Statistical Inference for Molecular Shapes

Irina Czogiel

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

December 2009

All things are made of atoms, and [...] *everything that living things can do can be understood in terms of the jiggling and wiggling of atoms.*

The Feynman Lectures in Physics

Abstract

This thesis is concerned with developing statistical methods for evaluating and comparing molecular shapes. Techniques from statistical shape analysis serve as a basis for our methods. However, as molecules are fuzzy objects of electron clouds which constantly undergo vibrational motions and conformational changes, these techniques should be modified to be more suitable for the distinctive features of molecular shape.

The first part of this thesis is concerned with the continuous nature of molecules. Based on molecular properties which have been measured at the atom positions, a continuous field–based representation of a molecule is obtained using methods from spatial statistics. Within the framework of reproducing kernel Hilbert spaces, a similarity index for two molecular shapes is proposed which can then be used for the pairwise alignment of molecules. The alignment is carried out using Markov chain Monte Carlo methods and posterior inference. In the Bayesian setting, it is also possible to introduce additional parameters (mask vectors) which allow for the fact that only part of the molecules may be similar. We apply our methods to a dataset of 31 steroid molecules which fall into three activity classes with respect to the binding activity to a common receptor protein. To investigate which molecular features distinguish the activity classes, we also propose a generalisation of the pairwise method to the simultaneous alignment of several molecules.

The second part of this thesis is concerned with the dynamic aspect of molecular shapes. Here, we consider a dataset containing time series of DNA configurations which have been obtained using molecular dynamic simulations. For each considered DNA duplex, both a damaged and an undamaged version are available, and the objective is to investigate whether or not the damage induces a significant difference to the the mean shape of the molecule. To do so, we consider bootstrap hypothesis tests for the equality of mean shapes. In particular, we investigate the use of a computationally inexpensive algorithm which is based on the Procrustes tangent space. Two versions of this algorithm are proposed. The first version is designed for independent configuration matrices while the second version is specifically designed to accommodate temporal dependence of the configurations within each group and is hence more suitable for the DNA data.

Acknowledgements

I'd like to thank my supervisors Ian Dryden and Chris Brignell for their support; with special thanks to Chris who was a great help in the pre–submission phase and might well have set a new world record in proof reading speed. I'm also enormously grateful to Dave Parkin who always fixed my computer and reassured me that computers are non–deterministic beings who just happen to dislike some people more than others. Huge thanks goes to Georgie and Bill plus cat and dog who made me feel at home in their house for a month during my visit in South Carolina. Thank you Tina for accidently studying stats and having been an amazing friend since pretty much our first day at uni – see you in Berlin!! The numerous tea breaks over the years were a big part of my time in Nottingham and I'd like to thank all my maths friends for making them my number one choice of procrastination and for letting me get the easy clues in the crossword. Last but definitely not least I'd like to thank my parents who always support me no matter what crazy ideas I come up with – Danke!!

Contents

List of Figures

List of Tables

Important Operators

The **vectorise operator** vec(.) of an $r \times c$ matrix $\boldsymbol{X}$ with column vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_c$ stacks the columns of $\boldsymbol{X}$ to give an rc–vector, i.e

$$\text{vec}(\boldsymbol{X}) = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_c^T)^T. \tag{0.1}$$

Let $\boldsymbol{x}$ denote an rc–vector. The **inverse vectorise operator** $\text{vec}_c^{-1}(.)$ then row–wise forms a matrix with c columns, i.e.

$$\text{vec}(\boldsymbol{X}) = \boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_c^T)^T \Leftrightarrow \text{vec}_c^{-1}(\boldsymbol{x}) = \boldsymbol{X} \tag{0.2}$$

Let $\boldsymbol{X}$ denote a symmetric $r \times r$ matrix with entries $x_{ij} = x_{ji}$ $(i, j = 1, \ldots r)$. The **vectorise–half operator** vech(.) then vectorises the $r(r+1)/2$ distinct elements of $\boldsymbol{X}$ which can be found in its upper triangle, i.e.

$$\text{vech}(\boldsymbol{X}) = (x_{11}, x_{12}, x_{22}, x_{13}, \ldots, x_{r-1\,r}, x_{rr})^T. \tag{0.3}$$

For more information about the above operators see Henderson & Searle (1979) who provide an account on their history, properties and also give many applications.

Chapter 1

Introduction

Chemical processes in general are largely governed by the structure (shape) of the involved molecules. Molecular shapes are therefore of great importance in many scientific areas such as rational drug design or molecular recognition. In particular, because molecules which are similar in shape can be expected to exhibit a similar biochemical behaviour, it often is of interest to determine the similarity between molecular structures. However, the notion of molecular shape is complex and there is no generally agreed algebraic expression for the similarity between molecular structures which is apparent in the vast number of similarity indices which have been proposed in the literature. Most of these indices are thereby obtained in a two–step procedure in which the molecules under consideration are first aligned as closely as possible with respect to a suitable objective function, and the actual similarity is then calculated based on the thus obtained optimal relative position.

Although the different approaches to structural alignment draw from a remarkably diverse range of mathematical concepts, statistical considerations have not been widely applied yet. This is somewhat surprising since data on a molecular level are often confounded with a considerable amount of uncertainty and, in general, accounting for measurement errors and similar flaws of the data often provides a deeper insight into underlying principles. The aim of the research presented in this thesis is to develop statistical methods for evaluating and comparing molecular shapes. These methods will draw from different areas of statistics. In particular statistical shape analysis, spatial

statistics, time series analysis and bootstrap methods will play an important role. However, the established statistical methods need to be combined and modified so that they can cope with the distinctive features of molecular shapes.

One of the challenges arises from the the fact that molecules are fuzzy objects which are diffused in space. For example the point–based methods from classical statistical shape analysis are therefore not ideal for capturing the true nature of molecular shapes. Another peculiarity is that a global expression for shape similarity may not be appropriate in the context of drug design since the entire molecular structure of a ligand is not usually involved in the interaction with the target molecule. The use of local shape similarities could therefore provide a better means for finding molecules with a desired biochemical activity. Moreover, a considerable challenge is that molecules constantly undergo vibrational motions and conformational changes so that it would be beneficial to take into account the dynamic aspect of the nuclear arrangement.

This thesis is divided into two main parts which consider different aspects of molecular shapes. In the first part (Chapters 3 and 4), we develop a framework for evaluating and comparing molecular shapes which is specifically designed to take into account the fuzzy nature of molecules. The molecules are assumed to be rigid in this part, and the focus lies on finding a suitable alignment method for continuous molecular shapes. The second part (Chapters 5 and 6) is complementary to this work. Here, the alignment is carried out using methods from classical statistical shape analysis, and the focus lies on incorporating the molecular dynamics information in the subsequent comparison.

1.1 Modelling and Comparing Continuous Molecular Shapes

The first part of this thesis has been motivated by the structural alignment problem in chemoinformatics where the main aim is to predict the drug potency of a molecule by comparing its shape to that of a known drug molecule. Some of the work presented in this part can be found in Czogiel *et al.* (2008) and Czogiel *et al.* (2009).

1.1.1 Background Information

A major goal in pharmaceutical research is the design of selective ligands for protein and DNA binding – a hard task because the space of ligands with a potential beneficial effect on the human body is vast. In fact, the number of small organic compounds which could potentially be orally administered as drugs has been estimated to exceed 10^{60} (e.g. Dobson, 2004).

Since in most practical cases the three–dimensional structure of the receptor is unknown, direct rational drug design techniques such as docking (e.g. Blanley & Dixon, 1993) are not generally applicable. A way to tackle this problem is to make use of the fact that any chemical binding process requires some complementarity between the ligand and its receptor. Ligands which bind to the same target can therefore be expected to possess a certain degree of shape similarity. When designing new drug molecules, the converse of this concept is exploited. Here, the underlying conjecture is that molecules of a similar shape exhibit a similar biochemical activity and hence drug potency.

One way of obtaining a numerical value for the shape similarity of two molecules is to align their structures to match each other as closely as possible with respect to an appropriate objective function. As this function is usually designed to measure the degree of similarity of the ligands dependent on their relative position, its value at the optimal relative position then provides a similarity measure for the ligands themselves which can then be exploited in several ways.

For example if the superimposed ligands are known to bind to the same target, then the optimal alignment approximates the binding geometry of the ligands. It is therefore possible to deduce the structural binding requirements by extracting the properties common to all or most of the aligned ligands at certain locations in space (e.g. Kim, 1995). Moreover, we can learn about the unknown receptor site from the negative imprint of the set of superimposed ligands (e.g. Crippen, 1987), and recently, Keiser *et al.* (2007) used the average pairwise similarity between two sets of ligands which bind to two distinct proteins to obtain a notion of the similarity of the proteins themselves.

Perhaps the most widespread application of the structural alignment method, however, is to use the resulting similarity measure as a scoring function in the screening of ligand databases. In this context, the alignment serves as a pre–filter for potential drug molecules, and ligands which are found to have a high degree of similarity to a known "lead" compound can then be further tested for beneficial bioactivity. Overviews of structural alignment techniques and their applications can be found in Good (1995), Lemmen & Lengauer (2000), and Bender & Glen (2004), and a summary of the most common concepts is also provided in Section 4.1.

1.1.2 Structural Alignment and Statistical Shape Analysis

Structural alignment of molecules filters out the information about their (usually arbitrarily recorded) relative position so that subsequent analyses can focus on their rotation/translation invariant properties. Similar problems are well–known in the field of statistical shape analysis which will be described in Section 2.1. In essence, classical statistical shape analysis is designed for the situation where each object is represented by a set of points (landmarks), and a shape–based measure of their similarity is obtained by rotating, translating and scaling the objects relative to each other so that the sum of the squared distances between corresponding landmarks is minimised.

Although the positions of the atoms in a molecule can serve as landmarks, it is not possible to directly apply methods from classical statistical shape analysis to the structural alignment problem because a one–to–one correspondence between atoms of different molecules is usually not known. A way to tackle this problem is to introduce a labelling matrix with binary entries which determines whether or not two atoms correspond to each other. This approach is pursued by Green & Mardia (2006), Dryden *et al.* (2007) and Schmidler (2007) who set up a Bayesian framework in which the labelling matrix is considered to be a random parameter which can be inferred about using posterior analysis. The main difference between the three papers is the way the nuisance parameters of rotation and translation are dealt with. More information about these approaches will be provided in Section 3.2.

The methods we propose in the first part of this thesis build on these previous applications of statistical shape analysis to the structural alignment of molecules. However, we will move away from a point–based representation of molecular shapes and generalise the concepts to a more realistic continuous representation.

1.1.3 The Steroid Dataset

The dataset we use to evaluate our methods was compiled by Cramer *et al.* (1988) and has been used before as a test bed for structural alignment techniques (e.g. Anzali *et al.*, 1998; Coats, 1998; Dryden *et al.*, 2007). It comprises of 31 steroid molecules which bind to the same corticosteroid binding globulin (CBG) receptor. For each molecule, the xyz–coordinates of the atom positions in Å (Ångström: 1Å=10^{-10} m) as well as the atom types (e.g. carbon, oxygen, ...), the associated van der Waals radii and the partial atomic charge values at the atom positions are provided.

Roughly speaking, the van der Waals radii define the range of the territory around each atom in which no other atom can intrude. They provide information about the steric (shape) properties of the molecules whereas the partial charge values within a molecule arise from asymmetries of the distribution of electrons in chemical bonds and are associated with the electrostatic properties of the molecules.

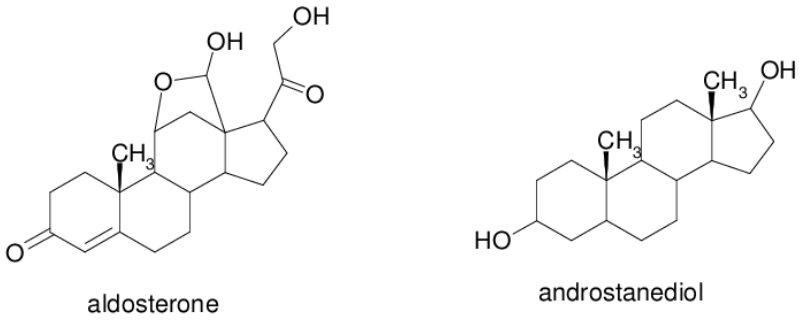

Figure 1.1: Two–dimensional representations of two steroid molecules from the dataset: The molecules are structurally similar in that their core structure consists of four carbon rings.

A major feature of the steroid dataset is that all molecules share a common core structure consisting of four carbon rings. Figure 1.1 displays the two steroid molecules aldosterone and androstanediol. In this two–dimensional representation, the common ring structure is clearly visible. Good *et al.* (1993) classified each steroid according to its binding activity towards the CBG receptor. This provides the opportunity to assess whether or not the obtained similarities are chemically meaningful in that they reflect the membership of the steroids to the different activity classes. The steroid dataset is publicly available from `http://www2.ccc.uni-erlangen.de/services/steroids/`.

1.2 Comparing Dynamic Molecular Shapes

The second part of this thesis is motivated by the question of whether or not damage significantly changes the shape of a DNA (DeoxyriboNucleid Acid) molecule. This question is important in the field of molecular recognition because significant shape differences between damaged and undamaged DNA strands could have an impact on the binding affinities of the DNA towards the corresponding repair protein.

1.2.1 Background Information

The DNA is a macromolecule which is found in the cells of living organisms. It is of vital importance as it contains the instructions needed for the organism to develop, survive and reproduce.

In the past decades, the DNA has received much attention from various research communities. In particular the characteristic double–helical structure of the DNA is of great interest as it transmits the genetic information and translates it into simple instructions for the cellular machinery. In fact, it was the discovery of this structure by Watson & Crick (1953) which triggered genetics, biochemistry and molecular biology, as understood at the beginning of the $21st$ century.

Chemically, a DNA molecule (or duplex) consists of two long polymers of repeating structural building blocks with backbones made of sugars and phosphate groups. These polymers are the two strands of the DNA. Attached to each sugar is one of four types of bases, namely cytosine (C), guanine (G), adenine (A) and thymine (T), and it is the sequence of these bases which codes the genetic information. The bases are also responsible for the characteristic shape of the DNA because they interact with each other in a way which stabilises the double–helical arrangement.

The double helix of a DNA molecule is not rigid as it – like every molecule – constantly undergoes vibrational motions and conformational changes. With the realisation that these internal motions play a functional role in that they contribute to the binding properties of the molecule, molecular dynamics (MD) modelling has become one of the most powerful tools for gaining atomic–level insight into nucleic acids.

The first MD simulation of a macromolecule of biological interest was published by McCammon *et al.* (1977). Since then, well–defined standards for simulation conditions and protocols have been established (cf. e.g. Olson & Zhurkin, 2000; Giudice & Levery, 2002; Orozco *et al.*, 2003), and today computer packages such as AMBER (Case *et al.*, 2005) are available which carry out all–atom simulations of several turns of double helix with surrounding solvent molecules. Roughly speaking, these simulations are based on deterministic models in which the atoms of the molecule are viewed as point masses which are attached to springs (bonds). Using the current atom positions, the equations of motion are solved to provide the positions at the next time point.

One application of MD simulations which has been of recent interest in the field of molecular recognition is to investigate the question of whether or not damage to DNA molecules has a significant impact on the shape of the duplexes which could explain why repair proteins have a larger binding affinity towards damaged than undamaged DNA strands (Jiranusronkul & Laughton, 2008).

Damage to DNA is in general caused by physical or chemical agents such as electromagnetic radiation or substances like nitrogen oxide (found in cigarette smoke) which change

the DNA molecules and can thus increase the frequency of DNA mutations above the natural background level. In particular, oxidative damage and the associated mutations are thought of as the major contributor to human cancer (Beckman & Ames, 1997).

Of the four nucleic acids guanine is the most prone to oxidation, and one of the most prevalent guanine–derived lesions is called FapydG. According to Wilson & Bohr (2007), FapydG is the most ubiquitous lesion associated with high mutagenicity in DNA. Its structure is very different from that of the original guanine base. In particular, it exhibits a considerably higher flexibility. This leads to the question how FapydG changes the overall structure of the DNA which could be connected to its mutagenic potential.

1.2.2 MD Simulations and Statistical Shape Analysis

The datasets which result from MD simulations are multivariate time series in the space of possible molecular configurations. However, they contain redundant information because the particular location of the molecule at each time step is irrelevant. When analysing MD time series, it is therefore advisable to employ methods which are invariant under the rotation and translation of the given molecular configurations. Like in the context of the structural alignment of ligands, a basis for analysing MD datasets from a statistical point of view is therefore given by the methods from the field of statistical shape analysis.

Previous applications of statistical shape analysis to MD simulations of DNA strands include Dryden *et al.* (2002, 2009) who consider estimating the configurational entropy of a duplex using a separable Gaussian model in size–and–shape and time, and Dryden & Zempléni (2006) who investigate the extreme size–and–shape behaviour of DNA sequences. The latter can be used to assess whether or not MD simulations have run long enough to sufficiently explore the configurational space. Another application of statistical shape analysis to MD data can be found in Preston & Wood (2009a) who construct non–parametric confidence regions for the mean atomic coordinates of a DNA sequence.

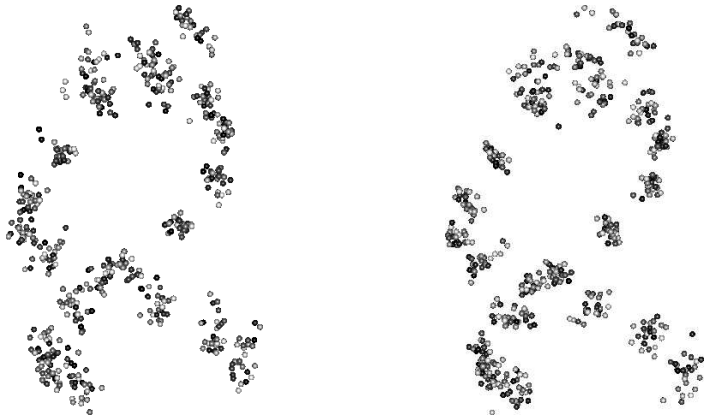

Figure 1.2: Thinned sequences for a normal/damaged pair of DNA duplexes: The displayed pair is AGG (left–hand side) and AFG (right–hand side). Every $100th$ configuration of each duplex is shown. The grey–level on each side corresponds to the MD iteration where lighter shades of grey show configurations obtained at later iterations.

Motivated by the problem of comparing MD trajectories of damaged and undamaged DNA strands, we propose a test procedure for the equality of mean shapes in the second part of this thesis. This test procedure is based on the model by Dryden *et al.* (2002, 2009) and can be applied for temporally evolving shape data in general.

1.2.3 The DNA Dataset

The dataset at hand is based on the data generated by Jiranusronkul & Laughton (2008) who apply MD simulations to identify the molecular perturbations to the normal DNA structure brought about by replacing guanine (G) by FapydG (F). Time series of the atomic xyz–coordinates of 22 phosphorus backbone atoms are provided for twelve different DNA strands (2,500 observation over time). Following Jiranusronkul & Laughton, these time series will be denoted as

AGA	AGC	AGG	TGA	TGC	TGT
AFA	AFC	AFG	TFA	TFC	TFT.

The twelve time series thereby come as six pairs, and each pair contains both an undamaged (top row) and a damaged version (bottom row) of the same duplex.

Figure 1.2 shows excerpts of the data for the normal/damaged DNA pair AGG/AFG. Every $100th$ configuration for each duplex is displayed. To ensure an unbiased representation, the combined data have been optimally aligned and scaled relative to each other before plotting, and Figure 1.2 shows the configurations in their optimal position and size. The displayed grey–levels thereby show the corresponding MD iteration where lighter shades show configurations obtained at later iterations.

From Figure 1.2, it is not possible to detect any clear distinction between the two duplexes. In Chapters 5 and 6 we investigate whether or not there are more subtle differences which can be detected numerically using a hypothesis test. To do so, we investigate the use of non–parametric bootstrap tests for the equality of mean shapes.

1.3 Thesis Outline

In Chapter 2, we provide an introduction to the methods which are applied in this thesis. As well as giving us tools for analysing the two datasets at hand, these methods provide starting points for the novel techniques developed. In particular, statistical shape analysis, spatial statistics and bootstrap methods will play an important role in this thesis.

From a statistical point of view, the steroid dataset is a set of unlabelled marked point sets, where "unlabelled" refers to the lack of one–to–one correspondences between the atoms (landmarks) and "marked" refers to the fact that additional information (e.g. partial charge values) is provided at each landmark. In Chapter 3, we propose a novel approach for aligning data of this kind which provides the possibility to counterbalance the lack of homologous landmarks with the spatial distribution of the given marks. Using spatial statistics, this idea leads to a continuous representation of the "shape" of a marked point set. An alignment of two objects can then be carried out using concepts from statistical shape analysis, reproducing kernel Hilbert spaces and Markov chain Monte Carlo. Our alignment algorithm is validated using a simulation study based on which we also formulate guidelines for a successful superposition.

In Chapter 4, the new alignment methodology is applied to the steroid data. In this application, it is also of interest to perform a multiple alignment of several molecules. We therefore propose an extension of the alignment algorithm to several objects which can be viewed as a continuous version of the generalised Procrustes analysis algorithm well–known in statistical shape analysis. Applying this extension to the steroid data then provides the possibility to post–process the alignment results – for example using exploratory t–tests.

Chapters 5 and 6 are concerned with the construction of a non–parametric hypothesis test for the equality of two population mean shapes. In Chapter 5, attention is restricted to the situation where the data at hand are sets of independent configuration matrices. Based on tangent projections of the observed data, a fast bootstrap algorithm is proposed whose performance is validated in a simulation study. This algorithm is then applied to the DNA dataset. However, as described above, the DNA data have been generated using MD simulations so that the observed molecular configurations exhibit some temporal dependence. In Chapter 6 we therefore propose an amendment of the bootstrap procedure to time series data. Based on simulated data the superiority of this amended version can be demonstrated so that test results based on this new bootstrap test should be more reliable when applied to the DNA data.

Finally, Chapter 7 concludes this thesis with a summary of the main results and discusses areas for further work.

All algorithms described are implemented using the statistical software package R (R Development Core Team, 2008).

Chapter 2

Applied Methods

In this chapter, we provide some background information about the statistical concepts applied in this thesis. The main topics thereby include statistical shape analysis, spatial statistics, and bootstrap methods.

2.1 Statistical Shape Analysis

Intuitively, the shape of an object can be characterised as all geometrical information which remains when translation, scaling and rotation are removed (e.g. Kendall, 1977). This invariance under the Euclidean similarity transformations implies that the space of all possible shapes is non–Euclidean in nature which makes defining a mathematical framework for the analysis of shapes not straightforward. Classical statistical techniques are often not appropriate and new methods have to be developed. In most cases, these methods are designed for the situation where an m–dimensional object is represented by a configuration matrix consisting of the position of k landmarks. Given such a matrix, the shape of the object can then be derived by removing the similarity transformations in turn. Since the pioneering papers by Kendall (1984) and Bookstein (1986), there have been several accounts on the field of statistical shape analysis including the books by Small (1996), Dryden & Mardia (1998) and Kendall *et al.* (1999). In this chapter, the treatment is largely based on the book by Dryden & Mardia (1998).

2.1.1 Shape – Removing the Similarity Transformations

Let $\boldsymbol{X}$ denote a $k \times m$ configuration matrix. To remove location, $\boldsymbol{X}$ can be pre–multiplied with a suitable matrix, e.g. the Helmert sub–matrix $\boldsymbol{H}$ or the centering matrix $\boldsymbol{C}$. The Helmert sub–matrix is the $(k-1) \times k$ matrix whose ith row has the form

$$(h_i, \ldots, h_i, -ih_i, 0 \ldots, 0),$$

where $h_i = -\{i(i+1)\}^{-1/2}$ is repeated i times $(i = 1, \ldots k-1)$, and the $(k \times k)$ centering matrix $\boldsymbol{C}$ can be written as $\boldsymbol{C} = \boldsymbol{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T$, where $\boldsymbol{I}_k$ denotes the identity matrix in k dimensions and $\mathbf{1}_k$ denotes the k–vector of ones. Note that the two matrices are related by $\boldsymbol{C} = \boldsymbol{H}^T\boldsymbol{H}$. Pre–multiplication with $\boldsymbol{H}$ or $\boldsymbol{C}$ yields

$$\boldsymbol{X}_H = \boldsymbol{H}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{X}_C = \boldsymbol{C}\boldsymbol{X} = \boldsymbol{H}^T\boldsymbol{X}_H,$$

i.e. the Helmertised and centred configuration matrix, respectively, which are invariant under the location of the original configuration matrix.

Having filtered out the translation information from the original landmarks, the scaling can be removed by normalising with respect to the Frobenius norm which is defined as

$$||\boldsymbol{X}_H|| = \sqrt{\text{tr}(\boldsymbol{X}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{X})} = \sqrt{\text{tr}(\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{X})} = ||\boldsymbol{X}_C|| =: S(\boldsymbol{X}).$$

The above expression is also called the *centroid size* of $\boldsymbol{X}$. It satisfies $S(a\boldsymbol{X}) = aS(\boldsymbol{X})$, where a is a positive scalar and therefore is a suitable measure of the size of $\boldsymbol{X}$. Geometrically, $S(\boldsymbol{X})$ is the square root of the sum of squared Euclidean distances from each landmark to the centroid. Using $S(\boldsymbol{X})$, the scale information can be removed from $\boldsymbol{X}_H$ and $\boldsymbol{X}_C$ yielding

$$\boldsymbol{Z} = \frac{\boldsymbol{X}_H}{||\boldsymbol{X}_H||} = \frac{\boldsymbol{H}\boldsymbol{X}}{||\boldsymbol{H}\boldsymbol{X}||} \quad \text{and} \quad \boldsymbol{Z}_C = \frac{\boldsymbol{X}_C}{||\boldsymbol{X}_C||} = \frac{\boldsymbol{C}\boldsymbol{X}}{||\boldsymbol{C}\boldsymbol{X}||} \tag{2.1}$$

which are invariant under the translation and scaling of the original configuration. The matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}_C$ are called the pre–shape and the centred pre–shape of $\boldsymbol{X}$, respectively. In this thesis, we will work in terms of pre–shapes which has the advantage that

they are of full rank. Since $||\boldsymbol{Z}|| = 1$, the space of all pre–shapes is $S^{(k-1)m-1}$, i.e. the hypersphere of unit radius in $(k-1)m$ real dimensions and it is commonly denoted as S_m^k. Formally, it is the orbit space of the non–coincident k–point configurations in $\mathbb{R}^m$ under the action of translation and isotropic scaling. The term *pre–shape* was coined by Kendall (1984) and indicates that only rotation remains to be removed to obtain the shape of the original configuration $\boldsymbol{X}$.

In order to remove rotation, all rotated versions of the pre–shape $\boldsymbol{Z}$ are identified with each other to form the equivalence class

$$[X] = \{\boldsymbol{Z}\boldsymbol{\Gamma} : \boldsymbol{\Gamma} \in SO(m)\}, \tag{2.2}$$

where $\boldsymbol{\Gamma}$ denotes an $m \times m$ rotation matrix. As a member of $SO(m)$, i.e. the special orthogonal group in m dimensions, $\boldsymbol{\Gamma}$ satisfies $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{I}_m$ and $|\boldsymbol{\Gamma}| = 1$, where $|.|$ denotes the determinant of a matrix. The shape of the $k \times m$ matrix $\boldsymbol{X}$ is the set (2.2), and the corresponding shape space is commonly denoted as Σ_m^k. Formally, Σ_m^k is the orbit space of the non–coincident k–point configurations in $\mathbb{R}^m$ under the action of the Euclidean similarity transformations. In relation to the pre–shape space, Σ_m^k is the quotient space of S_m^k under the action of $SO(m)$, and the equivalence classes of the form (2.2) are non–overlapping fibres on the pre–shape space. The dimension of Σ_m^k is

$$M = km - m - 1 - m(m-1)/2 \tag{2.3}$$

as the original configuration $\boldsymbol{X}$ has km coordinates, Helmertising then reduces the dimension by m, isotropic rescaling by one and finally, $m(m-1)/2$ dimensions are lost when the rotation information is removed.

2.1.2 Metrics on Shape Space

It is possible to define a metric on Σ_m^k in order to fully define the non–Euclidean shape metric space. Given two configurations $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with corresponding pre–shapes $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, a distance which is invariant under the Euclidean similarity transformations is

given by the full Procrustes distance

$$d_F(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \inf_{\substack{\boldsymbol{\Gamma} \in SO(m) \\ \beta > 0}} ||\boldsymbol{Z}_2 - \beta \boldsymbol{Z}_1 \boldsymbol{\Gamma}||,$$

Alternatively, the partial Procrustes distance can be used which is defined as

$$d_P(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \inf_{\boldsymbol{\Gamma} \in SO(m)} ||\boldsymbol{Z}_2 - \boldsymbol{Z}_1 \boldsymbol{\Gamma}||.$$

A third possible distance is given by the Riemmanian metric in shape space (Kendall, 1984) which is related to the above distance through

$$\rho(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \arcsin\big(d_F(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\big) = 2 \arcsin\big(d_P(\boldsymbol{Z}_1, \boldsymbol{Z}_2)/2\big). \tag{2.4}$$

Geometrically, $d_P(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ is the closest chordal distance on the pre–shape sphere between the rotated version of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, and $\rho(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ is the closest great circle distance. A further discussion of the above distances can be found in Kendall (1984, 1989), Le & Kendall (1993) and Small (1996).

2.1.3 Procrustes Analysis

Procrustes methods are (mainly descriptive) tools for analysing landmark data which use the similarity transformations to match configuration matrices as closely as possible with respect to a least–squares criterion. They can be traced back to Mosier (1939) and also find application in the comparison of (non–configuration) matrices Mardia *et al.* (e.g. 1979, p.416). Here we consider the case where the underlying perturbation model for the configuration matrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ at hand has the form

$$\boldsymbol{X}_i = \beta_i(\boldsymbol{\mu} + \boldsymbol{E}_i)\boldsymbol{\Gamma}_i + \mathbf{1}_k \boldsymbol{\gamma}_i^T, \quad i = 1, \ldots, n. \tag{2.5}$$

where the $\boldsymbol{E}_i$ are *i.i.d.* zero-mean $k \times m$ error matrices which follow an underlying km–variate distribution F_{E}. As (2.5) involves scaling, rotation and translation, Procrustes matching in this case involves the full set of similarity transformations.

Given two $k \times m$ configuration matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, the full ordinary Procrustes analysis (OPA) involves minimising

$$D^2_{\text{OPA}} = ||\boldsymbol{X}_2 - \beta \boldsymbol{X}_1 \boldsymbol{\Gamma} - \mathbf{1}_k \boldsymbol{\gamma}^T||^2 \tag{2.6}$$

over rotation $\boldsymbol{\Gamma} \in SO(m)$, scaling $\beta > 0$ and translation $\boldsymbol{\gamma} \in \mathbb{R}^m$. As described in Dryden & Mardia (1998, Chapter 5), the optimal values of the matching parameters can be found analytically, and the corresponding minimum has the form

$$\text{OSS}(\boldsymbol{X}_1, \boldsymbol{X}_2) = ||\boldsymbol{X}_2||^2 \sin^2 \rho(\boldsymbol{X}_1, \boldsymbol{X}_2),$$

where $\rho(\boldsymbol{X}_1, \boldsymbol{X}_2)$ denotes the Riemannian distance defined in (2.4). A variant of the full ordinary Procrustes analysis is the partial ordinary Procrustes analysis which involves minimising (2.6) over rotation and translation only. This does not change the optimal rotation matrix and translation vector, but the corresponding minimum then has the form

$$\text{OSS}_p(\boldsymbol{X}_1, \boldsymbol{X}_2) = ||\boldsymbol{X}_1||^2 + ||\boldsymbol{X}_2||^2 - 2||\boldsymbol{X}_1||||\boldsymbol{X}_2|| \cos \rho(\boldsymbol{X}_1, \boldsymbol{X}_2). \tag{2.7}$$

Note that it always holds that $\text{OSS}_p(\boldsymbol{X}_1, \boldsymbol{X}_2) = \text{OSS}_p(\boldsymbol{X}_2, \boldsymbol{X}_1)$, whereas $\text{OSS}(\boldsymbol{X}_1, \boldsymbol{X}_2) \neq \text{OSS}(\boldsymbol{X}_2, \boldsymbol{X}_1)$ unless the configurations are of the same size.

When a random sample of configuration matrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is available, a generalisation of the full OPA can be used to optimally rotate, translate and scale the configurations relative to each other. The idea of full generalised Procrustes analysis (GPA) was originally proposed by Kristof & Wingersky (1971); other work on this topic includes Gower (1975), Langron & Collins (1985), Goodall & Bose (1987) and Goodall (1991). Here, the appropriate least–squares criterion is

$$\text{G}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ||(\beta_i \boldsymbol{X}_i \boldsymbol{\Gamma}_i + \mathbf{1}_k \boldsymbol{\gamma}_i^T) - (\beta_j \boldsymbol{X}_j \boldsymbol{\Gamma}_j + \mathbf{1}_k \boldsymbol{\gamma}_j^T)||^2 \tag{2.8}$$

which is to be minimised over rotation, translation and scale subject to the constraint

$$S \left\{ \frac{1}{n} \sum_{i=1}^{n} (\beta_i \boldsymbol{X}_i \boldsymbol{\Gamma}_i + \mathbf{1}_k \boldsymbol{\gamma}_i^T) \right\} = 1.$$

For planar data, the minimum of the above expression can be found analytically (Kent, 1994). If $m \geq 3$, however, an iterative procedure has to be applied to minimise the matching parameters in turn. The corresponding algorithm is due to Gower (1975) and Ten Berge (1977) and is summarised in Appendix A (cf. also Dryden & Mardia, 1998, pp.90). It can be shown that

$$\inf_{\substack{\boldsymbol{\Gamma}_i \in SO(m) \\ \boldsymbol{\gamma}_i \in \mathbb{R}^m,\, \beta_i \in \mathbb{R}}} \mathrm{G}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \inf_{\boldsymbol{\mu}: S(\boldsymbol{\mu})=1} \sum_{i=1}^{n} \sin^2 \rho(\boldsymbol{X}_i, \boldsymbol{\mu}). \tag{2.9}$$

The shape of the minimising configuration, $[\hat{\mu}]$ say, therefore is the sample Fréchet mean of the shapes $[X_1], \ldots, [X_n]$ with respect to the full Procrustes distance, where the Fréchet mean is a generalisation of the expectation in Euclidean space. Given a density $f(.)$ on a general metric space $(\boldsymbol{M}, dist)$, then the general definition of the Fréchet mean is

$$\arg \inf_{x \in M} \int_M dist^2(x, y) f(y) dy,$$

see for example Le & Kume (2000). Existence and uniqueness thereby depend on the chosen metric. In our case (2.9), $[\hat{\mu}]$ is unique if the data is sufficiently concentrated in relation to the curvature of the corresponding shape space (Le, 1995). In that case it can serve as an estimate of the mean of the distribution $Q_{[X]}$ in shape space which is induced by $\boldsymbol{\mu}$ and the error distribution F_{E} in (2.5).

An important question is whether $[\hat{\mu}]$ is consistent for the population Fréchet mean

$$\left[\mu_{[X]}\right] = \arg \min_{[Y] \in \Sigma_m^k} \int_{\Sigma_m^k} \sin^2 \rho([X], [Y])\, dQ_{[X]}.$$

Le (1998) gives necessary and sufficient conditions for the consistency of $[\hat{\mu}]$ in the planar case; cf. also Kent & Mardia (1997) and Bhattacharya & Patrangenaru (2003). However, note that the shape of the mean configuration $\boldsymbol{\mu}$ in landmark space, cf. (2.5), is not always the same as the population Fréchet mean of the induced distribution in shape space, i.e. it is possible that $[\mu] \neq \left[\mu_{[X]}\right]$. In particular for $m \geq 3$ dimensions it is difficult to identify the mean in shape space which is induced by a distribution defined in landmark space. This will be further illustrated in Section 5.2.2, cf. Figure 5.1 .

2.1.4 Procrustes Tangent Space

The Procrustes tangent space is a Euclidean approximation of the shape space in the vicinity of a particular point in shape space (the pole of the tangent projection). It is of fundamental importance because it allows us to linearly approximate the non–Euclidean geometry of the shape space and facilitates the use of standard multivariate techniques to tackle many problems which arise in shape analysis. For planar data it was formulated by Kent (1994) and in higher dimensions by Dryden & Mardia (1993); see also Kent (1995), Small (1996) and Kendall *et al.* (1999, Chapter 6) for further discussion. In this thesis, we consider spaces that are tangent to the pre–shape sphere. It is also possible to formulate the procedure in terms of centred pre–shapes (Kent & Mardia, 2001).

Let $\boldsymbol{Z}_\mu$ be the $(k-1) \times m$ pre–shape corresponding to a $k \times m$ configuration matrix $\boldsymbol{\mu}$ and suppose that we are interested in the space tangent to the shape space at the point $[\mu]$. It can be expressed in terms of a linear subspace of the space tangent to the pre–shape sphere at $\boldsymbol{Z}_\mu$ which has the form

$$\mathcal{T}_{\boldsymbol{Z}_\mu}(S_m^k) = \big\{\boldsymbol{M} \in \mathbb{R}^{(k-1)\times m} : \mathrm{tr}\{\boldsymbol{Z}_\mu^T \boldsymbol{M}\} = 0\big\},$$

and hence contains all real–valued matrices $\boldsymbol{M}$ of the appropriate dimension which are orthogonal to the pole $\boldsymbol{Z}_\mu$.

The above space is too large for our purposes and for invariance under rotation further constraints have to be imposed. The resulting space is commonly called the *horizontal subspace* of $\mathcal{T}_{\boldsymbol{Z}_\mu}(S_m^k)$ (e.g. Kendall *et al.*, 1999, p.109) and has the form

$$\mathcal{H}_\mu(S_m^k) \;=\; \big\{\boldsymbol{M} \in \mathbb{R}^{(k-1)\times m} : \mathrm{tr}\{\boldsymbol{Z}_\mu^T \boldsymbol{M}\} = 0 \quad \text{and} \quad \boldsymbol{Z}_\mu^T \boldsymbol{M} = \boldsymbol{M}^T \boldsymbol{Z}_\mu\big\}, \tag{2.10}$$

where the symmetry constraint ensures that $\boldsymbol{M}$ is optimally rotated with respect to $\boldsymbol{Z}_\mu$ (e.g. Kent & Mardia, 2001). Considering all constraints, the dimensions of $\mathcal{H}_\mu(S_m^k)$ is $M = km - m - 1 - m(m-1)/2$, i.e. the same as the corresponding shape space Σ_m^k.

In practice, a pre–shape $\boldsymbol{Z}$ can be projected into $\mathcal{H}_\mu(S_m^k)$ by first rotating it as closely as possible to the pole using (2.6). The optimally rotated version $\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}$ can then be projected into $\mathcal{H}_\mu(S_m^k)$ using the vectorise operator vec(.) defined in (0.1) by

$$\begin{aligned} H_\mu : S_m^k &\rightarrow \mathcal{H}_\mu(S_m^k) \\ \boldsymbol{Z} &\mapsto \left(\boldsymbol{I}_{km-m} - \text{vec}(\boldsymbol{Z}_\mu)\text{vec}(\boldsymbol{Z}_\mu)^T\right)\text{vec}(\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}) = \tilde{\boldsymbol{v}}. \end{aligned} \tag{2.11}$$

The corresponding matrix can be obtained using the inverse vectorise operator (0.2), i.e. $\tilde{\boldsymbol{V}} = \text{vec}_m^{-1}(\tilde{\boldsymbol{v}})$. As it satisfies $\text{tr}\{\tilde{\boldsymbol{V}}^T\boldsymbol{Z}_\mu\} = 0$ and $\tilde{\boldsymbol{V}}^T\boldsymbol{Z}_\mu = \boldsymbol{Z}_\mu^T\tilde{\boldsymbol{V}}$, it is also an element of $\mathcal{H}_\mu(S_m^k)$. It can be shown that $||\tilde{\boldsymbol{V}}|| = d_F(\boldsymbol{Z}_\mu, \boldsymbol{Z})$, where $d_F(\boldsymbol{Z}_\mu, \boldsymbol{Z})$ is the full Procrustes distance between $\boldsymbol{Z}_\mu$ and $\boldsymbol{Z}$.

2.1.5 Geodesics in Shape Space and Exponential Map

In general, geodesics in a metric space are the curves which take the "shortest path" between two points. Here, we are interested in the shortest path between points $[Z_1]$ and $[Z_2]$ in Σ_m^k. As above, this can be defined in terms of optimally rotated pre–shapes. A geodesic on S_m^k between two orthogonal pre–shapes $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ can be defined as

$$\Gamma_{\boldsymbol{Z}_2}(s) = \cos s\, \boldsymbol{Z}_1 + \sin s\, \boldsymbol{Z}_2, \quad 0 < s \leq \pi/2, \tag{2.12}$$

where s denotes the Riemannian distance travelled from $\boldsymbol{Z}_1$ to $\Gamma_{\boldsymbol{Z}_2}(s)$. The direction of the geodesic at the starting point is given by $d\Gamma_{\boldsymbol{Z}_2}(s)/ds|_{s=0} = \boldsymbol{Z}_2$. Proposition 6.1 of Kendall *et al.* (1999) states that if a geodesic in S_m^k starts off in a horizontal direction, i.e. if the tangent vector at $s = 0$ is an element of (2.10), then its tangent vectors remain horizontal throughout its length. To ensure that the geodesic only contains optimally rotated pre–shapes, we therefore need to modify the endpoint in (2.12) to

$$\tilde{\boldsymbol{Z}}_2 = \frac{1}{\sin s_0}\left\{\boldsymbol{Z}_2\hat{\boldsymbol{\Gamma}} - \cos\, s_0\boldsymbol{Z}_1\right\},$$

where $\hat{\boldsymbol{\Gamma}}$ is the rotation matrix which optimally rotates $\boldsymbol{Z}_2$ to match $\boldsymbol{Z}_1$ and s_0 is the Riemannian distance between the two pre–shapes.

Being optimally rotated, $\tilde{\boldsymbol{Z}}_2$ has the property $\boldsymbol{Z}_1^T\tilde{\boldsymbol{Z}}_2 = \tilde{\boldsymbol{Z}}_2^T\boldsymbol{Z}_1$. Moreover, it satisfies $\mathrm{tr}\{\tilde{\boldsymbol{Z}}_2^T\boldsymbol{Z}_1\} = 0$ and $||\tilde{\boldsymbol{Z}}_2|| = 1$ as required. Replacing $\boldsymbol{Z}_2$ by $\tilde{\boldsymbol{Z}}_2$ in (2.12) then yields

$$\begin{aligned}\gamma_{\hat{\boldsymbol{\Gamma}}}(s) &= \cos s\,\boldsymbol{Z}_1 + \frac{\sin s}{\sin s_0}\left\{\boldsymbol{Z}_2\hat{\boldsymbol{\Gamma}} - \cos s_0\,\boldsymbol{Z}_1\right\} \\ &= \frac{1}{\sin s_0}\{\sin(s_0 - s)\,\boldsymbol{Z}_1 + \boldsymbol{Z}_2\hat{\boldsymbol{\Gamma}}\}, \qquad 0 < s \leq s_0. \end{aligned} \tag{2.13}$$

As $d\gamma_{\hat{\boldsymbol{\Gamma}}}(s)/ds|_{s=0} \in \mathcal{H}_{\boldsymbol{Z}_1}(S_m^k)$, (2.13) is a geodesic on S_m^k which only contains pre–shapes which are optimally rotated with respect to $\boldsymbol{Z}_1$ and therefore corresponds to the geodesic between the points $[Z_1]$ and $[Z_2]$ in Σ_m^k; cf. Kendall *et al.* (1999, Chapter 6).

Based on the geodesic it is possible to define the *exponential map* which – given a pole $\boldsymbol{Z}_\mu$ – identifies vectors in $\mathcal{H}_\mu(S_m^k)$ with (optimally rotated) pre–shapes whose Riemannian distance from $\boldsymbol{Z}_\mu$ is equal to the length of the tangent vector, i.e.

$$\begin{aligned}\exp: \quad \mathcal{H}_\mu(S_m^k) &\to S_m^k \\ \boldsymbol{M} &\mapsto \Gamma_{\boldsymbol{M}/||\boldsymbol{M}||}(||\boldsymbol{M}||) = \gamma_{\boldsymbol{I}}(||\boldsymbol{M}||).\end{aligned}$$

Using the inverse of the exponential map it is therefore possible to obtain tangent vectors in $\mathcal{H}_\mu(S_m^k)$ whose length is equal to the Riemannian distance of the corresponding pre–shape to the pole. The inverse exponential map can then be formulated as

$$\begin{aligned}\exp^{-1}: \quad S_m^k &\to \mathcal{H}_\mu(S_m^k) \\ \boldsymbol{Z} &\mapsto \rho(\boldsymbol{Z}, \boldsymbol{Z}_\mu)\cdot H_\mu(\boldsymbol{Z})/||H_\mu(\boldsymbol{Z})||,\end{aligned} \tag{2.14}$$

where $H_\mu(.)$ is defined in (2.11). The resulting vector $\boldsymbol{v}^\dagger = \exp^{-1}(\boldsymbol{Z})$ has the same direction as $\tilde{\boldsymbol{v}} = H_\mu(\boldsymbol{Z})$ but satisfies $||\boldsymbol{v}^\dagger|| = \rho(\boldsymbol{Z}, \boldsymbol{Z}_\mu) = \arcsin(||\tilde{\boldsymbol{v}}||)$.

2.2 Stochastic Processes

A stochastic process is a family of random variables $\{Z(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$ defined on a domain $\mathcal{D}$. In this thesis, we will consider random fields and discrete time series.

2.2.1 Spatial Statistics

In spatial statistics, the data at hand have the form $z(\boldsymbol{x}_1), \ldots z(\boldsymbol{x}_n)$, where $\boldsymbol{x}_i \in \mathcal{D}$ $(i = 1, \ldots, n)$ denotes a site within the domain, and $z(\boldsymbol{x}_i)$ denotes the value of a random variable $Z(\boldsymbol{x}_i)$ which has been observed at site $\boldsymbol{x}_i$. The main feature of spatial data is that the set $\{z(\boldsymbol{x}_i)\}_{i=1}^n$ does not represent a sample of size n. Instead, it is regarded as an incomplete observation of one realisation $\{z(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$ of the underlying random random field $\{Z(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$. In this probabilistic framework we are therefore not interested in trying to reconstruct the exact form of the deterministic function $z(.)$. Instead, the aim is to carry out inference about the underlying spatial process. The following treatment is largely based on Schabenberger & Gotway (2005). Other monographs on spatial statistics include Ripley (1981), Cressie (1993) and Wackernagel (2003), and for a theoretical account on random fields see Adler (1981).

2.2.1.1 Stationarity, Moments and Isotropy

A random field $Z(\boldsymbol{x})$ is called *strictly stationary* if its spatial distribution is invariant under translation of the coordinates, i.e. if

$$\begin{aligned}&\mathrm{P}\big(Z(\boldsymbol{x}_1) < z_1, Z(\boldsymbol{x}_2) < z_2 \ldots, Z(\boldsymbol{x}_n) < z_n\big) = \\ &\qquad \mathrm{P}\big(Z(\boldsymbol{x}_1 + \boldsymbol{h}) < z_1, Z(\boldsymbol{x}_2 + \boldsymbol{h}) < z_2 \ldots, Z(\boldsymbol{x}_n + \boldsymbol{h}) < z_n\big), \quad \forall n \in \mathbb{N},\, \boldsymbol{h} \in \mathcal{D}.\end{aligned}$$

However, strict stationarity is a very stringent assumption, and often it is sufficient to assume stationarity conditions only for the first and second moment of $Z(\boldsymbol{x})$. This weaker form of stationarity is called *second–order stationarity* and implies that

$$\begin{aligned}\mathrm{E}\big(Z(\boldsymbol{x})\big) &= \mu \qquad \forall\, \boldsymbol{x} \in \mathcal{D}, \\ \mathrm{Cov}\big(Z(\boldsymbol{x}), Z(\boldsymbol{x} + \boldsymbol{h})\big) &= \sigma(\boldsymbol{h}) \quad \forall\, \boldsymbol{x}, \boldsymbol{h} \in \mathcal{D},\end{aligned}$$

where $\mu \in \mathbb{R}$ denotes the constant mean and $\sigma(.)$ is the (auto–)covariance function of the random field which plays an important role in spatial modelling.

As $\sigma(.)$ only depends on the lag–vector $\boldsymbol{h}$ under the stationarity assumption, it follows that the variance $\mathrm{Cov}\big(Z(\boldsymbol{x}), Z(\boldsymbol{x})\big) = \sigma(\mathbf{0}) = \sigma^2$ is the same everywhere. Note that not every function on the domain can be used as a covariance function, and care must be taken to ensure that the choice of $\sigma(.)$ satisfies

$$\mathrm{Var}\big(\sum_{i=1}^{n} w_i Z(\boldsymbol{x}_i)\big) = \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \geq 0 \quad \forall\, \boldsymbol{w} \in \mathbb{R}^n; \boldsymbol{x}_i \in \mathcal{D}, \tag{2.15}$$

where $(\boldsymbol{\Sigma})_{ij} = \sigma(\boldsymbol{x}_i - \boldsymbol{x}_j)$. This property of valid covariance functions guarantees that any linear combination of any collection of sample points has a positive variance.

The covariance function of a random field determines many of its properties, e.g. the near–origin behaviour of $\sigma(.)$ determines the spatial continuity of $Z(\boldsymbol{x})$ in the mean–square sense, and the smoothness of $Z(\boldsymbol{x})$ depends on the number of times its covariance function is differentiable at the origin (cf. Schabenberger & Gotway, 2005, p.52). Another important property of the covariance function is that it determines the direction of the correlation structure of the random field. If the value $\sigma(\boldsymbol{h})$ only depends on the length of the lag vector $\boldsymbol{h}$, i.e. if $\sigma(\boldsymbol{h}) = \sigma(||\boldsymbol{h}||)$, then $\sigma(.)$ is called *isotropic*. Note that isotropy implies a rotation invariance of $Z(\boldsymbol{x})$ whereas stationarity implies an invariance under translation. See Abrahamsen (1997) for a list of isotropic covariance functions.

2.2.1.2 Covariance Estimation

In practice, the covariance function $\sigma(.)$ is unknown. Many methods in spatial statistics such as spatial prediction (cf. Section 2.2.1.3), however, require a functional descriptor of the covariance structure, so that estimating the covariance function is an important task in the spatial context. Instead of trying to estimate the covariance function directly, estimation is thereby often based on the *semivariogram*

$$\sigma^*(\boldsymbol{h}) = \tfrac{1}{2}\mathrm{Var}\big\{Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\big\} = \sigma(\mathbf{0}) - \sigma(\boldsymbol{h})$$

which has the practical benefit that $\sigma^*(.)$ is more robust against violations of the stationarity assumption.

Given the observed data $z(\boldsymbol{x}_1), \ldots z(\boldsymbol{x}_n)$, the semivariogram can be estimated using the *semivariogram cloud* which is a plot of the squared differences $\left\{z(\boldsymbol{x}_i) - z(\boldsymbol{x}_j)\right\}^2$ against the associated lag–vector. Provided the mean of the random field is constant, the $\left\{z(\boldsymbol{x}_i) - z(\boldsymbol{x}_j)\right\}^2$ unbiasedly estimate the semivariogram at lag $\boldsymbol{x}_i - \boldsymbol{x}_j$ (e.g. Schabenberger & Gotway, 2005, p.153). However, the number of pairwise differences can be very large so that the raw semivariogram cloud may be not informative. Matheron (1962) therefore suggests averaging the squared differences of points whose lag vector falls into the class $N(\boldsymbol{h}) = \boldsymbol{h} \pm \boldsymbol{\epsilon}$, where the choice of $\boldsymbol{\epsilon} \in \mathcal{D}$ is left to the user. If isotropy can be assumed, then these classes are groups of points whose distance falls into $N(||\boldsymbol{h}||) = ||\boldsymbol{h}|| \pm \epsilon$, where $\epsilon > 0$. The resulting estimator of $\sigma^*(||\boldsymbol{h}||)$ then has the form

$$\hat{\sigma}^*(||\boldsymbol{h}||) = \frac{1}{2\,|N(||\boldsymbol{h}||)|} \sum_{N(||\boldsymbol{h}||)} \left\{z(\boldsymbol{x}_i) - z(\boldsymbol{x}_j)\right\}^2,$$

where $|N(||\boldsymbol{h}||)|$ denotes the number of distinct pairs in $N(||\boldsymbol{h}||)$.

The above method yields unbiased estimates of $\sigma^*(.)$ for a discrete set of lag values $||\boldsymbol{h}||$, i.e. for the centres of the chosen distance classes $N(||\boldsymbol{h}||)$, and a plot of the resulting values $\sigma^*(||\boldsymbol{h}||)$ against the corresponding lag lengths $||\boldsymbol{h}||$ is called an *empirical semivariogram*. In order to obtain estimates of $\sigma^*(.)$ at any arbitrary lag, a parametric semivariogram model $\sigma^*(\boldsymbol{h}) = \sigma(\boldsymbol{0}) - \sigma(\boldsymbol{h})$ can be fitted to the empirical semivariogram, e.g. using a least–squares method. For more information see for example Olea (2006).

2.2.1.3 Spatial Prediction

A frequent objective in spatial statistics is to predict the value of the random field $Z(\boldsymbol{x})$ at some specified location $\boldsymbol{x}_0 \in \mathcal{D}$. Methods for spatial prediction are typically known as *kriging* – a term coined by Matheron in honour of D.G. Krige whose work laid the preliminary groundwork for the field of spatial statistics (Krige, 1951; Matheron, 1963). The derivation of a predictor commences with the choice of a loss function which measures the loss incurred by using a prediction $\hat{Z}^*(\boldsymbol{x}_0)$ instead of $Z(\boldsymbol{x}_0)$. The most common choice is the squared–error loss function under which the average loss is the

prediction mean squared error $\text{PMSE} = \text{E}\big[\big(\hat{Z}^*(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\big)^2\big]$. It can be shown that the predictor which miminises the PMSE is the conditional expectation of $Z(\boldsymbol{x}_0)$ given the data at hand (e.g. Schabenberger & Gotway, 2005, p.218). In most cases, however, this will be difficult to establish. It is therefore advisable to restrict the search for a good predictor to the class of linear predictors. In that case, the new objective is to find the *Best Linear Unbiased Predictor (BLUP)* under squared–error loss.

In this thesis we consider the case where the constant mean μ of a second–order stationary random field $Z(\boldsymbol{x})$ is known. In that case, a general linear predictor has the form

$$\hat{Z}^*_{\text{L}}(\boldsymbol{x}_0) = \mu + \sum_{i=1}^{n} \tilde{u}_i\big(Z(\boldsymbol{x}_i) - \mu\big), \tag{2.16}$$

so that the PMSE becomes a function of the weight vector $\tilde{\boldsymbol{u}} = (\tilde{u}_1, \ldots, \tilde{u}_n)^T$. For a given covariance function, the optimal weight vector can therefore be found by setting the gradient of the objective function to zero. As the mean of the random field is known and constant over the entire domain, a linear predictor of the form (2.16) is always unbiased so that no weight constraints have to be imposed to guarantee unbiasedness (e.g. Wackernagel, 2003, pp.24). The resulting equation system has the solution $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$, where $(\boldsymbol{\Sigma})_{ij} = \sigma(\boldsymbol{x}_i - \boldsymbol{x}_j)$ and $\boldsymbol{\sigma} = (\sigma(\boldsymbol{x}_1 - \boldsymbol{x}_0), \ldots, \sigma(\boldsymbol{x}_n - \boldsymbol{x}_0))^T$. The BLUP of $Z(\boldsymbol{x}_0)$ is then given by

$$\hat{Z}^*_{\text{BLUP}}(\boldsymbol{x}_0) = \mu + \boldsymbol{u}^T(\boldsymbol{Z} - \mu\mathbf{1}_n) = \mu + \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \mu\mathbf{1}_n), \tag{2.17}$$

where $\mathbf{1}_n$ denotes the n–vector of ones and $\boldsymbol{Z} = \big(Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)\big)^T$. As $\boldsymbol{\sigma}$ depends on $\boldsymbol{x}_0$, the optimal weights adapt to the location of interest. If the observed data vector $\boldsymbol{z} = \big(z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_n)\big)^T$ is inserted into (2.17), then $\hat{Z}_{\text{BLUP}}(\boldsymbol{x}_0) = \mu + \boldsymbol{u}^T(\boldsymbol{z} - \mu\mathbf{1}_n)$ is the predicted value of $Z(\boldsymbol{x}_0)$. This method of spatial prediction is called *simple kriging*.

2.2.2 Time Series Analysis and Autoregressive Models

A time series can be seen as a random field over the domain $\mathcal{D} = \mathbb{R}^+$, where $\mathcal{D}$ corresponds to the positive time line, i.e. it can be denoted as $\{X(t) : t \in \mathbb{R}^+\}$. In most

practical cases, however, the domain will be a discrete set of indices which represent equally spaced time points so that the time series will be denoted as $\{X(t) : t \in \mathbb{N}\}$ or equivalently $\{X_t\}_{t\geq 1}$. The following treatment will largely be based on the books by Chatfield (1996) and Cryer & Chan (2008).

2.2.2.1 Stationarity and Moments

Like in the case of spatial data, a frequently made assumption in the context of time series data is that of stationarity. Both strict and second–order stationarity thereby imply that an autocovariance function can be defined as

$$\gamma(s) = \mathrm{E}\big((X_t - \mu)(X_{t+s} - \mu)\big) = \mathrm{Cov}\big(X_t, X_{t+s}\big), \quad s \geq 0.$$

Dividing by the common variance $\sigma^2 = \gamma(0)$ then yields the autocorrelation function $\rho(s) = \gamma(s)/\gamma(0)$. The autocorrelation function provides valuable information about the underlying data generating process of $\{X_t\}_{t\geq 1}$. Tentatively assuming stationarity, it can be estimated from an observed data sequence $\{x_t\}_{t=1}^n$ by the sample autocorrelation function

$$r(s) = \frac{\sum_{t=s+1}^{n}(x_t - \bar{x})(x_{t-s} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}, \quad s = 1, \ldots n-1,$$

where $\bar{x} = n^{-1}\sum_{t=1}^{n} x_t$ denotes the mean of the observed values. The sequence of the sample autocorrelations coefficients $r(s)$ provides insight into the dependence structure of the underlying probability law. A plot of $r(s)$ versus lag s is called a *correlogram*.

2.2.2.2 Autoregressive Models

In many practical cases where a sequence of values $\{x_t\}_{t=1}^n$ has been observed, it may be beneficial to make a parametric assumption about the underlying process $\{X_t\}_{t\geq 1}$. One popular parametric model is the *autoregressive model* which was introduced by Yule

(1926). An autoregressive model of order p (AR(p) model) satisfies the equation

$$X_t = \psi_1 X_{t-1} + \ldots + \psi_p X_{t-p} + \epsilon_t, \tag{2.18}$$

i.e. the current value of the time series is a linear combination of the p most recent past values plus a random innovation term. The process $\{\epsilon_t\}_{t\geq 1}$ is thereby assumed to be a *white noise process*, i.e. the ϵ_t are assumed to be uncorrelated zero–mean random variables with unit variance.

Given an AR(p) process, it is of interest whether or not it is stationary. This can be investigated using the *characteristic equation* of (2.18) which has the form

$$\psi(x) = 1 - \psi_1 x - \ldots - \psi_p x^p = 0. \tag{2.19}$$

It can be shown (cf. e.g. Box *et al.*, 2008, Section 3.2.1) that the process is stationary iff the roots of the characteristic equation exceed one in absolute value, i.e. if they lie outside the unit circle in the complex plane. Let $G_1^{-1}, \ldots, G_p^{-1}$ denote the roots of (2.19). An explicit expression of the autocorrelation function of $\{X_t\}_{t\geq 1}$ can be formulated in terms of $G_1^{-1}, \ldots, G_p^{-1}$ and has the form

$$\rho(s) = A_1 G_1^s + \ldots + A_p G_p^s, \quad s \geq 0, \tag{2.20}$$

where the coefficients $A_1, \ldots, A_p$ are chosen to satisfy some conditions based on the properties $\rho(0) = 1$ and $\rho(s) = \rho(-s)$ of $\rho(.)$; see for example Chatfield (1996, p.38).

2.3 MCMC Simulations for Bayesian Inference

Whereas the unknown parameters of a distribution are considered as fixed in classical statistics, inference within the Bayesian framework is carried out in terms of probability statements. Here, we review Markov Chain Monte Carlo (MCMC) methods, a group of simulation methods which facilitate the application of Bayesian methods to many situations which are too complicated to work with analytically.

2.3.1 Bayes' Theorem

Let the data at hand be an n–vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, and let θ denote a parameter which determines the data generating process through a likelihood function $L(\boldsymbol{x}|\theta)$ which specifies the probability distribution of the underlying random variables *given a particular value of the parameter*. As θ is assumed to be random here, in addition to specifying a likelihood, Bayesian inference involves specifying a *prior distribution* $\pi(\theta)$. This prior distribution should capture any information about θ which is available before the data is observed. Within this framework, inference about θ can be based on the *posterior distribution* with density

$$\pi(\theta|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\theta)\pi(\theta)}{\int L(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \propto L(\boldsymbol{x}|\theta)\pi(\theta), \tag{2.21}$$

i.e. the posterior distribution is the conditional distribution of θ *given the observed data*. The above formula is referred to as *Bayes' Theorem*. The primary task of any application of the Bayesian framework is to develop the joint probability model $L(\boldsymbol{x}|\theta)\pi(\theta)$ and to perform the necessary computations to summarise $\pi(\theta|\boldsymbol{x})$ in appropriate ways (Gelman *et al.*, 2004, p.8).

2.3.2 Prior Distributions

There are various approaches for choosing a prior distribution (e.g. Kass & Wasserman, 1996). Here we mention the two approaches of choosing a prior we will employ in this thesis, namely the *conjugate prior* and the *non–informative prior*. The *conjugate prior* is a prior distribution for the parameter of interest which (depending on the likelihood) is chosen in a way that its posterior distribution belongs to the same parametric family as itself. Formally, if $\mathcal{F}$ is a class of likelihood functions $L(\boldsymbol{x}|\theta)$, and $\mathcal{P}$ is a class of prior distributions, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$\pi(\theta|\boldsymbol{x}) \in \mathcal{P} \quad \forall \quad L(.|\theta) \in \mathcal{F} \quad \text{and} \quad \pi(.) \in \mathcal{P},$$

see for example (Gelman *et al.*, 2004).

If no reliable prior information about the parameter of interest is available, then a *non–informative prior* should be chosen which does not favour one particular value of θ over others. Typical non–informative priors are the uniform distribution over the parameter space or *Jeffreys' prior* (Jeffreys, 1946). Non–informative priors are related to classical inference in that the posterior distribution of the parameter primarily contains information inherent in the data at hand.

2.3.3 Posterior Estimation

The posterior distribution represents a compromise between the prior model for the unknown parameter θ and the observed data. Once it has been determined, either a point estimate or an interval estimate of the unknown parameter can be obtained. Commonly used point estimates are the posterior mean or the posterior mode, i.e.

$$\hat{\theta}_{\text{mean}} = \mathrm{E}\big(\pi(\theta|\boldsymbol{x})\big) \quad \text{and} \quad \hat{\theta}_{\text{MAP}} = \arg\max_{\theta\in\Theta} \pi(\theta|\boldsymbol{x}) \tag{2.22}$$

where MAP is short for *Maximum A Posteriori*. To summarise posterior uncertainty, interval estimates of θ can be obtained based on the quantiles of the posterior distribution. For example

$$CI = \{\theta \in \Theta : \theta_{\alpha/2} \geq \theta \geq \theta_{1-\alpha/2}\}, \tag{2.23}$$

where $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$ denote the quantiles of $\pi(\theta|\boldsymbol{x})$, contains $100(1-\alpha)\%$ of the posterior probability. Posterior intervals like this are commonly called *credibility intervals*.

2.3.4 Markov Chain Monte Carlo

The use of Bayesian methods in applied problems greatly increased at the end of the $20th$ century. The reason for this recent popularity is the availability of fast computers combined with the development of Markov Chain Monte Carlo (MCMC) methods.

A *Markov chain* is a sequence of random variables $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$ which is generated over time. The evolution of a Markov chain is governed by a *transition kernel*

$$\mathrm{P}(\boldsymbol{x}, A) = \Pr(\boldsymbol{X}_{t+1} \in A | \boldsymbol{X}_t = \boldsymbol{x}) \quad \forall A \subset \Omega,\ \boldsymbol{x} \in \Omega,$$

where Ω denotes the set of possible values of each $\boldsymbol{X}_t$. The distribution of $\boldsymbol{X}_{t+1}$ therefore only depends on the current state $\boldsymbol{X}_t$ and not on the remainder $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{t-1}\}$ of the history of the chain. This is called the *Markov property* of the chain.

