Dimension Reduction for Regression: Theoretical and Methodological Developments

Benjamin Lee Jones

2023

School of Mathematics
Ysgol Mathemateg

Submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Abstract

This thesis has two themes: (1) the predictive potential of principal components in regression, and (2) methodological developments in sufficient dimension reduction.

For the first theme, several research papers have established a number of results showing that, under some uniformity assumptions, higher-ranking principal components of a predictor vector tend, across a range of datasets, to have greater squared correlation with a response variable than lower-ranking ones. This is despite the procedure being unsupervised. This thesis reviews these results and greatly extends them by showing that analogues hold in the setting where nonlinear principal component analysis with general predictors is applied.

For the second theme, research in the past 10 years has led to a measure-theoretic framework for sufficient dimension reduction, inspired by the measure-theoretic formulation of sufficient statistics, which permits nonlinear reductions. This thesis extends this framework to allow for some of the predictors to be categorical. A new estimator, partial generalised sliced inverse regression, is proposed and its properties and effectiveness are explored.

Acknowledgements

Thanks first goes to my supervisor and friend, Dr Andreas Artemiou. You have been a massive support and have provided invaluable guidance both in the drafting of this thesis and in the preparation of our published papers. A special thanks also goes to my secondary supervisor, Professor Karl Michael Schmidt, for his advice and input into the work.

My heartfelt gratitude goes to my funders, the EPSRC. Without their funding, my work would not have been possible. Thanks also goes to Cardiff University, in particular the School of Mathematics, for providing a rich and stimulating research environment.

My friends, Dr Julia Boelle and Dr Valeria Tolis, must also be thanked for being there during our many hours of working together. I will always cherish the many conversations we had and those trips to the local coffee shop.

Last but not least, I wish to thank my family for their continual love and support. Thank you to mum for cheering up my spirits whenever I needed it. Many thanks to dad for showing me the value of learning and encouraging me in my mathematical journey. Thanks finally goes to my brother for his unwavering faith in me.

Contents

List of Publications

B. Jones and A. Artemiou (2019). On Principal Components Regression with Hilbertian Predictors. *Annals of the Institute of Statistical Mathematics* 72, 627–644

B. Jones et al. (2020). On the Predictive Potential of Kernel Principal Components. *Electronic Journal of Statistics* 14.1, 1–23

B. Jones and A. Artemiou (2021). Revisiting the Predictive Power of Kernel Principal Components. *Statistics & Probability Letters* 171, 109019

Frequently used notation

The notation listed here is used throughout this thesis, however there are some occasions where a symbol is applied in a different way to that specified here. The context is used to determine which meaning is intended.

$\overset{\text{a.s.}\mathbb{P}}{=}$ $\mathbb{P}$ almost sure equality

$\mathcal{B}(S)$ the Borel σ-field on a topological space S

$\bigtimes_{i\in\mathcal{I}} S_i$ the Cartesian product of a family of sets indexed by $\mathcal{I}$

$(\perp\!\!\!\perp_{i\in\mathcal{I}} \mathcal{F}_i) \,|\mathcal{F}^*$ the family $(\mathcal{F}_i)_{i\in\mathcal{I}}$ of σ-fields is conditionally independent given $\mathcal{F}^*$

$(\perp\!\!\!\perp_{i\in\mathcal{I}} f_i) \,|f^*$ the stochastic process $(f_i)_{i\in\mathcal{I}}$ is conditionally independent given the random variable f^*

(a, b) an open interval on some linearly ordered set

$(M, \mathcal{F}_M)$ a measurable space

$(S, \mathcal{T})$ a topological space

$\mathbb{C}$ the complex numbers

Card (A) the cardinality of the set A

Cov $(A, B|G)$ the conditional cross-covariance of A and B given the σ-field G

$\mathbb{E}(A|G)$ the conditional expectation of a Banach random variable A given the σ-field G

Σ the covariance matrix/operator of the predictor X

$\mathrm{Cov}(A, B)$ the cross-covariance operator of A and B

$\mathbb{P}(A|G)$ the conditional probability of A given the σ-field G

$\overset{D}{=}$ equality in distribution

$\mathcal{F}_1 \trianglelefteq \mathcal{F}_2$ $\mathcal{F}_1$ is a sub-σ-field of $\mathcal{F}_2$

$\mathcal{H}$ a Hilbert space (generally over $\mathbb{C}$)

$\mathrm{Hom}(A, B)$ the set of all functions from A to B

$\int_\Omega f \, \mathrm{d}\mu$ the Bochner integral of a Banach random variable f (with respect to μ)

$\mathcal{L}(B_1, B_2)$ the bounded operators from B_1 to B_2 where B_1 and B_2 are complex Banach spaces

$\perp\!\!\!\perp_{i\in\mathcal{I}} \mathcal{F}_i$ the σ-fields $\mathcal{F}_i$ ($i \in \mathcal{I}$) are independent

$\perp\!\!\!\perp_{i\in\mathcal{I}} f_i$ the stochastic process $\{f_i\}_{i\in\mathcal{I}}$ is independent

$\mu \ll \nu$ μ is absolutely continuous with respect to ν

μ,ν measures defined on some measurable space

$\mathbb{N}$ the positive integers

$\mathbb{N}_m$ the first m positive integers

$\mathbb{P}_A$ the conditional probability measure defined by an event A

$\mathbb{P}_{f|G}$ a conditional distribution of the random variable f given the σ-field G

$\mathbb{P}_f$ the distribution of a random variable/stochastic process f

$\mathcal{P}_S$ the set of permutations on a set S

$\mathbb{E}_{\mathbb{P}}(f)$ the expectation of a Banach random variable f with respect to $\mathbb{P}$

$\phi_i : \bigtimes_{j\in\mathcal{I}} S_j \to S_i$ the projection from $\bigtimes_{j\in\mathcal{I}} S_j$ to S_i

$\mathcal{P}(A)$ the powerset of a set A

$(\Omega, \mathcal{F}, \mathbb{P})$ a probability space

$\mathbb{R}$ the real numbers

$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ the Radon Nikodym deriviative of μ with respect to ν

σ_A the relative σ-field on A

$[a, b]$ a closed interval on some linearly ordered set

$\bigotimes_{i\in\mathcal{I}} v_i$ product of elements of a complex unital Banach algebra over an arbitrary set

$\bigoplus_{i\in\mathcal{I}} v_i$ summation of elements of a complex Banach space over an arbitrary set

$\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i$ the tensor product σ-field

$\bigotimes_{i\in\mathcal{I}} \mathcal{H}_i$ tensor product of complex Hilbert spaces

$\bigotimes_{i\in\mathcal{I}} \mathbb{P}_i$ product of probability measures

$\bigotimes_{i\in\mathcal{I}} A_i$ the tensor product of operators

$\mathrm{Var}(A)$ the covariance operator of a Hilbertian random variable A

$\mathbb{Z}^*$ the counting numbers (the non-negative integers)

$a < b, a > b, a \leq b, a \geq b$ inequalities on a partially ordered set

$A \perp B$ A and B are orthogonal subsets of some Hilbert space

$A^{\dagger}$ Moore-Penrose inverse of an operator

B a Banach space

$D(G)$ the domain of some function G

d number of components extracted from a dimension reduction procedure

$HS(\mathcal{H}_1, \mathcal{H}_2)$ the Hilbert-Schmidt operators between two complex Hilbert spaces

n number of observations

p number of variables

X the predictor variable in a regression setting

Y the response variable in a regression setting

Chapter 1

Introduction

1.1 What is the need for dimension reduction?

When the number of variables (p) in a dataset is large relative to the number of observations (n), classical statistical methodology tends to break down. Take ordinary least squares regression for example. When $n < p$, the inversion in the equation for the estimated coefficients in a linear regression becomes impossible to perform as the matrix $A^T A$ (where $A \in \mathbb{R}^{n \times p}$ is the design matrix) is non-invertible.

This is an instance of the infamous 'curse of dimensionality', which is frequently encountered when dealing with high-dimensional data. This expression is used to describe an array of issues that may be encountered in Statistics. These include

1. statistical methods suffering from worse performance as the dimension increases,

2. statistical methods suffering from rapidly growing time complexity as the dimension increases,

3. statistical methods being impossible to perform because of theoretical limitations such as operator singularity or ill-conditioning,

4. the inherent difficulty in visualising high-dimensional data.

To perform statistical techniques with high-dimensional data then, this curse must be dealt with. This is where dimension reduction comes into play. By reducing the dimension of a dataset before applying classical statistical methodology, the aforementioned issues posed by high-dimensionality can be overcome.

1.2 What approaches to dimension reduction exist?

There are two main classes of dimension reduction methods. These are known respectively as feature selection and feature extraction. In practice, these approaches are often combined.

Feature selection, also called variable selection, works by choosing a proper subset of the variables, while discarding the rest. It does this by determining which variables are redundant or irrelevant, and discards them to leave a dimension reduced dataset. As feature extraction is the focus of this thesis, feature selection is not discussed further; a review of the literature can be found in Kumar and Minz (2014).

Feature extraction, on the other hand, seeks a collection of d functions of the original p variables where $d < p$, and makes use of these transformed variables instead. So far, 'dimension reduction' has meant either feature selection or feature extraction. As is conventional in the Statistics literature, the phrase is henceforth limited to feature extraction only.

The earliest methods for dimension reduction, which remain commonly used, sought only linear functions of the variables. For example, principal components analysis seeks linear combinations which have maximal variance, subject to the

coefficient vectors forming an orthonormal system. More recent approaches allow for nonlinear combinations by making use of the “kernel trick", most famous for its application to the support vector machine developed by Cortes and Vapnik (1995). For example, Schölkopf et al. (1998) use kernels to develop kernel principal components analysis. Note that, while resulting in further dimension reduction, these nonlinear approaches often sacrifice interpretability of the extracted components so may be inappropriate for some applications. Nonlinear approaches are also more prone to overfitting, though this can be controlled by using cross-validation methods to choose a kernel, from some parametrised family, which gives a function space of functions whose complexity is controlled.

Many commonly used dimension reduction procedures are unsupervised, and are often applied before a supervised learning task. For example, it is common practice, in a high-dimensional regression setting, to regress the response on the leading principal components of the predictors. There is also a supervised framework, sufficient dimension reduction, which is detailed in Chapter 4. Speaking loosely for now, methods in this framework take the response into account by requiring that the transformed predictors have the same predictive power for the response as the original predictors.

1.3 Non-classical data

The classical framework for statistical methodology assumes that the data, or the predictors in a regression setting, are vectors. This was so historically, but now researchers are able to collect more diverse types of data. These new data are broad in scope and include audio files, images, videos, tweets, curves, and surfaces. To handle such data, either classical statistical methodology needs

to be broadened to include them as subcases or entirely new methods need to be developed. New fields have emerged as a result of this development in data collection including image processing, audio processing, and natural language processing.

One particular type of data that has been of recent interest in the dimension reduction literature is Hilbertian data. A datum is said to be Hilbertian if it is an element of a (typically real) Hilbert space $\mathcal{H}$. The quintessential separable infinite-dimensional example is the space of square-integrable real-valued functions over the interval $[0, 1]$ where functions that are almost everywhere equal are considered equivalent. Such functions are common in Neuroscience, where brainwave data are being collected. They are also found in Economics and Finance, where functional processes (e.g.stock prices) are often discretely sampled in the form of time-series data.

As the Hilbert spaces used herein are typically infinite-dimensional, there are unique theoretical issues that do not arise for multivariate data. In particular, it is possible for relevant operators to be unbounded, non-compact, or non-nuclear (see Section 2.4). This statement notwithstanding, classical statistical methods are often adaptable to this setting by replacing the standard Euclidean inner product with the inner product of the Hilbert space.

Analogously to the situation with multivariate data, techniques for reducing the dimension of Hilbertian data can be developed. Indeed, many dimension reduction approaches have been extended to this case. For example, principal components analysis was extended to a functional data setting in Chapter 8 of Ramsay and Silverman (2005), while a more general Hilbertian data formulation is given in Chapter 9 of Hsing and Eubank (2015). In the sufficient dimension reduction framework, Ferré and Yao (2003) extended the sliced inverse regression method, developed by Li (1991), to Hilbertian data.

1.4 Thesis aims and structure

It has been noted already that, in high-dimensional regression, it is common to regress a response on the leading principal components of the predictor vector. As the principal components procedure is unsupervised, this practice is controversial. For any given dataset, there is no guarantee that the leading principal components will be more informative of the response than the trailing ones. Nevertheless, there are a number of research papers (see, e.g., Artemiou and Li (2009), Ni (2011), and Artemiou and Li (2013)) that establish that, across a range of datasets, higher-ranking components tend (i.e.with probability exceeding $1/2$) to have greater squared correlations with the response than lower-ranking ones. One of the aims of this thesis is to review these results, after giving an account of the principal components procedure, and to greatly extend them by proving analogues in the setting where nonlinear principal components analysis is used with a general predictor. This is the focus of Chapter 3.

More generally, the leading components obtained from an unsupervised dimension reduction procedure are not necessarily the most informative of the response. This makes it desirable that the response be taken into account, thus motivating the sufficient dimension reduction framework. In the past 10 years, this framework has been given (see Lee et al. (2013) and Li (2018)) a measure-theoretic formulation, herein called generalised sufficient dimension reduction, which allows for nonlinear reductions. This thesis extends this framework to allow for some of the predictors to be categorical. A new estimator, partial generalised sliced inverse regression, is proposed and its properties and effectiveness are explored. This is the focus of Chapter 4.

Chapter 2 provides the definitions used throughout, along with some supporting results. As it is rather long and dense, the author recommends reading from Chapter 3 and referring back whenever an unfamiliar term, notation, or result is

encountered. Chapter 5 closes with a review of the developments and provides ideas for future research.

Chapter 2

Definitions and supporting results

As the following sections overlap, some definitions in one section may seem to be more appropriately placed in another. The chosen order is designed to trade off between minimising forward cross-referencing as much as possible and placing supporting results in appropriate places.

2.1 Set Theory

Enough Set Theory is given in this section to be able to define arithmetic with cardinals. Though it may seem out of place in a Statistics thesis, this is done as infinite, maybe uncountable, sets are frequently used in Probability Theory and Functional Analysis. The novel usage of Set Theory in this thesis is to extend many classical definitions in Probability Theory to allow for uncountable sets in Section 2.2. Furthermore, the newly given proof of Theorem 2.3.1 (the result is not original, but the proof is) relies fundamentally on the arithmetic of cardinals. The axiom system used here is presumed to be the Zermelo–Fraenkel axiom system with the axiom of choice. The definitions are adapted from Jech (2006), except where it is stated that they are due to the author or otherwise referenced.

To begin, the formal definition of a function is given in a slightly different presentation to classically (see the remark following it).

Definition 2.1.1 (Relations and functions/maps, due to author)**.** Let A and B be sets. A subset $R \subseteq \{(a, b) : a \in A, b \in B\}$ is called a *binary relation between A and B* (if $A = B$, it is said to be a binary relation on A). A subset $f \subseteq \{(a, b) : a \in A, b \in B\}$ is called a *function/map with domain A and codomain B* if:

$$\forall x \in A \, \exists! y_x \in B : (x, y_x) \in f.$$

Write $f : A \to B$ if f is a function. For $x \in A$, let $f(x)$ be the unique element of B for which $(x, f(x)) \in f$ and call $f(x)$ the *image of x under f*, and write $x \mapsto f(x)$. For $S \in \mathcal{P}(A)$, where $\mathcal{P}(A)$ is the powerset of A, let $f(A) \coloneqq \{y \in B : \exists x \in A \, [y = f(x)]\}$ be called the *image of A under f*. For $T \in \mathcal{P}(B)$, let $f^{-1}(T) \coloneqq \{x \in A : f(x) \in T\}$ be called the *preimage of T under f*. Let $\mathrm{Hom}(A, B)$ be the set of all functions from A to B.

Remark 2.1.1. Unlike standard expositions, Definition 2.1.1 deliberately avoids the use of the term “Cartesian product". This version is given (despite being elementary, and using the same concepts) to avoid a circularity when defining, in Definition 2.1.17, the Cartesian product of an arbitrary number of sets in terms of a set of functions.

To start the development of the formal theory of numbers, it is necessary to begin by defining the counting numbers in set-theoretic terms.

Definition 2.1.2 (Von Neumann construction of the counting numbers)**.** Let $0 \coloneqq \emptyset$ and, for every set A, let $S(A) \coloneqq A \cup \{A\}$ be the *successor of A*. The set of *counting numbers* is the set containing 0 and all its successors. Denote this by $\mathbb{Z}^*$.

Remark 2.1.2. Throughout this thesis, 0 typically denotes the additive identity in a vector space over some field (typically either $\mathbb{R}$ or $\mathbb{C}$); the context makes it clear what meaning is to be understood. Some authors use $\mathbb{N}$ instead of $\mathbb{Z}^*$ for the counting numbers, but this thesis uses $\mathbb{N}$ to denote the positive integers.

Much of axiomatic set theory is concerned with formalising the conceptual framework for handling infinities, so now defined are finite and infinite sets.

Definition 2.1.3 (Finite and infinite sets)**.** For $n \in \mathbb{Z}^*$, a set A is said to have n elements if there is a bijection between n and A. If there exists $n \in \mathbb{Z}^*$ such that A has n elements, then A is called *finite*. A is called *infinite* if it is not finite.

The concept of order is one where a basic intuition is acquired at an early age for the integers; presented below is a formalisation of two types of orderings that may encountered when dealing with some particular set.

Definition 2.1.4 (Partial and linear orders)**.** A binary relation $<$ on a set S is said to be a *partial ordering on* S if

1. $\forall s \in S\,[s \not< s]$
2. $[(s_1 < s_2) \wedge (s_2 < s_3)] \implies s_1 < s_3$.

Moreover, if $<$ is a partial ordering on S, $<$ is called a *linear ordering on* S if additionally

$$\forall s_1, s_2 \in S\,[(s_1 < s_2) \vee (s_1 = s_2) \vee (s_2 < s_1)]\,.$$

If $<$ is a partial order on S, write $x \leq y$ if $x < y$ or $x = y$ for x and y in S.

Later, in Section 2.4, the following generalisation of a sequence is employed to define summations of vectors in some complex Banach space over arbitrary index sets.

Definition 2.1.5 (Directed sets and nets, see Willard (1970))**.** A set D is said to be *directed* if there is a binary relation $\leq$ on D such that: (1) $x \leq x$ for any $x \in D$, (2) if $x_1 \leq x_2$ and $x_2 \leq x_3$ then $x_1 \leq x_3$, and (3) if $x_1, x_2 \in D$ then there exists $x_3 \in D$ such that $x_1 \leq x_3$ and $x_2 \leq x_3$. A *net* is a function $f : D \to S$ where D is a directed set and S is an arbitrary set.

The following definition formalises the intuitions of “least", “greatest", and "bounds" at a high level of generality.

Definition 2.1.6 (Least and greatest elements, minimal and maximal elements, lower and upper bounds, infima and suprema)**.** If $(S, <)$ is a partially ordered set, P is a non-empty subset of S, and $a \in S$, then say

1. a is the *least element* of P if $a \in P$ and $\forall x \in P\ [a \leq x]$
2. a is the *greatest element* of P if $a \in P$ and $\forall x \in P\ [x \leq a]$
3. a is a *minimal element* of P if $a \in P$ and $\forall x \in P\ [x \not< a]$
4. a is a *maximal element* of P if $a \in P$ and $\forall x \in P\ [a \not< x]$
5. a is a *lower bound* of P if $\forall x \in P\ [a \leq x]$
6. a is an *upper bound* of P if $\forall x \in P\ [x \leq a]$
7. a is the *infimum* of P if a is the greatest lower bound (assuming it exists) of P. Write $a = \inf P$ if a is the infimum of P
8. a is the *supremum* of P if a is the least upper bound (assuming it exists) of P. Write $a = \sup P$ if a is the supremum of P.

Suprema and infima of bounded subsets of some set do not have to be elements of the set themselves; the classical example here being the rational numbers. The following definition delineates those which do have this property.

Definition 2.1.7 (The least upper bound and greatest lower bound properties, see Rudin (1976))**.** Let $(S, <)$ be a linearly ordered set. S is said to have the *least upper bound property* if the supremum of any non-empty bounded above (i.e. there is an upper bound) subset of S is itself an element of S. The *greatest lower bound property* is dually defined.

Remark 2.1.3. Theorem 1.11 of Rudin (1976) gives that a linearly ordered set with the least upper bound property also has the greatest lower bound property.

A further type of ordering that may be encountered is given below.

Definition 2.1.8 (Well-orderings)**.** A linear order $<$ on a set A is a *well-ordering* if every non-empty subset of A has a least element. A is said to be well-ordered by $<$.

Remark 2.1.4. It is known that the axiom of choice implies that every set can be well-ordered, and vice versa.

The approach taken here to define cardinals (which answers the question of how many elements are in a set) is to define them as ordinals (which answers the question of what position an element has within some ordered collection) with certain properties. To define ordinals then, the following is used.

Definition 2.1.9 (Transitive sets)**.** A set A is *transitive* if every element of A is a subset of A.

Definition 2.1.10 (Ordinals)**.** A set A is an *ordinal* if it is transitive and well-ordered by the membership relation $\in$. For any two ordinals α and γ, write $\alpha < \gamma$ if $\alpha \in \gamma$.

As is standard in Mathematics, when a new class of objects has been defined, it is important to consider structure-preserving maps between those objects. The following does this for partially ordered sets.

Definition 2.1.11 (Order isomorphisms)**.** Let $(A, <_A)$ and $(B, <_B)$ be partially ordered sets. A function $f : A \to B$ is said to be *order-preserving* if $x <_A y \implies f(x) <_B f(y)$. Now suppose that f has an inverse. f is said to be an *order-isomorphism* if both f and its inverse are order-preserving.

Theorem 2.1.1 (Theorem 2.12 of Jech (2006))**.** *Let $(A, <)$ be a well-ordered set. There exists a unique ordinal α such that $(A, <)$ and $(\alpha, \in)$ are order-isomorphic.*

As ordinals answer the question of what position an element has within some ordered collection, the notion of the "next" element can be defined as follows.

Definition 2.1.12 (Successor and limit ordinals)**.** Let α be an ordinal. If $\alpha = S(\gamma)$ (where S is as in Definition 2.1.2) for some ordinal γ, then α is called a *successor ordinal*. If α is not a successor ordinal then $\alpha = \sup\{\gamma : \gamma < \alpha\}$ is called a *limit ordinal*. 0 is also considered to be a limit ordinal and $\sup \emptyset := 0$.

Remark 2.1.5. Jech (2006) considers 0 to be a limit ordinal, but this is not typical. The least non-zero limit ordinal is denoted by ω. With this, it is seen that $\mathbb{Z}^*$ is the set of ordinals less than ω.

Now that ordinals have been defined, there is a need to specify their arithmetic and study its properties. To this end, the focus is now turned.

Definition 2.1.13 (Arithmetic with ordinals)**.** Let α be an ordinal. Define

1. $\alpha + 0 := \alpha$
2. for any ordinal γ, $\alpha + S(\gamma) := S(\alpha + \gamma)$
3. for any non-zero limit ordinal γ, $\alpha + \gamma := \sup\{\alpha + \eta : \eta < \gamma\}$
4. $\alpha \cdot 0 := 0$
5. for any ordinal γ, $\alpha \cdot S(\gamma) := \alpha \cdot \gamma + \alpha$

6. for any non-zero limit ordinal γ, $\alpha \cdot \gamma \coloneqq \sup\{\alpha \cdot \eta : \eta < \gamma\}$
7. $\alpha^0 \coloneqq 1$
8. for any ordinal γ, $\alpha^{S(\gamma)} \coloneqq \alpha^\gamma \cdot \alpha$
9. for any non-zero limit ordinal γ, $\alpha^\gamma \coloneqq \sup\{\alpha^\eta : \eta \in \gamma\}$

Remark 2.1.6. Lemma 2.21 of Jech (2006) gives that addition and multiplication of ordinals are both associative. However, they give examples to show that neither is commutative. Let α, γ, and η be ordinals. The statement of Exercise 2.8 of Jech (2006) gives that $\alpha \cdot (\gamma + \eta) = \alpha \cdot \gamma + \alpha \cdot \eta$, $\alpha^{\gamma+\eta} = \alpha^\gamma \cdot \alpha^\eta$, and $(\alpha^\gamma)^\eta = \alpha^{\gamma \cdot \eta}$.

Lemma 2.1.2 (Lemma 2.25 of Jech (2006))**.** *Let α, γ, and η be ordinals. Then*

1. *if $\gamma < \eta$ then $\alpha + \gamma < \alpha + \eta$*
2. *if $\alpha < \gamma$ then there is a unique ordinal δ such that $\alpha + \delta = \gamma$*
3. *if $\gamma < \eta$ and $0 < \alpha$ then $\alpha \cdot \gamma < \alpha \cdot \eta$*
4. *if $0 < \alpha$ then there is a unique ordinal ρ_1 and a unique ordinal $\rho_2 < \alpha$ such that $\eta = \alpha \cdot \rho_1 + \rho_2$*
5. *if $\gamma < \eta$ and $1 < \alpha$ then $\alpha^\gamma < \alpha^\eta$.*

Now to define cardinals, it is necessary to be able to say when two sets have the same number of elements. This motivates the following definition.

Definition 2.1.14 (Comparing sizes of sets)**.** Let A and B be sets. They are said to be *equinumerous* if there exists a bijection between them. A is said to be *no larger than B* if there is a injection from A to B. A is said to be *smaller than B* if it is both no larger than B and not equinumerous with B.

Remark 2.1.7. The term “equinumerous" is not used in Jech (2006), but is common in Set Theory.

Now cardinals are defined.

Definition 2.1.15 (Cardinals)**.** An ordinal α is called a *cardinal* if α is not equinumerous with any ordinal γ satisfying $\gamma < \alpha$. For any well-ordered set A, define its *cardinality* $\text{Card}(A)$ to be the least ordinal which is equinumerous with A. A is said to be *countably infinite* if $\text{Card}(A) = \omega$. A is said to be *countable* if it is either finite or countably infinite. For any cardinal λ, let λ^+ be the smallest cardinal larger than λ. For any two cardinals α_1 and α_2, write

1. $\alpha_1 = \alpha_2$ if there is a bijection between α_1 and α_2
2. $\alpha_1 \leq \alpha_2$ if there is an injection from α_1 to α_2
3. $\alpha_1 < \alpha_2$ if $\alpha_1 \leq \alpha_2$ but not $\alpha_1 = \alpha_2$.

As the particular interest in axiomatic set theory is for sets with infinite cardinality, it is worthwhile giving them their own name.

Definition 2.1.16 (Alephs)**.** The infinite ordinals which are also cardinals are called *alephs*. For any ordinal α, define

1. $\aleph_0 \coloneqq \omega$
2. $\aleph_{S(\alpha)} \coloneqq \aleph_\alpha^+$
3. if α is a non-zero limit ordinal $\aleph_\alpha \coloneqq \sup\left\{\aleph_\gamma : \gamma < \alpha\right\}$

Jech (2006) shows that this enumeration exhausts all alephs.

One of the most elementary set operations is that of taking the Cartesian product of a family of sets. The formal definition of this is now given in terms of what are often called “choice functions".

Definition 2.1.17 (Cartesian product and projections)**.** Let $\mathcal{I}$ be a set and let $(S_i)_{i\in\mathcal{I}}$ be a family of sets indexed by $\mathcal{I}$. The *Cartesian product* (see Cameron (1998)) is defined by

$$\bigtimes_{i\in\mathcal{I}} S_i \coloneqq \left\{ f : \mathcal{I} \to \bigcup_{i\in\mathcal{I}} S_i \;\middle|\; \forall i \in \mathcal{I}\ [f(i) \in S_i] \right\}.$$

Write $S = M^{\mathcal{I}}$ (which is equal to $\mathrm{Hom}\,(I, M)$) in the event that M is a set which satisfies $M = S_i$ for all $i \in \mathcal{I}$. In this case, write $M^{\mathcal{I}}$ as M^m if $\mathcal{I} = \mathbb{N}_m$ where $m \in \mathbb{N} \cup \{\omega\}$ and $\mathbb{N}_m \coloneqq \{x \in \mathbb{N} : [1 \leq x \leq m]\}$. Let $\phi_i : S \to S_i$ be the function defined by the mapping $f \mapsto f(i)$ and call it the *projection* (see Willard (1970)) onto S_i.

Remark 2.1.8. Definition 2.1.17 tends to be stated for $\mathcal{I} \in \mathcal{P}(\mathbb{N}) \setminus \{\emptyset\}$. When $\mathcal{I} \in \mathcal{P}(\mathbb{N}) \setminus \{\emptyset\}$, an arbitrary element f of the Cartesian product $\bigtimes_{i\in\mathcal{I}} S_i$ can be identified with $(f(x_1), f(x_2), \ldots)$ (if $\mathcal{I}$ is countably infinite) or $(f(x_1), \ldots, f(x_k))$ (if $\mathcal{I}$ has cardinality k where $k \in \mathbb{N}$) where x_i is the i^{th} largest element of $\mathcal{I}$. This means that the definition of a Cartesian product given here is a suitable generalisation of the classical definition for a countable non-empty collection of sets. Whenever it is permissible and reasonable to do so, the classical definition is favoured in this thesis.

At last, it is time to define arithmetic with cardinals which will be needed for the novel proof of Theorem 2.3.1

Definition 2.1.18 (Arithmetic with cardinals, see *Cardinal Arithmetic in nLab* (2022))**.** For a family of sets $(S_i)_{i\in\mathcal{I}}$ indexed by a non-empty set $\mathcal{I}$, the *sum* of the cardinalities of the elements in the family is defined by

$$\sum_{i\in\mathcal{I}} \mathrm{Card}\,(S_i) \coloneqq \mathrm{Card}\left(\biguplus_{i\in\mathcal{I}} S_i \right)$$

where,

$$\biguplus_{i\in\mathcal{I}} S_i \coloneqq \bigcup_{i\in\mathcal{I}} \{(i,a) : a \in S_i\}\,.$$

The *product* of the cardinalities of the elements in the family is defined by

$$\prod_{i\in\mathcal{I}} \mathrm{Card}\,(S_i) \coloneqq \mathrm{Card}\left(\bigtimes_{i\in\mathcal{I}} S_i\right).$$

Cardinal *exponentiation* of the cardinality of a set S to the power of the cardinality of a set T is defined by

$$\mathrm{Card}\,(S)^{\mathrm{Card}(T)} \coloneqq \mathrm{Card}\,(\mathrm{Hom}\,(T,S))\,.$$

Remark 2.1.9. Let α, γ, and η be cardinals. Jech (2006) states the following facts about cardinal arithmetic,

1. cardinal addition and multiplication are both associative and commutative
2. cardinal multiplication distributes over cardinal addition
3. $(\alpha\cdot\gamma)^\eta = \alpha^\eta\cdot\gamma^\eta$
4. $\alpha^{\gamma+\eta} = \alpha^\gamma\cdot\alpha^\eta$
5. $(\alpha^\gamma)^\eta = \alpha^{\gamma\cdot\eta}$
6. if $\alpha\le\gamma$ then $\alpha^\eta\le\gamma^\eta$
7. if $\gamma\le\eta$ and $0<\gamma$ then $\alpha^\gamma\le\alpha^\eta$
8. $\alpha^0 = 1$, $1^\alpha = 1$, and, if $0<\alpha$, $0^\alpha = 0$.

Remark 2.1.10. It is a trivial consequence of (3.14) of Jech (2006) that for $m\in\mathbb{N}$ and ordinals $(\alpha_i)_{i\in\mathbb{N}_m}$,

$$\sum_{i\in\mathbb{N}_m} \aleph_{\alpha_i} = \prod_{i\in\mathbb{N}_m} \aleph_{\alpha_i} = \max\left\{\aleph_{\alpha_i} : i\in\mathbb{N}_m\right\}.$$

This further implies that if $(c_i)_{i\in\mathbb{N}_m}$ satisfies $c_i \leq \aleph_{\alpha_i}$ for any $i \in \mathbb{N}_m$, then

$$\sum_{i\in\mathbb{N}_m} c_i \leq \max\left\{\aleph_{\alpha_i} : i \in \mathbb{N}_m\right\},$$
$$\prod_{i\in\mathbb{N}_m} c_i \leq \max\left\{\aleph_{\alpha_i} : i \in \mathbb{N}_m\right\}.$$

A result which is useful later in the proof of Theorem 2.3.1 is now given.

Lemma 2.1.3 (Lemmas 5.8 and 5.9 of Jech (2006))**.** *Let γ be an infinite cardinal and let $0 < \eta_i$ for each $i < \gamma$. Then*

$$\sum_{i<\gamma} \eta_i = \gamma \cdot \sup\{\eta_i : i < \gamma\}.$$

If additionally $i < j$ implies $\eta_i \leq \eta_j$, then

$$\prod_{i<\gamma} \eta_i = (\sup\{\eta_i : i < \gamma\})^{\gamma}.$$

To finish off this section, the formal definition of permutations is given which is of use when discussing exchangeability in Section 2.2

Definition 2.1.19 (Permutations)**.** Let S be a set. A *permutation* of S is a bijection from S to itself. Denote the set of all permutations of S by $\mathcal{P}_S$. If $m \in \mathbb{N} \cup \{\omega\}$ and $\text{Card}(S) = m$, write $\mathcal{P}_m$ instead of $\mathcal{P}_S$.

Remark 2.1.11. Definition 2.1.19 is not in Jech (2006), though can be found in most elementary set theory texts.

2.2 Measure Theory, Integration Theory, and Probability Theory

This section, along with Section 2.4 and Section 2.5, is fundamental to the theoretical and methodological developments in Chapter 3 and Chapter 4. The

definitions given here are adapted from Hoffmann-Jørgensen (1994), except where it is stated that they are due to the author or otherwise referenced. They are stated more abstractly than most presentations of them detail. A significant novelty in this section is that many definitions, like that of exchangeability, in Probability Theory are extended to settings where relevant sets may have uncountable cardinality. Proofs of any unreferenced results are given in Section 2.6.

First defined is absolute continuity of measures, which is a key hypothesis of the Radon-Nikodym theorem that is used to show the existence of conditional expectations in the real random variable setting.

Definition 2.2.1 (Absolute continuity of measures, see e.g.Li and Babu (2019))**.** Let $(M, \mathcal{F}_M)$ be a measurable space, with M some set and $\mathcal{F}_M$ some σ-field on M, and suppose that μ and ν are measures on $(M, \mathcal{F}_M)$. μ is said to be *absolutely continuous with respect to* ν if, for any $A \in \mathcal{F}_M$, $\nu(A) = 0 \implies \mu(A) = 0$. If this is the case, write $\mu \ll \nu$.

In Mathematics, it is commonplace to consider the structure-preserving maps between various entities of the same kind. For example, consider group homomorphisms between groups. The maps relevant to Probability Theory are random variables, defined below.

Definition 2.2.2 (Random variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ (with Ω some set, $\mathcal{F}$ a σ-field on Ω, and $\mathbb{P}$ a probability measure on $\mathcal{F}$) be a probability space and let $(M, \mathcal{F}_M)$ be a measurable space. A measurable function $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{F}_M)$ is called an *M-valued random variable*. Now let $\mathcal{I}$ be a non-empty set. Define $f^{\mathcal{I}} : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^{\mathcal{I}}, \mathcal{F}_M^{\mathcal{I}}\right)$ (this notation represents a product of measurable spaces, which is defined properly in Definition 2.2.12) to be the $M^{\mathcal{I}}$-valued random variable given by $f^{\mathcal{I}}(x) = g_x$ where $g_x : \mathcal{I} \to M$ is itself given by $g_x(i) = f(x)$.

Remark 2.2.1. Whenever the domain and codomain of a function are specified with σ-fields on them, it is assumed that the function is measurable.

Frequently, one will want to consider a collection of random variables – consider for example the setting of time series data where the variable(s) of interest being monitored across time are random variables. This motivates the definition of stochastic processes.

Definition 2.2.3 (Stochastic processes, this version due to the author)**.** Let $\mathcal{I}$ be a non-empty set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and, for $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i)$ be a measurable space. A *stochastic process* is a family $(f_i)_{i\in\mathcal{I}}$ where, for $i \in \mathcal{I}$, $f_i : (\Omega, \mathcal{F}, \mathbb{P}) \to (M_i, \mathcal{F}_i)$ is a M_i-valued random variable.

Remark 2.2.2. Definition 2.2.3 is more general than the definitions of a stochastic process given in any of the texts the author has searched through. They tend to have the random variables to have all the same codomain, the index set to be the non-negative reals or integers, or the random variables to be real-valued.

In Section 4.5, there is a need to condition on a particular categorical variable taking some fixed value so here is a definition of conditioning on an event that is useful for this task.

Definition 2.2.4 (Conditioning on an event)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$, define the *conditional probability measure* $\mathbb{P}_A : \mathcal{F} \to [0, 1]$ by the mapping $B \mapsto \mathbb{P}(B \cap A) / \mathbb{P}(A)$. An alternative notation is available by letting the function $\mathbb{P}(\cdot|A) : \mathcal{F} \to [0, 1]$ be given by $B \mapsto \mathbb{P}_A(B)$.

Sometimes, a subset of the sample space will be considered and so there is a need to consider a modification of the σ-field which is defined similarly to how relative topologies are defined; this motivates the below definition.

Definition 2.2.5 (Relative σ-field, due to the author)**.** Let $(M, \mathcal{F}_M)$ be a measurable space. For $A \subseteq M$, the *relative σ-field* σ_A is defined to be the set $\{A \cap E : E \in \mathcal{F}_M\}$.

Remark 2.2.3. Definition 2.2.5 is not original. It tends to be used without explicit terminology being given. It is similar to the definition of relative topologies.

Because of their primacy in Probability Theory, the attention is now turned to various ways to obtain σ-fields from given objects.

Definition 2.2.6 (σ-field generated by a set of sets)**.** Let Ω be a set. For any $S \subseteq \mathcal{P}(\Omega)$, define $\sigma(S)$, called *the σ-field generated by S*, to be the intersection of all σ-fields on Ω which contain S.

Remark 2.2.4. Let Ω be a set. For $S \subseteq \mathcal{P}(\Omega)$, a construction of $\sigma(S)$ can be given using ordinals. Let A_0 be the set of all elements of S along with their complements. Given an ordinal α with cardinality $\text{Card}(\alpha) \leq \aleph_0$, let A_α be the set of sets which are countable unions or countable intersections of elements belonging to $\bigcup_{\gamma<\alpha} A_\gamma$. Then, $\sigma(S)$ is the union of the sets A_α where the index runs over all ordinals with cardinality no larger than $\aleph_0$. This construction is given in Vestrup (2003).

Definition 2.2.7 (σ-field generated by a set of functions)**.** Let Ω and $\mathcal{I}$ be sets with $\mathcal{I}$ non-empty. For $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i)$ be a measurable space and let $f_i : \Omega \to (M_i, \mathcal{F}_i)$ be a M_i-valued function. Let $\mathcal{W} \coloneqq \{f_i : i \in \mathcal{I}\}$. Define $\sigma(\mathcal{W})$, called *the σ-field generated by $\mathcal{W}$*, as

$$\sigma(\mathcal{W}) \coloneqq \sigma\left(\left\{f_i^{-1}(S_i) : i \in \mathcal{I}, S_i \in \mathcal{F}_i\right\}\right).$$

Further, let $\sigma(f_i : i \in \mathcal{I}) \coloneqq \sigma(\{f_i : i \in \mathcal{I}\})$.

Definition 2.2.8 (Join of σ-fields, due to the author)**.** Let Ω and $\mathcal{I}$ be sets with $\mathcal{I}$ non-empty. For $i \in \mathcal{I}$, let $\mathcal{F}_i$ be a σ-field on Ω. The *join* of them is defined by

$$\bigvee_{i\in\mathcal{I}} \mathcal{F}_i \coloneqq \sigma\left(\bigcup_{i\in\mathcal{I}} \mathcal{F}_i\right).$$

Remark 2.2.5. Let Ω, $\mathcal{I}$, and $\mathcal{F}_i$ ($i \in \mathcal{I}$) be as in Definition 2.2.8. A π-system (a non-empty set of subsets of Ω which is closed under finite intersections) which generates $\bigvee_{i\in\mathcal{I}} \mathcal{F}_i$ is $\left\{\bigcap_{i\in\mathbb{N}_n} A_i : A_i \in \mathcal{F}_{\alpha_i}, \alpha_i \in \mathcal{I}, n \in \mathbb{N}\right\}$.

Remark 2.2.6. Definition 2.2.8 is not original. It tends to be used without the term "join" being explicitly defined.

Theorem 2.2.1 (Dynkin's π-λ theorem, see Billingsley (1995))**.** *Let Ω be a set. If P is a π-system on Ω and Λ is a λ-system on Ω (a set of subsets of Ω which contains Ω, is closed under complementation, and is closed under disjoint countable unions), then $P \subseteq \Lambda \implies \sigma(P) \subseteq \Lambda$.*

Definition 2.2.9 (Borel σ-field on a topological space)**.** Let $(S, \mathcal{T})$ be a topological space. Define $\mathcal{B}(S)$, *called the Borel σ-field on S*, by $\mathcal{B}(S) \coloneqq \sigma(\mathcal{T})$.

It is sometimes said (see, e.g., Loève (1977)) that the definition of independence is where Probability Theory comes into its own as opposed to being a part of Measure Theory. This fundamental notion is now defined.

Definition 2.2.10 (Independence of σ-fields, adapted from Williams (2018))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{I}$ be a non-empty set. Let $(\mathcal{F}_i)_{i\in\mathcal{I}}$ be a family of sub-σ-fields of $\mathcal{F}$. $(\mathcal{F}_i)_{i\in\mathcal{I}}$ is said to be *independent* (with respect to $\mathbb{P}$, though this will not usually be mentioned explicitly), denoted by $\perp\!\!\!\perp_{i\in\mathcal{I}} \mathcal{F}_i$, if for any finite subset $E \subseteq \mathcal{I}$ and, for $i \in E$, any $A_i \in \mathcal{F}_i$,

$$\mathbb{P}\left(\bigcap_{i\in E} A_i\right) = \prod_{i\in E} \mathbb{P}(A_i).$$

If $\mathcal{I} = \{1, 2\}$ and $(\mathcal{F}_1, \mathcal{F}_2)$ is independent, write $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2$.

Remark 2.2.7. For convenience, as it is extensively used in Chapter 4, the notation $A \trianglelefteq B$ is introduced. It means that A and B are σ-fields on some set with A a sub-σ-field of B.

Definition 2.2.11 (Independence of stochastic processes, adapted from Williams (2018))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{I}$ be a non-empty set, and, for $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i)$ be a measurable space. Let $f \coloneqq (f_i)_{i\in\mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, f_i is a M_i-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. f is said to be *independent*, denoted by $\perp\!\!\!\perp_{i\in\mathcal{I}} f_i$, if the family $(\sigma(f_i))_{i\in\mathcal{I}}$ is independent. Furthermore, if $(S, \mathcal{F}_S) = (M_i, \mathcal{F}_i)$ for any $i \in \mathcal{I}$ and there exists $\mathbb{P}^*$ such that $\forall i \in \mathcal{I}\,[\mathbb{P}^* = \mathbb{P}_i]$ (where $\mathbb{P}_i$ is the distribution of f_i, see Definition 2.2.13), then f is said to be *independent and identically distributed* (abbreviated i.i.d). If $\mathcal{I} = \{1, 2\}$ and (f_1, f_2) is independent, write $f_1 \perp\!\!\!\perp f_2$.

Remark 2.2.8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose $f_1 : (\Omega, \mathcal{F}, \mathbb{P}) \to (M_1, \mathcal{F}_1)$ and $f_2 : (\Omega, \mathcal{F}, \mathbb{P}) \to (M_2, \mathcal{F}_2)$ are such that (f_1, f_2) is an independent stochastic process. Then it is easy to see that, for any measurable functions $S : (M_1, \mathcal{F}_1) \to (M_S, \mathcal{F}_S)$ and $T : (M_2, \mathcal{F}_2) \to (M_T, \mathcal{F}_T)$, the stochastic process $(S \circ f_1, T \circ f_2)$ is independent. This follows immediately from $\sigma(S \circ f_1) \trianglelefteq \sigma(f_1)$ and $\sigma(T \circ f_2) \trianglelefteq \sigma(f_2)$. This remark is easily generalised to an arbitrary non-empty collection of random variables.

In analogy with the definition of Cartesian products in Set Theory, the taking of products of measurable spaces is a fundamental notion in Measure Theory and (consequently) Probability Theory.

Definition 2.2.12 (Product of measurable spaces)**.** Let $\mathcal{I}$ be a non-empty set and let $(M_i, \mathcal{F}_i)$ be a measurable space for any $i \in \mathcal{I}$. Let $M \coloneqq \bigtimes_{i\in\mathcal{I}} M_i$. For $i \in \mathcal{I}$, let $\phi_i : \bigtimes_{i\in\mathcal{I}} M_i \to (M_i, \mathcal{F}_i)$ be the projection onto M_i. Let $\mathcal{F} \coloneqq \bigotimes_{i\in\mathcal{I}} \mathcal{F}_i \coloneqq \sigma(\phi_i : i \in \mathcal{I})$ be the σ-field generated by the projections, and

call it the *tensor product σ-field*. Write $\mathcal{F} = G^{\mathcal{I}}$ in the event that: (1) there exists a set S which satisfies $S = M_i$ for any $i \in \mathcal{I}$, and (2) G is a σ-field satisfying $G = \mathcal{F}_i$ for any $i \in \mathcal{I}$. In this case, write $G^{\mathcal{I}}$ as G^m if $\mathcal{I} = \mathbb{N}_m$ where $m \in \mathbb{N} \cup \{\omega\}$. The pair $(M, \mathcal{F})$ is called the *product measurable space*.

Remark 2.2.9. The tensor product σ-field is the coarsest σ-field for which the projection maps are all measurable.

Before the definition of random variables, it was noted that they are essentially structure-preserving maps between probability spaces. There, however, was no probability measure given on their codomains. A natural way to take such a measure is given now as well as its generalisation to stochastic processes.

Definition 2.2.13 (Distribution of a random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(M, \mathcal{F}_M)$ be a measurable space. Let $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{F}_M)$ be a M-valued random variable. The *distribution of f*, denoted $\mathbb{P}_f$, is defined by the mapping $A \mapsto \mathbb{P}\left(f^{-1}(A)\right)$ where $A \in \mathcal{F}_M$.

Definition 2.2.14 (Distribution of a stochastic process, version due to author)**.** Let $\mathcal{I}$ be a non-empty set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and, for $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i)$ be a measurable space. Let $f \coloneqq (f_i)_{i \in \mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, f_i is a M_i-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The *distribution of f*, denoted $\mathbb{P}_f$, is defined to be the distribution of the random variable $f^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\bigtimes_{i \in \mathcal{I}} M_i, \bigotimes_{i \in \mathcal{I}} \mathcal{F}_i\right)$ given by $f^*(x) = h_x$ where $h_x : \mathcal{I} \to \bigcup_{i \in \mathcal{I}} M_i$ is itself given by $h_x(i) = f_i(x)$.

Lemma 2.2.2 (Relation between the independence of a stochastic process and the random variable defining its distribution)**.** *Let $\mathcal{I}$ be a non-empty set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and, for $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i)$ be a measurable space. Let $f \coloneqq (f_i)_{i \in \mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, f_i is a M_i-valued*

random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* $f^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\bigtimes_{i\in\mathcal{I}} M_i, \bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\right)$ *be the random variable given by* $f^*(x) = h_x$ *where* $h_x : \mathcal{I} \to \bigcup_{i\in\mathcal{I}} M_i$ *is itself given by* $h_x(i) = f_i(x)$. *Let* $\phi_i : \left(\bigtimes_{i\in\mathcal{I}} M_i, \bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\right) \to (M_i, \mathcal{F}_i)$ *be the projection onto* M_i. *If* f *is independent, then* f^* *satisfies that for any finite subset* $E \subseteq \mathcal{I}$ *and, for* $i \in \mathcal{I}$, *any* $A_i \in \sigma(\phi_i)$

$$\mathbb{P}\left(\bigcap_{i\in E} [f^*]^{-1}(A_i)\right) = \prod_{i\in E} \mathbb{P}\left([f^*]^{-1}(A_i)\right).$$

Given a product of measurable spaces each with a probability measure on their respective σ-fields, it is natural to wonder if there is a canonical probability measure on the product space. This is indeed the case, as given in the next definition.

Definition 2.2.15 (Product probability measures, adapted from Cohn (2013))**.** Let $\mathcal{I}$ be a non-empty set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and, for $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i, \mathbb{P}_i)$ be a probability space. Let $\left(\bigtimes_{i\in\mathcal{I}} M_i, \bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\right)$ be the product measurable space. For $i \in \mathcal{I}$, let $\phi_i : \left(\bigtimes_{i\in\mathcal{I}} M_i, \bigotimes_{i\in\mathcal{I}} \mathcal{F}_i\right) \to (M_i, \mathcal{F}_i)$ be the projection onto M_i. A measure $\bigotimes_{i\in\mathcal{I}} \mathbb{P}_i$ on $\bigotimes_{i\in\mathcal{I}} \mathcal{F}_i$ is called a *product probability measure* if it is a probability measure satisfying

1. the stochastic process $(\phi_i)_{i\in\mathcal{I}}$ is independent (with respect to $\bigotimes_{i\in\mathcal{I}} \mathbb{P}_i$).
2. for any $i \in \mathcal{I}$, the distribution of ϕ_i is $\mathbb{P}_i$.

If $(S, \mathcal{F}_S, \mathbb{P}) = (M_i, \mathcal{F}_i, \mathbb{P}_i)$ for all $i \in \mathcal{I}$, write $\bigotimes_{i\in\mathcal{I}} \mathbb{P}_i = \mathbb{P}^{\mathcal{I}}$.

Remark 2.2.10. Exercise 2 in Section 10.6 of Cohn (2013) states that the product probability measure, on the product of an arbitrary non-empty family of probability spaces, exists and is unique.

Lemma 2.2.3 (Relation between the distribution of an independent stochastic process and the product probability measure)**.** *Let* $\mathcal{I}$ *be a non-empty set, let*

$(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and, for $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i)$ be a measurable space. Let $f \coloneqq (f_i)_{i\in\mathcal{I}}$ be an independent stochastic process where, for $i \in \mathcal{I}$, f_i is a M_i-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, $\mathbb{P}_f = \bigotimes_{i\in\mathcal{I}} \mathbb{P}_{f_i}$.

