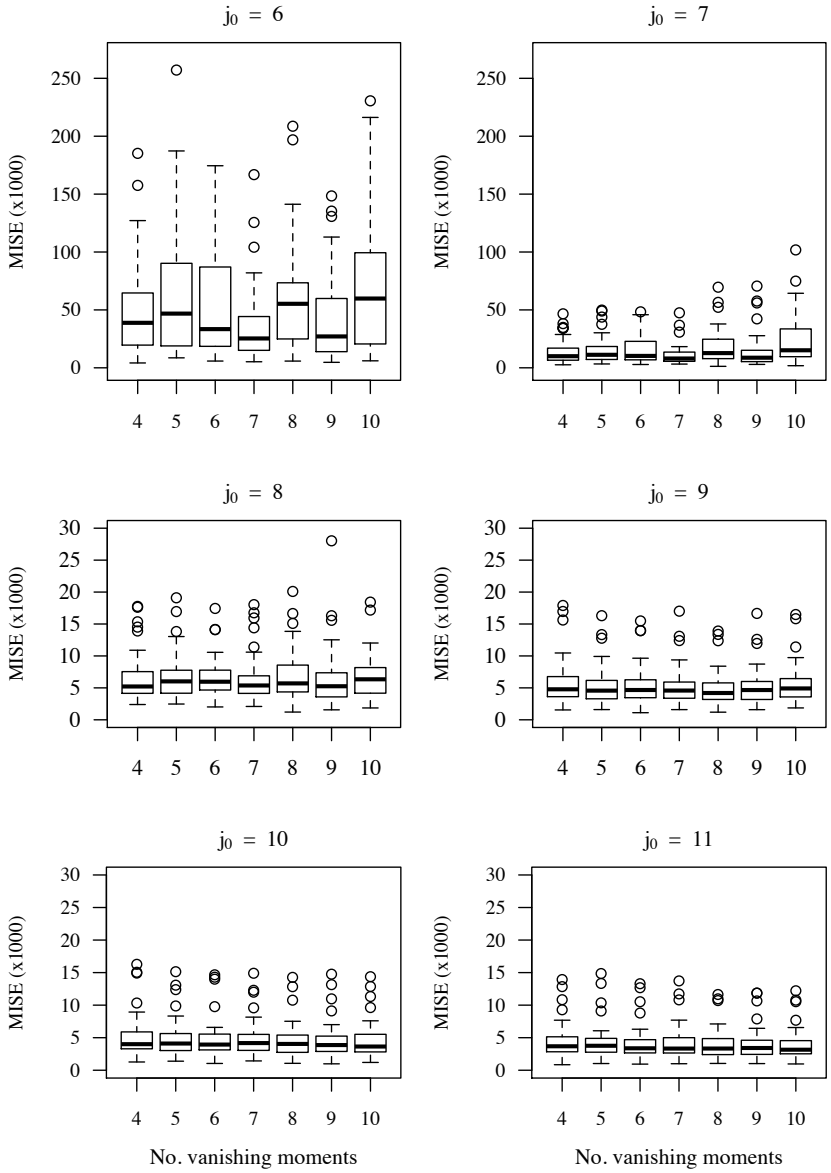

Figure 4.11: Box plots of Mean Integrated Square Error (MISE) (×1000) by number of vanishing moments (4 to 10) for primary resolution level (j_0, 6 to 11) using Daubechies Least Asymmetric wavelets.

We chose to use Daubechies least asymmetric wavelets for our analysis of the NMR data from the diet standardisation study. It is not obvious how many vanishing moments we should choose for the wavelets. We calculated the MISE for each of the 48 spectra and produced a box plot per combination of the number of vanishing moments (ranging from

4 to 10) and the primary resolution (ranging from 6 to 11). Apart from the median, we are interested in the upper extremes of each distribution, i.e. the maximum error we would encounter for any of the 48 spectra.

At primary resolutions 8 to 11 the MISE is very small (< 0.03) for any number of vanishing moments. At a primary resolution of $j_0 = 7$ the MISE is relatively small (< 0.1) for any number of vanishing moments except 10, where there is one outlier. At a primary resolution of $j_0 = 6$ a choice of seven or nine vanishing moments seem to produce the smallest MISEs (< 0.175).

In our data, the number of vanishing moments is not a crucial choice for primary resolution of 8 to 11. At a primary resolution of $j_0 = 7$ a choice of four, five or seven vanishing moments appear to produce somewhat smaller MISEs than other numbers of vanishing moments. At level $j_0 = 6$ the choice is more critical with seven or nine vanishing moments as the best options for our data.

4.4.3 Subset selection of wavelet coefficients across multiple signals

For an individual NMR spectrum from the diet standardisation study, the wavelet decomposition results in 16 383 wavelet coefficients on 14 levels (levels 0 to 13). The *SureShrink* thresholding procedure reduces the number of non-zero wavelet coefficients by shrinking many coefficients to zero. The number of coefficients remaining depends on the primary resolution. Furthermore this number may vary from spectrum to spectrum. For example, for SureShrink with $j_0 = 5$ a specific spectrum has 2 061 non-zero coefficients, but across the 48 spectra this number ranges from 1 500 to 2 576 with a median of 1 910. A number of these may be the same wavelet coefficients across all spectra, but some may be unique to one spectrum or present in a few spectra. For $j_0 = 5$ there are 286 wavelet coefficients that are present (non-zero) for all 48 spectra after thresholding. All together there are 3 944 different non-zero wavelet coefficients across the 48 spectra. Since $j_0 = 0$ there are 31 coefficients belonging to levels 0 to 4 that will be included among both the 286 and the 3 944 wavelet coefficients mentioned above. These numbers and corresponding numbers for $j_0 = 6 \ldots 11$ are shown in Table 4.1. Note that the number of non-zero wavelet coefficients present for at least one of the 48 spectra (second column) decreases from 4 455 at $j_0 = 11$ to 3944 at $j_0 = 5$. However, there is a drastic decrease in the number of wavelet coefficients present for all 48 spectra (column 6), from 2 049 at $j_0 = 11$ to only 286 at $j_0 = 5$.

The question of subset selection across spectra is "what criteria should be specified for the least number (say x) of spectra that should contain a specific wavelet coefficients for this coefficient to be included in mixed modelling of the 48 spectra?". Specifying $x = 1$ may result in violating model assumptions, whereas specifying $x = 48$ may be too strict in excluding any wavelet coefficient not present in all spectra. In Figure 4.12 subset selection amounts to choosing a cut-off point (from 0 to 48) along the x-axis and including all wavelet coefficients present (non-zero) for more spectra than the value of the cut-off point, to be included in modelling. Clearly the distribution of the number of wavelet coefficients

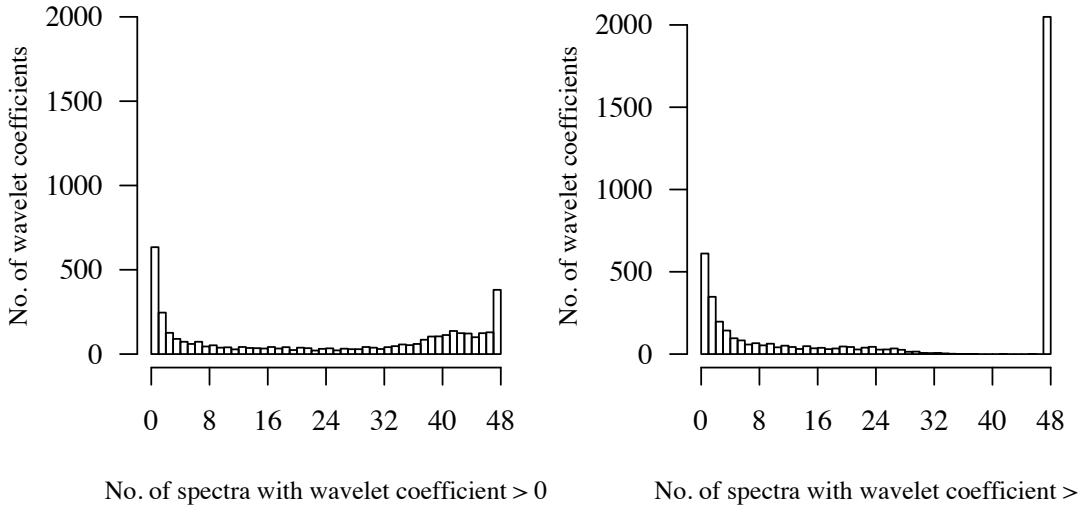

Figure 4.12: Histograms indicating the number of spectra (of 48) with wavelet coefficients larger than zero for a primary resolution of $j_0 = 7$ (left) and $j_0 = 11$ (right). Note that number of spectra with wavelet coefficients equal to zero is not shown (12 429 and 11890 respectively).

present for a certain number of spectra depends on the primary resolution (compare $j_0 = 7$ (left) with $j_0 = 11$ (right)). The most obvious difference in these two graphs is that the distribution around the higher end shifts dramatically from values just below 48 to 48 with the increase in j_0 from 7 to 11.

In Tables 4.2 and 4.3 we explore how the non-zero wavelet coefficients are distributed across resolution levels for respectively $x = 1$ and $x = 48$. Note that the last columns (Total) in Table 4.2 and Table 4.3 correspond to the second and the sixth columns in Table 4.1 respectively. Interestingly, in Table 4.2 all the wavelet coefficients in a resolution level are retained up to level 10, *regardless of the primary resolution* (compare with the maximum number of possible wavelet coefficients per level in the bottom row). Thus, only from level 11 there appear wavelet coefficients that are not present after thresholding for any spectra. In Table 4.3 relatively small numbers of wavelet coefficients within each resolution level are common among all spectra, e.g. 2 of a possible 64 at level 6, 23 of a possible 128 at level 7, etc. For levels 11 to 13 at most one wavelet coefficient is retained regardless of j_0. The increase in the total number of wavelet coefficients retained (last column) as j_0 increases, is mainly due to the number of wavelets coefficients 'forced' to be retained (second column) which increase from 31 at $j_0 = 5$ to 2 047 at $j_0 = 11$.

The combined effect of primary resolution and subset selection is demonstrated for a section of a single spectrum in Figure 4.13. Subset selection of wavelet coefficients present (non-zero) across all 48 spectra, combined with $j_0 = 7$ markedly 'excludes' a large number of peaks in estimation of the spectrum, even in this small section. Even at $j_0 = 9$ a number of peaks are still 'excluded' by subset selection. At $j_0 = 11$ there is no visual evidence of

Table 4.1: Number of wavelet coefficients selected, by coarsest level of thresholding (j_0) and various criteria for wavelet coefficient selection across spectra

j_0	At least x wavelet coefficients present across 48 spectra					No. of wavelet coefficients for $j = 0, 1, \ldots, j_0 - 1$
	x					
	1	16	32	44	48	
5	3944	2290	1761	778	286	31
6	3930	2310	1779	807	318	63
7	3950	2333	1817	864	382	127
8	3950	2342	1852	934	485	255
9	4031	2355	1869	1037	682	511
10	4142	2428	1893	1260	1080	1023
11	4455	2556	2072	2050	2049	2047

Table 4.2: Number of wavelet coefficients selected, by coarsest level of thresholding (j_0), for the criteria that a wavelet coefficient should be present (not equal to 0) for at least one of 48 individual spectra

j_0	# wavelet coefficients	No. of wavelet coefficients retained at thresholded levels									Total
	$j < j_0$	5	6	7	8	9	10	11	12	13	
5	31	32	64	128	256	512	1024	1554	342	1	3944
6	63	-	64	128	256	512	1024	1558	334	1	3930
7	127	-	-	128	256	512	1024	1572	330	1	3950
8	255	-	-	-	256	512	1024	1575	327	1	3950
9	511	-	-	-	-	512	1024	1601	382	1	4031
10	1023	-	-	-	-	-	1024	1656	438	1	4142
11	2047	-	-	-	-	-	-	1813	594	1	4455
Max. no. possible		32	64	128	256	512	1024	2048	4096	8192	

peaks being 'excluded' by subset selection in this small section of the spectrum. We suspect the possible reasons for this dramatic effect is possibly misalignment.

Finally, Figure 4.14 sheds light on subset selection. If we do not 'force' the inclusion of any wavelet coefficients in individual spectra by setting $j_0 = 0$, all 14 levels of each spectrum are thresholded. By counting the number of spectra (of 48) that have non-zero coefficients for a specific wavelet we obtain the values in Figure 4.14, displayed by resolution level. Wavelet coefficients where the count is zero (below the red dotted line at the bottom) are irrelevant to our model. Subset selection amounts to moving the criteria to include only wavelet coefficients present for all 48 spectra (red dotted line at the top) down to

Table 4.3: Number of wavelet coefficients selected, by coarsest level of thresholding (j_0), for the criteria that a wavelet coefficient should be present (not equal to 0) for all 48 individual spectra

j_0	# wavelet coefficients	No. of wavelet coefficients retained at thresholded levels									Total
	$j < j_0$	5	6	7	8	9	10	11	12	13	
5	31	0	2	23	63	114	52	1	0	0	286
6	63	-	2	23	63	114	52	1	0	0	318
7	127	-	-	23	63	116	52	1	1	0	382
8	255	-	-	-	62	115	52	1	0	0	485
9	511	-	-	-	-	117	53	1	0	0	682
10	1023	-	-	-	-	-	56	1	1	0	1080
11	2047	-	-	-	-	-	-	1	1	0	2049
Max. no. possible		32	64	128	256	512	1024	2048	4096	8192	

the chosen value for subset selection. By increasing the value of j_0, wavelet coefficients for levels $j < j_0$ (levels towards the righthand side of the graph) will be forced' to 'jump' to a value of 48 at the top of the graph and will be included for modelling regardless of the value chosen for subset selection (top red dotted line). Ideally j_0 should be chosen large enough to preserve 'low-frequency' terms, but to threshold noise where few spectra have the wavelet coefficients present. At the same time, the subset selection criteria should be chosen to preserve wavelet coefficients associated with features in the data while ensuring model assumptions are met by excluding wavelet coefficients with a low count of spectra.

To prioritise model assumptions we chose a subset selection criteria of 48 spectra. This required a high value for j_0 and we chose 11. Variations on these choices is of interest in future research.

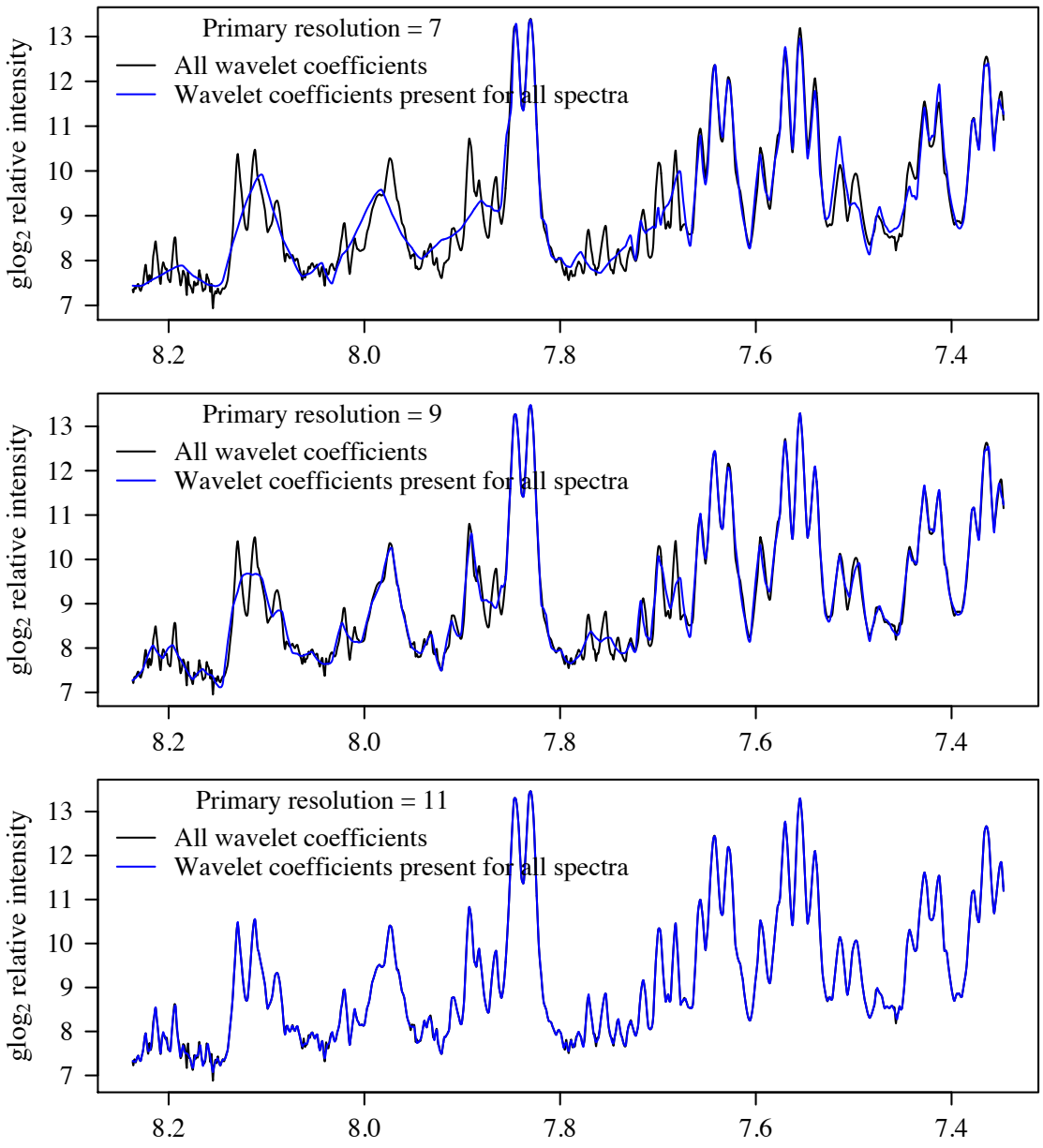

Figure 4.13: A section (8.25 to 7.35 ppm) of a single spectrum from the diet standardisation study. The effect of subset selection where all 48 wavelet coefficients are present, together with *SureShrink* thresholding for primary resolution levels 7 (top), 9 (middle) and 11 (bottom). The curve with no thresholding (black) is the inverse wavelet transform of the data.

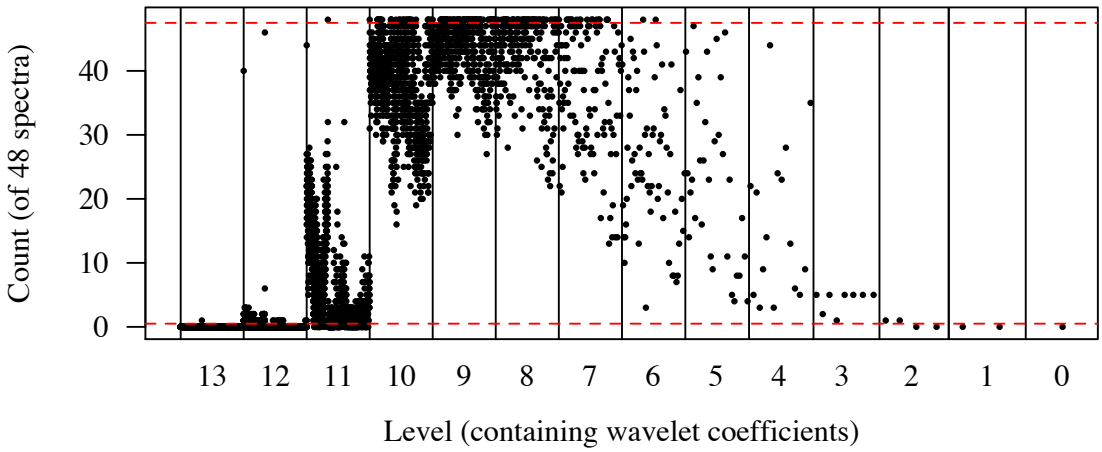

Figure 4.14: Number of spectra (of 48) for which wavelet coefficients are retained after SureShrink thresholding on all levels, by individual wavelet coefficients within level (13 to 0). Red dotted lines indicate number of spectra where either no wavelet coefficients 'survived' thresholding (below bottom line) or all 48 wavelet coefficients 'survived' thresholding (above top line).

5 NMR data and pre-processing

Nuclear magnetic resonance (NMR) spectra are complex. To obtain meaningful results from analysing NMR data, the ideal is to have spectral measurements that are without noise, errors and missing data. Furthermore, peaks from the same metabolite should line up across spectra; peaks should be comparable in intensity across spectra; variation in the intensity of a specific peak (across spectra) should be comparable across peaks; and, the intensity of a peak should reflect the abundance of the associated metabolite. In practice, NMR data rarely, if ever, comply with all these requirements and a substantial amount of pre-processing is required to prepare the data for analysis.

This chapter is mainly concerned with the pre-processing of NMR data, but first we discuss the structure and technical details of an NMR spectrum.

5.1 Technical details on NMR data

NMR spectroscopy of urine produces a complex 'fingerprint' with thousands of resolved peaks and typically 50 or more identifiable compounds (See Figure 2.1). Bouatra et al. (2013) recently identified 209 unique compounds in NMR urine spectra of 22 healthy individuals. Each compound consists of one or more peaks. The peaks from different compounds often overlap.

NMR data are measured in the time domain, where magnetic resonance signals are expressed as exponentially decaying sinusoidal waves. After Fourier transform, the resonances are expressed as Lorentzian peaks in the frequency domain (Figure 5.1).

The basic structure of an NMR spectrum is a vector $(y_1, \ldots, y_n)$ of resonance intensities measured at regularly spaced points on the chemical shift axis. Typically n is in the order of tens of thousands. By convention the chemical shift axis is labelled δ and decreases from left to right, typically with a range from 10 to 0 ppm. The unit for chemical shift is parts per million (ppm) and it is inversely related to frequency. The vector $\mathbf{y}$ is, in principle, strictly positive, but since it is observed with noise $\mathbf{y}$ can also take on values below zero.

An NMR spectrum consists of a large number of convolved peaks. The Lorentzian shape of each peak is better known in statistics as a Cauchy distribution with scale parameter

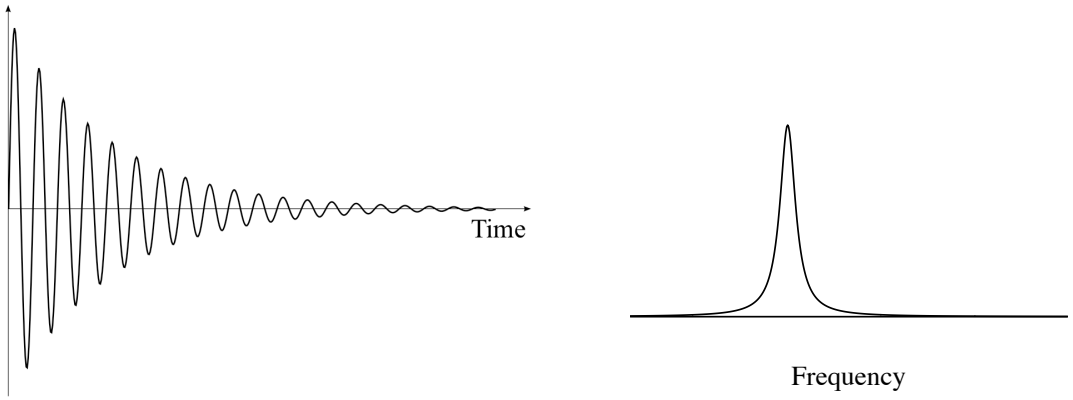

Figure 5.1: An exponentially decaying sinusoidal wave is Fourier transformed to a Lorentzian peak shape

$\gamma/2$. In NMR spectroscopy γ is called the *line width* or *full width half maximum* (FWHM) and indicates the width of the peak at half the maximum height.

^{1}H NMR is also called proton NMR and measures the resonance of hydrogen nuclei. The proton (hydrogen nuclei) resonates at specific frequencies. The resonant frequencies of a proton are determined by its bonding and the chemical structure of the molecule in which it is contained. The resonating frequencies, in turn, determine the specific chemical shift values of peaks in the ^{1}H NMR spectrum.

A metabolite (small molecule compound) displays an individual signature in an ^{1}H NMR spectrum: a convolution of Lorentzian peaks at specific chemical shift positions. A metabolite's peaks can appear in multiplets, defined by a specific number of peaks, relative peak heights and separations between these peaks.

In an NMR spectrum, a metabolite can be identified by its characteristic multiplet(s). The concentration of a metabolite in a mixture is proportional to the intensity of the peaks belonging to the metabolite. The area under a peak indicates the *relative* concentration of the associated metabolite.

5.2 Pre-processing

Goodacre et al. (2007) differentiate between pre-processing and pre-treatment:
Pre-processing is a "Generic term for methods to go from raw instrumental data to clean data for data analysis".
Pre-treatment is "Transforming the clean data to make them ready for data analysis".

The clean data are also described as the initial data matrix, X, with each row (i) containing one sample and each column (j) containing one feature (variable or chemical shift region).

Pre-processing includes (Goodacre et al., 2007):

- **Deconvolution** Resolving overlapping peaks in a spectrum;
- **Peak-picking** Selection of peaks to produce a table with ppm and corresponding intensities;
- **Target analysis** Peaks at specific chemical shift values are integrated and used in a peak table;
- **Alignment** Synchronisation of spectra (globally or in local regions) such that each metabolite signal has the same chemical shift in each sample;
- **Apodization function and weighting factors** Function and parameters used to multiply free induction decays (FIDs) before Fourier transformation to NMR spectra;
- **Phasing** To phase-correct peaks in Fourier transformed NMR spectra, manually or automatically by NMR software;
- **Baseline Correction** To address baseline tilts and drifts in Fourier transformed NMR spectra, automatically or semi-automatically; and
- **Bucketing** (or Binning) To define chemical shift bin sizes and integrate the bin intensities.

Pre-treatment includes (Goodacre et al., 2007):

- **Normalisation** Performed within or across rows (samples) to make the row profiles comparable in size;
- **Centring** Performed across the rows (samples) to translate the centre of gravity of the dataset;
- **Scaling** Performed within a column (variable) to make the column profiles more comparable; and
- **Transformation** Performed to linearise or otherwise change the scale of the data (total matrix), e.g., to remove heteroscedastic noise.

In the literature, the term ‘pre-processing’ (or sometimes ‘pre-treatment’) is often used to include what Goodacre et al. (2007) describes as ‘pre-processing’, as well as Alignment, Baseline-correction and/or Binning (Liland, 2011; Van den Berg et al., 2006; Bloemberg et al., 2013). In applied statistics and chemometrics, ‘pre-processing’ is generally used to describe all adjustments to the data, up to the start of analysis (Liland, 2011). We will refer to the spectra, obtained from the NMR instrument and already pre-processed to a certain degree (e.g. apodization and phasing), typically with instrument specific software, as ‘raw data’. This part of pre-processing falls outside the scope of this overview. We will refer to what Goodacre et al. (2007) call ’pre-treatment’, as well as alignment and baseline correction methods, using the term ‘pre-processing’.

Pre-processing of chemometric data requires substantial background knowledge, relating to the measurement platform, the biofluid being analysed, experimental conditions and biochemistry. For this reason, we provide an introduction and overview of the steps in the pre-processing of NMR data, from a chemometric point of view. To the statistician, this provides an understanding of the relevant issues in spectral data. Spectral pre-processing is a wide and expanding topic. We do not claim that this is a complete overview.

It should be noted that, although we focus on the pre-processing and analysis of NMR data, most of the aspects are relevant to other spectral and chromotographic data, e.g. mass spectrometry and infrared. We restrict the overview to what is called 'one-dimensional' (1-D) data in chemometrics, which is technically two-dimensional data, i.e. chemical shift (δ in ppm units) on the horizontal axis vs. relative intensity values on the vertical axis.

The typical steps in chemometric data pre-processing, in no specific order, are:

- Baseline correction;
- Alignment;
- Normalisation;
- Scaling and transformation; and
- Removing of spectral regions.

There is no consensus in the literature regarding the order of pre-processing steps. Engel et al. (2013) recently pointed out that the choice of pre-processing methods can strongly influence the results of subsequent data analysis. The problems related to certain pre-processing steps are more serious than others: alignment emerges as a crucial step, since peaks from the same chemical compound/metabolite should line up across spectra Wehrens (2011). Misalignment can result in inaccurate or even wrong results. Independent from chemometrics, the topic of alignment also emerged as a critical step and current topic in Functional Data Analysis (See Section 3.2).

Pre-processing of chemometric data is time intensive, often an iterative process, requires visual inspection of results, and is dependent on a number of subjective choices. In many ways, pre-processing is more an art than a science. Nevertheless, it is a critical process and no analysis can fix bad pre-processing. (Wehrens, 2011)

In the following sections we provide an overview of some of the steps in chemometric pre-processing. This review is not exhaustive, but is intended to give statisticians an idea of the broad range of methods used for pre-processing data in chemometrics.

5.3 Baseline correction

The baseline of a spectrum is supposed to be a horizontal line located at zero, i.e. no signal. This is rarely the case, with a baseline typically displaying some linear trend, curved shape

(especially towards the ends of the spectra) or other nonlinear effects. There is usually no pattern in how the baseline varies from spectrum to spectrum. (Liland et al., 2010)

Baseline correction (identification and removal) is often an automatic procedure performed on the NMR instrument. In addition, software can be used to remove any baseline structure remaining in the spectra. A baseline is undesired, since it influences the intensity of metabolites and thus the analysis. (Smolinska et al., 2012)

There are many methods for baseline correction, including, among others:

- B-splines and P-splines (B-splines with Penalisation) (Eilers and Marx, 1996);
- Locally weighted scatterplot smoothing (LOWESS) (Xi and Rocke, 2008); and
- Mixture models for baseline estimation (De Rooi and Eilers, 2012).

Evaluation of the fit of a baseline is typically done on a selected number of spectra and by visual inspection only. This is a very subjective way to choose an algorithm and parameters, and is not sufficiently systematic for statistical analysis. (Liland et al., 2010)

Liland et al. (2010) used the root mean squared error of cross-validation as a quality measure with multivariate regression to find optimal baseline algorithms and corresponding parameter values. This was applied to baseline-corrected and normalised data from two spectral data sets. The following algorithms were included in the search for optimal methods and parameter settings:

- Local medians (Friedrichs, 1995);
- Rolling ball (Kneen and Annegarn, 1996);
- Robust baseline estimation (closely related to LOWESS) (Ruckstuhl et al., 2001);
- Simultaneous peak detection and baseline correction (Coombes et al., 2003);
- Asymmetric Least Squares (Eilers, 2003) (regression using penalised least squares);
- Wavelets (Coombes et al., 2005); and
- Iterative polynomial-fitting (Gan et al., 2006; Lieber and Mahadevan-Jansen, 2003).

Instead of using a favourite baseline correction method with traditional parameter settings that provides a visually appealing result, the choice of algorithm and parameters should be optimised for each data set. In this way, simpler and more stable models can be obtained. Over-fitting is always a concern. No overall best algorithm was found, and the results were data-set dependent. The best baseline correction methods for one data set, turned out to be the worst for another data set. (Liland et al., 2010)

Komsta (2011) introduced two new automatic baseline methods for chromatographic signals, based on quantile regression (Koenker and Park, 1996):

- quantile polynomial regression; and
- quantile B-spline smoothing.

The methods are equally applicable to NMR spectra. The main advantage is fully automatic processing of spectra, without parameter setting. The two new quantile methods were compared with existing methods (each with a thresholding and reweighting approach) based on (Komsta, 2011):

- polynomial fitting (Gan et al., 2006);
- spline fitting (Eilers and Marx, 1996);
- LOWESS (Ruckstuhl et al., 2001); and
- Whittaker smoother (penalised regression) (Eilers and Boelens, 2005).

Komsta (2011) also introduced a new method to select curve flexibility in existing algorithms. It is based on the skewness of the residuals. The newly introduced quantile methods performed better than existing methods and required shorter computational time. The existing algorithms were comparable, but polynomial regression had shorter computational time than other existing methods. (Komsta, 2011)

5.4 Removal of specific spectral regions

In human metabolomics, the spectral regions smaller than 0.2 ppm and larger than 10.0 ppm are usually cut off, since they do not contain metabolites produced by the host organism (human). In biofluids, like urine and plasma, the water signal dominates the area from approximately 5.0 to 4.7 ppm, even when water suppression techniques are used. This water region is typically removed from the spectrum. Second to water, urea is the most abundant metabolite in urine. Water suppression as well as pH influences the urea signal. (Smolinska et al., 2012)

In urine, the spectral region from around 6.2 to 4.4 ppm, containing the water ($\sim$ 4.8 ppm) and urea ($\sim$ 5.8 ppm) signals, is frequently excluded.

5.5 Normalisation, Scaling and Transformation

Methods applicable to NMR spectra can be grouped (Zhang et al., 2009) into methods that

(i) remove unwanted sample-to-sample variation (normalisation)

(ii) adjust the variance of the different metabolites (scaling), including variance stabilisation methods and variable scaling methods.

Some methods, like Variance Stabilisation Normalisation (VSN, see section 5.5.2) combines normalisation (i) with variance stabilisation (ii). (Kohl et al., 2012)

In the rest of this chapter, let x_{ij} represent the intensity value for the i^{th} spectrum at position j on the chemical shift axis, where $i = 1 \ldots N$ and $j = 1 \ldots n$. Then, at each position i on the chemical shift axis, the estimated mean and standard deviation (across spectra) are respectively $\bar{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}$ and $s_j = \sqrt{\frac{\sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2}{N-1}}$. Let y_{ij} represent the data x_{ij} after the relevant normalisation, scaling or transformation method was applied.

5.5.1 Normalisation

Normalisation methods aim to remove unwanted sample–to–sample variation (Kohl et al., 2012). This includes correction for the overall concentrations of samples, which influences metabolite dilution (Smolinska et al., 2012).

Samples can display greatly varying concentrations of metabolites from subject to subject. A large part of these subject-to-subject variations are similar across the spectrum for each subject. Thus, all metabolites can be scaled based on some common measure, in order to obtain comparable samples, regardless of variations in concentration in the raw spectra. (Liland, 2011)

Normalisation is classically a multiplication of each NMR spectrum by a constant (Craig et al., 2006). The use of integral normalisation is quite common, but there are many other methods to calculate the normalisation constant for each spectrum (Smolinska et al., 2012).

These include the basic methods (Liland, 2011):

- mean;
- median;
- total standard deviation;
- length of the spectrum vector; and
- total area under the curve, also called total of sum normalisation, integral normalisation (Smolinska et al., 2012) or constant sum normalisation (Craig et al., 2006).

as well as some more refined methods (Smolinska et al., 2012):

- creatinine normalisation, i.e. normalisation using the area under the creatinine peak as reference (for NMR spectra of urine) (Holmes et al., 1994);
- probabilistic quotient normalisation (PQN) (Dieterle et al., 2006);

- histogram matching normalisation (Torgrip et al., 2008); and
- group aggregating normalisation (GAN) (Dong et al., 2011).

Kohl et al. (2012) compared a number of normalisation methods to Probabilistic Quotient Normalisation (Dieterle et al., 2006) on NMR-based metabolomics data. Probabilistic quotient normalisation (PQN) proceeds as follows (Kohl et al., 2012):

1. integral normalisation of every spectrum
2. calculate a reference spectrum (the best is a median spectrum of control samples)
3. for each variable, calculate the quotient of a given test spectrum and the reference spectrum
4. calculate the median of all quotients
5. divide all variables of the test spectrum by the median quotient.

The following methods were compared to PQN:

- Cyclic Loess Normalisation (Cleveland and Devlin, 1988; Dudoit et al., 2002);
- Contrast Normalisation (Åstrand, 2003);
- Quantile Normalisation (Bolstad et al., 2003) ;
- Linear Baseline Normalisation (Bolstad et al., 2003);
- Li-Wong Normalisation (Li and Wong, 2001);
- Cubic-Spline Normalisation (Workman et al., 2002); and
- Variance Stabilisation Normalisation (VSN) (Huber et al., 2002).

The reader is referred to Kohl et al. (2012) for an overview of these methods and the relevant equations in their Supplemental Table S1.