Within the MCMC framework, the aim is to find a transition kernel which (after many iterations) generates samples from a known distribution, namely the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. Luckily, this difficult task can be simplified by the fact that if

$$\pi(\tilde{\boldsymbol{\theta}}|\boldsymbol{x})p(\tilde{\boldsymbol{\theta}}, \breve{\boldsymbol{\theta}}) = \pi(\breve{\boldsymbol{\theta}}|\boldsymbol{x})p(\breve{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}), \quad \tilde{\boldsymbol{\theta}}, \breve{\boldsymbol{\theta}} \in \Theta, \tag{2.24}$$

where Θ denotes the parameter space and $p(\tilde{\boldsymbol{\theta}}, \breve{\boldsymbol{\theta}})$ is the probability $\big($determined by $\mathrm{P}(.,.)\big)$ of jumping from $\tilde{\boldsymbol{\theta}}$ to $\breve{\boldsymbol{\theta}}$, then the posterior distribution is the limiting distribution of $\mathrm{P}(.,.)$; see for example Tierney (1994). Equation (2.24) is called the *reversibility condition*. There are two important generic choices for $\mathrm{P}(.,.)$ which satisfy this condition, namely the Metropolis–Hastings algorithm and the Gibbs sampler.

The Metropolis–Hastings (MH) algorithm was first proposed by Metropolis *et al.* (1953) in the context of statistical physics and subsequently generalised by Hastings (1970). For the MH algorithm, a density $q(.|\boldsymbol{\theta}_t)$ needs to be defined which, conditional on the current parameter value $\boldsymbol{\theta}_t$, generates candidates $\boldsymbol{\phi}$ for the subsequent parameter value $\boldsymbol{\theta}_{t+1}$ in the Markov chain. Let $q(\boldsymbol{\theta}_t, \boldsymbol{\phi})$ denote the corresponding probability of jumping from $\boldsymbol{\theta}_t$ to $\boldsymbol{\phi}$. In most cases, $q(\boldsymbol{\theta}_t, \boldsymbol{\phi})$ will not satisfy condition (2.24). A convenient way to correct this condition is to adjust the transition probabilities $q(\boldsymbol{\theta}_t, \boldsymbol{\phi})$ and $q(\boldsymbol{\phi}, \boldsymbol{\theta}_t)$ by introducing acceptance probabilities $\alpha(\boldsymbol{\theta}_t, \boldsymbol{\phi})$ and $\alpha(\boldsymbol{\phi}, \boldsymbol{\theta}_t)$ for making the actual move. Thus, the new transition probabilities have the form

$$p_{\mathrm{MH}}(\boldsymbol{\theta}_t, \boldsymbol{\phi}) = q(\boldsymbol{\theta}_t, \boldsymbol{\phi})\alpha(\boldsymbol{\theta}_t, \boldsymbol{\phi}) \quad \text{and} \quad p_{\mathrm{MH}}(\boldsymbol{\phi}, \boldsymbol{\theta}_t) = q(\boldsymbol{\phi}, \boldsymbol{\theta}_t)\alpha(\boldsymbol{\phi}, \boldsymbol{\theta}_t),$$

where $\alpha(.,.)$ is to be determined.

It can be shown (e.g. Chib & Greenberg, 1995; Green, 2001) that the acceptance probability for a move from the current parameter vector $\boldsymbol{\theta}_t$ to a candidate $\boldsymbol{\phi}$ for $\boldsymbol{\theta}_{t+1}$ which ensures reversibility has the form

$$\alpha_{\text{HR}}(\boldsymbol{\theta}_t, \boldsymbol{\phi}) = \min\Big\{ \frac{\pi(\boldsymbol{\phi}|\boldsymbol{x})q(\boldsymbol{\phi}, \boldsymbol{\theta}_t)}{\pi(\boldsymbol{\theta}_t|\boldsymbol{x})q(\boldsymbol{\theta}_t, \boldsymbol{\phi})}, 1 \Big\} = \min\Big\{\text{HR}, 1\Big\}, \tag{2.25}$$

where $\text{HR} = \pi(\boldsymbol{\phi}|\boldsymbol{x})q(\boldsymbol{\phi}, \boldsymbol{\theta}_t)/\pi(\boldsymbol{\theta}_t|\boldsymbol{x})q(\boldsymbol{\theta}_t, \boldsymbol{\phi})$ is called the *Hastings ratio*. Under the regularity conditions of irreducibility and aperiodicity (e.g. Smith & Roberts, 1993), the Markov chain $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots\}$ generated by the MH algorithm converges to the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. Note that the MH algorithm does not require knowledge about the normalising constant of $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ because it appears in both the numerator and the denominator of the Hastings ratio.

An important special case of the MH algorithm is the Gibbs sampling approach which derives its name from Gibbs random fields, where it was used for the first time by Geman & Geman (1984). Here, we denote the unknown parameter vector as $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^l)^T$ to distinguish between components and iteration numbers. The idea of the Gibbs sampler is to directly connect the transition kernel to the target distribution and to use the *full–conditional distributions* as proposal distributions. The full conditional distributions $\pi_i(\theta^i|\boldsymbol{\theta}^{(-i)}, \boldsymbol{x})$ are defined as the conditional distributions of the components θ^i $(i = 1, \ldots, l)$ given the data and all the other elements $\boldsymbol{\theta}^{(-i)}$ of $\boldsymbol{\theta}$. Thus, given a current parameter vector $\boldsymbol{\theta}_t = (\theta_t^1, \ldots, \theta_t^l)^T$, the next vector $\boldsymbol{\theta}_{t+1}$ is simulated in l steps using

$$\theta_{t+1}^i \sim \pi_i(\theta^i|\boldsymbol{\theta}_{t+1}^1, \ldots, \boldsymbol{\theta}_{t+1}^{i-1}, \boldsymbol{\theta}_t^{i+1}, \ldots, \boldsymbol{\theta}_t^l), \quad i = 1, \ldots, l$$

as proposal distributions. It can be shown the that Hastings ratio in (2.25) is always equal to one if the full conditional distributions are used to generate candidate values.

If a Markov chain $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots\}$ is generated using either the MH algorithm or the Gibbs sampler, then the values $\boldsymbol{\theta}_t$ will eventually be approximate samples from the posterior distribution of interest. However, the speed of convergence varies from application to application which leads to the practical question how large an initial sample should be discarded for subsequent analysis.

A number of convergence diagnostics have been proposed in the literature (e.g. Gelman, 1996). In this thesis, we monitor convergence visually using *trace plots*. Trace plots are plots of the history of the parameter values over many iterations. A clear sign of non–convergence is for example a trend in the simulated data whereas a converged chain moves around the mode of the posterior distribution. Based on trace plots, it is possible to approximately determine the *burn–in* period of the chain, i.e. the period $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{t_0}\}$ where the generated parameter values cannot yet be considered as samples from the posterior distribution. Having determined the burn–in period, estimates of the characteristics of the posterior distribution such as (2.22) and (2.23) should only be based on the remainder $\{\boldsymbol{\theta}_{t_0+1}, \boldsymbol{\theta}_{t_0+2}, \boldsymbol{\theta}_{t_0+3}, \ldots\}$ of the chain.

2.4 Bootstrap Methods

The bootstrap is a modern approach to statistical inference which falls into the wider class of resampling methods. Resampling methods are computer–intensive techniques which create replicates of the original dataset to assess the uncertainty associated with a quantity of interest without making distributional assumptions. Some resampling techniques go back a long way (e.g. Quenouille, 1949), but it was Efron (1979) who unified ideas and established the theoretical underpinnings for simulation–based statistical analysis. Due to their popularity there are many books on bootstrap methods, including Efron & Tibshirani (1993), Davison & Hinkley (1997) and Chernick (1999). In this section, some of the notation and the general line of argumentation are based on the introductory chapter of Hall (1992) where the bootstrap is described as a direct application of the so–called plug–in (or Russian doll as he calls it) principle.

2.4.1 The Bootstrap as an Application of the Plug–In Principle

Let X be the random variable of interest which follows a certain distribution function F. In order to make quantitative statements about one of its characteristics $\theta = t(F)$ based

on a random sample $\boldsymbol{\mathcal{X}} = \{X_1, \ldots, X_n\}$, the basic idea of the plug–in principle is to estimate θ by first estimating the population distribution function F and then inserting the resulting estimate into the same functional $t(.)$ which calculates θ from F. In most cases F is estimated by the empirical distribution function of the sample at hand, i.e.

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}, \tag{2.26}$$

which assigns a probability mass of $1/n$ to each of the observed values in $\boldsymbol{\mathcal{X}}$. Due to the Gliwenko–Cantelli theorem (e.g. Chung, 1974, p.133), $\hat{F}_n$ has desirable asymptotic properties.

For most parameters $\theta = t(F)$, the functional $t(.)$ involves an integral over F which can be approximated by the corresponding integral over $\hat{F}$, i.e. conditioned on $\boldsymbol{\mathcal{X}}$. The bootstrap makes use of the fact that the uncertainty associated with $\hat{\theta} = t(\hat{F})$ can often also be posed in terms of an integral with respect F. For example if a symmetric 95% confidence interval for θ is to be constructed, a constant c_0 is sought which satisfies

$$\mathrm{P}\big(\hat{\theta} - c_0 \leq \theta \leq \hat{\theta} + c_0\big) = 0.95 \;\Leftrightarrow\; \mathrm{E}_F\big(I_{\{\hat{\theta} - c_0 \leq \theta \leq \hat{\theta} + c_0\}} - 0.95\big) = 0 \tag{2.27}$$

where I_E denotes the indicator function of an event E. To obtain the bootstrap estimate of c_0, the parameter $\theta = t(F)$ is replaced by $\hat{\theta} = t(\hat{F})$ and the expectation is evaluated with respect to $\hat{F}$ instead of F. To find an appropriate replacement for $\hat{\theta}$, a new (re)sampling process is introduced in which $\hat{F}$ takes over the role as population distribution function, i.e. like before, the population is replaced by the sample at hand and calculations are carried out conditional on $\boldsymbol{\mathcal{X}}$.

In a non–parametric setting (and when uniform resampling is applied), a resample $\boldsymbol{\mathcal{X}}^* = \{X_1^*, \ldots, X_n^*\}$ is an unordered collection of n items drawn from $\boldsymbol{\mathcal{X}}$ with replacement, so that each X_i^* has probability of $1/n$ of being equal to any given one of the X_j's, i.e.

$$\mathrm{P}(X_i^* = X_j \,|\, \boldsymbol{\mathcal{X}}) = n^{-1}, \quad 1 \leq i, j \leq n. \tag{2.28}$$

Conditional on $\boldsymbol{\mathcal{X}}$, the X_j^*'s are therefore independent and identically distributed.

Let $\hat{F}^*$ denote the distribution function of a (re)sample $\boldsymbol{\mathcal{X}}^*$ drawn from $\hat{F}$, and let $\hat{\theta}^* = t(\hat{F}^*)$ denote the "resample–approximation" of $\hat{\theta}$. Note that conditional on $\boldsymbol{\mathcal{X}}$, $\hat{\theta}$ is a fixed constant whereas $\hat{\theta}^*$ is a random variable. With these definitions, the bootstrap estimator of c_0 is

$$\hat{c}_0 = \inf\left\{c:\ \mathrm{P}\big(\hat{\theta}^* - c \le \hat{\theta} \le \hat{\theta}^* + c \,|\, \boldsymbol{\mathcal{X}}\big) \ge 0.95\right\},$$

i.e. the 95% quantile of the distribution of $|\hat{\theta} - \hat{\theta}^*|$ conditional on $\boldsymbol{\mathcal{X}}$, is the bootstrap estimator of c_0. The bootstrap confidence interval for θ therefore becomes $[\hat{\theta} - \hat{c}_0, \hat{\theta} + \hat{c}_0]$, and it has an approximate coverage of 0.95. This method for constructing a bootstrap confidence interval is called the *percentile method* by Hall (1992). For other, more complex bootstrap confidence intervals which correct for bias and skewness of F^* see for example Efron (1984) and Efron & Tibshirani (1986).

2.4.2 Monte Carlo Simulation for Approximating Bootstrap Estimates

One problem that frequently arises when using the bootstrap is that the number n^* of possible resamples $\boldsymbol{\mathcal{X}}^*$ from $\boldsymbol{\mathcal{X}}$ grows with n very quickly so that an exact calculation of the desired expected values is usually not feasible. Hence, employing a Monte Carlo simulation presents a practical solution. Within the bootstrap framework, Monte Carlo simulations involve taking B resamples $\{\boldsymbol{\mathcal{X}}_b^*,\, b = 1, \ldots, B\}$ from the original sample. For each of these resamples, a corresponding value $\hat{\theta}_b^*$ is computed and the required expected value is then approximated by an average over the iterations. For the example of constructing a symmetric 95% confidence interval for θ, the distribution of $|\hat{\theta}^* - \hat{\theta}|$ and hence its quantiles cannot be determined easily. Instead, B resamples are taken, and each of them results in a value of $|\hat{\theta}_b^* - \hat{\theta}|$. The value which approximates $\hat{c}_0$ then is the 95% quantile of the empirical distribution function of the $|\hat{\theta}_b^* - \hat{\theta}|$, i.e. the value

$$\hat{c}_0^B = \inf\left\{c:\ \frac{1}{B}\sum_{b=1}^{B} I_{\{|\hat{\theta}_b^* - \hat{\theta}| \le c\}} \le 0.95\right\}.$$

Using the methodology described above, the final bootstrap confidence interval is $[\hat{\theta} - \hat{c}_0^B, \hat{\theta} + \hat{c}_0^B]$ and has an approximate coverage of 0.95.

2.4.3 Improving the Coverage Error – Pivoting

The coverage error of the above general bootstrap interval is

$$\Delta = \Delta(\boldsymbol{\mathcal{X}}, B) = \mathrm{P}\big(\hat{\theta} - \hat{c}_0^B \leq \theta \leq \hat{\theta} + \hat{c}_0^B\big) - 0.95,$$

and it is determined by two sources of randomness, namely the randomness inherent in the initial random sample $\boldsymbol{\mathcal{X}}$ and the randomness caused by the Monte Carlo resampling. In some situations, the error due to the dependence on $\boldsymbol{\mathcal{X}}$ can be reduced by transforming the statistic of interest in a way that its (asymptotic) distribution does not depend on any unknown quantities, and such a statistic is called *(asymptotically) pivotal*.

A well–known example is the case where the distribution F is known to be normal but with unknown mean $\theta = t(F)$ and unknown standard deviation $\sigma = s(F)$. In this parametric case, resamples $\boldsymbol{\mathcal{X}}^*$ are drawn from $\mathrm{N}(\hat{\theta}, \hat{\sigma}^2)$, where $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \hat{\theta})^2}$. Conditional on $\boldsymbol{\mathcal{X}}$, the distribution of the resampled values $\sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ is known to be $\mathrm{N}(0,1)$ which depends on the sample at hand through $\hat{\sigma}$. To eliminate this dependence, a studentisation can be carried out which results in

$$\Big(\frac{\sqrt{n}(\hat{\theta}^* - \hat{\theta})}{\hat{\sigma}} \,|\, \boldsymbol{\mathcal{X}}\Big) \sim \mathrm{N}(0,1) \;\Rightarrow\; \Big(\frac{\sqrt{n}(\hat{\theta}^* - \hat{\theta})}{\hat{\sigma}^*}\Big) \sim t_{n-1}$$

where $\hat{\sigma}^*$ denotes the estimated variance of the (re)sample $\boldsymbol{\mathcal{X}}^*$ and t_{n-1} denotes the t–distribution with $n-1$ degrees of freedom. The studentised statistic $\sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ follows this distribution independently from $\boldsymbol{\mathcal{X}}$ and independently from the true value of θ. Additionally, the t–distribution does not depend on any unknown quantities so that in this case, a confidence interval with exact coverage accuracy can be constructed.

The concept of pivoting can also be applied to non–parametric settings where the test statistic can be transformed so that its asymptotic distribution has the pivotal property. Although the resulting confidence intervals will have a positive coverage error, pivoting usually improves its asymptotic behaviour in that the error decreases with n at a faster rate, e.g. for the non–parametric symmetric percentile intervals described earlier, pivoting increases the coverage accuracy from $O(n^{-1})$ to $O(n^{-2})$ (e.g. Hall, 1992, p.16).

2.4.4 Two–sample Problems

The bootstrap is not restricted to situations where the data are a simple random sample from a single distribution. For example let $\boldsymbol{\mathcal{X}} = \{X_1, \ldots, X_{n_X}\}$ and $\boldsymbol{\mathcal{Y}} = \{Y_1, \ldots, Y_{n_Y}\}$ be samples from two independent distributions F and G, and suppose the parameter of interest has the form $\theta = t(F, G)$. In this two–sample situation, the bootstrap estimator of θ can be obtained by inserting estimators of both F and G into the functional $t(.)$. A symmetric 95% bootstrap confidence interval for θ can then be obtained in much the same way as described for the one–sample case but using separate resamples $\boldsymbol{\mathcal{X}}^* = \{X_1^*, \ldots, X_{n_X}^*\}$ from $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}^* = \{Y_1^*, \ldots, Y_{n_Y}^*\}$ from $\boldsymbol{\mathcal{Y}}$. Also, once estimates for F and G have been obtained, Monte Carlo simulations can be used to generate combined resamples $\boldsymbol{\mathcal{Z}}_b^* = \{\boldsymbol{\mathcal{X}}_b^*, \boldsymbol{\mathcal{Y}}_b^*\}$ $(b = 1, \ldots, B)$ which can then be employed as described in Section 2.4.2 to approximate the confidence interval of interest.

2.4.5 Bootstrap Hypothesis Testing

In general, hypothesis testing and the construction of confidence intervals are intimately connected in that if $\mathcal{I}$ is a confidence interval for an unknown parameter θ with coverage probability α, then a $(1 - \alpha)$–level test of the null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ can be carried out by rejecting H_0 if $\theta_0 \notin \mathcal{I}$. This duality also holds in the bootstrap setting so that the statements made above in the context of bootstrap confidence intervals also apply when a bootstrap hypothesis test is to be constructed. In particular it is possible to improve the asymptotic coverage error of a bootstrap test by using an asymptotically pivotal test statistic which in fact is the first of the two guidelines Hall & Wilson (1991) suggest for the implementation of bootstrap hypothesis testing, the second guideline being that resampling should be carried out in a way that reflects H_0 even if the population fails to satisfy H_0. To see the rationale behind the second guideline, recall that hypothesis testing in general involves comparing the observed value of the test statistic with the distribution which would follow if the null hypothesis were true. To ensure a meaningful comparison in the bootstrap setting, the distribution of the resampled test statistic therefore needs to be a good approximation of the null distri-

bution which can only be achieved if the resampling is based on an empirical cumulative distribution function which satisfies H_0. If this is ignored, the results can be misleading, especially if H_0 is not met. The two guidelines by Hall & Wilson (1991) are therefore concerned with improving the coverage probability and the power, respectively.

If the hypothesis test is based on a confidence interval, it is an arbitrarily specified threshold level α which determines whether or not H_0 is rejected. More information about the data at hand can be obtained by investigating the actually attained significance level which can also be approximated using Monte Carlo resamples. For our example in Section 2.4.3, the estimated p–value has the form

$$\hat{p} = \frac{\#\{|\hat{\theta}_b^* - \hat{\theta}|/\hat{\sigma}_b^* > |\hat{\theta} - \theta_0|/\hat{\sigma}\} + 1}{B + 1}.$$

Note that we follow Davison & Hinkley (1997, p.161) in adding 1 to the numerator and denominator in the above formula for the estimated p–value.

2.4.6 Limitations of the Bootstrap

In general, a bootstrap procedure may be termed *consistent* if the distributions of $\hat{\theta} = t(\hat{F})$ and $\hat{\theta}^* = t(\hat{F}^*)$ agree in the limit (e.g. Bose & Politis, 1993). While this holds for a variety of situations where the data are independent, Singh (1981) points out the inadequacy of the *i.i.d.* bootstrap procedure described above in the context of dependent data. To give an example, he considers a sequence of m–dependent random variables $X_1, X_2, \ldots$ with $\mathrm{E}(X_1) = \mu$ and $\mathrm{E}(X_1^2) = \sigma^2 < \infty$, where *$m$–dependent* means that two subsequences of the form $\{X_1, \ldots, X_k\}$ and $\{X_{k+l+1}, \ldots, X_{2k+l}\}$ are independent for any $l \geq m$. Given a sample $\boldsymbol{\mathcal{X}} = \{X_1, \ldots, X_n\}$, the central limit theorem for m–dependent processes holds (cf. e.g. Lahiri, 2003, Appendix A) so that

$$\sqrt{n}\big(\bar{X} - \mu\big) \xrightarrow{\mathcal{D}} \mathrm{N}\big(0, \sigma^2 + \sum_{i=1}^{m-1} \mathrm{Cov}(X_1, X_{1+i})\big),$$

where $\bar{X} = 1/n \sum_{i=1}^{n} X_i$ denotes the sample mean. The bootstrap method described

above attempts to estimate the distribution of $\hat{\theta} = \sqrt{n}\big(\bar{X} - \mu\big)$ based on an *i.i.d* resample $\boldsymbol{\mathcal{X}}^* = \{X_1^*, \ldots, X_n^*\}$ from the original sample, and the bootstrap version of $\hat{\theta}$ is $\hat{\theta}^* = \sqrt{n}\big(\bar{X}^* - \bar{X}\big)$. Conditional on $\boldsymbol{\mathcal{X}}$, it holds under the *i.i.d.* bootstrap that

$$\sqrt{n}\big(\bar{X}^* - \bar{X}\big) \xrightarrow{\mathcal{D}} \mathrm{N}\big(0, \sigma^2\big),$$

so that the bootstrap approximation of the distribution of $\hat{\theta}$ is not consistent.

In the above example the reason for the inconsistency is that *i.i.d.* resampling from the given sample fails to account for the lag–covariance terms in the asymptotic variance. In other situations where the data are dependent, the bootstrap method as proposed by Efron (1979) will be inadequate for similar reasons because the total data "scrambling" (Politis, 2003) induced by the *i.i.d.* resampling loses all the dependence information. Since the paper by Singh (1981), there have been several attempts in the literature to extend Efron's (1979) *i.i.d.* bootstrap to the dependent case. The first attempts were model–based and focused on resampling of the approximately *i.i.d.* innovations (cf. e.g. Friedman, 1981, 1984). However, for situations where not enough prior knowledge is available to specify a parametric model, the breakthrough was achieved when resampling of single observations was replaced with block resampling where a block contains a number of consecutive observations. This idea was put forward by Hall (1985), Carlstein (1986), Künsch (1989), Politis & Romano (1992) and others in various forms. In this thesis we will use the circular block bootstrap by Politis & Romano (1992) which will be described in Section 6.2. For a detailed treatment of resampling methods for dependent data see the monograph by Lahiri (2003).

2.5 Reproducing Kernel Hilbert Spaces

A Reproducing Kernel Hilbert Space (RKHS) is a bijection which associates a positive definite kernel with a Hilbert space of functions. Despite having their origin in complex function theory, RKHSs have recently become widely used in more applied areas such as neural networks and machine learning. A well–known example for an application of

the RKHS theory is the Support Vector Machine method in pattern recognition where groups of data which cannot be separated by a linear function in their original space are transformed into a higher dimensional space, where a separating hyperplane can be found (cf. e.g. Vapnik, 1995; Schölkopf *et al.*, 1999). The book by Berlinet & Thomas-Agnan (2004) presents the theory of RKHSs together with examples of its use in probability and mathematical statistics.

As the name suggests, a RKHS is a Hilbert space. For a formal and comprehensive definition of Hilbert spaces see for example Promislow (2009, Chapter 4). The property of a general Hilbert space $\mathcal{H}$ which is of interest in this thesis is that it is endowed with an inner product $< ., . >_{\mathcal{H}}$ which in turn facilitates the definition of a norm $||.||_{\mathcal{H}}$. Within this framework, it also possible to define the angle θ_{xy} between two elements x and y via

$$< x, y >_{\mathcal{H}} = ||x||_{\mathcal{H}} \cdot ||y||_{\mathcal{H}} \cdot \cos\theta_{xy}. \tag{2.29}$$

Hilbert spaces are important because they provide powerful mathematical tools in many complicated settings. Due to their connection to Euclidean spaces (which are familiar to us from every day life and are in fact special cases of Hilbert spaces), they also provide geometric concepts on which our intuition can rest.

A RKHS is a Hilbert space of functions. A well–known example of another space which falls into that category is the space of Lebesgue square–integrable functions L_2 with inner product $< x(t), y(t) >_{L_2} = \int x(t)y(t)dt$. However, the inner product in L_2 can be very hard to evaluate. One benefit of RKHS theory is that it provides a way to restrict the space of functions in L_2 to those which allow to define a different, easier to calculate inner product. The tool for this restriction is a positive definite kernel. A symmetric function $K(.,.)$ on $\mathbb{R}^p \times \mathbb{R}^p$ is a positive definite kernel if for any L_2–function $f(.)$ (other than the zero function), it holds that

$$\int_{\mathbb{R}^p} \int_{\mathbb{R}^p} f(\boldsymbol{x}) K(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) \geq 0, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p, \tag{2.30}$$

see for example Christianini & Shawe-Taylor (2000, p.35). The above definition is a generalisation of the positive semi–definite definition for symmetric matrices which can

be recovered if $f(.)$ is chosen to be a weighted sum of delta functions on a set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with corresponding (scalar) weights $\{\alpha_1, \ldots, \alpha_n\}$. In that case, (2.30) reduces to

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0.$$

If this holds for all possible sets $\{\alpha_1, \ldots, \alpha_n\}$ of weights, then the $(n \times n)$ matrix $\boldsymbol{K}$ where $(\boldsymbol{K})_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semi–definite.

The Moore-Aronszajin theorem (Aronszajan, 1950) says that to any positive definite function $K(.,.)$ on $\mathbb{R}^p \times \mathbb{R}^p$ there corresponds a unique RKHS of real valued functions on $\mathbb{R}^p$ and *vice versa*. In practice, given a positive definite function $K(.,.)$, the corresponding RKHS $\mathcal{H}_K$ can be constructed as follows: let $K_{\boldsymbol{x}}(\boldsymbol{y}) = K(\boldsymbol{x}, \boldsymbol{y})$ denote the function of $\boldsymbol{y}$ obtained when $\boldsymbol{x}$ is fixed. The corresponding RKHS $\mathcal{H}_K$ then has the form

$$\mathcal{H}_K = \left\{ f \mid f(.) = \sum_{i=1}^{n} \alpha_i K_{\boldsymbol{x}_i}(.);\ n \in \mathbb{N}, \alpha_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^p \right\}. \tag{2.31}$$

Let $K_{\boldsymbol{x}_1}(.)$ and $K_{\boldsymbol{x}_2}(.)$ be two basic elements of $\mathcal{H}_K$. Their inner product in $\mathcal{H}_K$ is defined as $< K_{\boldsymbol{x}_1}, K_{\boldsymbol{x}_2} >_{\mathcal{H}_k} = K(\boldsymbol{x}_1, \boldsymbol{x}_2)$. This is the *reproducing property* of the kernel. Based on this property, the inner product of two general functions $f(.) = \sum_{i=1}^{n} \alpha_i K_{\boldsymbol{x}_i}(.)$ and $g(.) = \sum_{j=1}^{m} \beta_j K_{\boldsymbol{x}_j}(.)$ in $\mathcal{H}_K$ has the simple form

$$< f, g >_{\mathcal{H}_k} = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Note that functions $f(.) \in \mathcal{H}_K$ are also elements of the "bigger" space L_2. By restricting the space to functions of the form (2.31), however, the inner product can be evaluated without solving possibly high dimensional overlap integrals.

Chapter 3

Bayesian Alignment of Unlabelled Marked Point Sets

In many application areas, the objects of interest are given in form of marked point sets. In general, a marked point set can be described as a configuration of points in two– or three–dimensional Euclidean space where measurements (marks) are available at each point location. When the objective is to measure the similarity of two of these objects, this can be achieved by aligning the configurations as closely as possible while taking into account the associated marks. However, a frequent problem is that the given configurations are unlabelled in the sense that there is no natural correspondence between the points.

One area where the above data situation is frequently encountered is the structural alignment of molecules (cf. Section 1.1.1) where the configuration of each molecule is given by the set of xyz–coordinates of the atom locations, and the marks are additional measurements such as van der Waals radii or partial charge values measured at the atom positions. The configurations of different molecules are thereby unlabelled as in most cases, a one–to–one correspondence between the atoms of different molecules cannot be established.

The task of aligning unlabelled marked point sets has been of recent interest in statistical shape and image analysis. In Section 3.1, we provide a formal description of the problem and introduce some notation. To be able to point out the novelty in our methods, we

then briefly review the previously proposed statistical approaches in Section 3.2. In Sections 3.3 to 3.5, we describe our model for comparing unlabelled marked point sets. In order to validate our method, a simulation study is carried out in Section 3.6. Based on the results of this study, we are also able to formulate some prerequisites the data have to satisfy for the alignment to work well. A summary of the work presented in this chapter is provided in Section 3.7, and the next chapter describes an application of our methods to the steroid dataset and an extension for the alignment of multiple objects.

3.1 The Problem

A marked point set M can be represented as $M = \{z^{\mathrm{M}}(\boldsymbol{x}_1^{\mathrm{M}}), \ldots, z^{\mathrm{M}}(\boldsymbol{x}_{k_{\mathrm{M}}}^{\mathrm{M}})\}$, where k_{M} denotes the number of points in M, $\boldsymbol{x}_i^{\mathrm{M}} \in \mathbb{R}^m$ is the coordinate vector of the ith point in m dimensions, and $z^{\mathrm{M}}(\boldsymbol{x}_i^{\mathrm{M}})$ denotes the (scalar) mark observed at the ith point location. In this setting we wish to develop a measure of similarity between two objects A and B, say, which does not depend on their relative position, i.e. we wish to filter out rotations $\boldsymbol{\Gamma} \in SO(m)$ and translations $\boldsymbol{\gamma} \in \mathbb{R}^m$ between the corresponding configuration matrices $\boldsymbol{X}_{\mathrm{A}} \in \mathbb{R}^{k_{\mathrm{A}} \times m}$ and $\boldsymbol{X}_{\mathrm{B}} \in \mathbb{R}^{k_{\mathrm{B}} \times m}$, where $\boldsymbol{X}_{\mathrm{M}}$ row–wise contains the coordinate vectors $\boldsymbol{x}_i^{\mathrm{M}}$ $(M \in \{A, B\})$. As described in Section 2.1.1, the space $SO(m)$ contains the rotation (special orthogonal) matrices which satisfy $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{I}_m$ and $|\boldsymbol{\Gamma}| = 1$.

If the similarity between A and B in a certain relative position can be described by a similarity function of the general form

$$S_{\mathrm{AB}}\big(\boldsymbol{\Gamma}, \boldsymbol{\gamma}\big) = S\big(\{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}, \{z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_1^{\mathrm{B}} + \boldsymbol{\gamma}), \ldots, z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}} + \boldsymbol{\gamma})\}\big), \tag{3.1}$$

where a high value indicates a high similarity of the two point sets in the relative position defined by $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$, then a rotation/translation invariant similarity measure can be obtained by maximising (3.1) with respect to rotation and translation, i.e.

$$S(A, B) = \sup_{\substack{\boldsymbol{\Gamma} \in SO(m) \\ \boldsymbol{\gamma} \in \mathbb{R}^m}} S_{\mathrm{AB}}\big(\boldsymbol{\Gamma}, \boldsymbol{\gamma}\big). \tag{3.2}$$

To obtain $S(A, B)$, (3.1) has to be optimised with respect to rotation and translation, i.e. in $(m + m(m-1)/2)$–dimensional parameter space. This procedure bears a clear resemblance to the ordinary partial Procrustes analysis in statistical shape analysis where analytical methods are applied to superimpose two configuration matrices of the same dimension (cf. Section 2.1.3). However, when two objects of the form $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{x}_1^{\mathrm{B}}), \ldots, z^{\mathrm{B}}(\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}})\}$ are to be aligned, there usually are no clear one–to–one correspondences between the points. Moreover, not only the configuration matrices but also the observed values of the marks should be taken into account when aligning A and B. Maximising (3.2), therefore, can not in general be carried out analytically and will involve numerical methods.

There are three main statistical approaches to the problem of aligning two unlabelled point sets A and B, namely Green & Mardia (2006), Dryden *et al.* (2007) and Schmidler (2007) who formulate Bayesian models in which the required numerical calculations are carried out using Markov chain Monte Carlo (MCMC) methods (cf. Section 2.3.4). In all three cases, the alignment method proposed is primarily based on the configuration matrices $\boldsymbol{X}_{\mathrm{A}} \in \mathbb{R}^{k_{\mathrm{A}} \times m}$ and $\boldsymbol{X}_{\mathrm{B}} \in \mathbb{R}^{k_{\mathrm{B}} \times m}$ of the considered point sets, and the main tool for the alignment is a labelling matrix $\boldsymbol{\Lambda}$ with binary entries which determines which points of the two point sets correspond to each other. As these papers provide a starting point for the alignment methodology proposed in this thesis, the following section provides a brief summary of their main ideas.

3.2 Previous Point–Based Approaches

In order to match two unlabelled configuration matrices $\boldsymbol{X}_{\mathrm{A}} \in \mathbb{R}^{k_{\mathrm{A}} \times m}$ and $\boldsymbol{X}_{\mathrm{B}} \in \mathbb{R}^{k_{\mathrm{B}} \times m}$, Dryden *et al.* (2007) consider a $\big(k_{\mathrm{A}} \times (k_{\mathrm{B}}+1)\big)$–dimensional matrix with entries

$$(\boldsymbol{\Lambda})_{ij} = \begin{cases} I_{\{\text{point } \boldsymbol{x}_i^{\mathrm{A}} \text{ in } A \text{ matches point } \boldsymbol{x}_j^{\mathrm{B}} \text{ in } B\}} & , i = 1, \ldots, k_{\mathrm{A}};\ j = 1, \ldots, k_{\mathrm{B}} \\ I_{\{\text{point } \boldsymbol{x}_i^{\mathrm{A}} \text{ in } A \text{ does not match any point in } B\}} & , i = 1, \ldots, k_{\mathrm{A}};\ j = k_{\mathrm{B}} + 1 \end{cases}, \quad (3.3)$$

where $I_{\{\mathrm{E}\}}$ denotes the indicator function of an event E. The matrix $\boldsymbol{\Lambda}$ therefore defines

a correspondence between points in A and points in B. The actual labelling can be found in the first k_{B} columns. Moreover, as each row in $\boldsymbol{\Lambda}$ is constrained to sum to one, the number of zeros in the last column is equal to the number of points n_{M} in A which match points in B.

For a given $\boldsymbol{\Lambda}$, the configuration matrices of A and B can each be partitioned into two blocks containing the matching points and the non–matching points, respectively, i.e. $\boldsymbol{X}^{\mathrm{A}} = (\boldsymbol{X}^{\mathrm{A}}_{\mathrm{M}}, \boldsymbol{X}^{\mathrm{A}}_{\mathrm{N}})$ and $\boldsymbol{X}^{\mathrm{B}} = (\boldsymbol{X}^{\mathrm{B}}_{\mathrm{M}}, \boldsymbol{X}^{\mathrm{B}}_{\mathrm{N}})$. Both $\boldsymbol{X}^{\mathrm{A}}_{\mathrm{M}}$ and $\boldsymbol{X}^{\mathrm{B}}_{\mathrm{M}}$ are $(n_{\mathrm{M}} \times m)$–dimensional matrices which, without loss of generality, can be considered as ordered in a way that the ith row of $\boldsymbol{X}^{\mathrm{A}}_{\mathrm{M}}$ corresponds to the ith row of $\boldsymbol{X}^{\mathrm{B}}_{\mathrm{M}}$. The matching parts of both configurations therefore satisfy the data requirement needed for classical statistical shape analysis so that an appropriate rotation/translation invariant dissimilarity index is given by the ordinary partial Procrustes sum of squares $\mathrm{OSS}_p(\boldsymbol{X}^{\mathrm{A}}_{\mathrm{M}}, \boldsymbol{X}^{\mathrm{B}}_{\mathrm{M}})$ defined in (2.7).

Assuming that the point set A is random whereas B is a fixed reference point set, Dryden *et al.* (2007) formulate a likelihood for the configuration matrix $\boldsymbol{X}^{\mathrm{A}}$ as

$$L(\boldsymbol{X}^{\mathrm{A}}|\boldsymbol{\Lambda}, \tau, \boldsymbol{X}^{\mathrm{B}}) \propto |\mathcal{A}|^{n_{\mathrm{M}}-k_{\mathrm{A}}}(2\pi)^{-Q/2}\tau^{Q/2}\exp\big\{-\tfrac{\tau}{2}\,\mathrm{OSS}_p(\boldsymbol{X}^{\mathrm{A}}_{\mathrm{M}}, \boldsymbol{X}^{\mathrm{B}}_{\mathrm{M}})\big\}, \tag{3.4}$$

where $Q = mn_{\mathrm{M}} - m(m-1)/2$ and τ is a precision parameter. Moreover, $\mathcal{A}$ denotes a large bounded region in $\mathbb{R}^m$ with volume $|\mathcal{A}|$. In essence, (3.4) therefore defines an independent Gaussian/uniform mixture model for the configuration matrix $\boldsymbol{X}^{\mathrm{A}}$. The Gaussian part thereby implies that the likelihood for the points in A which match points in B increases as the rotation/translation invariant dissimilarity $\mathrm{OSS}_p(\boldsymbol{X}^{\mathrm{A}}_{\mathrm{M}}, \boldsymbol{X}^{\mathrm{B}}_{\mathrm{M}})$ decreases, whereas the uniform part arises from the assumption that the non–matching points in A do not have any preferred region on the domain $\mathcal{A}$.

A likelihood similar to (3.4) has been obtained by both Schmidler (2007) and Green & Mardia (2006). Like Dryden *et al.* (2007), Schmidler (2007) formulates the perturbations of the matching points of A and B in terms of their ordinary partial Procrustes sum of squares. However, Green & Mardia (2006) use a different starting point and consider both configurations $\boldsymbol{X}^{\mathrm{A}}$ and $\boldsymbol{X}^{\mathrm{B}}$ as noisy observations of a set of hidden reference points $\{\boldsymbol{\mu}_i\}$, where $\boldsymbol{\mu}_i \in \mathbb{R}^m$ denotes the coordinate vector of the ith hidden point. With this

assumption, the two given configuration matrices can be modelled as

$$\boldsymbol{x}_i^{\mathrm{A}} = \boldsymbol{\mu}_{\xi_i} + \epsilon_{\mathrm{A}i} \quad \text{and} \quad \boldsymbol{x}_j^{\mathrm{B}} = (\boldsymbol{\Gamma}\boldsymbol{\mu}_{\eta_j} + \gamma) + \epsilon_{\mathrm{B}j}, \qquad i = 1, \ldots, k_{\mathrm{A}};\, j = 1, \ldots, k_{\mathrm{B}}, \quad (3.5)$$

where $\boldsymbol{\Gamma} \in SO(m)$ is a rotation matrix and $\gamma \in \mathbb{R}^m$ denotes a translation vector. Moreover, $\{\epsilon_{\mathrm{A}i}\}$ and $\{\epsilon_{\mathrm{B}j}\}$ are sets of error terms with probability density functions f_{A} and f_{B}, respectively, and $\{\xi_i\}$ and $\{\eta_j\}$ denote sets of indexing arrays which define the mappings from the hidden point set to the observed coordinate vectors in $\boldsymbol{X}^{\mathrm{A}}$ and $\boldsymbol{X}^{\mathrm{B}}$. The mappings can be summarised in a labelling matrix of the form (3.3), but as multiple matches are excluded in this model the last column of $\boldsymbol{\Lambda}$ is omitted here, and both the rows and the columns are restricted to have a sum of either zero or one. The number of matches is therefore defined as $n_{\mathrm{M}} = \sum_{i=1}^{k_{\mathrm{A}}} \sum_{j=1}^{k_{\mathrm{B}}} (\boldsymbol{\Lambda})_{ij} \in \big\{0, \ldots, \min\{k_{\mathrm{A}}, k_{\mathrm{B}}\}\big\}$.

Assuming that the error terms are independent of each other and independent of the hidden reference points $\{\boldsymbol{\mu}_i\}$, the latter can be integrated out and it can be shown that the likelihood of the configuration matrices has the form

$$L(\boldsymbol{X}^{\mathrm{A}}, \boldsymbol{X}^{\mathrm{B}} | \, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\gamma}) = |\mathcal{A}|^{n_{\mathrm{M}} - (k_{\mathrm{A}} + k_{\mathrm{B}})} \prod_{i,j:(\boldsymbol{\Lambda})_{ij}=1} g(\boldsymbol{x}_i^{\mathrm{A}} - \boldsymbol{\Gamma}\boldsymbol{x}_j^{\mathrm{B}} - \boldsymbol{\gamma}), \quad (3.6)$$

where the function $g(\boldsymbol{z}) = \int f_{\mathrm{A}}(\boldsymbol{z} + \boldsymbol{u}) f_{\mathrm{B}}(\boldsymbol{u}) d\boldsymbol{u}$ denotes the density of $\epsilon_{\mathrm{A}i} - \epsilon_{\mathrm{B}j}$. If the error densities f_{A} and f_{B} can be assumed to be normal densities, then (3.6) reduces to a Gaussian/uniform mixture similar to (3.4). However, (3.6) is symmetric in that the configurations of both point sets A are B are considered as random and the perturbations between the matching points are formulated in configuration space directly.

The above shows that previous statistical approaches to the problem of aligning unlabelled point sets are based on a labelling matrix $\boldsymbol{\Lambda}$ which imposes a correspondence on the unlabelled configurations. As $\boldsymbol{\Lambda}$ is a likelihood parameter, it can be inferred about using MCMC simulations and posterior inference. In Dryden *et al.* (2007) and Schmidler (2007), the employed posterior estimate of $\boldsymbol{\Lambda}$ automatically determines the matching parameters as $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$ are optimised out within the Procrustes framework. Rather than being optimised out, $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$ are integrated out in Green & Mardia (2006) using simultaneous Bayesian inference about the transformation and the matching.

The above methods are designed to align unlabelled point sets. If additional information is provided at each point location and the objects of interest are unlabelled marked point sets of the form $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{x}_1^{\mathrm{B}}), \ldots, z^{\mathrm{B}}(\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}})\}$, Green & Mardia (2006) and Dryden *et al.* (2007) provide a way to extend their models by adding extra terms to the likelihood which increase the probability of matching points with similar marks.

One potential problem with the above methods is the need for imposing a correspondence structure on the points of the two objects. In particular in the molecular context where the points are associated with atom locations, such a correspondence might not exist in every application so that a method which does not rely on point correspondences seems to be preferable. Moreover, it would be desirable to be able to incorporate the marks in a more direct and natural way. With these goals in mind and the above methods as starting point, we develop an alignment methodology for unlabelled marked point sets which moves away from the point–based representation of the point sets and defines a similarity measure of the form (3.1) based on a continuous notion of their "shapes".

In essence, the novel idea developed in this thesis is to counterbalance the absence of point correspondences by assuming that the marks in both objects follow the same spatial distribution. With this assumption, a continuous version of (3.5) can be used where $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{x}_1^{\mathrm{B}}), \ldots, z^{\mathrm{B}}(\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}})\}$ are regarded as two (rotated and translated) noisy point samples of a common underlying hidden reference field $Z(\boldsymbol{x})$. In order to obtain a continuous representation of each point set which resembles $Z(\boldsymbol{x})$, methods from spatial statistics can then be used to predict the unobserved reference field and the alignment can be based on the predicted fields $\hat{Z}_{\mathrm{A}}(\boldsymbol{x})$ and $\hat{Z}_{\mathrm{B}}(\boldsymbol{x})$.

3.3 A Continuous Representation of Marked Point Sets

Consider a marked point set $M = \{z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_k)\}$, where the index M for the coordinates and marks is omitted for clarity in this section. In order to reconstruct the underlying reference field, the marks of M are interpolated into $\mathbb{R}^m$ using spatial pre-

diction. As described in Section 2.2.1, in the context of spatial statistics, the vector $\boldsymbol{z} = \big(z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_k)\big)^T$ is viewed as a sample of one realisation $z(\boldsymbol{x})$ of a random field $\big\{Z(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^m\big\}$ which in the following is assumed to be second–order stationary with a constant mean μ and a (known) covariance function $\sigma(\boldsymbol{h}) = \mathrm{Cov}\big(Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})\big)$ (cf. Section 2.2.1.1). As our main objective is the comparison of two or more point sets (which are assumed to stem from the same underlying reference field), the actual value of μ is not of interest in our application and can be set to zero without loss of generality.

With the above assumptions, simple kriging is appropriate for predicting the value of the random field at a location of interest $\boldsymbol{x}_0$. Given μ is set to zero, for a general location $\boldsymbol{x}$ this yields the predicted field

$$\hat{Z}(\boldsymbol{x}) = \boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}(\boldsymbol{x}) = \sum_{i=1}^{k} w_i \sigma(\boldsymbol{x}_i - \boldsymbol{x}) \tag{3.7}$$

where the vector of weights $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}\boldsymbol{z}$ is optimal in terms of the prediction mean squared error (PMSE) if the stationarity assumption is met; cf. Section 2.2.1.3. In this section, the subscript "BLUP" is omitted for clarity and we use $\hat{Z}(\boldsymbol{x})$ to refer to the predicted field obtained using simple kriging.

Given a marked point set $M = \{z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_k)\}$, the predicted field $\hat{Z}(\boldsymbol{x})$ combines the information about the geometry of the associated configuration matrix $\boldsymbol{X}_{\mathrm{M}}$ and the values of the associated marks. Moreover, with the assumption that M is a noisy pointwise observation of an underlying stationary hidden reference field $Z(\boldsymbol{x})$, $\hat{Z}(\boldsymbol{x})$ is the optimal representation of $Z(\boldsymbol{x})$ which can be obtained based on the given data. In the following, we will treat the predicted field as a continuous representation of M.

An important point in formula (3.7) is that the predicted field at a general location $\boldsymbol{x} \in \mathbb{R}^m$ can be expressed as a linear combination of versions of the covariance function $\sigma(.)$ which are centred at the point locations $\boldsymbol{x}_i$ of the considered marked point set M where (if the assumptions are correct) the weights are optimal with respect to squared–error loss. We now show why this representation will be very useful in the subsequent considerations.

In Section 2.2.1.1 we mentioned that not every function can be considered as a covariance function of a stationary random field. In particular, for a function $\sigma(.)$ to be a valid covariance function it must have the property that for any set of point locations $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$, the resulting covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma(\mathbf{0}) & \sigma(\boldsymbol{x}_1 - \boldsymbol{x}_2) & \ldots & \sigma(\boldsymbol{x}_1 - \boldsymbol{x}_{n-1}) & \sigma(\boldsymbol{x}_1 - \boldsymbol{x}_n) \\ \sigma(\boldsymbol{x}_2 - \boldsymbol{x}_1) & \sigma(\mathbf{0}) & \ldots & \sigma(\boldsymbol{x}_2 - \boldsymbol{x}_{n-1}) & \sigma(\boldsymbol{x}_2 - \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma(\boldsymbol{x}_{n-1} - \boldsymbol{x}_1) & \sigma(\boldsymbol{x}_{n-1} - \boldsymbol{x}_2) & \ldots & \sigma(\mathbf{0}) & \sigma(\boldsymbol{x}_{n-1} - \boldsymbol{x}_n) \\ \sigma(\boldsymbol{x}_n - \boldsymbol{x}_1) & \sigma(\boldsymbol{x}_n - \boldsymbol{x}_2) & \ldots & \sigma(\boldsymbol{x}_n - \boldsymbol{x}_{n-1}) & \sigma(\mathbf{0}) \end{pmatrix}$$

is positive semi–definite. Hence, if $\sigma\big(\boldsymbol{h}\big) = \mathrm{Cov}\big(Z(\boldsymbol{x}), Z(\boldsymbol{x}+\boldsymbol{h})\big)$ is considered as the value of a kernel function $\sigma_{\mathrm{K}}(.,.)$ on $\mathbb{R}^m \times \mathbb{R}^m$ evaluated at $(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h})$, then $\sigma_{\mathrm{K}}(.,.)$ satisfies the requirements for being a positive definite kernel (cf. Section 2.5). By virtue of the Moore-Aronszajin theorem (cf. Section 2.5) it therefore holds that for every valid covariance function $\sigma(.)$ there exists a unique reproducing kernel Hilbert space (RKHS) of functions which has the form

$$\mathcal{H}_\sigma = \left\{ f \,|\, f(.) = \sum_{i=1}^{n} \alpha_i \, \sigma_{\mathrm{K}}(., \boldsymbol{x}_i) \right\}. \tag{3.8}$$

Moreover, for two functions $f(.) = \sum_{i=1}^{n} \alpha_i \, \sigma_{\mathrm{K}}(., \boldsymbol{x}_i) \in \mathcal{H}_\sigma$ and $g(.) = \sum_{j=1}^{n'} \beta_j \, \sigma_{\mathrm{K}}(., \boldsymbol{x}_j) \in \mathcal{H}_\sigma$ the inner product is defined in terms of the covariance function as

$$< f, g >_{\mathcal{H}_\sigma} = \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j \, \sigma_{\mathrm{K}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j \, \sigma(\boldsymbol{x}_i - \boldsymbol{x}_j),$$

and the norm of a function $f(.) \in \mathcal{H}_\sigma$ has the form

$$||f||_{\mathcal{H}_\sigma} = < f, f >_{\mathcal{H}_\sigma}^{1/2} = \Big\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, \sigma_{\mathrm{K}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \Big\}^{1/2} = \Big\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, \sigma(\boldsymbol{x}_i - \boldsymbol{x}_j) \Big\}^{1/2}.$$

Note that our continuous representation (3.7) of a marked point set $M = \{z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_k)\}$ is a member of $\mathcal{H}_\sigma$, where the weights are chosen to optimally represent the hidden reference field $Z(\boldsymbol{x})$. This observation can be directly utilised for the alignment of two marked point sets.

3.4 Pairwise Similarity of Unlabelled Marked Point Sets

Now consider two point sets $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{x}_1^{\mathrm{B}}), \ldots, z^{\mathrm{B}}(\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}})\}$. To align the two point sets, we are interested in a similarity function of the form (3.1) which determines the similarity of A and B in a certain relative position. To define a suitable similarity function, we consider B as moveable, i.e.

$$B = \{z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_1^{\mathrm{B}} + \boldsymbol{\gamma}), \ldots, z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}} + \boldsymbol{\gamma})\},$$

where $\boldsymbol{\Gamma} \in SO(m)$ denotes a rotation matrix and $\boldsymbol{\gamma} \in \mathbb{R}^m$ denotes a translation vector. We can now use (3.7) to obtain the corresponding predicted fields

$$\hat{Z}_{\mathrm{A}}(\boldsymbol{x}) = \sum_{i=1}^{k_{\mathrm{A}}} w_i^{\mathrm{A}} \sigma\big(\boldsymbol{x}_i^{\mathrm{A}} - \boldsymbol{x}\big) \quad \text{and} \quad \hat{Z}_{\mathrm{B}}(\boldsymbol{x}) = \sum_{i=1}^{k_{\mathrm{B}}} w_i^{\mathrm{B}} \sigma\big((\boldsymbol{\Gamma}\boldsymbol{x}_i^{\mathrm{B}} + \boldsymbol{\gamma}) - \boldsymbol{x}\big)$$

which serve as continuous representations of A and B.

As both fields $\hat{Z}_{\mathrm{A}}(\boldsymbol{x})$ and $\hat{Z}_{\mathrm{B}}(\boldsymbol{x})$ are members of $\mathcal{H}_\sigma$, we can define a similarity function of the form (3.1) in terms of the inner product $< ., . >_{\mathcal{H}_\sigma}$ as

$$\begin{aligned} C_{\mathrm{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) &= \frac{< \hat{Z}_{\mathrm{A}}(\boldsymbol{x}), \hat{Z}_{\mathrm{B}}(\boldsymbol{x}) >_{\mathcal{H}_\sigma}}{||\hat{Z}_{\mathrm{A}}(\boldsymbol{x})||_{\mathcal{H}_\sigma} ||\hat{Z}_{\mathrm{B}}(\boldsymbol{x})||_{\mathcal{H}_\sigma}} \\ &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} w_i^{\mathrm{A}} w_j^{\mathrm{B}} \, \sigma\big(\boldsymbol{x}_i - (\boldsymbol{\Gamma}\boldsymbol{x}_j + \boldsymbol{\gamma})\big)}{||\hat{Z}_{\mathrm{A}}(\boldsymbol{x})||_{\mathcal{H}_\sigma} ||\hat{Z}_{\mathrm{B}}(\boldsymbol{x})||_{\mathcal{H}_\sigma}}. \end{aligned} \tag{3.9}$$

The above function is a variant of Pearson's correlation coefficient for continuous data. The numerator term measures the "overlap" of the fields (in a certain relative position) whereas the denominator is a rotation/translation invariant normalising constant which ensures that $C_{\mathrm{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) \in [-1, 1]$. Note that the variance parameter σ^2 of the applied covariance function cancels out. Also note that (3.9) can be interpreted as the cosine of the angle between the two predicted fields in a certain relative position; cf. (2.29).

We shall call the above similarity function the *Kernel Carbo function* as it is a modification of a similarity function proposed by Carbo *et al.* (1980) in the context of field–based

molecular alignment (cf. Section 4.1). The fields considered in the original paper are the electron densities of the two molecules under study, and the similarity is defined in terms of $< .,. >_{L_2}$, i.e. the inner product in the space of Lebesgue square–integrable functions L_2. As both fields in our setting are members of the RKHS $\mathcal{H}_\sigma$, the Carbo similarity function can be *kernelised* by replacing $< .,. >_{L_2}$ with $< .,. >_{\mathcal{H}_\sigma}$ which has the advantage that (3.9) does not require evaluating overlap integrals over $\mathbb{R}^m$.

Optimising (3.9) with respect to rotation and translation yields the *Kernel Carbo Index*

$$C(A,B) = \sup_{\substack{\boldsymbol{\Gamma}\in SO(m)\\ \boldsymbol{\gamma}\in\mathbb{R}^m}} C_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma}), \tag{3.10}$$

which is invariant under rigid–body transformations of A and B. In situations where a discrepancy rather than a similarity measure is required, (3.9) can be uniquely mapped into the appropriate codomain using

$$D_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma}) = \frac{1 - C_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma})}{1 + C_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma})} \in [0,\infty), \tag{3.11}$$

and applying the same transformation to (3.10) yields a rotation/translation invariant distance between two marked point sets.

Note that the denominator of (3.9) is invariant under rotation and translation of the point set B. The discrepancy measure (3.11) is therefore intimately linked to an alternative discrepancy measure defined as

$$\begin{aligned} \tilde{D}_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma}) &= ||\hat{Z}_{\mathrm{A}}(\boldsymbol{x}) - \hat{Z}_{\mathrm{B}}(\boldsymbol{x})||^2_{\mathcal{H}_\sigma} \\ &= ||\hat{Z}_{\mathrm{A}}(\boldsymbol{x})||^2_{\mathcal{H}_\sigma} + ||\hat{Z}_{\mathrm{B}}(\boldsymbol{x})||^2_{\mathcal{H}_\sigma} - 2 < \hat{Z}_{\mathrm{A}}(\boldsymbol{x}), \hat{Z}_{\mathrm{B}}(\boldsymbol{x}) >_{\mathcal{H}_\sigma} . \end{aligned}$$

Hence, the rigid–body parameters which minimise $D_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma})$ can also be obtained using $\tilde{D}_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma})$. However, throughout this thesis we will use $D_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma})$ as Carbo–based discrepancy measure. Also note that the actual kriging does not need to be performed to evaluate (3.9) or (3.11) because the fields $\hat{Z}_{\mathrm{A}}(\boldsymbol{x})$ and $\hat{Z}_{\mathrm{B}}(\boldsymbol{x})$ are not compared at individual locations $\boldsymbol{x}_0$. Instead, the Carbo–based indices provide global similarity measures of the point sets A and B which compare the associated fields in their totality.

3.5 MCMC for Aligning Unlabelled Marked Point Sets

In Section 3.3, we have developed a continuous representation of a marked point set which provides a natural way to incorporate both the geometry of its point configuration and the associated marks. If the dataset at hand contains two marked point sets $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_1^{\mathrm{B}} + \boldsymbol{\gamma}), \ldots, z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}} + \boldsymbol{\gamma})\}$ which are recorded in an arbitrary position, then the Kernel Carbo index (3.9) is a suitable objective function which measures the similarity of the point sets in a given relative position so that

$$(\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\gamma}}) = \arg \max_{\substack{\boldsymbol{\Gamma} \in SO(m) \\ \boldsymbol{\gamma} \in \mathbb{R}^m}} C_{\mathrm{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \arg \min_{\substack{\boldsymbol{\Gamma} \in SO(m) \\ \boldsymbol{\gamma} \in \mathbb{R}^m}} D_{\mathrm{AB}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma})$$

should provide the optimal alignment parameters. This optimisation, however, is difficult in practice. We therefore develop a Bayesian model for the alignment of two marked point sets. A rotation/translation invariant similarity index can then be obtained by inserting posterior point estimates of the rotation and translation parameters into (3.9).

Within the Bayesian framework, it also is possible to introduce extra parameters which can improve the alignment. In this thesis, we consider introducing a mask vector for each point set to allow for the possibility that they match only in parts whereas other parts may have been generated by different underlying reference fields or may be largely affected by noise.

3.5.1 The Likelihood

Let $\boldsymbol{\lambda}_{\mathrm{A}} \in \mathbb{R}^{k_{\mathrm{A}}}$ and $\boldsymbol{\lambda}_{\mathrm{B}} \in \mathbb{R}^{k_{\mathrm{B}}}$ denote the mask vectors. Each of their entries is defined to be an indicator function, i.e. $\lambda_i^{\mathrm{M}} \in \{0, 1\}$ which determines if the ith point of set M ($M \in \{A, B\}$) is considered to contribute to the matching parts ($\lambda_i^{\mathrm{M}} = 1$) or not ($\lambda_i^{\mathrm{M}} = 0$). Taking the mask vector into account, the predicted version of the common reference field based on M then has the form $\hat{Z}_{\mathrm{M}}(\boldsymbol{x}; \boldsymbol{\lambda}_{\mathrm{M}}) = \sum_{i:\lambda_i^{\mathrm{M}}=1} w_i^{\mathrm{M}}(\boldsymbol{\lambda}_{\mathrm{M}}) \sigma(\boldsymbol{x}_i^{\mathrm{M}} - \boldsymbol{x})$, and the resulting partial Kernel Carbo function for two masked fields $\hat{Z}_{\mathrm{A}}(\boldsymbol{x}; \boldsymbol{\lambda}_{\mathrm{A}})$ and

$\hat{Z}_{\mathrm{B}}(\boldsymbol{x};\boldsymbol{\lambda}_{\mathrm{B}})$ in a certain relative position becomes

$$C_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}}) = \sum_{i:\lambda_i^{\mathrm{A}}=1}\sum_{j:\lambda_j^{\mathrm{B}}=1} \tilde{w}_i^{\mathrm{A}}(\boldsymbol{\lambda}_{\mathrm{A}})\tilde{w}_j^{\mathrm{B}}(\boldsymbol{\lambda}_{\mathrm{B}})\,\sigma\big(\boldsymbol{x}_i^{\mathrm{A}} - (\boldsymbol{\Gamma}\boldsymbol{x}_j^{\mathrm{B}} + \boldsymbol{\gamma})\big), \tag{3.12}$$

where the tilde indicates that the kriging weights are normalised by the corresponding term in the normalising constant of the Carbo index, i.e. $\tilde{w}_i^{\mathrm{M}}(\boldsymbol{\lambda}_{\mathrm{M}}) = w_i^{\mathrm{M}}(\boldsymbol{\lambda}_{\mathrm{M}})/N_{\mathrm{M}}(\boldsymbol{\lambda}_{\mathrm{M}})$, where $N_{\mathrm{M}}(\boldsymbol{\lambda}_{\mathrm{M}}) = ||\hat{Z}_{\mathrm{M}}(\boldsymbol{x};\boldsymbol{\lambda}_{\mathrm{M}})||_{\mathcal{H}_\sigma}$.

With the assumption that the matching parts of the two point sets are noisy pointwise observations of the same underlying reference field, we define the likelihood of the two marked point sets $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}),\ldots,z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_1^{\mathrm{B}}+\boldsymbol{\gamma}),\ldots,z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}}+\boldsymbol{\gamma})\}$ in the relative position defined by $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$ as

$$L\big(A,B\,|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}},\tau\big) \propto \tau\exp\big\{-\tau\, D_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}})\big\},$$

where $\boldsymbol{\theta}$ denotes the vector of the Euler angles which specifies a rotation matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ and $\boldsymbol{\gamma}$ denotes a displacement vector between A and B. Further, $\tau \in \mathbb{R}^+$ is a precision parameter which determines the mean and variance of the model. The mask vectors in the above likelihood play a similar role as the labelling matrices in Green & Mardia (2006) and Dryden *et al.* (2007) except in our framework, there is no need to establish one–to–one or many–to–one correspondences between points in A and B. This becomes particularly clear from (3.12) as all pairs of matching points $\boldsymbol{x}_i^{\mathrm{A}}$ and $\boldsymbol{x}_j^{\mathrm{B}}$ are compared. The mask vectors are therefore defined separately for each point set.

The above likelihood is chosen in this thesis because it performed well in pilot simulations. Other possible choices include the half–normal likelihood

$$L\big(A,B\,|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}},\tau\big) \propto \tau^{1/2}\exp\big\{-\tau\, D^2_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}})\big\},$$

which is less accommodating of outliers and might be preferable in some situations. In both cases, the rigid–body parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are parameters in the likelihood so that our Bayesian framework is similar to that by Green & Mardia (2006) in that they will be integrated out (rather than optimised out as in Dryden *et al.*, 2007).

3.5.2 Prior Distributions

In order to set up a Bayesian framework, prior distributions for the unknown parameters $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}_{\mathrm{A}}$, $\boldsymbol{\lambda}_{\mathrm{B}}$ and τ need to be specified. As we do not have any prior information about the rigid–body parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, we choose uninformative priors for these parameters and treat them as *a priori* uniformly distributed on $SO(m)$ and on a large bounded region in $\mathbb{R}^m$, respectively.

For the translation vector, we therefore have $\pi(\boldsymbol{\gamma}) \propto 1$. The uniform density $f_U(\boldsymbol{\theta})$ on $SO(m)$ is more complicated and depends on the dimension m. In the two–dimensional case, $f_U(\boldsymbol{\theta}) \propto 1$. For $m = 3$, the appropriate measure depends on the parameterisation of $SO(3)$. In this thesis, we use the Euler angles in the so–called x–convention (e.g. Goldstein *et al.*, 2002, pp.150), where $\boldsymbol{\Gamma}$ is decomposed in the following elementary rotation matrices

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \begin{pmatrix} \cos\theta_3 & \sin\theta_3 & 0 \\ -\sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_2 & \sin\theta_2 \\ 0 & -\sin\theta_2 & \cos\theta_2 \end{pmatrix} \begin{pmatrix} \cos\theta_1 & \sin\theta_1 & 0 \\ -\sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

With the domains $-\pi \leq \theta_1, \theta_3 < \pi$ and $-\pi/2 \leq \theta_2 < \pi/2$, every $\boldsymbol{\Gamma} \in SO(3)$ is uniquely determined apart from a singularity at $\theta_2 = -\pi/2$ (e.g. Naimark, 1964, p.6), and the invariant probability measure is given by $d\boldsymbol{\Gamma} = (8\pi^2)^{-1}\cos(\theta_2)d\theta_1 d\theta_2 d\theta_3$ (e.g. Miles, 1965) so that $f_U(\boldsymbol{\theta}) \propto \cos(\theta_2)$ for $SO(3)$.

