Technische Universität München

TUM School of Computation, Information and Technology

Department of Mathematics

Canonical Correlation Analysis with Optimal Transport

Master's Thesis

by

Kian Saraf-Poor

Supervisor:	Prof. PhD Mathias Drton
Advisor:	PhD Hongjian Shi
Submission Date:	15.07.2024

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 15.07.2024

Abstract

Many studies involve the collection of multivariate data with an interest in understanding their dependencies. While some researchers aim to study all interrelations, others focus on the dependency between non-overlapping sets of random variables. Canonical Correlation Analysis (CCA) is a method for the latter, considering data partitioned into two groups. The goal is to find linear combinations of random variables in both groups that have the highest (canonical) correlation. Researchers can then study these combinations to explain the dependency within the dataset. Despite its established use, CCA assumes normality of the data. To overcome this restrictive assumption, generalizations have been proposed. One such generalization is the Gaussian Copula CCA (GCCCA) model, which allows the univariate marginals to have arbitrary continuous distributions while requiring that the joint dependency is described by a Gaussian Copula. Recently, the GCCCA model has been further generalized to the Cyclically Monotone CCA (CMCCA), allowing the joint marginals to follow arbitrary distributions. In this thesis, we will present all three models, including their estimation methods and establish their consistency. Given that the CMCCA model is based on optimal transport, we will provide an introduction to this topic, including cyclical monotonicity and gradients of convex functions. Additionally, we will suggest two Bayesian methods for the GCCCA and CMCCA models. Finally, we will present a simulation comparing the performance of all five methods.

Zusammenfassung

In vielen Studien werden multivariate Daten erhoben, wobei deren Abhängigkeit von Interesse ist. Einige sind an der Untersuchung aller Zusammenhänge interessiert, während andere sich auf die Abhängigkeit von nicht überlappenden Gruppen von Zufallsvariablen konzentrieren. Die Kanonische Korrelationsanalyse (CCA) gehört zu den letzteren und betrachtet Daten, die in zwei Gruppen aufgeteilt sind. Ziel ist es, Linearkombinationen von Zufallsvariablen aus beiden Gruppen zu finden, die die höchste (kanonische) Korrelation aufweisen. Der Forscher kann dann die Linearkombinationen mit der höchsten Korrelation untersuchen und die Abhängigkeit des Datensatzes mit diesen erklären. Heute ist die CCA eine etablierte Methode, die jedoch Normalität der Daten voraussetzt, sodass Verallgemeinerungen vorgeschlagen wurden, um diese restriktive Annahme zu überwinden. Eine dieser ist das Gauß'sche-Copula-CCA Modell (GCCCA), bei dem die univariaten Randverteilungen beliebige stetige Verteilungen haben können, aber die gemeinsame Abhängigkeit durch eine Gaußsche-Copula beschrieben werden muss. Neulich wurde das GCCCA-Modell weiter zur zyklisch monotonen CCA (CMCCA) verallgemeinert, so dass die gemeinsame Verteilung der beiden Gruppen beliebige Verteilungen haben können. Wir werden alle drei Modelle vorstellen, einschließlich eines Schätzers für jedes Modell, für welche wir die Konsistenz beweisen. Da das CMCCA-Modell auf optimalem Transport basiert, werden wir auch eine Einführung in dieses Thema geben und verwandte Begriffe, wie zyklische Monotonie und Gradienten von konvexen Funktionen erläutern. Außerdem werden wir zwei Bayes'sche Methoden für die GCCCA bzw. CMCCA vorschlagen. Zum Schluss wird eine Simulation vorgestellt, die die Genauigkeit aller fünf Methoden vergleicht.

Acknowledegements

I thank . . .

. . . Professor PhD Mathias Drton for letting my work and grow with this exciting project and for his phenomenal support during my work on the thesis.

. . . PhD Hongjian Shi for the time he invested in supporting me and the essential skills I learned from him.

. . . my mother for always believing in me and my study in mathematics.

. . . my father without whose incredible motivation my success would not have been possible.

Contents

1 Introduction

Until the 1930s, correlations and regressions were usually applied to one dimensional data, until Hotelling developed the theory of Canonical Correlation Analysis (CCA) in (Hotelling (1936)). Consider the following example for motivation. Let us say a researcher is interested in the dependency of two non-overlapping sets of random variables. One set is body characteristics of people given a fixed age, as height, weight, shoulder width, etc. The other variables are performances in sections of an IQ test, as processing speed, working memory, verbal comprehension and perceptual reasoning. Now, a typical approach at that point of time would have been to study all correlations which yields $3 \cdot 4 = 12$ numbers to explain the dependency. Although this approach is reasonable and 12 is not a too high number, Hotelling developed a better procedure to tackle this problem with CCA.

The idea of CCA is to study linear combinations of both variable sets which yield the highest correlation. These will be then called canonical variates and their correlation is the first canonical correlation. Thereafter, one will continue to find new linear combinations of both sets of random variables which are uncorrelated with the previous maximizing their correlation. Thereby, the requirement of being uncorrelated is there to ensure that the new linear combinations explain dependencies that have not been covered before. This procedure can be repeated as often the dimension of the smaller data set permits it, in our example 3 canonical correlations in total. This procedure is particularly advantageous in high-dimensional data sets as one can explain their dependency with a manageable number of canonical variates and canonical correlations instead of estimating a huge block of a correlation matrix.

Today, CCA has become an established method to examine the correlation of groups of random variables. However, one of its fundamental assumptions is the normality of the data which is obviously not satisfied in all cases. Therefore, there have been suggestions for generalizing the classical CCA model, such as the Gaussian Copula CCA and more recently the cyclically monotone CCA which allow more flexibility in the distribution of the data. The latter will be the main topic of this thesis.

The expected prerequisites for this thesis are measure theory, linear algebra, analysis and introductory courses to probability theory and statistics. We will begin with a chapter about measure-theoretic probability theory, as the proofs later on may need some results. Thereby, we will cover topics such as distributions, convergence of random vectors and tightness.

Then, Chapter 4 will provide an introduction into optimal transport as the cyclically monotone CCA model is based on it. Thereby, we will cover the function class of gradients of convex functions which can be interpreted as a generalization of increasing functions in higher dimensions. They are particularly interesting, as they are exactly our optimal transport maps. Further, we will learn about cyclical monotonicity which allows tackling discrete optimal transport problems by solving an optimal assignment problem. The theory has been extensively studied in (McCann, 1995) and (Rockafellar, 1966).

After introducing it in Section 4.1, we cover an example for the use of optimal transport in order to understand this concept better. We will present the center-outward distribution function, a concept invented in (Hallin et al., 2017) which allows to generalize the distribution function and quantiles to higher dimensions. This is based on so called rank statistics which can be interpreted as representants of samples with respect to another distribution. The application of rank statistics is a current research trend. For example, tests of equality and the independence of probability distributions based on rank statistics have been proposed in (Deb and Sen, 2019) and (Shi et al., 2022), respectively. We will later use them for Canonical Correlation Analysis.

In chapter 5, we will begin with a short recap of the multivariate normal distribution in section 5.1. Then, we will introduce the classical CCA theory in Section 5.2 following (Anderson, 2003). Thereby, we will present the estimation method and prove its consistency. Then, we will consider generalizations of the classical model, such as the Gaussian Copula CCA (GCCCA) in Section 5.3 which has also been proposed in Yoon et al. (2020). There, the univariate marginals of the data are allowed to follow arbitrary continuous distributions. One can then estimate the canonical correlations with Spearman's ρ. We will present this method and prove its consistency, as well.

In Section 5.4, we will then examine the main topic of this thesis, the cyclically monotone CCA model which is based on optimal transport. The idea behind it is that when the joint marginals of our observed data Y_1 and Y_2 follow some arbitrary distribution, we can transport them to the multivariate normal distribution and then apply the classical CCA estimator. This idea has been proposed and studied in Bryan et al. (2024). Also for this method we will prove the consistency which will be done in Section 5.5.

In chapter 6, will present two more estimation methods for the GCCCA and the CMCCA model, respectively. These will go in a different direction as the previous models and follow a Bayesian approach. Therefore, we begin the chapter with an introduction into Bayesian models and Markov Chain Monte Carlo algorithms, such as the Gibbs Sampler. After doing this in Section 6.1, we present the Bayesian methods in Section 6.2 and 6.3, respectively. These algorithms have been introduced in Hoff (2007b) and Bryan et al. (2024), respectively.

Finally, we will make a simulation in Chapter 7 in order to assess the performance of all five methods. These will compare the accuracy of all algorithms in different scenarios such as varying distributions of the data, changed dimensions and different sample sizes.

2 Notation

We will use the following notations in the thesis.

- The euclidean norm $\|\cdot\|_2$ will be denoted by $\|\cdot\|$, i.e. we have for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$,
$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$$
- For $n \in \mathbb{N}$, we will write
$$[n] \coloneqq \{1, \dots, n\},$$
and
$$[n]_0 \coloneqq [n] \cup \{0\}.$$
- We denote the set of all permutations of the set $[n]$ by S_n.
- Let X be a set. We denote the power set of X by $\mathfrak{P}(X) = \{M : M \subseteq X\}$. Further, let $A \in \mathfrak{P}(X)$ be a subset. Then, we will denote the complement of A in X by $A^c = X \setminus A$.
- Let X be a set. We denote the indicator function of the set by $\mathbb{I}_X$. We have
$$\mathbb{1}_X(x) = \begin{cases} 1, & \text{if } x \in X \\ 0, & \text{otherwise.} \end{cases}$$
Sometimes we will also use the notation $\mathbb{I}_Y$, when Y is a claim. Then, we will have
$$\mathbb{1}_Y = \begin{cases} 1, & \text{if } Y \text{ is true} \\ 0, & \text{otherwise.} \end{cases}$$
- For $d \in \mathbb{N}$, I_d will denote the identity matrix, i.e.
$$I_d = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{d \times d}.$$
- For $d \in \mathbb{N}$, we set $\mathbf{1}_d = (1, \dots, 1) \in \mathbb{R}^d$.

3 Fundamentals of Measure-Theoretic Probability

In order to present the theory for Canonical Correlation Analysis, we will need some results from measure-theoretic probability theory. This section explains some basic concepts. Introductions to the prerequisites, measure theory and calculus-based probability are provided in (Bauer, 2001) and (Georgii, 2013), respectively.

3.1 Preliminaries

We begin with some preliminaries, namely defining probability spaces, events, random variables/vectors, independence, distributions, and expectation. Further, we will establish some basic properties.

Definition 3.1.1 (Probability space)**.** Let Ω be a nonempty set. Further, let $\mathcal{F} \subseteq \mathfrak{P}(\Omega)$ be a σ-field on Ω, i.e.

(i) $\Omega \in \mathcal{F}$,

(ii) $A_1, A_2 \ldots \in \mathcal{F} \implies \bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{F}$, and

(iii) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$.

Then, we call the tuple $(\Omega, \mathcal{F})$ a *measurable space.* A map $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called *probability measure* on $(\Omega, \mathcal{F})$ if

(i) $\mathbb{P}(\emptyset) = 0$,

(ii) $\mathbb{P}(\Omega) = 1$, and

(iii) for all $A_1, A_2 \ldots \in \mathcal{F}$ pairwise disjoint, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a *probability space.* The elements of $\mathcal{F}$ are called *events.*

Definition 3.1.2 (Borel σ-field)**.** Consider the set of all open subsets of $\mathbb{R}^d$,

$$\mathcal{O} = \{A \subseteq \mathbb{R}^d : A \text{ open}\}.$$

We call

$$\mathcal{B}(\mathbb{R}^d) \coloneqq \sigma(\mathcal{O})$$

the *Borel σ-field* on $\mathbb{R}^d$, where $\sigma(\mathcal{O})$ is the generated σ-field of $\mathcal{O}$ on $\mathbb{R}^d$, i.e. the smallest σ-field on $\mathbb{R}^d$ containing $\mathcal{O}$.

For $E \in \mathcal{B}(\mathbb{R}^d)$, we define the σ-field

$$\mathcal{B}(E) \coloneqq \{E \cap F : F \in \mathcal{B}(\mathbb{R}^d)\},$$

which we call the *Borel σ-field* on E.

Example 3.1.3 (Dirac Measure)**.** Fix $x \in \mathbb{R}^d$. The *Dirac-Measure* with mass in x, defined as

$$\delta_x : \mathcal{B}(\mathbb{R}^d) \to [0,1], \delta_x(A) = \mathbb{1}_A(x),$$

is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Example 3.1.4. Let $\Omega = [0,1]$ and let $\mathcal{F} := \mathcal{B}([0,1])$. Set $\mathbb{P} = \lambda|_{[0,1]}$, where $\lambda|_{[0,1]}$ is the restriction of the Lebesgue-measure on $\mathcal{B}(\mathbb{R})$ to $\mathcal{B}([0,1])$. Then, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

In the following, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 3.1.5 (Random vector)**.** A measurable map

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$$

is called *random vector.* The set

$$\sigma(X) := \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R}^d)\}$$

is called the *generated σ-field* by X. If $d = 1$, we call X *random variable.*

Notation 3.1.6. Let X be a random vector in $\mathbb{R}^d$ and let $A \in \mathcal{B}(\mathbb{R}^d)$. We will use a common notation for events throughout the thesis. For example we define

$$\{X \in A\} := \{\omega \in \Omega : X(\omega) \in A\}$$
$$\{\|X\| \leq 1\} := \{\omega \in \Omega : \|X(\omega)\| \leq 1\}$$

Further, we will write $\mathbb{P}(X \in A)$ for $\mathbb{P}(\{X \in A\})$ and so on.

Definition 3.1.7 (Independence)**.** Random vectors $X_1, \dots, X_n$ in dimension $d_1, \dots, d_n$ are independent if for all $B_i \in \mathcal{B}(\mathbb{R}^{d_i})$

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \in B_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i).$$

An infinite collection of random vectors is called independent if every finite subset of random vectors is independent.

Theorem 3.1.8. *Suppose $X_1, \dots, X_n$ are independent random vectors with values in $\mathbb{R}^d$. Let $f_1, \dots, f_n : \mathbb{R}^d \to \mathbb{R}^d$ be measurable. Then, $f_1(X_1), \dots, f_n(X_n)$ are independent random vectors.*

Proof. This follows from (Durrett, 2019)[Theorem 2.1.10]. There the Theorem is presented in dimension 1. However, the proof in higher dimensions is analogous. □

Definition 3.1.9 (Distribution)**.** Let X be a random vector. Let μ_X be the push-forward measure of $\mathbb{P}$ by X,

$$\mu_X := X\#\mathbb{P} = \mathbb{P} \circ X^{-1},$$

i.e.

$$\mu_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) \qquad A \in \mathcal{B}(\mathbb{R}^d).$$

Then, μ_X is called the *distribution* of X and is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We write $X \sim \mu_X$.

Definition 3.1.10 (Distribution Function)**.** Let $X \sim \mu_X$ be a random variable in dimension 1. The function

$$F_X : \mathbb{R} \to [0,1], F_X(t) = \mu_X((-\infty, t]) = \mathbb{P}(X \leq t)$$

is called *distribution function.* If $X_1, \dots, X_n$ is a sample in dimension 1, the function

$$F^{(n)} : \mathbb{R} \to [0,1], F^{(n)}(t) = \frac{1}{n}\#\{k \in [n] : X_k \leq t\}$$

is called *empirical distribution function.*

Definition 3.1.11 (Continuous Distribution)**.** We say that a random variable X with values in $\mathbb{R}$ has a *continuous distribution* if its distribution function F_X is continuous.

Example 3.1.12. Consider the probability space from Example 3.1.4. Define

$$U : [0,1] \to \mathbb{R}, U(\omega) = \omega.$$

Then, U is uniformly distributed on the interval $(0,1)$. We write $U \sim \mathrm{Unif}(0,1)$. Also, U is continuous.

We can now introduce an important tool we will need later: the generalized inverse transform method.

Definition 3.1.13 (Generalized Inverse)**.** Let X be a random variable in dimension $d = 1$ with distribution function F_X. The *generalized inverse* of F_X is defined as

$$F_X^{-1} : (0,1) \to \mathbb{R}, u \mapsto \inf\{t \in \mathbb{R} : F(t) \geq u\}.$$

Note, that by the definition of F_X, F_X is increasing, and we have $\lim\limits_{t\to-\infty} F_X(t) = 0$ and $\lim\limits_{t\to\infty} F_X(t) = 1$. Hence, F_X^{-1} is well-defined with values in $\mathbb{R}$.

Proposition 3.1.14 (Proposition 5.2 in (McNeil et al., 2005))**.** *Let X be a random variable in dimension $d = 1$ with distribution function F_X and generalized inverse F_X^{-1}. Further, let $U \sim \mathrm{Unif}(0,1)$. We have*

(i) $F_X^{-1}(U) \sim \mu_X$*,*

(ii) F_X *is continuous* $\implies F_X(X) \sim \mathrm{Unif}(0,1)$*.*

Hence, for every one-dimensional distribution μ, there exists a random variable Y with $Y \sim \mu$ by Example 3.1.12. The same holds for all distribution functions.

Some other properties of the generalized inverse we may need later are the following:

Theorem 3.1.15. *Let X be a random variable in dimension 1 with distribution function F_X and generalized inverse F_X^{-1}.*

(i) We have $\mathbb{P}(X = F_X^{-1}(F_X(X))) = 1$*.*

(ii) If F_X is continuous, then F_X^{-1} is strictly increasing. Furthermore, we have $F_X(F_X^{-1}(u)) = u$ for all $u \in (0,1)$.

Proof. For (i) we refer to (McNeil et al., 2005)[Proposition A.4]. Claim (ii) can be found in (McNeil et al., 2005)[Proposition A.3 (viii)]. □

Now, we define densities and absolutely continuous distributions.

Definition 3.1.16 (Density)**.** Let X be a random vector with distribution μ_X. Let ν be a measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We say that X has a density with respect to ν, if μ_X has a density with respect to ν, i.e. there exists a $(\mathcal{B}(\mathbb{R}^d)/\mathcal{B}([0,\infty))$-measurable function $f_X : \mathbb{R}^d \to [0,\infty)$ such that for all $A \in \mathcal{B}(\mathbb{R}^d)$

$$\mu_X(A) = \int\limits_A f_X(x)\mathrm{d}\nu(x).$$

Definition 3.1.17 (Absolutely Continuous)**.** A probability measure μ is called *absolutely continuous* if it has a density w.r.t. the Lebesgue-measure on $\mathbb{R}^d$. We say a random vector X is absolutely continuous if its distribution is absolutely continuous.

Notation 3.1.18. In the following we denote the set of all probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ by $\mathcal{P}^d$. Further, we write $\mathcal{P}^d_{ac}$ for the set of all absolutely continuous distributions in $\mathbb{R}^d$.

Example 3.1.19. Let $X \sim \mathrm{Bin}(n,p)$ where $n \in \mathbb{N}$ and $p \in (0,1)$. Then, X has the density

$$f_X : \mathbb{R} \to [0,\infty), f_X(k) = \begin{cases} \binom{n}{k}p^k(1-p)^{n-k} & \text{if} \quad k \in \{0,\dots,n\} \\ 0 & \text{otherwise,} \end{cases}$$

with respect to the counting measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Further, we have $\mu_X \notin \mathcal{P}^1_{ac}$.

Next, we will define the expectation and variance of random variables and vectors starting in dimension 1.

Definition 3.1.20 (Expectation)**.** Let $X \geq 0$ be a non-negative random variable. Then, we define the expectation of X by

$$\mathbb{E}[X] = \int\limits_\Omega X\mathrm{d}\mathbb{P} \in [0,\infty].$$

For an arbitrary random variable Y, we say Y is *integrable*, if $\mathbb{E}[|Y|] < \infty$. For an integrable random variable Y, we define

$$Y^+ := \max\{Y, 0\} \qquad \text{and} \qquad Y^- := \max\{-Y, 0\}$$

and set

$$\mathbb{E}[Y] := \mathbb{E}[Y^+] - \mathbb{E}[Y^-].$$

Notation 3.1.21. We say that a random vector fulfills a property *almost surely*, shortly a.s., if it fulfills it $\mathbb{P}$-almost everywhere. For example $X \leq Y$ a.s., means that $X(\omega) \leq Y(\omega)$ for $\mathbb{P}$-a.a. $\omega \in \Omega$ or, equivalently, $\mathbb{P}(X \leq Y) = 1$.

Some elementary results about the expectation are the following.

Theorem 3.1.22 (Theorem 1.6.1 in (Durrett, 2019))**.** *Let $X, Y \in L^1$. Then, the following statements hold:*

(i) $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$,

(ii) $\forall a \in \mathbb{R} : \mathbb{E}[aX] = a\mathbb{E}[X]$

(iii) $X \leq Y$ *a.s.* $\implies \mathbb{E}[X] \leq \mathbb{E}[Y]$.

Note that (i)-(iii) are also true for not necessarily integrable random variables $X, Y \geq 0$. In that case one replaces $\forall a \in \mathbb{R}$ with $\forall a \geq 0$ in (ii).

Definition 3.1.23. Let X be a random variable and let $p \geq 1$. We say $X \in L^p$ if

$$\mathbb{E}[|X|^p] < \infty.$$

Lemma 3.1.24. *Let X be a random variable and let $1 \leq p \leq q$. We have $L^q \subseteq L^p$ i.e.,*

$$\mathbb{E}[|X|^q] < \infty \implies \mathbb{E}[|X|^p] < \infty.$$

Proof. Let $\mathbb{E}[|X|^q] < \infty$. Applying Theorem 3.1.22 yields

$$\mathbb{E}[|X|^p] = \mathbb{E}[|X|^p \mathbb{1}_{\{|X|\leq 1\}}] + \mathbb{E}[|X|^p \mathbb{1}_{\{|X|>1\}}] \leq 1 + \mathbb{E}[|X|^q] < \infty.$$

□

Theorem 3.1.25 (Markov's inequality)**.** *Let $X \geq 0$ be a random variable and let $c \geq 0$. Then,*

$$c \cdot \mathbb{P}(X \geq c) \leq \mathbb{E}[X].$$

Proof. Observe that $c\mathbb{1}_{\{X \geq c\}} \leq X$ a.s. and apply Theorem 3.1.22. □

Definition 3.1.26 (Variance)**.** Let $X \in L^2$ be a random variable. Then, its *variance*,

$$\mathbb{V}\mathrm{ar}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

is well-defined. Further, let $Y \in L^2$ be another random variable. Then, the *covariance*,

$$\mathbb{C}\mathrm{ov}[X, Y] := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

is well-defined. If both, $\mathbb{V}\mathrm{ar}[X] \neq 0$ and $\mathbb{V}\mathrm{ar}[Y] \neq 0$, we can define their *correlation* as

$$\mathbb{C}\mathrm{orr}[X, Y] := \frac{\mathbb{C}\mathrm{ov}[X, Y]}{\sqrt{\mathbb{V}\mathrm{ar}[X]\mathbb{V}\mathrm{ar}[Y]}}. \tag{3.1}$$

Now, we define the same terms for random vectors.

Definition 3.1.27. Let $X = (X_1, \dots, X_d)$ be a $\mathbb{R}^d$-valued random vector and let $p \geq 1$. If $X_i \in L^p$ for all i or, equivalently,

$$\mathbb{E}[\|X\|^p] < \infty,$$

we say $X \in L^p$. For $X \in L^1$, we define

$$\mathbb{E}[X] \coloneqq \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix} \in \mathbb{R}^d.$$

For $X \in L^2$, we define

$$\mathbb{V}\text{ar}[X] \coloneqq \begin{pmatrix} \mathbb{V}\text{ar}[X_1] & \mathbb{C}\text{ov}[X_1, X_2] & \cdots & \mathbb{C}\text{ov}[X_1, X_d] \\ \mathbb{C}\text{ov}[X_2, X_1] & \mathbb{V}\text{ar}[X_2] & \cdots & \mathbb{C}\text{ov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}[X_d, X_1] & \mathbb{C}\text{ov}[X_d, X_2] & \cdots & \mathbb{V}\text{ar}[X_d] \end{pmatrix} \in \mathbb{R}^{d \times d}.$$

If additionally $\mathbb{V}\text{ar}[X_i] \neq 0$ for all i, we set

$$\mathbb{C}\text{orr}[X] \coloneqq \begin{pmatrix} 1 & \mathbb{C}\text{orr}[X_1, X_2] & \cdots & \mathbb{C}\text{orr}[X_1, X_d] \\ \mathbb{C}\text{orr}[X_2, X_1] & 1 & \cdots & \mathbb{C}\text{orr}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{orr}[X_d, X_1] & \mathbb{C}\text{orr}[X_d, X_2] & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{d \times d}.$$

This concludes the preliminary section.

3.2 Convergence of Random Vectors

The next important topic is the convergence of random vectors. Therefore, we will define almost sure convergence and convergence in probability in this section, and present the convergence results we need.

Throughout this subsection, let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of $\mathbb{R}^{d_1}$- and $\mathbb{R}^{d_2}$-valued random vectors and let X and Y be $\mathbb{R}^{d_1}$- and $\mathbb{R}^{d_2}$-valued random vectors, respectively, all defined on the same probability space.

Definition 3.2.1 (Convergence)**.** The sequence $(X_n)_{n\in\mathbb{N}}$ converges against X

- *almost surely* if there exists $\tilde{\Omega} \in \mathcal{F}$ with $\mathbb{P}(\tilde{\Omega}) = 1$ such that for all $\omega \in \tilde{\Omega}$

  $$X_n(\omega) \longrightarrow X(\omega).$$

  We write

  $$X_n \xrightarrow{a.s.} X.$$

- *in probability* if for all $\varepsilon > 0$

  $$\mathbb{P}(\|X_n - X\| > \varepsilon) \longrightarrow 0.$$

  We write

  $$X_n \xrightarrow{P} X.$$

One helpful lemma for showing convergence is the following. It allows us to prove the convergence of a random vector componentwise.

Lemma 3.2.2. *We have*

(i) $X_n \xrightarrow{a.s.} X$ *and* $Y_n \xrightarrow{a.s.} Y \implies \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{a.s.} \begin{pmatrix} X \\ Y \end{pmatrix}$, *and*

(ii) $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y \implies \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{P} \begin{pmatrix} X \\ Y \end{pmatrix}$.

Proof. Claim (i) is trivial. For claim (ii), observe

$$\mathbb{P}\left(\left\|\begin{pmatrix} X_n \\ Y_n \end{pmatrix} - \begin{pmatrix} X \\ Y \end{pmatrix}\right\| > \varepsilon\right) \leq \mathbb{P}\left(\|X_n - X\| > \frac{\varepsilon}{\sqrt{2}}\right) + \mathbb{P}\left(\|Y_n - Y\| > \frac{\varepsilon}{\sqrt{2}}\right) \longrightarrow 0.$$

□

Convergence in probability and almost sure convergence are related in the following way:

Theorem 3.2.3. *We have*

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X.$$

Proof. The claim in dimension 1 is shown in (Shorack, 2017)[Theorem 5.7(i)]. With Lemma 3.2.2 the multivariate version follows. □

The next two convergence theorems we present are well known, namely, the Continuous Mapping Theorem and the Strong Law of Large Numbers.