Choice of normalisation methods

With regard to the basic normalisation methods Liland (2011) mentioned that the *median* is more robust than the *mean*, especially when sample–to–sample variation in the number of peaks is large. Given p variables, the *total area under the curve* is simply p times the *mean*. (Liland, 2011)

Quantification of metabolites in different samples may be more accurate when a *standard* is used for normalisation, implying that normalisation is not affected by the amount of peaks or other interfering effects. A stable standard that can be added to a sample for

NMR is trimethylsilyl propionate (TSP) or tetramethylsilane (TMS). Another chemical compound of known concentration (i.e. peak height) can also be used. If the absolute concentration in each sample is of relevance, normalising towards a *standard* can be a good choice. (Liland, 2011)

According to Zhang et al. (2010) *total area under the curve* normalisation and *creatinine* normalisation are widely used. In spectra from very disturbed systems (such as diabetes) or spectra with high concentration metabolites (e.g. glucose in blood), *total area under the curve* and *creatinine* normalisation display quantitative inaccuracy and these methods have been questioned (Zhang et al., 2009). *Probabilistic quotient normalisation* and *histogram matching normalisation* display advantages compared to total area under the curve normalisation and creatinine normalisation (Zhang et al., 2010).

The integral normalisation and vector length normalisation methods have constraints such as a total integral or a total vector length, respectively. If these constraints are not valid, the methods fail. The probabilistic quotient normalisation method has no such constraints. In a real-world metabonomic data set, the PQN outperformed the integral and vector normalisation methods by far, and compensated well for different urine dilutions. Both integral normalization and vector length normalisation, in particular, are hampered by extreme amounts of sample metabolites, such as glucose. The PQN is more robust than integral normalisation, but also more exact for control subjects with only low metabolic variations. The PQN is the better pre-processing method for all possible scenarios of NMR spectra from metabonomic studies, and benefits subsequent multivariate data analyses and quantifications of metabolites. (Dieterle et al., 2006)

PQN supposes that biologically interesting concentration changes affects the NMR spectrum only in certain parts, while dilution effects will influence all metabolite signals. Variations in fluid intake, for example, result in dilution of urine spectra (Kohl et al., 2012). Integral normalisation presupposes that the total integral, which covers all signals, is a function of dilution only. In contrast, the PQN assumes that the intensity of a majority of signals is a function of dilution only. (Dieterle et al., 2006)

Kohl et al. (2012) evaluated the performance of a number of normalisation methods on NMR-based metabolomics data:

Overall between-sample normalisation performance was best for PQN, with Quantile, Cyclic Loess, VSN, and Cubic Spline normalisation methods all performing very well compared to the only creatinine-normalised data. This performance evaluation was in one NMR urine dataset, with kidney disease patients and healthy volunteers. TSP referencing, equidistant binning and normalisation to creatinine were applied. Between-sample normalisation should, however, be balanced with reduction of the real biological signal in the data.

Next, Kohl et al. (2012) evaluated the performance with regards to identification of differentially produced metabolites and the estimation of fold changes in a 'spiked in' data set. TSP referencing, and equidistant binning was applied, but no creatine normalisation was required, since all samples came from the same matrix of pooled urine samples.

For the non-normalised data, Kohl et al. (2012) found that good separation was achieved between spiked and non-spiked data points. The same results were found for the PQN and the Linear Baseline methods. For the Cyclic Loess, Quantile, Cubic Spline, Contrast and VSN methods, there was somewhat less, but still good separation between spike-in and non-spiked data points. (Kohl et al., 2012)

With regard to intensity-dependent bias, Quantile and Cubic Spline performed well, with Cyclic Loess, PQN and VSN evening out most, but not all, of the bias.

For correcting dynamic range (ratio of the largest to the smallest detectable peaks in a spectrum), Quantile and VSN methods performed the best with Cubic Spline and PQN still doing better than creatine-only normalised data as well as spiked-in data.

Concerning standard deviation relative to dynamic spectrum, the standard deviation decreases with feature intensity, is relatively low, and also performs similar or better than the creatinine– or non-normalised data for PQN, Cyclic Loess, Quantile, Linear baseline and Cubic Spline. The VSN improves by keeping the standard deviation relatively constant over the feature intensity range.

In terms of classification performance (by a support vector machine with nested cross validation) on the creatinine–normalised data, normalisation methods are strongly dependent on sample size in the training set. Although the authors claim that Quantile normalisation had the best classification for data sets with over 50 samples and Cubic Splines for smaller data sets, the results were obtained from one data set (ADPKD) and need to be verified on more data sets.

Furthermore, their recommendations are based on the average AUC values for classification, but fail to take into account the confidence intervals around these estimates. When this is taken into account, Quantile, Cubic Spline, VSN and PQN performs similarly from $n = 20$ to $n = 60$, with Quantile and Cubic Spline occasionally having larger CIs, and VSN smaller CIs, than other methods. Creatinine also performs similar at $n = 40$ and $n = 60$, but worse at $n = 20$. At $n = 80$, VSN, Quantile and Cubic spline perform equally well and at $n = 100$, Quantile slightly outperforms VSN and Cubic spline. These results can be interpreted broadly to say VSN is the most consistent over all samples sizes investigated here; that PQN performs comparable up to $n = 60$; from $n = 80$ Quantile performs the best, with cubic spline still doing better than VSN. Still, these results should be verified on data sets from different studies and classification groups (e.g. diseases or treatments).

The authors concluded that inappropriate normalisation methods could considerably damage the data. Although they concluded that widely used normalisation methods were outperformed by Quantile Normalisation (especially for $n \leq 50$ samples), and Cubic Spline Normalisation as an alternative (Kohl et al., 2012), their results should be interpreted with caution. We interpret their results to indicate that VSN is not only a reasonable choice, but may be the preferred method, together with PQN.

In the end, the choice of normalisation method will depend on the application and the known variations between samples (Liland, 2011).

5.5.2 Scaling

Intensity of metabolites can range over orders of magnitude. Furthermore, metabolites with high intensity will often have high variation and thus the greatest effect on the analysis. Scaling is done to prevent selection of the metabolites with the largest intensity as significant. (Smolinska et al., 2012)

Scaling methods are aimed at adjusting the variance of the different metabolites. These include variable scaling and variance stabilisation approaches. (Kohl et al., 2012)

Transformation methods can be included under scaling, but we discuss them in section 5.5.3. Mean-centring is not technically a scaling method, but is described in this section, since it is a pre-processing step applied per variable across samples. We list the different scaling methods according to category, before discussing each method in more detail.

1. Variable scaling methods divide each variable by a scaling factor determined individually for each variable. Variable scaling can be divided into two subclasses (Van den Berg et al., 2006), namely methods that use

   A measure of data dispersion as a scaling factor

   - Auto scaling;
   - Pareto scaling (Wold, 1995);
   - Vast scaling (Keun et al., 2003); and
   - Range scaling (Smilde et al., 2005).

   A size measure as a scaling factor

   - Level scaling.

2. Variance stabilisation methods which reduce heteroscedasticity

   - Variance Stabilisation Normalisation (Huber et al., 2002; Parsons et al., 2007).

Auto scaling

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Auto scaling is also called unit variance (uv) scaling and the standard deviation of the data is used as a scaling factor. In short, the data are first mean-centred across spectra (i.e. by feature), then divided by the standard deviation of each feature. After Auto scaling all features in the data set are considered equally important, but the effect of noise will be increased. (Kohl et al., 2012)

By Auto scaling unit variance is attained, therefore data are then analysed based on the correlations instead of covariances (Smolinska et al., 2012). Between-sample variation, caused by different dilution of samples, is not removed by Auto scaling. In urine samples, dilution can be due to differences in fluid intake. (Kohl et al., 2012)

Wehrens (2011) recommended that Auto scaling not be used for spectral data, since the noise is enlarged to similar importance than the section of the spectra that include the important information.

Pareto scaling

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j}}$$

Pareto scaling is similar to Auto scaling, but uses the square root of the standard deviation as a scaling factor, instead of the standard deviation (Wold, 1995).

The scaling effect of Pareto scaling is not as strong as for Auto scaling, i.e. after Pareto scaling the data remain closer in value to the original data. Pareto scaling is less likely to inflate noisy background data and to diminish the importance of large fold changes compared to small ones. Huge fold changes may, nevertheless, still display a dominating effect. (Kohl et al., 2012)

Pareto scaling is popular in biomarker identification. Specifically in metabolomics, the aim is to identify metabolites that behave differently in two populations, and the interest is focused on high-intensity variables. (Wehrens, 2011)

Vast scaling

$$y_{ij} = \frac{(x_{ij} - \bar{x}_j)}{s_j} \times \frac{\bar{x}_j}{s_j}$$

Vast scaling is an extension of Auto scaling (Keun et al., 2003). Vast is an acronym for variance stability scaling (Van den Berg et al., 2006). The method concentrates on metabolites with small variations, i.e. metabolites that are stable. (Smolinska et al., 2012)

Range scaling

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{(max(x_j) - min(x_j))}$$

In Range scaling, the range of each metabolite is used as the scaling factor (Smilde et al., 2005). Range scaling is sensitive to outliers (Smolinska et al., 2012). For spectral data, the natural lower bound is zero and in this way Range scaling only considers the maximum (Wehrens, 2011).

Level scaling

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{\bar{x}_j}$$

The mean is used as scaling factor in Level scaling. Level scaling is relevant when large relative changes are of interest. (Smolinska et al., 2012)

Variance Stabilisation Normalisation

Variance Stabilisation Normalisation (VSN) is a set of non-linear transformations that aim to keep the variance constant over the entire data range (Huber et al., 2002; Parsons et al., 2007). VSN addresses the problem of non-constant coefficient of variation by using the inverse hyperbolic sine. The data are returned on a generalised logarithm (glog) scale to base 2. For large values, this transformation approaches the logarithm, thus removing heteroscedasticity. For small values, it approaches the linear transformation, and the variance remains unchanged. (Kohl et al., 2012)

Mean-centring

As mentioned previously, mean-centring is not technically a scaling method. It can be applied as a pre-processing step or as part of statistical analysis (Liland, 2011).

$$y_{ij} = x_{ij} - \bar{x}_j$$

Mean-centring is applied per variable across samples (Liland, 2011). It converts all values to vary around zero instead of around the mean, thus regulating for differences between high-intensity and low-intensity chemical compounds (metabolites). Mean-centring does not remove heteroscedasticity. The method is often used in combination with other scaling methods. (Smolinska et al., 2012)

5.5.3 Transformation

Apart from scaling methods, transformation methods can be utilised as a step in pre-processing. When metabolite concentrations vary in orders of magnitude, it is wise to do a transformation in order to avoid the statistical analysis emphasising only metabolites with larger concentrations. The disadvantage of transforming spectra is the potential increase in the noise. (Liland, 2011)

Log transformation

$$y_{ij} = \log_{10}(x_{ij})$$

Large dominant features in the data can get in the way of analysis or pre-processing, e.g. in alignment (Wehrens, 2011). In logarithmic (log) transformation, large dominant features (peaks) in the data are reduced relatively more than smaller features (Smolinska et al., 2012) and noise is made more constant over the whole range (Wehrens, 2011). In this way, the log transform removes heteroscedasticy from data, provided that the relative standard deviation is constant (Smolinska et al., 2012). Note that y_{ij} does not exist for $x_{ij} \leq 0$.

When noise is multiplicative rather than additive, i.e. the level of variation depends on the signal strength, log-transformation of the data is appropriate. (Wehrens, 2011)

Square root transformation

$$y_{ij} = \sqrt{x_{ij}}$$

The square root transformation is typically used for spectral data where the data can be seen as generated from a Poisson process, e.g. ion count, in time-of-flight (TOF) mass spectrometry (MS) (Liland, 2011). The square root transformation is sometimes referred to as the power transformation (Van den Berg et al., 2006), although power transformations other than 1/2 are possible. These are described under Box-Cox transformations.

Box–Cox transformation

$$y_{ij}^{(\lambda)} = \begin{cases} \frac{x_{ij}^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0, \\ log(x_{ij}) & \text{if } \lambda = 0. \end{cases}$$

The Box–Cox transformation (Box and Cox, 1964) is a parametric power transformation technique. It reduces the effect of non-normality and heteroscedasticity (Sakia, 1992) and can be used for pre-processing (Smolinska et al., 2012).

Glog transform

The generalised logarithm (glog) transform (Durbin et al., 2002) is a variance stabilising method and makes use of a transform parameter, λ. If x represents the untransformed data, then y is the glog transformed data:

$$y_{ij} = ln(x_{ij} + \sqrt{x_{ij}^2 + \lambda}) \tag{5.1}$$

Parsons et al. (2007) extended the glog to suppress noise. The extended glog transform is given by

$$y_{ij} = ln((x_{ij} - x_0) + \sqrt{(x_{ij} - x_0)^2 + \lambda}) \tag{5.2}$$

where x_0 shifts the glog function to suppress noise and x_0 is dependent on the choice of λ.

The reader is referred to Parsons et al. (2007) for details on optimisation of λ and x_0 in the glog and extended glog transforms.

Choice of scaling/transformation methods

The scaling method used can have a huge influence on the result of an analysis (Wehrens, 2011). Van den Berg et al. (2006) compared Centring, Auto Scaling, Range Scaling, Pareto Scaling, Vast Scaling, Level Scaling, Log Transformation and Power Transformation to GC-MS based metabolomics data. They found that the selection of the scaling/transformation method depended on (i) the biological question; (ii) the data set's general properties; and, (iii) the statistical methodology following the pre-processing. Principal Component

Analysis (PCA) was used to evaluate the effect of scaling/transformation methods on data analysis and PCA score plots were judged visually based on distance within as well as between clusters belonging to different groups. On one of the data sets Range Scaling and Auto Scaling performed best: clear clustering was visible in PCA score plots and the dependence of the rank of metabolites on the average concentration and the magnitude of fold changes were removed. (Van den Berg et al., 2006)

Kohl et al. (2012) evaluated normalisation and scaling methods, for their application to NMR-based metabolite fingerprinting, as mentioned in section 5.5.1. Concerning the scaling methods (auto, Pareto and VSN), VSN outperformed auto and Pareto scaling. VSN improved on other scaling/normalisation methods by keeping the standard deviation relatively constant over the feature intensity range. VSN performed consistently well over all sample sizes investigated (n=20 to 100). VSN also compared very well with normalisation methods. Apart from performing very well on overall between-sample normalisation, VSN attained good separation between spike-in and non-spiked data points, evened out most of the intensity-dependent bias, and performed excellently for correcting dynamic range.

Parsons et al. (2007) compared the glog and extended glog transforms to Auto scaling, Pareto scaling and unscaled NMR metabolomics data. For three NMR datasets glog and extended glog transforms attained the best, or equal to the best, classification accuracy. Classification accuracy was determined by PCA followed by linear discriminant analysis (LDA) on the first two PCs. Sensitivity, specificity and cross-validation accuracy were calculated. Furthermore, the glog transform was considerably better at discovering metabolic biomarkers that can discriminate between sample classes. This was based on the top five peaks in the corresponding PCA loadings plots. Note that spectra were grouped in 0.005 ppm bins, residual water and urea (where relevant) sections were removed and spectra were normalised to a total spectral area of 1. (Parsons et al., 2007)

5.6 Alignment

The problem of misalignment is also referred to as peak shift. It originates when, for different samples (spectra), the same molecule displays peaks at different chemical shift positions (on the horizontal axis). A single sample displays in the order of hundreds to thousands of peaks. When a data set contains many samples, it will be unclear which peaks, from among the many possibilities, should be aligned between different spectra. (Torgrip et al., 2010)

Considering statistical analysis, the minor peak shifts between different NMR spectra are disturbing (Ebbels et al., 2011): it is assumed that, for a certain molecule, peak intensity is contained in a unique column of the data; in the case of peak shifts, the relevant intensities are not limited to a unique column; thus it is more likely to miss potential biomarkers (interesting molecules); and given systematic peak shifts, it is possible to detect false (positive) markers (Torgrip et al., 2010). In short, minor but important peak shifts can

make it impossible to detect patterns that exist in the spectra (Smolinska et al., 2012) and for this reason peak alignment is critical.

Similar to NMR data, many other types of analytical chemistry data are subject to minor misalignments (Wehrens, 2011) and, as such, the peak shift problem is platform independent (Torgrip et al., 2010). However, NMR data are more complex in the sense that peak shifts are not at all uniform over the chemical shift axis: many peaks can stay at their initial positions, a number of peaks can shift with varying distances, and, to complicate matters, in different directions (on the chemical shift axis) (Wehrens, 2011).

In alignment of peaks between two different spectra, the match will, for many peaks, be quite obvious. For some peaks, however, there will be more than one peak in the other spectrum with which they can be aligned and it will be unclear which is correct. There are essentially three instances of ambiguous peak alignment (Torgrip et al., 2010):

1. A single peak from spectrum A match either of two peaks in spectrum B
2. Two peaks in spectrum A match two peaks in spectrum B, but with minor or no shifting the last peak in A matches the first peak in B
3. Peaks alter their order between samples A and B.

Most alignment methods only consider the first two cases of ambiguous peak alignment.

In chemometrics the aim of alignment is to obtain individual profiles (spectra) that look alike as far as possible, though the area and shape of peaks should preferably not be changed. Thus, alterations should best be done in the spectra's baseline. This is only possible if the spectra are reasonably similar (Skov et al., 2006). Moreover, and often not explicitly stated, peaks from the same molecule should be aligned (Torgrip et al., 2010).

In terms of terminology, the process of alignment is also called warping or occasionally (Torgrip et al., 2010) synchronisation. Misalignment is also referred to as peak shift, positional uncertainty or sometimes (Torgrip et al., 2010) unsynchronised data. Torgrip et al. (2010) described the alignment of chemical analytes within the context of a general problem formulation, known as the 'inter-sample correspondence problem'.

5.6.1 Reasons for peak shift (misalignment)

There can be a number of reasons for peak shifts in NMR (Torgrip et al., 2010):

(a) instrumental drift

(b) physio-chemistry of the sample

(c) random variation

(d) post-processing artifacts.

Instrumental drift

For any instrumental technique, e.g. NMR or MS, a component of instrumental drift is ubiquitous in the data. The amount of drift relative to the duration of the experiment's measurement time will determine if the instrumental drift is seen as random or systematic. Misalignment due to instrumental drift should usually be reasonably small. If not, experimental protocol should be revised. (Torgrip et al., 2010)

Physicochemical properties of the sample

In NMR data, the peak shifts (misalignment) along horizontal axis (chemical shift) are due to either the sample's physio-chemical properties or changes in these properties: (Torgrip et al., 2010; Smolinska et al., 2012; Fan and Lane, 2008):

- Overall dilution;
- Changes in temperature;
- Changes in pH;
- Changes in salt concentration; and/or
- Relative concentration of specific ions.

It is standard procedure to add a buffer to reduce pH variations in the sample (Fan and Lane, 2008) and to manage temperature throughout data acquisition. Even when samples are buffered, this does not guarantee peak alignment. For example, citrate peaks are sensitive to salt and pH and are well known for shifting in ^{1}H NMR spectra, regardless of buffering. To complicate matters, different metabolites are affected in very different ways by changes in the physio-chemical properties of the sample. (Torgrip et al., 2010)

Pre-processing

Peaks can be misaligned due to pre-processing, e.g. NMR peaks can change in shape after Gaussian line-broadening, i.e. free induction decays (FIDs) (convolution) for noise repression and smoothing. (Torgrip et al., 2010)

Random shift

Random shift is the portion of peak misalignment that cannot be ascribed to a systematic nature, e.g. instrumental drift or chemistry. Supposedly, the random shift is minor and can be dealt with using almost any current alignment method. (Torgrip et al., 2010)

5.6.2 Alignment methods

As a first step in alignment of NMR spectra, *spectral referencing* is used. An internal reference (or internal standard), is added to each sample before chemical analysis. Typically tetramethylsilane (TMS) or trimethylsilyl propionate (TSP) is used. Spectral referencing fixes the internal reference signal, e.g. TSP, to 0 ppm. This is a global alignment method that shifts the entire spectrum on the chemical shift axis. However, this method is not sufficient to address local alignment issues and should be followed by another alignment method. (Smolinska et al., 2012)

In a recent review of alignment methods for NMR spectra Vu and Laukens (2013) included 18 alignment methods, and discussed these according to methodological variations, i.e.:

- Alignment using extracted peaks (peak picking) vs. full spectra;
- Pairwise alignment to a reference spectrum vs. inter-sample methods with no reference spectrum (compare section 5.6.3);
- Alignment of entire spectra vs. spectral segments;
- Different 'target functions' for optimisation of alignment, e.g. Pearson correlation coefficient, (squared) Euclidian distances, FFT cross-correlation for segment alignment or other method-specific 'target functions';
- Correction of misalignment via shifting and/or stretching/compression, or a model, e.g. polynomial or Bayesian;
- Evaluation of quality of alignment (compare section 5.7); and
- Method complexity, including computational time and the number of user-defined parameters (excluding parameters required for peak extraction

In a recent tutorial on warping methods for spectroscopic and chromatographic signal alignment, Bloemberg et al. (2013) provided a critical introduction to what they consider the most important warping methods. A number of methods were demonstrated on an NMR example.

An in-depth description of these alignment methods falls outside the scope of this work and the reader is referred to the abovementioned review and tutorial.

5.6.3 Selection of a Reference Spectrum

Fundamental to the alignment process is the challenge of selecting a suitable reference spectrum to align to (Skov et al., 2006). The reference profile should preferably be representative of all chemical compounds (metabolites) in the data and their associated intensity peaks. (Veselkov et al., 2009) (Supporting Information)

One solution is to select the mean spectrum over the data as the reference spectrum. Another option is to select the first principal component loadings. Neither of these is ideal, since they can have profoundly deformed peaks, negatively impacting on the success of the alignment. (Veselkov et al., 2009) (Supporting Information)

In a study by Giskeodegard et al. (2010), a bad choice of reference had a larger influence on the correlations between spectra, than the particular warping method used.

Similarity index

Skov et al. (2006) presented a similarity index to select a reference profile for chromatographic data. It is equally relevant in spectroscopy.

This similarity index is the product of Pearson correlation coefficients (CCs) between a test spectrum $\mathbf{x}_T$ and all other spectra of interest $\mathbf{x}_i$:

$$\text{similarity index} = \prod_{i=1}^{N} |cc(\mathbf{x}_T, \mathbf{x}_i)| \tag{5.3}$$

where the correlation coefficient (cc) is given by (5.8) (Skov et al., 2006).

For each spectrum in the data set the similarity index will be less than or equal to one. The spectrum with the largest similarity index is chosen as the most suitable reference spectrum for the specific data set. This spectrum will be the most similar to all others. (Skov et al., 2006)

The index does not give a good indication of the similarity among spectra, because the correlation coefficient between spectra is disproportionally influenced by the difference in peak heights, as well as by the covariance of the highest peaks. (Veselkov et al., 2009) (Supporting Information) (also see section 5.7.1)

Closeness index

To address the problem of undue influence of a few large peaks on the similarity index (5.3), Veselkov et al. (2009) (Supporting Information) scaled the local areas to an equal variance prior to calculating the segment-wise correlation coefficient cc_{bin} (5.7). They re-defined the similarity index as the closeness index:

$$\text{closeness index} = \prod_{i=1}^{N} cc_{\alpha}(x_T, x_i) \tag{5.4}$$

where cc is the correlation coefficient between the variance-scaled potential reference (or target), x_T, and the i^{th} spectrum, x_i.

The subscript α, for example $\alpha = 0.02$ indicates that spectral segments down to step size of 0.02 ppm are scaled to unit variance. The value of α should be chosen as equivalent to

the size of an average peak. This will provide the same influence of small as well as large peaks. The spectrum with the highest closeness index is selected as a reference. (Veselkov et al., 2009)

A variety of reference spectra

Giskeodegard et al. (2010) recommended the use of various reference spectra for alignment. Attempting a number of references is not a large effort, yet it will create awareness of the results' variability and will possibly give a result near the optimum (Giskeodegard et al., 2010).

The authors used a selection of 10 different reference spectra (Giskeodegard et al., 2010):

- the spectrum with the highest average correlation with all other spectra (say, this spectrum is from class 1). This choice appears to always give good results, but it is not certain that it is the optimal result.

For data sets with two (or more) classes, alignment will possibly be influenced by the class to which the reference belongs. Trying references from both classes may therefore be advisable. Apart from the reference spectrum mentioned above, also:

- the (on average) second most highly correlated spectrum from class 1;
- two random spectra from class 1;
- the most highly correlated spectrum from class 2;
- the second most highly correlated spectrum from class 2; and
- two random spectra from class 2.

Using a 'central' spectrum as a reference is also an option:

- the overall mean spectrum; and
- the overall median spectrum.

For data sets with large misalignments, the mean or median spectrum may be a bad reference spectrum since it can have wide peaks and will not look like any one of the real spectra. Aligning the data with the mean or median spectrum and then recalculating the mean or median as a reference spectrum, may solve the problem in an iterative way. (Giskeodegard et al., 2010)

Veselkov et al. (2009) also recommended the use several reference profiles per data set. When substantial metabolic changes occur due to a treatment, they advise a separate reference spectrum per treatment group, with the two (or more) reference spectra aligned before aligning other spectra to them. (Veselkov et al., 2009) (Supporting Information)

Variable reference alignment

MacKinnon et al. (2012) proposed a variable reference alignment, as opposed to a single reference alignment (see 5.6.3, 5.6.3 and 5.6.3).

Variable reference alignment generates spectral segments that share a common target spectrum. The goal of this approach is to perform local alignment corrections on spectral regions, which share a common 'most similar' spectrum. Spectral segments are automatically defined in the process. (MacKinnon et al., 2012)

The segmentation and construction of a composite reference spectrum is done by identifying spectral segments sharing a common reference spectrum, as calculated according to (5.4) (with e.g. $\alpha = 0.02$ppm). The segments are generated by incremental growth of a test segment followed by calculation of the reference spectrum, repeated until a different reference spectrum is identified. The segment boundaries are thus defined and incremental growth of a new segment occurs in an identical fashion. Subsequently the segments are individually subjected to an alignment algorithm (Section 5.6.2), with alignment taking place toward the segment–specific reference spectrum. Finally, the fully aligned spectrum is reassembled. (MacKinnon et al., 2012)

Parameters to be specified include the maximum inter-peak distance (e.g. 20 Hz) and the maximum shift threshold (e.g. 25 Hz). Alignment of a segment is accepted subject to an improvement in the alignment quality parameter (5.7.1 with $\alpha = 0.05$ppm). As always, the specific alignment algorithm and respective parameters should be chosen with caution, and are reliant on the data set as well as the user's requirements. (MacKinnon et al., 2012)

MacKinnon et al. (2012) used variable reference alignment with both the *i*coshift (Savorani et al., 2010) and RSPA (Veselkov et al., 2009) alignment algorithms. For *i*coshift, the automated selection of spectral segments with non-constant length resulted in better alignment. For RSPA, the alignment toward a segment specific reference spectrum provided improved alignment. Variable reference alignment showed improved quality of alignment in ^{1}H NMR data sets that exhibit large inter-sample compositional variation (e.g. ionic strength, pH). (MacKinnon et al., 2012)

5.7 Evaluation of Alignment

In a review of NMR alignment methods Giskeodegard et al. (2010) included five different measures to assess the results of alignment, namely: correlation, simplicity value, peak factor, classification and visual inspection (See sections 5.7.1, 5.7.2, 5.7.3, 5.7.4 and 5.7.5 respectively). The simplicity value and peak factor were described in detail by Skov et al. (2006). They combined these two measures to form the warping effect (section 5.7.3).

Over a large chemical shift interval, major peaks will be very influential on the variance and consequently on the correlation. Minor peaks will not have much influence (section 5.7.1). Additionally, correlation, as an evaluation criterion for alignment, only performs well after

scaling. To overcome this problem of large peaks dominating the correlation, Veselkov et al. (2009) developed the Alignment Quality measure (section 5.7.1).

Even more recently than the review by Giskeodegard, a similar set of criteria was used by MacKinnon et al. (2012) to assess alignment, namely, the mean correlation coefficient (section 5.7.1), the alignment quality parameter (section 5.7.1), the simplicity value (section 5.7.2) and the peak factor (section 5.7.3). Additionally the authors also used scree plots (section 5.7.2) and a pseudo-variable importance to projection (VIP) score (section 5.7.2) from PCA. The pseudo VIP score is a qualitative measure of improved information recovery (MacKinnon et al., 2012).

Zhang et al. (2012) assessed the alignment quality by correlation maps (section 5.7.5). They also used the mean of the mean correlation coefficients (mcc) (section 5.7.1) between the reference spectrum and the spectra to be aligned. Certain alignment methods change peak shapes. Zhang et al. (2012) quantified changes in peak area and used the mean relative change in area (mrca) (section 5.7.3) to evaluate the ability to maintain peak shapes during alignment.

In addition to the Correlation coefficient and the Alignment Quality measure, Veselkov et al. (2009) also utilised 1D-STOCSY covariance plots (section 5.7.5) to evaluate the success of alignment relating to details of the molecular structures of complex biological mixtures.

We have classified the evaluation criteria mentioned above according to the type of measure used for evaluating the spectral alignment. Sections 5.7.1 to 5.7.5 below respectively cover measures of correlation, explained variance, peak shape, classification and plots/maps for visual assessment.

5.7.1 Measures of Correlation

Correlation

Following alignment, spectra should be more similar and thus have a greater correlation. Calculation of the correlation between spectra, both before and after alignment, gives a basic criterion for evaluation of spectral alignment. (Giskeodegard et al., 2010)

The correlation coefficient (cc) between two spectra, $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ is given by

$$cc(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{Cov(\mathbf{x}_i, \mathbf{x}_{i'})}{\sqrt{Var(\mathbf{x}_i)Var(\mathbf{x}_{i'})}} \tag{5.5}$$

As mentioned in section 5.7 correlation requires similarity in spectra's sample composition, small peaks are down weighed, and a few large peaks can dominate the correlation coefficient (Veselkov et al., 2009).

Mean correlation coefficients (mcc)

Zhang et al. (2012) utilised the mean of the mean correlation coefficients (MCC). This measure is calculated between the reference spectrum and the spectra to be aligned:

$$mcc(\mathbf{x}_T, \mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{n}(\mathbf{x}_{Tj}-\bar{\mathbf{x}}_T)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)}{\sqrt{\sum_{j=1}^{n}(\mathbf{x}_{Tj}-\bar{\mathbf{x}}_T)^2}\sqrt{\sum_{j=1}^{n}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)^2}}\right) \tag{5.6}$$

where $\mathbf{x}_T$ is a vector of the target (or reference) signal and $\mathbf{X}$ is a matrix with elements x_{ij} and each row of $\mathbf{X}$ is a vector $\mathbf{x}_i$ of a spectrum to be aligned. (Zhang et al., 2012)

Veselkov et al. (2009) removed the difference in metabolite concentrations by variance-scaling spectra, followed by the mean of correlation coefficients between spectra, to evaluate quality of alignment (section 5.7.1).

Alignment quality measure ($aq_{\mathbf{bin}}$)

One or more high peaks can dominate the correlation coefficient (cc) (section 5.7.1). To address this issue, Veselkov et al. (2009) divides the spectra into segments and scales each segment by mean centring and adjusting to unit variance. They defined the correlation coefficient, cc_{bin}, on the abovementioned segments (called bins) :

$$cc_{\text{bin}}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{Cov(\mathbf{x}_{i,\text{bin}}, \mathbf{x}_{i',\text{bin}})}{\sqrt{Var(\mathbf{x}_{i,\text{bin}})Var(\mathbf{x}_{i',\text{bin}})}} \tag{5.7}$$

The cc_{bin} reflects, on a specific segment, the similarity of peaks between any two spectra, $\mathbf{x}_i$ and $\mathbf{x}_{i'}$.

$$cc(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{Cov(\mathbf{x}_i, \mathbf{x}_{i'})}{\sqrt{Var(\mathbf{x}_i)Var(\mathbf{x}_{i'})}} \tag{5.8}$$

Veselkov et al. (2009) used bin sizes of $\delta = 0.02$ or $\delta = 0.08$ ppm. The smaller bin size (0.02 ppm) was chosen to allow the equal influence of both minor and major peaks in the alignment quality measure, and was selected as a multiple of the full width half maximum of a typical peak. The larger bin size (0.08 ppm) is useful for assessing alignment of major peaks, since the bin size is large enough to minimise the role of minor peaks.

Next, the authors calculated the mean of all pairwise cc_{bin} values, i.e. the values below the main diagonal of the correlation matrix:

$$aq_{\text{bin}} = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{i'=1}^{i-1} cc_{\text{bin}}(\mathbf{x}_i, \mathbf{x}_{i'}) \tag{5.9}$$

where $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ are the ith and i'th spectra, respectively. Veselkov et al. (2009) used aq_{bin} to evaluate the peak alignment quality across all spectra. For a data set, the value

of aq_{bin} can range from zero (non-aligned) to one (completely aligned). (Veselkov et al., 2009)

5.7.2 Measures of explained variance

Simplicity value

Skov et al. (2006) initially developed the simplicity value for chromatographic data, but it has also been applied successfully to spectral data (Giskeodegard et al., 2010).

The simplicity value is connected, via singular value decomposition (SVD) to principal component analysis (PCA) (Giskeodegard et al., 2010).

The original data $\mathbf{X}$ can be decomposed as:

$$\mathbf{X} = \mathbf{USV}^T \tag{5.10}$$

where $\mathbf{S}$ is a diagonal matrix containing the singular values equal to the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$. $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal matrices, where the columns in $\mathbf{U}$ are the eigenvectors of $\mathbf{XX}^T$ and the columns of $\mathbf{V}$ the eigenvectors of $\mathbf{X}^T\mathbf{X}$. (Skov et al., 2006)

In SVD of a matrix, the sum of squared singular values is equal to the total sum of squares of all the original data entries in the uncentred data matrix, $\mathbf{X}$ (Skov et al., 2006).

The sum of the first R squared singular values (scaled to a total sum of squares of one) is a measure of how much of the variation is explained by the corresponding R components:

$$\text{Explained variance} = \sum_{r=1}^{R} \left(\text{SVD}\left(\frac{\mathbf{X}}{\sqrt{\sum_i \sum_j x_{ij}^2}} \right) \right)^2 \tag{5.11}$$

where SVD(M) denotes the singular value for a given component r. The above expression is by definition equal to one if all singular values are retained, and, as such, this sum cannot be used to evaluate pre-processing and the effect of alignment.

The simplicity value ($0 \leq \text{simplicity} \leq 1$) of a matrix is defined as the sum of all singular values of the matrix–scaled to a total sum of squares of one–taken to the fourth power:

$$\text{Simplicity} = \sum_{r=1}^{R} \left(\text{SVD}\left(\frac{\mathbf{X}}{\sqrt{\sum_i \sum_j x_{ij}^2}} \right) \right)^4 \tag{5.12}$$

Aligned spectra will have more variance explained by the first components, and thus the simplicity value will be higher. In general the simplicity value will be smaller if the spectra are not well aligned. For perfectly alignment spectra, the simplicity value will be close to, but not necessarily equal to, one. (Skov et al., 2006)

Principal Component loadings

Veselkov et al. (2009) evaluated alignment of spectral data by the increase of explained variance by principal components (PCs), after alignment.

In unscaled data, the largest peaks contribute more to overall variance (section 5.7), and this poses the same challenge as for the *cc* analysis (section 5.7.1). Principal component analysis (PCA) on unscaled data is skewed towards the variation of the highest peaks. To solve this problem, each variable is scaled to unit variance. In PCA of either scaled or unscaled data, increase in explained variance should reflect the variation in chemical composition, but not variation in variable peak positions – and should be investigated. (Veselkov et al., 2009)

If PC scores explain the variation in chemical composition, the line shapes of PC loadings will resemble NMR spectral peaks. However, if PC scores explain variation caused by distortions in phase, variable peak position and peak line shape, the line shapes of PC loadings will be distorted. For unit-variance scaled models, unlike unscaled models, the line shapes of loadings are not interpretable and no direct identification of peaks is possible. However, for unit-variance (*uv*) scaled models, a loading value can be transformed by multiplying it by the standard deviation of an original spectral variable for interpretation of the main source of the variance contribution into PC scores, similar to the unscaled case. The transformed *uv*–loadings can be plotted using a colour code corresponding to the weight value obtained from a unit–variance PC. (Veselkov et al., 2009)

VIP scores from Principal Components Analysis

MacKinnon et al. (2012) utilised a pseudo-variable importance to projection (VIP) score after they did PCA on the unaligned and aligned data sets. For the PCA they mean-centred the data.