Let $\Lambda_{k_{\mathrm{M}}}$ denote the space of all k_{M}–vectors with entries of either zero or one. To prevent the situation where only very few points are used in the field comparison, we introduce a (fixed) penalty parameter $\zeta > 1$ and define the joint prior density of the mask vectors as

$$\pi(\boldsymbol{\lambda}_{\mathrm{A}}, \boldsymbol{\lambda}_{\mathrm{B}}|\zeta) \propto \zeta^{\sum_i \lambda_i^{\mathrm{A}} + \sum_i \lambda_i^{\mathrm{B}}}, \quad (\boldsymbol{\lambda}_{\mathrm{A}}^T, \boldsymbol{\lambda}_{\mathrm{B}}^T) \in \Lambda_{k_{\mathrm{A}}} \times \Lambda_{k_{\mathrm{B}}}.$$

The penalty parameter therefore inherently comprises prior assumptions about the extent of the matching parts of A and B. Moreover, we choose a Gamma prior for the precision

parameter, i.e.

$$\pi(\tau|\alpha,\beta) \propto \tau^{\alpha-1}\exp(-\beta\tau), \quad \tau \geq 0,$$

where $\alpha > 0$ is a shape parameter and $\beta > 0$ is a scale parameter. This choice of prior distribution is generic in that it is conjugate to the likelihood (cf. Section 3.5.2). With the further assumptions that all unknown parameters are independent *a priori*, their joint posterior density is

$$\begin{aligned}\pi\big(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}},\tau|A,B,\alpha,\beta,\zeta\big)& \\ \propto \tau^{\alpha}\exp\big\{-\tau\left(D_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}})+\beta\right)\big\}\cdot\zeta^{\sum_i\lambda_i^{\mathrm{A}}+\sum_i\lambda_i^{\mathrm{B}}}\cdot f_U(\boldsymbol{\theta}),& \quad (3.13)\end{aligned}$$

Note that this can be regarded as a mixture model over $\Lambda_{k_{\mathrm{A}}} \times \Lambda_{k_{\mathrm{B}}}$.

3.5.3 Posterior Sampling

We use MCMC to sample from the posterior distribution (cf. Section 2.3.4). The resulting point estimates for the rigid–body parameters and the mask vectors can then be substituted into $D_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}})$ to yield a rotation/translation invariant point estimate of the distance

$$\hat{D}(A,B) = D_{\mathrm{AB}}(\hat{\boldsymbol{\Gamma}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\lambda}}_{\mathrm{A}},\hat{\boldsymbol{\lambda}}_{\mathrm{B}}). \quad (3.14)$$

Within the MCMC scheme, τ is updated with a Gibbs step, i.e. we use samples from the full–conditional distribution

$$\pi(\tau|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}},A,B) \sim \Gamma\big(\alpha+1, D_{\mathrm{AB}}(\boldsymbol{\Gamma},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}})+\beta\big) \quad (3.15)$$

to propose updates of τ which are then accepted at every iteration. Updated versions of the other parameters are obtained in four blocks, each using a Metropolis–Hastings step. For the rigid–body parameters, we use random walk proposals with normally distributed noise and standard deviations η_1 and η_2 for the Euler angles and the translation parameters, respectively.

A proposal distribution for the masks vectors $\boldsymbol{\lambda}_{\mathrm{A}}$ and $\boldsymbol{\lambda}_{\mathrm{B}}$ can be obtained by choosing an entry at random and then switching its value from zero to one or *vice versa*. The algorithm we use ensures that the defined Markov chain is irreducible and aperiodic so that it will eventually converge to the posterior distribution (3.13).

Due to the symmetry of the proposal distributions, convergence to and sampling from the limiting distribution in practice results in an approximate stochastic minimisation of the discrepancy term, and this behaviour can be emphasised by choosing a prior distribution with a large mean for τ. To see this, note that (ignoring the cosine term in the posterior distribution for $m = 3$), the Hastings ratio (HR) defined in (2.25) for the considered Metropolis–Hastings steps can take the following forms

$$\mathrm{HR} = \begin{cases} \exp\{(D_{\mathrm{AB}} - D^*_{\mathrm{AB}})\}^{\tau}, & \text{for an update of } \gamma \text{ or } \theta, \\ \zeta \cdot \exp\{(D_{\mathrm{AB}} - D^*_{\mathrm{AB}})\}^{\tau}, & \text{if a point } \boldsymbol{x}_i^{\mathrm{A}} \text{ or } \boldsymbol{x}_i^{\mathrm{B}} \text{ is included}, \\ 1/\zeta \cdot \exp\{(D_{\mathrm{AB}} - D^*_{\mathrm{AB}})\}^{\tau}, & \text{if a point } \boldsymbol{x}_i^{\mathrm{A}} \text{ or } \boldsymbol{x}_i^{\mathrm{B}} \text{ is deleted}, \end{cases} \tag{3.16}$$

where D^*_{AB} denotes the distance which results from the new proposed set of parameter values, and D_{AB} denotes the distance at the previous step. Moreover, deleting or adding a point from the matching parts of A and B is associated with updating the corresponding entry in $\boldsymbol{\lambda}_{\mathrm{A}}$ or $\boldsymbol{\lambda}_{\mathrm{B}}$. As described in Section 2.3.4, a proposed set of parameters will be accepted with probability $\alpha_{\mathrm{HR}} = \min\{\mathrm{HR}, 1\}$. From (3.16) it follows that

$$\mathrm{HR} > 1 \Leftrightarrow \begin{cases} D^*_{\mathrm{AB}} < D_{\mathrm{AB}}, & \text{for an update of } \gamma \text{ or } \theta \\ D^*_{\mathrm{AB}} < D_{\mathrm{AB}} + \frac{\log \zeta}{\tau}, & \text{if a point } \boldsymbol{x}_i^{\mathrm{A}} \text{ or } \boldsymbol{x}_i^{\mathrm{B}} \text{ is included} \\ D^*_{\mathrm{AB}} < D_{\mathrm{AB}} - \frac{\log \zeta}{\tau}, & \text{if a point } \boldsymbol{x}_i^{\mathrm{A}} \text{ or } \boldsymbol{x}_i^{\mathrm{B}} \text{ is deleted}. \end{cases} \tag{3.17}$$

Updates of the rigid–body parameters are therefore always accepted if they decrease the discrepancy term. When updating the mask vectors, however, the penalty parameter comes into play. It can be seen from (3.17) that $\zeta > 0$ encourages the inclusion of points in the matching parts of A and B as an increase of the discrepancy term up to $(\log \zeta)/\tau$ is tolerated if a point is included whereas an exclusion must decrease the discrepancy by at least $(\log \zeta)/\tau$. As expected, the larger the value of the penalty parameter, the more points will therefore be included in the matching parts of the point sets.

The precision parameter τ does not only influence the updating procedure of the mask vectors. From both (3.16) and (3.17) it follows that large values of τ in general encourage updates which yield small values of the discrepancy term. In particular, if $\mathrm{HR} < 1$ and updates are not automatically accepted, then the acceptance probability α_{HR} decreases with τ in all three cases. From that it is possible to predict how the discrepancy term and the penalty parameter will interact in course of the algorithm: given that τ is updated using its full conditional distribution (3.15) with

$$\mathrm{E}(\tau|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}},A,B) = \frac{\alpha+1}{D_{\mathrm{AB}}+\beta} \quad \text{and} \quad \mathrm{Var}(\tau|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}},A,B) = \frac{\alpha+1}{(D_{\mathrm{AB}}+\beta)^2}, \tag{3.18}$$

it follows that smaller values of D_{AB} are likely to increase the value of the precision parameter which in turn is likely to result in smaller discrepancy values.

The MCMC algorithm we propose to superimpose two unlabelled marked point sets $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}),\ldots,z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_1^{\mathrm{B}}+\boldsymbol{\gamma}),\ldots,z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}}+\boldsymbol{\gamma})\}$ will therefore usually progress as follows: at the start of the sampling procedure where the discrepancy between the point sets is usually large due to a poor superposition, small precision values will be proposed which allows the algorithm to accept many "uphill moves" in terms of the discrepancy. At this initial stage, the parameter space can therefore be explored thoroughly. However, once the goodness of the superposition increases and small discrepancy values are obtained, the value of the precision parameter will increase and thus prevent the algorithm from accepting superposition which result in a large discrepancy. In that, our MCMC algorithm is similar to the simulated annealing algorithm proposed by (Kirkpatrick *et al.*, 1983) which simulates from

$$\pi\big(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}_{\mathrm{A}},\boldsymbol{\lambda}_{\mathrm{B}},\tau|A,B,\alpha,\beta,\zeta\big)^{1/T},$$

where $T > 0$ is slowly reduced deterministically.

From the above considerations, it is also possible to formulate a guideline for the choice of the hyperparameters α and β. In essence, large values of α will result in large values of the precision parameter τ which could prevent the algorithm from adequately exploring the parameter space so that it is likely to get stuck in a local maximum of the posterior

distribution. Values which are too small on the other hand will prevent the algorithm from homing in on a good superposition. The same principle applies for β. In particular, we want β to be small enough to not mask the impact of the discrepancy value on the proposed values of τ. In any practical situation, choosing adequate values of α and β therefore has to be a balance between the two extremes.

3.6 Simulation Study

To evaluate the performance of our alignment method, we simulate unlabelled marked point sets $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{x}_1^{\mathrm{B}}), \ldots, z^{\mathrm{B}}(\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}})\}$ which share a common underlying reference field. This reference field determines the optimal relative position of A and B. Using the MCMC algorithm described above, the optimal alignment can be estimated by means of (post burn–in) posterior summary statistics of the accepted Euler angles and translation vectors. The performance of our alignment method can therefore be assessed by the deviation of the final relative position from the optimal alignment. If some contamination points which are not related to the underlying reference field are also included in the point sets, then posterior summary statistics of the mask vectors $\boldsymbol{\lambda}_{\mathrm{A}}$ and $\boldsymbol{\lambda}_{\mathrm{B}}$ provide a further way to validate our method.

3.6.1 Obtaining Marked Point Sets With a Common Reference Field

As a reference field we use a realisation of a zero–mean Gaussian random field. We generate this by defining a grid of 961 (31×31) regularly spaced points $\boldsymbol{y}_i$ within the unit square and simulating from

$$\tilde{\boldsymbol{Z}} = (\tilde{Z}(\boldsymbol{y}_1) \ldots \tilde{Z}(\boldsymbol{y}_{961}))^T \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $(\boldsymbol{\Sigma})_{ij} = \sigma_{\mathrm{W}}(||\boldsymbol{y}_i - \boldsymbol{y}_j||)$ is the value of a Whittle covariance function with unit variance and range parameter $\rho = 0.2$.

The Whittle covariance function is a member of the class of Matérn covariance functions which has the general form

$$\sigma_{\mathrm{M}}(\boldsymbol{h}) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}||\boldsymbol{h}||}{\rho} \right)^{\nu} K_{\nu} \left(\frac{2\nu^{1/2}||\boldsymbol{h}||}{\rho} \right), \tag{3.19}$$

where $\sigma^2 = \sigma(\mathbf{0})$ denotes the variance of the random field and $\rho > 0$ is a range parameter which determines how quickly the covariance between $Z(\boldsymbol{x})$ and $Z(\boldsymbol{x}+\boldsymbol{h})$ decreases with $||\boldsymbol{h}||$. Moreover, $\Gamma(.)$ denotes the Gamma function and $K_{\nu}(.)$ is the modified Bessel function of the third kind of order ν (e.g. Abramowitz & Stegun, 1964, Section 9.6). In this context, ν is a smoothness parameter as it determines the number of times $\sigma_{\mathrm{M}}(.)$ is differentiable at the origin (cf. e.g. Haskard, 2007). The above parameterisation has been suggested by Handcock & Wallis (1994) because it has the appealing property that the resulting correlation functions are comparable for different values of ν as the correlation at a separation distance of $||\boldsymbol{h}|| = \rho\sqrt{2}$ takes a value of approximately $\exp(-2)$. Using this parameterisation, (3.19) reduces to the exponential covariance function $\sigma_{\exp}(\boldsymbol{h}) = \sigma^2 \exp\{-\sqrt{2}\,||\boldsymbol{h}||/\rho\}$ for $\nu = 1/2$, and the Whittle covariance function which is associated with $\nu = 1$. Figure 3.1 displays several examples of the Matérn covariance function which are relevant in this simulation study. The red line shows the Whittle covariance function with range $\rho = 0.2$. The other lines show the exponential covariance function with $\rho = 0.2$ (yellow), $\rho = 0.3$ (blue) and $\rho = 0.1$ (green). In all cases, σ^2 is chosen to be one.

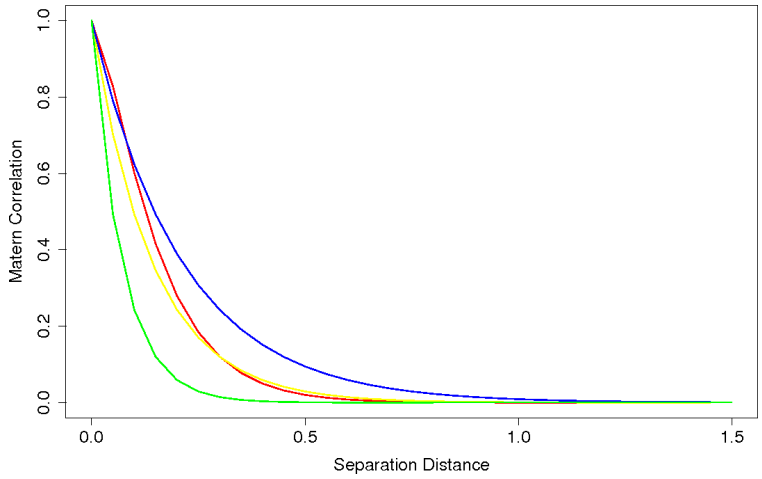

Figure 3.1: Examples of Matérn covariance functions: The red line shows the Whittle covariance function with $\rho = 0.2$. The other lines show the exponential correlation function with $\rho = 0.2$ (yellow), $\rho = 0.3$ (blue) and $\rho = 0.1$ (green). In all cases $\sigma^2 = 1$.

Figure 3.2 shows two realisations $\tilde{\boldsymbol{z}}$ of $\tilde{\boldsymbol{Z}}$. To be able to assess the performance of the mask vectors, we define the marked point sets in two parts, i.e. $A = \{A^{\text{true}}, A^{\text{cont}}\}$ and $B = \{B^{\text{true}}, B^{\text{cont}}\}$, where "true" denotes the part of each point set which stems from the underlying reference field $\tilde{\boldsymbol{z}}$ (and should therefore be included in the alignment procedure) and "cont" denotes the contaminated part. We obtain B^{true} by randomly choosing $k_{\text{B}}^{\text{true}}$ entries $i_1, \ldots, i_{k_{\text{B}}^{\text{true}}}$ from $\tilde{\boldsymbol{z}}$ and adding some Gaussian noise with standard deviation σ_ϵ to the corresponding marks, i.e.

$$B^{\text{true}} = \{z^{\text{B}}(\boldsymbol{x}_1^{\text{B}}), \ldots, z^{\text{B}}(\boldsymbol{x}_{k_{\text{B}}^{\text{true}}}^{\text{B}})\} = \{\tilde{z}(\boldsymbol{y}_{i_1}) + \epsilon_1^{\text{B}}, \ldots, \tilde{z}(\boldsymbol{y}_{i_{k_{\text{B}}^{\text{true}}}}) + \epsilon_{k_{\text{B}}^{\text{true}}}^{\text{B}}\}.$$

For B^{cont}, $k_{\text{B}}^{\text{cont}} = k_{\text{B}} - k_{\text{B}}^{\text{true}}$ locations on the (31×31) grid are chosen at random and the corresponding marks are random values from a uniform distribution on $[-c, c]$.

To obtain A^{true}, we introduce a nearness parameter $\kappa \in \mathbb{N}$ and define a set of grid points $\mathcal{U}_\kappa$ as the union of neighbourhoods around the points $\boldsymbol{x}_i^{\text{B}}$ $(i = 1, \ldots, k_{\text{B}}^{\text{true}})$, where each neighbourhood contains the vertically, horizontally and diagonally adjacent grid points in a $(2\kappa+1)\times(2\kappa+1)$–box around the corresponding $\boldsymbol{x}_i^{\text{B}}$. The points $\boldsymbol{x}_i^{\text{A}}$ $(i = 1, \ldots, k_{\text{A}}^{\text{true}})$ are then chosen at random from $\mathcal{U}_\kappa$ and A^{true} is defined as

$$A^{\text{true}} = \{z^{\text{A}}(\boldsymbol{x}_1^{\text{A}}), \ldots, z^{\text{A}}(\boldsymbol{x}_{k_{\text{A}}^{\text{true}}}^{\text{A}})\} = \{\tilde{z}(\boldsymbol{x}_1^{\text{A}}) + \epsilon_1^{\text{A}}, \ldots, \tilde{z}(\boldsymbol{x}_{k_{\text{A}}^{\text{true}}}^{\text{A}}) + \epsilon_{k_{\text{A}}^{\text{true}}}^{\text{A}}\}.$$

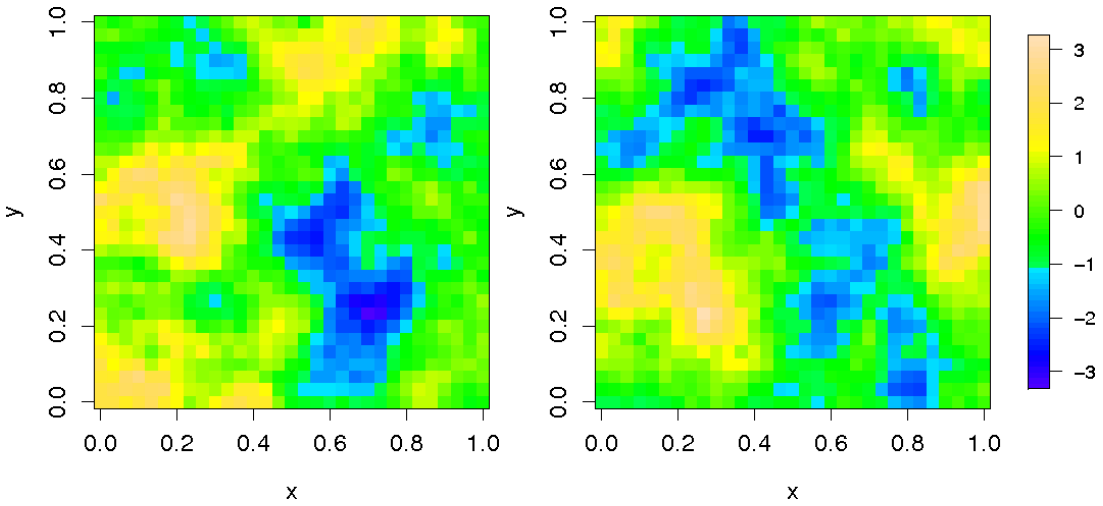

Figure 3.2: Examples of underlying reference fields: As reference fields we use realisations of a zero–mean isotropic Gaussian random field with a Matérn covariance function ($\nu = 1$ and $\rho = 0.2$, $\sigma^2 = 1$).

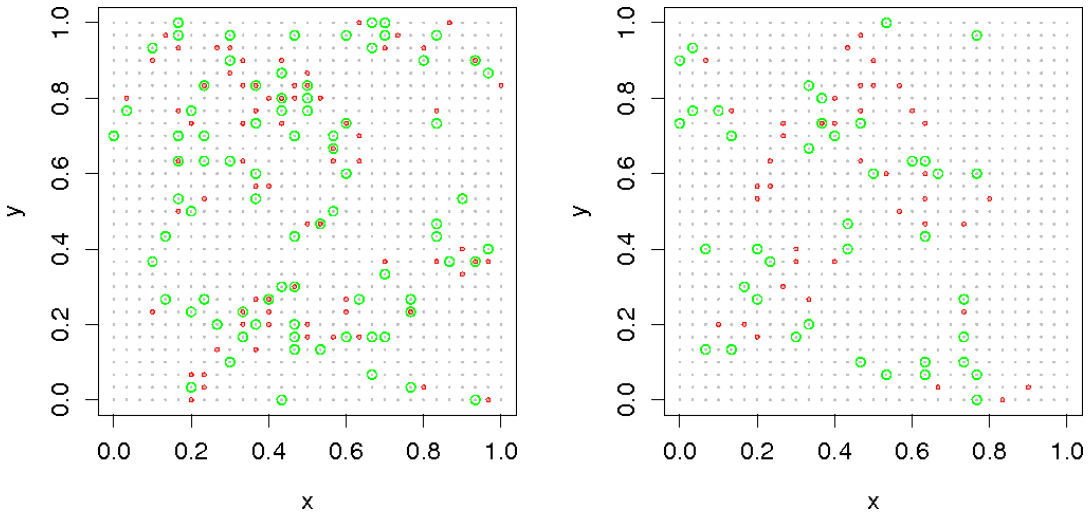

Figure 3.3: Two examples of sampling schemes: On both sides, the chosen points for B^{true} are shown as big green circles, and the points for A^{true} are shown as small red circles. On the left–hand side, we chose $k_{\text{B}}^{\text{true}} = k_{\text{B}}^{\text{true}} = 80$ and $\kappa = 1$ whereas the right–hand side shows the case $k_{\text{B}}^{\text{true}} = k_{\text{B}}^{\text{true}} = 40$ and $\kappa = 4$.

The $k_{\text{A}}^{\text{cont}} = k_{\text{A}} - k_{\text{A}}^{\text{true}}$ points in A^{cont} are obtained in the same way as the contamination points in B. Note that this sampling scheme does not create correspondences between points in A^{true} and B^{true}. This is further demonstrated in Figure 3.3, where the left–hand side shows an example of a sampling scheme with $k_{\text{B}}^{\text{true}} = k_{\text{A}}^{\text{true}} = 80$ and $\kappa = 1$ whereas $k_{\text{B}}^{\text{true}} = k_{\text{A}}^{\text{true}} = 40$ and $\kappa = 4$ on the right–hand side. The big green circles show the points in B^{true} and the small red circles show the points in A^{true}. The impact of κ is clearly visible. For our simulation study we consider three realisations of $\tilde{\boldsymbol{Z}}$, and for each of these realisations we define 12 different pairs of marked point sets by varying the parameters $k^{\text{true}} = k_{\text{A}}^{\text{true}} = k_{\text{B}}^{\text{true}} \in \{40, 80\}$, $k^{\text{cont}} = k_{\text{A}}^{\text{cont}} = k_{\text{B}}^{\text{cont}} \in \{0.05k^{\text{true}}, 0.1k^{\text{true}}, 0.15k^{\text{true}}\}$ and $\kappa \in \{1, 4\}$. Moreover, we choose $c = 7$ and $\sigma_\epsilon = \sqrt{0.02}$. Generated as above, the 36 pairs A and B are recorded in the optimal relative position, and the optimal mask vectors are $\boldsymbol{\lambda}_{\text{A}}^T = (\mathbf{1}_{k_{\text{A}}^{\text{true}}}^T, \mathbf{0}_{k_{\text{A}}^{\text{cont}}}^T)$ and $\boldsymbol{\lambda}_{\text{B}}^T = (\mathbf{1}_{k_{\text{B}}^{\text{true}}}^T, \mathbf{0}_{k_{\text{B}}^{\text{cont}}}^T)$.

3.6.2 Hyperparameter Settings

To obtain a starting point for the algorithm, we randomly rotate and translate B away from this position, and the MCMC algorithm should ideally reconstruct the original

alignment as closely as possible. For each pairwise superposition 50,000 MCMC iterations are carried out which each contain five blocks updating the rotation parameter (proposal sd: 0.75°), the translation vector (proposal sd for each entry: 0.01), the precision parameter, and the two mask vectors, respectively.

The Kernel Carbo similarity calculations are based on the exponential kernel, i.e. (3.19) with $\nu = 0.5$. Initially we use $\rho = 0.6$ but within the first 1,000 iterations, this value is dynamically reduced to $\rho = 0.2$. This initial phase allows the algorithm to home in on a good alignment even if the two points sets are far away from their optimal relative position. After the initial phase, $\rho = 0.2$ is kept fixed for all iterations. The corresponding covariance function is shown as the yellow line in Figure 3.1. The covariance estimation described in Section 2.2.1.2 is not applied here as the contamination points distort the empirical semivariogram (cf. also Appendix B where we describe an *ad–hoc* approach to alleviate this problem). However, in the following we will show that choosing the exact right form of the covariance kernel is not essential for a good alignment.

The hyperparameters which govern the full conditional distribution of τ are α and β. We choose $\beta = 0.05$ which yielded good results in pilot runs: larger values for β mask the impact of the discrepancy on the proposed values for τ whereas smaller values increase the full conditional mean of τ and therefore the probability of getting trapped in a local mode (cf. Section 3.5.3). The same reasoning applies for the chosen value of $\alpha = 200$.

In the pilot runs it became obvious that the value for the penalty parameter ζ has a big impact on the alignment result. We therefore include ζ as a variable parameter in our simulation study and consider $\zeta = \{10, 50, 90\}$.

Simultaneous inference about the rigid–body parameters, the precision parameter and the mask vectors is a difficult task and it is not surprising that the MCMC algorithm sometimes gets trapped in a local mode. To overcome this difficulty, we propose a big change of the rigid–body parameters by increasing the standard deviations of the random walk proposals to 60° (rotation) and 0.3 (translation) every 125 iterations. Moreover, we restart the algorithm if the Carbo distance exceeds 0.3 after 7,500 iterations.

3.6.3 Results

For each of the 36 pairs A and B, we use all three values of ζ so that we consider 108 MCMC runs. For each run, the starting position of the movable point set B is obtained by rotating and translating it using $\mathbf{\Gamma}(\theta_0)$ and $\boldsymbol{\gamma}_0$, where θ_0 and $\gamma_{0i}\,(i = 1, 2)$ are uniformly distributed on $[-20°, 20°]$ and $[-0.1, 0.1]$, respectively. Moreover, both mask vectors are initiated using $\lambda_i^{\mathrm{M}} \sim \text{Bernoulli}(0.5)$ $(i = 1, \ldots, k_{\mathrm{M}};\, M = A, B)$. Figures 3.4-3.6 show the typical output of a successful run. As described in Section 2.3.4, the convergence of the MCMC algorithm can be monitored by the trace plots of the involved parameters. The trace plots in Figures 3.4 and 3.5 indicate that the algorithm converges quickly. This is due to the interplay between the precision parameter τ and the Kernel Carbo distance described in Section 3.5.3 which can be observed in the left–hand side and right–hand side plot of the bottom row of Figure 3.4. Obviously, this interplay leads to a steady increase of the posterior distribution (3.13).

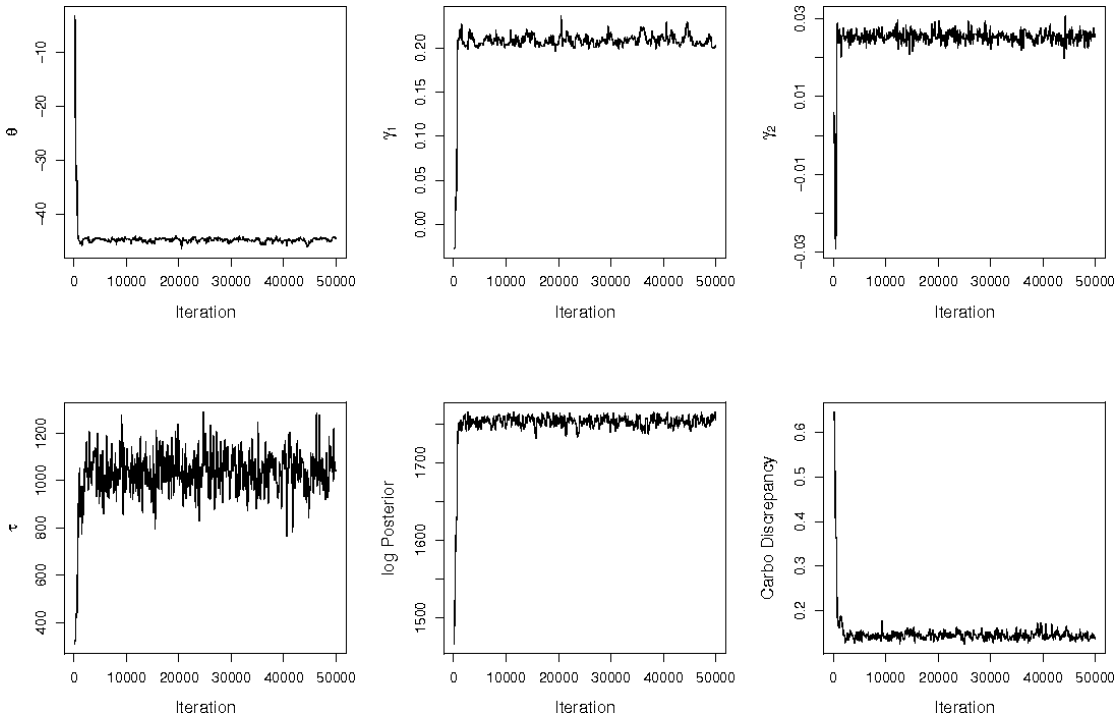

Figure 3.4: Top Row: Trace plot of the rigid–body parameters (in terms of the initial relative position of the two points sets under consideration). Bottom row: Trace plots of the precision parameter, the log–posterior (up to a constant) and the Kernel Carbo discrepancy. In all plots, every $100th$ simulated value is displayed.

The top row of Figure 3.5 shows the trace plots for the number of points $\sum_{i=1}^{k_{\mathrm{A}}} \lambda_i^{\mathrm{A}}$ and $\sum_{i=1}^{k_{\mathrm{B}}} \lambda_i^{\mathrm{B}}$ which are involved in the field calculation and are hence considered to belong to A^{true} and B^{true}, respectively. Like the other parameters, these values also converge quickly. A (post burn–in) summary of the two mask vectors is displayed in the bottom row of Figure 3.5. The big circles show the mean values of the binary entries over all post–burn in iterations. As the entries of the corresponding mean vectors $\bar{\boldsymbol{\lambda}}_{\mathrm{A}}$ and $\bar{\boldsymbol{\lambda}}_{\mathrm{B}}$ take values between zero and one, they can be interpreted as the estimated posterior probability of the corresponding point belonging to the matching part of the point sets. Moreover, the small circles display the mask vectors which are observed at the maximum a posteriori (cf. Section 2.3.3) iteration. For the example considered here, the true mask vectors are $\boldsymbol{\lambda}_{\mathrm{M}}^T = (\mathbf{1}_{80}^T, \mathbf{0}_{12}^T)$ $(M = A, B)$, and the algorithm is able to reconstruct the mask vector very well for both point sets.

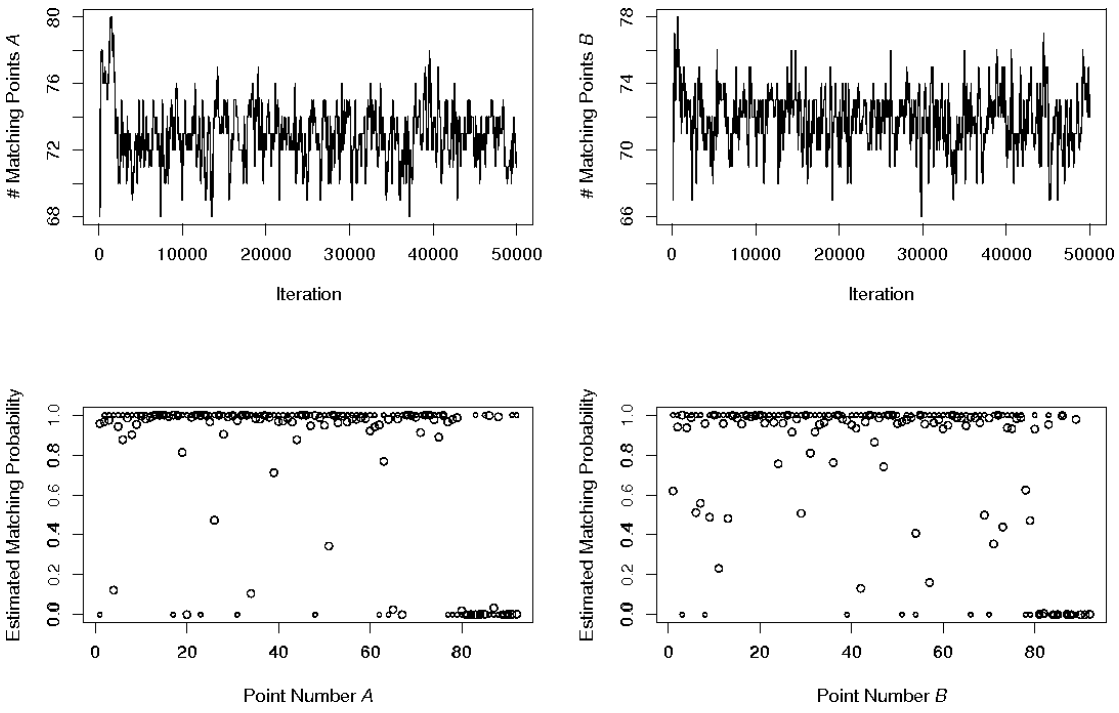

Figure 3.5: Top Row: Trace plots of the number of points involved in the field calculation. Bottom Row: Two possible point estimates for the mask vectors of A (left) and B (right). The big circles show the mean values of the (0,1)–entries for the masks vectors (which can be interpreted as the estimated posterior probability for the corresponding point to belong to the common reference field), and the small circles display the observed mask vectors at the MAP iteration. The total number of points in A and B is 92, and the last 12 points in each set are contamination points.

One way to assess the performance of the alignment procedure numerically is to use the MAP estimates of the rigid–body parameters to transform the moveable point set B from its initial position to the MAP position and to determine the closeness of the MAP position to the position which was originally generated as described in Section 3.6.1. The closeness can be determined by the root mean squared deviation (RMSD) of the corresponding configuration matrices $\boldsymbol{X}_{\mathrm{B}}^{\mathrm{MAP}} \in \mathbb{R}^{k_{\mathrm{B}} \times 2}$ and $\boldsymbol{X}_{\mathrm{B}}^{\mathrm{orig}} \in \mathbb{R}^{k_{\mathrm{B}} \times 2}$, i.e.

$$\mathrm{RMSD}\left(\boldsymbol{X}_{\mathrm{B}}^{\mathrm{MAP}}, \boldsymbol{X}_{\mathrm{B}}^{\mathrm{orig}}\right) = \sqrt{\frac{1}{k_{\mathrm{B}}} \sum_{i=1}^{k_{\mathrm{B}}} ||\boldsymbol{x}_{i,\mathrm{MAP}}^{\mathrm{B}} - \boldsymbol{x}_{i,\mathrm{orig}}^{\mathrm{B}}||^2},$$

where $\boldsymbol{x}_{i,\mathrm{MAP}}^{\mathrm{B}}$ and $\boldsymbol{x}_{i,\mathrm{orig}}^{\mathrm{B}}$ denote the xy–coordinate vectors of the ith point in the corresponding configuration matrix. For our example, Figure 3.6 shows the initial (left) and the MAP (right) position of B. In both cases, the original position is displayed in grey. The MCMC algorithm is able to reproduce the original position very well. Here, the RMDS–value is reduced from 0.479 to 0.032 by the alignment algorithm.

In the following, we choose a RMSD–value of 0.1 as a benchmark for a successful alignment. With this benchmark, 76% of the runs can be considered as successful. As expected, the best performance (89% success rate) is achieved with $k^{\mathrm{true}} = 80$ and $k^{\mathrm{cont}} = 4$. If the nearness parameter is in addition chosen as $\kappa = 1$, 100% of the MCMC runs are

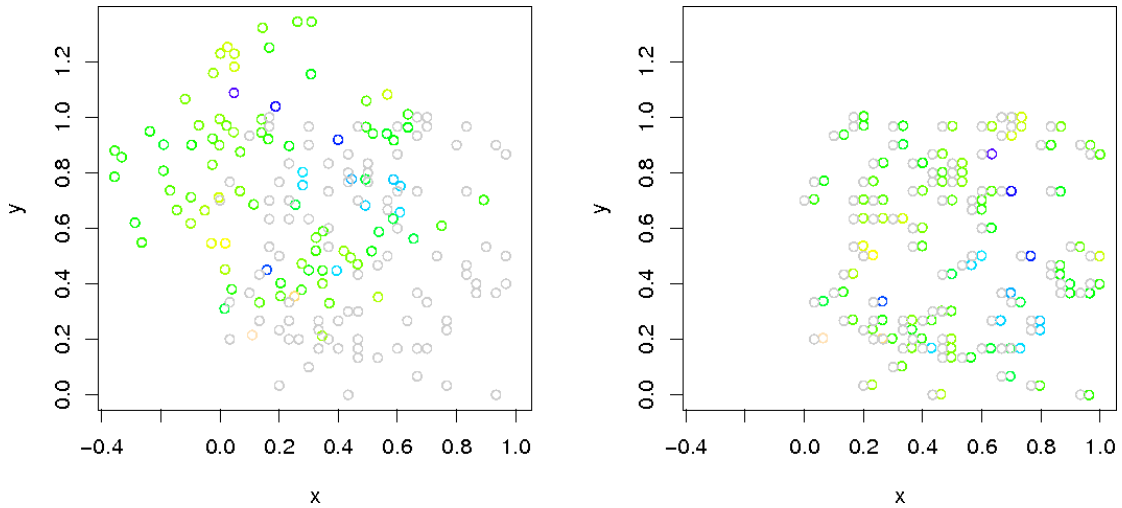

Figure 3.6: Successful alignment: The left–hand side shows the initial position of point set B and the right–hand side shows the position of B at the MAP iteration. Like in Figure 3.2 the colours correspond to the values of the marks. The original position is displayed in grey on both sides.

successful. The most difficult setting is $k^{\text{true}} = 40$ and $k^{\text{cont}} = 6$ with a success rate of 44%. If in addition $\kappa = 4$, this combination yields a success rate of only 22%. Overall, the number of contamination points has the biggest impact on the results as the success rate drops from 92% for $k^{\text{cont}} = 0.05k^{\text{true}}$ to 53% for $k^{\text{cont}} = 0.15k^{\text{true}}$. The impacts of the nearness parameter and the penalty parameter in this setting are considerable as well: 85% for $\kappa = 1$ and 67% for $\kappa = 4$, and 61% for $\zeta = 10$, 81% for $\zeta = 50$ and 86% for $\zeta = 90$.

The above results point out that a satisfactory alignment can be obtained if the number of non–contamination points is large enough to represent the main features of the underlying reference field and large relative to the number of contamination points. Moreover, especially when the number of points is small and the sampling of the reference field is sparse, it is important that the non–contamination points in A and B represent the same features of the reference field (which is not always the case if $k^{\text{true}} = 40$ and $\kappa = 4$). From an algorithmic point of view, large values for the penalty value ζ are favourable as they prevent the algorithm from converging to solutions with a low Kernel Carbo distance mainly by dismissing relevant points from the field calculation.

All the above trends can be emphasised by rerunning the experiments using $\theta \sim U_{[-60°,60°]}$ and $\gamma_i \sim U_{[-0.3,0.3]}$ $(i = 1, 2)$ to obtain the starting position of B. In this more challenging setting, 48% of the 108 runs can be classified as successful, and the sampling scheme for the point sets drastically influences the success rate as it ranges from 83% ($k^{\text{true}} = 80$ and $k^{\text{cont}} = 4$) to 17% ($k^{\text{true}} = 40$ and $k^{\text{cont}} = 6$). The impact of the penalty parameter in this setting can be summarised as: 33% for $\zeta = 10$, 47% for $\zeta = 50$ and 61% for $\zeta = 90$.

In both settings, the performance of our alignment procedure can be much improved if there are some points in A and B which can be identified as non–contamination points prior to the alignment because in that case, the corresponding entries of the mask vectors can be fixed to one. For our examples, identifying some relevant points (on average 12 per point set) improves the overall success rate from 76% to 93% in the first setting and from 48% to 78% in the second setting. In many applications it may be possible to identify some relevant points so that the possibility of incorporating this knowledge is a valuable tool to improve the alignment in practice.

Finally, we rerun the above experiments with different values for the range parameter ρ. For example with $\rho = 0.3$ (displayed as the blue line in Figure 3.1), overall success rates of 77% in the first and 48% in the second, more challenging setting are achieved, and for $\rho = 0.1$ (displayed as the green line in Figure 3.1), the corresponding success rates are 77% and 52%. These results demonstrate that choosing the correct covariance function for the spatial interpolation is not crucial for the performance of the algorithm.

3.7 Summary

In this chapter we proposed a novel method for aligning two unlabelled marked point sets which is based on the assumption that both point sets are noisy pointwise observations of a common hidden reference field. The main difference between our methodology and previous approaches to the problem is that we use spatial interpolation of the given marks to obtain a continuous representation of the point sets. Within the framework of reproducing kernel Hilbert spaces and by using the Carbo index from structural bioinformatics, a similarity index of the two points sets can be formulated in terms of the predicted fields. This has the advantage that point correspondences do not need to be estimated.

The actual alignment is carried out within the MCMC framework. This enables us to incorporate mask vectors which automatically determine the matching regions of the considered point sets whilst ignoring the rest which helps to reduce the level of noise in the alignment procedure. Our alignment method works well in a simulation study – in particular if the point sets satisfy a certain nearness criterion which measures whether or not they represent the same features of the hidden reference field.

In the next chapter, we will apply the algorithm to the steroid data described in Section 1.1.3. In the context of structural alignment of molecules, it is also of interest to align several molecules simultaneously. We will therefore also propose an extension of the method described in this chapter which can carry out an alignment of multiple unlabelled marked point sets.

Chapter 4

Bayesian Alignment of Continuous Molecular Shapes

As described in Chapter 1, the concept of molecular similarity is of great importance in rational drug design because similar molecules can be expected to exhibit a similar drug potency. As molecular data are often given in the form of unlabelled marked point sets where the individual points represent the position of atoms within the molecules and the marks are some additional properties such as partial charge values which have been measured at the atom locations, the (partial) Kernel Carbo index and the MCMC scheme developed in the previous chapter can directly be utilised to obtain a shape–based similarity index for molecules.

In particular if all molecules of the given dataset bind to the same receptor, the assumption of a common underlying reference field is suitable for this application because the underlying reference field can in that case be interpreted as a negative imprint of the binding pocket of the receptor. The MCMC scheme described in Section 3.5 then determines the parts of each molecule which fit into the binding pocket and aligns the molecules based on these parts only so that the resulting relative position should reproduce the relative binding positions of the molecules. Moreover, using a field–based representation of the molecules is not only beneficial in that correspondences between atoms of different molecules do not need to estimated. It also provides a possibility to account for the continuous, fuzzy nature of a molecule.

Section 4.1 provides a brief literature review of previously proposed molecular alignment techniques and points out similarities and differences to our method. In Section 4.2, we apply our methodology to the steroid dataset and show that resulting similarity values are chemically relevant in that they are associated with the differences in the binding activity to the common receptor protein. To investigate this fact further and assess where around the molecular skeletons the differences occur, we propose an extension of our superposition algorithm which can perform an alignment of multiple marked point sets in Section 4.3. We apply this extension to the steroid data in Section 4.4. Section 4.5 provides a summary of the main points in this chapter.

4.1 Structural Alignment of Molecules – Literature Review

As molecular recognition is inherently a three–dimensional phenomenon, most similarity indices and their associated methods for structural molecular alignment are based on comparing the three–dimensional geometrical features of the molecules of interest. These methods generally fall into three categories, namely atom–based methods, methods which are based on hard–sphere representations of the involved molecules, and field–based methods.

Atom–based methods mainly utilise the configuration matrix $\boldsymbol{X}^{\mathrm{M}}$ of each molecule for determining a suitable superposition. Additional information from the marks is not necessarily required but can be used to improve the superposition results. From a statistical point of view, atom–based alignment of molecules therefore is the problem of aligning unlabelled point sets. The relevant approaches have already been described in Section 3.2, and the links to our alignment methodology have been pointed out.

Hard–sphere models for molecular shape treat a molecule as a set of intersecting spheres centred at the atom position. The most common choice for the associated radii is the van der Waals radii. In a hard–sphere model, the volume of a molecule is usually defined as the volume of the union of the van der Waals spheres which can be calculated based

on the inclusion–exclusion principle (e.g. Hall, 1998, pp.8) as

$$V(\mathrm{M}) = V(\cup\, \mathrm{S}_i^{\mathrm{M}}) = \sum_{i=1} V(\mathrm{S}_i^{\mathrm{M}}) - \sum_{i<j} V(\mathrm{S}_i^{\mathrm{M}} \cap \mathrm{S}_j^{\mathrm{M}}) + \sum_{i<j<k} V(\mathrm{S}_i \cap \mathrm{S}_j^{\mathrm{M}} \cap \mathrm{S}_k^{\mathrm{M}}) - \ldots, \quad (4.1)$$

where $\mathrm{S}_i^{\mathrm{M}}$ denotes the van der Waals sphere of the the ith atom in M whose radius is equal to the corresponding van der Waals radius r_i^{M} $(i = 1, \ldots, k_{\mathrm{M}})$. Based on their van der Waals volumes, the similarity of two molecules A and B in a certain relative position can then be defined as the volume of their overlapping parts

$$V_{\mathrm{AB}}\big(\boldsymbol{\Gamma}, \boldsymbol{\gamma}\big) = V(\mathrm{A} \cap \mathrm{B}) = V(\mathrm{A}) + V(\mathrm{B}) - V(\mathrm{A} \cup \mathrm{B}). \quad (4.2)$$

This overlap volume is a special case of (3.1) where the marks of the two molecules (point sets) are the van der Waals radii. A rotation/translation invariant similarity can therefore be obtained by maximising (4.2) with respect to rotation and translation (e.g. Masek *et al.*, 1993).

It can be argued (e.g. Mezey, 1995) that both the atom–based and the hard–sphere methods do not reflect the true nature of the involved molecules which are in fact fuzzy bodies of electronic clouds whose electron density fades away gradually with the distance from the molecular skeleton. To account for the fuzziness of molecular bodies, Grant & Pickup (1995) define a molecular density as

$$\rho_{\mathrm{M}}^{\mathrm{G}}(\boldsymbol{x}) = \sum_{i} \rho_{\mathrm{M},i}^{\mathrm{G}}(\boldsymbol{x}) - \sum_{i<j} \rho_{\mathrm{M},i}^{\mathrm{G}}(\boldsymbol{x})\rho_{\mathrm{M},j}^{\mathrm{G}}(\boldsymbol{x}) + \sum_{i<j<k} \rho_{\mathrm{M},i}^{\mathrm{G}}(\boldsymbol{x})\rho_{\mathrm{M},j}^{\mathrm{G}}(\boldsymbol{x})\rho_{\mathrm{M},k}^{\mathrm{G}}(\boldsymbol{x}) - \ldots, \quad (4.3)$$

where $\rho_{\mathrm{M},i}^{\mathrm{G}}(\boldsymbol{x}) = \gamma_i^{\mathrm{M}} \exp(-\alpha_i^{\mathrm{M}} ||\boldsymbol{x} - \boldsymbol{x}_i^{\mathrm{M}}||^2)$ is an isotropic Gaussian function centred at the ith atom position $\boldsymbol{x}_i^{\mathrm{M}}$. With this definition, the modified molecular volume becomes $V^{\mathrm{G}}(\mathrm{M}) = \int \rho_{\mathrm{M}}^{\mathrm{G}}(\boldsymbol{x}) d\boldsymbol{x}$ which is a direct generalisation of $V(\mathrm{M})$ as (4.1) can be written in the same form as (4.3) but using the step functions $\rho_{\mathrm{M},i}^{\mathrm{HS}}(\boldsymbol{x}) = I_{\{||\boldsymbol{x}-\boldsymbol{x}_i^{\mathrm{M}}|| \leq r_i^{\mathrm{M}}\}}$. Grant & Pickup (1995) choose the parameters γ_i^{M} and α_i^{M} of each $\rho_{\mathrm{M},i}^{\mathrm{G}}(.)$ so that the new "volume" of the ith atom matches that of the corresponding van der Waals sphere, i.e.

$$\int \rho_{\mathrm{M},i}^{\mathrm{G}}(\boldsymbol{x}) d\boldsymbol{x} = \int \rho_{\mathrm{M},i}^{\mathrm{HS}}(\boldsymbol{x}) d\boldsymbol{x} = V(\mathrm{S}_i^{\mathrm{M}}) = \tfrac{4}{3}\pi {r_i^{\mathrm{M}}}^3, \quad i = 1, \ldots, k_{\mathrm{M}}.$$

The Gaussian version $V^{\mathrm{G}}(\mathrm{M})$ of the molecular volume therefore resembles the hard–

sphere version $V(\mathrm{M})$. However, $V^{\mathrm{G}}(\mathrm{M})$ is based on a softer description of the molecular density which is more in line with the true nature of a molecule. In a follow–up paper, Grant *et al.* (1996) use the above definitions to obtain a similarity measure for two molecules A and B in a certain relative position as

$$V^{\mathrm{G}}_{\mathrm{AB}}(\mathbf{\Gamma}, \boldsymbol{\gamma}) = \int \rho^{\mathrm{G}}_{\mathrm{A}}(\boldsymbol{x}) \rho^{\mathrm{G}}_{\mathrm{B}}(\boldsymbol{x}) d\boldsymbol{x},$$

which again is a special case of (3.1) so that optimising over rotation and translation provides a rotation/translation invariant similarity index.

The work by Grant *et al.* (1996) can be viewed as a link between molecular alignment techniques which are based on hard–sphere representations of molecular shapes and the family of field–based methods where each molecule M ($M \in \{A, B\}$) is represented as a field $P_{\mathrm{M}}(\boldsymbol{x})$ of a molecular property P over $\mathbb{R}^3$. One possible use of the field representations is to obtain sets of isosurfaces which can be compared using topological considerations (e.g. Mezey, 1993). More commonly, however, the fields are compared over the entire space $\mathbb{R}^3$ using overlap–based functions such as the (L_2–)Carbo function

$$C^{L_2}_{\mathrm{AB}}(\mathbf{\Gamma}, \boldsymbol{\gamma}) = \frac{\int P_{\mathrm{A}}(\boldsymbol{x}) P_{\mathrm{B}}(\boldsymbol{x}) d\boldsymbol{x}}{\big(\int P^2_{\mathrm{A}}(\boldsymbol{x}) d\boldsymbol{x}\big)^{1/2} \big(\int P^2_{\mathrm{B}}(\boldsymbol{x}) d\boldsymbol{x}\big)^{1/2}} \tag{4.4}$$

whose kernelised version we utilise in our alignment method (cf. Section 3.4). For an overview of other field–based similarity indices see for example Petke (1993).

The (L_2–)Carbo index has originally been proposed to assess the similarity of two molecules with respect to their electron density. Despite having the virtue of being firmly grounded in quantum chemistry, the electron density of a molecule M is hard to calculate and was soon to be replaced by an approximation of the form

$$P^{\mathrm{Q}}_{\mathrm{M}}(\boldsymbol{x}) = \sum_{i=1}^{k_{\mathrm{M}}} \frac{q^{\mathrm{M}}_i}{||\boldsymbol{x} - \boldsymbol{x}^{\mathrm{M}}_i||}, \tag{4.5}$$

where $\boldsymbol{x}^{\mathrm{M}}_i$ denotes position of the ith atom of M and q^{M}_i denotes the associated partial charge value. Similar to the kriging–based evaluation of fields described in Section 3.3, the field (4.5) is obtained as a linear combination of the given marks (i.e. the partial

charge values in this case). However, the weights in (4.5) are the inverse distance weights which contain less information than the kriging weights in (3.7). Moreover, if the fields $P_{\mathrm{A}}^{\mathrm{Q}}(\boldsymbol{x})$ and $P_{\mathrm{B}}^{\mathrm{Q}}(\boldsymbol{x})$ of two molecules A and B are inserted into (4.4), the overlap integral in the numerator cannot be evaluated without expensive numerical calculations.

To overcome the latter drawback, Good *et al.* (1992) propose a further approximation to the electron density and replace the inverse distance weights in (4.5) by a series of isotropic Gaussian functions, i.e.

$$\tilde{P}_{\mathrm{M}}^{\mathrm{Q}}(\boldsymbol{x}) = \sum_{i=1}^{k_{\mathrm{M}}} q_i^{\mathrm{M}} \left(\tilde{\gamma}_1^{\mathrm{M}} \exp\{-\tilde{\alpha}_1^{\mathrm{M}} ||\boldsymbol{x} - \boldsymbol{x}_i^{\mathrm{M}}||^2\} + \ldots + \tilde{\gamma}_{n_{\mathrm{G}}}^{\mathrm{M}} \exp\{-\tilde{\alpha}_{n_{\mathrm{G}}}^{\mathrm{M}} ||\boldsymbol{x} - \boldsymbol{x}_i^{\mathrm{M}}||^2\} \right). \quad (4.6)$$

For a given order n_{G} of the Gaussian expansion, the coefficients $\tilde{\alpha}_k^{\mathrm{M}}$ and $\tilde{\gamma}_k^{\mathrm{M}}$ are chosen to optimally fit the inverse distance terms in a least–squares sense. The resulting values for $n_{\mathrm{G}} \leq 3$ can be found in Good (1995). If the above approximation of the electron density is inserted into (4.4), then both the numerator and the denominator reduce to a series of two-centre Gaussian overlap integrals which can be solved analytically. The required optimisation over rotation and translation can therefore be carried out using gradient–based methods (McMahon & King, 1997).

The above shows that Gaussian functions play an important role in the field–based structural alignment of two molecules as the overlap integral of two Gaussians can be evaluated analytically. Another method which makes use of Gaussian functions is the SEAL (Steric and Electrostatic Alignment) method proposed by Kearsley & Smith (1990) where two molecules A and B are aligned by maximising the similarity index

$$S_{\mathrm{AB}}^{\mathrm{SEAL}}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \sum_{i=1}^{k_{\mathrm{A}}} \sum_{j=1}^{k_{\mathrm{B}}} w_{ij} \exp\left(-\alpha^{\mathrm{SEAL}} ||\boldsymbol{x}_i^{\mathrm{A}} - (\boldsymbol{\Gamma} \boldsymbol{x}_j^{\mathrm{B}} + \boldsymbol{\gamma})||^2\right) \quad (4.7)$$

with respect to rotation and translation. The weights w_{ij} are thereby chosen to be weighted averages of the electrostatic and steric properties of atom i in A and atom j in B, i.e. $\omega_{ij} = w_Q q_i^{\mathrm{A}} q_j^{\mathrm{B}} + w_S v_i^{\mathrm{A}} v_j^{\mathrm{B}}$, where q_i^{M} denotes the partial charge value associated with the ith atom position in molecule M and v_i^{M} denotes some power of the corresponding van der Waals radius r_i^{M}.

Despite not being explicitly based on the overlap of two molecular fields in a certain relative position, it is the SEAL function which is most directly related to our proposed kriging–based Kernel Carbo similarity index (3.9). To see this consider two unlabelled marked point sets $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_1^{\mathrm{B}}+\boldsymbol{\gamma}), \ldots, z^{\mathrm{B}}(\boldsymbol{\Gamma}\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}}+\boldsymbol{\gamma})\}$. If the kriging is performed based on the Gaussian covariance function

$$\sigma^{\mathrm{G}}(\boldsymbol{h}) = \sigma^2 \exp\{-||\boldsymbol{h}||^2/\rho^2\}, \tag{4.8}$$

(which is another special case of a Matérn covariance function (3.19) with $\nu \to \infty$), then the kriged fields have the form

$$\hat{Z}_{\mathrm{A}}(\boldsymbol{x}) = \sigma^2 \sum_{i=1}^{k_{\mathrm{A}}} w_i^{\mathrm{A}} \exp\{||\boldsymbol{x}_i^{\mathrm{A}} - \boldsymbol{x}||/\rho^2\} \text{ and } \hat{Z}_{\mathrm{B}}(\boldsymbol{x}) = \sigma^2 \sum_{i=1}^{k_{\mathrm{B}}} w_i^{\mathrm{B}} \exp\{||(\boldsymbol{\Gamma}\boldsymbol{x}_i^{\mathrm{B}} + \boldsymbol{\gamma}) - \boldsymbol{x}||/\rho^2\},$$

where the weight w_i^{M} denotes the ith element of the weight vector $\boldsymbol{w}_{\mathrm{M}} = \boldsymbol{\Sigma}_{\mathrm{M}}^{-1}\boldsymbol{z}_{\mathrm{M}}$ with $(\boldsymbol{\Sigma})_{ij}^{\mathrm{M}} = \sigma^2 \exp\{-||\boldsymbol{x}_i^{\mathrm{M}} - \boldsymbol{x}_j^{\mathrm{M}}||^2/\rho^2\}$ ($M \in \{A, B\}$). In that case the (L_2–)Carbo index becomes

$$C_{\mathrm{AB}}^{L_2}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^{k_{\mathrm{A}}} \sum_{j=1}^{k_{\mathrm{B}}} w_i^{\mathrm{A}} w_j^{\mathrm{B}} \exp\left\{-||\boldsymbol{x}_i^{\mathrm{A}} - (\boldsymbol{\Gamma}\boldsymbol{x}_j^{\mathrm{B}} + \boldsymbol{\gamma})||^2/(2\rho^2)\right\}}{N^{\mathrm{A}} N^{\mathrm{B}}}, \tag{4.9}$$

where $N^{\mathrm{M}} = \left(\sum_{i=1}^{k_{\mathrm{M}}} \sum_{j=1}^{k_{\mathrm{M}}} w_i^{\mathrm{M}} w_j^{\mathrm{M}} \exp\{-||\boldsymbol{x}_i^{\mathrm{M}} - \boldsymbol{x}_j^{\mathrm{M}}||^2/(2\rho^2)\}\right)^{1/2}$ $M \in \{A, B\}$. Note that if the Gaussian covariance function is used, then the (L_2–)Carbo index of the kriged fields (4.9) is almost identical to its kernelised version which is of the same form but without the factor two in the denominator term within the exponential function. This is a special feature of the Gaussian covariance function which in turn makes the numerator of both the (L_2–)Carbo index and its kernelised version very similar to the SEAL objective function (4.7) if kriging is used to construct the molecular fields.

The SEAL method is well–established in the structural alignment community so that the similarity of its objective function to our kriging–based (partial) Kernel Carbo index is reassuring. While it does not allow for the possibility that only parts of the molecules match, the SEAL method does provide the opportunity to incorporate two molecular properties, namely the steric properties in the form of the van der Waals radii and the electrostatic properties in the form of the partial charge values. The same concept of

using a weighted average of multiple properties can also be applied to the (partial) Kernel Carbo index, and for the following application we introduce a multivariate version of (3.12) which is obtained by first calculating the univariate indices separately, and then calculating a weighted average where the weights are positive and normalised to sum to one. The resulting multivariate partial Kernel Carbo index therefore takes values between minus one and one like its univariate equivalent and can therefore be directly transformed to a distance and utilised within the MCMC scheme in the same way.

4.2 Application to the Steroid Molecules

We now consider the application of our alignment method for unlabelled marked point sets to the steroid data. As described in Section 1.1.3, the (numerical) marks provided at the atom location for each of the 31 steroid molecules in this dataset are the van der Waals radii and the partial charge values. Moreover, the unit of the xyz–coordinates is Å (Ångström). As the alignment carried out by the MCMC algorithm is asymmetric in the sense that molecule A is treated as fixed and the other molecule B as moveable, we carry out each of the 930 (31·30) possible pairwise superpositions.

4.2.1 Hyperparameter Settings

For each superposition, 10,000 MCMC iterations are used, and each iteration contains five blocks updating rotation, translation, precision, and the two mask vectors, respectively. In an initial phase of the MCMC algorithm, we use the information about both the partial charge values and the (cubed) van der Waals radii by calculating a bivariate partial Kernel Carbo index as described above. Both univariate indices are thereby based on the Gaussian covariance function (4.8). As the variance parameters cancel out, they do not need to be estimated. Assuming that the electrostatic field which gives rise to the partial charge values of the molecules has the same covariance structure across the given steroids, the range parameter ρ for the electrostatic field is estimated by visual

inspection of a pooled empirical semivariogram function of all 31 considered molecules (cf. Section 2.2.1.2) where "pooled" means that the semivariogram clouds of all molecules are combined before the distance classes are obtained. Doing so yields a range parameter of $\rho_{\mathrm{Q}}^2 = 40.33$, and the range of the steric field is taken to be the largest van der Waals radius in the data set, i.e. $\rho_{\mathrm{S}}^2 = 8.67$.

The initial phase for each pairwise superposition comprises $n_{\mathrm{I}} = 2,000$ MCMC iterations during which the relative weights of the univariate partial Kernel Carbo indices are chosen dynamically as

$$w_{\mathrm{Q}} = \frac{n_{\mathrm{I}} - i}{n_{\mathrm{I}}} \quad \text{and} \quad w_{\mathrm{S}} = \frac{i}{n_{\mathrm{I}}}, \quad i = 1, \ldots, n_{\mathrm{I}}. \tag{4.10}$$

The electrostatic field are therefore only used for an approximate alignment and their impact fades out as the algorithm proceeds. This has a similar effect to decreasing the range of a univariate partial Kernel Carbo index dynamically as we did in the simulation study in Section 3.6 and helps the algorithm home in on a good solution. In the molecular context, however, the above bivariate method has the advantage that it directly mimics real–life molecular recognition where the long–range electrostatic attraction governs the initial approach of the molecules whereas the short–range repulsive steric forces gradually take over and become the chief manipulator for the binding affinity (e.g. Richards, 1993). After the initial 2,000 iterations, the alignment is adjusted using the steric fields only.

For reasons outlined in Section 3.6, we use $\alpha = 31$ and $\beta = 0.04$ which worked well in pilot runs. Based on these pilot runs we also choose the penalty parameter value $\zeta = 3$. As standard deviations of the proposal distributions we use $\eta_1 = 3.25°$ for the rotation parameters and $\eta_2 = 0.5$Å for the translation parameters, and these values ensure acceptance rates between 20% and 40%. The standard deviation for the rotation parameters is thereby in line with previously described proposal distributions for rotation parameters in the molecular context (e.g. Green & Mardia, 2006). We define the initial relative position of the two molecules by first aligning both molecules along their principal axes. We then translate and rotate the random test molecule using $\boldsymbol{\gamma}_0$ and $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$ where $\gamma_{0i}\,(i = 1, 2, 3)$ and $\theta_{0i}\,(i = 1, 2, 3)$ are uniformly distributed on $[-5\text{Å}, 5\text{Å}]$ and $[-90°, 90°]$, respectively.

In the majority of the 930 cases, the algorithm converges quickly. However, like in the simulation study, the algorithm can sometimes get trapped in a local mode (which mostly corresponds to an alignment along the wrong principal axes in this application) so that a restart is necessary. We restart the algorithm if the sum of the 10% smallest distances between atoms of the test and reference molecule exceeds 400 Å after 1,500 iterations or if the mean of the Carbo distance values between iteration 3,000 and 4,000 exceeds 0.1. The latter can thereby be interpreted as a convergence criterion whereas the first is merely used as an early detector for an alignment along the wrong principal axes.

4.2.2 Example Run

Figure 4.1 shows an example result where aldosterone has successfully been superimposed onto androstanediol (cf. also Figure 1.1). The top row shows orthographic views of the initial relative position of the two molecules, and the relative position according to the

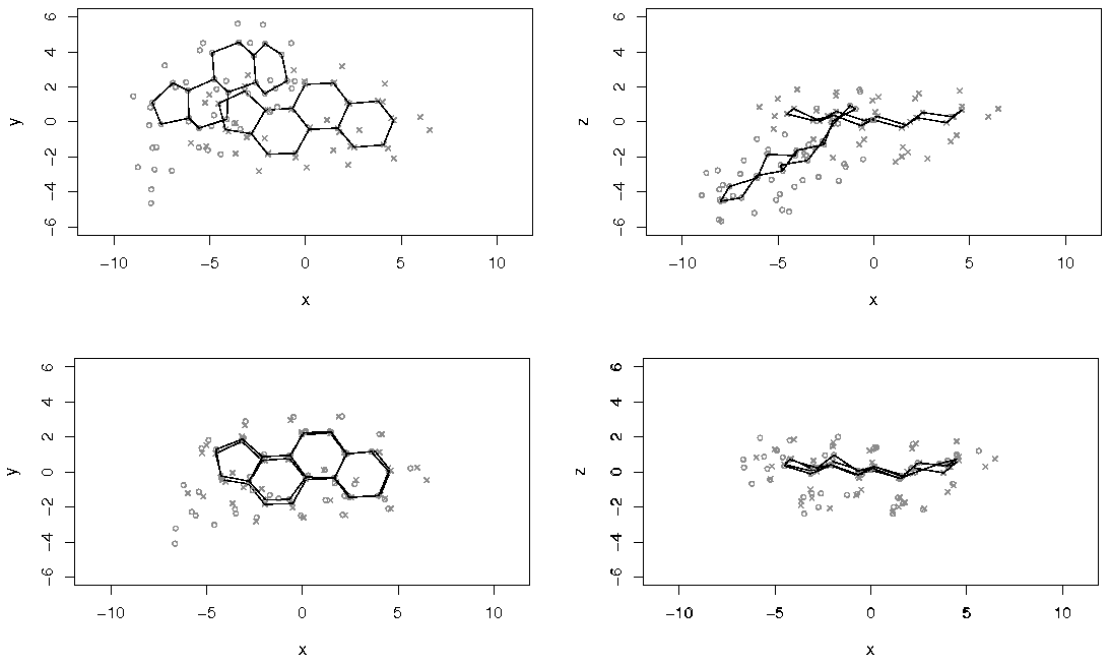

Figure 4.1: Successful alignment of two steroid molecules: Orthographic views of the starting position (top row) and the MAP position (bottom row) for the alignment of aldosterone and androstanediol are shown. The carbon rings are displayed as solid lines, and the remaining atoms are shown as circles (aldosterone) and crosses (androstanediol). The unit of all axes is Ångström (Å).