Theorem 3.2.4 (Continuous Mapping Theorem)**.** *Let* $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ *be continuous. We have*

(i) $X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X)$

(ii) $X_n \xrightarrow{P} X \implies g(X_n) \xrightarrow{P} g(X)$

Proof. Claim (i) is trivial. Further, Claim (ii) can be shown using the subsequence criterion in (Shorack, 2017)[Theorem 3.1 eq. (15)]. As it is provided in dimension 1 one can show the claim componentwise and then apply Lemma 3.2.2. □

Corollary 3.2.5. *Now suppose all random vectors are defined in the same dimension* $d = d_1 = d_2$ *and let* $a, b \in \mathbb{R}$. *We have*

(i) $X_n \xrightarrow{a.s.} X$ *and* $Y_n \xrightarrow{a.s.} Y \implies aX_n + bY_n \xrightarrow{a.s.} aX + bY$

(ii) $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y \implies aX_n + bY_n \xrightarrow{P} aX + bY$

Furthermore, we have in dimension $d = 1$,

(iii) $X_n \xrightarrow{a.s.} X$ *and* $Y_n \xrightarrow{a.s.} Y \implies X_nY_n \xrightarrow{a.s.} XY$

(iv) $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y \implies X_nY_n \xrightarrow{P} XY$

Proof. If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{a.s.} \begin{pmatrix} X \\ Y \end{pmatrix}$ by Lemma 3.2.2. Now, note that $f : \mathbb{R}^{2d} \to \mathbb{R}^d, f(x,y) = ax + by$ and $g : \mathbb{R}^2 \to \mathbb{R}, g(x,y) = xy$ are continuous. Hence, (i) and (iii) follow from Theorem 3.2.4. Finally, (ii) and (iv) can be shown analogously for convergence in probability. □

Theorem 3.2.6 (Strong Law of Large Numbers)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed (i.i.d.) random vectors with $X_1 \in L^1$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}[X_1].$$

Proof. For the SLLN in dimension 1, we refer to (Durrett, 2019)[Theorem 2.4.1]. This can be generalized to the multivariate version with Lemma 3.2.2 and Lemma 3.1.8. □

A consequence of the Strong Law of Large Numbers is the Glivenko-Cantelli theorem:

Theorem 3.2.7 (Theorem 19.1 in van der Vaart (1998))**.** *Let $\mu \in \mathcal{P}^1$ be an arbitrary one-dimensional distribution and let $X_1, X_2, \ldots \overset{iid}{\sim} \mu$ be an i.i.d. sample. Further, let F be the distribution function corresponding to μ and let $F^{(n)}$ be the (random) empirical distribution function obtained from the first n samples, $X_1, \ldots, X_n$. Then, we have*

$$||F^{(n)} - F||_\infty = \sup_{t\in\mathbb{R}} |F^{(n)}(t) - F(t)| \longrightarrow 0 \qquad a.s.$$

The next result we would like to present is the Dominated Convergence Theorem. The common form, in which it is stated, has the assumption that the random vectors are jointly bounded by an integrable random variable. However, there exists a stronger version which only requires uniform integrability which we define next.

Definition 3.2.8 (Uniformly Integrable)**.** The sequence $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable if

$$\lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}[\|X_n\| \mathbb{1}_{\{\|X_n\|\geq K\}}] = 0.$$

A useful criterion for showing uniform integrability is the following:

Theorem 3.2.9. *If*

$$\sup_{n\in\mathbb{N}} \mathbb{E}[\|X_n\|^p] < \infty \tag{3.2}$$

for some $p > 1$ then, $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable.

Proof. We show the claim by using Theorem 5.6 in (Shorack, 2017). It states that a family of integrable random variables $(Z_n)_{n\in\mathbb{N}}$ in uniformly inetgrable if there exists a convex function $G : [0,\infty) \to [0,\infty)$ with

(i) $G(0) = 0$,

(ii) $\lim\limits_{x\to\infty} \frac{G(x)}{x} = \infty$, and

(iii) $\sup_{n\in\mathbb{N}} \mathbb{E}[G(|Z_n|)] < \infty$.

Condition (3.2) implies that $(\|X_n\|)_{n\in\mathbb{N}}$ is a family of integrable random variables by Lemma 3.1.24. Setting $G(x) = x^p$, shows that $(\|X_n\|)_{n\in\mathbb{N}}$ is uniformly integrable. Therefore, $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable. □

Now, we can present the theorem:

Theorem 3.2.10 (Extended Dominated Convergence Theorem)**.** *If* $X_n \xrightarrow{P} X$ *and* $(X_n)_{n\in\mathbb{N}}$ *is uniformly integrable, then,*

$$\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$

Proof. See Vitali's Theorem (Theorem 5.5 in (Shorack, 2017)). □

Another notion we will encounter is conditional expectation which we define next.

Definition 3.2.11 (Conditional Expectation)**.** Let $X \in L^1$ be a random variable in dimension $d = 1$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-field on Ω. Then, a random variable $X_0 \in L^1$ is a version of the conditional expectation of X given $\mathcal{G}$ if

(i) X_0 is $\mathcal{G}$-measurable, and

(ii) for all $A \in \mathcal{G}$, we have

$$\mathbb{E}[X_0 \mathbb{1}_A] = \mathbb{E}[X \mathbb{1}_A].$$

We write $X_0 = \mathbb{E}[X|\mathcal{G}]$.

For a random variable Y or a family of random variables $(Y_i)_{i\in I}$, we define

$$\mathbb{E}[X|Y] \coloneqq \mathbb{E}[X|\sigma(Y)] \qquad \text{and} \qquad \mathbb{E}[X|Y_i, i \in I] \coloneqq \mathbb{E}[X|\sigma(Y_i, i \in I)],$$

respectively, where I is an arbitrary index set and $\sigma(Y_i, i \in I)$ is the smallest σ-field such that all Y_i are measurable.

For a random vector $X = (X_1, \dots, X_d) \in L^1$, we define

$$\mathbb{E}[X|\mathcal{G}] \coloneqq \begin{pmatrix} \mathbb{E}[X_1|\mathcal{G}] \\ \vdots \\ \mathbb{E}[X_d|\mathcal{G}] \end{pmatrix}.$$

The definition is justified by the following theorem.

Theorem 3.2.12 (Theorem 9.2 in (Williams, 1991))**.** *Let* $X \in L^1$ *and let* $\mathcal{G} \subseteq \mathcal{F}$ *be a* σ*-field. Then, the conditional expectation* $\mathbb{E}[X|\mathcal{G}]$ *exists and is a.s.-unique, i.e. if* $Y = \mathbb{E}[X|\mathcal{G}]$ *and* $Z = \mathbb{E}[X|\mathcal{G}]$*, then* $Y = Z$ *almost surely.*

There are many results about conditional expectation, but the most important one to us is the Bounded Convergence Theorem.

Theorem 3.2.13 (Bounded Convergence for Conditional Expectation)**.** *Suppose*

$$X_n \xrightarrow{a.s.} X.$$

Assume that there exists $C > 0$ such that for all $n \in \mathbb{N}$

$$\|X_n\| \leq C.$$

Now, let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-field on Ω. Then, we have

$$\mathbb{E}[X_n|\mathcal{G}] \xrightarrow{a.s.} \mathbb{E}[X|\mathcal{G}].$$

Proof. See (Williams, 1991)[Section 9.7(g)]. □

This concludes the section of the convergence of random vectors.

3.3 Weak Convergence, Tightness and Prokhorov's Theorem

The final preliminary topics we discuss are weak convergence, tightness and Prokhorov's Theorem.

The most results in this subsection will be about distributions, so let $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ be sequences of probability measures in $\mathcal{P}^{d_1}$ and $\mathcal{P}^{d_2}$, respectively, throughout this subsection. Further, let $\mu \in \mathcal{P}^{d_1}$ and $\nu \in \mathcal{P}^{d_2}$.

Also, we will have some theorems about random vectors, so let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of random vectors in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Further, let X and Y be random vectors in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. For notional convenience, we set $d := d_1$.

Definition 3.3.1 (Weak Convergence)**.** The sequence $(\mu_n)_{n\in\mathbb{N}}$ converges *weakly* to μ if

$$\int\limits_{\mathbb{R}^d} f(x)\mathrm{d}\mu_n(x) \longrightarrow \int\limits_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x) \tag{3.3}$$

for all $f : \mathbb{R}^d \to \mathbb{R}$ continuous and bounded. We write

$$\mu_n \xrightarrow{w} \mu.$$

The sequence $(X_n)_{n\in\mathbb{N}}$ converges *weakly* against X if $\mu_{X_n} \xrightarrow{w} \mu_X$, or equivalently,

$$\mathbb{E}[f(X_n)] \longrightarrow \mathbb{E}[f(X)]$$

for all $f : \mathbb{R}^d \to \mathbb{R}$ continuous and bounded. We write

$$X_n \xrightarrow{w} X.$$

A useful criterion for weak convergence is the following:

Lemma 3.3.2. *We have* $\mu_n \xrightarrow{w} \mu$ *if and only if*

$$\mu_n(A) \longrightarrow \mu(A)$$

for all $A \in \mathcal{B}(\mathbb{R}^d)$ *with* $\mu(\partial A) = 0$*, where* ∂A *is the boundary of* A*.*

Hence, we have $X_n \xrightarrow{w} X$ *if and only if*

$$\mathbb{P}(X_n \in A) \longrightarrow \mathbb{P}(X \in A)$$

for all $A \in \mathcal{B}(\mathbb{R}^d)$ *with* $\mathbb{P}(X \in \partial A) = 0$*.*

In dimension $d = 1$*, we have* $X_n \xrightarrow{w} X$ *if and only if*

$$F_{X_n}(t) \longrightarrow F_X(t)$$

for all $t \in \mathbb{R}$*, where* F_X *is continuous.*

Proof. For the first claim, see (Billingsley, 1999)[page 26]. For the second claim see (Durrett, 2019)[page 116 and Theorem 3.2.9]. □

The weak convergence of random vectors is related to the convergence in probability in the following way.

Theorem 3.3.3. *We have*

$$X_n \xrightarrow{P} X \implies X_n \xrightarrow{w} X.$$

If X *is a almost surely constant random vector, i.e. there exists some* $a \in \mathbb{R}^d$ *such that* $\mathbb{P}(X = a) = 1$*, we have*

$$X_n \xrightarrow{w} X \implies X_n \xrightarrow{P} X.$$

Proof. For the first claim suppose $X_n \xrightarrow{P} X$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous and bounded. Then, we have

$$f(X_n) \xrightarrow{P} f(X)$$

by the Continuous Mapping Theorem (Theorem 3.2.4). Since the $f(X_n)$ are uniformly bounded by the supremum norm of f, i.e.

$$|f(X_n)| \leq \|f\|_\infty$$

for all $n \in \mathbb{N}$, the collection $(f(X_n))_{n\in\mathbb{N}}$ is uniformly integrable. We conclude from the extended dominated convergence theorem (Theorem 3.2.10) that

$$\mathbb{E}[f(X_n)] \longrightarrow \mathbb{E}[f(X)].$$

Hence, $X_n \xrightarrow{w} X$.

For the second claim suppose $X_n \xrightarrow{w} X$ such that $\mathbb{P}(X = a) = 1$ for some $a \in \mathbb{R}^d$. Let $\varepsilon > 0$. Define the set

$$A \coloneqq \{x \in \mathbb{R}^d : \|x - a\| > \varepsilon\} \in \mathcal{B}(\mathbb{R}^d).$$

Then,

$$\partial A = \{x \in \mathbb{R}^d : \|x - a\| = \varepsilon\}$$

and we have $\mathbb{P}(X \in \partial A) = 0$. By Theorem 3.3.2, we have

$$\mathbb{P}(\|X_n - X\| > \varepsilon) = \mathbb{P}(X_n \in A) \longrightarrow \mathbb{P}(X \in A) = 0,$$

which shows $X_n \xrightarrow{P} X$. □

Hence, we get the following order of convergences: Almost sure convergence is the strongest implying convergence in probability which implies weak convergence. All modes of convergence fulfill a continuous mapping theorem. We have established it for the first two. Next, we introduce the continuous mapping theorem for weak convergence.

Theorem 3.3.4. *Let $g : \mathbb{R}^d \to \mathbb{R}^m$ be a measurable function and let D_g be the set of discontinuities of g. If $\mathbb{P}(X \in D_g) = 0$ then,*

$$X_n \xrightarrow{w} X \implies g(X_n) \xrightarrow{w} g(X)$$

Proof. This is shown in (Billingsley, 1999)[page 26]. □

Other important definitions for a collection of distributions are tightness and relatively compactness.

Definition 3.3.5 (Tight)**.** The sequence $(\mu_n)_{n\in\mathbb{N}}$ is *tight* if

$$\forall \varepsilon > 0 : \exists M > 0 : \forall n \in \mathbb{N} : \mu_n([-M, M]^d) \geq 1 - \varepsilon.$$

The sequence $(X_n)_{n\in\mathbb{N}}$ is tight if $(\mu_{X_n})_{n\in\mathbb{N}}$ is tight or, equivalently, if

$$\forall \varepsilon > 0 : \exists M > 0 : \forall n \in \mathbb{N} : \mathbb{P}(\|X_n\| \leq M) \geq 1 - \varepsilon.$$

Definition 3.3.6 (Relatively Compact)**.** The sequence $(\mu_n)_{n\in\mathbb{N}}$ is *relatively compact* if for all subsequences $(\mu_{n_k})_{k\in\mathbb{N}}$, there exists a subsubsequence $(\mu_{n_{k_l}})_{l\in\mathbb{N}}$ and $\tilde{\mu} \in \mathcal{P}^d$ such that

$$\mu_{n_{k_l}} \xrightarrow{w} \tilde{\mu}.$$

The sequence $(X_n)_{n\in\mathbb{N}}$ is relatively compact if $(\mu_{X_n})_{n\in\mathbb{N}}$ is relatively compact or, equivalently, if for all subsequences $(X_{n_k})_{k\in\mathbb{N}}$ there exists a subsubsequence $(X_{n_{k_l}})_{l\in\mathbb{N}}$ and a random vector $\tilde{X}$ such that

$$X_{n_{k_l}} \xrightarrow{w} \tilde{X}.$$

Prokhorov's Theorem shows that tightness and relatively compactness are equivalent:

Theorem 3.3.7 (Theorem 5.1 and 5.2 in (Billingsley, 1999))**.** *It holds*

$$(\mu_n)_{n\in\mathbb{N}} \textit{ is tight} \iff (\mu_n)_{n\in\mathbb{N}} \textit{ is relatively compact.}$$

Hence,

$$(X_n)_{n\in\mathbb{N}} \textit{ is tight} \iff (X_n)_{n\in\mathbb{N}} \textit{ is relatively compact.}$$

Tightness can be very useful to show convergence in some scenarios. With Theorem 3.3.7 we obtain the following criterion for tightness:

Corollary 3.3.8. *Suppose*

$$\mu_n \xrightarrow{w} \mu.$$

Then, $(\mu_n)_{n\in\mathbb{N}}$ is tight.

Equivalently, if

$$X_n \xrightarrow{w} X,$$

then, $(X_n)_{n\in\mathbb{N}}$ is tight.

Proof. Since the whole sequence $(\mu_n)_{n\in\mathbb{N}}$ converges, it is relatively compact. By Theorem 3.3.7, $(\mu_n)_{n\in\mathbb{N}}$ is tight. The same holds for random vectors. □

One scenario in which tightness can help to show weak convergence is the following:

Corollary 3.3.9 (Corollary of Theorem 5.1 in (Billingsley, 1999))**.** *If $(\mu_n)_{n\in\mathbb{N}}$ is tight and every subsequence that converges weakly, converges against the same distribution μ, then, the entire sequence converges against μ:*

$$\mu_n \xrightarrow{w} \mu$$

The last two lemmata of this section are rather specific and are tailored to proofs later throughout the thesis.

Definition 3.3.10. We define $\Gamma(\mu,\nu)$ as the set of all probability measures in $\mathcal{P}^{d_1+d_2}$ with marginals μ and ν on the first d_1 and last d_2 coordinates, respectively.

Lemma 3.3.11. *Suppose $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ are tight. Then, any sequence of probability measures $(\gamma_n)_{n\in\mathbb{N}}$ in $\mathcal{P}^{d_1+d_2}$ with $\gamma_n \in \Gamma(\mu_n,\nu_n)$ for all $n\in\mathbb{N}$ is tight.*

Proof. Let $\varepsilon > 0$. There exist $M_1, M_2 > 0$ such that for all $n \in \mathbb{N}$

$$\mu_n([-M_1,M_1]^{d_1}) \geq 1-\varepsilon/2 \qquad \text{and} \qquad \nu_n([-M_2,M_2]^{d_2}) \geq 1-\varepsilon/2,$$

respectively. With $M := \max\{M_1,M_2\}$, we have for all $\gamma_n \in \Gamma(\mu_n,\nu_n)$ and all $n\in\mathbb{N}$

$$\gamma_n\left(\left([-M,M]^{d_1+d_2}\right)^c\right) \leq \mu_n\left(\left([-M,M]^{d_1}\right)^c\right) + \nu_n\left(\left([-M,M]^{d_2}\right)^c\right) < \varepsilon,$$

which shows the claim. □

Definition 3.3.12 (o_P and O_P)**.** We introduce the following notations:

(i) $X_n \in o_P(1) :\Longleftrightarrow X_n \xrightarrow{P} 0,$

(ii) $X_n \in O_P(1) :\Longleftrightarrow (X_n)_{n\in\mathbb{N}}$ is tight.

Lemma 3.3.13. *Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of random variables in $\mathbb{R}$ so that*

$$X_n \in o_P(1) \qquad \textit{and} \qquad Y_n \in O_P(1).$$

Then,

$$X_nY_n \in o_P(1).$$

Proof. Let $\varepsilon, \delta > 0$. Let $M > 0$ such that

$$\sup_{n\in\mathbb{N}} \mathbb{P}(|Y_n| > M) < \frac{\delta}{2}.$$

Further, let $n_0 \in \mathbb{N}$ such that for all $\mathbb{N} \ni n \geq n_0$

$$\mathbb{P}\left(|X_n| > \frac{\varepsilon}{2M}\right) \leq \frac{\delta}{2}.$$

Then, we have for all $\mathbb{N} \ni n \geq n_0$

$$\begin{aligned}
\mathbb{P}(|X_nY_n| > \varepsilon) &= \mathbb{P}\left(|X_n||Y_n|\mathbb{1}_{\{|Y_n|\leq M\}} + |X_n||Y_n|\mathbb{1}_{\{|Y_n|\leq M\}} > \varepsilon\right) \\
&\leq \mathbb{P}\left(|X_n||Y_n|\mathbb{1}_{\{|Y_n|\leq M\}} > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|X_n||Y_n|\mathbb{1}_{\{|Y_n|> M\}} > \frac{\varepsilon}{2}\right) \\
&\leq \mathbb{P}\left(|X_n| \cdot M\mathbb{1}_{\{|Y_n|\leq M\}} > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\mathbb{1}_{\{|Y_n|> M\}} \neq 0\right) \\
&= \mathbb{P}\left(|X_n| > \frac{\varepsilon}{2M}\right) + \mathbb{P}(|Y_n| > M) \leq \delta.
\end{aligned}$$

□

4 Optimal Transport

In the last section we have introduced the necessary mathematical background we need for the theory later. This section will provide a brief introduction into optimal transport which our method for Canonical Correlation Analysis will be based on.

4.1 Cyclical Monotonicity

We begin by considering Monge's problem (see also (Monge, 1781)): How should one move given piles of sand to fill up given holes of the same total finite volume? For a mathematical formulation of this problem, let us assume w.l.o.g. that the total volume of the piles and holes is equal to 1, respectively. Then, one can imagine the piles and holes as probability measures μ_1 and μ_2 on $(\mathbb{R}^3, \mathcal{B}(\mathbb{R}^3))$, respectively. Now, the objective is to find a (measurable) transport map $T^* : \mathbb{R}^3 \to \mathbb{R}^3$ solving the problem

$$\inf_T \int_{\mathbb{R}^3} \|x - T(x)\| \mathrm{d}\mu_1 \qquad \text{subject to} \qquad T\#\mu_1 = \mu_2.$$

Thereby, the function T assigns every $x \in \mathbb{R}^3$ the point $T(x) \in \mathbb{R}^3$, where it gets transported to. Obviously, we are only interested in points x which lie in the support of μ_1, i.e. where sand is present. The integral calculates the entire distance the sand has been transported with T. Finally, the condition $T\#\mu_1 = \mu_2$ is there to ensure that when we use the transport map T, all the sand holes have been filled with the sand piles. The following picture from (Williams, 2020) illustrates the problem.

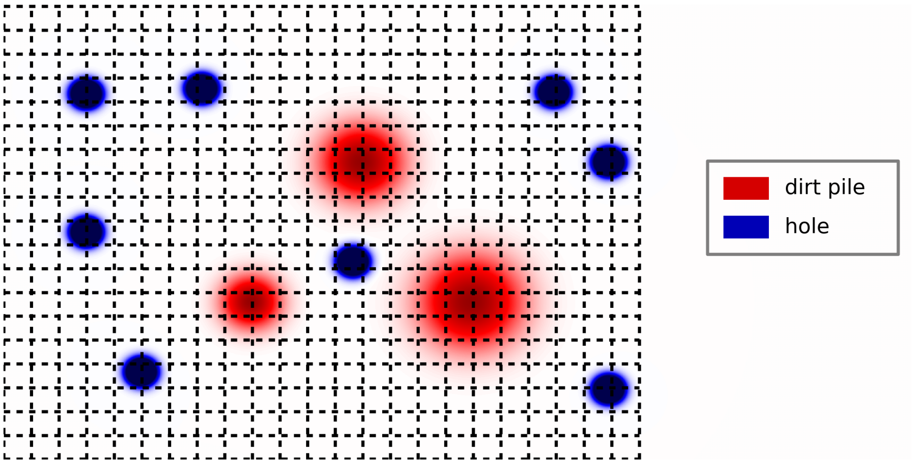

Figure 4.1: Example for an optimal transport problem

One could also view the problem in a discrete way. Let n green and n red points be given. What is the best one-to-one way to connect each red point to a corresponding green point? Thereby, one aims to minimize the sum of distances of the resulting pairs of points. Figure 4.2 illustrates the problem in dimension 2.

More generally, the optimal transport problem can be formulated in d dimensions with probability measures μ_1 and μ_2 on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and a measurable loss $L : \mathbb{R}^{2d} \to [0, \infty]$:

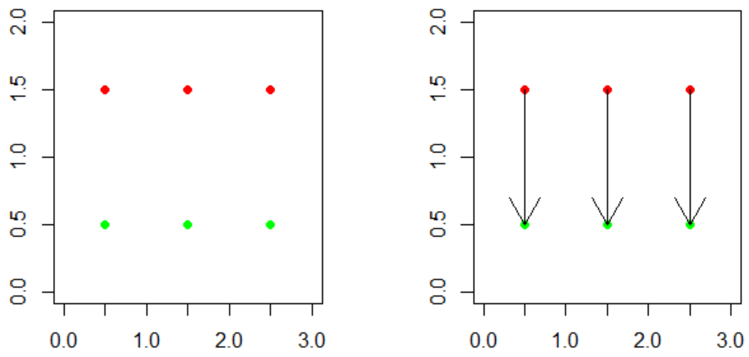

Figure 4.2: Example for a discrete optimal transport problem (solution on the right)

$$\inf_{T} \int_{\mathbb{R}^d} L(x, T(x)) \mathrm{d}\mu_1(x) \qquad \text{subject to} \qquad T\#\mu_1 = \mu_2. \tag{OT}$$

Although this problem looks simple, it has not been solved for a long time. For the squared error loss $L(x, T(x)) = \|x - T(x)\|^2$, it has been shown that if μ_1 and μ_2 are absolutely continuous with finite second moments (i.e. $\mathbb{E}[\|X_1\|^2] < \infty$ and $\mathbb{E}[\|X_2\|^2] < \infty$ for $X_1 \sim \mu_1$ and $X_2 \sim \mu_2$, respectively) then, the solution to (OT) exists, is a.e. unique and the gradient of a convex function (see (Villani, 2009)[Theorem 9.4]).

Why does this problem matter for us, especially when our main topic is Canonical Correlation Analysis? We will later see that transport maps can help us overcome the traditional assumption of normality in CCA. Thereby, the mentioned function class of gradients of convex functions will be particularly interesting for us as these functions will be exactly our transport maps. Note that the property of being a gradient of a convex function can be interpreted as a generalization of increasing functions in higher dimensions. Let us provide some examples for optimal transport maps.

Example 4.1.1 (Transport Maps or Gradients of Convex Functions)**.**

(i) Let $A \in \mathbb{R}^{d \times d}$ be symmetric and positive semidefinite. Then, for all $b \in \mathbb{R}^d$, the function

$$\varphi : \mathbb{R}^d \to \mathbb{R}, \varphi(z) = \frac{1}{2} z^T A z + b^T z$$

is convex. Hence, its gradient

$$\nabla\varphi : \mathbb{R}^d \to \mathbb{R}^d, \varphi(z) = Az + b$$

is an optimal transport map.

(ii) Let $f_1, \dots, f_d : \mathbb{R} \to \mathbb{R}$ be increasing functions. Then, the function

$$G : \mathbb{R}^d \to \mathbb{R}^d, z = (z_1, \dots, z_d) \mapsto (f_1(z_1), \dots, f_d(z_d))$$

is the gradient of a convex function.

(iii) Let $f : (0, \infty) \to [0, \infty)$ be an increasing function. Then, the function

$$J : \mathbb{R}^d \setminus \{0\} \to \mathbb{R}^d, J(z) = \frac{f(\|z\|)}{\|z\|}z$$

is the gradient of a convex function.

(iv) Let $\Psi : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable function with symmetric and positive semidefinite Jacobian matrix $D\Psi(z)$ for all $z \in \mathbb{R}^d$. Then, Ψ is the gradient of a convex function.

(v) Let $H : \mathbb{R}^d \to \mathbb{R}^d$ be the gradient of a convex function, let $b \in \mathbb{R}^d$, and let $B \in \mathbb{R}^{d\times d}$ be an arbitrary matrix. Then, the function

$$\tilde{H} : \mathbb{R}^d \to \mathbb{R}^d, \tilde{H}(z) = B^T H(Bz + b)$$

is the gradient of a convex function.

Proof.

(i) Since, the Hessian

$$\nabla^2\varphi : \mathbb{R}^d \to \mathbb{R}^{d\times d}, z \mapsto A$$

is positive semidefinite on $\mathbb{R}^d$, φ is convex.