The pseudo-VIP score provides a measure of the quality of improved information recovery. The only change from the standard VIP score is, for the pseudo-VIP score, an unsupervised multivariate method (i.e., PCA) was used, as opposed to a supervised multivariate analysis technique (e.g., PLS). (MacKinnon et al., 2012)

The sum of the weighted latent variable loadings is, in essence, the VIP score. Each loading is weighted by the fraction of variation described by the latent variable. The average VIP over all variables is 1. Therefore, instead of the average, the total number of variables with a pseudo-VIP score greater than 1 (i.e., regarded as significant), was calculated for the unaligned and the aligned PCA models, respectively. (MacKinnon et al., 2012)

5.7.3 Measures of Peak shape

Certain alignment methods can alter the shape of peaks. To investigate an alignment method's ability to preserve the shape of peaks, the differences in peak shape before and after alignment can be quantified. (Zhang et al., 2012)

Peak factor

Giskeodegard et al. (2010) used the peak factor to evaluate alignment or NMR spectra. The peak factor was originally described by Skov et al. (2006) for chromatograms. A peak factor of 1, implies there is no change in peak shape (Giskeodegard et al., 2010).

Skov et al. (2006) quantified the change in peak shape and named it peak factor. The peak factor can range from 0 to 1 and indicates how much the collection of samples has changed.

$$\text{Peak factor} = \frac{\sum_{i=1}^{N}(1 - \min(c_i, 1)^2)}{N} \tag{5.13}$$

where

$$c_i = \left| \frac{\|\mathbf{y}_i\| - \|\mathbf{x}_i\|}{\|\mathbf{x}_i\|} \right| \tag{5.14}$$

and

$$\|\mathbf{x}_i\| = \sqrt{\sum_{j=1}^{n} x_{ij}^2} \tag{5.15}$$

is the norm (Euclidian length) for $\mathbf{x}_i$; $\mathbf{x}_i$ and $\mathbf{y}_i$ is the same spectrum, respectively before and after alignment. If the norm stays the same in (5.14), the relative change is 0, and the total contribution for that sample is 1 in Equation (5.13). If there is little change in the spectrum after alignment, c_i (5.14) will be between 0 and 1, and the total contribution for the sample will be smaller than 1 in (5.13). When the aligned spectrum is very distorted, c_i will be larger, and the sample's total contribution (in (5.13)) will be 0. Higher peaks will have relatively more influence on the peak factor, because of the use of the norm. (Skov et al., 2006)

Peak factor values can be plotted with their simplicity values for the data. High simplicity values, but with 'low' peak factor values, does not indicate good alignment. (Skov et al., 2006)

Mean relative change in area (MRCA)

Zhang et al. (2012) adopted the mean relative change in area to evaluate the changes in area during alignment.

They defined the mean relative change in area (MRCA) as

$$mrca = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{|\sum_{j=1}^{n} y_{ij} - \sum_{j=1}^{n} x_{ij}|}{\sum_{j=1}^{n} x_{ij}} \right) \tag{5.16}$$

where x_{ij} are elements of N row vectors $\mathbf{x}_i$ with j elements each. Each vector, or spectrum, $\mathbf{x}_i$ need to be aligned. The y_{ij} are elements of N row vectors $\mathbf{y}_i$ after alignment.

Considering both the mean correlation coefficient (mcc, section 5.7.1) and the MRCA of two alignment methods: if the MCC for the second method is larger than for the first method, but the MRCA for the first method is low, the reliability of the large MCC for the second method, obtained at the cost of peak shapes (large MRCA), is questionable. (Zhang et al., 2012)

Zhang et al. (2012) compared alignment methods with regard to: alignment quality (section 5.7.1), changes in the shapes of peaks (MCRA), speed of the method (time required by calculations) and the best trade-off between speed and alignment quality.

The warping effect

The warping effect combines the simplicity measure (section 5.7.2) and the peak factor value (section 5.7.3) and was introduced by Skov et al. (2006)

$$\text{warping effect} = \text{simplicity} + \text{peak factor} \tag{5.17}$$

The warping effect can take on values from 0 to 2.

5.7.4 Measures of Classifiability

Classification

After alignment classification results should improve, especially for spectra influenced by random shifts. On the other hand, different classes of spectra may contain information based on biological differences and alignment may destroy this biological information. (Giskeodegard et al., 2010)

Partial Least Squares-Discriminant Analysis (PLS-DA) uses latent variables (LVs) to maximise the covariance between the spectra and an outcome variable and aims to discriminate between classes. In their review of alignment methods, Giskeodegard et al. (2010) applied PLS-DA to assess the classifiability of aligned and unaligned spectra. They also utilised the warping path or warping parameters as input to investigate possible shift information. (Giskeodegard et al., 2010)

5.7.5 Visual inspection of plots and maps

Quantitative measures of alignment (sections 5.7.1, 5.7.2, 5.7.3) are helpful for comparing specific characteristics of large data sets at a glance, but are not without limitations. The human eye-brain combination still excels in tasks of pattern recognition. (Giskeodegard et al., 2010)

Giskeodegard et al. (2010) used visual inspection of end results to assess alignment quality and detect artefacts. They emphasised that this is 'an absolute necessity'. Quantitative measures can produce good results, regardless of artefacts. Nevertheless, visual inspection on its own is open to personal opinion and not reliable. (Giskeodegard et al., 2010)

Heat maps

Peak-position variation in raw data and subsequent alignment of peaks can be displayed in heat maps of peak intensity (with ppm on the x-axis and sample number on the y-axis) of all samples combined with their spectral plot (intensity by ppm) (Veselkov et al., 2009).

Correlation maps

The results from various alignment methods, applied to the same data set, can be presented with between-sample correlation maps where the colour of a block represents the strength of the correlation (See section 5.7.1) between two spectra. The correlation maps will indicate which methods can improve similarity between samples. The size of between-sample correlation coefficients can be visually compared for different alignment methods. A method that aligns peaks more accurately will display the colours associated with higher correlation coefficients. (Zhang et al., 2012)

1D-STOCSY covariance plots

NMR spectroscopy in biofluids gives detailed information on molecular structures and peaks in observed ^{1}H NMR data are statistical correlated. These correlations can be produced by both biological relationships and intra-molecular connectivity of proton nuclei. In general, the structural correlations are stronger than the biological correlations. (Cloarec et al., 2005)

Statistical total correlation spectroscopy (STOCSY) analysis exploits this multicolinearity to unveil structural correlations between certain peaks in a group of spectra. STOCSY investigates the correlation matrix of a group of ^{1}H NMR spectra

$$\mathbf{R} = \frac{1}{N-1}\mathbf{X}\mathbf{X}' \tag{5.18}$$

where N is the number of samples, $\mathbf{X}$ is a data matrix of ^{1}H NMR spectra (columns are intensity variables scaled to unit variance; rows are spectral observations) and $\mathbf{R}$ is a matrix of pairwise correlations (cc) between intensity variables, but not between spectra. It is typical to study correlations of one specific variable to all other intensity variables, and this is called 1D-STOCSY. (Cloarec et al., 2005)

The influence of spectral misalignment on 1D–STOCSY can be displayed by plotting the covariance between intensity variables, colour coded by their correlations. Line shapes of covariance patterns should look like peaks in an NMR spectrum, but will be deformed by misalignment. Identification of biological as well as structural correlations will be enhanced by effective alignment. (Veselkov et al., 2009)

5.8 Conclusion

In Chapter 5 we provided an introductory overview to some of the most important steps and issues involved in pre-processing of so called one-dimensional chemometric data. Statisticians are often oblivious to these methods commonly used in chemometrics. We do not claim that this is a complete overview of this complex topic. Although our focus is on nuclear magnetic resonance data in metabolomics, most of the methods in Chapter 5 are also applicable to other spectral and chemometric data and other applications.

The choice of pre-processing methods (section 5.2) and order of baseline correction (section 5.3), normalisation (section 5.5.1), scaling (section 5.5.2), transformation (section 5.5.3) and alignment (section 5.6) are, in general, subjective and dependent on the analyst. These choices and the order in which these pre-processing methods are applied can individually or combined have a strong influence on the results of subsequent data analysis Engel et al. (2013). This is a serious concern regarding the art of chemometric pre-processing.

6

Functional registration subject to constraints

Registration of functional data refers to the process of transforming the time argument so that features in the data are more aligned. The process is often carried out by estimating a time-warping function for each curve and then applying these warping functions to the smoothed curves prior to statistical analysis. The warping functions are typically estimated by minimising the difference between the warped functions while controlling the roughness of the warping function. The registration process, in essence, separates the *phase* and *amplitude* variation. However, warping may destroy essential properties of the observed data, i.e. internal data structure, possibly originating from physical constraints in the system generating the data. This aspect of warping is often ignored, for example when NMR spectral curves are continuously warped without preserving the shape of peaks.

Statistics of Time Warpings and Phase Variations are current topics in Functional Data Analysis and were the focus of a workshop at the Mathematical Biosciences Institute (MBI) in November 2012 in Columbus, Ohio. The juggling data referred to below were provided as part of this MBI workshop. Results from the workshop, including our Paper III, Tolver et al. (2014), have been accepted for publication as a Special Section in the Electronic Journal of Statistics.

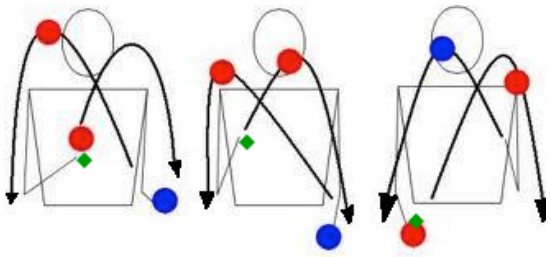

Figure 6.1: Diagram of a three-ball juggling cycle. The green diamond indicates the approximate hand/finger position during three stages of a typical juggling cycle. Adapted from Steve (2014)

The juggling data are from an experiment where a juggler juggled three balls (Figure 6.1).

The three-dimensional position of the tip of the juggler's right index finger was recorded 200 times per second. The juggler performed ten juggling trials. Each trial lasted ten seconds and contained 11 to 13 juggling cycles. A juggling cycle began with throwing a ball and ended with catching another ball. (Ramsay and Silverman, 2002; Ramsay et al., 2014)

Considerable pre-processing was done before we received the data. The data were lightly smoothed to fill in missing values. Furthermore, the data were centred, rotated and trimmed (Ramsay et al., 2014). Centring and rotation was done in the following way:

A coordinate system was defined by smoothing chest coordinates while preserving gentle and slow changes in chest position and orientation. Large-scale upper body movements were removed by averaging the three smoothed chest coordinates at each time point to create a mean chest curve. This curve was subtracted from right index finger coordinates. Next each coordinate was zero centred by subtracting the mean of the coordinate over the entire trial (all position measurements from different body parts). The coordinates were rotated (Figure 6.2) in such a way that coordinate 1 displays the greatest variation in the horizontal plane and corresponds mainly to lateral movement across the body plane, with zero at the body midline and from the viewer's standpoint positive moment is to the left. The second rotated coordinate reflected mostly forward-backward movement with forward corresponding to positive values. The third coordinate, in the vertical direction, was left unchanged. (Ramsay et al., 2014; Ramsay and Gribble, 1999)

In Paper III we considered the pre-processed data described above. We registered the juggling trials to allow comparison among trials, and among cycles within trials. Our approach is to estimate a warping function for each trial and then optimise the fit of the

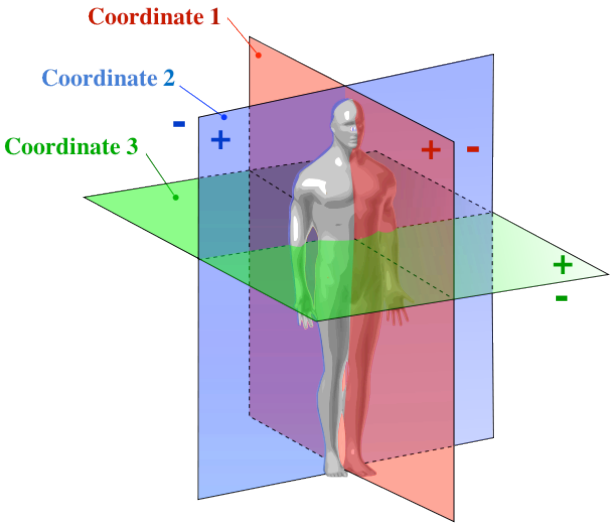

Figure 6.2: Direction of the three Cartesian coordinates for the juggling data, as defined by Ramsay et al. (2014). The point of origin is dependent on the data. Diagram adapted from Richfield (2014)

warped trial to an idealised model. We suggest that the appropriate way to represent phase variation in the juggling system, is to decompose it into two periodic components, where one periodic component is of approximate constant length. In this way, we incorporate the physical constraints of the biomechanic system, namely regular joint movement and fixed length of a limb, into the registration process.

To create an ideal curve to warp to, we conceptualised an electromechanical juggling robot as a basic mathematical model of human juggling. The model consists of a periodic joint movement and a periodic position vector (from the joint to the 'fingertip'). The position vector has approximately constant length along the observed trajectory.

For each trial we followed the following procedure:
We used a well-known idea of warping cycles towards each other and then obtained a periodic average of the warped cycles by projection onto a high-dimensional space of periodic functions. The periodic average was then decomposed according to the idealised model of juggling: a periodic joint movement and a periodic deviance (i.e. the position vector) from the periodic average, subject to the deviance having approximately constant length along the trajectory.

In more general terms, our approach consists of the following steps:

1. Define the class of idealised average juggling cycles for an imagined 'juggling robot'
2. Define a class of warping functions
3. For each warping function, w, compute the average, f_{per}, of the warped function $f(w(t))$ over the juggling cycles, then compute a measure of the deviance between f_{per} and the best function from the idealised model of average juggling cycles
4. Estimate the warping function by minimising the deviance measure in step 3 over the class of warping functions in step 2

We did not address the problem of optimisation w over a class of warping functions, which admittedly will cause problems for the current implementation.

We demonstrated that the ten juggling trials can be registered in such a way that the structural average over all cycles is appropriately described by the idealised model.

Apart from addressing the challenge of functional registration subject to constraints, the solution suggested in Tolver et al. (2014) also addressed the challenge of finding a more suitable coordinate system for the juggling data. No natural coordinate system exists for the juggling data. The Cartesian coordinate system is merely convenient. Finger and wrist movements are influenced by variation of the angle at the elbow, the angle at the shoulder and movements of the body. Therefore it is likely that the coordinate system of the juggling data varies with time over the duration of a trial (Ramsay et al., 2014), i.e. the relevant coordinate system can move around. This complicates the registration problem. A coordinate system that is not fixed and takes the constraints and mechanics of the human body into account will possibly be more meaningful and give better results

(Ramsay and Gribble, 1999). We established an elliptical coordinate system on an arbitrary plane, subject to biomechanical constraints of the human body (Tolver et al., 2014).

The idea that physical constraints should be taken into account in registration has broader application than only registration over time and the juggling data. Apart from registration over time, curves can be registered over other measures like distance or, for example, chemical shift in the case of NMR data (Sections 5.1 and 5.6). The physical constraints can be anatomical and biomechanical, as in the case of the juggling data. In the case of NMR spectra, the constraints relate to peak shape (Lorentizian) (Section 5.1), area under a peak and minimum detectable peak width based on the specifications of the spectrometer. If the Lorentzian shape of NMR peaks are not preserved during registration, inherent structure in the data could be destroyed and the heart shapes in phase-plane plots (see Muller and Ramsay (2014)) could disappear. Constraints should be carefully considered in any registration procedure and should be modelled in a way that does not destroy the structure of the data and that results in interpretable parameters.

7

Conclusions and Perspectives

This thesis describes the use of functional data analysis as a method to analyse spectral data in metabolomics. Functional data analysis takes advantage of the smoothness underlying the data which are in the form of curves. There is no requirement to assume that adjacent data points on a curve are independent. These are distinct advantages over many standard chemometric methods. Additionally, functional data analysis can unlock the information hidden in the derivatives of spectra.

The motivating example focused on NMR data from a human metabolomics study, specifically a diet standardisation study. However, the methods are also applicable to other spectral data, like mass spectrometry or infrared. The data are not required to originate from a human metabolomics study, but can come from, for example, plant metabolomics. The applications also range much wider than only diet standardisation studies and may include diet or other clinical interventions.

To summarise the results: In Paper I we successfully applied wavelet-based functional mixed models together with bootstrap-based inference on functions to detect the influence of covariates on specific metabolites or spectral regions. To our knowledge, this is the first time that wavelet-based functional mixed models have been applied in metabolomics. In Paper II we illustrated the rich nature of functional derivatives in simulated nuclear magnetic peaks with characteristic Lorentzian line shapes. Using phase-plane plots to explore the anatomy of NMR peaks, we introduced the novelty of heart plots for spectral data. In Paper III we applied functional registration in the context of biomechanics, specifically to data from a juggling experiment. The novelty of this work is that the registration is done towards an idealised biomechanical model. In this way, the warping is performed subject to biomechanical constraints. Additionally, Paper IV, demonstrated the value of classical mixed model methodology in the context of targeted metabolomics.

In order to build a stronger connection between the worlds of statistics and chemometrics, we gave a glimpse of the essential and complex data pre-processing that is well known to chemometricians, but is generally unknown to statisticians. We also touched on the important aspect of registration, also called warping or alignment, which emerges from both the chemometric and statistical perspectives.

We offer a number of perspectives on future work.

A natural next step in terms of application would be to analyse the DiOGenes dietary intervention data using wavelet-based functional mixed models. This study investigated

the effect of diets with differential protein levels and glycaemic index loads on obese and overweight individuals over six months. It will be interesting to compare these results with published results which used more standard methodology.

We have some concerns related to the influence of 'residual misalignment' in terms of data that could be 'well enough' aligned for standard chemometric analysis. We suspect that wavelet-based functional mixed models are sensitive to very small misalignments, especially in the base of peaks and the valleys between peaks. We plan to investigate this potential sensitivity of our method to small perturbations in alignment on simulated data.

On a methodological level there is a need to investigate subset selection of wavelet coefficients for input to mixed modelling, together with the selection of primary resolution. The abovementioned issue of sensitivity to small misalignments is also relevant in this context. The use of alignment methods from the functional data analysis literature, specifically k-means clustering and alignment, as well as simultaneous alignment and modelling may provide interesting avenues to investigate in the context of NMR metabolomics spectra.

Hearts plots provide a new way to view spectral peaks and the possibilities for future analyses are exciting: principal component analysis of hearts and hypothesis testing of hearts would be only the first statistical steps in exploring the anatomy of spectral hearts.

With regards to registration subject to constraints, it would be of interest to extend the procedure to include the optimisation of each individual warping function over a class of warping functions.

Overall this thesis gives an indication of the huge possibilities for functional data analysis in metabolomics and chemometrics. Spectral data are inherently functional in nature. Functional data analysis provides access to many functional equivalents of methods currently used in chemometrics, with the benefits of no strong assumptions regarding neighbouring observations. Functional data analysis also provides access to the data's derivatives and opens up the ability to analyse information that is otherwise locked away in the data.

On a health research level, nutritional metabolomics shows great potential for the discovery of novel biomarkers of food consumption, personal nutritional status and metabolic phenotype. The use of functional data analysis in metabolomics can make a valuable contribution to the emerging technology in personalised medicine and health care, including personalised nutrition for prevention and treatment.

Bibliography

Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics & Data Analysis*, 22(4):351–361.

Alsberg, B. K. (1993). Representation of spectra by continuous functions. *Journal of Chemometrics*, 7(3):177–193.

Antoniadis, A., Bigot, J., and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: A comparative simulation study. *Journal of Statistical Software*, 6(6):1–83.

Astarita, G. and Langridge, J. (2013). An emerging role for metabolomics in nutrition science. *Journal of Nutrigenetics and Nutrigenomics*, 6(4-5):179–198.

Astle, W., De Iorio, M., Richardson, S., Stephens, D., and Ebbels, T. (2012). A bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, 107(500):1259–1271.

Åstrand, M. (2003). Contrast normalization of oligonucleotide arrays. *Journal of Computational Biology*, 10(1):95–102.

Barding, G. A., J., Salditos, R., and Larive, C. K. (2012). Quantitative NMR for bioanalysis and metabolomics. *Analytical and Bioanalytical Chemistry*, 404(4):1165–79.

Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11):2692–2703.

Berk, M., Ebbels, T., and Montana, G. (2011). A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics*, 27(14):1979–85.

Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R., and Buydens, L. M. C. (2013). Warping methods for spectroscopic and chromatographic signal alignment: A tutorial. *Analytica Chimica Acta*, 781(0):14–32.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., Bjorndahl, T. C., Krishnamurthy, R., Saleem, F., Liu, P., Dame, Z. T., Poelzer, J., Huynh, J., Yallou, F. S., Psychogios, N., Dong, E., Bogumil, R., Roehring, C., and Wishart, D. S. (2013). The human urine metabolome. *PLoS ONE*, 8(9):e73076.

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

Brennan, L. (2008). Session 2: Personalised nutrition metabolomic applications in nutritional research. *Proceedings of the Nutrition Society*, 67(04):404–408.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., and Holmes, E. (2005). Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry*, 77(5):1282–1289.

Coombes, K. R., Fritsche, H. A., Clarke, C., Chen, J.-N., Baggerly, K. A., Morris, J. S., Xiao, L.-C., Hung, M.-C., and Kuerer, H. M. (2003). Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clinical Chemistry*, 49(10):1615–1623.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C., and Kuerer, H. M. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., and Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267.

Dalskov, S.-M., Muller, M., Ritz, C., Damsgaard, C. T., Papadaki, A., Saris, W. H., Astrup, A., Michaelsen, K. F., and Mølgaard, C. (2014). Effects of dietary protein and glycemic index on biomarkers of bone turnover in children. *British Journal of Nutrition*, 111(7):1253–1262.

Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61. Society for Industrial and Applied Mathematics.

Daviss, B. (2005). Growing pains for metabolomics. *The Scientist*, 19(8):25–28.

De Rooi, J. J. and Eilers, P. H. C. (2012). Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, 117(0):56–60.

Dettmer, K. and Hammock, B. D. (2004). Metabolomics–a new exciting field within the "omics" sciences. *Environmental Health Perspectives*, 112(7):A396.

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1H NMR metabonomics. *Analytical Chemistry*, 78(13):4281–4290.

Dong, J., Cheng, K.-K., Xu, J., Chen, Z., and Griffin, J. L. (2011). Group aggregating normalization method for the preprocessing of NMR-based metabolomic data. *Chemometrics and Intelligent Laboratory Systems*, 108(2):123–132.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):301–369.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, 12(1):111–140.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl 1):S105–S110.

Ebbels, T. M., Lindon, J. C., and Coen, M. (2011). *Processing and modeling of nuclear magnetic resonance (NMR) metabolic profiles*, pages 365–388. Springer.

Echeverry, G., Hortin, G. L., and Rai, A. J. (2010). *Introduction to Urinalysis: Historical Perspectives and Clinical Application*, volume 641 of *Methods in Molecular Biology*, chapter 1, pages 1–12. Springer.

Eilers, P. H. C. (2003). A perfect smoother. *Analytical Chemistry*, 75(14):3631–3636.

Eilers, P. H. C. and Boelens, H. F. M. (2005). Baseline correction with asymmetric least squares smoothing. Technical report, Leiden University Medical Centre.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with *b*-splines and penalties. *Statistical Science*, 11(2):89–102.

Eknoyan, G. (2007). Looking at the urine: The renaissance of an unbroken tradition. *American Journal of Kidney Diseases*, 49(6):865–872.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., and Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50(0):96–106.

Fan, T. W. M. and Lane, A. N. (2008). Structure-based profiling of metabolites and isotopomers by NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 52(2–3):69–117.

Friedrichs, M. (1995). A model-free algorithm for the removal of baseline artifacts. *Journal of Biomolecular NMR*, 5(2):147–153.

Gan, F., Ruan, G., and Mo, J. (2006). Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82(1–2):59–65.

Gauss, C. (1809). *Theoria motus corporum celestium*. Perthes et Besser, Hamburg.

Gavaghan, C. L., Holmes, E., Lenz, E., Wilson, I. D., and Nicholson, J. K. (2000). An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the c57bl10j and alpk:apfcd mouse. *FEBS Letters*, 484(3):169–174.

Giskeodegard, G. F., Bloemberg, T. G., Postma, G., Sitter, B., Tessem, M. B., Gribbestad, I. S., Bathen, T. F., and Buydens, L. M. (2010). Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Analytica Chimica Acta*, 683(1):1–11.

Goodacre, R., Broadhurst, D., Smilde, A., Kristal, B., Baker, J. D., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., Ebbels, T., Kell, D., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjöström, M., Trygg, J., and Wulfert, F. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3):231–241.

Hall, P. and Penev, S. (2001). Cross-validation for choosing resolution level for nonlinear wavelet curve estimators. *Bernoulli*, 7(2):317–341.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer series in Statistics. Springer, New York, 2nd edition.

Hedrick, V., Dietrich, A., Estabrooks, P., Savla, J., Serrano, E., and Davy, B. (2012). Dietary biomarkers: advances, limitations and future directions. *Nutrition Journal*, 11(1):109.

Heinzmann, S. S., Merrifield, C. A., Rezzi, S., Kochhar, S., Lindon, J. C., Holmes, E., and Nicholson, J. K. (2011). Stability and robustness of human metabolic phenotypes in response to sequential food challenges. *Journal of Proteome Research*, 11(2):643–655.

Holmes, E., Foxall, P. J. D., Nicholson, J. K., Neild, G. H., Brown, S. M., Beddell, C. R., Sweatman, B. C., Rahr, E., Lindon, J. C., Spraul, M., and Neidig, P. (1994). Automatic data reduction and pattern recognition methods for analysis of 1h nuclear magnetic resonance spectra of human urine from normal and pathological states. *Analytical Biochemistry*, 220(2):284–296.

Hore, P. and Compton, R. (1995). *Nuclear Magnetic Resonance: Oxford Chemistry Primers*. Oxford University Press, New York.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl 1):S96–S104.

Jenab, M., Slimani, N., Bictash, M., Ferrari, P., and Bingham, S. (2009). Biomarkers in nutritional epidemiology: applications, needs and new horizons. *Human Genetics*, 125(5-6):507–525.

Keun, H. C., Ebbels, T., Antti, H., Bollard, M. E., Beckonert, O., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2003). Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Analytica chimica acta*, 490(1):265–276.

Kim, S. B., Wang, Z., Oraintara, S., Temiyasathit, C., and Wongsawat, Y. (2008). Feature selection and classification of high-resolution NMR spectra in the complex wavelet transform domain. *Chemometrics and Intelligent Laboratory Systems*, 90(2):161–168.

Kneen, M. A. and Annegarn, H. J. (1996). Algorithm for fitting xrf, sem and pixe x-ray spectra backgrounds. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 109–110(0):209–213.

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283.

Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R., and Gronwald, W. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8:146–160.

Komsta, L. (2011). Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia*, 73(7-8):721–731.

Kouba, E., Wallen, E. M., and Pruthi, R. S. (2007). Uroscopy by hippocrates and theophilus: Prognosis versus diagnosis. *The Journal of Urology*, 177(1):50–52.

Lavine, B. and Workman, J. (2006). Chemometrics. *Analytical Chemistry*, 78(12):4137–4145.

Lavine, B. and Workman, J. (2008). Chemometrics. *Analytical Chemistry*, 80(12):4519–4531.

Lavine, B. and Workman, J. (2010). Chemometrics. *Analytical Chemistry*, 82(12):4699–4711.

Lavine, B. and Workman, J. J. (2004). Chemometrics. *Analytical Chemistry*, 76(12):3365–3372.

Lavine, B. K. (1998). Chemometrics. *Analytical Chemistry*, 70(12):209–228.

Lavine, B. K. (2000). Chemometrics. *Analytical Chemistry*, 72(12):91–98.

Lavine, B. K. and Workman, J. (2002). Chemometrics. *Analytical Chemistry*, 74(12):2763–2770.

Lavine, B. K. and Workman, J., J. (2013). Chemometrics. *Analytical Chemistry*, 85(2):705–14.

Legendre, A. (1805). *Nouvelles Methodés pur la Détermination des Orbites des Cométes*. Courcier, Paris.

Lenz, E. M. and Wilson, I. D. (2006). Analytical strategies in metabonomics. *Journal of Proteome Research*, 6(2):443–458.

Levitin, D. J., Nuzzo, R. L., Vines, B. W., and Ramsay, J. O. (2007). Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne*, 48(3):135–155.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36.

Lieber, C. A. and Mahadevan-Jansen, A. (2003). Automated method for subtraction of fluorescence from biological raman spectra. *Applied Spectroscopy*, 57(11):1363–1367.

Liland, K. H. (2011). Multivariate methods in metabolomics–from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry*, 30(6):827–841.

Liland, K. H., Almøy, T., and Mevik, B.-H. (2010). Optimal choice of baseline correction for multivariate calibration of spectra. *Applied Spectroscopy*, 64(9):1007–1016.

Llorach, R., Garcia-Aloy, M., Tulipani, S., Vazquez-Fresno, R., and Andres-Lacueva, C. (2012). Nutrimetabolomic strategies to develop new biomarkers of intake and health effects. *Journal of Agricultural and Food Chemistry*, 60(36):8797–8808.

MacKinnon, N., Ge, W., Khan, A. P., Somashekar, B. S., Tripathi, P., Siddiqui, J., Wei, J. T., Chinnaiyan, A. M., Rajendiran, T. M., and Ramamoorthy, A. (2012). Variable reference alignment: an improved peak alignment protocol for NMR spectral data with large intersample variation. *Analytical Chemistry*, 84(12):5372–9.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693.

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H., and Patil, P. (1998). Exact risk analysis of wavelet regression. *Journal of Computational and Graphical Statistics*, 7(3):278–309.

Martin, F.-P. J., Rezzi, S., Peré-Trepat, E., Kamlage, B., Collino, S., Leibold, E., Kastler, J., Rein, D., Fay, L. B., and Kochhar, S. (2009). Metabolic effects of dark chocolate consumption on energy, gut microbiota, and stress-related metabolism in free-living subjects. *Journal of Proteome Research*, 8(12):5568–5579.

McNiven, E. M. S., German, J. B., and Slupsky, C. M. (2011). Analytical metabolomics: nutritional opportunities for personalized health. *The Journal of nutritional biochemistry*, 22(11):995–1002.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64(2):479–89.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):179–199.

Muller, M. and Ramsay, J. O. (2014). Heart plots for spectral data.

Muller, M. and Tolver, A. (2014). Analysis of nutri-metabolomics NMR-spectra using wavelet-based functional mixed models.

Nason, G. (2013). *wavethresh: Wavelets statistics and transforms*. R package version 4.6.6.

Nason, G. P. (2008). *Wavelet methods in statistics with R*. Springer, New York.

Nicholson, J. K., Holmes, E., Kinross, J. M., Darzi, A. W., Takats, Z., and Lindon, J. C. (2012). Metabolic phenotyping in clinical and surgical environments. *Nature*, 491(7424):384–392.

Nicholson, J. K. and Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature*, 455(7216):1054–1056.

Ogden, R. T. (1997a). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhauser, Boston.

Ogden, R. T. (1997b). On preconditioning the data for the wavelet transform when the sample size is not a power of two. *Communications in Statistics - Simulation and Computation*, 26(2):467–486.

O'Gorman, A., Gibbons, H., and Brennan, L. (2013). Metabolomics in the identification of biomarkers of dietary intake. *Computational and Structural Biotechnology Journal*, 4:e201301004.

O'Sullivan, A., Gibney, M. J., and Brennan, L. (2011). Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies. *The American journal of clinical nutrition*, 93(2):314–321.

Pardalidis, N., Kosmaoglou, E., Diamantis, A., and Sofikitis, N. (2008). Uroscopy in byzantium (330–1453 ad). *The Journal of Urology*, 179(4):1271–1276.

Parsons, H., Ludwig, C., Gunther, U., and Viant, M. (2007). Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, 8(1):234.

Percival, D. B. and Walden, A. T. (2006). *Wavelet methods for time series analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Potischman, N. (2003). Biologic and methodologic issues for nutritional biomarkers. *The Journal of Nutrition*, 133(3):875S–880S.

Primrose, S., Draper, J., Elsom, R., Kirkpatrick, V., Mathers, J. C., Seal, C., Beckmann, M., Haldar, S., Beattie, J. H., Lodge, J. K., Jenab, M., Keun, H., and Scalbert, A. (2011). Metabolomics and human nutrition. *British Journal of Nutrition*, 105(08):1277–1283.

Ramsay, J. and Gribble, P. (1999). Functional data analysis in action. In *Proceedings of the American Statistical Association*, pages 30–36.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer-Verlag, New York, 2nd edition.

Ramsay, J. O., Gribble, P., and Kurtek, S. (2014). Functional data analysis of juggling trajectories: Data description and processing. *Special Section, Electronic Journal of Statistics*.

Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. Springer, New York.

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.

Rezzi, S., Ramadan, Z., Fay, L. B., and Kochhar, S. (2007). Nutritional metabonomics:? applications and perspectives. *Journal of Proteome Research*, 6(2):513–525.

Richfield, D. (2014). Anatomical planes in a human [figure]. http://en.wikipedia.org/wiki/Anatomical_plane [Online; accessed 20 September 2014].

Rubio-Aliaga, I., Kochhar, S., and Silva-Zolezzi, I. (2012). Biomarkers of nutrient bioactivity and efficacy: A route toward personalized nutrition. *Journal of Clinical Gastroenterology*, 46(7):545–554.

Ruckstuhl, A. F., Jacobson, M. P., Field, R. W., and Dodd, J. A. (2001). Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193.

Saeys, W., De Ketelaere, B., and Darius, P. (2008). Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics*, 22(5):335–344.

Sakia, R. M. (1992). The box-cox transformation technique: A review. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(2):169–178.

Savorani, F., Rasmussen, M. A., Mikkelsen, M. S., and Engelsen, S. B. (2013). A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. *Food Research International*, 54(1):1131–1145.

Savorani, F., Tomasi, G., and Engelsen, S. B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202.

Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B., van Ommen, B., Pujos-Guillot, E., Verheij, E., Wishart, D., and Wopereis, S. (2009). Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5(4):435–458.

Skov, T., van den Berg, F., Tomasi, G., and Bro, R. (2006). Automated alignment of chromatographic data. *Journal of Chemometrics*, 20(11-12):484–497.

Smilde, A. K., van der Werf, M. J., Bijlsma, S., van der Werff-van der Vat, B. J., and Jellema, R. H. (2005). Fusion of mass spectrometry-based metabolomics data. *Analytical chemistry*, 77(20):6729–6736.

Smolinska, A., Blanchet, L., Buydens, L. M., and Wijmenga, S. S. (2012). NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Analytica Chimica Acta*, 750:82–97.

Sørensen, H., Goldsmith, J., and Sangalli, L. M. (2013). An introduction with medical applications to functional data analysis. *Statistics in Medicine*, 32(30):5222–5240.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.

Steve (2014). Juggling instructions - illustrated juggling tutorial [figure]. http://juggling-for-beginners.com/how-to-juggle/juggling-instructions/ [Online; accessed 20 September 2014].

Tolver, A., Sørensen, H., Muller, M., and Mousavi, S. N. (2014). Analysis of juggling data: Registration subject to biomechanical constraints. *Electronic Journal of Statistics*, 8(2):1856–1864.

Torgrip, R. J. O., Åberg, K. M., Alm, E., Schuppe-Koistinen, I., and Lindberg, J. (2008). A note on normalization of biofluid 1D 1H-NMR data. *Metabolomics*, 4(2):114–121.

Torgrip, R. O., Alm, E., and Åberg, K. M. (2010). Warping and alignment technologies for inter-sample feature correspondence in 1D H-NMR, chromatography-, and capillary electrophoresis-mass spectrometry data. *Bioanalytical Reviews*, 1(2-4):105–116.

Trygg, J., Holmes, E., and Lundstedt, T. (2006). Chemometrics in metabonomics. *Journal of Proteome Research*, 6(2):469–479.

Van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., and van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7(1):142.

Van der Greef, J. and Smilde, A. K. (2005). Symbiosis of chemometrics and metabolomics: past, present, and future. *Journal of Chemometrics*, 19(5-7):376–386.

Veselkov, K. A., Lindon, J. C., Ebbels, T. M. D., Crockford, D., Volynkin, V. V., Holmes, E., Davies, D. B., and Nicholson, J. K. (2009). Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Analytical Chemistry*, 81(1):56–66.

Vidakovic, B. (1999). *Statistical modeling by wavelets*. John Wiley & Sons, New York.

Voswinckel, P. (2000). From uroscopy to urinalysis. *Clinica Chimica Acta*, 297(1–2):5–16.

Vu, T. N. and Laukens, K. (2013). Getting your peaks in line: A review of alignment methods for NMR spectral data. *Metabolites*, 3(2):259–276.

Walker, J. S. (2008). *A primer on wavelets and their scientific applications*. CRC press, Boca Raton, 2nd edition.