MAP estimates of the rigid–body parameters after a burn–in period of 3,500 iterations are displayed in the bottom row. The trace plots for this superposition are shown in Figures 4.2 and 4.3. The MCMC chain converges quickly and the trace plots show a similar behaviour as the corresponding plots obtained in the simulation study (cf. Figures 3.4 and 3.5). In this example, the acceptance rate for the rotation parameters is 37.95%, proposed translation vectors were accepted for 21.50% of the iterations, and the acceptance rates for the mask vectors $\boldsymbol{\lambda}_{\mathrm{A}}$ and $\boldsymbol{\lambda}_{\mathrm{B}}$ are 34.40% and 34.81%, respectively. Moreover, inserting the MAP estimates of the rigid–body parameters and the mask vectors into the Kernel Carbo discrepancy (3.14) yields $\hat{D}_{\mathrm{MAP}}(A, B) = 0.027$.

In order to obtain a similar value $\hat{D}_{\mathrm{mean}}(A, B)$ based on the estimates of the posterior mean values of the rigid–body and mask parameters, a threshold must be defined for the entries of the (post burn–in) mean mask vectors $\bar{\boldsymbol{\lambda}}_{\mathrm{A}}$ and $\bar{\boldsymbol{\lambda}}_{\mathrm{B}}$ which are displayed as big circles in the bottom row of Figure 4.2. Based on the observation that entries $\bar{\lambda}_i^{\mathrm{M}}$ $(M \in \{A, B\},\ i \in \{1, \ldots, k_{\mathrm{M}}\})$ below a threshold of $p_{\mathrm{crit}} = 0.7$ appear as outliers

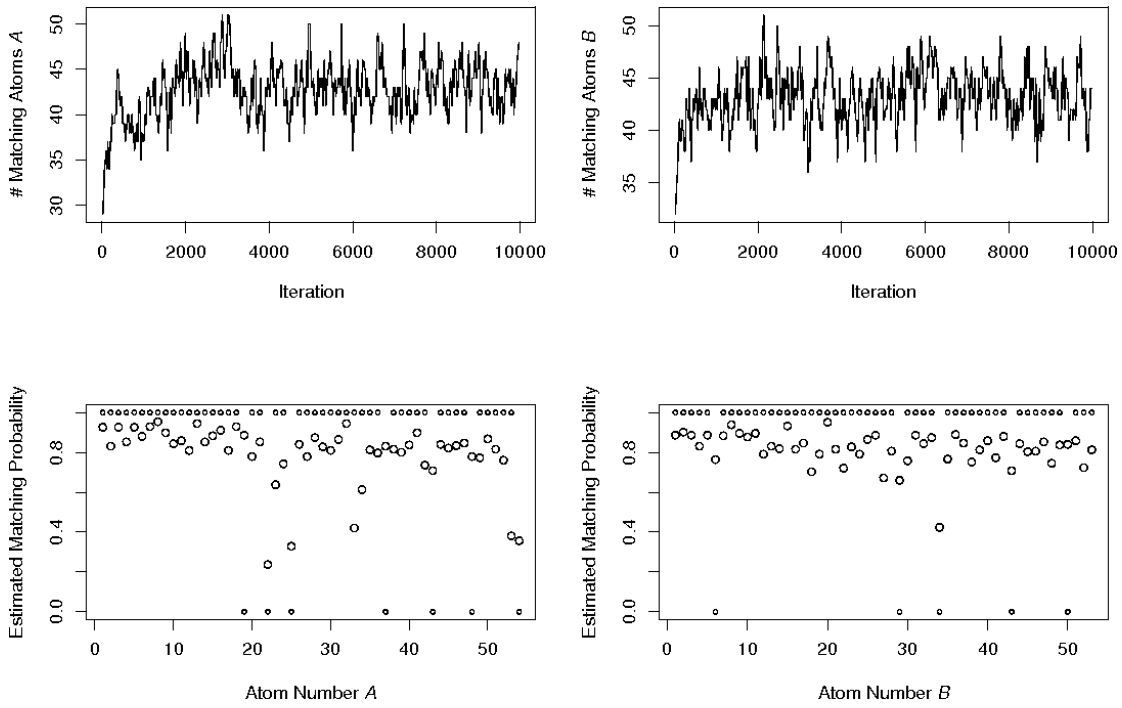

Figure 4.2: Trace plots and (post burn–in) posterior summary statistics of the mask vectors for the superposition of aldosterone and androstanediol: The top row shows the trace plots of the number of matching atoms in both molecules, and the bottom row shows the MAP (small circles) and the posterior mean (big circles) estimates of the corresponding mask vectors.

in most of the 930 performed superpositions, we set all entries below 0.7 to zero and all entries above 0.7 to one to obtain thresholded posterior mean estimates of the mask vectors which can then be inserted into (3.14). Doing so for the considered example yields $\hat{D}_{\text{mean}}(A, B) = 0.012$. From a decision theoretical point of view (cf. Appendix C), choosing a threshold of $p_{\text{crit}} = 0.7$ for the mean mask vectors thereby indicates that we consider a false inclusion of an atom as worse than a false exclusion which is readily justified by the fact that falsely including atoms can distort an alignment more severely than falsely omitting relevant atoms.

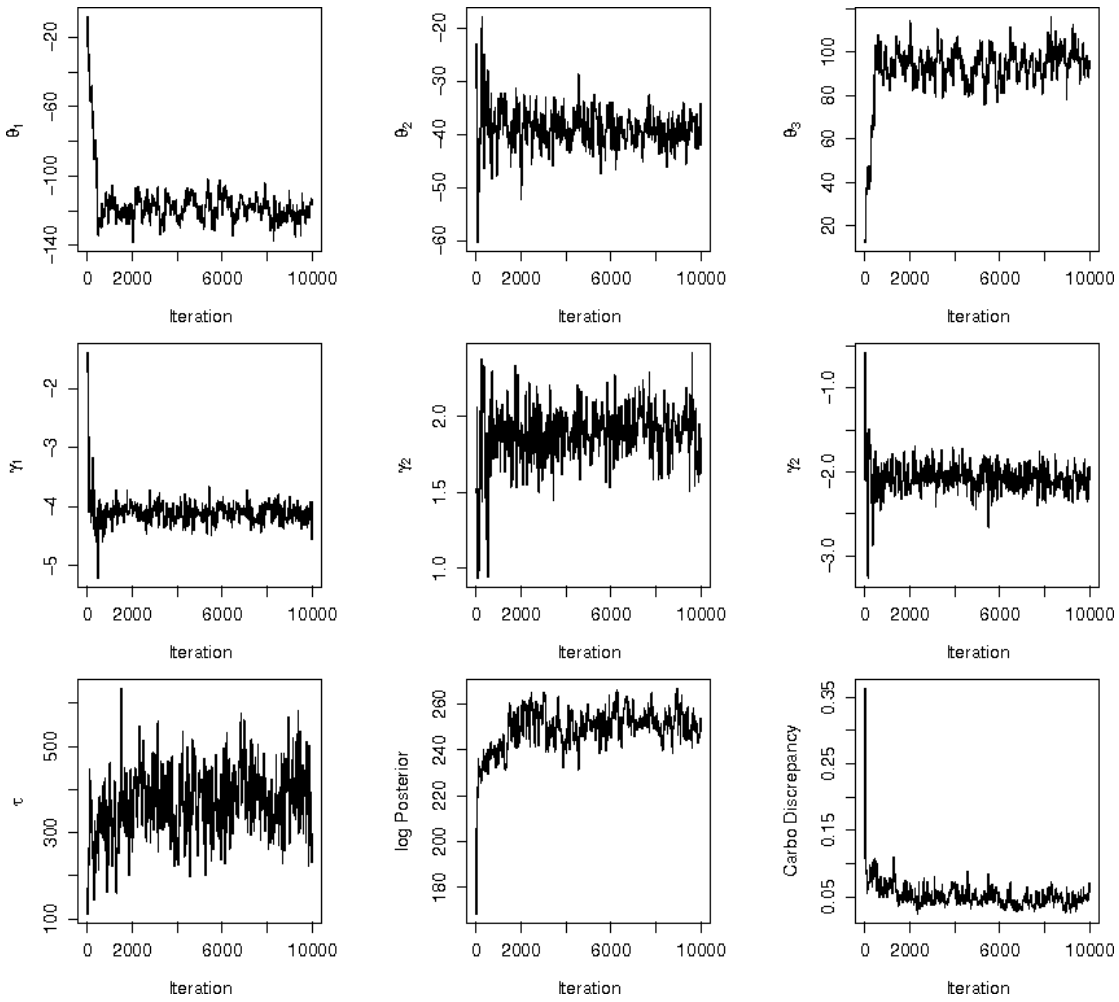

Figure 4.3: Trace Plots of the scalar parameters for the steroid application: The top two rows show the trace plots for the Euler angles and the translation parameters, respectively. The bottom row shows the plots for the precision parameter, the resulting log–posterior density values and the Kernel Carbo discrepancy. Like in Figure 3.4, the interplay between the discrepancy and the precision parameter is clearly visible.

Table 4.1: Prior sensitivity of the alignment of aldosterone and androstanediol: The impact of the penalty parameter (top part) and α (bottom part) on the marginal posterior distribution of the parameters of interest. The credibility intervals are based on every $20th$ value of MCMC period.

ζ	95% CI for τ	95% CI for $\sum_j \lambda_j^{\mathrm{A}}$	95% CI for $\sum_j \lambda_j^{\mathrm{B}}$
2	(226.62, 543.78)	(34, 46)	(34, 45)
3	(230.93, 543.30)	(37, 49)	(38, 48)
4	(250.69, 562.65)	(40, 51)	(40, 49)
5	(244.67, 548.41)	(41, 51)	(42, 51)
α	95% CI for τ	95% CI for $\sum_j \lambda_j^{\mathrm{A}}$	95% CI for $\sum_j \lambda_j^{\mathrm{B}}$
21	(102.53, 315.95)	(36, 48)	(37, 48)
31	(221.14, 515.13)	(38, 49)	(38, 49)
41	(344.68, 770.30)	(38, 48)	(39, 49)
51	(432.36, 1010.77)	(35, 48)	(37, 50)

4.2.3 Prior Sensitivity

To investigate the sensitivity of the alignment to the prior distributions, we again consider the alignment of aldosterone and androstanediol. The top part of Table 4.1 shows how different values of the penalty parameter ζ affect the empirical (post burn–in) 95% credibility intervals of the number of included atoms for both molecules; cf. Section 2.3.3. As expected, the total number of included atoms increases with ζ. As the two molecules in the example run are structurally very similar, they can be aligned more closely if more atoms are included so that the credibility interval for the precision parameter τ is shifted towards higher values as ζ increases. After a certain threshold, however, even larger values for the penalty parameter force the algorithm to include more atoms in the similarity calculations than desired and the precision decreases. Moreover, the bottom part of Table 4.1 shows that – in terms of the number of included atoms – the algorithm is robust against changes of α. Also, as the posterior mean and variance of the precision parameter directly depend on α, the credibility intervals for τ become wider and get shifted towards higher values as α increases.

We do not include a prior sensitivity analysis of β as decreasing β has the same effect on the algorithm as increasing α and *vice versa*. These contrary effects become clear from the

first and second posterior moments of the precision parameter τ, cf. (3.18). The bottom part of Table 4.1 therefore inherently covers a prior sensitivity analysis with respect to β. However, as mentioned in Section 3.5.3, it is vital that β is not substantially larger than discrepancy values which result from a good superposition because the interplay between the precision parameter and the discrepancy could not take place in such a case.

4.2.4 Chemical Relevance of the Results

As mentioned in Section 1.1.3, Good *et al.* (1993) classified each steroid according to its binding activity towards the CBG receptor as 1 (high), 2 (intermediate), or 3 (low). The pairwise distances which result from the 930 superpositions can therefore be regarded as chemically meaningful if they reflect the membership of the steroid molecules to the three activity classes, i.e. if steroids within an activity class can be aligned more closely than those from different activity classes. In terms of our assumption about a common underlying reference field, such a result would indicate that there are actually three different reference fields which exhibit different small scale variations and hence different abilities to fit into the protein binding pocket.

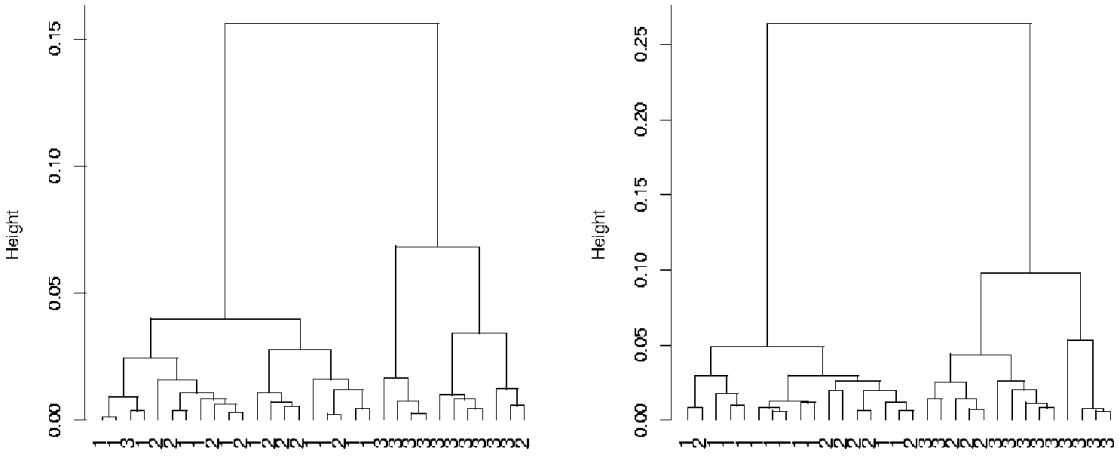

Figure 4.4: Dendrograms of the partial Kernel Carbo distances for the steroid molecules: The left–hand side dendrogram is based on $D_{\text{mean}}(.)$, and the dendrogram on the right–hand side is calculated using $D_{\text{MAP}}(.)$. The labels correspond to the activity classes of the steroids (1=high, 2=intermediate, 3=low).

We assess the chemical relevance of our results by performing two cluster analyses using Ward's (1963) method. To account for the asymmetry in our alignment method, the applied pairwise dissimilarity measures for two molecules A and B are thereby based on both the MCMC run which superimposes A on B and the MCMC run which superimposes B on A. In particular, we use

$$\tilde{D}_{\text{mean}}(A,B) = \sqrt{\hat{D}^{\text{mean}}_{A\to B}\,\hat{D}^{\text{mean}}_{B\to A}} \quad \text{and} \quad \tilde{D}_{\text{MAP}}(A,B) = \sqrt{\hat{D}^{\text{MAP}}_{A\to B}\,\hat{D}^{\text{MAP}}_{B\to A}},$$

where the arrow denotes the direction of the superposition and, as above, "mean" and "MAP" indicate which type of (post burn–in) point estimate for the parameters is inserted into the Carbo distance (3.14).

Figure 4.4 shows the dendrograms resulting from the cluster analyses. The graph on the left–hand side is based on $\tilde{D}_{\text{mean}}(.)$, and the right–hand side shows the dendrogram calculated using $\tilde{D}_{\text{MAP}}(.)$. The labels on both sides correspond to the activity classes of the steroid molecules. It is notable that both distance measures lead to a very good separation of high and low activity steroids. In particular, the cluster analysis based on $\tilde{D}_{\text{MAP}}(.)$ is at the highest level able to separate these two activity classes completely. Overall, our distance can separate the activity classes as well as the distance which Dryden *et al.* (2007) found to have the highest separation power, and it clearly outperforms the other distances defined in their paper.

4.3 Multiple Alignment of Unlabelled Marked Point Sets

The above dendrograms indicate that it is plausible to assume that there are at least two different reference fields underlying the steric properties of the steroid data. It is therefore of interest to determine these fields and examine where differences occur as they could give rise to the different binding activities. In this section, we therefore propose an extension of our pairwise alignment method for unlabelled marked points based on which the mean fields of the different activity classes can be determined.

In the multiple alignment problem, the objective is to simultaneously superimpose n unlabelled marked point sets $M_1, \dots M_n$ which are recorded in a certain position, i.e.

$$M_i = \{z^{\mathrm{M}_i}(\boldsymbol{\Gamma}_i \boldsymbol{x}_1^i + \boldsymbol{\gamma}_i), \dots, z^{\mathrm{M}_i}(\boldsymbol{\Gamma}_i \boldsymbol{x}_{k_i}^i + \boldsymbol{\gamma}_i)\}, \quad i = 1, \dots, n,$$

where $\boldsymbol{x}_l^i$ denotes the coordinate vector of the lth point in M_i ($l = 1, \dots, k_i$), $z^{\mathrm{M}_i}(\boldsymbol{x}_l^i)$ denotes the corresponding mark, and $\boldsymbol{\Gamma}_i \in SO(m)$ and $\boldsymbol{\gamma}_i \in \mathbb{R}^m$ define the position of M_i. Recall that $\boldsymbol{\Gamma}_i = \boldsymbol{\Gamma}(\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ is a $(m(m-1)/2)$–vector of Euler angles. Previous approaches to simultaneously aligning several unlabelled marked point sets include Dryden *et al.* (2007) and Ruffieux & Green (2009) which provide generalisations of the methods described in Section 3.2. Here, we adapt the generalised Procrustes analysis (GPA) algorithm for discrete landmark data (cf. Appendix A) to our field–based approach.

To do so, let $\boldsymbol{\lambda}_i \in \Lambda_{k_i}$ denote a fixed mask vector for the ith point set where, as before, Λ_{k_i} denotes the space of k_i–vectors with entries of either zero or one. Further suppose that simple kriging is performed using a positive definite covariance function $\sigma(.)$. The corresponding normalised predicted field then has the form

$$\tilde{Z}_i(\boldsymbol{x}; \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i) = \sum_{l:\lambda_l^i = 1} \tilde{w}_i^l(\boldsymbol{\lambda}_i) \sigma\big((\boldsymbol{\Gamma}_i \boldsymbol{x}_l^i + \boldsymbol{\gamma}_i) - \boldsymbol{x}\big)$$

where λ_l^i denotes the lth entry of $\boldsymbol{\lambda}_i$, and $\tilde{w}_i^l(\boldsymbol{\lambda}_i)$ denotes the corresponding normalised kriging weight which is defined using the norm $||.||_{\mathcal{H}_\sigma}$ as in (3.12).

In the classical GPA context, the aim is to find an alignment of the given objects (configuration matrices) which minimises the sum of their pairwise distances as measured by (2.8), and if the objects are commensurate in scale, then a partial GPA can be carried out where the scaling parameters are fixed to one. A similar goodness of fit criterion for the multiple superposition of n unlabelled marked point sets in terms of their predicted (masked) fields can be formulated as

$$C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \;=\; \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left\{ \sum_{l:\lambda_l^i = 1} \sum_{l':\lambda_{l'}^j = 1} \tilde{w}_i^l(\boldsymbol{\lambda}_i) \tilde{w}_j^{l'}(\boldsymbol{\lambda}_j)\, \sigma\big((\boldsymbol{\Gamma}_i \boldsymbol{x}_l^i + \boldsymbol{\gamma}_i) - (\boldsymbol{\Gamma}_j \boldsymbol{x}_{l'}^j + \boldsymbol{\gamma}_j)\big) \right\}$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} < \tilde{Z}_i(\boldsymbol{x}; \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i), \tilde{Z}_j(\boldsymbol{x}; \boldsymbol{\lambda}_j, \boldsymbol{\theta}_j, \boldsymbol{\gamma}_j) >_{\mathcal{H}_\sigma}, \tag{4.11}$$

where $\boldsymbol{\lambda}^T = (\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_n^T) \in \boldsymbol{\Lambda}_{\sum_i k_i}$, $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_n^T) \in \mathbb{R}^{m(m-1)n/2}$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_n^T) \in \mathbb{R}^{mn}$ denote the stacked vectors of the involved mask, rotation and translation parameters, respectively. For the multiple alignment of $M_1, \ldots, M_n$ we want to maximise (4.11) with respect to the $m(m-1)n/2 + mn + \sum_i k_i$ parameters.

From the bilinearity property of an inner product it follow that

$$\begin{aligned}
C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} < \tilde{Z}_i(\boldsymbol{x}; \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i), \tilde{Z}_j(\boldsymbol{x}; \boldsymbol{\lambda}_j, \boldsymbol{\theta}_j, \boldsymbol{\gamma}_j) >_{\mathcal{H}_\sigma} \\
&= \frac{1}{2} \sum_{i=1}^{n} \sum_{j \neq i} < \tilde{Z}_i(\boldsymbol{x}; \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i), \tilde{Z}_j(\boldsymbol{x}; \boldsymbol{\lambda}_j, \boldsymbol{\theta}_j, \boldsymbol{\gamma}_j) >_{\mathcal{H}_\sigma} \\
&= \frac{1}{2} \sum_{i=1}^{n} < \tilde{Z}_i(\boldsymbol{x}; \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i), \sum_{j \neq i} \tilde{Z}_j(\boldsymbol{x}; \boldsymbol{\lambda}_j, \boldsymbol{\theta}_j, \boldsymbol{\gamma}_j) >_{\mathcal{H}_\sigma} \\
&\propto \frac{1}{n} \sum_{i=1}^{n} < \tilde{Z}_i(\boldsymbol{x}; \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i), \tilde{Z}_{(i)}(\boldsymbol{x}; \boldsymbol{\lambda}_{(i)}, \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)}) >_{\mathcal{H}_\sigma},
\end{aligned}$$

where $\tilde{Z}_{(i)}(\boldsymbol{x}; \boldsymbol{\lambda}_{(i)}, \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)})$ is a "normalised mean field" of all but the ith point set, i.e.

$$\tilde{Z}_{(i)}(\boldsymbol{x}; \boldsymbol{\lambda}_{(i)}, \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)}) = \frac{1}{n-1} \sum_{j \neq i} \sum_{l: \lambda_l^j = 1} \tilde{w}_l^j(\boldsymbol{\lambda}_j) \sigma\big((\boldsymbol{\Gamma}_j \boldsymbol{x}_l^j + \boldsymbol{\gamma}_j) - \boldsymbol{x}\big),$$

where $\boldsymbol{\theta}_{(i)}^T = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_{i-1}^T, \boldsymbol{\theta}_{i+1}^T, \ldots \boldsymbol{\theta}_n^T)$, $\boldsymbol{\gamma}_{(i)}^T = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_{i-1}^T, \boldsymbol{\gamma}_{i+1}^T, \ldots \boldsymbol{\gamma}_n^T)$ and $\boldsymbol{\lambda}_{(i)}^T = (\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_{i-1}^T, \boldsymbol{\lambda}_{i+1}^T, \ldots \boldsymbol{\lambda}_n^T)$. It therefore follows that (4.11) can be decomposed as

$$C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \propto \frac{1}{n} \sum_{i=1}^{n} C_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i; \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)}, \boldsymbol{\lambda}_{(i)}),$$

where $C_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i; \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)}, \boldsymbol{\lambda}_{(i)})$ denotes the partial Kernel Carbo index between the normalised field $\tilde{Z}_i(\boldsymbol{x}; \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i)$ of M_i and $\tilde{Z}_{(i)}(\boldsymbol{x}; \boldsymbol{\lambda}_{(i)}, \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)})$. Due to this decomposition, the optimisation of the overall partial Kernel Carbo index $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ can therefore be carried out stepwise by maximising $C_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i; \boldsymbol{\theta}_{(i)}, \boldsymbol{\gamma}_{(i)}, \boldsymbol{\lambda}_{(i)})$ in turn. The vectors $\boldsymbol{\theta}_{(i)}$, $\boldsymbol{\gamma}_{(i)}$ and $\boldsymbol{\lambda}_{(i)}$ are thereby kept fixed at each step.

An optimisation of $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ is difficult so we replace it by posterior inference within the MCMC scheme developed for the pairwise alignment. As before, the choice of the prior distribution for the precision parameter τ determines how much the algorithm pushes the estimates of the other model parameters towards the posterior mode. An iterative stochastic optimisation of the normalised fields $\tilde{Z}_i(\boldsymbol{x})$ can therefore be formulated by employing a "large precision version" of the MCMC algorithm for the pairwise alignments and then using the obtained MAP estimates to determine a new mean field. This procedure will in practice decrease $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ at every step and can be repeated until a convergence criterion is met.

Algorithm 4.1 summarises our field GPA algorithm. As the objective of the multiple alignment is to find the features common to all or most of the objects, the algorithm superimposes each point set on the smallest (in terms of the number of points) one in the data set as a first step. Contrary to the pairwise alignment which started at a random place in the parameter space, this initialisation will be close to the global optimum which justifies the use of the large prior mean for the precision values.

4.4 Simultaneous Alignment of the Steroid Molecules

In this Section, we apply Algorithm 4.1 to the steroid molecules with the aim of obtaining the mean steric fields for each of the three activity groups. As a first step, the algorithm is applied to the entire set of the 31 steroids which is useful to determine the overall optimal relative position of the molecules. The pairwise superpositions carried out in step 1 are thereby performed as described before but with $\zeta = 2$ to incorporate the knowledge that the reference molecule in all superpositions has a small number of atoms. The superpositions on the mean fields (step 7) are obtained using only the discrepancies of the steric fields $\big($i.e. $w_{\mathrm{Q}} = 0$ in (4.10)$\big)$. As the initial molecular fields obtained in step 1 are good approximations of the fields which minimise the multiple Kernel Carbo index, we use $\alpha = 600$ and $\beta = 0.0001$ to ensure that the full conditional distribution of the precision parameter has a large mean value at each iteration, and we reduce the standard deviations of the proposal distributions for the rigid–body parameters to $\eta_1 = 0.75$ Å

Algorithm 4.1 Stochastic GPA for multiple unlabelled marked point sets

1: choose the smallest molecule as reference molecule and superimpose the $n-1$ remaining molecules onto it
2: define $d \leftarrow d_0$, where $d_0 > tol$ and tol is a positive tolerance threshold
3: calculate the multiple Carbo index $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$
4: **while** $d > tol$ **do**
5: **for** i in $(1:n)$ **do**
6: using the current parameter values for rotation, translation and mask vectors, calculate a normalised mean field $\tilde{Z}_{(i)}(\boldsymbol{x})$ omitting the ith molecule
7: based on the discrepancy $D_{(i)}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\lambda}_i)$, superimpose the ith molecular field onto $\tilde{Z}_{(i)}(\boldsymbol{x})$; $\tilde{Z}_{(i)}(\boldsymbol{x})$ thereby takes the role of the reference molecule and $\boldsymbol{\lambda}_{(i)}$, $\boldsymbol{\theta}_{(i)}$ and $\boldsymbol{\gamma}_{(i)}$ are treated as fixed
8: record the MAP estimates for position and mask of the ith molecule
9: **end for**
10: calculate the updated $C^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$
11: $d \leftarrow C^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) - C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$
12: $C(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \leftarrow C^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$
13: **end while**

and $\eta_2 = 0.03°$. Moreover, we set the number of iterations for each MCMC run in step 7 to 500, and the tolerance value to $tol = 0.0001$. The algorithm is therefore used as a stochastic optimiser.

The algorithm converges after the 9th field GPA iteration (cf. Figure F.1 in Appendix F). Figure 4.5 shows orthographic views of the resulting overlays. The superposition after step 1 of the field GPA algorithm is displayed in the top row, and the bottom row shows the final overlay. For clarity, the random starting positions of the steroids are not displayed in this picture. However, the top row of Figure 4.1 gives an indication of how far from the optimal overlay the algorithm started.

The relative positions obtained in the field GPA provide the best overall alignment of the 31 steroid molecules. To explore where the differences between the steric mean fields of the three activity groups are most pronounced, we perform the generalised field matching within each group separately to obtain mask vectors which reflect the steric properties common to all molecules within a group but with the features of the individual molecules removed. Using these mask vectors and the relative positions obtained in the overall field GPA, we then calculate the mean fields for each group. Figure 4.6 displays xy–cross–sections of the three mean fields for different values of z. Light points thereby

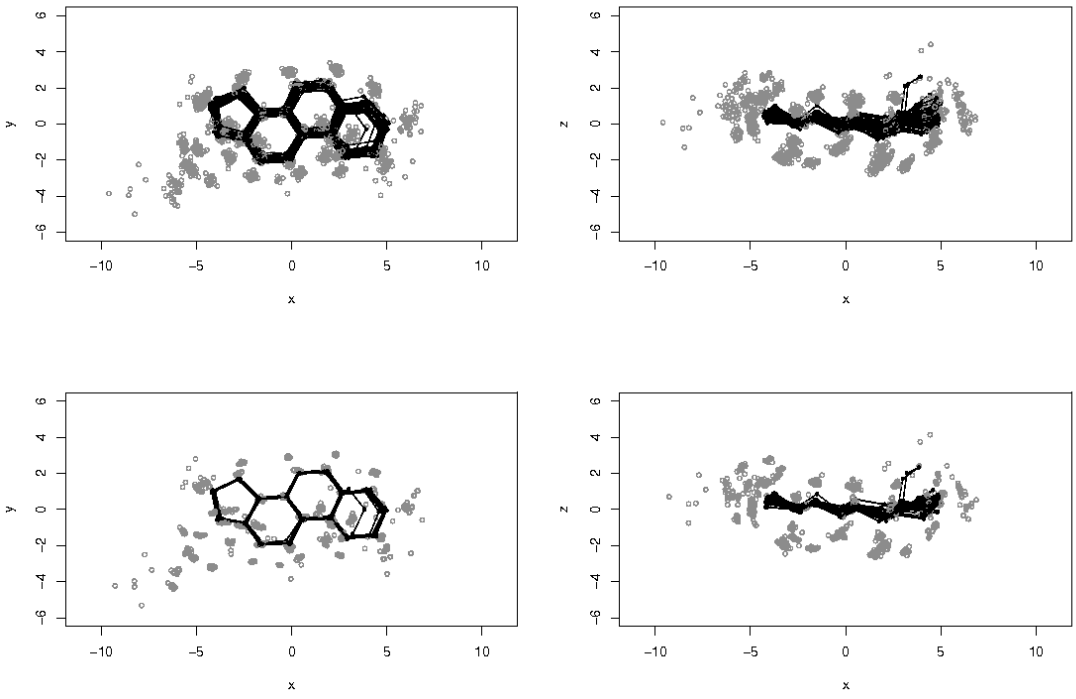

Figure 4.5: Overlay of the 31 steroid molecules obtained with the field GPA: Orthographic projections of the the relative position of the 31 steroid molecules after step 1 are shown in the top row. The bottom row shows orthographic projections of the final relative position.

correspond to locations where the displayed steric field takes a large value whereas dark points show field values close to zero.

Due to the fact that the common ring structure of the molecules is almost planar, the middle row ($z = 0$) essentially depicts the ring atoms of the mean fields and is similar for all three activity groups. At $z = 1.5$ and $z = -1.5$, however, differences occur and, as expected, the observed differences are most pronounced between the mean field of the high and low activity groups. To assess the differences for each pair (C_a, C_b) of activity classes $(a, b = 1, 2, 3;\ a \neq b)$ numerically, we consider a (two sample) t–field of the form

$$t_{ab}(\boldsymbol{x}) = \frac{\bar{Z}_a(\boldsymbol{x}) - \bar{Z}_b(\boldsymbol{x})}{s^*_{\text{pool}}(\boldsymbol{x})\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}, \quad \boldsymbol{x} \in \mathbb{R}^3, \tag{4.12}$$

where n_a and n_b denote the number of molecules in activity class C_a and C_b, respectively, $\bar{Z}_a(\boldsymbol{x})$ and $\bar{Z}_b(\boldsymbol{x})$ denote the corresponding mean fields, and $s^{*2}_{\text{pool}}(\boldsymbol{x}) = s^2_{\text{pool}}(\boldsymbol{x}) + d$, is the

pooled variance (with $d = 0.001$ a small offset to avoid spuriously large values in regions far away from the centre). For each pairwise comparison we define a three–dimensional grid G and calculate a t–value of the form (4.12) at a large number $(142, 598)$ of points. Here we use (4.12) as an exploratory tool rather than a formal test to see where the most pronounced differences occur.

Figure 4.7 shows the regions in which the (absolute) t–field for each comparison exceeds a threshold of eight. The main feature which distinguishes the high activity class from the other two classes is that the very active molecules commonly extend to the right of the ring structure much more than the other molecules. From the original data it can be seen that the associated atoms are oxygen and carbon atoms. Another interesting difference is located at the top left–hand side of the molecules where the low activity class differs from the other two classes in the location of oxygen atoms. These findings are

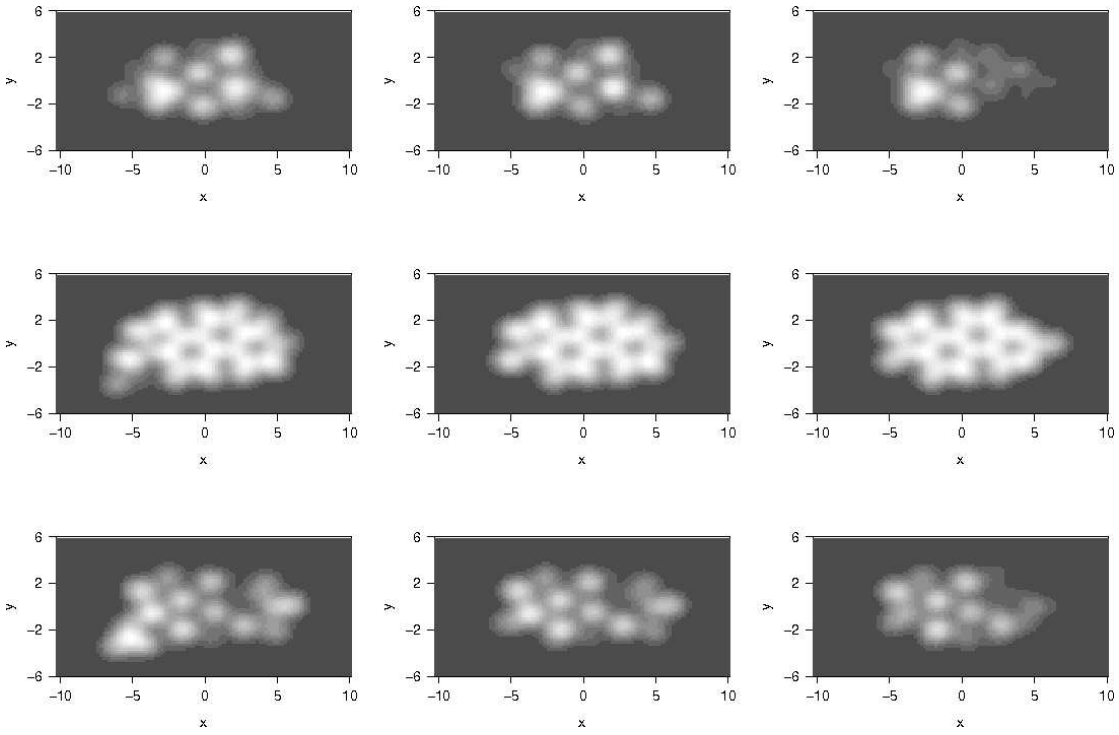

Figure 4.6: Cross–sections of the mean steric fields of the three activity groups: The left column shows the mean field of the high activity group, the middle column that of the medium activity group and the right column shows the mean field of the low activity group. The different rows display cross sections at $z = -1.5$ (top row), $z = 0$ (medium row), and $z = 1.5$ (bottom row). Light points correspond to locations with large values of the displayed field whereas dark values show points with values close to zero.

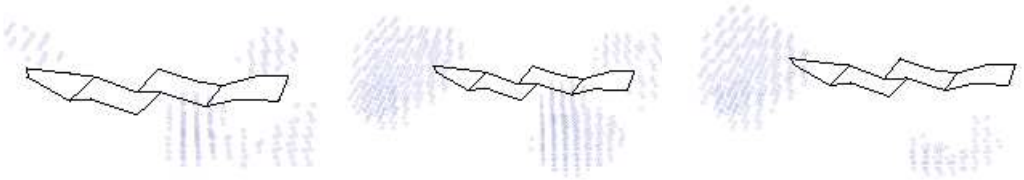

Figure 4.7: Thresholded t–fields resulting from pairwise comparisons of the steric mean fields of the three activity classes: Left–Hand Side: Low vs. Medium Activity Class, Middle: Low vs. High Activity Class, Right–Hand Side: Medium vs. High Activity Class. The shaded areas display regions where the t–field takes absolute values of larger than eight.

in line with Figure 4.6 and they are also supported by Figure 9 in Dryden *et al.* (2007) and support the conjecture that the steric properties of the steroid molecules have a discriminating effect with respect to the binding affinity towards the CBG receptor.

4.5 Summary

In this chapter, we pointed out similarities and differences of our field–based alignment method for two unlabelled marked point sets with previously proposed methods. We saw that it is related to a number of well–established structural alignment techniques but has the considerable advantage that mask vectors can be incorporated in the optimisation procedure which makes the alignment less susceptible to outliers.

We then applied the pairwise alignment to the steroid data set. A good superposition could be achieved for all pairs of steroids, and using a cluster analysis of the resulting partial Kernel Carbo distances, it could also be seen that the distances are chemically relevant in that they are associated with the different binding activities of the steroids.

In order to assess where the main steric differences between the activity classes can

be found, we proposed an extension of the pairwise field–based alignment to multiple unlabelled marked point sets. When applied to the steroid data, the mean steric fields of the activity classes could be obtained. Using a t–field as an exploratory tool, the regions around the molecular skeletons could then be identified where the steric properties differ the most between the activity classes.

Chapter 5

Fast Bootstrap Hypothesis Testing for Independent Configuration Matrices

In the previous two chapters we developed methods for the Bayesian alignment of unlabelled marked point sets which can be applied to the comparison of two or more (rigid) molecules. However, the methodology developed so far does not take into account that every molecule constantly undergoes vibrational motions and conformational changes. Information about the dynamic behaviour of molecules can be obtained using molecular dynamic (MD) simulations (cf. Section 1.2.1). In datasets obtained by MD simulations, each molecule is given in the form of a time series of its atom positions, i.e. an essential feature of such datasets is that a group of temporally dependent configuration matrices is provided for each object.

There are many other situations where the dataset at hand comprises two or more groups of configuration matrices, e.g. landmark data for human faces which can be divided into age groups (cf. Evison & Vorder Bruegge, 2008; Preston & Wood, 2009b) or data on skull shapes of male and female monkeys (e.g. Amaral *et al.*, 2007; Dryden & Mardia, 1998, Chapter 1). A frequent objective in such a situation is to investigate whether or not the underlying data generating processes of the two groups can be considered to be equal. One approach of tackling this problem is to employ a bootstrap hypothesis test where the null hypothesis is the equality of the underlying distributions. However, as location and scale of the given data are not of interest here, conventional multivariate bootstrap tests cannot be applied.

In this and the following chapter, we investigate the use of a fast bootstrap methodology which operates in the Procrustes tangent space of the combined data (cf. Section 2.1.4). Our approach is novel in that the Procrustes tangent space is only calculated once prior to the actual resampling – an approximation which facilitates a straightforward way of transforming the data to the null hypothesis (cf. Section 2.4.5) and reduces the computational cost of the resampling procedure.

We first focus on the situation where the configuration matrices in each group are assumed to be independent which will be the case in many applications. In Section 5.1, the underlying problem is stated in a general form and a literature review about the existing techniques is given. We describe our fast bootstrap algorithm in Section 5.2, and a simulation study is carried out to evaluate its performance. In Section 5.3, the method is applied to the dataset described in Amaral *et al.* (2007) where it is of interest to investigate sexual dimorphism in the mean shape of male and female chimpanzees. In this application, the independence assumption is met. In Section 5.4, we also apply the fast bootstrap algorithm to the DNA dataset described in Section 1.2.3. Here, the objective is to investigate whether or not distributional differences between damaged and undamaged duplexes can be found as this could explain why the damaged DNA molecules have a higher binding affinity towards the corresponding repair proteins. However, the independence assumption is not met for the DNA data so that methods which assume independence are not ideal. This is demonstrated in Section 5.5. Motivated by this problem, a more appropriate bootstrap test for molecular dynamics data is proposed in the next chapter. Section 5.6 concludes this chapter with a summary of the main results.

5.1 Hypothesis Tests in Shape Analysis – Literature Review

Consider the situation where the given data consist of two samples $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_\mathrm{X}}\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_\mathrm{Y}}\}$ of $(k \times m)$ configuration matrices, where $k \geq m + 1$. To test whether or not the underlying data generating processes can be considered as equal, various approaches have been proposed which require different levels of assumptions. An assumption common to all methods proposed so far, however, is that the groups are

independent from each other and that the configuration matrices within each group form a simple random sample. The employed shape models in landmark space can therefore be formulated as

$$\begin{aligned} \boldsymbol{X}_i &= \beta_i(\boldsymbol{\mu}_{\mathrm{X}} + \boldsymbol{E}_i)\boldsymbol{\Gamma}_i + \mathbf{1}_k\boldsymbol{\gamma}_i^T, \quad i = 1, \ldots, n_{\mathrm{X}} \\ \text{and} \qquad & \qquad (5.1) \\ \boldsymbol{Y}_j &= \beta_j(\boldsymbol{\mu}_{\mathrm{Y}} + \boldsymbol{E}_j)\boldsymbol{\Gamma}_j + \mathbf{1}_k\boldsymbol{\gamma}_j^T, \quad j = 1, \ldots, n_{\mathrm{Y}}, \end{aligned}$$

where the βs are positive scale factors, the $\boldsymbol{\Gamma}$s are rotation matrices and the $\boldsymbol{\gamma}$s are translation vectors. Moreover, $\boldsymbol{\mu}_{\mathrm{X}}$ and $\boldsymbol{\mu}_{\mathrm{Y}}$ are the population mean configurations, and the error matrices are assumed to be mutually independent and

$$\mathrm{vec}(\boldsymbol{E}_i) \overset{i.i.d}{\sim} F_{\mathrm{X}} \quad \text{and} \quad \mathrm{vec}(\boldsymbol{E}_j) \overset{i.i.d}{\sim} F_{\mathrm{y}}, \quad i = 1, \ldots, n_{\mathrm{X}}; j = 1, \ldots, n_{\mathrm{Y}}. \tag{5.2}$$

The km–variate error distributions F_{X} and F_{Y} are thereby unknown but assumed to have the zero–vector as their mean.

As described in Section 2.1.3, the mean configurations $\boldsymbol{\mu}_{\mathrm{X}}$ and $\boldsymbol{\mu}_{\mathrm{Y}}$ in combination with the error distributions induce certain distributions $Q_{[X]}$ and $Q_{[Y]}$ on Σ_m^k with population mean shapes $[\mu_{[X]}]$ and $[\mu_{[Y]}]$, respectively. The following paragraphs provide a short summary of two–sample hypothesis testing for the problem

$$H_0 : [\mu_{[X]}] = [\mu_{[Y]}] \quad \text{vs.} \quad H_1 : [\mu_{[X]}] \neq [\mu_{[Y]}]. \tag{5.3}$$

Again, note that the mean shapes $[\mu_{[X]}]$ and $[\mu_{[Y]}]$ of the distributions in Σ_m^k are not necessarily equal to the shapes of the mean configurations $\boldsymbol{\mu}_{\mathrm{X}}$ and $\boldsymbol{\mu}_{\mathrm{Y}}$.

5.1.1 Parametric Approaches

Most parametric approaches are based on the Procrustes tangent space approximation to shape space. To obtain a common tangent space for both groups, generalised Procrustes analysis (GPA, cf. Section 2.1.3) is performed on all $n_{\mathrm{X}} + n_{\mathrm{Y}}$ configuration matrices. The

pre–shape $\boldsymbol{Z}_{\hat{\mu}}$ of the minimising configuration $\hat{\boldsymbol{\mu}}$ of (2.9) is then taken as the pole and the tangent vectors for both groups are calculated using the projections (2.11) or (2.14). Both projections could be used here but in the following, we demonstrate the methods with the tangent vectors obtained from the inverse exponential map.

Let $\boldsymbol{v}_1^\dagger, \ldots, \boldsymbol{v}_{n_\mathrm{X}}^\dagger$ and $\boldsymbol{w}_1^\dagger, \ldots, \boldsymbol{w}_{n_\mathrm{Y}}^\dagger$ denote the resulting tangent vectors for the two groups. To test (5.3), multivariate normal models for the tangent vectors can be proposed, e.g.

$$\boldsymbol{v}_i^\dagger \sim \mathrm{N}(\boldsymbol{\xi}_\mathrm{X}, \boldsymbol{\Sigma}) \quad \text{and} \quad \boldsymbol{w}_j^\dagger \sim \mathrm{N}(\boldsymbol{\xi}_\mathrm{Y}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n_\mathrm{X}; j = 1, \ldots, n_\mathrm{Y}, \tag{5.4}$$

where the $\boldsymbol{v}_i^\dagger$ and $\boldsymbol{w}_j^\dagger$ are all mutually independent. As the null hypothesis (5.3) implies $\boldsymbol{\xi}_\mathrm{X} = \boldsymbol{\xi}_\mathrm{Y}$, the classical two–sample Hotelling's T^2–test (e.g. Mardia *et al.*, 1979, Chapter 3) can be carried out. For that, the Mahalanobis squared distance

$$D^2 = (\bar{\boldsymbol{v}}^\dagger - \bar{\boldsymbol{w}}^\dagger)^T \left(\frac{n_\mathrm{X} \boldsymbol{S}_v^\dagger + n_\mathrm{Y} \boldsymbol{S}_w^\dagger}{n_\mathrm{X} + n_\mathrm{Y} - 2} \right)^- (\bar{\boldsymbol{v}}^\dagger - \bar{\boldsymbol{w}}^\dagger), \tag{5.5}$$

is calculated where $\bar{\boldsymbol{v}}^\dagger = n_\mathrm{X}^{-1} \sum_{i=1}^{n_\mathrm{X}} \boldsymbol{v}_i^\dagger$, $\bar{\boldsymbol{w}}^\dagger = n_\mathrm{Y}^{-1} \sum_{i=1}^{n_\mathrm{Y}} \boldsymbol{w}_i^\dagger$, $\boldsymbol{S}_v^\dagger$ and $\boldsymbol{S}_w^\dagger$ are the maximum likelihood estimators of $\boldsymbol{\Sigma}$ based on the two groups (i.e. with divisors n_X and n_Y, respectively), and A^- denotes the generalised inverse of the matrix A. Assuming that $n_\mathrm{X} + n_\mathrm{Y} > M + 2$, where M is the dimension of the corresponding shape space $\big($cf. (2.3)$\big)$, it is well–known that under H_0

$$T^2 = \frac{n_\mathrm{X} n_\mathrm{Y} (n_\mathrm{X} + n_\mathrm{Y} - M - 1)}{(n_\mathrm{X} + n_\mathrm{Y})(n_\mathrm{X} + n_\mathrm{Y} - 2) M} \, D^2 \sim F_{M, n_\mathrm{X} + n_\mathrm{Y} - M - 1}, \tag{5.6}$$

where $F_{M, n_\mathrm{X} + n_\mathrm{Y} - M - 1}$ denotes an F–distribution with M and $n_\mathrm{X} + n_\mathrm{Y} - M - 1$ degrees of freedom. Here, H_0 is rejected for large values of T^2.

In case the assumption of equal covariance matrices in model (5.4) is questionable, i.e. in the (multivariate) Behrens–Fisher case, T^2 can be modified to

$$T_J^{\dagger 2} = (\bar{\boldsymbol{v}}^\dagger - \bar{\boldsymbol{w}}^\dagger)^T \left(\frac{1}{n_\mathrm{X}} \boldsymbol{S}_v^\dagger + \frac{1}{n_\mathrm{Y}} \boldsymbol{S}_w^\dagger \right)^- (\bar{\boldsymbol{v}}^\dagger - \bar{\boldsymbol{w}}^\dagger). \tag{5.7}$$

This statistic was proposed by James (1954) as a multivariate generalisation of the Welch test (Welch, 1947). The distribution of (5.7) is not easy to specify, but as $(\boldsymbol{S}_v^\dagger / n_\mathrm{X} +$

$\boldsymbol{S}_w^\dagger/n_\mathrm{Y}$) is a consistent estimator for the variance of $(\bar{\boldsymbol{v}}^\dagger - \bar{\boldsymbol{w}}^\dagger)$, it can be shown that its asymptotic distribution under H_0 is χ^2_M (e.g. Seber, 1984, p.115). Note that T^2 and T_J^2 are proportional for equal sample sizes n_X and n_Y. Also note that (5.7) can also be defined in terms of the unbiased estimators of the covariance matrices (i.e. with divisors $n_\mathrm{X} - 1$ and $n_\mathrm{Y} - 1$, respectively) which does not change its asymptotic distribution.

For situations in which the error distributions F_X and F_Y in (5.1) can be assumed to be isotropic normal, Goodall (1991) proposes a test for (5.3) based on

$$F = \frac{n_\mathrm{X} + n_\mathrm{Y} - 2}{n_\mathrm{X}^{-1} + n_\mathrm{Y}^{-1}} \, \frac{d_F^2(\hat{\boldsymbol{\mu}}_\mathrm{X}, \hat{\boldsymbol{\mu}}_\mathrm{Y})}{\sum_{i=1}^{n_\mathrm{X}} d_F^2(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_\mathrm{X}) + \sum_{j=1}^{n_\mathrm{Y}} d_F^2(\boldsymbol{Y}_j, \hat{\boldsymbol{\mu}}_\mathrm{Y})} \overset{\text{appr.}}{\sim} F_{M,(n_\mathrm{X}+n_\mathrm{Y}-2)M}. \quad (5.8)$$

The approximate F–distribution thereby results from approximate χ^2–distributions of the involved Procrustes distances and their approximate mutual independence.

5.1.2 Non–parametric Approaches

The above distributional assumptions can be relaxed if a bootstrap test is applied (cf. Section 2.4.5). As pointed out by Hall & Wilson (1991), the achieved type I error and power of a bootstrap test are more satisfactory if an (asymptotically) pivotal test statistic is used and resampling is performed under the null hypothesis. Unfortunately, adhering to either of these suggestions is not straightforward in the context of shape analysis due to the nuisance parameters of location and scaling. The papers by Amaral *et al.* (2007) and Preston & Wood (2009a,b) are concerned with this problem and investigate the use of bootstrap hypothesis testing in shape analysis for the planar shape and the higher dimensional reflection–shape case, respectively.

5.1.2.1 Planar Shape Data

In the two–dimensional case, the algebra for computing angles and distances between complex pre–shapes on the complex pre–shape sphere has a direct analogy to the cal-

culation of angles and distances between unsigned unit vectors in $\mathbb{R}^d$ (e.g. Dryden & Mardia, 1998, p.69). This analogy is exploited by Amaral *et al.* (2007). Based on the paper by Fisher *et al.* (1996), who propose pivotal methods for constructing a confidence region for the mean direction or the mean polar axis of one sample of directional or axial data, Amaral *et al.* (2007) transfer the methodologies to the two–dimensional shape case and extend them to multi–sample problems.

Consider the two–sample problem and let the sets $\boldsymbol{\mathcal{X}} = \{\boldsymbol{Z}_1^{\mathrm{X}}, \ldots, \boldsymbol{Z}_{n_{\mathrm{X}}}^{\mathrm{X}}\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Z}_1^{\mathrm{Y}}, \ldots, \boldsymbol{Z}_{n_{\mathrm{Y}}}^{\mathrm{Y}}\}$ denote two independent random samples of complex pre–shapes in d dimensions with population mean pre–shapes $\boldsymbol{m}_{\mathrm{X}}$ and $\boldsymbol{m}_{\mathrm{Y}}$, respectively. Further, let $\widehat{\boldsymbol{m}}_{\mathrm{X}}$ and $\widehat{\boldsymbol{m}}_{\mathrm{Y}}$ be the sample mean pre–shapes which can be obtained analytically in the planar case (Kent, 1994). Here, the null hypothesis in (5.3) can be written as $H_0 : \boldsymbol{m}_{\mathrm{X}} = e^{i\psi}\boldsymbol{m}_{\mathrm{Y}}$, where $\psi \in (0, 2\pi]$ denotes an arbitrary angle. Under H_0, both $\boldsymbol{m}_{\mathrm{X}}$ and $\boldsymbol{m}_{\mathrm{Y}}$ are therefore members of an equivalence class of the form $[m_0] = \{e^{i\theta}\,\boldsymbol{m}_0 : \theta \in (0, 2\pi]\}$. A rotation–invariant estimator of $\boldsymbol{m}_0$ can be obtained using

$$\begin{aligned} \widehat{\boldsymbol{m}}_0 &= \arg \min_{\boldsymbol{m}:||\boldsymbol{m}||=1} 2(n_{\mathrm{X}} + n_{\mathrm{Y}})\, \boldsymbol{m}^* \big(\widehat{\boldsymbol{M}}_{\mathrm{X}}^* \widehat{\boldsymbol{G}}_{\mathrm{X}}^{-1} \widehat{\boldsymbol{M}}_{\mathrm{X}} + \widehat{\boldsymbol{M}}_{\mathrm{Y}}^* \widehat{\boldsymbol{G}}_{\mathrm{Y}}^{-1} \widehat{\boldsymbol{M}}_{\mathrm{Y}}\big)\boldsymbol{m} \\ &= \arg \min_{\boldsymbol{m}:||\boldsymbol{m}||=1} T_0(\boldsymbol{m}), \end{aligned} \tag{5.9}$$

where $\widehat{\boldsymbol{M}}_{\mathrm{X}}$ and $\widehat{\boldsymbol{M}}_{\mathrm{Y}}$ are $\big((d-1) \times d\big)$ complex matrices which project onto the tangent space of the pre–shape sphere at $\widehat{\boldsymbol{m}}_{\mathrm{X}}$ and $\widehat{\boldsymbol{m}}_{\mathrm{Y}}$, respectively. Further, $\widehat{\boldsymbol{G}}_{\mathrm{X}}$ is a consistent estimator of the asymptotic covariance matrix of $n_{\mathrm{X}}^{1/2}\widehat{\boldsymbol{M}}_{\mathrm{X}}\boldsymbol{m}_0$ and $\widehat{\boldsymbol{G}}_{\mathrm{Y}}$ is defined accordingly. The pooled estimator for $\boldsymbol{m}_0$ is therefore given by the eigenvector of $\widehat{\boldsymbol{A}}_0 = (n_{\mathrm{X}} + n_{\mathrm{Y}})\big(\widehat{\boldsymbol{M}}_{\mathrm{X}}^* \widehat{\boldsymbol{G}}_{\mathrm{X}}^{-1} \widehat{\boldsymbol{M}}_{\mathrm{X}} + \widehat{\boldsymbol{M}}_{\mathrm{Y}}^* \widehat{\boldsymbol{G}}_{\mathrm{Y}}^{-1} \widehat{\boldsymbol{M}}_{\mathrm{Y}}\big)$ corresponding to its smallest eigenvalue, and an estimator of the common mean shape $[\boldsymbol{m}_0]$ is the equivalence class of pre–shapes $\{e^{i\psi}\,\widehat{\boldsymbol{m}}_0 : \psi \in (0, 2\pi]\}$.

Let $\lambda_{\min} = T_0(\widehat{\boldsymbol{m}}_0)$ denote the smallest eigenvalue of $\widehat{\boldsymbol{A}}_0$. Under some fairly mild conditions (e.g. existence of a well–defined common mean shape $\boldsymbol{m}_0$) it holds that

$$n_{\mathrm{X}}^{1/2}\widehat{\boldsymbol{M}}_{\mathrm{X}}\boldsymbol{m}_0 \xrightarrow{\mathcal{D}} \mathrm{CN}_{d-1}(\boldsymbol{0}, \boldsymbol{G}_{\mathrm{X}}) \quad \text{and} \quad n_{\mathrm{Y}}^{1/2}\widehat{\boldsymbol{M}}_{\mathrm{Y}}\boldsymbol{m}_0 \xrightarrow{\mathcal{D}} \mathrm{CN}_{d-1}(\boldsymbol{0}, \boldsymbol{G}_{\mathrm{Y}}),$$

where CN_{d-1} denotes the complex normal distribution in $d-1$ complex dimensions.

Consequently, $\lambda_{\min} \xrightarrow{\mathcal{D}} \chi^2_{2(d-1)}$ under the null hypothesis so that it can be used as an asymptotically pivotal test statistic. Amaral *et al.* (2007) also introduce a method to adhere to Hall & Wilson's (1991) second guideline. They propose transformations of the pre–shapes at hand which result in an equality of their sample mean pre–shapes while changing $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ as little as possible. Resampling from the transformed samples then satisfies the recommendation that resampling should be performed under H_0.

Using Monte Carlo simulations, Amaral *et al.* (2007) compare the performance of the $\lambda_{\min}$–test with those based on the test statistics (5.6), (5.13) and (5.8). In each case, both the bootstrap version of the test and the version based on the theoretical distribution of the test statistic are considered. Generally, the bootstrap versions outperform the tabular versions and overall, the bootstrap test based on $\lambda_{\min}$ yields the best results and is the recommended test.

5.1.2.2 Shape Data in Higher Dimensions

Motivated by the above results, Preston & Wood (2009a,b) investigate the use of bootstrap test procedures for shape data in $m \geq 3$ dimensions. In this case, estimating population mean shapes in the Procrustes framework requires an iterative algorithm (cf. Appendix A) which, combined with the computer–intensive nature of the bootstrap, can be very time–consuming. Preston & Wood (2009a,b) therefore make use of the computationally more appealing multidimensional scaling (MDS) approach to shape analysis (cf. Kent, 1994, and Dryden *et al.*, 2008), where inference is based on $\boldsymbol{Z}\boldsymbol{Z}^T$[1] and $\boldsymbol{Z}$ denotes a (Helmertised) pre–shape of dimension $(k-1) \times m$ (cf. Section 2.1.1).

The rationale behind the MDS approach is that $\boldsymbol{Z}\boldsymbol{Z}^T$ is invariant under rotation and reflection so that the relevant reflection size–and–shape space can be represented as

$$\mathcal{P}_m(k-1) = \{\boldsymbol{P} \in \mathcal{P}(k-1) \,|\, 1 \leq \operatorname{rank}(\boldsymbol{P}) \leq m \,\&\, \operatorname{tr}(\boldsymbol{P}) = 1\},$$

where $\mathcal{P}(k)$ is the space of $k \times k$ positive semi–definite symmetric matrices.

[1] Preston & Wood (2009a,b) work with transposed versions of the pre–shapes defined in this thesis.

The mean of a random pre–shape $\boldsymbol{Z}$ can be then defined using the spectral decomposition $\boldsymbol{\Xi} = \mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}^T) = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^T$ where $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_{k-1})$ and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1})$ (e.g. Mardia *et al.*, 1979, p.469), and the so–called mean ϕ–shape of $\boldsymbol{Z}$ is defined as

$$\phi(\boldsymbol{\Xi}) = \frac{1}{\delta_1 + \ldots + \delta_{k-1}} \sum_{i=1}^{m} \delta_i \boldsymbol{u}_i \boldsymbol{u}_i^T \in \mathcal{P}_m(k-1),$$

and it is unique if $\delta_m > \delta_{m+1}$. Its sample analogue for a sample $\boldsymbol{\mathcal{X}} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$ of pre–shapes is defined by $\phi(\widehat{\boldsymbol{\Xi}})$, where $\widehat{\boldsymbol{\Xi}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^T$.

In the one–sample case the null hypothesis of interest is $H_0 : \phi(\boldsymbol{\Xi}) = \boldsymbol{M}$, where $\boldsymbol{M} \in \mathcal{P}_m(k-1)$. To test H_0, Preston & Wood (2009a) propose two statistics of the form

$$\widetilde{T}(\boldsymbol{M}) = n \left(\widetilde{H}_{\widehat{\boldsymbol{M}}}(\boldsymbol{M})\right)^T \widehat{\boldsymbol{G}}^{-1}_{\boldsymbol{M},\widehat{\boldsymbol{M}}} \widetilde{H}_{\widehat{\boldsymbol{M}}}(\boldsymbol{M}), \tag{5.10}$$

where $\widetilde{H}_{\widehat{\boldsymbol{M}}}(.)$ denotes a function similar to (2.11) so that $\widetilde{H}_{\widehat{\boldsymbol{M}}}(\boldsymbol{M})$ is a vectorised version of the projection of the hypothesised matrix $\boldsymbol{M}$ onto the tangent space $\mathcal{T}_{\hat{\boldsymbol{M}}}(\mathcal{P}_m(k-1))$ of $\mathcal{P}_m(k-1)$ at $\widehat{\boldsymbol{M}} = \phi(\widehat{\boldsymbol{\Xi}})$, and $\widehat{\boldsymbol{G}}_{\boldsymbol{M},\widehat{\boldsymbol{M}}}$ denotes the asymptotic covariance matrix of $n^{1/2}\widetilde{H}_{\widehat{\boldsymbol{M}}}(\boldsymbol{M})$. For both versions of (5.10), it can be shown that $n^{1/2}\widetilde{H}_{\widehat{\boldsymbol{M}}}(\boldsymbol{M}) \xrightarrow{\mathcal{D}} \mathrm{N}(\boldsymbol{0}, \widehat{\boldsymbol{G}}_{\boldsymbol{M},\widehat{\boldsymbol{M}}})$ under H_0 (cf. also Dryden *et al.*, 2008). The proposed statistics are therefore asymptotically pivotal with a limiting χ^2–distribution and can be seen as direct generalisations of the $\lambda_{\min}$–statistic in the one–sample case (Amaral *et al.*, 2007).

In a second paper, Preston & Wood (2009b) also explore the use of bootstrap testing in the two–sample case. In the MDS–setting, the test problem (5.3) becomes

$$H_0 : \phi(\boldsymbol{\Xi}_{\mathrm{X}}) = \phi(\boldsymbol{\Xi}_{\mathrm{Y}}) \quad \text{vs.} \quad H_1 : \phi(\boldsymbol{\Xi}_{\mathrm{X}}) \neq \phi(\boldsymbol{\Xi}_{\mathrm{Y}}), \tag{5.11}$$

where $\phi(\boldsymbol{\Xi}_{\mathrm{X}})$ and $\phi(\boldsymbol{\Xi}_{\mathrm{Y}})$ denote the population mean ϕ–shapes of the samples $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$. In this case a direct generalisation of the $\lambda_{\min}$–statistic would have the form

$$\widetilde{T}_0(\widehat{\boldsymbol{M}}) = n_{\mathrm{X}}\widetilde{T}_{\mathrm{X}}(\widehat{\boldsymbol{M}}) + n_{\mathrm{Y}}\widetilde{T}_{\mathrm{Y}}(\widehat{\boldsymbol{M}}), \quad \text{where} \quad \widehat{\boldsymbol{M}} = \arg \min_{\boldsymbol{M} \in \mathcal{P}_m(k-1)} \widetilde{T}_0(\boldsymbol{M}),$$

and $\widetilde{T}_{\mathrm{X}}(.)$ and $\widetilde{T}_{\mathrm{Y}}(.)$ are defined as in (5.10) but using the sample mean ϕ–shapes of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ as the pole for the tangent projections.

Unfortunately, such a direct generalisation does not appear feasible. As complex algebra (where orthogonal transformations can be carried out as multiplications) cannot be used here, the corresponding optimisation problems become too complicated to be performed repeatedly within a bootstrap procedure. Preston & Wood (2009b) therefore explore the use of three computationally simpler test statistics which are defined in $\mathcal{T}_{\widehat{\boldsymbol{M}}_p}(\mathcal{P}_m(k-1))$, where $\widehat{\boldsymbol{M}}_p = \phi(\widehat{\boldsymbol{\Xi}}_p)$ and

$$\widehat{\boldsymbol{\Xi}}_p = \frac{1}{n_{\mathrm{X}} + n_{\mathrm{Y}}} \left\{ \left(\sum_{i=1}^{n_{\mathrm{X}}} \boldsymbol{Z}_i^{\mathrm{X}} \boldsymbol{Z}_i^{\mathrm{X}T} \right) + \left(\sum_{j=1}^{n_{\mathrm{Y}}} \boldsymbol{Z}_j^{\mathrm{Y}} \boldsymbol{Z}_j^{\mathrm{Y}T} \right) \right\}.$$

The test statistics they consider are the James statistic T_J as defined in (5.13), a regularised variant of (5.13), and a statistic based on the empirical likelihood by Owen (2001). All three test statistics are asymptotically pivotal under mild conditions with limiting distributions based on the χ^2–distribution and outperform their tabular version regarding their achieved significance value in simulations.