For the work in Chapter 3, the notion of an exchangeable stochastic process is fundamental, hence is defined here.

Definition 2.2.16 (Exchangeable stochastic process, due to the author)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{I}$ be a non-empty set, and let $(M, \mathcal{F}_M)$ be a measurable space. Let $f \coloneqq (f_i)_{i\in\mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, f_i is a M-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. f is said to be *exchangeable* if

$$\begin{aligned}
&\forall T \in \Big\{ \mathbb{N}_{n_T} : [\text{Card}\,(\mathcal{I}) \geq \aleph_0 \implies n_T \in \mathbb{N}] \\
&\qquad\qquad \wedge [\text{Card}\,(\mathcal{I}) = k < \aleph_0 \implies n_T \in \mathbb{N}_{k+1}] \Big\} \\
&\forall m \in \{n \in \mathbb{N} : n < n_T\} \\
&\forall k_1, \ldots, k_m \in T \\
&\left[\left(\forall i, j \in \mathbb{N}_m \left[i \neq j \implies k_i \neq k_j \right] \right) \implies \left(\mathbb{P}_{f_m^\dagger} = \mathbb{P}_{f_m^*} \right) \right]
\end{aligned}$$

where $f_m^\dagger : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^m, \mathcal{F}_M^m\right)$ is given by $f_m^\dagger(x) = g_x$ where $g_x : \mathbb{N}_m \to M$ is itself given by $g_x(i) = f_i(x)$, and $f_m^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^m, \mathcal{F}_M^m\right)$ is given by $f_m^*(x) = h_x$ where $h_x : \mathbb{N}_m \to M$ is itself given by $h_x(i) = f_{k_i}(x)$.

Remark 2.2.11. Definition 2.2.16 is, to the best of the author's knowledge, the most general definition (as far as the cardinality of $\mathcal{I}$ is concerned) of exchangeability to date. The concept can be found in Kallenberg (1988, 1992, 2000, 2005, 2021).

Lemma 2.2.4 (i.i.d stochastic processes are exchangeable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{I}$ be a non-empty set, and let $(M, \mathcal{F}_M)$ be a measurable space. If the stochastic process $f \coloneqq (f_i)_{i\in\mathcal{I}}$ (where, for any $i \in \mathcal{I}$, f_i is a M-valued random variable) is i.i.d, then it is exchangeable.*

Exchangeability is essentially a permutation invariance assumption. Another notion of invariance is that of contractability, now defined. This notion is less fundamental for this thesis, but is included to explore its relation to exchangeability.

Definition 2.2.17 (Contractable stochastic processes, due to the author)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{I}$ be a non-empty set, and let $(M, \mathcal{F}_M)$ be a measurable space. Let $f \coloneqq (f_i)_{i\in\mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, f_i is a M-valued random variable. f is said to be *contractable* if

$$\begin{aligned}
&\forall T \in \Big\{\mathbb{N}_{n_T} : [\text{Card}\,(\mathcal{I}) \geq \aleph_0 \implies n_T \in \mathbb{N}] \\
&\qquad\qquad \wedge [\text{Card}\,(\mathcal{I}) = k < \aleph_0 \implies n_T \in \mathbb{N}_{k+1}]\Big\} \\
&\forall m \in \{n \in \mathbb{N} : n < n_T\} \\
&\forall k_1, \ldots, k_m \in T \\
&\left[(1 \leq k_1 < \ldots < k_m \leq n_T) \implies \left(\mathbb{P}_{f_m^\dagger} = \mathbb{P}_{f_m^*}\right)\right]
\end{aligned}$$

where $f_m^\dagger : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^m, \mathcal{F}_M^m\right)$ is given by $f_m^\dagger(x) = g_x$ where $g_x : \mathbb{N}_m \to M$ is itself given by $g_x(i) = f_i(x)$, and $f_m^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^m, \mathcal{F}_M^m\right)$ is given by $f_m^*(x) = h_x$ where $h_x : \mathbb{N}_m \to M$ is itself given by $h_x(i) = f_{k_i}(x)$.

Remark 2.2.12. Definition 2.2.17 is, to the best of the author's knowledge, the most general definition of contractability (as far as the cardinality of $\mathcal{I}$ is concerned) to date. The concept can be found in Kallenberg (1988, 1992, 2000, 2005, 2021). It is trivial to see that exchangeability implies contractability.

Lemma 2.2.5 (Some equivalences)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{I}$ be a non-empty set, and let $(M, \mathcal{F}_M)$ be a measurable space. Let $f \coloneqq (f_i)_{i\in\mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, f_i is a M-valued random variable. Let $f^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^{\mathcal{I}}, \mathcal{F}_M^{\mathcal{I}}\right)$ be given by $f^*(x) = h_x$ where $h_x : \mathcal{I} \to M$ is itself given by $h_x(i) = f_i(x)$. Let $\phi \coloneqq (\phi_i)_{i\in\mathcal{I}}$ be the stochastic process where, for $i \in \mathcal{I}$, $\phi_i : \left(M^{\mathcal{I}}, \mathcal{F}_M^{\mathcal{I}}, \mathbb{P}_{f^*}\right) \to (M, \mathcal{F}_M)$ is the projection onto M. Then*

1. *f is independent if and only if ϕ is independent*
2. *f is identically distributed if and only if ϕ is identically distributed*
3. *f is contractable if and only if ϕ is contractable*
4. *f is exchangeable if and only if ϕ is exchangeable.*

The following theorem gives an implication of contractability/exchangeability, where the exchangeability case is useful for the proof of Theorem 2.2.11 which plays a crucial role in the developments in Chapter 3.

Theorem 2.2.6 (An implication of contractability and exchangeability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{I}$ be a non-empty set, and let $(M, \mathcal{F}_M)$ be a measurable space. Let $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{F}_M)$ be a M-valued random variable. Let $V \coloneqq (v_i)_{i\in\mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, v_i is a M-valued random variable. Let $V^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^{\mathcal{I}}, \mathcal{F}_M^{\mathcal{I}}\right)$ be the random variable given by $V^*(x) = g_x$ where $g_x : \mathcal{I} \to M$ is itself given by $g_x(i) = v_i(x)$. Suppose that the stochastic process (V^*, f) is independent (written $V^* \perp\!\!\!\perp f$). Let $z \coloneqq (z_i)_{i\in\mathcal{I}}$ be the stochastic process where, for $i \in \mathcal{I}$, $z_i : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^2, \mathcal{F}_M^2\right)$ is given by $z_i(x) = l_x^i$ where $l_x^i : \{1, 2\} \to M$ is itself given by $l_x^i(1) = v_i(x)$ and $l_x^i(2) = f(x)$. Let $T : \left(M^2, \mathcal{F}_M^2\right) \to (N, \mathcal{F}_N)$ be a measurable function where $(N, \mathcal{F}_N)$ is a measurable space. Let $W \coloneqq (w_i)_{i\in\mathcal{I}}$ be the stochastic process where, for $i \in \mathcal{I}$, $w_i : (\Omega, \mathcal{F}, \mathbb{P}) \to (N, \mathcal{F}_N)$ is given by $w_i(x) = [T \circ z_i](x)$. Claim 1: if V is exchangeable then so is W. Claim 2: if V is contractable then so is W.*

Remark 2.2.13. It is convenient to introduce some seemingly abusive notation here. For $i \in \mathcal{I}$, define $v_i(f) \coloneqq w_i$. If $M = \mathcal{H}$ for some Hilbert space over $\mathbb{C}$, $\mathcal{F}_M = \mathcal{B}(\mathcal{H})$, $N = \mathbb{C}$, $\mathcal{F}_N = \mathcal{B}(\mathbb{C})$, and $T = \langle\cdot,\cdot\rangle_{\mathcal{H}}$, then define, for $i \in \mathcal{I}$, $\langle v_i, f\rangle_{\mathcal{H}} \coloneqq w_i$. It is straightforward to see that, for $i \in \mathcal{I}$ and $x \in \Omega$, $[\langle v_i, f\rangle_{\mathcal{H}}](x) = \langle v_i(x), f(x)\rangle_{\mathcal{H}}$.

As empirical measures are fundamental to Statistics, and they depend on a sample, they are random measures – this notion is thus important to define here.

Definition 2.2.18 (Random measures and random probability measures, adapted from Klenke (2008))**.** Let $(M, \mathcal{B}(M))$ be a measurable space where M is a metric space, let $\mathcal{M}^*$ be the set of all measures on $\mathcal{B}(M)$, and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{M}$ be the set of all locally finite measures in $\mathcal{M}^*$; that is $\mathcal{M} \coloneqq \{\mu \in \mathcal{M}^* : \mu(B) < \infty \text{ for all bounded } B \in \mathcal{B}(M)\}$. For any bounded $B \in \mathcal{B}(M)$, let $I_B : \mathcal{M}^* \to [0, \infty]$ be given by $\mu \mapsto \mu(B)$. Let $\mathcal{B}(M)^* \coloneqq \sigma(I_B : B \in \mathcal{B}(M), B \text{ is bounded})$ define a σ-field on $\mathcal{M}^*$, and let $\mathcal{B}(M)^\dagger \coloneqq \sigma(I_B|_{\mathcal{M}} : B \in \mathcal{B}(M), B \text{ is bounded})$ define a σ-field on $\mathcal{M}$. A *random measure on* $\mathcal{B}(M)$ is a $\mathcal{M}^*$-valued random variable $\mu : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\mathcal{M}^*, \mathcal{B}(M)^*\right)$ which, almost surely, takes values in $\mathcal{M}$. For the definition of a random probability measure, replace "measure(s)" with "probability measure(s)" in the above.

Being random, it is natural to wonder if it is possible to talk about the expectation of a random measure; this can indeed be done with the following definition of an intenstity measure.

Definition 2.2.19 (Intensity measure, adapted from Kallenberg (2017))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(M, \mathcal{B}(M))$ be a measurable space where M is a metric space. Let μ be a random measure on $\mathcal{B}(M)$. The *intensity measure of* μ is the function $\mathbb{E}(\mu) : \mathcal{B}(M) \to [0, \infty]$ given by $[\mathbb{E}(\mu)](B) = \mathbb{E}(\mu(B))$. See Definition 2.2.35 for the definition of the Lebesgue integral, which is used to define the expectation used on the right hand side of the previous sentence (note it does not always exist).

The attention is now turned to generalised notions of cumulative distribution functions, quantiles, and medians. This generalisation is at a high level of

abstraction, which is useful for the work in Chapter 3.

Definition 2.2.20 (Cumulative distribution function, see Simons (1974))**.** Let $(S, <_S)$ be a linearly ordered set. $A \subseteq S$ is called an *initial* if it holds that $\forall y \in A\,[x <_S y \implies x \in A]$ and a *terminal* if $\forall y \in A\,[y <_S x \implies x \in A]$. An *interval* is an intersection of an inital and a terminal. A set of the form $\{y \in S : y \leq_S x\}$ is called a *closed initial* and, dually, a set of the form $\{y \in S : x \leq_S y\}$ is called a *closed terminal*. The coarsest σ-field $\mathcal{F}_{<_S}$ on S which contains the intervals is called the *order σ-field*. Let $\mathbb{P}_S$ be a probability measure on $\mathcal{F}_{<_S}$. The *cumulative distribution function of* $\mathbb{P}_S$ is the function $F_S : S \to [0,1]$ given by $F_S(x) = \mathbb{P}_S(\{y \in S : y \leq_S x\})$. Now let $f : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(S, \mathcal{F}_{<_S}, <_S\right)$ be a random variable, and define the *cumulative distribution function of f* to be the function $F_f : S \to [0,1]$ given by $F_f(x) = \mathbb{P}(\{y \in \Omega : f(y) \leq_S x\})$.

Definition 2.2.21 (Open initials and terminals, see Simons (1974))**.** Let $(S, <_S)$ be a linearly ordered set. An initial is *open* if its complement is a closed terminal. A terminal is *open* if its complement is a closed initial.

Definition 2.2.22 (Left and right continuity, see Simons (1974))**.** Let $(S, <_S)$ be a linearly ordered set and let $f : S \to \mathbb{R}$ be some function. f is said to be *left continuous at a point x* if, for each $\epsilon > 0$, there exists an open terminal A such that $x \in A$ and $\forall u \in A\,[f(x) - \epsilon < f(u)]$. f is said to be *left continuous* if it is left continuous at each $x \in S$ for which $\{y \in S : x \leq_S y\} \neq S$. Right continuity at a point and right continuity are dually defined. *Total continuity* (or just *continuity*) of f means that f is both left and right continuous.

Definition 2.2.23 (Order topology and order σ-field, see Gálvez-Rodríguez and Sánchez-Granero (2019))**.** Let $(S, <_S)$ be a linearly ordered set. The *order topology on S* is the topology $\mathcal{T}_{<_S}$ generated (see Definition 2.3.5) by the open

initials and the open terminals. The *order σ-field* $\mathcal{B}(S)$ is then defined to be the σ-field generated by $\mathcal{T}_{<_S}$.

Definition 2.2.24 (Upper/lower semicontinuity, see Stromberg (2015))**.** Let $(S, \mathcal{T})$ be a topological space. Let $f : S \to [-\infty, \infty]$ be some function. f is said to be *lower semicontinuous at a point $x \in S$* (also known as *left $\mathcal{T}$-continuous at a point $x \in S$*) if, for each $y \in \mathbb{R}$ with $y < f(x)$, there exists a neighbourhood (see Definition 2.3.1) U of x such that $f(z) > y$ for each $z \in U$. f is said to be *lower semicontinuous* (also known as *left $\mathcal{T}$-continuous*) if it is lower semicontinuous at each $x \in S$. *Upper semicontinuity at a point $x \in S$* (*right $\mathcal{T}$-continuity at $x \in S$*) and *upper semicontinuity* (*right $\mathcal{T}$-continuity*) are defined similarly by reversing the inequalities.

Lemma 2.2.7 (Relation between types of left/right continuity)**.** *Let $\left(S, \mathcal{T}_{<_S}, <_S\right)$ be a linearly ordered topological space with the order topology, which has the least upper bound and greatest lower bound properties. A function $f : S \to \mathbb{R}$ is left $\mathcal{T}_{<_S}$-continuous (respectively right $\mathcal{T}_{<_S}$-continuous) if it is left continuous (respectively right continuous).*

Theorem 2.2.8 (Properties of a cumulative distribution function on a linearly ordered topological space)**.** *Let $\left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ be a linearly ordered topological space with the order topology and the order σ-field. Let $\mathbb{P}_S$ be a probability measure on $\mathcal{B}(S)$ and let $F_S : S \to [0, 1]$ be its cumulative distribution function. Then*

1. *F_S is non-decreasing*
2. *F_S is right continuous*
3. *if S does not have a minimal element, then* $\inf F_S(S) = 0$
4. $\sup F_S(S) = 1$.

Theorem 2.2.9 (A restatement of Theorem 7.7 of Gálvez-Rodríguez and Sánchez–Granero (2020))**.** *Let* $\left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ *be a linearly ordered topological space with the order topology and the order* σ*-field. Suppose that* $\left(S, \mathcal{T}_{<_S}, <_S\right)$ *is separable and has the least upper bound and greatest lower bound properties. Let* $F : S \to [0, 1]$ *be a non-decreasing and right continuous function satisfying* $\sup F(S) = 1$*. Then there exists a unique probability measure* $\mathbb{P}_S$ *on* $\mathcal{B}(S)$ *whose cumulative distribution function coincides with* F*.*

Remark 2.2.14. That Theorem 2.2.9 is a restatement of Theorem 7.7 of Gálvez-Rodríguez and Sánchez-Granero (2020) follows from combining point 7 in Section 39 of Part II in Steen and Seebach (1978) with their statement of Theorem 7.7, and applying Lemma 2.2.7.

Remark 2.2.15. Gálvez-Rodríguez and Sánchez-Granero (2022) give a way to construct a linearly ordered topological space from a so-called “fractal structure". While it is beyond the author’s objectives to discuss this notion technically, what is important to note here is that these objects are abundant so there is no shortage of instances where the previous results are applicable.

Definition 2.2.25 (Joint cumulative distribution function, due to the author)**.** Let $\mathcal{I}$ be a countable non-empty set and, for $i \in \mathcal{I}$, let $\left(S_i, \mathcal{B}(S_i), \mathbb{P}_i, \mathcal{T}_{<_i}, <_i\right)$ be a linearly ordered topological space with the order topology and the order σ-field that is equipped with a probability measure. Let $\left(S, \mathcal{B}(S), \mathbb{P}_S, \mathcal{T}_{<_S}, <_S\right)$ be the product of such spaces where $S = \bigtimes_{i\in\mathcal{I}} S_i$, $\mathbb{P}_S = \bigotimes_{i\in\mathcal{I}} \mathbb{P}_i$, $<_S$ is the product order, $\mathcal{T}_{<_S}$ is the product topology of the order topologies, and $\mathcal{B}(S)$ is the Borel σ-field generated by $\mathcal{T}_{<_S}$. Define the *joint cumulative distribution function of* $\mathbb{P}$ to be the function $F_S : S \to [0, 1]$ given by $F_S(x) = \mathbb{P}_S(\{y \in S : y \leq x\}) = \mathbb{P}_S\left(\bigcap_{i\in\mathcal{I}} \{y \in S : y(i) \leq_i x(i)\}\right)$. Now suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and let $f := (f_i)_{i\in\mathcal{I}}$ be a stochastic process

where, for $i \in \mathcal{I}$, f_i is a S_i-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Define the *joint cumulative distribution function of* f to be the function $F_f : S \to [0,1]$ given by $F_f(x) = \mathbb{P}\left(\bigcap_{i \in \mathcal{I}} \{y \in \Omega : f_i(y) \leq_i x(i)\}\right)$.

Remark 2.2.16. By considering Theorem 2.2.9, it is seen that, if the spaces are separable and have the least upper bound and greatest lower bound properties, then the joint cumulative distribution function of some stochastic process characterises the product probability measure provided that the process is independent. Note that, by Tychonoff's theorem (see Theorem 17.8 of Willard (1970)) and Point 7 in Section 39 of Part II in Steen and Seebach (1978), the product space is also separable and has the least upper bound and greatest lower bound properties.

Definition 2.2.26 (Quantile functions and medians, due to the author)**.** Let $\left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ be a linearly ordered topological space with the order topology and the order σ-field. Suppose that $\left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ is separable and has the least upper bound and greatest lower bound properties. It is convenient here to introduce the notation $[a, b] \coloneqq \{x \in S : a \leq_S x \leq_S b\}$ for $a, b \in S$ with $a \leq_S b$. Let $\mathbb{P}_S$ be a probability measure on $(S, \mathcal{B}(S))$ and let F_S be its cumulative distribution function. Define the *quantile function of* $\mathbb{P}_S$ to be the function $Q_S : [0,1] \to \mathcal{P}(S)$ given by

$$Q_S(p) = \begin{cases} \emptyset & p = 0 \\ \left[\sup\{x \in S : F_S(x) < p\}, \sup\{x \in S : F_S(x) \leq p\}\right] & p \in (0,1) \\ S & p = 1 \end{cases}$$

A *median of* $\mathbb{P}_S$ is any member of $Q_S(0.5)$. Now let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $f : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ be a random variable with cumulative distribution function F_f. The *quantile function of* f is defined to be

the function $Q_f : [0,1] \to \mathcal{P}(S)$ given by

$$Q_f(p) = \begin{cases} \emptyset & p = 0 \\ \left[\sup\left\{x \in S : F_f(x) < p\right\}, \sup\left\{x \in S : F_f(x) \leq p\right\}\right] & p \in (0,1) \\ S & p = 1 \end{cases}$$

A *median of* f is any member of $Q_f(0.5)$.

Theorem 2.2.10 (Probability integral transform)**.** *Let* $\left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ *be a linearly ordered topological space with the order topology and the order* σ*-field. Suppose that* $\left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ *is separable and has the least upper bound and greatest lower bound properties. Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $f : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(S, \mathcal{B}(S), \mathcal{T}_{<_S}, <_S\right)$ *be a* S*-valued random variable with cumulative distribution function* F_f*. Suppose* F_f *is continuous. Then* $g \coloneqq F_f \circ f$ *has a standard uniform distribution.*

A standard way to model multivariate distributions in Statistics is to consider a collection of standard uniform random variables coupled with what is known as a copula. This notion is now defined and it is conjectured that they can be used in the more general setting than real random vectors which was previously worked in.

Definition 2.2.27 (Copulas, see Sklar (1973))**.** Let $n \in \mathbb{N}$. An *n-interval* is a set of the form $\bigtimes_{i \in \mathbb{N}_n} [x(i), y(i)]$ for some $x, y \in \mathbb{R}^n$ with, for $i \in \mathbb{N}_n$, $x(i) \leq y(i)$. An *n-place real function* is a function whose domain is a non-empty subset of $[-\infty, \infty]^n$ and whose range is a subset of $\mathbb{R}$. Let G be an n-place real function. Suppose $N = \bigtimes_{i \in \mathbb{N}_n} [x(i), y(i)]$ is an n-interval whose points are in $D(G)$ (the domain of G) . The *G-volume of* N is the sum

$$V_G(N) \coloneqq \sum_{z \in N} \alpha(z) G(Z)$$

where

$$\alpha(z) \coloneqq \begin{cases} 1 & \text{if } z(i) = x(i) \text{ for an even number of } i\text{'s} \\ -1 & \text{if } z(i) = x(i) \text{ for an odd number of } i\text{'s} \end{cases}$$

G is said to be *n-increasing* if $V_G(N) \geq 0$ for any n-interval whose points lie in $D(G)$. An *n-copula* is an n-place real function with $D(G) = [0,1]^n$, $G(D(G)) = [0,1]$, and which satisfies

1. $\forall m \in \mathbb{N}_n \forall i \in \mathbb{N}_n \forall x \in D(G)\ [[(i \neq m) \wedge (x(i) = 1)] \implies C(x) = x(m)]$
2. $C(x) = 0$ if $x(i) = 0$ for some $i \in \mathbb{N}_n$
3. C is n-increasing.

Conjecture 2.2.1 (A generalisation of Sklar's theorem). Let $\mathcal{I} = \mathbb{N}_n$ for some $n \in \mathbb{N}$ and, for $i \in \mathcal{I}$, let $\left(S_i, \mathcal{B}(S_i), \mathbb{P}_i, \mathcal{T}_{<_i}, <_i\right)$ be a linearly ordered topological space with the order topology and the order σ-field that is equipped with a probability measure. Let $\left(S, \mathcal{B}(S), \mathbb{P}_S, \mathcal{T}_{<_S}, <_S\right)$ be the product of such spaces where $S = \bigtimes_{i\in\mathcal{I}} S_i$, $\mathbb{P}_S = \bigotimes_{i\in\mathcal{I}} \mathbb{P}_i$, $<_S$ is the product order, $\mathcal{T}_{<_S}$ is the product topology of the order topologies, and $\mathcal{B}(S)$ is the Borel σ-field generated by $\mathcal{T}_{<_S}$. For $i \in \mathcal{I}$, let F_i be the cumulative distribution function of $\mathbb{P}_i$. Let F_S be the joint cumulative distribution function of $\mathbb{P}$. Then there exists an n-copula C such that

$$F_S(x) = C(x^*) \tag{2.1}$$

where $x^* : \mathbb{N}_n \to [0,1]$ is given by $x^*(i) = F_i(x(i))$. Furthermore, if, for $i \in \mathcal{I}$, F_i is continuous, then C is unique; if not, C is uniquely determined on $\bigtimes_{i\in\mathcal{I}} F_i(S_i)$. Conversely, if C is an n-copula and, for $i \in \mathcal{I}$, F_i is a function satisfying the properties of a cumulative distribution function given in Theorem 2.2.8, then the function F_S determined by Equation (2.1) is a joint cumulative distribution function.

The following result is a significant generalisation of Lemma 3.1 from Artemiou and Li (2009), which is used to prove a major result of Chapter 3.

Theorem 2.2.11 (A unique median result)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, let* $\mathcal{I}$ *be a non-empty set, and let* $(M, \mathcal{B}(M), \mathcal{T}_M)$ *be a topological vector space, over some field, with the Borel* σ*-field. Let* $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{B}(M), \mathcal{T}_M)$ *be a non-zero* M*-valued random variable. Let* $V \coloneqq (v_i)_{i \in \mathcal{I}}$ *be a stochastic process where, for* $i \in \mathcal{I}$*,* v_i *is a non-zero* M*-valued random variable on* $(\Omega, \mathcal{F}, \mathbb{P})$*. Let* $\mathcal{T}_M^{\mathcal{I}}$ *be the product topology (see Definition 2.3.6) and let* $V^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^{\mathcal{I}}, [\mathcal{B}(M)]^{\mathcal{I}}, \mathcal{T}_M^{\mathcal{I}}\right)$ *be the random variable given by* $V^*(x) = g_x$ *where* $g_x : \mathcal{I} \to M$ *is itself given by* $g_x(i) = v_i(x)$*. Suppose that the stochastic process* (V^*, f) *is independent (written* $V^* \perp\!\!\!\perp f$*). Let* $z \coloneqq (z_i)_{i \in \mathcal{I}}$ *be the stochastic process where, for* $i \in \mathcal{I}$*,* $z_i : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^2, [\mathcal{B}(M)]^2, \mathcal{T}_M^2\right)$ *is given by* $z_i(x) = (v_i(x), f(x))^T$ *(viewing* M^2 *as the vector space of ordered pairs of elements of* M*). Let* $T : \left(M^2, \mathcal{F}_M^2\right) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *be a bilinear map where any pair with both entries non-zero is mappped to a non-zero real number. Let* $i, j \in \mathcal{I}$ *be distinct. Suppose* $\mathbb{P}(\{x \in \Omega : f(x) \in G\}) > 0$ *for any non-empty open set* $G \in \mathcal{T}_M$*. Suppose* $\left(v_i, v_j\right)$ *is exchangeable and that* $v_i(x)$ *and* $v_j(x)$ *are linearly independent for any* $x \in \Omega$*. Then* $\frac{(T \circ z_i)^2}{(T \circ z_j)^2}$ *has a unique median of* 1.

Notions of elliptical symmetry, spherical symmetry, and unitary invariance of real Hilbert space valued random variables are now defined. These are essential for Chapter 3, and were either used in Jones and Artemiou (2019, 2021) and Jones et al. (2020) or originally defined in those papers.

Definition 2.2.28 (Characteristic function)**.** The *characteristic function* $\Phi_A : \mathcal{H} \to \mathbb{C}$ of an $\mathcal{H}$-valued random variable A on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{H}$ is a real Hilbert space, is defined by $\Phi_A(f) \coloneqq \mathbb{E}\left(\exp\left(i \langle f, A \rangle_{\mathcal{H}}\right)\right)$. See

Definition 2.5.6 for the definition of expectation for a B-valued random variable where B is a Banach space over $\mathbb{C}$.

Definition 2.2.29 (Elliptically symmetric distribution)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ be a measurable space where $\mathcal{H}$ is a Hilbert space over $\mathbb{R}$. An $\mathcal{H}$-valued random variable A on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to have an *elliptically symmetric distribution* if there exists $\mu \in \mathcal{H}$, a nuclear, non-negative definite, self-adjoint operator (see Section 2.4 for the definitions) $\Psi : \mathcal{H} \to \mathcal{H}$, and a function $\varphi : [0, \infty) \to \mathbb{R}$ such that the characteristic function of $A - \mu$ has the form:

$$\Phi_{A-\mu}(f) = \varphi\left(\langle f, \Psi f \rangle_{\mathcal{H}}\right)$$

for all $f \in \mathcal{H}$.

Remark 2.2.17. Definition 2.2.29 is taken from Li (2007b). For such an A, Li (2007b) shows that if $\mathbb{E}(A)$ and $\mathrm{Var}(A)$ exist (see Section 2.5 for the definitions) then $\mathbb{E}(A) = \mu$ and Ψ is a non-negative multiple of $\mathrm{Var}(A)$. φ must be real-valued; to see this, let $f \in \mathcal{H}$ then consider that $\Phi_{A-\mu}(-f) = \overline{\Phi_{A-\mu}(f)}$ and $\Phi_{A-\mu}(-f) = \varphi\left(\langle -f, \Psi(-f)\rangle_{\mathcal{H}}\right) = \varphi\left(\langle f, \Psi(f)\rangle_{\mathcal{H}}\right) = \Phi_{A-\mu}(f)$. Definition 2.2.30 now defines spherical symmetry for an $\mathcal{H}$-valued random variable, where $\mathcal{H}$ is a real Hilbert space, when the space is finite-dimensional.

Definition 2.2.30 (Spherically symmetric distribution)**.** Suppose $\mathcal{H}$ is a finite-dimensional Hilbert space over $\mathbb{R}$. An $\mathcal{H}$-valued random variable A on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to have an *spherically symmetric distribution* if it has an elliptically symmetric distribution with $\mu = 0$ and Ψ being a non-negative multiple of the identity operator.

Remark 2.2.18. Definition 2.2.30 requires the Hilbert space $\mathcal{H}$ to be finite-dimensional as the identity operator is not nuclear on infinite-dimensional spaces.

Remark 2.2.19. Definition 2.2.30 is adapted from Boente et al. (2014). They remarked that a real random vector f (here viewed as a stochastic process, but there is an equivalence) is spherically distributed if and only if its distribution is invariant under orthogonal transformations. This observation motivates Definition 2.2.31.

Definition 2.2.31 (Unitarily invariant random variables, due to the author originally in Jones et al. (2020))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{H}$ be a finite-dimensional Hilbert space over $\mathbb{R}$. A random variable $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is said to be *unitarily invariant* if, for any unitary operator (see Definition 2.4.11) $U : \mathcal{H} \to \mathcal{H}$, $\mathbb{P}_{U \circ f} = \mathbb{P}_f$.

Lemma 2.2.12 (Spherically distributed random variables have spherically distributed Fourier coefficients, first appeared in Jones et al. (2020))**.** *Suppose $\mathcal{H}$ is a p-dimensional ($p \in \mathbb{N}$) Hilbert space over $\mathbb{R}$ and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is spherically symmetric then, for any orthonormal basis $\{v_1, \ldots, v_p\}$, so is the random vector $\langle v, f \rangle_{\mathcal{H}} : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ (viewing $\mathbb{R}^p$ in its classical form as the real vector space of p-tuples of real numbers) given by $[\langle v, f \rangle_{\mathcal{H}}](x) = \left(\langle v_1, f \rangle_{\mathcal{H}}, \ldots, \langle v_p, f \rangle_{\mathcal{H}}\right)^T$. Furthermore, the stochastic process $\left(\langle v_1, f \rangle_{\mathcal{H}}, \ldots, \langle v_p, f \rangle_{\mathcal{H}}\right)$ is exchangeable.*

Lemma 2.2.13 (A relation between unitary operators, first appeared in Jones et al. (2020))**.** *Let $\mathcal{H}$ be a real Hilbert space with $\dim \mathcal{H} = p$ for some $p \in \mathbb{N}$. Suppose that $\{v_1, \ldots, v_p\}$ forms an orthonormal basis for $\mathcal{H}$ and let $T : \mathbb{R}^p \to \mathbb{R}^p$ be a unitary operator (see Definition 2.4.11). Then the operator $U : \mathcal{H} \to \mathcal{H}$ defined by*

$$U(h) \coloneqq \sum_{j \in \mathbb{N}_p} T_j(C) v_j$$

is a unitary operator. C is the coordinate of h and $T_j(C)$ denotes the j^{th} component of $T(C)$.

Lemma 2.2.14 (A relation between unitary operators, first appeared in Jones et al. (2020))**.** *Let $\mathcal{H}$ be a real Hilbert space with* $\dim \mathcal{H} = p$ *for some $p \in \mathbb{N}$. Suppose that $\{v_1, \ldots, v_p\}$ forms an orthonormal basis for $\mathcal{H}$ and let $U : \mathcal{H} \to \mathcal{H}$ be a unitary operator. Let h be an arbitrary element of $\mathcal{H}$ with coordinate C, and let D be the coordinate of $U(h)$. Define the operator $T : \mathbb{R}^p \to \mathbb{R}^p$ by $T(C) \coloneqq D$. Then T is a unitary operator on $\mathbb{R}^p$.*

Theorem 2.2.15 (Relation between unitary invariance and spherical symmetry, first appeared in Jones et al. (2020))**.** *Suppose $\mathcal{H}$ is a real Hilbert space with* $\dim \mathcal{H} = p$ *for some $p \in \mathbb{N}$ and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An $\mathcal{H}$-valued random variable $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is unitarily invariant if and only if, for any orthonormal basis $\{v_i\}_{i \in \mathbb{N}_p}$, the random variable $\langle v, f \rangle_{\mathcal{H}} : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ given by $[\langle v, f \rangle_{\mathcal{H}}](x) = \left(\langle v_1, f(x) \rangle_{\mathcal{H}}, \ldots, \langle v_p, f(x) \rangle_{\mathcal{H}} \right)^T$ is a spherically distributed $\mathbb{R}^p$-valued random variable.*

Now turned to are definitions of random operators, with a spectral decomposition, which satisfy some kind of invariance assumption. Like the notions recently defined, these are essential to the work in Chapter 3.

Definition 2.2.32 (Orientationally uniform random operator, see Artemiou and Li (2009))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{H}$ be a real p-dimensional ($p \in \mathbb{N}$) Hilbert space. A random operator $\Sigma : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}(\mathcal{H}, \mathcal{H}), \mathcal{B}(\mathcal{L}(\mathcal{H}, \mathcal{H})))$ (see Definition 2.4.1) is said to have an *orientationally uniform distribution* if there exists positive distinct random variables $(\lambda_i)_{i \in \mathbb{N}_p}$ and $\mathcal{H}$-valued random variables $(v_i)_{i \in \mathbb{N}_p}$ such that

1. $\Sigma \overset{\text{a.s.}\mathbb{P}}{=} \sum_{i \in \mathbb{N}_p} \lambda_i (v_i \otimes v_i)$ (see Definition 2.5.7)
2. $(\lambda_i)_{i \in \mathbb{N}_p}$ is exchangeable

3. $(v_i)_{i\in\mathbb{N}_p}$ is exchangeable and, for any $x \in \Omega$, $\{v_i(x)\}_{i\in\mathbb{N}_p}$ is an orthonormal basis of $\mathcal{H}$

4. the stochastic process $\left((\lambda_i)_{i\in\mathbb{N}_p}, (v_i)_{i\in\mathbb{N}_p}\right)$ is independent.

Remark 2.2.20. Definition 2.2.32 is a generalisation of the definition of an orientationally uniform random $p \times p$ matrix given in Artemiou and Li (2009). Their definition required that the distribution of the random vector $(\lambda_1, \ldots, \lambda_p)$ be absolutely continuous with respect to Lebesgue measure, though that assumption does not appear to be used in subsequent developments in their work.

Definition 2.2.33 (Unitarily invariant random operators, originally in Jones et al. (2020))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{H}$ be a real p-dimensional ($p \in \mathbb{N}$) Hilbert space. A random self-adjoint operator (see Section 2.4 for the definitions) $\Sigma : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}(\mathcal{H}, \mathcal{H}), \mathcal{B}(\mathcal{L}(\mathcal{H}, \mathcal{H})))$ (see Definition 2.4.8) is said to be *unitarily invariant* if, for any unitary operator $U : \mathcal{H} \to \mathcal{H}$, $\Sigma \overset{D}{=} U\Sigma U^{-1}$ where $U\Sigma U^{-1} : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}(\mathcal{H}, \mathcal{H}), \mathcal{B}(\mathcal{L}(\mathcal{H}, \mathcal{H})))$ is given by $\left[U\Sigma U^{-1}\right](x) = U\Sigma(x)U^{-1}$.

Finally for this section, the standard definition of the Lebesgue integral for complex-valued measurable functions is recalled.

Definition 2.2.34 (Simple real-valued functions, see Bass (2016))**.** Let $(M, \mathcal{F}_M)$ be a measurable space. A *simple real-valued function on M* is a function $g : M \to \mathbb{R}$ of the form

$$g(x) = \sum_{i\in\mathbb{N}_m} a_i \mathbf{1}_{E_i}(x)$$

for some $m \in \mathbb{N}$, $a_i \in \mathbb{R}$ ($i \in \mathbb{N}_m$), and $E_i \in \mathcal{F}_M$.

Definition 2.2.35 (Lebesgue integral, see Bass (2016))**.** Let $(M, \mathcal{F}_M, \mu)$ be a measure space. If $s : (M, \mathcal{F}_M, \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a non-negative measurable

simple function defined on M with representation $s = \sum_{i\in\mathbb{N}_m} a_i \mathbf{1}_{E_i}$, define the *Lebesgue integral of s* to be

$$\int_M s \, \mathrm{d}\mu \coloneqq \sum_{i\in\mathbb{N}_m} a_i \mu(E_i).$$

Here, the convention $0 \cdot \infty = 0$ is adopted. If $f : (M, \mathcal{F}_M, \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a non-negative measurable function, define the *Lebesgue integral of f* to be

$$\int_M f \, \mathrm{d}\mu \coloneqq \sup\left\{\int_M s \, \mathrm{d}\mu : 0 \le s \le f, s \text{ is simple}\right\}.$$

For measurable $g : (M, \mathcal{F}_M, \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, let $g^+ \coloneqq \max\{g, 0\}$ and $g^- \coloneqq \max\{-g, 0\}$. Provided $\int_M g^+ \, \mathrm{d}\mu$ and $\int_M g^- \, \mathrm{d}\mu$ are not both infinite, define the *Lebesgue integral of g* to be

$$\int_M g \, \mathrm{d}\mu \coloneqq \int_M g^+ \, \mathrm{d}\mu - \int_M g^- \, \mathrm{d}\mu.$$

For measurable $h : (M, \mathcal{F}_M, \mu) \to (\mathbb{C}, \mathcal{B}(\mathbb{C}))$, define the *Lebesgue integral* to be

$$\int_M h \, \mathrm{d}\mu \coloneqq \int_M \Re(h) \, \mathrm{d}\mu + i \int_M \Im(h) \, \mathrm{d}\mu$$

provided that $\int_M \Re(h) \, \mathrm{d}\mu$ and $\int_M \Im(h) \, \mathrm{d}\mu$ both exist, where $\Re(h) : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is given by $[\Re(h)](x) = \Re(h(x))$ and $\Im(h)$ is defined analogously.

Remark 2.2.21. The standard properties of the Lebesgue integral are given in virtually every measure-theoretic probability textbook, so are not explicitly noted here. Taylor (2006) shows that the integral does not depend on the representation of the simple functions.

2.3 Topology

This section is entirely dedicated to Theorem 2.3.1, with all definitions (bar net convergence, which is used in Section 2.4) being needed in the proof.

Definition 2.3.1 (Neighbourhoods, see Willard (1970))**.** Let $(S, \mathcal{T})$ be a topological space. For $x \in S$, a *neighbourhood of* x is a subset $U \subseteq S$ which contains an open set containing x.

Definition 2.3.2 (Net convergence, see Bass (2016))**.** Let $(S, \mathcal{T})$ be a topological space and $(D, \leq)$ be a directed set. A net $f : D \to S$ is said to *converge to* $y \in S$ if, for each $G \in \mathcal{T}$ with $y \in G$, there exists $\alpha_0 \in D$ such that $f(\alpha) \in G$ whenever $\alpha_0 \leq \alpha$.

Definition 2.3.3 (Topological bases, see Bass (2016))**.** Let $(S, \mathcal{T})$ be a topological space. $\mathcal{C} \subseteq \mathcal{T}$ is called a *base* for $\mathcal{T}$ if every element of $\mathcal{T}$ can be written as a union of elements in $\mathcal{C}$.

Definition 2.3.4 (Topological subbases, see Bass (2016))**.** Let $(S, \mathcal{T})$ be a topological space. $\mathcal{C} \subseteq \mathcal{T}$ is called a *subbase* for $\mathcal{T}$ if the set of finite intersections of elements of $\mathcal{C}$ is a base for $\mathcal{T}$.

Definition 2.3.5 (Topology generated by a set of sets, see Bass (2016))**.** Let S be a set and let $\mathcal{C} \subseteq \mathcal{P}(S)$. Define the *topology generated by* $\mathcal{C}$, written $G(\mathcal{C})$, to be the intersection of all topologies on S which have $\mathcal{C}$ as a subbase.

Remark 2.3.1. Let $(S, \mathcal{T})$ be a topological space. Referring to the construction given by ordinals in Remark 2.2.4, it is seen that if $\mathcal{C}$ is a countable base for $\mathcal{T}$ then $\sigma(\mathcal{T}) = \sigma(\mathcal{C})$. Furthermore, if $\mathcal{D}$ is a countable family of subsets of S then $\sigma(\mathcal{D}) = \sigma(G(\mathcal{D}))$.

Definition 2.3.6 (Product topology, see Engelking (1989))**.** Let $\mathcal{I}$ be a non-empty set and let $(S_i, \mathcal{T}_i)$ be a topological space for any $i \in \mathcal{I}$. The *product topology* $\mathcal{T}$ on $S \coloneqq \bigtimes_{i \in \mathcal{I}} S_i$ is defined to the topology generated by $\left\{\phi_i^{-1}(U_i) : i \in \mathcal{I}, U_i \in \mathcal{T}_i\right\}$ where ϕ_i is the projection from S to S_i. Define $\mathcal{B}(S)$ to be $\sigma(\mathcal{T})$.

Remark 2.3.2. The product topology is the coarsest topology for which the projection maps are all continuous.

Theorem 2.3.1 (Relation between the Borel σ-field on the product space with the tensor product σ-field)**.** *Let $m \in \mathbb{N} \cup \{\omega\}$ and let (S_i, d_i) be a separable metric space for each $i \in \mathbb{N}_m$. Let $S \coloneqq \bigtimes_{i \in \mathbb{N}_m} S_i$. Then $(S, \mathcal{T})$ is also a separable metric space where $\mathcal{T}$ is the product topology. Furthermore,*

$$\mathcal{B}(S) = \bigotimes_{i \in \mathbb{N}_m} \mathcal{B}(S_i).$$

2.4 Functional Analysis

The definitions given in this section are adapted from Hsing and Eubank (2015), except where it is stated that they are due to the author or otherwise referenced. Proofs of any unreferenced lemmas are given at the end of the chapter. The purpose of this section is to give the necessary theory from Functional Analysis that is required for Section 2.5, Chapter 3, and Chapter 4.

Definition 2.4.1 (L^p spaces, due to the author)**.** Let $(M, \mathcal{F}_M, \mu)$ be a measure space and let $p \in \{x \in \mathbb{R} : x \geq 1\}$. Let $L^p\left((M, \mathcal{F}_M, \mu), B\right)$ be the set of all B-valued (where B is a complex Banach space) measurable functions on M for which the integral $\int \|f\|_B \, \mathrm{d}\mu$ exists where functions which are almost surely equal are considered equivalent. If B is defined over the reals and $B = \mathbb{R}$, write $L^p\left((M, \mathcal{F}_M, \mu)\right)$ and, when there is no potential ambiguity, write $L^p(\mu)$ as shorthand. Define the p-norm

$$\|f\|_p \coloneqq \left(\int \|f\|_B^p \right)^{1/p}$$

where $f \in L^p\left((M, \mathcal{F}_M, \mu), B\right)$.

Remark 2.4.1. Hsing and Eubank (2015) only gives Definition 2.4.1 for the case $B = \mathbb{R}$. They give a proof that $L^p\left(M, \mathcal{F}_M, \mu\right)$ is complete which can be trivially adapted (by replacing the absolute value with the B norm) to show that $L^p\left(\left(M, \mathcal{F}_M, \mu\right), B\right)$ is complete.

Definition 2.4.2 (Subsets which are orthogonal, see Muscat (2014))**.** Subsets A and B of a complex Hilbert space $\mathcal{H}$ are said to be *orthogonal*, denoted by $A \perp B$, if $\langle f, g\rangle_{\mathcal{H}} = 0$ for any $f \in A$ and $g \in B$.

Remark 2.4.2. For a complex Hilbert space $\mathcal{H}$, it is supposed that the inner product is linear in the first entry and conjugate linear in the second.

Definition 2.4.3 (Orthonormal subsets and bases, see Conway (1990))**.** An *orthonormal subset* of a complex Hilbert space $\mathcal{H}$ is a subset E with the properties: (1) for any $e \in E$, $\|e\|_{\mathcal{H}} = 1$; (2) for any two distinct $e_1, e_2 \in E$, $\langle e_1, e_2\rangle_{\mathcal{H}} = 0$. An *orthonormal basis for* $\mathcal{H}$ is a maximal orthonormal subset (where the ordering is given by set inclusion).

Remark 2.4.3. Proposition 4.14 in Chapter 1 of Conway (1990) gives that all orthonormal bases for a complex Hilbert space have the same cardinality. Proposition 4.16 of the same chapter in the same text gives that a complex infinite-dimensional Hilbert space is separable if and only if the cardinality of any orthonormal basis is equal to $\aleph_0$.

Definition 2.4.4 (Summations over arbitrary sets, due to the author)**.** Let B be a complex Banach space and let $\mathcal{I}$ be a set. Suppose that $(v_i)_{i\in\mathcal{I}}$ is a family of elements of B. Let $\mathcal{F}$ be the set of all finite subsets of $\mathcal{I}$ and order $\mathcal{F}$ by inclusion so that it becomes a directed set. Define the *summation* $\bigoplus_{i\in\mathcal{I}} v_i$ to be the limit, if it exists, of the net $f : \mathcal{F} \to B$ given by $f(F) = \sum_{i\in F} v_i$.

Remark 2.4.4. The concept behind Definition 2.4.4 is used in Conway (1990).

Definition 2.4.5 (Products over arbitrary sets, due to the author)**.** Let B be a complex unital Banach algebra and let $\mathcal{I}$ be a set. Suppose that $(v_i)_{i\in\mathcal{I}}$ is a family of elements of B. Let $\mathcal{F}$ be the set of all finite subsets of $\mathcal{I}$ and order $\mathcal{F}$ by inclusion so that it becomes a directed set. Define the *product* $\bigotimes_{i\in\mathcal{I}} v_i$ to be the limit, if it exists, of the net $f : \mathcal{F} \to B$ given by $f(F) = \prod_{i\in F} v_i$.

Remark 2.4.5. Definition 2.4.5 is inspired by Definition 2.4.4.

Remark 2.4.6. Let $\mathcal{H}$ be a complex Hilbert space and let $\mathcal{I}$ be a set. Let $(v_i)_{i\in\mathcal{I}}$ be a family of elements of $\mathcal{H}$. If $\mathcal{I} = \mathbb{N}_m$ for some $m \in \mathbb{N}$ then it is straightforward to see that $\bigoplus_{i\in\mathcal{I}} v_i = \sum_{i\in\mathcal{I}} v_i$. If $\mathcal{I} = \mathbb{N}$ then the statement of Exercise 10 in Section 4 of Chapter 1 of Conway (1990) says that if $\bigoplus_{i\in\mathcal{I}} v_i$ exists then so does $\sum_{i\in\mathcal{I}} v_i$ and they agree; however the existence of $\sum_{i\in\mathcal{I}} v_i$ does not necessarily imply the existence of $\bigoplus_{i\in\mathcal{I}} v_i$.

Remark 2.4.7. Theorem 4.13 in Chapter 1 of Conway (1990) gives that if E is an orthonormal basis for a complex Hilbert space $\mathcal{H}$ then, for any $h \in \mathcal{H}$, h is equal to $\bigoplus_{e\in E} \langle h, e\rangle_{\mathcal{H}}\, e$. Furthermore, Corollary 4.9 of the same chapter in the same text gives that only countably many of the Fourier coefficients are nonzero. Therefore, it is often reasonable to assume that $\mathcal{H}$ is separable.

Definition 2.4.6 (Dimension, see Conway (1990))**.** The *dimension* of a Hilbert space $\mathcal{H}$, denoted by $\dim(\mathcal{H})$, is defined to be the cardinality of any orthonormal basis for $\mathcal{H}$.

Definition 2.4.7 (Tensor product of Hilbert spaces, see Conway (1990))**.** Let $\mathcal{I}$ be a non-empty set and let $\{\mathcal{H}_i : i \in \mathcal{I}\}$ be a set of complex Hilbert spaces. Define their *tensor product* $\mathcal{H} = \bigotimes_{i\in\mathcal{I}} \mathcal{H}_i$ to be

$$\left\{ h : \mathcal{I} \to \bigcup_{i\in\mathcal{I}} \mathcal{H}_i \;\middle|\; (\forall i \in \mathcal{I}\,[h(i) \in \mathcal{H}_i]) \wedge \left(\bigoplus_{i\in\mathcal{I}} \|h(i)\|^2_{\mathcal{H}_i} < \infty \right) \right\}.$$

$\mathcal{H}$ is made into a complex Hilbert space by equipping it with the inner product $\langle f, g\rangle_{\mathcal{H}} \coloneqq \bigoplus_{i\in\mathcal{I}} \langle f(i), g(i)\rangle_{\mathcal{H}_i}$.

Remark 2.4.8. What is here called the tensor product, Conway (1990) actually calls a direct sum. This terminology is avoided here to reduce the risk of confusion with that of the usual sum of vector spaces.

Definition 2.4.8 (Bounded operators)**.** Let $\left(B_1, \|\cdot\|_{B_1}\right)$ and $\left(B_2, \|\cdot\|_{B_2}\right)$ be complex Banach spaces. A linear transformation $A : B_1 \to B_2$ is called a *bounded operator* if there exists a positive constant C such that, for all $x \in B_1$, $\|Ax\|_{B_2} \leq C \|x\|_{B_1}$. Let $\mathcal{L}(B_1, B_2)$ denote the collection of all such operators.

Remark 2.4.9. Proposition 8.2 of Muscat (2014) gives that every bounded operator between two complex Banach spaces is continuous and vice versa.

Remark 2.4.10. For complex Banach spaces B_1 and B_2, Theorem 3.1.3 of Hsing and Eubank (2015) gives that $\mathcal{L}(B_1, B_2)$ is also a complex Banach space when considered to be equipped with the *operator norm*

$$\|A\|_{\mathcal{L}(B_1,B_2)} \coloneqq \sup_{x\in B_1, \|x\|_{B_1}=1} \|Ax\|_{B_2} .$$

where $A \in \mathcal{L}(B_1, B_2)$. Throughout this thesis, it is assumed that the space of bounded operators between complex Banach spaces is considered with the operator norm. This implies that $\mathcal{B}(\mathcal{L}(B_1, B_2))$ is the Borel σ-field generated by the topology induced by this norm.