Walsh, M. C., Brennan, L., Malthouse, J. P. G., Roche, H. M., and Gibney, M. J. (2006). Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. *The American Journal of Clinical Nutrition*, 84(3):531–539.

Wehrens, R. (2011). *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Use R! Springer, New York.

Wishart, D. S. (2007). Current progress in computational metabolomics. *Briefings in Bioinformatics*, 8(5):279–293.

Wishart, D. S. (2009). *Computational Approaches to Metabolomics*, volume 593 of *Methods in Molecular Biology*, chapter 14, pages 283–313. Humana Press.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., and Scalbert, A. (2013). Hmdb 3.0—the human metabolome database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807.

Wittern-Sterzel, R. (1999). Diagnosis: the doctor and the urine glass. *The Lancet*, 354(suppl):SIV13.

Wold, S. (1995). *PLS for multivariate linear modelling*, volume 2, chapter Multivariate Data Analysis of Chemical and Biological Data, page 201. Verlag Chemie, Weinheim, Germany.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H.-H., Nielsen, C., Brunak, S., and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in dna microarray experiments. *Genome biol*, 3(9):1–16.

Xi, Y. and Rocke, D. M. (2008). Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, 9:1–10.

Zeisel, S. H., Waterland, R. A., Ordovás, J. M., Muoio, D. M., Jia, W., and Fodor, A. (2013). Highlights of the 2012 research workshop: Using nutrigenomics and metabolomics in clinical nutrition research. *Journal of Parenteral and Enteral Nutrition*, 37(2):190–200.

Zhang, J.-T. (2013). *Analysis of Variance for Functional Data*. Chapman and Hall, London.

Zhang, S., Gowda, G. N., Ye, T., and Raftery, D. (2010). Advances in NMR-based biofluid analysis and metabolite profiling. *Analyst*, 135(7):1490–1498.

Zhang, S., Zheng, C., Lanza, I. R., Nair, K. S., Raftery, D., and Vitek, O. (2009). Interdependence of signal processing and analysis of urine 1H NMR spectra for metabolic profiling. *Analytical Chemistry*, 81(15):6080–6088.

Zhang, Z. M., Liang, Y. Z., Lu, H. M., Tan, B. B., Xu, X. N., and Ferro, M. (2012). Multiscale peak alignment for chromatographic datasets. *Journal of Chromatography A*, 1223:93–106.

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012). Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617.

Zivkovic, A. M. and German, J. B. (2009). Metabolomics for assessment of nutritional status. *Current Opinion in Clinical Nutrition & Metabolic Care*, 12(5):501–507.

Papers

I

Analysis of Nutri-metabolomics NMR-spectra using Wavelet-Based Functional Mixed Models

Martha Muller and Anders Tolver
Department of Mathematical Sciences
University of Copenhagen

Publication details

Manuscript (in preparation for submission).

Analysis of Nutri-metabonomics NMR-Spectra using Wavelet-Based Functional Mixed Models

Martha Muller, Anders Tolver

Department of Mathematical Sciences
University of Copenhagen
e-mail: muller.martie@gmail.com; tolver@math.ku.dk

Abstract: In this article we apply wavelet-based functional mixed model methodology to analyse nuclear magnetic resonance spectrometry data. The application is a diet standardisation study in human nutrition metabolomics, where participants provided three repeated measurements. We use bootstrap-based inference to estimate the difference in means between groups in the longitudinal functional model. This approach allows us to respect the study design, while modelling the NMR spectra as functions. We model nonparametric fixed and random effect functions that enable us to incorporate covariates and repeated measurements in one model. We investigate NMR spectral regions that are significantly different for gender and diet culture groups.

Keywords and phrases: functional data analysis, functional mixed model, metabolomics, nuclear magnetic resonance, nutrition, wavelets.

1. Introduction

Nutrition plays a crucial role in preventing disease and, therefore, in establishing and maintaining an acceptable overall population health status. In contrast to many other areas of health care, nutritional therapy is fairly easy and inexpensive. It can potentially be used for preventive and therapeutic treatment of many diseases. In most western countries, obesity, type-2 diabetes, cardiovascular disease and cancer are affecting a growing number of people. Although these diseases have been studied extensively, much remains unknown regarding the direct and indirect impact of nutritional interventions on health status at the individual as well as the population level. There is reason to believe that this shortcoming in knowledge potentially has a major socio-economic impact, specifically in the context of constantly growing health care budgets (Moore et al., 2000; Biel, Evans and Clarke, 2009).

Our level of understanding as well as our means of monitoring dietary exposure have changed in a revolutionary way over the last decades. This field is embedded in a multidisciplinary context bringing biochemistry, human nutrition, preventive medicine, systems biology, bioinformatics and statistics together. Emerging high-throughput screening techniques, like metabolomics, have played a crucial role in initiating this transition. In parallel to high-throughput screening techniques within the 'omics', new data analytic concepts and methods are

key factors in the pursuit of extracting useful knowledge from metabolomics data (Wishart, 2007). However, there are two major challenges. First, large amounts of noisy data that are rapidly amassed and, second, the complexity and dynamics of connected and interrelated data structures. Recent technological improvements in mass spectrometry and nuclear magnetic resonance spectroscopy have led to greatly enhanced experimental capabilities in metabolomics. However, the development of better adapted statistical concepts and tools for pre-processing and data analysis in this field has been identified as a major bottleneck. The elimination of this bottleneck will add the final link in the technology infrastructure and will allow the rapid enhancement of general nutrition knowledge and health status. Furthermore, it will be a step towards personalised dietary monitoring and personalised nutritional interventions.

The influx of metabolomics in human nutrition is based on the understanding that metabolomics offers a powerful approach for reconstructing dietary influences on biological systems Favé et al. (2009). Several authors have emphasised that metabolomics offers a truly holistic perception and mode of thinking of biological processes (Kell, 2004; Quackenbush, 2007; Wishart, 2007). In terms of both biological understanding and technological improvement, metabolomics has resulted in a landslide of important results (e.g. Holmes, Wilson and Nicholson (2008); Favé et al. (2009)). The potential of metabolomics in the context of human nutrition has been well-established in the literature over the last decade (Whitfield, German and Noble, 2004; Goodacre et al., 2004; Gibney et al., 2005). However, in terms of statistical methodology inference methods often produce simplistic per-metabolite conclusions that are more exploratory than confirmatory and that are difficult to combine into an overall characterisation of metabolomic status.

We present a wavelet-based mixed-model approach for handling high-dimensional data while respecting the study design in the process of dimension reduction. The key idea is to approach wavelet regression methods from a functional data analytic perspective (e.g. Ramsay (2002, 2005)) and to re-cast the entire statistical methodology in a framework ideal for metabolomics data analysis. Wavelets were introduced into statistics in the 1990s (Nason, 1996; Johnstone and Silverman, 1997) almost exclusively with applications in time series analysis in mind but since then they have disappeared out of the mainstream of statistical methods and had a little renaissance related to proteomics (Morris et al., 2003, 2008), electrophysiology (Davidson, 2009; Pigoli and Sangalli, 2012), human vision (Ogden and Greene, 2010) and transcriptome analysis (Clement et al., 2012). In order to re-cast the wavelet theory in a useful statistical framework, mixed model methodology will be extended to the nutritional metabolomics data structures. Related extensions of mixed models concepts to other types of complex data structures were considered by Guo (2002); Qin and Guo (2006); Morris (2006) and Scheipl, Staicu and Greven (2014).

Although our main purpose is to apply the methodology in a nutritional metabolomics context, we also provide some methodological advances.

In section 2 we describe the data from a human nutritional metabolomics study which motivated this work. Section 3 covers the wavelet-based functional

mixed model. Bootstrap-based inference follows in section 4. Details on the implementation and application of the methodology to the example data are covered in sections 5 and 6 respectively. Results of the analysis of the example data are in the last section before the discussion.

2. Motivating example

The diet standardisation study investigated the effect of a specified daily meal composition on the human urine metabolome over three days (Rasmussen et al. (2011), study B).

The study included 16 healthy non-smoking human subjects, aged 22 to 39 years. Participants were of Danish or Italian nationality. Table 2 provides a summary of the number of subjects by gender and nationality. Nationality was used as an indicator of diet culture.

The diet was standardised, i.e. the composition of the three daily meals was fixed. However, the amount to be consumed was only fixed for certain food items (tuna, mackerel, salmon fillet, smoked pork, broccoli, onion, red pepper, tomato, pesto, raisins, almonds and dark chocolate), and the amount of other food items could be freely adjusted (oatmeal, soya milk, rye bread, white bread, pasta, apple, banana, orange, water, black tea, sugar, salt and pepper). Participants collected all food, pre-weighted and packed, from the research site. The details of the standardised diet are described elsewhere (Rasmussen et al., 2011). Activity level among participants and days were standardised by prohibition of any strenuous physical activity. All 16 participants collected 24-hour urine samples (from 08:00 till 08:00 the next day) on the three consecutive days while they were on the standardised diet.

The diet standardisation study was a pilot study for the Diet, Obesity, and Genes (DiOGenes) dietary intervention trial (Larsen et al., 2010; Aston et al., 2010; Moore et al., 2010). The objective of the diet standardisation study was to investigate the effect of a three-day standardised diet on urine metabolomics, but also to investigate the influence of other factors like diet culture and gender.

2.1. Pre-processing of ^{1}H NMR data from urine samples

The urine samples were analysed by ^{1}H NMR spectroscopy at 500.13 MHz. Technical details on the acquisition of the spectra can be found in Rasmussen et al. (2011). The resulting spectra were referenced to the TSP (3-(TrimethylSilyl)-Propionic acid-d_4) peak at 0.00 ppm and automatically baseline corrected using TopSpin^{TM} (Bruker BioSpin), software related to the spectrometer.

TABLE 1
Diet standardisation study: number of participants

	Female	Male	Total
Danish	7	3	10
Italian	2	4	6
Total	9	7	16

The pre-processing steps for the data are described in Rasmussen et al. (2011). In short, the NMR regions (15.21 to 9.20 ppm, 6.34 to 4.09 ppm and 0.62 to -5.61 ppm) were removed due to being noise-only regions or due to the strong influence of the residual water peak. Spectra were aligned using the interval-based *i*coshift algorithm (Savorani, Tomasi and Engelsen, 2010). Spectra were normalised according to the sum of the squared value of all variables for the given sample (2-Norm) (Craig et al., 2006; Dieterle et al., 2006). Each individual spectrum consisted of 19 930 values after pre-processing.

2.2. Presentation of the data

A typical NMR spectra from the diet standardisation study is displayed in Figure 1, first with the original relative intensity scale on the y-axis (top) and then after variance stabilisation normalisation (VSN) transform, on the generalised logarithm (base 2) scale ($glog_2$) (below). Chemical shift (on the x-axis) is conventionally called δ, measured in parts per million (ppm) and ordered from large to small. Here (δ) runs from approximately 9 parts per million (ppm) to 1 ppm, with the region between 6.5 and 4.1 ppm removed. Many smaller peaks that are not visible on the relative intensity scale become clearly visible on the $glog_2$ scale. We visually inspect the dependence of standard deviation (or variance) on the mean, over all samples (per chemical shift value), before and after VSN transform (Figure 2). The red dots are the running median (window-width 10%). An approximate horizontal line of red dots would indicate no variance-mean dependence, and thus variance stabilisation. The VSN transform removes the dependence of the variance on the mean.

Figure 3 displays three daily spectra (blue, red, green), on the $glog_2$ scale, for each of four individuals. A number of features are noticeable: there is no clear pattern from day to day across the four subjects, there is daily variation within each subject, as well as variation between subjects. It is unclear how much of the variation is due to gender and/or diet culture (nationality) effects.

A simple approach would be, for each day separately, to treat the relative intensity at each ppm-point on the chemical shift axis as a dependent variable and use a linear model containing gender, nationality and their interaction as predictors. Thus, each daily model contains tens of thousands of linear models - one linear model at each ppm-value. This approach provides predicted spectra (relative intensity at each ppm-value) per gender-nationality group for each of the three days separately (Figure 4, left column). The variance in the data, that cannot be explained by gender and nationality, is represented by the residual errors. The estimated standard deviation of the residual errors of the three daily models (Figure 4, right column) reveal a number of aspects: large variance in residual errors is not necessarily at chemical shift positions where there are large peaks (this is due to the VSN transform's variance stabilisation), per day there are certain patterns (peaks) in the residual error variances that match groups of peaks in the estimated means (left column), e.g. doublets in 3.90 to 3.85 ppm, as well as 3.65 to 3.60 ppm. Related to diet standardisation over the three days

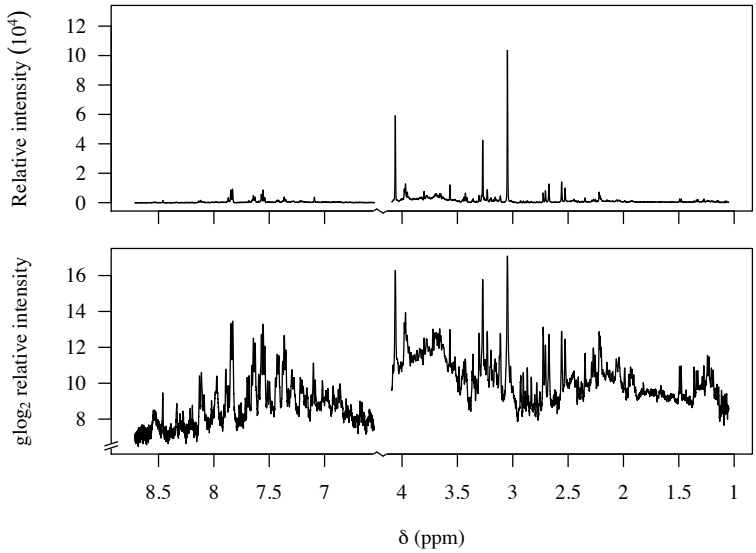

FIG 1. *An NMR spectrum from the diet standardisation study: chemical shift (δ) in parts per million (ppm) and NMR relative intensity (top), VSN transformed relative intensity on the generalised logarithmic (base 2) scale (bottom)*

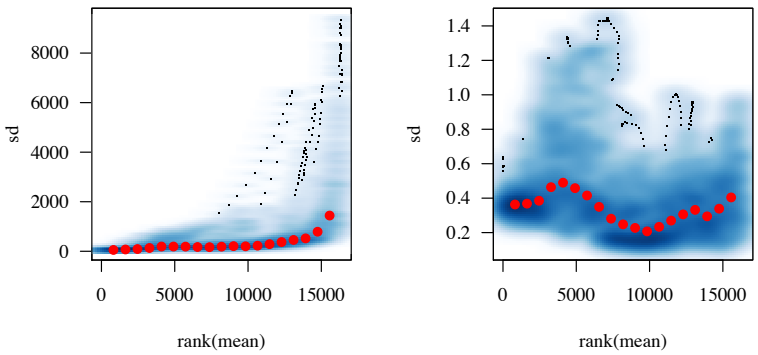

FIG 2. *Standard deviation (sd) versus ranked mean for all NMR spectra in the diet standardisation study, on the original scale (left) and after VSN transform (right).*

(rows in Figure 4), there is no obvious reduction in residual error variance across the three days of the study.

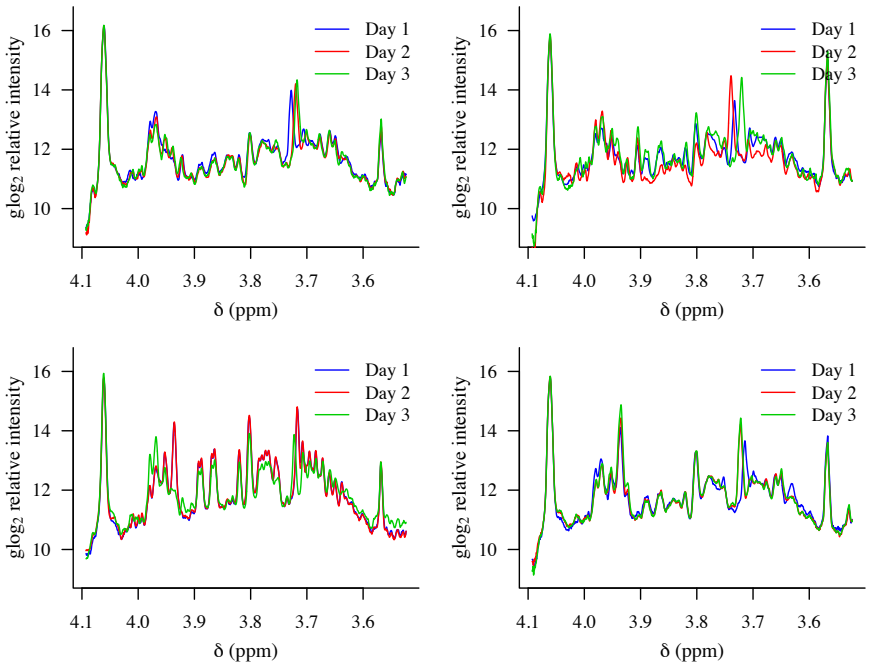

FIG 3. *Relative intensity (on the glog$_2$ scale) for a section of the NMR spectrum (4.1 - 3.5 ppm) for four individual subjects, for day 1 (blue), day 2 (red) and day 3 (green). A Danish male (top left), an Italian male (top right), a Danish female (bottom left) and an Italian female (bottom right).*

2.3. Challenges related to the data

The data contain 48 urine samples. Each urine sample, after ^{1}H NMR analysis, generates a spectrum containing 32 768 data points. This number was almost halved to 19 930 points by cutting out three regions containing mostly noise or the residual water signal. Nevertheless, this remains a so-called small-n-large-p problem where the number of the data points per individual spectrum far exceeds the number of individuals.

The spectral data contain noise which can obscure smaller peaks in the data and can influence the results of analyses.

The relative intensity values of the data (Figure 1, top) range from a minimum (per spectrum) of approximately -500 to -50, to a maximum of approximately 84 500 to 116 500. The median in each spectrum ranges from approximately 450 to 830. The data from each spectrum are clearly skewed to the right, due to a small number of peaks with very large values. Large values can potentially display large variation, whereas small values will tend to have small variation, i.e. heterogeneity of variance can be a problem.

There is correlation across an individual spectrum, since each peak consists

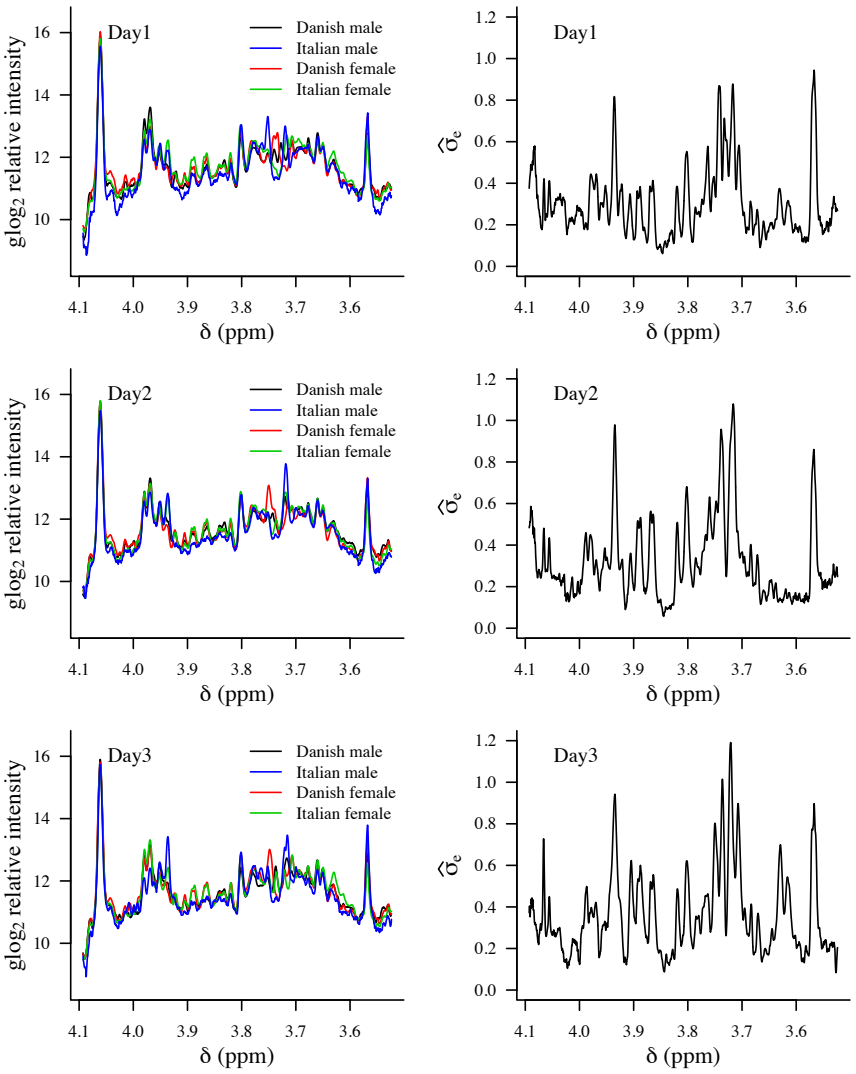

FIG 4. *For days 1 to 3 (rows), based on three separate models: sections (4.1 - 3.5 ppm) of predicted spectra (left column) for Danish males (black), Italian males (blue), Danish females (red) and Italian females (green) and the square root of the estimated variances of the random errors (right column).*

of a number of neighbouring data points. Furthermore, different peaks can be correlated, since a single metabolite consists of one or multiple peaks. There is also within-subject correlation since spectra from the same individual (repeated measurements on day 1, 2 and 3) will tend to be more similar.

To pre-process the data in a meaningful way, subject-specific expertise is essential, i.e. understanding the NMR platform and measurement procedures as well as knowledge of 'what the data should look like' and how to correct the data if they do not conform to expectations. Pre-processing was conducted by an experienced chemometrician. Some pre-processing steps, e.g. baseline adjustment, were performed on the NMR platform using platform-specific software. Other pre-processing steps, e.g. normalisation and alignment, were performed using custom developed software that is publicly available.

There is no established standard for pre-processing these types of data (Engel et al., 2013) i.e. the order of pre-processing steps and the specific methods chosen depends on the chemometrician conducting the pre-processing. The parameters used in pre-processing are often not reported and not available. Thus, pre-processing can be irreversible and also not reproducible, depending on the methods used. Different chemometricians may choose different pre-processing methods and, even when choosing the same methods, may choose different parameters. Even the same chemometrician pre-processing the same data set twice with the same methods, may choose different parameters. Furthermore, the order of pre-processing steps can influence each other, e.g. alignment/misalignment will influence normalisation and vice versa. The specific data pre-processing methods and order of steps can have a huge influence on the results of the analysis (Engel et al., 2013).

Considering the above issues, the inherent complexity of these data is evident.

In the next sections we describe our approach to address some of the complexities in the data through:

- a wavelet transform to smooth the data (remove noise) and to reduce the dimension of the data
- a mixed model with random effect for individual participants, to take into account correlation related to repeated measurements on the same participant
- bootstrap-based inference together with a functional data approach in order to address correlation across the spectrum.

3. Functional mixed models

In this section we first describe how functional observations fit into the classical mixed model. Second, we distinguish our work from existing literature on the topic.

For univariate observations y_i, $i = 1, \ldots, N$, a mixed model is expressed as

$$y_i = x_i^T \beta + z_i^T u + \varepsilon_i, \tag{1}$$

where x_i and β are p-dimensional vectors, z_i and u are q-dimensional vectors and ε_i are iid $\sim N(0, \sigma^2)$. Here x_i and z_i are vectors describing the observed covariates and experimental design, and the random effects are modelled as $u \sim N_q(0, \Sigma)$. In (1) any dependence between y_i's are described using the random effect term $z_i^T u$.

Concatenating the y_i's into an N-dimensional vector $\mathbf{y}$, and introducing the matrices $X_{N\times p}$ and $Z_{N\times q}$ with rows given by x_i^T and z_i^T we can write the mixed model (1) for $\mathbf{y}$ as

$$\mathbf{y} = X\beta + Zu + \varepsilon, \tag{2}$$

where $\varepsilon \sim N_N(0, \sigma^2 I_N)$ and (as before) $u \sim N_q(0, \Sigma)$.

In the case of functional observations, we typically observe the value $\mathbf{y}(t_j)$ of N curves at $j = 1, \ldots, n$ timepoints $t = (t_1, \ldots, t_n)$. We can use a mixed model for each $\mathbf{y}(t_j)$ but a full model specification requires a description of the correlation structure between $\mathbf{y}(t_j)$ for different $j = 1, \ldots n$. The mixed model for all $\mathbf{y}(t_j)$ is specified by stacking observations for different t_j's into a vector $\mathbf{y}$ of length $N \cdot n$ and by using a model of the form (2). However, the functional nature of the data often imposes structure between $\mathbf{y}(t_j)$. This should be reflected in the structure of the design matrices and the model for the random terms.

Morris (2006) assumes that the design matrices X and Z do not depend on the time argument, t, and model random effects $u(t)$ as well as residual errors $\varepsilon(t)$ by multivariate Gaussian processes with N-dimensional cross covariance functions parameterised by the product of an $N \times N$ matrix Γ and covariance surface Σ. For a fixed grid of time points $t = (t_1, \ldots, t_n)$ this amounts to a tensor product structure on the covariance matrix for vectors obtained by stacking the random effects $u(t_j)$ and the error terms $\varepsilon(t_j)$ respectively.

Instead of modelling the raw functional observations $\mathbf{y}(t_j)$, Morris (2006) models the wavelet coefficients obtained from the discrete wavelet transform of each functional observation. Since this transformation acts independently on every function (curve), the correlations between curves remain unchanged after the discrete wavelet transformation. In particular, a functional mixed model with a tensor product structure on the covariance matrix on the level of raw data will lead to a tensor product structure on the covariance matrix for the wavelet coefficients. Morris (2006) assumes the wavelet coefficients within a given curve are independent across wavelet scale and location making the column covariances for both the random effects and the residual errors diagonal. This structure accommodates non-stationarity (e.g. curve-to-curve variances and smoothness in curve-to-curve deviations both to vary over t) and allows the wavelet space model (2) to be fitted one column (wavelet coefficient) at a time.

The Morris (2006) model described above does not assume independent random-effect functions. The between-curve correlation matrices can be chosen to accommodate different covariance structures between curves that may be suggested by the experimental design. These include simple random-effects, structures for functional data from nested designs, split-plot designs, sub-sampling designs and designs involving repeated functions over time (Morris, 2006).

We apply a mixed effect model to each coefficient of the discrete wavelet transform. However, no model assumptions are made about the correlations between different wavelet coefficients. Instead, a nonparametric bootstrap procedure (Crainiceanu et al., 2012) is used to account for correlations between coefficients within a curve.

Regularisation is a central issue when applying the mixed model to functional data. The estimation procedure should adapt to the smoothness of the functional fixed effects. In the notation of (2) we essentially need to estimate a p-dimensional parameter $\beta(t)$ for any time argument t. The most common approach is to used penalised optimisation (Guo, 2002; Chen and Wang, 2011; Krafty, Hall and Guo, 2011; Scheipl, Staicu and Greven, 2014). However, by working on the discrete wavelet transform of the functional observations we take advantage of the sparse representations that wavelets yield for smooth curves with a varying number of local features such as sharp peaks. By shrinking all wavelet coefficients towards zero, many of the small wavelet coefficients take on a zero value. In this way, we obtain a sparse approximation of the wavelet transform. After back transformation this sparse approximation provides a smoothed version of the original functional observations. Ogden and Greene (2010) also used a thresholding approach to shrink wavelet coefficients in the functional mixed model context.

Morris (2006) use a Bayesian prior with point mass at zero for the wavelet coefficients to regularise the parameters of the functional mixed model. Here we use a two-step procedure instead: first, a hybrid version of the SureShrink procedure Donoho (1995) is applied to shrink the wavelet coefficients of each curve. Second, a subset of sample-wise non-zero wavelet coefficients are retained as input to the mixed model. The bootstrap procedure addresses the variability of the estimates from the mixed model while taking into account the subset selection step.

The wavelet based functional mixed model can be fitted using various approaches: a Bayesian approach with Markov chain Monte Carlo (MCMC) simulation (Morris, 2006; Morris et al., 2008), a faster empirical Bayes method (Clement et al., 2012) or frequentist approaches, either focusing on functional hypothesis testing (Abramovich and Angelini, 2006; Antoniadis, 2007) or estimation of fixed and random effects (Ogden and Greene, 2010). We use a frequentist approach for the estimation of differences in fixed effects with joint confidence intervals.

4. Wavelet-based functional mixed models

The three basic steps for the nonparametric wavelet-based approach to fit a functional mixed model are (Morris, 2006) :

Step 1 Decompose the N observed spectra to obtain empirical wavelet coefficients, by using the DWT on each spectrum. This is a projection of the observed spectra from the data space to the wavelet space.

Step 2 Model the empirical wavelet coefficients using a wavelet space version of the functional mixed model.

Step 3 Transform the wavelet space model estimates back to the data space, by using the IDWT, and use these estimates for inference in the original data space.

We expand step 2 in the following way:

Step 2a Use the hybrid SureShrink procedure Donoho (1995) to shrink the wavelet coefficients of each curve.

Step 2b Select a subset of sample-wise non-zero wavelet coefficients to be retained as input to the mixed model.

Step 2c Apply the wavelet-based functional mixed model to empirical wavelet coefficients

Step 2d Bootstrap the wavelet coefficients from Step 2a while keeping the structure of the data intact (e.g. repeated measurements) and repeat Steps 2b and 2c.

Step 1

Step 2 focuses on constructing the functional mixed model in the wavelet space. Right multiplication of both sides of model (2) yields a wavelet space model:

$$\mathbf{d} = X\beta^* + Zu^* + \varepsilon^* \tag{3}$$

where $X_{N\times p}$ and $Z_{N\times q}$ are the design matrices, $\beta^* = \beta W^T$ is a $p \times n$ matrix whose rows contain the wavelet coefficients for the p fixed effect functions on the grid t, $u^* = uW^T$ is a $q \times n$ matrix whose rows contain the wavelet coefficients for the q random effect functions and $\varepsilon^* = \varepsilon W^T$ is a $N \times n$ matrix consisting of the residual errors in the wavelet space. Like $\mathbf{d}$, the columns of β^*, u^* and ε^* are all double indexed by the wavelet coefficients' scale and location. Note that the between-row covariance structure is retained when projecting into the wavelet space; only the column covariances change. (Morris, 2006)

In Step 3 the wavelet model results from Step 2 are projected back into the data space, by using the inverse discrete wavelet transform $\mathbf{y} = \mathbf{d}W$.

Our approach in Step 2 combines bootstrap resampling of wavelet coefficients and subset selection. This yields model estimates for each bootstrap sample. After inverse discrete wavelet transformation back from the wavelet space to the original data space the bootstrap estimates can be used to estimate the uncertainty in the original data space mixed model. The bootstrap procedure is described in the next section.

5. Bootstrap based inference

Crainiceanu et al. (2012) described a general statistical framework for bootstrap-based inference for correlated functional processes where the data Y are defined as functions $Y_{ik}(t)$ with $t = t_1, \ldots, t_n$, i is the individual subject for whom the function is measured and k is the index associated with the correlated functional process, e.g. longitudinal observations, repeated measurements or matched pairs. Let

$$Y_{ik}(t) = \eta(t, X_{ik}) + V_{ik} \tag{4}$$

where X_{ik} is a vector of covariates, $\eta(t, X_{ik})$ is the population-level mean of the functional process $Y_{ik}(t)$ and V_{ik} is the residual process and can have a complex correlation structure. In a longitudinal study X_{ik} may depend on subject i only, or on the subject i and observation k within the subject i. $\eta(t, X_{ik})$ can take on

many forms, e.g. $X_{ik}\beta$ (standard parametric linear regression), $\mu(t)+X_{ik}\beta$ ($\mu(t)$ modelled parametrically or nonparametrically) or $\mu_A(t)I\{t \in A\} - \mu_B(t)I\{t \in B\}$, where groups A and B have mean functions $\mu_A(t)$ and $\mu_B(t)$ respectively.

Crainiceanu et al. (2012) proposed the following approach:

- Bootstrap subjects and obtain estimators of population-level parameters $\eta(t, X_{ik})$ under the assumption of independence, i.e. $V_{ik}(t)$ are i.i.d. mean zero and homoscedastic random variables.
- Conduct inference about $\eta(t, X_{ik})$ by using the empirical bootstrap distribution of $\eta(t, X_{ik})$, namely $\hat{\eta}^b(t, X_{ik})$ for $b = 1, \ldots, B$.

The bootstrap procedure used should preserve the correlation structure specific to the study design, e.g. the individual correlation in a longitudinal study. The initial estimation (to obtain a mean function from data values) as well as the bootstrap procedure can be performed in various ways.

With regard to bootstrap methods for estimating uncertainty in parameters of linear mixed effect models Thai et al. (2013) compared a variety of different parametric and non-parametric bootstrap approaches. The paired bootstrap, also called the case bootstrap, performed as well as the bootstraps of both random effects and residuals. The case bootstrap is a nonparametric bootstrap where entire subjects are resampled with replacement. An entire subject would consist of the joint vector of design variables and corresponding responses for subject i (for all observations k) from the original data before fitting a model, i.e. (X_i, Z_i, y_i). No assumptions are made about the model. Observations within subjects are not resampled (Thai et al., 2013). Although the above results are for the classical mixed effects model, we can apply it in the functional context where individual wavelet coefficients are modelled using classical mixed effect models. The case bootstrap corresponds to the nonparametric bootstrap in Crainiceanu et al. (2012).

Our method is conceptually similar to the 'nonparametric estimation using nonparametric bootstrap approach' in Crainiceanu et al. (2012). However, we do not use a penalised spline approach for nonparametric smoothing (of either the entire dataset for each group (over time) or the empirical means for each group (over time)), but a wavelet-based approach on each individual (over time). Additionally, instead of calculating bootstrap estimates of differences, we estimated fixed effects from a mixed model on the wavelet coefficients. Our data do not consist of matched pairs, but of repeated measurements of individuals over time. Our method entails the following:

1. Instead of obtaining estimators of the mean function (in each group k) under the independence assumption and then calculating the difference in group mean functions, we use a functional mixed model to obtain estimators of the fixed effect functions (in the wavelet space).
2. Use nonparametric bootstrap that keeps the structure of the data intact, i.e. repeated measurements. We used $B = 501$ bootstrap samples.
3. Fit the mixed model on each bootstrap sample $b = 1, \ldots, B$ and obtain estimators of $\eta(t, X_{ik})$ under the assumption of independence, i.e. under the

assumption that $V_{ik}(t)$ are i.i.d. zero-mean homoscedastic random variables. For NMR spectra a functional mixed model as in section 3 will translate to

$$\eta(t, X_{ik}) = \mu(t) + X_{ik}\beta \tag{5}$$

where we account for the correlation between repeated measurements within individual subjects i by defining a multi-level functional model for the error process as

$$V_{ik}(t) = A_i(t) + \epsilon_{ik}(t) \tag{6}$$

where $A_i(t)$ is a random functional process and $\epsilon_{ik}(t)$ is the individual residual error, assumed to be a Gaussian process with variance σ^2.

4. Transform the bootstrap model estimates $\hat{\eta}^b(t, X_{ik})$ back from the wavelet space to the original data space

$$\hat{\zeta}^b(t, X_{ik}) = \hat{\eta}^b(t, X_{ik})W \qquad \text{for } b = 1 \dots B \tag{7}$$

5. Contruct 95% pointwise confidence intervals for $\zeta(t, X_{ik}(t))$ based on the empirical bootstrap distribution obtained from $\hat{\zeta}^b(t, X_{ik})$ for $b = 1, \dots, B$ where $b = 1, \dots, B$, i.e. bootstrap percentile 95% confidence intervals ($q_{0.025}(t)$, $q_{0.975}(t)$) for all t.
6. Contruct 95% joint confidence intervals based on the empirical bootstrap distribution: assume $\hat{\zeta}^b(t, X_{ik})$ is an estimator for $\zeta_(t, X_{ik})$ for bootstrap b. The pointwise estimators for the mean and standard deviation of the mean for $\zeta_(t, X_{ik})$ are $\bar{\zeta}(t, X_{ik}) = \sum_{b=1}^{B} \zeta_b(t, X_{ik})/B$ and $s_{\bar{\zeta}(}t, X_{ik}) = \sqrt{\sum_{b=1}^{B}\{\zeta_b(t, X_{ik} - \bar{\zeta}(t, X_{ik}))\}^2/B}$ respectively. We construct random variable realisations $M_b = max_t|\zeta_b(t, X_{ik} - \bar{\zeta}(t, X_{ik}))|/s_{\bar{\zeta}(}t, X_{ik})$, the maximum over the entire range of t values of the standardised mean realisations. Then a $100(1-\alpha)\%$ joint confidence interval for $\zeta_(t, X_{ik})$ will take the form $\bar{\zeta}(t, X_{ik}) \pm q_{1-\alpha}s_{\bar{\zeta}(}t, X_{ik})$, where $q_{1-\alpha}$ is the $1-\alpha$ quantile of M_b for $b = 1, \dots, B$.