Regarding Hall & Wilson's (1991) second guideline, Preston & Wood (2009b) consider three methods of resampling under the null hypothesis. The first two of these approaches (resampling the centred tangent vectors and resampling the tangent vectors with appropriate resampling probabilities) involve fixing the tangent space to $\mathcal{T}_{\widehat{\boldsymbol{M}}_p}(\mathcal{P}_m(k-1))$ whereas the tangent space in the third approach is calculated anew for each resample. Like in Amaral *et al.* (2007), this last method of transforming to the null hypothesis changes the given data as little as possible. Preston & Wood (2009b) find that despite involving fewer approximations, the third method of resampling from the null hypothesis does not outperform the other ones and overall, centering the data in the fixed tangent space yields the best results on the grounds of both computational costs and accuracy.

5.2 Fast Bootstrap Test in Procrustes Tangent Space

The simulation studies by Amaral *et al.* (2007) (for planar shapes) and Preston & Wood (2009a,b) (using the MDS approach to shape analysis) indicate a superior performance

of bootstrap hypothesis tests over classical hypothesis tests. This leads to the question of whether bootstrap tests also perform well for $m \geq 3$ dimensions when the data are projected onto the Procrustes tangent space (cf. Section 2.1.4). This is important as Procrustes methods are more widely used in shape analysis than MDS methods. Especially when the reflection information of the given data should be retained, Procrustes methods are vital tools as any MDS–based calculation is inherently invariant under reflection.

To answer the above question, a Monte Carlo simulation study is carried out. We thereby focus on the case where the "observed tangent space" is fixed once it has been calculated based on the original data as this is much faster than carrying out a new GPA at every bootstrap iteration. Fixing the tangent space adds another level of approximation to the bootstrap procedure because it effectively reduces (5.3) to the standard multivariate problem of testing the equality of the mean vectors of two populations. If it can be demonstrated that this approximation works as well in the Procrustes as in the MDS setting, then the use of GPA in combination with bootstrap hypothesis testing becomes a practicable method for shape data in $m \geq 3$ dimensions.

5.2.1 Fast Bootstrap Algorithm

We now describe the fast bootstrap algorithm we propose in this thesis. The algorithm is designed for the data situation described at the beginning of Section 5.1, i.e. we are dealing with two samples $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_\mathrm{X}}\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_\mathrm{Y}}\}$ of independent configuration matrices, and we are interested in the test problem (5.3). A pseudo–code for our algorithm is given in Algorithm 5.1.

5.2.1.1 Remove the information about position and scale of data

GPA is carried out using the entire set $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_\mathrm{X}}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_\mathrm{Y}}\}$ of configuration matrices in order to obtain (an icon of) the sample Fréchet mean shape, $[\hat{\mu}_p]$ say, of the combined sample (cf. Section 2.1.3). The data are then projected into $\mathcal{H}_{\hat{\mu}_p}(S_m^k)$,

Algorithm 5.1 Fast bootstrap algorithm for testing the null hypothesis of equal mean shapes when the configuration matrices are independent

1: carry out GPA on the entire set of configuration matrices $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_\mathrm{X}}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_\mathrm{Y}}\}$
2: obtain the tangent vectors $\widetilde{\boldsymbol{v}}_1, \ldots, \widetilde{\boldsymbol{v}}_{n_\mathrm{X}}, \widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_{n_\mathrm{Y}}$ by projecting the optimally rotated, translated and scaled data onto the observed tangent space $\mathcal{H}_{\hat{\mu}_p}(S_m^k)$
3: eliminate the redundant dimensions and obtain tangent vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_\mathrm{X}}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n_\mathrm{Y}}$ using the the projections (5.12)
4: calculate the observed value $T^2_{J,\mathrm{obs}}$
5: transform to the null hypothesis by centering to give $\boldsymbol{\mathcal{X}}^c = \{\boldsymbol{v}_1^c, \ldots, \boldsymbol{v}_{n_\mathrm{X}}^c\}$ and $\boldsymbol{\mathcal{Y}}^c = \{\boldsymbol{w}_1^c, \ldots, \boldsymbol{w}_{n_\mathrm{Y}}^c\}$
6: **for** b in $(1:B)$ **do**
7: select random samples of size n_X and n_Y with replacement from $\boldsymbol{\mathcal{X}}^c$ and $\boldsymbol{\mathcal{Y}}^c$, respectively
8: calculate the bootstrap value $T^{2*}_{J,b}$ of the test statistic
9: **end for**
10: calculate the estimated p–value $\hat{p} = \big(1 + \sum_{b=1}^{B} I_{\{T^2_{J,\mathrm{obs}} > T^{2*}_{J,b}\}}\big)/(B+1)$

i.e. the horizontal subspace of the tangent space of the pre–shape sphere at a pre–shape which corresponds to $[\hat{\mu}_p]$. As described in Section 2.1.4, the resulting tangent vectors are invariant under location and scale of the original data. Here we use the inverse exponential map (2.14) which has the advantage that it preserves the natural, intrinsic distance of the corresponding shape space Σ_m^k between the pole and the observations.

This first step of the algorithm transforms the shape data into multivariate data in Euclidean space, and the bootstrap procedure we propose is formulated conditional on $\mathcal{H}_{\hat{\mu}_p}(S_m^k)$, i.e. the original configuration matrices are only used to obtain a suitable Euclidean approximation of the shape space. All of the following steps are carried out in $\mathcal{H}_{\hat{\mu}_p}(S_m^k)$ which has the advantage that the computer–intensive GPA only has to be carried out once in the course of the algorithm.

5.2.1.2 Eliminate the Redundant Dimensions

Let $\boldsymbol{v}_1^\dagger, \ldots, \boldsymbol{v}_{n_\mathrm{X}}^\dagger$ and $\boldsymbol{w}_1^\dagger, \ldots, \boldsymbol{w}_{n_\mathrm{Y}}^\dagger$ denote the tangent vectors of the configuration matrices of the first and second group, respectively, and let $\boldsymbol{D}^\dagger = \big(\boldsymbol{v}_1^\dagger, \ldots, \boldsymbol{v}_{n_\mathrm{X}}^\dagger, \boldsymbol{w}_1^\dagger, \ldots, \boldsymbol{w}_{n_\mathrm{Y}}^\dagger\big)^T$ denote the combined data matrix. Due to its invariance under location and scale of the original configuration matrices, the dimension of $\mathcal{H}_{\hat{\mu}_p}(S_m^k)$ is $M = k(m-1) - m(m-1)/2 - 1$ whereas the length of each tangent vector is $(k-1)m > M$.

To avoid problems with singularity in the course of the algorithm, the tangent vectors are projected into an appropriate M–dimensional subspace of $\mathbb{R}^{(k-1)m}$. The desired projection can thereby be determined based on the sample covariance matrix, $\boldsymbol{S}^\dagger = (n_\mathrm{X}+n_\mathrm{Y})^{-1}\boldsymbol{D}^{\dagger T}\boldsymbol{C}\boldsymbol{D}^\dagger$, where $\boldsymbol{C}$ denotes the centering matrix in $n_\mathrm{X}+n_\mathrm{Y}$ dimensions. As $\boldsymbol{S}^\dagger$ is singular, its spectral decomposition can be written as

$$\boldsymbol{S}^\dagger = \begin{pmatrix}\boldsymbol{P}_1 & \boldsymbol{P}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{\Lambda} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{P}_1^T\\ \boldsymbol{P}_2^T\end{pmatrix} = \boldsymbol{P}_1\boldsymbol{\Lambda}\boldsymbol{P}_1^T,$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1,\ldots,\lambda_M\}$ denotes the diagonal matrix of the non–zero eigenvalues of $\boldsymbol{S}^\dagger$, and the $\boldsymbol{0}$s denote matrices of zeros. The matrix of eigenvectors $\boldsymbol{P} = \begin{pmatrix}\boldsymbol{P}_1\,\boldsymbol{P}_2\end{pmatrix} \in O\big(k(m-1)\big)$ can therefore be decomposed into two matrices $\boldsymbol{P}_1 \in V\big((k-1)m, M\big)$ and $\boldsymbol{P}_2 \in V\big((k-1)m, (k-1)m-M\big)$, where $V(r,c)$ denotes the space of $(r\times c)$–matrices with the property $V(r,c) = \big\{\boldsymbol{A}\in\mathbb{R}^{r\times c} : \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_c\big\}$. The transformations

$$\boldsymbol{v}_i = \boldsymbol{P}_1^T\boldsymbol{v}_i^\dagger \quad\text{and}\quad \boldsymbol{w}_j = \boldsymbol{P}_1^T\boldsymbol{w}_j^\dagger \tag{5.12}$$

therefore result in a new $\big((n_\mathrm{X}+n_\mathrm{Y})\times M\big)$–dimensional data matrix $\boldsymbol{D} = \boldsymbol{D}^\dagger\boldsymbol{P}_1$ whose sample covariance matrix $\boldsymbol{P}_1^T\boldsymbol{S}^\dagger\boldsymbol{P}_1 = \boldsymbol{\Lambda}$ contains the entire variability of the original tangent vectors, i.e. the M–dimensional subspace of interest is spanned by the eigenvectors of $\boldsymbol{S}^\dagger$ associated with the non–zero eigenvalues, and the transformations (5.12) project the original tangent vectors into this space (cf. also Díaz-García *et al.*, 1997).

5.2.1.3 Choice of the Test Statistic

The tangent vectors $\boldsymbol{\mathcal{V}} = \{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{n_\mathrm{X}}\}$ and $\boldsymbol{\mathcal{W}} = \{\boldsymbol{w}_1,\ldots\boldsymbol{w}_{n_\mathrm{Y}}\}$ form our original sample and are used to calculate the observed value of the test statistic. Here, we choose the James statistic

$$T_J^2 = (\bar{\boldsymbol{v}}-\bar{\boldsymbol{w}})^T\left(\frac{1}{n_\mathrm{X}}\boldsymbol{S}_v + \frac{1}{n_\mathrm{Y}}\boldsymbol{S}_w\right)^{-1}(\bar{\boldsymbol{v}}-\bar{\boldsymbol{w}}), \tag{5.13}$$

where $\bar{\boldsymbol{v}} = n_\mathrm{X}^{-1}\sum_{i=1}^{n_\mathrm{X}}\boldsymbol{v}_i$, $\bar{\boldsymbol{w}} = n_\mathrm{Y}^{-1}\sum_{i=1}^{n_\mathrm{Y}}\boldsymbol{w}_i$, $\boldsymbol{S}_v$ and $\boldsymbol{S}_w$ are the sample covariance matrices of the groups (with divisors n_X and n_Y, respectively). This statistic is essentially the

same as (5.7), but after transformations (5.12) the generalised inverse can be replaced by the usual inverse. Assuming we have continuous data, the multivariate central limit theorem (e.g. Mardia *et al.*, 1979, p.51) holds, and (5.13) is asymptotically pivotal with a limiting χ^2_M–distribution for the same reasons as (5.7).

5.2.1.4 Transformation to the Null Hypothesis and Resampling

To investigate how extreme the observed value is under the null hypothesis, repeated resamples from both $\boldsymbol{\mathcal{V}} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_\mathrm{X}}\}$ and $\boldsymbol{\mathcal{W}} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n_\mathrm{Y}}\}$ are taken to approximate the distribution of (5.13) under H_0. Note that conditioning on the observed samples $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{W}}$ of tangent vectors implies conditioning on the observed samples $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ of configuration matrices *and* the observed tangent space $\mathcal{H}_{\hat{\mu}_p}(S^k_m)$. For our algorithm the chosen method of resampling under the null hypothesis is that of centering the (projected) tangent vectors. This is the natural choice for multivariate Euclidean data and the method which performed best in the MDS setting (Preston & Wood, 2009b). Resamples are therefore taken from $\boldsymbol{\mathcal{V}}^c = \{\boldsymbol{v}^c_1, \ldots, \boldsymbol{v}^c_{n_\mathrm{X}}\}$ and $\boldsymbol{\mathcal{W}}^c = \{\boldsymbol{w}^c_1, \ldots, \boldsymbol{w}^c_{n_\mathrm{Y}}\}$ where

$$\boldsymbol{v}^c_i = \boldsymbol{v}_i - \bar{\boldsymbol{v}} \quad (i = 1, \ldots, n_\mathrm{X}) \quad \text{and} \quad \boldsymbol{w}^c_j = \boldsymbol{w}_j - \bar{\boldsymbol{w}} \quad (j = 1, \ldots, n_\mathrm{Y}).$$

Each resample results in a bootstrap value $T^{2*}_{J,b}$ and as described in Section 2.4.5, a Monte Carlo estimate of the corresponding p–value can be calculated using

$$\hat{p} = \frac{\#\{T^{2*}_{J,b} > T^2_{J,\mathrm{obs}}\} + 1}{B + 1}, \tag{5.14}$$

where B denotes the number of Monte Carlo resamples.

5.2.2 Evaluation – A Monte Carlo Simulation Study

We carry out a Monte Carlo simulation study in which the bootstrap procedure described above is repeated a large number of times under the same conditions. The data are generated in landmark space. In particular, we use two multivariate normal models

to simulate independent 4×3 configuration matrices. The models thereby differ in the underlying dependence structure of the coordinates, i.e. in the (12×12)–dimensional covariance matrix $\widetilde{\boldsymbol{\Sigma}}_C$[1]. The first model assumes isotropy, i.e. we simulate from

Shape Model 1:

$$\mathrm{vec}(\boldsymbol{X}) \sim \mathrm{N}\big(\mathrm{vec}(\boldsymbol{\mu}), \sigma_c^2 \boldsymbol{I}\big), \tag{5.15}$$

where $\boldsymbol{\mu}$ denotes the mean configuration matrix. To define a non–isotropic model we use the factorisation $\widetilde{\boldsymbol{\Sigma}}_C = \boldsymbol{\Sigma}_m \otimes \boldsymbol{\Sigma}_k$ which allows us to model separately the variation identical at each landmark (summarised in $\boldsymbol{\Sigma}_m$) and the covariance between the landmarks (summarised in $\boldsymbol{\Sigma}_k$) (cf. Dryden & Mardia, 1998, p.167). Here we use

Shape Model 2:

$$\mathrm{vec}(\boldsymbol{X}) \sim \mathrm{N}\left(\mathrm{vec}(\boldsymbol{\mu}), \sigma_c^2 \begin{pmatrix} 1 & 1/4 & 1/4 \\ 1/4 & 1 & 1/4 \\ 1/4 & 1/4 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1/2 & 0 & 1/4 \\ 1/2 & 1 & 1/8 & 0 \\ 0 & 1/8 & 1 & 0 \\ 1/4 & 0 & 0 & 1 \end{pmatrix}\right). \tag{5.16}$$

5.2.2.1 Assessing the Performance of the Fast Bootstrap Tests and Problems

Let n_{sim} denote the number of Monte Carlo iterations, and let $\hat{p}_1, \ldots, \hat{p}_{n_{\mathrm{sim}}}$ denote the corresponding estimated p–values. As the theoretical p–value of a test is a random variable which follows a uniform distribution on $[0, 1]$ under H_0, the empirical distribution of the n_{sim} estimated p–values under H_0 is a good indicator for the performance of the test: if it is close to uniform, then the applied test statistic and the involved approximations are appropriate for the problem at hand. Of special interest is thereby the lower tail of this distribution as it has a direct impact on the achieved (empirical) significance value $\hat{\alpha} = \frac{1}{n_{\mathrm{sim}}} \sum_{i=1}^{n_{\mathrm{sim}}} I^{H_0}_{\{\hat{p}_i \leq \alpha\}}$ of the test for small (and hence typical) values of α. If $\hat{\alpha} \approx \alpha$, then the applied test is good in terms of its achieved significance level.

[1] The reason for this notation will become clear in Section 6.1.1.

Under the alternative, the distribution of the theoretical p–value is skewed to the right (e.g. Bhattacharya & Habtzghi, 2002). This should be reflected in the empirical distribution of $\hat{p}$ under H_1 so that the estimated power $\hat{\beta} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} I^{H_1}_{\{\hat{p}_i \leq \alpha\}}$ usually has a larger value than the specified α. The power thereby obviously depends on the degree of deviance from H_0, e.g. for test problem (5.3), the power depends on the distance (cf. Section 2.1.2) between the two population mean shapes, i.e. $\hat{\beta} = \hat{\beta}\left(\rho\big([\mu_{[X]}], [\mu_{[Y]}]\big)\right)$.

While the above facts provide straightforward guidelines for assessing the performance of a test when the data at hand are Euclidean, it is more difficult in the shape context due to the nuisance parameter of rotation, location and scale. In particular, for $m \geq 3$ dimensions it is difficult to control the distribution in shape space which is induced by a certain model in configuration space $\big($cf. also Section 2.1.3 and the comment after (5.3)$\big)$. Figure 5.1 demonstrates this statement. Here, shape model 2 is used to generate $50 \times 1,000$ configuration matrices for each standard deviation $\sigma_c \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and each time the mean configuration $\boldsymbol{\mu}$ is taken to be the icon of the regular tetrahedron which is denoted as $\breve{\boldsymbol{X}}_0$ in (5.17). For each set of 1,000 configuration matrices, GPA is carried out so that we have 50 icons $\hat{\boldsymbol{\mu}}_1^{\sigma_c}, \ldots, \hat{\boldsymbol{\mu}}_{50}^{\sigma_c}$ per standard deviation whose shapes

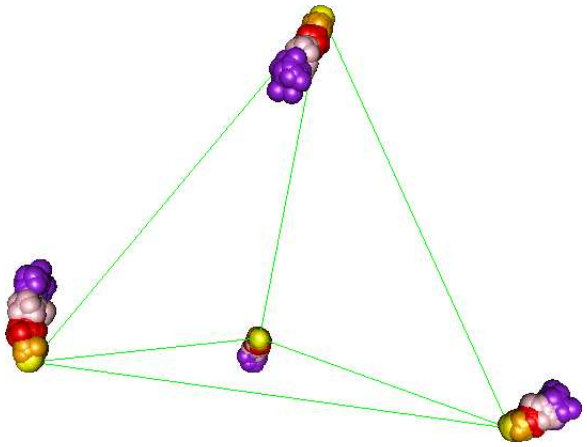

Figure 5.1: Impact of the standard deviation in shape model 2 on the mean shape: The displayed configurations are optimally rotated, translated and scaled icons of estimated mean shapes. Each estimate has been calculated using GPA on 1,000 configuration matrices which were generated from shape model 2 with the same mean configuration (displayed in green) but with different standard deviations. The icons are colour–coded corresponding to the employed standard deviation $\big($$\sigma_c = 0.1$ (yellow), $\sigma_c = 0.2$ (orange), $\sigma_c = 0.3$ (red), $\sigma_c = 0.4$ (pink), and $\sigma_c = 0.5$ (purple)$\big)$.

estimate the corresponding population mean shape $[\mu_{[X]}^{\sigma_c}]$. On these icons, a new GPA is carried out. Figure 5.1 shows the resulting optimally rotated, translated and scaled configurations. The colours correspond to the associated standard deviation $\big(\sigma_c = 0.1$ (yellow), $\sigma_c = 0.2$ (orange), $\sigma_c = 0.3$ (red), $\sigma_c = 0.4$ (pink), and $\sigma_c = 0.5$ (purple)$\big)$. A clear trend is visible which indicates that $[\mu_{[X]}^{\sigma_c}]$ depends on the standard deviation.

The above shows that it is difficult to simulate configuration matrices $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_X}\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_Y}\}$ in a way that the corresponding shapes satisfy the null hypothesis in (5.3) but exhibit different dependence structures. When assessing the empirical significance level of the fast bootstrap test, we will therefore concentrate on the case where all configuration matrices are simulated using the same model $\big($either (5.15) or (5.16)$\big)$ in landmark space. Despite this difficulty, we use the Riemannian distance between $[\mu_X]$ and $[\mu_Y]$ to assess the effect of $\rho([\mu_{[X]}], [\mu_{[Y]}])$ on the power of the test. This is reasonable as $\rho([\mu_X], [\mu_Y]) \approx \rho([\mu_{[X]}], [\mu_{[Y]}])$, and when assessing the power, it is not essential to know the exact distance between the population mean shapes so that $\rho([\mu_X], [\mu_Y])$ can provide valuable information about the underlying deviance from the null hypothesis.

For empirical power calculations, we therefore want to choose the mean configurations $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ in a way that their shapes exhibit a certain Riemannian distance. As described in Section 2.1.5, this can be achieved using a geodesic of the form (2.13): starting at a pre–shape $\boldsymbol{Z}_{\mu_X}$ associated with $\boldsymbol{\mu}_X$, another pre–shape $\boldsymbol{Z}_{\mu_Y}$ can be generated whose shape $[\mu_Y]$ exhibits a certain Riemannian distance s from $[\mu_X]$. If the mean configuration of the second group $\boldsymbol{\mu}_Y$ is chosen to be an icon of $[\mu_Y]$, then $\rho([\mu_X], [\mu_Y]) = s$ as desired.

5.2.2.2 Simulated Data

In our simulation study, the mean configurations $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ we consider for the shape models (5.15) and (5.16) are icons from the geodesic path which connects the shape of the regular tetrahedron $\big($the configuration matrix of a corresponding icon is given by $\breve{\boldsymbol{X}}_0$ in (5.17)$\big)$ with the shape of the configuration $\breve{\boldsymbol{Y}}$ which results from moving the first

landmark in $\breve{\boldsymbol{X}}_0$ to the position $(1, 1, -1)$, i.e. the considered geodesic is defined as the shortest path in shape space which connects the shapes of

$$\breve{\boldsymbol{X}}_0 = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \quad \text{and} \quad \breve{\boldsymbol{Y}} = \begin{pmatrix} 1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix}. \tag{5.17}$$

Figure 5.2 visualises this geodesic in terms of optimally rotated, translated and scaled icons. The green configuration shows the regular tetrahedron, the black sequence displays icons along the geodesic and the blue points correspond to configurations $\breve{\boldsymbol{X}}_1, \breve{\boldsymbol{X}}_2, \breve{\boldsymbol{X}}_3$ whose shapes exhibit distances $\rho_1 = \pi/16$, $\rho_2 = \pi/8$ and $\rho_3 = \pi/4$ from $[\breve{X}_0]$.

The mean configuration of the first group is kept fixed at $\boldsymbol{\mu}_{\mathrm{X}} = \breve{\boldsymbol{X}}_0$ in all cases. To assess the achieved significance level and power, the configurations $\breve{\boldsymbol{X}}_0, \ldots, \breve{\boldsymbol{X}}_3$ are used in turn as mean configuration $\boldsymbol{\mu}_{\mathrm{Y}}$ for the second group. Different values of σ_c and different sample sizes are considered for both shape models, namely $\sigma_c \in \{0.1, 0.2, 0.3, 0.5\}$ and $n_{\mathrm{X}}, n_{\mathrm{Y}} \in \{20, 50, 100\}$. In all cases, the number of bootstrap iterations is fixed at $B = 200$ and each scenario is repeated in $n_{\mathrm{sim}} = 2,500$ Monte Carlo iterations.

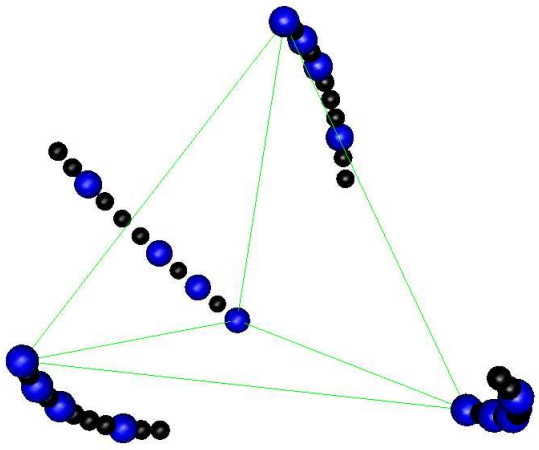

Figure 5.2: Geodesic between $\breve{\boldsymbol{X}}_0$ and $\breve{\boldsymbol{Y}}$: The green configuration is the regular tetrahedron which is taken as the starting point of the geodesic, the black configurations are icons along the path and the blue configurations correspond to shapes with Riemannian distances of $\rho_1 = \pi/16$, $\rho_2 = \pi/8$ and $\rho_3 = \pi/4$ from the regular tetrahedron.

5.2.2.3 Results for Shape Model 1

Consider shape model (5.15). We first simulate configurations $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_\mathrm{X}}\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_\mathrm{Y}}\}$ whose shapes satisfy the null hypothesis of our test problem (5.3), i.e. we use $\boldsymbol{\mu}_\mathrm{X} = \boldsymbol{\mu}_\mathrm{Y} = \breve{\boldsymbol{X}}_0$ and the same value for the standard deviation σ_c for both groups (cf. Section 5.2.2.1). Moreover, we generate the same number of matrices for each group, i.e. $n_\mathrm{X} = n_\mathrm{Y} = n$ and we consider all combinations of standard deviations $\sigma_c \in \{0.1, 0.2, 0.3, 0.5\}$ and sample sizes $n \in \{20, 50, 100\}$. In all cases the empirical distribution of the estimated p–values follows the uniform distribution closely. Figure 5.3 illustrates this observation for the challenging case $\sigma_c = 0.5$.

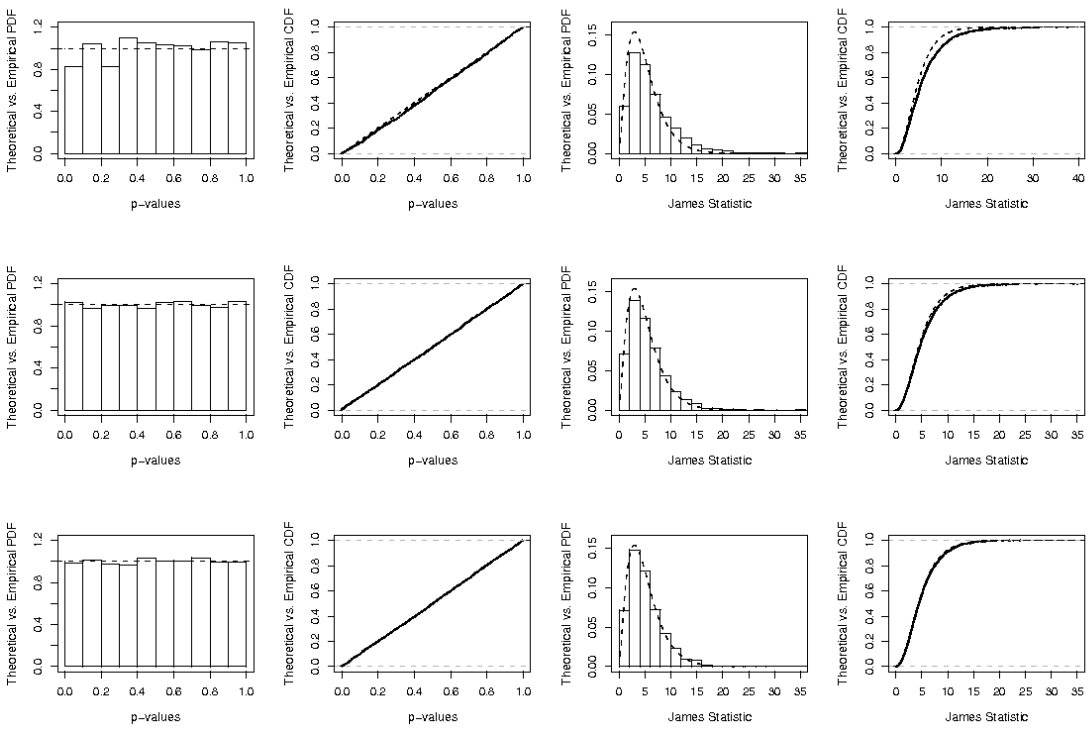

Figure 5.3: Null distribution of the n_sim estimated p–values and the observed values of the James statistic for data simulated according to (5.15): The three rows correspond to the sample sizes (top: $n_\mathrm{X} = n_\mathrm{Y} = 20$, middle: $n_\mathrm{X} = n_\mathrm{Y} = 50$, bottom: $n_\mathrm{X} = n_\mathrm{Y} = 100$), and $\sigma_c = 0.5$ for all cases. The first and second column show histograms and empirical distribution functions of the estimated p–values. In columns three and four histograms and the empirical distribution functions of the test statistics are displayed. The null distribution of the estimated p–values is very close to uniform (dashed line in columns one and two) and the empirical distribution of the James statistic approaches its asymptotic χ^2_5–distribution (dashed lines in columns three and four) as the sample size grows.

The rows in Figure 5.3 correspond to the different sample sizes (top: $n_X = n_Y = 20$, middle: $n_X = n_Y = 50$, bottom: $n_X = n_Y = 100$). It can be seen that even for a large value of $\sigma_c = 0.5$, the distribution of the estimated p–values (solid line) is close to that of the uniform distribution (dashed line) for all sample sizes. The approximations inherent in the considered test procedure therefore do not seem to have a negative effect on the achieved significance value. The empirical distribution of the James statistic (solid line) is also displayed. As the dimension of the shape space for (4×3)–dimensional configuration matrices is $M = 5$, its limiting distribution is the χ^2_5–distribution here. It can be seen the the empirical distribution of the test statistic approaches this limiting distribution (dashed line) as the sample size grows.

In Table 5.1, we compare the empirical significance levels $\hat{\alpha}$ of the fast bootstrap test with the empirical significance levels $\hat{\alpha}_{\text{tab}}$ of the tabular test (where the estimated p–values are calculated based on the quantiles of the χ^2_5–distribution) for different nominal significance levels α. It can be seen that $\hat{\alpha}$ is close to the nominal value in most cases whereas $\hat{\alpha}_{\text{tab}}$ dramatically exceeds the nominal level for small n. In those cases, the tabular test is too liberal and tends to detect spurious differences between the population mean shapes $[\mu_{[X]}]$ and $[\mu_{[Y]}]$. The standard deviation does not seem to have a big impact on the achieved significance levels of both tests. At least for the small sample size $n = 20$, this is quite surprising as $\sigma_c = 0.5$ in combination with a mean configuration of $\breve{\boldsymbol{X}}_0$ entails a very large amount of variability in the generated data.

We now concentrate on a nominal significance value of $\alpha = 0.05$. The left–hand side of Table 5.2 shows the resulting achieved significance level and power for all considered combinations of sample sizes and standard deviations. As described in Section 5.2.2.2, the mean configuration of the first groups is thereby kept fixed at $\boldsymbol{\mu}_X = \breve{\boldsymbol{X}}_0$ whereas the mean configuration of the second group varies according to the geodesic path displayed in Figure 5.2, i.e. $\boldsymbol{\mu}_Y \in \{\breve{\boldsymbol{X}}_0, \ldots, \breve{\boldsymbol{X}}_3\}$, and the columns in Table 5.2 correspond to the resulting Riemannian distances $\rho(\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y) = \rho([\breve{X}_0], [\breve{X}_i])$ $(i = 0, \ldots, 3)$. It can be seen that the power of the test is very good. In most cases it increases quickly with the true distance of the mean shapes of the two populations, and the rate of this increase depends on the sample size and the standard deviation: the power increases faster with large sample sizes and small standard deviations.

Table 5.1: Comparison of the bootstrap and the tabular significance levels for shape model 1: The achieved significance level $\hat{\alpha}$ of the fast bootstrap test is close to the nominal level α in all cases whereas the achieved significance level of the tabular test $\hat{\alpha}_{\text{tab}}$ can dramatically exceed the nominal level.

		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
n	σ_c	$\hat{\alpha}$	$\hat{\alpha}_{\text{tab}}$	$\hat{\alpha}$	$\hat{\alpha}_{\text{tab}}$	$\hat{\alpha}$	$\hat{\alpha}_{\text{tab}}$
20	0.1	0.01	0.05	0.046	0.135	0.098	0.206
	0.2	0.009	0.045	0.038	0.112	0.08	0.185
	0.3	0.009	0.046	0.038	0.118	0.078	0.185
	0.5	0.01	0.046	0.041	0.123	0.083	0.192
50	0.1	0.012	0.018	0.049	0.073	0.101	0.136
	0.2	0.014	0.02	0.052	0.066	0.089	0.118
	0.3	0.014	0.02	0.052	0.07	0.092	0.13
	0.5	0.018	0.022	0.057	0.076	0.103	0.135
100	0.1	0.013	0.015	0.057	0.063	0.11	0.13
	0.2	0.013	0.012	0.047	0.058	0.096	0.105
	0.3	0.014	0.016	0.056	0.067	0.105	0.122
	0.5	0.014	0.012	0.052	0.064	0.099	0.112

Table 5.2: Achieved significance level and power for a nominal significance value of $\alpha = 0.05$ based on configurations generated using shape model 1 (left–hand side) and shape model 2 (right–hand side): In most cases the power increases quickly with the deviance from the null hypothesis. However, when the sample size is small relative to the standard deviation, the fast bootstrap test becomes less powerful.

		Shape Model 1 $\hat{\alpha}$ and $\hat{\beta}$				**Shape Model 2** $\hat{\alpha}$ and $\hat{\beta}$			
n	σ_c	0	$\pi/16$	$\pi/8$	$\pi/4$	0	$\pi/16$	$\pi/8$	$\pi/4$
20	0.1	0.046	0.995	1	1	0.034	1	1	1
	0.2	0.038	0.448	0.984	1	0.035	0.626	0.999	1
	0.3	0.038	0.162	0.6	0.965	0.028	0.245	0.826	1
	0.5	0.041	0.06	0.122	0.22	0.039	0.072	0.185	0.417
50	0.1	0.049	1	1	1	0.05	1	1	1
	0.2	0.052	0.949	1	1	0.045	0.995	1	1
	0.3	0.052	0.519	0.99	1	0.05	0.705	0.999	1
	0.5	0.057	0.114	0.352	0.658	0.049	0.175	0.565	0.908
100	0.1	0.057	1	1	1	0.046	1	1	1
	0.2	0.047	1	1	1	0.051	1	1	1
	0.3	0.056	0.859	1	1	0.047	0.975	1	1
	0.5	0.052	0.224	0.677	0.948	0.055	0.353	0.899	0.999

5.2.2.4 Results for Shape Model 2

To assess the performance of Algorithm 5.1 for the case of a non–isotropic dispersion structure, all above calculations are repeated with configuration matrices simulated according to (5.16). The results are very similar to the isotropic case. Figure 5.4 is the equivalent to Figure 5.3 for shape model 2 (i.e. for $\sigma_c = 0.5$). Again, the empirical distribution of the estimated p–values is close to uniform in all cases so that our fast bootstrap test should perform very well in terms of its achieved significance level. Also, the empirical distribution of the James statistic approaches its limiting distribution as the sample size grows. The corresponding figures for smaller values of σ_c show similarly good results. A summary of the performance of Algorithm 5.1 in terms of its achieved significance value can be found Table 5.3.

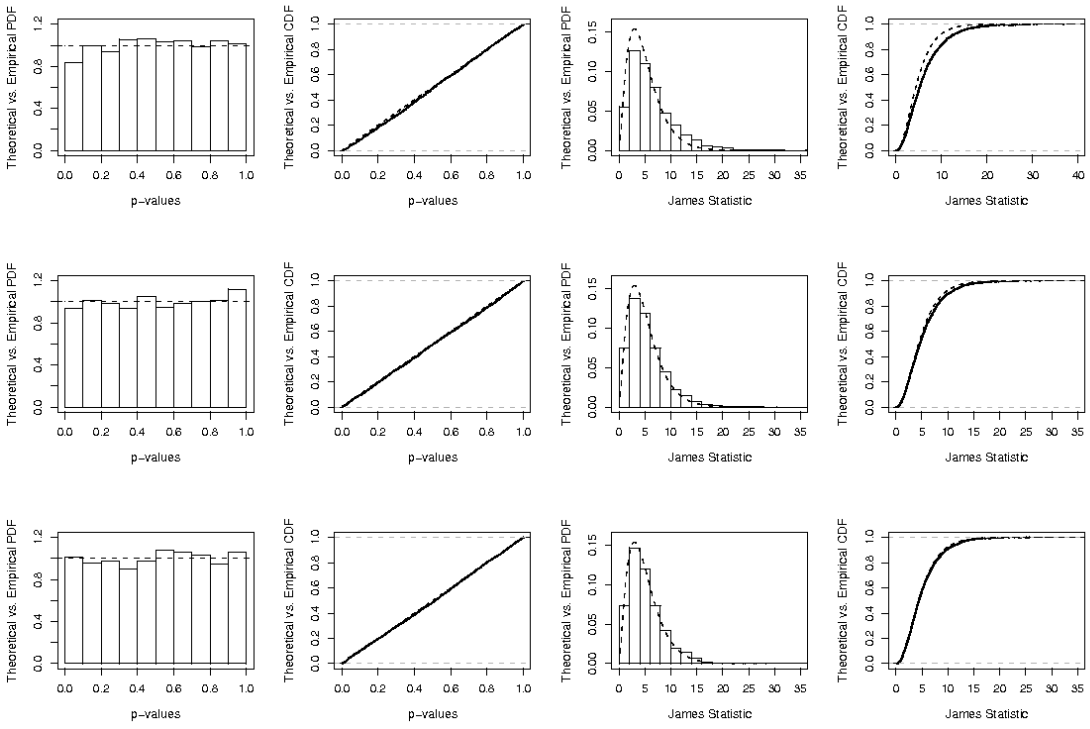

Figure 5.4: Null distribution of the n_{sim} estimated p–values and the observed values of the James statistic for data simulated according to shape model 2: The three rows correspond to the sample sizes (top: $n_X = n_Y = 20$, middle: $n_X = n_Y = 50$, bottom: $n_X = n_Y = 100$), and $\sigma_c = 0.5$ for all cases. Like in the isotropic case, the null distribution of the estimated p–values is very close to uniform and the empirical distribution of the James statistic approaches its asymptotic χ^2_5–distribution as the sample size grows.

Table 5.3: Comparison of the bootstrap and the tabular significance levels for shape model 2: Like in the isotropic case (cf. Table 5.1), the fast bootstrap outperforms the tabular test in the majority of cases as the tabular test tends to be too liberal.

		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
		$\hat{\alpha}$	$\hat{\alpha}_{\text{tab}}$	$\hat{\alpha}$	$\hat{\alpha}_{\text{tab}}$	$\hat{\alpha}$	$\hat{\alpha}_{\text{tab}}$
20	0.1	0.006	0.043	0.034	0.128	0.083	0.198
	0.2	0.004	0.044	0.036	0.12	0.078	0.188
	0.3	0.004	0.039	0.029	0.117	0.075	0.175
	0.5	0.005	0.043	0.039	0.122	0.083	0.195
50	0.1	0.008	0.023	0.05	0.077	0.101	0.132
	0.2	0.008	0.018	0.046	0.074	0.092	0.132
	0.3	0.01	0.021	0.05	0.08	0.101	0.147
	0.5	0.01	0.019	0.049	0.077	0.094	0.129
100	0.1	0.01	0.014	0.046	0.056	0.093	0.113
	0.2	0.008	0.012	0.051	0.065	0.097	0.119
	0.3	0.01	0.014	0.047	0.054	0.094	0.111
	0.5	0.008	0.014	0.055	0.064	0.101	0.119

Table 5.3 shows that Algorithm 5.1 performs very well. Only for $n = 20$ is the achieved significance level systematically too small. The small quantiles of the empirical distribution of $\hat{p}$ in that case do not, therefore, accurately represent the corresponding quantiles of the uniform distribution (cf. also the top row of Figure 5.4). However, the deviance of the achieved significance level from the nominal level for the tabular test is larger in all cases so that Algorithm 5.1 clearly outperforms the tabular test.

Finally, the right–hand side of Table 5.2 shows the achieved significance level and power for the case $\boldsymbol{\mu}_{\text{X}} = \breve{\boldsymbol{X}}_0$ and $\boldsymbol{\mu}_{\text{Y}} \in \{\breve{\boldsymbol{X}}_0, \ldots, \breve{\boldsymbol{X}}_3\}$ in shape model 2. It can be seen that the bootstrap test performs very well in terms of its power, unless the standard deviation is large relative to the sample size. Compared to the left–hand side of Table 5.2, the bootstrap test in the non–isotropic case is slightly more conservative.

5.2.2.5 Speed Comparison

We use the name "fast bootstrap" for the bootstrap algorithm described in this chapter because Algorithm 5.1 avoids determining a new tangent space at each bootstrap iter-

ation. Instead, the given data are projected only once into the observed tangent space $\mathcal{H}_{\hat{\mu}_p}(S_m^k)$, and resamples are taken from the resulting (centred) tangent vectors. As it deals with the nuisance parameters of rotation, translation and scale prior to resampling, this procedure effectively transforms the shape problem into a multivariate test problem. In the above simulation study we show that the inherent approximations do not seem to have a negative effect on the achieved significance level and power of the test, Here, we will quantify the gain in speed over the version of Algorithm 5.1 which takes resamples from the original configuration matrices $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_\mathrm{X}}\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_\mathrm{Y}}\}$ and calculates a new tangent space $\mathcal{H}_{\hat{\mu}^*_{p_b}}(S_m^k)$ with pole

$$\hat{\boldsymbol{\mu}}^*_{p_b} = \arg \min_{\boldsymbol{\mu}: S(\boldsymbol{\mu})=1} \left\{ \sum_{i=1}^{n_\mathrm{X}} \sin^2 \rho(\boldsymbol{X}^*_{i_b}, \boldsymbol{\mu}) + \sum_{j=1}^{n_\mathrm{Y}} \sin^2 \rho(\boldsymbol{Y}^*_{j_b}, \boldsymbol{\mu}) \right\}$$

at each iteration. As described in Section 2.1.3, determining $\hat{\boldsymbol{\mu}}^*_{p_b}$ requires the use of an iterative algorithm for $m \geq 3$ dimensions so that we expect a considerable increase in computational cost compared to Algorithm 5.1.

We first compare the speed for the case where each group consists of 20 configuration matrices. As before, we use 200 bootstrap iterations to obtain an estimated p–value. We run both version of the algorithm 100 times on a high performance GRID computer. The average running time of Algorithm 5.1 is 0.84 seconds with a standard deviation (sd) of 0.02 seconds. The slow version takes on average 110.08 seconds (sd: 0.54 seconds) to complete. This effect is amplified if the sample size in each group is increased to 100. In that case, the running time of Algorithm 5.1 is still fast with an average of 2.81 seconds (sd: 0.03 seconds) whereas the average running time of the slow version is 566.35 seconds (sd: 26.01 seconds). Although speed comparisons like this obviously depend on the exact implementation and can vary between programmers, the gain in speed achieved by using Algorithm 5.1 is substantial.

Another advantage of Algorithm 5.1 over its slow counterpart is that fixing the tangent space and centering the resulting tangent vectors presents a natural and successful way to adhere to Hall & Wilson's (1991) second guideline. For the slow bootstrap version, the original configuration matrices would have to be transformed to the null hypothesis.

Due to the non–homogeneity of the shape space for $m \geq 3$ dimensions, it is not clear how that can be done. The benefit of the fast bootstrap algorithm proposed in this section is therefore twofold.

5.3 Application to the Skull Data

Algorithm 5.1 is designed for the situation where the data follow the general shape model (5.1), i.e. it assumes independence of the given objects. Before we apply it to the DNA, we first consider an application where this assumption is met. In particular, we consider the dataset analysed in Amaral *et al.* (2007) (cf. also O'Higgins & Dryden, 1993) which contains landmark data of skull shapes for male and female chimpanzees. The objective in this application is to examine whether or not male and female chimpanzees have different mean skull shapes. The dataset comprises $k = 8$ landmarks in $m = 2$ dimensions in the midline of the cranium of 28 male and 26 female chimpanzees. Figure 5.5 shows the landmarks for both sexes which were registered using partial GPA (i.e. involving rotation and translation only).

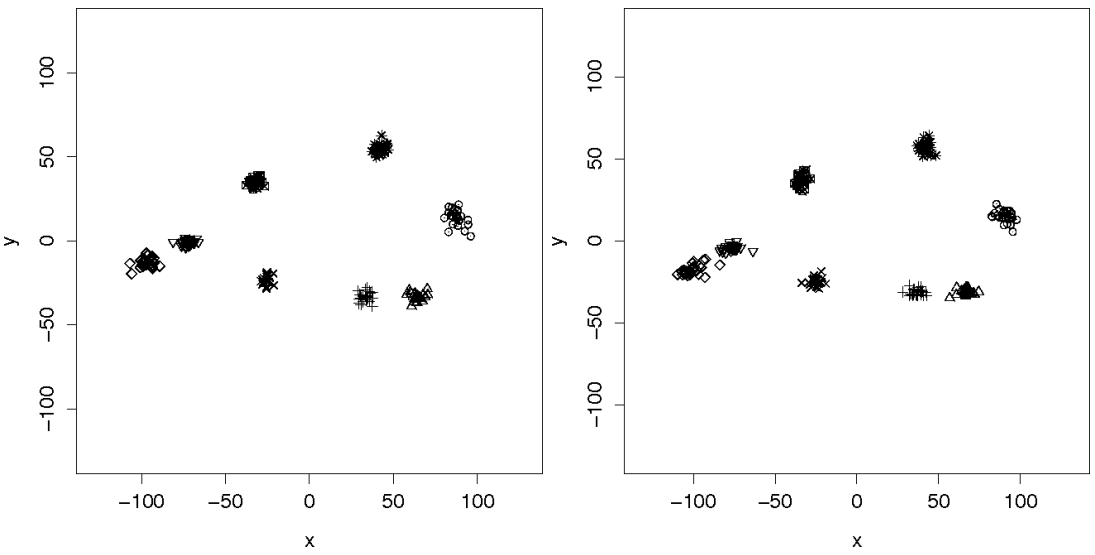

Figure 5.5: Landmark data for skulls of female and male chimpanzees: The configurations within both groups were registered using the partial GPA algorithm. The left–hand side shows the landmarks of the skulls for the 26 female apes, and the right–hand side shows the landmarks of the skulls for the 28 male apes.

To investigate the specificity of the test, we first randomly partition each group into two equally sized subsamples and apply Algorithm 5.1 to the subsamples within the groups (using B=1,000 bootstrap iterations). Note that the null distribution of the James statistic is the χ^2_{12}–distribution for this application. The resulting estimated p–values are $\hat{p} = 0.886$ ($T^2_{J,\text{obs}} = 13.552$) for the male chimpanzees and $\hat{p} = 0.704$ ($T^2_{J,\text{obs}} = 26.954$) for the female chimpanzees. Both tests therefore correctly find no evidence against the null hypothesis of equal mean shapes.

We then use Algorithm 5.1 to compare the full set of configuration matrices of the two groups (again using $B = 1,000$). The resulting estimated p–value is $\hat{p} = 0.245$ with an observed value of the James statistic of $T^2_{J,\text{obs}} = 24.356$. As expected, the estimated p-value is smaller than the ones obtained when configurations within either group are compared, but no evidence can be found which supports the conjecture that the mean shapes of the skulls for male and female chimpanzees are significantly different. These results are in line with those in Amaral *et al.* (2007) where $\hat{p} = 0.227$. However, note that the observed value of the James statistic in their paper is slightly different ($\tilde{T}^2_{J,\text{obs}} = 23.456$) as it was obtained using the version of the James statistic which is based on the unbiased estimators of the covariance matrices (cf. Section 5.1.1).

5.4 Application to the DNA Data

We now consider the DNA data described in Section 1.2.3 where the question of interest is whether or not the guanine lesion FapydG (F) induces a significant change in the shape of a DNA duplex when it is compared with its undamaged counterpart. Here, we restrict our attention to potential differences between the mean shapes. Figure 5.6 shows the sample mean shapes of the twelve DNA duplexes in terms of pairwise optimally rotated, translated and scaled icons. For each pair, the grey configuration shows the mean shape of the undamaged molecule, and the black configuration shows the mean shape of its damaged version. Differences can be seen, and in order to assess whether or not these differences are above the noise level we apply Algorithm 5.1 to each of the pairs, i.e. the groups in this application consist of the 2,500 configurations for each duplex.

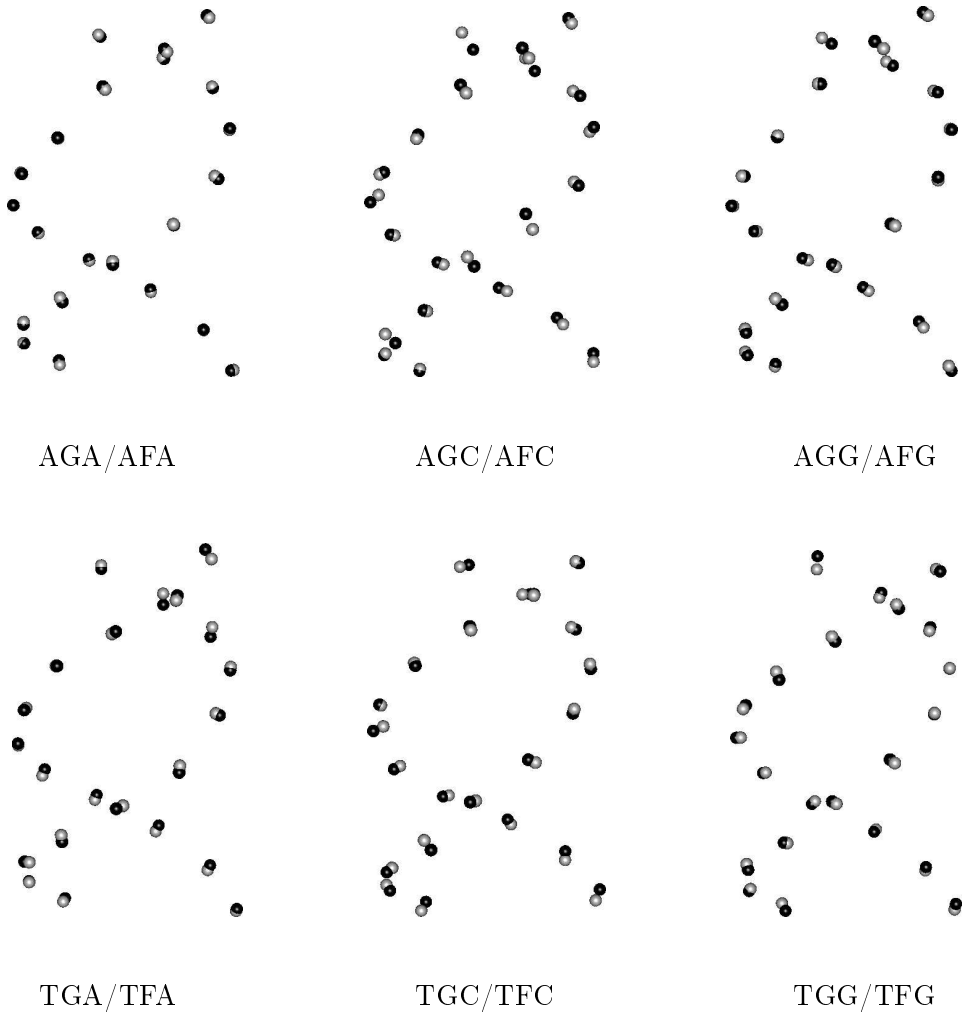

Figure 5.6: Optimally aligned and scaled icons of the sample mean shapes of the twelve DNA duplexes: The black configurations correspond to damaged DNA molecules and the grey configurations correspond to molecules where guanine has been replaced by FapydG.

As the data have been obtained using molecular dynamics simulations where the configuration at each iteration (time point) is obtained based on the configuration at the previous iteration (cf. Section 1.2.1), the configurations within each molecule (group) cannot be considered as independent. Figure 5.7 shows an example of this temporal dependence. To obtain this figure, partial GPA has been carried out on the entire set of 30,000 (12 duplexes $\times$ 2,500 time points) configurations in the dataset. Each configuration has then been projected onto the Procrustes tangent space of the resulting overall sample mean shape. Doing so yields twelve multivariate time series of 2,500 time points.

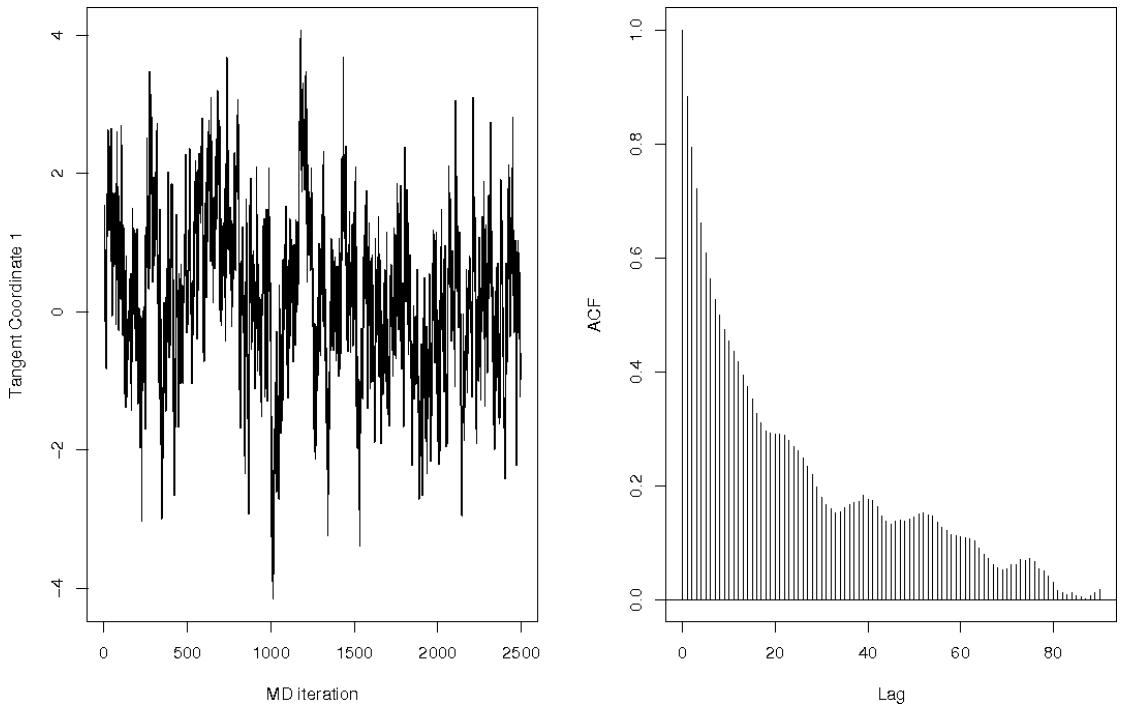

Figure 5.7: Time series of a tangent coordinate for the AFA duplex: Partial GPA has been performed on the entire set of 30,000 configuration matrices, and all configurations have been projected onto the Procrustes tangent space at the overall sample mean shape. The left hand side shows the time series of the first tangent coordinate for the AFA duplex and the right–hand side shows the corresponding correlogram. It can be seen that the data are heavily correlated.

The time series on the left–hand side in Figure 5.7 shows the temporal dependence of the first tangent coordinate of the AFA duplex, and the right–hand side shows the corresponding correlogram (cf. Section 2.2.2.1). It can be seen that the data within each molecule are heavily correlated.

To obtain approximately independent configurations in each group, thinning can be applied where only a fraction of the data is used. Ideally, the degree of thinning is thereby minimal while eliminating most of the temporal dependence. Figure 5.7 indicates that the thinning needs to be considerable in order to obtain approximately independent observations in each group. However, in this application the dimension of the shape space is large, i.e. $M = 3k - 7 = 59$ so that we need $n_{\mathrm{X}} + n_{\mathrm{Y}} > 61$ configurations in the pooled sample to calculate the observed value of the James statistic (5.13). Moreover, to ensure that the covariance estimate in (5.13) is not ill–conditioned for the bootstrap samples, the pooled sample size needs to be even bigger than 61 due to repetition of observations

Table 5.4: Estimated p–values and observed values of the James statistic for tests for the equality of mean shapes of the six pairs of (thinned) DNA data: Different values of thinning are applied which yield similar results. Only for the pair AGA/AFA is the evidence against the null hypothesis of equal mean shapes ambiguous.

	every $30th$		every $40th$		every $50th$		every $60th$	
pair	$\hat{p}$	$T^2_{J,\text{obs}}$	$\hat{p}$	$T^2_{J,\text{obs}}$	$\hat{p}$	$T^2_{J,\text{obs}}$	$\hat{p}$	$T^2_{J,\text{obs}}$
A.A	0.007	54.57	0.016	62.22	0.016	73.51	0.254	56.38
A.C	0.001	416.09	0.001	403.52	0.001	318.58	0.001	348.70
A.G	0.001	269.04	0.001	201.76	0.001	185.41	0.001	236.08
T.A	0.001	293.03	0.001	242.35	0.001	268.72	0.001	195.67
T.C	0.001	287.85	0.001	303.14	0.001	222.69	0.001	223.70
T.G	0.001	226.68	0.001	227.55	0.001	199.87	0.001	179.95

in the resamples. For the DNA data, an appropriate degree of thinning therefore has to strike a compromise between these two requirements and it is not clear which value will work best. To investigate the impact of eliminating observations, we apply different degrees of thinning, namely including every $30th$, every $40th$, every $50th$ and every $60th$ observation of each DNA duplex. Note that using every $60th$ configuration leaves only 82 observations in the pooled sample. We therefore add a small constant (10^{-6}) to the diagonal elements of the covariance estimate in (5.13) before carrying out the inversion. This computationally avoids the above mentioned problem of singularity.

Table 5.4 shows the resulting estimated p–values and observed values of the James statistic for all damaged/undamaged pairs of DNA duplexes (the dot represents either G or F in the left–hand column). All values are based on $B = 1,000$ bootstrap iterations. It can be seen that, based on Algorithm 5.1, there is very strong evidence against the null hypothesis of equal mean shapes for most pairs of duplexes. Note that an estimated p–value of $1/1001 \approx 0.001$ indicates that none of the resampled bootstrap values of the test statistic is smaller than the observed value, cf. (5.14). For most damaged/undamaged pairs, the degree of thinning does not change this results. Only for the AGA/AFA pair, does it have an impact on the estimated p–value: while $\hat{p}$ suggests very strong evidence against the null hypothesis when every $30th$, every $40th$ or every $50th$ configuration of both AGA and AFA are included in the test procedure, it increases if the test is based on only every $60th$ configuration. The corresponding observed values of the test statistic, however, do not change substantially. As they are much smaller than the observed values

Table 5.5: Estimated p–values and observed values of the James statistic for tests for the equality of mean shapes within each (thinned) duplex: Here every $30th$ configuration of each duplex is used. In each case, the test correctly finds no evidence against the null hypothesis of equal mean shapes at the 5% significance level.

duplex	$\hat{p}$	$T^2_{J,\text{obs}}$	duplex	$\hat{p}$	$T^2_{J,\text{obs}}$	duplex	$\hat{p}$	$T^2_{J,\text{obs}}$
AGA	0.668	40.88	AGC	0.371	52.06	AGG	0.160	65.93
AFA	0.277	54.51	AFC	0.533	45.28	AFG	0.298	53.72
TGA	0.809	36.43	TGC	0.435	48.14	TGG	0.446	48.71
TFA	0.606	41.23	TFC	0.079	65.73	TFG	0.262	59.74

of the test statistic for the other pairs, it can be concluded that the mean shapes of all but the AGA/AFA pair are sufficiently different to yield very large observed values of the test statistic so that the thinning does not have an impact on the result. Based on Table 5.4, the mean shapes of the AGA/AFA pair are not as different as those for the other pairs. The corresponding observed values of the test statistic are moderately large so that the applied degree of thinning can affect the results.

Note that Figure 5.6 supports the findings summarised in Table 5.4 as the mean shapes of the AGA/AFA pair do not appear as different as the mean shapes of the other pairs. Moreover, Table 5.5 shows that Algorithm 5.1 correctly finds no evidence against the null hypothesis of equal mean shapes at the 5% level if it is applied to configurations within the same duplex. Here, every $30th$ configuration for each molecule is used, and to obtain the two groups the configurations were randomly split into subsamples of equal size. While these results are reassuring, Figure 5.7 suggests that the applied level of thinning is not sufficient for the data to meet the independence assumption. The following section shows how even small temporal correlations can distort the results.

5.5 Problems with Temporally Dependent Data

To assess the performance of Algorithm 5.1 for temporally dependent shape data, we simulate time series of (4×3)–dimensional configuration matrices in landmark space using the Time Orthogonal Principal Component (TOPC) model proposed by Dryden

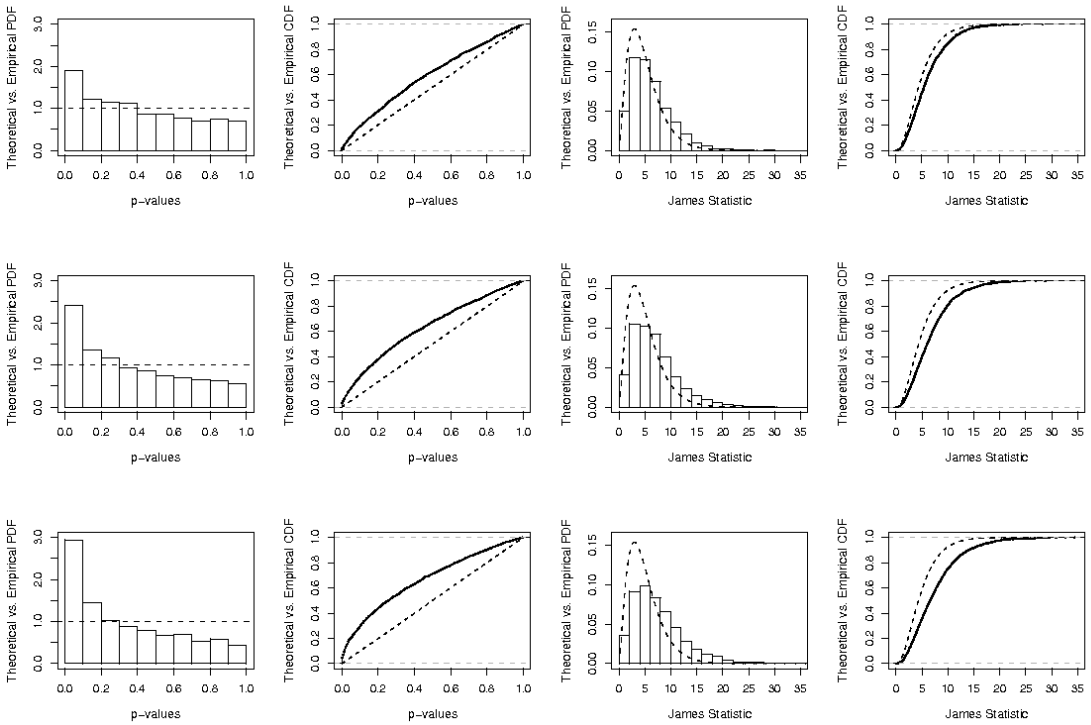

Figure 5.8: Empirical distribution of the estimated p–values and the observed values of the James statistic for dependent data with small correlations: The data were simulated from the separable TOPC–AR(1) model (cf. Section 6.1.1.3) with $\psi = 0.1$ (top), $\psi = 0.15$ (middle), $\psi = 0.2$ (bottom). Even for these small correlations, Algorithm 5.1 produces unreliable results. In particular, based on the empirical distribution of the estimated p–values, the test is very liberal with a large type I error.

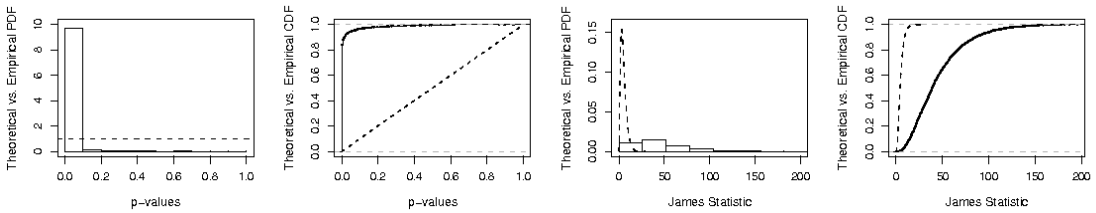

Figure 5.9: Empirical distribution of the estimated p–values and the observed values of the James statistic for dependent data with a large correlation of 0.8: Although the data were simulated under the null hypothesis of equal mean shapes, Algorithm 5.1 would reject the null hypothesis in almost every case.

et al. (2009) which will be described in detail in the next chapter. In essence, the TOPC model introduces some temporal correlation to multivariate Gaussian data in a way that each principal component (PC) is permitted to have a general dependence structure but distinct PCs are assumed to be independent. Here, we simulate 750 configuration matrices for each group using a first order autoregressive dependence structure (cf. Section 2.2.2.2) for each PC. Configuration $\breve{\boldsymbol{X}}_0$ from (5.17) is thereby chosen as the mean configuration for both groups so that the data are simulated using the exact same model (and in particular satisfy the null hypothesis of equal mean shapes).