Definition 2.4.9 (Compact operators)**.** Let B_1, B_2 be complex Banach spaces. The operator $T \in \mathcal{L}(B_1, B_2)$ is called a *compact operator* if, for any bounded sequence $\{x_i\}_{i\in\mathbb{N}}$ of elements of B_1, it holds that $\{Tx_i\}_{i\in\mathbb{N}}$ contains a convergent subsequence in B_2.

Definition 2.4.10 (Isomorphisms between Hilbert spaces, see Conway (1990))**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be complex Hilbert spaces. A bounded linear bijection $U : \mathcal{H}_1 \to \mathcal{H}_2$ is said to be an *isomorphism* if $\langle Uf, Ug\rangle_{\mathcal{H}_2} = \langle f, g\rangle_{\mathcal{H}_1}$ for any $f, g \in \mathcal{H}_1$.

Remark 2.4.11. Theorem 5.4 in Chapter 1 of Conway (1990) gives that there is an isomorphism between two complex Hilbert spaces if and only if they have the same dimension.

Definition 2.4.11 (Unitary operators on a Hilbert space, see Conway (1990))**.** Let $\mathcal{H}$ be a complex Hilbert space. A bounded linear bijection $U : \mathcal{H} \to \mathcal{H}$ is said to be a *unitary operator* if $\langle Uf, Ug\rangle_{\mathcal{H}} = \langle f, g\rangle_{\mathcal{H}}$ for any $f, g \in \mathcal{H}$.

Definition 2.4.12 (Adjoint operators, see Conway (1990))**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be complex Hilbert spaces. Let $A \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$. The *adjoint* of A is the unique (by the Riesz representation theorem) $A^* \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$ such that

$$\forall f \in \mathcal{H}_1, g \in \mathcal{H}_2 \left[\langle f, A^* g\rangle_{\mathcal{H}_1} = \langle Af, g\rangle_{\mathcal{H}_2}\right].$$

When $\mathcal{H}_1 = \mathcal{H}_2$, A is said to be *self-adjoint* if $A = A^*$.

Remark 2.4.12. Proposition 2.5 in Chapter 1 of Conway (1990) gives that $U \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ is an isomorphism if and only if $U^{-1} = U^*$. As a trivial corollary, if $\mathcal{H}_1 = \mathcal{H}_2$ then U is a unitary operator if and only if $U^{-1} = U^*$.

Definition 2.4.13 (Tensor product of bounded operators, adapted from Conway (1990))**.** Let $\mathcal{I}$ be a non-empty set and let $\{\mathcal{H}_i : i \in \mathcal{I}\}$ and $\{\mathcal{G}_i : i \in \mathcal{I}\}$ be sets of complex Hilbert spaces. Let $\mathcal{H} := \bigotimes_{i\in\mathcal{I}} \mathcal{H}_i$ and $\mathcal{G} := \bigotimes_{i\in\mathcal{I}} \mathcal{G}_i$. For $i \in \mathcal{I}$, let $A_i \in \mathcal{L}(\mathcal{H}_i, \mathcal{G}_i)$ and suppose that $\sup_{i\in\mathcal{I}} \|A_i\|_{\mathcal{L}(\mathcal{H}_i,\mathcal{G}_i)} < \infty$. The operator $A \in \mathcal{L}(\mathcal{H}, \mathcal{G})$ which satisfies, for any $h \in \mathcal{H}$ and $i \in \mathcal{I}$, $(Ah)(i) = A_i h(i)$ is called the *tensor product* of the operators A_i $(i \in \mathcal{I})$. It is denoted by $A = \bigotimes_{i\in\mathcal{I}} A_i$. It has norm $\|A\|_{\mathcal{L}(\mathcal{H},\mathcal{G})} = \sup_{i\in\mathcal{I}} \|A_i\|_{\mathcal{L}(\mathcal{H}_i,\mathcal{G}_i)}$.

Remark 2.4.13. What is here called the tensor product, Conway (1990) actually calls a direct sum. This terminology is avoided here to reduce the risk of confusion with that of the usual sum of bounded operators.

Theorem 2.4.1 (The tensor product of closed ranges is the closed range of the tensor product operators)**.** *Let $\mathcal{I}$ be a non-empty set and let $\{\mathcal{H}_i : i \in \mathcal{I}\}$ and $\{\mathcal{G}_i : i \in \mathcal{I}\}$ be sets of complex Hilbert spaces. Let $\mathcal{H} \coloneqq \bigotimes_{i\in\mathcal{I}} \mathcal{H}_i$ and $\mathcal{G} \coloneqq \bigotimes_{i\in\mathcal{I}} \mathcal{G}_i$. For $i \in \mathcal{I}$, let $A_i \in \mathcal{L}(\mathcal{H}_i, \mathcal{G}_i)$ and suppose that $\sup_{i\in\mathcal{I}} \|A_i\|_{\mathcal{L}(\mathcal{H}_i,\mathcal{G}_i)} < \infty$. Then*

$$\bigotimes_{i\in\mathcal{I}} \overline{\operatorname{Ran}(A_i)} = \overline{\operatorname{Ran}\left(\bigotimes_{i\in\mathcal{I}} A_i\right)}$$

Definition 2.4.14 (Definiteness of bounded operators)**.** Let $\mathcal{H}$ be a complex Hilbert space and let $A \in \mathcal{L}(\mathcal{H}, \mathcal{H})$. A is said to be

1. *non-negative definite* if $\langle h, Ah\rangle_{\mathcal{H}} \geq 0$ for any nonzero $h \in \mathcal{H}$
2. *positive definite* if $\langle h, Ah\rangle_{\mathcal{H}} > 0$ for any nonzero $h \in \mathcal{H}$
3. *non-positive definite* if $\langle h, Ah\rangle_{\mathcal{H}} \leq 0$ for any nonzero $h \in \mathcal{H}$
4. *negative definite* if $\langle h, Ah\rangle_{\mathcal{H}} < 0$ for any nonzero $h \in \mathcal{H}$.

Definition 2.4.15 (Hilbert-Schmidt operators, adapted from Gretton et al. (2005))**.** Let $\mathcal{I}$ be a non-empty set and let $\mathcal{H}_1$ and $\mathcal{H}_2$ be complex Hilbert spaces with dimension equal to the cardinality of $\mathcal{I}$. Let $\{u_i\}_{i\in\mathcal{I}}$ and $\{v_i\}_{i\in\mathcal{I}}$ be orthonormal bases for $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. A linear operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ is called a *Hilbert-Schmidt operator* if

$$\bigoplus_{i,j\in\mathcal{I}} \langle Tu_i, v_j\rangle^2_{\mathcal{H}_2} < \infty$$

Remark 2.4.14. Definition 2.4.15 turns out to be independent of the choice of orthonormal bases for $\mathcal{H}_1$ and $\mathcal{H}_2$. To see this, consider first that Parseval's identity implies

$$\bigoplus_{i,j\in\mathcal{I}} \langle Tu_i, v_j\rangle_{\mathcal{H}_2}^2 = \bigoplus_{i\in\mathcal{I}} \|Tu_i\|_{\mathcal{H}_2}^2 .$$

Hence the sum is independent of the choice of orthonormal basis for $\mathcal{H}_2$. Similarly, by the definition of the adjoint operator and Parseval's identity again,

$$\bigoplus_{i,j\in\mathcal{I}} \langle Tu_i, v_j\rangle_{\mathcal{H}_2}^2 = \bigoplus_{i,j\in\mathcal{I}} \langle u_i, T^* v_j\rangle_{\mathcal{H}_1}^2 = \bigoplus_{j\in\mathcal{I}} \|T^* v_j\|_{\mathcal{H}_1}^2$$

The sum is therefore also independent of the choice of orthonormal basis for $\mathcal{H}_1$. Let $HS(\mathcal{H}_1, \mathcal{H}_2)$ be the vector space of all Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. Define an inner product on $HS(\mathcal{H}_1, \mathcal{H}_2)$ by

$$\langle S, T\rangle_{HS(\mathcal{H}_1,\mathcal{H}_2)} \coloneqq \bigoplus_{i,j\in\mathcal{I}} \langle Su_i, v_j\rangle_{\mathcal{H}_2} \langle Tu_i, v_j\rangle_{\mathcal{H}_2}$$

This definition can, similarly to above, be shown to be independent of the choice of orthonormal bases for $\mathcal{H}_1$ and $\mathcal{H}_2$. According to Gretton et al. (2005), this inner product makes $HS(\mathcal{H}_1, \mathcal{H}_2)$ into a separable complex Hilbert space provided that $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable.

Definition 2.4.16 (Trace-class operators, see Conway (1990))**.** Let $\mathcal{I}$ be a non-empty set and let $\mathcal{H}$ be a complex Hilbert space with dimension equal to the cardinality of $\mathcal{I}$. Define the vector space of *trace-class operators* (also called *nuclear operators*) to be $\mathrm{Tr}(\mathcal{H}) \coloneqq \{AB : A, B \in HS(\mathcal{H}, \mathcal{H})\}$. The statement of Exercise 20 in Section 2 of Chapter 9 of Conway (1990) gives that $\bigoplus_{i\in\mathcal{I}} |\langle Av_i, v_i\rangle_{\mathcal{H}}| < \infty$ for any $A \in \mathrm{Tr}(\mathcal{H})$ and any orthonormal basis $\{v_i\}_{i\in\mathcal{I}}$ of $\mathcal{H}$, with the sum being independent of the choice of orthonormal basis. For $A \in \mathrm{Tr}(\mathcal{H})$, $\mathrm{Tr}(A) \coloneqq \bigoplus_{i\in\mathcal{I}} \langle Av_i, v_i\rangle_{\mathcal{H}}$ is called the *trace* of A. Define the *trace norm* to be the function given by $\|A\|_{\mathrm{Tr}(\mathcal{H})} \coloneqq \mathrm{Tr}\left([A^*A]^{1/2}\right)$. This norm makes $\mathrm{Tr}(\mathcal{H})$ into a Banach space.

Definition 2.4.17 (Moore-Penrose inverse)**.** Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be complex Hilbert spaces. Let $A \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ and let $A^{\perp}$ be the restriction of A to $\text{Ker}(A)^{\perp}$. The *Moore-Penrose (generalised) inverse of* A is the linear transformation $A^{\dagger} : \mathcal{H}_2 \to \text{Ker}(A)^{\perp}$ with domain $D\left(A^{\dagger}\right) = \text{Ran}(A) + \text{Ran}(A)^{\perp}$ which is specified by

$$A^{\dagger} f := \begin{cases} \left(A^{\perp}\right)^{-1} f & f \in \text{Ran}(A) \\ 0 & f \in \text{Ran}(A)^{\perp} \end{cases}$$

2.5 Banach/Hilbertian data analysis

The definitions given in this section are adapted from Hsing and Eubank (2015), except where it is stated that they are due to the author or otherwise referenced. Proofs of any unreferenced results are given at the end of the chapter. The purpose of this section, like the previous one, is to give the essential theory from Banach/Hilbertian data analysis that is used in Chapter 3 and Chapter 4. It could be argued that this section is classifiable as part of Functional Analysis – it is separate as it takes on a more probabilistic flavour. A particularly important part of this section is towards the end where the theory of kernels is discussed, in particular the definition of conditional kernel mean embeddings and the Hilbert-Schmidt Conditional Independence Criterion.

Definition 2.5.1 (Banach and Hilbertian random variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *Banach random variable* f is a measurable function $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (B, \mathcal{B}(B))$ where B is a complex Banach space. A *Hilbertian random variable* g is a measurable function $g : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ where $\mathcal{H}$ is a complex Hilbert space.

Definition 2.5.2 (Random operators)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Γ is called a *random operator* if $\Gamma : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}(B_1, B_2), \mathcal{B}(\mathcal{L}(B_1, B_2)))$ is a measurable function where B_1 and B_2 are complex Banach spaces.

Definition 2.5.3 (Simple Banach space valued functions)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let B be a complex Banach space. A B-valued function $f : (\Omega, \mathcal{F}, \mu) \to B$ is said to be *simple* if it satisfies:

$$\forall \omega \in \Omega,\ f(\omega) = \sum_{j=1}^{k} \mathbf{1}_{F_j}(\omega) g_j$$

for some $k \in \mathbb{N}$ where, for any $j \in \mathbb{N}_k$, $F_j \in \mathcal{F}$ and $g_j \in B$.

Definition 2.5.4 (Bochner integral for simple functions)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let B be a complex Banach space. A B-valued simple function $f : (\Omega, \mathcal{F}, \mu) \to B$ with representation $f(\omega) = \sum_{j=1}^{k} \mathbf{1}_{F_j}(\omega) g_j$ is said to be *Bochner integrable* if, for all $j \in \mathbb{N}_k$, $\mu(F_j) < \infty$. The Bochner integral of such a function is defined as:

$$\int_\Omega f \, \mathrm{d}\mu \coloneqq \sum_{j=1}^{k} \mu(F_j) g_j$$

Remark 2.5.1. Hsing and Eubank (2015) remark that Definition 2.5.4 does not depend on the particular representation of f.

Definition 2.5.5 (Bochner integral for measurable functions)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $(B, \mathcal{B}(B))$ be a measurable space where B is a complex Banach space. A B-valued measurable function $f : (\Omega, \mathcal{F}, \mu) \to (B, \mathcal{B}(B))$ is said to be *Bochner integrable* if there exists a sequence $\{f_i\}_{i \in \mathbb{N}}$ where: (1) for any $i \in \mathbb{N}$, $f_i : (\Omega, \mathcal{F}, \mu) \to (B, \mathcal{B}(B))$ is a simple Bochner integrable function, and (2) $\lim_{i \to \infty} \int_\Omega \| f_i - f \|_B \, \mathrm{d}\mu = 0$. The Bochner integral of such a function is defined as:

$$\int_\Omega f \, \mathrm{d}\mu \coloneqq \lim_{i \to \infty} \int_\Omega f_i \, \mathrm{d}\mu.$$

When the context is unambiguous, Ω is suppressed in the notation. Suppose that f is Bochner integrable and, for $A \in \mathcal{F}$, define

$$\int_A f \, \mathrm{d}\mu \coloneqq \int \mathbf{1}_A f \, \mathrm{d}\mu.$$

Remark 2.5.2. The notation $\int f(x)\ \mathrm{d}\mu(x)$ is also sometimes used for the Bocher integral, if it exists, of f.

Theorem 2.5.1 (Theorem 36 of Section 1 of Chapter 1 of Dinculeanu (2000))**.** *Let $(M, \mathcal{F}, \mu)$ be a measure space and let $(B_1, \mathcal{B}(B_1))$ and $(B_2, \mathcal{B}(B_2))$ be measurable spaces where B_1 and B_2 are complex Banach spaces. Let $T \in \mathcal{L}(B_1, B_2)$. If $f : (M, \mathcal{F}, \mu) \to (B_1, \mathcal{B}(B_1))$ is Bochner integrable then so is $T \circ f$ with*

$$T \int f \,\mathrm{d}\mu = \int T \circ f \,\mathrm{d}\mu.$$

Remark 2.5.3. Dinculeanu (2000) gives a number of properties of the Bochner integral which are similar to those for the Lebesgue integral, so they will not be listed here.

Remark 2.5.4. Let $(M, \mathcal{F}, \mu)$ be a measure space and let $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ be a measurable space where $\mathcal{H}$ is a complex Hilbert space. Let $f : (M, \mathcal{F}, \mu) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ be a Bochner integrable function. As a straightforward corollary of Theorem 2.5.1, it holds that, for any $g \in \mathcal{H}$,

$$\int \langle f, g \rangle_{\mathcal{H}}\ \mathrm{d}\mu = \left\langle \int f \,\mathrm{d}\mu, g \right\rangle_{\mathcal{H}}.$$

Definition 2.5.6 (Expectation of a Banach random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(B, \mathcal{B}(B))$ be a measurable space where B is a complex Banach space. Suppose that $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (B, \mathcal{B}(B))$ is Bochner integrable. Define the *expectation* $\mathbb{E}_{\mathbb{P}}(f)$ of f by

$$\mathbb{E}_{\mathbb{P}}(f) \coloneqq \int f \,\mathrm{d}\mathbb{P}.$$

When the context is unambiguous, $\mathbb{P}$ is suppressed in the notation for the expectation.

Remark 2.5.5. If $\mathcal{H}$ is a separable complex Hilbert space and f is a $\mathcal{H}$-valued random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Theorem 2.6.5 of Hsing and Eubank (2015) implies that $\mathbb{E}\left(\|f\|_{\mathcal{H}}\right) < \infty$ is sufficient for the existence of $\mathbb{E}\left(f\right)$.

Lemma 2.5.2 (Expectation and inner products)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ be a measurable space where $\mathcal{H}$ is a complex Hilbert space. Suppose that $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ and $g : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ are independent and are both Bochner integrable. Then*

$$\mathbb{E}\left(\langle f, g\rangle_{\mathcal{H}}\right) = \langle \mathbb{E}\left(f\right), \mathbb{E}\left(g\right)\rangle_{\mathcal{H}}$$

Definition 2.5.7 (Tensor product operator)**.** Let x_1, x_2 be elements of complex Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. The tensor product operator $(x_1 \otimes x_2) : \mathcal{H}_1 \to \mathcal{H}_2$ is defined by:

$$(x_1 \otimes x_2)\, y \coloneqq \langle y, x_1\rangle_{\mathcal{H}_1} x_2$$

for $y \in \mathcal{H}_1$. It is straightforward to see that $x_1 \otimes x_2 = (x_2 \otimes x_1)^*$.

Remark 2.5.6. A tensor product operator is Hilbert-Schmidt with Hilbert-Schmidt norm equal to the product of the norms of its parts (see (3) in Gretton et al. (2005)). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{H}_1$ and $\mathcal{H}_2$ be complex Hilbert spaces, $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$ and $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ be random variables. The random operator $A \otimes B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2), \mathcal{B}(\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)))$ is given by $(A \otimes B)(x) = A(x) \otimes B(x)$.

Definition 2.5.8 (Cross-covariance operators)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{H}_1$ and $\mathcal{H}_2$ be complex Hilbert spaces. Let $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$ and $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ be random variables. The *cross-covariance operator* $\Sigma_{BA} : \mathcal{H}_1 \to \mathcal{H}_2$ (also written $\mathrm{Cov}\left(A, B\right)$) is defined to be the expectation

$\mathbb{E}((A - \mathbb{E}(A)) \otimes (B - \mathbb{E}(B)))$. Also define the *covariance operator* $\Sigma_A \coloneqq \Sigma_{AA} \coloneqq \mathrm{Var}(A) \coloneqq \mathrm{Cov}(A, A)$.

Remark 2.5.7. Assuming that the cross-covariance operator Σ_{BA} of Hilbertian random variables A and B exists, it can equivalently be written as

$$\Sigma_{BA} = \mathbb{E}(A \otimes B) - (\mathbb{E}(A) \otimes \mathbb{E}(B)).$$

Remark 2.5.8. A sufficient condition, due to Theorem 2.6.5 of Hsing and Eubank (2015), for the existence of the cross-covariance operator between Hilbertian random variables A and B, when the spaces are separable, is that $\mathbb{E}\left(\|A\|_{\mathcal{H}_1} \|B\|_{\mathcal{H}_2}\right) < \infty$.

Remark 2.5.9. Cross-covariance operators are Hilbert-Schmidt (see Lemma 4 of Fukumizu et al. (2007)). Covariance operators are non-negative definite (by a straightforward application of Lemma 2.5.6) and self-adjoint.

Definition 2.5.9 (Conditional expectation, see Dinculeanu (2000))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let B be a complex Banach space. Let $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (B, \mathcal{B}(B))$ be a B-valued random variable such that $\mathbb{E}(A)$ exists. Let G be a sub-σ-field of $\mathcal{F}$. The conditional expectation $\mathbb{E}(A|G)$ is defined to be any G-measurable Bochner integrable function $g : (\Omega, \mathcal{F}, \mathbb{P}) \to (B, \mathcal{B}(B))$ such that, for any $H \in G$, $\mathbb{E}(\mathbf{1}_H g) = \mathbb{E}(\mathbf{1}_H A)$. If $M : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_M, \mathcal{F}_M)$ is a Ω_M-valued random variable with $(\Omega_M, \mathcal{F}_M)$ being a measurable space, then $\mathbb{E}(A|M) \coloneqq \mathbb{E}(A|\sigma(M))$. For a non-empty family $(\mathcal{F}_i)_{i \in \mathcal{I}}$ of sub-σ-fields of $\mathcal{F}$, let $\mathbb{E}(A|\mathcal{F}_i : i \in \mathcal{I}) \coloneqq \mathbb{E}(A|\bigvee_{i \in \mathcal{I}} \mathcal{F}_i)$ and, if $\mathcal{I} = \mathbb{N}_m$ for some $m \in \mathbb{N}$, let $\mathbb{E}(A|\mathcal{F}_1, \ldots, \mathcal{F}_m) \coloneqq \mathbb{E}(A|\bigvee_{i \in \mathcal{I}} \mathcal{F}_i)$.

Remark 2.5.10. The existence and almost sure uniqueness of the conditional expectation is established in Dinculeanu (2000). Taking $H = \Omega$ in the definition of conditional expectation gives the law of total expectation for Banach random variables.

Remark 2.5.11. Dinculeanu (2000) gives a number of properties of the conditional expectation which are alike those for the classical real-valued case, so they will not be presented here.

Definition 2.5.10 (Conditional probability, see Dinculeanu (2000))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For $G \trianglelefteq \mathcal{F}$ and $A \in \mathcal{F}$, let $\mathbb{P}(A|G) \coloneqq \mathbb{E}(\mathbf{1}_A|G)$ be the *conditional probability of A given G*.

Remark 2.5.12. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ and $G \trianglelefteq \mathcal{F}$. The solution to Problem 34.4(a) of Billingsley (1995) gives that $\mathbb{P}_A(B|G) \overset{\text{a.s.}\mathbb{P}_A}{=} \frac{\mathbb{P}(A \cap B|G)}{\mathbb{P}(A|G)}$ for any $B \in \mathcal{F}$.

Definition 2.5.11 (Conditional cross-covariance, see Dinculeanu (2000))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$ and $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ be Hilbertian random variables (the spaces are defined over $\mathbb{C}$), and let $G \trianglelefteq \mathcal{F}$. Suppose that $\Sigma_{BA} = \mathrm{Cov}(A, B)$ exists. Define the *conditional cross-covariance* $\Sigma_{BA|G} \coloneqq \mathrm{Cov}(A, B|G) \coloneqq \mathbb{E}((A - \mathbb{E}(A|G)) \otimes (B - \mathbb{E}(B|G))|G)$. If G is generated by a random variable M, let $\Sigma_{BA|M} \coloneqq \Sigma_{BA|G}$. Also, define the *conditional covariance* of A by $\Sigma_{A|G} \coloneqq \Sigma_{AA|G} \coloneqq \mathrm{Var}(A|G) \coloneqq \mathrm{Cov}(A, A|G)$.

Remark 2.5.13. In the literature on kernel mean embeddings of distributions (see Muandet et al. (2016), a different, non-equivalent and more restrictive, definition of conditional cross-covariance operators is typically used. Definition 2.5.11 is more alike the definition of conditional cross-covariance operators in the kernel framework given in Park and Muandet (2020), where they discuss the advantages of the measure-theoretic definition over the commonly used one.

Remark 2.5.14. In Classical Statistics, the law of total covariance follows from the law of total expectation. This still holds here with the proof being essentially the same, hence is omitted.

Lemma 2.5.3 (A property of conditional expectation)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\mathcal{H}_1$ *and* $\mathcal{H}_2$ *be complex Hilbert spaces. Let* $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$ *and* $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ *be random variables such that* $\mathbb{E}(A \otimes B)$ *exists. Let* $G \trianglelefteq \mathcal{F}$*. Then*

$$\mathbb{E}(A \otimes \mathbb{E}(B|G)|G) \overset{a.s.\mathbb{P}}{=} \mathbb{E}(A|G) \otimes \mathbb{E}(B|G).$$

Lemma 2.5.4 (A property of conditional expectation)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\mathcal{H}_1$ *and* $\mathcal{H}_2$ *be complex Hilbert spaces. Let* $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$ *and* $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ *be random variables such that* $\mathbb{E}(A \otimes B)$ *exists. Let* $G \trianglelefteq \mathcal{F}$*. Then*

$$\mathbb{E}(A \otimes \mathbb{E}(B|G)) = \mathbb{E}(\mathbb{E}(A|G) \otimes B) = \mathbb{E}(\mathbb{E}(A|G) \otimes \mathbb{E}(B|G)).$$

Corollary 2.5.5 (A property of conditional cross-covariance)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\mathcal{H}_1$ *and* $\mathcal{H}_2$ *be complex Hilbert spaces. Let* $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$ *and* $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ *be random variables such that* $\mathrm{Cov}(A, B)$ *exists. Let* $G \trianglelefteq \mathcal{F}$*. Then*

$$\mathrm{Cov}(A, \mathbb{E}(B|G)) \overset{a.s.\mathbb{P}}{=} \mathrm{Cov}(\mathbb{E}(A|G), \mathbb{E}(B|G)) \overset{a.s.\mathbb{P}}{=} \mathrm{Cov}(\mathbb{E}(A|G), B).$$

Lemma 2.5.6 (Conditional cross-covariance operators and inner products)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\mathcal{H}_1$ *and* $\mathcal{H}_2$ *be complex Hilbert spaces. Let* $A : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$ *and* $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ *be Hilbertian random variables such that* Σ_{AB} *exists. Let* $G \trianglelefteq \mathcal{F}$*. For any* $f \in \mathcal{H}_1$ *and* $g \in \mathcal{H}_2$

$$\mathrm{Cov}\left(\langle f, A\rangle_{\mathcal{H}_1}, \langle B, g\rangle_{\mathcal{H}_2} \middle| G\right) \overset{a.s.\mathbb{P}}{=} \left\langle f, \Sigma_{AB|G} g\right\rangle_{\mathcal{H}_1} \overset{a.s.\mathbb{P}}{=} \left\langle \Sigma_{BA|G} f, g\right\rangle_{\mathcal{H}_2}.$$

where $\langle f, A\rangle_{\mathcal{H}_1} : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{C}, \mathcal{B}(\mathbb{C}))$ *is given by* $x \mapsto \langle f, A(x)\rangle_{\mathcal{H}_1}$ *and* $\langle B, g\rangle_{\mathcal{H}_2} : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{C}, \mathcal{B}(\mathbb{C}))$ *is given by* $x \mapsto \langle B(x), g\rangle_{\mathcal{H}_2}$.

Definition 2.5.12 (Conditional independence of σ-fields, adapted from Li (2018))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{I}$ be a non-empty set. Let $(\mathcal{F}_i)_{i\in\mathcal{I}}$ be a family of sub-σ-fields of $\mathcal{F}$ and also let $\mathcal{F}^*$ be a sub-σ-field of $\mathcal{F}$. The family is said to be *conditionally independent* of $\mathcal{F}^*$ (with respect to $\mathbb{P}$, though this will not normally be explicitly mentioned) if for any finite subset $E \subseteq \mathcal{I}$ and $A_i \in \mathcal{F}_i$ $(i \in E)$

$$\mathbb{P}\left(\bigcap_{i\in E} A_i \,\middle|\, \mathcal{F}^*\right) \overset{\text{a.s.}\mathbb{P}}{=} \prod_{i\in E} \mathbb{P}\left(A_i|\mathcal{F}^*\right)$$

Denote the conditional independence by $\left(\perp\!\!\!\perp_{i\in\mathcal{I}} \mathcal{F}_i\right)|\mathcal{F}^*$. If $\mathcal{I} = \{1, 2\}$ and $(\mathcal{F}_1, \mathcal{F}_2)$ is conditionally independent of $\mathcal{F}^*$, write $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2|\mathcal{F}^*$.

Definition 2.5.13 (Conditional independence of stochastic processes, adapted from Li (2018))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{I}$ be a non-empty set, and, for $i \in \mathcal{I}$, let $(M_i, \mathcal{F}_i)$ be a measurable space. Let $\left(M^*, \mathcal{F}^*_M\right)$ be a measurable space. Let $(f_i)_{i\in\mathcal{I}}$ be a stochastic process where, for $i \in \mathcal{I}$, f_i is a M_i-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Let f^* be an M^*-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The stochastic process is said to be *conditionally independent* of f^* if the family $(\sigma(f_i))_{i\in\mathcal{I}}$ is conditionally independent of $\sigma(f^*)$. Denote the conditional independence by $\left(\perp\!\!\!\perp_{i\in\mathcal{I}} f_i\right)|f^*$. If $\mathcal{I} = \{1, 2\}$ and (f_1, f_2) is conditionally independent of f^*, write $f_1 \perp\!\!\!\perp f_2|f^*$.

Lemma 2.5.7 (Proposition 2.1 of Li (2018))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3 \trianglelefteq \mathcal{F}$. The following statements are equivalent*

1. $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2|\mathcal{F}_3$
2. $\forall A \in \mathcal{F}_1 \left[\mathbb{P}\left(A|\mathcal{F}_2, \mathcal{F}_3\right) \overset{a.s.\mathbb{P}}{=} \mathbb{P}\left(A|\mathcal{F}_3\right)\right]$.

Lemma 2.5.8 (Theorem 2.1. of Li (2018))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4 \trianglelefteq \mathcal{F}$. Then*

1. $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | \mathcal{F}_3 \implies \mathcal{F}_2 \perp\!\!\!\perp \mathcal{F}_1 | \mathcal{F}_3$
2. $\mathcal{F}_1 \perp\!\!\!\perp (\mathcal{F}_2, \mathcal{F}_3) | \mathcal{F}_4 \implies \mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | \mathcal{F}_4$
3. $\mathcal{F}_1 \perp\!\!\!\perp (\mathcal{F}_2, \mathcal{F}_3) | \mathcal{F}_4 \implies \mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | (\mathcal{F}_3, \mathcal{F}_4)$
4. $[\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | (\mathcal{F}_3, \mathcal{F}_4)] \wedge [\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_3 | \mathcal{F}_4] \implies \mathcal{F}_1 \perp\!\!\!\perp (\mathcal{F}_2, \mathcal{F}_3) | \mathcal{F}_4$.

Lemma 2.5.9 (A property of conditional independence)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. Let* $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4 \trianglelefteq \mathcal{F}$*. Then*

$$[\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | (\mathcal{F}_3, \mathcal{F}_4)] \wedge [\mathcal{F}_2 \perp\!\!\!\perp \mathcal{F}_3 | \mathcal{F}_4]$$

is equivalent to

$$[\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | \mathcal{F}_4] \wedge [\mathcal{F}_2 \perp\!\!\!\perp \mathcal{F}_3 | (\mathcal{F}_1, \mathcal{F}_4)] .$$

Lemma 2.5.10 (A property of conditional independence)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. Let* $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4, \mathcal{F}_5 \trianglelefteq \mathcal{F}$*. Suppose that* $\mathcal{F}_4 \trianglelefteq \mathcal{F}_5 \trianglelefteq \mathcal{F}_2$ *and* $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | (\mathcal{F}_3, \mathcal{F}_4)$*. Then* $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | (\mathcal{F}_3, \mathcal{F}_5)$.

Corollary 2.5.11 (A property of conditional independence)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. Let* $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4 \trianglelefteq \mathcal{F}$*. Suppose that* $\mathcal{F}_3 \trianglelefteq \mathcal{F}_4 \trianglelefteq \mathcal{F}_2$ *and* $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | \mathcal{F}_3$*. Then* $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | \mathcal{F}_4$.

Definition 2.5.14 (Conditional distribution, adapted from Klenke (2020))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(M, \mathcal{F}_M)$ be a measurable space. Let $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{F}_M)$ be a M-valued random variable. Let $G \trianglelefteq \mathcal{F}$. A *conditional distribution* is any function $\mathbb{P}_{f|G} : \mathcal{F}_M \times \Omega \to [0, 1]$ where, for $x \in \Omega$, $\mathbb{P}_{f|G}(\cdot, x) \coloneqq \mathbb{P}\left(f^{-1}(\cdot) \middle| G\right)(x)$ and, for $A \in \mathcal{F}_M$, $\mathbb{P}_{f|G}(A, \cdot) \coloneqq \mathbb{P}\left(f^{-1}(A) \middle| G\right)(\cdot)$. A conditional distribution $\mathbb{P}_{f|G}$ is said to be *regular* if, for each $x \in \Omega$, $\mathbb{P}_{f|G}(\cdot, x)$ is a probability measure on $\mathcal{F}_M$.

Definition 2.5.15 (Borel spaces, see Klenke (2020))**.** A measurable space $(M, \mathcal{F}_M)$ is called a *Borel space* if there exists $B \in \mathcal{B}(\mathbb{R})$ such that there exists an invertible measurable function $f : (M, \mathcal{F}_M) \to (B, \mathcal{B}(B))$ ($\mathcal{B}(B)$ is the Borel σ-field on B generated by the B-relative topology) with measurable inverse.

Remark 2.5.15. Theorem 8.36 of Klenke (2020) gives that a separable topological space for which there exists a complete metric inducing the topology is a Borel space when equipped with the Borel σ-field.

Theorem 2.5.12 (Theorem 8.37 of Klenke (2020))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $G \trianglelefteq \mathcal{F}$. Suppose that $(M, \mathcal{F}_M)$ is a Borel space and let $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{F}_M)$ be a M-valued random variable. Then there exists a regular conditional distribution of f given G.*

Theorem 2.5.13 (A restatement of Theorem 8.38 of Klenke (2020))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(M, \mathcal{F}_M)$ be a Borel space, and let $f : (\Omega, \mathcal{F}, \mathbb{P}) \to (M, \mathcal{F}_M)$ be a M-valued random variable. Let $G \trianglelefteq \mathcal{F}$. Let $g : (M, \mathcal{F}_M) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a measurable function such that $\mathbb{E}(g \circ f)$ exists. Let $\mathbb{P}_{f|G}$ be a regular conditional distribution of f given G. Then*

$$\mathbb{E}(g \circ f | G)(x) = \int g \, \mathrm{d}\mathbb{P}_{f|G}(\cdot, x)$$

for almost every $x \in \Omega$.

Definition 2.5.16 (Reproducing kernels, see Berlinet and Thomas-Agnan (2004))**.** Let E be some set and let $\mathcal{H}$ be a complex Hilbert space of $\mathbb{C}$-valued functions on E. A function $\kappa : E \times E \to \mathbb{C}$ is called a *reproducing kernel of $\mathcal{H}$* if

1. for any $x \in E$, it holds that $\kappa(\cdot, x) \in \mathcal{H}$
2. for any $x \in E$ and $f \in \mathcal{H}$, it holds that $\langle f, \kappa(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

Remark 2.5.16. In the field of Statistical Learning, $\mathcal{H}$ is generally a real Hilbert space of $\mathbb{R}$-valued functions on E. This complex version is given as there has been recent interest (see, e.g., Schuld and Killoran (2019) and Kübler et al. (2019)) in the connection between reproducing kernel Hilbert spaces and Quantum Mechanics which the author intends to explore deeper in further research.

Theorem 2.5.14 (Theorem 1 of Berlinet and Thomas-Agnan (2004))**.** *Let E be some set and let $\mathcal{H}$ be a complex Hilbert space of $\mathbb{C}$-valued functions on E. $\mathcal{H}$ has a reproducing kernel if and only if, for every $x \in E$, the evaluation functional $\phi_x : \mathcal{H} \to \mathbb{C}$ given by $\phi_x(f) = f(x)$ is continuous.*

Corollary 2.5.15 (Corollary 1 of Berlinet and Thomas-Agnan (2004))**.** *Let E be some set and let $\mathcal{H}$ be a complex Hilbert space of $\mathbb{C}$-valued functions on E which has a reproducing kernel κ. Then, every norm convergent sequence of functions in $\mathcal{H}$ is pointwise convergent.*

Definition 2.5.17 (Non-negative definite functions, see Berlinet and Thomas-Agnan (2004))**.** Let E be some set. A function $\kappa : E \times E \to \mathbb{C}$ is called a *non-negative definite function* (some authors say positive instead of non-negative) if

$$\forall n \in \mathbb{N} \, \forall \, (a_1, \ldots, a_n)^T \in \mathbb{C}^n \, \forall (x_1, \ldots, x_n) \in E^n \left[\sum_{i,j \in \mathbb{N}_n} a_i \overline{a_j} \kappa \left(x_i, x_j \right) \in [0, \infty) \right].$$

Lemma 2.5.16 (Lemma 1 of Berlinet and Thomas-Agnan (2004))**.** *Let E be some set and let $\mathcal{H}$ be a complex Hilbert space. Let $\psi : E \to \mathcal{H}$ be some function. The function $\kappa : E \times E \to \mathbb{C}$ given by $\kappa(x, y) \coloneqq \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$ is non-negative definite.*

Lemma 2.5.17 (Lemma 2 of Berlinet and Thomas-Agnan (2004))**.** *Any function which is a reproducing kernel for some complex Hilbert space of $\mathbb{C}$-valued functions on some set is necessarily non-negative definite.*

Theorem 2.5.18 (The Moore-Aronszajn theorem, see Theorem 3 of Berlinet and Thomas-Agnan (2004))**.** *Let E be some set and let $\kappa : E \times E \to \mathbb{C}$ be a non-negative definite function. Let $\mathcal{H}_0 \coloneqq \operatorname{Span}\{\kappa(\cdot, x) : x \in E\}$ and equip $\mathcal{H}_0$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0} : \mathcal{H}_0 \times \mathcal{H}_0 \to \mathbb{C}$ given by*

$$\left\langle \sum_{i \in \mathbb{N}_m} a_i \kappa(\cdot, x_i), \sum_{j \in \mathbb{N}_n} b_j \kappa(\cdot, y_j) \right\rangle_{\mathcal{H}_0} \coloneqq \sum_{i \in \mathbb{N}_m} \sum_{j \in \mathbb{N}_n} a_i \overline{b_j} \kappa(y_j, x_i)$$

The completion $\mathcal{H}$ of $\mathcal{H}_0$ is a reproducing kernel Hilbert space of $\mathbb{C}$-valued functions on E with reproducing kernel κ. Furthermore, no other Hilbert space of $\mathbb{C}$-valued functions on E has κ as a reproducing kernel.

Definition 2.5.18 (Kernel mean embedding, see Muandet et al. (2016))**.** Let $(M, \mathcal{F}_M)$ be a measurable space and let $\mathcal{M}$ be the set of all probability measures on $\mathcal{F}_M$. Let $\kappa : M \times M \to \mathbb{C}$ be a reproducing kernel which generates the reproducing kernel Hilbert space $\mathcal{H}$. Suppose that, for every $\mathbb{P} \in \mathcal{M}$, the function $\int_M \kappa(\cdot, x) \, \mathrm{d}\mathbb{P}(x)$ exists. Define the function $\mu : \mathcal{M} \to \mathcal{H}$ by $\mu(\mathbb{P}) \coloneqq \int_M \kappa(\cdot, x) \, \mathrm{d}\mathbb{P}(x)$ and call it the *kernel mean embedding of* $\mathbb{P}$.

Definition 2.5.19 (Maximum mean discrepancy. see Muandet et al. (2016))**.** Let $(M, \mathcal{F}_M)$, $\mathcal{M}$, κ, $\mathcal{H}$, μ be as in Definition 2.5.18. The pseudo-metric $d : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ given by $d(\mathbb{P}, \mathbb{Q}) \coloneqq \|\mu(\mathbb{P}) - \mu(\mathbb{Q})\|_{\mathcal{H}}$ is called the *maximum mean discrepancy*. It can equivalently be written as $d(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left[\int_M f(x) \, \mathrm{d}\mathbb{P}(x) - \int_M f(x) \, \mathrm{d}\mathbb{Q}(x) \right]$, hence the name.

Definition 2.5.20 (Characteristic kernel, see Muandet et al. (2016))**.** Let $(M, \mathcal{F}_M)$, $\mathcal{M}$, κ, $\mathcal{H}$, μ be as in Definition 2.5.18. If μ is injective, then κ is called a *characteristic kernel*. If this is the case, the maximum mean discrepancy becomes a metric on $\mathcal{M}$.

Theorem 2.5.19 (When is a kernel characteristic, see Proposition 5 of Fukumizu et al. (2009))**.** *Let $(M, \mathcal{F}_M)$ be a measurable space. A kernel $\kappa : M \times M \to \mathbb{C}$ is*

characteristic if and only if, for any probability measure $\mathbb{P}$ on $\mathcal{F}_M$, the vector space sum of the reproducing kernel Hilbert space $\mathcal{H}$ generated by κ and the $\mathbb{P}$ almost surely constant $\mathbb{C}$-valued functions on M is dense in $L^2(\mathbb{P}, \mathbb{C})$. Equivalently (see Li (2018)), a kernel is characteristic when, for any probability measure $\mathbb{P}$ on $\mathcal{F}_M$, $\mathcal{H}$ is dense modulo $\mathbb{P}$ almost sure constants in $L^2(\mathbb{P}, \mathbb{C})$ meaning that, for any $f \in L^2(\mathbb{P}, \mathbb{C})$, there exists a sequence $\{f_i\}_{i \in \mathbb{N}}$ of elements of $\mathcal{H}$ such that $\mathrm{Var}_{\mathbb{P}}(f - f_i) \to 0$ as $i \to \infty$.

Definition 2.5.21 (Kernel induced cross-covariance operators, see Muandet et al. (2016))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_X, \mathcal{F}_X)$ and $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_Y, \mathcal{F}_Y)$ be random variables. Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be reproducing kernel Hilbert spaces of $\mathbb{C}$-valued functions on Ω_X and Ω_Y respectively with measurable kernels κ_X and κ_Y. Suppose that $\mathbb{E}_{\mathbb{P}}(\kappa_X(X, X)) < \infty$ and $\mathbb{E}_{\mathbb{P}}(\kappa_Y(Y, Y)) < \infty$. Let $Z_1, Z_2 \in \{X, Y\}$. The following *kernel induced cross-covariance operator* exists

$$\Sigma_{Z_1 Z_2} \coloneqq \mathbb{E}_{\mathbb{P}}\left(\left[\kappa_{Z_1}(\cdot, Z_1) - \mathbb{E}_{\mathbb{P}}\left(\kappa_{Z_1}(\cdot, Z_1)\right)\right] \otimes \left[\kappa_{Z_2}(\cdot, Z_2) - \mathbb{E}_{\mathbb{P}}\left(\kappa_{Z_2}(\cdot, Z_2)\right)\right]\right)$$

Remark 2.5.17. $\Sigma_{Z_1 Z_2}$ is really the cross-covariance operator of $\kappa_{Z_2}(\cdot, Z_2)$ and $\kappa_{Z_1}(\cdot, Z_1)$. It is, however, common to call it the cross-covariance operator of Z_2 and Z_1. While related, this is not to be confused with the cross-covariance operators defined in Definition 2.5.8.

Definition 2.5.22 (Hilbert-Schmidt independence criterion, see Muandet et al. (2016))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$, X, Y, $\mathcal{H}_X$, $\mathcal{H}_Y$, κ_X, and κ_Y be as in Definition 2.5.21. The *Hilbert-Schmidt independence criterion between X and Y* is defined by

$$HSIC(X, Y) \coloneqq \|\Sigma_{XY}\|_{HS(\mathcal{H}_Y, \mathcal{H}_X)}$$

Theorem 2.5.20 (Unlabelled result given in Muandet et al. (2016))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$, X, Y, $\mathcal{H}_X$, $\mathcal{H}_Y$, κ_X, and κ_Y be as in Definition 2.5.21. Suppose that κ_X and*

κ_Y *are characteristic and, furthermore, the product kernel* $\kappa_{XY} : (\Omega_X \times \Omega_Y) \times (\Omega_X \times \Omega_Y) \to \mathbb{C}$ *given by* $\kappa_{XY}((x_1, y_1), (x_2, y_2)) \coloneqq \kappa_X(x_1, x_2)\,\kappa_Y(y_1, y_2)$ *is characteristic on* $(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y)$. *Then* $X \perp\!\!\!\perp Y$ *if and only if* $HSIC(X, Y) = 0$.

Remark 2.5.18. Theorem 2.5.20 has been used as the basis for an independence test of two random variables, see Gretton et al. (2005, 2007).

Definition 2.5.23 (Conditional mean embedding and maximum conditional mean discrepancy, adapted from Park and Muandet (2020))**.** Let $(\Omega, \mathcal{F})$ be a measurable space. Let $X : (\Omega, \mathcal{F}) \to (\Omega_X, \mathcal{F}_X)$ be a measurable function. Let $G \trianglelefteq \mathcal{F}$. Let $\mathcal{M}$ be the set of all probability measures on $\mathcal{F}$. Let $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{C}$ be a measurable kernel for which, for each $\mathbb{P} \in \mathcal{M}$, $\mathbb{E}_{\mathbb{P}}(\kappa_X(\cdot, X))$ exists and let $\mathcal{H}_X$ be its associated reproducing kernel Hilbert space. Let $\mathbb{P} \in \mathcal{M}$ and define the *kernel conditional mean embedding of* $\mathbb{P}$ *given* G to be the random variable $\mu_{\mathbb{P}|G} \coloneqq \mathbb{E}_{\mathbb{P}}(\kappa_X(\cdot, X)|G)$. For $\mathbb{P}, \mathbb{Q} \in \mathcal{M}$, define the *maximum conditional mean discrepancy of* $\mathbb{P}$ *and* $\mathbb{Q}$ *given* G to be the random variable $d(\mathbb{P}, \mathbb{Q}|G) \coloneqq \left\|\mu_{\mathbb{P}|G} - \mu_{\mathbb{Q}|G}\right\|_{\mathcal{H}_X}$.

Definition 2.5.24 (Hilbert-Schmidt conditional independence criterion, adapted from Park and Muandet (2020))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$, X, Y, $\mathcal{H}_X$, $\mathcal{H}_Y$, κ_X, and κ_Y be as in Definition 2.5.21. Let $G \trianglelefteq \mathcal{F}$. The *Hilbert-Schmidt conditional independence criterion between* X *and* Y *given* G is the random variable

$$HSCIC(X, Y|G) \coloneqq \left\|\Sigma_{XY|G}\right\|_{HS(\mathcal{H}_Y, \mathcal{H}_X)}$$

Theorem 2.5.21 (Conditional independence and the Hilbert-Schmidt conditional independence criterion)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$, X, Y, $\mathcal{H}_X$, $\mathcal{H}_Y$, κ_X, *and* κ_Y *be as in Definition 2.5.21. Let* $G \trianglelefteq \mathcal{F}$. *Suppose that* κ_X *and* κ_Y *are characteristic and that* κ_{XY}, *the product kernel, is characteristic on* $(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y)$. *Suppose that* $\mathbb{P}(\cdot|G)$ *admits a regular version. Then* $Y \perp\!\!\!\perp X|G$ *if and only*

if $HSCIC\left(X,Y|G\right) \overset{a.s.\mathbb{P}}{=} 0$. *Note that this is also equivalent to saying that* $\mathbb{E}_{\mathbb{P}}\left(HSCIC\left(X,Y|G\right)\right) = 0$ *or* $\Sigma_{XY|G} \overset{a.s.\mathbb{P}}{=} 0$ *or* $\mathbb{E}\left(\Sigma_{XY|G}\right) = 0$.

Remark 2.5.19. In future work, the author intends to use Theorem 2.5.21 as a basis for developing a test of conditional independence of two random variables given a σ-field generated by some further random variable. Furthermore, Theorem 2.5.21 can be used as the basis for a nonlinear Sufficient Dimension Reduction method. It can also be used to devise a test for the number of components to extract from a nonlinear Sufficient Dimension Reduction procedure in order to characterise conditional independence.