(ii) Let $f_1, \dots, f_d : \mathbb{R} \to \mathbb{R}$ have antiderivatives $F_1, \dots, F_d : \mathbb{R} \to \mathbb{R}$ which must be convex. Now, let

$$F : \mathbb{R}^d \to \mathbb{R}, F(z) = \sum_{i=1}^{d} F_i(z_i).$$

Then, trivially $\nabla F = G$ and we have for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$

$$F(\lambda x + (1 - \lambda)y) = \sum_{i=1}^{d} F_i(\lambda x_i + (1 - \lambda)y_i) \leq \sum_{i=1}^{d}(\lambda F_i(x_i) + (1 - \lambda)F_i(y_i))$$
$$=\lambda F(x) + (1 - \lambda)F(y)$$

Hence, F is convex and G is its gradient.

(iii) Let F be an antiderivative of f. Since $f \geq 0$, F is increasing and convex. By the convexity of the norm the function,

$$\hat{J} : \mathbb{R}^d \setminus \{0\} \to \mathbb{R}, \hat{J}(z) = F(\|z\|)$$

is convex. Further, we have $\nabla\hat{J} = J$ which shows the claim.

(iv) Since $D\Psi$ is symmetric on the convex domain $\mathbb{R}^d$ we may conclude from the Poincaré Lemma that Ψ has a potential. This potential must also be convex since $D\Psi$ is positive semidefinite on the whole domain.

(v) Let ψ be a convex potential of H. Set

$$\tilde{\psi} : \mathbb{R}^d \to \mathbb{R}, z \mapsto \psi(Bz + b)$$

Then, $\nabla\tilde{\psi} = H$ and we have for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$

$$\begin{aligned}\tilde{\psi}(\lambda x + (1-\lambda)y) &= \psi(B(\lambda x + (1-\lambda)y) + b) = \psi(\lambda(Bx + b) + (1-\lambda)(By + b)) \\ &\leq \lambda\psi(Bx + b) + (1-\lambda)\psi(By + b) = \lambda\tilde{\psi}(x) + (1-\lambda)\tilde{\psi}(y),\end{aligned}$$

showing the convexity of $\tilde{\psi}$. Hence, $\tilde{H}$ is a gradient of a convex function.

□

The transport of probability distributions with gradients of convex functions has been extensively studied in (McCann, 1995). The following proposition summarizes the main results.

Proposition 4.1.2 (Proposition 2.1 in (Deb et al., 2023)). *Let μ and ν be absolutely continuous probability measures in $\mathbb{R}^d$. Then, there exist gradients of convex functions $T, S : \mathbb{R}^d \to \mathbb{R}^d$ such that $T\#\mu = \nu$ and $S\#\nu = \mu$, respectively. Furthermore, T and S are unique, μ and ν a.e., respectively. We also have $(T \circ S)(x) = x$ for μ-a.a. x and $(S \circ T)(y) = y$ for ν-a.a. y, respectively.*

Hence, given two absolutely continuous distributions μ and ν, we can find a μ-a.e.-unique transport map T transporting μ to ν. Therefore, when $Y \sim \mu$, we have $Z := T(Y) \sim \nu$. Further, when given a sample $Y_1, \ldots, Y_n \overset{iid}{\sim} \mu$, the random variables $Z_1, \ldots, Z_n$, defined as $Z_i := T(Y_i)$, can be interpreted as representants of the Y-sample with respect to the distribution ν. We will later see why this is useful. The next question we deal with is: What if the distribution μ, and hence also T is unknown? How can we transport a sample $Y_1, \ldots, Y_n$ from an unknown distribution to its representants with respect to a known distribution ν? To do this, we need to discretize the mentioned concepts.

So far we have transported a distribution μ to another distribution ν. Now, when given a sample $Y_1, \ldots, Y_n \overset{iid}{\sim} \mu$ with μ unknown, we need to transport a random set of vectors representing μ to another random set representing ν. Hence, a discretization of ν will be helpful:

Definition 4.1.3 (Grid). Let $\nu \in \mathcal{P}^d$. A triangular array of random vectors $(X_k^{(n)})_{n \geq k \geq 1}$ is called *grid* of ν if

$$\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k^{(n)}} \xrightarrow{w} \nu \qquad a.s.$$

Or equivalently, there exists $\tilde{\Omega} \in \mathcal{F}$ with $\mathbb{P}(\tilde{\Omega}) = 1$ such that for all $\omega \in \tilde{\Omega}$

$$\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k^{(n)}(\omega)} \xrightarrow{w} \nu.$$

Hence, for a grid $(X_k^{(n)})_{n \geq k \geq 1}$ of ν and a large $n \in \mathbb{N}$, the empirical uniform distribution on the random set $\{X_1^{(n)}, \ldots, X_n^{(n)}\}$ is a random probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$

approximating ν. For some grids, $X_k^{(n)}$ will be independent of n and we will simply write $(X_n)_{n\in\mathbb{N}}$. The following theorem provides an easy general method to generate grids.

Theorem 4.1.4. *Let $\nu \in \mathcal{P}^d$ and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random vectors with $X_1 \sim \nu$. Then, $(X_n)_{n\in\mathbb{N}}$ is a grid of ν.*

Proof. Let $A \in \mathcal{B}(\mathbb{R}^d)$ with $\nu(\partial A) = 0$. Then, for all $n \in \mathbb{N}$, the random variable $Y_n := \mathbb{1}_{\{X_n \in A\}}$ follows a Bernoulli-distribution:

$$Y_n \sim \mathrm{Ber}(\mathbb{P}(X_1 \in A))$$

Clearly, $(Y_n)_{n\in\mathbb{N}}$ is a sequence of independent (by Lemma 3.1.8), identically distributed and integrable random variables. Hence, by the Strong Law of Large Numbers (Theorem 3.2.6), there exists $\tilde{\Omega} \in \mathcal{F}$ with $\mathbb{P}(\tilde{\Omega}) = 1$ such that for all $\omega \in \tilde{\Omega}$

$$\frac{1}{n}\sum_{k=1}^{n} Y_k(\omega) \longrightarrow \mathbb{P}(X_1 \in A).$$

Hence, we have for all $\omega \in \tilde{\Omega}$

$$\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k(\omega)}(A) = \frac{1}{n}\sum_{k=1}^{n} Y_k(\omega) \longrightarrow \mathbb{P}(X_1 \in A) = \nu(A).$$

Hence, by Lemma 3.3.2 we have shown

$$\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k} \xrightarrow{w} \nu \qquad a.s.$$

and the claim is proven. □

Example 4.1.5. Let $\nu \sim \mathrm{Unif}[0,1]$. Then, when $(X_n)_{n\in\mathbb{N}}$ is a sequence of identically distributed random variables with $X_1 \sim \nu$, $(X_n)_{n\in\mathbb{N}}$ is a grid of ν by Theorem 4.1.4.

Another choice of a grid could be the following. For $1 \le k \le n$, set $Y_k^{(n)}$ to be deterministic with $Y_k^{(n)} = \frac{k}{n}$. Let us show that $(Y_k^{(n)})_{n\ge k\ge 1}$ is a grid of ν. We have that the uniform distribution on the set

$$M_n := \{Y_1^{(n)}, \dots, Y_n^{(n)}\} = \left\{\frac{1}{n}, \frac{2}{n}, \dots, 1\right\}$$

converges weakly against ν. This is because we have for the distribution function F_n of $\mathrm{Unif}(M_n)$

$$F_n(t) = \max\left\{\frac{k}{n} : \frac{k}{n} \le t, k \in [n]_0\right\} \longrightarrow t,$$

for $t \in [0,1]$. Hence, the convergence follows from Lemma 3.3.2.

Still, our goal is to transport a sample with unknown distribution $Y_1, \dots, Y_n \sim \mu$ to another known distribution ν. With the help of grids we know where we can transport our sample to, namely to $X_1^{(n)}, \dots, X_n^{(n)}$. But what would be the transport map? We also need a discrete version of gradients of convex functions, in particular an optimal way to assign the Y_i's to the $X_i^{(n)}$'s. Therefore, we introduce the concept of cyclical monotonicity.

Definition 4.1.6 (Cyclical Monotonicity)**.** A subset $S \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is called cyclically monotone if, for all finite collection of points $\{(z_1, y_1), \dots, (z_n, y_n)\} \subseteq S$,

$$\sum_{i=1}^n \|y_i - z_i\|^2 \leq \sum_{i=1}^n \|y_i - z_{\sigma(i)}\|^2 \tag{4.1}$$

for all permutations $\sigma \in S_n$. This condition can be understood as optimal assignment of the y_i's to the z_i's with respect to the squared error loss. Equivalently one can require that for all $\{(z_1, y_1), \dots, (z_n, y_n)\} \subseteq S$,

$$\sum_{i=1}^n \langle y_i, z_i \rangle \geq \sum_{i=1}^n \langle y_i, z_{\sigma(i)} \rangle$$

for all permutations $\sigma \in S_n$. Another equivalent condition found in literature is that for all $\{(z_1, y_1), \dots, (z_n, y_n)\} \subseteq S$,

$$\sum_{i=1}^n \langle y_i, z_{i+1} - z_i \rangle \leq 0 \qquad z_{n+1} \coloneqq z_1.$$

Proof. We show that these definitions are equivalent. Let $\{(z_1, y_1), \dots, (z_n, y_n)\} \subseteq S$. We have for all $\sigma \in S_n$

$$\begin{aligned}
& \sum_{i=1}^n \|y_i - z_i\|^2 \leq \sum_{i=1}^n \|y_i - z_{\sigma(i)}\|^2 \\
\iff & \sum_{i=1}^n \left(\|y_i\|^2 + \|z_i\|^2 - 2\langle y_i, z_i \rangle\right) \leq \sum_{i=1}^n \left(\|y_i\|^2 + \|z_{\sigma(i)}\|^2 - 2\langle y_i, z_{\sigma(i)} \rangle\right) \\
\iff & \sum_{i=1}^n \left(\|y_i\|^2 + \|z_i\|^2 - 2\langle y_i, z_i \rangle\right) \leq \sum_{i=1}^n \left(\|y_i\|^2 + \|z_i\|^2 - 2\langle y_i, z_{\sigma(i)} \rangle\right) \\
\iff & \sum_{i=1}^n \langle y_i, z_i \rangle \geq \sum_{i=1}^n \langle y_i, z_{\sigma(i)} \rangle,
\end{aligned}$$

which shows the equivalence of the first two definitions. Next, we show that the second defintion implies the third. When we have

$$\sum_{i=1}^n \langle y_i, z_i \rangle \geq \sum_{i=1}^n \langle y_i, z_{\sigma(i)} \rangle,$$

for all permutations σ we can set σ to be the one with $\sigma(i) = i + 1$ for $i \in [n-1]$ and $\sigma(n) = 1$. Then, we have

$$\sum_{i=1}^n \langle y_i, z_i \rangle \geq \sum_{i=1}^n \langle y_i, z_{i+1} \rangle \implies \sum_{i=1}^n \langle y_i, z_{i+1} - z_i \rangle \leq 0,$$

where we set $z_{n+1} = z_1$. For the last implication, assume that for all $\{(z_1, y_1), \dots, (z_n, y_n)\} \subseteq S$, we have

$$\sum_{i=1}^n \langle y_i, z_{i+1} - z_i \rangle \leq 0.$$

Now, fix a set $\{(z_1,y_1),\dots,(z_n,y_n)\} \subseteq S$ and a permutation $\sigma \in S_n$. We need to show that

$$\sum_{i=1}^{n} \langle y_i, z_{\sigma(i)} - z_i \rangle \leq 0.$$

We show this first under the assumption that σ is cyclical which means that successive application of σ would take every element $i \in [n]$ through the whole set, i.e.

$$\forall i \in [n] : \{\sigma^k(i) : k \in \mathbb{N}\} = [n],$$

where $\sigma^k(i) = (\sigma \circ \cdots \circ \sigma)(i)$ is the k-times composition of σ. Now, we reindex the set with

$$(z_i', y_i') = (z_{\sigma^{i-1}(1)}, y_{\sigma^{i-1}(1)})$$

for $i \in [n]$, where σ^0 is the identity, i.e. $\sigma^0(1) = 1$. Since σ is cyclical we have $\{(z_1',y_1'),\dots,(z_n',y_n')\} = \{(z_1,y_1),\dots,(z_n,y_n)\} \subseteq S$ and therefore,

$$\begin{aligned} 0 \geq \sum_{i=1}^{n} \langle y_i', z_{i+1}' - z_i' \rangle &= \sum_{i=1}^{n} \langle y_{\sigma^{i-1}(1)}, z_{\sigma^{i}(1)} - z_{\sigma^{i-1}(1)} \rangle \\ &= \sum_{i=1}^{n} \langle y_{\sigma^{i-1}(1)}, z_{\sigma(\sigma^{i-1}(1))} \rangle - \sum_{i=1}^{n} \langle y_{\sigma^{i-1}(1)}, z_{\sigma^{i-1}(1)} \rangle \\ &= \sum_{i=1}^{n} \langle y_i, z_{\sigma(i)} \rangle - \sum_{i=1}^{n} \langle y_i, z_i \rangle = \sum_{i=1}^{n} \langle y_i, z_{\sigma(i)} - z_i \rangle. \end{aligned}$$

For the second to last equality we used that σ is cyclical. Now, consider an arbitrary σ. Let $C_1,\dots,C_k$ be the cycles of σ, i.e. all the sets of the form $\{\sigma^k(i) : k \in \mathbb{N}\}$ for $i \in [n]$. Further, let σ_j be the restriction of σ to C_j. Then, σ_j is cyclical and we have $\{(z_i,y_i) : i \in C_j\} \subseteq \{(z_1,y_1),\dots,(z_n,y_n)\} \subseteq S$. Hence, we have by our previous result

$$\sum_{i=1, i \in C_j}^{n} \langle y_i, z_{\sigma_j(i)} - z_i \rangle = \sum_{i=1, i \in C_j}^{n} \langle y_i, z_{\sigma(i)} - z_i \rangle \leq 0.$$

Summing over the cycles yields

$$0 \geq \sum_{j=1}^{k} \sum_{i=1, i \in C_j}^{n} \langle y_i, z_{\sigma(i)} - z_i \rangle = \sum_{i=1}^{n} \langle y_i, z_{\sigma(i)} - z_i \rangle$$

which finalizes the proof. □

Suppose we have chosen a way to assign the Y_i's to the $X_i^{(n)}$'s on the grid with a permutation $\sigma \in S_n$. That is, for $i \in [n]$, we transport Y_i to $X_{\sigma(i)}^{(n)}$. Then, the set

$$\left\{ (Y_1, X_{\sigma(1)}^{(n)}), (Y_2, X_{\sigma(2)}^{(n)}), \dots, (Y_n, X_{\sigma(n)}^{(n)}) \right\}$$

being cyclically monotone would imply that this assignment is optimal with respect to the squared error loss, i.e. we would have

$$\sum_{i=1}^{n} \left\| Y_i - X_{\sigma(i)}^{(n)} \right\|^2 = \min_{\pi \in S_n} \sum_{i=1}^{n} \left\| Y_i - X_{\pi(i)}^{(n)} \right\|^2.$$

Hence, we have found a discrete version for optimal transport maps. But how is this cyclical monotonicity related to gradients of convex functions? The following theorem yields the answer.

Definition 4.1.7 (Cyclically Monotone Function)**.** A function $G : \mathbb{R}^d \to \mathbb{R}^d$ is called *cyclically monotone* if its graph

$$\mathrm{Gr}(G) \coloneqq \{(z, G(z)) : z \in \mathbb{R}^d\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$$

is cyclically monotone.

Theorem 4.1.8 (Theorem 1 and Corollary 1 in (Rockafellar, 1966))**.** *A function $G : \mathbb{R}^d \to \mathbb{R}^d$ is cyclically monotone if and only if it is the gradient of a convex function.*

With the help of the following lemma we can show our solution to the problem.

Lemma 4.1.9. *Fix $y_1, \dots, y_n \in \mathbb{R}^d$ and $z_1, \dots, z_n \in \mathbb{R}^d$. Then, there exists a (not necessarily unique) permutation $\sigma \in S_n$ such that the set $\{(z_{\sigma(1)}, y_1), \dots, (z_{\sigma(n)}, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone.*

Proof. Take a permutation $\sigma \in S_n$ minimizing $\sum\limits_{i=1}^{n} \|y_i - z_{\sigma(i)}\|^2$. □

Definition 4.1.10 (Empirical Ranks)**.** Let $\mu, \nu \in \mathcal{P}_{ac}^d$. Further, let $(X_k^{(n)})_{n \geq k \geq 1}$ be a grid of ν and let $Y_1, \dots, Y_n \overset{iid}{\sim} \mu$ be a sample. Let $\sigma_n \in S_n$ be a permutation such that the set

$$\left\{(X_{\sigma_n(1)}^{(n)}, Y_1), \dots, (X_{\sigma_n(n)}^{(n)}, Y_n)\right\}$$

is cyclically monotone. We define the so called *empirical rank statistics* (or empirical representants) of the sample $Y_1, \dots, Y_n$ with respect to ν as

$$\hat{R}_\nu^n(Y_k) = X_{\sigma_n(k)}^{(n)}$$

for $k \in [n]$ (which are not necessarily unique). Further, let G be the μ-a.e. unique cyclically monotone function pushing μ forward to ν. Then, the *true rank statistics* (or true representants) of the sample $Y_1, \dots, Y_n$ with respect to ν are defined as

$$R_\nu^\mu(Y_k) = G(Y_k)$$

for $k \in [n]$ (which are defined μ-a.e.).

By the characterization of cyclical monotonicity in (4.1), we observe that the optimal transport becomes an optimal assignment problem in the discrete case.

Example 4.1.11. Let $\nu = \mathrm{Unif}[0,1]$ and let $\mu \in \mathcal{P}_{ac}^1$. Further, let $Y_1, \dots, Y_n \overset{iid}{\sim} \mu$ be an observed sample. The μ-a.e unique cyclically monotone function transporting μ to ν is the distribution function F corresponding to μ by Theorem 3.1.14. Hence, the true rank statistics of $Y_1, \dots, Y_n$ with respect to ν are $F(Y_1), \dots, F(Y_n)$. Equivalently, we have

$$R_\nu^\mu(Y_i) = F(Y_i).$$

Now, fix the grid of ν,

$$\left(\left\{\frac{1}{n},\frac{2}{n},\dots,1\right\}\right)_{n\in\mathbb{N}}$$

from Example 4.1.5. Then, the empirical rank statistics of $Y_1,\dots,Y_n$ with respect to ν are given by

$$\hat{R}^n_\nu(Y_i) = F_n(Y_i). \tag{4.2}$$

Thereby, F_n is the empirical distribution function

$$F_n : \mathbb{R} \to \left\{0,\frac{1}{n},\frac{2}{n},\dots,1\right\}, F_n(t) = \frac{1}{n}\#\{X_i : X_i \leq t, i \in [n]\}.$$

The claim in (4.2) follows since F_n is increasing, hence a cyclically monotone function and therefore, its graph is a cyclically monotone set, and $\{(X_1,F_n(X_1)),\dots,(X_n,F_n(X_n))\}$ is cyclically monotone.

This concludes our introduction into optimal transport, gradients of convex functions and cyclical monotonicity. Further literature we recommend for interested readers, are McCann (1995) and Rockafellar (1966).

Let us summarize the results so far. Given two absolutely continuous probability distributions $\mu,\nu \in \mathcal{P}^d_{ac}$, there is a μ-a.e. unique transport map T from μ to ν. This means, when given a sample $Y_1,\dots,Y_n \overset{iid}{\sim} \mu$, there are μ-a.e. unique "representants" (or rank statistics) $R^\mu_\nu(Y_1),\dots,R^\mu_\nu(Y_n)$ with respect to distribution ν. When μ is unknown, we can work with empirical rank statistics $\hat{R}^n_\nu(Y_1),\dots,\hat{R}^n_\nu(Y_n)$. But what can rank statistics/representants be used for? We provide an example in the next subsection.

4.2 Example: Quantiles and Ranks in $\mathbb{R}^d$

What is the idea of optimal transport that we are going to use? Often when we work with random variables/vectors, they may follow some arbitrary distribution. Then, we would like to make inference, e.g. estimating their correlation. Sometimes this is easier when working with some other specific distribution. Consider the following example: The Pearson's correlation defined as in (3.1) is a well established way to measure the dependency of two random variables. However, it has some disadvantages, e.g. it is undefined for random variables with infinite variance. Also, it only measures the linear dependency and one can therefore lose a part of the interval $[-1,1]$. This holds for example for a Exp(1)- and a Unif$[0,1]$-random variable. An alternative to Pearson's correlation is Spearman's ρ.

For random variables X and Y with distribution functions F_X and F_Y, Spearman's ρ is defined as

$$\rho_{X,Y} = \mathbb{C}\mathrm{orr}[F_X(X),F_Y(Y)].$$

So what does Spearman's ρ do? In the case where X and Y are absolutely continuous, we know by Theorem 3.1.14 that $F_X(X)$ and $F_Y(Y)$ have Unif$[0,1]$-distribution. Hence, Spearman's ρ calculates the correlation of the Unif$[0,1]$ versions of X and Y. In other words, one can transport X and Y to Uniform distributions on $[0,1]$ with the cyclically monotone distribution functions $F := F_X$ and $G := F_Y$, and determine the correlation $\mathbb{C}\mathrm{orr}[F(X), G(Y)]$. The transport to Unif$[0, 1]$ has two advantages: Firstly, we do not need to worry about the existence of second moments and secondly, Spearman's ρ is invariant under increasing transformations. Thirdly, values over the whole interval $[-1, 1]$ are always attainable for absolutely continuous random variables.

We will use this idea of transporting to a more convenient distribution to work with, when we introduce our new method for Canonical Correlation Analysis. In order to understand this concept better, we provide another example beforehand, namely the Centre-Outward Distribution Function introduced in (Hallin et al., 2017). The theory around it suggests a way to generalize the distribution function, the empirical distribution function, the median, rank statistics and quantiles to higher dimensions. Note that the main goal of this section is not to introduce this concept, but to comprehend the use of optimal transport and the definitions in the previous section illustrated at this example, both in the distribution and in the empirical case.

Consider the following problem: Suppose μ is a distribution in dimension 1 with corresponding distribution function F,

$$F : \mathbb{R} \to [0, 1], F(t) = \mu((-\infty, t]).$$

We are now interested in the quantiles of μ. As we are working in dimension 1, we can simply define the q-th quantile as $F^{-1}(q)$ for $q \in (0, 1)$, where F^{-1} is the generalized inverse of F defined in Definition 3.1.13. Similarly, the median is given by $F^{-1}(1/2)$. But what if we were working in dimension $d > 1$? The quantile can not be analogously defined in higher dimensions as we do not have a canonical ordering of $\mathbb{R}^d$ as in the one-dimensional case.

The empirical case is also interesting. When given a sample $X_1, \ldots, X_n \sim \mu$, the order statistics are defined such that

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

Then, one can easily define the median as the value in the middle (in the n odd case). Due to the canonical ordering of $\mathbb{R}$ it is also very easy to define the empirical quantiles. But what if we were working in a higher dimension? Also, what is the empirical distribution function in higher dimensions?

Transport maps can help us suggest a solution this problem. As established in Example 4.1.11, the distribution function is a transport map from a distribution μ to the Unif$[0, 1]$ distribution. Similarly, in the empirical case the empirical distribution function transports the sample to the set $\left\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}$ which approximates Unif$[0, 1]$. The problem in higher dimensions is that it is difficult to work with the set $[0, 1]$ equipped with Unif$[0, 1]$ on which the theory in dimension 1 is based on. Instead we will work with the unit ball in higher dimensions equipped with a type uniform distribution. We will later see how this can be done.

For now, let us begin with some intuition. Let $X \sim \mu$ with $\mu \in \mathcal{P}^1_{ac}$ be absolutely continuous and have distribution function F. Consider the increasing function $F_\pm := 2F - 1$, which we for now call center-outward distribution function. It is cyclically monotone since it is increasing, and it transports μ to the uniform distribution on the unit ball $U_1 = \mathrm{Unif}(-1, 1)$. The question is: Why is this transformation useful? The answer is: It is not very helpful in dimension 1, but in higher dimensions it is. The center-outward distribution $F_\pm$ contains the same information as F and we can define familiar and new terms:

- The medians are given by the elements of $F_\pm^{-1}(\{0\})$.
- The center-outward quantile regions are defined as $\mathbb{C}(q) := F_\pm^{-1}([-q, q])$ for $q \in (0, 1)$.
- The center-outward quantile contours are defined as $\mathcal{C}(q) := F_\pm^{-1}(\{-q, q\})$ for $q \in (0, 1)$.

We observe the difference between the classical distribution function and the center-outward distribution function. As the name already suggests the center-outward distribution function defines quantiles from a center-outward perspective. Everything is centered around 0. We will later see that we can generalize this to higher dimensions. Next, let us consider the empirical case in dimension 1.

Suppose we are given a sample $X_1, \dots, X_n \overset{iid}{\sim} \mu$. Then, the empirical rank statistics[1] $R_1^{(n)}, \dots, R_n^{(n)}$ are defined as

$$R_i^{(n)} = \#\{k \in [n] : X_k \leq X_i\} \qquad i \in [n].$$

Further, the classical empirical distribution is given by

$$F^{(n)}(t) = \frac{1}{n}\#\{k \in [n] : X_k \leq t\}.$$

Observe that only the values of $F^{(n)}$ on the sample $\{X_1, \dots, X_n\}$ are uniquely defined, as $F^{(n)}$ can be defined arbitrarily in between as long as it remains increasing and we have $\lim_{t \to -\infty} F^{(n)}(t) = 0$ and $\lim_{t \to \infty} F^{(n)}(t) = 1$. We can also say that $F^{(n)}$ maps the order statistics

$$X_{(1)}, X_{(2)}, \dots, X_{(n)} \qquad \text{to} \qquad \frac{1}{n}, \frac{2}{n}, \dots, 1.$$

We can now define the empirical center-outward distribution $F_\pm^{(n)}$ analogously by mapping

$$X_{(1)}, X_{(2)}, \dots, X_{(n)} \qquad \text{to} \qquad -\frac{n-1}{n+1}, -\frac{n-3}{n+1}, \dots, \frac{n-1}{n+1},$$

where we assume n to be odd[2]. Outside the sample $F_\pm^{(n)}$ can be defined such that it is increasing and we have $\lim_{t \to -\infty} F_\pm^{(n)}(t) = -1$ and $\lim_{t \to \infty} F_\pm^{(n)}(t) = 1$. Again we can define the familiar and new terms:

[1] Note that both the statistics from Definition 4.1.10 and here are called rank statistics. The ones from the definition are with respect to a distribution. As the rank statistics from here take values in $[n]$, they may be interpreted as empirical rank statistics with respect to the uniform distribution on $[n]$. This can be shown with a similar argument as in Example 4.1.11.

[2] The case, n even, is also studied in (Hallin et al., 2017). It is very similar, but with slightly more unconvenient notation. As we are mainly interested in the idea behind it than introducing the concept we do not consider the n even case.