Algorithm 5.1 is applied to test for differences in the mean shape. In Figure 5.8 the resulting empirical distributions of the estimated p–values and the James statistic are displayed for different values of the temporal correlation: $\psi = 0.1$ (top), $\psi = 0.15$ (middle) and $\psi = 0.2$ (bottom). It can be seen that the distribution of the estimated p–values ceases to follow the uniform distribution even for these small correlations, and the effect of this becomes worse very quickly. Figure 5.9 shows the effect of a large correlation ($\psi = 0.8$). In this case, almost all $\hat{p}$ are concentrated at low values. The fast bootstrap test described in Algorithm 5.1 will therefore reject the null hypothesis in almost all cases, even if it is true. Given the remaining correlation of the DNA data after thinning, this yields the question of how this drawback can be rectified, and we will investigate this in the following chapter.

5.6 Summary

In this chapter we proposed a fast bootstrap algorithm which carries out a hypothesis test for the equality of the population mean shapes of two groups of landmark data. As opposed to the bootstrap procedures proposed by Amaral *et al.* (2007), our algorithm does not use complex algebra and can be applied to landmark data of any dimension. It is based on the Procrustes tangent space approximation to shape space and can be seen as complementary to the procedures described by Preston & Wood (2009b) which are formulated in context of the MDS approach to shape analysis.

The simulation study in Section 5.2.2 shows that our algorithm yields very good results in terms of both achieved significance level and power if the data within each group are independent which will be the case in most applications.

When Algorithm 5.1 is applied to the skull data, it yields similar results to those described in Amaral *et al.* (2007). When applied to the (thinned) DNA data described in Section 1.2.3, the results suggest that the oxidative guanine lesion FapydG induces significant changes of the duplexes under study in terms of their mean shapes. However, we briefly demonstrated in Section 5.5 that Algorithm 5.1 does not yield reliable results for cases where the data within each group exhibit some temporal dependence: even for small correlations, the estimated p–values are systematically too small so that the results in Table 5.4 are questionable.

The shortcomings of Algorithm 5.1 in the context of temporal data are not surprising as both the applied test statistic and the resampling procedure are designed for independent data (cf. Sections 2.4.6 and 6.1.2.1). In the next chapter, we will investigate the question of how Algorithm 5.1 can be amended to accommodate temporal dependence of the configuration matrices within each group, and an alternative bootstrap procedure will be proposed which is specifically designed to test for mean differences of temporally evolving shape data.

Chapter 6

Bootstrap Hypothesis Testing for Temporally Dependent Configuration Matrices

Like the previous chapter, this chapter is concerned with developing a bootstrap test for the equality of the underlying population mean shapes of two groups of configuration matrices. Motivated by the problem of comparing the mean shapes of two DNA duplexes which evolve over time, we propose an amendment of Algorithm 5.1 which is specifically designed to accommodate time series of configuration matrices. The amendment is concerned with both the applied test statistic and the resampling procedure. As before, the location and scale of the data will be eliminated using a Procrustes tangent projection prior to resampling so that we seek a suitable test statistic and resampling algorithm for multivariate Euclidean data.

6.1 Amending the Test Statistic

In this section, we derive a test statistic for the equality of the population means of two groups of temporally dependent multivariate Euclidean data. This statistic is based on the Time Orthogonal Principal Component (TOPC) model by Dryden *et al.* (2009) and the likelihood ratio test (LRT) procedure (cf. Appendix D). We show that the new test statistic can be seen as a direct generalisation of the James statistic (5.13) if the sample sizes of the two groups are equal.

6.1.1 Gaussian Models for Random Matrices

The TOPC model is a special case of a Gaussian model for dependent multivariate data. It can be formulated as a Gaussian model for random matrices. Before we describe the TOPC model, we briefly review Gaussian models for random matrices in general.

6.1.1.1 Independent Rows

Let $\boldsymbol{X}$ be the $(n \times p)$–matrix which results from row–wise stacking p–dimensional random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. If the vectors can be assumed to be independent and multivariate Gaussian with mean $\boldsymbol{\mu}$ and $(p \times p)$ covariance matrix $\widetilde{\boldsymbol{\Sigma}}_C$, then their joint density has the familiar form

$$f(\boldsymbol{X}; \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^{np} |\widetilde{\boldsymbol{\Sigma}}_C|^n}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \widetilde{\boldsymbol{\Sigma}}_C^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right\}, \tag{6.1}$$

where $\boldsymbol{\theta}^T = \left(\boldsymbol{\mu}^T, \mathrm{vech}(\widetilde{\boldsymbol{\Sigma}}_C)^T\right)$ denotes the vector of all involved parameters and vech(.) denotes the vectorise–half operator defined in (0.3). Let $\mathcal{S}_p$ denote the $(p(p+1)/2)$–dimensional space of parameters which form a symmetric and positive semi–definite $(p \times p)$–matrix. The entire parameter space can then be written as $\Theta = \{\mathbb{R}^p \times \mathcal{S}_p\}$.

6.1.1.2 Factored Covariance Model – General Case

If the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ exhibit some dependence, this can be captured by introducing an additional $(n \times n)$ covariance matrix $\widetilde{\boldsymbol{\Sigma}}_R$ to describe the covariance structure between the rows of $\boldsymbol{X}$. The covariance matrix for the entire random matrix $\boldsymbol{X}$ then becomes

$$\widetilde{\boldsymbol{\Sigma}} = \mathrm{E}\big(\mathrm{vec}(\boldsymbol{X} - \boldsymbol{M})\mathrm{vec}(\boldsymbol{X} - \boldsymbol{M})^T\big) = \widetilde{\boldsymbol{\Sigma}}_C \otimes \widetilde{\boldsymbol{\Sigma}}_R,$$

where $\otimes$ denotes the Kronecker product (e.g. Mardia *et al.*, 1979, p.459), and $\boldsymbol{M} = \mathrm{E}(\boldsymbol{X})$ denotes the mean matrix. With the definition $\boldsymbol{\mu}_i = \mathrm{E}(\boldsymbol{x}_i)$ $(i = 1 \ldots, n)$, $\boldsymbol{M}$ has the form

$\boldsymbol{M}^T = \big(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n\big)$, and the density for $\boldsymbol{X}$ can be written as

$$\begin{aligned} f(\boldsymbol{X};\boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^{np}|\widetilde{\boldsymbol{\Sigma}}_C \otimes \widetilde{\boldsymbol{\Sigma}}_R|}} \exp\left\{-\frac{1}{2}\mathrm{vec}(\boldsymbol{X}-\boldsymbol{M})^T(\widetilde{\boldsymbol{\Sigma}}_C \otimes \widetilde{\boldsymbol{\Sigma}}_R)^{-1}\mathrm{vec}(\boldsymbol{X}-\boldsymbol{M})\right\} \\ &= \frac{1}{\sqrt{(2\pi)^{np}|\widetilde{\boldsymbol{\Sigma}}_C|^n|\widetilde{\boldsymbol{\Sigma}}_R|^p}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big[\widetilde{\boldsymbol{\Sigma}}_C^{-1}(\boldsymbol{X}-\boldsymbol{M})^T\widetilde{\boldsymbol{\Sigma}}_R^{-1}(\boldsymbol{X}-\boldsymbol{M})\big]\right\}. \quad (6.2) \end{aligned}$$

To accommodate the greater generality of (6.2), some additional parameters are necessary. Here, $\boldsymbol{\theta}^T = \big(\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_n^T, \mathrm{vech}(\widetilde{\boldsymbol{\Sigma}}_R)^T, \mathrm{vech}(\widetilde{\boldsymbol{\Sigma}}_C)^T\big)$ and $\Theta = \{\mathbb{R}^{np} \times \mathcal{S}_n \times \mathcal{S}_p\}$. The above model is commonly called the *matrix normal model* (cf. e.g. Arnold, 1981, p.312).

6.1.1.3 Time–Dependent Rows

Note that the independence model (6.1) is a special case of (6.2) with $\widetilde{\boldsymbol{\Sigma}}_R = \boldsymbol{I}_n$ and $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_n = \boldsymbol{\mu}$. Another special case arises when the dependence between the rows of $\boldsymbol{X}$ is temporal. In this thesis, we consider first and second order autoregressive (AR) models for the between–row dependence, and we let $\boldsymbol{\Sigma}_T$ denote the corresponding $(n \times n)$ between-row correlation matrix. In the AR(2) case, the elements of $\boldsymbol{\Sigma}_T$ represent correlations between observations of the form

$$Y_t = \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \epsilon_t,$$

cf. Section 2.2.2.2. Assuming $\mathrm{Var}(\epsilon_t) = 1\,\forall t$, it can be shown that

$$\mathrm{Var}(Y_t) = \frac{1-\psi_2}{(1+\psi_2)\big((1-\psi_2)^2 - \psi_1^2\big)} = \sigma_a^{-2},$$

where the notation σ_a^{-2} is chosen to be consistent with that by Dryden *et al.* (2009). The between–row covariance matrix therefore has the form $\widetilde{\boldsymbol{\Sigma}}_T = \sigma_a^{-2}\boldsymbol{\Sigma}_T$. Let $\widetilde{\boldsymbol{\Sigma}}_C$ denote the between–column covariance matrix. The overall covariance matrix of $\boldsymbol{X}$ then is

$$\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Sigma}}_C \otimes \widetilde{\boldsymbol{\Sigma}}_T = \widetilde{\boldsymbol{\Sigma}}_C \otimes \sigma_a^{-2}\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_C \otimes \boldsymbol{\Sigma}_T, \quad (6.3)$$

where $\widetilde{\boldsymbol{\Sigma}}_C = \sigma_a^2\boldsymbol{\Sigma}_C$. As the individual matrices in factored covariance models are defined only up to a constant, working with the temporal correlation matrix is a sensible choice.

It is clear that assuming an AR(2) dependence structure for the sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ reduces the number of parameters in $\boldsymbol{\Sigma}$ to $2 + p(p+1)/2$. If the underlying AR–process is assumed to be stationary, then the number of parameters in the mean part of the model can also be reduced because in that case, the Gaussian model for $\boldsymbol{X}$ becomes

$$f(\boldsymbol{X};\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^{np}|\boldsymbol{\Sigma}_C|^n|\boldsymbol{\Sigma}_T|^p}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Sigma}_C^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}^T)^T\boldsymbol{\Sigma}_T^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}^T)\big]\right\}, \quad (6.4)$$

where $\boldsymbol{\mu}$ denotes the marginal mean of the $\boldsymbol{x}_i$. The parameter vector of this model is $\boldsymbol{\theta}^T = (\boldsymbol{\mu}^T, \psi_1, \psi_2, \mathrm{vech}(\boldsymbol{\Sigma}_C)^T)$, and the corresponding parameter space has the form $\Theta = \mathbb{R}^p \times \mathcal{T}_2^{\mathrm{AR}} \times \mathcal{S}_p$, where

$$\mathcal{T}_2^{\mathrm{AR}} = \left\{ \boldsymbol{\psi} = (\psi_1, \psi_2)^T \in \mathbb{R}^2 : \left\{ \begin{array}{c} \psi_1 + \psi_2 < 1 \\ \psi_1 - \psi_2 < 1 \\ |\psi_2| < 1 \end{array} \right\} \right. \quad (6.5)$$

denotes the stationarity region of an AR(2) process. As mentioned in Section 2.2.2.2, the form of $\mathcal{T}_2^{\mathrm{AR}}$ can be obtained using the characteristic equation (2.19).

Under the stationarity conditions, Siddiqui (1958) shows that the inverse of the temporal correlation matrix is given by the persymmetric (symmetric about both diagonals) and pentadiagonal matrix

$$\boldsymbol{\Sigma}_T^{-1} = \sigma_a^{-2} \begin{pmatrix} 1 & -\psi_1 & -\psi_2 & 0 & \ldots & 0 & 0 \\ -\psi_1 & 1+\psi_1^2 & -\psi_1(1-\psi_2) & -\psi_2 & \ldots & 0 & 0 \\ -\psi_2 & -\psi_1(1-\psi_2) & 1+\psi_1^2+\psi_2^2 & -\psi_1(1-\psi_2) & \ldots & 0 & 0 \\ 0 & -\psi_2 & -\psi_1(1-\psi_2) & 1+\psi_1^2+\psi_2^2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1+\psi_1^2 & -\psi_1 \\ 0 & 0 & 0 & 0 & \ldots & -\psi_1 & 1 \end{pmatrix} \quad (6.6)$$

Moreover, its determinant is given by

$$|\boldsymbol{\Sigma}_T^{-1}| = \sigma_a^{-2n}\big((1-\psi_2^2)^2 - (1+\psi_2)^2\psi_1^2\big). \quad (6.7)$$

These results are useful when $\boldsymbol{\theta}$ is to be estimated using the maximum likelihood method.

The above model is a special case of the TOPC model introduced by Dryden *et al.* (2009). In particular, model (6.4) is a separable AR(2) version of the TOPC model. If ψ_2 is set to zero in all equations, then this reduces to the separable TOPC–AR(1) model.

6.1.2 Likelihood Ratio Test for Dependent Gaussian Observations

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_\mathrm{X}}$ and $\boldsymbol{y}_1, \ldots \boldsymbol{y}_{n_\mathrm{Y}}$ be two groups of random vectors. Based on model (6.4), we will derive a LRT for the test problem

$$H_0 : \boldsymbol{\mu}_\mathrm{X} = \boldsymbol{\mu}_\mathrm{Y} \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_\mathrm{X} \neq \boldsymbol{\mu}_\mathrm{Y}, \tag{6.8}$$

where $\boldsymbol{\mu}_\mathrm{X} = \mathrm{E}(\boldsymbol{x}_t)$ $(t = 1, \ldots, n_\mathrm{X})$ and $\boldsymbol{\mu}_\mathrm{Y} = \mathrm{E}(\boldsymbol{y}_t)$ $(t = 1, \ldots, n_\mathrm{Y})$. Under the assumption $\widetilde{\boldsymbol{\Sigma}}_\mathrm{X} = \widetilde{\boldsymbol{\Sigma}}_\mathrm{Y}$, where $\widetilde{\boldsymbol{\Sigma}}_\mathrm{X}$ and $\widetilde{\boldsymbol{\Sigma}}_\mathrm{Y}$ are defined as in (6.3), the corresponding LR statistic is a direct generalisation of the Mahalanobis squared distance which in fact can be derived from a LR statistic under (6.1).

6.1.2.1 The LRT for Independent Vectors

Assuming an equal covariance structure in both groups, the joint likelihood of the data under (6.1) has the form

$$L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{(n_\mathrm{X}+n_\mathrm{Y})p/2} |\widetilde{\boldsymbol{\Sigma}}_C|^{(n_\mathrm{X}+n_\mathrm{Y})/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n_\mathrm{X}} (\boldsymbol{x}_i - \boldsymbol{\mu}_\mathrm{X})^T \widetilde{\boldsymbol{\Sigma}}_C^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_\mathrm{X}) \right\}$$
$$\times \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n_\mathrm{Y}} (\boldsymbol{y}_j - \boldsymbol{\mu}_\mathrm{Y})^T \widetilde{\boldsymbol{\Sigma}}_C^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_\mathrm{Y}) \right\}. \tag{6.9}$$

Here, the joint parameter vector $\boldsymbol{\theta}^T = \left(\boldsymbol{\mu}_\mathrm{X}^T, \boldsymbol{\mu}_\mathrm{Y}^T, \mathrm{vech}(\widetilde{\boldsymbol{\Sigma}}_C)^T\right)$ is an element of $\Theta = (\mathbb{R}^{2p} \times \mathcal{S}_p)$ which can be divided into

$$\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\mu}_\mathrm{X} = \boldsymbol{\mu}_\mathrm{Y}\} \quad \text{and} \quad \Theta_1 = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\mu}_\mathrm{X} \neq \boldsymbol{\mu}_\mathrm{Y}\}.$$

Let $\hat{\boldsymbol{\theta}}_h^T = (\hat{\boldsymbol{\mu}}_{\mathrm{X},h}^T, \hat{\boldsymbol{\mu}}_{\mathrm{Y},h}^T, \mathrm{vech}(\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,h})^T)$ denote the vector which maximises (6.9) within Θ_h $(h = 0, 1)$. Under the null hypothesis, $\hat{\boldsymbol{\mu}}_{\mathrm{X},0} = \hat{\boldsymbol{\mu}}_{\mathrm{Y},0} = (n_{\mathrm{X}}+n_{\mathrm{Y}})^{-1}(n_{\mathrm{X}}\bar{\boldsymbol{x}}+n_{\mathrm{Y}}\bar{\boldsymbol{y}})$ whereas the mean vectors under the alternative are estimated separately as $\hat{\boldsymbol{\mu}}_{\mathrm{X},1} = \bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\mu}}_{\mathrm{Y},1} = \bar{\boldsymbol{y}}$. Moreover, it can be shown that the estimate of $\widetilde{\boldsymbol{\Sigma}}_{C,h}$ $(h = 1, 2)$ has the general form

$$\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,h} = \frac{1}{(n_{\mathrm{X}} + n_{\mathrm{Y}})} \left\{ \sum_{i=1}^{n_{\mathrm{X}}} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{x,h})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{x,h})^T + \sum_{j=1}^{n_{\mathrm{Y}}} (\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_{y,h})(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_{y,h})^T \right\}.$$

If $\hat{\boldsymbol{\theta}}_h$ is inserted in (6.9), then the LR statistic becomes

$$\lambda(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} f(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta_1} f(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta})} = \left\{ \frac{|\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,0}|}{|\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,1}|} \right\}^{-(n_{\mathrm{X}}+n_{\mathrm{Y}})/2}. \tag{6.10}$$

Note that $\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,0}$ captures the total variation in the data whereas $\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,1}$ summarises the within–group variation so that $\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,0}$ can be decomposed as $\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,0} = \widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,1} + \boldsymbol{B}$, where

$$\boldsymbol{B} = \frac{n_{\mathrm{X}} n_{\mathrm{Y}}}{(n_{\mathrm{X}} + n_{\mathrm{Y}})^2} (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})^T$$

estimates the between–group variation. Using this decomposition formula (6.10) can be simplified to

$$\lambda(\boldsymbol{X}, \boldsymbol{Y}) \;=\; \left\{ 1 + \frac{n_{\mathrm{X}} n_{\mathrm{Y}}}{(n_{\mathrm{X}} + n_{\mathrm{Y}})^2} (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})^T \widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,1}^{-1} (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}) \right\}^{-(n_{\mathrm{X}}+n_{\mathrm{Y}})/2}. \tag{6.11}$$

The above test is well–known in multivariate statistics (e.g. Srivastava, 2002, pp.109). As $\widehat{\widetilde{\boldsymbol{\Sigma}}}_{C,1}$ is proportional to the pooled estimator of $\widetilde{\boldsymbol{\Sigma}}_C$, the LR statistic can also be formulated in terms of the Mahalanobis squared distance (5.5). In fact, (6.11) is a monotone transformation of (5.5) so that both statistics lead to the same test result within a bootstrap procedure. Moreover, if $n_{\mathrm{X}} = n_{\mathrm{Y}} = n$, then (6.11) also is a monotone transformation of the James statistic (5.13) so that the assumption of equal covariances in the two group can be relaxed in that case. Relaxing the independence assumption, however, is less straightforward and requires the use of different models.

We now show how the above LRT can be generalised to dependent situations using model (6.4). To simplify the treatment, we thereby concentrate on the case where $n_{\mathrm{X}} = n_{\mathrm{Y}} = n$.

6.1.2.2 Time–Dependent Rows

If the vectors within each group can adequately be modelled using (6.4), then – assuming a common covariance structure – the joint likelihood has the form

$$L(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{np}|\boldsymbol{\Sigma}_C|^n|\boldsymbol{\Sigma}_T|^p}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_C^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{X}}^T)^T\boldsymbol{\Sigma}_T^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{X}}^T)\right]\right\}$$
$$\times\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_C^{-1}(\boldsymbol{Y}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{Y}}^T)^T\boldsymbol{\Sigma}_T^{-1}(\boldsymbol{Y}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{Y}}^T)\right]\right\}, \quad (6.12)$$

where $\boldsymbol{\theta}^T = (\boldsymbol{\mu}_{\mathrm{X}}^T, \boldsymbol{\mu}_{\mathrm{Y}}^T, \psi_1, \psi_2, \mathrm{vech}(\boldsymbol{\Sigma}_C)^T)$. The corresponding parameter space is $\Theta = \mathbb{R}^{2p} \times \mathcal{T}_2^{\mathrm{AR}} \times \mathcal{S}_p$, and the test problem at hand divides Θ into

$$\Theta_0 = \{\boldsymbol{\theta}\in\Theta : \boldsymbol{\mu}_{\mathrm{X}} = \boldsymbol{\mu}_{\mathrm{Y}}\} \quad \text{and} \quad \Theta_1 = \{\boldsymbol{\theta}\in\Theta : \boldsymbol{\mu}_{\mathrm{X}} \neq \boldsymbol{\mu}_{\mathrm{Y}}\}. \quad (6.13)$$

Dryden *et al.* (2009) describe the ML estimation in the one–sample case. To obtain the desired LR statistic, this has to be extended to (6.12) taking into account the hypotheses.

It can be shown that the MLEs of the mean vectors have the form

$$\hat{\boldsymbol{\mu}}_{\mathrm{X},0} = \frac{(\boldsymbol{X}+\boldsymbol{Y})^T\widehat{\boldsymbol{\Sigma}}_{T,0}^{-1}\mathbf{1}_n}{2\mathbf{1}_n^T\widehat{\boldsymbol{\Sigma}}_{T,0}^{-1}\mathbf{1}_n} = \hat{\boldsymbol{\mu}}_{\mathrm{Y},0}, \quad \hat{\boldsymbol{\mu}}_{\mathrm{X},1} = \frac{\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_{T,1}^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\widehat{\boldsymbol{\Sigma}}_{T,1}^{-1}\mathbf{1}_n}, \quad \text{and} \quad \hat{\boldsymbol{\mu}}_{\mathrm{Y},1} = \frac{\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}_{T,1}^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\widehat{\boldsymbol{\Sigma}}_{T,1}^{-1}\mathbf{1}_n}.$$

Define $\hat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\Sigma}}_{T,0}^{-1}\mathbf{1}_n/(2\mathbf{1}_n^T\widehat{\boldsymbol{\Sigma}}_{T,0}^{-1}\mathbf{1}_n)$ and $\hat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}_{T,1}^{-1}\mathbf{1}_n/(\mathbf{1}_n^T\widehat{\boldsymbol{\Sigma}}_{T,1}^{-1}\mathbf{1}_n)$. The above estimators can then be written as weighted means of the vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ and $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$, i.e.

$$\hat{\boldsymbol{\mu}}_{\mathrm{X},0} = \sum_{t=1}^n \hat{\alpha}_t(\boldsymbol{x}_t+\boldsymbol{y}_t) = \hat{\boldsymbol{\mu}}_{\mathrm{Y},0}, \quad \hat{\boldsymbol{\mu}}_{\mathrm{X},1} = \sum_{t=1}^n \hat{\beta}_t\boldsymbol{x}_t, \quad \text{and} \quad \hat{\boldsymbol{\mu}}_{\mathrm{Y},1} = \sum_{t=1}^n \hat{\beta}_t\boldsymbol{y}_t.$$

As both $\widehat{\boldsymbol{\Sigma}}_{T,0}^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_{T,1}^{-1}$ have the general form (6.6), all but four rows within each matrix have the same sum. If n is large, then these end effects can be neglected and the above mean estimators can therefore be approximated well by

$$\hat{\boldsymbol{\mu}}_{\mathrm{X},0} = \frac{\bar{\boldsymbol{x}}+\bar{\boldsymbol{y}}}{2} = \hat{\boldsymbol{\mu}}_{\mathrm{Y},0}, \quad \hat{\boldsymbol{\mu}}_{\mathrm{X},1} = \bar{\boldsymbol{x}}, \quad \text{and} \quad \hat{\boldsymbol{\mu}}_{\mathrm{Y},1} = \bar{\boldsymbol{y}}. \quad (6.14)$$

These approximations are asymptotically efficient (Grenander & Rosenblatt, 1957).

Let $\hat{\boldsymbol{\psi}}_h$ denote the MLEs for the AR(2)–parameters under H_h $(h = 0, 1)$. Given $\hat{\boldsymbol{\psi}}_h$, the corresponding correlation matrix $\widehat{\boldsymbol{\Sigma}}_{T,h}$ is fully determined, and it can be shown that

$$\widehat{\boldsymbol{\Sigma}}_{C,h} = \frac{1}{2n}\big\{(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{X},h}^T)^T\widehat{\boldsymbol{\Sigma}}_{T,h}^{-1}(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{X},h}^T) + (\boldsymbol{Y} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{Y},h}^T)^T\widehat{\boldsymbol{\Sigma}}_{T,h}^{-1}(\boldsymbol{Y} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{Y},h}^T)\big\} \quad (6.15)$$

The part in the exponential term of (6.12) therefore reduces to np under both hypotheses. Regarding $\widehat{\boldsymbol{\Sigma}}_{C,h}$ as a function of $\boldsymbol{\psi}$ and using (6.7), it follows that

$$\sup_{\boldsymbol{\theta}\in\Theta_h} L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}) = \sup_{\boldsymbol{\psi}\in\mathcal{T}_2^{\mathrm{AR}}} c \cdot |\widehat{\boldsymbol{\Sigma}}_{C,h}|^{-n}\sigma_a^{-2np}\big((1-\psi_2^2)^2 - (1+\psi_2)^2\psi_1^2\big)^p,$$

where $c = (2\pi)^{-np}\exp(-np)$. With the definition $f_h(\boldsymbol{\psi}) = |\widehat{\boldsymbol{\Sigma}}_{C,h}|^{-n}\sigma_a^{-2np}\big((1-\psi_2^2)^2 - (1+\psi_2)^2\psi_1^2\big)^p$ the LR statistic (D.2) then becomes

$$\lambda(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sup_{\boldsymbol{\psi}\in\mathcal{T}_2^{\mathrm{AR}}} f_0(\boldsymbol{\psi})}{\sup_{\boldsymbol{\psi}\in\mathcal{T}_2^{\mathrm{AR}}} f_1(\boldsymbol{\psi})}, \quad (6.16)$$

and the corresponding LRT can be seen as a generalisation of the well–known LRT based on (6.11). A more detailed derivation of the above statistic is provided in Appendix E.

6.2 Amending the Resampling Procedure

As mentioned in Section 2.4.6, the reason for the inadequacy of Efron's (1979) *i.i.d.* bootstrap in the context of dependent data is that by using the resampling scheme (2.28), all information about the dependence structure is lost. One way to preserve this information is to use a block bootstrap method where blocks of (consecutive) observations instead of single observations are resampled. There are different versions of block bootstrap methods, e.g. the moving block bootstrap (Künsch, 1989; Liu & Singh, 1992) and the non–overlapping block bootstrap (Carlstein, 1986). Here, we adapt the circular block bootstrap (CBB) by Politis & Romano (1992) to the two–sample situation.

Consider the situation where the data at hand are two samples $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of consecutive multivariate observations from some underlying

(strictly) stationary temporal processes $\{\boldsymbol{X}\}_{j\geq 1}$ and $\{\boldsymbol{Y}\}_{j\geq 1}$ which follow distributions F and G, respectively. In this two–sample situation, the parameter of interest has the form $\theta = t(F, G)$. To estimate θ, the idea of the CBB is to "wrap" the data around in a "circle" by defining new time series $\{\boldsymbol{x}_{i_0}\}_{i_0\geq 1}$ and $\{\boldsymbol{y}_{i_0}\}_{i_0\geq 1}$, where $i_0 = j$ if $i_0 = mn+j$ for some integers $m \geq 0$ and $1 \leq j \leq n$, e.g. the $(n+1)st$ observation of $\{\boldsymbol{x}_{i_0}\}_{i_0\geq 1}$ corresponds to the first observation in $\{\boldsymbol{x}_j\}_{j=1}^n$. Let l be an integer satisfying $1 < l < n$ and define blocks of length l by $\mathcal{B}_{\mathrm{X}}(i_0, l) = \{\boldsymbol{x}_{i_0}, \ldots, \boldsymbol{x}_{i_0+l-1}\}$ and $\mathcal{B}_{\mathrm{Y}}(i_0, l) = \{\boldsymbol{y}_{i_0}, \ldots, \boldsymbol{y}_{i_0+l-1}\}$. Moreover, define $n_B = \lceil n/l \rceil$, where $\lceil x \rceil$ denote the largest integer not exceeding x. To create resamples $\boldsymbol{\mathcal{X}}^*$ and $\boldsymbol{\mathcal{Y}}^*$ of the given data, n_B blocks are selected at random from the sets $\{\mathcal{B}_{\mathrm{X}}(1, l), \ldots, \mathcal{B}_{\mathrm{X}}(n, l)\}$ and $\{\mathcal{B}_{\mathrm{Y}}(1, l), \ldots, \mathcal{B}_{\mathrm{Y}}(n, l)\}$, respectively, i.e.

$$\boldsymbol{\mathcal{X}}^* = \{\mathcal{B}_{\mathrm{X}}(I_1^{\mathrm{X}}, l), \ldots, \mathcal{B}_{\mathrm{X}}(I_{n_B}^{\mathrm{X}}, l)\} \quad \text{and} \quad \boldsymbol{\mathcal{Y}}^* = \{\mathcal{B}_{\mathrm{Y}}(I_1^{\mathrm{Y}}, l), \ldots, \mathcal{B}_{\mathrm{Y}}(I_{n_B}^{\mathrm{Y}}, l)\},$$

where $I_1^{\mathrm{X}}, \ldots, I_{n_B}^{\mathrm{X}}$ and $I_1^{\mathrm{Y}}, \ldots, I_{n_B}^{\mathrm{Y}}$ are conditionally *i.i.d* random variables following a uniform distribution on $\{1 \ldots, n\}$, i.e.

$$\mathrm{P}(I_j^{\mathrm{X}} = i \,|\, \boldsymbol{\mathcal{X}}) = n_B^{-1} \quad \text{and} \quad \mathrm{P}(I_j^{\mathrm{Y}} = i \,|\, \boldsymbol{\mathcal{Y}}) = n_B^{-1}; \quad 1 \leq i, j \leq n_B.$$

Based on these resamples, the underlying distributions can be estimated, and the block bootstrap estimator of θ is $\hat{\theta}_{\mathrm{B}}^* = t(\hat{F}, \hat{G})$. Note that for $l = 1$, the CBB reduces to the *i.i.d.* bootstrap described in Section 2.4.

One of the advantages of the CBB is that each of the original observations receives equal weight in the resampling procedure, e.g. each observation $\boldsymbol{x}_j$ from $\{\boldsymbol{x}_j\}_{j=1}^n$ appears exactly l times in the collection of blocks $\{\mathcal{B}_{\mathrm{X}}(1, l), \ldots, \mathcal{B}_{\mathrm{X}}(n, l)\}$ which in turn are resampled with equal probabilities. This property distinguishes the CBB from the other block bootstrap methods. In particular, this means that the conditional expectation of the bootstrap sample mean equals the sample mean of the original sample $\{\boldsymbol{x}_j\}_{j=1}^n$ (e.g. Lahiri, 2003, p.34). For our two–sample situation, it therefore holds under the CBB that

$$\mathrm{E}(\bar{\boldsymbol{X}}^* | \, \boldsymbol{\mathcal{X}}) = \bar{\boldsymbol{x}} \quad \text{and} \quad \mathrm{E}(\bar{\boldsymbol{Y}}^* | \, \boldsymbol{\mathcal{Y}}) = \bar{\boldsymbol{y}},$$

where $\bar{\boldsymbol{X}}^*$ denotes the sample mean of a resample $\boldsymbol{\mathcal{X}}^* = \{\mathcal{B}_{\mathrm{X}}(I_1^{\mathrm{X}}, l), \ldots, \mathcal{B}_{\mathrm{X}}(I_{n_B}^{\mathrm{X}}, l)\} = \{\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_{l\,n_B}^*\}$ and $\bar{\boldsymbol{Y}}^*$ is defined in the same way.

In general, as the properties of a block bootstrap estimator $\hat{\theta}^*_{\text{B}} = \hat{\theta}^*_{\text{B}}(l)$ depend on the block length l, the choice of l is an important issue. In large sample considerations, it is typically required that l increases with the sample size so that any finite–dimensional joint distribution of the underlying processes $\{\boldsymbol{X}_j\}_{j\geq 1}$ and $\{\boldsymbol{Y}_j\}_{j\geq 1}$ can eventually be recovered from the resampled sequences. For a fixed, finite sample size, the bias of a block bootstrap estimator usually decreases with the block length whereas its variance increases. Thus, there is an optimal value of l that balances the trade off between the bias and the variance.

In some cases, it is possible to obtain the optimal block length for a given data situation analytically (cf. e.g. Hall *et al.*, 1995; Lahiri, 2003, Chapter 5) using an expansion of the mean squared error $\text{MSE}\big(\hat{\theta}^*_{\text{B}}(l)\big) = \big\{\text{E}\big(\hat{\theta}^*_{\text{B}}(l)\big) - \theta\big\}^2 + \text{Var}\big\{\hat{\theta}^*_{\text{B}}(l)\big\}^2$ which can then be minimised with respect to l, i.e. $l_{\text{opt}} = \arg\min_{1\leq l\leq n} \text{MSE}\big(\hat{\theta}^*_{\text{B}}(l)\big)$. However, the block bootstrap estimator of interest in this thesis is a block–based version of the LR statistic (6.16) which has a rather complicated form. We therefore apply a different method to choose the block length for the test problem at hand (cf. step 5 in Algorithm 6.1), and within a simulation study it is possible to assess the performance of this method based on the performance criteria for hypothesis tests (cf. Section 5.2.2.1).

6.3 Bootstrap Test for Temporally Dependent Configuration Matrices

In this section, we show how the methods described in Sections 6.1 and 6.2 can be utilised for testing the equality of the underlying mean shapes of two groups of temporally dependent configuration matrices. Given two samples $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ of $(k \times m)$ configuration matrices, the considered shape model in landmark space can in this context be formulated as

$$\boldsymbol{X}_i = \beta_i(\boldsymbol{\mu}_{\text{X}} + \boldsymbol{E}_i)\boldsymbol{\Gamma}_i + \boldsymbol{1}_k\boldsymbol{\gamma}_i^T \quad \text{and} \quad \boldsymbol{Y}_j = \beta_j(\boldsymbol{\mu}_{\text{Y}} + \boldsymbol{E}_j)\boldsymbol{\Gamma}_j + \boldsymbol{1}_k\boldsymbol{\gamma}_j^T,$$

where $\{\text{vec}(\boldsymbol{E}_i)\}_{i=1}^n$ and $\{\text{vec}(\boldsymbol{E}_j)\}_{j=1}^n$ follow a stationary zero–mean stochastic process.

Algorithm 6.1 Bootstrap algorithm for testing the null hypothesis of equal mean shapes when the observations are temporally dependent configuration matrices

1: carry out GPA on the entire set of configuration matrices
2: obtain the sequences of tangent vectors $\{\widetilde{\boldsymbol{w}}_t\}_{t=1}^n$ and $\{\widetilde{\boldsymbol{v}}_t\}_{t=1}^n$ by projecting the optimally rotated, translated and scaled data onto the observed tangent space $\mathcal{H}_{\hat{\mu}_p}(S_m^k)$
3: eliminate the redundant dimensions to obtain the sequences $\{\boldsymbol{w}_t\}_{t=1}^n$ and $\{\boldsymbol{v}_t\}_{t=1}^n$ and the corresponding data matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ using (5.12)
4: calculate the observed value $\lambda^{\text{obs}}(\boldsymbol{V}, \boldsymbol{W})$ as well as $\hat{\boldsymbol{\psi}}^{H_0}$ and $\hat{\boldsymbol{\psi}}^{H_1}$ for the observed samples
5: select block length l and the number of blocks n_B using the autocorrelation function based on $\hat{\boldsymbol{\psi}}^{H_1}$ as $l = \arg\min_{l \in \mathbb{N}_n} = |l - l_{\text{crit}}|$, where $\mathbb{N}_n = \{l \in \mathbb{N} : n/l \in \mathbb{N}\}$, $l_{\text{crit}} = \min\{l \in \mathbb{N} : |\rho(l)| < \rho_{\text{crit}}\}$ and $n_B = n/l$
6: transform to the null hypothesis by centering to yield $\{\boldsymbol{w}_t^c\}_{t=1}^n$ and $\{\boldsymbol{v}_t^c\}_{t=1}^n$
7: periodically extend time series to yield $\{\boldsymbol{w}_{0t}^c\}_{t\geq 1}$ and $\{\boldsymbol{v}_{0t}^c\}_{t\geq 1}$
8: **for** b in $(1:B)$ **do**
9: select random random staring indices $I_1^{\text{X}}, \ldots, I_{n_B}^{\text{X}}$ and $I_1^{\text{Y}}, \ldots, I_{n_B}^{\text{Y}}$
10: form resamples $\boldsymbol{\mathcal{V}}_c^* = \{\mathcal{B}_{\text{X}}(I_1^{\text{X}}, l), \ldots, \mathcal{B}_{\text{X}}(I_{n_B}^{\text{X}}, l)\}$, $\boldsymbol{\mathcal{W}}_c^* = \{\mathcal{B}_{\text{Y}}(I_1^{\text{Y}}, l), \ldots, \mathcal{B}_{\text{Y}}(I_{n_B}^{\text{Y}}, l)\}$ and corresponding data matrices $\boldsymbol{V}_c^*$, $\boldsymbol{W}_c^*$ from $\{\boldsymbol{w}_{0t}^c\}_{t\geq 1}$ and $\{\boldsymbol{v}_{0t}^c\}_{t\geq 1}$, respectively
11: calculate the bootstrap value $\lambda^{(b)}(\boldsymbol{V}_c^*, \boldsymbol{W}_c^*)$ of the test statistic
12: **end for**
13: calculate the estimated p–value $\hat{p} = \big(1 + \sum_{b=1}^B I_{\{\lambda^{\text{obs}}(\boldsymbol{V},\boldsymbol{W}) > \lambda^{(b)}(\boldsymbol{V}_c^*, \boldsymbol{W}_c^*)\}}\big)/\big(B+1\big)$

As before, the βs are positive scale factors, the $\boldsymbol{\Gamma}$s are rotation matrices, the $\boldsymbol{\gamma}$s are translation vectors and $\boldsymbol{\mu}_{\text{X}}$ and $\boldsymbol{\mu}_{\text{Y}}$ denote the population mean configurations. Like in the independent case, the mean configurations in combination with the error distributions induce certain distributions $Q_{[X]}$ and $Q_{[Y]}$ on the corresponding shape space Σ_m^k, and the considered test problem is

$$H_0 : \big[\mu_{[X]}\big] = \big[\mu_{[Y]}\big] \quad \text{vs.} \quad H_1 : \big[\mu_{[X]}\big] \neq \big[\mu_{[Y]}\big], \tag{6.17}$$

where $\big[\mu_{[X]}\big]$ and $\big[\mu_{[Y]}\big]$ denote the population mean shapes.

6.3.1 The Algorithm

Algorithm 6.1 summarises the amended bootstrap algorithm for testing (6.17) when the configuration matrices within each group are temporally dependent. Steps 1–3 are thereby identical to the corresponding steps in Algorithm 5.1. However, the resulting

projected tangent vectors in this context form two multivariate time series $\{\boldsymbol{w}_t\}_{t=1}^n$ and $\{\boldsymbol{v}_t\}_{t=1}^n$ in M dimensions. Row–wise stacking the vectors $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$ then yields two $(n \times M)$–matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ based on which the observed value of the test statistic (6.16) can be calculated (cf. step 4). As described in Section 6.1.2.2, obtaining $\lambda^{\text{obs}}(\boldsymbol{V}, \boldsymbol{W})$ involves estimating the weight parameters ψ_1 and ψ_2 of the assumed underlying AR(2) process under both the null hypothesis and the alternative. Let

$$\hat{\boldsymbol{\psi}}^{H_h} = \arg \max_{\boldsymbol{\psi} \in \mathcal{T}_2^{\text{AR}}} f_h(\boldsymbol{\psi})$$

denote the MLE of $\boldsymbol{\psi} = (\psi_1, \psi_2)^T$ under H_h $(h = 0, 1)$. Note that $\hat{\boldsymbol{\psi}}^{H_1}$ is obtained under a less restrictive model which allows for separate mean vectors of the two groups (for the considered data situation, Θ_0 is $(M + 2 + M(M+1)/2)$–dimensional whereas Θ_1 contains $(2M + 2 + M(M+1)/2)$–dimensional vectors), so that it will fit the observed data more closely than $\hat{\boldsymbol{\psi}}^{H_0}$ – especially if the data do not satisfy the null hypothesis.

Step 5 of algorithm 6.1 is concerned with choosing the block length l for the CBB. To do so, $\hat{\boldsymbol{\psi}}^{H_1}$ is inserted into the formula for the autocorrelation function $\rho(.)$ of the corresponding AR(2) process; cf. (2.20). Regardless of the specific values $\hat{\psi}_1^{H_1}$ and $\hat{\psi}_2^{H_1}$, $\rho(k)$ will thereby decrease exponentially as the lag k increases. For every value $\rho_{\text{crit}} > 0$, there will therefore be a lag value k_{crit} for which $|\rho(k_{\text{crit}})| < \rho_{\text{crit}}$, and k_{crit} obviously depends on the strength of the dependence. To take into account the dependence structure of the given data, we select l according to

$$\arg \min_{l \in \mathbb{N}_n} = |l - l_{\text{crit}}|, \quad \text{where} \quad \mathbb{N}_n = \{l \in \mathbb{N} : n/l \in \mathbb{N}\}, \; l_{\text{crit}} = \min\{l \in \mathbb{N} : |\rho(l)| < \rho_{\text{crit}}\},$$

i.e. l is the integer divisor of the sample size n which is closest to the lag at which the autocorrelation falls below a certain critical value. With this choice, the number of blocks required for each resample is $n_B = n/l$. If ρ_{crit} is chosen to be small, then the blocks $\{\mathcal{B}_{\text{X}}(1, l), \ldots, \mathcal{B}_{\text{X}}(n, l)\}$ and $\{\mathcal{B}_{\text{Y}}(1, l), \ldots, \mathcal{B}_{\text{Y}}(n, l)\}$ contain almost the entire information about the dependence structure of the data.

Steps 6 and 7 are preprocessing steps for the actual CBB algorithm. In step 6, the time series $\{\boldsymbol{w}_t\}_{t=1}^n$ and $\{\boldsymbol{v}_t\}_{t=1}^n$ in both groups are centred. As before, this step is important

to ensure a meaningful comparison of the observed test statistic with its (estimated) distribution under the null hypothesis of equal means. In step 7, the centred time series $\{\boldsymbol{w}_t^c\}_{t=1}^n$ and $\{\boldsymbol{v}_t^c\}_{t=1}^n$ are periodically extended to yield $\{\boldsymbol{w}_{0t}^c\}_{t\geq 1}$ and $\{\boldsymbol{v}_{0t}^c\}_{t\geq 1}$, respectively.

Steps 8–12 correspond to the CBB algorithm as described in Section 6.2: using conditionally independent random variables $I_1^{\mathrm{X}}, \ldots, I_{n_B}^{\mathrm{X}}$ and $I_1^{\mathrm{Y}}, \ldots, I_{n_B}^{\mathrm{Y}}$ as starting indices, resamples $\boldsymbol{\mathcal{V}}_c^*$ and $\boldsymbol{\mathcal{W}}_c^*$ are generated from $\{\mathcal{B}_{\mathrm{X}}(1,l), \ldots, \mathcal{B}_{\mathrm{X}}(n,l)\}$ and $\{\mathcal{B}_{\mathrm{Y}}(1,l), \ldots, \mathcal{B}_{\mathrm{Y}}(n,l)\}$. Based on the corresponding data matrices $\boldsymbol{V}_c^*$, $\boldsymbol{W}_c^*$, the bootstrap value $\lambda^{(b)}(\boldsymbol{V}_c^*, \boldsymbol{W}_c^*)$ of the test statistic can then be calculated at each bootstrap iteration.

Finally, step 13 comprises of calculating the estimated p–value.

6.3.2 A Monte Carlo Simulation Study

In this section, a simulation study is carried out, and the Monte–Carlo based performance criteria for hypothesis tests (cf. Section 5.2.2.1) are used to assess the benefits of the employed amendments in Algorithm 6.1.

6.3.2.1 Simulating Dependent Configuration Matrices

To obtain a stationary sequence $\{\boldsymbol{X}_t\}_{t=1}^n$ of $(k \times m)$ configuration matrices with a marginal mean configuration $\boldsymbol{\mu} \in \mathbb{R}^{k\times m}$ and a separable TOPC-AR(2) dependence structure, we want to simulate from the matrix normal model

$$\tilde{\boldsymbol{X}} \sim \mathrm{N}\big(\mathbf{1}_n \mathrm{vec}(\boldsymbol{\mu})^T, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_C\big), \tag{6.18}$$

where $\tilde{\boldsymbol{X}}$ row–wise contains $\mathrm{vec}(\boldsymbol{X}_1), \ldots, \mathrm{vec}(\boldsymbol{X}_n)$, $\boldsymbol{\Sigma}_T$ denotes the temporal correlation matrix between the rows of $\tilde{\boldsymbol{X}}$, and $\boldsymbol{\Sigma}_C$ denotes a multiple of the between column covariance matrix $\widetilde{\boldsymbol{\Sigma}}_C$, i.e. $\boldsymbol{\Sigma}_C = \sigma_a^{-2}\widetilde{\boldsymbol{\Sigma}}_C$ as described in Section 6.1.1.3.

If the underlying AR(2) parameters fall within the stationarity region (6.5), it holds that

$$\mathrm{vec}(\boldsymbol{X}_i) \sim \mathrm{N}\big(\mathrm{vec}(\boldsymbol{\mu}), \widetilde{\boldsymbol{\Sigma}}_C\big), \quad i = 1, \ldots, n,$$

i.e. each row of $\tilde{\boldsymbol{X}}$ marginally follows a normal model for a $(k \times m)$ configuration matrix such as (5.15) or (5.16) in the previous chapter. However, here we also introduce the correlation matrix $\boldsymbol{\Sigma}_T$ which determines the temporal dependence between the $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$.

To simulate from (6.18), we start with an $\big(n \times (km)\big)$ matrix $\boldsymbol{Z}$ with $(\boldsymbol{Z})_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0,1)$ $(i = 1, \ldots, n; j = 1, \ldots, km)$. For a given vector $\boldsymbol{\psi} = (\psi_1, \psi_2)^T$ of AR(2) parameters, we then calculate the inverse of the temporal correlation matrix $\boldsymbol{\Sigma}_T^{-1}$ given by (6.6). As $\boldsymbol{\Sigma}_T^{-1}$ is an $(n \times n)$ matrix, it can be large. However, many of its entries are zero and we can use the *sparse* option of the *Matrix* package in R to calculate the square root of its inverse. For a given $(km \times km)$ between column covariance matrix $\widetilde{\boldsymbol{\Sigma}}_C$ and a mean vector $\boldsymbol{\mu}$, we can then simulate a normally distributed trajectory of configuration matrices using

$$\underbrace{\boldsymbol{\Sigma}_T^{1/2} \boldsymbol{Z} \boldsymbol{\Sigma}_C^{1/2} + \mathbf{1}_n \mathrm{vec}(\boldsymbol{\mu})^T}_{\tilde{\boldsymbol{X}}} \sim \mathrm{N}\big(\mathbf{1}_n \mathrm{vec}(\boldsymbol{\mu})^T, \boldsymbol{\Sigma}_T, \boldsymbol{\Sigma}_C\big), \tag{6.19}$$

where $\tilde{\boldsymbol{X}}$ is defined as in (6.18); see also Arnold (1981, p.312).

6.3.2.2 Simulated Data

Like in the previous chapter, this simulation study is based on configuration matrices which contain $k = 4$ landmarks in $m = 3$ dimensions, i.e. effectively we generate data in $M = 3k - 7 = 5$ dimensions. Here, we use an isotropic between column covariance matrix $\widetilde{\boldsymbol{\Sigma}}_C = \sigma_c^2 \boldsymbol{I}_{12}$. The overall dependence structure of a trajectory $\{\boldsymbol{X}_t\}_{t=1}^n$ or $\{\boldsymbol{Y}_t\}_{t=1}^n$ therefore has the form $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_{12} \otimes \boldsymbol{\Sigma}_T$, where $\sigma^2 = \sigma_c^2 \sigma_a^{-2}$.

To determine $\boldsymbol{\Sigma}_T$, we need a vector $\boldsymbol{\psi} = (\psi_1, \psi_2)^T$ of AR(2) parameters. Initially, we choose $\psi_2 = 0$ and simulate configuration matrices from a separable TOPC–AR(1) model with three different values of ψ_1, namely $\psi_1 \in \{0.2, 0.5, 0.8\}$. Figure 6.1 displays the

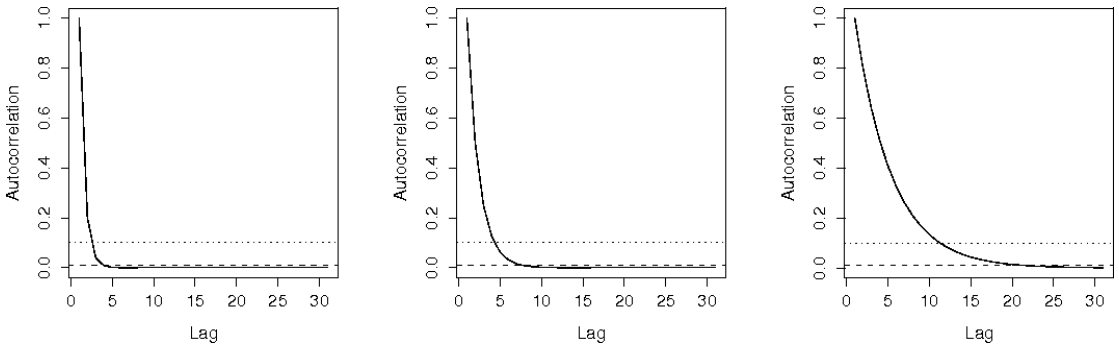

Figure 6.1: Autocorrelation functions of the employed AR(1) models: Three different AR(1) parameters are chosen, namely $\psi_1 = 0.2$ (left), $\psi_1 = 0.5$ (middle), and $\psi_1 = 0.8$ (right). The dependence structure induced on the data varies considerably between these choices. In each plot, the dotted line is the constant function $f(\text{Lag}) = 0.1$, and the dashed line corresponds to $g(\text{Lag}) = 0.01$.

correlation structures these choices of ψ_1 induce on the data. The horizontal lines show the constant functions $f(s) = 0.1$ (dotted) and $g(s) = 0.01$ (dashed), where s denotes the lag between two observations. These lines are added to the graphs as 0.1 and 0.01 are the choices we consider for the hyperparameter ρ_{crit} in Algorithm 6.1 (cf. step 5) so that they give an impression about the selected block lengths.

We consider three values of standard deviations and three sample sizes, namely $\sigma \in \{0.1, 0.3, 0.5\}$ and $n \in \{150, 500, 750\}$. Like in Chapter 5, the mean configurations $\boldsymbol{\mu}_{\text{X}}$ and $\boldsymbol{\mu}_{\text{Y}}$ are chosen as icons whose shapes lie on the geodesic path displayed in Figure 5.2. In particular, we fix $\boldsymbol{\mu}_{\text{X}} = \breve{\boldsymbol{X}}_0$ as the mean configuration for the sequence $\{\boldsymbol{X}_t\}_{t=1}^n$ in all cases and use either $\boldsymbol{\mu}_{\text{Y}} = \breve{\boldsymbol{X}}_0$ or $\boldsymbol{\mu}_{\text{Y}} = \breve{\boldsymbol{X}}_2$ as the mean configuration for $\{\boldsymbol{Y}_t\}_{t=1}^n$.

Overall, we therefore consider 54 scenarios, i.e. 27 scenarios (3 AR(1) models $\times$ 3 standard deviations $\times$ 3 sample sizes) under H_0 and the same number of scenarios under H_1. Moreover, in order to assess the impact of the hyperparameter ρ_{crit} on the results, each of these 54 parameter combinations is repeated twice using $\rho_{\text{crit}} = 0.1$ and $\rho_{\text{crit}} = 0.01$, respectively. In all cases, the number of bootstrap iterations is fixed at $B = 150$, and each scenario is repeated for $n_{\text{sim}} = 500$ Monte Carlo iterations. The reason for reducing B and n_{sim} compared to the corresponding values in the previous chapter is the increased computational cost brought about by modifying both the test statistic and the resampling procedure of Algorithm 5.1.

6.3.2.3 Results

We first focus on the case where the hyperparameter is chosen as $\rho_{\text{crit}} = 0.1$. The results will subsequently be compared with those obtained using $\rho_{\text{crit}} = 0.01$.

Achieved Significance Level and Power

For the 27 parameter combinations which simulate $\{\boldsymbol{X}_t\}_{t=1}^n$ and $\{\boldsymbol{Y}_t\}_{t=1}^n$ under H_0, we can assess the performance of Algorithm 6.1 in terms of the distribution of the estimated p–values (cf. Section 5.2.2.1). The solid lines in Figure 6.2 show the histograms and empirical distribution functions of the estimated p–values for the most and least challenging of the considered simulation scenarios. The case where $n = 750$ is combined with $\sigma = 0.1$

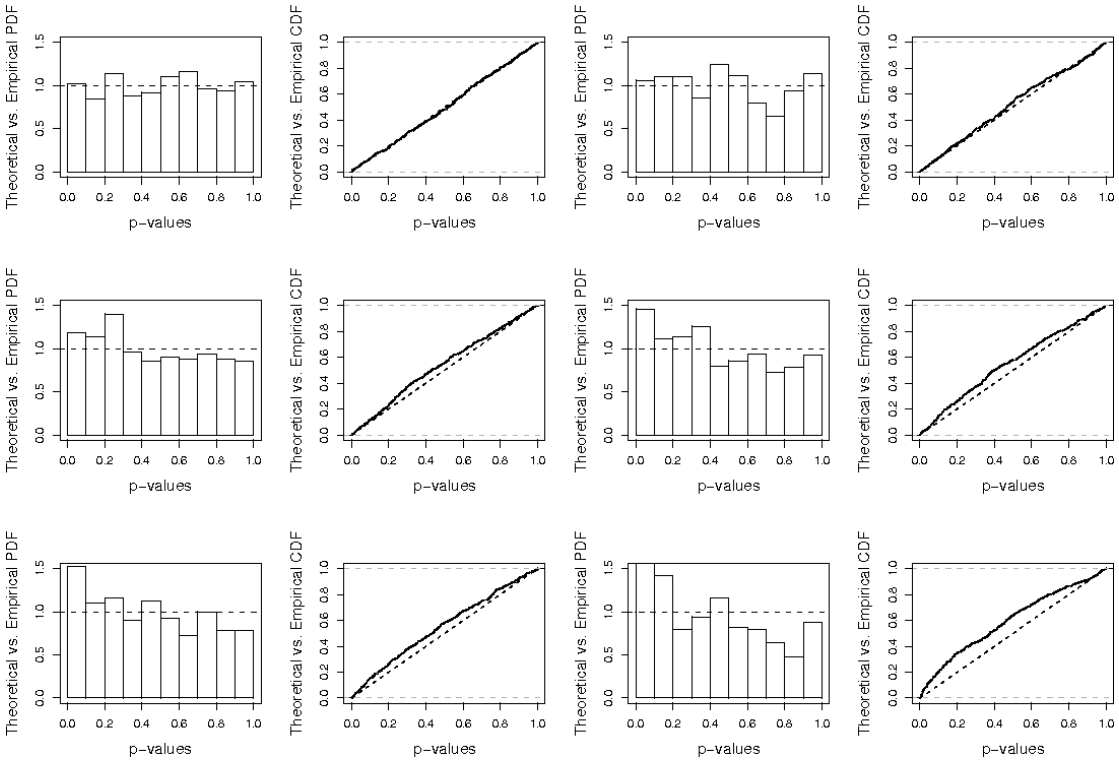

Figure 6.2: Null distribution of the estimated p–values for sequences of configuration matrices simulated according to separable TOPC–AR(1) models and $\rho_{\text{crit}} = 0.1$: The three rows correspond to the different AR(1) models (top: $\psi_1 = 0.2$, middle: $\psi_1 = 0.5$, bottom: $\psi_1 = 0.8$). The first and second column show histograms and empirical distribution functions of the estimated p–values (solid lines) for the case $n = 750$ and $\sigma = 0.1$. In columns three and four, the same graphs are displayed for the challenging case $n = 150$ and $\sigma = 0.5$.

Table 6.1: Achieved significance level and power for the considered TOPC–AR(1) models when $\rho_{\text{crit}} = 0.1$: The nominal significance level is taken as 0.05. The test yields good results in most cases although it can be too liberal in the challenging situation where both the dependence and the variability of the data are large.

		$\psi = 0.2$		$\psi = 0.5$		$\psi = 0.8$	
n	σ	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
150	0.1	0.034	1	0.05	1	0.06	1
	0.3	0.056	1	0.056	0.998	0.12	0.818
	0.5	0.056	0.776	0.06	0.528	0.124	0.362
500	0.1	0.054	1	0.052	1	0.082	1
	0.3	0.054	1	0.07	1	0.102	1
	0.5	0.056	1	0.074	0.984	0.112	0.778
750	0.1	0.056	1	0.052	1	0.074	1
	0.3	0.052	1	0.07	1	0.13	1
	0.5	0.058	1	0.074	0.984	0.15	0.91

is shown in the first and second column and the third and fourth column show the case where $n = 150$ is combined with $\sigma = 0.5$. The dashed lines show the corresponding graphs for the uniform distribution, and the rows in this figure correspond to the different AR(1) models (top: $\psi_1 = 0.2$, middle: $\psi_1 = 0.5$, and bottom: $\psi_1 = 0.8$). The estimated p–values follow the uniform distribution quite closely in most cases, although a trend is visible indicating that the lower tail of the distribution starts to become too large as the dependence increases. However, in comparison with the corresponding graphs in Chapter 5, i.e. the bottom row of Figure 5.8 ($\psi_1 = 0.2$) and Figure 5.9 ($\psi_1 = 0.8$) which were in fact generated with a small $\sigma = 0.1$, it can be seen that the amendments to Algorithm 5.1 considerably improve the test procedure.

Table 6.1 summarises the achieved significance level $\hat{\alpha}$ and power $\hat{\beta}$ for a nominal significance level $\alpha = 0.05$. The results are consistent across different sample sizes. Moreover, for $\psi_1 = 0.2$ and $\psi_1 = 0.5$ the achieved significance level $\hat{\alpha}$ is reasonably close to 0.05 in all cases. For $\psi_1 = 0.8$, however, $\hat{\alpha}$ depends on the standard deviation: it takes roughly the desired value for $\sigma = 0.1$ but increases for larger standard deviations. If both the dependence and the variability of the data are high, then Algorithm 6.1 will therefore be too liberal. In terms of its power, the test is very good with most values of $\hat{\beta}$ being close to one. Only if the sample size is small relative to the standard deviation, does the power take smaller values. This phenomenon has already been observed in Table 5.2.

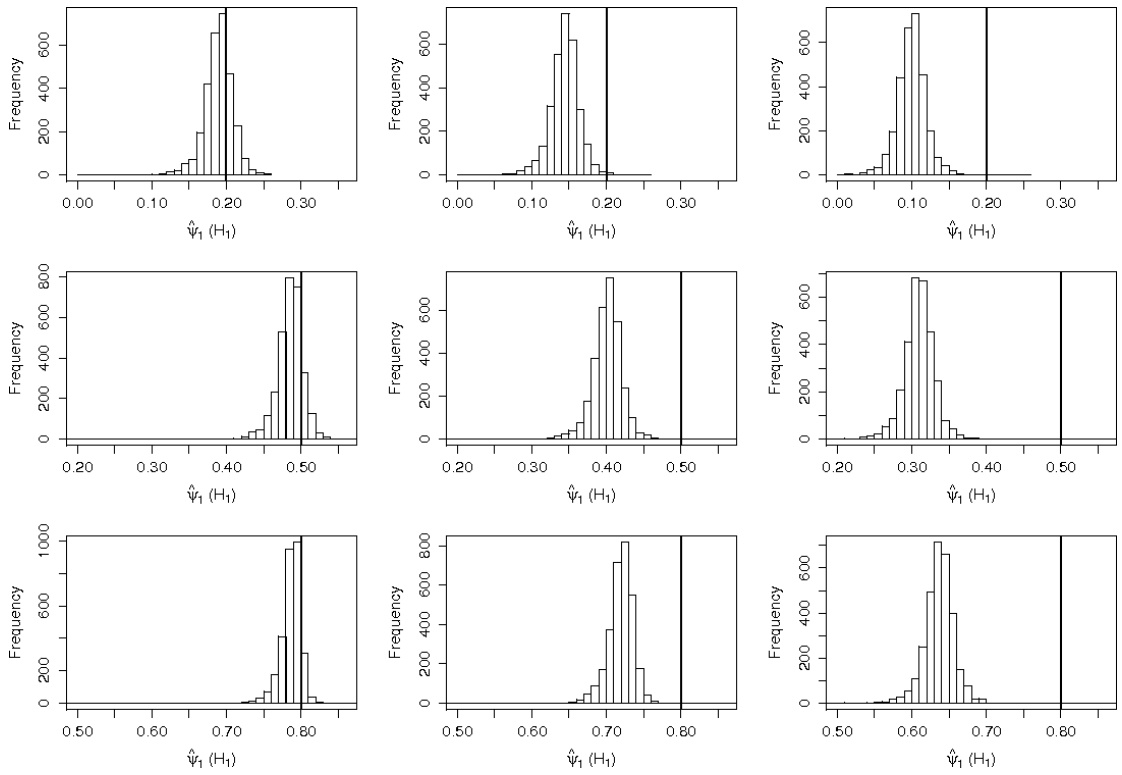

Figure 6.3: Histograms of the estimated AR(1) parameters under the alternative: Each histogram displays 3,000 estimates of $\hat{\psi}_1^{H_1}$ obtained in step 4 of Algorithm 6.1. The rows correspond to $\psi_1 = 0.2$ (top), $\psi_1 = 0.5$ (middle), and $\psi_1 = 0.8$ (bottom), and the columns correspond to $\sigma = 0.1$ (left), $\sigma = 0.3$ (middle) and $\sigma = 0.5$ (right). For small σ, the histograms are roughly centred around the AR(1) parameter according to which the configuration matrices were simulated. For larger values of σ, $\hat{\psi}_1^{H_1}$ take smaller values.

Estimated AR Coefficients

As we use 500 Monte Carlo iterations for each of the 54 considered parameter combinations, Algorithm 6.1 is carried out 27,000 times, and each time step 4 yields an estimate of ψ_1 under both H_0 and H_1. (In this first part of the simulation study, ψ_2 is kept fixed at zero when the optimisation (6.16) is carried out.) Of particular interest are thereby the estimates $\hat{\psi}_1^{H_1}$ obtained under the alternative as they should reflect the true dependence structure of the tangent vectors better than the estimates $\hat{\psi}_1^{H_0}$ which are obtained under a more restrictive model. Figure 6.3 shows histograms of the $\hat{\psi}_1^{H_1}$ which are grouped according to the true value of ψ_1 in landmark space and according to the employed standard deviation. The rows thereby correspond to $\psi_1 = 0.2$ (top), $\psi_1 = 0.5$ (middle), and $\psi_1 = 0.8$ (bottom), and the columns correspond to $\sigma = 0.1$ (left), $\sigma = 0.3$ (middle)

and $\sigma = 0.5$ (right). Each histogram is therefore based on 3,000 values of $\hat{\psi}_1^{H_1}$. It can be seen that the estimates are roughly centred around the true parameter in landmark space (displayed as a bold vertical line) for $\sigma = 0.1$. For larger values of the standard deviation, $\hat{\psi}_1^{H_1}$ tends to be smaller than the employed ψ_1. Figure F.2 in Appendix F shows that the tangent projection for highly dispersed data reduces the correlation of the data so that this effect is not necessarily a flaw in the estimation procedure.