2.6 Proofs of supporting results

Proof of Lemma 2.2.2. Consider

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i\in E}\left[f^*\right]^{-1}\left(A_i\right)\right) &= \mathbb{P}\left(\bigcap_{i\in E}\left\{x\in\Omega : f^*(x)\in A_i\right\}\right)\\
&= \mathbb{P}\left(\bigcap_{i\in E}\left\{x\in\Omega : h_x\in A_i\right\}\right)\\
&= \mathbb{P}\left(\bigcap_{i\in E}\left(\bigcap_{j\in\mathcal{I}}\left\{x\in\Omega : f_i(x)\in\left\{y(j) : y\in A_i\right\}\right\}\right)\right)\\
&= \mathbb{P}\left(\bigcap_{i\in E}\left(\bigcap_{j\in\mathcal{I}} f_i^{-1}\left(\left\{y(i) : y\in A_i\right\}\right)\right)\right)\\
&= \mathbb{P}\left(\bigcap_{i\in E} f_i^{-1}\left(\bigcap_{j\in\mathcal{I}}\left\{y(j) : y\in A_i\right\}\right)\right)\\
&= \prod_{i\in E}\mathbb{P}\left(f_i^{-1}\left(\bigcap_{j\in\mathcal{I}}\left\{y(j) : y\in A_i\right\}\right)\right)\\
&= \prod_{i\in E}\mathbb{P}\left(\bigcap_{j\in\mathcal{I}} f_i^{-1}\left(\left\{y(j) : y\in A_i\right\}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \prod_{i \in E} \mathbb{P}\left(\bigcap_{j \in \mathcal{I}} \{x \in \Omega : f_i(x) \in \{y(j) : y \in A_i\}\} \right) \\
&= \prod_{i \in E} \mathbb{P}\left(\{x \in \Omega : f^*(x) \in A_i\} \right) \\
&= \prod_{i \in E} \mathbb{P}\left([f^*]^{-1}(A_i) \right).
\end{aligned}
$$

□

Proof of Lemma 2.2.3. Let $f^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(\bigtimes_{i \in \mathcal{I}} M_i, \bigotimes_{i \in \mathcal{I}} \mathcal{F}_i \right)$ be given by $f^*(x) = h_x$ where $h_x : \mathcal{I} \to \bigcup_{i \in \mathcal{I}} M_i$ is itself given by $h_x(i) = f_i(x)$. Consider the stochastic process $(\phi_i)_{i \in \mathcal{I}}$ where, for $i \in \mathcal{I}$, $\phi_i : \left(\bigtimes_{i \in \mathcal{I}} M_i, \bigotimes_{i \in \mathcal{I}} \mathcal{F}_i \right) \to (M_i, \mathcal{F}_i)$ is the projection onto M_i. As the product measure is unique, it needs to be shown that: (1) $(\phi_i)_{i \in \mathcal{I}}$ is independent (with respect to $\mathbb{P}_f = \mathbb{P}_{f^*}$), and (2) the distribution of ϕ_i equals $\mathbb{P}_{f_i}$ for any $i \in \mathcal{I}$. For the first, let E be a finite subset of $\mathcal{I}$ and, for $i \in E$, let $A_i \in \sigma(\phi_i)$. Consider

$$
\begin{aligned}
\mathbb{P}_f\left(\bigcap_{i \in E} A_i \right) &= \mathbb{P}_{f^*}\left(\bigcap_{i \in E} A_i \right) \\
&= \mathbb{P}\left([f^*]^{-1}\left(\bigcap_{i \in E} A_i \right) \right) \\
&= \mathbb{P}\left(\bigcap_{i \in E} [f^*]^{-1}(A_i) \right) \\
&= \prod_{i \in E} \mathbb{P}\left([f^*]^{-1}(A_i) \right) \\
&= \prod_{i \in E} \mathbb{P}_{f^*}(A_i) \\
&= \prod_{i \in E} \mathbb{P}_f(A_i).
\end{aligned}
$$

This gives that $(\phi_i)_{i \in \mathcal{I}}$ is independent. For the second claim, let $i \in \mathcal{I}$ and let

$A \in \mathcal{F}_M$. Consider

$$\begin{aligned}
\mathbb{P}_{f_i}(A) &= \mathbb{P}\left(f_i^{-1}(A)\right) \\
&= \mathbb{P}\left(\left[\phi_i \circ f^*\right]^{-1}(A)\right) \\
&= \mathbb{P}\left(\left[f^*\right]^{-1}\left(\phi_i^{-1}(A)\right)\right) \\
&= \mathbb{P}_{f^*}\left(\phi_i^{-1}(A)\right) \\
&= \mathbb{P}_{\phi_i}(A).
\end{aligned}$$

This gives the claim, and so concludes the proof. □

Proof of Lemma 2.2.4. Let $\mathbb{P}_f^*$ be the common distribution of the f_i's where $i \in \mathcal{I}$. Let $T \coloneqq \mathbb{N}_{n_T}$ for some $n_T \in \mathbb{N}$ (if $\mathcal{I}$ is infinite) or $n_T \in \mathbb{N}_{k+1}$ (if $\mathcal{I}$ has k elements for some $k \in \mathbb{N}$), let m be a natural number less than n_T, and let $k_1, \ldots, k_m$ be distinct elements of T. Let $f_m^\dagger$, f_m^*, g_x, and h_x be as in Definition 2.2.16. Let $A \in \mathcal{F}_M^m$ and, for $i \in \mathbb{N}_m$, let $A_i \coloneqq \{x(i) : x \in A\}$. Consider

$$\begin{aligned}
\mathbb{P}_{f_m^\dagger}(A) &= \mathbb{P}\left(\left[f_m^\dagger\right]^{-1}(A)\right) \\
&= \mathbb{P}\left(\left\{x \in \Omega : f_m^\dagger(x) \in A\right\}\right) \\
&= \mathbb{P}\left(\{x \in \Omega : g_x \in A\}\right) \\
&= \mathbb{P}\left(\bigcap_{i \in \mathbb{N}_m} \{x \in \Omega : f_i(x) \in A_i\}\right) \\
&= \prod_{i \in \mathbb{N}_m} \mathbb{P}\left(\{x \in \Omega : f_i(x) \in A_i\}\right) \\
&= \prod_{i \in \mathbb{N}_m} \mathbb{P}_{f_i}(A_i) \\
&= \prod_{i \in \mathbb{N}_m} \mathbb{P}_f^*(A_i).
\end{aligned}$$

Also

$$\begin{aligned}
\mathbb{P}_{f_m^*}(A) &= \mathbb{P}\left(\left[f_m^*\right]^{-1}(A)\right) \\
&= \mathbb{P}\left(\left\{x \in \Omega : f_m^*(x) \in A\right\}\right) \\
&= \mathbb{P}\left(\{x \in \Omega : h_x \in A\}\right) \\
&= \mathbb{P}\left(\bigcap_{i \in \mathbb{N}_m} \left\{x \in \Omega : f_{k_i}(x) \in A_i\right\}\right) \\
&= \prod_{i \in \mathbb{N}_m} \mathbb{P}\left(\left\{x \in \Omega : f_{k_i}(x) \in A_i\right\}\right) \\
&= \prod_{i \in \mathbb{N}_m} \mathbb{P}_{f_{k_i}}(A_i) \\
&= \prod_{i \in \mathbb{N}_m} \mathbb{P}_f^*(A_i).
\end{aligned}$$

Putting these together concludes the proof. □

Proof of Lemma 2.2.5. Let $A \in \mathcal{F}_M^{\mathcal{I}}$ and, for $i \in \mathcal{I}$, let $A_i \coloneqq \{x(i) : x \in A\}$. Let $\phi^* : \left(M^{\mathcal{I}}, \mathcal{F}_M^{\mathcal{I}}, \mathbb{P}_{f^*}\right) \to \left(M^{\mathcal{I}}, \mathcal{F}_M^{\mathcal{I}}\right)$ be the $M^{\mathcal{I}}$-valued random variable given by $\phi^*(x) = g_x$ where $g_x : \mathcal{I} \to M$ is itself given by $g_x(i) = \phi_i(x)$. Consider

$$\begin{aligned}
\mathbb{P}_\phi(A) &= \mathbb{P}_{\phi^*}(A) \\
&= \mathbb{P}_{f^*}\left(\phi^{-1}(A)\right) \\
&= \mathbb{P}\left(\left[f^*\right]^{-1}\left(\phi^{-1}(A)\right)\right) \\
&= \mathbb{P}\left(\left[\phi \circ f^*\right]^{-1}(A)\right) \\
&= \mathbb{P}\left(\left\{x \in \Omega : \left[\phi \circ f^*\right](x) \in A\right\}\right) \\
&= \mathbb{P}\left(\left\{x \in \Omega : k_{f_x^*} \in A\right\}\right) \\
&= \mathbb{P}\left(\bigcap_{i \in \mathcal{I}} \left\{x \in \Omega : \phi_i\left(f^*(x)\right) \in A_i\right\}\right) \\
&= \mathbb{P}\left(\bigcap_{i \in \mathcal{I}} \left\{x \in \Omega : \left[\phi_i \circ f^*\right](x) \in A_i\right\}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(\bigcap_{i\in\mathcal{I}} \{x \in \Omega : f_i(x) \in A_i\}\right) \\
&= \mathbb{P}\left(\{x \in \Omega : f^*(x) \in A\}\right) \\
&= \mathbb{P}\left(\left[f^*\right]^{-1}(A)\right) \\
&= \mathbb{P}_{f^*}(A) \\
&= \mathbb{P}_f(A).
\end{aligned}
$$

In a similar manner, it is shown that, for any $B \in \mathcal{F}_m$, $\mathbb{P}_{\phi_i}(B) = \mathbb{P}_{f_i}(B)$. The given equivalences then follow by pairing $(\Omega, \mathcal{F}, \mathbb{P})$ with $\left(M^{\mathcal{I}}, \mathcal{F}_M^{\mathcal{I}}, \mathbb{P}_f\right)$, pairing $\mathbb{P}_\phi$ with $\mathbb{P}_f$, and pairing (for $i \in \mathcal{I}$) $\mathbb{P}_{\phi_i}$ with $\mathbb{P}_{f_i}$. □

Proof of Theorem 2.2.6. For the first claim, begin by supposing that V is exchangeable. Let $S \coloneqq \mathbb{N}_{n_S}$ for some $n_S \in \mathbb{N}$ (if $\mathcal{I}$ is infinite) or $n_S \in \mathbb{N}_{k+1}$ (if $\mathcal{I}$ has k elements for some $k \in \mathbb{N}$), let m be a natural number less than n_S, and let $k_1, \ldots, k_m$ be distinct elements of S. Let $W_m : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(N^m, \mathcal{F}_N^m\right)$ be given by $W_m(x) = a_x$ where $a_x : \mathbb{N}_m \to N$ is itself given by $a_x(i) = w_i(x)$, and let $W_m^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(N^m, \mathcal{F}_N^m\right)$ be given by $W_m^*(x) = b_x$ where $b_x : \mathbb{N}_m \to N$ is itself given by $b_x(i) = w_{k_i}(x)$. The aim is to show that $\mathbb{P}_{W_m} = \mathbb{P}_{W_m^*}$. To this end, let $A \in \mathcal{F}_N^m$ and, for $i \in \mathbb{N}_m$, let $A_i \coloneqq \{x(i) : x \in A\}$. Consider

$$
\begin{aligned}
\mathbb{P}_{W_m}(A) &= \mathbb{P}\left(\{x \in \Omega : W_m(x) \in A\}\right) \\
&= \mathbb{P}\left(\bigcap_{i\in\mathbb{N}_m} \{x \in \Omega : w_i(x) \in A_i\}\right) \\
&= \mathbb{P}\left(\bigcap_{i\in\mathbb{N}_m} \{x \in \Omega : [T \circ z_i](x) \in A_i\}\right) \\
&= \mathbb{P}\left(\bigcap_{i\in\mathbb{N}_m} \{x \in \Omega : T(z_i(x)) \in A_i\}\right) \\
&= \mathbb{P}\left(\bigcap_{i\in\mathbb{N}_m} \left\{x \in \Omega : z_i(x) \in T^{-1}(A_i)\right\}\right). \qquad (2.2)
\end{aligned}
$$

Now, for $i \in \mathbb{N}_m$, let

$$B_i^* \coloneqq \left\{x \in M : \exists y \in T^{-1}(A_i)\,[x = y(1)]\right\}$$
$$B_i^{**} \coloneqq \left\{x \in M : \exists y \in T^{-1}(A_i)\,[x = y(2)]\right\}$$

With this, Equation (2.2) can be rewritten as

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i\in\mathbb{N}_m}\left\{x \in \Omega : z_i(x) \in T^{-1}(A_i)\right\}\right) &= \mathbb{P}\left(\bigcap_{i\in\mathbb{N}_m}\left[v_i^{-1}(B_i^*) \cap f^{-1}(B_i^{**})\right]\right) \\ &= \mathbb{P}\left(\left[\bigcap_{i\in\mathbb{N}_m} v_i^{-1}(B_i^*)\right] \cap f^{-1}\left(\bigcap_{i\in\mathbb{N}_m} B_i^{**}\right)\right) \qquad (2.3)\end{aligned}$$

Now let $V_m : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^m, \mathcal{F}_M^m\right)$ be given by $V_m(x) = g_x$ where $g_x : \mathbb{N}_m \to M$ is itself given by $g_x(i) = v_i(x)$, let $V_m^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^m, \mathcal{F}_M^m\right)$ be given by $V_m^* = h_x$ where $h_x : \mathbb{N}_m \to M$ is itself given by $h_x(i) = v_{k_i}(x)$, and let $B \coloneqq \left\{x \in M^m : \forall i \in \mathbb{N}_m \exists y_i \in B_i^*\,[x(i) = y]\right\}$. Equation (2.3) can now be rewritten as

$$\begin{aligned}\mathbb{P}\left(\left[\bigcap_{i\in\mathbb{N}_m} v_i^{-1}(B_i^*)\right] \cap f^{-1}\left(\bigcap_{i\in\mathbb{N}_m} B_i^{**}\right)\right) &= \mathbb{P}\left(V_m^{-1}(B) \cap f^{-1}\left(\bigcap_{i\in\mathbb{N}_m} B_i^{**}\right)\right) \\ &= \mathbb{P}\left(V_m^{-1}(B)\right)\mathbb{P}\left(f^{-1}\left(\bigcap_{i\in\mathbb{N}_m} B_i^{**}\right)\right) \\ &= \mathbb{P}\left(\left[V_m^*\right]^{-1}(B)\right)\mathbb{P}\left(f^{-1}\left(\bigcap_{i\in\mathbb{N}_m} B_i^{**}\right)\right)\end{aligned}$$

Repeating the above argument with i replaced by k_i where needed yields the result. For the second claim, replace the requirement that $k_1, \ldots, k_m$ be distinct elements of S with the requirement that $1 \leq k_1 < \ldots < k_m \leq n_S$ and repeat the above argument. □

Proof of Lemma 2.2.7. Only the left version is proven as the case for the right is similar. Let $x \in S$ and let $y \in \mathbb{R}$ with $y < f(x)$. The goal is to show that there exists a neighbourhood U of x such that $f(z) > y$ for $z \in U$. For

$a, b \in S$ with $a \leq_S b$, let $(a, b) \coloneqq \{\alpha \in S : a <_S \alpha <_S b\}$. Taking $U \coloneqq (\inf \{\alpha \in S : f(\alpha) < f(x)\}, \sup \{\alpha \in S : y < f(\alpha)\})$ suffices for the proof. □

Proof of Theorem 2.2.8. Only Property 2 is shown as the others are established in the proof of Proposition 8 in Gálvez-Rodríguez and Sánchez-Granero (2019). By definition, $F_S : S \to [0, 1]$ is right continuous if for each $x \in S$ for which $\{y \in S : y \leq_S x\} \neq S$ and for each $\epsilon > 0$, there exists an open initial A such that $x \in A$ and $\forall u \in A\, [F_S(u) < F_S(x) + \epsilon]$. So let $x \in S$ with $\{y \in S : y \leq_S x\} \neq S$ and let $\epsilon > 0$. Choosing $z \coloneqq \inf \{\alpha \in S : F_S(\alpha) \geq F_S(x) + \epsilon\}$ and letting $A \coloneqq \{y \in S : y <_S z\}$ suffices for the proof. □

Proof of Theorem 2.2.10. Define the *extension of* S to be the set $S^* \coloneqq S \cup \{-\infty, \infty\}$ where $-\infty$ and ∞ are formal symbols conventionally satisfying $\forall \alpha \in S\, [(-\infty <_S \alpha) \wedge (\alpha <_S \infty)]$. If S has no lower bounds, let $\inf S \coloneqq -\infty$. Similarly, if S has no upper bounds, let $\sup S \coloneqq \infty$. Define the function $\lambda : [0, 1] \to S^*$ by

$$\lambda(p) = \begin{cases} \inf S & p = 0 \\ \inf \left\{\alpha \in S : F_f(\alpha) \geq p\right\} & p \in (0, 1) \\ \sup S & p = 1 \end{cases}$$

Let F_g be the cumulative distribution function of g. Let $y \in [0, 1]$ and consider

$$\begin{aligned} F_g(y) &= \mathbb{P}(\{x \in \Omega : g(x) \leq y\}) \\ &= \mathbb{P}\left(\left\{x \in \Omega : F_f(f(x)) \leq y\right\}\right) \\ &= \mathbb{P}(\{x \in \Omega : f(x) \leq \lambda(y)\}) \\ &= F_f(\lambda(y)) = y. \end{aligned}$$

Thus F_g has the form of a cumulative distribution function for a standard uniform random variable. □

Proof of Theorem 2.2.11. Let $Q : [0,1] \to \mathcal{P}((0,\infty))$ be the quantile function of $\frac{(T\circ z_i)^2}{(T\circ z_j)^2}$. Recall that a median of $\frac{(T\circ z_i)^2}{(T\circ z_j)^2}$ is any member of $Q(0.5)$. Let $F : (0,\infty) \to [0,1]$ be the cumulative distribution function of $\frac{(T\circ z_i)^2}{(T\circ z_j)^2}$. To show that 1 is a median of $\frac{(T\circ z_i)^2}{(T\circ z_j)^2}$, it needs to be shown that

$$1 \in \left[\sup\left\{x \in (0,\infty) : F(x) < \frac{1}{2}\right\}, \sup\left\{x \in (0,\infty) : F(x) \leq \frac{1}{2}\right\}\right].$$

More explicitly, it needs to be shown that

$$1 \in \left[\begin{array}{l} \sup\left\{x \in (0,\infty) : \mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq x\right) < \frac{1}{2}\right\}, \\ \sup\left\{x \in (0,\infty) : \mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq x\right) \leq \frac{1}{2}\right\}\end{array}\right].$$

As (v_i, v_j) is exchangeable and $V^* \perp\!\!\!\perp f$, $\left([T\circ z_i]^2, [T\circ z_j]^2\right)$ is exchangeable by Theorem 2.2.6. It is now shown that 1 is a median for the target random variable. Consider

$$\begin{aligned}\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq 1\right) &= \mathbb{P}\left(\frac{(T\circ z_j)^2}{(T\circ z_i)^2} \leq 1\right) \\ &= 1 - \mathbb{P}\left(\frac{(T\circ z_j)^2}{(T\circ z_i)^2} > 1\right) \\ &= 1 - \mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} < 1\right).\end{aligned}$$

So

$$\begin{aligned}&\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} < 1\right) \leq 1 - \mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} < 1\right), \\ \text{and}\quad &\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq 1\right) \geq 1 - \mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq 1\right).\end{aligned}$$

Hence

$$\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} < 1\right) \leq \frac{1}{2} \leq \mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq 1\right).$$

Hence, 1 is a median of $\frac{(T\circ z_i)^2}{(T\circ z_j)^2}$ as claimed. It is now shown that 1 is a unique median. In other words, for any $c_1 \in (0,1)$ and $c_2 \in (1,\infty)$, it needs to be shown that

$$\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq c_1\right) < \frac{1}{2}$$
$$\text{and} \quad \mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} < c_2\right) > \frac{1}{2}.$$

Only the first inequality will be shown as the second is similarly proven. Let $c_1 \in (0,1)$ and observe that

$$\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq c_1\right) = \mathbb{E}\left(\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq c_1 \middle| v_i, v_j\right)\right)$$

It thus suffices to show that, for any linearly independent $\alpha_1, \alpha_2 \in M$,

$$\mathbb{P}\left(\frac{(T\circ z_i)^2}{(T\circ z_j)^2} \leq c_1 \middle| (v_i, v_j) = (\alpha_1, \alpha_2)\right) < \frac{1}{2}. \tag{2.4}$$

Let $\alpha_1, \alpha_2 \in M$ be linearly independent. Let $k \in \{1,2\}$. Let $\delta_k : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(M^2, [\mathcal{F}_M]^2, \mathcal{T}_M^2\right)$ be given by $\delta_k(x) = m_x^k$ where $m_x^k : \{1,2\} \to M$ is itself given by $m_x^k(1) = \alpha_k$ and $m_x^k(2) = f(x)$. As $V^* \perp\!\!\!\perp f$, Equation (2.4) is equivalent to

$$\mathbb{P}\left(\frac{[T\circ\delta_1]^2}{[T\circ\delta_2]^2} \leq c_1\right) < \frac{1}{2}.$$

Let $c_3 \in (c_1, 1)$. Consider $L : M \to \mathbb{R}^2$ given by $L(u) \coloneqq (T(\alpha_1,u), T(\alpha_2,u))^T$. L is linear as T is linear in the second entry. Also, L has rank 2 as α_1 and α_2 are linearly independent. Thus, the system

$$T(\alpha_1, u) = \sqrt{c_3}$$
$$T(\alpha_2, u) = 1$$

has a unique solution, call it u^*. As the mapping $u \to \frac{[T(\alpha_1,u)]^2}{[T(\alpha_2,u)]^2}$ is continuous, there exists an open neighbourhood G of u^* such that $u \in G \implies \frac{[T(\alpha_1,u)]^2}{[T(\alpha_2,u)]^2} \in (c_1, 1)$.

By assumption, $\mathbb{P}(f \in G) > 0$. Thus $\mathbb{P}\left(\frac{[T(\alpha_1,f)]^2}{[T(\alpha_2,f)]^2} \in (c_1, 1)\right) > 0$. Combining this with the fact that 1 is a median of $\frac{[T(v_i,f)]^2}{[T(v_j,f)]^2}$ gives the result. □

Proof of Lemma 2.2.12. The first part of the result follows from Theorem 4 of Li (2007b). Now the exchangeability of $\left(\langle v_1, f\rangle_{\mathcal{H}}, \ldots \langle v_p, f\rangle_{\mathcal{H}}\right)$ is equivalent to saying that, for any permutation $\sigma \in \mathcal{P}_p$, $\left(\langle v_1, f\rangle_{\mathcal{H}}, \ldots, \langle v_p, f\rangle_{\mathcal{H}}\right)^T \overset{D}{=} \left(\langle v_{\sigma(1)}, f\rangle_{\mathcal{H}}, \ldots, \langle v_{\sigma(p)}, f\rangle_{\mathcal{H}}\right)^T$. By Remark 2.2.19, the distribution of the left hand side is invariant under orthogonal transformations. Hence, it is also invariant after being transformed by a permutation matrix thus giving the claim. □

Proof of Lemma 2.2.13. Since T is a unitary operator on $\mathbb{R}^p$, T and U are invertible with the inverse of U being given by

$$U^{-1}(h) = \sum_{j \in \mathbb{N}_p} T_j^{-1}(C)(v_j).$$

Let g be a member of $\mathcal{H}$ with coordinate representation $A = (\alpha_1, \ldots, \alpha_p)^T$ with respect to $\{v_1, \ldots, v_p\}$. Then

$$\begin{aligned}
\langle g, U(h)\rangle_{\mathcal{H}} &= \sum_{j \in \mathbb{N}_p} \alpha_j T_j(C) \\
&= A^T T(C) \\
&= (T^{-1}(A))^T C \\
&= \sum_{j \in \mathbb{N}_p} T_j^{-1}(A) C_j \\
&= \left\langle U^{-1} g, h\right\rangle_{\mathcal{H}}.
\end{aligned}$$

Thus $U^* = U^{-1}$, as desired. □

Proof of Lemma 2.2.14. Suppose that U is a unitary operator on $\mathcal{H}$ and let $g \in \mathcal{H}$

have coordinate representation $A = (\alpha_1, \ldots, \alpha_p)^T$. Then

$$\begin{aligned} C^T T(A) &= \sum_{j \in \mathbb{N}_p} c_j T_j(A) \\ &= \langle h, U(g) \rangle_{\mathcal{H}} \\ &= \left\langle U^{-1}(h), g \right\rangle_{\mathcal{H}} \\ &= \sum_{j \in \mathbb{N}_p} T_j^{-1}(C) \alpha_j \\ &= (T^{-1}(C))^T A. \end{aligned}$$

Thus $T^* = T^{-1}$, as desired. □

Proof of Theorem 2.2.15. "only if": Suppose that $U : \mathcal{H} \to \mathcal{H}$ is a unitary operator and f is a unitarily invariant $\mathcal{H}$-valued random variable with coordinate $A = (\alpha_1, \ldots, \alpha_k)^T$ with respect to an orthonormal basis $\{v_1, \ldots, v_p\}$ of $\mathcal{H}$. For each $i \in \mathbb{N}_p$ and unitary $U : \mathcal{H} \to \mathcal{H}$:

$$\alpha_i = \langle f, v_i \rangle_{\mathcal{H}} \overset{D}{=} \langle U(f), v_i \rangle_{\mathcal{H}} \tag{2.5}$$

Let $T : \mathbb{R}^p \to \mathbb{R}^p$ be a unitary operator on $\mathbb{R}^p$ and now let U be as in Lemma 2.2.13. Then, for each $i \in \mathbb{N}_p$

$$\langle U(f), v_i \rangle_{\mathcal{H}} = T_i(A). \tag{2.6}$$

From Equation (2.5) and Equation (2.6), $A \overset{D}{=} T(A)$. So A is a unitarily invariant $\mathbb{R}^n$-valued random variable. Identifying T with an orthogonal matrix then gives that A is a spherically distributed $\mathbb{R}^p$-valued random variable. "if": Suppose $A = (\alpha_1, \ldots, \alpha_p)^T$ is a spherically distributed $\mathbb{R}^p$-valued random variable, and let $v_1, \ldots, v_k$ be an orthonormal basis of $\mathcal{H}$. Define $f \coloneqq \sum_{i \in \mathbb{N}_p} \alpha_i v_i$ and let $U : \mathcal{H} \to \mathcal{H}$ be a unitary operator. Let T be as in Lemma 2.2.14. Then, as T is unitary and by identifying it with an orthogonal matrix, $A \overset{D}{=} T(A)$. Furthermore,

for each $i \in \mathbb{N}_p$, $\alpha_i \overset{D}{=} T_i(A)$. Consequently

$$U(f) = \sum_{i \in \mathbb{N}_p} T_i(A) v_i \overset{D}{=} \sum_{i \in \mathbb{N}_p} \alpha_i v_i = f$$

which means f is unitarily invariant. □

Proof of Theorem 2.3.1. Let $i \in \mathbb{N}_m$. Let $\mathcal{T}_i$ be the topology induced by d_i. By Theorem 22.2 of Willard (1970), there exists a metric d_i^* on S_i which also induces $\mathcal{T}_i$ and satisfies

$$\exists\, C_i \geq 0 : x, y \in S_i \implies d_i^*(x, y) \leq C_i.$$

It may be assumed that $C_i = 1$. For each $x \in S_i$ and $\epsilon > 0$, let the open ball centred at x with radius ϵ be defined as $B_i(x, \epsilon) \coloneqq \left\{ y \in S_i : d_i^*(x, y) < \epsilon \right\}$. Now recall that $(S_i, \mathcal{T}_i)$ being separable means that there exists a countable subset $E_i \subseteq S_i$ such that $\overline{E_i} = S_i$ where $\overline{E_i}$ is the intersection of all closed sets which contain E_i. Let E_i be any such set. Following the proof of Proposition 20.7 in Bass (2016), it is seen that $\mathcal{C}_i \coloneqq \{B_i(x, r) : r \in \{z \in \mathbb{Q} : z > 0\},\ x \in E_i\}$ forms a countable base for $\mathcal{T}_i$. Recalling Remark 2.3.1, this gives that $\sigma(\mathcal{C}_i) = \sigma(G(\mathcal{C}_i)) = \sigma(\mathcal{T}_i)$.

Now let $\mathcal{T}$ be the product topology on S. By Proposition 2.3.1 of Engelking (1989),

$$\mathcal{D} \coloneqq \left\{ \bigtimes_{i \in \mathbb{N}_m} A_i \;\middle|\; [\forall i \in \mathbb{N}_m \, [A_i \in \mathcal{T}_i]] \wedge [\mathrm{Card}(\{i \in \mathbb{N}_m : A_i \neq S_i\}) < \aleph_0] \right\}$$

and

$$\mathcal{D}^* \coloneqq \left\{ \bigtimes_{i \in \mathbb{N}_m} A_i \in D \;\middle|\; j \in \{i \in \mathbb{N}_m : A_i \neq S_i\} \implies A_j \in \mathcal{C}_j \right\}$$

are bases for $\mathcal{T}$. It is now shown that $\mathcal{D}^*$ is countable. Note that $\mathrm{Card}(\mathcal{C}_i) \leq \aleph_0$ and $0 < \mathrm{Card}(\mathcal{C}_i)$. Let $\mathfrak{E}_m$ be the set of those functions in $\mathrm{Hom}(\mathbb{N}_m, \{0, 1\})$ which have finitely many inputs with image equal to 1. For $f \in \mathfrak{E}_m$, let $\left(F_{if}\right)_{i \in \mathbb{N}_m}$ be a family of sets with $\mathrm{Card}\left(F_{if}\right) = f(i)$ for all $i \in \mathbb{N}_m$. For $f \in \mathfrak{E}_m$, let $\mathcal{I}_f$ be

the finite subset of $\mathbb{N}_m$ for which $f(i) = 1$ whenever $i \in \mathcal{I}_f$. It is seen that

$$\begin{aligned}
\sum_{f\in\mathfrak{E}_m}\prod_{i\in\mathbb{N}_m}\mathrm{Card}\left(\mathcal{C}_i\right)^{f(i)} &= \sum_{f\in\mathfrak{E}_m}\prod_{i\in\mathbb{N}_m}\mathrm{Card}\left(\mathcal{C}_i\right)^{\mathrm{Card}\left(F_{if}\right)} \\
&= \sum_{f\in\mathfrak{E}_m}\prod_{i\in\mathbb{N}_m}\mathrm{Card}\left(\mathrm{Hom}\left(F_{if},\mathcal{C}_i\right)\right) \\
&= \sum_{f\in\mathfrak{E}_m}\mathrm{Card}\left(\bigtimes_{i\in\mathbb{N}_m}\mathrm{Hom}\left(F_{if},\mathcal{C}_i\right)\right) \\
&= \mathrm{Card}\left(\bigsqcup_{f\in\mathfrak{E}_m}\left(\bigtimes_{i\in\mathbb{N}_m}\mathrm{Hom}\left(F_{if},\mathcal{C}_i\right)\right)\right) \\
&= \mathrm{Card}\left(\left\{(f,a)\;\middle|\; f\in\mathfrak{E}_m,\ a\in\bigtimes_{i\in\mathbb{N}_m}\mathrm{Hom}\left(F_{if},\mathcal{C}_i\right)\right\}\right).
\end{aligned}$$

Let $\Gamma \coloneqq \left\{(f,a) \;\middle|\; f\in\mathfrak{E}_m,\ a\in\bigtimes_{i\in\mathbb{N}_m}\mathrm{Hom}\left(F_{if},\mathcal{C}_i\right)\right\}$. Consider the function $M : \mathcal{D}^* \to \Gamma$ given by the mapping

$$\bigtimes_{i\in\mathbb{N}_m} A_i \mapsto (f,a)$$

where $f : \mathbb{N}_m \to \{0,1\}$ is the function given by $f(i) = 1$ if $i \in \left\{j \in \mathbb{N}_m : A_j \neq S_j\right\}$ and $f(i) = 0$ otherwise, and a is the function which is specified by $a(i) = \emptyset$ for $i \notin \left\{j \in \mathbb{N}_m : A_j \neq S_j\right\}$ (for any set Ω, $\emptyset$ is itself a function from $\emptyset$ to Ω) and, for $i \in \left\{j \in \mathbb{N}_m : A_j \neq S_j\right\}$, $a(i) = A_i$. It is claimed that M is a bijection. To see that it is first an injection, suppose

$$M\left(\bigtimes_{i\in\mathbb{N}_m} A_i\right) = M\left(\bigtimes_{i\in\mathbb{N}_m} B_i\right)$$

and let $(f_1,a_1) \coloneqq M\left(\bigtimes_{i\in\mathbb{N}_m} A_i\right)$ and $(f_2,a_2) \coloneqq M\left(\bigtimes_{i\in\mathbb{N}_m} B_i\right)$. As two ordered pairs are equal if and only if both of their corresponding entries are equal, it holds that $f_1 = f_2$ and $a_1 = a_2$. The first of these implies that $\{i \in \mathbb{N}_m : A_i \neq S_i\} = \{i \in \mathbb{N}_m : B_i \neq S_i\}$. This along with $a_1 = a_2$ implies that $A_i = B_i$ for $i \in \mathbb{N}_m$.

Hence

$$\begin{aligned}\bigtimes_{i\in\mathbb{N}_m} A_i &= \left\{ g : \mathbb{N}_m \to \bigcup_{i\in\mathbb{N}_m} A_i \;\middle|\; \forall i \in \mathbb{N}_m \left[g(i) \in A_i \right] \right\} \\ &= \left\{ g : \mathbb{N}_m \to \bigcup_{i\in\mathbb{N}_m} B_i \;\middle|\; \forall i \in \mathbb{N}_m \left[g(i) \in B_i \right] \right\} \\ &= \bigtimes_{i\in\mathbb{N}_m} B_i.\end{aligned}$$

To see that M is also a surjection, let $(f, a) \in \Gamma$ and consider a family of sets $(A_i)_{i\in\mathbb{N}_m}$ where $A_i = S_i$ for $i \notin \mathcal{I}_f$ and $A_i \in \mathcal{C}_i$ for $i \in \mathcal{I}_f$. It is immediate that $M\left(\bigtimes_{i\in\mathbb{N}_m} A_i\right) = (f, a)$. This all implies that Γ and $\mathcal{D}^*$ are equinumerous. Thus

$$\operatorname{Card}(\mathcal{D}^*) = \sum_{f\in\mathfrak{E}_m} \prod_{i\in\mathbb{N}_m} \operatorname{Card}(\mathcal{C}_i)^{f(i)}$$

Let $b_m \coloneqq \operatorname{Card}(\mathbb{N}_m)$ and $c_m \coloneqq \operatorname{Card}(\mathfrak{E}_m)$. For $j < c_m$, let g_j be the function in $\mathfrak{E}_m$ given by $g_j(1+i) = 1$ whenever $1 + i \in \mathcal{I}_{g_j}$ and $g_j(1+i) = 0$ otherwise. The above can be rewritten as

$$\sum_{j<c_m} \prod_{i<b_m} \operatorname{Card}(\mathcal{C}_{1+i})^{g_j(1+i)}.$$

If $m \in \mathbb{N}$, this is a finite sum of a finite product of terms with countable cardinality so, by Remark 2.1.10, is itself countable. If $m = \omega$ and $j < c_m$, let $x_j \coloneqq \prod_{i<b_m} \operatorname{Card}(\mathcal{C}_{1+i})^{g_j(1+i)}$. Infinitely many of the terms in the product defining x_j are equal to 1 and finitely many terms have countable (non-zero) cardinality; all the terms being one of the two. Hence, by Remark 2.1.10, x_j is countable and non-zero. Now, by Lemma 2.1.3,

$$\sum_{j<c_m} x_j = c_m \cdot \sup\left\{x_j : j < c_m\right\}$$

It is readily verified that $c_m = \aleph_0$. As $x_j \leq \aleph_0$ for each $j < c_m$, it holds that $\sup\left\{x_j : j < c_m\right\} \leq \aleph_0$. Summarising all this, $\mathcal{D}^*$ is a countable base for $\mathcal{T}$.

Now let $d_S : S \times S \to \{x \in \mathbb{R} : 0 \leq x\}$ be defined by:

$$d_S(x, y) = \sum_{i \in \mathbb{N}_m} \frac{d_i^*(x(i), y(i))}{2^i}.$$

It is readily verified that d_S is a metric on S. Following the proof of Theorem 22.3 of Willard (1970) gives that it induces $\mathcal{T}$. Proposition 20.7 of Bass (2016) then implies that $(S, \mathcal{T})$ is separable. Thus it has been shown that $(S, \mathcal{T})$ is a separable metric space.

For $i \in \mathbb{N}_m$, let ϕ_i be the projection from S to S_i and let

$$A \coloneqq \left\{\phi_i^{-1}(U_i) : i \in \mathbb{N}_m, U_i \in \mathcal{C}_i\right\}$$
$$A^* \coloneqq \left\{\phi_i^{-1}(U_i) : i \in \mathbb{N}_m, U_i \in \mathcal{T}_i\right\}$$

As $\mathcal{C}_i$ is a base for $\mathcal{T}_i$, it holds that $G(A) = G(A^*)$. By definition, $\mathcal{T} = G(A^*)$. Thus $\mathcal{T} = G(A)$. This implies that $\sigma(\mathcal{T}) = \sigma(G(A))$. Furthermore, as A is countable, $\sigma(A) = \sigma(G(A))$. Thus $\sigma(\mathcal{T}) = \sigma(A)$. Now as $A \subseteq \bigotimes_{i \in \mathbb{N}_m} \sigma(\mathcal{C}_i)$ and $\sigma(\mathcal{C}_i) = \sigma(\mathcal{T}_i)$, $\sigma(A) \trianglelefteq \bigotimes_{i \in \mathbb{N}_m} \sigma(\mathcal{T}_i)$. By definition, $\bigotimes_{i \in \mathbb{N}_m} \sigma(\mathcal{T}_i)$ is the coarsest σ-field on S such that the projection maps are all measurable, hence $\bigotimes_{i \in \mathbb{N}_m} \sigma(\mathcal{T}_i) \trianglelefteq \sigma(\mathcal{T})$.

This allows the conclusion $\sigma(\mathcal{T}) = \sigma(A) \trianglelefteq \bigotimes_{i \in \mathbb{N}_m} \sigma(\mathcal{T}_i) \trianglelefteq \sigma(\mathcal{T})$. There is thus equality throughout, which finishes the proof. □

Proof of Theorem 2.4.1. The claim is equivalent to

$$\bigotimes_{i \in \mathcal{I}} \operatorname{Ker}\left(A_i^*\right) = \operatorname{Ker}\left(\left[\bigotimes_{i \in \mathcal{I}} A_i\right]^*\right).$$

Before establishing this, it is first shown that $\left[\bigotimes_{i \in \mathcal{I}} A_i\right]^* = \bigotimes_{i \in \mathcal{I}} A_i^*$. To see this,

let $f \in \mathcal{H}$ and $g \in \mathcal{G}$ and consider

$$\begin{aligned}
&\left\langle f, \left[\bigotimes_{i\in\mathcal{I}} A_i\right]^* g\right\rangle_{\mathcal{H}} \\
&= \left\langle \left[\bigotimes_{i\in\mathcal{I}} A_i\right] f, g\right\rangle_{\mathcal{G}} \\
&= \bigoplus_{j\in\mathcal{I}} \left\langle \left[\left[\bigotimes_{i\in\mathcal{I}} A_i\right] f\right](j), g(j)\right\rangle_{\mathcal{G}_j} \\
&= \bigoplus_{j\in\mathcal{I}} \left\langle A_j f(j), g(j)\right\rangle_{\mathcal{G}_j} \\
&= \bigoplus_{j\in\mathcal{I}} \left\langle f(j), A_j g(j)^*\right\rangle_{\mathcal{G}_j} \\
&= \left\langle f, \left[\bigotimes_{i\in\mathcal{I}} A_i^*\right] g\right\rangle_{\mathcal{H}}.
\end{aligned}$$

Now let $h \in \bigotimes_{i\in\mathcal{I}} \operatorname{Ker}\left(A_i^*\right)$. Then, for any $f \in \mathcal{H}$,

$$\begin{aligned}
&\left\langle f, \left[\bigotimes_{i\in\mathcal{I}} A_i\right]^* h\right\rangle_{\mathcal{H}} \\
&= \bigoplus_{j\in\mathcal{I}} \left\langle f(j), \left[\left[\bigotimes_{i\in\mathcal{I}} A_i\right]^* h\right](j)\right\rangle_{\mathcal{H}_j} \\
&= \bigoplus_{j\in\mathcal{I}} \left\langle f(j), A_j^* h(j)\right\rangle_{\mathcal{H}_j} \\
&= \bigoplus_{j\in\mathcal{I}} \langle f(j), 0\rangle_{\mathcal{H}_j} = 0.
\end{aligned}$$

Hence $h \in \operatorname{Ker}\left(\left[\bigotimes_{i\in\mathcal{I}} A_i\right]^*\right)$. For the other inclusion, begin by letting $h \in \operatorname{Ker}\left(\left[\bigotimes_{i\in\mathcal{I}} A_i\right]^*\right)$ so that, for any $f \in \mathcal{H}$,

$$\left\langle f, \left[\bigotimes_{i\in\mathcal{I}} A_i\right]^* h\right\rangle_{\mathcal{H}} = \bigoplus_{j\in\mathcal{I}} \left\langle f(j), A_j^* h(j)\right\rangle_{\mathcal{H}_j} = 0.$$

Take $f = \left[\bigotimes_{i\in\mathcal{I}} A_i\right]^* h$ to obtain

$$\left\|\left[\bigotimes_{i\in\mathcal{I}} A_i\right]^* h\right\|_{\mathcal{H}}^2 = \bigoplus_{j\in\mathcal{I}} \left\|A_j^* h(j)\right\|_{\mathcal{H}_j}^2 = 0.$$

This gives that $h(j) \in \mathrm{Ker}\left(A_j^*\right)$ for $j \in \mathcal{I}$, thus giving the desired inclusion. □

Proof of Lemma 2.5.2. First see that $\mathbb{E}\left(\langle f, g\rangle_{\mathcal{H}}\right) = \mathbb{E}\left(\mathbb{E}\left(\langle f, g\rangle_{\mathcal{H}}|g\right)\right)$. Now, for any $g^* \in \mathcal{H}$, $\mathbb{E}\left(\langle f, g\rangle_{\mathcal{H}}|g = g^*\right) = \mathbb{E}\left(\langle f, g^*\rangle_{\mathcal{H}}\right) = \langle\mathbb{E}\left(f\right), g^*\rangle_{\mathcal{H}}$. Hence, as f and g are independent, $\mathbb{E}\left(\langle f, g\rangle_{\mathcal{H}}|g\right) \overset{\text{a.s.}\mathbb{P}}{=} \langle\mathbb{E}\left(f\right), g\rangle_{\mathcal{H}}$. Thus $\mathbb{E}\left(\langle f, g\rangle_{\mathcal{H}}\right) = \mathbb{E}\left(\langle\mathbb{E}\left(f\right), g\rangle_{\mathcal{H}}\right) = \langle\mathbb{E}\left(f\right), \mathbb{E}\left(g\right)\rangle_{\mathcal{H}}$. □

Proof of Lemma 2.5.3. Let $f \in \mathcal{H}_1$. Consider

$$\begin{aligned}
\left[\mathbb{E}\left(A|G\right) \otimes \mathbb{E}\left(B|G\right)\right] f &\overset{\text{a.s.}\mathbb{P}}{=} \langle f, \mathbb{E}\left(A|G\right)\rangle_{\mathcal{H}_1} \mathbb{E}\left(B|G\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\langle f, A\rangle_{\mathcal{H}_1}\middle|G\right) \mathbb{E}\left(B|G\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\langle f, A\rangle_{\mathcal{H}_1} \mathbb{E}\left(B|G\right)\middle|G\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left[A \otimes \mathbb{E}\left(B|G\right)\right] f|G\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left[A \otimes \mathbb{E}\left(B|G\right)\right]|G\right) f.
\end{aligned}$$

□

Proof of Lemma 2.5.4. Using the law of total expectation

$$\begin{aligned}
\mathbb{E}\left(A \otimes \mathbb{E}\left(B|G\right)\right) &= \mathbb{E}\left(\mathbb{E}\left(A \otimes \mathbb{E}\left(B|G\right)|G\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(A|G\right) \otimes \mathbb{E}\left(B|G\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\mathbb{E}\left(A|G\right) \otimes B|G\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(A|G\right) \otimes B\right).
\end{aligned}$$

□

Proof of Corollary 2.5.5. Using Lemma 2.5.4 and the law of total expectation

$$\begin{aligned}\operatorname{Cov}\left(A, \mathbb{E}\left(B|G\right)\right) &\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(A \otimes \mathbb{E}\left(B|G\right)\right) - \left[\mathbb{E}\left(A\right) \otimes \mathbb{E}\left(B\right)\right] \\ &\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\mathbb{E}\left(A|G\right) \otimes B\right) - \left[\mathbb{E}\left(A\right) \otimes \mathbb{E}\left(B\right)\right] \\ &\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(\mathbb{E}\left(A|G\right), B\right).\end{aligned}$$

Furthermore

$$\operatorname{Cov}\left(A, \mathbb{E}\left(B|G\right)\right) \overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(A, \mathbb{E}\left(\mathbb{E}\left(B|G\right)|G\right)\right) \overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(\mathbb{E}\left(A|G\right), \mathbb{E}\left(B|G\right)\right).$$

Putting these together completes the proof. □

Proof of Lemma 2.5.6. Let $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. Then

$$\begin{aligned}&\operatorname{Cov}\left(\left\langle f, A\right\rangle_{\mathcal{H}_1}, \left\langle B, g\right\rangle_{\mathcal{H}_2} \middle| G\right) \\ &\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left[\left\langle f, A\right\rangle_{\mathcal{H}_1} - \mathbb{E}\left(\left\langle f, A\right\rangle_{\mathcal{H}_1} \middle| G\right)\right]\left[\left\langle B, g\right\rangle_{\mathcal{H}_2} - \mathbb{E}\left(\left\langle B, g\right\rangle_{\mathcal{H}_2} \middle| G\right)\right] \middle| G\right) \\ &\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left[\left\langle f, A\right\rangle_{\mathcal{H}_1} - \left\langle f, \mathbb{E}\left(A|G\right)\right\rangle_{\mathcal{H}_1}\right]\left[\left\langle B, g\right\rangle_{\mathcal{H}_2} - \left\langle \mathbb{E}\left(B|G\right), g\right\rangle_{\mathcal{H}_2}\right] \middle| G\right) \\ &\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left\langle f, A - \mathbb{E}\left(A|G\right)\right\rangle_{\mathcal{H}_1} \left\langle B - \mathbb{E}\left(B|G\right), g\right\rangle_{\mathcal{H}_2} \middle| G\right) \\ &\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left\langle \left\langle f, A - \mathbb{E}\left(A|G\right)\right\rangle_{\mathcal{H}_1} \left[B - \mathbb{E}\left(B|G\right)\right], g\right\rangle_{\mathcal{H}_2} \middle| G\right) \\ &\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left\langle \left[\left(A - \mathbb{E}\left(A|G\right)\right) \otimes \left(B - \mathbb{E}\left(B|G\right)\right)\right] f, g\right\rangle_{\mathcal{H}_2} \middle| G\right) \\ &\overset{\text{a.s.}\mathbb{P}}{=} \left\langle \mathbb{E}\left(\left[A - \mathbb{E}\left(A|G\right)\right] \otimes \left[B - \mathbb{E}\left(B|G\right)\right]|G\right) f, g\right\rangle_{\mathcal{H}_2} \\ &\overset{\text{a.s.}\mathbb{P}}{=} \left\langle \Sigma_{BA|G} f, g\right\rangle_{\mathcal{H}_2}.\end{aligned}$$

The second final equality follows from Corollary 37 in Section 1 of Chapter 1 of Dinculeanu (2000). The other relation follows from the fact that $\Sigma^*_{BA|G} \overset{\text{a.s.}\mathbb{P}}{=} \Sigma_{AB|G}$. □

Proof of Lemma 2.5.9. Begin by supposing that

$$\left[\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2 | \left(\mathcal{F}_3, \mathcal{F}_4\right)\right] \wedge \left[\mathcal{F}_2 \perp\!\!\!\perp \mathcal{F}_3 | \mathcal{F}_4\right].$$

By part 4 of Lemma 2.5.8, it follows that $\mathcal{F}_2 \perp\!\!\!\perp (\mathcal{F}_1, \mathcal{F}_3) \,|\mathcal{F}_4$. The desired expression then follows from parts 2 (giving $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2|\mathcal{F}_4$) and 3 (giving that $\mathcal{F}_2 \perp\!\!\!\perp \mathcal{F}_3| \,(\mathcal{F}_1, \mathcal{F}_4)$). The converse is proven in a similar manner. □

Proof of Lemma 2.5.10. As $\mathcal{F}_4 \trianglelefteq \mathcal{F}_5 \trianglelefteq \mathcal{F}_2$ by assumption, it holds that $\mathcal{F}_2 = (\mathcal{F}_2, \mathcal{F}_5)$ and $(\mathcal{F}_3, \mathcal{F}_5) = (\mathcal{F}_3, \mathcal{F}_4, \mathcal{F}_5)$. By substituting, $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2| \,(\mathcal{F}_3, \mathcal{F}_4)$ becomes $\mathcal{F}_1 \perp\!\!\!\perp (\mathcal{F}_2, \mathcal{F}_5) \,|\, (\mathcal{F}_3, \mathcal{F}_4)$. By part 3 of Lemma 2.5.8, this implies that

$$\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2| \,(\mathcal{F}_5, (\mathcal{F}_3, \mathcal{F}_4)) \,.$$

This is equivalent to $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2| \,(\mathcal{F}_3, \mathcal{F}_4, \mathcal{F}_5)$ which, by substitution, can be rewritten as $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2| \,(\mathcal{F}_3, \mathcal{F}_5)$. □

Proof of Corollary 2.5.11. As $\mathcal{F}_3 \trianglelefteq \mathcal{F}_4 \trianglelefteq \mathcal{F}_2$ by assumption, it holds that $\mathcal{F}_2 = (\mathcal{F}_2, \mathcal{F}_4)$ and $\mathcal{F}_4 = (\mathcal{F}_3, \mathcal{F}_4)$. By substituting, it is seen that $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2|\mathcal{F}_3$ can be rewritten as $\mathcal{F}_1 \perp\!\!\!\perp (\mathcal{F}_2, \mathcal{F}_4) \,|\mathcal{F}_3$. By part 3 of Lemma 2.5.8, this implies that $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2| \,(\mathcal{F}_3, \mathcal{F}_4)$ which, by substitution, is equivalent to $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2|\mathcal{F}_4$. □

Proof of Theorem 2.5.21. By applying the isometric isomorphism T of Park and Muandet (2020) to their $\mu_{P_{XY|\mathbb{Z}}}$ and $\mu_{P_{X|Z}} \otimes \mu_{P_{Y|Z}}$ when evaluated at $x \in \Omega$, the proof of their Theorem 5.4 yields the proof of Theorem 2.5.21. Note that their proof does not depend on Z being a random variable, and can be replaced with any $G \trianglelefteq \mathcal{F}$. □

Chapter 3

The predictive potential of principal components in regression

3.1 Outline of chapter

In this chapter, the predictive potential of principal components in regression is explored. First, the principal components procedure is described technically for the multivariate and Hilbertian settings. Then a nonlinear extension is technically described for the general predictor setting. This is followed by a review of the results in the existing literature regarding the predictive potential of principal components for multivariate predictors. These results are then greatly extended to the case when nonlinear principal components analysis is applied with general predictors. All results are proven in Section 3.5.4 at the end of the chapter.

Remark 3.1.1. The results in this chapter build on (and subsume) those published in Jones and Artemiou (2019), Jones et al. (2020), and Jones and Artemiou (2021).

3.2 The principal components analysis procedure

Principal components analysis (PCA) is the earliest, most well-known, and most commonly used procedure for unsupervised dimension reduction. An authoritative work on the subject is Jolliffe (2002). As a comprehensive presentation is provided in that text, the procedure given here is only a basic technical description.

3.2.1 Multivariate setting

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ be a $\mathbb{R}^p$-valued random variable. Suppose that $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ exist. It is assumed that $\mathbb{E}(X) = 0$. Let $\mathcal{S}^p$ be the set of unit norm vectors in $\mathbb{R}^p$. The first principal direction v_1 is defined to be any element of:

$$\left\{ v \in \mathcal{S}^p : \forall w \in \mathcal{S}^p, \mathrm{Var}\left(v^T X\right) \geq \mathrm{Var}\left(w^T X\right) \right\}.$$

The first principal component is $v_1^T X$. The k^{th} $(k > 1)$ principal direction v_k is then defined to be any unit norm variance maximiser subject to the additional constraint of being orthogonal to the previous principal directions. The k^{th} principal component is $v_k^T X$.