- The empirical median is given by $F_{\pm}^{(n)^{-1}}(\{0\})$.
- The center-outward ranks $R_{\pm,i}^{(n)}$ are defined as $R_{\pm,i}^{(n)} = \left| R_i^{(n)} - \frac{n+1}{2} \right|$ for $i \in [n]$.
- The center-outward signs $S_{\pm,i}^{(n)}$ are defined as $S_{\pm,i}^{(n)} = \operatorname{sign}\left(R_i^{(n)} - \frac{n+1}{2} \right)$ for $i \in [n]$.
- The empirical quantile regions $\mathbb{C}_{\pm}^{(n)}(q) = \left\{ X_i : R_{\pm,i}^{(n)} \leq \frac{q(n+1)}{2} \right\}$ for $q \in (0,1)$.
- The empirical quantile contours $\mathcal{C}_{\pm}^{(n)}\left(\frac{2j}{n+1}\right) = \left\{ X_i : R_{\pm,i}^{(n)} \leq j \right\}$ for $j \in \left[\frac{n+1}{2}\right]$.

Everything is again defined in a center-outward perspective. Also, note that the set $\left\{-\frac{n-1}{n+1}, -\frac{n-3}{n+1}, \dots, \frac{n-1}{n+1}\right\}$ is a grid of Unif$[-1,1]$, when interpreted as a triangular array of fixed points increasing with n.

This concludes the definitions for dimension 1. We are now able to examine the interesting part namely, generalizing these concepts to the higher dimension. Thereby, we will be working with the unit ball equipped with a uniform type distribution which we introduce next:

Definition 4.2.1 (Uniform Distribution on the Unit Ball)**.** Let S be uniformly distributed on the unit sphere $\mathcal{S}_{d-1}$ and let U be uniformly distributed on $[0,1]$. Then, we call the distribution of US the uniform distribution U_d on the d-dimensional (open) unit ball $\mathbb{S}_d$. Note that this is not the actual uniform distribution on the unit ball in dimensions $d \geq 2$.

In dimension 1 the center-outward distribution function was the μ-a.e. unique function transporting μ to U_1. This property will carry over to the higher dimension, as we will be working with $\mathbb{S}_d$ equipped with U_d. We can now define the center-outward distribution function based on Proposition 4.1.2.

Definition 4.2.2 (Center-Outward Distribution Function)**.** Let $X \sim \mu \in \mathcal{P}_{ac}^d$.

- The *center-outward distribution function* of X is the unique gradient of a convex function $F_{\pm}$ pushing μ forward to U_d.
- The corresponding *center-outward quantile function* $Q_{\pm}$ is defined as the unique gradient of a convex function pushing U_d forward to μ.
- The *quantile regions* of order $q \in (0,1)$ are defined as $\mathbb{C}(q) = Q_{\pm}(q\bar{\mathbb{S}}_d)$
- The *quantile contours* of order $q \in (0,1)$ are defined as $\mathcal{C}(q) = Q_{\pm}(q\mathcal{S}_{d-1})$.

This definition is slightly different to the corresponding Definition 4.1 in (Hallin et al., 2017). There, the center-outward distribution function is defined for non-vanishing distributions which is a stronger requirement:

Definition 4.2.3 (Non-Vanishing)**.** An absolutely continuous distribution $\mu \in \mathcal{P}_{ac}^d$ is called non-vanishing if its density f satisfies the following property: For all $D > 0$ there exist $0 < \lambda_D < \Lambda_D$ such that

$$\lambda_D < f(x) < \Lambda_D$$

for all x with $\|x\| \leq D$. We write $\mu \in \mathcal{P}_{nv}^d$.

For non-vanishing distributions, Definition 4.2.2 becomes particularly meaningful. In that case, Proposition 4.2 in (Hallin et al., 2017) shows that the quantile regions $\mathbb{C}(q)$ have boundaries $\mathcal{C}(q)$, and are connected and nested as q increases from 0 to 1. Further, the set $F_{\pm}^{-1}(\{0\})$ is a compact set of measure zero and can be defined as the set of center-outward medians. Note that the center-outward medians can not be well-defined in the absolutely continuous case. Proposition 4.1.2 only provides the existence of transport maps almost everywhere.

In summary, the transport map from an arbitrary non-vanishing distribution $\mu \in \mathcal{P}_{nv}^d$ to the uniform distribution on the unit ball U_d enabled us to generalize the distribution function, the median, and quantiles to the higher dimension. Optimal transport theory will also be very useful in the empirical case, as we will see next.

In order to generalize our concepts we will need to find a grid for the uniform distribution on the unit ball U_d:

Definition 4.2.4. We will formulate the grid visualized in Figure 4.3 mathematically. Let $(u_n)_{n\in\mathbb{N}} \overset{iid}{\sim} \mathrm{Unif}(\mathcal{S}_{d-1})$ fix a sequence of directions. Next, factorize n as

$$n = n_R n_S + n_0, \quad n_R, n_S, n_0 \in \mathbb{N}, \quad 0 \leq n_0 < \min(n_R, n_S),$$

where $n_R, n_S \to \infty$ when $n \to \infty$. Now, we obtain a grid of $n_R n_S$ points in the unit ball at the intersection of the directions $(u_1, \dots, u_{n_S})$ and n_R hyperspheres centered at the origin with radii $1/(n_R+1), \dots, n_R/(n_R+1)$. The remaining points are n_0 indistinguishable copies of the origin if $n_0 > 0$.

Let us denote this collection of points by $Z_1^{(n)}, \dots, Z_n^{(n)}$. We define the discrete distribution which assigns the probability mass $1/n$ to every point but the origin, and the probability mass n_0/n to the origin, i.e.

$$U_d^{(n)} \sim \frac{1}{n}\sum_{i=1}^{n} Z_i^{(n)}.$$

We call $U_d^{(n)}$ the *uniform distribution over the augmented grid.* By (Hallin et al., 2017), the array $(Z_k^{(n)})_{n\geq k\geq 1}$ is in fact a grid of U_d in the sense of Definition 4.1.3, i.e. we have

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^{(n)} \xrightarrow{w} U_d \qquad a.s.$$

Now, the grid for U_d enables us to define the empirical center-outward distribution function. Recall that in dimension 1, we were only interested in the restriction of the empirical center-outward distribution function to the sample on which it was cyclically monotone. This will also carry over to higher dimensions.

Definition 4.2.5 (Empirical Center-Outward Distribution Function)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} \mu$ be a sample with $\mu \in \mathcal{P}^d$. Let $(Z_k^{(n)})_{n\geq k\geq 1}$ be the grid of U_d as in Definition 4.2.4. We define the empirical center-outward distribution function as

$$F_{\pm}^{(n)} : \{X_1, \dots, X_n\} \to \{Z_1^{(n)}, \dots, Z_n^{(n)}\}, F_{\pm}^{(n)}(X_i) = \hat{R}_{U_d}^{n}(X_i),$$

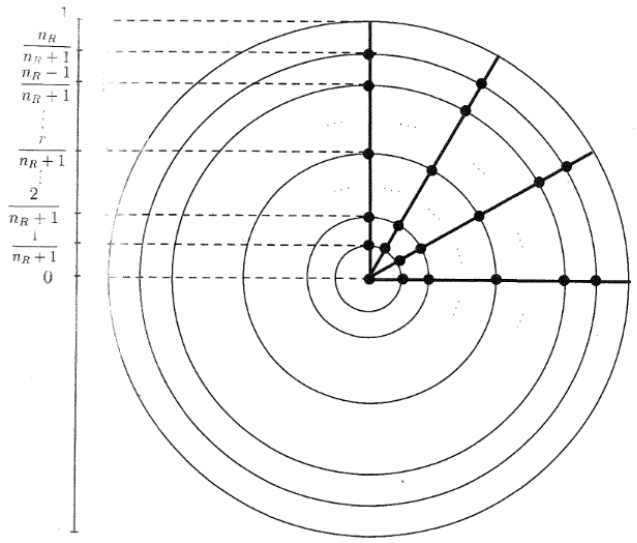

Figure 4.3: The augmented grid for U_2 (taken from page 15 in (Hallin et al., 2017)).

where the empirical rank statistics are with respect to the grid $(Z_k^{(n)})_{n\geq k\geq 1}$. Equivalently, we have

$$F_{\pm}^{(n)}(X_i) = Z_{\sigma(i)}$$

for an optimal permutation σ which satisfies

$$\sum_{i=1}^{n} \left\| X_i - Z_{\sigma(i)} \right\|^2 = \min_{\pi \in S_n} \sum_{i=1}^{n} \left\| X_i - Z_{\pi(i)} \right\|^2 .$$

In particular, the set

$$\left\{ \left(X_1, F_{\pm}^{(n)}(X_1) \right), \ldots, \left(X_n, F_{\pm}^{(n)}(X_n) \right) \right\}$$

is cyclically monotone.

The following picture illustrates the center-outward distribution of a sample.

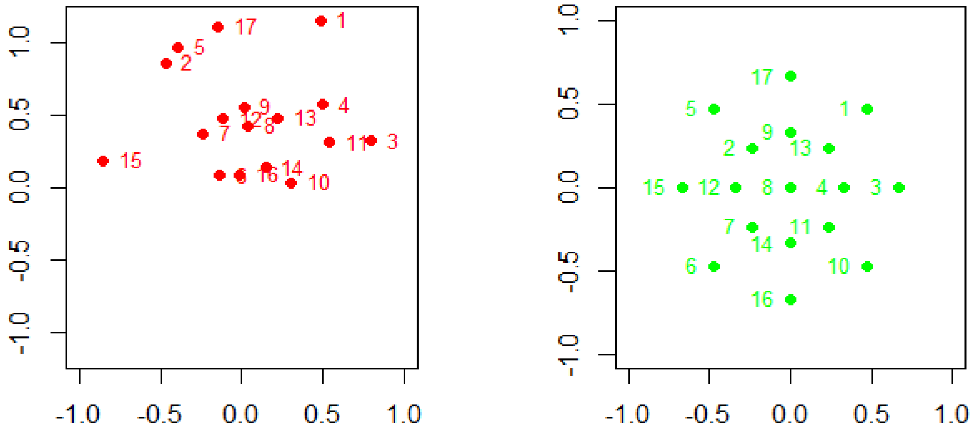

Figure 4.4: Center-Outward Distribution function of a sample of $\mathrm{Exp}(1) \otimes \mathcal{N}(0, 1)$

Example 4.2.6. What can we see in the plots? On the left hand side there is a $n = 17$ sample of the distribution $\mu_1 \otimes \mu_2$, where $\mu_1 \sim \text{Exp}(1)$ and $\mu_2 \sim \mathcal{N}(0,1)$ are independent. On the right hand side we see the grid of U_2 as in Definition 4.2.4. Further, we see which sample point gets mapped to which point on the grid, e.g. the observed point with number 8 is the empirical center-outward median of the sample.

With the empirical distribution function we can now define the familiar concepts in a higher dimension:

Definition 4.2.7. Let $X_1, \dots, X_n \overset{iid}{\sim} \mu$ be a sample with $\mu \in \mathcal{P}^d$ and empirical center-outward distribution function $F_{\pm}^{(n)}$. Then, we define

- center-outward ranks: $R_{\pm,i}^{(n)} = (n_R + 1)\|F_{\pm}^{(n)}(X_i)\|$ for $i \in [n]$.
- center-outward signs: $S_{\pm,i}^{(n)} = \frac{F_{\pm}^{(n)}(X_i)}{\|F_{\pm}^{(n)}(X_i)\|}$ for $i \in [n]$ and 0 if undefined.
- center-outward quantile regions: $\mathbb{C}_{\pm}^{(n)}(q) = \{X_i \mid R_{\pm,i}^{(n)} \leq q n_R\}$ for $q \in (0,1)$.
- center-outward quantile contours: $\mathcal{C}_{\pm}^{(n)}(j/n_R) = \{X_i \mid R_{\pm,i}^{(n)} = j\}$ for $j \in [n_R]$.

Next, we show examples for the mentioned terms.

Example 4.2.8. Consider again the sample from Example 4.2.6 with $n = 17$. We have $n_0 = 1$, $n_R = 2$ and $n_S = 8$. Some calculations:

$$R_{\pm,12}^{(17)} = 1, \quad R_{\pm,1}^{(17)} = 2, \quad S_{\pm,14}^{(17)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad S_{\pm,5}^{(17)} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

The following are examples for quantile regions and quantile contours.

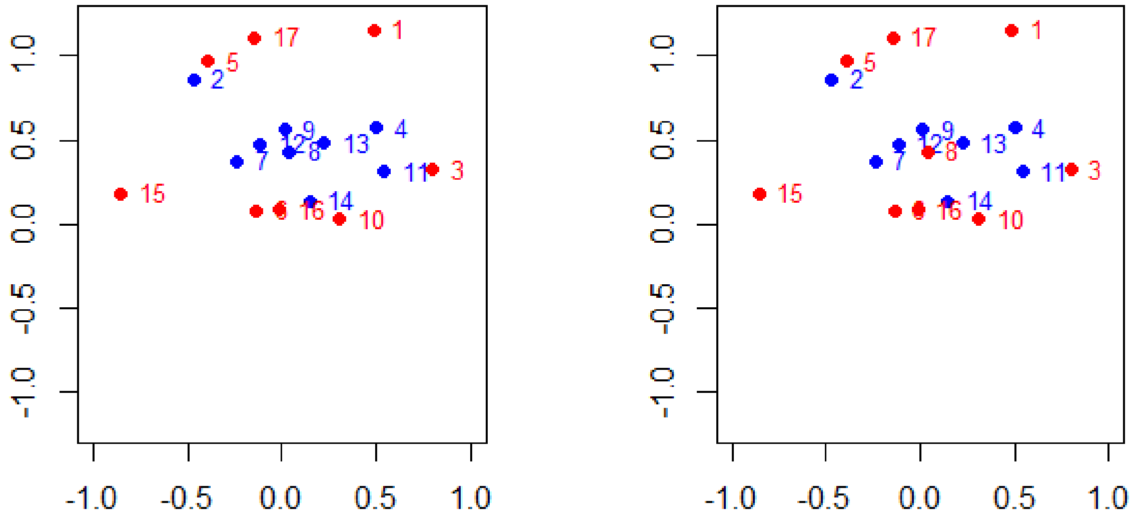

Figure 4.5: Quantile Region $\mathbb{C}_{\pm}^{(17)}(1/2)$ (left) and Quantile Contour $\mathcal{C}_{\pm}^{(17)}(1/2)$ (right)

Finally, the center-outward distribution has a Glivenko-Cantelli type property:

Theorem 4.2.9 (Theorem 5.1 in (Hallin et al., 2017))**.** *Let $(X_n)_{n\in\mathbb{N}} \overset{iid}{\sim} \mu$ be a sample with $\mu \in \mathcal{P}^d_{nv}$ and empirical center-outward distribution function $F^{(n)}_{\pm}$. Further, let $F_{\pm}$ be the actual center-outward distribution function with respect to μ. Then,*

$$\lim_{n\to\infty} \max_{i=1,\dots,n} \left\| F^{(n)}_{\pm}(X_i) - F_{\pm}(X_i) \right\| = 0 \qquad a.s.$$

When rewriting the statement with the rank statistics from Definition 4.1.10, we have

$$\lim_{n\to\infty} \max_{i=1,\dots,n} \left\| \hat{R}^n_{U_d}(X_i) - R^{\mu}_{U_d}(X_i) \right\| = 0 \qquad a.s.$$

Hence, the empirical rank statistics of the X_i's on U_d converge uniformly against the true rank statistics on U_d.

Let us summarize the results of this subsection. Although we see how strong the theory of the center-outward distribution function is with the defined concepts in Definition 4.2.7 and the property in Theorem 4.2.9, this was not the main purpose of this subsection. The idea was to get a better understanding for the use of optimal transport. So what can we learn from this example? When we have theory around some distribution ν, but we are working with a $Y \sim \mu$, we can transport μ to ν. Then, we can use the theory for ν with the rank statistic $R^{\mu}_{\nu}(Y)$. When working with a sample $Y_1, \dots, Y_n \overset{iid}{\sim} \mu$ and μ is unknown one needs a grid of ν and can use the empirical rank statistics $\hat{R}^n_{\nu}(Y_1), \dots, \hat{R}^n_{\nu}(Y_n)$ instead.

This concept has been a focal point of research in recent years. For example, tests of equality and the independence of probability distributions based on rank statistics have been proposed in (Deb and Sen, 2019) and (Shi et al., 2022), respectively. Also, rank statistics can be used to define multivariate analogues of Spearman's ρ and Kendall's τ as suggested in (Shi et al., 2024). Overall, there any many possible applications for rank statistics. We will use them for estimation in Canonical Correlation Analysis later.

5 Canonical Correlation Analysis

Now, we have introduced the measure-theoretic probability and optimal transport theory we need. This chapter will deal with Canonical Correlation Analysis (CCA). We begin with a recap of the multivariate normal distribution as the classical CCA theory is based on it. Thereafter, we continue with the motivation, history, and the model of CCA. Next, the Gaussian Copula CCA, a generalization to the classical model, will be introduced. Finally, we will present our main topic, the Cyclically Monotone CCA model. Throughout the chapter we discuss the formulations of all three models, present an estimation method for each and prove their consistency.

5.1 Recap: Multivariate Normal Distribution

Definition 5.1.1 (Normal Distribution)**.**

- A random variable Z follows a univariate standard normal distribution if Z has the density
$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad z \in \mathbb{R}$$
with respect to the Lebesgue-measure in $\mathbb{R}$. We write $Z \sim \mathcal{N}(0,1)$.
- A random variable Y follows a univariate normal distribution if $Y = m + \sigma Z$ for some $m \in \mathbb{R}, \sigma \geq 0$ and $Z \sim \mathcal{N}(0,1)$. We write $Y \sim \mathcal{N}(m, \sigma^2)$.
- A random vector $X = (X_1, \dots, X_d)$ follows a multivariate normal distribution, denoted by $X \sim \mathcal{N}_d(m, \Sigma)$, if for all $a \in \mathbb{R}^d$, $a^T X$ is univariate normal with $a^T X \sim \mathcal{N}(a^T m, a^T \Sigma a)$, where $m \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is positive semi-definite.

Definition 5.1.2. A quadratic and symmetric matrix $A \in \mathbb{R}^{d \times d}$ is called positive-semidefinite if for all $x \in \mathbb{R}^d$, $x^T A x \geq 0$. A positive-semidefinite matrix is called positive definite if it is also invertible.

The next important property of the normal distribution is that all of its moments exist:

Proposition 5.1.3. *Let $Y \sim \mathcal{N}(m, \sigma^2)$ be normally distributed with $m \in \mathbb{R}$ and $\sigma \geq 0$. Then, all moments of Y exist and we have $\mathbb{E}[Y] = m$ and $\mathbb{V}\mathrm{ar}[Y] = \sigma^2$. Hence, by the definition of the multivariate normal distribution, all its marginal and product moments exist, too. In particular, we have $\mathbb{E}[X] = m$ and $\mathbb{V}\mathrm{ar}[X] = \Sigma$ for $X \sim \mathcal{N}_d(m, \Sigma)$.*

Proof. We refer to page 45 and Proposition 5.1 in (Bilodeau and Brenner, 1999). □

Other properties we are interested in are linear transformations, independence, and conditional distributions.

Proposition 5.1.4 (Proposition 5.2 and 5.3 in (Bilodeau and Brenner, 1999))**.** *Let $X \sim \mathcal{N}_d(m, \Sigma)$, let $b \in \mathbb{R}^n$ and let $A \in \mathbb{R}^{n \times d}$. Then, $AX + b \sim \mathcal{N}_n(Am + b, A\Sigma A^T)$. In particular, if*
$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_d\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$
then, $X_1 \sim \mathcal{N}_{d_1}(m_1, \Sigma_{11})$, where d_1 is the length of m_1.

Proposition 5.1.5 (Proposition 5.4 in (Bilodeau and Brenner, 1999))**.** *Let $X \sim \mathcal{N}_d(m, \Sigma)$ with*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_d\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Then, X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$. In that case we write $X_1 \perp\!\!\!\perp X_2$.

Proposition 5.1.6 (Proposition 5.6 in (Bilodeau and Brenner, 1999))**.** *Let $X \sim \mathcal{N}_d(m, \Sigma)$ with Σ positive definite and*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_d\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Then, the conditional distribution of X_1 given the event $X_2 = x_2$ for some $x_2 \in \mathbb{R}^{d_2}$ is given by

$$X_1 | X_2 = x_2 \sim \mathcal{N}_{d_1}\left(m_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - m_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right).$$

Thereby, the matrix $\Sigma_{11.2} \coloneqq \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ is called the Schur-complement.

Proposition 5.1.7. *Let $m \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ be positive definite. Then, there exists a random vector X with $X \sim \mathcal{N}_d(m, \Sigma)$.*

Proof. By Theorem 3.2.4 in (Tong, 1990), we have $X \sim \mathcal{N}_d(m, \Sigma)$ if there exist $Z_1, \dots, Z_d \overset{iid}{\sim} \mathcal{N}(0, 1)$ and $A \in \mathbb{R}^{d \times d}$ such that $AZ + m = X$ and $AA^T = \Sigma$. By generalized inverse transform in Theorem 3.1.14, there exists $Z_1 \sim \mathcal{N}(0, 1)$. It is an elementary result from probability theory that we can always construct independent random variables. Hence, there exist $Z_1, \dots, Z_d \overset{iid}{\sim} \mathcal{N}(0, 1)$. By the Cholesky factorization in (Horn and Johnson, 2013, Corollary 7.2.9), we can find $L \in \mathbb{R}^{d \times d}$, so that $LL^T = \Sigma$. Letting $Z = (Z_1, \dots, Z_d)$ and $X = LZ + m$ shows the claim. □

This finalizes our recap for the multivariate normal distribution. We can now begin with the history of Canonical Correlation Analysis.

5.2 Classical CCA

5.2.1 History and Motivation

The original theory for Canonical Correlation Analysis (CCA) was developed by Harold Hotelling in (Hotelling, 1936). He examined the idea of applying correlations and regressions not just to one-dimensional random variables.

An example for the theory of correlation is marksmen, side by side, firing shots at a target. The deviations of their shots are partly due to individual errors and partly to a common cause such as wind. When one wanted to investigate this association with approaches available at that point of time, one could calculate the correlation of one-dimensional variables. For example the correlation of the x-deviation of the shots and the x-velocity of the wind could have been determined. However, studying only this dependence would lead to an incomplete analysis as there are other variables which could be analyzed, such as correlation of the x-,y- and z-velocity of the wind, and the x- and y-deviation of the shots. But if one aims to make a more detailed analysis involving all mentioned variables, the problem becomes more complicated. Of course one could estimate a 2×3 block of a correlation matrix in this case, providing 6 correlation values for interpretation. However, in a higher dimensional data set this option would not be suitable as there are too many correlations to work with.

One example could be the following scenario: Suppose one is interested in studying the association of returns of two baskets of assets in a financial market. Thereby, one aims to study the dependence between the two sets of assets rather than the correlation inside each set. When the size of each basket is not small, let us say $n \geq 30$, then, the corresponding block of their correlation matrix contains at least 900 elements which is not very informative. Canonical Correlation Analysis provides a solution to overcome this problem. The approach is to find one linear combination of assets for each basket which have the highest correlation. Thereafter, one can find two new linear combinations of assets which are uncorrelated with the previous which have the highest correlation and so on. The investigator would then observe that the interrelation of the two sets can almost entirely be explained by a few of such linear combinations.

Today, CCA has become an established method to analyze the correlation of non-overlapping sets of random variables. Let us provide another example: Suppose one is interested in the academic success of students based on their personality. Variables for academic success could be grades in subjects, like:

- X_1: Languages
- X_2: Mathematics
- X_3: Humanities
- X_4: Science

For Personality Traits one could take the ones from the Five-Factor Model:

- Y_1: Agreeableness (being compassionate and polite)
- Y_2: Conscientiousness (being industrious and orderly)
- Y_3: Extraversion (being enthusiastic and assertive)
- Y_4: Neuroticism (affinity towards negative emotion, withdrawal and volatility)
- Y_5: Openness (intellect and openness to experience)

The variables $X = (X_1, \ldots, X_4)$ and $Y = (Y_1, \ldots, Y_5)$ are taken from (Journals, 2021). Now, instead of considering all pairs of correlations, the approach of CCA is to learn which combination of personality traits is correlated with strengths in which combination of subjects.

The goal is to find linear combinations $\alpha_1^T X$ and $\gamma_1^T Y$ which have the maximum correlation $\lambda_1 := \mathbb{C}\text{orr}[\alpha_1^T X, \gamma_1^T Y]$ with $\alpha_1 \in \mathbb{R}^4$ and $\gamma_1 \in \mathbb{R}^5$. Then, λ_1 is called the first canonical correlation of X and Y, and $\alpha_1^T X$ & $\gamma_1^T Y$ are called canonical variables. Hypothetically, one pair of such linear combinations could look like this: High conscientiousness combined with low neuroticism is positively correlated with good grades in mathematics and science. Note that as described before we do not study the dependency within each set. A relationship that CCA does not cover is e.g., the correlation of good grades in languages and humanities with the performance in mathematics and science.

After identifying the first canonical variables $\alpha_1^T X$ and $\gamma_1^T Y$, one can then continue to find new linear combinations $\alpha_2^T X$ and $\gamma_2^T Y$ maximizing the correlation $\lambda_2 := \mathbb{C}\mathrm{orr}[\alpha_2^T X, \gamma_1^T Y]$ with $\alpha_2 \in \mathbb{R}^4$ and $\beta_2 \in \mathbb{R}^5$ subject to being uncorrelated with the first canonical variables $\alpha_1^T X$ and $\gamma_1^T Y$. Thereby, this constraint is there to ensure that the new pair of linear combinations explains dependencies that have not been covered before. Then, λ_2 is called the second canonical correlation of X and Y and so on. We will later see that this could be done 4 times, as this the number of variables in the smaller data set. In practice one would stop after a satisfying number of canonical correlations, in particular when most of the interrelations have been explained.

For further reading about Canonical Correlation Analysis we refer to (Hotelling, 1936), (Anderson, 2003), and (Johnson and Wichern, 2007). In the following we will introduce CCA mainly following (Anderson, 2003).

5.2.2 The Classical CCA Model

In the end of this section we will provide the main theorem about CCA, namely Theorem 5.2.1 which describes the theory concisely. For now let us begin with some intuitive steps for the derivation of it.