The pointwise confidence intervals imply that at *each* chemical shift position in repeated samples, the true function will be covered by the pointwise confidence interval $100(1-\alpha)\%$ of the time. The joint confidence intervals imply that at *all* chemical shift positions in repeated samples, the true function will be covered by the joint confidence intervals $100(1-\alpha)\%$ of the time.

Although the joint confidence intervals should be used for formal hypothesis testing, pointwise confidence intervals can be used in exploratory analysis for biomarker discovery, but should be followed by validation.

6. Implementation

The analysis was done in R (R Core Team, 2014). We used the VSN package from Bioconductor (Huber et al., 2002) to transform the data and the R WaveThresh

package (Nason, 2013) to perform the discrete wavelet transform, SureShrink thresholding and the inverse discrete wavelet transform. We modified procedures from the R boot package (Canty and Ripley, 2014) to perform bootstrapping of the estimated functions. We used the R lme4 package (Bates et al., 2014) to fit mixed models on the wavelet coefficients, running the models on the bootstrap samples in parallel using the R package snow (Tierney et al., 2014).

7. Analysis of the example data

We chose to use the pre-processed data as provided, in order to facilitate comparisons of our functional approach with published results (Rasmussen et al., 2011) and to preserve chemometric expertise embedded in the pre-processing. It is known that different approaches to pre-processing spectra can seriously influence the results from subsequent data analysis (Engel et al., 2013). We briefly described the pre-processing in section 2.1.

We did, however, perform some additional steps to tailor the data for the wavelet based functional approach. We reduced the spectrum length to the largest power of two smaller than the initial length, by cutting the ends of spectral sections without removing areas that may contain meaningful peaks. Relative intensity data were transformed to reduce skewness - the variance stabilisation normalisation transform was used. We did not adjust the baseline or the alignment of the spectra, but relied on the pre-processing already conducted.

We use wavelets to estimate the unknown true NMR functions, from the noisy observations. More specifically, we use Daubechies Least Assymetric wavelets (Daubechies, 1992) with four vanishing moments, periodic boundary handling and a primary resolution of 11. After SureShrink thresholding of wavelet coefficients, we selected a subset of wavelet coefficients that have non-zero values for all spectra.

An NMR spectrum or signal consists of relative resonance intensities $\boldsymbol{y} = (y_1, \ldots, y_n)$, which are discrete observations at equally spaced values $t = (t_1, \ldots, t_n)$ on the chemical shift axis (δ) and we write $\boldsymbol{y}(t)$.

In our application, we write $\boldsymbol{y}_i(t)$ to indicate the pre-processed NMR spectrum belonging to individual i, where $i = 1 \ldots 48$. The vector t consists of equally spaced chemical shift values t_j with $j = 1, \ldots, 16384$. The elements of X_{ih}, the fixed effect design matrix are covariates $h = 1, \ldots, p$, e.g. day, gender and nationality (diet culture) and interactions of these, for individual i. In this model the covariates are discrete (but the model allow for continuous covariates). β_h is the functional coefficients for predictor h over all chemical shift values t of the NMR spectrum. The elements, Z_{ik} with $k = 1, 2, 3$, of the random effect design matrix are used to model correlation among spectra, e.g. for repeated spectra per individual i, an individual-level random effect function is specified.

Our aim is to assess whether levels of relative resonance intensity are predicted by gender, diet culture, number of days on the standardised diet and/or any interaction of these terms, and whether these relationships depend on spectral position. We formulate a mixed model that is conditional on the thresholding

used and on all ($N = 48$) values of a specific wavelet coeffcient being non-zero after thresholding. This model can be written as a classical linear mixed model per coefficient in the wavelet space. For our data there are 2049 wavelet coefficients where all 48 values are non-zero.

What we expect in a diet standardisation study

It is important to note that the aim of a diet standardisation intervention is to eliminate individual differences in metabolites, where these differences are due to diet culture and food intake. It is to be expected that some differences in metabolites, due to gender and other known or unknown biological or other sources could remain. Taking these considerations into account, we expect to find mostly spectral regions with no significant differences between diet culture groups. We also expect differences in gender groups to be reduced, where differences were influenced by male or female diet preferences. Some gender differences that are biological in nature are expected to remain.

Motivation for considering specific spectral regions

Prior to the diet standardisation pilot study, a smaller pilot study was conducted (Rasmussen et al. (2011), study A). Here we refer to it as the non-intervention study, since participants had no dietary standardisation and followed their habitual diets. This study included seven of the 16 individuals who were later included in the diet standardisation study. The data from this non-intervention study are not analysed here. Nevertheless, the spectral regions identified as potentially discriminating for gender groups and diet culture groups, are used here as a guide to investigate regions with potential differences in metabolites in the diet standardisation study.

8. Results

We considered different contrasts from the model: contrasts for both main effects and interaction terms (Table 2). Each contrast consists of a vector with the same length as the spectrum ($n = 16384$). Only a small number of points were 95% jointly significant over the spectrum (ranging from 0 to 73, depending on the contrast).

To summarise the results, we calculated for each contrast the minimum p-value that would ensure that a 100(1-p)% joint confidence interval would contain no difference in the contrast over the entire spectrum. These minimum p-values are relevant since the study focused on diet standardisation and we are interested in the reduction of metabolic differences that can be ascribed to diet.

For example, a p-value of < 0.01 indicates that at least a 99% joint confidence interval is required to ensure that zero (no difference) is contained in the confidence interval for the contrast, over the entire range of the spectrum. Similarly, a p-value of < 0.05 indicates that a joint confidence interval wider than that associated with a 95% significance level is required to contain zero values for the contrast over the entire range of the spectrum.

For the minimum p-values in Table 2 we did not adjust for multiple testing across contrasts, although the joint confidence intervals that the numbers are

TABLE 2
Wavelet-based functional mixed model contrasts

Contrast		Min. p-value	95% jointly significant	
		(over spectrum)	No. of points	No. of regions
Day 2	Day 1	<0.001	19	3
Day 3	Day 2	0.108	25	3
Day 3	Day 1	0.008	0	0
Female	Male	<0.001	54	10
Italian	Danish	<0.001	73	16
Female Day 2	Female Day 1	0.040	7	2
Female Day 3	Female Day 2	0.036	3	1
Female Day 3	Female Day 1	0.030	2	1
Male Day 2	Male Day 1	0.042	14	3
Male Day 3	Male Day 2	0.052	1	1
Male Day 3	Male Day 1	0.048	0	0
Italian Day 2	Italian Day 1	<0.001	22	7
Italian Day 3	Italian Day 2	0.054	15	3
Italian Day 3	Italian Day 1	0.002	0	0
Danish Day 2	Danish Day 1	0.046	4	2
Danish Day 3	Danish Day 2	0.058	4	1
Danish Day 3	Danish Day 1	0.040	0	0
Day 1 Female	Day 1 Male	0.006	13	2
Day 2 Female	Day 2 Male	0.002	65	7
Day 3 Female	Day 3 Male	0.002	14	2
Day 1 Italian	Day 1 Danish	0.020	40	5
Day 2 Italian	Day 2 Danish	0.006	41	7
Day 3 Italian	Day 3 Danish	0.008	30	6

based on were adjusted for multiple testing across the spectrum. In the same table we present the corresponding number of 95% jointly significant points and regions (neighbouring points) for each contrast.

Table 2 serves as an informative summary. The smallest minimum p-values (< 0.001) were obtained for the differences between respectively day 1 and 2, female and male, Italian and Danish, and Italians on day 1 and 2. No 95% jointly significant points were obtained in the difference between respectively day 1 and 3, males on day 1 and 3, and Danes on day 1 and 3. The largest number of 95% jointly significant regions were obtained for differences between Italians and Danes, and females and males respectively.

We discuss some specific regions of interest related to gender differences, diet culture differences and differences over days of diet standardisation in the next sections. The same can be done for regions of interest related to interaction terms, but that is beyond the scope of this paper.

8.1. Differences related to gender

We found no 95% jointly significant points in the spectral areas for creatinine, (singlet at 4.06 ppm), citrate (two doublets at 2.72 - 2.64 and 2.57 - 2.51 ppm, respectively) or alanine (doublet at 1.50 - 1.47 ppm), displayed in Figure 5.

Our results support Rasmussen's findings: after diet standardisation differ-

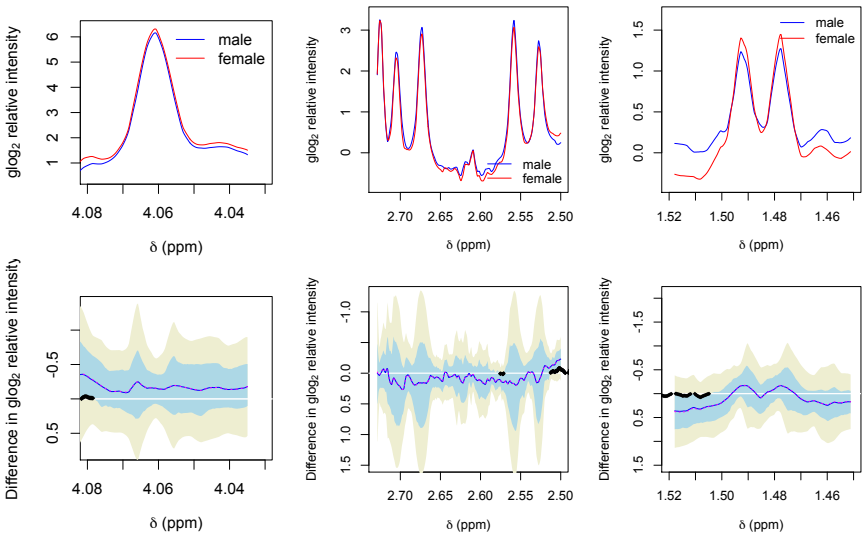

FIG 5. *Top row: Estimated mean curves for males (blue) and females (red) for (1) creatinine (4.06 ppm), (2) citrate (two doublets, centred at respectively 2.70 and 2.55 ppm) and (3) alanine (doublet centred at 1.49 ppm). Bottom row: estimated mean differences (blue line) between females and males, with 95% pointwise (light blue areas) and 95% joint (light yellow areas) bootstrap confidence intervals for (1) creatinine (2) citrate and (3) alanine. Areas marked in black indicate regions of 95% pointwise significant differences.*

ences in mean estimates for gender groups are not jointly or pointwise significant for spectral areas related to creatinine, citrate or alanine peaks.

8.2. Differences related to diet culture

Differences between Italians and Danes are discussed below as differences in 'dietary culture'. These differences are presumably due to different dietary habits caused by a different diet cultures (Rasmussen et al., 2011).

In the spectral region from 8.0 - 7.3 ppm, we found no 95% jointly significant points for difference in diet culture (Figure 6). Several spectral areas are 95% pointwise significant and these areas correspond to a number of peaks (including a doublet and triplets belonging to phenylalanine and hippurate) in the areas where diet culture differences were observed in the non-intervention study (Rasmussen et al., 2011).

In the spectral regions from 3.9 - 3.6 ppm (Figure 7) and 3.5 - 3.1 ppm (Figure 8) we found, respectively, two neighbouring points (at 3.846 ppm) and four neighbouring points (at 3.222 - 3.221 ppm) to be 95% jointly significant for difference in diet culture. Both of these small spectral areas fall within the valley between two peaks, but also within a larger spectral area that is 95% pointwise significant. On inspection of the individual spectra, these two areas

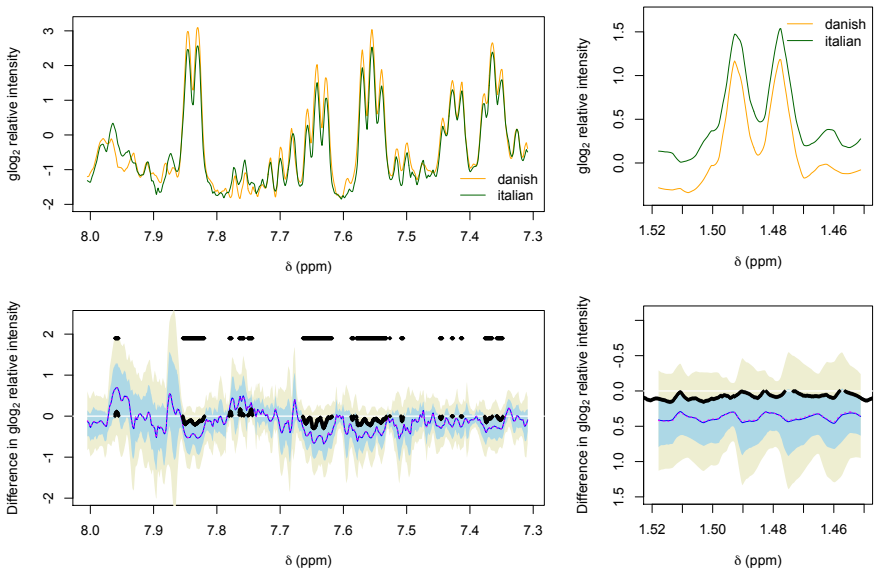

FIG 6. *Top row: Estimated mean curves for Danes (orange) and Italians (green) for (1) spectral region from 8.0 - 7.3 ppm containing a number of peaks (including a doublet and triplets belonging to phenylalanine and hippurate) (2) alanine (doublet centered at 1.49 ppm). Bottom row: estimated mean differences (blue line) between Italians and Danes, with 95% pointwise (light blue areas) and 95% joint (light yellow areas) bootstrap confidence intervals for the same chemical shift regions as in the top row. Areas marked in black indicate regions of 95% pointwise significant differences.*

seem to indicate either (a) misalignment in the areas surrounding 3.846 ppm and 3.222 ppm, (b) that the region of highest significant difference in a peak is located at the foot on one side of the peak, or (c) both of the above.

We did not find a 95% jointly significant difference between diet culture groups for the alanine doublet (1.5 - 1.47 ppm) (Figure 6). The differences were, however, 95% pointwise significant for almost all points in the alanine peak region.

The spectral features of the most discriminative regions (for diet culture) were 7.9 - 7.5 ppm, 3.92 - 3.82 ppm and 3.45 - 3.1 ppm (Rasmussen et al., 2011). Rasmussen reported enhanced signal intensities for Italian subjects in the spectral region 3.90 -3.60 ppm, but these differences from the non-intervention study were no longer significant after diet standardisation. The specific area of 3.92 - 3.82 ppm that was one of the most discriminative spectral areas for diet culture in the non-intervention study (Rasmussen et al., 2011), but disappeared after diet standardisation, was also not 95% jointly significant in our analysis.

Rasmussen reported that the spectral region containing the alanine doublet (1.5 - 1.47 ppm) was promising in discriminating between the two diet cultures in the non-intervention study. For the diet standardisation study, differences

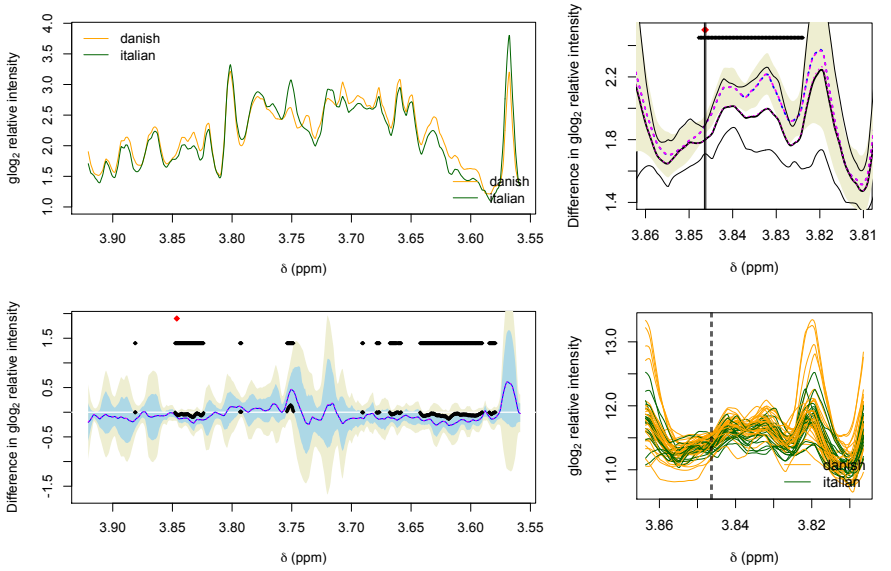

FIG 7. *Left, top: Estimated mean curves for Danes (orange) and Italians (green) for (1) spectral region from 3.95 - 3.55 ppm containing a number of peaks (including a singlet for glycine and a number of peaks for mannitol). Left, bottom: estimated mean differences (blue line) between Italians and Danes, with 95% pointwise (light blue areas) and 95% joint (light yellow areas) bootstrap confidence intervals for the same chemical shift regions as in the top row. Areas marked in black indicate regions of 95% pointwise significant differences; the red dot at the top of the figure indicates a region of 95% joint significant difference. Right, top: enlargement of the jointly significant region at 3.846 ppm with estimated mean curves and 95% joint bootstrap confidence intervals for Italians (black lines) and Danes (dotted pink line with light yellow regions), 95% pointwise (black) and joint (red) significant regions marked at the top of the figure. Right, bottom: wavelet estimates (thresholded) of the individual spectra in the same region (Danes - orange, Italians - green). Vertical lines correspond to points on the chemical shift axis where the difference is jointly significant.*

in alanine excretion separating Italians from Danes remained and were unexplained, and possibly due to chance (Rasmussen et al., 2011). Our results differ in that we did not find any joint significant differences in the alanine region, but this supports their suggestion that their finding may be due to chance.

8.3. Differences related to diet standardisation

We found 95% jointly significant points for the difference between Day 1 and Day 2 in three spectral areas: at approximately 8.18 ppm, 3.65 ppm and 3.49 ppm as displayed in Figure 9. We also found 95% jointly significant points for the difference between Day 2 and Day 3 in three other spectral areas: at approximately 7.785 ppm, 3.84 ppm and 3.065 ppm as displayed in Figure 10. The identification of the associated metabolites falls outside the scope of this study.

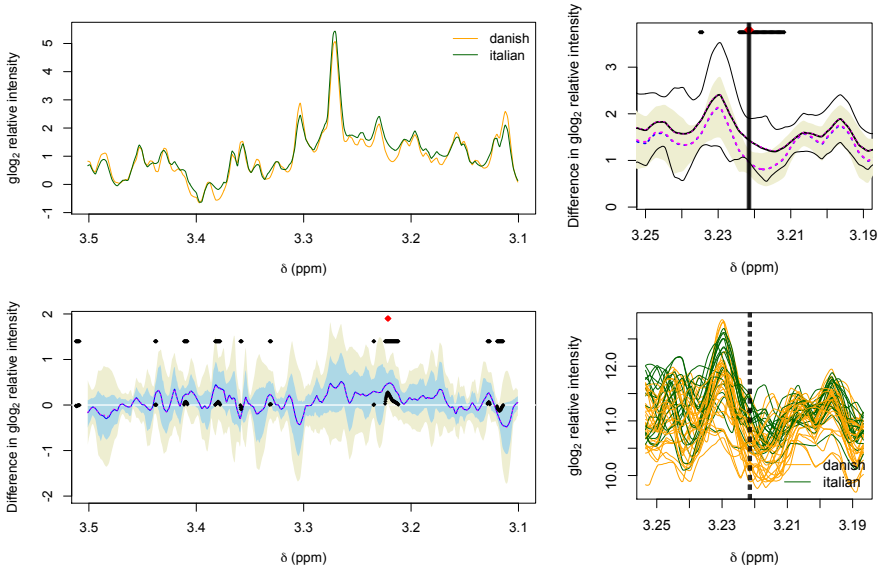

FIG 8. *Left, top: Estimated mean curves for Danes (orange) and Italians (green) for (1) spectral region from 3.5 - 3.1 ppm containing a number of peaks (including peaks from taurine, phenyalanine, TMAO, carnitine and DMS). Left, bottom: estimated mean differences (blue line) between Italians and Danes, with 95% pointwise (light blue areas) and 95% joint (light yellow areas) bootstrap confidence intervals for the same chemical shift regions as in the top row. Areas marked in black indicate regions of 95% pointwise significant differences; the red dot at the top of the figure indicates a region of 95% joint significant difference. Right, top: enlargement of the jointly significant region at 3.846 ppm with estimated mean curves and 95% joint bootstrap confidence intervals for Italians (black lines) and Danes (dotted pink line with light yellow regions), 95% pointwise (black and joint (red) significant regions marked at the top of the figure.Right, bottom: wavelet estimates (thresholded) of the individual spectra in the same region (Danes - orange, Italians - green). Vertical lines correspond to points on the chemical shift axis where the difference is jointly significant.*

Our results support Rasmussen's findings in the sense that, after diet standardisation, differences in mean estimates for gender groups are not jointly significant for spectral areas related to creatinine, citrate or alanine peaks. However, we identified six distinct spectral areas with jointly significant differences related to contrasts between days: three spectral areas related to the difference between Day 1 and Day 2, and three other areas related to Day 2 and Day 3.

8.4. Methodological results

The wavelet-based functional mixed effect model enabled us to investigate fixed effect contrasts for main effects like gender or day, as well as for interactions, e.g. gender differences on a specific day or day differences for a specific gender (Table 2). We took into consideration the design of the study by incorporating

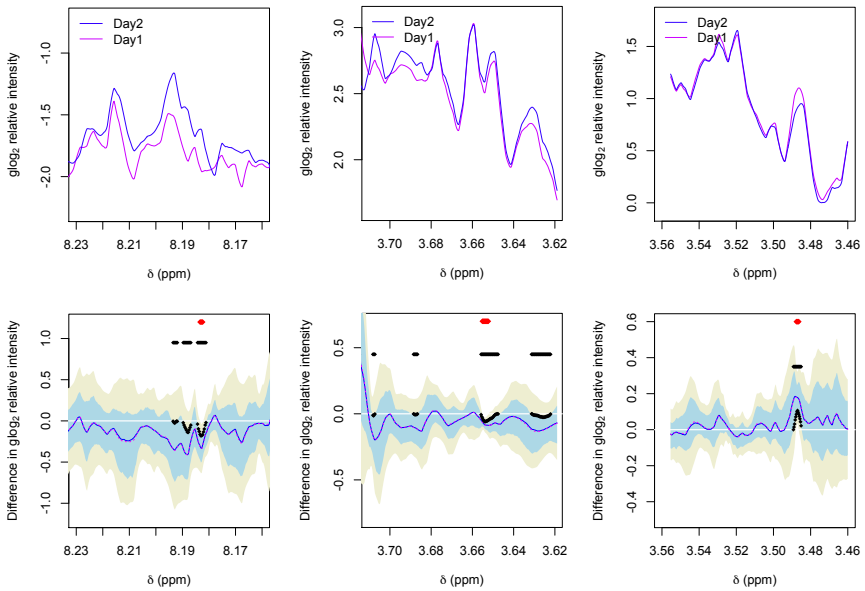

FIG 9. *Top: Estimated mean curves for Day 1 (pink) and Day 2 (purple) for the three different spectral regions that contain jointly significant points. Bottom: estimated mean differences (blue line) between Day 2 and Day 1, with 95% pointwise (light blue areas) and 95% joint (light yellow areas) bootstrap confidence intervals for the same chemical shift regions as in the top row. Areas marked in black indicate regions of 95% pointwise significant differences; the red dot at the top of the figure indicates a region of 95% joint significant difference.*

random effects per individual for repeated measurements over three days. Additionally to these advantages, the results are displayed as mean effects or mean differences across the range of the chemical shift. This is a major advantage for the interpretation of the results, in the sense that estimated effects and differences can be related to metabolites at the same position on the chemical shift axis.

Apart from mean estimates and estimates of mean differences we also obtained pointwise and joint confidence intervals over the entire range of the chemical shift axis. For our model the pointwise confidence intervals imply that at *each* chemical shift position in repeated samples over three days, the true mean (or difference in means) function will be covered by the pointwise confidence interval 95% of the time. The joint confidence intervals imply that at *all* chemical shift positions in repeated samples over three days, the true mean (or difference in means) function will be covered by the joint confidence intervals 95% of the time.

We used the joint confidence intervals for formal hypothesis testing. We can use the pointwise confidence intervals in an exploratory analysis for discovery of potential biomarkers, but validation will be necessary.

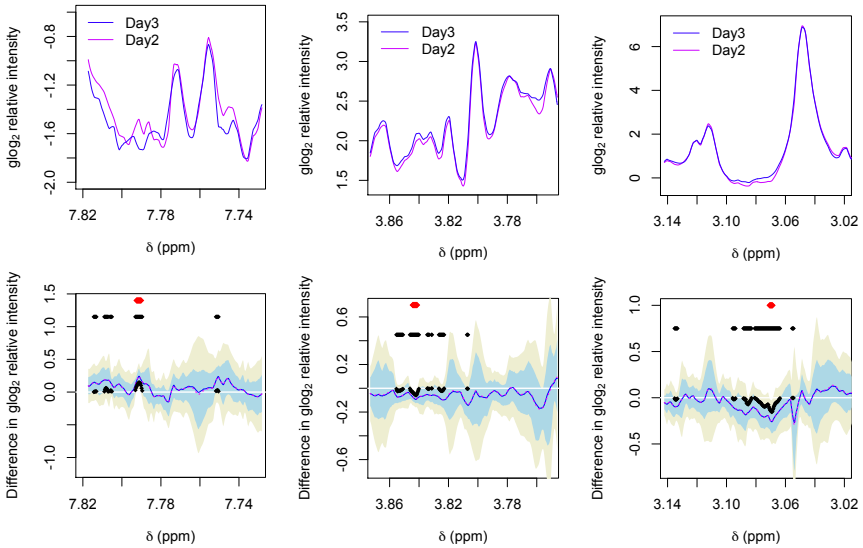

FIG 10. *Top: Estimated mean curves for Day 2 (pink) and Day 3 (purple) for the three different spectral regions that contain jointly significant points. Bottom: estimated mean differences (blue line) between Day 3 and Day 2, with 95% pointwise (light blue areas) and 95% joint (light yellow areas) bootstrap confidence intervals for the same chemical shift regions as in the top row. Areas marked in black indicate regions of 95% pointwise significant differences; the red dot at the top of the figure indicates a region of 95% joint significant difference.*

The data we analysed in this paper were from a small pilot study and not representative of any population and results should not be generalised. It serves merely as an example data set.

In terms of comparison of results from our method with published results using standard chemometric methods, there are different scenarios that can play out per spectral area (metabolite) and we offer a possible interpretation:

- we obtain no pointwise significant results and no differences were reported (published) - the results agree
- we obtain pointwise significant results and no differences were reported - this suggest a possible biomarker using our method
- we obtain jointly significant results and no differences were reported - we found a difference that remained otherwise obscured
- we obtain no pointwise significant difference but significant differences were reported - taking into account the study design may cause significant differences to disappear
- we obtain pointwise significant differences and significant differences were reported - differences suggest possible biomarkers, but are not jointly significant
- we obtain jointly significant results and significant differences are reported

- the results agree

9. Discussion

We applied the wavelet-based functional mixed model to NMR nutri-metabolomics data from a diet standardisation study. This approach allows us to respect the study design, while modelling the NMR spectra as functions. We modelled non-parametric fixed and random effect functions that enable us to incorporate covariates and repeated measurements in one model. We also demonstrated that it is possible to model interactions. The adaptive regularisation obtained through wavelet shrinkage is well suited to the type of data obtained from NMR where there are many sharp peaks at different locations. We used bootstrap-based inference to calculate 95% pointwise and joint confidence intervals for differences in gender, diet culture and number of days on the diet.

The method can accommodate more complex sampling designs.

This pilot study is not representative of any population and caution should be taken in interpreting results. However, the data serve to illustrate the methodology and the potential of wavelet-based functional mixed models in diet standardisation studies. The methodology is also applicable in intervention studies in nutritional metabolomics and other metabolomics studies like cancer metabolomics and plant metabolomics.

The diet standardisation study was a pilot study for the Diet, Obesity, and Genes (DiOGenes) six-month pan-European, multi-centre, randomised, controlled, dietary-intervention trial in obese and overweight families. In DiOGenes the five dietary interventions consisted of four combinations of low- or high-Glycaemic Index (LGI or HGI) and low- or high-protein (LP or HP) respectively, and a control diet. (Larsen et al., 2010; Aston et al., 2010; Moore et al., 2010)

The DiOGenes trial provides the context for future analyses using the wavelet based functional mixed model that we demonstrate here using the diet standardisation study. The adults participating in the DiOGenes study via the Danish centre were selected for a metabolomic analysis to investigate the influence of different dietary patterns on the urine metabolome. Of the 109 participants, 77 collected 24-hour urine samples at all four time points in the study (Rasmussen et al., 2012a,b). The analysis of these data is not included here. It will, however, be a natural next step to analyse these data using the wavelet based functional mixed model approach, now that proof of concept has been established on metabolomics data from the diet standardisation study.

It is interesting that some of the 95% joint significant areas lie in the shoulders of peaks, where the rest of the peak is potentially pointwise significant. We have some concerns related to the influence of 'residual misalignment' in terms of data that could be 'well enough' aligned for standard chemometric analysis. We suspect that wavelet-based functional mixed models are sensitive to very small misalignments, especially in the base of peaks and the valleys between peaks. We plan to investigate this potential sensitivity of our method to small perturbations in alignment on simulated data.

On a methodological level there is a need to investigate subset selection of wavelet coefficients for input to mixed modelling, together with the selection of primary resolution. The abovementioned issue of sensitivity to small misalignments is also relevant in this context. The use alignment methods from the functional data analysis literature, specifically k-means clustering and alignment, as well as simultaneous alignment and modelling may provide interesting avenues to investigate in the contact of NMR metabolomics spectra.

Acknowledgements

The authors are grateful to the following collaborators: Christian Ritz for initialising the project and for initial supervision; Lone Graasbøl Rasmussen for making the data available and for answering many questions; Francesco Savorani for providing the data and for very helpful discussions regarding NMR data and pre-processing.

References

Abramovich, F. and Angelini, C. (2006). Testing in mixed-effects FANOVA models. *Journal of Statistical Planning and Inference* **136** 4326-4348.

Antoniadis, T. Anestis; Sapatinas (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis* **51** 4793-4813.

Aston, L. M., Jackson, D., Monsheimer, S., Whybrow, S., Handjieva-Darlenska, T., Kreutzer, M., Kohl, A., Papadaki, A., Martinez, J. A., Kunova, V., Van Baak, M. A., Astrup, A., Saris, W. H. M., Jebb, S. A. and Lindroos, A. K. (2010). Developing a methodology for assigning glycaemic index values to foods consumed across Europe. *Obesity Reviews* **11** 92–100.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 R package version 1.1-7.

Biel, M., Evans, S. H. and Clarke, P. (2009). Forging links between nutrition and healthcare using community-based partnerships. *Fam Community Health* **32** 196–205.

Canty, A. and Ripley, B. D. (2014). boot: Bootstrap R (S-Plus) Functions R package version 1.3-13.

Chen, H. and Wang, Y. (2011). A Penalized Spline Approach to Functional Mixed Effects Model Analysis. *Biometrics* **67** 861-870.

Clement, L., De Beuf, K., Thas, O., Vuylsteke, M., Irizarry, R. A. and Crainiceanu, C. M. (2012). Fast wavelet based functional models for transcriptome analysis with tiling arrays. *Stat Appl Genet Mol Biol* **11** Article 4.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K. and Lindon, J. C. (2006). Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets. *Analytical Chemistry* **78** 2262-2267.

Crainiceanu, C. M., Staicu, A. M., Ray, S. and Punjabi, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Stat Med* **31** 3223-40.

Daubechies, I. (1992). *Ten lectures on wavelets* **61**. Society for Industrial and Applied Mathematics (SIAM).

Davidson, D. J. (2009). Functional Mixed-Effect Models for Electrophysiological Responses. *Neurophysiology* **41** 71-79.

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006). Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Analytical Chemistry* **78** 4281-4290.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory* **41** 613-627.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L. and Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry* **50** 96-106.

Favé, G., Beckmann, M., Draper, J. and Mathers, J. (2009). Measurement of dietary exposure: a challenging problem which may be overcome thanks to metabolomics? *Genes & Nutrition* **4** 135-141. 10.1007/s12263-009-0120-y.

Gibney, M. J., Walsh, M., Brennan, L., Roche, H. M., German, B. and van Ommen, B. (2005). Metabolomics in human nutrition: opportunities and challenges. *The American Journal of Clinical Nutrition* **82** 497-503.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G. and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol* **22** 245–252.

Guo, W. (2002). Functional mixed effect models. *Biometrics* **58** 121-128.

Holmes, E., Wilson, I. D. and Nicholson, J. K. (2008). Metabolic phenotyping in health and disease. *Cell* **134** 714–717.

Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A. and Vingron, M. (2002). Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* **18 Suppl. 1** S96-S104.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet Threshold Estimators for Data with Correlated Noise. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 319–351.

Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol* **7** 296–307.

Krafty, R. T., Hall, M. and Guo, W. (2011). Functional mixed effects spectral analysis. *Biometrika* **98** 583-598.

Larsen, T. M., Dalskov, S., Van Baak, M., Jebb, S., Kafatos, A., Pfeiffer, A., Martinez, J. A., Handjieva-Darlenska, T., Kunešová, M., Holst, C., Saris, W. H. M. and Astrup, A. (2010). The Diet, Obesity and Genes (Diogenes) Dietary Study in eight European countries – a comprehensive design for long-term intervention. *Obesity Reviews* **11** 76–91.

Moore, H., Adamson, A. J., Gill, T. and Waine, C. (2000). Nutrition

and the health care agenda: a primary care perspective. *Family Practice* **17** 197-202.

Moore, C. S., Lindroos, A. K., Kreutzer, M., Larsen, T. M., Astrup, A., Van Baak, M. A., Handjieva-Darlenska, T., Hlavaty, P., Kafatos, A., Kohl, A., Martinez, J. A., Monsheimer, S., Jebb, S. A. and of Diogenes, O. B. (2010). Dietary strategy to manipulate ad libitum macronutrient intake, and glycaemic index, across eight European countries in the Diogenes Study. *Obesity Reviews* **11** 67–75.

Morris, R. J. Jeffrey S; Caroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society B* **68** 179-199.

Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003). Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis. *Journal of the American Statistical Association* **98** 573-583.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A. and Coombes, K. R. (2008). Bayesian Analysis of Mass Spectrometry Proteomic Data Using Wavelet-Based Functional Mixed Models. *Biometrics* **64** 479–489.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society B* **58** 463-479.

Nason, P. Guy (2013). WaveThresh 4.6.6.

Ogden, R. T. and Greene, E. (2010). Wavelet modeling of functional random effects with application to human vision data. *Journal of Statistical Planning and Inference* **140** 3797-3808.

Pigoli, D. and Sangalli, L. M. (2012). Wavelets in functional data analysis: Estimation of multidimensional curves and their derivatives. *Computational Statistics & Data Analysis* **56** 1482-1498.

Qin, L. and Guo, W. (2006). Functional mixed-effects model for periodic data. *Biostatistics* **7** 225-234.

Quackenbush, J. (2007). Extracting biology from high-dimensional biological data. *Journal of Experimental Biology* **210** 1507-1517.

R Core Team, (2014). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, B. W. J. O. & Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag, New York.

Ramsay, B. W. J. O. & Silverman (2005). *Functional Data Analysis*, 2 ed. Springer-Verlag, New York.

Rasmussen, L., Savorani, F., Larsen, T., Dragsted, L., Astrup, A. and Engelsen, S. (2011). Standardization of factors that influence human urine metabolomics. *Metabolomics* **7** 71-83. 10.1007/s11306-010-0234-7.

Rasmussen, L. G., Winning, H., Savorani, F., Ritz, C., Engelsen, S. B., Astrup, A., Larsen, T. M. and Dragsted, L. O. (2012a). Assessment of dietary exposure related to dietary GI and fibre intake in a nutritional metabolomic study of human urine. *Genes & Nutrition* **7** 281-293.