An alternative, equivalent, formulation is to perform an eigendecomposition of the covariance matrix $\mathrm{Var}(X)$. Let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ denote the eigenvalues of $\mathrm{Var}(X)$. The k^{th} $(k \geq 1)$ principal direction v_k is any normalised eigenvector of $\mathrm{Var}(X)$ which corresponds to λ_k. The requirement that the principal directions have unit norm ensures that $\mathrm{Var}\left(v_i^T X\right) = \lambda_i$.

The orthogonality constraint implies that the principal components have zero covariance, and so are uncorrelated. To see this, let i and j be distinct and consider

$$\mathrm{Cov}\left(v_i^T X, v_j^T X\right) = v_i^T \mathrm{Var}(X)\, v_j = \lambda_i v_i^T v_j = 0.$$

Because of this, the procedure is commonly used to replace a set of correlated variables with uncorrelated ones. If X has a multivariate Gaussian distribution, this further implies that the principal components are independent.

3.2.2 Hilbertian setting

PCA is extended to data that lie in a real separable Hilbert space $\mathcal{H}$ by replacing the $\mathbb{R}^p$ inner product with the $\mathcal{H}$ inner product. That is, $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is now a $\mathcal{H}$-valued random variable. It is supposed that $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ exist, and that $\mathbb{E}(X) = 0$. Let $\mathcal{S}$ denote the unit norm elements of $\mathcal{H}$. The first principal direction v_1 is defined to be any element of:

$$\{v \in \mathcal{S} : \forall w \in \mathcal{S}, \mathrm{Var}(\langle v, X\rangle_{\mathcal{H}}) \geq \mathrm{Var}(\langle w, X\rangle_{\mathcal{H}})\}.$$

The first principal component is $\langle v_1, X\rangle_{\mathcal{H}}$. The k^{th} $(k > 1)$ principal direction v_k is then defined to be any unit norm variance maximiser subject to the additional constraint of being orthogonal to the previous principal directions. The k^{th} principal component is $\langle v_k, X\rangle_{\mathcal{H}}$.

Like for multivariate data, this procedure is equivalent to performing an eigendecomposition of $\mathrm{Var}(X)$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be the eigenvalues of $\mathrm{Var}(X)$. The k^{th} $(k \geq 1)$ principal direction v_k is any normalised eigenvector of $\mathrm{Var}(X)$ which corresponds to λ_k. Again, in this setting, it holds that: (1) the variance of the k^{th} $(k \geq 1)$ principal component equals λ_k, and (2) any two distinct principal components have zero covariance.

See Chapter 9 of Hsing and Eubank (2015) for further details of PCA for Hilbertian data.

3.3 A nonlinear version of principal components analysis with general predictors

To formulate nonlinear principal components analysis with general predictors, suppose that $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_X, \mathcal{F}_X)$ is a Ω_X-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume $\Omega_X = X(\Omega)$. Let $\mathcal{G}$ be a real separable Hilbert space whose members are measurable real-valued functions defined on Ω_X, such that there exists a unique (up to permutations of $\mathcal{G}$) surjective function $f : \left(\Omega_X, \sigma_{\Omega_X}, \mathbb{P}_X\right) \to (\mathcal{G}, \mathcal{B}(\mathcal{G}))$ for which the covariance operator $\mathrm{Var}(f \circ X)$ exists and, without loss of generality, $\mathbb{E}(f \circ X) = 0$. Let $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ denote the inner product in $\mathcal{G}$, $\|\cdot\|_{\mathcal{G}}$ denote the induced norm, and $\mathcal{S}$ be the set of unit norm functions in $\mathcal{G}$.

$\mathcal{G}$ is often taken to be a reproducing kernel Hilbert space derived from some measurable kernel function $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$. Though this is common, it is not necessary for the results in this thesis. Indeed the space of square-integrable real valued functions on some probability space, where almost surely equal functions are considered equivalent, is not generally a reproducing kernel Hilbert space (see Berlinet and Thomas-Agnan (2004)).

Remark 3.3.1. If $\mathcal{G}$ is generated by a measurable kernel $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$, the function f is the function which satisfies $f \circ X = \kappa_X(\cdot, X)$. $\mathrm{Var}(f \circ X)$ exists provided that $\mathbb{E}(\kappa_X(X, X))$ exists. Kernels satisfying this assumption are known to exist (see, e.g., Virta et al. (2022)) for when $(\Omega_X, \mathcal{F}_X)$ is a separable complete metric space with the Borel σ-field. Therefore the requirements for $\mathcal{G}$ given above can be satisfied in the kernel setting.

Remark 3.3.2. In this general setup, Ω_X need not be a vector space so it is natural to ask what is meant by "dimension reduction" in this setting. What it means is that the extracted components generate a sub-σ-field of $\sigma(X)$.

At population level, nonlinear principal components is described as follows. The first nonlinear principal direction v_1 is any function in

$$S \coloneqq \left\{u \in \mathcal{S} : \forall w \in \mathcal{S} \left[\text{Var}\left(\langle u, f \circ X\rangle_{\mathcal{G}}\right) \geq \text{Var}\left(\langle w, f \circ X\rangle_{\mathcal{G}}\right)\right]\right\}$$

The first nonlinear principal component is $\langle v_1, f \circ X\rangle_{\mathcal{G}}$. For $k = 2, 3, \ldots$, the k^{th} nonlinear principal direction v_k is any unit norm variance maximiser subject to the constraints

$$\text{Cov}\left(\langle v, f \circ X\rangle_{\mathcal{G}}, \langle v_i, f \circ X\rangle_{\mathcal{G}}\right) = 0, \;\; i = 1, \ldots, k-1.$$

The k^{th} principal component is $\langle v_k, f \circ X\rangle_{\mathcal{G}}$. This is much more general than the classical (linear) principal components because the maximisation is carried out among all functions in $\mathcal{S}$, not just linear functions of the form $a^T X$.

Similarly to classical principal components analysis, the nonlinear version can be represented as an eigendecomposition task. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be the eigenvalues of $\text{Var}(f \circ X)$. The k^{th} ($k \geq 1$) principal direction v_k is any normalised eigenvector of $\text{Var}(f \circ X)$ which corresponds to λ_k.

3.4 Literature review: the predictive potential of principal components in regression with multivariate data

As mentioned in Chapter 1, it is common, in high-dimensional regression, to regress the response on the leading principal components of the predictor. This practice, called principal component regression, is controversial. As Cox (1968) notes

> "A difficulty seems to be that there is no logical reason why the dependent variable should not be closely tied to the least important principal component."

This issue arises as the principal components are extracted without making use of the response Y (i.e.the procedure is unsupervised). This means that there is no guarantee that, for any given dataset, the higher-ranking components are more informative of Y than lower-ranking ones. However, one wonders whether, across a range of datasets, the higher-ranking components will tend to be more informative than the lower-ranking ones. Cook (2007) gives a historical account of the debate surrounding the use of principal component regression, highlighting the views of both advocates and opponents of the practice.

In their comments on Cook (2007), Li (2007a) proposed a conjecture which suggests a probabilistic justification for the practice. The conjecture roughly (the exact form is more technical than outlined here) states

> If nature (the original quote author's term) uniformly randomly selects a covariance matrix Σ for the predictor X and independently randomly selects a linear relation between X and the response Y then, conditioning on Σ and the regression coefficients, the first principal component of X is the most likely, among all of the principal components, to have the largest absolute correlation with Y.

Motivated by this conjecture, Artemiou and Li (2009) gave empirical evidence in support of its claim by examining 33 datasets and comparing how often the first component has the largest absolute correlation with the response against the second component. They also proved that, in a linear regression setting, the i^{th} principal component tends (i.e.with probability greater than $1/2$) to have greater squared correlation with the response than the j^{th} component (where $i < j$). This

was done under a permutation invariance assumption on the eigenvalues and eigenvectors of the randomly chosen covariance matrix Σ of X. Following up on that result, Ni (2011) tacitly used a rotational invariance assumption on Σ to obtain an exact form for the probability of this phenomenon in terms of the eigenvalues of the covariance matrix. They also used a spherical symmetry assumption on the randomly chosen regression coefficients to derive a similar result when the conditioning is on the regression coefficients as opposed to the covariance matrix. Artemiou and Li (2013) generalised these results to a conditional independence model, which subsumes the linear regression model as a subcase.

Spherical distributions were central to the work of Artemiou and Li (2013). For data in a real separable infinite-dimensional Hilbert space, this notion cannot be generalised. This is because Ψ in Definition 2.2.29 cannot be the identity operator as it is not nuclear for such spaces. This implies that it is impossible for the characteristic function of $A - \mathbb{E}(A)$ (A is as in Definition 2.2.29) to depend only on the norm of f. As a result of these considerations, the notion of a spherically symmetric distribution cannot be extended to the entirety of an infinite-dimensional space. In subsequent developments, it is thus necessary to assume that the space is finite-dimensional. That said, it is seen that some results can still be established for the infinite-dimensional setting.

Having outlined their history, the discussed results are now presented formally. Artemiou and Li (2009) gave a matrix version of Definition 2.2.32 to define what is meant by uniformly randomly selecting a covariance matrix. It means, intuitively, that the relative positions of the eigenvalues and eigenvectors of Σ can be freely permutated without changing the distribution of Σ. If Σ is an orientationally uniform random matrix and if the random vector X satisfies $\mathbb{E}(X|\Sigma) \overset{\text{a.s.}\mathbb{P}}{=} 0$ and $\text{Var}(X|\Sigma) \overset{\text{a.s.}\mathbb{P}}{=} \Sigma$, then any random variable among $v_1^T X, \ldots, v_p^T X$ is equally likely to be the 1^{st}, 2^{nd}, $\ldots$, or p^{th} principal component of X. This follows from

Theorem 2.2.6.

Using the matrix version of Definition 2.2.32, Artemiou and Li (2009) proved Theorem 3.4.1.

Theorem 3.4.1 (Predictive potential under orientationally uniform covariance matrix, see Artemiou and Li (2009))**.** *Let X and β be p-dimensional real random vectors on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathrm{Var}(X)$ existing, and let Y be a real random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that:*

1. *Σ is an orientationally uniform random matrix*
2. *$\mathbb{E}(X|\Sigma) \overset{a.s.\mathbb{P}}{=} 0$ and $\mathrm{Var}(X|\Sigma) \overset{a.s.\mathbb{P}}{=} \Sigma$*
3. *$Y = \beta^T X + \epsilon$ where $\beta \perp\!\!\!\perp (X, \Sigma)$, $\epsilon \perp\!\!\!\perp (X, \beta, \Sigma)$, ϵ is square-integrable, $\mathbb{E}(\epsilon) = 0$, and $\mathrm{Var}(\epsilon) < \infty$*
4. *$\mathbb{P}(\beta \in G) > 0$ for any nonempty open set $G \subseteq \mathbb{R}^p$*

Write Σ as $\sum_{i \in \mathbb{N}_p} \lambda_i v_i v_i^T$. Let $\rho_k(\beta, \Sigma) \coloneqq \mathrm{Corr}^2\left(Y, v_k^T X \middle| \beta, \Sigma\right)$. Reorder the eigenvalues as $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(p)} > 0$, let $\left(v_{(1)}, \ldots v_{(p)}\right)$ be the corresponding normalised eigenvectors, and let $\rho_{(i)}(\beta, \Sigma)$ be the conditional squared correlation corresponding to $v_{(i)}$. Then, for $i < j \leq p$,

$$\mathbb{P}\left(\rho_{(i)}(\beta, \Sigma) \geq \rho_{(j)}(\beta, \Sigma)\right) > \frac{1}{2}.$$

Under Assumption 3.4.1 and Assumption 3.4.2 respectively, Ni (2011) proved Theorem 3.4.2 and Theorem 3.4.3. These give an explicit form for the probability of the phenomenon. Notice that the conditioning is different in each theorem. The first conditions on a random covariance matrix, while the second conditions on random regression coefficients. Assumption 3.4.2 was used tacitly.

Assumption 3.4.1. The p-dimensional random vector β has a spherically symmetric distribution.

Assumption 3.4.2. The $p \times p$ random matrix Σ is symmetric and has the same distribution as $U^T \Sigma U$ for any $p \times p$ orthogonal matrix U. Moreover, the eigenvalues (which are themselves random variables) of Σ are almost surely positive and distinct.

Theorem 3.4.2 (Ni (2011))**.** *Let X and β be p-dimensional real random vectors, with* $\mathrm{Var}(X)$ *existing, and let Y be a real random variable. Suppose that:*

1. *β satisfies Assumption 3.4.1*
2. $\mathbb{E}(X|\Sigma) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}(X|\Sigma) \overset{a.s.\mathbb{P}}{=} \Sigma$
3. *$Y = \beta^T X + \epsilon$ where $\beta \perp\!\!\!\perp (X, \Sigma)$, $\epsilon \perp\!\!\!\perp (X, \beta, \Sigma)$, ϵ is square-integrable,* $\mathbb{E}(\epsilon) = 0$*, and* $\mathrm{Var}(\epsilon) < \infty$

Write Σ as $\sum_{i \in \mathbb{N}_p} \lambda_i v_i v_i^T$ alike in Definition 2.2.32. Suppose that the eigenvalues of Σ are almost surely positive and distinct. Define $\rho_k(\Sigma) \coloneqq \mathrm{Corr}^2\left(Y, v_k^T X \middle| \Sigma\right)$. Reorder the eigenvalues as $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(p)} > 0$, let $\left(v_{(1)}, \ldots v_{(p)}\right)$ be the corresponding normalised eigenvectors, and let $\rho_{(i)}(\Sigma)$ be the conditional squared correlation corresponding to $v_{(i)}$. Then, for $i < j \leq p$,

$$\mathbb{P}\left(\rho_{(i)}(\Sigma) \geq \rho_{(j)}(\Sigma)\right) = \frac{2}{\pi}\mathbb{E}\left(\arctan\left(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\right)\right) > \frac{1}{2}.$$

Theorem 3.4.3 (Ni (2011))**.** *Let X and β be p-dimensional real random vectors and Y be a real random variable. Suppose that:*

1. *Σ satisfies Assumption 3.4.2*
2. $\mathbb{E}(X|\Sigma) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}(X|\Sigma) \overset{a.s.\mathbb{P}}{=} \Sigma$
3. *$Y = \beta^T X + \epsilon$ where $\beta \perp\!\!\!\perp (X, \Sigma)$, $\epsilon \perp\!\!\!\perp (X, \beta, \Sigma)$, ϵ is square-integrable,* $\mathbb{E}(\epsilon) = 0$*, and* $\mathrm{Var}(\epsilon) < \infty$.

Write Σ *as* $\sum_{i\in\mathbb{N}_p} \lambda_i v_i v_i^T$. *Define* $\rho_k(\beta) \coloneqq \mathrm{Corr}^2\left(Y, v_k^T X \middle| \beta\right)$ *Reorder the eigenvalues as* $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(p)} > 0$, *let* $\left(v_{(1)}, \ldots v_{(p)}\right)$ *be the corresponding normalised eigenvectors, and let* $\rho_{(i)}(\beta)$ *be the conditional squared correlation corresponding to* $v_{(i)}$. *Then, for* $i < j \leq p$

$$\mathbb{P}\left(\rho_{(i)}(\beta) \geq \rho_{(j)}(\beta)\right) = \frac{2}{\pi}\mathbb{E}\left(\arctan\left(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\right)\right) > \frac{1}{2}.$$

In the above results, the authors assumed a linear regression setting. Artemiou and Li (2013) explored the probabilistic tendency under a conditional independence model which includes the linear model as a subcase. The most general result they showed was Theorem 3.4.4.

Theorem 3.4.4 (Artemiou and Li (2013))**.** *Let* X *and* β *be* p*-dimensional random vectors. Let* Y *be a* m*-dimensional random vector. Let* $f : \mathbb{R}^m \to \mathbb{R}$ *be a measurable function. Suppose that*

1. $Y \perp\!\!\!\perp X | (\beta^T X, \beta, \Sigma)$
2. $\mathrm{Var}\left(f(Y)|\beta, \Sigma\right) < \infty$ *almost surely*
3. $\mathrm{Cov}\left(f(Y), \beta^T X \middle| \beta, \Sigma\right) \neq 0$ *and* $\mathrm{Cov}\left(f(Y), \beta^T X \middle| \beta, \Sigma\right) < \infty$ *almost surely*
4. $\mathbb{E}\left(X|\Sigma\right) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}\left(X|\Sigma\right) \overset{a.s.\mathbb{P}}{=} \Sigma$
5. $\beta \perp\!\!\!\perp (X, \Sigma)$
6. $\mathbb{E}\left(X \middle| \beta^T X, \beta, \Sigma\right)$ *is linear in* $\beta^T X$
7. *either Assumption 3.4.1 or Assumption 3.4.2 holds*
8. *the eigenvalues of* Σ *are almost surely positive and distinct.*

Write Σ *as* $\sum_{i\in\mathbb{N}_p} \lambda_i v_i v_i^T$. *Define* $\rho_k(\beta, \Sigma) \coloneqq \mathrm{Corr}^2\left(f(Y), v_k^T X \middle| \beta, \Sigma\right)$. *Reorder the eigenvalues as* $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(p)} > 0$, *let* $\left(v_{(1)}, \ldots v_{(p)}\right)$ *be the*

corresponding normalised eigenvectors, and let $\rho_{(i)}(\beta, \Sigma)$ *be the conditional squared correlation corresponding to* $v_{(i)}$*. Then, for* $i < j \leq p$*,*

$$\mathbb{P}\left(\rho_{(i)}(\beta, \Sigma) \geq \rho_{(j)}(\beta, \Sigma)\right) = \frac{2}{\pi}\mathbb{E}\left(\arctan\left(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\right)\right) > \frac{1}{2}.$$

The presentation of Theorem 3.4.4 in Artemiou and Li (2013) has $m = p$, so the presentation here is slightly more general.

3.5 The predictive potential of nonlinear principal components with general predictors

In this section and subsequent sections in this chapter, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_X, \mathcal{F}_X)$ is a random variable called the *predictor*, $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_Y, \mathcal{F}_Y)$ is a random variable, $g : (\Omega_Y, \mathcal{F}_Y) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a measurable function, $\mathcal{G}$ and f are as described in Section 3.3. $g(Y)$ represents the response variable — in the setting of multivariate Y, each component could represent the number of crimes of different kinds while $g(Y)$ could represent their total.

The central question of this section is whether nonlinear principal components analysis possesses a similar predictive tendency to that held by classical principal components when general predictors are used. This question is addressed at two different levels. The first is the conditional independence model

$$g(Y) \perp\!\!\!\perp X | \langle h, f \circ X \rangle_{\mathcal{G}} \tag{3.1}$$

where $h : \Omega_X \to \mathbb{R}$ is an arbitrary function in $\mathcal{G}$. The question is formulated by asking if, given a randomly selected function h and an independently randomly selected operator Σ where $f \circ X$ satisfies $\text{Var}\,(f \circ X | \Sigma) \overset{\text{a.s.}\mathbb{P}}{=} \Sigma$, nonlinear principal

components analysis enjoys a similar predictive tendency to that classically possessed?

The second level is the most general. Suppose Y and X are dependent but the dependence is not restricted by any model, parametric or nonparametric. Then, given a randomly selected regular conditional distribution for $g(Y)|X$ and an independently randomly selected operator Σ where $f \circ X$ satisfies $\text{Var}\,(f \circ X|\Sigma) \overset{\text{a.s.}\mathbb{P}}{=} \Sigma$, do nonlinear principal components possess the predictive tendency?

For technical reasons owing to the non-existence of orientationally uniform random operators, unitarily invariant random variables, and unitarily invariant random operators in infinite-dimensional spaces, it is assumed that $\mathcal{G}$ is finite-dimensional. This assumption is not restrictive as one can take as large a dimension as desired. Nevertheless, workarounds in the infinite-dimensional case are considered.

First explored is the first level where an orientationally uniform random operator is selected for the covariance operator independently of the randomly chosen h. Then considered is the first level with a unitarily invariant random h chosen independently of the randomly chosen covariance operator. Finally considered is the second level with a either an orientationally uniform or unitarily invariant random operator for the covariance operator.

Both levels are answered in the affirmative. While the work at the second level is more general, the first level is not technically a special case. This is because the conditions assumed for each of the levels are different.

3.5.1 Main results

In this section, it is assumed that $\mathcal{G}$ is q-dimensional for some $q \in \mathbb{N}$.

3.5.1.1 Orientationally uniform random operator

The following theorem is useful in the proofs of Theorem 3.5.2 and Theorem 3.5.4.

Theorem 3.5.1. *Suppose*

1. Σ *is a random positive definite operator on* $\mathcal{G}$ *with* $f \circ X$ *satisfying* $\mathbb{E}\left(f \circ X | \Sigma\right) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}\left(f \circ X | \Sigma\right) \overset{a.s.\mathbb{P}}{=} \Sigma$

2. h *is a random function in* $\mathcal{G}$ *satisfying* $h \perp\!\!\!\perp (X, \Sigma)$

3. $\mathbb{P}\left(h = 0\right) = 0$

4. *for any* $v \in \mathcal{G}$*, there exists a constant* c_v *such that*

$$\mathbb{E}\left(\left\langle v, f \circ X \right\rangle_{\mathcal{G}} \middle| h, \left\langle h, f \circ X \right\rangle_{\mathcal{G}}, \Sigma\right) \overset{a.s.\mathbb{P}}{=} c_v \left\langle h, f \circ X \right\rangle_{\mathcal{G}}.$$

Then

$$\mathbb{E}\left(f \circ X \middle| h, \left\langle h, f \circ X \right\rangle_{\mathcal{G}}, \Sigma\right) \overset{a.s.\mathbb{P}}{=} \left[\frac{1}{\left\langle h, \Sigma h \right\rangle_{\mathcal{G}}} \Sigma \left[h \otimes h\right]\right] \left(f \circ X\right).$$

Remark 3.5.1. Item 4 is called the *linearity condition*. Li (2007b) has shown that it holds if $f \circ X$ has an elliptically symmetric distribution. Even if $f \circ X$ is not elliptically symmetric, by expanding in an orthonormal basis for $\mathcal{G}$, the results of Hall and Li (1993) can be used, provided that h has unit norm (which can be assumed without loss of generality), to give probabilistic bounds in terms of q of the deviation from holding. As $q \to \infty$, there is convergence in probability of the deviation to zero. As q can be arbitrarily large, this condition is therefore mild.

Theorem 3.5.2. *Suppose*

1. Σ *is an orientationally uniform random operator with* $f \circ X$ *satisfying* $\mathbb{E}\left(f \circ X | \Sigma\right) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}\left(f \circ X | \Sigma\right) \overset{a.s.\mathbb{P}}{=} \Sigma$

2. $g(Y) \perp\!\!\!\perp X | \left(\left\langle h, f \circ X \right\rangle_{\mathcal{G}}, h, \Sigma\right)$ *where* $h \perp\!\!\!\perp (X, \Sigma)$

3. $\mathbb{P}(h = 0) = 0$

4. $\mathbb{P}(h \in G) > 0$ *for any nonempty open set* $G \subseteq \mathcal{G}$

5. $\mathrm{Var}(g(Y)|h, \Sigma) < \infty$ *almost surely*

6. $\mathrm{Cov}\left(g(Y), \langle h, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right)$ *is nonzero almost surely*

7. *for any* $v \in \mathcal{G}$*, there exists a constant* c_v *such that*

$$\mathbb{E}\left(\langle v, f \circ X\rangle_{\mathcal{G}} \middle| h, \langle h, f \circ X\rangle_{\mathcal{G}}, \Sigma\right) \overset{a.s.\mathbb{P}}{=} c_v \langle h, f \circ X\rangle_{\mathcal{G}}.$$

Write Σ *as* $\sum_{i \in \mathbb{N}_q} \lambda_i (v_i \otimes v_i)$*. Reorder the eigenvalues as* $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(q)} > 0$ *and let* $\left(v_{(1)}, \ldots v_{(q)}\right)$ *be the corresponding normalised eigenvectors. Let* $\rho_i(h, \Sigma) \coloneqq \mathrm{Corr}^2\left(g(Y), \langle v_i, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right)$*. Then, whenever* $i < j \leq q$*,* $\mathbb{P}\left(\rho_{(i)}(h, \Sigma) \geq \rho_{(j)}(h, \Sigma)\right) > 1/2$*.*

3.5.1.2 Unitarily invariant random functions

To establish the results using unitary invariance, the following lemma is needed.

Lemma 3.5.3. *Suppose that* v_1, v_2 *are random functions in* $\mathcal{G}$ *such that (1)* $\langle v_1, v_2\rangle_{\mathcal{G}} \overset{a.s.\mathbb{P}}{=} 0$ *and (2) for any unitary operator* $U : \mathcal{G} \to \mathcal{G}$*,* $(v_1, v_2) \overset{D}{=} (U(v_1), U(v_2))$*. Then, for any (nonrandom) function* $h \in \mathcal{G} \setminus \{0\}$*, the ratio* $\langle h, v_1\rangle_{\mathcal{G}} / \langle h, v_2\rangle_{\mathcal{G}}$ *has a standard Cauchy distribution.*

Theorem 3.5.4. *Suppose*

1. Σ *is a random operator with* $f \circ X$ *satisfying* $\mathbb{E}(f \circ X|\Sigma) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}(f \circ X|\Sigma) \overset{a.s.\mathbb{P}}{=} \Sigma$

2. *almost surely, each nonzero eigenvalue of* Σ *has multiplicity* 1

3. h *is a unitarily invariant random function satisfying* $h \perp\!\!\!\perp (X, \Sigma)$

4. $\mathbb{P}(h = 0) = 0$

5. $g(Y) \perp\!\!\!\perp X | \left(\langle h, f \circ X\rangle_{\mathcal{G}}, h, \Sigma\right)$

6. $\mathrm{Var}\left(g(Y)|h, \Sigma\right) < \infty$ *almost surely*

7. $\mathrm{Cov}\left(g(Y), \langle h, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right)$ *is nonzero almost surely*

8. *for any* $v \in \mathcal{G}$*, there exists a constant* c_v *such that*

$$\mathbb{E}\left(\langle v, f \circ X\rangle_{\mathcal{G}} \middle| h, \langle h, f \circ X\rangle_{\mathcal{G}}, \Sigma\right) \overset{a.s.\mathbb{P}}{=} c_v \langle h, f \circ X\rangle_{\mathcal{G}}.$$

Write Σ *as* $\sum_{i \in \mathbb{N}_q} \lambda_i (v_i \otimes v_i)$*. Reorder the eigenvalues as* $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(q)} > 0$ *and let* $\left(v_{(1)}, \ldots v_{(q)}\right)$ *be the corresponding normalised eigenvectors. Let* $\rho_i(h, \Sigma) \coloneqq \mathrm{Corr}^2\left(g(Y), \langle v_i, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right)$*. Then, whenever* $i < j \leq q$*,* $\mathbb{P}\left(\rho_{(i)}(h, \Sigma) \geq \rho_{(j)}(h, \Sigma)\right) = \frac{2}{\pi}\mathbb{E}\left(\arctan\left(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\right)\right)$.

3.5.1.3 Orientationally uniform random operators, unitarily invariant random operators, and random regular conditional distributions

The second more general level is now addressed. This is the general situation where X and Y are dependent, but the dependence is not restricted to any model.

While no model is assumed for the relation between X and Y, the following conditional independence is needed

$$g(Y) \perp\!\!\!\perp \Sigma | X. \tag{3.2}$$

This means that $g(Y)$ depends on X only through the value of X itself, and not its covariance operator.

Recall that, by Theorem 2.5.12, there exists a regular conditional distribution of $g(Y)|X$. Let $\mathcal{K}$ be the collection of all such mappings. For simplicity, it is assumed that $\mathcal{G}$ is rich enough to contain all bounded measurable functions of

X, so that, for each $\kappa \in \mathcal{K}$, and each $A \in \mathcal{B}(\mathbb{R})$, $\kappa(A, \cdot) \in \mathcal{G}$. Define a random element in $\mathcal{K}$ to be a mapping

$$\nu : \Omega \to \mathcal{K}, \;\; \omega \mapsto \nu_\omega(\cdot, \cdot),$$

such that, for each $A \in \mathcal{B}(\mathbb{R})$, the function $\Omega \to \mathcal{G}$, $\omega \mapsto \nu_\omega(A, \cdot)$ is measurable. The notation $g(Y)|(X, \nu) \sim \nu$ is used to indicate that a ν is chosen from $\mathcal{K}$ to be the regular conditional distribution of $Y|X$.

If, for each $A \in \mathcal{B}(\mathbb{R})$, $\kappa(A, X)$ is almost surely constant, then κ represents a regular conditional distribution under which X and $g(Y)$ are independent. Let $\mathcal{K}_0$ be the collection of all such κ. Since the tendency described in this section occurs only when X and $g(Y)$ are related in some way, the case of independence needs to be excluded from consideration. This is formulated with the condition $\mathbb{P}(\nu \in \mathcal{K}_0) = 0$.

Theorem 3.5.5. *Suppose*

1. Σ *is an orientationally uniform random operator with* $f \circ X$ *satisfying* $\mathbb{E}(f \circ X|\Sigma) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}(f \circ X|\Sigma) \overset{a.s.\mathbb{P}}{=} \Sigma$

2. ν *is a random element of* $\mathcal{K}$ *with* $\mathbb{P}(\nu \in \mathcal{K}_0) = 0$ *and which satisfies* $g(Y)|(X, \nu) \sim \nu$, $\nu \perp\!\!\!\perp (X, \Sigma)$, $g(Y) \perp\!\!\!\perp \Sigma|(X, \nu)$

3. *the random function* $m_\nu(\cdot) \coloneqq \int g \, \mathrm{d}\nu(\cdot, \omega)$ *almost surely belongs to* $\mathcal{G}$

4. $\mathrm{Var}(g(Y)|\nu, \Sigma) < \infty$ *almost surely*

5. *almost surely,* $\mathrm{Cov}(g(Y), m_\nu(X)|\nu, \Sigma)$ *is both nonzero and finite.*

6. $\mathbb{P}(m_\nu \in G) > 0$ *for any nonempty open set* $G \subseteq \mathcal{G}$

Write Σ *as* $\sum_{i \in \mathbb{N}_q} \lambda_i (v_i \otimes v_i)$. *Reorder the eigenvalues as* $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(q)} > 0$ *and let* $\left(v_{(1)}, \ldots v_{(q)}\right)$ *be the corresponding normalised eigenvectors. Let* $\rho_i(\nu, \Sigma)$ *be defined to be the squared conditional correlation*

$\mathrm{Corr}^2\left(g(Y), \langle v_i, f \circ X\rangle_{\mathcal{G}} \middle| \nu, \Sigma\right)$. *Then, for any* $i < j \leq q$,

$$\mathbb{P}\left(\rho_{(i)}(\nu, \Sigma) \geq \rho_{(j)}(\nu, \Sigma)\right) > \frac{1}{2}.$$

Although Theorem 3.5.5 subsumes Theorem 3.5.2, the latter is included as its proof makes use of Theorem 3.5.1 which is a significant result in its own right. Before giving this result, it is worth noting that a sufficient condition for Σ to be unitarily invariant is that each of the eigenvectors are unitarily invariant.

Theorem 3.5.6. *Suppose*

1. Σ *is a unitarily invariant random operator with* $f \circ X$ *satisfying* $\mathbb{E}\left(f \circ X | \Sigma\right) \overset{a.s.\mathbb{P}}{=} 0$ *and* $\mathrm{Var}\left(f \circ X | \Sigma\right) \overset{a.s.\mathbb{P}}{=} \Sigma$

2. *almost surely, each nonzero eigenvalue of* Σ *has multiplicity* 1

3. ν *is a random element of* $\mathcal{K}$ *with* $\mathbb{P}\left(\nu \in \mathcal{K}_0\right) = 0$ *and which satisfies* $g(Y)|(X, \nu) \sim \nu$, $\nu \perp\!\!\!\perp (X, \Sigma)$, $g(Y) \perp\!\!\!\perp \Sigma|(X, \nu)$

4. *the random function* $m_\nu(\cdot) \coloneqq \int g \, \mathrm{d}\nu\left(\cdot, \omega\right)$ *almost surely belongs to* $\mathcal{G}$

5. $\mathrm{Var}\left(g(Y)|\nu, \Sigma\right) < \infty$ *almost surely*

6. *almost surely,* $\mathrm{Cov}\left(g(Y), m_\nu(X)|\nu, \Sigma\right)$ *is both nonzero and finite.*

Write Σ *as* $\sum_{i \in \mathbb{N}_q} \lambda_i \left(v_i \otimes v_i\right)$. *Reorder the eigenvalues as* $\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(q)} > 0$ *and let* $\left(v_{(1)}, \ldots v_{(q)}\right)$ *be the corresponding normalised eigenvectors. Let* $\rho_i(\nu, \Sigma)$ *be defined to be the squared conditional correlation* $\mathrm{Corr}^2\left(g(Y), \langle v_i, f \circ X\rangle_{\mathcal{G}} \middle| \nu, \Sigma\right)$. *Then, for any* $i < j \leq q$,

$$\mathbb{P}\left(\rho_{(i)}(\nu, \Sigma) \geq \rho_{(j)}(\nu, \Sigma)\right) = \frac{2}{\pi}\mathbb{E}\left(\arctan\left(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\right)\right).$$

3.5.2 The infinite-dimensional case

The difficulty of extending the historical results to nonlinear principal components with general predictors where $\mathcal{G}$ may have infinite-dimension arises from the fact that orientationally uniform operators, unitarily invariant functions, and unitarily invariant operators cannot be defined on infinite-dimensional spaces. However, one can have these in any finite-dimensional subspace hence this section presents workarounds which allow similar results to those in Section 3.5.1 to be given for the infinite-dimensional setting.

The assumptions one can make are as follows.

Assumption 3.5.1. Let Σ be a random compact self-adjoint positive definite operator on G. This Σ is taken to be the covariance operator of $f \circ X$. It is assumed that there exists a finite subvector $V \coloneqq (x_1, \ldots, x_a)$ of $\mathbb{N}$ such that $\Sigma^* \coloneqq \sum_{i \in \mathbb{N}_a} \lambda_{x_i} \left[v_{x_i} \otimes v_{x_i} \right]$ is orientationally uniform.

Assumption 3.5.2. Let Σ be a random compact self-adjoint positive definite operator on G. This Σ is taken to be the covariance operator of $f \circ X$. It is assumed that there exists a finite subvector $V \coloneqq (x_1, \ldots, x_a)$ of $\mathbb{N}$ such that $\Sigma^* \coloneqq \sum_{i \in \mathbb{N}_a} \lambda_{x_i} \left[v_{x_i} \otimes v_{x_i} \right]$ is unitarily invariant.

Assumption 3.5.3. Let Σ be a random compact self-adjoint positive definite operator on $\mathcal{G}$. This Σ is taken to be the covariance operator of $f \circ X$. Let v_i be the i^{th} eigenvector of Σ. It is assumed that h is a random element of $\mathcal{G}$, with $h \perp\!\!\!\perp (X, \Sigma)$, such that there exists a finite subvector $V \coloneqq (x_1, \ldots, x_a)$ of $\mathbb{N}$ such that the sequence $\left(\left\langle h, v_{x_l} \right\rangle_{\mathcal{G}} \right)_{l \in \mathbb{N}_a}$ is spherically distributed.

With these assumptions replacing their counterparts, results analogous to those in the previous section can be obtained by having Σ^* replacing Σ, v_{x_i} and v_{x_j} replacing v_i and v_j respectively, λ_{x_i} and λ_{x_j} replacing λ_i and λ_j respectively,

and a replacing q. The proofs are essentially the same after these modifications are made, hence are omitted.

3.5.3 Summary

In this section, the predictive utility of nonlinear principal components in regression with general predictors was explored at two levels. The first is the conditional independence model and the second is an arbitrary X-$g(Y)$ relation. For both levels, it was demonstrated that the predictive tendency held by classical principal components remains valid.

3.5.4 Proofs

Proof of Theorem 3.5.1. First observe that

$$\begin{aligned}&\mathbb{E}\left(\langle v, f\circ X\rangle_{\mathcal{G}}\middle| h, \langle h, f\circ X\rangle_{\mathcal{G}}, \Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} c_v \langle h, f\circ X\rangle_{\mathcal{G}} \iff\\ &\left\langle v, \mathbb{E}\left(f\circ X\middle| h, \langle h, f\circ X\rangle_{\mathcal{G}}, \Sigma\right)\right\rangle_{\mathcal{G}} \overset{\text{a.s.}\mathbb{P}}{=} c_v \langle h, f\circ X\rangle_{\mathcal{G}}.\end{aligned}$$

To show the result, it suffices to show that, for any $A \in \sigma\left(h, \langle h, f\circ X\rangle_{\mathcal{G}}, \Sigma\right)$ and $v \in \mathcal{G}$,

$$\begin{aligned}&\mathbb{E}\left(\mathbf{1}_A \left\langle v, \mathbb{E}\left(f\circ X\middle| h, \langle h, f\circ X\rangle_{\mathcal{G}}, \Sigma\right)\right\rangle_{\mathcal{G}}\right)\\ &= \mathbb{E}\left(\mathbf{1}_A \left\langle v, \frac{1}{\langle h, \Sigma h\rangle_{\mathcal{G}}}\Sigma\left[h\otimes h\right](f\circ X)\right\rangle_{\mathcal{G}}\right)\end{aligned}$$

So let $v \in \mathcal{G}$ and let $A \in \sigma\left(h, \langle h, f\circ X\rangle_{\mathcal{G}}, \Sigma\right)$. By the definition of conditional expectation and the linearity assumption, there exists a constant c_v such that,

$$\begin{aligned}&\mathbb{E}\left(\mathbf{1}_A \left\langle v, \mathbb{E}\left(f\circ X\middle| h, \langle h, f\circ X\rangle_{\mathcal{G}}, \Sigma\right)\right\rangle_{\mathcal{G}}\right)\\ &= \mathbb{E}\left(\mathbf{1}_A c_v \langle h, f\circ X\rangle_{\mathcal{G}}\right) = \mathbb{E}\left(\mathbf{1}_A \langle v, f\circ X\rangle_{\mathcal{G}}\right)\end{aligned}$$

Consider

$$\begin{aligned}
&\mathbb{E}\left(\mathbf{1}_A \left\langle v, \frac{1}{\langle h, \Sigma h\rangle_{\mathcal{G}}} \Sigma \left[h \otimes h\right] (f \circ X)\right\rangle_{\mathcal{G}}\right) \\
&= \mathbb{E}\left(\frac{\mathbf{1}_A \left\langle v, \langle h, f \circ X\rangle_{\mathcal{G}} \Sigma h\right\rangle_{\mathcal{G}}}{\langle h, \Sigma h\rangle_{\mathcal{G}}}\right) \\
&= \mathbb{E}\left(\frac{\mathbf{1}_A \langle v, \Sigma h\rangle_{\mathcal{G}} \langle h, f \circ X\rangle_{\mathcal{G}}}{\langle h, \Sigma h\rangle_{\mathcal{G}}}\right) \qquad (3.3)
\end{aligned}$$

Now observe

$$\begin{aligned}
\langle v, \Sigma h\rangle_{\mathcal{G}} &\overset{\text{a.s.}\mathbb{P}}{=} \langle v, \mathbb{E}\left(\left[f \circ X\right] \otimes \left[f \circ X\right] | \Sigma\right) h\rangle_{\mathcal{G}} \\
&\overset{\text{a.s.}\mathbb{P}}{=} \langle v, \mathbb{E}\left(\left(\left[f \circ X\right] \otimes \left[f \circ X\right]\right) h | h, \Sigma\right)\rangle_{\mathcal{G}} \\
&\overset{\text{a.s.}\mathbb{P}}{=} \left\langle v, \mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}} \left(f \circ X\right) \middle| h, \Sigma\right)\right\rangle_{\mathcal{G}} \\
&\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\left\langle v, \langle h, f \circ X\rangle_{\mathcal{G}} \left(f \circ X\right)\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}} \langle v, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right)
\end{aligned}$$

By a similar argument, $\langle h, \Sigma h\rangle_{\mathcal{G}} \overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}}^2 \middle| h, \Sigma\right)$. So Equation (3.3) can be rewritten as

$$\begin{aligned}
&\mathbb{E}\left(\frac{\mathbf{1}_A \mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}} \langle v, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \langle h, f \circ X\rangle_{\mathcal{G}}}{\mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}}^2 \middle| h, \Sigma\right)}\right) \\
&= \mathbb{E}\left(\frac{\mathbb{E}\left(\mathbf{1}_A \middle| h, \langle h, f \circ X\rangle_{\mathcal{G}}, \Sigma\right) \mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}} \langle v, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \langle h, f \circ X\rangle_{\mathcal{G}}}{\mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}}^2 \middle| h, \Sigma\right)}\right) \\
&= \mathbb{E}\left(\frac{\mathbf{1}_A \langle h, f \circ X\rangle_{\mathcal{G}} \mathbb{E}\left(\mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}} \langle v, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \middle| h, \langle h, f \circ X\rangle_{\mathcal{G}}, \Sigma\right)}{\mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}}^2 \middle| h, \Sigma\right)}\right) \\
&= \mathbb{E}\left(\frac{\mathbf{1}_A \langle h, f \circ X\rangle_{\mathcal{G}} \mathbb{E}\left(\mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}} \langle v, f \circ X\rangle_{\mathcal{G}} \middle| h, \langle h, f \circ X\rangle_{\mathcal{G}}, \Sigma\right) \middle| h, \Sigma\right)}{\mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}}^2 \middle| h, \Sigma\right)}\right) \\
&= \mathbb{E}\left(\frac{\mathbf{1}_A \langle h, f \circ X\rangle_{\mathcal{G}} \mathbb{E}\left(c_v \langle h, f \circ X\rangle_{\mathcal{G}}^2 \middle| h, \Sigma\right)}{\mathbb{E}\left(\langle h, f \circ X\rangle_{\mathcal{G}}^2 \middle| h, \Sigma\right)}\right)
\end{aligned}$$

$$= \mathbb{E}\left(\mathbf{1}_A c_v \langle h, f \circ X\rangle_{\mathcal{G}}\right)$$

This completes the proof. □

Proof of Theorem 3.5.2. By the definition of conditional correlation,

$$\operatorname{Corr}^2\left(g(Y), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} \frac{\operatorname{Cov}^2\left(g(Y), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right)}{\operatorname{Var}\left(g(Y)|h, \Sigma\right) \operatorname{Var}\left(\left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right)}$$

Now as $h \perp\!\!\!\perp (X, \Sigma)$, $\operatorname{Var}\left(\left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Var}\left(\left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| \Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} \lambda_{(i)}$. Consider now

$$\begin{aligned}
&\operatorname{Cov}\left(g(Y), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(g(Y), \mathbb{E}\left(\left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma, X\right) \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(\mathbb{E}\left(g(Y)|h, \Sigma, X\right), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(\mathbb{E}\left(g(Y) \middle| h, \Sigma, \langle h, f \circ X\rangle_{\mathcal{G}}\right), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(g(Y), \mathbb{E}\left(\left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma, \langle h, f \circ X\rangle_{\mathcal{G}}\right) \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(g(Y), \left\langle v_{(i)}, \mathbb{E}\left(f \circ X \middle| h, \Sigma, \langle h, f \circ X\rangle_{\mathcal{G}}\right)\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(g(Y), \left\langle v_{(i)}, \frac{1}{\langle h, \Sigma h\rangle_{\mathcal{G}}} \Sigma\left[h \otimes h\right](f \circ X)\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \frac{1}{\langle h, \Sigma h\rangle_{\mathcal{G}}} \operatorname{Cov}\left(g(Y), \left\langle \Sigma v_{(i)}, \left[h \otimes h\right](f \circ X)\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \frac{\lambda_{(i)}\left\langle v_{(i)}, h\right\rangle_{\mathcal{G}}}{\langle h, \Sigma h\rangle_{\mathcal{G}}} \operatorname{Cov}\left(g(Y), \langle h, f \circ X\rangle_{\mathcal{G}} \middle| h, \Sigma\right)
\end{aligned}$$

After substituting this into the above relation for conditional correlation, doing a similar analysis with i replaced by j, and applying some elementary algebra, this implies that

$$\frac{\operatorname{Corr}^2\left(g(Y), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right)}{\operatorname{Corr}^2\left(g(Y), \left\langle v_{(j)}, f \circ X\right\rangle_{\mathcal{G}} \middle| h, \Sigma\right)} \overset{\text{a.s.}\mathbb{P}}{=} \frac{\lambda_{(i)}\left\langle v_{(i)}, h\right\rangle_{\mathcal{G}}^2}{\lambda_{(j)}\left\langle v_{(j)}, h\right\rangle_{\mathcal{G}}^2}.$$

Therefore

$$\mathbb{P}\left(\rho_{(i)}(h,\Sigma) \geq \rho_{(j)}(h,\Sigma)\right) = \mathbb{P}\left(\frac{\lambda_{(i)}\left\langle v_{(i)}, h\right\rangle_{\mathcal{G}}^2}{\lambda_{(j)}\left\langle v_{(j)}, h\right\rangle_{\mathcal{G}}^2} \geq 1\right) = \mathbb{P}\left(\frac{\left\langle v_{(j)}, h\right\rangle_{\mathcal{G}}^2}{\left\langle v_{(i)}, h\right\rangle_{\mathcal{G}}^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}}\right). \tag{3.4}$$

By the law of total probability, the final expression in Equation (3.4) can be rewritten as

$$\sum_{k\neq l} \mathbb{P}\left(\frac{\left\langle v_{(j)}, h\right\rangle_{\mathcal{G}}^2}{\left\langle v_{(i)}, h\right\rangle_{\mathcal{G}}^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}}\middle| \lambda_{(i)} = \lambda_k, \lambda_{(j)} = \lambda_l\right) \mathbb{P}\left(\lambda_{(i)} = \lambda_k, \lambda_{(j)} = \lambda_l\right).$$

Each of the unconditional probabilities in this summation is equal to $\binom{q}{2}^{-1}$. Hence, after seeing that $\left(v_{(i)}, v_{(j)}\right) = (v_k, v_l)$ when the event $\left(\lambda_{(i)} = \lambda_k, \lambda_{(j)} = \lambda_l\right)$ is conditioned on, this summation can be rewritten as

$$\binom{q}{2}^{-1} \sum_{k\neq l} \mathbb{P}\left(\frac{\langle v_l, h\rangle_{\mathcal{G}}^2}{\langle v_k, h\rangle_{\mathcal{G}}^2} \leq \frac{\lambda_k}{\lambda_l}\middle| \lambda_{(i)} = \lambda_k, \lambda_{(j)} = \lambda_l\right). \tag{3.5}$$

Reexpress each term in this summation as

$$\mathbb{E}\left(\mathbb{P}\left(\frac{\langle v_l, h\rangle_{\mathcal{G}}^2}{\langle v_k, h\rangle_{\mathcal{G}}^2} \leq \frac{\lambda_k}{\lambda_l}\middle| \lambda_{(i)} = \lambda_k, \lambda_{(j)} = \lambda_l, \lambda_k, \lambda_l\right)\middle| \lambda_{(i)} = \lambda_k, \lambda_{(j)} = \lambda_l\right). \tag{3.6}$$

By the definition of an orientationally uniform random operator,

$$\begin{aligned}
&(v_k, v_l) \perp\!\!\!\perp \left(\lambda_1, \ldots, \lambda_q\right) \\
&\implies (v_k, v_l) \perp\!\!\!\perp \left(\lambda_1, \ldots, \lambda_q, \lambda_{(1)}, \ldots, \lambda_{(q)}\right) \\
&\implies (v_k, v_l) \perp\!\!\!\perp \left(\lambda_k, \lambda_l, \lambda_{(i)}, \lambda_{(j)}\right) \\
&\implies (v_k, v_l) \perp\!\!\!\perp \left(\lambda_{(i)}, \lambda_{(j)}\right) \mid (\lambda_k, \lambda_l).
\end{aligned}$$

Thus the expression in 3.6 can be rewritten as

$$\mathbb{E}\left(\mathbb{P}\left(\frac{\langle v_l, h\rangle_{\mathcal{G}}^2}{\langle v_k, h\rangle_{\mathcal{G}}^2} \leq \frac{\lambda_k}{\lambda_l}\middle| \lambda_k, \lambda_l\right)\middle| \lambda_{(i)} = \lambda_k, \lambda_{(j)} = \lambda_l\right). \tag{3.7}$$

As $(h, v_k, v_l) \perp\!\!\!\perp (\lambda_k, \lambda_l)$, it holds that for $s > t > 0$,

$$\mathbb{P}\left(\frac{\langle v_l, h\rangle_{\mathcal{G}}^2}{\langle v_k, h\rangle_{\mathcal{G}}^2} \leq \frac{s}{t}\middle|\lambda_k = s, \lambda_l = t\right) = \mathbb{P}\left(\frac{\langle v_l, h\rangle_{\mathcal{G}}^2}{\langle v_k, h\rangle_{\mathcal{G}}^2} \leq \frac{s}{t}\right) > \frac{1}{2}.$$

where the inequality follows from Theorem 2.2.11. As the event $\lambda_k = \lambda_l$ has zero probability, it follows that the expression in 3.7, and hence 3.5, exceeds $1/2$. □

Proof of Lemma 3.5.3. Since U^{-1} is also a unitary operator,

$$(v_1, v_2) \overset{D}{=} (U^{-1}(v_1), U^{-1}(v_2)).$$

Consequently,

$$\begin{aligned}\left(\langle h, v_1\rangle_{\mathcal{G}}, \langle h, v_2\rangle_{\mathcal{G}}\right) &\overset{D}{=} \left(\left\langle h, U^{-1}(v_1)\right\rangle_{\mathcal{G}}, \left\langle h, U^{-1}(v_2)\right\rangle_{\mathcal{G}}\right) \\ &= \left(\langle U(h), v_1\rangle_{\mathcal{G}}, \langle U(h), v_2\rangle_{\mathcal{G}}\right).\end{aligned}$$

Thus, the distribution of $\left(\langle h, v_1\rangle_{\mathcal{G}}, \langle h, v_2\rangle_{\mathcal{G}}\right)$ depends on h only through $\|h\|_{\mathcal{G}} = a > 0$. Let $\tilde{h}$ be a $\mathcal{G}$-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ which is independent of (v_1, v_2) and uniformly distributed on the sphere $\mathcal{S}(a) = \left\{l \in \mathcal{G} : \|l\|_{\mathcal{G}} = a\right\}$. Then, for any $A \in \mathcal{B}(\mathbb{R})$, and any nonrandom function h_0,

$$\mathbb{P}\left(\left\langle \tilde{h}, v_1\right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2\right\rangle_{\mathcal{G}} \in A \middle| \tilde{h} = h_0\right) = \mathbb{P}\left(\langle h_0, v_1\rangle_{\mathcal{G}} / \langle h_0, v_2\rangle_{\mathcal{G}} \in A\right). \tag{3.8}$$