One fundamental assumption of classical CCA is multivariate normality which also implies invariance of the canonical correlations under linear combinations of the variables. Let two multivariate normal data sets $Y_1 \sim \mathcal{N}_{p_1}(m_1, \Sigma_{11}), Y_2 \sim \mathcal{N}_{p_2}(m_2, \Sigma_{22})$ with $p_1 \leq p_2$ be given such that they are jointly normal

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}_p \left(m := \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \Sigma := \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Thereby, we assume $\Sigma \in \mathbb{R}^{p \times p}$ to be positive definite. As described we are interested in finding linear combinations maximizing the correlation, i.e. solving the problem

$$\max_{\alpha_1 \in \mathbb{R}^{p_1}, \gamma_1 \in \mathbb{R}^{p_2}} \lambda_1 := \mathbb{C}\mathrm{orr}[\alpha_1^T Y_1, \gamma_1^T Y_2]. \tag{P1}$$

If this problem can be solved, we want to find new linear combinations $\alpha_2^T Y_1$ and $\gamma_2^T Y_2$ maximizing their correlation subject to being independent[3] of all previous linear combinations, i.e.

$$\begin{aligned} &\max_{\alpha_2 \in \mathbb{R}^{p_1}, \gamma_2 \in \mathbb{R}^{p_2}} \lambda_2 := \mathbb{C}\mathrm{orr}[\alpha_2^T Y_1, \gamma_2^T Y_2] \\ &\text{subject to} \qquad (\alpha_2^T Y_1, \gamma_2^T Y_2) \perp\!\!\!\perp (\alpha_1^T Y_1, \gamma_1^T Y_2). \end{aligned} \tag{P2}$$

If this problem can be solved, we again aim to find new linear combinations $\alpha_3^T Y_1$ and $\gamma_3^T Y_2$ maximizing their correlation subject to being independent of all previous linear

[3] Note that independence and having zero correlation are equivalent for normal random vectors by Proposition 5.1.5.

combinations, i.e.

$$\begin{aligned}\max_{\alpha_3\in\mathbb{R}^{p_1},\gamma_3\in\mathbb{R}^{p_2}} \lambda_3 &:= \mathbb{C}\mathrm{orr}[\alpha_3^T Y_1, \gamma_3^T Y_2] \\ \text{subject to} \quad & (\alpha_3^T Y_1, \gamma_3^T Y_2) \perp\!\!\!\perp (\alpha_1^T Y_1, \gamma_1^T Y_2, \alpha_2^T Y_1, \gamma_2^T Y_2).\end{aligned} \tag{P3}$$

Let us rewrite these problems for simplicity. Since Σ is positive definite (p.d.), Σ_{11} and Σ_{22} must be p.d., too. Hence, they have p.d. square roots $\Sigma_{11}^{1/2}$ and $\Sigma_{22}^{1/2}$, respectively. Now consider the random variable

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \Sigma_{11}^{-1/2}(Y_1 - m_1) \\ \Sigma_{22}^{-1/2}(Y_2 - m_2) \end{pmatrix} \sim \mathcal{N}_p\left(0, \begin{pmatrix} I_{p_1} & \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} \\ \Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2} & I_{p_2} \end{pmatrix}\right),$$

where we derived its distribution with Proposition 5.1.4. Now, (P1) can be rewritten to

$$\max_{\alpha_1\in\mathbb{R}^{p_1},\gamma_1\in\mathbb{R}^{p_2}} \lambda_1 := \mathbb{C}\mathrm{orr}[\alpha_1^T(\Sigma_{11}^{1/2} Z_1 + m_1), \gamma_1^T(\Sigma_{22}^{1/2} Z_2 + m_2)].$$

Since $\Sigma_{11}^{1/2}$ and $\Sigma_{22}^{1/2}$ are invertible, and the correlation is invariant under the addition of constants, we conclude that by reparametrization (P1) is equivalent to

$$\max_{\xi_1\in\mathbb{R}^{p_1},\zeta_1\in\mathbb{R}^{p_2}} \lambda_1 := \mathbb{C}\mathrm{orr}[\xi_1^T Z_1, \zeta_1^T Z_2].$$

Further, w.l.o.g. we can require our linear combinations to have unit variance. This condition can be expressed as

$$\begin{aligned} 1 &= \mathbb{V}\mathrm{ar}[\xi_1^T Z_1] = \mathbb{E}[\xi_1^T Z_1 Z_1^T \xi_1] = \xi_1^T \mathbb{E}[Z_1 Z_1^T]\xi_1 = \xi_1^T \xi_1 \\ 1 &= \mathbb{V}\mathrm{ar}[\zeta_1^T Z_2] = \mathbb{E}[\zeta_1^T Z_2 Z_2^T \zeta_1] = \zeta_1^T \mathbb{E}[Z_2 Z_2^T]\zeta_1 = \zeta_1^T \zeta_1. \end{aligned}$$

With $W := \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$, we then have

$$\mathbb{C}\mathrm{orr}[\xi_1^T Z_1, \zeta_1^T Z_2] = \frac{\mathbb{C}\mathrm{ov}[\xi_1^T Z_1, \zeta_1^T Z_2]}{\sqrt{\mathbb{V}\mathrm{ar}[\xi_1^T Z_1]\mathbb{V}\mathrm{ar}[\zeta_1^T Z_2]}} = \frac{\mathbb{E}[\xi_1^T Z_1 Z_2^T \zeta_1]}{\sqrt{\xi^T\xi \cdot \zeta^T\zeta}} = \xi_1^T W \zeta_1,$$

by Proposition 5.1.4. Finally, (P1) can be rewritten as

$$\begin{aligned}\max_{\xi_1\in\mathbb{R}^{p_1},\zeta_1\in\mathbb{R}^{p_2}} \lambda_1 &= \xi_1^T W \zeta_1 \\ \text{subject to} \quad & \xi_1^T \xi_1 = 1 \\ & \zeta_1^T \zeta_1 = 1.\end{aligned} \tag{P1*}$$

With Lagrange multipliers this problem can be solved, see Section 12.2 in (Anderson, 2003), and there exist maximizers ξ_1, ζ_1, and λ_1 to (P1*) such that $\lambda_1 = \mathbb{C}\mathrm{orr}[\xi_1^T Z_1, \zeta_1^T Z_2]$. One can easily observe that the maximizers for (P1) are then given by λ_1, $\alpha_1 = \Sigma_{11}^{-1/2}\xi_1$ and $\gamma_1 = \Sigma_{22}^{-1/2}\zeta_1$.

Next, we rewrite (P2). The new condition $(\alpha_2^T Y_1, \gamma_2^T Y_2) \perp\!\!\!\perp (\alpha_1^T Y_1, \gamma_1^T Y_2)$ is equivalent to $(\xi_2^T Z_1, \zeta_2^T Z_2) \perp\!\!\!\perp (\xi_1^T Z_1, \zeta_1^T Z_2)$. Since these linear combinations are normally distributed by Proposition 5.1.4 and independence for normal variables is equivalent to having 0 covariance by Proposition 5.1.5, the condition is equivalent to all four pairs having covariance 0, i.e.

$$\begin{aligned}0 &= \mathbb{E}[\xi_2^T Z_1 Z_1 \xi_1] = \xi_2^T \xi_1\\ 0 &= \mathbb{E}[\xi_2^T Z_1 Z_2 \zeta_1] = \xi_2^T W \zeta_1\\ 0 &= \mathbb{E}[\zeta_2^T Z_2 Z_1^T \xi_1] = \zeta_2^T W \xi_1\\ 0 &= \mathbb{E}[\zeta_2^T Z_2 Z_2 \zeta_1] = \zeta_2^T \zeta_1.\end{aligned}$$

Finally, under the unit variance assumption, (P2) can be rewritten as

$$\begin{aligned}\max_{\xi_2\in\mathbb{R}^{p_1},\zeta_2\in\mathbb{R}^{p_2}} \quad & \lambda_2 = \xi_2^T W \zeta_2 \qquad \text{(P2*)}\\ \text{subject to} \quad & \xi_2^T \xi_2 = 1\\ & \zeta_2^T \zeta_2 = 1\\ & \xi_1^T \xi_2 = 0\\ & \zeta_1^T \zeta_2 = 0\\ & \xi_2^T W \zeta_1 = 0\\ & \zeta_2^T W \xi_1 = 0\end{aligned}$$

It is shown in (Anderson, 2003) that this problem has a solution as well. In fact the procedure of finding linear combinations can be done as long as the dimension permits it, that is, p_1 times. Assume that this is done and we obtain canonical correlations $1 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p_1} \geq 0$ and vectors for linear combinations $\xi_1, \ldots, \xi_{p_1}, \zeta_1, \ldots, \zeta_{p_1}$. If we summarize the λ_i's in a diagonal matrix Λ in decreasing order and the vectors in matrices, $\Xi = (\xi_1, \ldots, \xi_{p_1})$ and $\Psi = (\zeta_1, \ldots, \zeta_{p_1})$, the conditions of our optimization problems imply that Ξ and Ψ must be orthogonal and we have $\Xi^T W \Psi = \Lambda$, similar to the singular value composition of W. This raises the question: Can canonical correlations be obtained through a singular value composition of W?

The answer to that question is yes and is shown in a very long proof in section 12.2 of (Anderson, 2003). We summarize the results in the following theorem.

Theorem 5.2.1 (Canonical Correlation Analysis). *Let $Y_1 \sim \mathcal{N}_{p_1}(m_1, \Sigma_{11}), Y_2 \sim \mathcal{N}_{p_2}(m_2, \Sigma_{22})$ be given such that they are jointly normal*

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}_p\left(m := \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \Sigma := \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \qquad \text{(CCA)}$$

with Σ positive definite and $p_1 \leq p_2$. Let Z be the multivariate normal random vector with

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \Sigma_{11}^{-1/2}(Y_1 - m_1) \\ \Sigma_{22}^{-1/2}(Y_2 - m_2) \end{pmatrix} \sim \mathcal{N}_p\left(0, \begin{pmatrix} I_{p_1} & W \\ W^T & I_{p_2} \end{pmatrix}\right),$$

where $W = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} \in \mathbb{R}^{p_1 \times p_2}$. Further, let W have singular value composition

$$W = Q_1 \Lambda Q_2^T, \qquad \Lambda = \begin{pmatrix} \lambda_1^* & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^* & \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_{p_1}^* & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{p_1 \times p_2},$$

where $Q_1Q_1^T = I_{p_1}$ and $Q_2Q_2^T = I_{p_2}$. Then, λ_k^* is called the k-th canonical correlation of Y_1 and Y_2. Further, let $A := \Sigma_{11}^{-1/2}Q_1$ and $\Gamma := \Sigma_{22}^{-1/2}Q_2$ have columns $A = (\alpha_1^*|\cdots|\alpha_{p_1}^*)$ and $\Gamma = (\gamma_1^*|\cdots|\gamma_{p_2}^*)$, respectively. We call the triple (A, Γ, Λ) the CCA parameters of Y. In particular, for $k \in [p_1]$, α_k^* and γ_k^* solve the optimization problem

$$\begin{aligned} &\max_{\alpha_k \in \mathbb{R}^{p_1}, \gamma_k \in \mathbb{R}^{p_2}} \lambda_k = \mathbb{C}\mathrm{orr}[\alpha_k^T Y_1, \gamma_k^T Y_2] \\ \textit{subject to} \quad &(\alpha_k^T Y_1, \gamma_k^T Y_2) \perp\!\!\!\perp \left(\alpha_1^{*T} Y_1, \gamma_1^{*T} Y_2, \ldots, \alpha_{k-1}^{*T} Y_1, \gamma_{k-1}^{*T} Y_2\right) \end{aligned}$$

with optimal value $\lambda_k = \lambda_k^*$. Finally, we have for $k \in [p_2] \setminus [p_1]$

$$\gamma_k^{*T} Y_2 \perp\!\!\!\perp \left(\alpha_1^{*T} Y_1, \gamma_1^{*T} Y_2, \ldots, \alpha_{p_1}^{*T} Y_1, \gamma_{p_1}^{*T} Y_2, \gamma_{p_1+1}^{*T} Y_2, \ldots, \gamma_{k-1}^{*T} Y_2\right).$$

Let us provide an example for this theorem.

Example 5.2.2. Suppose Y_1 and Y_2 have distributions

$$Y_1 \sim \mathcal{N}_2\left(\begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}\right), \quad \text{and} \quad Y_2 \sim \mathcal{N}_2\left(\begin{pmatrix} -3 \\ 4 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

with joint distribution

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{2,1} \\ Y_{2,2} \end{pmatrix} \sim \mathcal{N}_4\left(\begin{pmatrix} 2 \\ -1 \\ -3 \\ 4 \end{pmatrix}, \begin{pmatrix} 5 & 4 & 1.75 & 0.625 \\ 4 & 5 & 1.25 & 0.875 \\ 1.75 & 1.25 & 4 & 0 \\ 0.625 & 0.875 & 0 & 1 \end{pmatrix}\right)$$

What are the CCA parameters of Y? The corresponding optimization problems are

$$\max_{\alpha_1 \in \mathbb{R}^2, \gamma_1 \in \mathbb{R}^2} \lambda_1 := \mathbb{C}\mathrm{orr}[\alpha_1^T Y_1, \gamma_1^T Y_2],$$

and

$$\begin{aligned} &\max_{\alpha_2 \in \mathbb{R}^2, \gamma_2 \in \mathbb{R}^2} \lambda_2 := \mathbb{C}\mathrm{orr}[\alpha_2^T Y_1, \gamma_2^T Y_2] \\ \text{subject to} \quad &(\alpha_2^T Y_1, \gamma_2^T Y_2) \perp\!\!\!\perp (\alpha_1^T Y_1, \gamma_1^T Y_2). \end{aligned}$$

We use the approach from Theorem 5.2.1. Now, one can compute

$$\Sigma_{11}^{-1/2} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}^{-1/2} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \quad \text{and} \quad \Sigma_{22}^{-1/2} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}^{-1/2} = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then, we have

$$W = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1.75 & 0.625 \\ 1.25 & 0.875 \end{pmatrix}\begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.375 & 0.125 \\ 0.125 & 0.375 \end{pmatrix}$$

Then, one can calculate the singular value decomposition

$$W = Q_1 \Lambda Q_2^T = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0.5 & 0 \\ 0 & 0.25 \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

Finally, one can compute

$$A = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \frac{1}{3\sqrt{2}} \begin{pmatrix} 1 & -3 \\ 1 & 3 \end{pmatrix}$$

and

$$\Gamma = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0.5 & -0.5 \\ 1 & 1 \end{pmatrix}.$$

The CCA parameters of Y are (A, Γ, Λ), but how can we interpret these results? Firstly, note that the correlation of the linear combinations is invariant under scaling with positive constants. Hence, linear combinations of Y_1 and Y_2 with maximum correlation are given by

$$Y_{1,1} + Y_{1,2} \qquad \text{and} \qquad Y_{2,1} + 2Y_{2,2}$$

with correlation 0.5 which is the first canonical correlation. The second pair of linear combinations explaining the dependency of the set is given by

$$-Y_{1,1} + Y_{1,2} \qquad \text{and} \qquad - Y_{2,1} + 2Y_{2,2}$$

with second canonical correlation 0.25. In summary, these two pairs explain the dependency of the random vectors Y_1 and Y_2 and we have canonical correlations of 0.5 and 0.25.

The next question is: How can the CCA parameters be estimated in practice, i.e. when given a sample?

5.2.3 Estimation of the CCA Parameters

In this subsection we provide the estimation method for the classical CCA model and prove asymptotic results. Let us begin with the notation which is used in the following.

Notation 5.2.3. Let $Y^{(1)}, Y^{(2)}, \ldots \overset{iid}{\sim} \mathcal{N}_p(m, \Sigma)$ be given as in (CCA). We denote the data matrix based on the first n samples by $\boldsymbol{Y}^{(n)}$. Thereby, $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$ are the rows of $\boldsymbol{Y}^{(n)}$ and we use the partitioned notation

$$\boldsymbol{Y}^{(n)} = \begin{pmatrix} \boldsymbol{Y}_1^{(n)} & \boldsymbol{Y}_2^{(n)} \end{pmatrix} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{pmatrix} \in \mathbb{R}^{n\times p}, \qquad \boldsymbol{Y}_k^{(n)} = \begin{pmatrix} Y_k^{(1)} \\ \vdots \\ Y_k^{(n)} \end{pmatrix} \in \mathbb{R}^{n\times p_k}.$$

Some times we will also suppress the index (n) in the exponent.

The determination of the CCA parameters is straightforward by plug-in estimation. Note that the canonical correlations are obtained by the singular value decomposition of the matrix $W = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$. Hence, we need to estimate blocks of the covariance matrix Σ. The following theorem provides a consistent method.

Theorem 5.2.4. *In the setting of Notation 5.2.3, define*

$$\tilde{\boldsymbol{Y}}_k^{(n)} = \boldsymbol{Y}_k^{(n)} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{Y}_k^{(n)}.$$

It holds for all $k, l = 1, 2$,

$$\frac{1}{n}\tilde{\boldsymbol{Y}}_k^{(n)T}\tilde{\boldsymbol{Y}}_l^{(n)} \xrightarrow{a.s.} \Sigma_{kl}.$$

Proof. We define the sample mean,

$$\overline{Y}_k^{(n)} = \frac{1}{n}\sum_{i=1}^{n} Y_k^{(i)} = \frac{1}{n}\mathbf{1}_n^T \boldsymbol{Y}_k^{(n)}.$$

Note that we enter the samples as row vectors, hence, the sample mean $\overline{Y}_k^{(n)} \in \mathbb{R}^{1\times p_k}$ and the true mean $m_k \in \mathbb{R}^{1\times p_k}$ are row vectors, as well. We have

$$\begin{aligned}
\frac{1}{n}\tilde{\boldsymbol{Y}}_k^{(n)^T}\tilde{\boldsymbol{Y}}_l^{(n)} &= \frac{1}{n}\left(\boldsymbol{Y}_k^{(n)} - \mathbf{1}_n\overline{Y}_k^{(n)}\right)^T\left(\boldsymbol{Y}_l^{(n)} - \mathbf{1}_n\overline{Y}_l^{(n)}\right)\\
&= \frac{1}{n}\left(\boldsymbol{Y}_k^{(n)} - \mathbf{1}_n m_k + \mathbf{1}_n m_k - \mathbf{1}_n\overline{Y}_k^{(n)}\right)^T\left(\boldsymbol{Y}_l^{(n)} - \mathbf{1}_n m_l + \mathbf{1}_n m_l - \mathbf{1}_n\overline{Y}_l^{(n)}\right)\\
&= \frac{1}{n}\left(\boldsymbol{Y}_k^{(n)} - \mathbf{1}_n m_k\right)^T\left(\boldsymbol{Y}_l^{(n)} - \mathbf{1}_n m_l\right) - \frac{1}{n}\left(\overline{Y}_k^{(n)} - m_k\right)^T\mathbf{1}_n^T\left(\boldsymbol{Y}_l^{(n)} - \mathbf{1}_n m_l\right)\\
-\frac{1}{n}&\left(\boldsymbol{Y}_k^{(n)} - \mathbf{1}_n m_k\right)^T\mathbf{1}_n\left(\overline{Y}_l^{(n)} - m_l\right) + \frac{1}{n}\left(\overline{Y}_k^{(n)} - m_k\right)^T\mathbf{1}_n^T\mathbf{1}_n\left(\overline{Y}_l^{(n)} - m_l\right)\\
&= \frac{1}{n}\left(\boldsymbol{Y}_k^{(n)} - \mathbf{1}_n m_k\right)^T\left(\boldsymbol{Y}_l^{(n)} - \mathbf{1}_n m_l\right) - \left(\overline{Y}_k^{(n)} - m_k\right)^T\left(\overline{Y}_l^{(n)} - m_l\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(Y_k^{(i)} - m_k\right)^T\left(Y_l^{(i)} - m_l\right) - \left(\overline{Y}_k^{(n)} - m_k\right)^T\left(\overline{Y}_l^{(n)} - m_l\right).
\end{aligned}$$

By the Strong Law of Large Numbers (Theorem 3.2.6) and the fact that the matrix product is continuous, the first and the second term converge to Σ_{kl} and 0, respectively. Hence, the assertion is proven. □

Now, we would like to plug the consistent estimators of Theorem 5.2.4 into the equations of Theorem 5.2.1, but this involves the calculation of a matrix inverse and square root. We shortly address their well-definedness in the following lemma and present the algorithm thereafter.

Lemma 5.2.5. *In the setting of Notation 5.2.3, we have that both,* $\tilde{\boldsymbol{Y}}_1^T\tilde{\boldsymbol{Y}}_1$ *and* $\tilde{\boldsymbol{Y}}_2^T\tilde{\boldsymbol{Y}}_2$ *are almost surely positive definite if* $n \geq p_2 + 1$.

Proof. For $k = 1, 2$, clearly $\tilde{\boldsymbol{Y}}_k^T\tilde{\boldsymbol{Y}}_k$ is positive semi-definite, so let us show that it is also invertible. By Complement 3.4.5 (c) in (Mardia et al., 1979), $\tilde{\boldsymbol{Y}}_k^T\tilde{\boldsymbol{Y}}_k \sim \mathcal{W}_{p_k}(n-1, \Sigma)$ follows a Wishart distribution which we introduce in Definition 6.2.2. By Proposition 6.2.3, $\tilde{\boldsymbol{Y}}_k^T\tilde{\boldsymbol{Y}}_k$ is invertible with probability 1 if $n \geq p_k + 1$. □

Algorithm 1 CCA Estimation

```
Input: Data Matrix $\boldsymbol{Y} = \boldsymbol{Y}^{(n)}$ with $n \geq p_2 + 1$
Output: CCA parameters $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{\Gamma}}$, and $\hat{\boldsymbol{\Lambda}}$.
Set $\tilde{\boldsymbol{Y}}_1 = \boldsymbol{Y}_1 - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{Y}_1$ and $\tilde{\boldsymbol{Y}}_2 = \boldsymbol{Y}_2 - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{Y}_2$, respectively.
Set $\hat{W} = \left(\tilde{\boldsymbol{Y}}_1^T\tilde{\boldsymbol{Y}}_1\right)^{-1/2}\tilde{\boldsymbol{Y}}_1^T\tilde{\boldsymbol{Y}}_2\left(\tilde{\boldsymbol{Y}}_2^T\tilde{\boldsymbol{Y}}_2\right)^{-1/2}$
Let $\hat{W}$ have singular value composition $\hat{W} = \hat{Q}_1\hat{\Lambda}\hat{Q}_2^T$.
Set $\hat{A} = \left(\tilde{\boldsymbol{Y}}_1^T\tilde{\boldsymbol{Y}}_1\right)^{-1/2}\hat{Q}_1$ and $\hat{\Gamma} = \left(\tilde{\boldsymbol{Y}}_2^T\tilde{\boldsymbol{Y}}_2\right)^{-1/2}\hat{Q}_2$, respectively.
```

Algorithm 1 is consistent for the estimation of the CCA parameters. We can show this with Theorem 5.2.4 and the following lemmas.

Lemma 5.2.6. *The matrix inverse and square root are continuous on the space of positive definite matrices.*

Proof. The entries of a matrix inverse can be written as a continuous function of the entries of the matrix and the determinant. Since the determinant is a polynomial of the matrix entries, the inverse is continuous. For a proof of the continuity of the matrix square root, we refer to (Wihler, 2009). □

Lemma 5.2.7 (Theorem 3 and Theorem 6 in (Nathanson and Ross, 2023))**.** *Let*

$$R : \mathbb{R}^{n+1} \setminus (\mathbb{R}^n \times \{0\}) \to \mathbb{C}^n$$

be the function that assigns each $(n+1)$-tuple $(a_0, \ldots, a_n)$ the roots of the polynomial $f(z) = \sum\limits_{i=0}^{n} a_i z^i$ in increasing order. Then, R is continuous, i.e. the roots of a polynomial of fixed degree n depend continuously on its coefficients.

Corollary 5.2.8 (Consistency of CCA)**.** *Let $Y^{(1)}, Y^{(2)}, \ldots \overset{iid}{\sim} \mathcal{N}_p(m, \Sigma)$ be as in* (CCA)*. Then, algorithm 1 is consistent for the estimation of W and the canonical correlations $\lambda_1, \ldots, \lambda_{p_1}$. That, is when $\hat{W}^{(n)}$ and $\hat{\Lambda}^{(n)}$ with diagonal entries $\lambda_1^{(n)}, \ldots, \lambda_{p_1}^{(n)}$ are obtained from algorithm 1 with the first n samples $Y^{(1)}, \ldots, Y^{(n)}$, we have*

$$\hat{W}^{(n)} \xrightarrow{a.s.} W$$

and for all $k \in [p_1]$,

$$\lambda_k^{(n)} \xrightarrow{a.s.} \lambda_k.$$

Proof. Combining Theorem 5.2.4 with Lemma 5.2.6 and the Continuous Mapping Theorem yields

$$\hat{W}^{(n)} = \left(\frac{1}{n}\tilde{\boldsymbol{Y}}_1^T\tilde{\boldsymbol{Y}}_1\right)^{-1/2}\left(\frac{1}{n}\tilde{\boldsymbol{Y}}_1^T\tilde{\boldsymbol{Y}}_2\right)\left(\frac{1}{n}\tilde{\boldsymbol{Y}}_2^T\tilde{\boldsymbol{Y}}_2\right)^{-1/2} \xrightarrow{a.s.} \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} = W$$

showing the consistency for W. Now, note that for $k \in [p_1]$, λ_k is the k-th biggest root of the polynomial

$$f(\lambda) = \det\left(\hat{W}^{(n)^T}\hat{W}^{(n)} - \lambda I_{p_2}\right).$$

By Lemma 5.2.7, the continuity of the determinant and the Continuous Mapping Theorem, we conclude

$$\lambda_k^{(n)} \xrightarrow{a.s.} \lambda_k.$$

□

Another interesting result about the classical CCA is the asymptotic normality of the squared canonical correlations.

Theorem 5.2.9 (Asymptotic Normality of CCA)**.** *Let $Y_1, Y_2, \ldots \overset{iid}{\sim} \mathcal{N}_p(m, \Sigma)$ as in* (CCA) *such that $\lambda_1 \neq 1$ and the nonzero canonical correlations are pairwise distinct. Then, we have for all $k \in [p_1]$ with $\lambda_k \neq 0$*

$$\sqrt{n}\frac{\left(\lambda_k^{(n)}\right)^2 - \lambda_k^2}{2\lambda_k(1-\lambda_k^2)} \xrightarrow{w} \mathcal{N}(0,1),$$

where $\lambda_k^{(n)}$ is the k'th canonical correlation obtained from Algorithm 1.

Proof. See Section 12.5 in (Anderson, 2003). □

This concludes our section about the classical Canonical Correlation Analysis. Next, we will consider models generalizing the classical CCA.

5.3 Gaussian Copula CCA

5.3.1 The Gaussian Copula CCA Model

Obviously, one drawback of the classical CCA is the normality assumption which may not be fulfilled in many scenarios, so that there have been proposed models relaxing this assumption. In this section, we introduce the Gaussian Copula model which allows the marginals of the data to follow an arbitrary absolutely continuous distribution. However, their dependency must be described by a Gaussian Copula. This model has also been presented in (Yoon et al., 2020). Further reading about semiparametric CCA models we recommend are (Agniel and Cai, 2017) and (Zoh et al., 2016).

We begin this section by defining the model setup, continue by discussing the implications of the model assumptions and explain how the Gaussian Copula model generalizes the classical CCA. Further, we present an estimation method of the Gaussian Copula CCA (GCCCA) parameters and prove its consistency.