Rasmussen, L. G., Winning, H., Savorani, F., Toft, H., Larsen, T. M., Dragsted, L. O., Astrup, A. and Engelsen, S. B. (2012b). Assessment of the Effect of High or Low Protein Diet on the Human Urine Metabolome as Measured by NMR. *Nutrients* **4** 112–131.

Savorani, F., Tomasi, G. and Engelsen, S. B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance* **202** 190 - 202.

Scheipl, F., Staicu, A.-M. and Greven, S. (2014). Functional Additive Mixed Models. *Journal of Computational and Graphical Statistics*.

Thai, H.-T., Mentr, F., Holford, N. H. G., Veyrat-Follet, C. and Comets, E. (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical Statistics* **12** 129-140.

Tierney, L., Rossini, A. J., Li, N. and Sevcikova, H. (2014). snow: Simple Network of Workstations R package version 0.3-13.

Whitfield, P. D., German, A. J. and Noble, P. J. (2004). Metabolomics: an emerging post-genomic tool for nutrition. *Br J Nutr* **92** 549–555.

Wishart, D. S. (2007). Current Progress in computational metabolomics. *Briefings in Bioinformatics* **8** 279-293.

II

Heart plots for Spectral Data

Martha Muller
Department of Mathematical Sciences
University of Copenhagen

James O. Ramsay
Department of Psychology (Professor Emeritus)
McGill University

Publication details

Manuscript (in preparation for submission).

Heart plots for spectral data

Martha Muller[a], James O. Ramsay[b]

[a]Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100, Copenhagen , Denmark
[b]McGill University, Montreal, Canada

Abstract

We present the concept of heart plots in spectral data, specifically for nuclear magnetic resonance peak of Lorentzian shape, i.e. with a Cauchy distribution. Heart plots provide a unique and useful view of the anatomy of spectral peaks. By using derivatives in FDA we extend the range of simple graphical exploratory methods and enable the development of more detailed methodology. Further statistical analysis may, among other methods, include principal component analysis of peak hearts and hypothesis testing.

Keywords: spectra, NMR, phase-plane plots, functional data analysis

1. Introduction

Analytical chemists are confronted with huge quantities of data and routinely use multivariate approaches, especially for pattern recognition.

There is a huge body of statistical literature, of which only a small portion will eventually be useful to chemists. Modern statistical approaches are not common in mainstream chemistry, and very few of the recent developments in statistics will make their way into the chemometricians' toolbox. As is the case in many disciplines, ideas are typically dissipated separately in chemometrics and statistics. Generally speaking, there is quite a gap between statisticians and chemometricians. ([1], Chap. 1) In this article we take a first step in terms of bridging the gap between statisticians and chemometricians.

This paper will appeal to (1) analytical chemists and chemometricians interested in modern statistical techniques applied to the analysis of spectral data, but also to (2) applied statisticians who desire to acquire an understanding of spectral data and the application of functional data analysis in chemometrics.

In terms of mathematical novelty, new theory is not the only sign of innovation. In fact, much of science involves connecting ideas ([1], Chap. 1). In this article we connect the theory of functional data analysis (from statistics) with spectral data (analytical chemistry).

The defining feature of Functional Data Analysis (FDA) is that a sample, or record, is a function as opposed to a single data point [2]. The functions are often smooth curves and can be observed over time or other dimensions such as space, frequency or chemical shift. Statistical analyses are then carried out on these functions and their derivatives, as opposed to analyses of the data points as repeated measurements (in statistics) or as multiple variables (in chemometrics).

In terms of spectral data, each function will belong to a single (chemical) sample and will be measured over chemical shift (frequency) for NMR data. Although we focus on NMR spectra in this article, many of the concepts are applicable to other spectral data, e.g. infrared and mass spectrometry.

As is often the case in different disciplines, different terminology is used to describe the same concept. In chemometrics and functional data analysis/statistics this is also the case. We provide a glossary of basic terms that are equivalent in the two fields (Table 1).

Although FDA has not yet featured in recent reviews of the most significant developments in the field of chemometrics [3, 4], it has been applied in chemometrics [5].

Our objective is to build a bridge between the fields of chemometrics and functional data analysis by introducing heart plots. In this article we formulate spectral data as mathematical functions, in order to reveal important and often hidden features of peak anatomy in the data. We plot these peak anatomy features, namely the slope (first derivative) against the curvature (second derivative), in a phase-plane plot, called a 'heart plot'. These heart plots are best utilised per interval and we name these *i*heart plots. Heart plots are powerful tools that reveal variability in spectral data.

In the first section we present simulated data. Second, we present a brief overview of the theory and methods of Functional Data Analysis relevant to spectral data in chemometrics. In the third section, we use simulated data, to illustrate the concept of heart plots. Finally we discuss the significance of our results and the value of *i*heart plots as a chemometric tool.

2. Materials and Methods

We use a number of simulated peaks to illustrate the concept of heart plots.

The chemical shift (x-axis) scale that we use is 3000 points for 1 ppm (i.e. 1 point = 0.0003 ppm). This is typically what we can expect from real data acquired at 400.13 MHz. We simulate the data over 1000 points, which spans an interval of 0.3 ppm.

Email addresses: m.muller@math.ku.dk (Martha Muller), ramsay@psych.mcgill.ca (James O. Ramsay)

Table 1: Glossary of equivalent terms

NMR chemometrics	**Functional Data Analysis**
intensity	amplitude
chemical shift (δ) (in *parts per million (ppm))*	location, phase
sample	record, curve, function
smoothing	regularisation
alignment	registration, warping
line	peak
line shape	peak shape, distribution
Lorentzian line shape	Cauchy distribution $X \sim \text{Cauchy}(\theta, \gamma)$
full width half maximum (FWHM)	2γ in the Cauchy distribution
multiplet	a group of (multiple) peaks
singlet	one peak
doublet	two peaks (1:1)
triplet	three peaks (1:2:1)
quartet	four peaks (1:3:3:1)
quintet	five peaks (1:4:6:4:1)
sextet	six peaks (1:5:10:10:5:1)
septet	seven peaks (1:6:15:20:15:6:1)
J-coupling	distance between peaks
coupling constant	magnitude of the splitting (difference in frequency) between peaks
intensity ratio	height ratio (of peaks in a group)
pattern recognition	classification
calibration	regression
multivariate curve resolution	deconvolution

It is known that NMR line shapes are Lorentzian. Typical linewidths (FWHM) in ^{1}H NMR spectra of small molecules in solution are around 0.2Hz [6] (p.7). Therefore we use Lorentzian line shapes (Cauchy distributions for peaks), with full width half maximum (FWHM) of 0.2 Hz, i.e. (0.2 Hz / 400.13 MHz =) 0.0005 ppm = 0.83 points. The scale parameter of the Cauchy distributions is thus 0.005 ppm = 1.66 points.

A typical value for the coupling constant would be 7 Hz. We simulate multiplets with coupling constants of 7 Hz = 0.0175 ppm = 58.33 points and intensity ratios of 1:1 (doublet), 1:2:1 (triplet), 1:3:3:1 (quartet), 1:4:6:4:1 (quintet), 1:5:10:10:5:1 (sextet), and 1:6:15:20:15:6:1 (septet).

2.1. Functional data analysis

The field of FDA has developed rapidly over the last two decades, both in terms of theory and diverse fields of application. We provide a brief overview of the basics of Functional data analysis (FDA) [2] as it is applicable to functional representation of NMR spectra. The concepts can be applied to other spectra, e.g. mass spectrometry.

In FDA each sample, or record, is a *function* or curve as opposed to a single data point. The functions are often smooth curves and can be observed over time or other dimensions such as space, frequency or chemical shift. Functional data are often measured at equally spaced intervals, but this is not a requirement. In each functional sample there is a finite set of numbers that reflect smooth variation in intensity and can be assessed at any value in the defined range. Functional data are often of high dimension, and for practical purposes each record consists of a fully observed function over the defined range. One of the aims of FDA is to separate amplitude variation (on the y-axis) and phase variation (on the x-axis) by *curve registration*. This process has some overlap with *alignment* of spectra in chemometrics. Statistical analyses are performed on the functions and their derivatives.

NMR spectra are represented as NMR intensity values over equally spaced intervals on the chemical shift axis (in ppm units). In each functional record, i.e. each spectrum, there is a finite set of numbers that reflect smooth variation in the intensity. The intensity can be assessed at any chemical shift (ppm) value in the range. The data are of high dimension, typical of functional data, and for practical purposes each record consist of a fully observed function over the defined chemical shift range.

Typically we will apply the following steps to each functional record:

1. Smoothing (specification of a basis system, building of Functional Data objects);
2. Registration / Feature alignment:
3. Calculate first and second derivatives of functional objects;
4. Phase-plane plots; and
5. Functional modelling.

In phase-plane plots we plot the second derivative (acceleration) against first derivative (velocity). A phase-plane plots is a powerful tool for exploring harmonic variation even in data where we do not ordinarily think of cyclic variation. Essentially, it is a graphical analogue of a second-order differential equation.

3. Results

3.1. Simulated data

In Figure 1, on the left, five identical Lorentzian lines (Cauchy peaks) are displayed (solid lines). Five factors are varied one by one (broken lines). On the righthand side of Figure 1 the corresponding heart plots are displayed. The following concepts are illustrated in (a) to (e):

The heart plot of a Lorentzian line is

a. invariant to the position of the peak (chemical shift)
b. invariant to a constant baseline added to the peak
c. invariant, with regard to shape and size, to a non-constant baseline added to the peak, but the baseline translates into a horizontal shift of the heart

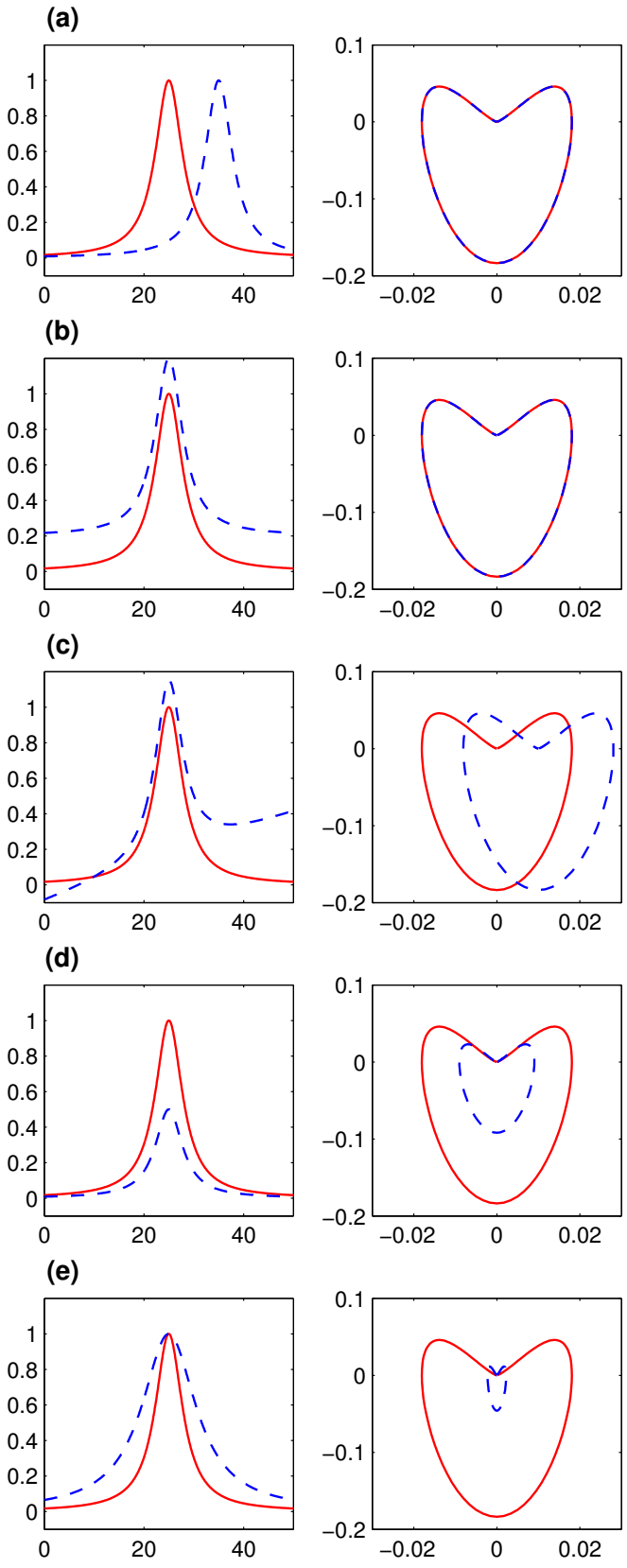

Figure 1: Lorentzian lines (Cauchy peaks) (solid red lines, graphs on the left) and variations (broken blue lines) with corresponding heart plots (on the right). Variations in (a) position / chemical shift, (b) constant baseline added t, (c) non-constant baseline added, (d) height (e) line width/FWHM (scale parameter).

d. invariant in shape, with regard to the height of (i.e. area under) a peak, but not with regard to size: a smaller peak height results in a smaller heart

e. an increase in line width results in a reduction in the size of the heart, but the shape of the heart is invariant to the line width (FWHM, scale parameter of the Cauchy distribution).

In Figure 2, (a) on the left, a Lorentzian line (Cauchy peak), identical to those in Figure 1, is displayed (solid red line) with a Gaussian peak of similar FWHM (broken blue line). Figure 2, (b-e) on the left displays Lorentzian shape doublets (two peaks) with decreasing coupling constants of (b) 7 and 2.8 Hz, (c) 1.4 and 1.05 Hz, (d) 0.77 and 0.56 Hz, (e) 0.42 and 0.28 Hz. Corresponding hearts plots follow on the righthand side of each graph.

The following concepts are illustrated in (a) to (e):

a. changing the line shape from Lorentzian to Gaussian results in a differently shaped heart: much 'fatter' i.e. shorter and wider, compared to the Lorentzian heart

b. a doublet, with a coupling constant of equivalent to 7 Hz, has a practically identical heart plot to that of a singlet of the same height (see Fig. 1(a)), apart from having two hearts on top of each other. Even when the coupling constant is reduced to 40% of the original value (blue dashed line) the heart plot of the doublet is not distinguishable from a singlet of the same height

c. when the coupling constant decreases further (to 20% and 15% of the original value), the two hearts of a doublet becomes distinguishable and the top centre of the smaller 'heart' swells upward

d. when the coupling constant decreases even further (to 11% and 8% of the original value), the smaller of the two doublet 'heart' shapes becomes more tear shaped and the larger one breaks into two heart points at the bottom

e. when the coupling constant decreases so much (to 6% and 4% of the original value) that the doublet appears as a singlet, firstly (for 6%) the smaller of the two 'hearts' (that became tear shaped) disappears into the larger one that breaks further into two 'points' at the bottom and then (for 4%) become one large 'heart' where the bottom point has been cut off.

In Figure 3, on the left, one large Cauchy peak (red broken line) is displayed in all five graphs (a to e) and a small Cauchy peak (20% of the height of the large peak) (solid lines) approaching the larger peak. The dynamic range is decreasing the distance between the peaks from (b) 23.4 points to 11.7 to (c) 8.775 and 7.02 to (d) 5.265 and 4.095 to (e) 2.925 points.

We observe that the small peak approaches the large peak, but is still clearly distinguishable (b) before it comes so close that it becomes a 'shoulder' on the large peak (c). In (d) and (e) the small peak is absorbed within the large peak. The heart plots show that:

a. the large Cauchy peak forms a heart plot (for reference)

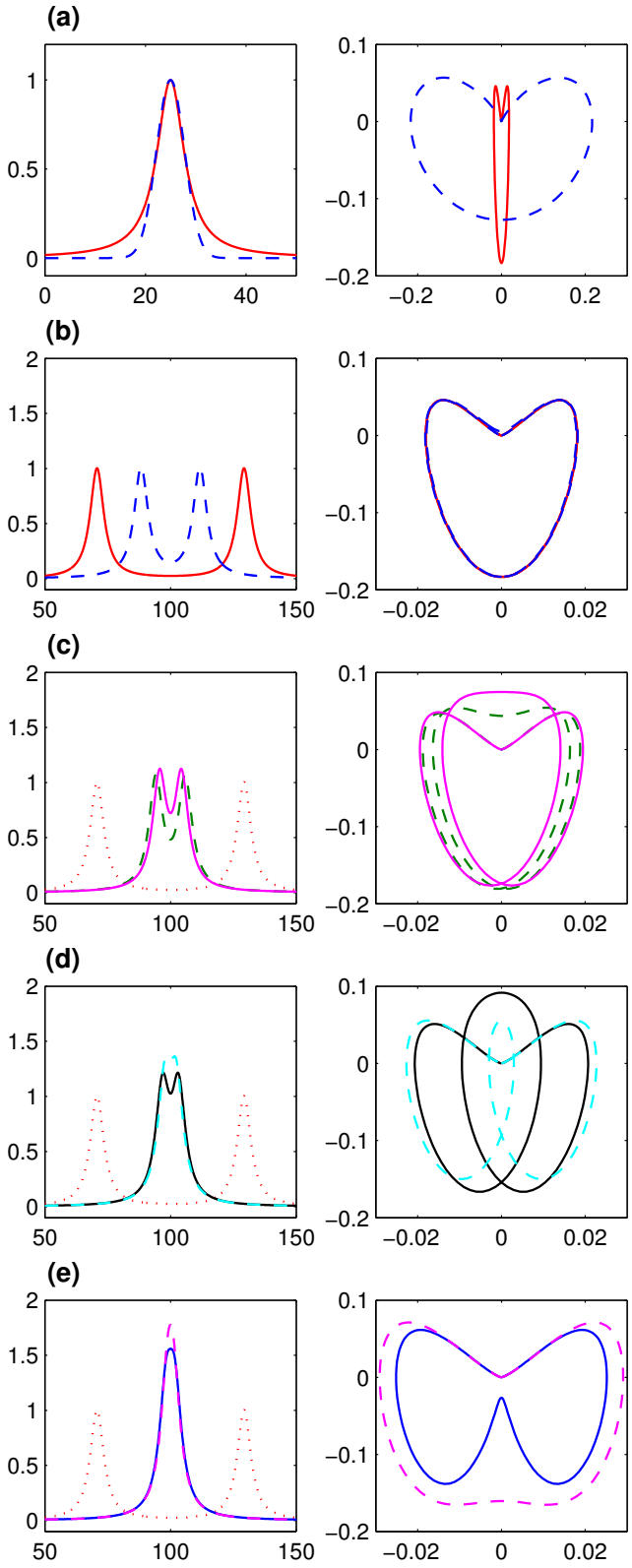

Figure 2: Singlets and doublets (graphs on the left) with corresponding heart plots (on the right). Variations in (a) the shape of the peak and (b-e) size of the coupling constant of the doublet.

b. the small peak appears as a small heart inside the large heart, which is identical to the singlet heart in (a). As the smaller peak moves closer to the large peak the heart has a slight and asymmetric swelling in the top centre
c. the more the small peak becomes a shoulder of the large peak, the more the smaller 'heart' deforms, migrates towards the right and the right upper lobe of the heart swells
d. when the small peak is absorbed by the large peak the small 'heart' shrinks and deforms, the previous swelling of the right upper lobe diminishes and the righthand side of the large heart also shrinks
e. in the final stage of totally absorbing the small 'heart', the large heart deforms by diminishing the entire right lobe and swelling the entire left lobe and in so doing tilts the bottom point of the heart slightly to the left

In Figure 4, on the left, we show (a) a singlet (for reference), followed by (b-e) a triplet (peak height ratios of 1:2:1), quartet (peak height ratios of 1:3:3:1) , quintet (peak height ratios of 1:4:6:4:1) and sextet (peak height ratios of 1:5:10:10:5:1). Corresponding heart plots are show on the righthand side.

The following is displayed:

a. the heart plot of a singlet (for reference)
b. the triplet heart plot with one heart identical to that of the singlet in (a) with two additional and identical (in shape) smaller hearts inside
c. the quartet heart plot with two pairs of identically shaped hearts, one pair identical in size to the singlet heart in (a), with another pair of smaller hearts inside. Note the quartet's inner hearts are smaller than that of the doublet in (b)
d. the quintet heart plot results in a heart identical to a singlet, but with two pairs of identically shaped and smaller hearts inside
e. the sextet heart plot has six identically shaped hearts, in three pairs of two identically sized hearts each, with the largest pair identical in size to the singlet heart. Note the two pairs of smaller hearts are smaller than those of the quintet in (d).

In Figure 5, we demonstrate the effect of square root transformation on Cauchy peaks. On the left, in Fig. 5(a) a singlet Cauchy peak (red) with its square root transformation (blue) is displayed and enlarged in (b). In spectral data, peak heights are typically of different orders of magnitude (c and d) e.g. from right to left (all in red) heights of 100 000, 10 000, 1000 (barely visible in (c)), 100 (invisible in (c)) and 10 (invisible in (c) and barely visible in (d)) with the corresponding square root transformation displayed (all in blue in (c and d)).

From the heart plots on the right of Fig. 5(a – e) we see:

a. the heart plot of a square root transformed Cauchy peak (blue) is much smaller than the original (red)
b. the shape of the square root transformed Cauchy peak (blue) is more compressed in vertical dimension and the two top lobes of the heart are stretched towards the outside (compare with (a))

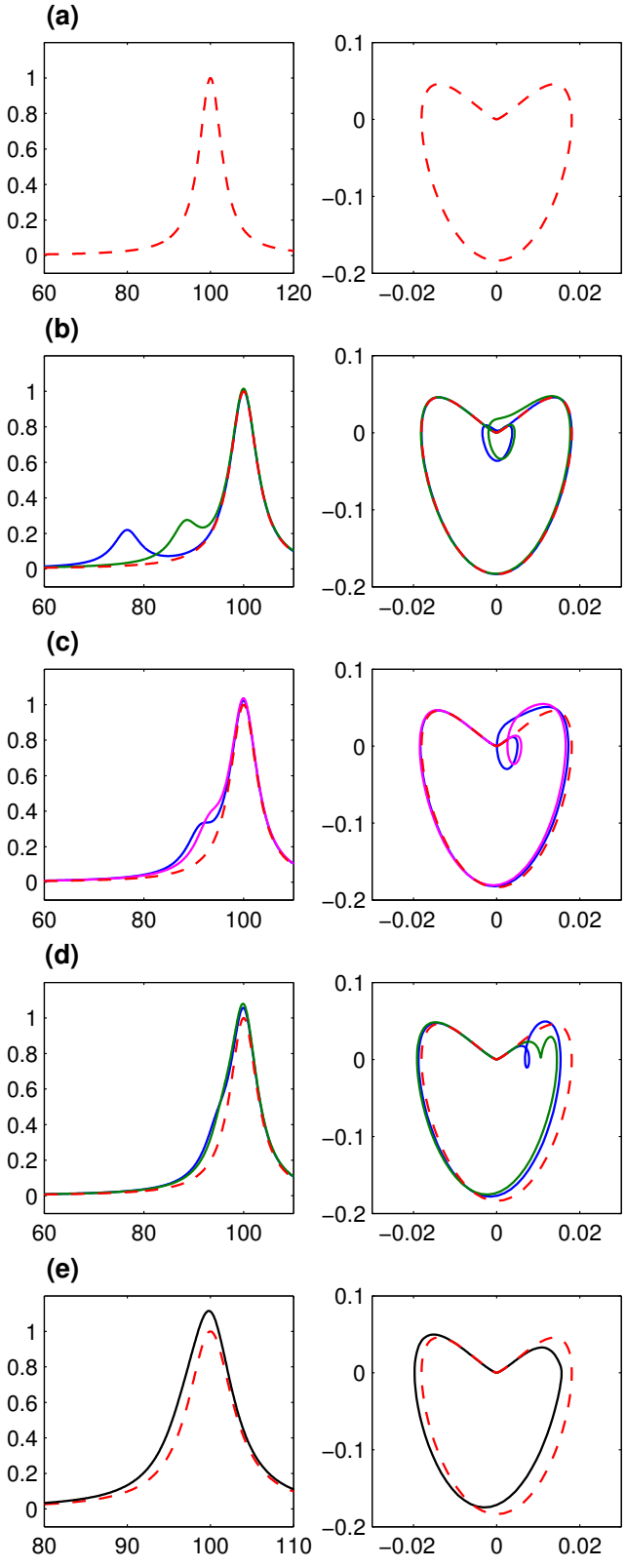

Figure 3: A large Cauchy peak (red broken line, graphs on the left) and a small Cauchy peak (solid lines) with decreasing dynamic range between the two peaks. Corresponding heart plots (on the right). (a) is the reference large Cauchy peak; with decreasing dynamic range between the two peaks: (b) , (c) , (d) and (e) s.

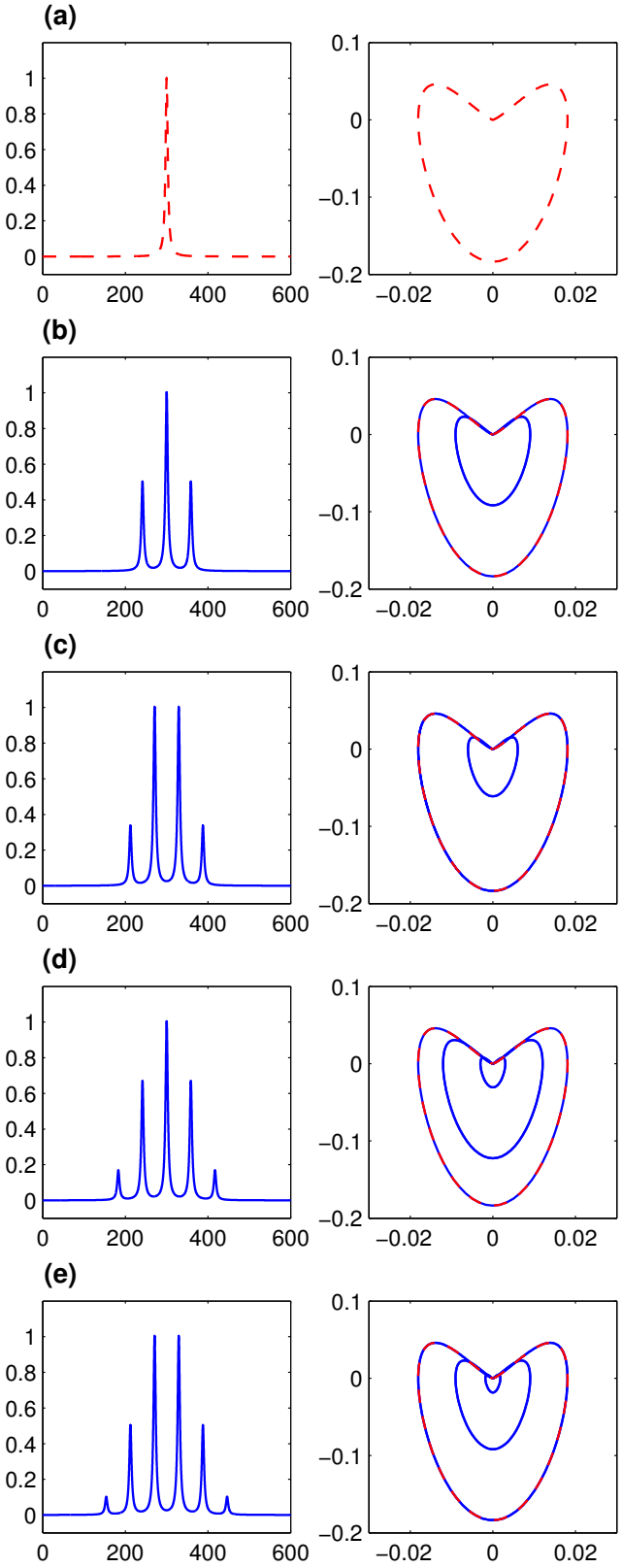

Figure 4: Multiplets (solid lines, graphs on the left) with corresponding heart plots (on the right). Variations in the number of peaks in the multiplet: (a) single, (b) triplet, (c) quartet, (d) quintet and (e) sextet.

c. on the original scale only the heart plots of the largest two peaks (heights of 100 000 and 1000) are visible without zooming in; the three smaller peaks 'disappear' on the heart plot

d. for the square root transformed peaks, there is less difference in size among hearts originating from peak heights with different orders of magnitude in their height and three of the hearts are clearly visible without zooming in

4. Discussion

Heart plots provide a unique and useful view of the anatomy of spectral peaks. A number of questions arise with the perspective for future analysis of spectral hearts

- Where in the spectra is variability the largest?
- Are there individual records with distinctive curves?
- Why do peaks show up as hearts?
- What information does the size of the heart convey?
- Does inter-record variability correspond to (chemical) energy?
- Why do certain sections of the phase-plane plot show up on the right vs. the left?

By using derivatives in FDA we extend the range of simple graphical exploratory methods and enable the development of more detailed methodology. Further statistical analysis may, among other methods, include principal component analysis of peak hearts and hypothesis testing.

5. References

[1] R. G. Brereton, Chemometrics: data analysis for the laboratory and chemical plant, John Wiley & Sons, 2003.

[2] J. O. Ramsay, B. W. Silverman, Functional Data Analysis, Verlag, New York, 1997.

[3] B. K. Lavine, J. Workman, J., Chemometrics, Anal Chem 85 (2013) 705–14. Lavine, Barry K Workman, Jerome Jr Anal Chem. 2013 Jan 15;85(2):705-14. doi: 10.1021/ac303193j. Epub 2012 Dec 3.

[4] B. Lavine, J. Workman, Chemometrics, Analytical chemistry 82 (2010) 4699–4711.

[5] W. Saeys, B. De Ketelaere, P. Darius, Potential applications of functional data analysis in chemometrics, Journal of Chemometrics 22 (2008) 335–344.

[6] P. Hore, R. Compton, Nuclear Magnetic Resonance: Oxford Chemistry Primers, Oxford University Press, New York, 1995.

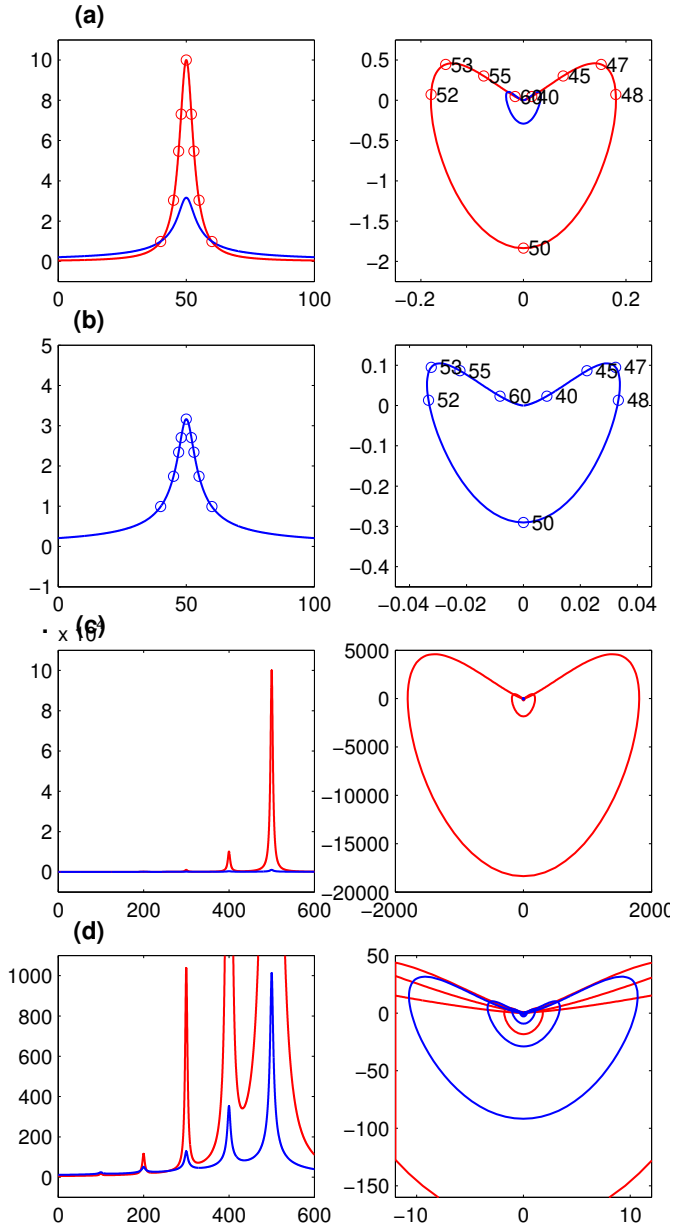

Figure 5: Comparison of Cauchy peaks (red lines, graphs on the left) and their square root transformation (blue lines, graphs on the left) with corresponding heart plots (on the right). Variations in the order of magnitude of the peak height (c and d): from right to left 100 000, 10 000, 1000, 100 and 10 on the original scale.

III

Analysis of Juggling Data: Registration Subject to Biomechanical Constraints

Anders Tolver, Helle Sørensen, Martha Muller and Seyed Nourollah Mousavi
Department of Mathematical Sciences
University of Copenhagen

Publication details

Published (2014) in *Electronic Journal of Statistics*, Special Section on Statistics of Time Warpings and Phase Variations. Vol. 8, No. 2, 1856–1864.

Electronic Journal of Statistics
Vol. 8 (2014) 1856–1864
ISSN: 1935-7524
DOI: 10.1214/14-EJS937F

Analysis of juggling data: Registration subject to biomechanical constraints*

Anders Tolver, Helle Sørensen, Martha Muller and Seyed Nourollah Mousavi

Department of Mathematical Sciences
University of Copenhagen
e-mail: tolver@math.ku.dk; helle@math.ku.dk;
m.muller@math.ku.dk; nourollah@math.ku.dk

Abstract: We illustrate how physical constraints of a biomechanical system can be taken into account when registering functional data from juggling trials. We define an idealized model of juggling, based on a periodic joint movement in a low-dimensional space and a periodic position vector (from an undefined joint to the finger tip) of approximately constant length along the observed trajectory. Our registration procedure first warps the cycles in the trial to each other and computes a periodic average, and then estimates the joint movement and the position vector of the abovementioned model.

Keywords and phrases: Biomechanical constraints, decomposition, functional data analysis, juggling trajectories, periodic average, registration, warping.

Received August 2013.

1. Introduction

Functional data are often unsynchronized in their raw form, either due to the sampling process or due to random phase variation (or both). This makes analysis on the raw data problematic since, for example, cross-sectional sample statistics can be misleading. Registration is the process of mapping unsynchronized curves into a synchronized class of functions, with the purpose of effectively filtering out noise before subsequent statistical analyses [1].

At best, registration should use any knowledge of the data generating system, in particular the shape of the underlying signal as well as the nature of possible pertubations. In this paper we discuss registration for functional data from juggling, taking into account simple biomechanical considerations.

Ideally, biomechanics of juggling may be described mathematically by nonlinear dynamical systems, but feedback and feedforward motor control mechanisms are necessary to overrule any disturbed dynamics and thereby impose desired movements or dynamics. We consider data from juggling cycles within in trial as pertubated versions of an idealized periodic movement. The periodic curve represents the average dynamics of the juggling process, whereas the deviations

*Main article 10.1214/14-EJS937.

between the observed data and the idealized signal reflect the complex feedback mechanism between the brain and the motor control system [4].

In conceptualizing an appropriate idealized mathematical model of human juggling, we consider the creation of an electromechanical juggling robot. How would we build and program such a robot? As a minimum, we would construct a rotating finger or hand limb and attach it with a joint to a fixed bar (representing an arm). We could conveniently label the two ends of the hand limb as 'finger tip' and 'joint'.

As a first attempt, we keep the position of the joint fixed and let the position vector from joint to finger tip be periodic. Regarded from a fixed external coordinate frame the position of the finger tip of the robot would trace a trajectory described by

$$f(t) = f_0(t) + c_0$$

where $c_0 \in \mathbb{R}^3$ corresponds to the fixed position of the joint and $f_0 : I \to \mathbb{R}^3$ is the periodic position vector function. Assuming that the robot is a rigid body introduces the geometric constraint that f_0 has constant length, d, such that $|f_0(t)| = d$ for all $t \in I$.

The juggling robot can be improved by allowing the position of the joint to follow a periodic curve. This gives a decomposition of the form

$$f(t) = f_0(t) + c_0(t), \tag{1}$$

where $c_0 : I \to \mathbb{R}^3$ is the trajectory of the joint, while f_0 still describes the vector from joint to finger tip and satisfies $|f_0(t)| = d$ for all $t \in I$ for some d. For identification purposes we assume that c_0 has a simple structure meaning that it belongs to a lower dimensional function space.