This implies

$$\mathbb{P}\left(\left\langle \tilde{h}, v_1\right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2\right\rangle_{\mathcal{G}} \in A \middle| \tilde{h}\right) = \mathbb{P}\left(\left\langle \tilde{h}, v_1\right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2\right\rangle_{\mathcal{G}} \in A\right) \tag{3.9}$$

The right hand side can be rewritten as

$$\mathbb{E}\left(\mathbb{P}\left(\left\langle \tilde{h}, v_1\right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2\right\rangle_{\mathcal{G}} \in A \middle| v_1, v_2\right)\right)$$

Because $\tilde{h} \perp\!\!\!\perp (v_1, v_2)$, $\tilde{h}$ is unitarily invariant when conditioning on (v_1, v_2). Then, by Theorem 2.2.15 and Theorem 1 of Arnold and Brockett (1992), the ratio

$\left\langle \tilde{h}, v_1 \right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2 \right\rangle_{\mathcal{G}}$ has a standard Cauchy distribution regardless of the value of (v_1, v_2). This means that the ratio $\left\langle \tilde{h}, v_1 \right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2 \right\rangle_{\mathcal{G}}$ is independent of (v_1, v_2), and therefore has a standard Cauchy distribution unconditionally. Hence

$$\mathbb{P}\left(\left\langle \tilde{h}, v_1 \right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2 \right\rangle_{\mathcal{G}} \in A \right) = \mathbb{P}_C(A),$$

where $\mathbb{P}_C(A)$ is the probability of A under the standard Cauchy distribution. However, by Equation (3.8) and Equation (3.9) equalities and the discussion preceding them,

$$\begin{aligned}
&\mathbb{P}\left(\langle h, v_1 \rangle_{\mathcal{G}} / \langle h, v_2 \rangle_{\mathcal{G}} \in A \right) \\
&= \mathbb{P}\left(\langle h_0, v_1 \rangle_{\mathcal{G}} / \langle h_0, v_2 \rangle_{\mathcal{G}} \in A \right) \\
&= \mathbb{P}\left(\left\langle \tilde{h}, v_1 \right\rangle_{\mathcal{G}} / \left\langle \tilde{h}, v_2 \right\rangle_{\mathcal{G}} \in A \right) = \mathbb{P}_C(A).
\end{aligned}$$

That is, $\langle h, v_1 \rangle_{\mathcal{G}} / \langle h, v_2 \rangle_{\mathcal{G}}$ has a standard Cauchy distribution, which proves the result. □

Proof of Theorem 3.5.4. The first part of the proof is identical to that for Theorem 3.5.2 up to Equation (3.4). Note now that

$$\mathbb{P}\left(\frac{\left\langle v_{(j)}, h \right\rangle_{\mathcal{G}}^2}{\left\langle v_{(i)}, h \right\rangle_{\mathcal{G}}^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}} \right) = \mathbb{P}\left(-\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}} \leq \frac{\left\langle v_{(j)}, h \right\rangle_{\mathcal{G}}}{\left\langle v_{(i)}, h \right\rangle_{\mathcal{G}}} \leq \sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}} \right)$$

By Theorem 1 of Arnold and Brockett (1992), $\left\langle v_{(j)}, h \right\rangle_{\mathcal{G}} / \left\langle v_{(i)}, h \right\rangle_{\mathcal{G}}$ has a standard Cauchy distribution. Hence, for any $s > t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(-\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}} \leq \frac{\left\langle v_{(j)}, h \right\rangle_{\mathcal{G}}}{\left\langle v_{(i)}, h \right\rangle_{\mathcal{G}}} \leq \sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}} \middle| \lambda_{(i)} = s, \lambda_{(j)} = t \right) \\
&= \mathbb{P}\left(-\sqrt{\frac{s}{t}} \leq \frac{\left\langle v_{(j)}, h \right\rangle_{\mathcal{G}}}{\left\langle v_{(i)}, h \right\rangle_{\mathcal{G}}} \leq \sqrt{\frac{s}{t}} \right) \\
&= \frac{2}{\pi} \arctan\left(\sqrt{\frac{s}{t}} \right)
\end{aligned}$$

As the event $\lambda_{(j)} = \lambda_{(i)}$ has probability zero, it follows that

$$\mathbb{P}\left(-\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}} \leq \frac{\left\langle v_{(j)}, h\right\rangle_{\mathcal{G}}}{\left\langle v_{(i)}, h\right\rangle_{\mathcal{G}}} \leq \sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\middle|\lambda_{(i)}, \lambda_{(j)}\right) \overset{\text{a.s.}\mathbb{P}}{=} \frac{2}{\pi}\arctan\left(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\right)$$

Taking the expectation of both sides gives the result. □

Proof of Theorem 3.5.5. Note that

$$\operatorname{Cov}\left(g(Y), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\nu, \Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(\mathbb{E}\left(g(Y)|\nu, \Sigma, X\right), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\nu, \Sigma\right)$$

Since $g(Y) \perp\!\!\!\perp \Sigma|(X, \nu)$,

$$\mathbb{E}\left(g(Y)|\nu, \Sigma, X\right) \overset{\text{a.s.}\mathbb{P}}{=} \mathbb{E}\left(g(Y)|\nu, X\right) \overset{\text{a.s.}\mathbb{P}}{=} m_\nu(X).$$

Since $\nu \perp\!\!\!\perp (X, \Sigma)$, $m_\nu \perp\!\!\!\perp (X, \Sigma)$. Hence, for any $\kappa \in \mathcal{K}$,

$$\begin{aligned}\operatorname{Cov}\left(m_\nu(X), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\nu = \kappa, \Sigma\right) &\overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Cov}\left(m_\kappa(X), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\Sigma\right)\\ &\overset{\text{a.s.}\mathbb{P}}{=} \left\langle m_\kappa, \Sigma v_{(i)}\right\rangle_{\mathcal{G}}\\ &\overset{\text{a.s.}\mathbb{P}}{=} \lambda_{(i)}\left\langle m_\kappa, v_{(i)}\right\rangle_{\mathcal{G}}.\end{aligned}$$

This implies

$$\operatorname{Cov}\left(m_\nu(X), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\nu, \Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} \lambda_{(i)}\left\langle m_\nu, v_{(i)}\right\rangle_{\mathcal{G}}.$$

Similarly, by $\nu \perp\!\!\!\perp (X, \Sigma)$

$$\operatorname{Var}\left(\left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\nu, \Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} \operatorname{Var}\left(\left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\Sigma\right) \overset{\text{a.s.}\mathbb{P}}{=} \lambda_{(i)}.$$

It follows that

$$\frac{\operatorname{Corr}^2\left(g(Y), \left\langle v_{(i)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\nu, \Sigma\right)}{\operatorname{Corr}^2\left(g(Y), \left\langle v_{(j)}, f \circ X\right\rangle_{\mathcal{G}}\middle|\nu, \Sigma\right)} \overset{\text{a.s.}\mathbb{P}}{=} \frac{\lambda_{(i)}\left\langle m_\nu, v_{(i)}\right\rangle_{\mathcal{G}}^2}{\lambda_{(j)}\left\langle m_\nu, v_{(j)}\right\rangle_{\mathcal{G}}^2}.$$

Since $m_\nu \perp\!\!\!\perp (v_{(i)}, v_{(j)}, \lambda_{(i)}, \lambda_{(j)})$, it follows that $m_\nu \perp\!\!\!\perp (v_{(i)}, v_{(j)})|(\lambda_{(i)}, \lambda_{(j)})$. Hence, for any $\kappa \in \mathcal{K}$,

$$\begin{aligned}
&\mathbb{P}\left(\left(\frac{\langle m_\nu, v_{(j)}\rangle_{\mathcal{G}}}{\langle m_\nu, v_{(i)}\rangle_{\mathcal{G}}}\right)^2 < \frac{\lambda_{(i)}}{\lambda_{(j)}}\middle| \nu = \kappa, \lambda_{(i)}, \lambda_{(j)}\right) \\
&\overset{\text{a.s.}\mathbb{P}}{=} \mathbb{P}\left(\left(\frac{\langle m_\kappa, v_{(j)}\rangle_{\mathcal{G}}}{\langle m_\kappa, v_{(i)}\rangle_{\mathcal{G}}}\right)^2 < \frac{\lambda_{(i)}}{\lambda_{(j)}}\middle| \lambda_{(i)}, \lambda_{(j)}\right)
\end{aligned}$$

By an argument similar to that for the proof of Theorem 3.5.2, the right hand side almost surely exceeds $1/2$. Taking the expectation on both sides of the above equality completes the proof. □

Proof of Theorem 3.5.6. The proof is similar to that for Theorem 3.5.5, except that the last paragraph is replaced with saying that the probability preceeding it is almost surely equal to $(2/\pi)\arctan\left(\left(\lambda_{(i)}/\lambda_{(j)}\right)^{1/2}\right)$. Thus it has been shown that

$$\mathbb{P}\left(\left(\frac{\langle m_\nu, v_{(j)}\rangle_{\mathcal{G}}}{\langle m_\nu, v_{(i)}\rangle_{\mathcal{G}}}\right)^2 < \frac{\lambda_{(i)}}{\lambda_{(j)}}\middle| \nu, \lambda_{(i)}, \lambda_{(j)}\right) \overset{\text{a.s.}\mathbb{P}}{=} \frac{2}{\pi}\arctan\left(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\right).$$

Taking the expectation on both sides of the above equality completes the proof. □

Chapter 4

Methodological developments in sufficient dimension reduction

4.1 Outline of chapter

In this chapter, methodological developments in sufficient dimension reduction are given. First, an overview of this supervised framework for dimension reduction is provided. A brief presentation of two commonly used procedures for linear sufficient dimension reduction is given for context. The interest in this thesis is the scenario where some of the predictors are categorical, so an account of the existing literature for this situation is given in the following. The main methodological developments are then provided; in brief, a nonlinear approach for sufficient dimension reduction with some categorical predictors is developed.

4.2 Overview of sufficient dimension reduction

This thesis has so far focussed on principal components analysis, an early unsupervised dimension reduction method, and explored its predictive utility for

regression. As the discussion indicated (because the results in Chapter 3 are of a probabilistic nature), it is desirable that dimension reduction procedures take the response variable into account. The discussion is thus turned to a supervised framework, known as *sufficient dimension reduction* (SDR), which combines dimension reduction with notions from classical statistics, particularly that of a sufficient statistic (see Fisher (1922)).

Classically, SDR assumes that the predictor X is a random vector in $\mathbb{R}^p$, for some $p \in \mathbb{N}$, and the response Y is a random variable in $\mathbb{R}$. In this setting, the idea is to find a d-dimensional subspace $\mathcal{S}$, called an *SDR subspace*, of $\mathbb{R}^p$ ($d \leq p$) on which to project X such that the projection $P_{\mathcal{S}}X$ of X onto $\mathcal{S}$ retains the predictive power that X has for Y. This preservation is characterised via the conditional independence $Y \perp\!\!\!\perp X|P_{\mathcal{S}}X$. There can be many such subspaces, so SDR seeks those with minimal d. Under some mild assumptions, Cook (1998) showed that the intersection of all SDR subspaces, denoted by $\mathcal{S}_{Y|X}$, is itself an SDR subspace. $\mathcal{S}_{Y|X}$ is called the *central subspace* and its dimension is called the *structural dimension*. As it is the intersection of all SDR subspaces, the central subspace is unique. Some of the key methods and results for this setting are given in: Li (1991), Cox (1968), Li et al. (2005), Cook (1998), Cook (1994), Cook and Ni (2005), and Cook and Forzani (2009). See Li (2018) for a comprehensive presentation of the most commonly used procedures for linear SDR, two of which are discussed in Section 4.3.

The assumption that the extracted manifold in SDR is a linear subspace of $\mathbb{R}^p$ has been relaxed as the field has evolved. Most of the approaches for finding nonlinear manifolds have been developed by applying the "kernel trick", most well-known for its application to the support vector machine of Cortes and Vapnik (1995), to extend the classical methods. This is done by allowing for nonlinear projections of X which are constrained to be functions of X in some reproducing

kernel Hilbert space (see Aronszajn (1950)). This technique relies on the fact that many of the methods only require the computation of inner products. This nonlinear approach allows for even greater reduction of dimension. Some of the key literature for this setting are: Fukumizu et al. (2004), Li et al. (2011), Artemiou and Dong (2016), and Yeh et al. (2009).

Many of the methods developed for the setting of multivariate data have been broadened to situations where the predictor is a random function or, more generally, a random variable in some real separable Hilbert space $\mathcal{H}$. These extensions are achieved by replacing the $\mathbb{R}^p$ inner product with the inner product in $\mathcal{H}$. Some alterations are made in estimation procedures as the possible infinite-dimensionality necessitates careful consideration regarding the nature of various operators particularly relating to their boundedness, compactness, and if they are trace-class. For an overview of the field of functional data analysis, see Hsing and Eubank (2015). Some methods for this setting are given in: Ferré and Yao (2003), Wang et al. (2013), Lian and Li (2014), and Wang et al. (2015).

A significant development in SDR has arisen by exploiting its similarities (see Li and Song (2017), Li (2018), and Lee et al. (2013)) with the measure-theoretic formulation of sufficient statistics. This theory, referred to in this thesis as *generalised SDR*, allows: (1) the response to be categorical or multivariate, (2) the predictor or the response to be random functions, and (3) the extracted manifold to be nonlinear. Generalised SDR is characterised in terms of sub-σ-fields of the σ-field generated by X ($\sigma(X)$). Any sub-σ-field G of $\sigma(X)$ which satisfies the conditional independence relation $Y \perp\!\!\!\perp X|G$ is called a *SDR σ-field*. Under a mild assumption, given in Lee et al. (2013), the intersection of all SDR σ-fields is itself a SDR σ-field known as the *central subfield*. This theory is built on in Section 4.5 to develop a nonlinear approach to SDR when some of the predictors are categorical.

4.3 Literature review: two commonly used methods for linear sufficient dimension reduction

In this section, X is a p-dimensional random vector and Y is a real random variable. The setting is that of linear sufficient dimension reduction as described in the previous section.

4.3.1 Sliced inverse regression

The earliest method for linear SDR, sliced inverse regression (SIR), was introduced by Li (1991). It relies on the following linearity assumption.

Assumption 4.3.1. Let $\beta \in \mathbb{R}^{p\times d}$ be such that the columnspace of β coincides with the central subspace. Assume that $\mathbb{E}\left(X\middle|\beta^T X\right)$ is a linear function of $\beta^T X$.

This assumption is known to hold when X has an elliptical distribution (see Eaton (1986)). With this assumption, the following theorem gives that the SIR estimator is unbiased.

Theorem 4.3.1 (SIR is unbiased, see Li (2018))**.** *Suppose* $\Sigma = \mathrm{Var}\,(X)$ *exists and is nonsingular. Then*

$$\Sigma^{-1}\left[\mathbb{E}\,(X|Y) - \mathbb{E}\,(X)\right] \in \mathcal{S}_{Y|X}.$$

Note that this is how the result is presented in Li (2018) – the author of this thesis believes that the fact that $\mathbb{E}\,(X|Y)$ is random means there should be a clarification that the result holds almost surely. This point holds throughout this section.

The term “inverse regression" comes from the fact that it is the conditional expectation of X given Y which is of interest, rather than the other way around. Notably, this is easier to estimate as Y is univariate. The term “sliced" comes

from the way that the inverse regression curve is estimated. In the following, the span of a matrix is defined to be its columnspace (to use the same notation as Li (2018)).

Corollary 4.3.2 (A property of the SIR estimator, see Li (2018))**.** *Under the assumptions of the previous theorem*

$$\text{Span}\left\{\Sigma^{-1}\,\text{Var}\left(\mathbb{E}\left(X|Y\right)\right)\Sigma^{-1}\right\} \subseteq \mathcal{S}_{Y|X}$$

Let $\Lambda_{\text{SIR}} \coloneqq \text{Var}\left(\mathbb{E}\left(X|Y\right)\right)$. The above corollary implies that the columnspace (denoted $\mathcal{S}_{\text{SIR}}$) of $\Sigma^{-1}\Lambda_{\text{SIR}}\Sigma^{-1}$ can be used to estimate at least part of the central subspace. Before examining how to estimate this space, note that Li (2018) gives an example to show that, in general, SIR is unable to recover the entirety of the central subspace. More problematically, a regression function (in a multi-index model) which is symmetric about the origin is only able to recover the zero vector which is of no use for dimension reduction.

The target columnspace can be found by solving the following optimisation problem. Let α_1 solve

$$\begin{gathered}\text{Max } \alpha^T\Lambda_{\text{SIR}}\alpha \\ \text{Subject to } \alpha^T\Sigma\alpha = 1.\end{gathered} \tag{4.1}$$

For $k = 2, \ldots, p$, let α_k be the solution to 4.1 subject to the extra condition

$$\alpha^T\Sigma\alpha_l = 0 \text{ for } l = 1, \ldots, k-1$$

This can be converted into an eigenvalue problem by the substitution $\beta \coloneqq \Sigma^{1/2}\alpha$. This gives that the SIR directions are given by the vectors $\Sigma^{-1/2}u_1, \ldots, \Sigma^{-1/2}u_r$ where r is the rank of $\Sigma^{-1/2}\Lambda_{\text{SIR}}\Sigma^{-1/2}$ and u_i is the i^{th} most dominant normalised eigenvector of that matrix.

Now the sample level estimation procedure is outlined. Let (X_1, Y_1), …, (X_n, Y_n) be an i.i.d sample of (X, Y) with sample size $n \in \mathbb{N}$. The SIR estimation procedure is as follows.

1. Compute $\hat{\mu} \coloneqq \frac{1}{n} \sum_{i \in \mathbb{N}_n} X_i$ and $\hat{\Sigma} \coloneqq \frac{1}{n} \sum_{i \in \mathbb{N}_n} (X_i - \hat{\mu}) (X_i - \hat{\mu})^T$.

2. Compute the standardised random vectors $Z_i \coloneqq \hat{\Sigma}^{-1/2} (X_i - \hat{\mu})$ for each $i \in \mathbb{N}_n$

3. Let $h \in \mathbb{N}$ and let $J_1 \coloneqq [\min_{i \in \mathbb{N}_n} Y_i, j_1)$, $J_2 \coloneqq [j_1, j_2)$, $\ldots$, $J_h \coloneqq [j_{h-1}, \max_{i \in \mathbb{N}_n} Y_i]$ for some $\min_{i \in \mathbb{N}_n} Y_i < j_1 < j_2 < \ldots < j_{h-1} < \max_{i \in \mathbb{N}_n} Y_i$. In practice, the intervals (known as "slices" in the literature, hence the name) are chosen to have approximately the same number of observations taking values in them.

4. For $i \in \mathbb{N}_h$, estimate $\mathbb{E}(Z|Y \in J_i)$ by
$$\mathbb{E}_n (Z|Y \in J_i) \coloneqq \frac{1}{p_i} \left(\frac{1}{n} \sum_{j \in \mathbb{N}_n} Z_j \mathbf{1}_{Y \in J_i} \right)$$
where p_i is the empirical estimate of $\mathbb{P}(Y \in J_i)$.

5. Estimate $\text{Var}(\mathbb{E}(Z|Y))$ by
$$\hat{\Lambda} \coloneqq \sum_{i \in \mathbb{N}_h} p_i \mathbb{E}_n (Z|Y \in J_i) \, \mathbb{E}_n \left(Z^T \,\middle|\, Y \in J_i \right).$$

6. Let $\hat{v}_1$, $\ldots$, $\hat{v}_r$ be the first r most dominant eigenvectors of $\hat{\Lambda}$ and let $\hat{\beta}_k \coloneqq \hat{\Sigma}^{-1/2} \hat{v}_k$ ($k \in \mathbb{N}_r$). The sufficient predictors are $\hat{\beta}_k^T (X_1 - \hat{\mu})$, $\ldots$, $\hat{\beta}_k^T (X_n - \hat{\mu})$.

Determining the number of components r to take is an ongoing problem, though there are some methods in the literature: see Bura and Yang (2011) for a unified approach to dimension estimation.

4.3.2 Sliced average variance estimation

To overcome the limitations of SIR, Cook and Weisberg (1991) introduced *sliced average variance estimation* (SAVE). This is a second-order conditional moment

method, which is able to capture a subspace which contains the SIR subspace and is itself a subspace of the central subspace (possibly, and ideally, equal to it). It relies on Assumption 4.3.2 as well as on Assumption 4.3.1, so has stronger conditions than just SIR.

Assumption 4.3.2. Let β be as in Assumption 4.3.1. It is furthermore assumed that $\mathrm{Var}\left(X\middle|\beta^T X\right)$ is non-random.

It is known (see Proposition 5.1 of Li (2018)) that Assumption 4.3.2 holds when X has a multivariate Gaussian distribution and a nonsingular covariance matrix.

Corollary 4.3.3 (Corollary 5.1 of Li (2018))**.** *If the random vector X satisfies Assumption 4.3.1 and Assumption 4.3.2, then*

$$\mathrm{Var}\left(X\middle|\beta^T X\right) \overset{a.s.\mathbb{P}}{=} \Sigma\left(I_{d\times d} - \beta\left(\beta^T\Sigma\beta\right)^{-1}\beta^T\Sigma\right)$$

Theorem 4.3.4 (Theorem 5.1 of Li (2018))**.** *Let β be as in Assumption 4.3.1 and suppose that assumption holds as well as Assumption 4.3.2, Then*

$$\mathrm{Span}\left\{\Sigma - \mathrm{Var}\left(X|Y\right)\right\} \subseteq \Sigma\mathcal{S}_{Y|X}$$

where $\Sigma\mathcal{S}_{Y|X} \coloneqq \left\{\Sigma v : v \in \mathcal{S}_{Y|X}\right\}$.

This theorem is applied to the standardised random vector $Z \coloneqq \Sigma^{-1/2}\left(X - \mu\right)$ to obtain that, under the same assumptions,

$$\mathrm{Span}\left\{I_{p\times p} - \mathrm{Var}\left(Z|Y\right)\right\} \subseteq \mathcal{S}_{Y|Z}.$$

This result, combined with an application of Proposition 5.2 of Li (2018), gives that

$$\mathrm{Span}\left\{\mathbb{E}\left(\left(I_{p\times p} - \mathrm{Var}\left(Z|Y\right)\right)^2\right)\right\} \subseteq \mathcal{S}_{Y|Z}.$$

Let the expectation inside the span be denoted by Λ_{SAVE}. By this observation and Theorem 2.2 of Li (2018) (which relates $\mathcal{S}_{Y|X}$ wth $\mathcal{S}_{Y|Z}$), it holds that

$$\Sigma^{-1/2} \operatorname{Span}\{\Lambda_{\text{SAVE}}\} \subseteq \mathcal{S}_{Y|X},$$

where $\Sigma^{-1/2} \operatorname{Span}\{\Lambda_{\text{SAVE}}\} \coloneqq \left\{\Sigma^{-1/2} v : v \in \operatorname{Span}\{\Lambda_{\text{SAVE}}\}\right\}$. Thus if r is the rank of Λ_{SAVE}, then

$$\operatorname{Span}\left\{\Sigma^{-1/2} v_1, \ldots, \Sigma^{-1/2} v_r\right\} \subseteq \mathcal{S}_{Y|X},$$

where v_i is the i^{th} most dominant eigenvector of Λ_{SAVE}. This is the basis for the SAVE estimation procedure, which is now described.

As in SIR, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d sample of (X, Y) with sample size $n \in \mathbb{N}$. The SAVE estimation procedure is as follows.

1. Perform the first three steps of the SIR estimation procedure to obtain $\hat{\mu}$, $\hat{\Sigma}$, Z_i $(i \in \mathbb{N}_n)$, $h \in \mathbb{N}$, and J_i $(i \in \mathbb{N}_h)$ as in that algorithm.

2. For $i \in \mathbb{N}_h$, compute

$$\operatorname{Var}_n(Z|Y \in J_i) \coloneqq \frac{1}{p_i}\left(\frac{1}{n}\sum_{j \in \mathbb{N}_n} Z_j Z_j^T \mathbf{1}_{Y \in J_i}\right),$$

   where p_i is the empirical estimate of $\mathbb{P}(Y \in J_i)$.

3. Calculate the sample version of Λ_{SAVE} as

$$\hat{\Lambda}_{\text{SAVE}} \coloneqq \frac{1}{h}\sum_{i \in \mathbb{N}_h} p_i \left(I_{p\times p} - \operatorname{Var}_n(Z|Y \in J_i)\right)^2$$

4. Let $\hat{v}_1, \ldots, \hat{v}_r$ be the first r most dominant normalised eigenvectors of Λ_{SAVE}. Compute $\hat{\beta}_k \coloneqq \hat{\Sigma}^{-1/2} v_k$ for $k \in \mathbb{N}_r$, and obtain the sufficient predictors $\hat{\beta}_k^T (X_i - \hat{\mu})$ where $i \in \mathbb{N}_n$ and $k \in \mathbb{N}_r$.

Theorem 5.2 of Li (2018) gives that the SAVE subspace contains the SIR subspace, even when either Assumption 4.3.1 or Assumption 4.3.2 fails to hold (though both are needed to ensure that both subspaces are contained in the central subspace).

The question of when the SAVE subspace coincides with the central subspace is addressed in Section 5.6, particularly Theorem 5.3, of Li (2018).

4.4 Literature review: linear sufficient dimension reduction with categorical predictors

In many real-world situations, some of the predictors can be categorical. In experimental settings for example, a categorical predictor could be the group, control or experiment, to which subjects are assigned. The presence of such variables has some implications for sufficient dimension reduction regarding how to take these variables into account. Note that the categorical variable need not be given in advance, and can be found by clustering; though this will affect the estimation procedure for SDR.

As any vector of categorical variables is itself a categorical variable, it can be assumed that only a single such predictor W is present. Chiaromonte et al. (2002) describe three approaches which can be taken to handle this variable: (1) marginalise it out, (2) use it to constrain the dimension reduction procedure, or (3) perform dimension reduction within each subpopulation. These approaches are respectively called *marginal sufficient dimension reduction*, *partial sufficient dimension reduction*, and *conditional sufficient dimension reduction*.

In the classical multivariate linear setting, these situations are technically the following.

Marginal Find the *marginal central subspace* $\mathcal{S}_{Y|X}$.

Partial Find the *partial central subspace* $\mathcal{S}^{(W)}_{Y|X}$. This is the intersection of all subspaces $\mathcal{S}$ of $\mathbb{R}^p$ such that the projection $P_{\mathcal{S}}X$ of X onto $\mathcal{S}$ satisfies

$$Y \perp\!\!\!\perp X | \left(P_{\mathcal{S}}X, W\right).$$

Conditional For each categorical label w, find the *w-conditional central subspace* $\mathcal{S}_{Y_w|X_w}$. This is defined similarly to the central subspace, except now the predictor and the response are constrained by the categorical label w.

The marginal and w-conditional central subspaces can be estimated using any estimation procedure for sufficient dimension reduction. The goal then is to estimate the partial central subspace. Letting C be the number of categorical labels for W, Chiaromonte et al. (2002) showed the following result.

Theorem 4.4.1 (Chiaromonte et al. (2002))**.** *The partial central subspace is the vector space sum of the w-conditional central subspaces. Symbolically,*

$$\mathcal{S}^{(W)}_{Y|X} = \sum_{w \in \mathbb{N}_C} \mathcal{S}_{Y_w|X_w}$$

This result shows the relationship between the partial central subspace and the w-conditional central subspaces. It suggests that the partial central subspace can be estimated by combining estimators of the w-conditional central subspaces. With this insight, Chiaromonte et al. (2002) use it as the basis for extending the sliced inverse regression procedure of Li (1991) to accommodate categorical predictors. They give the name of *partial sliced inverse regression* to their method. Later, Shao et al. (2009) developed *partial sliced average variance estimation* to derive a subspace which captures a greater share of the partial central subspace.

Chiaromonte et al. (2002) also explore the relationship between the partial central subspace and the marginal central subspace, in particular the conditions

under which one is a subspace of the other. They showed the following two results.

Theorem 4.4.2 (Chiaromonte et al. (2002))**.** *If* $W \perp\!\!\!\perp X | P_{\mathcal{S}^W_{Y|X}} X$ *or* $W \perp\!\!\!\perp Y | P_{\mathcal{S}^W_{Y|X}} X$*, then* $\mathcal{S}_{Y|X} \subseteq \mathcal{S}^{(W)}_{Y|X}$.

Theorem 4.4.3 (Chiaromonte et al. (2002))**.** *If* $W \perp\!\!\!\perp Y | X$*, then* $\mathcal{S}^{(W)}_{Y|X} \subseteq \mathcal{S}_{Y|X}$.

Cook and Critchley (2000) gave the following result which relates the regression of Y on X with the regressions of W upon X (containing "joining information") and Y_w upon X_w (containing "conditional regression information").

Theorem 4.4.4 (Cook and Critchley (2000))**.** *Let* $\mathcal{S}_{W|X}$ *be the central subspace, assuming it exists, for the regression of* W *on* X*. Then,*

$$S_{Y|X} \subseteq S_{W|X} + \left(\sum_{w \in \mathbb{N}_C} S_{Y_w | X_w} \right)$$

Sliced inverse regression also has a functional predictor X counterpart developed by Ferré and Yao (2003). It is natural then to apply the ideas from Chiaromonte et al. (2002) to extend functional sliced inverse regression to the setting where there are categorical predictors. This was done by Wang (2017), who developed similar results to the above.

4.5 A nonlinear approach to sufficient dimension reduction with categorical predictors

The existing results, provided in the previous section, in the literature are limited to the case where the extracted manifold is assumed to be linear. In this section, a nonlinear approach for sufficient dimension reduction with categorical predictors is developed. This is accomplished by combining the generalised sufficient

dimension reduction theory, as set out by Li (2018), with the approach taken by Chiaromonte et al. (2002). Chiaromonte et al. (2002) extended the sliced inverse regression estimator of Li (1991); in a similar vein here, the generalised sliced inverse regression estimator, a kernel-based method described by Li (2018), is extended to accommodate categorical predictors.

The rest of this section is structured as follows. In Section 4.5.1, notation is provided. Some preliminary results are given in Section 4.5.2. In Section 4.5.3, measure-theoretic formulations of marginal, partial, and conditional sufficient dimension reduction are provided, and the relationships between these approaches are explored. Kernel mappings, covariance operators, and cross-covariance operators are used to develop, in Section 4.5.6, the partial generalised sliced inverse regression estimator. Section 4.5.4 is therefore devoted to defining these notions and exploring their relevant properties. As numerical implementation of these notions requires the use of vectors and matrices rather than functions and operators, coordinate representation of linear operators is recapped in Section 4.5.5. The proposed estimator is applied to two real-world datasets in Section 4.5.7. Section 4.5.8 gives a closing summary. Proofs are supplied in Section 4.5.9.

4.5.1 Notation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. As any vector of categorical random variables is itself categorical, it is assumed that there is a single such predictor $W : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{N}_C, \mathcal{P}(\mathbb{N}_C))$ where $C \in \mathbb{N}$ is the number of possible categorical labels. Let $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$ be measurable spaces, and define the predictor and response to be some random variables $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_X, \mathcal{F}_X)$ and $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega_Y, \mathcal{F}_Y)$. Let $w \in \mathbb{N}_C$ and define $E_w \coloneqq \{\omega \in \Omega : W(\omega) = w\}$. Notice that $\{E_w : w \in \mathbb{N}_C\}$ is a partition of Ω. It is assumed that $\mathbb{P}(E_w) > 0$ which allows the probability measure $\mathbb{P}_w : \mathcal{F} \to [0, 1]$ given by $\mathbb{P}_w(A) \coloneqq \mathbb{P}(A|E_w)$

to be defined. Observe that if $A \in \mathcal{F}$ is a $\mathbb{P}$-null set ($\mathbb{P}(A) = 0$) then it is also a $\mathbb{P}_w$-null set ($\mathbb{P}_w(A) = 0$), meaning that $\mathbb{P}_w \ll \mathbb{P}$. Recall that $\mathbb{P}(A) = \mathbb{E}_{\mathbb{P}}(\mathbf{1}_A)$. It is seen that, for each $A \in \mathcal{F}$,

$$\mathbb{P}_w(A) = \frac{\mathbb{P}(A \cap E_w)}{\mathbb{P}(E_w)} = \frac{\mathbb{E}_{\mathbb{P}}(\mathbf{1}_A \mathbf{1}_{E_w})}{\mathbb{E}_{\mathbb{P}}(\mathbf{1}_{E_w})} = \mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_A \left(\frac{\mathbf{1}_{E_w}}{\mathbb{E}_{\mathbb{P}}(\mathbf{1}_{E_w})}\right)\right).$$

This gives the Radon-Nikodym derivative $\frac{d\mathbb{P}_w}{d\mathbb{P}} \overset{a.s.\mathbb{P}}{=} \frac{\mathbf{1}_{E_w}}{\mathbb{E}_{\mathbb{P}}(\mathbf{1}_{E_w})}$. The induced probability measures $\mathbb{P}_{X|w} : \mathcal{F}_X \to [0,1]$, $\mathbb{P}_{Y|w} : \mathcal{F}_Y \to [0,1]$, and $\mathbb{P}_{W|w} : \mathcal{P}(\mathbb{N}_C) \to [0,1]$ are defined by the respective mappings $A \mapsto \mathbb{P}_w\left(X^{-1}(A)\right)$, $A \mapsto \mathbb{P}_w\left(Y^{-1}(A)\right)$, and $A \mapsto \mathbb{P}_w\left(W^{-1}(A)\right)$. Notice that for $A \in \mathcal{P}(\mathbb{N}_C)$, $\mathbb{P}_{W|w}(A) = 0$ if $w \notin A$ and $\mathbb{P}_{W|w}(A) = 1$ if $w \in A$. It also holds that $\mathbb{P}_{X|w} \ll \mathbb{P}_X$, $\mathbb{P}_{Y|w} \ll \mathbb{P}_Y$, and $\mathbb{P}_{W|w} \ll \mathbb{P}_W$. By applying Theorem 16.13 of Billingsley (1995), it is seen that the Radon-Nikodym derivatives are related to $\frac{d\mathbb{P}_w}{d\mathbb{P}}$ via $\frac{d\mathbb{P}_w}{d\mathbb{P}} \overset{a.s.\mathbb{P}}{=} \frac{d\mathbb{P}_{X|w}}{d\mathbb{P}_X} \circ X$, $\frac{d\mathbb{P}_w}{d\mathbb{P}} \overset{a.s.\mathbb{P}}{=} \frac{d\mathbb{P}_{Y|w}}{d\mathbb{P}_Y} \circ Y$, and $\frac{d\mathbb{P}_w}{d\mathbb{P}} \overset{a.s.\mathbb{P}}{=} \frac{d\mathbb{P}_{W|w}}{d\mathbb{P}_W} \circ W$.

It is assumed that the codomain of a tuple of random variables is the cartesian product of the respective codomains equipped with the tensor product σ-field. Any remaining notation used has already been defined in Chapter 2.

4.5.2 Preliminary results

The following results are used in the proofs of those in Section 4.5.3. They are included here for clarity.

Lemma 4.5.1. *For any $A \in \mathcal{F}, G \trianglelefteq \mathcal{F}$, and $w \in \mathbb{N}_C$*

$$\mathbb{P}_w(A|G) \overset{a.s.\mathbb{P}_w}{=} \mathbb{P}_w(A|G, W).$$

Corollary 4.5.2. *For any $A \in \mathcal{F}, G \trianglelefteq \mathcal{F}$, and $w \in \mathbb{N}_C$*

$$\mathbb{P}_w(A|G) \overset{a.s.\mathbb{P}_w}{=} \mathbb{P}(A|G, W).$$

The second of these is the main result. It essentially says that conditioning on a particular value of the categorical variable followed by a sub-σ-field of $\mathcal{F}$ gives a w-conditional probability which is equivalent, with respect to $\mathbb{P}_w$, to a conditional probability where the conditioning is on the same σ-field and the categorical variable.

4.5.3 Marginal, partial, and conditional approaches for the categorical predictors

4.5.3.1 Formulation

In the general theory of sufficient dimension reduction, as described by Li (2018), a sub-σ-field G (called an *SDR σ-field*) of $\sigma(X)$ which satisfies the conditional independence $Y \perp\!\!\!\perp X|G$ is sought. As in the classical formulation, there can be many SDR σ-fields so the intersection of them is the target of estimation. Under a mild condition, given in Theorem 1 in Lee et al. (2013), this intersection is also a SDR σ-field.

In the present setting, W also needs to be considered. Taking inspiration from the work of Chiaromonte et al. (2002), the following three approaches to handling this variable are given.

Marginal: Find $G \trianglelefteq \sigma(X)$ such that $Y \perp\!\!\!\perp X|G$. Any such G is called a *marginal SDR σ-field*.

Partial: Find $G^{(W)} \trianglelefteq \sigma(X)$ such that $Y \perp\!\!\!\perp X| \left(G^{(W)}, \sigma(W)\right)$. Any such $G^{(W)}$ is called a *partial SDR σ-field.*

Conditional: For each $w \in \mathbb{N}_C$, find $G_w \trianglelefteq \sigma(X)$ such that $Y \overset{w}{\perp\!\!\!\perp} X|G_w$. This means that, for any $A \in \sigma(X)$ and $B \in \sigma(Y)$,

$$\mathbb{P}_w(A \cap B|G_w) \overset{\text{a.s.}\mathbb{P}}{=} \mathbb{P}_w(A|G_w)\,\mathbb{P}_w(B|G_w)$$

Any such G_w is called a *w-conditional SDR σ-field*.

In each approach, the intersection of all sub-σ-fields of $\sigma(X)$ that satisfy the appropriate relation is sought. It is assumed that these intersections themselves satisfy the relevant relations. These sub-σ-fields are respectively called the *marginal central σ-field* (denoted by $\mathcal{M}$), the *partial central σ-field* (denoted by $\mathcal{M}^{(W)}$), and the *w-conditional central σ-field* (denoted by $\mathcal{M}_w$).

4.5.3.2 Relationships between the three approaches

The relationship between the partial central σ-field and the w-conditional central σ-fields is now explored.

Lemma 4.5.3. *$G \trianglelefteq \sigma(X)$ is a partial SDR σ-field if and only if it is a w-conditional SDR σ-field for each $w \in \mathbb{N}_C$.*

This implies that, for any $w \in \mathbb{N}_C$, $\mathcal{M}_w \trianglelefteq G^{(W)}$ for any partial SDR σ-field $G^{(W)}$. In particular, $\mathcal{M}_w \trianglelefteq \mathcal{M}^{(W)}$. This implies that $(\mathcal{M}_1, \ldots, \mathcal{M}_C) \trianglelefteq \mathcal{M}^{(W)}$. Theorem 4.5.4 gives the stronger result that these σ-fields are equal.

Theorem 4.5.4. $\mathcal{M}^{(W)} = (\mathcal{M}_1, \ldots, \mathcal{M}_C)$.

The relationship between the partial central σ-field and the marginal central σ-field is now explored. The first two of the following results give sufficient conditions for one to be a sub-σ-field of the other.

Theorem 4.5.5. *If $W \perp\!\!\!\perp X|\mathcal{M}^{(W)}$ or $W \perp\!\!\!\perp Y|\mathcal{M}^{(W)}$, then $\mathcal{M} \trianglelefteq \mathcal{M}^{(W)}$.*

Theorem 4.5.6. *If $W \perp\!\!\!\perp Y|X$, then $\mathcal{M}^{(W)} \trianglelefteq \mathcal{M}$.*

Theorem 4.5.7. *$\mathcal{M} \trianglelefteq \left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right)$ where $\mathcal{M}_{(W)} \trianglelefteq \sigma(X)$ is the central σ-field, assumed to exist, for the conditional independence $W \perp\!\!\!\perp X|G$.*

4.5.4 Covariance operators and their properties

The target σ-fields are abstract concepts which are not immediately estimable. To overcome this, what are sought instead are subspaces of reproducing kernel Hilbert spaces of measurable functions. To this end, suppose that $\mathcal{H}_X$ and $\mathcal{H}_Y$ are reproducing kernel Hilbert spaces of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$-valued measurable functions on $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_Y, \mathcal{F}_Y)$ which are generated by measurable kernels $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$ and $\kappa_Y : \Omega_Y \times \Omega_Y \to \mathbb{R}$ respectively. It is assumed that κ_X and κ_Y are characteristic. Define the product kernel $\kappa_{XY} : (\Omega_X \times \Omega_Y) \times (\Omega_X \times \Omega_Y) \to \mathbb{R}$ by $\kappa_{XY}((x_1, y_1), (x_2, y_2)) \coloneqq \kappa_X(x_1, x_2)\,\kappa_Y(y_1, y_2)$. It is also assumed that κ_{XY} is characteristic.

It is assumed that $\mathcal{H}_X$ and $\mathcal{H}_Y$ are separable. A sufficient condition, by Theorem 2.7.5 of Hsing and Eubank (2015), for this is that

1. Ω_X and Ω_Y are separable metric spaces.

2. $\mathcal{F}_X$ and $\mathcal{F}_Y$ are the Borel σ-fields generated by the induced topologies on Ω_X and Ω_Y.

3. κ_X and κ_Y are continuous.

If the first and second conditions are assumed, then, for any $G_X \trianglelefteq \mathcal{F}_X$ and $G_Y \trianglelefteq \mathcal{F}_Y$, there exists regular conditional distributions $\mathbb{P}_{X|G_X}$ and $\mathbb{P}_{Y|G_Y}$. In light of Theorem 2.5.21, these conditions are therefore assumed. Virta et al. (2022) developed generalised sliced inverse regression for predictors and responses taking values in separable metric spaces.

For any $G \trianglelefteq \sigma(X)$, let $\mathcal{H}_X(G) \coloneqq \{f \in \mathcal{H}_X : f(X) \text{ is } G\text{-measurable}\}$. and let $\mathcal{H}_X^G \coloneqq \{f \in \mathcal{H}_X : f(X) \text{ is } (G, W)\text{-measurable}\}$. Name $\mathcal{H}_X(\mathcal{M})$ the *marginal central class*, $\mathcal{H}_X(\mathcal{M}_w)$ the w-conditional central class ($w \in \mathbb{N}_C$), and $\mathcal{H}_X^{(W)} \coloneqq \mathcal{H}_X^{\mathcal{M}^{(W)}}$ the partial central class. The marginal central class and, for each $w \in \mathbb{N}_C$,

the w-conditional central class can be estimated using the generalised sliced inverse regression procedure as described by Li (2018). More generally, they can be estimated using any kernel based method for marginal or w-conditional generalised sufficient dimension reduction.

Lemma 4.5.8. *For any $G \trianglelefteq \sigma(X)$, $\mathcal{H}_X(G)$ and $\mathcal{H}_X^G$ are both closed subspaces of $\mathcal{H}_X$.*

Note it appears to have been assumed in Li (2018) that there is a one-to-one correspondence between $\mathcal{C} \coloneqq \{G \trianglelefteq \sigma(X)\}$ and $\{\mathcal{H}_X(G) : G \in \mathcal{C}\}$. The author of this thesis believes this to be true, though leaves the proof to future research.

The goal is to find some operator from $\mathcal{H}_Y$ to $\mathcal{H}_X$ whose closed range is equal, under some assumptions to be given, to the partial central class. To this end, the kernel mean embeddings of $\mathbb{P}_X$ and $\mathbb{P}_Y$ are now defined. To do this, Assumption 4.5.1 is made so that, by Theorem 2.6.5 of Hsing and Eubank (2015), $\kappa_X(\cdot, X)$ and $\kappa_Y(\cdot, Y)$ are Bochner $\mathbb{P}$-integrable. The kernel mean embeddings μ_X and μ_Y of $\mathbb{P}_X$ and $\mathbb{P}_Y$ respectively are defined to be the expectations $\mathbb{E}_{\mathbb{P}}(\kappa_X(\cdot, X))$ and $\mathbb{E}_{\mathbb{P}}(\kappa_Y(\cdot, Y))$. See Muandet et al. (2016) for a comprehensive review of the theory of kernel mean embeddings of distributions.

Assumption 4.5.1. $\mathbb{E}_{\mathbb{P}}(\kappa_X(X, X)) < \infty$ and $\mathbb{E}_{\mathbb{P}}(\kappa_Y(Y, Y)) < \infty$.

By using the law of total probability, it is seen that this assumption implies that, for $w \in \mathbb{N}_C$, $\mathbb{E}_{\mathbb{P}_w}(\kappa_X(X, X)) < \infty$ and similarly for Y. This thus allows defining the kernel mean embeddings of $\mathbb{P}_{X|w}$ and $\mathbb{P}_{Y|w}$, for $w \in \mathbb{N}_C$, to be the functions $\mu_{X|w} \coloneqq \mathbb{E}_{\mathbb{P}_w}(\kappa_X(\cdot, X))$ and $\mu_{Y|w} \coloneqq \mathbb{E}_{\mathbb{P}_w}(\kappa_Y(\cdot, Y))$.

By using the reproducing property, it is seen that, for any $f \in \mathcal{H}_X$, $f(X) = \langle f, \kappa_X(\cdot, X)\rangle_{\mathcal{H}_X}$ and so $\mathbb{E}_{\mathbb{P}}(f(X)) = \langle f, \mu_X\rangle_{\mathcal{H}_X}$. Similar considerations apply for Y and when the expectation is taken with respect to $\mathbb{P}_w$ instead for some $w \in \mathbb{N}_C$.

By Theorem 2.5.19, $\mathcal{H}_X$ is a subset of $L^2(\mathbb{P}_X)$ such that its vector space sum with the space of $\mathbb{P}_X$ almost surely constant real-valued functions on Ω_X is dense in $L^2(\mathbb{P}_X)$. Furthermore, $\mathcal{H}_X$ is a subset of $L^2\left(\mathbb{P}_{X|w}\right)$ such that its vector space sum with the space of $\mathbb{P}_{X|w}$ almost surely constant real-valued functions on Ω_X is dense in $L^2\left(\mathbb{P}_{X|w}\right)$.

Now observe that $L^2(\mathbb{P}_X) = \bigcap_{w\in\mathbb{N}_C} L^2\left(\mathbb{P}_{X|w}\right)$ as, for $f \in L^2(\mathbb{P}_X)$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_X}\left(f^2\right) &= \int f^2 \mathrm{d}\mathbb{P}_X \\
&= \int f^2 \mathrm{d}\left(\sum_{w\in\mathbb{N}_C} \mathbb{P}(E_w)\,\mathbb{P}_{X|w}\right) \\
&= \sum_{w\in\mathbb{N}_C} \mathbb{P}(E_w) \int f^2 \mathrm{d}\mathbb{P}_{X|w} \\
&= \sum_{w\in\mathbb{N}_C} \mathbb{P}(E_w)\,\mathbb{E}_{\mathbb{P}_{X|w}}\left(f^2\right).
\end{aligned}$$

As the first expectation is finite, so must be all the expectations in the summation hence $L^2(\mathbb{P}_X) \subseteq \bigcap_{w=1}^{C} L^2\left(\mathbb{P}_{X|w}\right)$. Conversely, if all the expectations in the summation are finite then the first expectation is also finite thus the reverse inclusion holds. By similar reasoning, $L^2(\mathbb{P}_Y) = \bigcap_{w\in\mathbb{N}_C} L^2\left(\mathbb{P}_{Y|w}\right)$.

For $G \trianglelefteq \sigma(X)$, define $L^2(\mathbb{P}_X|G) \coloneqq \left\{f \in L^2(\mathbb{P}_X) : f(X) \text{ is } G\text{-measurable}\right\}$. Define $L^2(\mathbb{P}_Y|G)$ and, for $w \in \mathbb{N}_C$, $L^2\left(\mathbb{P}_{X|w}|G\right)$, and $L^2\left(\mathbb{P}_{Y|w}|G\right)$ similarly.

Assumption 4.5.2. Let $G \trianglelefteq \sigma(X)$. It is assumed that $\mathcal{H}_X(G)$ is dense modulo $\mathbb{P}_X$ almost sure constants in $L^2(\mathbb{P}_X)$. Furthermore, for each $w \in \mathbb{N}_C$, it is assumed that $\mathcal{H}_X(G)$ is dense modulo $\mathbb{P}_{X|w}$ almost sure constants in $L^2\left(\mathbb{P}_{X|w}\right)$. Yet further, it is assumed that $\mathcal{H}_X^G$ is dense modulo $\mathbb{P}_X$ almost sure constants in $\left\{f \in L^2(\mathbb{P}_X) : f(X) \text{ is } (G,W)\text{-measurable}\right\}$.

This assumption is similar to an assumption used in Lee et al. (2013) and Li (2018) for marginal generalised sliced inverse regression. An analogue of Theorem 13.3 of Li (2018) would imply that this assumption automatically holds,

but the author of this thesis believes that there is an invalid step in the given proof and has been contact with the author of that text. Specifically, the line in that text (page 217) which reads "Because $\mathcal{A} \setminus \mathcal{A}_{\mathcal{G}}$ contains only functions that are not measurable with respect to $\mathcal{G}$, the only $s^{(2)}$ that satisfies the above equality (see the text) is 0 (no constants other than 0 are in $\mathcal{H}_X$)" is a non-sequiter, because (1) $s^{(2)}$ is only known to belong to $L^2(\mathbb{P}_X)$ and not neccessarily to $\mathcal{H}_X$, (2) the sum of non-$\mathcal{G}$-measurable functions may be $\mathcal{G}$-measurable, and (3) the author seems to implicitly assume that $\mathrm{Var}\left(\mathbb{E}\left(s^{(2)}(X)\middle|\mathcal{G}\right)\right) = 0$ without justification. Furthermore, there are gaps in previous steps of the given argument but these may be filled in by application of the law of total variance.