Definition 5.3.1 (Gaussian Copula Model)**.** A random vector $Y = (Y_1, \ldots, Y_p)$ satisfies the Gaussian Copula Model if there exists an underlying multivariate normal random vector $Z = (Z_1, \ldots, Z_p) \sim \mathcal{N}_p(0, C)$ and increasing functions $f_1, \ldots, f_p : \mathbb{R} \to \mathbb{R}$ so that

$$f_k(Y_k) = Z_k \tag{GCCCA}$$

for all $k \in [p]$. Thereby, we assume C to be a correlation matrix which means being positive definite and having diagonal entries 1. We write $Y \sim \mathrm{NPN}_p(C, f)$, where $f = (f_1, \ldots, f_p)$. This is sometimes referred to as the *nonparanormal distribution.*

Remark 5.3.2. Note that the index notation in this section is slightly different to the last section. Previously, we had $Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ to denote the two partitions. In this section we write $Y = (Y_1, \ldots, Y_p)$ in order to emphasize the one-dimensional univariate marginals as we will mostly focus on them in this section.

Definition 5.3.3 (Gaussian Copula CCA)**.** Suppose $Y \sim \mathrm{NPN}_p(C, f)$ satisfies the Gaussian Copula Model with an underlying $Z \sim \mathcal{N}_p(0, C)$. If Y is partitioned into $(Y_1, \ldots, Y_{p_1})$ and $(Y_{p_1+1}, \ldots, Y_p)$ with $p_1 \leq p_2 := p - p_1$ then, we define the GCCCA parameters of Y as the CCA parameters of Z as in Theorem 5.2.1.

There are three aspects that we will discuss for the model formulation. We begin with the implications of the model assumptions.

What do the model assumptions imply? Another way this question can be formulated is: If $Y \sim \mathrm{NPN}_p(C, f)$, what can we deduce about the distribution of Y or its marginals? In fact, we show that $Y \sim \mathrm{NPN}_p(C, f)$ implies that the marginals Y_k can follow arbitrary continuous distributions.

Theorem 5.3.4. *Let $F_1, \dots, F_p$ be continuous distribution functions and let C be a correlation matrix. Then, there exists $Y \sim \mathrm{NPN}_p(C, f)$, so that $Y_k \sim F_k$ for all k.*

Proof. Let $Z \sim \mathcal{N}_p(0, C)$ be an underlying multivariate normal random vector which exists by Proposition 5.1.7. Let Φ be the distribution function of a $\mathcal{N}(0,1)$ random variable. We define

$$Y_k := F_k^{-1}(\Phi(Z_k)).$$

By the generalized inverse transform (Theorem 3.1.14), we have $\Phi(Z_k) \sim \mathrm{Unif}(0,1)$ and therefore, $Y_k = F_k^{-1}(\Phi(Z_k)) \sim \mu_k$. Further, since F_k is continuous, we have by Theorem 3.1.15 (ii), $\Phi^{-1}(F_k(Y_k)) = Z_k$. Finally, note that $f_k := \Phi^{-1} \circ F_k$ is increasing, so that $Y = (Y_1, \dots, Y_p) \sim \mathrm{NPN}_p(C, f)$ satisfies the Gaussian Copula Model. □

The converse is also true.

Theorem 5.3.5. *Let $Y = (Y_1, \dots, Y_p) \sim \mathrm{NPN}_p(C, f)$ follow the Gaussian Copula model. Then, for all k, the function f_k has the form $f_k = \Phi^{-1} \circ F_k$, where F_k is the continuous distribution function of Y_k. In particular, we have*

$$Y_k = F_k^{-1}(\Phi(Z_k)) \qquad a.s. \tag{5.1}$$

for all k.

Proof. Let $k \in [p]$. Note that since $f_k(Y_k) = Z_k \sim \mathcal{N}(0,1)$, f_k must be surjective. By assumption it is also increasing and hence, continuous. Now, let us define the continuous function $F_k := \Phi \circ f_k$. We show that F_k is the distribution function of Y_k. We have for all $t \in \mathbb{R}$

$$\mathbb{P}(Y_k \leq t) \leq \mathbb{P}(f_k(Y_k) \leq f_k(t)) = \Phi(f_k(t)) = F_k(t)$$
$$\mathbb{P}(Y_k \leq t) \geq \mathbb{P}(\{Y_k \leq t\} \cup \{Y_k > t, f_k(Y_k) = f_k(t)\}) \geq \mathbb{P}(f_k(Y_k) \leq f_k(t)) = \Phi(f_k(t)) = F_k(t).$$

Hence, Y_k has the continuous distribution function f_k. Finally, note that

$$f_k(Y_k) = Z_k \implies F_k(Y_k) = \Phi(Z_k) \implies Y_k = F_k^{-1}(\Phi(Z_k)) \qquad a.s.,$$

where we applied Theorem 3.1.15 (i) for the last implication. □

Hence, the Gaussian Copula CCA allows exactly all continuous distributions as univariate marginals. The next question we address is the parametrization of the Gaussian Copula.

Why this parametrization? Note that for $Y \sim \mathrm{NPN}_p(C, f)$, the multivariate normal Z is parametrized with mean 0 and a correlation matrix C instead of an arbitrary mean

$m \in \mathbb{R}^p$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. Working with a correlation matrix will turn out to be very useful, but for now let us address why this parametrization is not a restriction.

Theorem 5.3.6. *Suppose $Y = (Y_1, \dots, Y_p)$ is a multivariate normal random vector, so that there exist $Z = (Z_1, \dots, Z_p) \sim \mathcal{N}_p(m, \Sigma)$ with $\Sigma \in \mathbb{R}^{p \times p}$ an arbitrary positive definite covariance matrix and increasing functions $f_1, \dots, f_p$, so that*

$$f_k(Y_k) = Z_k$$

for all k. Then, Y satisfies the Gaussian Copula model. In particular, there exists a correlation matrix C and increasing functions $\tilde{f}_1, \dots, \tilde{f}_p$, so that $Y \sim \mathrm{NPN}_p(C, \tilde{f})$.

Proof. Since $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ is positive definite, all its diagonal entries are greater than 0. Define the matrix $C = (c_{ij}) \in \mathbb{R}^{p \times p}$ with $c_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$. Further, set $\tilde{f}_k := (f_k - m_k)/\sqrt{\sigma_{kk}}$ and define

$$\tilde{Z}_k := \tilde{f}_k(Y_k) = \frac{Z_k - m_k}{\sqrt{\sigma_{kk}}}.$$

Then, $\tilde{Z} = (\tilde{Z}_1, \dots, \tilde{Z}_p) \sim \mathcal{N}_p(0, C)$ and C is positive definite with diagonal entries 1 by Proposition 5.1.4. Hence, $Y \sim \mathrm{NPN}_p(C, \tilde{f})$. □

Which models are covered? As the Gaussian Copula relaxes the normality assumption of the CCA and allows arbitrary marginals, it should intuitively be larger than the CCA. In fact, this is true and we have already shown it.

Corollary 5.3.7. *Let Y be a multivariate random vector satisfying* (CCA)*. Then, Y satisfies the Gaussian Copula Model* (GCCCA)*.*

Proof. If Y satisfies (CCA), then this implies that Y fulfills the assumptions of Theorem 5.3.6, where we set $Z_k = Y_k$ and f_k to be the identity. Hence, Y satisfies the Gaussian Copula Model. □

This concludes our discussion of the Gaussian Copula CCA model. The next question we deal with is: How can we estimate the GCCCA parameters when observing Y?

5.3.2 Estimation of the GCCCA parameters

When solely observing $Y \sim \mathrm{NPN}_p(C, f)$, the estimation of the GCCCA parameters is nontrivial, as they are based on the correlation matrix C of the unobserved Z. If we knew C then, we could proceed as in Algorithm 1 and estimate a block of a covariance matrix and make the singular value decomposition. Thankfully, there exists a lemma which lets us make inference on C by using Spearman's ρ. In this section we will present this estimation method and prove its consistency. For now, let us begin with the notation and setting throughout the section.

Notation 5.3.8. Let $Y^{(1)}, Y^{(2)}, \ldots \overset{iid}{\sim} \mathrm{NPN}_p(C, f)$ be observed, so that $Z^{(1)}, Z^{(2)}, \ldots \overset{iid}{\sim} \mathcal{N}_p(0, C)$ are normal with a correlation matrix C and increasing functions $f_1, \ldots, f_p$ which satisfy

$$f_k(Y_k^{(n)}) = Z_k^{(n)}$$

for all $k \in [p]$ and all $n \in \mathbb{N}$. We denote the data matrix based on the first n samples by $\boldsymbol{Y}^{(n)}$. Thereby, $Y^{(1)}, \ldots, Y^{(n)}$ enter the matrix as row vectors:

$$\boldsymbol{Y}^{(n)} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{pmatrix} = \begin{pmatrix} \gamma_1^{(n)} & \gamma_2^{(n)} & \cdots & \gamma_p^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

The vectors $\gamma_j^{(n)} = (Y_j^{(1)}, \ldots, Y_j^{(n)})$ denote the columns of $\boldsymbol{Y}^{(n)}$. As we will compute the Spearman's correlation coefficients, having a notation for the columns will be helpful.

Next, let us define Spearman's ρ.

Definition 5.3.9 (Spearman's ρ)**.** Let X and Y be random variables with distribution functions F_X and F_Y. Then, their *Spearman's correlation coefficient* $\rho_{X,Y}$ is defined as

$$\rho_{X,Y} = \mathbb{C}\mathrm{orr}[F_X(X), F_Y(Y)].$$

The estimation method relies on the following relationship of the Sperman's correlation and the entries of the correlation matrix:

Lemma 5.3.10 (Lemma 3.1 in (Liu et al., 2012))**.** *Let $Y = (Y_1, \ldots, Y_d) \sim \mathrm{NPN}_d(C, f)$ with $C = (c_{ij}) \in \mathbb{R}^{p \times p}$ and let $\rho_{i,j}$ be the Spearman's correlation coefficient of Y_i and Y_j. That is,*

$$\rho_{i,j} = \mathbb{C}\mathrm{orr}[F_i(Y_i), F_j(Y_j)],$$

where F_i and F_j are the distribution functions of Y_i and Y_j, respectively. Then, the following relationship between the Spearman's correlation coefficient and the correlation matrix holds:

$$c_{ij} = 2 \sin\left(\frac{\pi}{6} \rho_{i,j}\right) \tag{5.2}$$

Hence, we can estimate the entries of C with Spearman's ρ without having to know the increasing functions $f_1, \ldots, f_p$ or the latent $Z^{(i)}$'s which is very convenient. Before we can present the algorithm we need to define the empirical version of Spearman's ρ.

Definition 5.3.11 (Empirical Spearman's ρ)**.** Let $X_1^{(1)}, \ldots, X_1^{(n)} \overset{iid}{\sim} \mu_1$ and $X_2^{(1)}, \ldots, X_2^{(n)} \overset{iid}{\sim} \mu_2$ be two samples in $\mathbb{R}$ with empirical distribution functions $F_1^{(n)}$ and $F_2^{(n)}$, respectively. We use the notation

$$\overline{F}_1^{(n)} := \frac{1}{n} \sum_{i=1}^{n} F_1^{(n)}(X_1^{(i)}) \qquad \text{and} \qquad \overline{F}_2^{(n)} := \frac{1}{n} \sum_{i=1}^{n} F_2^{(n)}(X_2^{(i)}).$$

The empirical Spearman's ρ of the samples is defined as

$$\hat{\rho}_{X_1,X_2}^{(n)} := \frac{\sum_{i=1}^{n} \left(F_1^{(n)}(X_1^{(i)}) - \overline{F}_1^{(n)}\right)\left(F_2^{(n)}(X_2^{(i)}) - \overline{F}_2^{(n)}\right)}{\sqrt{\sum_{i=1}^{n} \left(F_1^{(n)}(X_1^{(i)}) - \overline{F}_1^{(n)}\right)^2 \cdot \sum_{i=1}^{n} \left(F_2^{(n)}(X_2^{(i)}) - \overline{F}_2^{(n)}\right)^2}}$$

Now, we can plug the empirical Spearman's ρ of the marginals into (5.2) and estimate the correlation matrix C. Then, we can apply the same steps as in classical CCA (Theorem 5.2.1). This yields the following straightforward algorithm:

Algorithm 2 GCCCA Estimation

Input: Data Matrix $\boldsymbol{Y} = (\gamma_1 | \dots | \gamma_p)$

Output: GCCCA Parameters $\hat{W}$ and $\hat{\Lambda}$

```
for $i = 1, \dots, p$ do
    for $j = i, \dots, p$ do
        Determine the Spearman's correlation coefficient $\hat{\rho}_{i,j}$ of $\gamma_i$ and $\gamma_j$.
        Set $\hat{c}_{ij} = \hat{c}_{ji} = 2 \sin\left(\frac{\pi}{6}\hat{\rho}_{i,j}\right)$.
Let $\hat{C}$ have block structure $\hat{C} = \begin{pmatrix} \hat{C}_{11} & \hat{C}_{12} \\ \hat{C}_{21} & \hat{C}_{22} \end{pmatrix}$.
Set $\hat{W} = \hat{C}_{11}^{-1/2} \hat{C}_{12} \hat{C}_{22}^{-1/2}$.
Compute $\hat{\Lambda}$ with the singular value decomposition of $\hat{W} = \hat{Q}_1 \hat{\Lambda} \hat{Q}_2^T$.
```

We can show that Algorithm 2 is consistent for the estimation of the GCCCA parameters. To do this we first show the consistency of the empirical Spearman's ρ.

Theorem 5.3.12. *Let $X_1^{(1)}, X_1^{(2)}, \dots \overset{iid}{\sim} F_1$ and $X_2^{(1)}, X_2^{(2)}, \dots \overset{iid}{\sim} F_2$ be i.i.d samples in $\mathbb{R}$. Let $\hat{\rho}^{(n)}_{X_1,X_2}$ be the empirical Spearman's correlation coefficient of the first n samples, i.e.*

$$\hat{\rho}^{(n)}_{X_1,X_2} = \frac{\sum\limits_{i=1}^{n} \left(F_1^{(n)}(X_1^{(i)}) - \overline{F}_1^{(n)}\right)\left(F_2^{(n)}(X_2^{(i)}) - \overline{F}_2^{(n)}\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(F_1^{(n)}(X_1^{(i)}) - \overline{F}_1^{(n)}\right)^2 \cdot \sum\limits_{i=1}^{n} \left(F_2^{(n)}(X_2^{(i)}) - \overline{F}_2^{(n)}\right)^2}},$$

where $F_1^{(n)}$ and $F_2^{(n)}$ are the empirical distribution functions. Further, let

$$\rho_{X_1,X_2} = \mathbb{C}\mathrm{orr}[F_1(X_1^{(1)}), F_2(X_2^{(1)})]$$

denote the true Spearman's correlation of the two distributions. Then, $\hat{\rho}^{(n)}_{X_1,X_2}$ is consistent for ρ_{X_1,X_2}, i.e. we have

$$\hat{\rho}^{(n)}_{X_1,X_2} \xrightarrow{a.s.} \rho_{X_1,X_2}.$$

Note that this is a general result, not specific to the Gaussian Copula. I.e. Spearman's ρ is always consistent for i.i.d. samples.

Proof. We show the convergence for the numerator and the two factors of the denominator separately. For indices $k, l = 1, 2$, define

$$m_k = \mathbb{E}[F_k(X_k^{(1)})] \qquad \overline{F}_k = \frac{1}{n}\sum_{i=1}^{n} F_k(X_k^{(i)})$$

and consider the statistic

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(F_k(X_k^{(i)})-\overline{F}_k\right)\left(F_l(X_l^{(i)})-\overline{F}_l\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\left(F_k(X_k^{(i)})-m_k\right)\left(F_l(X_l^{(i)})-m_l\right)+\left(\overline{F}_k-m_k\right)\left(\overline{F}_l-m_l\right).
\end{aligned}$$

By the Strong Law of Large Numbers, the first term and second term converge almost surely to $\mathbb{C}\mathrm{ov}[F_k(X_k^{(1)}),F_l(X_l^{(1)})]$ and 0, respectively. Further, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left(F_k^{(n)}(X_k^{(i)})-\overline{F}_k^{(n)}\right)\left(F_l^{(n)}(X_l^{(i)})-\overline{F}_l^{(n)}\right)-\left(F_k(X_k^{(i)})-\overline{F}_k\right)\left(F_l(X_l^{(i)})-\overline{F}_l\right)\right|\\
=&\left|\frac{1}{n}\sum_{i=1}^{n}\left(F_k^{(n)}(X_k^{(i)})F_l^{(n)}(X_l^{(i)})-\overline{F}_k^{(n)}\overline{F}_l^{(n)}\right)-\left(F_k(X_k^{(i)})F_l(X_l^{(i)})-\overline{F}_k\overline{F}_l\right)\right|\\
\leq&\left|\frac{1}{n}\sum_{i=1}^{n}F_k^{(n)}(X_k^{(i)})F_l^{(n)}(X_l^{(i)})-F_k(X_k^{(i)})F_l(X_l^{(i)})\right|+\left|\overline{F}_k^{(n)}\overline{F}_l^{(n)}-\overline{F}_k\overline{F}_l\right|\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\left|F_k^{(n)}(X_k^{(i)})F_l^{(n)}(X_l^{(i)})-F_k^{(n)}(X_k^{(i)})F_l(X_l^{(i)})\right|+\\
&+\frac{1}{n}\sum_{i=1}^{n}\left|F_k^{(n)}(X_k^{(i)})F_l(X_l^{(i)})-F_k(X_k^{(i)})F_l(X_l^{(i)})\right|+\left|\overline{F}_k^{(n)}\overline{F}_l^{(n)}-\overline{F}_k^{(n)}\overline{F}_l\right|+\left|\overline{F}_k^{(n)}\overline{F}_l-\overline{F}_k\overline{F}_l\right|\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\left|F_k^{(n)}(X_k^{(i)})\right|\left|F_l^{(n)}(X_l^{(i)})-F_l(X_l^{(i)})\right|+\frac{1}{n}\sum_{i=1}^{n}\left|F_l(X_l^{(i)})\right|\left|F_k^{(n)}(X_k^{(i)})-F_k(X_k^{(i)})\right|\\
&+\left|\overline{F}_k^{(n)}\right|\left|\frac{1}{n}\sum_{i=1}^{n}F_l^{(n)}(X_l^{(i)})-F_l(X_l^{(i)})\right|+\left|\overline{F}_l\right|\left|\frac{1}{n}\sum_{i=1}^{n}F_k^{(n)}(X_k^{(i)})-F_k(X_k^{(i)})\right|\\
\leq&2\cdot\left\|F_k^{(n)}-F_k\right\|_\infty+2\cdot\left\|F_l^{(n)}-F_l\right\|_\infty\xrightarrow{a.s.}0,
\end{aligned}$$

where we applied the Glivenko-Cantelli-Theorem (Theorem 3.2.7) in the last step. Combining these two statements yields that

$$\frac{1}{n}\sum_{i=1}^{n}\left(F_k^{(n)}(X_k^{(i)})-\overline{F}_k^{(n)}\right)\left(F_l^{(n)}(X_l^{(i)})-\overline{F}_l^{(n)}\right)\xrightarrow{a.s.}\mathbb{C}\mathrm{ov}[F_k(X_k^{(1)}),F_l(X_l^{(1)})].$$

Since the numerator and the factors of the denominator converge as desired, we have by the Continuous Mapping Theorem

$$\hat{\rho}_{X_1,X_2}^{(n)}\xrightarrow{a.s.}\frac{\mathbb{C}\mathrm{ov}[F_1(X_1^{(1)}),F_2(X_2^{(1)})]}{\sqrt{\mathbb{V}\mathrm{ar}[F_1(X_1^{(1)})]\mathbb{V}\mathrm{ar}[F_2(X_2^{(1)})]}}=\mathbb{C}\mathrm{orr}[F_1(X_1^{(1)}),F_2(X_2^{(1)})]=\rho_{X_1,X_2}$$

and the assertion is proven. □

Finally, we have that Algorithm 2 is consistent for the estimation of the GCCCA parameters.

Corollary 5.3.13. *Let* $Y^{(1)}, Y^{(2)}, \ldots, \overset{iid}{\sim} \mathrm{NPN}_p(C, f)$ *be as in* (GCCCA)*. Then, Algorithm 2 is consistent for the estimation of* W *and* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{p_1})$*. That is, when* $\hat{W}^{(n)}$ *and* $\hat{\Lambda}^{(n)} = \mathrm{diag}(\lambda_1^{(n)}, \ldots, \lambda_{p_1}^{(n)})$ *are obtained from Algorithm 2 based on the first n samples, we have*

$$\hat{W}^{(n)} \xrightarrow{a.s.} W$$

and for all k

$$\lambda_k^{(n)} \xrightarrow{a.s.} \lambda_k.$$

Proof. By Theorem 5.3.12, we have that

$$\hat{\rho}_{i,j}^{(n)} \xrightarrow{a.s.} \rho_{i,j},$$

where $\hat{\rho}_{i,j}^{(n)}$ is the Spearman's correlation of $\gamma_i^{(n)}$ and $\gamma_j^{(n)}$ as in line 3 of Algorithm 2. Hence, we have by (5.2) and the Continuous Mapping Theorem

$$\hat{c}_{ij}^{(n)} \xrightarrow{a.s.} c_{ij}.$$

Now that one has the consistency for the correlation matrix C, one can proceed with the same steps as in the proof of the consistency of classical CCA (Theorem 5.2.8) and show the consistency for W and Λ. □

This finalizes our discussion of the Gaussian Copula CCA model. We continue with the cyclically monotone CCA.

5.4 Cyclically Monotone CCA

In this section we introduce our model for Canonical Correlation Analysis based on optimal transport. We begin with some intuition for the model, define it, discuss the assumptions and the parametrization, explain how it generalizes the previous methods and then, suggest an estimation method. The proof of its consistency is long and presented in Section 5.5. For further reading about this model we refer to (Bryan et al., 2024).

5.4.1 The Cyclically Monotone CCA Model

The motivation behind the cyclically monotone CCA is as follows. Suppose one observes a partitioned random vector $Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ and aims to examine the dependency of Y_1 and Y_2 with Canonical Correlation Analysis. Then, one could use the classical CCA to do so, but what if the normality assumption of Y is not appropriate? One could also try the Gaussian Copula CCA, but this assumes that the univariate marginals of the data are increasing functions of a multivariate normal. Hence, the GCCCA does not consider joint transformations of multiple margins which may be the case. How can we include more complicated transformations in our model?

At this point our main optimal transport result, Proposition 4.1.2, can help: For any two absolutely continuous distributions μ and ν, there exist unique transport maps S and T, so that $S\#\mu = \nu$ and $T\#\nu = \mu$. Hence, if we observe $Y_1 \sim \mu_1$ and $Y_2 \sim \mu_2$ with for now absolutely continuous distributions, we know that there exist $Z_1 \sim \mathcal{N}_{p_1}(0, I_{p_1})$ and $Z_2 \sim \mathcal{N}_{p_2}(0, I_{p_2})$, so that $G_1(Z_1) = Y_1$ and $G_2(Z_2) = Y_2$ for cyclically monotone functions G_1 and G_2. Equivalently formulated, we can transport the observed Y_1 and Y_2 to the multivariate normal Z_1 and Z_2 and then apply the classical CCA. The only assumption we need is that Z_1 and Z_2 are jointly normal.

We summarize this idea in the following definition.

Definition 5.4.1 (Cyclically Monotone CCA)**.** Let $Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ be a partitioned random vector, where Y_1 and Y_2 take values in $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively, with $p_1 \leq p_2$. Then, Y satisfies the *Cyclically Monotone CCA model* if there exists a multivariate normal random vector $Z = \begin{pmatrix} Z_1 & Z_2 \end{pmatrix}$ and cyclically monotone functions G_1 and G_2, so that

$$\begin{aligned} &Z_1 \sim \mathcal{N}_{p_1}(0, I_{p_1}) \\ &Z_2 \sim \mathcal{N}_{p_2}(0, I_{p_2}) \\ &Y_1 = G_1(Z_1) \\ &Y_2 = G_2(Z_2) \\ &Z = \begin{pmatrix} Z_1 & Z_2 \end{pmatrix} \sim \mathcal{N}_p \left(0, C := \begin{pmatrix} I_{p_1} & W \\ W^T & I_{p_2} \end{pmatrix} \right). \end{aligned} \tag{CMCCA}$$

Thereby, we require C to be positive definite. Then, the CMCCA parameters of Y are defined as the CCA parameters of Z.

Remark 5.4.2. Note that we again change our index notation from the previous section. This time we primarily work with the joint marginals, so that we will write $Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ instead of emphasizing the univariate marginals with $Y = (Y_1, \ldots, Y_p)$ as in Section 5.3.

The definition of the cyclically monotone CCA model raises multiple questions which we examine next. We begin with discussing the model assumptions.

What do the model assumptions imply? To answer this question we present the following Proposition first.

Proposition 5.4.3 (Proposition 2.1 in (Bryan et al., 2024))**.** *Let $\mu \in \mathcal{P}^d$ and let $\nu = \mathcal{N}_d(0, I_d)$. Then, there exists a unique cyclically monotone function G, so that $G\#\nu = \mu$.*

This Proposition is an extension to Proposition 4.1.2. So far we knew of the existence of unique pairs of invertible transport maps for two absolutely continuous distributions. Here, one distribution can be arbitrary and we still have the existence of a unique optimal transport map in one direction. However, it does not have to be invertible. We conclude:

Corollary 5.4.4. *Let $\mu_1 \in \mathcal{P}^{p_1}$ and $\mu_2 \in \mathcal{P}^{p_2}$ be two arbitrary distributions. Then, there exists $Y = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix}$ satisfying the cyclically monotone CCA model with $Y_1 \sim \mu_1$ and $Y_2 \sim \mu_2$. Equivalently, the two joint marginals of the cyclically monotone CCA model can have arbitrary distributions.*

This shows that the cyclically monotone CCA allows more flexibility in the distributions of the marginals than the Gaussian Copula model. There, the univariate marginals could follow continuous distributions, but their dependency must have been described by a Gaussian Copula. Here, we can have arbitrary joint distributions as marginals which makes many more scenarios admissible. This particularly includes cases where the univariate marginals have some complicated joint distribution which can not be described by a Gaussian Copula.

Next, we discuss the parametrization of the CMCCA model.

Why this parametrization? We observe that Z is parametrized with mean 0 and a specific block correlation matrix C. This assumption is more convenient for estimation as will compute square roots of inverses of an empirical covariance matrix later. These matrices will then be "closer" to the identity matrix. Further, this parametrization is not a restriction which we show next.