The Test Statistic

As mentioned in Section 6.3.1, the null hypothesis of equal mean shapes induces M linear constraints on the parameter vector. According to Wilks' theorem (cf. Appendix D), the asymptotic null distribution of $-2\log(\lambda(\boldsymbol{V}, \boldsymbol{W}))$ would be χ^2_M if the data were independent. Figure 6.4 shows that the χ^2–approximation is very good for our case as

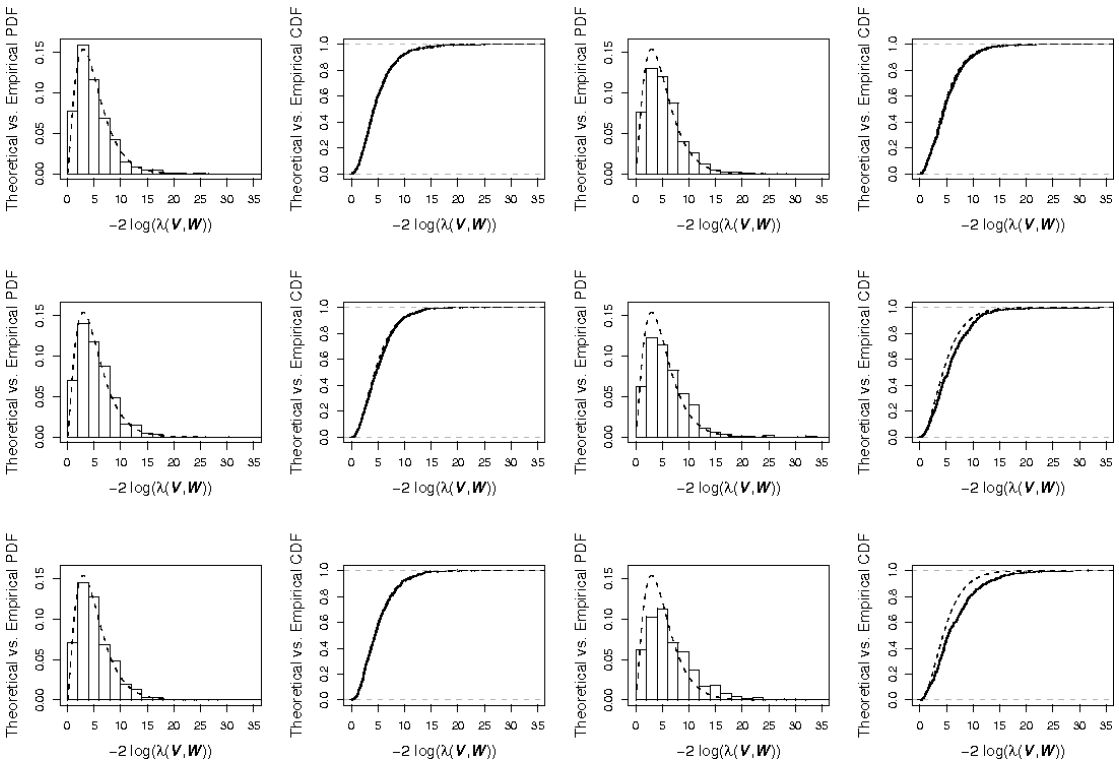

Figure 6.4: Null distribution of the observed values of the test statistic for sequences of configuration matrices simulated according to separable TOPC–AR(1) models: The rows correspond to the different AR(1) models (top: $\psi_1 = 0.2$, middle: $\psi_1 = 0.5$, bottom: $\psi_1 = 0.8$). The first and second column show histograms and empirical distribution functions of the observed values of the LR statistic (solid lines) for the case $n = 750$ and $\sigma = 0.1$. They are compared with the density and distribution function of χ^2_5–distribution (dashed lines). Columns three and four correspond to $n = 150$ and $\sigma = 0.5$.

Table 6.2: Achieved significance level and power for the considered TOPC–AR(1) models when $\rho_{\text{crit}} = 0.01$: The nominal significance level is taken as 0.05. The results are the same as those in Table for $\psi_1 = 0.2$ and very similar for $\psi_1 = 0.5$. For highly correlated data, the choice of the hyperparameter ρ_{crit} has an impact on the achieved significance level and power, and $\rho_{\text{crit}} = 0.01$ yields better results than $\rho_{\text{crit}} = 0.1$

		$\psi = 0.2$		$\psi = 0.5$		$\psi = 0.8$	
n	σ	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
150	0.1	0.034	1	0.056	1	0.072	1
	0.3	0.056	1	0.062	0.996	0.064	0.73
	0.5	0.056	0.776	0.06	0.528	0.098	0.296
500	0.1	0.054	1	0.052	1	0.058	1
	0.3	0.054	1	0.07	1	0.088	1
	0.5	0.056	1	0.074	0.984	0.104	0.716
750	0.1	0.056	1	0.052	1	0.058	1
	0.3	0.052	1	0.07	1	0.06	1
	0.5	0.058	1	0.074	0.984	0.112	0.886

well. This is true for all cases we consider (including the ones described later where the configuration matrices are simulated according to a non–normal model) and suggests that our test statistic is asymptotically pivotal. However, the precise condition for this to hold need further investigation.

Comparison With the Results Obtained Using $\rho_{\text{crit}} = 0.01$

As mentioned earlier, we repeat the above simulations using $\rho_{\text{crit}} = 0.01$ instead of $\rho_{\text{crit}} = 0.1$ in step 5 of Algorithm 6.1. This should increase the block length l of the resamples $\boldsymbol{\mathcal{V}}_c^* = \{\mathcal{B}_{\text{X}}(I_1^{\text{X}}, l), \ldots, \mathcal{B}_{\text{X}}(I_{n_B}^{\text{X}}, l)\}$, $\boldsymbol{\mathcal{W}}_c^* = \{\mathcal{B}_{\text{Y}}(I_1^{\text{Y}}, l), \ldots, \mathcal{B}_{\text{Y}}(I_{n_B}^{\text{Y}}, l)\}$ which are obtained during the block bootstrap procedure, and as it can be seen from Figure 6.1, the change should be most noticeable for the cases where the sequences $\{\boldsymbol{X}_t\}_{t=1}^n$ and $\{\boldsymbol{Y}_t\}_{t=1}^n$ have been generated using $\psi_1 = 0.8$. In fact, for $\psi_1 = 0.8$ the maximal of the 9,000 block lengths (18 parameter combinations with $\psi_1 = 0.8 \times 500$ Monte Carlo iterations) obtained with $\rho_{\text{crit}} = 0.1$ is $l = 10$ whereas the maximal block length obtained using $\rho_{\text{crit}} = 0.01$ is $l = 25$, and 3,716 times the block length increases by nine or more. For $\psi_1 = 0.5$, only eight block lengths increase by five or more when $\rho_{\text{crit}} = 0.01$ is used, and for $\psi_1 = 0.2$ all block lengths stay the same at $l = 5$. Substantial changes in performance are therefore only to be expected for $\psi_1 = 0.8$ – in particular as we use the same seed for the resampling procedure to allow for a direct comparison.

Table 6.2 summarises the performance of Algorithm 6.1 when $\rho_{\text{crit}} = 0.01$ is used. For reasons outlined above, the results for $\psi_1 = 0.2$ and $\psi_1 = 0.5$ are very similar to those in Table 6.1. For $\psi_1 = 0.8$, however, decreasing ρ_{crit} in general has a beneficial effect on the achieved significance level. Compared with the respective columns of Table 6.1, it can be seen that the block bootstrap procedure with the increased block lengths is more robust against large standard deviations. Although the achieved significance levels for the combination of $\psi_1 = 0.8$ and $\sigma = 0.5$ are still higher than desired, the results have in particular improved for combinations of $\psi_1 = 0.8$ and $\sigma = 0.3$.

6.3.2.4 Additional Simulations

We carry out some more simulations to assess the performance of Algorithm 6.1 under more challenging conditions.

Non–Normal Data

We repeat all of the above calculation for (4×3)–dimensional temporally dependent configuration matrices which have been generated as described in Section 6.3.2.1 but using as a starting point an $\big(n \times 12\big)$ matrix $\boldsymbol{Z}^*$ with $(\boldsymbol{Z}^*)_{ij} \overset{i.i.d.}{\sim} t_3$ $(i = 1, \ldots, n; j = 1, \ldots, 12)$, where t_3 denotes the t–distribution with three degrees of freedom. We choose this distribution as it is rather extreme in its large tails but still has a finite variance. The left–hand side of Table 6.3 provides a summary of the results for a nominal significance level of $\alpha = 0.05$. The top half corresponds to Table 6.1 where the hyperparameter value $\rho_{\text{crit}} = 0.1$ was used, and the bottom half corresponds to Table 6.2 where $\rho_{\text{crit}} = 0.01$. Overall, the results are similar to those for the normal trajectories. However, the achieved significance levels for small and moderate correlations ($\psi = 0.2$ and $\psi = 0.5$) are more scattered around the desired value of 0.05, and for the large correlation of $\psi = 0.8$, the tendency to yield too large values of $\hat{\alpha}$ is slightly worse. Moreover, the resulting values for the power are also slightly worse which is probably due to the large tails of the t_3–distribution. The two groups of tangent vectors in the t_3–case will therefore exhibit a larger overlap than in the normal case. Like before, decreasing the value of ρ_{crit} (and hence increasing the block length) has in general a beneficial effect on the results.

Table 6.3: Achieved significance level and power for the more challenging cases of t_3–based configuration matrices and AR(2) dependence structures: The left–hand side shows the results of the simulations which are based on the t_3–distribution with an AR(1) dependence structure, and the right–hand side shows the results of the simulations which are based on configurations generated according to a TOPC–AR(2) model. The top half in both cases corresponds to Table 6.1 where $\rho_{\text{crit}} = 0.1$ and the bottom half corresponds to Table 6.2 where $\rho_{\text{crit}} = 0.01$.

		t_3 & $\psi_1 = 0.2$		t_3 & $\psi_1 = 0.5$		t_3 & $\psi_1 = 0.8$		$\boldsymbol{\psi} = (0.2, -0.5)^T$		$\boldsymbol{\psi} = (0.4, 0.1)^T$		$\boldsymbol{\psi} = (-1.0, -0.75)^T$	
n	σ	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
150	0.1	0.054	1	0.056	1	0.068	1	0.06	1	0.052	1	0.188	1
	0.3	0.052	0.968	0.076	0.742	0.132	0.384	0.09	1	0.052	1	0.356	1
	0.5	0.05	0.3	0.06	0.23	0.09	0.17	0.078	0.896	0.066	0.514	0.25	0.878
500	0.1	0.062	1	0.052	1	0.068	1	0.092	1	0.057	1	0.234	1
	0.3	0.058	1	0.07	1	0.124	0.862	0.1	1	0.062	1	0.344	1
	0.5	0.026	0.848	0.044	0.558	0.124	0.282	0.096	1	0.052	0.972	0.296	1
750	0.1	0.036	1	0.052	1	0.098	1	0.072	1	0.067	1	0.198	1
	0.3	0.048	1	0.07	1	0.132	0.95	0.076	1	0.076	1	0.386	1
	0.5	0.04	0.964	0.044	0.558	0.108	0.384	.076	1	0.05	1	0.308	1
150	0.1	0.054	1	0.068	1	0.084	1	0.057	1	0.037	1	0.166	1
	0.3	0.052	0.968	0.082	0.704	0.096	0.314	0.068	1	0.042	0.99	0.29	1
	0.5	0.05	0.3	0.06	0.23	0.098	0.126	0.08	0.906	0.04	0.514	0.268	0.884
500	0.1	0.062	1	0.052	1	0.064	1	0.045	1	0.035	1	0.134	1
	0.3	0.058	1	0.07	1	0.08	0.78	0.056	1	0.042	1	0.338	1
	0.5	0.026	0.848	0.044	0.558	0.084	0.258	0.086	1	0.062	0.972	0.296	1
750	0.1	0.036	1	0.052	1	0.086	1	0.07	1	0.047	1	0.108	1
	0.3	0.048	1	0.07	1	0.104	0.928	0.056	1	0.052	1	0.36	1
	0.5	0.04	0.964	0.044	0.558	0.098	0.332	0.076	1	0.07	1	0.308	1

It can be concluded that the performance of Algorithm 6.1 gets slightly worse if the trajectories do not follow a normal distribution. However, for the considered t_3–based trajectories, the results are still very good for small and moderate correlations.

Normal Data with AR(2) Dependence

Finally, we consider the case where the trajectories $\{\boldsymbol{X}_t\}_{t=1}^n$ and $\{\boldsymbol{Y}_t\}_{t=1}^n$ are sequences of (4×3)–configuration matrices whose vectorised versions follow a separable TOPC–AR(2) model, i.e. both trajectories are generated as described in Section 6.3.2.1 but here we use a temporal correlation matrix which is based on a vector $\boldsymbol{\psi} = (\psi_1, \psi_2)^T$ of AR(2) parameters. In particular, we consider three cases of temporal dependence structures, namely $\boldsymbol{\psi} = (0.2, -0.5)^T$, $\boldsymbol{\psi} = (0.4, 0.1)^T$ and $\boldsymbol{\psi} = (-1.0, -0.75)^T$. Figure 6.5 shows the corresponding autocorrelation functions. It can be seen that the last case implies a very strong correlation and hence provides the most challenging scenario. Here, the horizontal lines represent the constant functions $f(s) = \pm 0.1$ (dotted) and $g(s) = \pm 0.01$ (dashed), where s denotes the lag between two observations.

The right–hand side of Table 6.3 summarises the results. As before, the top half corresponds to $\rho_{\text{crit}} = 0.1$ and the bottom half corresponds to $\rho_{\text{crit}} = 0.01$. Here, the choice of ρ_{crit} does not have an overall positive impact on the results. The smaller value $\rho_{\text{crit}} = 0.01$ does tend to decrease the achieved significance level but in some cases this results in $\hat{\alpha}$

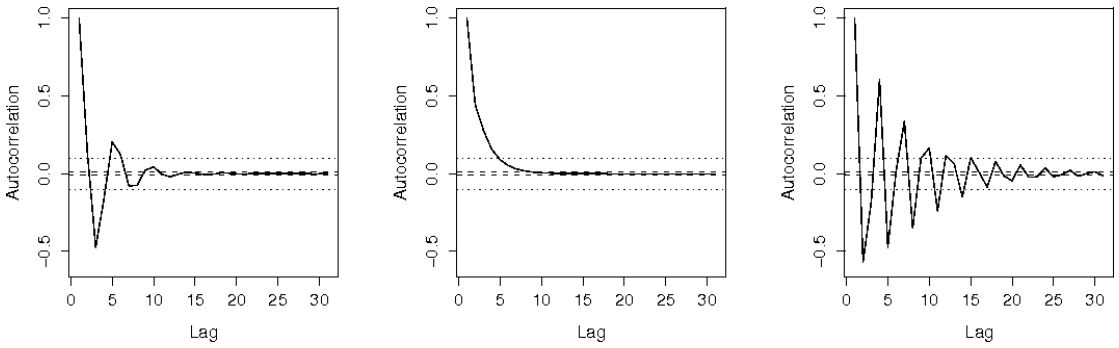

Figure 6.5: Autocorrelation functions of the employed AR(2) models: Three different combinations of AR(2) parameters are chosen, namely $\boldsymbol{\psi} = (0.2, -0.5)^T$ (left), $\boldsymbol{\psi} = (0.4, 0.1)^T$ (middle), and $\boldsymbol{\psi} = (-1.0, -0.75)^T$ (right). The dotted lines in each plot are the constant functions $f(\text{Lag}) = \pm 0.1$, and the dashed lines corresponds to $g(\text{Lag}) = \pm 0.01$.

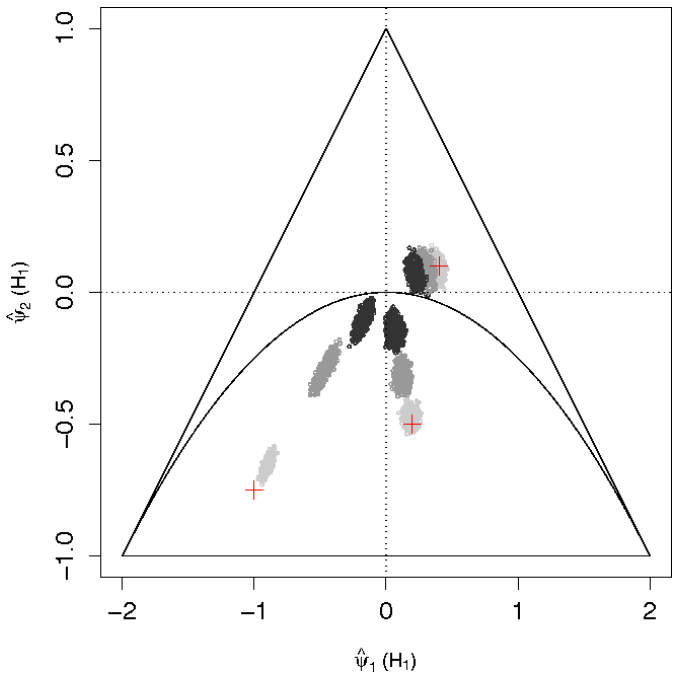

Figure 6.6: Plot of the estimated AR(2) parameters under the alternative: Each point cloud displays 3,000 estimates of $\hat{\boldsymbol{\psi}}^{H_1}$ obtained from the observed data in step 4 of Algorithm 6.1. The different shades of grey correspond to the standard deviation (lightgrey: $\sigma = 0.1$, grey: $\sigma = 0.3$, black: $\sigma = 0.5$). The true values of the AR(2) parameters are displayed in red. The triangle correspond to the stationarity region of AR(2) processes (cf. (6.5)), and the curve corresponds to the cases where $\psi_2 = -\psi_1^2/4$. Parameter combinations below this curve correspond to complex roots of the characteristic equation of an AR(2) process (cf. Section 2.2.2.2).

being below the nominal level $\alpha = 0.05$. It can be seen that the strength of the underlying correlation has the biggest impact on the results. For both choices $\boldsymbol{\psi} = (0.4, 0.1)^T$ and $\boldsymbol{\psi} = (0.2, -0.5)^T$, the results are reasonably good. Unfortunately, this changes for the very strong correlation implied by $\boldsymbol{\psi} = (-1.0, -0.75)^T$ where the achieved significance levels are dramatically larger than 0.05, and this effect is particularly strong for data with low concentration (i.e. large values of σ).

To further investigate the impact of the standard deviation and the strength of the correlation on the results, we consider their effect on the estimated AR(2) parameter vector obtained under H_1 in step 4 of Algorithm 6.1. Here, we restrict our attention to the case where $\rho_{\text{crit}} = 0.1$. As before in Section 6.3.2.3, we therefore inspect the estimates $\hat{\boldsymbol{\psi}}^{H_1} = (\hat{\psi}_1^{H_1}, \hat{\psi}_2^{H_1})^T$ obtained in 27,000 (54 parameter combination $\times$ 500 Monte Carlo iterations) runs of Algorithm 6.1. Figure 6.6 shows these estimates. The true AR(2)

parameters in landmark space are displayed in red here, and the shades of grey correspond to the different values of the standard deviation (lightgrey: $\sigma = 0.1$, grey: $\sigma = 0.3$, black: $\sigma = 0.5$). Figure 6.6 is therefore the counterpart of Figure 6.3. As before, it can be seen that the standard deviation has a big effect on the estimates in that larger standard deviations lead to estimates $\hat{\boldsymbol{\psi}}^{H_1}$ which are closer to the centre of the stationarity region which implies independence. Like Figure F.2, Figure F.3 demonstrates that the tangent projection plays an important role in this effect.

One peculiarity of Figure 6.6 is that the obtained estimates $\hat{\boldsymbol{\psi}}^{H_1}$ even for the small standard deviation of $\sigma = 0.1$ are not centred around the true AR(2) parameter vector in landmark space for $\boldsymbol{\psi} = (-1.0, -0.75)^T$ (note that $\boldsymbol{\psi} = (-1.0, -0.75)^T$ implies a larger correlation than $\psi_1 = 0.8$). We rerun some of the simulations (i.e. those with $n = 750$, $\rho_{\text{crit}} = 0.1$ and $\boldsymbol{\mu}_Y \in \{\breve{\boldsymbol{X}}_0, \breve{\boldsymbol{X}}_2\}$) with a smaller value of standard deviation, namely $\sigma = 0.05$. Figure F.4 in Appendix F indicates that in that case the correlation of the data in the observed tangent space is the same as in landmark space because the resulting estimates $\hat{\boldsymbol{\psi}}^{H_1}$ are now centred around $\boldsymbol{\psi} = (-1.0, -0.75)^T$. The corresponding achieved significance level is $\hat{\alpha} = 0.08$ which suggests that Algorithm 5.1 is able to deal with very large correlations if the variability in the data is sufficiently small. In fact, with $\sigma = 0.03$, the achieved significance level reduces to $\hat{\alpha} = 0.044$ which is very close to the nominal value $\alpha = 0.05$.

6.4 Application to the DNA Data

We now apply Algorithm 6.1 to the DNA data described in Section 1.2.3. As we saw in the previous section, a very high correlation between the configurations within each group causes the test to reject the null hypothesis of equal mean shapes even if it is true. To avoid this effect, we slightly thin the data and use every $5th$ molecular configuration within each time series. This reduces the number of configurations within each group from 2,500 to 500 which is still large and compared to thinned data we considered in Section 5.4, the groups will contain much more information about the underlying mean shapes which is obviously desirable.

We first consider the AGA/AFA pair of DNA duplexes and perform GPA on the pooled set of 1,000 configuration matrices of the thinned data. After the projection (5.12), we then have two groups $\boldsymbol{\mathcal{V}} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{500}\}$ and $\boldsymbol{\mathcal{W}} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{500}\}$ of temporally dependent tangent vectors in $M = 59$ dimensions which provide the basis for the subsequent block bootstrap procedure. To assess how adequate the underlying assumptions of Algorithm 6.1 are for the DNA data, we estimate the AR(2) parameters and the principal components of the between column covariance matrix separately for each of the two groups using a one–sample version of the ML procedure described in Section 6.1.2.2.

For a general sample $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of n temporally dependent vectors in p dimensions, the resulting MLEs of the AR(2) parameters thereby satisfy

$$\hat{\boldsymbol{\psi}} = \arg \min_{\boldsymbol{\psi} \in \mathcal{T}_2^{\mathrm{AR}}} |\widehat{\boldsymbol{\Sigma}}_C|^{-n/2} \{\sigma_a^{-2n} \big((1-\psi_2^2)^2 - (1+\psi_2)^2 \psi_1^2\big)\}^{p/2}, \tag{6.20}$$

where $\widehat{\boldsymbol{\Sigma}}_C = 1/n\big(\boldsymbol{X} - \mathbf{1}_n \bar{\boldsymbol{x}}^T\big)^T \boldsymbol{\Sigma}_T^{-1} \big(\boldsymbol{X} - \mathbf{1}_n \bar{\boldsymbol{x}}^T\big)$ and $\boldsymbol{X}$ is the $(n \times p)$ matrix of the stacked vectors in the sample (Dryden *et al.*, 2009). Carrying out this estimation for the considered DNA pair yields $\hat{\boldsymbol{\psi}}_{\mathrm{V}} = (0.410, 0.197)^T$ for the AGA duplex and $\hat{\boldsymbol{\psi}}_{\mathrm{W}} = (0.380, 0.179)^T$ for the AFA duplex. The similarity of these estimates is reassuring for our assumption of a pooled covariance structure. Moreover, both vectors are well within the stationarity region of an AR(2) process (cf. Figure 6.6) which is also reassuring. Table 6.4 shows that this can be observed for all pairs of damaged/undamaged duplexes. Moreover, the estimated vectors of AR(2) parameters are similar for all molecules.

Table 6.4: Maximum likelihood estimates of the underlying AR(2) parameters of each duplex under a separable TOPC–AR(2) model: For each damaged/undamaged pair, GPA on the pooled sample of the (thinned) configuration matrices was performed, and the resulting tangent vectors were projected into a 59–dimensional subspace using (5.12). For each duplex, the MLE (6.20) was then calculated separately.

duplex	$\hat{\boldsymbol{\psi}} = (\hat{\psi}_1, \hat{\psi}_2)^T$	duplex	$\hat{\boldsymbol{\psi}} = (\hat{\psi}_1, \hat{\psi}_2)^T$
AGA	$(0.410, 0.197)^T$	AFA	$(0.380, 0.179)^T$
AGC	$(0.409, 0.195)^T$	AFC	$(0.393, 0.171)^T$
AGG	$(0.437, 0.219)^T$	AFG	$(0.412, 0.182)^T$
TGA	$(0.415, 0.201)^T$	TFA	$(0.395, 0.167)^T$
TGC	$(0.383, 0.164)^T$	TFC	$(0.392, 0.170)^T$
TGG	$(0.412, 0.186)^T$	TFG	$(0.441, 0.202)^T$

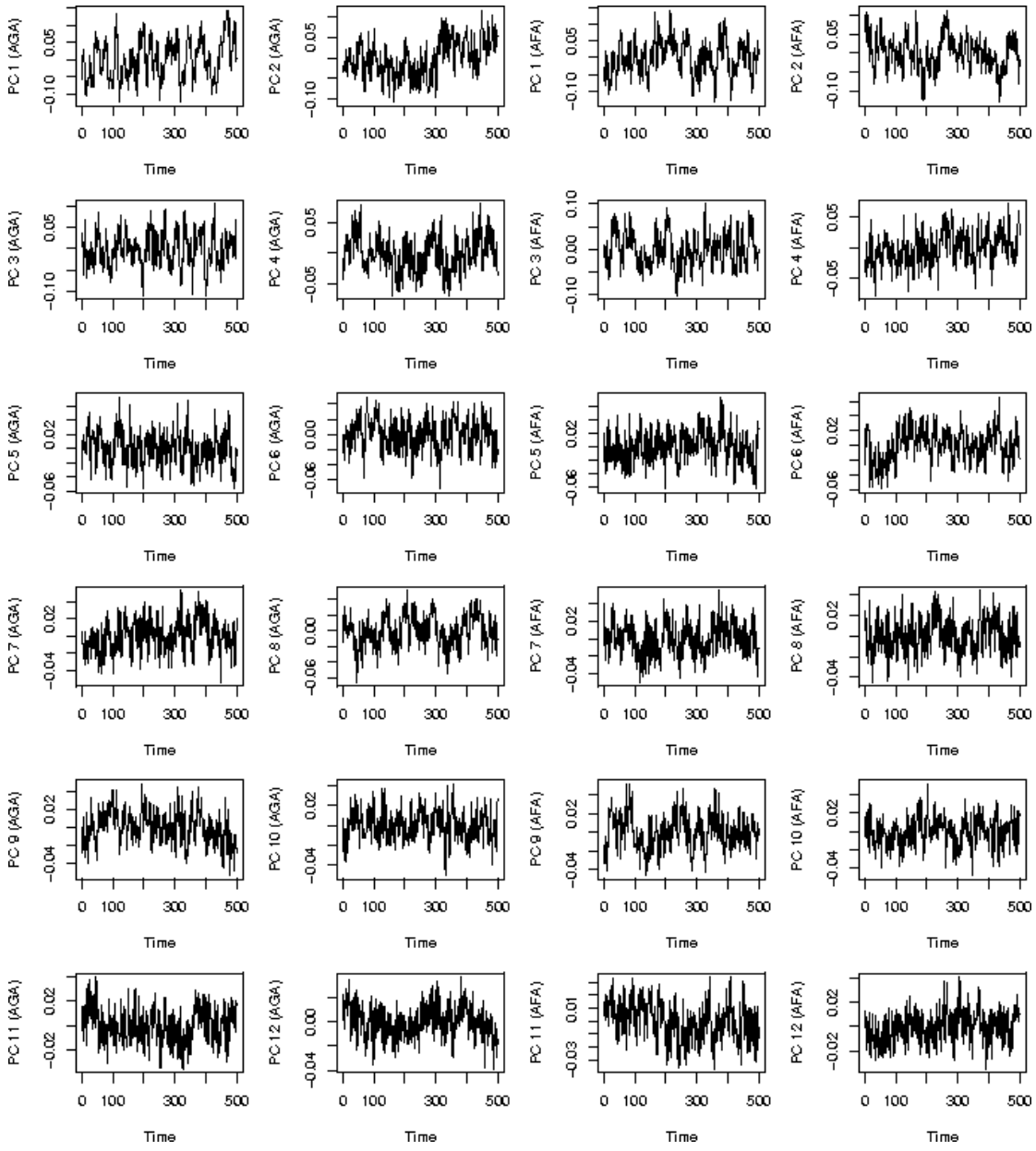

Figure 6.7: Time series of the principal components of shape for the AGA/AFA pair of DNA duplexes: GPA was performed on the pooled sample of 1,000 configuration matrices of the two (thinned) times series. For each group of projected tangent data, the principal components of shape were obtained separately using the ML estimation under the separable TOPC–AR(2) model. The left–hand side shows the time series of the first twelve PC scores for the AGA duplex, and the right–hand side shows the corresponding time series for the AFA duplex.

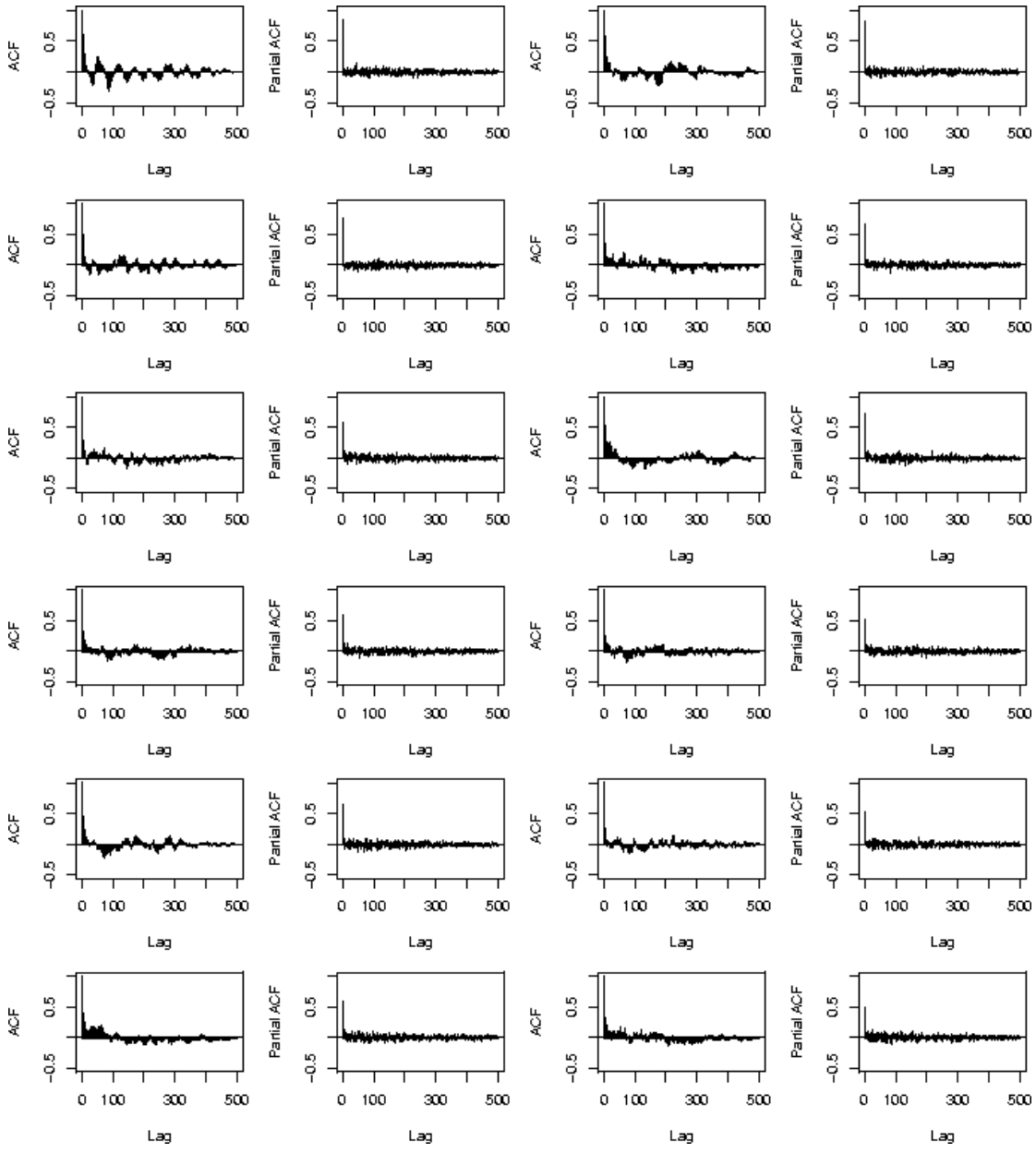

Figure 6.8: Autocorrelation and partial autocorrelation for shape PC scores of the AGA duplex: The first row shows the correlation structure of the scores on the first two shape PCs, the second row those of the third and fourth shape PC and so on. It can be seen that the correlation structure is somewhat different on each shape PC. Each partial autocorrelation function shows only a few spikes which is in line with a low order AR(p) model.

Based on $\hat{\boldsymbol{\psi}}_{\mathrm{V}}$ and $\hat{\boldsymbol{\psi}}_{\mathrm{W}}$, the between column covariance matrices can be estimated as described above for both molecules AGA and AFA. Dryden *et al.* (2009) call the corresponding eigenvalues *principal components of shape* as the between–column structure of the tangent vectors summarises the dependence between the shape dimensions. Figure 6.7 shows the time series of the first twelve shape PC scores for both duplexes. The left–hand side corresponds to the AGA duplex, and the right–hand side corresponds to the AFA duplex.

Figure 6.8 row–wise shows the corresponding empirical autocorrelation functions and empirical partial autocorrelation functions for AGA. The plot of the empirical partial autocorrelation function thereby is a plot of the lag s against the sth estimated AR coefficient, $\hat{\psi}_{ss}$ say, which would occur in an AR(s) process. This plot helps to determine the order of an AR(p) process: if $\hat{\psi}_{ss}$ is close to zero for all $s > \tilde{s}$, then the order of the autoregressive process is likely to be $\tilde{s}$ (e.g. Box *et al.*, 2008, pp.66).

It can be seen that the autocorrelation on each shape PC is somewhat different which is not ideal for our assumption of separability. However, modelling an AR(2) dependence is in line with the partial autocorrelation functions which exhibit only a few spikes. The corresponding plots for AFA are provided in Figure F.5 in Appendix F. An observation which can be made in both Figure 6.8 and Figure F.5 is that the correlation decreases for higher PCs, and the estimates $\hat{\boldsymbol{\psi}}_{\mathrm{V}} = (0.410, 0.197)^T$ and $\hat{\boldsymbol{\psi}}_{\mathrm{W}} = (0.380, 0.179)^T$ strike a compromise between the different observed correlations.

Table 6.5: Estimated p–Values, observed values of the (transformed) test statistic and other parameters of Algorithm 6.1 when applied to the DNA data: The tests show very strong evidence against the null hypothesis of equal mean shapes for all damaged/undamaged pairs of DNA duplexes. The pooled estimates of the AR(2) parameters obtained under the alternative are in line with those in Table 6.4.

pair	$\hat{p}$	$-2\log\lambda^{\mathrm{obs}}$	l	$\hat{\boldsymbol{\psi}}^{H_1} = (\hat{\psi}_1^{H_1}, \hat{\psi}_2^{H_1})^T$
A.A	0.001	140.976	10	$(0.400, 0.187)^T$
A.C	0.001	742.554	10	$(0.416, 0.190)^T$
A.G	0.001	353.909	10	$(0.431, 0.203)^T$
T.A	0.001	553.790	10	$(0.414, 0.189)^T$
T.C	0.001	507.125	10	$(0.396, 0.172)^T$
T.G	0.001	379.815	10	$(0.434, 0.198)^T$

Table 6.6: Estimated p–values, observed values of the (transformed) test statistic and other parameters of Algorithm 6.1 when applied within the DNA duplexes: In all cases, the result correctly indicates no evidence against the null hypothesis of equal mean shapes at the 5% significance level.

duplex	$\hat{p}$	$-2\log\lambda^{\text{obs}}$	duplex	$\hat{p}$	$-2\log\lambda^{\text{obs}}$	duplex	$\hat{p}$	$-2\log\lambda^{\text{obs}}$
AGA	0.138	86.562	AGC	0.456	62.949	AGG	0.125	88.90
AFA	0.4	66.096	AFC	0.237	76.181	AFG	0.27	73.857
TGA	0.742	49.788	TGC	0.188	83.862	TGG	0.107	87.134
TFA	0.246	74.555	TFC	0.273	74.363	TFG	0.571	58.093

We now apply Algorithm 6.1 to test for the equality of the mean shapes for all pairs of damaged/undamaged DNA molecules. Due to the overall superiority of $\rho_{\text{crit}} = 0.01$ we choose this hyperparameter value for determining the block length. Table 6.5 summarises the results. All tests are thereby based on $B = 1,000$ bootstrap iterations. Like before in Section 5.4, the results indicate very strong evidence against the null hypothesis of equal mean shapes for all pairs. Even for the AGA/AFA pair for which the lowest value of the observed test statistic occurs, does the estimated p–value take the smallest possible value for $B = 1,000$. The fourth column of Table 6.5 shows the pooled MLEs of the AR(2) parameters which are obtained under the alternative in step 4 of Algorithm 6.1. It can be seen that they are in line with the corresponding estimates in Table 6.4. The resulting block length is 10 in each case.

We saw before that small estimated p–values can also be a result of model misspecification. We therefore apply Algorithm 6.1 to the data within each DNA strand. To do so, we divide the (thinned) time series for each duplex into ten parts of 50 iterations and form the two groups by assigning alternating group membership to the ten parts, i.e. group 1 of each DNA consists of observations $1, \ldots, 50, 101, \ldots, 150, \ldots, 401, \ldots, 450$. Table 6.6 shows that Algorithm 6.1 correctly finds no evidence against the equality of equal mean shapes for these within–DNA data. Although the assumption of separability is not met for the DNA application, the proposed test procedure is therefore able to distinguish between cases where the null hypothesis is true and those where it is not.

The test results are in line with those in the previous chapter. Overall, it therefore is to be concluded that the oxidative guanine damage FapydG does induce significant changes

in the mean shapes of the DNA molecules. To investigate further changes in the DNA structure and dynamics brought about by the damage is left for further work. Some ideas are mentioned in Section 7.3.

6.5 Summary

In this chapter we proposed a bootstrap algorithm for testing the equality of the underlying population mean shapes for two groups of temporally dependent configuration matrices. This algorithm is based on the Procrustes tangent space of the pooled sample. It can be seen as a direct generalisation of the bootstrap algorithm proposed in the previous chapter as the applied test statistic is a generalisation of a quantity which is proportional to the Mahalanobis squared distance (which for equal sample sizes is turn proportional to the James statistics and hence yields the same results in a bootstrap procedure), and the used resampling method is a generalisation of the independent bootstrap applied in Algorithm 5.1.

In a simulation study, the superiority of the amended bootstrap algorithm over Algorithm 5.1 is demonstrated. Whereas Algorithm 5.1 breaks down even for very small correlations within the groups, the new algorithm works well in most cases. Only for very large correlations and large standard deviations does it become too liberal and tends to spuriously reject the null hypothesis.

When applied to the DNA data, the results are consistent with those obtained in the previous chapter and suggest very strong evidence against the null hypothesis of equal mean shapes for the damaged/undamaged pairs of DNA molecules. This is interesting in that it could be linked to the binding activity of the duplexes towards the repair protein which changes the lesion FapydG back to guanine.

Chapter 7

Discussion and Further Work

In this thesis we developed statistical methods for modelling and comparing molecular shapes. In particular, the fuzzy nature of molecules and the fact that molecules constantly undergo conformational changes are important features in molecular modelling and cannot be addressed with methods from classical statistical shape analysis. In two separate parts, we therefore developed novel techniques which are specifically designed to incorporate these two molecular properties.

7.1 Modelling and Comparing Continuous Molecular Shapes

In Chapters 3 and 4, we considered the situation where each molecule is given in the form of a marked point set where the points represent the atom positions and the marks are values of a molecular property measured at these positions. In order to obtain a continuous representation of molecular shape, we used kriging of the given marks. This yields a predicted molecular field based on which a comparison of different molecules can be carried out. Although kriging has been mentioned before in the chemoinformatics literature (e.g. Fang *et al.*, 2004, and Pen *et al.*, 2006, use kriging to introduce a correlation structure for the errors when a linear model is set up to predict a molecular property such as the boiling point from topological measurements of molecules), its application to the prediction of a molecular field provides a novel tool in structural alignment.

We assumed that the marked point sets of the given molecules are noisy pointwise observations of a common underlying reference field which cannot be observed. This assumption is particularly reasonable for molecules which bind to the same target as the underlying reference field can in that case be interpreted as the negative imprint of the binding pocket of the receptor. With the additional assumption that the reference field is stationary, the constant mean can be set to zero for the purpose of molecular alignment. The predicted fields therefore take a form which allows us to view them as members of the reproducing kernel Hilbert space associated with the employed covariance function.

In order to compare two molecular fields, we proposed a modification of the Carbo similarity index which is well–established in the structural alignment community. The original (L_2–)Carbo index essentially generalises Pearson's correlation coefficient to continuous functions and measures the similarity of two fields by an overlap integral, i.e. by the inner product in the space of Lebesgue square–integrable functions L_2. In this thesis, we introduced a kernelised version of the (L_2–)Carbo index which has the advantage that the field overlap can be calculated without expensive numerical integration.

The alignment of two molecules with respect to the Kernel Carbo index was carried out within a Bayesian framework. Markov chain Monte Carlo sampling and posterior inference were used to obtain a rotation/translation invariant notion of molecular similarity. As the rigid–body parameters are integrated out, our alignment method is similar to that of Green & Mardia (2006). However, it avoids estimating correspondences between the atoms of different molecules which poses a substantial difference to previously proposed methods. With our field–based approach, the absence of atom correspondences is counterbalanced by the spatial distribution of the marks so that it is only necessary to determine whether or not an atom belongs to the matching part of the molecules. This is a considerable advantage as correspondences do not exist in every application. In the simulation study in Chapter 3, we also demonstrated that the mask vectors can identify contamination points so that our alignment method is somewhat robust to outliers.

Another approach for the pairwise alignment of unlabelled point sets which does not require correspondences has been formulated by Durrleman *et al.* (2007) in the context of aligning brain shapes. They view the given sets of point coordinates as segmented

lines and formulate a distance between the point sets in terms of a distance between the lines using "currents" and reproducing kernel Hilbert spaces. The resulting similarity index bears some algebraic similarities to the Kernel Carbo index. However, Durrleman *et al.* (2007) do not incorporate marks or the possibility that only subsets of the given point sets match but they do use non–rigid deformations.

Our alignment method worked well for pairwise alignments of the steroid molecules in Chapter 4. We also demonstrated that the resulting rotation/translation invariant discrepancy values are chemically meaningful in that they are associated with the binding activities of the steroids towards a common receptor protein. In order to assess which parts of the molecular fields differ between the activity classes, we developed an extension of the pairwise alignment to the simultaneous superposition of several unlabelled marked point sets. This extension can be seen as a field–based version of the generalised partial Procrustes algorithm in statistical shape analysis as it determines the optimal matching parameters for each point set in turn. It is related to the model proposed by Dryden *et al.* (2007) where an iterative optimisation of the matching parameters is carried out with respect to an unknown reference configuration. Contrary to that, a hidden reference configuration is integrated out in the fully model–based Bayesian approach of Ruffieux & Green (2009) which is an extension of the pairwise method of Green & Mardia (2006).

The fact that our field–based approach provides the opportunity to naturally incorporate additional information is of particular advantage in the multiple comparison setting because the resulting mean fields allow straightforward post–processing. For example in the steroid application we used an exploratory t–test to determine the regions where the steric mean field of the three activity classes differ the most.

7.2 Comparing Dynamic Molecular Shapes

Chapters 5 and 6 were concerned with comparing the sample mean shapes of DNA molecules with the aim of investigating whether or not the oxidative guanine lesion FapydG significantly changes the mean shape of a DNA molecule. Motivated by the success of

bootstrap hypothesis tests for comparing the means shapes of two groups of configuration matrices as demonstrated by Amaral *et al.* (2007) and Preston & Wood (2009b) in the context of planar shapes and the multidimensional scaling (MDS) approach to shape analysis, respectively, we considered the bootstrap framework for this problem.

In Chapter 5, we proposed a fast bootstrap algorithm for testing the equality of the underlying mean shapes of two groups of configurations matrices for the case where the data within each group are independent. Our algorithm is based on the Procrustes tangent space of the pooled sample. Once obtained, this space is fixed and resampling is carried out conditional on the "observed tangent space" which considerably reduces the computational cost of the algorithm. Moreover, as the tangent space is a Euclidean approximation of the shape space it offers a natural way of transforming the data to the null hypothesis by centering both groups of tangent vectors. We use the James statistic in tangent space which is asymptotically pivotal so that our bootstrap algorithm meets both guidelines Hall & Wilson (1991) suggested for general bootstrap hypothesis tests. Both, centering the tangent vectors and the use of the James statistic were also considered by Preston & Wood (2009b) in the context of the MDS approach to shape analysis. Our Procrustes–based work can therefore be viewed as complementary to their paper, and in particular if the reflection information of the data should be retained, it provides a valuable tool for testing for the equality of mean shapes.

In a simulation study we showed that our fast bootstrap algorithm works very well in terms of both achieved significance level and power if the data are independent. However, if even a small temporal correlation is present the test tends to be too liberal. This shortcoming is not surprising as both the resampling procedure and the test statistic are designed under the independence assumption. Motivated by the DNA data which are highly correlated within each group (molecule), we therefore extended our fast bootstrap algorithm to accommodate temporal dependence in Chapter 6.

The new test statistic is based on the separable time–orthogonal principal component (TOPC) model of Dryden *et al.* (2009). Assuming an AR(2) dependence structure on each principal component (PC), we proposed a likelihood ratio statistic. This statistic can be seen as a direct generalisation of the Mahalanobis squared distance which can be

derived as a monotone transformation of a likelihood ratio statistic in the independent case. To generalise the resampling procedure, we applied the circular block bootstrap of Politis & Romano (1992) where blocks of consecutive observations are resampled to preserve the dependence structure in the data. We chose the circular block bootstrap for our algorithm because it does not suffer from edge effects so that all observations have equal probabilities to be part of a resampled block. To determine the block length we used an *ad–hoc* criterion which ensures that the block length depends on the estimated AR(2) correlation structure of the data.

Simulations showed that the amended bootstrap test algorithm works well in most situations. This includes cases where the configuration matrices have been simulated based on the heavy tailed t_3–distribution. Only if the data exhibit very large correlations in combination with a large standard deviation, does it become too liberal. Overall, the amended bootstrap procedure is able to cope much better with temporal correlations than the algorithm proposed in Chapter 5. One of the advantages of this is that no or only a small degree of thinning needs to be applied before the test can be carried out so that the results are based on more information. Unfortunately, this improvement comes with an increased computational cost because two optimisations have to be carried out at each bootstrap iteration.

When applied to the DNA data, both the independent and the dependent version of our bootstrap test indicated that all pairs of damaged/undamaged DNA duplexes have significantly different mean shapes. This is interesting as it could be linked to the binding activity towards the repair protein which changes the damage FapydG back to the original base guanine. To investigate whether or not these differences are consistent across all pairs, we carried out further significance tests to compare all pairs of undamaged and all pairs of damaged DNA. If, for example, the damaged versions do not significantly differ from each other, this could be a clue as to how the repair protein recognises the damage. Unfortunately, no consistent differences between the mean shapes of the damaged and undamaged DNA could be found. In fact, for all possible pairs of the twelve duplexes, both bootstrap tests yielded very small estimated p–values which suggests that all duplexes have significantly different mean shapes.

Our analysis in Chapter 6 did, however, reveal a potentially interesting difference between the damaged and undamaged molecules: Table 6.4 suggests that, overall, the temporal dependence of the configurations of the damaged duplexes is slightly smaller than that of the undamaged molecules. A possible explanation for this is that the hydrogen bonding of FapydG to the base in the complementary strand is weaker than the one between guanine and the corresponding base (Jiranusronkul & Laughton, 2008). This observation is also in line with Table 1 in their paper which shows that the average fluctuation of the damaged molecules about the starting configuration is consistently higher (in terms of the root mean square deviation) than that of the undamaged molecules.

Finally, as mentioned by Dryden *et al.* (2009), assuming an AR(2) dependence structure is a reasonable starting point for the DNA application as the molecular dynamics simulations are based on Newtonian mechanics where the position of a particle can be determined based on its speed and acceleration (cf. Section 1.2.1). Therefore, given the past, the previous observation could be used to estimate the speed and the two previous observations could be used to estimate the acceleration.

7.3 Further Work

Both parts of this thesis generate questions which could be further investigated. In this section we outline some ideas for further work. In the context of aligning unlabelled marked point sets, a possible amendment of our methodology is the use of cokriging for cases where each coordinate in a point set is associated with a vector of marks rather than a scalar. While the dynamic weighted average approach for the steric and electrostatic fields (cf. Section 4.2.1) seems appropriate in the molecular context as it mimics real–life molecular recognition, it might be beneficial to account for covariances between the predicted fields in other examples. In that case, it would also be of interest to introduce separate mask vectors for each type of mark to allow for the possibility that the regions of high similarity differ between the different types of fields. Conceptually, this can easily be incorporated in the Bayesian framework. It would, however, be computationally very demanding.

In the simulation study described in Section 3.6, we showed that using the true covariance function of the underlying reference field in the Kernel Carbo calculations is not essential for good alignment results, and that it is more important that the point sets are samples from similar parts of the underlying reference field. For the steroid data, it is the common core structure of the four carbon rings which satisfies this requirement so that the alignment works well although the isotropic Gaussian covariance function for the field predictions might not be the correct choice. For general applications where it is not known whether or not the nearness–requirement is satisfied, it may be more important to estimate the covariance function of the underlying reference field correctly. However, if outliers or other contamination points are present, then the results can be distorted. Applying an outlier detection method will therefore be a beneficial pre–processing step before the (pooled) empirical semivariogram is obtained. In Appendix B we describe an *ad–hoc* method to do so based on a leave–one–out procedure which also yields adequate starting values for the mask parameters in the MCMC algorithm. We did not extensively study the performance of this method so that this is subject to further work.

Our MCMC algorithm does sometimes get stuck in a local maximum of the posterior distribution. This problem could potentially be alleviated by using "soft" mask vectors whose entries take values between zero and one instead of being binary. The predicted field for each labelled point set would then be based on all points at all iterations, and the current soft mask vectors could be incorporated by multiplying the kriging weights by the corresponding entries. This idea is based on Rangarajan *et al.* (1997) who describe a soft matching algorithm for unlabelled point sets using Procrustes analysis and a soft labelling matrix. For our case, softening the mask vectors would have the additional computational advantage that the kriging weights do not have to be calculated anew each time a new mask vector is accepted.

While it would be good to incorporate molecular dynamics in the alignment procedure, this would be computationally very demanding. A simpler way to account for molecular flexibility is described by Schmidler (2009) who extends the notion of shape as described in Section 2.1.1 to a notion of *flexible* shape based on changepoint analysis. For two labelled configuration matrices $\boldsymbol{X} \in \mathbb{R}^{k \times m}$ and $\boldsymbol{Y} \in \mathbb{R}^{k \times m}$ with corresponding rows $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ $(i = 1, \ldots k)$, a changepoint allows for hinge–like motions of parts of the con-

figurations. It is defined as an index j ($j = 1, \ldots, k-1$) such that the transformation between $\boldsymbol{x}_j$ and $\boldsymbol{y}_j$ is different from the transformation between $\boldsymbol{x}_{j+1}$ and $\boldsymbol{y}_{j+1}$. For example, if a single changepoint is present, then two sets $(\boldsymbol{\Gamma}_1, \boldsymbol{\gamma}_1)$ and $(\boldsymbol{\Gamma}_2, \boldsymbol{\gamma}_2)$ of matching parameters need to be estimated which describe the transformation between the first j rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the last $k-j$ rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Schmidler (2009) formulates a Bayesian framework to determine the number of changepoints and their locations and successfully applies his method to the structural alignment of proteins. A similar extension of our field–based approach would be desirable.

There are also several potential areas for further work which arise from the second part of this thesis. Firstly, the employed test statistic in Algorithm 6.1 has been derived under quite a restrictive model so that generalisations are desirable. For example, one could incorporate the possibility of unequal covariance structures of the two groups. Moreover, we saw for the DNA data that the assumption of separability is not met (cf. Figures 6.8 and F.5). For the one–sample case, Dryden *et al.* (2009) account for this by defining a non–separable version of the TOPC model where the correlation structure on each PC is allowed to follow a different model. They also describe an iterative algorithm to obtain approximate ML estimates, and it may be worth investigating its use for calculating a LR statistic of the form (6.16). However, within the bootstrap framework, two of these optimisations would have to be carried out at each bootstrap iteration which might be prohibitively slow.

With respect to the resampling procedure in Algorithm 6.1, a possible improvement lies in the choice of the block length. Step 4 is an *ad–hoc* method which seems to work well in most cases considered in this thesis. For the circular block bootstrap as proposed by Politis & Romano (1992), however, the chosen block length does not need to be an integer divisor of the sample size. This could refine the choice of the block length and improve the results. Moreover, our method of determining the block length could be compared with the method proposed by Hall *et al.* (1995) where (as a step prior to applying a block bootstrap algorithm to the entire sample) the performance of different block lengths is assessed by an empirical version of the mean squared error evaluated in terms of the estimates obtained from subsamples of the given time series and the estimate obtained from the entire sample.

Further analysis of the DNA data could also be carried out. Our test results from both Chapters 5 and 6 indicate that the oxidative guanine lesion FapydG does induce significant changes to the mean shapes of the given DNA duplexes. It would be interesting to find out what other changes are brought about by replacing G with F. Table 6.4 indicates that there might be a consistent difference of the dependence structure between the damaged and undamaged duplexes. One could investigate if this, combined with the different mean shapes, leads to different possible extreme configurations. A particularly relevant question is whether or not the damaged versions can assume the configuration required for the binding complex with the repair protein more easily than their undamaged counterparts. However, to fully investigate this, it might be necessary to incorporate more atoms than just the phosphorus atoms of the DNA backbones.

Like the steroid molecules, the DNA duplexes are obviously also continuous objects. Despite the absence of marks in the DNA dataset, one could account for this by describing the two strands of each duplex as continuous curves which interpolate the given coordinates of the phosphorus atoms. This approach has been applied in the context of modelling facial shapes by Barry & Bowman (2008) who fit B–splines to a set of landmarks which describe the faces of children with cleft lip and palate and a similarly aged control group. Barry & Bowman (2008) then use the spline coefficients to compare the facial shapes of the two groups over time. These coefficients are inherently invariant under rotation and translation so that they can be used instead of the tangent coordinates of the original landmarks. A similar approach could be used for the DNA data although the test statistic in Algorithm 6.1 would be harder to interpret in that case.

Finally, it would be interesting to apply our methods to other datasets. As the field–based matching described in Chapters 3 and 4 does not require any predefined point–by–point correspondences, it could be an approach to resolve the alignment problem for a fairly broad range of applications. Examples include matching organs in medical images (Rangarajan *et al.*, 1997) or matching two views of the same object from different cameras (Cross & Hancock, 1998). Moreover, if the objects are represented by closed outlines which may be occluded by other objects in an image, then a similar approach can be used to perform a partial matching of the outlines. In fact, this has successfully been done in Cao *et al.* (2009). The bootstrap algorithm proposed in Chapter 5 can be

applied to all situations where two groups of independent configuration matrices are to be compared with respect to their mean shapes. This is a frequently occurring situation and examples include landmark data for human faces which are divided into age groups (cf. Evison & Vorder Bruegge, 2008; Preston & Wood, 2009b) or comparing the brain shapes of schizophrenia and normal patients (Bookstein, 1996). If the configurations are observed over time and can be assumed to follow a stationary process, then the bootstrap procedure proposed in Chapter 6 should be applied.

Bibliography

Abrahamsen, P. (1997): A review of Gaussian random fields and correlation functions. *Technical report*, Norwegian Computing Center, Oslo.

Abramowitz, M. & Stegun, I. A. (1964): *Handbook of Mathematical Functions*. Washington D.C.: National Bureau of Standards.

Adler, R. J. (1981): *The Geometry of Random Fields*. Chichester: Wiley.

Amaral, G. J. A., Dryden, I. L., & Wood, A. T. A. (2007): Pivotal bootstrap methods for k–sample problems in directional statistics and shape analysis. *Journal of the American Statistcial Association*, 102, 695–707.

Anzali, S., Gasteiger, J., Holzgrabe, U., Polanski, J., Sadowski, J., Teckentrup, A., & Wagener, M. (1998): The use of self–organising neural networks in drug design. In: H. Kubinyi, G. Folkers, & Y. C. Martin (eds.) *3D QSAR in Drug Design*, 273–299. London: Kluwer/ESCOM, 2nd edition.

Arnold, S. F. (1981): *The Theory of Linear Models and Multivariate Analysis*. Chichester: Wiley.

Aronszajan, N. (1950): Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 337–404.

Barry, S. J. E. & Bowman, A. W. (2008): Linear mixed models for longitudinal shape data with applications to facial modeling. *Biostatistics*, 9, 555–565.

Beckman, K. B. & Ames, B. N. (1997): Oxidative decay of DNA. *Journal of Biological Chemistry*, 272, 19633–19636.

Bender, A. & Glen, R. C. (2004): Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry*, 2, 3204–3218.

Berlinet, A. & Thomas-Agnan, C. (2004): *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. London: Kluwer Academic Press.

Bhattacharya, B. & Habtzghi, D. (2002): Median of the p value under the alternative hypothesis. *The American Statistician*, 56, 202–206.

Bhattacharya, R. & Patrangenaru, V. (2003): Large–sample theory on intrinsic and extrinsic sample means on manifolds, I. *Annals of Statistics*, 31, 1–29.

Blanley, J. M. & Dixon, J. S. (1993): A good ligand is hard to find: Automated docking methods. *Perspectives in Drug Discovery and Design*, 1, 301–319.

Bookstein, F. L. (1986): Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Sciences*, 1, 181–242.

Bookstein, F. L. (1996): Biometrics, biomathematics and the morphometric synthesis. *Bulletin of Mathematical Biology*, 58, 313–365.

Bose, A. & Politis, D. N. (1993): A review of the bootstrap for dependent samples. *Technical report*, Department of Statistics, Purdue University.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008): *Time Series Analysis*. Hoboken: Wiley, 4th edition.

Cao, Y., Zhang, Z., Czogiel, I., Dryden, I., & Wang, S. (2009): 2D nonrigid partial shape matching using MCMC and contour subdivision. *Submitted for Publication*.

Carbo, R., Leyda, L., & Arnau, M. (1980): An electron density measure of the similarity between two compounds. *International Journal of Quantum Chemistry*, 17, 1185–1189.

Carlstein, E. (1986): The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *The Annals of Statistics*, 14, 1171–1179.

Case, D. A., Cheatham III, T. E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K. M., Onufriev, A., Simmerling, C., Wang, B., & Woods, B. (2005): The Amber biomolecular simulation program. *Journal of Computational Chemistry*, 26, 1668–1688.

Chatfield, C. (1996): *The Analysis of Time Series (An Introduction)*. London: Chapman & Hall, 5th edition.

Chernick, M. R. (1999): *Bootstrap Methods*. Chichester: Whiley.

Chib, S. & Greenberg, E. (1995): Understanding the Metropolis–Hastings algorithm. *The American Statistician*, 49, 327–335.

Christianini, N. & Shawe-Taylor, J. (2000): *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.

Chung, K. L. (1974): *A Course in Probability Theory*. New York: Academic Press, 2nd edition.

Coats, E. A. (1998): The CoMFA steroids as a benchmark data set for development of 3D QSAR methods. *Perspectives in Drug Discovery and Design*, 12, 119–213.

Cramer, R. D., III, Patterson, D. E., & Bunce, J. D. (1988): Comparative molecular field analysis (CoMFA). 1. effect of shape on binding of steroids on carrier proteins. *Journal of the American Chemical Society*, 110, 5959–5967.

Cressie, N. A. C. (1993): *Statistics for Spatial Data*. Chichester: Wiley.

Crippen, G. M. (1987): Voronoi binding site models. *Journal of Computational Chemistry*, 8, 943–955.

Cross, A. D. J. & Hancock, E. R. (1998): Graph matching with a dual–step EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 1236–1253.

Cryer, D. J. & Chan, K.-S. (2008): *Time Series Analysis with Applications in R*. Berlin: Springer, 2nd edition.

Czogiel, I., Dryden, I. L., & Brignell, C. J. (2008): Bayesian alignment of continuous molecular shapes using random fields. In: S. Barber, P. D. Baxter, A. Gusnato, & K. V. Mardia (eds.) *The Art and Science of Statistical Bioinformatics*, 85–88. Leeds: Leeds University Press.

Czogiel, I., Dryden, I. L., & Brignell, C. J. (2009): Bayesian alignment of unlabelled marked point sets using random fields. *Submitted for Publication*.

Davison, A. C. & Hinkley, D. V. (1997): *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.

Díaz-García, J. A., Gutierrey Jáimez, R., & Mardia, K. V. (1997): Wishart and pseudo–Wishart distributions and some applications to shape theory. *Journal of Multivariate Analysis*, 63, 73–87.

Dobson, C. M. (2004): Chemical space and biology. *Nature*, 432, 824–828.

Dryden, I. L., Hirst, J. D., & Melville, J. M. (2007): Statistical analysis of unlabelled point sets: comparing molecules in chemoinformatics. *Biometrics*, 63, 237–251.

Dryden, I. L., Kume, A., Le, H., & Wood, A. T. A. (2008): A multidimensional scaling approach to shape analysis. *Biometrika*, 95, 779–798.

Dryden, I. L., Kume, A., Le, H., & Wood, A. T. A. (2009): Statistical inference for functions of the covariance matrix in the stationary Gaussian time–orthogonal principal components model. *Annals of the Institute of Statistical Mathematics*, to appear.

Dryden, I. L., Kume, A., Le, H., Wood, A. T. A., & Laughton, C. A. (2002): Size–and–shape analysis of DNA molecular dynamic simulations. In: K. V. Mardia, R. D. Aykroyed, & P. McDonnel (eds.) *Proceedings in Statistics of Large Datasets*. University of Leeds.

Dryden, I. L. & Mardia, K. V. (1993): Multivariate shape analysis. *Sankhyā, Series A*, 55, 460–480.

Dryden, I. L. & Mardia, K. V. (1998): *Statistical Shape Analysis*. Chichester: Wiley.

Dryden, I. L. & Zempléni, A. (2006): Extreme shape analysis. *Journal of the Royal Statistical Society, Series C*, 55, 103–121.

Durrleman, S., Pennec, X., Trouvé, A., & Ayache, N. (2007): Measuring brain variability via sulcal lines registration: a diffeomorphic approach. In: N. Ayache, S. Ourselin, & A. Maeder (eds.) *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 4791 of LNCS. Brisbane.

Efron, B. (1979): Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.

Efron, B. (1984): Better boostrap confidence intervals. *Technical report*, Stanford University, Department of Statistics.

Efron, B. & Tibshirani, R. J. (1986): Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54–75.

Efron, B. & Tibshirani, R. J. (1993): *An Introduction to the Bootstrap*. London: Chapman & Hall.

Evison, M. P. & Vorder Bruegge, R. W. (2008): The Magna Database: a database of three–dimensional facial images for research in human identification and recognition. *Forensic Science Communications*, 10, [Web].

Fang, K.-T., Ying, H., & Liang, Y.-Z. (2004): New approach by kriging models to problems in QSAR. *Journal of Chemical Information and Computer Sciences*, 44, 2106–2113.

Fisher, N. I., Hall, P., Jing, B.-Y., & Wood, A. T. A. (1996): Improved pivotal methods for constructing confidence regions with directional data. *Journal of the American Statistical Association*, 91, 1062–1070.

Friedman, D. A. (1981): Bootstrapping regression models. *Annals of Statistics*, 9, 1218–1228.

Friedman, D. A. (1984): On bootstrapping two–stage least–squares estimates in stationary linear models. *Annals of Statistics*, 12, 827–842.

Gelman, A. (1996): Inference and monitoring convergence. In: W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (eds.) *Markov chain Monte Carlo in Practice*, 131–144. London: Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004): *Bayesian Data Analysis*. London: Chapman & Hall, 2nd edition.

Geman, S. & Geman, D. (1984): Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Giudice, E. & Levery, R. (2002): Simulations of nucleic acids and their complexes. *Accounts of Chemical Research*, 35, 350–357.

Goldstein, H., Poole, C., & Safko, J. (2002): *Classical Mechanics*. London: Addison Wesley, 3rd edition.

Good, A. C. (1995): 3D molecular similarity indices and their application in QSAR studies. In: P. M. Dean (ed.) *Molecular Similarity in Drug Design*, 24–56. London: Blackie Academic & Professional.

Good, A. C., Hodgkin, E. E., & Richards, W. G. (1992): The utilisation of Gaussian functions for the rapid evaluation of molecular similarity. *Journal of Chemical Information and Computer Sciences*, 32, 188–191.

Good, A. C., So, S., & Richards, W. G. (1993): Structure–activity relationships from molecular similarity matrices. *Journal of Medicinal Chemistry*, 36, 433–438.

Goodall, C. R. (1991): Procrustes methods in the statistical analysis of shape (with discussion). *Journal of the Royal Statistical Society, Series B*, 53, 285–339.

Goodall, C. R. & Bose, A. (1987): Models and Procrustes methods for the analysis of shape differences. In: R. M. Heiberger (ed.) *Proceedings of the 19th INTERFACE Symposium*, 445–454. Fairfax Station, Interface Foundation.