Define the random operators

$$\begin{aligned}
\varphi_{XX} &\coloneqq \kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X) \\
\varphi_{YY} &\coloneqq \kappa_Y(\cdot, Y) \otimes \kappa_Y(\cdot, Y) \\
\varphi_{XY} &\coloneqq \kappa_X(\cdot, X) \otimes \kappa_Y(\cdot, Y) \\
\varphi_{YX} &\coloneqq \kappa_Y(\cdot, Y) \otimes \kappa_X(\cdot, X).
\end{aligned}$$

Define now the covariance and cross-covariance operators

$$\begin{aligned}
\Sigma_{XX} &\coloneqq \mathbb{E}_{\mathbb{P}}(\varphi_{XX}) - (\mu_X \otimes \mu_X) \\
\Sigma_{YY} &\coloneqq \mathbb{E}_{\mathbb{P}}(\varphi_{YY}) - (\mu_Y \otimes \mu_Y) \\
\Sigma_{XY} &\coloneqq \mathbb{E}_{\mathbb{P}}(\varphi_{XY}) - (\mu_X \otimes \mu_Y) \\
\Sigma_{YX} &\coloneqq \mathbb{E}_{\mathbb{P}}(\varphi_{YX}) - (\mu_Y \otimes \mu_X).
\end{aligned}$$

Note that the effective codomain of these Hilbert-Schmidt operators is their range, which is not necessarily closed. If the codomain were a real Hilbert space, this would imply that the spaces are finite-dimensional; the fact that the codomain is generally not closed circumvents this issue.

Let $w \in \mathbb{N}_C$. Define the operators

$$\begin{aligned}
\Sigma_{XX|w} &:= \mathbb{E}_{\mathbb{P}_w}(\varphi_{XX}) - (\mu_{X|w} \otimes \mu_{X|w}) \\
\Sigma_{YY|w} &:= \mathbb{E}_{\mathbb{P}_w}(\varphi_{YY}) - (\mu_{Y|w} \otimes \mu_{Y|w}) \\
\Sigma_{XY|w} &:= \mathbb{E}_{\mathbb{P}_w}(\varphi_{XY}) - (\mu_{X|w} \otimes \mu_{Y|w}) \\
\Sigma_{YX|w} &:= \mathbb{E}_{\mathbb{P}_w}(\varphi_{YX}) - (\mu_{Y|w} \otimes \mu_{X|w}).
\end{aligned}$$

The following lemma relates the range of the w-conditional cross-covariance operators with the w-conditional covariance operators.

Lemma 4.5.9. *Let $w \in \mathbb{N}_C$. It holds that*

$$\begin{aligned}
\operatorname{Ran}(\Sigma_{XY|w}) &\subseteq \overline{\operatorname{Ran}(\Sigma_{XX|w})} \\
\operatorname{Ran}(\Sigma_{YX|w}) &\subseteq \overline{\operatorname{Ran}(\Sigma_{YY|w})}.
\end{aligned}$$

Recall (see, for example, Theorem 3.3.7 of Hsing and Eubank (2015)) that, for any bounded operator $A \in \mathcal{L}(\mathcal{H}_X, \mathcal{H}_X)$, $\mathcal{H}_X = \operatorname{Ker}(A) \oplus \operatorname{Ker}(A)^\perp$ and $\operatorname{Ker}(A)^\perp = \overline{\operatorname{Ran}(A^*)}$. This, coupled with the fact that Σ_{XX} is self-adjoint, implies that $\mathcal{H}_X = \operatorname{Ker}(\Sigma_{XX}) \oplus \overline{\operatorname{Ran}(\Sigma_{XX})}$. Similarly, $\mathcal{H}_Y = \operatorname{Ker}(\Sigma_{YY}) \oplus \overline{\operatorname{Ran}(\Sigma_{YY})}$. Furthermore, for each $w \in \mathbb{N}_C$, $\mathcal{H}_X = \operatorname{Ker}(\Sigma_{XX|w}) \oplus \overline{\operatorname{Ran}(\Sigma_{XX|w})}$ and $\mathcal{H}_Y = \operatorname{Ker}(\Sigma_{YY|w}) \oplus \overline{\operatorname{Ran}(\Sigma_{YY|w})}$.

For the purposes of marginal or partial sufficient dimension reduction, non-zero functions which are constant almost surely with respect to $\mathbb{P}_X$ and $\mathbb{P}_Y$ respectively can be discarded. Now the elements in $\mathcal{H}_X$ which are almost surely constant with respect to $\mathbb{P}_X$ are those in $\operatorname{Ker}(\Sigma_{XX})$. To see this, consider that if $f \in \operatorname{Ker}(\Sigma_{XX})$ then $\operatorname{Var}_{\mathbb{P}}(f(X)) = \langle f, \Sigma_{XX} f \rangle_{\mathcal{H}_X} = \langle f, 0 \rangle_{\mathcal{H}_X} = 0$ which implies that f is almost surely $\mathbb{P}_X$-constant. Conversely, if $f \in \mathcal{H}_X$ is almost surely $\mathbb{P}_X$-constant then $\operatorname{Var}_{\mathbb{P}}(f(X)) = 0$ so $\Sigma_{XX} f = 0$. Analogous considerations apply for $\mathcal{H}_Y$. To perform this discarding then, the following assumption is made.

Assumption 4.5.3. It is assumed that

$$\text{Ker}\left(\Sigma_{XX}\right) = \{0\}$$
$$\text{Ker}\left(\Sigma_{YY}\right) = \{0\}.$$

Analogously, for the purposes of w-conditional sufficient dimension reduction, non-zero functions which are constant almost surely with respect to $\mathbb{P}_{X|w}$ and $\mathbb{P}_{Y|w}$ respectively may be discarded with. These are seen, in a similar way to above, to be the functions in $\text{Ker}\left(\Sigma_{XX|w}\right)$ and $\text{Ker}\left(\Sigma_{YY|w}\right)$ respectively. Thus the following assumption is made.

Assumption 4.5.4. For each $w \in \mathbb{N}_C$, it is assumed that

$$\text{Ker}\left(\Sigma_{XX|w}\right) = \{0\}$$
$$\text{Ker}\left(\Sigma_{YY|w}\right) = \{0\}.$$

Under Assumption 4.5.3, $\mathcal{H}_X = \overline{\text{Ran}\left(\Sigma_{XX}\right)}$ and $\mathcal{H}_Y = \overline{\text{Ran}\left(\Sigma_{YY}\right)}$. Furthermore, under Assumption 4.5.4, it holds that, for $w \in \mathbb{N}_C$, $\mathcal{H}_X = \overline{\text{Ran}\left(\Sigma_{XX|w}\right)}$ and $\mathcal{H}_Y = \overline{\text{Ran}\left(\Sigma_{YY|w}\right)}$. Now it turns out that $\overline{\text{Ran}\left(\Sigma_{XX}\right)}$ and $\overline{\text{Ran}\left(\Sigma_{YY}\right)}$ have explicit expressions, as given in the following lemma. This lemma and its proof are given in Li (2018), though the proof is repeated here for completeness (and to address a typo in their version).

Lemma 4.5.10. *It holds that*

$$\overline{\text{Ran}\left(\Sigma_{XX}\right)} = \overline{\text{Span}}\left\{\kappa_X\left(\cdot, x\right) - \mu_X : x \in \Omega_X\right\}$$
$$\overline{\text{Ran}\left(\Sigma_{YY}\right)} = \overline{\text{Span}}\left\{\kappa_Y\left(\cdot, y\right) - \mu_Y : y \in \Omega_Y\right\}.$$

Making the appropriate modifications, the proof of this lemma can be changed to prove that, for each $w \in \mathbb{N}_C$,

$$\overline{\text{Ran}\left(\Sigma_{XX|w}\right)} = \overline{\text{Span}}\left\{\kappa_X\left(\cdot, x\right) - \mu_{X|w} : x \in \Omega_X\right\}$$
$$\overline{\text{Ran}\left(\Sigma_{YY|w}\right)} = \overline{\text{Span}}\left\{\kappa_Y\left(\cdot, y\right) - \mu_{Y|w} : y \in \Omega_Y\right\}.$$

The operator for partial GSIR will soon be defined as a combination of the operators for w-conditional GSIR, which are themselves used to obtain an optimisation problem paralleling that for w-conditional classical sliced inverse regression. In analogy to the marginal GSIR operator $\Sigma_{XX}^{-1}\Sigma_{XY}$, and its counterpart $\Sigma_{YY}^{-1}\Sigma_{YX}$, used in Li (2018) (see that text for the assumptions required for these to be defined and compact), the w-conditional ($w \in \mathbb{N}_C$) GSIR operator is defined to be $\Sigma_{XX|w}^{-1}\Sigma_{XY|w}$ and its counterpart part is $\Sigma_{YY|w}^{-1}\Sigma_{YX|w}$. In order for these to be defined, the following assumption is made.

Assumption 4.5.5. For $w \in \mathbb{N}_C$, it is assumed that

$$\begin{aligned}\text{Ran}\left(\Sigma_{XY|w}\right) &\subseteq \text{Ran}\left(\Sigma_{XX|w}\right)\\ \text{Ran}\left(\Sigma_{YX|w}\right) &\subseteq \text{Ran}\left(\Sigma_{YY|w}\right).\end{aligned}$$

These operators are also assumed to be compact.

Assumption 4.5.6. For $w \in \mathbb{N}_C$, it is assumed that $\Sigma_{XX|w}^{-1}\Sigma_{XY|w}$ and $\Sigma_{YY|w}^{-1}\Sigma_{YX|w}$ are compact operators.

The interpretation of these assumptions is similar to that for the marginal GSIR operator which is discussed in Li (2018). Essentially, they are assumptions on the smoothness of the relationship between X and Y within the category indexed by w.

For convenience, let (for $w \in \mathbb{N}_C$) $R_{YX|w} \coloneqq \Sigma_{YY|w}^{-1}\Sigma_{YX|w}$ and $R_{XY|w} \coloneqq \Sigma_{XX|w}^{-1}\Sigma_{YX|w}$.

Lemma 4.5.11. *Let $w \in \mathbb{N}_C$. It holds that, for any $f \in \mathcal{H}_X$,*

$$(R_{YX|w}f)(Y) \overset{a.s.\mathbb{P}_w}{=} \mathbb{E}_{\mathbb{P}_w}\left(f(X)|Y\right) + \mathbb{E}_{\mathbb{P}_w}\left(\left(R_{YX|w}\right)(Y)\right) - \mathbb{E}_{\mathbb{P}_w}\left(f(X)\right).$$

Theorem 4.5.12. *Let $w \in \mathbb{N}_C$. It holds that*

$$\overline{\text{Ran}\left(\Sigma_{XX|w}^{-1}\Sigma_{XY|w}\right)} \subseteq \mathcal{H}_X\left(\mathcal{M}_w\right).$$

Observe that, for any invertible bounded linear operator A_w from $\mathcal{H}_Y$ to itself, $\overline{\text{Ran}\left(R_{XY|w}A_wR^*_{XY|w}\right)} = \overline{\text{Ran}\left(R_{XY|w}\right)}$ (see the proof for Theorem 3.3.7 of Hsing and Eubank (2015) for the case when A_w is the identity operator). By similar reasoning to that used by Li (2018), taking $A_w \coloneqq \Sigma^{-1}_{YY|w}$ yields the optimisation problem

$$\begin{aligned} &\text{Max } \text{Var}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(\langle f, \kappa_X(\cdot, X)\rangle_{\mathcal{H}_X}\middle|Y\right)\right) \\ &\text{Subject to } \text{Var}_{\mathbb{P}_w}\left(\langle f, \kappa_X(\cdot, X)\rangle_{\mathcal{H}_X}\right) = 1, f \perp \text{Span}\{f_1, \ldots, f_k\} \end{aligned} \tag{4.2}$$

where $f_1, \ldots f_k$ are the previous solutions to this problem. This parallels the optimisation problem for classical sliced inverse regression. Hence, for all $w \in \mathbb{N}_C$, take $A_w \coloneqq \Sigma^{-1}_{YY|w}$. Define the w-conditional GSIR estimator to be $\Lambda_{w\text{-GSIR}} \coloneqq \Sigma^{-1}_{XX|w}\Sigma_{XY|w}\Sigma^{-1}_{YY|w}\Sigma_{YX|w}\Sigma^{-1}_{XX|w}$. The closed range of this operator is contained in $\mathcal{H}_X$ ($\mathcal{M}_w$). In analogy to how, in Chiaromonte et al. (2002), the PSIR matrix is taken to be a probability weighted sum of the w-SIR matrices, define $\Lambda_{\text{PGSIR}} \coloneqq \sum_{w\in\mathbb{N}_C}\Lambda_{w\text{-GSIR}}$. Note that there is no need to probability weight this operator because, if this was done, one would obtain an operator with the same closed range. This is the operator for which an eigenvalue decomposition will be used to derive the partial sufficient predictors.

Theorem 4.5.13. *It holds that*

$$\overline{\text{Ran}\left(\Lambda_{PGSIR}\right)} \subseteq \mathcal{H}_X^{(W)}.$$

In the marginal generalised sliced inverse regression setting, Li (2018) showed the exhaustiveness (see Theorem 13.2 of that text) of $\Sigma^{-1}_{XX}\Sigma_{XY}$ (required assumptions given in that text) under a so-called "completeness" assumption – an analogue of which is shortly given. Significantly, they demonstrate that it is not a strong assumption by showing, e.g., it holds in the setting of nonparametric regression (see Proposition 12.7 of that text).

Definition 4.5.1. Let $w \in \mathbb{N}_C$. $G \trianglelefteq \sigma(X)$ is $\mathbb{P}_w$-*complete* for $\sigma(Y)$ if, for all $f \in L^2\left(\mathbb{P}_{X|w}|G\right)$,

$$\left[\mathbb{E}_{\mathbb{P}_w}\left(f(X)|Y\right) \overset{\text{a.s.}\mathbb{P}_w}{=} c_f\right] \implies \left[f(X) \overset{\text{a.s.}\mathbb{P}_w}{=} c_f\right]$$

where c_f is a real constant.

Theorem 4.5.14. *For each $w \in \mathbb{N}_C$, suppose that $\mathcal{M}_w$ is $\mathbb{P}_w$-complete for $\sigma(Y)$. Then, it holds that*

$$\overline{\text{Ran}\left(\Sigma_{XX|w}^{-1}\Sigma_{XY|w}\right)} = \mathcal{H}_X\left(\mathcal{M}_w\right).$$

The author of this thesis believes that it should additionally be possible to prove the exhaustiveness of the PGSIR estimator under some kind of completeness assumption. For time purposes, this is left to future research.

4.5.5 Coordinate representation

For numerical implementation, use has to be made of vectors and matrices instead of functions and operators. This section therefore recaps the theory of coordinate representation of functions in finite-dimensional spaces and linear operators between such spaces.

Suppose $\mathcal{H}_1, \mathcal{H}_2$, and $\mathcal{H}_3$ are finite-dimensional Hilbert spaces with spanning systems $\mathcal{B}_1 = \left\{b_1^{(1)}, \ldots, b_{m_1}^{(1)}\right\}$, $\mathcal{B}_2 = \left\{b_1^{(2)}, \ldots, b_{m_2}^{(2)}\right\}$ and $\mathcal{B}_3 = \left\{b_1^{(3)}, \ldots, b_{m_3}^{(3)}\right\}$ respectively. Let $k \in \{1, 2, 3\}$ and let $[f]_{\mathcal{B}_k}$ denote the coordinate vector of f with respect to $\mathcal{B}_k$. Let $\mathcal{G}_{\mathcal{B}_k}$ denote the matrix whose (i, j)-th element is $\left\langle b_i^{(k)}, b_j^{(k)}\right\rangle_{\mathcal{H}_k}$. Let $l \in \{1, 2, 3\}$ and suppose that $A : \mathcal{H}_k \to \mathcal{H}_l$ is a linear operator. Denote by $[A]_{\mathcal{B}_k}^{\mathcal{B}_l}$ the matrix whose (i, j)-th element is given by $\left(\left[Ab_j^{(k)}\right]_{\mathcal{B}_l}\right)_i$ where $i \in \mathbb{N}_{m_2}$ and $j \in \mathbb{N}_{m_1}$.

Theorem 4.5.15 (Lemma 12.3 of Li (2018))**.** *It holds that:*

1. *if* $h \in \mathcal{H}_k$ *and* $T : \mathcal{H}_k \to \mathcal{H}_l$ *is a linear operator, then*

$$[Th]_{\mathcal{B}_l} = \left([T]_{\mathcal{B}_k}^{\mathcal{B}_l}\right)[h]_{\mathcal{B}_k}.$$

2. *if* $f_1, f_2 \in \mathcal{H}_k$ *and* $\alpha_1, \alpha_2 \in \mathbb{R}$*, then*

$$[\alpha_1 f_1 + \alpha_2 f_2]_{\mathcal{B}_k} = \alpha_1 [f_1]_{\mathcal{B}_k} + \alpha_2 [f_2]_{\mathcal{B}_k}.$$

Furthermore if $T_1, T_2 : \mathcal{H}_k \to \mathcal{H}_l$ *are linear operators then*

$$[\alpha_1 T_1 + \alpha_2 T_2]_{\mathcal{B}_k}^{\mathcal{B}_l} = \alpha_1 \left([T_1]_{\mathcal{B}_k}^{\mathcal{B}_l}\right) + \alpha_2 \left([T_2]_{\mathcal{B}_k}^{\mathcal{B}_l}\right).$$

3. *suppose that* $j \in \{1, 2, 3\}$*. If* $T_1 : \mathcal{H}_j \to \mathcal{H}_k$ *and* $T_2 : \mathcal{H}_k \to \mathcal{H}_l$*, then*

$$[T_2 T_1]_{\mathcal{B}_j}^{\mathcal{B}_l} = \left([T_2]_{\mathcal{B}_k}^{\mathcal{B}_l}\right)\left([T_1]_{\mathcal{B}_j}^{\mathcal{B}_k}\right).$$

4. *if* $f_1, f_2 \in \mathcal{H}_k$*, then*

$$\langle f_1, f_2 \rangle_{\mathcal{H}_k} = [f_1]_{\mathcal{B}_k}^T \mathcal{G}_{\mathcal{B}_k} [f_2]_{\mathcal{B}_k}.$$

5. *if* $f \in \mathcal{H}_k$ *and* $g \in \mathcal{H}_l$*, then*

$$[g \otimes f]_{\mathcal{B}_k}^{\mathcal{B}_l} = [g]_{\mathcal{B}_l} [f]_{\mathcal{B}_k}^T \mathcal{G}_{\mathcal{B}_k}.$$

4.5.6 Partial generalised sliced inverse regression

The work in this and the next subsection represent a first attempt at implementing partial generalised sliced inverse regression numerically and testing it on two real-world datasets. The author believes that other approaches may be possible and preferable, and is actively working on them.

Let $(W_1, X_1, Y_1), \ldots, (W_n, X_n, Y_n)$ be an independent and identically distributed sample for (W, X, Y). For $w \in \mathbb{N}_C$, let $\mathcal{I}_w := \{i \in \mathbb{N}_n : W_i = w\}$ and let $n_w := \text{Card}(\mathcal{I}_w)$. It holds that $\sum_{w \in \mathbb{N}_C} n_w = n$. Let $\mathcal{T} := \{w \in \mathbb{N}_C : n_w > 0\}$.

This allows for the possibility that at least one of the possible categorical labels is not observed in the sample. Henceforth, $w \in \mathcal{T}$. Let $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$ and $\kappa_Y : \Omega_Y \times \Omega_Y \to \mathbb{R}$ be measurable kernel functions satisfying those assumptions regarding the kernels that were given previously. Let $\mathcal{H}_X \coloneqq \text{Span}\{\kappa_X(\cdot, X_i) : i \in \mathbb{N}_n\}$ and $\mathcal{H}_Y \coloneqq \text{Span}\{\kappa_Y(\cdot, Y_i) : i \in \mathbb{N}_n\}$ be the kernel spaces generated by the sample. For $w \in \mathcal{T}$, let $(X_1^{(w)}, Y_1^{(w)}), \ldots (X_{n_w}^{(w)}, Y_{n_w}^{(w)})$ be the subsample corresponding to $W = w$. For each $w \in \mathcal{T}$, define the empirical estimators:

$$\begin{aligned}
\hat{\mu}_{X|w} &\coloneqq \frac{1}{n_w}\sum_{i=1}^{n_w} \kappa_X(\cdot, X_i^{(w)}) \\
\hat{\mu}_{Y|w} &\coloneqq \frac{1}{n_w}\sum_{i=1}^{n_w} \kappa_Y(\cdot, Y_i^{(w)}) \\
\hat{B}_{XX|w} &\coloneqq \frac{1}{n_w}\sum_{i=1}^{n_w} \kappa_X(\cdot, X_i^{(w)}) \otimes \kappa_X(\cdot, X_i^{(w)}) \\
\hat{B}_{XY|w} &\coloneqq \frac{1}{n_w}\sum_{i=1}^{n_w} \kappa_X(\cdot, X_i^{(w)}) \otimes \kappa_Y(\cdot, Y_i^{(w)}) \\
\hat{B}_{YX|w} &\coloneqq \frac{1}{n_w}\sum_{i=1}^{n_w} \kappa_Y(\cdot, Y_i^{(w)}) \otimes \kappa_X(\cdot, X_i^{(w)}) \\
\hat{B}_{YY|w} &\coloneqq \frac{1}{n_w}\sum_{i=1}^{n_w} \kappa_Y(\cdot, Y_i^{(w)}) \otimes \kappa_Y(\cdot, Y_i^{(w)}) \\
\hat{\Sigma}_{XX|w} &\coloneqq \hat{B}_{XX|w} - (\hat{\mu}_{X|w} \otimes \hat{\mu}_{X|w}) \\
\hat{\Sigma}_{XY|w} &\coloneqq \hat{B}_{XY|w} - (\hat{\mu}_{X|w} \otimes \hat{\mu}_{Y|w}) \\
\hat{\Sigma}_{YX|w} &\coloneqq \hat{B}_{YX|w} - (\hat{\mu}_{Y|w} \otimes \hat{\mu}_{X|w}) \\
\hat{\Sigma}_{YY|w} &\coloneqq \hat{B}_{YY|w} - (\hat{\mu}_{Y|w} \otimes \hat{\mu}_{Y|w}).
\end{aligned}$$

Notice that the last four of these can be written as:

$$\begin{aligned}
\hat{\Sigma}_{XX|w} &= \frac{1}{n_w}\sum_{i=1}^{n_w} \left[\left(\kappa_X(\cdot, X_i^{(w)}) - \hat{\mu}_{X|w}\right) \otimes \left(\kappa_X(\cdot, X_i^{(w)}) - \hat{\mu}_{X|w}\right)\right] \\
\hat{\Sigma}_{XY|w} &= \frac{1}{n_w}\sum_{i=1}^{n_w} \left[\left(\kappa_X(\cdot, X_i^{(w)}) - \hat{\mu}_{X|w}\right) \otimes \left(\kappa_Y(\cdot, Y_i^{(w)}) - \hat{\mu}_{Y|w}\right)\right]
\end{aligned}$$

$$\hat{\Sigma}_{YX|w} = \frac{1}{n_w}\sum_{i=1}^{n_w}\left[\left(\kappa_Y(\cdot, Y_i^{(w)}) - \hat{\mu}_{Y|w}\right)\otimes\left(\kappa_X(\cdot, X_i^{(w)}) - \hat{\mu}_{X|w}\right)\right]$$
$$\hat{\Sigma}_{YY|w} = \frac{1}{n_w}\sum_{i=1}^{n_w}\left[\left(\kappa_Y(\cdot, Y_i^{(w)}) - \hat{\mu}_{Y|w}\right)\otimes\left(\kappa_Y(\cdot, Y_i^{(w)}) - \hat{\mu}_{Y|w}\right)\right].$$

For each $w \in \mathcal{T}$, the subspaces $\overline{\text{Ran}}\left(\hat{\Sigma}_{XX|w}\right)$ and $\overline{\text{Ran}}\left(\hat{\Sigma}_{YY|w}\right)$ are spanned by (see the point following Lemma 4.5.10):

$$\mathcal{B}_{X|w} \coloneqq \left\{\kappa_X(\cdot, X_i^{(w)}) - \hat{\mu}_{X|w} : i \in \mathbb{N}_{n_w}\right\}$$
$$\mathcal{B}_{Y|w} \coloneqq \left\{\kappa_Y(\cdot, Y_i^{(w)}) - \hat{\mu}_{Y|w} : i \in \mathbb{N}_{n_w}\right\}.$$

For $i \in \mathbb{N}_{n_w}$, let $b_{X|w}^{(i)} \coloneqq \kappa_X(\cdot, X_i^{(w)}) - \hat{\mu}_{X|w}$ and $b_{Y|w}^{(i)} \coloneqq \kappa_Y(\cdot, Y_i^{(w)}) - \hat{\mu}_{Y|w}$ giving that $\mathcal{B}_{X|w} = \left\{b_{X|w}^{(1)}, \ldots, b_{X|w}^{(n_w)}\right\}$ and $\mathcal{B}_{Y|w} = \left\{b_{Y|w}^{(1)}, \ldots, b_{Y|w}^{(n_w)}\right\}$.

Theorem 4.5.16. *The following coordinate representations hold:*

$$\left[\hat{\Sigma}_{XX|w}\right]_{\mathcal{B}_{X|w}}^{\mathcal{B}_{X|w}} = \frac{1}{n_w}\mathcal{G}_{\mathcal{B}_{X|w}}$$
$$\left[\hat{\Sigma}_{YY|w}\right]_{\mathcal{B}_{Y|w}}^{\mathcal{B}_{Y|w}} = \frac{1}{n_w}\mathcal{G}_{\mathcal{B}_{Y|w}}$$
$$\left[\hat{\Sigma}_{XY|w}\right]_{\mathcal{B}_{Y|w}}^{\mathcal{B}_{X|w}} = \frac{1}{n_w}\mathcal{G}_{\mathcal{B}_{Y|w}}$$
$$\left[\hat{\Sigma}_{XX|w}\right]_{\mathcal{B}_{X|w}}^{\mathcal{B}_{Y|w}} = \frac{1}{n_w}\mathcal{G}_{\mathcal{B}_{X|w}}.$$

Notice now that for $i, j \in \mathbb{N}_{n_w}$:

$$\begin{aligned}\left\langle b_{X|w}^{(i)}, b_{X|w}^{(j)}\right\rangle_{\mathcal{H}_X} &= \left\langle \kappa_X(\cdot, X_i^{(w)}) - \hat{\mu}_{X|w}, \kappa_X(\cdot, X_j^{(w)}) - \hat{\mu}_{X|w}\right\rangle_{\mathcal{H}_X} \\ &= \kappa_X(X_i^{(w)}, X_j^{(w)}) - \frac{1}{n_w}\sum_{l=1}^{n_w}\kappa_X(X_i^{(w)}, X_l^{(w)}) \\ &\quad - \frac{1}{n_w}\sum_{k=1}^{n}\kappa_X(X_j^{(w)}, X_k^{(w)}) + \frac{1}{n_w^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\kappa_X(X_k^{(w)}, X_l^{(w)}).\end{aligned}$$

Let $m \in \mathbb{N}$, I_m be the $m \times m$ identity matrix, and let $K_{X|w}$ be the $n_w \times n_w$ matrix whose (i, j)-th entry is $\kappa_X(X_i^{(w)}, X_j^{(w)})$. Let $Q_w \coloneqq I_{n_w} - \frac{1}{n_w}\mathbf{1}_{n_w}\mathbf{1}_{n_w}^T$

where $\mathbf{1}_{n_w}$ is the n_w-dimensional vector whose components are all equal to 1. The above is the (i, j)-th entry of $Q_w K_{X|w} Q_w$, hence $\mathcal{G}_{\mathcal{B}_{X|w}} = Q_w K_{X|w} Q_w$. Similarly, $\mathcal{G}_{\mathcal{B}_{Y|w}} = Q_w K_{Y|w} Q_w$ where $K_{Y|w}$ is the matrix whose (i, j)-th entry is $\kappa_Y(Y_i^{(w)}, Y_j^{(w)})$.

Let $\hat{A}_w$ be an invertible operator from $\overline{\text{Ran}}\left(\hat{\Sigma}_{YY|w}\right)$ to $\overline{\text{Ran}}\left(\hat{\Sigma}_{YY|w}\right)$. By applying Theorem 4.5.15 and Theorem 4.5.16, it is seen that, ignoring constants, the empirical form of the optimisation problem (4.2) is given by:

$$\begin{aligned} &\text{Maximise } [f]^T_{\mathcal{B}_{X|w}} \mathcal{G}_{\mathcal{B}_{X|w}} \mathcal{G}_{\mathcal{B}_{Y|w}} \left[\hat{A}_w\right]^{\mathcal{B}_{Y|w}}_{\mathcal{B}_{Y|w}} \mathcal{G}_{\mathcal{B}_{X|w}} [f]_{\mathcal{B}_{X|w}} \\ &\text{Subject to } [f]^T_{\mathcal{B}_{X|w}} \mathcal{G}^2_{\mathcal{B}_{X|w}} [f]_{\mathcal{B}_{X|w}} = 1, \quad f \perp \hat{\mathcal{S}}_{k-1} \end{aligned} \tag{4.3}$$

where $\hat{\mathcal{S}}_0 \coloneqq \emptyset$, $\hat{\mathcal{S}}_k \coloneqq \text{Span}\{f_1, \ldots, f_k\}$ for $k \in \mathbb{N}$, and $f_1, \ldots, f_k$ are the k previous solutions to this constrained optimisation problem. Let $v \coloneqq \mathcal{G}_{\mathcal{B}_{X|w}} [f]_{\mathcal{B}_{X|w}}$ so $[f]_{\mathcal{B}_{X|w}} = \mathcal{G}^{\dagger}_{\mathcal{B}_{X|w}} v$. The optimisation problem (4.3) becomes:

$$\begin{aligned} &\text{Maximise } v^T \left(\mathcal{G}^{\dagger}_{\mathcal{B}_{X|w}} \mathcal{G}_{\mathcal{B}_{X|w}} \mathcal{G}_{\mathcal{B}_{Y|w}} \left[\hat{A}_w\right]^{\mathcal{B}_{Y|w}}_{\mathcal{B}_{Y|w}} \mathcal{G}_{\mathcal{B}_{X|w}} \mathcal{G}^{\dagger}_{\mathcal{B}_{X|w}}\right) v \\ &\text{Subject to } v^T v = 1, \quad v^T v_i = 0, \quad i \in \mathbb{N}_{k-1}. \end{aligned} \tag{4.4}$$

where $v_1, \ldots, v_{k-1}$ are the $k-1$ previous solutions to (4.4). Alike in Li (2018), replace the Moore-Penrose inverse $\mathcal{G}^{\dagger}_{\mathcal{B}_{X|w}}$ by the Tychonoff-regularised inverse $\left(\mathcal{G}_{\mathcal{B}_{X|w}} + \eta_{X|w} I_{n_w}\right)^{-1}$, where $\eta_{X|w} > 0$ is a tuning constant, in order to control overfitting. To endow $\eta_{X|w}$ with appropriate scale, let $\eta_{X|w} = \epsilon_{X|w} \lambda_{\max}\left(\mathcal{G}_{\mathcal{B}_{X|w}}\right)$ where $\lambda_{\max}\left(\mathcal{G}_{\mathcal{B}_{X|w}}\right)$ is the largest eigenvalue of $\mathcal{G}_{\mathcal{B}_{X|w}}$. For simplicity, let $\epsilon_{X|w} = \epsilon_X$ for each $w \in \mathcal{T}$. This then gives that v_i is the i^{th} eigenvector, corresponding to the i^{th} largest eigenvalue, of the matrix:

$$\Lambda_w \coloneqq \left(\mathcal{G}_{\mathcal{B}_{X|w}} + \eta_{X|w} I_{n_w}\right)^{-1} \mathcal{G}_{\mathcal{B}_{X|w}} \mathcal{G}_{\mathcal{B}_{Y|w}} \left[\hat{A}_w\right]^{\mathcal{B}_{Y|w}}_{\mathcal{B}_{Y|w}} \mathcal{G}_{\mathcal{B}_{X|w}} \left(\mathcal{G}_{\mathcal{B}_{X|w}} + \eta_{X|w} I_{n_w}\right)^{-1}.$$

With some choices of $\hat{A}_w$ (say $\hat{A}_w = \hat{\Sigma}^{-1}_{YY|w}$), $\mathcal{G}^{\dagger}_{\mathcal{B}_{Y|w}}$ appears in the coordinate representation. If this occurs when Y is categorical, it is not replaced with a

Tychonoff-regularisation. If, however, it occurs when Y is a random vector then replace it with $\left(\mathcal{G}_{\mathcal{B}_{Y|w}} + \eta_{Y|w} I_{n_w}\right)^{-1}$ where $\eta_{Y|w}$ is analogous to $\eta_{X|w}$ above.

In order to be able to merge these matrices, which are $n_w \times n_w$, assume that, for $w \in \mathcal{T}$, $n_w = c > 0$ where c is some constant. Ways to handle when this does not hold, using the SMOTE algorithm, are discussed in Section 4.5.7. These operators are then merged as follows:

$$\Lambda^{(W)} := \frac{1}{C} \sum_{w \in \mathcal{T}} \Lambda_w$$

The partial generalised sliced inverse regression procedure is then summarised as follows.

1. [Optional] Standardise $X_1, \ldots, X_n$ marginally. If Y is a random vector, then also marginally standardise $Y_1, \ldots, Y_n$.

2. From each of the larger categories, randomly sample c observations where c is the number of observations in the smallest observed category.

3. Choose kernel functions κ_X and κ_Y. If Y is categorical, say with $\Omega_Y = \mathbb{N}_D$ where $D \in \mathbb{N}$ and the integers represent the categorical labels, then the discrete kernel $\kappa_Y(i, j) = \delta_{ij}$ should be used. If κ_X, or κ_Y, is chosen to be the Gaussian radial basis function, then choose the kernel parameter γ_X (respectively γ_Y) according to the procedure described in Section 13.7 of Li (2018).

4. For each $w \in \mathcal{T}$, specify an invertible operator $\hat{A}_w$ from $\overline{\text{Ran}}\left(\hat{\Sigma}_{YY|w}\right)$ to $\overline{\text{Ran}}\left(\hat{\Sigma}_{YY|w}\right)$ whose coordinate representation $\left[\hat{A}_w\right]_{\mathcal{B}_{Y|w}}^{\mathcal{B}_{Y|w}}$ is known.

5. Choose the tuning parameter ϵ_X and, if needed, also choose ϵ_Y according to the procedure described in Section 13.7 of Li (2018).

6. For $w \in \mathcal{T}$, calculate $\mathcal{G}_{\mathcal{B}_{X|w}} = QK_{X|w}Q$ and $\mathcal{G}_{\mathcal{B}_{Y|w}} = QK_{Y|w}Q$, where $Q = I_c - \frac{1}{c}\mathbf{1}_c\mathbf{1}_c^T$.

7. For $w \in \mathcal{T}$, find

$$\Lambda_w = \left(\mathcal{G}_{\mathcal{B}_{X|w}} + \eta_{X|w}I_c\right)^{-1} \mathcal{G}_{\mathcal{B}_{X|w}}\mathcal{G}_{\mathcal{B}_{Y|w}} \left[\hat{A}_w\right]_{\mathcal{B}_{Y|w}}^{\mathcal{B}_{Y|w}} \mathcal{G}_{\mathcal{B}_{X|w}} \left(\mathcal{G}_{\mathcal{B}_{X|w}} + \eta_{X|w}I_c\right)^{-1}.$$

8. Find

$$\Lambda^{(W)} = \frac{1}{C}\sum_{w\in\mathcal{T}} \Lambda_w$$

9. Find the first d eigenvectors $v_1, \ldots, v_d$ of $\Lambda^{(W)}$ corresponding to the d largest eigenvalues.

10. Calculate the i^{th} sufficient predictor as $v_i^T(\mathcal{G}_{\mathcal{B}_{X|w}} + \eta_{X|w}I_c)^{-1}QK_{X|w}$.

4.5.7 Application to two real-world datasets

In this subsection, the results of partial generalised sliced inverse regression are compared to those for its marginal counterpart on two real datasets from the Machine Learning repository of UC Irvine. These are the abalone dataset and the autoMPG dataset. For both datasets, it seen that generalised sliced inverse regression captures something meaningful. The advantage of partial generalised sliced invere regression is that it captures the differences between the different values for the categorical predictors. When there were differing numbers of observations within each category, c observations are randomly sampled from each of the larger categories where c is the number of observations in the smallest observed category. The operators $\hat{A}_w$ are all taken to be the identity.

4.5.7.1 Abalone dataset

The predictors X in the abalone dataset consist of observations of 7 physical characteristics for 4177 abalones. These characteristics are length, diameter,

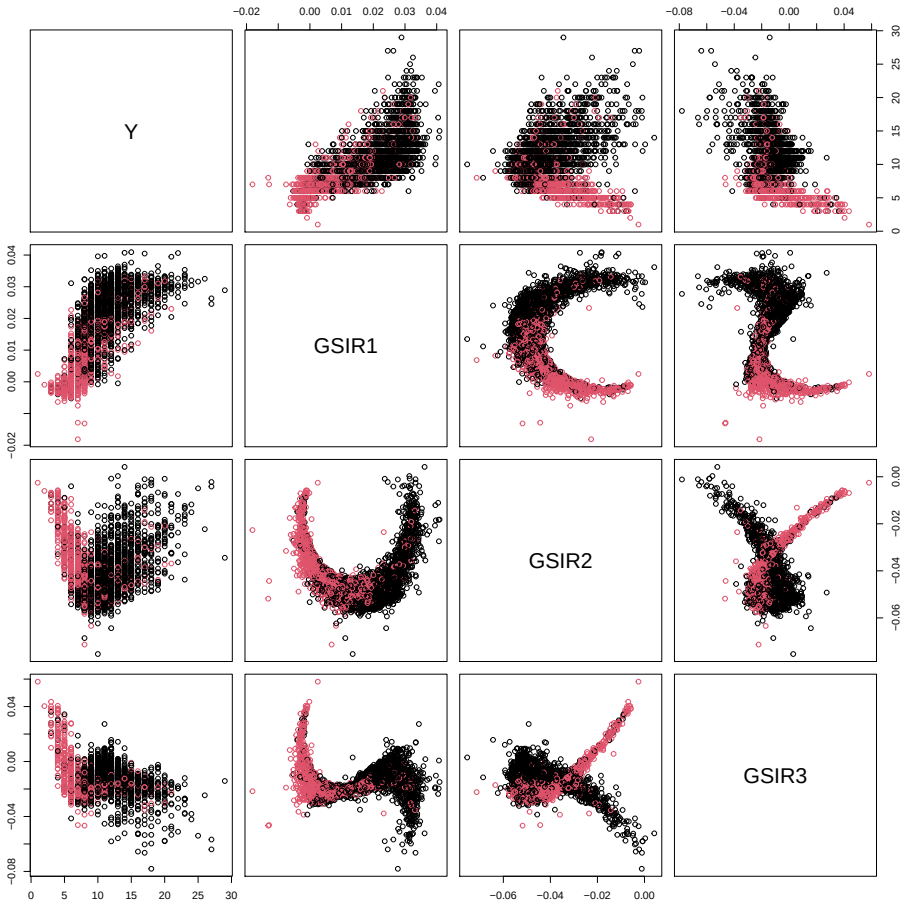

Figure 4.1: Generalised sliced inverse regression applied to the abalone dataset when males and females are merged into adults. Red represents the infants and black represents the adults.

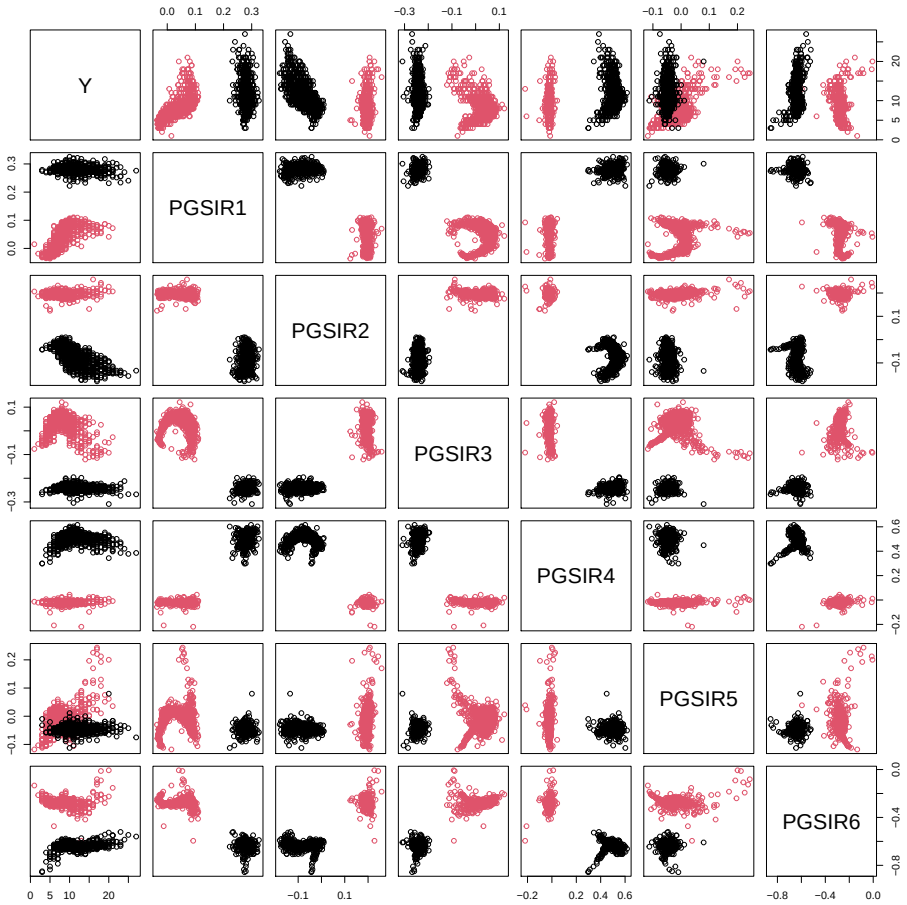

Figure 4.2: Partial generalised sliced inverse regression applied to the abalone dataset when males and females are merged into adults. Red represents the infants and black represents the adults.

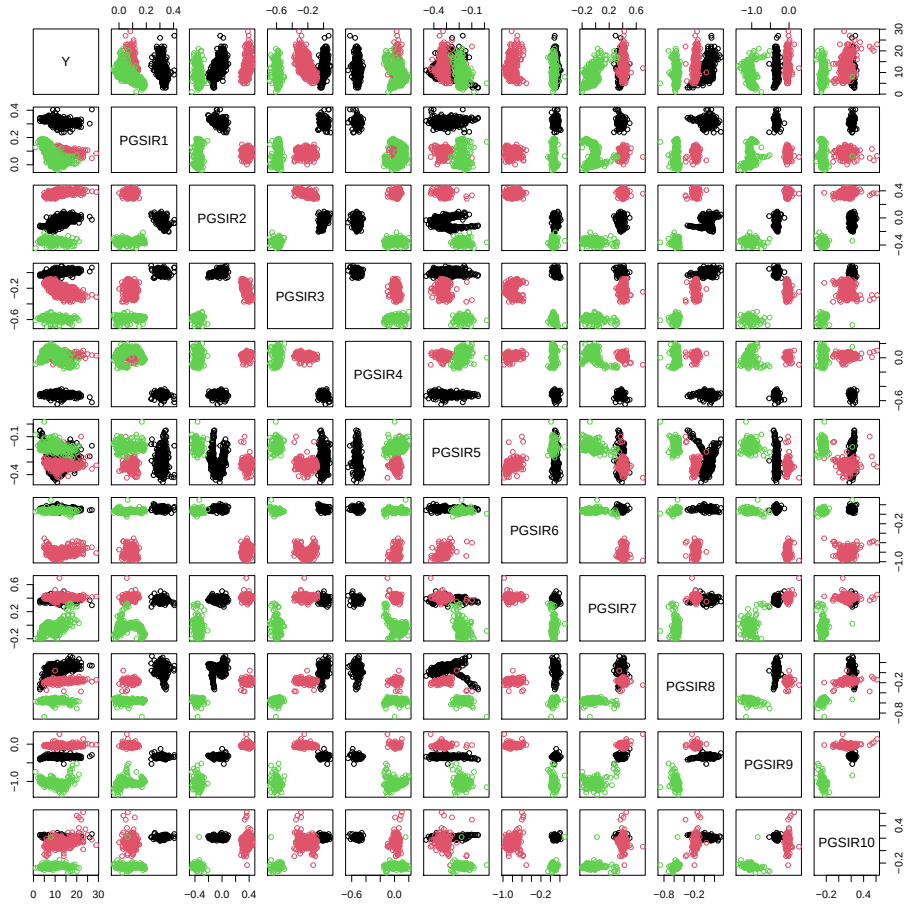

Figure 4.3: Partial generalised sliced inverse regression applied to the abalone dataset when all three categories are included.

height, weight, shell weight, gut weight, and meat weight. The response variable Y is the number of rings of the abalone, which is known to be related with the age of the abalone (age in years is equal to the number of rings plus 1.5). The categorical predictor W represents the sex. There are three categories: male, female and infant. For interpretability purposes, the males and females are combined into one category called adults. Figure 4.1 shows that the first direction of generalised sliced inverse regression captures a relatively linear relationship with the number of rings, while the other 2 capture some nonlinear relationships.

Adults are colored in black and infants in red. As expected, as W is not used, the two groups of points overlap. In Figure 4.2, the results obtained by partial generalised sliced inverse regression are shown. It is seen that the first direction captures the variability for the infants while the second direction captures it for the adults. Furthermore, the curvature for the infants is captured in the third direction, while the fourth direction captures the curvature for the adults. For comparison, the results of running partial generalised sliced inverse regression using all three categories (male, female, and infant) are shown in Figure 4.3. It is seen that the first three directions capture variability in each class, while the next three directions capture the curvature.

4.5.7.2 AutoMPG dataset

The predictors X in the autoMPG dataset consist of 5 variables for 398 cars. These are the number of cylinders (treated as a numeric variable rather than a categorical one), the displacement, the horsepower, the weight, and the acceleration. The response Y is the fuel consumption in miles per gallon (mpg). There are two categorical predictors: the location of manufacture (labelled 1, 2, and 3) and the year of manufacture (which ranges from 1970 to 1982). Here, W is taken to be only the location of manufacture.

As can be seen in Figure 4.4, generalised sliced inverse regression captures an almost linear relationship between the the first direction and the response, and captures nonlinear relationships with the second and third. As W is not considered, there is, as expected, overlap between the three locations of manufacture. The results of applying partial generalised sliced inverse regression are seen in Figure 4.5. The first 5 directions are presented. The first direction clearly separates the 3 locations of interest. The others, except direction 4 which is similar to direction 1, reorder the three groups in different arrangements.

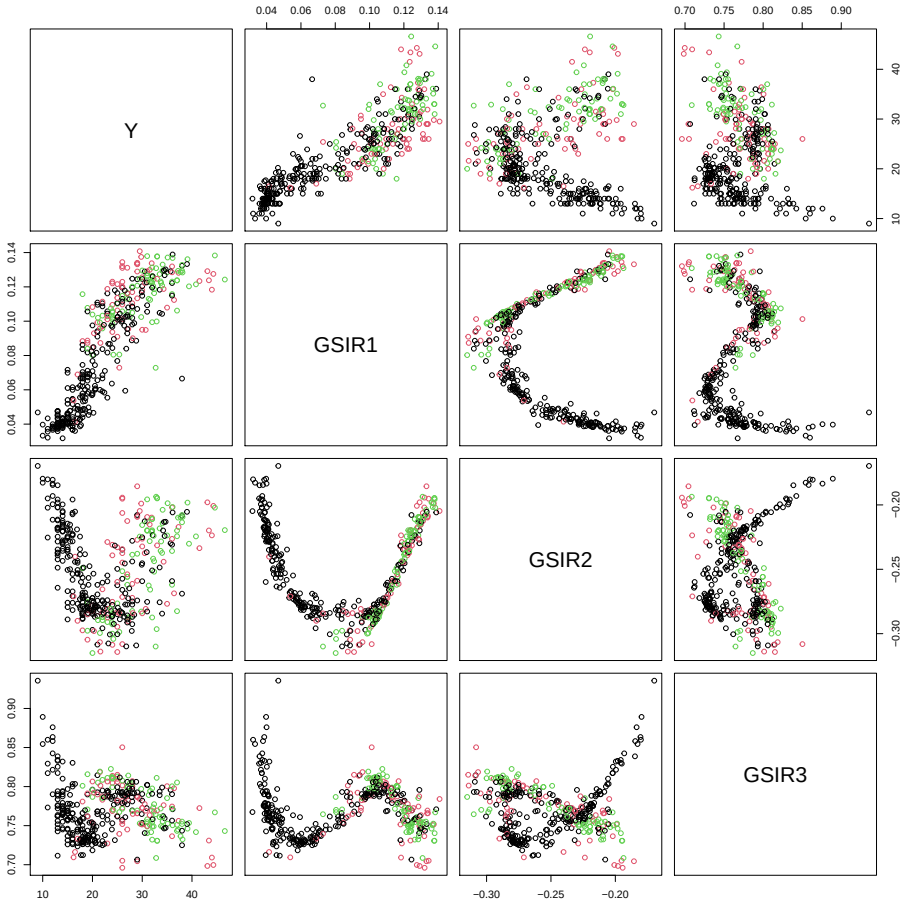

Figure 4.4: Generalised sliced inverse regression applied to the autoMPG dataset.

4.5.8 Summary

In this section, the general theory of nonlinear sufficient dimension reduction as developed by Lee et al. (2013), Li and Song (2017), and Li (2018) was extended to the setting where some of the predictors are categorical. This was accomplished by defining marginal, partial, and w-conditional sufficient dimension reduction in a measure-theoretic manner and exploring their relationships. A new estimator, partial generalised sliced inverse regression, was proposed. The effectiveness of this estimator was seen in practice on two real-world datasets.