In this paper, decompositions of the type (1) will be regarded as idealized juggling signals, and we will demonstrate how to register the observed data towards such idealized signals, i.e. demonstrate that is it is possible to warp and filter the juggling trials such that the resulting curves allow a decomposition of the form (1).

Sections 2 and 3 give a complete description of the registration procedure and details about implementation. In Section 4 we display the results of applying the procedure to the ten trials from the juggling data. Finally, in Section 5 we evaluate the perspectives of combining phase registration and biomechanical constraints.

2. Data and registration procedure

The pre-processed data [2] (lightly smoothed, centered, rotated and trimmed) is the starting point of our analysis, and is referred to as "observed data" or "raw data" in the remainder of the paper. The data indicate the position of the right index finger during juggling and is thus composed of three coordinates. We write $f(t) = (f_1(t), f_2(t), f_3(t))$, and let n denote the number of cycles. There are 10 signals/trials, all collected from the same person. The number of cycles per trial varies from 11 to 13.

The suggested registration procedure is applied to each trial separately, but on all three dimensions and all cycles simultaneously. The implementation details are described in Section 3, but, in short, the complete procedure is split into three steps:

1. **Warping** The observed signal consisting of several cycles is converted into a warped version $f \circ h$, where cycles are warped towards each other using a periodic average function as target for the registration procedure.
2. **Averaging** Based on the warped signal, $f \circ h$, a periodic average, denoted by $\mathcal{P}f$, is computed as a projection onto the (high-dimensional) space of periodic functions.
3. **Decomposition** The periodic average $\mathcal{P}f$ is decomposed into two periodic terms: a joint movement $\mathcal{J}$ belonging to a low-dimensional space, V, and a remainder $\mathcal{P}f - \mathcal{J}f$ with approximately constant length along the trajectory.

The complete procedure involves estimation of a warping function h, a periodic average, and a joint movement $\mathcal{J}f$. Notice that $\mathcal{P}f$ and $\mathcal{J}f$ are periodic per construction, and thus have no between-cycle variation. In particular, we only need to plot the curves on the interval corresponding to one cycle. On the other hand, the warped, but not averaged, curve $f \circ h$ may potentially show amplitude variation between cycles, but presumably only little phase variation, since that has been diminished by warping.

The second step involves projection onto a space of periodic three-dimensional functions. If this projection is denoted by Q_{per}, then $\mathcal{P}f = Q_{per}(f \circ h)$. If $\|\cdot\|$ is the standard L^2-norm and g is a three-dimensional curve, then

$$\frac{\|Q_{per}g\|}{\|g\|} = \sqrt{\frac{\|g\|^2 - \|g - Q_{per}g\|^2}{\|g\|^2}} = \sqrt{1 - \frac{\|g - Q_{per}g\|^2}{\|g\|^2}} \tag{2}$$

takes values in $[0, 1]$ and is a natural measure of the degree of periodicity in g. When data from different cycles are warped against each other as in step 1, we would expect a larger degree of periodicity compared to the raw data. Hence, comparison of $\frac{\|Q_{per}f\|}{\|f\|}$ and $\frac{\|Q_{per}(f \circ h)\|}{\|(f \circ h)\|}$ can be used to quantify the effect of warping on periodicity (see Section 4).

3. Implementation

This section describes technical details of the implementation of our registration procedure. The emphasis is on the decomposition step, since warping and averaging rely on existing techniques and software.

Let f denote a signal consisting of n complete juggling cycles. The duration of each cycle within a trial is rescaled to $[0, 1]$, then the same implementation can be used for all trials, even though the number of cycles are different.

Warping First, we expressed f in terms of 201 Fourier basis functions, and computed the orthogonal projection f_{per} on the space of periodic functions $L_{per,n}$ containing n replications of the same signal. Due to the Fourier basis

representation this amounts to keeping coefficients corresponding to harmonics of order $n, 2n, 3n, \ldots, Kn$ (where K is the largest K such that $Kn \leq 100$). Second, a time warping function h maximizing the coherence between $f \circ h$ and f_{per} was estimated. We used the minimal eigenvalue of a cross-product matrix with a roughness penalty on curvature of h as estimation criterion, see [3, Section 7.6]. In order to ensure a sufficient degree of smoothness of the warped signal $f \circ h$ we restricted h to the space spanned by 101 B-splines of order 5 with equally spaced break points. The roughness of the warping functions were controlled by penalizing the squared integral of second order derivatives. The robostness to the value of the penalty parameter λ was examined and for the results presented below we used $\lambda = 10^{-11}$ based on visual inspection.

Averaging The warped function $f \circ h$ was projected onto $L_{per,n}$ (see the paragraph on the warping step above). Hence, we obtain a periodic average of $f \circ h$, denoted $\mathcal{P}f$ and spanned by periodic harmonics.

Decomposition To implement the estimation of $\mathcal{J}f$ in step 3 it was convenient to expand all functions in terms of orthogonal complex exponentials. Denoting by a_k and $b_k, k = 1, 2, 3$, the three coordinate functions of the periodic average $\mathcal{P}f$ (known) and joint movement $\mathcal{J}f$ (to be estimated), we have expansions

$$a_k(t) = \sum_{j=-m}^{m} a_{k,j} \exp(i\omega jt), \quad b_k(t) = \sum_{j=-l}^{l} b_{k,j} \exp(i\omega jt)$$

and hence

$$a_k'(t) = \sum_{j=-m}^{m} i\omega j a_{k,j} \exp(i\omega jt), \quad b_k'(t) = \sum_{j=-l}^{l} i\omega j b_{k,j} \exp(i\omega jt).$$

Here $\omega = 2\pi n$ where n is the number of cycles.

We emphasize that $\mathcal{P}f$ has already been expressed in a finite Fourier basis, thus m and $a_{k,j}$ are all fixed and known at this point of the analysis, whereas the coefficients b_k should be estimated. For $l < m$ fixed, we collect the unknown parameters in θ:

$$\theta = \{b_{k,j} | k = 1, 2, 3, j = -l, \ldots, l\}$$

Some comments on the choice of l: The regularization assumption $l < m$ is necessary for identification, i.e., for the decomposition (1) to be unique since otherwise we could just let $\mathcal{J}f = \mathcal{P}f - c_0$ with $c_0 \in \mathbb{R}^3$ any fixed vector. For $l < m$ the joint movement $\mathcal{J}f$ belongs to a subspace of lower dimension than $\mathcal{P}f$, and the idea is to choose a small l, such that the joint movement is simple.

Recall that we aim at finding $\mathcal{J}f$ such that $\mathcal{P}f - \mathcal{J}f$ has approximately constant length; hence we want the derivative of the squared length to be approximately zero for all t:

$$D|\mathcal{P}f(t) - \mathcal{J}f(t)|^2 \approx 0.$$

This leads to the following criterion function to be minimized:

$$C(\theta) = \int_0^1 \left[D|\mathcal{P}f(t) - \mathcal{J}f(t)|^2\right]^2 dt \tag{3}$$
$$= \int_0^1 \left[D\sum_{k=1}^{3}(a_k(t) - b_k(t))^2\right]^2 dt$$
$$= 4\int_0^1 \left[\sum_{k=1}^{3} D(a_k(t) - b_k(t)) \cdot (a_k(t) - b_k(t))\right]^2 dt.$$

If we introduce the notation $e_{k,j} = a_{k,j} - b_{k,j}$ (with $b_{k,j} = 0, |k| > l$) for the Fourier coefficients of the difference $\mathcal{P}f - \mathcal{J}f$, and furthermore $c_{j_1,j_2} = \{\sum_{k=1}^{3} j_2 e_{k,j_1} e_{k,j_2}\}$ and let $j \in I_s$ if $j, s-j \in \{-m, \ldots, m\}$, then

$$C(\theta) = \int_0^1 \left[\sum_{s=-2m}^{2m} i\omega \sum_{j\in I_s} c_{s-j,j} \exp(i\omega st)\right]^2 dt.$$

Finally, if we let $d_s = \sum_{j\in I_s} c_{s-j,j}$ and use that $d_{-s} = -\overline{d_s}$ (complex conjugate), then we end up with the following simple formula for the criterion function

$$C(\theta) = -4\omega^2 \sum_{s=-2m}^{2m} d_s d_{-s} = 4\omega^2 \left\{|d_0|^2 + 2\sum_{s=1}^{2m} |d_s|^2\right\}. \tag{4}$$

The representation (4) makes it feasible to compute numerically the value and the gradient of the objective function as a function of θ to be used for the minimization algorithm. Since we are looking for a real valued estimate of the joint movement $\mathcal{J}f$, we found it convenient to reparameterize the problem in terms of a basis of sines and cosines. For the results below we used $l = 1$ corresponding to the joint movement being expressed in terms of first order harmonics only.

4. Results

We applied the registration procedure described above to each of the ten juggling trials. We will use trial 8 for detailed illustration, because the effect of the warping step was largest for this trial.

Warping and averaging Figure 1 shows the effect of steps 1 and 2 (warping and averaging) on trial 8. The vertical coordinate (z)of the raw data (dashed) is shown together with vertical coordinate of the periodic signal $\mathcal{P}f$ (solid). The raw signal does not exhibit much misalignment but the signal is indeed warped slightly. Notice how the warping is more pronounced towards the ends of the trial. The average curve $\mathcal{P}f$ for trial 8 is shown for each coordinate separately in the left part of Figure 2, and as a 3d-curve in the right part of the figure (solid curve).

For the raw data the degree of periodicity, cf. definition (2), was 88.0%, whereas for the warped data this number increased to 98.6%. All other trials had degrees of periodicity of 94.3% to 97.2% before warping and between 97.5%

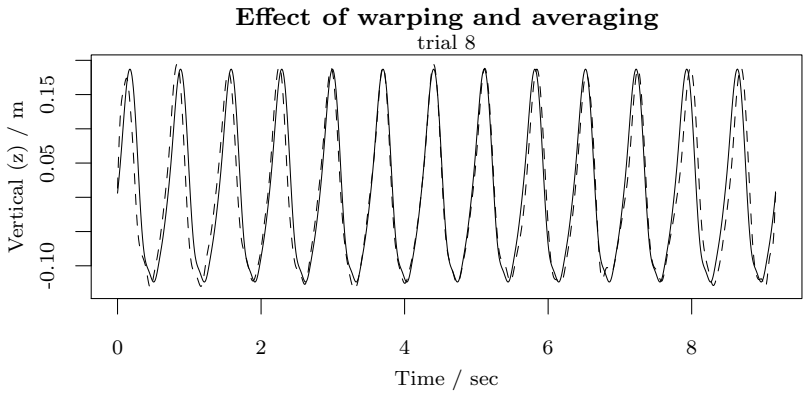

FIG 1. *Warping and averaging for trial 8. The dashed curve shows the z coordinate of the observed data, while the solid curve shows the z coordinate of* $\mathcal{P}f$.

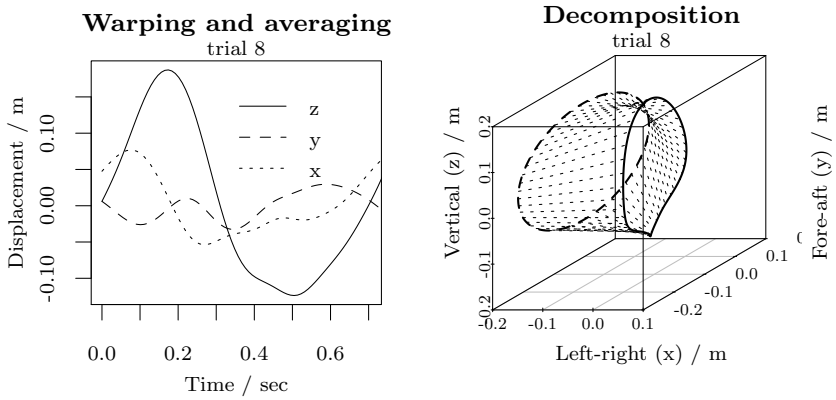

FIG 2. *Left: The three directions of the warped and averaged curve* $\mathcal{P}f$ *for trial 8. Right: 3d-illustration of the decomposition for trial 8. The solid curve shows the average* $\mathcal{P}f$*, the dashed curve shows the estimated joint movement curve* $\mathcal{J}f$*, and the dotted lines illustrate the trajectory of the difference* $\mathcal{P}f - \mathcal{J}f$ *(each dotted line correspond to a specific time point.)*

and 99.2% after warping. Hence, in general, only a limited amount of warping towards the periodic template was necessary. Visually, the raw and averaged trials were almost indistinguishable, except for trial 8 (see Figure 1).

The upper left, upper right and lower left plots of Figure 3 show the three coordinates of the warped curves $f \circ h$ for all ten trials, split into cycles and rescaled to the unit interval. The curves are coloured according to trial (but note that curves from different trials have not been aligned). In general, cycles within a trial are well aligned. Therefore the projection onto $L_{per,n}$ is a good representation of a trial. Note that the projections are similar across trials (-see the lower right part of Figure 3). The warping criterion gives less weight to coordinates with lower amplitude variation. This may explain why most misalignment is present in the y direction.

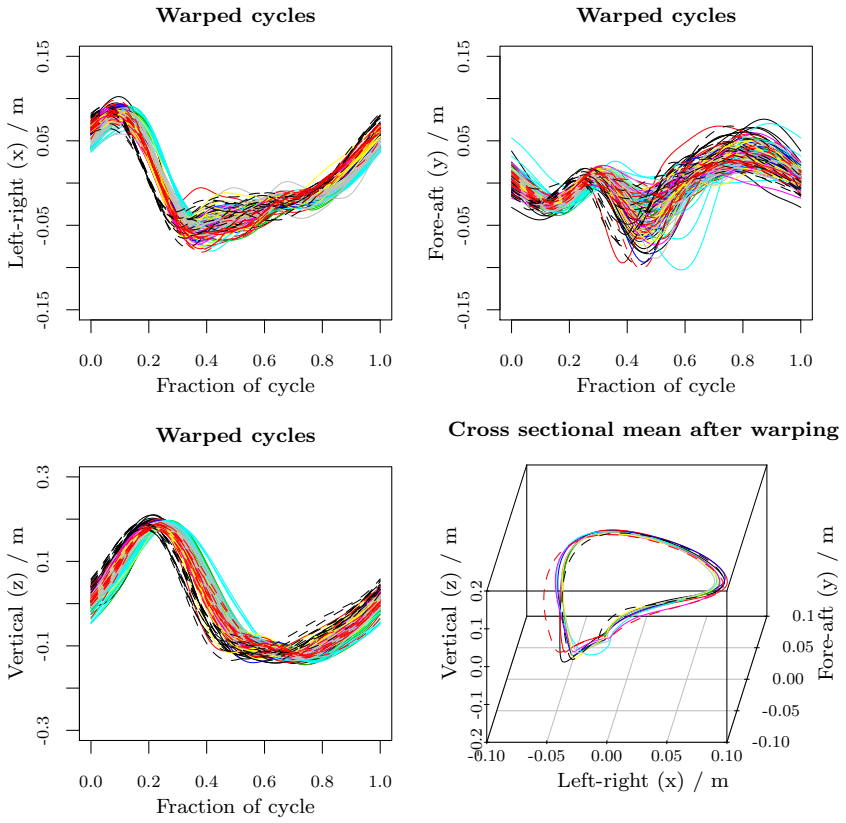

FIG 3. *Upper left, upper right and lower left: The three coordinates of the warped curves $f \circ h$ cut into individual cycles for each trial. For a trial with n cyc les, the complete curve was simply divided into n pieces of the same length, which was then rescaled to the unit interval. Cycles of the same colour and line type stem from the same trial. Lower right: 3d-scatterplot of the periodic average $\mathcal{P}f$ for all trials.*

Decomposition The estimated joint movement $\mathcal{J}f$ for trial 8 is shown as a dashed curve in the right part of Figure 2. Recall that the estimation procedure seeks the curve $\mathcal{J}f$ such that the vector $\mathcal{P}f - \mathcal{J}f$ has approximately constant length over the trajectory. This vector is illustrated by the dotted lines between the two curves, and its length varies from 0.179 m to 0.182 m for trial 8.

The decompositions for all curves are illustrated in Figure 4. The left part shows the length $|\mathcal{P}f - \mathcal{J}f|$ over the trajectories (scaled to the unit interval), and the right part shows the joint movements $\mathcal{J}f$. We make the following immediate observations from Figure 4: First, for all ten trials it was possible to obtain a function $\mathcal{J}f \in V$ such that the distance $|\mathcal{P}f - \mathcal{J}f|$ is approximately constant over time. This indicates that our simplistic biomechanical considerations leading to equation (1) characterizes some of the main features of the data generating mechanism. Second, the estimated length varies from 0.077 m

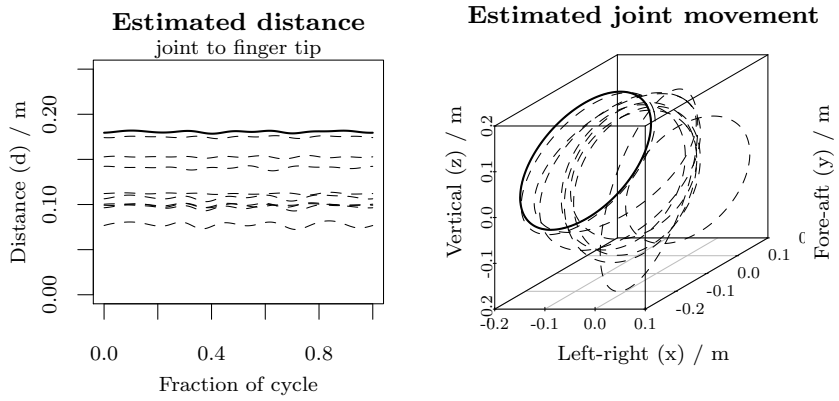

FIG 4. *Left: Estimated trajectory of distances, $|\mathcal{P}f - \mathcal{J}f|$, for all 10 trials. Right: Estimated joint movement, $\mathcal{J}f$, for all ten trials. In both plots the estimate correponding to trial 8 is shown as a solid curve.*

.

to 0.181 m across the ten trials. This is somewhat disappointing as we had hoped for an interpretation of this length as the length of a part of the hand or arm of the juggler. Third, the variation between the estimated joint movement curves is substantial. The decomposition restricts $\mathcal{J}f$ to be spanned by first order harmonics in all three directions. Allthough the curves are approximately elliptic they are different regarding angle and position.

5. Discussion

The purpose of the paper was to illustrate how the physical nature of a biomechanical system could be taken into account when removing phase variation of functional data from juggling. We have demonstrated that it is possible to warp all ten juggling trials such that the resulting structural mean over all cycles allows a decomposition as in (1).

The most striking observation is that the estimated distance from finger tip to joint, which should be an internal constant of the body anatomy, varies substantially across the ten trials. This complicates the physical interpretation of the estimated decomposition. Looking more carefully at the curves in the left part of Figure 4, there seems to be some common patterns in the deviations from constancy. Curves with low values of d seem to have peaks and valleys at the same time points (for example around 0.38 and 0.82), i.e. at the same time points of the juggling cycle. This indicates that our simple model might not have captured all features in the data.

A possible extension of the model would be to allow for more flexibility in the space V for the joint movement, i.e. by introducing harmonics of higher order in the basis for $\mathcal{J}f$. However, it seems more likely that adjustments from the idealized set-up given by (1) is taking place around the finger tip (far from the corpus) rather than at joints closer to the corpus. This suggest to relax the

focus on constant length of $\mathcal{P}f - \mathcal{J}f$. For example, the criterion function $C(\theta)$ in the decomposition step, see (3) and (4), could be adjusted to have a time-varying penalty on deviations from constancy. This would, however, complicate the optimization problem substantially.

In this connection, it should be mentioned that the numerical optimization problem for estimating the decomposition was more challenging than expected. The algorithm we used produced reliable estimates but was slow. This part of the implementation could be improved.

It is important to realize that amplitude and phase variation are bound to be intertwined, as an adjustment via a change in speed (phase) will most likely also change the amplitude. In relation to this, the complicated interplay between the estimation the warping function (step 1) and the estimation of the joint movement (step 3) should also be noticed. In particular, the space V for the joint movement is not invariant to warping (i.e. $g \in V$ does not imply that $g \circ h \in V$ for a warping function h). Too much warping of f may destroy the interpretation of the decomposition. This could be avoided by simultaneously estimating the warping function and the decomposition, i.e. to incorporate the warping (and averaging) step into the decomposition step.

Apart from the suggestions mentioned above, it would be interesting to examine the robustness of the registration. Simulations could clarify the importance of the explicit form of the underlying signal on the performance of the registration procedure. Moreover, it would be interesting to fit a common joint movement curve to all s, and see the effect on the corresponding position vectors $\mathcal{P}f - \mathcal{J}f$ and their lengths.

Acknowledgements

We acknowledge the Mathematical Biosciences Institute, Ohio, for supporting our participation in the workshop on "Statistics of Time Warpings and Phase Variations".

References

[1] Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103**, 483, 1155–1165. http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000517. MR2528838

[2] Ramsay, J. O., Gribble, P., and Kurtek, S. (2014). Description and processing of functional data arising from juggling trajectories. *Electron. J. Statist.* **8**, 1811–1816, Special Section on Statistics of Time Warpings and Phase Variations.

[3] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Second ed. Springer, New York. MR2168993

[4] Schaal, S., Atkeson, C. G., and Sternad, D. (1996). One-handed juggling: A dynamical approach to a rhythmic movement task. *Journal of Motor Behavior* **28**, 2, 165–183.

IV

Effects of Dietary Protein and Glycaemic Index on Biomarkers of Bone Turnover in Children

Stine-Mathilde Dalskov[1], Martha Muller[2], Christian Ritz[1], Camilla T Damsgaard[1], Angeliki Papadaki[3], Wim H.M. Saris[4], Arne Astrup[1], Kim Fleischer Michaelsen[1] and Christian Mølgaard[1]
on behalf of the Diet, Obesity and Genes (DiOGenes) project

[1]Department of Nutrition, Exercise and Sports
University of Copenhagen

[2]Department of Mathematical Sciences
University of Copenhagen

[3]Department of Social Medicine
University of Crete

[4]NUTRIM, Department of Human Biology
Maastricht University Medical Center

Publication details

Published (2014) in *British Journal of Nutrition*. Vol. 111, No. 7, 1253–1262.

British Journal of Nutrition (2014), **111**, 1253–1262
doi:10.1017/S0007114513003760

Effects of dietary protein and glycaemic index on biomarkers of bone turnover in children

Stine-Mathilde Dalskov[1]*, Martha Müller[2], Christian Ritz[1], Camilla T. Damsgaard[1], Angeliki Papadaki[3], Wim H. M. Saris[4], Arne Astrup[1], Kim Fleischer Michaelsen[1] and Christian Mølgaard[1] on behalf of DiOGenes†

[1]*Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Roligheds vej 30, 1958 Frederiksberg C, Denmark*
[2]*Department of Mathematical Sciences, Faculty of Science, University of Copenhagen, Copenhagen, Denmark*
[3]*Department of Social Medicine, Preventive Medicine and Nutrition Clinic, University of Crete, Heraklion, Greece*
[4]*Department of Human Biology, NUTRIM, Maastricht University Medical Center, Maastricht, The Netherlands*

(Submitted 25 February 2013 – Final revision received 16 October 2013 – Accepted 16 October 2013 – First published online 6 February 2014)

Abstract

For decades, it has been debated whether high protein intake compromises bone mineralisation, but no long-term randomised trial has investigated this in children. In the family-based, randomised controlled trial DiOGenes (Diet, Obesity and Genes), we examined the effects of dietary protein and glycaemic index (GI) on biomarkers of bone turnover and height in children aged 5–18 years. In two study centres, families with overweight parents were randomly assigned to one of five *ad libitum*-energy, low-fat (25–30 % energy (E%)) diets for 6 months: low protein/low GI; low protein/high GI; high protein/low GI; high protein/high GI; control. They received dietary instructions and were provided all foods for free. Children, who were eligible and willing to participate, were included in the study. In the present analyses, we included children with data on plasma osteocalcin or urinary N-terminal telopeptide of collagen type I (U-NTx) from baseline and at least one later visit (month 1 or month 6) (n 191 in total, n 67 with data on osteocalcin and n 180 with data on U-NTx). The level of osteocalcin was lower (29·1 ng/ml) in the high-protein/high-GI dietary group than in the low-protein/high-GI dietary group after 6 months of intervention (95 % CI 2·2, 56·1 ng/ml, $P=0·034$). The dietary intervention did not affect U-NTx ($P=0·96$) or height ($P=0·80$). Baseline levels of U-NTx and osteocalcin correlated with changes in height at month 6 across the dietary groups ($P<0·001$ and $P=0·001$, respectively). The present study does not show any effect of increased protein intake on height or bone resorption in children. However, the difference in the change in the level of osteocalcin between the high-protein/high-GI group and the low-protein/high-GI group warrants further investigation and should be confirmed in other studies.

Key words: Children: Bone turnover: Dietary glycaemic: Osteocalcin: Dietary protein: index: DiOGenes: Randomised controlled trials

Optimal growth and skeletal development during childhood and young adulthood is crucial for avoiding low bone mass and osteoporosis later in life. The influence of dietary protein on bone status has been debated for decades, but remains controversial. Different study designs have been used to investigate this, but conflicting results have been reported[1–4]. Experimental studies on the effects of dietary protein on Ca excretion and absorption have been carried out in adults. Based on these studies, high protein intake, especially that of animal origin, has been hypothesised to affect bone mineralisation adversely by increasing bone resorption and thereby urinary Ca excretion[5–10]. However, some studies in adults[11–14], but not all[5,6,15,16], have shown compensatory increased Ca absorption with increasing intake of dietary protein. When looking at measures of bone status, observational studies in adults[2,17] and children[18–22] and protein supplementation trials in adults[2] have shown a small positive effect of dietary protein on bone status. In a 7 d intervention study in 8-year-old boys, Budek *et al.*[23] found that a high intake of protein from milk, but not from meat, decreased

Abbreviations: DiOGenes, Diet, Obesity and Genes; E%, percentage of energy; GI, glycaemic index; HGI, high glycaemic index; HP, high protein; LGI, low glycaemic index; LP, low protein; U-NTx, urinary N-terminal telopeptide of collagen type I.

* **Corresponding author:** S. Dalskov, fax +45 35 33 24 83, email smd@life.ku.dk

† DiOGenes is the acronym of the project 'Diet, Obesity and Genes'.

British Journal of Nutrition

bone turnover as measured by serum osteocalcin and serum C-terminal telopeptides of type I collagen. So far, no long-term trial in children has been conducted to assess the effect of dietary protein on bone turnover or bone status.

Bone turnover can be assessed by biomarkers in the blood and urine. Osteocalcin is a non-collagenous extracellular matrix protein produced by osteoblasts. It contains three glutamic acid residues, which are post-translationally carboxylated to increase their affinity for mineral ions. In contrast, partial or no carboxylation makes osteocalcin more susceptible to be released from osteoblasts into the circulation[24]. Traditionally, osteocalcin measured in serum or plasma has been considered as a marker of bone formation. However, genetic knockout studies have indicated no direct relationship between osteocalcin and mineral deposition events, but have rather shown that osteocalcin participates in the regulation of the mineralisation process[25]. Urinary N-terminal telopeptide of collagen type I (U-NTx) is a breakdown product released during the resorption of bone, and is used as a marker of bone resorption. Biomarkers of bone formation and resorption are normally closely related, and the balance between them may reflect whether a higher turnover results in increased or reduced bone mass.

The primary aim of DiOGenes (Diet, Obesity and Genes), a large-scale, European randomised intervention trial, was to examine the effects of diets varying in protein content and glycaemic index (GI) on weight maintenance in adults after a weight-loss period. However, the children of these adults were also included in the study. To assess whether a high-protein diet could be detrimental to bone health in children, the bone markers osteocalcin and U-NTx were analysed in the children's blood and urine samples, respectively. The possible positive effects of protein on body-weight regulation and the risk markers of CVD in adults[26] should be weighed up against concerns about the safety of high-protein diets. The question then arises: what about the GI part of the DiOGenes study – does that mean anything to bone health? Since the initiation of the study, several studies have examined the connection between energy metabolism (including insulin signalling) and bone metabolism[27]. Looking at some of the findings in these studies, we postulate that a diet with a low GI might benefit not only body-weight regulation, but also bone growth.

The aim of the present paper was to examine the effects of dietary protein and GI on bone turnover based on blood (osteocalcin) and urine (U-NTx) analyses in children from two of the participating centres in the DiOGenes study. To elucidate the relationship between osteocalcin/U-NTx and bone growth, we also examined the relationship between the baseline levels of osteocalcin and U-NTx and the following changes in height across dietary groups, and examined the relationship between dietary group and changes in height during the intervention. All analyses presented in the study are *post hoc* analyses.

Experimental methods

Experimental design

Children and their parents were enrolled at eight centres across Europe. In the present study, only data from the centres in Copenhagen and Maastricht were included. These two centres (the so-called 'shop centres') did run a more strictly controlled version of the intervention, providing all families with foods for free from specially designed shops, and dietary data indicated that the intervention was only successful among children at these centres.

The study was conducted according to the guidelines in the Declaration of Helsinki, and all procedures involving human subjects were approved by the local ethical committees in the respective countries. Written informed consent was obtained from all custody holders of the child and from the child, when considered mature enough to understand the procedure. During the screening visit, children and their parents were asked to choose between participation in all planned examinations ('full' protocol) or only take part in some of the examinations, excluding blood and urine samples. Only children accepting the full protocol were included in the present study. The trial was registered in the Clinical Trials database (ClinicalTrials.gov no. NCT00390637).

In brief, families with at least one child aged 5–18 years and one or two overweight or obese parents reaching an initial weight loss of ≥8 % of their body weight after an 8-week low-energy diet (3347 kJ (800 kcal)) were randomised to one of five intervention diets for 6–12 months: low protein (LP)/low GI (LGI); LP/high GI (HGI); high protein (HP)/LGI; HP/HGI; control. The randomisation was stratified according to centre, the number of eligible parents in each family and the number of parents with a BMI > 34 kg/m^2 in each family. The five intervention diets were all *ad libitum* (no restriction on total energy intake), low-fat (25–30 E%) diets. The target dietary differences were 15 GI units between the LGI and HGI diets and 13 E% points from protein between the LP (10–15 E%) and HP (23–28 E%) diets. Families randomised to the control diet were instructed to eat according to some general dietary guidelines: eat fruit and vegetables several times per d; eat fish several times per week; eat potatoes, rice or pasta and whole-grain bread every day; limit the sugar intake especially from liquids, candy and cakes; eat less fat especially from dairy products and meat; eat varied food and keep the weight stable. The dietetic counselling was focused on fat quality and amount, and less on carbohydrate intake and sources, to prevent the control group from becoming just another LP/LGI group.

The participating families were provided with free foods from a specially designed shop during 6 months of intervention. For more details about the study design and the dietary intervention strategies used, see Larsen *et al.*[28] and Moore *et al.*[29].

At baseline, two examination days were planned for the children: one before and one after their parents' low-energy diet. For logistic reasons, the majority of children had these two visits combined in one visit around the scheduled second examination day. In the present study, the term 'baseline' refers to latest of the two visits, whenever two separate visits were made. In addition to the baseline visit, examinations were scheduled for 1 month and 6 months after the start of the intervention.

British Journal of Nutrition

Study subjects

Children were excluded from the study if they used prescription medication, suffered from diseases or conditions that might influence the outcome of the study, followed a special diet (e.g. vegetarian or lactose free) or practised elite sports. Children with data from baseline and from at least one of two subsequent visits (month 1 or month 6) were included in the present analyses.

Examinations

Examinations were carried out in the morning after the child had fasted (except for 350–500 ml water) for at least 4 h. Height (to the nearest 0·5 cm) and body weight (to the nearest 0·1 kg) were recorded at each examination day. Children were weighed wearing light clothing. Sex- and age-specific *z*-scores for height and BMI were calculated using WHO AnthroPlus software[30,31].

On the examination days, the children delivered a spot urine sample, avoiding the first morning urine. A blood sample was drawn from an antecubital vein. It was not possible to perform blood sampling and urine collection at exactly the same hour in the morning each time a child came in for examination. For the analysis of osteocalcin, Li-heparinised blood was centrifuged within 1 h after collection at 2500 ***g*** for 15 min at 4°C, and plasma was stored at −80°C until analysis. Osteocalcin was measured on an Immulite 2500 using a solid-phase, two-site chemiluminescent immunometric assay (Siemens Medical Solutions Diagnostics, DPC Scandinavia). U-NTx was analysed using an ELISA (Osteomark NTx Urine kit; Wampole Laboratories, Orion Diagnostica). Urinary creatinine was measured by a colorimetric assay on a Vitros 950 analyser (Ortho-Clinical Diagnostics, Johnson & Johnson Medical). To adjust for the concentration of the urine, U-NTx, expressed in nM-bone collagen equivalents, was divided by urinary creatinine in mM. Intra- and inter-assay CV were 3·5 and 5·4 % (osteocalcin) and 4·0 and 7·6 % (U-NTx), as reported by the manufacturer. No information was available for creatinine.

Children and their parents were instructed to register the dietary intake of the children for three consecutive days (two weekdays and one weekend day) at baseline, month 1 and month 6. They were equipped with weighing scales (Soehnle 1208 Actuell Backnang; Leifheit AG), and were instructed to weigh all foods and beverages consumed during the registration periods and to provide cooking methods and recipes for composite meals. When weighing was not possible, the children and their parents were instructed to record the dietary intake in household measures. If the children were not able to perform the dietary registrations themselves, their parents were asked to assist them. The principles of analysis of dietary records in DiOGenes have been described elsewhere[28,32]. Intakes of protein, carbohydrates and fat were expressed as E%. Since energy intake is dependent on sex, age and body size, it was evaluated relative to an estimated BMR calculated using the formulas suggested by Henry[33].

Statistical methods

Children with data from baseline and from at least one later visit (month 1, month 6 or both) were included in the present analyses.

Baseline characteristics are presented for children included in the analyses of osteocalcin, U-NTx and either or both of these. For children whose 6-month data were available, changes in height and BMI *z*-score over this 6-month period are also given.

Dietary intake was compared at baseline, month 1 and month 6 between the five dietary groups. Raw data are presented as medians and interquartile ranges. Data were analysed using ANCOVA, with centre as a random effect detecting variations between centres. Outcomes were transformed if necessary to meet model requirements. *P* values based on likelihood ratio tests are reported for the overall group effect; in addition, *P* values, estimates and 95 % CI are given for selected pairwise comparisons.

ANCOVA was used for evaluating differences between diets over time. Initially, the effects of dietary protein and GI were assumed to modify how the levels of osteocalcin and U-NTx have changed linearly over time since randomisation (effect modification). To evaluate whether diet effects were modified by sex, an additional sex×diet interaction term was included in the model. To adjust for anticipated child-specific differences, baseline values of the bone markers, BMI *z*-score at each of the time points and sex-specific linear and quadratic trends in age were included in the model. The adjustment for the BMI *z*-score in the present analyses was made so that the results were not primarily caused by differences in weight change based on the different diets. Cluster effects were addressed by means of random effects that were included for children, families and centres. Thus, multi-level linear-mixed ANCOVA models were used. Model checking was based on residual plots and normal probability plots. If needed, data were logarithmically transformed to meet model assumptions and, subsequently, estimates were transformed to the original scale. Likelihood ratio tests were used to assess the combined effects and interaction terms, whereas approximate *t* tests were used for pairwise comparisons between the dietary groups. Adjustment for multiple *P* values was based on the single-step method[34]. Likewise, a linear-mixed ANCOVA model was used to examine whether diet influenced height. However, no adjustment for the BMI *z*-score over time was made since height is an integral part of BMI. Finally, a linear model was used to investigate whether baseline levels of osteocalcin and U-NTx could predict height at month 6 across all the dietary groups, when adjusting for height at baseline. Estimates and 95 % CI for significant effects of the linear relationships are reported.

The significance level was set at $P<0{\cdot}05$ (two-sided). The statistical environment R version 2.15.1[35] and, in particular, the extension packages lme4 and multcomp, as well as STATA 12.0[36] were used for the analyses.

Results

Data from 191 children were included in the present paper: *n* 67, osteocalcin analyses; *n* 180, U-NTx analyses; *n* 56,

both the analyses. The progress of the study participants from screening to month 6 and the selection criteria for the analyses of osteocalcin are illustrated in Fig. 1.