Gower, J. C. (1975): Generalized Procrustes analysis. *Psychometrika*, 40, 33–50.

Grant, J. A., Gallardo, M. A., & Pickup, B. T. (1996): A fast method of molecular shape comparison: A simple application fo a Gaussian description of molecular shape. *Journal of Computational Chemistry*, 17, 1653–1666.

Grant, J. A. & Pickup, B. T. (1995): A Gaussian description of molecular shape. *Journal of Physical Chemistry*, 99, 3503–3510.

Green, P. J. (2001): A primer on Markov Chain Monte Carlo. In: O. E. Barndorff-Nielsen, D. R. Cox, & C. Klüppelberg (eds.) *Complex Stochastic Systems*, 1–62. London: Chapman & Hall.

Green, P. J. & Mardia, K. V. (2006): Bayesian alignment using hierarchical models, with application in protein bioinformatics. *Biometrika*, 93, 235–254.

Grenander, U. & Rosenblatt, M. (1957): *Statistical Analysis of StationaryTime Series*. Chichester: Wiley.

Hall, M. (1998): *Combinatorial Theory*. Chichester: Wiley, 2nd edition.

Hall, P. (1985): Resampling a coverage pattern. *Stochastic Processes and Their Applications*, 20, 231–246.

Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*. Berlin: Springer.

Hall, P., Horowitz, J. L., & Jing, B. (1995): On blocking rules for the bootstrap with dependent data. *Biometrika*, 82, 561–574.

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

Handcock, M. S. & Wallis, J. R. (1994): An approach to statistical spatial–temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association*, 89, 368–390.

Haskard, K. A. (2007): *An anisotropic Matérn spatial covariance model: REML estimation and properties*. Ph.D. thesis, School of Agriculture, Food and Wine, University of Adelaide.

Hastings, W. K. (1970): Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Henderson, H. V. & Searle, S. R. (1979): Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian Journal of Statistics*, 7, 65–81.

James, G. S. (1954): Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika*, 41, 19–43.

Jeffreys, H. (1946): An invariant form of the prior probability in estiamtion problems. *Proceedings of the Royal Society of London: Series A (Mathematical and Physical Sciences)*, 186, 453–461.

Jiranusronkul, S. & Laughton, C. A. (2008): Destabilisation on DNA duplexes by oxidative damage at guanine: implications for lesion recognition and repair. *Journal of the Royal Society Interface*, 5, 191–198.

Kass, R. E. & Wasserman, L. (1996): The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343–1370.

Kearsley, S. K. & Smith, G. M. (1990): An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlaps. *Tetrahedron Computer Methodology*, 3, 315–633.

Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., & Shoichet, B. K. (2007): Relating protein pharmocology by ligand chemistry. *Nature Biotechnology*, 25, 197–206.

Kendall, D. G. (1977): The diffusion of shape. *Advances in Applied Probability*, 9, 428–430.

Kendall, D. G. (1984): Shape manifolds, Procrustean metrix and complex projective spaces. *Bulletin of the London Mathematical Society*, 16, 81–121.

Kendall, D. G. (1989): A survey of the statistical theory of shape. *Statistical Science*, 4, 87–120.

Kendall, D. G., Barden, D., Carne, T. K., & Le, H. (1999): *Shape and Shape Theory*. Chichester: Wiley.

Kent, J. T. (1994): The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society, Series B*, 56, 285–299.

Kent, J. T. (1995): Current issues for statistical inference in shape analysis. In: K. V. Mardia & C. A. Gill (eds.) *Current Issues in Statistical Shape Analysis*, 167–175. University of Leeds.

Kent, J. T. & Mardia, K. V. (1997): Consistency of Procrustes estimators. *Journal of the Royal Statistical Society, Series B*, 59, 281–290.

Kent, J. T. & Mardia, K. V. (2001): Shape, Procrustes tangent projections and bilateral systems. *Biometrika*, 88, 496–485.

Kim, K. H. (1995): Comparative molecular field analysis (CoMFA). In: P. M. Dean (ed.) *Molecular Similarity in Drug Design*, 291–331. London: Blackie Academic & Professional.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983): Optimization by simulated annealing. *Science*, 220, 671–680.

Krige, D. G. (1951): A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of Chemical, Metallurgical, and Mining Society of South Africa*, 52, 119–139.

Kristof, W. & Wingersky, B. (1971): Generalization of the orthogonal Procrustes rotation procedure to more than two matrices. In: *Proceedings of the 79th Annual Convention of the American Psychological Association*, 89–90.

Künsch, H. R. (1989): The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17, 1217–1261.

Lahiri, S. N. (2003): *Resampling Methods for Dependent Data*. Berlin: Springer.

Langron, S. P. & Collins, A. J. (1985): Perturbation theory for generalized Procrustes analysis. *Journal of the Royal Statistical Society, Series B*, 47, 277–284.

Le, H.-L. (1995): Mean size–and–shapes and mean shapes: a geometric point of view. *Advances in Applied Probability*, 27, 44–55.

Le, H.-L. (1998): On the consistency of Procrustean mean shapes. *Advances in Applied Probability*, 30, 53–63.

Le, H.-L. & Kendall, D. G. (1993): The Riemmanian structure of Euclidean shape spaces: a novel environment for statistics. *Annals of Statistics*, 21, 1225–1271.

Le, H.-L. & Kume, A. (2000): The Fréchet mean shape and the space of the means. *Advances in Applied Probability*, 32, 101–113.

Lemmen, C. & Lengauer, T. (2000): Computational methods for the structural alignment of molecules. *Journal of Computer-Aided Molecular Design*, 14, 215–232.

Liu, R. Y. & Singh, K. (1992): Moving blocks jackknife and bootstrap capture weak dependence. In: R. Lapage & L. Billard (eds.) *Exploring the Limits of Bootstrap*. New York: Wiley.

Mardia, K. V., Kent, J. T., & Bibby, J. (1979): *Multivariate Analysis*. London: Academic Press.

Masek, B. B., Merchant, A., & Matthew, J. B. (1993): Molecular shape comparison of angiotensin II receptor antagonists. *Journal of Medicinal Chemistry*, 36, 1230–1238.

Matheron, G. (1962): Traite de Geostatistique Appliquee, Tome I. *Memoires du Bureau de Recherches Geologiques et Minieres*, No. 14, Editions Technip, Paris, 1230–1238.

Matheron, G. (1963): Principles of geostatistics. *Economic Geology*, 58, 1246–1266.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977): Dynamics of folded proteins. *Nature*, 267, 585–590.

McMahon, A. J. & King, P. M. (1997): Optimization of Carbó molecular similarity index using gradient methods. *Journal of Computational Chemistry*, 18, 151–158.

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

Mezey, P. G. (1993): *Shape in Chemistry*. Cambridge: VHC Publishers.

Mezey, P. G. (1995): Methods of molecular shape similarity and topological shape design. In: P. M. Dean (ed.) *Molecular Similarity in Drug Design*, 241–268. London: Blackie Academic & Professional.

Miles, R. E. (1965): On random rotations in $\mathbb{R}3$. *Biometrika*, 52, 636–639.

Mosier, C. I. (1939): Determining a simple structure when loadings for certain tests are known. *Psychometrika*, 4, 149–162.

Naimark, M. A. (1964): *Linear Representations of the Lorentz Group*. Oxford: Pergamon Presss.

O'Higgins, P. & Dryden, I. L. (1993): Sexual dimorphism in hominoids: further studies of craniofacial shape differences in *pan*, *gorilla*, and *pongo*. *Journal of Human Evolution*, 24, 183–205.

Olea, R. M. (2006): A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment*, 20, 307–318.

Olson, W. & Zhurkin, V. B. (2000): Modeling DNA deformations. *Current Opinion in Structural Biology*, 10, 182–197.

Orozco, M., Pérez, A., Noy, A., & Luque, F. J. (2003): Theoretical methods for the simulation of nucleic acids. *Chemical Society Reviews*, 32, 350–364.

Owen, A. B. (2001): *Empirical Likelihood*. London: Chapman & Hall.

Pen, X.-L., Ying, H., Li, R., & Fang, K.-T. (2006): The application of Kriging and empirical Kriging based on the variables selected by SCAD. *Analytica Chimica Acta*, 578, 178–185.

Petersen, K. B. & Pedersen, M. S. (2008): The matrix cookbook. URL `http://www2.imm.dtu.dk/pubdb/p.php?3274`.

Petke, J. D. (1993): Cumulative and discrete similarity analysis of electrostatic potentials and fields. *Journal of Computational Chemistry*, 14, 928–932.

Politis, D. & Romano, J. P. (1992): A circular block resampling procedure of stationary data. In: R. Lapage & L. Billard (eds.) *Exploring the Limits of Bootstrap*. New York: Wiley.

Politis, D. N. (2003): The impact of bootstrap methods on time series analysis. *Statistical Science*, 18, 219–230.

Preston, S. P. & Wood, A. T. A. (2009a): Bootstrap inference for mean reflection shape and size–and–shape from three–dimensional labelled landmark data. *Submitted for Publication*.

Preston, S. P. & Wood, A. T. A. (2009b): Two–sample bootstrap hypothesis tests for three–dimensional labelled landmark data. *Submitted for Publication*.

Promislow, S. D. (2009): *A first course in functional analysis*. Hoboken: Wiley.

Quenouille, M. H. (1949): Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11, 18–84.

R Development Core Team (2008): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rangarajan, A., Chui, H., & Bookstein, F. L. (1997): The softassign Procrustes algorithm. In: J. Duncan & G. Gindi (eds.) *Information Processing in Medical Imaging*, 29–52. Berlin: Springer.

Richards, W. G. (1993): Computers in drug design. *Pure and Applied Chemistry*, 65, 231–234.

Ripley, B. D. (1981): *Spatial Statistics*. Chichester: Wiley.

Ruffieux, Y. & Green, P. J. (2009): Alignment of multiple configurations using hierarchical models. *Journal of Computational and Graphical Statistics*.

Schabenberger, O. & Gotway, C. A. (2005): *Statistical Analysis for Spatial Data*. London: Chapman & Hall.

Schmidler, S. C. (2007): Fast Bayesian shape matching using geometric algorithms. In: J. M. Bernado, D. Herckerman, J. O. Berger, & A. P. Dawid (eds.) *Bayesian Statistics*, 8. Oxford University Press.

Schmidler, S. C. (2009): Bayesian flexible shape matching with applications to structural proteomics. *Submitted for Publication*.

Schölkopf, B., Burges, C. J. C., & Schmola, A. J. (1999): *Advances in kernel methods. Support vector learning*. Cambridge: MIT Press.

Seber, G. A. F. (1984): *Multivariate Observations*. Chichester: Wiley.

Siddiqui, M. M. (1958): On the inversion of the sample covariance matrix in a stationary autoregressive model. *Annals of Mathematical Statistics*, 29, 585–588.

Silvey, S. D. (1975): *Statistical Inference*. Harmondsworth: Penguin Books.

Singh, K. (1981): On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics*, 9, 1187–1195.

Small, C. G. (1996): *The Statistical Theory of Shape*. Berlin: Springer.

Smith, A. & Roberts, G. O. (1993): Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 55, 3–24.

Srivastava, M. S. (2002): *Methods of Multivariate Statistics*. New York: Wiley.

Ten Berge, J. M. F. (1977): Orthogonal Procrustes rotation for two or more matrices. *Psychometrika*, 42, 267–276.

Tierney, L. (1994): Markov chains for exloring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701–1762.

Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory*. Berlin: Springer.

Wackernagel, H. (2003): *Multivariate Geostatistics*. Berlin: Springer, 3rd edition.

Ward, J. H., Jr. (1963): Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.

Watson, J. D. & Crick, F. H. (1953): A structure for deoxyribose nucleic acids. *Nature*, 171, 737–738.

Welch, B. L. (1947): The generalisation of "student's" problem when several different population variances are involved. *Biometrika*, 34, 28–35.

Wilks, S. S. (1938): The large–sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9, 60–62.

Wilson III, D. M. & Bohr, V. A. (2007): The mechanics of base exicion repair, and its relationship to aging and disease. *DNA Repair*, 6, 544–559.

Yule, G. U. (1926): Why do we sometimes get nonsense–correlations between time–series? – A study in sampling and the nature of time–series. *Journal of the Royal Statistical Society*, 89, 1–63.

Appendix A

The Generalised Procrustes Algorithm

Algorithm A.1 shows the pseudo–code for the generalised Procrustes algorithm for landmark data. For more information see Dryden & Mardia (1998, pp.90).

Algorithm A.1 Generalised Procrustes Algorithm

center the configuration matrices to give $\boldsymbol{X}_{\mathrm{C}1} \dots \boldsymbol{X}_{\mathrm{C}n}$ (cf.(2.1))
set $\boldsymbol{X}^P_{\mathrm{C}\,i} \leftarrow \boldsymbol{X}_{\mathrm{C}\,i}$
define $d \leftarrow d_0$, where $d_0 > tol_1$ and tol_1 is a positive tolerance threshold
while $d \geq tol_1$ **do**
    calculate $G = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ||\boldsymbol{X}^P_{\mathrm{C}\,i} - \boldsymbol{X}^P_{\mathrm{C}\,j}||^2$
    define $e \leftarrow e_0$, where $e_0 > tol_2$ and tol_2 is a positive tolerance threshold
    while $e \geq tol_2$ **do**
        for i in $(1:n)$ **do**
            calculate $\bar{\boldsymbol{X}}_{(i)} = \frac{1}{n-1} \sum_{j \neq i} \boldsymbol{X}^P_{\mathrm{C}\,j}$, i.e. the mean of all but the ith configuration matrix
            optimise $||\bar{\boldsymbol{X}}_{(i)} - \boldsymbol{X}^P_{\mathrm{C}\,i}\boldsymbol{\Gamma}||^2$ over rotation
            set $\boldsymbol{X}^P_{\mathrm{C}\,i} \rightarrow \boldsymbol{X}^P_{\mathrm{C}\,i}\hat{\boldsymbol{\Gamma}}$, where $\hat{\boldsymbol{\Gamma}}$ is the optimal rotation matrix
        end for
        calculate $G^* = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ||\boldsymbol{X}^P_{\mathrm{C}\,i} - \boldsymbol{X}^P_{\mathrm{C}\,j}||^2$
        set $e \leftarrow G - G^*$ and $G \leftarrow G^*$
    end while
    for i in $(1:n)$ **do**
        calculate $\hat{\beta}_i = \left(\frac{\sum_{k=1}^{n} ||\boldsymbol{X}^P_{\mathrm{C}\,k}||^2}{||\boldsymbol{X}^P_{\mathrm{C}\,i}||^2} \right)^{1/2} \phi_i$, where ϕ_i is the ith component of the eigenvector $\boldsymbol{\phi}$ corresponding to the largest eigenvalue of the $(n \times n)$ correlation matrix $\boldsymbol{\Phi}$ of the $\mathrm{vec}(\boldsymbol{X}^P_{\mathrm{C}\,i})$
        set $\boldsymbol{X}^P_{\mathrm{C}\,i} \leftarrow \hat{\beta}_i \boldsymbol{X}^P_{\mathrm{C}\,i}$
    end for
    calculate $G^* = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ||\boldsymbol{X}^P_{\mathrm{C}\,i} - \boldsymbol{X}^P_{\mathrm{C}\,j}||^2$
    set $d \leftarrow G - G^*$ and $G \leftarrow G^*$
end while

Appendix B

Leave–One–Out Method for Identifying Contamination Points

Here, we describe an *ad–hoc* approach for estimating the common underlying covariance structure of two unlabelled point sets A and B which also yields a method to identify contamination points. With the assumption that $A = \{z^{\mathrm{A}}(\boldsymbol{x}_1^{\mathrm{A}}), \ldots, z^{\mathrm{A}}(\boldsymbol{x}_{k_{\mathrm{A}}}^{\mathrm{A}})\}$ and $B = \{z^{\mathrm{B}}(\boldsymbol{x}_1^{\mathrm{B}}), \ldots, z^{\mathrm{B}}(\boldsymbol{x}_{k_{\mathrm{B}}}^{\mathrm{B}})\}$ are noisy pointwise observations of the same underlying reference field, it is appropriate to consider a pooled version of the empirical semivariogram described in Section 2.2.1.2, where "pooled" means that the semivariogram cloud for each point set is determined separately, but both clouds are combined before the distance classes are obtained.

Assuming isotropy, the pooled estimate of the common underlying semivariance function at a separation distance $||\boldsymbol{h}||$ then has the form

$$\hat{\sigma}_{\mathrm{P}}^{*}(||\boldsymbol{h}||) = \frac{1}{2\,|N(||\boldsymbol{h}||)|} \sum_{N(||\boldsymbol{h}||)} \left\{ \left\{ z^{\mathrm{A}}(\boldsymbol{x}_i^{\mathrm{A}}) - z^{\mathrm{A}}(\boldsymbol{x}_j^{\mathrm{A}}) \right\}^2 + \left\{ z^{\mathrm{B}}(\boldsymbol{x}_{i'}^{\mathrm{B}}) - z^{\mathrm{B}}(\boldsymbol{x}_{j'}^{\mathrm{B}}) \right\}^2 \right\}, \quad \text{(B.1)}$$

where, as before, $|N(||\boldsymbol{h}||)|$ denotes the number of distinct pairs in the distance class $N(||\boldsymbol{h}||)$ centred around $||\boldsymbol{h}||$.

If the point sets contain contamination points, then the resulting pooled empirical semivariogram can be a poor estimate of the underlying semivariance function. To demonstrate this, we choose one of the pairs A and B which were considered in the simulation

study in Section 3.6, namely the case where each point set contains 92 points with $k^{\text{true}} = k_{\text{A}}^{\text{true}} = k_{\text{B}}^{\text{true}} = 80$ matching points and $k^{\text{cont}} = k_{\text{A}}^{\text{cont}} = k_{\text{B}}^{\text{cont}} = 12$ contamination points. Moreover, $\kappa = 4$ in this case, and the true underlying covariance function is the Whittle covariance function with $\rho = 0.2$ and $\sigma^2 = 1$. The left–hand side of Figure B.1 shows the resulting pooled empirical semivariogram for the case that the distance classes are chosen as $N(0.05)$, $N(0.1)$, $N(0.15)$, ... It can be seen that it is a rather poor estimate of the true semivariance function which is shown as the solid line. Due to the large number of contamination points, however, this is not surprising.

In order to improve the estimate, we want to identify the contamination points and remove them from (B.1). To do so, we employ a leave–one–out procedure in which the pooled empirical semivariogram is calculated $k_{\text{A}} + k_{\text{B}}$ times, and each time one of the points in either A or B is omitted.

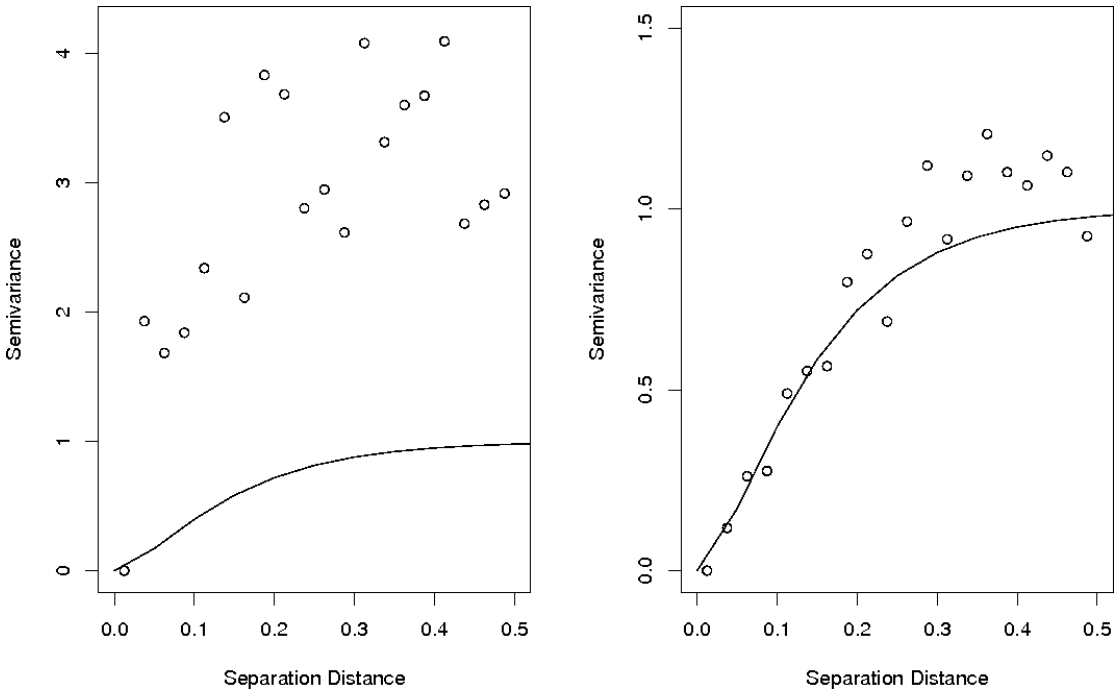

Figure B.1: Pooled empirical semivariograms for unlabelled marked point sets: The left–hand side shows the pooled empirical semivariogram for the raw data, and the right–hand side shows the pooled empirical semivariogram after the high–impact points were removed. Both versions are compared with the true underlying semivariogram (Whittle with $\rho = 0.2$ and $\sigma^2 = 1$) which is shown as the solid line. It can be seen that the considered leave–one–out method considerably improves the estimate.

The impact of each point $\boldsymbol{x}_i^{\mathrm{M}}$ ($M \in \{A, B\}$, $i = 1, \ldots k_{\mathrm{M}}$) on the semivariogram estimate can then be assessed by calculating the mean of the $k_{\mathrm{A}} + k_{\mathrm{B}} - 1$ empirical semivariograms where $\boldsymbol{x}_i^{\mathrm{M}}$ is involved in the estimation. Let $\mathcal{H}_{||\boldsymbol{h}||}$ denote the set of considered centres $||\boldsymbol{h}||$ for the distance classes, i.e. for our example $\mathcal{H}_{||\boldsymbol{h}||} = \{0.05, 0.1, 0.15, \ldots\}$. If the resulting values of $\hat{\sigma}^*_{\mathrm{P,mean}(i)}(||\boldsymbol{h}||)$ ($||\boldsymbol{h}|| \in \mathcal{H}_{||\boldsymbol{h}||}$) substantially differ from the empirical variogram values $\hat{\sigma}^*_{\mathrm{P},(-i)}(||\boldsymbol{h}||)$ where $\boldsymbol{x}_i^{\mathrm{M}}$ is deleted, then this is an indication that $\boldsymbol{x}_i^{\mathrm{M}}$ is a contamination point. To obtain a numerical value for the impact of $\boldsymbol{x}_i^{\mathrm{M}}$, we then sum these differences over the considered distance classes, i.e. we consider

$$I(\boldsymbol{x}_i^{\mathrm{M}}) = \frac{1}{|\mathcal{H}_{||\boldsymbol{h}||}|} \sum_{||\boldsymbol{h}|| \in \mathcal{H}_{||\boldsymbol{h}||}} \hat{\sigma}^*_{\mathrm{P,mean}(i)}(||\boldsymbol{h}||) - \hat{\sigma}^*_{\mathrm{P},(-i)}(||\boldsymbol{h}||)$$

to assess the impact of each point $\boldsymbol{x}_i^{\mathrm{M}}$. If the absolute value of $I(\boldsymbol{x}_i^{\mathrm{M}})$ exceeds a certain threshold I_{crit}, then we consider $\boldsymbol{x}_i^{\mathrm{M}}$ to be a contamination point and delete it from the calculation when we obtain a new, final empirical semivariogram.

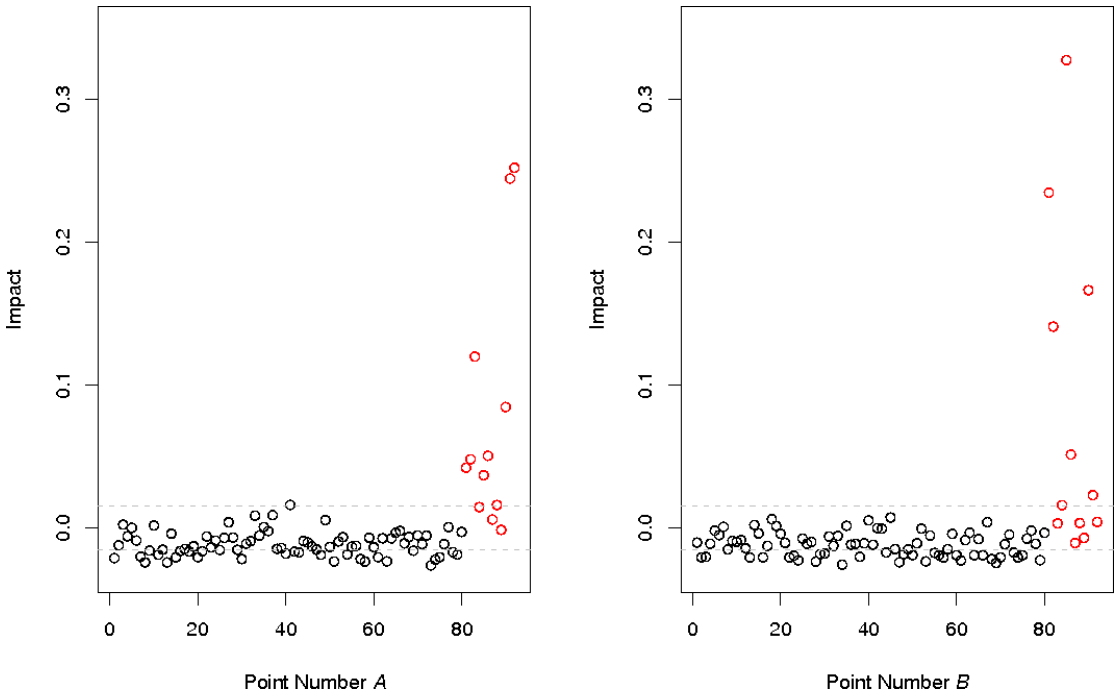

Figure B.2: Determining points with high impact on the pooled empirical semivariogram: For each point $\boldsymbol{x}_i^{\mathrm{M}}$, the corresponding impact value $I(\boldsymbol{x}_i^{\mathrm{M}})$ is displayed. The last twelve points in each point set are the real contamination points, and their impact in shown in red. The threshold is set as $I_{\mathrm{crit}} = 0.015$ here (shown as grey dashed lines), and points whose absolute impact is higher than this value are deleted from the subsequent calculation of the pooled empirical semivariogram.

For our example, Figure B.2 shows the value of $I(\boldsymbol{x}_i^{\mathrm{M}})$ for each of the 184 points $\boldsymbol{x}_i^{\mathrm{M}}$. The actual contamination points are the last twelve points of each point set, and their impact is shown in red. It can be seen that they tend to have a higher impact than the points which were generated from the underlying reference field as described in Section 3.6.1. In this example, we choose $I_{\mathrm{crit}} = 0.015$ as the critical value, and points whose absolute impact exceeds this threshold are deleted when a new, final pooled empirical semivariogram is calculated. This new estimate is shown on the right–hand side of Figure B.1. Although our leave–one–out methods does not identify the contamination points perfectly, the pooled empirical semivariogram where the high–impact points have been removed is considerably closer to the real underlying semivariance model.

We did not extensively investigate this method, but as demonstrated, first results indicate that it can lead to more reliable variogram estimates. Moreover, the threshold for the impact leads to starting points for the mask vectors $\boldsymbol{\lambda}_{\mathrm{A}}$ and $\boldsymbol{\lambda}_{\mathrm{B}}$ which are more informative than the Bernoulli–generated starting masks which are currently applied.

Appendix C

Decision Theoretical Interpretation of Choosing a Threshold for the Posterior Mean Mask Vectors

In Section 4.2.2, we mentioned that choosing a threshold of $p_{\text{crit}} = 0.7$ for setting the entries of the mean mask vectors $\bar{\boldsymbol{\lambda}}_{\text{A}}$ and $\bar{\boldsymbol{\lambda}}_{\text{B}}$ to one implies that we consider the error of falsely excluding a point from the partial Kernel Carbo calculation as less severe than that of including a point which does not in fact belong to the matching parts of the predicted fields. To see that, we adapt the decision theoretical considerations for choosing a labelling matrix described in Green & Mardia (2006) to our situation.

Let $\boldsymbol{\lambda}^T = (\boldsymbol{\lambda}_{\text{A}}^T, \boldsymbol{\lambda}_{\text{B}}^T)$ be the $(k_{\text{A}} + k_{\text{B}})$–vector of the combined true mask entries for both point sets. Within the decision theory framework, we define a loss function $L(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})$ which specifies the cost that arises from declaring the combined mask vector to be $\hat{\boldsymbol{\lambda}}$ ($\hat{\lambda}_i \in \{0, 1\},\ i = 1, \ldots, k_{\text{A}} + k_{\text{B}}$) when it is in fact $\boldsymbol{\lambda}$. Following Green & Mardia (2006), we use a component–wise loss function of the form

$$L(\hat{\lambda}_i, \lambda_i) = \begin{cases} l_{01}, & \text{if } \hat{\lambda}_i = 1 \text{ but } \lambda_i = 0, \\ l_{10}, & \text{if } \hat{\lambda}_i = 0 \text{ but } \lambda_i = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The aim now is to find the vector $\hat{\boldsymbol{\lambda}}$ which minimises the marginal posterior expected loss $\mathrm{E}\big(L(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) | \,\text{data}\big)$ where the expectation is taken over the joint marginal posterior distribution of the two mask vectors $\boldsymbol{\lambda}_{\text{A}}$ and $\boldsymbol{\lambda}_{\text{B}}$.

As described in Chapter 3, the ith entry $\bar{\lambda}_i^{\mathrm{M}}$ of $\bar{\boldsymbol{\lambda}}_{\mathrm{M}}$ ($M \in \{A, B\}$) can be considered as the marginal posterior estimate for the corresponding point to belong to the matching part of the point set. In terms of the combined mask vector $\boldsymbol{\lambda}^T = (\boldsymbol{\lambda}_{\mathrm{A}}^T, \boldsymbol{\lambda}_{\mathrm{B}}^T)$ this means

$$\hat{\mathrm{P}}(\lambda_i = 1 | \,\mathrm{data}) = \bar{\lambda}_i \quad i = 1, \ldots, k_{\mathrm{A}} + k_{\mathrm{B}},$$

so that $\mathrm{E}\big(L(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) | \,\mathrm{data}\big)$ can be estimated component–wise as

$$\hat{\mathrm{E}}\big(L(\hat{\lambda}_i, \lambda_i) | \,\mathrm{data}\big) = l_{10}\, \bar{\lambda}_i \cdot I_{\{\hat{\lambda}_i = 0\}} + l_{01}\, (1 - \bar{\lambda}_i) \cdot I_{\{\hat{\lambda}_i = 1\}}, \quad i = 1, \ldots, k_{\mathrm{A}} + k_{\mathrm{B}}.$$

If the risk for the combined mask vector is calculated cumulatively, then it follows that

$$\begin{aligned}
\hat{\mathrm{E}}\big(L(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) | \,\mathrm{data}\big) &= \sum_{i:\hat{\lambda}_i = 0} l_{10}\, \bar{\lambda}_i + \sum_{i:\hat{\lambda}_i = 1} l_{01}\, (1 - \bar{\lambda}_i) \\
&= (l_{10} + l_{01}) \left\{ \sum_{i:\hat{\lambda}_i = 0} \frac{l_{10}}{l_{10} + l_{01}}\, \bar{\lambda}_i + \sum_{i:\hat{\lambda}_i = 1} \frac{l_{01}}{l_{10} + l_{01}}\, (1 - \bar{\lambda}_i) \right\} \\
&= (l_{10} + l_{01}) \left\{ \sum_{i:\hat{\lambda}_i = 0} \frac{l_{10}}{l_{10} + l_{01}}\, \bar{\lambda}_i \right\} \\
&\qquad + (l_{10} + l_{01}) \left\{ \sum_{i:\hat{\lambda}_i = 1} \frac{l_{01}}{l_{10} + l_{01}} - \sum_{i:\hat{\lambda}_i = 1} \left(1 - \frac{l_{10}}{l_{10} + l_{01}}\right) \bar{\lambda}_i \right\} \\
&= (l_{10} + l_{01}) \left\{ \sum_{i} \frac{l_{10}}{l_{10} + l_{01}}\, \bar{\lambda}_i + \sum_{i:\hat{\lambda}_i = 1} \left(\frac{l_{01}}{l_{10} + l_{01}} - \bar{\lambda}_i \right) \right\}
\end{aligned}$$

For a given cost ratio $K = l_{01}/(l_{10} + l_{01}) \in [0, 1]$, we have therefore shown that the combined mask vector which minimises the estimated marginal posterior risk is

$$\hat{\boldsymbol{\lambda}}_{\mathrm{opt}} = \arg \min_{\hat{\boldsymbol{\lambda}} \in \Lambda_{k_{\mathrm{A}} + k_{\mathrm{B}}}} \hat{\mathrm{E}}\big(L(\hat{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) | \,\mathrm{data}\big) = \arg \max_{\hat{\boldsymbol{\lambda}} \in \Lambda_{k_{\mathrm{A}} + k_{\mathrm{B}}}} \sum_{i:\hat{\lambda}_i = 1} (\bar{\lambda}_i - K).$$

Setting all mask elements whose entries are larger than $p_{\mathrm{crit}} = 0.7$ to one and all others to zero can therefore be interpreted as choosing the optimal mask vectors which minimises the estimated marginal posterior risk for a cost ratio of $K = 0.7$. Note that $K > 0.5$ implies that false inclusions are less desirable than false omissions.

Appendix D

Likelihood Ratio Test

The likelihood ratio test (LRT) is a generic test procedure which can be applied when a parametric model for the data has been specified. Here, we consider the multivariate case. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_x}$ and $\boldsymbol{y}_1, \ldots \boldsymbol{y}_{n_y}$ be two groups of p–dimensional vectors with underlying distributions G and H, respectively. The objective of the LRT is to test for distributional differences between G and H. In particular, if G and H stem from the same parametric family, then it can be used to test for differences between the underlying parameters $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$. The general strategy thereby is to maximise the joint likelihood under both the null hypothesis and the alternative and assess the ratio of the two resulting values.

Let $\boldsymbol{X} \in \mathbb{R}^{n_x \times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_y \times p}$ be the matrices which result from stacking the vectors within each group, and let $L_x(\boldsymbol{X}; \boldsymbol{\theta}_x)$ and $L_y(\boldsymbol{Y}; \boldsymbol{\theta}_y)$ denote the corresponding likelihood functions. As we allow the vectors within each group to be dependent in our main application of the LRT (cf. Chapter 6), the likelihood functions are here defined in terms of the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ instead of the individual vectors. The two groups are assumed to be independent from each other. The overall likelihood of the data is therefore

$$L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}) = L_x(\boldsymbol{X}; \boldsymbol{\theta}_x) L_y(\boldsymbol{Y}; \boldsymbol{\theta}_y), \tag{D.1}$$

where $\boldsymbol{\theta}$ denotes the overall parameter vector whose elements are a subset of the elements of $\boldsymbol{\theta}_{xy}^T = (\boldsymbol{\theta}_x^T, \boldsymbol{\theta}_y^T)$. The specific subset thereby depends on the additional assumptions. For example if the covariances of the two groups are assumed to be equal, then the corresponding elements of $\boldsymbol{\theta}_x^T$ and $\boldsymbol{\theta}_y^T$ occur only once in the combined parameter vector.

The parameter space of $\boldsymbol{\theta}$ depends on the considered hypothesis. Let Θ_0 and Θ_1 denote the parameter spaces under H_0 and H_1, respectively. The LR statistic then has the form

$$\lambda(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_1} L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta})},$$

and the null hypothesis is rejected for small values of $\lambda(\boldsymbol{X}, \boldsymbol{Y})$. Moreover, under some regularity conditions, Wilks' Theorem (Wilks, 1938) holds which states that $\lambda(\boldsymbol{X}, \boldsymbol{Y})$ is a pivotal statistic in the sense that $-2 \log \lambda(\boldsymbol{X}, \boldsymbol{Y})$ has an asymptotic χ^2–distribution under the null hypothesis. The degrees of freedom thereby correspond to the difference between the dimensions of Θ_0 and Θ_1. For a sketch of the proof of Wilks' theorem see for example Silvey (1975, pp.113).

Appendix E

Derivation of the LRT Statistic for the TOPC-AR(2) Model

Here we provide a more detailed derivation of the LRT statistic for the separable TOPC model with an AR(2) dependence structure. As described in Section 6.1.2.2, the joint likelihood of the two p–dimensional time series $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_j\}_{j=1}^n$ is given by

$$L(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{np}|\boldsymbol{\Sigma}_{\mathrm{C}}|^n|\boldsymbol{\Sigma}_{\mathrm{T}}|^p}\exp\left\{-\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{X}}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{X}}^T)\big]\right\}$$
$$\times\exp\left\{-\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\boldsymbol{Y}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{Y}}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{Y}-\mathbf{1}_n\boldsymbol{\mu}_{\mathrm{Y}}^T)\big]\right\}, \quad \text{(E.1)}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $(n\times p)$ matrices which row–wise contain the observations in the two groups and $\boldsymbol{\theta}^T = (\boldsymbol{\mu}_{\mathrm{X}}^T, \boldsymbol{\mu}_{\mathrm{Y}}^T, \psi_1, \psi_2, \mathrm{vech}(\boldsymbol{\Sigma}_{\mathrm{C}})^T)$ denotes the parameter vector which results from assuming that both groups exhibit the same dependence structure.

The test problem at hand divides the parameter space $\Theta = \mathbb{R}^{2p}\times\mathcal{T}_2^{\mathrm{AR}}\times\mathcal{S}_p$ into

$$\Theta_0 = \{\boldsymbol{\theta}\in\Theta : \boldsymbol{\mu}_{\mathrm{X}} = \boldsymbol{\mu}_{\mathrm{Y}}\} \quad \text{and} \quad \Theta_1 = \{\boldsymbol{\theta}\in\Theta : \boldsymbol{\mu}_{\mathrm{X}} \neq \boldsymbol{\mu}_{\mathrm{Y}}\},$$

and to obtain the LRT statistic, the joint likelihood has to be maximed over Θ_0 and Θ_1, respectively. To do so, some matrix calculus has to be applied. The next section provides the relevant results which can be found in Petersen & Pedersen (e.g. 2008, Chapter 2).

Some Matrix Calculus

Let $\boldsymbol{a}$ and $\boldsymbol{w}$ be n–vectors and $\boldsymbol{W}$ an $(n \times n)$–matrix, then

Result 1:

$$f(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a} \quad \Rightarrow \quad \frac{\partial f}{\partial \boldsymbol{a}} = (\boldsymbol{W} + \boldsymbol{W}^T)\boldsymbol{a}$$

Result 2:

$$f(\boldsymbol{a}) = \boldsymbol{w}^T \boldsymbol{a} \quad \Rightarrow \quad \frac{\partial f}{\partial \boldsymbol{a}} = \boldsymbol{w}$$

Let $\boldsymbol{A}$ be an unstructured, invertible matrix, then

Result 3:

$$f(\boldsymbol{A}) = \mathrm{tr}\big[\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{C}\big] \quad \Rightarrow \quad \frac{\partial f}{\partial \boldsymbol{A}} = -\big(\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{B}\boldsymbol{A}^{-1}\big)^T$$

for two general matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ of appropriate dimensions and

Result 4:

$$f(\boldsymbol{A}) = \log|\boldsymbol{A}| \quad \Rightarrow \quad \frac{\partial f}{\partial \boldsymbol{A}} = (\boldsymbol{A}^T)^{-1}.$$

If $\boldsymbol{A}$ exhibits some structure (e.g. symmetric or diagonal), this has to be taken into account when the derivatives are obtained, e.g. if $\boldsymbol{A}$ is symmetric, it can be shown that

Result 5:

$$\frac{df}{d\boldsymbol{A}} = \left[\frac{\partial f}{\partial \boldsymbol{A}}\right] + \left[\frac{\partial f}{\partial \boldsymbol{A}}\right]^T - \mathrm{diag}\left[\frac{\partial f}{\partial \boldsymbol{A}}\right],$$

where d denotes the derivative with the structure taken into account whereas ∂ ignores the structure.

Deriving the Mean Estimators

Under H_0 the data follow the same distribution and we have $\boldsymbol{\mu}_{x,0} = \boldsymbol{\mu}_{y,0} =: \boldsymbol{\mu}_0$ in (E.1). If the values of ψ_1, ψ_2 and $\boldsymbol{\Sigma}_{\mathrm{C}}$ are known, then the MLE for this parameter is

$$\hat{\boldsymbol{\mu}}_0 = \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \underbrace{\operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{X}-\mathbf{1}_n\boldsymbol{\mu}^T)+\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\boldsymbol{Y}-\mathbf{1}_n\boldsymbol{\mu}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{Y}-\mathbf{1}_n\boldsymbol{\mu}^T)\right\}}_{=:f_0(\boldsymbol{\mu})}$$

Expanding $f_0(.)$ leads to

$$\begin{aligned} f_0(\boldsymbol{\mu}) &= c_0 - \operatorname{tr}\left\{\left(\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}+\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}\boldsymbol{Y}^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\right)\mathbf{1}_n\boldsymbol{\mu}^T\right\} \\ &\qquad -\operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\mathbf{1}_n\boldsymbol{\mu}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{X}+\boldsymbol{Y})\right\}+2\operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\mathbf{1}_n\boldsymbol{\mu}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n\boldsymbol{\mu}^T\right\} \\ &= c_0 - 2\cdot\mathbf{1}_n^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{X}+\boldsymbol{Y})\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}\boldsymbol{\mu}+2\cdot\operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\mathbf{1}_n\boldsymbol{\mu}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n\boldsymbol{\mu}^T\right\}, \end{aligned}$$

where c_0 is a constant not dependent on $\boldsymbol{\mu}$.

From Results 1 and 2 it follows that

$$\frac{\partial}{\partial\boldsymbol{\mu}}\mathbf{1}_n^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{X}+\boldsymbol{Y})\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}\boldsymbol{\mu} = \boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\boldsymbol{X}+\boldsymbol{Y})^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n$$

and

$$\frac{\partial}{\partial\boldsymbol{\mu}}\operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\mathbf{1}_n\boldsymbol{\mu}^T)^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n\boldsymbol{\mu}^T\right\} = \mathbf{1}_n^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n\frac{\partial}{\partial\boldsymbol{\mu}}\operatorname{tr}\left\{\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T\right\} = 2\mathbf{1}_n^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}\boldsymbol{\mu}.$$

Setting the gradient equal to zero then yields

$$\begin{aligned} \frac{\partial f_0}{\partial\boldsymbol{\mu}} = 0 \quad &\Leftrightarrow \quad 2\mathbf{1}_n^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n\boldsymbol{\Sigma}_{\mathrm{C}}^{-1}\hat{\boldsymbol{\mu}}_0 = \boldsymbol{\Sigma}_{\mathrm{C}}^{-1}(\boldsymbol{X}+\boldsymbol{Y})^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n \\ &\Leftrightarrow \quad \hat{\boldsymbol{\mu}}_0 = \frac{(\boldsymbol{X}+\boldsymbol{Y})^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n}{2\mathbf{1}_n^T\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}\mathbf{1}_n}. \end{aligned}$$

If all rows of $\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}$ have the same sum, s say, then

$$\hat{\boldsymbol{\mu}}_0 = \frac{s(\boldsymbol{X}+\boldsymbol{Y})^T\mathbf{1}_n}{2sn} = \frac{\bar{\boldsymbol{x}}+\bar{\boldsymbol{y}}}{2}.$$

Due to the particular structure of $\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}$ for AR(2) models $\big($cf. (6.6)$\big)$, this is approximately true for our case and the approximation improves as n grows. Under H_1, the estimators for the individual mean vectors in both groups can be done in a similar manner.

Deriving the Estimators of the Covariance Matrices

For two given estimates $\hat{\psi}_{1,h}$ and $\hat{\psi}_{2,h}$ of the AR(2) parameters under H_h $(h = 0, 1)$, let $\widehat{\boldsymbol{\Sigma}}_{\mathrm{T},h}$ be the corresponding estimate of the temporal covariance matrix. Further, let $\hat{\boldsymbol{\mu}}_{\mathrm{X},h}$ and $\hat{\boldsymbol{\mu}}_{\mathrm{Y},h}$ be the MLEs of the mean vectors of the two groups. To obtain the MLE of $\boldsymbol{\Sigma}_{\mathrm{C},h}$ in this general case, further define

$$\boldsymbol{M}_{\mathrm{X},h} = (\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{X},h}^T)^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{T},h}^{-1}(\boldsymbol{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{X},h}^T) \text{ and } \boldsymbol{M}_{\mathrm{Y},h} = (\boldsymbol{Y} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{Y},h}^T)^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{T},h}^{-1}(\boldsymbol{Y} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_{\mathrm{Y},h}^T).$$

With these definitions, the joint log–likelihood for $\boldsymbol{\Sigma}_{\mathrm{C},h}$ becomes

$$l(\boldsymbol{\Sigma}_{\mathrm{C},h}) = -np\log(2\pi) - n\log(|\boldsymbol{\Sigma}_{\mathrm{C},h}|) - p\log(|\widehat{\boldsymbol{\Sigma}}_{\mathrm{T},h}|) - \tfrac{1}{2}\mathrm{tr}\big\{\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h})\big\}.$$

From Results 3–5 it therefore follows that

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\Sigma}_{\mathrm{C},h}} &= -n\left\{2\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1} - \mathrm{diag}(\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1})\right\} \\ &\qquad -\tfrac{1}{2}\left\{-2\big[\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h})\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}\big] + \mathrm{diag}\big[\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h})\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}\big]\right\} \\ &= \big[-2n\boldsymbol{I} + \boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h})\big]\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1} + \mathrm{diag}\left\{\big[n\boldsymbol{I} - \tfrac{1}{2}\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h})\big]\boldsymbol{\Sigma}_{\mathrm{C},h}^{-1}\right\}.\end{aligned}$$

Hence the critical point of $l(\boldsymbol{\Sigma}_{\mathrm{C},h})$ is given by

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\Sigma}_{\mathrm{C},h}} = 0 \quad &\Leftrightarrow \quad -2n\boldsymbol{I} + \widehat{\boldsymbol{\Sigma}}_{\mathrm{C},h}^{-1}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h}) = 0 \\ &\Leftrightarrow \quad \widehat{\boldsymbol{\Sigma}}_{\mathrm{C},h} = \tfrac{1}{2n}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h})\end{aligned}$$

as stated in (6.15). When $\widehat{\boldsymbol{\Sigma}}_{\mathrm{C},h}$ is inserted into the joint likelihood (6.12), then

$$\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}_{\mathrm{C},h}^{-1}(\boldsymbol{M}_{\mathrm{X},h} + \boldsymbol{M}_{\mathrm{Y},h})\right]\right\} = \exp\{-np\}.$$

so that

$$\sup_{\boldsymbol{\theta}\in\Theta_h} L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\theta}) = \sup_{\boldsymbol{\psi}\in\mathcal{T}_2^{\mathrm{AR}}} c\cdot|\widehat{\boldsymbol{\Sigma}}_{\mathrm{C},h}|^{-n}\sigma_a^{-2np}\big((1-\psi_2^2)^2 - (1+\psi_2)^2\psi_1^2\big)^p = \sup_{\boldsymbol{\psi}\in\mathcal{T}_2^{\mathrm{AR}}} f_h(\boldsymbol{\psi}),$$

where $c = (2\pi)^{-np}\exp(-np)$. Inserting the optimising values of the AR(2) parameters into (6.6) then gives an estimate of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{T},h}^{-1}$.

Appendix F

Additional Figures

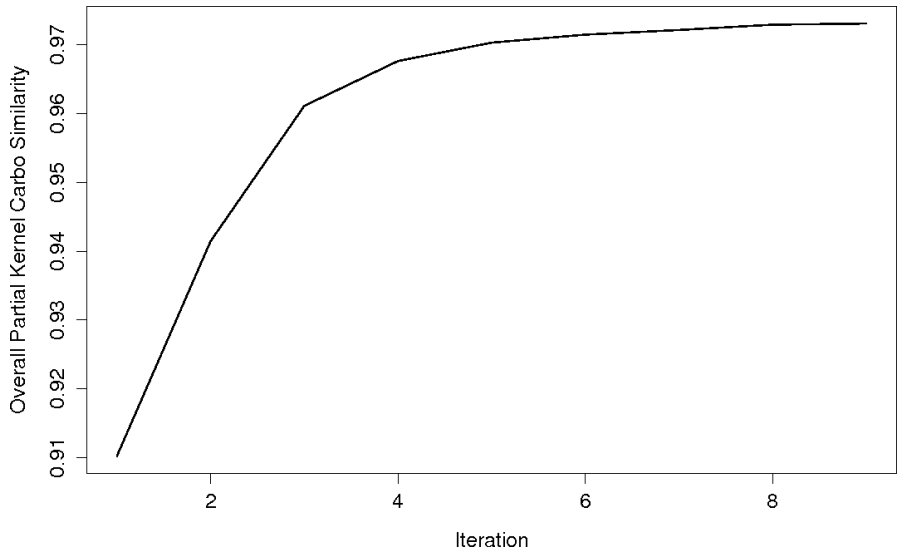

Figure F.1: Sequence of the overall partial Kernel Carbo similarities obtained in course of the field GPA algorithm: Algorithm 4.1 converges quickly, and after 9 iterations of the field GPA, the improvement of the overall Kernel Carbo Index ceases to exceed a tolerance threshold of 0.0001.

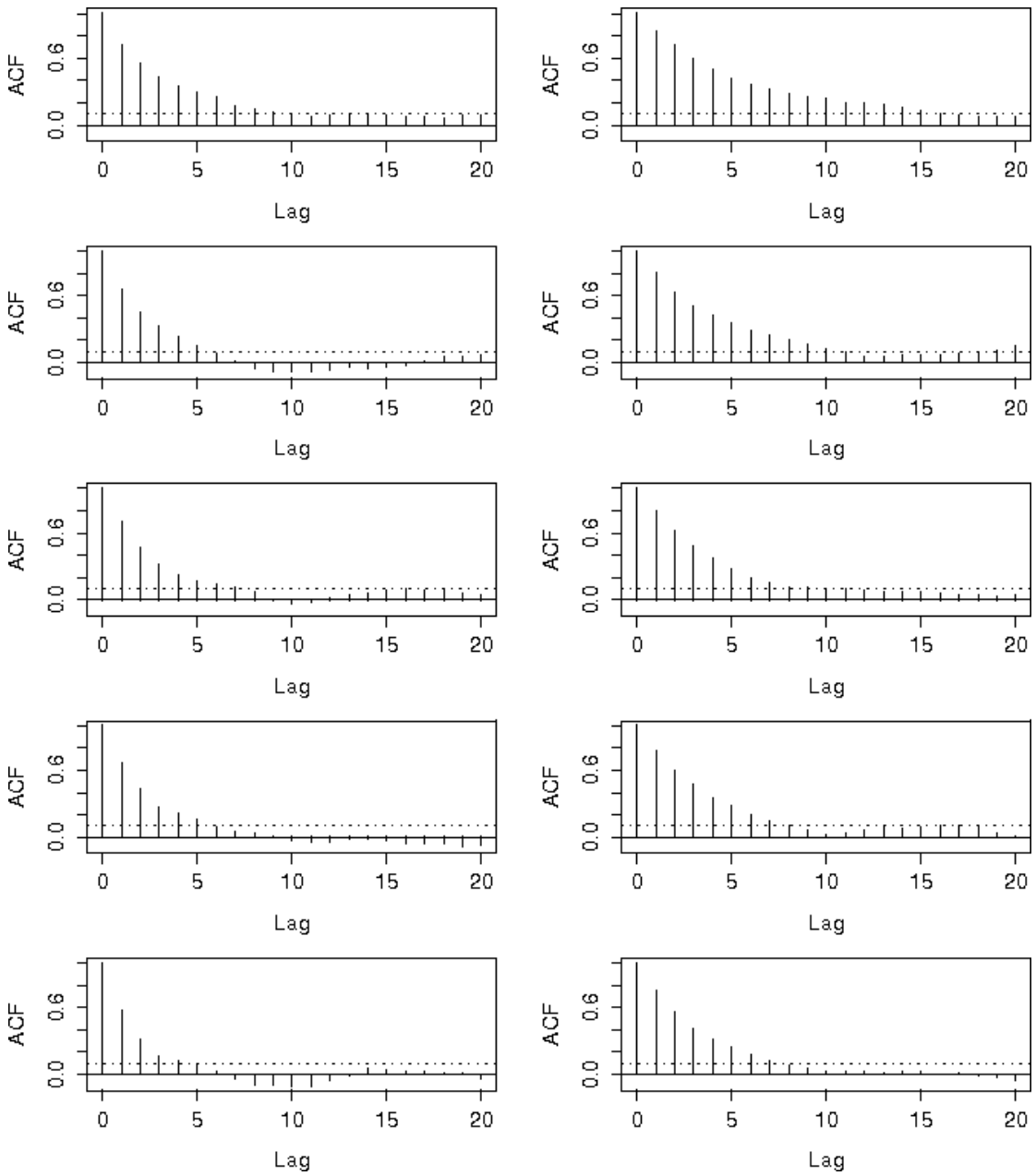

Figure F.2: Impact of the variance in landmark space on the correlation structure in tangent space: Two trajectories of (4×3)–configuration matrices were generated according to (6.19) with an AR(1) dependence structure ($\psi_1 = 0.8$) and $\sigma = 0.1$. GPA was carried out and the tangent vectors were projected using (5.12). The right–hand side shows the correlograms of the PC scores of the resulting five–dimensional projected tangent vectors. The correlation structure closely resembles that of the underlying AR(1) process in landmark space. The procedure was repeated using the same seed but $\sigma = 0.5$. The simulated configurations are therefore proportional to the previous ones. The left–hand side shows the correlograms of the five PC scores of the resulting projected tangent vectors. It can be seen that the tangent projection for highly dispersed data reduces the correlation structure. The horizontal dotted lines show the constant function $f(\text{Lag}) = 0.1$ on both sides.

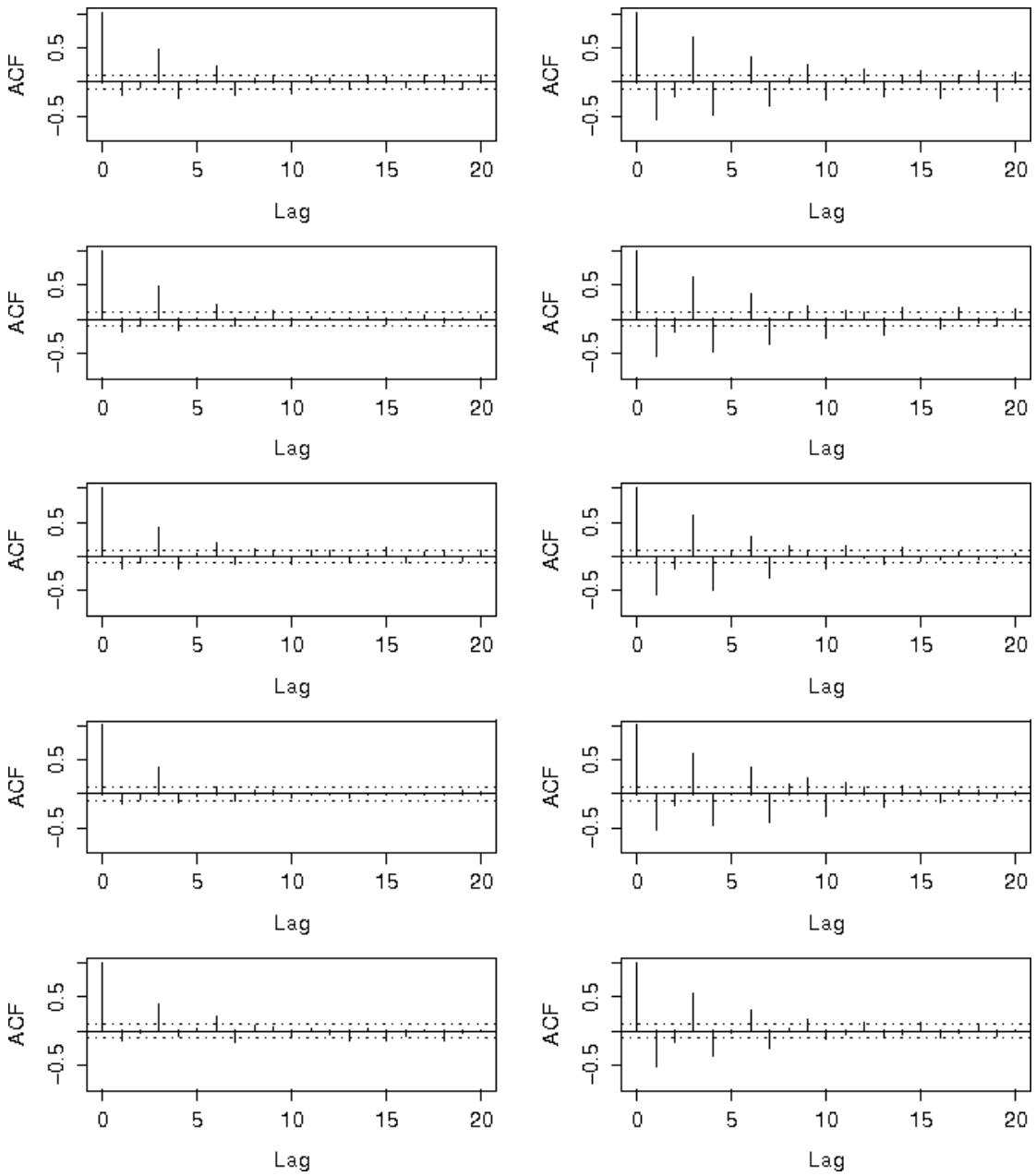

Figure F.3: Impact of the variance in landmark space on the correlation structure in tangent space: Two trajectories of (4×3)–configuration matrices were generated according to (6.19) with an AR(2) dependence structure ($\boldsymbol{\psi}_1 = (-1, -0.75)^T$) and $\sigma = 0.1$. GPA was carried out and the tangent vectors were projected using (5.12). The right–hand side shows the correlograms of the PC scores of the resulting five–dimensional projected tangent vectors. The correlation structure closely resembles that of the underlying AR(2) process in landmark space. The procedure was repeated using the same seed but $\sigma = 0.5$. The simulated configurations are therefore proportional to the previous ones. The left–hand side shows the correlograms of the five PC scores of the resulting projected tangent vectors. Like for the AR(1) example in Figure F.2, it can be seen that the tangent projection for highly dispersed data reduces the correlation structure. The horizontal dotted lines show $f(\text{Lag}) = \pm 0.1$ on both sides.

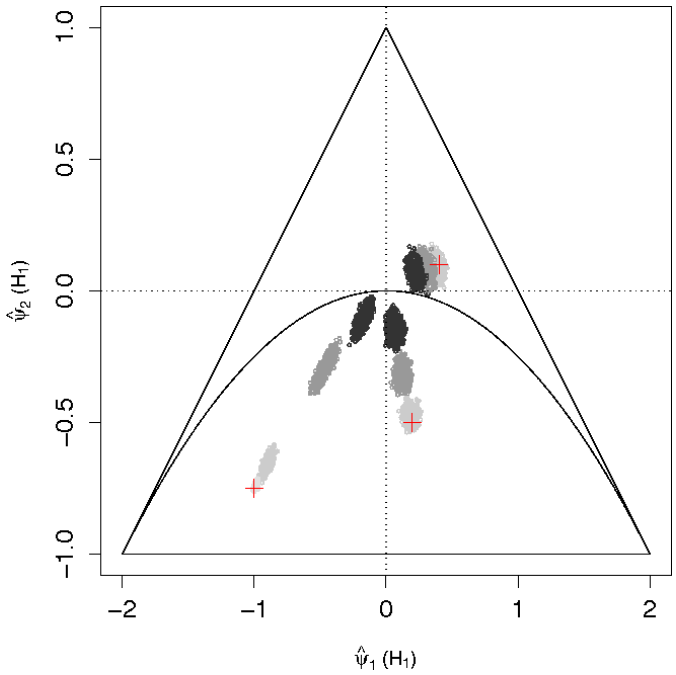

Figure F.4: Plot of the estimated AR(2) parameters under the alternative including some small variance examples: This figure is identical to Figure 6.6 except for the points around $\boldsymbol{\psi} = (-1, -0.75)^T$ which show the estimates $\hat{\boldsymbol{\psi}}^{H_1}$ obtained for configuration matrices which were generated with a very low standard deviation of $\sigma = 0.05$. In this case the correlation structure in tangent space reflects the correlation structure in landmark space.

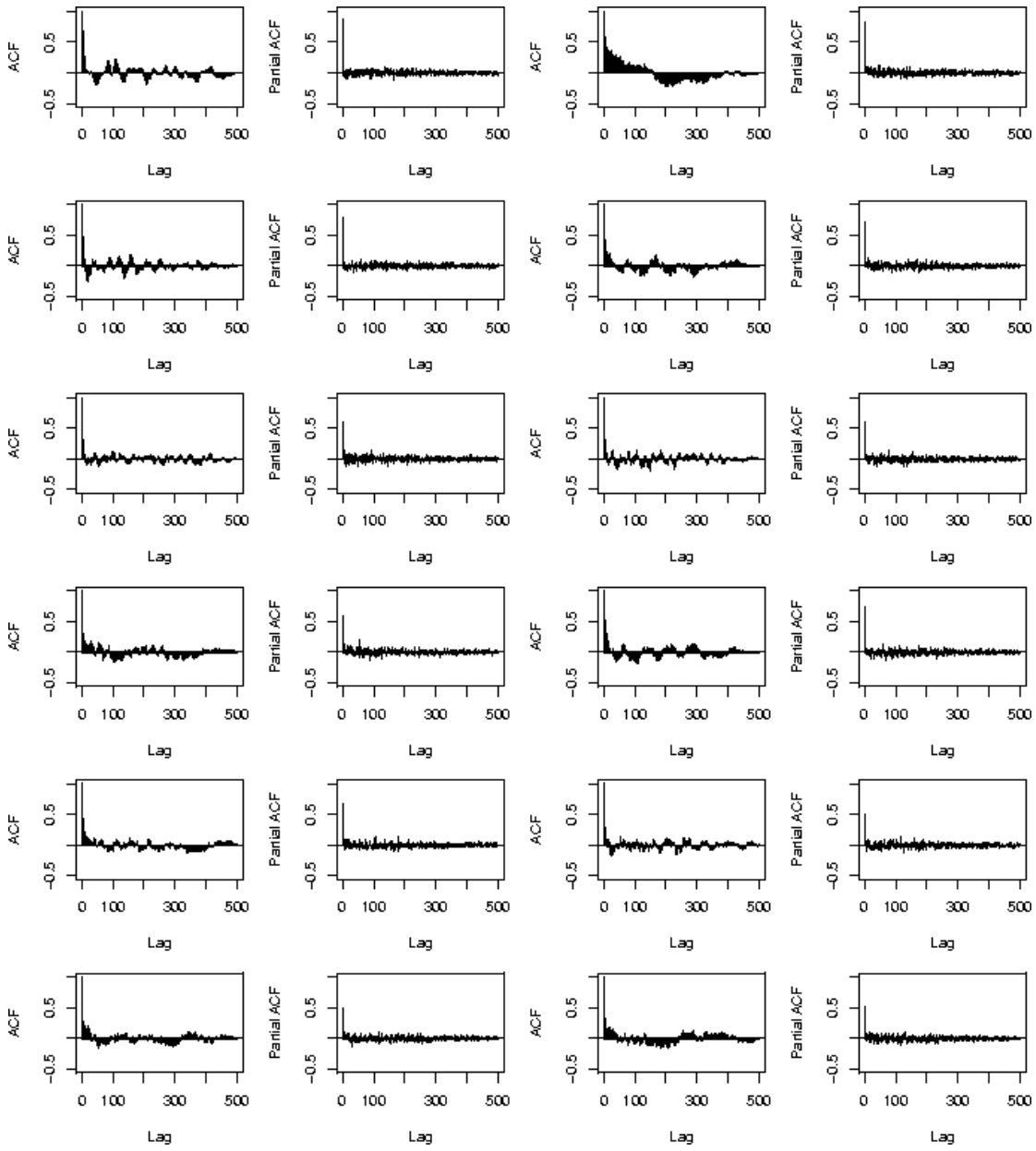

Figure F.5: Autocorrelation and partial autocorrelation for shape PC scores of the AGA duplex: The first row shows the correlation structure of the scores on the first two shape PCs, the second row those of the third and fourth shape PC and so on. It can be seen that the correlation structure is somewhat different on each PC. Each partial autocorrelation function shows only a few spikes which is in line with a low order AR(p) model.