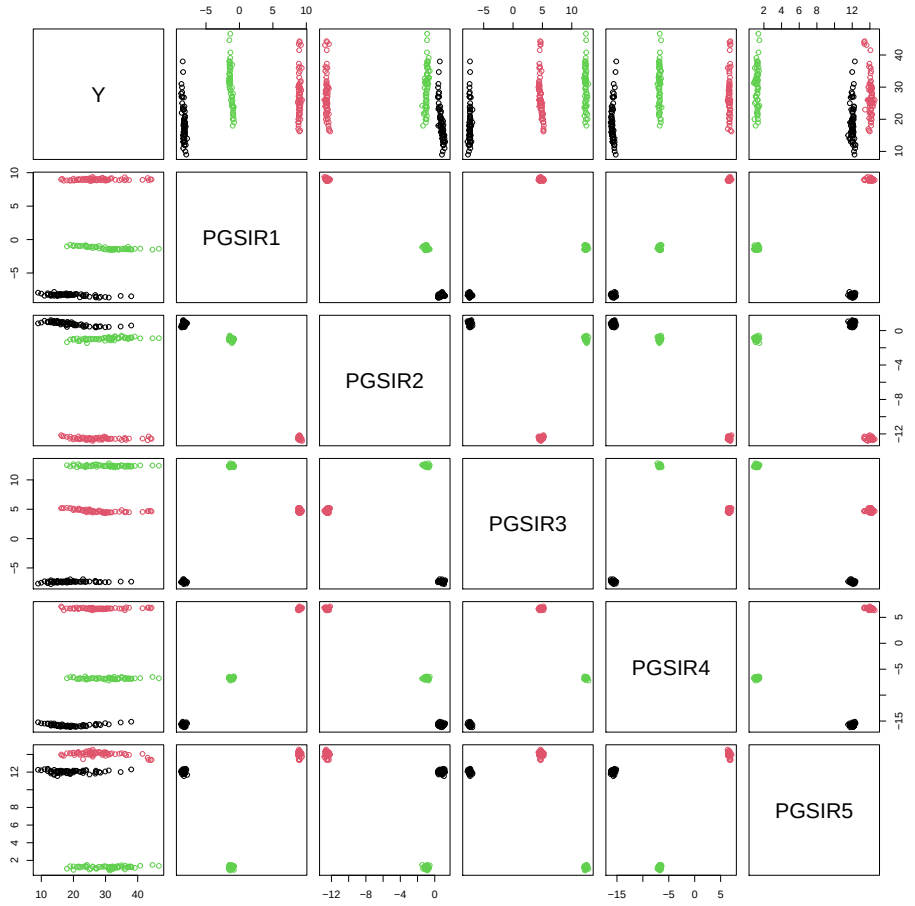

Figure 4.5: Partial generalised sliced inverse regression applied to the autoMPG dataset.

4.5.9 Proofs

Proof of Lemma 4.5.1. As $G \trianglelefteq (G, \sigma(W))$, $\mathbb{P}_w(A|G)$ is $(G, \sigma(W))$-measurable. It is required to show that for $H \in (G, \sigma(W))$,

$$\mathbb{E}_{\mathbb{P}_w}(\mathbf{1}_H \mathbb{P}_w(A|G)) = \mathbb{E}_{\mathbb{P}_w}(\mathbf{1}_H \mathbf{1}_A).$$

As $(G, \sigma(W))$ is generated by the π-system $P \coloneqq \{C \cap D : C \in G, D \in \sigma(W)\}$, it suffices to show the above for $H \in P$. Observe now that $\sigma(W) = \{\emptyset\} \cup \{\bigcup_{i \in I} E_i : I \in \mathcal{P}(\mathbb{N}_C) \setminus \{\emptyset\}\}$. This gives three cases: (1) $D = \emptyset$, (2) $D = \bigcup_{i \in I} E_i$

for some $I \in \mathcal{P}(\mathbb{N}_C) \setminus \{\emptyset\}$ and $w \notin I$, and (3) $D = \bigcup_{i \in I} E_i$ for some $I \in \mathcal{P}(\mathbb{N}_C) \setminus \{\emptyset\}$ and $w \in I$. The result is trivial in the first case as both sides equal 0. For the other cases, let $H = C \cap D$ for some $C \in G$ and $D \in \sigma(W) \setminus \{\emptyset\}$ and consider that:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_w}\left(\mathbf{1}_H \mathbb{P}_w(A|G)\right) &= \mathbb{E}_{\mathbb{P}_w}\left(\mathbf{1}_C \mathbf{1}_D \mathbb{E}_{\mathbb{P}_w}(\mathbf{1}_A|G)\right) \\
&= \mathbb{E}_{\mathbb{P}_w}\left(\mathbf{1}_C \mathbb{E}_{\mathbb{P}_w}(\mathbf{1}_D|G)\,\mathbf{1}_A\right) \\
&= \mathbb{E}_{\mathbb{P}_w}\left(\mathbf{1}_C \left(\frac{\mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_D \mathbf{1}_{E_w}|G\right)}{\mathbb{E}_{\mathbb{P}}\left(\mathbf{1}_{E_w}|G\right)}\right)\mathbf{1}_A\right).
\end{aligned}$$

The second equality follows from Lemma 2.5.4 and the fact that $\mathbf{1}_C$ is G-measurable as $C \in G$. In the second case, $D \cap E_w = \emptyset$ so this expectation is equal to 0. Furthermore, $\mathbb{P}_w(A \cap C \cap D) = 0$ as $\mathbb{P}(A \cap C \cap D \cap E_w) = \mathbb{P}(\emptyset) = 0$. In the final case, $D \cap E_w = E_w$ so the above simplifies to $\mathbb{E}_{\mathbb{P}_w}(\mathbf{1}_C \mathbf{1}_A)$. This equals $\mathbb{P}_w(A \cap C)$. Now this is equal to $\mathbb{P}_w(A \cap C \cap D)$ as $\mathbb{P}(A \cap C \cap E_w) = \mathbb{P}(A \cap C \cap D \cap E_w)$. □

Proof of Corollary 4.5.2. For any $w \in \mathbb{N}_C$, $\mathbf{1}_{E_w}$ is $(G, \sigma(W))$-measurable as it is $\sigma(W)$-measurable because $E_w \in \sigma(W)$. This implies that

$$\mathbb{P}(A \cap E_w|G, W) \overset{\text{a.s.}\mathbb{P}}{=} \mathbf{1}_{E_w} \mathbb{P}(A|G, W)$$

and $\mathbb{P}(E_w|G, W) \overset{\text{a.s.}\mathbb{P}}{=} \mathbf{1}_{E_w}$. As these hold almost surely $\mathbb{P}$ and $\mathbb{P}_w \ll \mathbb{P}$, they also hold almost surely $\mathbb{P}_w$ for any $w \in \mathbb{N}_C$. Applying Lemma 4.5.1 and taking the ratio (noting that $\mathbb{P}_w\left(\{\omega \in \Omega : \mathbf{1}_{E_w}(\omega) = 0\}\right) = 0$) gives the result. □

Proof of Lemma 4.5.3. Let $A \in \sigma(Y)$ and $B \in \sigma(X)$. Suppose that G is a w-conditional SDR σ-field for each $w \in \mathbb{N}_C$. Applying Corollary 4.5.2 gives that

$$\mathbb{P}_w(A \cap B|G) \overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{P}_w(A|G)\,\mathbb{P}_w(B|G)$$

can be rewritten as

$$\mathbb{P}(A \cap B|G, W) \overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{P}(A|G, W)\,\mathbb{P}(B|G, W).$$

As this holds for each $w \in \mathbb{N}_C$, this implies that

$$\mathbb{P}(A \cap B|G, W) \overset{\text{a.s.}\mathbb{P}}{=} \mathbb{P}(A|G, W)\,\mathbb{P}(B|G, W).$$

G is therefore a partial SDR σ-field. For the converse, suppose G is a partial SDR σ-field so $\mathbb{P}(A \cap B|G, W) \overset{\text{a.s.}\mathbb{P}}{=} \mathbb{P}(A|G, W)\,\mathbb{P}(B|G, W)$. As this holds almost surely $\mathbb{P}$, it also holds almost surely $\mathbb{P}_w$ (owing to the fact that $\mathbb{P}_w \ll \mathbb{P}$) for any $w \in \mathbb{N}_C$. Applying Corollary 4.5.2 yields the result. □

Proof of Theorem 4.5.4. Showing that $(\mathcal{M}_1, \ldots, \mathcal{M}_C)$ is a partial SDR σ-field implies that $\mathcal{M}^{(W)} \trianglelefteq (\mathcal{M}_1, \ldots, \mathcal{M}_C)$, yielding the result. For each $w \in \mathbb{N}_C$, it holds that $\mathcal{M}_w \trianglelefteq (\mathcal{M}_1, \ldots, \mathcal{M}_C) \trianglelefteq \sigma(X)$ so, by Corollary 2.5.11 with $\overset{w}{\perp\!\!\!\perp}$ replacing $\perp\!\!\!\perp$, $(\mathcal{M}_1, \ldots, \mathcal{M}_C)$ is a w-conditional SDR σ-field. Applying Lemma 4.5.3 yields the result. □

Proof of Theorem 4.5.5. It suffices to show that $W \perp\!\!\!\perp X|\mathcal{M}^{(W)}$ implies $\mathcal{M} \trianglelefteq \mathcal{M}^{(W)}$ as the other implication is similarly shown by interchanging the roles of X and Y. In Lemma 2.5.9, let $\mathcal{F}_1 = \sigma(W), \mathcal{F}_2 = \sigma(X), \mathcal{F}_3 = \sigma(Y), \mathcal{F}_4 = \mathcal{M}^{(W)}$. The statement in the lemma becomes

$$\left[W \perp\!\!\!\perp X|\left(Y, \mathcal{M}^{(W)}\right)\right] \wedge \left[X \perp\!\!\!\perp Y|\mathcal{M}^{(W)}\right]$$

is equivalent to

$$\left[W \perp\!\!\!\perp X|\mathcal{M}^{(W)}\right] \wedge \left[X \perp\!\!\!\perp Y|\left(W, \mathcal{M}^{(W)}\right)\right].$$

In the latter form, the first conjunct holds by assumption and the second holds by the definition of $\mathcal{M}^{(W)}$. This gives that the second conjunct in the first form holds. This is $X \perp\!\!\!\perp Y|\mathcal{M}^{(W)}$, which implies that $\mathcal{M} \trianglelefteq \mathcal{M}^{(W)}$. □

Proof of Theorem 4.5.6. In Lemma 2.5.9, let $\mathcal{F}_1 = \sigma(W), \mathcal{F}_2 = \sigma(Y), \mathcal{F}_3 = \sigma(X), \mathcal{F}_4 = \mathcal{M}$. The statement in the lemma becomes:

$$[W \perp\!\!\!\perp Y|(X, \mathcal{M})] \wedge [Y \perp\!\!\!\perp X|\mathcal{M}]$$

is equivalent to:

$$[W \perp\!\!\!\perp Y|\mathcal{M}] \wedge [Y \perp\!\!\!\perp X|\,(W, \mathcal{M})]\,.$$

As $\mathcal{M} \trianglelefteq \sigma\,(X)$, it holds that $(X, \mathcal{M}) = \sigma\,(X)$. This implies that the first conjunct in the first form is the same as $W \perp\!\!\!\perp Y|X$, which holds by the assumption. The second conjunct in the first form holds by the definition of $\mathcal{M}$. Thus, the second conjunct in the second form holds. This is $Y \perp\!\!\!\perp X|\,(W, \mathcal{M})$. By the definition of $\mathcal{M}^{(W)}$, this gives that $\mathcal{M}^{(W)} \trianglelefteq \mathcal{M}$. □

Proof of Theorem 4.5.7. Note that $\left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right) \trianglelefteq \sigma\,(X)$ as $\mathcal{M}_{(W)} \trianglelefteq \sigma\,(X)$ and $\mathcal{M}^{(W)} \trianglelefteq \sigma\,(X)$. Showing that $Y \perp\!\!\!\perp X|\left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right)$ suffices because of the definition of $\mathcal{M}$. By the definition of $\mathcal{M}_{(W)}$, $X \perp\!\!\!\perp W|\mathcal{M}_{(W)}$. As $\mathcal{M}_{(W)} \trianglelefteq \left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right) \trianglelefteq \sigma\,(X)$, applying Corollary 2.5.11 with $\mathcal{F}_1 = \sigma\,(W)\,, \mathcal{F}_2 = \sigma\,(X)\,, \mathcal{F}_3 = \mathcal{M}_{(W)}$, and $\mathcal{F}_4 = \left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right)$ implies that X is conditionally independent of W given $\left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right)$. Now by the definition of $\mathcal{M}^{(W)}$, it is known that $Y \perp\!\!\!\perp X|\left(\mathcal{M}^{(W)}, W\right)$. Note that $\mathcal{M}^{(W)} \trianglelefteq \left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right) \trianglelefteq \sigma\,(X)$ also holds. Applying Lemma 2.5.10 with $\mathcal{F}_1 = \sigma\,(Y)\,, \mathcal{F}_2 = \sigma\,(X)\,, \mathcal{F}_3 = \sigma\,(W)\,, \mathcal{F}_4 = \mathcal{M}^{(W)}$, and $\mathcal{F}_5 = \left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right)$ gives that

$$Y \perp\!\!\!\perp X|\left(W, \left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right)\right).$$

In the statement of Lemma 2.5.9, let $\mathcal{F}_1 = \sigma\,(Y)\,, \mathcal{F}_2 = \sigma\,(X)\,, \mathcal{F}_3 = \sigma\,(W)$, and $\mathcal{F}_4 = \left(\mathcal{M}_{(W)}, \mathcal{M}^{(W)}\right)$. The conditional independences just established form the conjuncts of the first form of the equivalent statement, thus the second form must hold. The desired result is the first conjunct in this second form. □

Proof of Lemma 4.5.8. Let $G \trianglelefteq \sigma\,(X)$. Consider first $\mathcal{H}_X\,(G)$. By Theorem 13.3 of Billingsley (1995), finite linear combinations of G-measurable real-valued functions are themselves G-measurable thus $\mathcal{H}_X(G)$ is closed under finite linear combinations. It remains to show that $\mathcal{H}_X(G)$ is topologically

closed. To this end, suppose that $\{f_i\}_{i\in\mathbb{N}} \subseteq \mathcal{H}_X(G)$ is a $\mathcal{H}_X$-norm convergent sequence with limit $f \in \mathcal{H}_X$. By Theorem 2.7.6 of Hsing and Eubank (2015), this implies that, for any $x \in \Omega_X$, $|f_i(x) - f(x)| \to 0$ as $i \to \infty$. Thus, for $\omega \in \Omega$, $|(f_i(X))(\omega) - (f(X))(\omega)| \to 0$ as $i \to \infty$. Applying Theorem 13.4 of Billingsley (1995) then gives that $f(X)$ is also G-measurable, hence $f \in \mathcal{H}_X(G)$ as desired. That $\mathcal{H}_X^G$ is a closed subspace of $\mathcal{H}_X$ is shown similarly. □

Proof of Lemma 4.5.9. Only the first relation is proven as the second follows similarly by interchanging X and Y. It suffices to show that

$$\overline{\text{Ran}\left(\Sigma_{XX|w}\right)}^{\perp} \subseteq \text{Ran}\left(\Sigma_{XY|w}\right)^{\perp}$$

This is equivalent to $\text{Ker}\left(\Sigma_{XX|w}\right) \subseteq \text{Ker}\left(\Sigma_{YX|w}\right)$. Let $f \in \text{Ker}\left(\Sigma_{XX|w}\right)$. This implies that $\text{Var}_{\mathbb{P}_w}(f(X)) = \left\langle f, \Sigma_{XX|w} f\right\rangle_{\mathcal{H}_X} = 0$ hence $f(X)$ is $\mathbb{P}_w$ almost surely constant. This implies that for any $g \in \mathcal{H}_Y$, $\text{Cov}_{\mathbb{P}_w}(f(X), g(Y)) = \left\langle \Sigma_{YX|w} f, g\right\rangle_{\mathcal{H}_Y} = 0$. Taking $g = \Sigma_{YX|w} f$ gives

$$\left\langle \Sigma_{YX|w} f, \Sigma_{YX|w} f\right\rangle_{\mathcal{H}_Y} = \left\|\Sigma_{YX|w} f\right\|_{\mathcal{H}_Y}^2 = 0.$$

Hence $\Sigma_{YX|w} f = 0$ so $f \in \text{Ker}\left(\Sigma_{YX|w}\right)$ as desired. □

Proof of Lemma 4.5.10. Only the first relation is shown as the second follows similarly. A member f of $\mathcal{H}_X$ is orthogonal to the subspace on the right-hand side if and only if $f \perp \kappa_X(\cdot, x) - \mu_X$ for all $x \in \Omega_X$. That is, $\langle f, \kappa_X(\cdot, x) - \mu_X\rangle_{\mathcal{H}_X} = 0$. By the reproducing property and the definition of μ_X, this is equivalent to $f(x) = \mathbb{E}_{\mathbb{P}}(f(X))$ for all $x \in \Omega_X$ which means that $\text{Var}_{\mathbb{P}}(f(X)) = 0$ hence $\Sigma_{XX} f = 0$. Thus

$$\text{Ker}\left(\Sigma_{XX}\right) = \overline{\text{Span}}\left\{\kappa_X(\cdot, x) - \mu_X : x \in \Omega_X\right\}^{\perp}.$$

which implies the desired equality because Σ_{XX} is self-adjoint. □

Proof of Lemma 4.5.11. Let $w \in \mathbb{N}_C$ and $f \in \mathcal{H}_X$. It is first shown that, for $g \in L^2\left(\mathbb{P}_{Y|w}\right)$,

$$\mathrm{Cov}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(f(X)|Y\right) - \left(R_{YX|w}f\right)(Y), g(Y)\right) = 0.$$

As $\left(R_{YX|w}f\right)(Y)$ is $\sigma(Y)$-measurable, this can be rewritten as

$$\mathrm{Cov}_{\mathbb{P}_w}\left(f(X) - \left(R_{YX|w}f\right)(Y), g(Y)\right) = 0. \tag{4.5}$$

As $\mathcal{H}_Y$ is dense modulo $\mathbb{P}_{Y|w}$ almost sure constants in $L^2\left(\mathbb{P}_{Y|w}\right)$ (see Theorem 2.5.19), there is a sequence $\{g_k\}_{k\in\mathbb{N}} \subseteq \mathcal{H}_Y$ such that $\mathrm{Var}_{\mathbb{P}_w}(G_k(Y)) \to 0$ as $k \to \infty$ where $G_k(Y) \coloneqq g(Y) - g_k(Y)$. With any such sequence

$$\begin{aligned}\mathrm{Cov}_{\mathbb{P}_w}\left(\left(R_{YX|w}f\right)(Y), g(Y)\right) = {} & \mathrm{Cov}_{\mathbb{P}_w}\left(\left(R_{YX|w}f\right)(Y), g_k(Y)\right) \\ & + \mathrm{Cov}_{\mathbb{P}_w}\left(\left(R_{YX|w}f\right)(Y), G_k(Y)\right)\end{aligned} \tag{4.6}$$

The first term on the right-hand side of (4.6) is equal to $\left\langle R_{YX|w}f, \Sigma_{YY|w}g_k\right\rangle_{\mathcal{H}_Y}$. Now, as $\Sigma_{YY|w}$ is self-adjoint,

$$\begin{aligned}\left\langle R_{YX|w}f, \Sigma_{YY|w}g_k\right\rangle_{\mathcal{H}_Y} &= \left\langle \Sigma_{YX|w}f, g_k\right\rangle_{\mathcal{H}_Y} \\ &= \mathrm{Cov}_{\mathbb{P}_w}\left(f(X), g_k(Y)\right) \\ &= \mathrm{Cov}_{\mathbb{P}_w}\left(f(X), g(Y)\right) - \mathrm{Cov}_{\mathbb{P}_w}\left(f(X), G_k(Y)\right).\end{aligned}$$

Hence (4.6) can be rewritten as

$$\begin{aligned}\mathrm{Cov}_{\mathbb{P}_w}\left(\left(R_{YX|w}f\right)(Y), g(Y)\right) = {} & \mathrm{Cov}_{\mathbb{P}_w}\left(f(X), g(Y)\right) \\ & + \mathrm{Cov}_{\mathbb{P}_w}\left((R_{YX|w}f)(Y) - f(X), G_k(Y)\right).\end{aligned} \tag{4.7}$$

Now apply the Cauchy-Schwarz inequality to see that

$$\begin{aligned}\left|\mathrm{Cov}_{\mathbb{P}_w}\left((R_{YX|w}f)(Y) - f(X), G_k(Y)\right)\right| \leq {} & \sqrt{\mathrm{Var}_{\mathbb{P}_w}\left((R_{YX|w}f)(Y) - f(X)\right)} \\ & \times \sqrt{\mathrm{Var}_{\mathbb{P}_w}\left(G_k(Y)\right)}\end{aligned}$$

This converges to 0 as $k \to \infty$, hence the second term of (4.7) converges to 0. In the limit, this leaves

$$\mathrm{Cov}_{\mathbb{P}_w}\left(\left(R_{YX|w}f\right)(Y), g(Y)\right) = \mathrm{Cov}_{\mathbb{P}_w}\left(f(X), g(Y)\right)$$

This then implies (4.5), which yields that $\left(R_{YX|w}f\right)(Y) \overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{E}_{\mathbb{P}_w}\left(f(X)|Y\right) + c_f$ where c_f is a real constant. By taking the $\mathbb{P}_w$ expectation of both sides, it is seen that $c_f = \mathbb{E}_{\mathbb{P}_w}\left(\left(R_{YX|w}f\right)(Y)\right) - \mathbb{E}_{\mathbb{P}_w}\left(f(X)\right)$ which gives the result. □

Proof of Theorem 4.5.12. Let $\Sigma_{XX|w}\mathcal{H}_X\left(\mathcal{M}_w\right) \coloneqq \left\{\Sigma_{XX|w}f : f \in \mathcal{H}_X\left(\mathcal{M}_w\right)\right\}$. It is first shown that:

$$\overline{\mathrm{Ran}\left(\Sigma_{XY|w}\right)} \subseteq \Sigma_{XX|w}\mathcal{H}_X\left(\mathcal{M}_w\right). \tag{4.8}$$

This is equivalent to showing that $\left(\Sigma_{XX|w}\mathcal{H}_X\left(\mathcal{M}_w\right)\right)^{\perp} \subseteq \mathrm{Ker}\left(\Sigma_{YX|w}\right)$. Let $f \in \left(\Sigma_{XX|w}\mathcal{H}_X\left(\mathcal{M}_w\right)\right)^{\perp}$. Then, for all $g \in \mathcal{H}_X\left(\mathcal{M}_w\right)$,

$$\left\langle f, \Sigma_{XX|w}g\right\rangle_{\mathcal{H}_X} = \mathrm{Cov}_{\mathbb{P}_w}\left(f(X), g(X)\right) = 0.$$

As $g(X)$ is $\mathcal{M}_w$-measurable, $g(X) \overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{E}_{\mathbb{P}_w}\left(g(X)|\mathcal{M}_w\right)$. By Lemma 2.5.4,

$$\mathrm{Cov}_{\mathbb{P}_w}\left(f(X), \mathbb{E}_{\mathbb{P}_w}\left(g(X)|\mathcal{M}_w\right)\right) = \mathrm{Cov}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(f(X)|\mathcal{M}_w\right), g(X)\right).$$

Now as $\mathcal{H}_X\left(\mathcal{M}_w\right)$ is dense modulo $\mathbb{P}_{X|w}$ almost sure constants in $L^2\left(\mathbb{P}_{X|w}|\mathcal{M}_w\right)$, there exists a sequence $\{f_k\}_{k\in\mathbb{N}} \subseteq \mathcal{H}_X\left(\mathcal{M}_w\right)$ such that, as $k \to \infty$,

$$\mathrm{Var}_{\mathbb{P}_w}\left(f_k(X) - \mathbb{E}_{\mathbb{P}_w}\left(f(X)|\mathcal{M}_w\right)\right) \to 0.$$

Using this fact and that, for any $k \in \mathbb{N}$, $\mathrm{Cov}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(f(X)|\mathcal{M}_w\right), f_k(X)\right) = 0$, implies that

$$\begin{aligned}
&\mathrm{Cov}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(f(X)|\mathcal{M}_w\right), f_k(X)\right) \\
&\to \mathrm{Cov}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(f(X)|\mathcal{M}_w\right), \mathbb{E}_{\mathbb{P}_w}\left(f(X)|\mathcal{M}_w\right)\right) \\
&= \mathrm{Var}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(f(X)|\mathcal{M}_w\right)\right) = 0.
\end{aligned}$$

Hence $\mathbb{E}_{\mathbb{P}_w}(f(X)|\mathcal{M}_w) \overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{E}_{\mathbb{P}_w}(f(X))$. As $Y \overset{w}{\perp\!\!\!\perp} X|\mathcal{M}_w$, it holds that $\mathbb{E}_{\mathbb{P}_w}(f(X)|\mathcal{M}_w) \overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{E}_{\mathbb{P}_w}(f(X)|Y, \mathcal{M}_w)$. Because this is $\mathbb{P}_w$ almost surely constant, the tower property of conditional expectation implies that $\mathbb{E}_{\mathbb{P}_w}(f(X)|Y)$ is also $\mathbb{P}_w$ almost surely constant. By Lemma 4.5.11, this implies that $\Sigma^{-1}_{YY|w}\Sigma_{YX|w}f$ is $\mathbb{P}_{Y|w}$ almost surely constant. As $\operatorname{Ker}(\Sigma_{YY|w}) = \{0\}$, this gives that $\Sigma^{-1}_{YY|w}\Sigma_{YX|w}f = 0$ which implies that $f \in \operatorname{Ker}(\Sigma_{YX|w})$ as desired.

Now Equation (4.8) implies that

$$\operatorname{Ran}(\Sigma_{XY|w}) \subseteq \Sigma_{XX|w}\mathcal{H}_X(\mathcal{M}_w).$$

This then implies that

$$\Sigma^{-1}_{XX|w}\operatorname{Ran}(\Sigma_{XY|w}) \subseteq \mathcal{H}_X(\mathcal{M}_w).$$

Now consider that

$$\begin{aligned}\Sigma^{-1}_{XX|w}\operatorname{Ran}(\Sigma_{XY|w}) &= \left\{\Sigma^{-1}_{XX|w}f : f = \Sigma_{XY|w}g, g \in \mathcal{H}_Y\right\}\\ &= \left\{\Sigma^{-1}_{XX|w}\Sigma_{XY|w}g : g \in \mathcal{H}_Y\right\}\\ &= \operatorname{Ran}\left(\Sigma^{-1}_{XX|w}\Sigma_{XY|w}\right).\end{aligned}$$

Now as $\mathcal{H}_X(\mathcal{M}_w)$ is closed, taking the closure of this range gives the result. □

Proof of Theorem 4.5.13. As $\mathcal{H}_X^{(W)}$ is closed, it suffices to show that

$$\operatorname{Ran}(\Lambda_{\text{PGSIR}}) \subseteq \mathcal{H}_X^{(W)}.$$

As $\Lambda_{\text{PGSIR}} = \sum_{w\in\mathbb{N}_C}\Lambda_{w\text{-GSIR}}$, $\mathcal{M}^{(W)} = (\mathcal{M}_1, \ldots, \mathcal{M}_C)$, and by Theorem 4.5.12, it follows that

$$\begin{aligned}\operatorname{Ran}(\Lambda_{\text{PGSIR}}) &= \bigplus_{w\in\mathbb{N}_C}\operatorname{Ran}(\Lambda_{w\text{-GSIR}})\\ &\subseteq \bigplus_{w\in\mathbb{N}_C}\mathcal{H}_X(\mathcal{M}_w) \subseteq \mathcal{H}_X^{(W)}.\end{aligned}$$

□

Proof of Theorem 4.5.14. It suffices to show $\mathcal{H}_X(\mathcal{M}_w) \subseteq \overline{\text{Ran}\left(\Sigma_{XX|w}^{-1}\Sigma_{XY|w}\right)}$, which is equivalent to $\text{Ker}\left(\Sigma_{YX|w}\Sigma_{XX|w}^{-1}\right) \subseteq \mathcal{H}_X(\mathcal{M}_w)^\perp$. Let f be in the kernel of $\Sigma_{YX|w}\Sigma_{XX|w}^{-1}$. This implies that $(\Sigma_{YY|w}^{-1}\Sigma_{YX|w}\Sigma_{XX|w}^{-1}f)(Y) = 0$ hence, by Lemma 4.5.11, there is a real constant c_f such that $\mathbb{E}_{\mathbb{P}_w}\left(\left(\Sigma_{XX|w}^{-1}f\right)(X)\middle|Y\right) \overset{\text{a.s.}\mathbb{P}_w}{=} c_f$. By the tower property of conditional expectation and the definition of $\mathcal{M}_w$, it follows that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_w}\left(\left(\Sigma_{XX|w}^{-1}f\right)(X)\middle|Y\right)\\
&\overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{E}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(\left(\Sigma_{XX|w}^{-1}f\right)(X)\middle|Y,\mathcal{M}_w\right)\middle|Y\right)\\
&\overset{\text{a.s.}\mathbb{P}_w}{=} \mathbb{E}_{\mathbb{P}_w}\left(\mathbb{E}_{\mathbb{P}_w}\left(\left(\Sigma_{XX|w}^{-1}f\right)(X)\middle|\mathcal{M}_w\right)\middle|Y\right).
\end{aligned}$$

Now $\mathcal{M}_w$ is $\mathbb{P}_w$-complete for $\sigma(Y)$, so $\mathbb{E}_{\mathbb{P}_w}\left(\Sigma_{XX|w}^{-1}f(X)\middle|\mathcal{M}_w\right) \overset{\text{a.s.}\mathbb{P}_w}{=} c_f$. It follows that for $g \in \mathcal{H}_X(\mathcal{M}_w)$,

$$\begin{aligned}
\langle f, g\rangle_{\mathcal{H}_X} &= \left\langle \Sigma_{XX|w}^{-1}f, \Sigma_{XX|w}g\right\rangle_{\mathcal{H}_X}\\
&= \text{Cov}_{\mathbb{P}_w}\left(\left(\Sigma_{XX|w}^{-1}f\right)(X), g(X)\right)\\
&= \text{Cov}_{\mathbb{P}_w}\left(\left(\Sigma_{XX|w}^{-1}f\right)(X), \mathbb{E}_{\mathbb{P}_w}(g(X)|\mathcal{M}_w)\right)\\
&= \text{Cov}_{\mathbb{P}_w}\left(\left(\mathbb{E}_{\mathbb{P}_w}\left(\Sigma_{XX|w}^{-1}f\middle|\mathcal{M}_w\right)\right)(X), g(X)\right) = 0
\end{aligned}$$

Hence $f \in \mathcal{H}_X(\mathcal{M}_w)^\perp$ as desired. □

Proof of Theorem 4.5.16. The third coordinate representation is shown as the rest follow by similar reasoning. For $k \in \mathbb{N}_{n_w}$, let e_k be the vector with 1 in the k-th component and 0 elsewhere. Applying Theorem 4.5.15,

$$\begin{aligned}
\left[\hat{\Sigma}_{XY|w}\right]_{\mathcal{B}_{Y|w}}^{\mathcal{B}_{X|w}} &= \left[\frac{1}{n_w}\sum_{k=1}^{n_w}\left(b_{X|w}^{(k)} \otimes b_{Y|w}^{(k)}\right)\right]_{\mathcal{B}_{Y|w}}^{\mathcal{B}_{X|w}}\\
&= \frac{1}{n_w}\sum_{k=1}^{n_w}\left[b_{X|w}^{(k)} \otimes b_{Y|w}^{(k)}\right]_{\mathcal{B}_{Y|w}}^{\mathcal{B}_{X|w}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n_w} \sum_{k=1}^{n_w} \left[b_{X|w}^{(k)} \right]_{\mathcal{B}_{X|w}} \left[b_{Y|w}^{(k)} \right]_{\mathcal{B}_{Y|w}}^{T} \mathcal{G}_{\mathcal{B}_{Y|w}} \\
&= \frac{1}{n_w} \left(\sum_{k=1}^{n_w} e_k e_k^T \right) \mathcal{G}_{\mathcal{B}_{Y|w}} \\
&= \frac{1}{n_w} I_{n_w} \mathcal{G}_{\mathcal{B}_{Y|w}} \\
&= \frac{1}{n_w} \mathcal{G}_{\mathcal{B}_{Y|w}}.
\end{aligned}
$$

□

Chapter 5

Discussion

5.1 Summary of the contributions of this thesis

This thesis has provided a number of novel developments in Probability Theory proper, Functional Analysis, Banach/Hilbertian data analysis, the predictive potential of nonlinear principal components analysis for general predictors, and methodological developments in nonlinear sufficient dimension reduction when categorical predictors are present.

Chapter 2 gave definitions and results, some classical and others novel, spanning several branches of Mathematics. A major development in that chapter is the extension of many classical definitions in Probability Theory to allow for relevant sets to be uncountable. The novel results from that chapter are as follows:

1. Lemma 2.2.2 relates the independence of stochastic process with the random variable that defines its distribution.

2. Lemma 2.2.3 relates the distribution of an independent stochastic process with the product probability measure.

3. Lemma 2.2.4 says that i.i.d stochastic processes are exchangeable, using the provided generalised definitions of these notions.

4. Lemma 2.2.5 relates the independence (respectively identical distribution, exchangeability, contractability) of a particular kind of stochastic process with that for the process comprised of projections.

5. Theorem 2.2.6 gives a significant implication of the exchangeability or contractability of a stochastic process which is used to prove Theorem 2.2.11, a substantial generalisation of Lemma 3.1 from Artemiou and Li (2009) that is used in Chapter 3.

6. Existing literature on generalised notions of cumulative distribution functions was extended. Lemma 2.2.7 related two notions of continuity with each other. Theorem 2.2.8 gave some properties of a cumulative distribution function on a linearly ordered topological space. Generalised notions of joint cumulative distribution functions, quantile functions, and medians were provided. Theorem 2.2.10 gave a generalised version of of the probability integral transform, which was historically used for simulating random variables with a given distribution using standard uniform variates which were easier to generate. A generalised version of Sklar's Theorem, the basis for modelling multivariate dependencies with copulas, was conjectured in Conjecture 2.2.1.

7. A novel proof, using the arithmetic of cardinals, for Theorem 2.3.1 was provided.

8. Theorem 2.4.1 demonstrated the relationship between tensor products of closed ranges of bounded linear operators with the closed range of the tensor product of the operators. This result, to the author's knowledge,

has not been given in the Functional Analysis literature though it is not difficult to prove. Note that the notions of tensor products used here are not used in other chapters, but are nevertheless included as novel results were provided and they formed the basis for an earlier attempt at the novel work in Chapter 4.

9. Expectations, inner products, conditional expectations, tensor products, and conditional cross-covariance were related via Lemma 2.5.2, Lemma 2.5.3, Lemma 2.5.4, Corollary 2.5.5, and Lemma 2.5.6.

10. Some properties of conditional independence were given in Lemma 2.5.9, Lemma 2.5.10, and Corollary 2.5.11. These were useful for establishing the theoretical results in Section 4.5.3.

11. Theorem 2.5.21 relates conditional independence with the Hilbert-Schmidt conditional independence criterion.

Chapter 3 generalised the results of Jones and Artemiou (2019), Jones et al. (2020), and Jones and Artemiou (2021) to when nonlinear principal components analysis is applied with general predictors. It was established that, under some uniformity assumptions, higher-ranking nonlinear principal components tend (with probability exceeding $1/2$, with the probability being quantified when unitary invariance is used) to be more informative (in terms of conditional squared correlation) of a univariate measurable transformation of a response variable than the lower-ranking components. This was done under the conditional independence model and in a model-free setting.

Chapter 4 generalised the measure-theoretic framework for sufficient dimension reduction that was developed by Lee et al. (2013), Li and Song (2017), and Li (2018) to the setting where some categorical predictors are present. A new estimator was proposed and its properties and effectiveness were explored.

5.2 Ideas for further research

There are a wide variety of avenues for future research related to the work in this thesis, only a subset of which are included here.

Conjecture 2.2.1 could be proven. As noted in Remark 2.5.19, Theorem 2.5.21 could be used to (1) develop a test for conditional independence, (2) develop a novel nonlinear sufficient dimension reduction method, and (3) develop a procedure for determining the number of components to take from a nonlinear sufficient dimension reduction method.

The conjecture of Li (2007a) that, under some uniformity assumptions on the randomly chosen regression coefficients and the covariance matrix, the first principal component of some predictor vector X is the most likely, among all the principal components, to have the greatest absolute correlation with some univariate response Y remains unproven. Further research could examine this conjecture and generalisations thereof. Two possible generalisations that could be explored are: (1) the conjecture can be modified for when a nonlinear variant of principal components is used with general predictors, and (2) the conjecture can be extended to claim that the k^{th} principal component of X (either the classical or nonlinear version) is the k^{th} most likely to have the k^{th} largest absolute correlation (alternatively squared correlation) with some univariate response Y. Another avenue of investigation is to change the measure of how informative the principal components are of the response from conditional squared correlation to some other measure. Some developments along this line of thought have been given by Artemiou (2021) who used conditional mutual information to show similar results to the classical ones for the multivariate data setting, under the linear regression model with Gaussian predictors. The author of this thesis believes that a promising approach to make use of characterstic kernels to derive a measure which fully characterises the predictive power that the principal components

have for the transformation of the response. The predictive potential of other unsupervised dimension methods could also be explored.

In the kernel-based nonlinear sufficient dimension reduction framework, the suspected one-to-one correspondence between sub-σ-fields of the σ-field generated by the predictor and the closed subspaces of the form $\mathcal{H}_X(G)$ remains to be proven. It also remains to be shown that, under some assumptions, the PGSIR estimator is exhaustive. For any method based on covariance and cross-covariance operators, the norm is to take the classic unbiased estimator; alternative approaches based on shrinkage could be used and are discussed in Zhou et al. (2019). As the existing methods for nonlinear sufficient dimension reduction require eigendecomposition of an $n \times n$ (n being the sample size) matrix, they may be computationally expensive when n is large. To othercome this limitation, kernel approximation methods can be used: see Gauthier and Suykens (2018) and Hutchings and Gauthier (2022) for examples of such approximation approaches. It has been demonstrated in this thesis that measure-theoretic approaches to sufficient dimension reduction are fruitful in the setting where categorical predictors are present. It is reasonable to suspect that they would also be useful in other settings such as with time series data, with extremes, and with tensors. It should be possible to generalise also the generalised sliced average variance estimator, as developed in Lee et al. (2013) and further in Li (2018), to the categorical predictor setting. Li et al. (2011) and Artemiou and Dong (2016) develop support vector machine based approaches to nonlinear sufficient dimension reduction with multivariate predictors and univariate responses; the author of this thesis believes such approaches can be generalised to have both the predictor and response lie in separable metric spaces. The last idea that is mentioned in this thesis for future work is to examine statistical inference when conducted post nonlinear sufficient dimension reduction, thus continuing the line of research began in Kim et al.

(2020) who did it for linear sufficient dimension reduction.

Bibliography

Arnold, B. C. and Brockett, P. L. (1992). On Distributions Whose Component Ratios Are Cauchy. *American Statistician* 46.1, 25–26.

Aronszajn, N. (1950). Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*.

Artemiou, A. (2021). Using Mutual Information to Measure the Predictive Power of Principal Components. In: *Festschrift in Honor of R. Dennis Cook*. Ed. by E. Bura and B. Li. Springer, Cham, 1–16.

Artemiou, A. and Dong, Y. (2016). Sufficient Dimension Reduction via Principal Lq Support Vector Machine. *Electronic Journal of Statistics* 10, 783–805.

Artemiou, A. and Li, B. (2009). On Principal Components Regression: A Statistical Explanation of a Natural Phenomenon. *Statistica Sinica* 19, 1557–1565.

Artemiou, A. and Li, B. (2013). Predictive Power of Principal Components for Single-Index Model and Sufficient Dimension Reduction. *Journal of Multivariate Analysis* 119, 176–184.

Bass, R. (2016). *Real Analysis for Graduate Students*.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, MA: Springer US.

Billingsley, P. (1995). *Probability and Measure*. New York: John Wiley & Sons.

Boente, G., Salibián Barrera, M., and Tyler, D. E. (2014). A Characterization of Elliptical Distributions and Some Optimality Properties of Principal Components for Functional Data. *Journal of Multivariate Analysis* 131, 254–264.

Bura, E. and Yang, J. (2011). Dimension Estimation in Sufficient Dimension Reduction: A Unifying Approach. *Journal of Multivariate Analysis* 102.1, 130–142.

Cameron, P. (1998). *Sets, Logic and Categories*. Springer Undergraduate Mathematics Series. London: Springer-Verlag.

Cardinal Arithmetic in nLab (2022). nLab. URL: `https://ncatlab.org/nlab/show/cardinal+arithmetic` (visited on 09/23/2022).

Chiaromonte, F., Cook, R., and Li, B. (2002). Sufficient Dimension Reduction in Regressions with Categorical Predictors. *The Annals of Statistics* 30.2, 475–497.

Cohn, D. (2013). *Measure Theory*. Birkhäuser Advanced Texts Basler Lehrbücher. New York, NY: Springer.

Conway, J. (1990). *A Course in Functional Analysis*. Vol. 96. Graduate Texts in Mathematics. New York: Springer.

Cook, R. (1994). Using Dimension-Reduction Subspaces to Identify Important Inputs in Models of Physical Systems. *Proceedings of the Section on Physical and Engineering Sciences*. American Statistical Association, 18–25.

Cook, R. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: John Wiley & Sons.

Cook, R. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science* 22.1, 1–26.

Cook, R. and Critchley, F. (2000). Identifying Regression Outliers and Mixtures Graphically. *Journal of the American Statistical Association* 95.451, 781–794.

Cook, R. and Forzani, L. (2009). Likelihood-Based Sufficient Dimension Reduction. *Journal of the American Statistical Association*.

Cook, R. and Ni, L. (2005). Sufficient Dimension Reduction via Inverse Regression. *Journal of the American Statistical Association*.

Cook, R. and Weisberg, S. (1991). Sliced Inverse Regression for Dimension Reduction: Comment. *Journal of the American Statistical Association* 86.414, 328–332.

Cortes, C. and Vapnik, V. (1995). Support Vector Networks. *Machine Learning* 20, 273–297.

Cox, D. R. (1968). Notes on Some Aspects of Regression Analysis. *Journal of the Royal Statistical Society. Series A (General)* 131.3, 265–279.

Dinculeanu, N. (2000). *Vector Integration and Stochastic Integration in Banach Spaces*. John Wiley & Sons.

Eaton, M. L. (1986). A Characterization of Spherical Distributions. *Journal of Multivariate Analysis* 20.2, 272–276.

Engelking, R. (1989). *General Topology*. Berlin: Heldermann Verlag.

Ferré, L. and Yao, A. F. (2003). Functional Sliced Inverse Regression Analysis. *Statistics* 37.6, 475–488.

Fisher, R. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society* 222.594-604, 309–368.

Fukumizu, K., Bach, F., and Jordan, M. (2004). Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research* 5, 106–8569.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical Consistency of Kernel Canonical Correlation Analysis. *Journal of Machine Learning Research* 8.14, 361–383.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel Dimension Reduction in Regression. *The Annals of Statistics* 37.4.

Gálvez-Rodríguez, J. and Sánchez-Granero, M. (2020). The Distribution Function of a Probability Measure on the Dedekind-MacNeille Completion. *Topology and its Applications* 275, 107010.

Gálvez-Rodríguez, J. F. and Sánchez-Granero, M. (2022). Constructing a Linearly Ordered Topological Space from a Fractal Structure: A Probabilistic Approach. *Mathematics* 10.23.

Gálvez-Rodríguez, J. F. and Sánchez-Granero, M. Á. (2019). The Distribution Function of a Probability Measure on a Linearly Ordered Topological Space. *Mathematics* 7.9, 864.

Gauthier, B. and Suykens, J. A. K. (2018). Optimal Quadrature-Sparsification for Integral Operator Approximation. *SIAM Journal on Scientific Computing* 40.5, A3636–A3674.

Gretton, A. et al. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. *International Conference on Algorithmic Learning Theory*. Springer, 63–77.

Gretton, A. et al. (2007). A Kernel Statistical Test of Independence. *Advances in Neural Information Processing Systems*. NeurIPS Proceedings, 8.

Hall, P. and Li, K.-C. (1993). On Almost Linearity of Low Dimensional Projections from High Dimensional Data. *The Annals of Statistics* 21.2.

Hoffmann-Jørgensen, J. (1994). *Probability with a View toward Statistics*. Vol. 1. 2 vols. Chapman and Hall Probability Series. Chapman & Hall.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. West Sussex: Wiley.

Hutchings, M. and Gauthier, B. (2022). *Local Optimisation of Nyström Samples through Stochastic Gradient Descent*. URL: `http://arxiv.org/abs/2203.13284` (visited on 11/24/2022). preprint.

Jech, T. (2006). *Set Theory: The Third Millennium Edition, Revised and Expanded*. 3rd ed. Springer Monographs in Mathematics. Springer.

Jolliffe, I. (2002). *Principal Component Analysis*. Aberdeen: Springer.

Jones, B. and Artemiou, A. (2019). On Principal Components Regression with Hilbertian Predictors. *Annals of the Institute of Statistical Mathematics* 72, 627–644.

Jones, B. and Artemiou, A. (2021). Revisiting the Predictive Power of Kernel Principal Components. *Statistics & Probability Letters* 171, 109019.

Jones, B., Artemiou, A., and Li, B. (2020). On the Predictive Potential of Kernel Principal Components. *Electronic Journal of Statistics* 14.1, 1–23.

Kallenberg, O. (1988). Spreading and Predictable Sampling in Exchangeable Sequences and Processes. *The Annals of Probability* 16.2, 508–534.

Kallenberg, O. (1992). Symmetries on Random Arrays and Set-Indexed Processes. *Journal of Theoretical Probability* 5.4, 727–765.

Kallenberg, O. (2000). Spreading-Invariant Sequences and Processes on Bounded Index Sets. *Probability Theory and Related Fields* 118.2, 211–250.

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. 1st ed. Probability and Its Applications. Springer New York.

Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Vol. 77. Probability Theory and Stochastic Modelling. Cham: Springer International Publishing.

Kallenberg, O. (2021). *Foundations of Modern Probability*. Probability Theory and Stochastic Modelling. Springer Cham.

Kim, K. et al. (2020). On Post Dimension Reduction Statistical Inference. *The Annals of Statistics* 48.3, 1567–1592.

Klenke, A. (2008). *Probability Theory*. 1st ed. London: Springer.

Klenke, A. (2020). *Probability Theory: A Comprehensive Course*. 3rd ed. Universitext. Cham: Springer.

Kübler, J. M., Muandet, K., and Schölkopf, B. (2019). Quantum Mean Embedding of Probability Distributions. *Physical Review Research* 1.3, 033159.

Kumar, V. and Minz, S. (2014). Feature Selection : A Literature Review. *Smart Computing Review* 4.3, 211–229.

Lee, K.-Y., Li, B., and Chiaromonte, F. (2013). A General Theory for Nonlinear Sufficient Dimension Reduction: Formulation and Estimation. *The Annals of Statistics* 41.1, 221–249.

Li, B. (2007a). Comment: Fisher Lecture: Dimension Reduction in Regression. *Statistical Science* 22.1, 32–35.

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. Boca Raton: CRC Press.

Li, B., Artemiou, A., and Li, L. (2011). Principal Support Vector Machines for Linear and Nonlinear Sufficient Dimension Reduction. *The Annals of Statistics* 39.6, 3182–3210.

Li, B. and Babu, G. J. (2019). *A Graduate Course on Statistical Inference*. Springer Texts in Statistics. New York: Springer.

Li, B. and Song, J. (2017). Nonlinear Sufficient Dimension Reduction for Functional Data. *The Annals of Statistics* 45.3, 1059–1095.

Li, B., Zha, H., and Chiaromonte, F. (2005). Contour Regression: A General Approach to Dimension Reduction. *Annals of Statistics*.

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* 86.414, 316–327.

Li, Y. (2007b). *A Note on Hilbertian Elliptically Contoured Distributions*.

Lian, H. and Li, G. (2014). Series Expansion for Functional Sufficient Dimension Reduction. *Journal of Multivariate Analysis*.

Loève, M. (1977). *Probability Theory Vol 1*. 4th ed. Graduate Texts in Mathematics. New York: Springer.

Muandet, K. et al. (2016). Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning* 10.1-2, 1–141.

Muscat, J. (2014). *Functional Analysis*. Springer International Publishing.

Ni, L. (2011). Principal Component Regression Revisited. *Statistica Sinica* 21, 741–747.

Park, J. and Muandet, K. (2020). A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, 21247–21259.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Rudin, W. (1976). *Principles of Mathematical Analysis*. 3rd ed. International Series in Pure and Applied Mathematics. New York: McGraw-Hill.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* 10.5, 1299–1319.

Schuld, M. and Killoran, N. (2019). Quantum Machine Learning in Feature Hilbert Spaces. *Physical Review Letters* 122.4, 040504.

Shao, Y., Cook, R., and Weisberg, S. (2009). Partial Central Subspace and Sliced Average Variance Estimation. *Journal of Statistical Planning and Inference* 139.3, 952–961.

Simons, G. (1974). Generalized Distribution Functions: The Linearly Ordered Case with Applications to Nonparametric Statistics. *The Annals of Probability* 2.3, 501–508.

Sklar, A. (1973). Random Variables, Joint Distribution Functions, and Copulas. *Kybernetika* 9.6, 449–460.

Steen, L. A. and Seebach, J. A. (1978). *Counterexamples in Topology*. 2nd ed. New York: Springer-Verlag.

Stromberg, K. (2015). *An Introduction to Classical Real Analysis*. Reprint. Providence, Rhode Island: American Mathematical Society.

Taylor, M. (2006). *Measure Theory and Integration*. Graduate Studies in Mathematics 76. American Mathematical Society.

Vestrup, E. (2003). *The Theory of Measures and Integration*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Virta, J., Lee, K.-Y., and Li, L. (2022). *Sliced Inverse Regression in Metric Spaces*. URL: `http://arxiv.org/abs/2206.11511` (visited on 09/25/2022). preprint.

Wang, G. (2017). Dimension Reduction in Functional Regression with Categorical Predictor. *Computational Statistics* 32.2, 585–609.

Wang, G., Lin, N., and Zhang, B. (2013). Functional Contour Regression. *Journal of Multivariate Analysis*.

Wang, G. et al. (2015). The Hybrid Method of FSIR and FSAVE for Functional Effective Dimension Reduction. *Computational Statistics and Data Analysis*.

Willard, S. (1970). *General Topology*. Addison-Wesley Publishing Company, Inc.

Williams, D. (2018). *Probability with Martingales*. Cambridge Mathematical Textbooks. Cambridge: Cambridge University Press.

Yeh, Y. R., Huang, S. Y., and Lee, Y. J. (2009). Nonlinear Dimension Reduction with Kernel Sliced Inverse Regression. *IEEE Transactions on Knowledge and Data Engineering*.

Zhou, Y. et al. (2019). A Class of Optimal Estimators for the Covariance Operator in Reproducing Kernel Hilbert Spaces. *Journal of Multivariate Analysis* 169, 166–178.