Theorem 5.4.5. *The parametrization of* (CMCCA) *with 0 mean and block correlation matrix C is not a restriction. That is, if $Y' = \begin{pmatrix} Y_1' & Y_2' \end{pmatrix}$ satisfies the more general CMCCA model assumptions with*

$$\begin{aligned}
&Z_1' \sim \mathcal{N}_{p_1}(m_1, \Sigma_{11})\\
&Z_2' \sim \mathcal{N}_{p_2}(m_2, \Sigma_{22})\\
&Y_1' = H_1(Z_1')\\
&Y_2' = H_2(Z_2')\\
&Z' = \begin{pmatrix} Z_1' & Z_2' \end{pmatrix} \sim \mathcal{N}_p\left(m := \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \Sigma := \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right),
\end{aligned}$$

where Σ is positive definite, then Y' satisfies (CMCCA)*.*

Proof. We denote the distributions of Y_1' and Y_2' by μ_1 and μ_2. Now, let us define the random vectors $Z_1 := \Sigma_{11}^{-1/2}(Z_1' - m_1)$ and $Z_2 := \Sigma_{22}^{-1/2}(Z_2' - m_2)$. Then, $Z = \begin{pmatrix} Z_1 & Z_2 \end{pmatrix} \sim \mathcal{N}_p(0, C)$ is jointly normal, where C satisfies the block structure from (CMCCA). Now, by Proposition 5.4.3, there exist unique cyclically monotone functions G_1 and G_2, so that $Y_1 := G_1(Z_1) \sim \mu_1$ and $Y_2 := G_2(Z_2) \sim \mu_2$ are equal in distribution to Y_1' and Y_2', respectively. We can now assume that we observed Y as its CMCCA parameters are equal to the CMCCA parameters of Y'. This is, because the CCA parameters of Z and Z' are the same. □

Finally, we discuss how the CMCCA generalizes the two previous models.

Which models are covered? The cyclically monotone CCA model has very strong implications as we argued before. Intuitively, it should contain the Gaussian Copula CCA and therefore, the classical CCA. This is in fact the case. We capture this result in the following theorem.

Theorem 5.4.6. *If $Y \sim \mathrm{NPN}_p(C, f)$ follows the Gaussian Copula Model, then it satisfies the cyclically monotone CCA assumptions. In summary, the cyclially monotone CCA contains the Gaussian Copula CCA which contains the classical CCA.*

Proof. The fact that the GCCCA contains the classical CCA was already shown in Corollary 5.3.7. We show that the CMCCA generalizes the GCCCA. Let $Y = (Y_1, \ldots, Y_p) \sim \mathrm{NPN}_p(C, f)$ follow the GCCCA model. As an exception to the notation in the rest of the section, the index of Y_k represents the univariate marginals. Further, let F_k be the distribution function of Y_k. By assumption, there exists $Z = (Z_1, \dots, Z_p) \sim \mathcal{N}_p(0, C)$, so that

$$Y_k = F_k^{-1}(\Phi(Z_k)),$$

where we use the representation from (5.1). Now, let us the define the functions

$$\begin{aligned} G_1 : \mathbb{R}^{p_1} \to \mathbb{R}^{p_1}, G_1(z_1, \dots, z_{p_1}) &= (F_1^{-1}(\Phi(z_1)), \dots, F_{p_1}^{-1}(\Phi(z_{p_1}))) \\ G_2 : \mathbb{R}^{p_2} \to \mathbb{R}^{p_2}, G_2(z_{p_1+1}, \dots, z_p) &= (F_{p_1+1}^{-1}(\Phi(z_{p_1+1})), \dots, F_p^{-1}(\Phi(z_p))). \end{aligned}$$

These functions are cyclically monotone by Example 4.1.1 (ii). Further, we have

$$(Y_1, \dots, Y_{p_1}) = G_1(Z_1, \dots, Z_{p_1}), \qquad \text{and} \qquad (Y_{p_1+1}, \dots, Y_p) = G_1(Z_{p_1+1}, \dots, Z_p).$$

Hence, Y satisfies the more general CMCCA assumptions, where the correlation matrix does not necessarily satisfy the block structure from (CMCCA). However, we know that by Theorem 5.4.5 that this parametrization is not a restriction, so that Y satisfies (CMCCA). □

This finalizes our theoretical discussion of the cyclically monotone CCA. The next question we address is: How can the CMCCA parameters be estimated in practice?

5.4.2 Estimation of the CMCCA parameters

In this section we present a plug-in estimation method in the case that the joint marginals Y_1 and Y_2 are absolutely continuous. Before we discuss its intuition, let us address the notation and setup in this section.

Notation 5.4.7. Suppose we observe $Y^{(1)}, \dots, Y^{(n)}$ i.i.d. satisfying the cyclically monotone CCA model. We summarize these in the data matrix $\boldsymbol{Y}^{(n)}$, where we enter the samples as row vectors and use the partitioned notation:

$$\boldsymbol{Y}^{(n)} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{pmatrix} = \begin{pmatrix} Y_1^{(1)} & Y_2^{(1)} \\ \vdots & \vdots \\ Y_1^{(n)} & Y_2^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

Further, denote the distributions of $Y_1^{(1)}$ and $Y_2^{(1)}$ by μ_1 and μ_2, respectively. As mentioned before, we **additionally require** that μ_1 and μ_2 are **absolutely continuous** throughout the remainder of the chapter. By the CMCCA model assumptions there exist $Z^{(1)}, \dots, Z^{(n)} \overset{iid}{\sim} \mathcal{N}_p(0, C)$ and cyclically monotone functions G_1 and G_2, so that

$$\begin{pmatrix} Y_1^{(1)} & Y_2^{(1)} \\ \vdots & \vdots \\ Y_1^{(n)} & Y_2^{(n)} \end{pmatrix} = \begin{pmatrix} G_1(Z_1^{(1)}) & G_2(Z_2^{(1)}) \\ \vdots & \vdots \\ G_1(Z_1^{(n)}) & G_2(Z_2^{(n)}) \end{pmatrix}.$$

Further, we have that

$$C = \begin{pmatrix} I_{p_1} & W \\ W^T & I_{p_2} \end{pmatrix}$$

is positive definite. We also use the notations $\nu_1 := \mathcal{N}_{p_1}(0, I_{p_1})$ and $\nu_2 := \mathcal{N}_p(0, I_{p_2})$ in the following.

Now, our goal is to determine the CCA parameters which eventually means inferring W. Then, we can make the singular value decomposition of W which yields the canonical correlations. The problem is that we only observe $Y^{(1)}, \dots, Y^{(n)}$ and the normal $Z^{(1)}, \dots, Z^{(n)}$, and the cyclically monotone transformations G_1 and G_2 are unknown. If we knew the $Z^{(i)}$'s, then we could easily estimate the block of its covariance matrix with the estimator from Section 5.2.

Let us try to estimate the two parts $Z_1^{(i)}$ and $Z_2^{(i)}$, separately. The information about the $Z_j^{(i)}$'s that we have is that they have been optimally transported to the $Y_j^{(i)}$'s with a cyclically monotone function. Now, we would like to "transport the $Y_j^{(i)}$'s back" to the $Z_j^{(i)}$'s. We learned in chapter 4 that the $Z_j^{(i)}$'s are rank statistics with respect to the distribution μ_j, i.e.

$$Z_j^{(i)} = R_{\nu_j}^{\mu_j}(Y_j^{(i)})$$

for all $i = 1, \dots, n$ and $j = 1, 2$. Hence, we can estimate them with the empirical ranks by defining a grid of ν_j. We use the canonical choice for a grid from Lemma 4.1.4 and sample

$$X_j^{(1)}, \dots, X_j^{(n)} \overset{iid}{\sim} \mathcal{N}_{p_j}(0, I_{p_j}).$$

Now, we have grid points where we can transport the $Y_j^{(1)}, \dots, Y_j^{(n)}$ to. In chapter 4, we learned that we do this by solving an optimal assignment problem. Thus, we need to find a permutation σ^* minimizing

$$\sum_{i=1}^{n} \left\| Y_j^{(i)} - X_{\sigma(j)}^{(i)} \right\|^2.$$

Then, we can define our estimators by the empirical rank statistics with respect to this grid by

$$\hat{Z}_j^{(i)} := \hat{R}_{\nu_j}^{n}(Y_j^{(i)}) = X_{\sigma^*(j)}^{(i)}.$$

Now that we have the estimators for the $Z^{(i)}$'s, the remaining steps are straightforward as we can apply the classical CCA to estimate W. We summarize our estimators in the matrix

$$\hat{\boldsymbol{Z}}_j := \begin{pmatrix} \hat{Z}_j^{(1)} \\ \vdots \\ \hat{Z}_j^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times p_j}$$

and define

$$\hat{\hat{\boldsymbol{Z}}}_j = \hat{\boldsymbol{Z}}_j - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\hat{\boldsymbol{Z}}_j.$$

Then, we can define $\hat{W}$ by

$$\hat{W} = \left(\hat{\hat{\boldsymbol{Z}}}_1^T \hat{\hat{\boldsymbol{Z}}}_1\right)^{-1/2} \hat{\hat{\boldsymbol{Z}}}_1^T \hat{\hat{\boldsymbol{Z}}}_2 \left(\hat{\hat{\boldsymbol{Z}}}_2^T \hat{\hat{\boldsymbol{Z}}}_2\right)^{-1/2}$$

and calculate the singular value decomposition. Note that these expressions are well-defined for $n \geq p_2 + 1$, which follows from applying Lemma 5.2.5. The whole procedure is summarized in the following algorithm which uses the preceding notation.

Algorithm 3 Cyclically Monotone CCA Estimation

```
Input: Data Matrix $\boldsymbol{Y}^{(n)}$ with $n \geq p_2 + 1$
Output: CMCCA Parameters $\hat{W}$ and $\hat{\Lambda}$
for $j = 1, 2$ do
    Sample grid points $X_j^{(1)}, \dots, X_j^{(n)} \overset{iid}{\sim} \mathcal{N}_{p_j}(0, I_{p_j})$.
    Define $\hat{Z}_j^{(i)} = \hat{R}_{\nu_j}^n(Y_j^{(i)})$ as the empirical rank w.r.t. the grid for all $i = 1, \dots, n$.
Set $\hat{W} = \left(\hat{\tilde{\boldsymbol{Z}}}_1^T \hat{\tilde{\boldsymbol{Z}}}_1\right)^{-1/2} \hat{\tilde{\boldsymbol{Z}}}_1^T \hat{\tilde{\boldsymbol{Z}}}_2 \left(\hat{\tilde{\boldsymbol{Z}}}_2^T \hat{\tilde{\boldsymbol{Z}}}_2\right)^{-1/2}$.
Compute the singular value decomposition $\hat{W} = \hat{Q}_1 \hat{\Lambda} \hat{Q}_2^T$.
```

This concludes our presentation of the CMCCA estimation method. We will prove its consistency in the next section.

5.5 Consistency of CMCCA

In order to prove the consistency of our method we need some more theorems on optimal transport. The central result we require is a convergence theorem of empirical rank statistics to true rank statistics. In order to present it we will begin the section with some lemmas which help us with the proof of this theorem. Thereafter, we state the general rank statistic convergence theorem and prove it. At that point we will have all the results we need in order to show the consistency of our method. However, there will be an estimator for every sample size n which makes the notation more difficult. Thus, we will precisely define the notation and setup and finally, prove the consistency.

Recall that for distributions $\mu \in \mathcal{P}^{d_1}$ and $\nu \in \mathcal{P}^{d_2}$, the set $\Gamma(\mu, \nu)$ is defined as the set of all probability measures in $\mathcal{P}^{d_1+d_2}$ with marginals μ and ν on the first d_1 and last d_2, respectively. The following lemmas are tailored to the steps in the proof of the convergence theorem.

Lemma 5.5.1 (Theorem 6 and Corollary 14 in (McCann, 1995))**.** *Let $\mu, \nu \in \mathcal{P}_{ac}^d$. Then, there exists a unique $\gamma \in \Gamma(\mu, \nu)$ with cyclically monotone support. This means that there exists a cyclically monotone set $S \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ such that $\gamma(S) = 1$.*

Lemma 5.5.2 (Lemma 9 in (McCann, 1995))**.** *Let $(\gamma_n)_{n \in \mathbb{N}}$ be a sequence of probability measures in $\mathcal{P}^{2d}$ such that*[4]

$$\gamma_n \xrightarrow{w} \gamma,$$

where $\gamma \in \mathcal{P}^{2d}$. Then,

(i) If γ_n has cyclically monotone support for all $n \in \mathbb{N}$, then so does γ.

(ii) If the left and right marginals of $\gamma_n \in \Gamma(\mu_n, \nu_n)$ converge weakly to limits $\mu, \nu \in \mathcal{P}^d$, respectively, then, $\gamma \in \Gamma(\mu, \nu)$.

[4] This version of the lemma is a bit weaker as the version in (McCann, 1995). We require weak convergence to a limit in $\mathcal{P}^{2d}$, while the original version only requires weak*-convergence as in functional analysis (which effectively means that (3.3) needs to hold for a smaller set of test functions f). Also, the limit may lie outside of $\mathcal{P}^{2d}$ in the original version.

With Lemma 5.5.1 and Lemma 5.5.2, we can show the following lemma.

Lemma 5.5.3. *Let $\mu, \nu \in \mathcal{P}_{ac}^d$ be absolutely continuous. Further, let $Y_1, Y_2, \ldots \overset{iid}{\sim} \mu$ be a sample and let $(X_k^{(n)})_{n \geq k \geq 1}$ be a grid of ν. Let $\hat{R}_\nu^n(Y_1), \ldots, \hat{R}_\nu^n(Y_n)$ be the empirical ranks of $Y_1, \ldots, Y_n$ with respect to the random grid points $X_1^{(n)}, \ldots, X_n^{(n)}$. Additionally, define the distribution $\gamma :\sim (Y_1, R_\nu^\mu(Y_1))$. Then, we have*

$$\frac{1}{n}\sum_{i=1}^n \delta_{(Y_i, \hat{R}_\nu^n(Y_i))} =: \gamma_n \xrightarrow{w} \gamma \qquad a.s.$$

The weak convergence almost surely is due to the fact that γ_n is a random probability measure.

Proof. Assume that all the random variables are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let us break down the proof in multiple steps.

Step 1: The left and right marginals of $(\gamma_n)_{n\in\mathbb{N}}$, which we denote by $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ converge weakly against μ and ν almost surely, respectively.

For $(\mu_n)_{n\in\mathbb{N}}$ this fact is immediate from Lemma 4.1.4. It implies that there exists $\tilde{\Omega} \in \mathcal{F}$ such that $\mathbb{P}(\tilde{\Omega}) = 1$ and for all $\omega \in \tilde{\Omega}$

$$\frac{1}{n}\sum_{i=1}^n \delta_{Y_i(\omega)} = \mu_n(\omega) \xrightarrow{w} \mu. \tag{5.3}$$

For $(\nu_n)_{n\in\mathbb{N}}$, note that the empirical distribution on the set $\{\hat{R}_\nu^n(Y_j)\}_{j\in[n]}$ equals the empirical distribution on the grid $\{X_j^{(n)}\}_{j\in[n]}$. Hence, the convergence follows from the definition of a grid. That is, there exists $\hat{\Omega} \in \mathcal{F}$ such that $\mathbb{P}(\hat{\Omega}) = 1$ and for all $\omega \in \hat{\Omega}$

$$\frac{1}{n}\sum_{i=1}^n \delta_{\hat{R}_\nu^n(Y_i(\omega))} = \frac{1}{n}\sum_{k=1}^n \delta_{X_k^{(n)}(\omega)} = \nu_n(\omega) \xrightarrow{w} \nu. \tag{5.4}$$

Now, fix $\omega \in \tilde{\Omega} \cap \hat{\Omega}$.

Step 2: The sequence $(\gamma_n(\omega))_{n\in\mathbb{N}}$ is tight.

By (5.3), (5.4) and Corollary 3.3.8, both $(\mu_n(\omega))_{n\in\mathbb{N}}$ and $(\nu_n(\omega))_{n\in\mathbb{N}}$ are tight. Hence, $(\gamma_n(\omega))_{n\in\mathbb{N}}$ is tight by Lemma 3.3.11.

Step 3: For any subsequence $x := (n_k)_{k\in\mathbb{N}}$, so that $\gamma_{n_k}(\omega) \xrightarrow{w} \gamma_x(\omega)$ converges to some $\gamma_x(\omega) \in \mathcal{P}^{2d}$, we have $\gamma_x(\omega) = \gamma$.

Suppose $x := (n_k)_{k\in\mathbb{N}}$ is such a subsequence and we have

$$\gamma_{n_k}(\omega) \xrightarrow{w} \gamma_x(\omega). \tag{5.5}$$

By (5.3) and (5.4) and Lemma 5.5.2 (ii), we must have $\gamma_x(\omega) \in \Gamma(\mu, \nu)$. Now, note that for all $k \in \mathbb{N}$,

$$\gamma_{n_k}(\omega) = \frac{1}{n_k}\sum_{i=1}^{n_k} \delta_{\left(Y_i(\omega), \hat{R}_\nu^{n_k}(Y_i(\omega))\right)}$$

has cyclically monotone support by the definition of empirical ranks. By (5.5) and Lemma 5.5.2 (i), we conclude that $\gamma_x(\omega)$ has cyclically monotone support. Since μ and ν are absolutely continuous, $\Gamma(\mu, \nu)$ contains only one measure with cyclically monotone support by Lemma 5.5.1. Since $\gamma \in \Gamma(\mu, \nu)$ has cyclically monotone support, we deduce that $\gamma_x(\omega) = \gamma$.

Finally, we obtain from Corollary 3.3.9 that for all $\omega \in \tilde{\Omega} \cap \hat{\Omega}$,

$$\gamma_n(\omega) \xrightarrow{w} \gamma.$$

Since $\mathbb{P}(\tilde{\Omega} \cap \hat{\Omega}) = 1$, we have

$$\gamma_n \xrightarrow{w} \gamma \qquad a.s.$$

and the assertion is proven.

□

Now, we can present the convergence theorem. It states that under certain integrability conditions the empirical ranks converge against the true rank statistics, where the mode of convergence is the average powered deviation. The claim is derived from Theorem 2.1 in (Deb et al., 2023). There the theorem is more general involving continuous joint transformations of the ranks and allowing a pooled sample from two different distributions. We tailored the statement to our proof of the consistency later. This also involved adding the power parameter α which is not included in the original, but useful for us later.

Theorem 5.5.4. *Let $\mu, \nu \in \mathcal{P}_{ac}^d$ be absolutely continuous. Further, let $Y_1, Y_2, \ldots \overset{iid}{\sim} \mu$ be a sample and let $(X_k^{(n)})_{n \geq k \geq 1}$ be a grid of ν. Let $\hat{R}_\nu^n(Y_1), \ldots, \hat{R}_\nu^n(Y_n)$ be the empirical ranks of $Y_1, \ldots, Y_n$ with respect to the random grid points $X_1^{(n)}, \ldots, X_n^{(n)}$. If for some $\alpha' > 0$*

$$\sup_{n \in \mathbb{N}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left\|X_i^{(n)}\right\|^{1+\alpha'}\right] < \infty \qquad \textit{and} \qquad \mathbb{E}\left[\left\|R_\nu^\mu(Y_1)\right\|^{1+\alpha'}\right] < \infty,$$

then for all $0 \leq \alpha < \alpha'$

$$\frac{1}{n} \sum_{i=1}^{n} \left\|\hat{R}_\nu^n(Y_i) - R_\nu^\mu(Y_i)\right\|^{1+\alpha} \xrightarrow{P} 0.$$

For the proof we will need Alexandroff's Theorem:

Theorem 5.5.5 ((Alexandroff, 1939))**.** *Let $f : U \to \mathbb{R}$ be a convex function where $U \subseteq \mathbb{R}^d$ open. Then f has a second derivative Lebesgue-a.e. in U.*

Proof. For this proof we may assume that all random variables are on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We begin by defining the function

$$g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, g(y, z) = \left\|R_\nu^\mu(y) - z\right\|^{1+\alpha}.$$

Note that R_ν^μ is the gradient of a convex function. Therefore, by Theorem 5.5.5, g is continuous almost everywhere. Now, let γ be the distribution of $(Y_1, R_\nu^\mu(Y_1))$. By Lemma 5.5.3 we have

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{(Y_i, \hat{R}_\nu^n(Y_i))} =: \gamma_n \xrightarrow{w} \gamma \qquad a.s.$$

For $n \in \mathbb{N}$, let $(C_n, D_n) \sim \gamma_n$. Now, we would like to apply the Continuous Mapping Theorem for weak convergence (Theorem 3.3.4) on g. We denote the set of discontinuities of g and R_ν^μ, by D_g and $D_{R_\nu^\mu}$, respectively. Then, we have

$$\gamma(D_g) = \mathbb{P}((Y_1, R_\nu^\mu(Y_1)) \in D_g) \leq \mathbb{P}(Y_1 \in D_{R_\nu^\mu}) = 0$$

Thereby, we used the fact that any discontinuity of g must come from a discontinuity of R_ν^μ since the other functions are all continuous. As g is continuous Lebesgue-a.e. and μ is absolutely continuous by assumption, we could conclude that g is continuous γ-a.e. Hence, the assumptions of Theorem 3.3.4 are satisfied and we have

$$g(C_n, D_n) \xrightarrow{w} g(Y_1, R_\nu^\mu(Y_1)) = 0 \qquad a.s.$$

Since weak convergence to a constant implies convergence in probability (Theorem 3.3.3), we have

$$g(C_n, D_n) \xrightarrow{P} g(Y_1, R_\nu^\mu(Y_1)) = 0 \qquad a.s.$$

Equivalently, for all $\varepsilon > 0$

$$\mathbb{P}\left(g(C_n, D_n) > \varepsilon \,\middle|\, Y_1, \ldots, Y_n, X_1^{(n)}, \ldots, X_n^{(n)}\right) \longrightarrow 0 \qquad a.s.$$

Next, we apply the bounded convergence theorem for conditional expectation (Theorem 3.2.13) and we obtain

$$\begin{aligned} 0 &= \lim_{n\to\infty} \mathbb{P}\left(g(C_n, D_n) > \varepsilon \,\middle|\, Y_1, \ldots, Y_n, X_1^{(n)}, \ldots, X_n^{(n)}\right) \\ &= \lim_{n\to\infty} \mathbb{P}\left(g(C_n, D_n) > \varepsilon \,\middle|\, (Y_i)_{i\in\mathbb{N}}, (X_i^{(j)})_{j\geq i\geq 1}\right) \\ &= \lim_{n\to\infty} \mathbb{E}\left[\mathbb{1}_{\{g(C_n,D_n)>\varepsilon\}} \,\middle|\, (Y_i)_{i\in\mathbb{N}}, (X_i^{(j)})_{j\geq i\geq 1}\right] \\ &= \mathbb{E}\left[\lim_{n\to\infty} \mathbb{1}_{\{g(C_n,D_n)>\varepsilon\}} \,\middle|\, (Y_i)_{i\in\mathbb{N}}, (X_i^{(j)})_{j\geq i\geq 1}\right], \end{aligned}$$

where the equalities hold almost surely. Applying expectations on both sides yields

$$\begin{aligned} 0 &= \mathbb{E}\left[\mathbb{E}\left[\lim_{n\to\infty} \mathbb{1}_{\{g(C_n,D_n)>\varepsilon\}} \,\middle|\, (Y_i)_{i\in\mathbb{N}}, (X_i^{(j)})_{j\geq i\geq 1}\right]\right] \\ &= \mathbb{E}\left[\lim_{n\to\infty} \mathbb{1}_{\{g(C_n,D_n)>\varepsilon\}}\right] \\ &= \lim_{n\to\infty} \mathbb{E}\left[\mathbb{1}_{\{g(C_n,D_n)>\varepsilon\}}\right] \\ &= \lim_{n\to\infty} \mathbb{P}\left(g(C_n, D_n) > \varepsilon\right), \end{aligned}$$

where we used the dominated convergence theorem (Theorem 3.2.10) in the third equality. Hence, we have shown $g(C_n, D_n) \xrightarrow{P} 0$. Next, define

$$\begin{aligned} V_n &= \frac{1}{n}\sum_{i=1}^{n} \left\| R_\nu^\mu(Y_i) - \hat{R}_\nu^n(Y_i) \right\|^{1+\alpha} \\ &= \mathbb{E}\left[g(C_n, D_n) \middle| Y_1, \ldots, Y_n, X_1^{(n)}, \ldots, X_n^{(n)} \right] \end{aligned}$$

for $n \in \mathbb{N}$. To finish the proof, we need to show $V_n \xrightarrow{P} 0$. We have

$$\begin{aligned} \mathbb{P}(V_n > \varepsilon) &= \mathbb{P}\left(\mathbb{E}\left[g(C_n, D_n) \middle| Y_1, \ldots, Y_n, X_1^{(n)}, \ldots, X_n^{(n)} \right] > \varepsilon \right) \\ &\leq \varepsilon^{-1} \mathbb{E}\left[\mathbb{E}\left[g(C_n, D_n) \middle| Y_1, \ldots, Y_n, X_1^{(n)}, \ldots, X_n^{(n)} \right] \right] \\ &= \varepsilon^{-1} \mathbb{E}[g(C_n, D_n)] \end{aligned}$$

for $\varepsilon > 0$ by Markov's inequality (Theorem 3.1.25). Next, we show that

$$(g(C_n, D_n))_{n\in\mathbb{N}} \quad \text{is uniformly integrable.} \tag{5.6}$$

This will prove that $\mathbb{E}[g(C_n, D_n)] \longrightarrow 0$ by Theorem 3.2.10, since we have already shown that $g(C_n, D_n) \xrightarrow{P} 0$. To show (5.6), let us rewrite $g(C_n, D_n)$:

$$\begin{aligned} g(C_n, D_n) &= \left\| R_\nu^\mu(C_n) - D_n \right\|^{1+\alpha} \\ &\leq 2\left\| R_\nu^\mu(C_n) \right\|^{1+\alpha} + 2\left\| D_n \right\|^{1+\alpha} \\ &=: 2G_n + 2H_n. \end{aligned}$$

Therefore, it is sufficient to show that $(G_n)_{n\in\mathbb{N}}$ and $(H_n)_{n\in\mathbb{N}}$ are uniformly integrable. We use the criterion from Theorem 3.2.9. Let $p := \frac{1+\alpha'}{1+\alpha} > 1$. We have

$$\begin{aligned} \sup_{n\in\mathbb{N}} \mathbb{E}[G_n^p] &= \sup_{n\in\mathbb{N}} \mathbb{E}\left[\mathbb{E}\left[G_n^p | Y_1, \ldots, Y_n \right] \right] \\ &= \sup_{n\in\mathbb{N}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left\| R_\nu^\mu(Y_i) \right\|^{1+\alpha'} \right] \\ &= \mathbb{E}\left[\left\| R_\nu^\mu(Y_1) \right\|^{1+\alpha'} \right] < \infty \end{aligned}$$

by assumption. Similarly, we have

$$\begin{aligned} \sup_{n\in\mathbb{N}} \mathbb{E}[H_n^p] &= \sup_{n\in\mathbb{N}} \mathbb{E}\left[\mathbb{E}\left[H_n^p \middle| Y_1, \ldots, Y_n, X_1^{(n)}, \ldots, X_n^{(n)} \right] \right] \\ &= \sup_{n\in\mathbb{N}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left\| F(\hat{R}_\nu^n(Y_i)) \right\|^{1+\alpha'} \right] < \infty \\ &= \sup_{n\in\mathbb{N}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left\| F(X_i^{(n)}) \right\|^{1+\alpha'} \right] < \infty \end{aligned}$$

by assumption. Hence, both, $(G_n)_{n\in\mathbb{N}}$ and $(H_n)_{n\in\mathbb{N}}$ are uniformly integrable by Theorem 3.2.9. Therefore, $(g(C_n, D_n))_{n\in\mathbb{N}}$ is uniformly integrable and the assertion is proven. □

Now, we have the convergence theorem for rank statistics that we needed in order to prove the consistency. As we will have an estimator for every sample size n for the consistency proof, the notation will be slightly more complicated. In order to be precise we define the notation and the setup of this section next. This will be similar to Notation 5.4.7, but more rigorous with entries of the matrices and additional indices referring to the sample size. Also we apply Algorithm 3 and define notations for the estimators resulting from it.