Characteristics

Baseline characteristics and 6-month changes in height-for-age *z*-scores and BMI-for-age *z*-scores are presented in Table 1 for the groups of children included in the analyses of either osteocalcin or U-NTx, osteocalcin, U-NTx or both. The median BMI-for-age *z*-score for the children included in either of the two analyses was 1·13, which was above the cut-off (1·0) for overweight according to the WHO growth reference[30]. Having a median height-for-age *z*-score of 0·77, the children were not only thicker, but also taller than the WHO growth reference. None of the children was underweight, which is defined as a BMI-for-age *z*-score less than −2, and none of the children was stunted that is defined

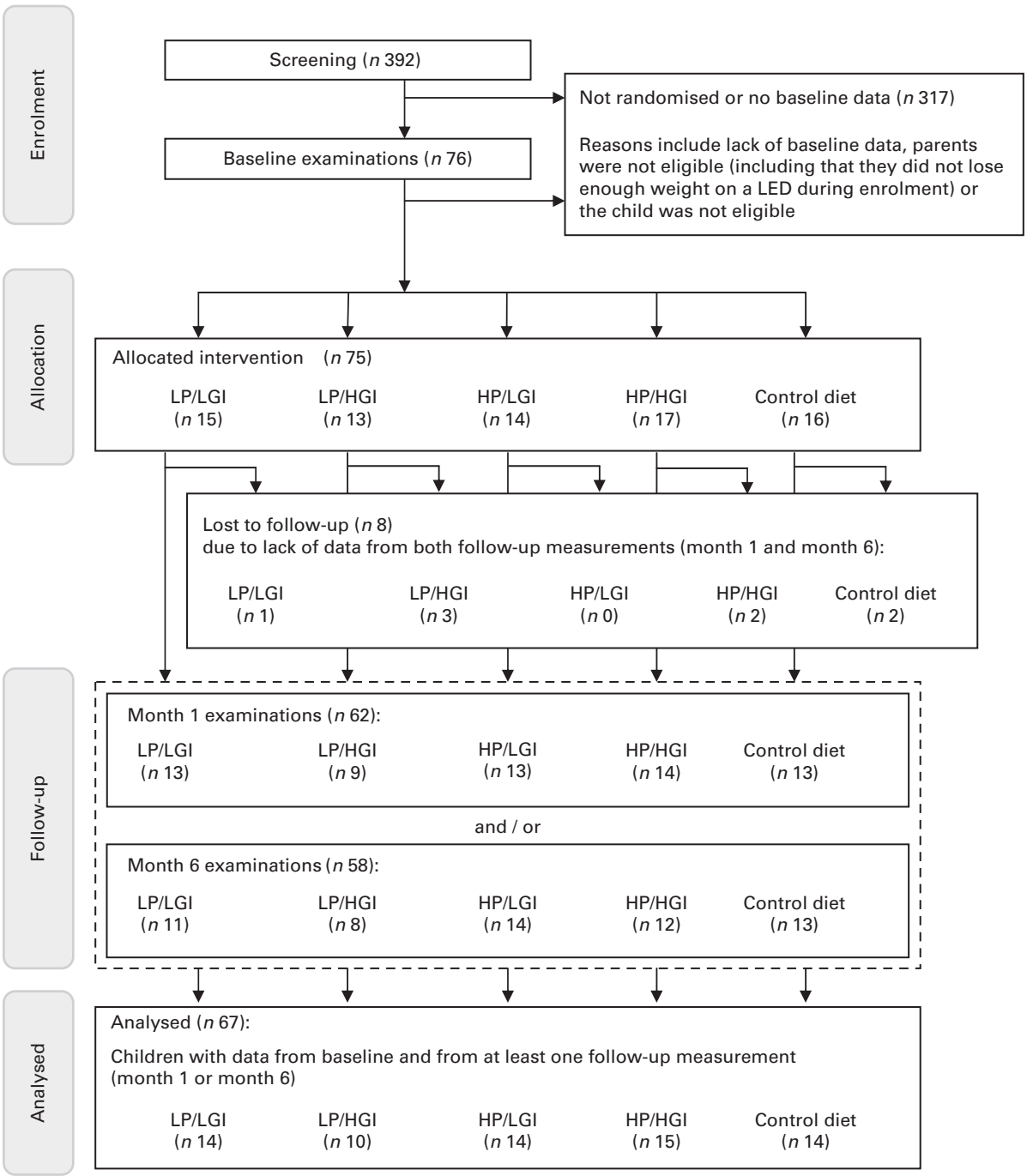

Fig. 1. Flow diagram illustrating the progress of the study participants from screening to month 6, and the selection criteria for the analyses of osteocalcin. LED, low-energy diet; LP, low protein; LGI, low glycaemic index; HGI, high glycaemic index; HP, high protein.

British Journal of Nutrition

British Journal of Nutrition

Table 1. Characteristics of the study participants included in the analyses of either osteocalcin or urinary N-terminal telopeptide of collagen type I (U-NTx), osteocalcin, U-NTx or both (Number of participants, median values and interquartile ranges (IQR))

	Either			Osteocalcin			U-NTx			Both		
	n	Median	IQR	*n*	Median	IQR	*n*	Median	IQR	*n*	Median	IQR
Baseline												
Age (years)	191	11·9	9·6–14·3	67	12·4	10·6–15·2	180	12·0	9·5–14·4	56	12·9	10·6–15·2
Height-for-age *z*-score	191	0·77	0·28–1·43	67	1·00	0·49–1·53	180	0·76	0·24–1·42	56	0·96	0·48–1·47
BMI-for-age *z*-score	191	1·13	0·21–1·93	67	1·39	0·50–2·07	180	1·12	0·20–1·90	56	1·37	0·42–2·00
U-NTx (nM-BCE/mM-u-crea)	181	320	186–534	57	210	136–335	180	320	182–535	56	217	133–347
Osteocalcin (ng/ml)	70	52·3	27·2–75·5	67	54·9	27·2–76·3	59	47·3	26·8–75·5	56	52·3	27·0–76·0
6-month changes												
Δ Height-for-age *z*-score	163	−0·03	−0·13–0·09	60	−0·02	−0·13–0·12	154	−0·03	−0·15–0·10	51	−0·02	−0·16–0·13
Δ BMI-for-age *z*-score	162	−0·06	−0·36–0·12	60	−0·20	−0·50–0·07	153	−0·05	−0·34–0·12	51	−0·23	−0·49–0·07

BCE, bone collagen equivalents; u-crea, urinary creatinine.

as a height-for-age *z*-score less than −2. Median changes in the height-for-age *z*-score during the 6-month intervention period were close to 0, and thus it could be considered within normal limits. Baseline median values of osteocalcin and U-NTx for boys and girls at different ages are given in Figs. 2 and 3, respectively. The levels of the biomarkers of bone turnover were lowest among the oldest children.

Dietary intakes

Among the included children, 85, 83 and 45% registered their dietary intake at baseline, month 1 and month 6, respectively. For the four dietary groups whose baseline characteristics are given in Table 1, these numbers varied from 85 to 91% at baseline, 82 to 85% at month 1 and 44 to 54% at month 6.

Dietary intakes in the different dietary groups were not different at baseline (Table 2). Dietary GI was higher at both month 1 (8·3 (95% CI 6·1, 10·5) GI units, $P<0·001$) and month 6 (7·2 (95% CI 4·5, 9·9) GI units, $P<0·001$) in the HP/HGI group compared with the HP/LGI group, while the GI was higher only at month 1 (5·8 (95% CI 3·7, 7·8) GI units, $P<0·001$) in the LP/HGI group compared with the LP/LGI group. The E% from protein was higher at both month 1 (5·2 (95% CI 3·6, 6·7)% points, $P<0·001$) and month 6 (6·3 (95% CI 3·6, 9·1)% points, $P<0·001$) in the HP/LGI group compared with the LP/LGI group, and the same was the case when comparing the HP/HGI and LP/HGI groups at month 1 (4·0 (95% CI 2·4, 5·5)% points, $P<0·001$) and month 6 (6·5 (95% CI 3·9, 9·1)% points, $P<0·001$).

Osteocalcin

A total of sixty-seven children were included in the osteocalcin analyses (Fig. 1). Of these, fifty-four children provided follow-up data from both month 1 and month 6, nine children from month 1 only and four children from month 6 only. After 6 months of intervention, a close-to-significant change in the level of osteocalcin of −16·5 (95% CI −33·7, 0·74) ng/ml ($P=0·06$) was found in the HP/HGI group, whereas the corresponding change in the level of osteocalcin of 12·6 ng/ml in the LP/HGI group was not different from 0 (95% CI −8·2, 33·4) ng/ml ($P=0·23$) (Fig. 4). Consequently, after 6 months of intervention, the overall difference in the level of osteocalcin between the HP/HGI and LP/HGI groups was 29·1 (95% CI 2·2, 56·1) ng/ml ($P=0·034$). There were no differences between the LP/HGI and LP/LGI ($P=0·45$), HP/LGI and LP/LGI ($P=0·40$) and HP/HGI and HP/LGI ($P=0·46$) groups after 6 months of intervention. There was no effect modification of diet × sex on osteocalcin ($P=0·71$).

Urinary N-terminal telopeptide of collagen type I

A total of 180 children were included in the U-NTx analyses. Of these, 123 children provided follow-up data from both month 1 and month 6, thirty-seven children from month 1 only and twenty children from month 6 only. There was no effect modification of diet on U-NTx ($P=0·96$).

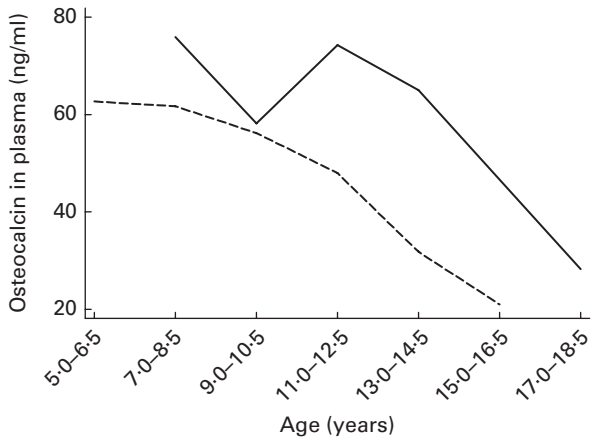

Fig. 2. Median values of osteocalcin for boys (——) and girls (----) at different ages. Subjects were categorised into seven age categories due to the low number of subjects at some ages.

Height

There was no effect of diet on height ($P=0{\cdot}80$). Baseline levels of both osteocalcin and U-NTx were strongly correlated with height at month 6, adjusted for baseline height across the dietary groups ($P<0{\cdot}001$ and $P=0{\cdot}001$, respectively). For every 10 ng/ml increase in the level of osteocalcin at baseline, children grew on average 0·3 cm more during the following 6 months, and for every 100 nM-bone collagen equivalents/mM-creatinine increase in the level of U-NTx at baseline, children grew on average 0·2 cm more during the following 6 months.

Discussion

The present sub-study of the DiOGenes study is the first randomised controlled trial to assess the effects of dietary protein and GI on bone turnover in children. The observed difference in the effects of the HP/HGI and LP/HGI diets on the bone marker osteocalcin (but not between the corresponding LGI diets) could point to a modulating effect of the GI on the effects of dietary protein on bone turnover. However, the diet had no effect on bone resorption and height.

To the best of our knowledge, only one randomised trial has investigated the relationship between dietary protein intake and bone turnover in children. It has shown that an increased intake of protein from milk during 7 d decreased bone turnover in 8-year-old boys as measured by serum osteocalcin and serum C-terminal telopeptides of type I collagen (a measure of bone resorption) when compared with a similar increase in protein intake from meat. Thus, the decrease in bone turnover was not due to protein as such, but to milk proteins or some other component in milk, e.g. Ca[23]. However, in the present study, all children were instructed to eat or drink dairy products corresponding to 0·5 litres of milk daily, and thus the achieved difference in protein between the HP and LP groups is expected to be derived primarily from non-dairy products (meat, nuts and cereals). As in the study in 8-year-old boys by Budek *et al.*[23], studies in postmenopausal women have not found any effect of meat protein on markers of bone turnover[11,37].

In an observational study of 17-year-old children, Budek *et al.*[38] found that milk protein was positively associated with size-adjusted bone mineral content, while no association was observed for meat protein. In another observational study, Remer *et al.*[39] found that urinary N excretion (a biomarker for protein intake) in 6 to 18-year-old children was a positive predictor of forearm bone mineral content, cortical area, strength strain index and periosteal circumference, but not of bone mineral density based on peripheral quantitative computed tomography.

Considering the apparently different effects of milk protein and meat protein on bone turnover, it cannot be excluded that a decrease in bone turnover due to the intake of dairy products is responsible for the beneficial effects of dairy protein or total protein on bone status in the aforementioned observational studies. If that is the case, then a decline in level of osteocalcin in the HP/HGI group may not be detrimental to bone health (maybe even the opposite). However, we wonder whether the observed decline in the level of osteocalcin without a corresponding decrease in the level of U-NTx indicates a decreased bone turnover, or rather an unbalanced bone turnover with a decrease in the formation part of the modelling and remodelling processes. The latter could have detrimental effects on bone health in these children.

Biomarkers of bone turnover have the advantage that they are more sensitive to short time exposure than measures of bone status. In the present study, the first post-baseline measurement was after 1 month of intervention. According to the literature, one should expect to detect changes in the measures of bone resorption before changes in the measures of bone formation. The full response is typically seen within 1–3 months for the markers of bone resorption *v.* within 6–9 months for those of bone formation[40]. Thus, the lack of the effect of the dietary intervention in the present study on the bone resorption marker U-NTx cannot be due to a too short follow-up. The different mediums used to measure

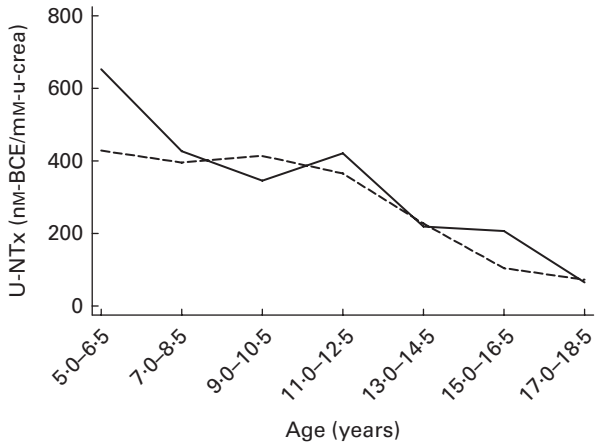

Fig. 3. Median values of urinary N-terminal telopeptide of collagen type I (U-NTx) for boys (——) and girls (----) at different ages. Subjects were categorised into seven age categories due to the low number of subjects at some ages. BCE, bone collagen equivalents; u-crea, urinary creatinine.

British Journal of Nutrition

Table 2. Dietary intakes at baseline, month 1 and month 6 across the dietary groups
(Number of participants, median values and interquartile ranges (IQR))

		Baseline				Month 1				Month 6			
	Dietary group	*n*	Median	IQR	*P*	*n*	Median	IQR	*P*	*n*	Median	IQR	*P*
EI:BMR*	LP/LGI	38	1·45	1·23–1·58		33	0·94	0·74–1·23		11	1·29	0·94–1·41	
	LP/HGI	30	1·31	1·08–1·57		31	1·11	0·80–1·28		12	1·24	1·09–1·47	
	HP/LGI	33	1·42	1·15–1·59	0·633	28	1·11	0·96–1·22	0·069	21	1·24	1·09–1·47	0·217
	HP/HGI	27	1·32	1·09–1·47		29	0·90	0·70–1·05		17	0·94	0·78–1·16	
	Ctr	35	1·31	1·15–1·58		32	1·11	0·85–1·27		21	1·17	0·97–1·34	
Carbohydrates (E%)	LP/LGI	38	55·5	50·8–57·6		34	62·8[a]	57·4–66·3		11	63·0[a]	54·1–68·0	
	LP/HGI	30	52·2	49·9–56·7		32	59·7[a,c]	53·6–65·2		13	57·5[a]	53·9–62·1	
	HP/LGI	33	54·1	48·9–57·1	0·274	28	50·4[b]	47·8–58·1	<0·001	21	50·7[b,c]	48·3–54·6	<0·001
	HP/HGI	27	50·7	46·6–56·7		30	53·6[b]	49·6–55·6		20	48·0[c]	44·3–54·8	
	Ctr	35	53·3	50·5–56·5		34	56·7[c]	52·2–62·7		21	54·4[b]	47·8–59·7	
Fat (E%)	LP/LGI	38	30·9	28·1–35·6		34	21·7	18·8–27·4		11	21·9	17·9–25·6	
	LP/HGI	30	32·4	28·5–34·4		32	24·3	20·4–31·2		13	27·7	22·5–29·9	
	HP/LGI	33	32·7	28·8–36·2	0·355	28	27·4	24·5–30·8	0·054	21	28·6	24·7–31·0	0·368
	HP/HGI	27	34·5	28·2–37·5		30	26·5	21·8–29·6		20	28·8	25·9–33·5	
	Ctr	35	32·4	28·5–35·8		34	25·8	22·5–28·6		21	29·9	20·5–34·7	
Protein (E%)	LP/LGI	38	14·2	12·2–15·7		34	15·3[a]	13·9–17·3		11	13·9[a]	12·5–18·0	
	LP/HGI	30	14·5	13·3–16·6		32	16·0[a]	13·2–17·7		13	14·3[a]	12·0–18·1	
	HP/LGI	33	13·4	11·5–15·8	0·220	28	18·4[b]	16·8–24·3	<0·001	21	17·7[b]	16·1–23·6	<0·001
	HP/HGI	27	14·4	12·9–17·5		30	20·1[b]	17·5–23·3		20	20·4[b]	19·7–24·8	
	Ctr	35	14·3	11·5–16·1		34	15·6[a]	13·5–19·4		21	17·1[c]	14·9–19·8	
GI (units)	LP/LGI	38	61·6	59·6–63·9		34	57·3[a]	53·6–60·8		11	59·8[a,c]	55·5–61·9	
	LP/HGI	30	63·2	60·7–65·6		32	62·7[b]	60·2–64·8		13	62·1[a,b]	59·6–64·3	
	HP/LGI	33	61·2	58·3–64·4	0·267	28	56·1[a]	54·8–59·1	<0·001	21	55·9[c]	53·3–59·8	<0·001
	HP/HGI	27	62·4	59·6–65·8		30	64·4[c]	60·8–69·7		20	63·9[b]	61·9–65·5	
	Ctr	35	61·8	58·7–64·6		34	61·4[b]	58·1–64·8		21	60·5[a]	57·4–63·6	

EI, energy intake; LP, low protein; LGI, low glycaemic index; HGI, high glycaemic index; HP, high protein; Ctr, control; E%, percentage of energy.
[a,b,c]Adjusted median values within a column with unlike superscript letters were significantly different at month 1 and month 6 ($P<0·05$; ANCOVA).
* BMR could not be estimated for five children at month 1 and four children at month 6 because of missing data for height and weight.

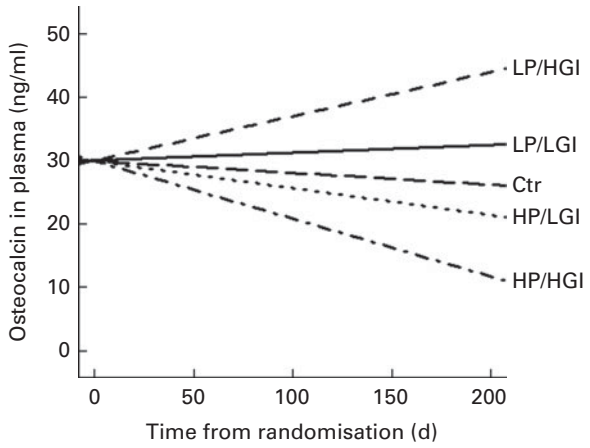

Fig. 4. Osteocalcin over time in the different dietary groups. There was a significant difference between the low-protein (LP)/high-glycaemic index (HGI) group and the high-protein (HP)/HGI group (P=0·034). LGI, low glycaemic index; Ctr, control; HP, high protein.

the levels of osteocalcin and U-NTx (blood *v.* urine) could be an explanation for the different results obtained for U-NTx and osteocalcin. U-NTx can be measured in both urine and blood. In the DiOGenes study, more children were willing to participate in the urine sampling than in the blood sampling, and thus the U-NTx results reflected a larger fraction of the children. However, the larger variability of measures in the urine than in the blood may offset this larger representativeness of the U-NTx data.

When comparing the levels of U-NTx and osteocalcin for age in this population with those found in other studies, the overall pattern is similar. Equivalent to the study by Mora *et al.*[41], we found that the U-NTx: creatinine ratio is approximately stable between 5 and 12 years, and then after about 12 years of age, it falls abruptly. Also, the absolute values are very similar in the two populations. With regard to osteocalcin, the present dataset is too small for comparisons of the effects of sex and age with those found in other populations such as that of van der Sluis *et al.*[42].

Results on the biomarkers of bone turnover are difficult to interpret, particularly in growing children. Concentrations cannot be directly translated into amounts of bone gained or lost, and it is not known whether the different biomarkers mainly reflect growth in size, growth in mass or both[43]. A high bone turnover in late adulthood is considered unfavourable as it results in net bone loss, while in children, a high bone turnover may simply be the result of a high growth velocity. Finally, changes in measures of bone status may not even presuppose changes in biomarkers of bone turnover as indicated by a study by Cadogan *et al.*[44], where a milk intervention increased bone mass accretion in 12-year-old girls without affecting bone turnover markers.

We found that baseline levels of both U-NTx and osteocalcin were strongly correlated with changes in height during the following 6 months, and thus they indeed seem to be measures of bone growth in children. However, as previously reported among DiOGenes children[38], diets did not affect these changes in height. Previous studies on osteocalcin[45] and U-NTx[41] have shown that these markers do not only depend on age and sex, but also on pubertal development stage. Unfortunately, pubertal status was not assessed in the present study.

In a 1-year lifestyle intervention based on exercise, behaviour and nutrition therapy in sixty obese children, Reinehr & Roth[46] found a significant negative correlation between changes in total osteocalcin and changes in the homeostasis model of assessment for insulin resistance index. Since the initiation of the DiOGenes study, several studies in children have linked bone metabolism with energy metabolism[46–52]. Osteocalcin is among the bone turnover markers that has attracted most attention. Mechanistic studies in rodents have pointed to an endocrine bone–pancreas loop, through which insulin signalling in the osteoblasts stimulates osteocalcin production, which in turn increases pancreatic insulin secretion and insulin sensitivity to control glucose homeostasis. Thus, on the one hand, osteocalcin-deficient mice have shown decreased insulin secretion and decreased insulin sensitivity – effects that can be reversed by infusions with osteocalcin[27], while, on the other hand, mice lacking the insulin receptor in the osteoblasts have shown reduced postnatal bone acquisition[53]. Based on these studies, it would appear that not only the protein component of the DiOGenes dietary intervention may have an influence on bone metabolism, but also the GI component – through the interplay between osteocalcin and insulin. This was also what we observed in the present analyses. The effect of protein on osteocalcin was only evident within the HGI groups. It is possible that the effect of protein on osteocalcin depends on a concurrent high level of insulin. In the present study, only total osteocalcin was measured – not undercarboxylated and carboxylated osteocalcin. On the one hand, it is possible that a decrease in total osteocalcin, primarily caused by a decrease in carboxylated osteocalcin, may pose a threat to bone health. On the other hand, a decrease in total osteocalcin, primarily caused by a decrease in undercarboxylated osteocalcin, may possibly not be harmful to bone health[24], but could have unfavourable effects on insulin sensitivity[54]. Future research should take into account the possible interaction with insulin, when examining the relationship between protein intake, GI and bone turnover in children.

The median intakes of about 18–20 E% protein in the HP groups were lower than that aimed for these groups (23–28 E%), while the average intakes of 14–16 E% protein in the LP groups were slightly higher or within the intended range for these groups (10–15 E%). Similarly, only approximately one-half (approximately 6–8 GI units) of the aimed difference of 15 GI units between the LGI and HGI groups was achieved, and the difference was not even significant between the LP/HGI and LP/LGI groups at month 6. Thus, the effects of more extreme intakes of protein and GI on bone turnover in children are still unknown. As usually observed in relation to dietary recording, under-reporting was very common (median energy intake:BMR 0·90–1·45). We chose measures for dietary intake that we expected to be less dependent on age and sex of the child and not to be so sensitive to general under-reporting, e.g. E% of protein,

British Journal of Nutrition

fat and carbohydrate instead of using grams. However, it is possible that the study participants were more likely to under-report certain food items than others.

In a mixed diet as used in the present study, other components than protein and the GI such as Ca, vitamin D, vitamin K, P and Na, as well as the sources of protein (dairy products/animal sources other than dairy products/vegetables) may determine whether protein and the GI influence bone turnover or not. Also, it is possible that besides the differences in dietary groups, differences in these other dietary components were actually the reason for an effect on bone markers – not GI or total protein *per se*. However, we did not find that 3 d dietary records were sufficient to determine protein sources and intakes of specific micronutrients, and for this reason, these dietary components were not included in the analyses.

In conclusion, the present study does not show any effect of increased protein intake on height or bone resorption in children. However, the difference in changes in the level of osteocalcin between the HP/HGI group and the LP/HGI group warrants further investigation and should be confirmed in other studies.

Acknowledgements

The DiOGenes project was supported by a contract (FP6-2005-513946) from the European Commission Food Quality and Safety Priority of the Sixth Framework Program. Local sponsors made financial contributions to the shop centres, which also received a number of foods free of charge from food manufacturers. A full list of these sponsors is available online (www.diogenes-eu.org/sponsors/). The European Commission and the local sponsors had no role in the design, analysis or writing of this article.

The authors' contributions are as follows: W. H. M. S. and A. A. designed the study; W. H. M. S., A. A. and A. P. conducted the study; M. M., C. R. and S.-M. D. conducted the statistical analyses; S.-M. D., C. T. D., K. F. M. and C. M. wrote the paper; S.-M. D. had primary responsibility for the final content. All authors read and approved the final manuscript.

The Department of Nutrition, Exercise and Sports at the University of Copenhagen has received research support from more than 100 food companies for this and other studies. W. H. M. S. is part-time employed by DSM, Inc., The Netherlands. A. A. is currently a member of the following scientific advisory boards: Global Dairy Platform, USA; Jenny Craig, USA; Pathway Genomics, USA; McDonald's, USA. K. F. M. received grants from Arla Foods Ingredients, Denmark and the US Dairy Export Council for studies focusing on undernutrition in low-income countries. S.-M. D., M. M., C. R., C. T. D., A. P. and C. M. declare no conflicts of interest.

References

1. Cao JJ & Nielsen FH (2010) Acid diet (high-meat protein) effects on calcium metabolism and bone health. *Curr Opin Clin Nutr Metab Care* **13**, 698–702.
2. Darling AL, Millward DJ, Torgerson DJ, *et al.* (2009) Dietary protein and bone health: a systematic review and meta-analysis. *Am J Clin Nutr* **90**, 1674–1692.
3. Heaney RP & Layman DK (2008) Amount and type of protein influences bone health. *Am J Clin Nutr* **87**, 1567S–1570S.
4. Jesudason D & Clifton P (2011) The interaction between dietary protein and bone health. *J Bone Miner Metab* **29**, 1–14.
5. Hegsted M & Linkswiler HM (1981) Long-term effects of level of protein intake on calcium metabolism in young adult women. *J Nutr* **111**, 244–251.
6. Hegsted M, Schuette SA, Zemel MB, *et al.* (1981) Urinary calcium and calcium balance in young men as affected by level of protein and phosphorus intake. *J Nutr* **111**, 553–562.
7. Margen S, Chu JY, Kaufmann NA, *et al.* (1974) Studies in calcium metabolism. I. The calciuretic effect of dietary protein. *Am J Clin Nutr* **27**, 584–589.
8. Pannemans DLE, Schaafsma G & Westerterp KR (1997) Calcium excretion, apparent calcium absorption and calcium balance in young and elderly subjects: influence of protein intake. *Br J Nutr* **77**, 721–729.
9. Schofield FA & Morrell E (1960) Calcium, phosphorus and magnesium. *Fed Proc* **19**, 1014–1016.
10. Zemel MB, Schuette SA, Hegsted M, *et al.* (1981) Role of the sulfur-containing amino acids in protein-induced hypercalciuria in men. *J Nutr* **111**, 545–552.
11. Cao JJ, Johnson LK & Hunt JR (2011) A diet high in meat protein and potential renal acid load increases fractional calcium absorption and urinary calcium excretion without affecting markers of bone resorption or formation in postmenopausal women. *J Nutr* **141**, 391–397.
12. Chu JY, Margen S & Costa FM (1975) Studies in calcium metabolism. II. Effects of low calcium and variable protein intake on human calcium metabolism. *Am J Clin Nutr* **28**, 1028–1035.
13. Kerstetter JE, O'Brien KO & Insogna KL (1998) Dietary protein affects intestinal calcium absorption. *Am J Clin Nutr* **68**, 859–865.
14. Kerstetter JE, O'Brien KO, Caseria DM, Wall DE, *et al.* (2005) The impact of dietary protein on calcium absorption and kinetic measures of bone turnover in women. *J Clin Endocrinol Metab* **90**, 26–31.
15. Allen LH, Oddoye EA & Margen S (1979) Protein-induced hypercalciuria: a longer term study. *Am J Clin Nutr* **32**, 741–749.
16. Johnson NE, Alcantara EN & Linkswiler H (1970) Effect of level of protein intake on urinary and fecal calcium and calcium retention of young adult males. *J Nutr* **100**, 1425–1430.
17. Skov AR, Haulrik N, Toubro S, *et al.* (2002) Effect of protein intake on bone mineralization during weight loss: a 6-month trial. *Obes Res* **10**, 432–438.
18. Alexy U, Remer T, Manz F, *et al.* (2005) Long-term protein intake and dietary potential renal acid load are associated with bone modeling and remodeling at the proximal radius in healthy children. *Am J Clin Nutr* **82**, 1107–1114.
19. Chevalley T, Bonjour JP, Ferrari S, *et al.* (2008) High-protein intake enhances the positive impact of physical activity on BMC in prepubertal boys. *J Bone Miner Res* **23**, 131–142.
20. Hoppe C, Molgaard C & Michaelsen KF (2000) Bone size and bone mass in 10-year-old Danish children: effect of current diet. *Osteoporos Int* **11**, 1024–1030.
21. Libuda L, Wudy SA, Schoenau E, *et al.* (2011) Comparison of the effects of dietary protein, androstenediol and forearm muscle area on radial bone variables in healthy prepubertal children. *Br J Nutr* **105**, 428–435.

British Journal of Nutrition

22. Vatanparast H, Bailey DA, Baxter-Jones AD, *et al.* (2007) The effects of dietary protein on bone mineral mass in young adults may be modulated by adolescent calcium intake. *J Nutr* **137**, 2674–2679.
23. Budek AZ, Hoppe C, Michaelsen KF, *et al.* (2007) High intake of milk, but not meat, decreases bone turnover in prepubertal boys after 7 days. *Eur J Clin Nutr* **61**, 957–962.
24. Bugel S (2008) Vitamin K and bone health in adult humans. *Vitam Horm* **78**, 393–416.
25. Gundberg CM (2003) Matrix proteins. *Osteoporos Int* **14**, Suppl. 5, S37–S40.
26. Hu FB (2005) Protein, body weight, and cardiovascular health. *Am J Clin Nutr* **82**, 242S–247S.
27. Ng KW (2011) Regulation of glucose metabolism and the skeleton. *Clin Endocrinol (Oxf)* **75**, 147–155.
28. Larsen TM, Dalskov S, van Baak M, *et al.* (2010) The Diet, Obesity and Genes (Diogenes) Dietary Study in eight European countries – a comprehensive design for long-term intervention. *Obes Rev* **11**, 76–91.
29. Moore CS, Lindroos AK, Kreutzer M, *et al.* (2010) Dietary strategy to manipulate *ad libitum* macronutrient intake, and glycaemic index, across eight European countries in the Diogenes Study. *Obes Rev* **11**, 67–75.
30. de OM, Onyango AW, Borghi E, *et al.* (2007) Development of a WHO growth reference for school-aged children and adolescents. *Bull World Health Organ* **85**, 660–667.
31. World Health Organization (2012) WHO Anthroplus macros for STATA2012. http://www.who.int/growthref/tools/en/ (accessed October 2012).
32. Aston LM, Jackson D, Monsheimer S, *et al.* (2010) Developing a methodology for assigning glycaemic index values to foods consumed across Europe. *Obes Rev* **11**, 92–100.
33. Henry CJ (2005) Basal metabolic rate studies in humans: measurement and development of new equations. *Public Health Nutr* **8**, 1133–1152.
34. Hothorn T, Bretz F & Westfall P (2008) Simultaneous inference in general parametric models. *Biom J* **50**, 346–363.
35. R Core Team (2012) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/
36. StataCorp (2011) *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.
37. Roughead ZK, Johnson LK, Lykken GI, *et al.* (2003) Controlled high meat diets do not affect calcium retention or indices of bone status in healthy postmenopausal women. *J Nutr* **133**, 1020–1026.
38. Budek AZ, Hoppe C, Ingstrup H, *et al.* (2007) Dietary protein intake and bone mineral content in adolescents – The Copenhagen Cohort Study. *Osteoporos Int* **18**, 1661–1667.
39. Remer T, Manz F, Alexy U, *et al.* (2011) Long-term high urinary potential renal acid load and low nitrogen excretion predict reduced diaphyseal bone mass and bone size in children. *J Clin Endocrinol Metab* **96**, 2861–2868.
40. Christenson RH (1997) Biochemical markers of bone metabolism: an overview. *Clin Biochem* **30**, 573–593.
41. Mora S, Prinster C, Proverbio MC, *et al.* (1998) Urinary markers of bone turnover in healthy children and adolescents: age-related changes and effect of puberty. *Calcif Tissue Int* **63**, 369–374.
42. van der Sluis IM, de Ridder MA, Boot AM, *et al.* (2002) Reference data for bone density and body composition measured with dual energy X ray absorptiometry in white children and young adults. *Arch Dis Child* **87**, 341–347.
43. Szulc P, Seeman E & Delmas PD (2000) Biochemical measurements of bone turnover in children and adolescents. *Osteoporos Int* **11**, 281–294.
44. Cadogan J, Eastell R, Jones N, *et al.* (1997) Milk intake and bone mineral acquisition in adolescent girls: randomised, controlled intervention trial. *BMJ* **315**, 1255–1260.
45. van der Sluis IM, Hop WC, van Leeuwen JP, *et al.* (2002) A cross-sectional study on biochemical parameters of bone turnover and vitamin D metabolites in healthy Dutch children and young adults. *Horm Res* **57**, 170–179.
46. Reinehr T & Roth CL (2010) A new link between skeleton, obesity and insulin resistance: relationships between osteocalcin, leptin and insulin resistance in obese children before and after weight loss. *Int J Obes (Lond)* **34**, 852–858.
47. Afghani A & Goran MI (2009) The interrelationships between abdominal adiposity, leptin and bone mineral content in overweight Latino children. *Horm Res* **72**, 82–87.
48. Lawlor DA, Sattar N, Sayers A, *et al.* (2012) The association of fasting insulin, glucose, and lipids with bone mass in adolescents: findings from a cross-sectional study. *J Clin Endocrinol Metab* **97**, 2068–2076.
49. Pollock NK, Bernard PJ, Gower BA, *et al.* (2011) Lower uncarboxylated osteocalcin concentrations in children with prediabetes is associated with beta-cell function. *J Clin Endocrinol Metab* **96**, E1092–E1099.
50. Rochefort GY, Rocher E, Aveline PC, *et al.* (2011) Osteocalcin–insulin relationship in obese children: a role for the skeleton in energy metabolism. *Clin Endocrinol (Oxf)* **75**, 265–270.
51. Sayers A, Timpson NJ, Sattar N, *et al.* (2010) Adiponectin and its association with bone mass accrual in childhood. *J Bone Miner Res* **25**, 2212–2220.
52. Sayers A, Lawlor DA, Sattar N, *et al.* (2012) The association between insulin levels and cortical bone: findings from a cross-sectional analysis of pQCT parameters in adolescents. *J Bone Miner Res* **27**, 610–618.
53. Fulzele K, Riddle RC, DiGirolamo DJ, *et al.* (2010) Insulin receptor signaling in osteoblasts regulates postnatal bone acquisition and body composition. *Cell* **142**, 309–319.
54. Ducy P (2011) The role of osteocalcin in the endocrine cross-talk between bone remodelling and energy metabolism. *Diabetologia* **54**, 1291–1297.

British Journal of Nutrition