Notation 5.5.6. Let $Y^{(1)}, Y^{(2)}, \dots$ i.i.d. satisfying the CMCCA model be observed, where we summarize the first n samples in the data matrix

$$\boldsymbol{Y}^{(n)} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{pmatrix} = \begin{pmatrix} Y_1^{(1)} & Y_2^{(1)} \\ \vdots & \vdots \\ Y_1^{(n)} & Y_2^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

We denote the distributions of $Y_1^{(1)}$ and $Y_2^{(1)}$ by μ_1 and μ_2, respectively. Again, we assume μ_1 and μ_2 to be absolutely continuous. Further, let $Z^{(1)}, Z^{(2)}, \dots \overset{iid}{\sim} \mathcal{N}_p(0, C)$ be given, where

$$C = \begin{pmatrix} I_{p_1} & W \\ W^T & I_{p_2} \end{pmatrix}$$

is positive definite. By the CMCCA assumptions we have $G_j(Z_j^{(i)}) = Y_j^{(i)}$ for cyclically monotone functions G_1 and G_2. We summarize the first n of the $Z^{(i)}$'s in the data matrix

$$\boldsymbol{Z}^{(n)} = \begin{pmatrix} \boldsymbol{Z}_1^{(n)} & \boldsymbol{Z}_2^{(n)} \end{pmatrix} =$$

$$\begin{pmatrix} Z^{(1)} \\ \vdots \\ Z^{(n)} \end{pmatrix} = \begin{pmatrix} Z_1^{(1)} & Z_2^{(1)} \\ \vdots & \vdots \\ Z_1^{(n)} & Z_2^{(n)} \end{pmatrix} = \begin{pmatrix} Z_{1,1}^{(1)} & \cdots & Z_{1,p_1}^{(1)} & Z_{2,1}^{(1)} & \cdots & Z_{2,p_2}^{(1)} \\ \vdots & & \vdots & \vdots & & \vdots \\ Z_{1,1}^{(n)} & \cdots & Z_{1,p_1}^{(n)} & Z_{2,1}^{(n)} & \cdots & Z_{2,p_2}^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

We will later show a componentwise convergence, so that we need this notation for the entries of the matrix. Further, we will need the following terms later:

$$\bar{Z}_j^{(n)} = \frac{1}{n} \mathbf{1}_n^T \boldsymbol{Z}_j^{(n)} = \frac{1}{n} \sum_{i=1}^{n} Z_j^{(i)} = \begin{pmatrix} \bar{Z}_{j,1}^{(n)} & \cdots & \bar{Z}_{j,p_j}^{(n)} \end{pmatrix} \in \mathbb{R}^{1 \times p_j}$$

$$\tilde{\boldsymbol{Z}}_j^{(n)} = \boldsymbol{Z}_j^{(n)} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \boldsymbol{Z}_j^{(n)} = \begin{pmatrix} \tilde{Z}_j^{(1)} \\ \vdots \\ \tilde{Z}_j^{(n)} \end{pmatrix} = \begin{pmatrix} Z_{j,1}^{(1)} - \bar{Z}_{j,1}^{(n)} & \cdots & Z_{j,p_j}^{(1)} - \bar{Z}_{j,p_j}^{(n)} \\ \vdots & & \vdots \\ Z_{j,1}^{(n)} - \bar{Z}_{j,1}^{(n)} & \cdots & Z_{j,p_j}^{(n)} - \bar{Z}_{j,p_j}^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times p_j}$$

Next, we proceed as in Algorithm 3 and sample two grids $X_j^{(1)}, X_j^{(2)}, \dots \overset{iid}{\sim} \mathcal{N}_{p_j}(0, I_{p_j}) = \nu_j$ for $j = 1, 2$, independent from each other. Then, for $n \in \mathbb{N}$, $i \in [n]$ and $j = 1, 2$, we define

$$\hat{Z}_j^{(i,n)} = \hat{R}_{\nu_j}^n (Y_j^{(i)}),$$

where the empirical ranks are with respect to the corresponding grid points $X_j^{(1)}, \dots, X_j^{(n)}$. These will be summarized in the matrix

$$\hat{\boldsymbol{Z}}^{(n)} = \begin{pmatrix} \hat{\boldsymbol{Z}}_1^{(n)} & \hat{\boldsymbol{Z}}_2^{(n)} \end{pmatrix} =$$

$$\begin{pmatrix} \hat{Z}^{(1,n)} \\ \vdots \\ \hat{Z}^{(n,n)} \end{pmatrix} = \begin{pmatrix} \hat{Z}_1^{(1,n)} & \hat{Z}_2^{(1,n)} \\ \vdots & \vdots \\ \hat{Z}_1^{(n,n)} & \hat{Z}_2^{(n,n)} \end{pmatrix} = \begin{pmatrix} \hat{Z}_{1,1}^{(1,n)} & \cdots & \hat{Z}_{1,p_1}^{(1,n)} & \hat{Z}_{2,1}^{(1,n)} & \cdots & \hat{Z}_{2,p_2}^{(1,n)} \\ \vdots & & \vdots & \vdots & & \vdots \\ \hat{Z}_{1,1}^{(n,n)} & \cdots & \hat{Z}_{1,p_1}^{(n,n)} & \hat{Z}_{2,1}^{(n,n)} & \cdots & \hat{Z}_{2,p_2}^{(n,n)} \end{pmatrix} \in \mathbb{R}^{n\times p}.$$

Further, we set

$$\hat{\bar{Z}}_j^{(n)} = \frac{1}{n}\sum_{i=1}^n \hat{Z}_j^{(i)} = \begin{pmatrix} \hat{\bar{Z}}_{j,1}^{(n)} & \cdots & \hat{\bar{Z}}_{j,p_j}^{(n)} \end{pmatrix} \in \mathbb{R}^{1\times p_j}$$

$$\hat{\tilde{\boldsymbol{Z}}}_j^{(n)} = \hat{\boldsymbol{Z}}_j^{(n)} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\hat{\boldsymbol{Z}}_j^{(n)} = \begin{pmatrix} \hat{\tilde{Z}}_j^{(1,n)} \\ \vdots \\ \hat{\tilde{Z}}_j^{(n,n)} \end{pmatrix} = \begin{pmatrix} \hat{Z}_{j,1}^{(1,n)} - \hat{\bar{Z}}_{j,1}^{(n)} & \cdots & \hat{Z}_{j,p_j}^{(1,n)} - \hat{\bar{Z}}_{j,p_j}^{(n)} \\ \vdots & & \vdots \\ \hat{Z}_{j,1}^{(n,n)} - \hat{\bar{Z}}_{j,1}^{(n)} & \cdots & \hat{Z}_{j,p_j}^{(n,n)} - \hat{\bar{Z}}_{j,p_j}^{(n)} \end{pmatrix} \in \mathbb{R}^{n\times p_j}.$$

Finally, we define

$$\hat{W}^{(n)} = \left(\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)}\right)^{-1/2}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)}\left(\hat{\tilde{\boldsymbol{Z}}}_2^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)}\right)^{-1/2}$$

with singular value decomposition $\hat{W}^{(n)} = \hat{Q}_1^{(n)}\hat{\Lambda}^{(n)}\hat{Q}_2^{(n)^T}$, whereby $\hat{\Lambda}^{(n)} = \operatorname{diag}(\hat{\lambda}_1^{(n)}, \ldots, \hat{\lambda}_{p_1}^{(n)})$.

Now, Theorem 5.5.4 yields the intermediate step:

Corollary 5.5.7. *In the setting of Notation 5.5.6 we have*

$$\frac{1}{n}\sum_{i=1}^n \left\|\hat{Z}_k^{(i,n)} - Z_k^{(i)}\right\|^2 = \frac{1}{n}\sum_{i=1}^n \left\|\hat{R}_{\nu_k}^n(Y_k^{(i)}) - R_{\nu_k}^{\mu_k}(Y_k^{(i)})\right\|^2 \xrightarrow{P} 0$$

for $k = 1, 2$.

Proof. The claim follows from Theorem 5.5.4 when checking the integrability conditions with $\alpha' = 2$ and $\alpha = 1$. We have for $k = 1, 2$

$$\sup_{n\in\mathbb{N}} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \left\|X_k^{(i)}\right\|^3\right] = \mathbb{E}\left[\left\|X_k^{(i)}\right\|^3\right] < \infty$$

and

$$\mathbb{E}\left[\left\|R_{\nu_k}^{\mu_k}(Y_k^{(i)})\right\|^3\right] = \mathbb{E}\left[\left\|Z_k^{(1)}\right\|^3\right] < \infty.$$

This holds since all moments of the normal distribution exist (Theorem 5.1.3). □

Finally, we can show the consistency of our method:

Theorem 5.5.8. *In the setting of Notation 5.5.6, Algorithm 3 is consistent for the estimation of W. That is, we have*

$$\left(\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)}\right)^{-1/2}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)}\left(\hat{\tilde{\boldsymbol{Z}}}_2^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)}\right)^{-1/2} = \hat{W}^{(n)} \xrightarrow{P} W$$

and for all $k \in [p_1]$,

$$\hat{\lambda}_k^{(n)} \xrightarrow{P} \lambda_k.$$

Proof. The first step of the proof is to prove that for $k, l = 1, 2$, we have

$$\frac{1}{n}\left\|\hat{\tilde{\boldsymbol{Z}}}_k^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_l^{(n)} - \tilde{\boldsymbol{Z}}_k^{(n)^T}\tilde{\boldsymbol{Z}}_l^{(n)}\right\| \xrightarrow{P} 0, \tag{5.7}$$

where $\|\cdot\|$ is any norm. We show this componentwise. It holds for $i \in [p_k]$ and $j \in [p_l]$

$$\begin{aligned}
&\frac{1}{n}\left|\left(\hat{\tilde{\boldsymbol{Z}}}_k^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_l^{(n)}\right)_{ij} - \left(\tilde{\boldsymbol{Z}}_k^{(n)^T}\tilde{\boldsymbol{Z}}_l^{(n)}\right)_{ij}\right| \\
=&\frac{1}{n}\left|\left(\sum_{m=1}^{n}\left(\hat{Z}_{k,i}^{(m,n)} - \hat{\bar{Z}}_{k,i}^{(n)}\right)\left(\hat{Z}_{l,j}^{(m,n)} - \hat{\bar{Z}}_{l,j}^{(n)}\right)\right) - \sum_{m=1}^{n}\left(\left(Z_{k,i}^{(m)} - \bar{Z}_{k,i}^{(n)}\right)\left(Z_{l,j}^{(m)} - \bar{Z}_{l,j}^{(n)}\right)\right)\right| \\
=&\left|\left(\frac{1}{n}\sum_{m=1}^{n}\left(\hat{Z}_{k,i}^{(m,n)}\hat{Z}_{l,j}^{(m,n)}\right)\right) - \hat{\bar{Z}}_{k,i}^{(n)}\hat{\bar{Z}}_{l,j}^{(n)} - \left(\frac{1}{n}\sum_{m=1}^{n}\left(Z_{k,i}^{(m)}Z_{l,j}^{(m)}\right)\right) - \bar{Z}_{k,i}^{(n)}\bar{Z}_{l,j}^{(n)}\right| \\
\leq&\left(\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\hat{Z}_{l,j}^{(m,n)} - Z_{k,i}^{(m)}Z_{l,j}^{(m)}\right|\right) + \left|\hat{\bar{Z}}_{k,i}^{(n)}\hat{\bar{Z}}_{l,j}^{(n)}\right| + \left|\bar{Z}_{k,i}^{(n)}\bar{Z}_{l,j}^{(n)}\right|
\end{aligned}$$

To show that the last two summands converge to 0, note that we have by the SLLN

$$\begin{aligned}
\bar{Z}_k^{(n)} &= \frac{1}{n}\sum_{m=1}^{n} Z_k^{(m)} \xrightarrow{a.s.} 0 \\
\text{and} \qquad \hat{\bar{Z}}_k^{(n)} &= \frac{1}{n}\sum_{m=1}^{n}\hat{Z}_k^{(m)} = \frac{1}{n}\sum_{m=1}^{n} X_k^{(m)} \xrightarrow{a.s.} 0,
\end{aligned}$$

where the index k can be interchanged with l. Now, let us consider the first summand. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\hat{Z}_{l,j}^{(m,n)} - Z_{k,i}^{(m)}Z_{l,j}^{(m)}\right| \\
=&\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\hat{Z}_{l,j}^{(m,n)} - \hat{Z}_{k,i}^{(m,n)}Z_{l,j}^{(m)} + \hat{Z}_{k,i}^{(m,n)}Z_{l,j}^{(m)} - Z_{k,i}^{(m)}Z_{l,j}^{(m)}\right| \\
\leq&\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\right|\left|\hat{Z}_{l,j}^{(m,n)} - Z_{l,j}^{(m)}\right| + \frac{1}{n}\sum_{m=1}^{n}\left|Z_{l,j}^{(m)}\right|\left|\hat{Z}_{k,i}^{(m,n)} - Z_{k,i}^{(m)}\right| \\
\leq&\sqrt{\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\right|^2\cdot\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{l,j}^{(m,n)} - Z_{l,j}^{(m)}\right|^2} + \sqrt{\frac{1}{n}\sum_{m=1}^{n}\left|Z_{l,j}^{(m)}\right|^2\cdot\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)} - Z_{k,i}^{(m)}\right|^2}
\end{aligned}$$

by applying the Cauchy-Schwarz inequality. At this point we can apply Lemma 3.3.13 in order to show convergence to 0. That is, we show that in each square root one factor is $o_P(1)$ and the other is $O_P(1)$, so that their product is $o_P(1)$ and hence, converges to 0. Firstly, note that by the Strong Law of Large Numbers, we have

$$\frac{1}{n}\sum_{m=1}^{n}\left|Z_{l,j}^{(m)}\right|^2 \xrightarrow{a.s.} 1$$

$$\text{and} \qquad \frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\right|^2 = \frac{1}{n}\sum_{m=1}^{n}\left|X_{k,i}^{(m)}\right|^2 \xrightarrow{a.s.} 1.$$

As almost sure convergence implies weak convergence by Theorem 3.2.3 and Theorem 3.3.3, we can conclude that both sequences are tight by Corollary 3.3.8. Hence, we have

$$\frac{1}{n}\sum_{m=1}^{n}\left|Z_{l,j}^{(m)}\right|^2 \in O_p(1) \qquad \text{and} \qquad \frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\right|^2 \in O_p(1).$$

For the other factors, note that we have by Corollary 5.5.7

$$\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{l,j}^{(m,n)} - Z_{l,j}^{(m)}\right|^2 \in o_p(1) \qquad \text{and} \qquad \frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{j,i}^{(m,n)} - Z_{j,i}^{(k)}\right|^2 \in o_P(1).$$

Thus, we have shown that

$$\sqrt{\frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)}\right|^2 \cdot \frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{l,j}^{(m,n)} - Z_{l,j}^{(m)}\right|^2} + \sqrt{\frac{1}{n}\sum_{m=1}^{n}\left|Z_{l,j}^{(m)}\right|^2 \cdot \frac{1}{n}\sum_{m=1}^{n}\left|\hat{Z}_{k,i}^{(m,n)} - Z_{k,i}^{(m)}\right|^2}$$

converges to 0 in probability, where we also applied the Continuous Mapping Theorem (Theorem 3.2.4). Hence, (5.7) follows. Intuitively, this means that the CMCCA estimator based on the empirical ranks is asymptotically equivalent, to the CCA estimator if we knew the latent $Z^{(m)}$'s. An application of the triangle inequality yields

$$\frac{1}{n}\left\|\hat{\tilde{\boldsymbol{Z}}}_k^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_l^{(n)} - C_{kl}\right\| \leq \frac{1}{n}\left\|\hat{\tilde{\boldsymbol{Z}}}_k^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_l^{(n)} - \tilde{\boldsymbol{Z}}_k^{(n)^T}\tilde{\boldsymbol{Z}}_l^{(n)}\right\| + \frac{1}{n}\left\|\tilde{\boldsymbol{Z}}_k^{(n)^T}\tilde{\boldsymbol{Z}}_l^{(n)} - C_{kl}\right\| \xrightarrow{P} 0,$$

where $C_{11} = I_{p_1}$, $C_{22} = I_{p_2}$ and $C_{12} = W$. Thereby, the convergence of the second summand follows from Theorem 5.2.4 as in classical CCA. Hence, we have shown

$$\frac{1}{n}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)} \xrightarrow{P} I_{p_1}, \qquad \frac{1}{n}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)} \xrightarrow{P} I_{p_2}, \qquad \frac{1}{n}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)} \xrightarrow{P} W.$$

Now, Lemma 5.2.6 implies

$$\hat{W}^{(n)} = \left(\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)}\right)^{-1/2}\hat{\tilde{\boldsymbol{Z}}}_1^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)}\left(\hat{\tilde{\boldsymbol{Z}}}_2^{(n)^T}\hat{\tilde{\boldsymbol{Z}}}_2^{(n)}\right)^{-1/2} \xrightarrow{P} W.$$

By the same reasoning as in the proof of Theorem 5.2.8, the estimated canonical correlations $\hat{\lambda}_1^{(n)}, \ldots, \hat{\lambda}_{p_1}^{(n)}$ continuously depend on $W^{(n)}$, so that by the Continuous Mapping Theorem (Theorem 3.2.4),

$$\hat{\lambda}_k^{(n)} \xrightarrow{P} \lambda_k$$

for all $k \in [p_1]$. □

This finalizes our discussions of different CCA models. In the next chapter we will introduce another type of estimation method for the GCCCA and CMCCA, respectively.

6 Bayesian Methods for CCA

In this chapter we will present two more estimation methods for Canonical Correlation Analysis. These go in a different direction and are both Bayesian methods which use a Markov Chain Monte Carlo algorithm. In the first section of this chapter we will present a short introduction into Bayesian models and explain how they motivate the use of Markov Chains. Further, we will introduce the Gibbs Sampling algorithm, a classical Markov Chain Monte Carlo method. Then, we will present the two Bayesian Methods for the GCCCA and the CMCCA in Section 6.2 and 6.3, respectively.

For more detailed information on Bayesian Models and MCMC methods we refer to (Gelman et al., 2014).

6.1 Bayesian Models and Markov Chain Monte Carlo

When working with Bayesian models, one usually specifies prior and posterior distributions which have densities or probability mass functions. Further, one will also encounter conditional distributions and densities. Therefore, the notation can easily become confusing. In order to be precise, the following definition defines not only Bayesian Models, but also the notation for this chapter.

Definition 6.1.1 (Bayesian Model Formulation)**.** Suppose θ is a parameter (possibly a vector or a matrix). In Bayesian Models we assume that θ is random and we have some prior knowledge on it in form of a distribution. We will write

$$\theta \sim \pi(\theta),$$

where $\pi(\theta)$ is called the *prior distribution* of θ. Thereby, we will often assume that $\pi(\theta)$ has a density or probability mass function which we also denote by $\pi(\theta)$.

Now suppose we observe a sample $x = (x_1, \dots, x_n)$ which has a joint density or probability mass function depending on the parameter θ. We denote it by $\pi(x|\theta)$. Sometimes we will refer to the function $\pi(x|\theta)$ for a fixed a x depending on θ, as the *likelihood function* and write $l(\theta) = \pi(x|\theta)$.

Finally, we would like to make inference on θ using the information obtained from the sample x. As θ is random we can do this by the distribution of θ given x which we denote by $\pi(\theta|x)$. It can be expressed by Bayes Theorem

$$\pi(\theta|x) = \frac{\pi(\theta, x)}{\pi(x)} = \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)},$$

where $\pi(\theta, x)$ denotes the joint density of θ and x, and $\pi(x)$ the density of x. Then, $\pi(\theta|x)$ is called the *posterior distribution* of θ. Sometimes we will write

$$\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta),$$

to indicate proportionality as the normalization constant is not relevant in order to identify the posterior distribution.

One important case of Bayesian Models is when the prior distribution of θ and the distribution of x given θ are such that the posterior is from the same family as the prior. Such prior distributions will be called conjugate.

Definition 6.1.2 (Conjugate Prior Distributions)**.** Suppose $\mathscr{P}$ denotes a family of prior distributions and $\mathscr{F}$ denotes a family of distributions for the sample x. Then, we call $\mathscr{P}$ conjugate to $\mathscr{F}$, if for every prior $\pi(\theta) \in \mathscr{P}$ and every $x \sim \pi(x|\theta) \in \mathscr{F}$, we have $\pi(\theta|x) \in \mathscr{P}$.

Let us provide an example for a Bayesian Model and conjugacy.

Example 6.1.3. Suppose we are interested in inferring a mean parameter θ of a normal distribution $\mathcal{N}(\theta, 1)$. Thereby, we have prior knowledge about θ, which is also a normal distribution, given by $\pi(\theta) \sim \mathcal{N}(m, \sigma^2)$, where $m \in \mathbb{R}$ and $\sigma > 0$. Now, we observe a sample $x = (x_1, \dots, x_n) \overset{iid}{\sim} \mathcal{N}(\theta, 1)$, so that we can determine the posterior distribution of θ. In the following let $\bar{x} = \frac{1}{n} \sum\limits_{i=1}^{n} x_i$ denote the sample mean.

$$\begin{aligned}
\pi(\theta|x) &\propto \pi(x|\theta)\pi(\theta) \\
&= \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2 \right) \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\theta - m)^2 \right) \\
&\propto \exp\left(-\frac{1}{2} \left(\sum_{i=1}^{n} (\theta - x_i)^2 \right) - \frac{1}{2\sigma^2}(\theta - m)^2 \right) \\
&\propto \exp\left(-\frac{1}{2} \left(n\theta^2 - 2\theta \cdot n\bar{x} + \frac{\theta^2}{\sigma^2} - 2\theta \frac{m}{\sigma^2} \right) \right) \\
&= \exp\left(-\frac{1}{2} \left(\frac{n\sigma^2 + 1}{\sigma^2} \theta^2 - 2\theta \cdot \left(n\bar{x} + \frac{m}{\sigma^2} \right) \right) \right) \\
&= \exp\left(-\frac{1}{2 \cdot \frac{\sigma^2/n}{\sigma^2 + 1/n}} \left(\theta^2 - 2\theta \cdot \left(\bar{x} \frac{\sigma^2}{\sigma^2 + 1/n} + m \frac{1/n}{\sigma^2 + 1/n} \right) \right) \right) \\
&\propto \exp\left(-\frac{1}{2 \cdot \frac{\sigma^2/n}{\sigma^2 + 1/n}} \left(\theta - \left(\bar{x} \frac{\sigma^2}{\sigma^2 + 1/n} + m \frac{1/n}{\sigma^2 + 1/n} \right) \right)^2 \right)
\end{aligned}$$

Note that we were interested in identifying the posterior distribution and thus could rewrite the density up to proportionality without considering normalization constants. We conclude that the posterior distribution of θ is given by

$$\pi(\theta|x) \sim \mathcal{N}\left(\bar{x} \frac{\sigma^2}{\sigma^2 + 1/n} + m \frac{1/n}{\sigma^2 + 1/n}, \frac{\sigma^2/n}{\sigma^2 + 1/n} \right)$$

One can also observe that the posterior distribution is normal like the prior. Hence, we have shown that the family of priors $\mathscr{P} = \{\mathcal{N}(m, \sigma^2) : m \in \mathbb{R}, \sigma > 0\}$ is conjugate to the family $\mathscr{F} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$.

For readers who are new to the topic a valid question might be: How are Bayesian models related to Markov Chains? The answer is that Markov Chains can help us to sample from a complicated posterior distribution. Let us provide a short introduction into Markov Chains. For more detailed information we refer to (Feller, 1971) and (Meyn and Tweedie, 2009).

Definition 6.1.4 (Markov Chain)**.** Suppose $S \in \mathcal{B}(\mathbb{R}^d)$ is a state space. Then, we call a sequence of random vectors $(\theta^{(n)})_{n\in\mathbb{N}}$ with values in S a *stochastic process*.

A stochastic process is called *Markov Chain* if the future states are independent of the past states given the present state. That is, for all n, $x \in S$ and all $A_1, \dots, A_{n-1}, A_{n+1} \in \mathcal{B}(S)$, we have

$$\mathbb{P}(\theta^{(n+1)} \in A_{n+1}|\theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \dots, \theta^{(1)} \in A_1) = \mathbb{P}(\theta^{(n+1)} \in A_{n+1}|\theta^{(n)} = x).$$

If the probability on the right-hand side is independent of n for all $A = A_{n+1} \in \mathcal{B}(S)$ and all $x \in S$, then $(\theta^{(n)})_{n\in\mathbb{N}}$ is called a *homogeneous Markov Chain.*

For a stochastic process on a finite state space S, the following condition is sufficient to show that it is a Markov Chain: It holds for all $x_1, \dots, x_{n+1} \in S$,

$$\mathbb{P}(\theta^{(n+1)} = x_{n+1}|\theta^{(n)} = x_n, \theta^{(n-1)} = x_{n-1}, \dots, \theta^{(1)} = x_1) = \mathbb{P}(\theta^{(n+1)} = x_{n+1}|\theta^{(n)} = x_n)$$

The probabilities on the right-hand side are then called transition probabilities.

The most important definition for the algorithms later is the stationary distribution.

Definition 6.1.5 (Stationary Distribution)**.** Let $(\theta^{(n)})_{n\in\mathbb{N}}$ be a homogeneous Markov Chain on the state space $S \in \mathcal{B}(\mathbb{R}^d)$. A probability measure π on $(S, \mathcal{B}(S))$ is called *stationary distribution* of the Markov Chain $(\theta^{(n)})_{n\in\mathbb{N}}$ if for all $A \in \mathcal{B}(S)$

$$\pi(A) = \int\limits_S \mathbb{P}(\theta^{(n+1)} \in A|\theta^{(n)} = x)\mathrm{d}\pi(x). \tag{6.1}$$

Condition (6.1) can be interpreted as follows. If the distribution of the Markov Chain at time n is π, it also at time $n + 1$. Hence, if a Markov Chain reaches a stationary distribution it is in an equilibrium.

For Markov Chain on a finite state space condition (6.1) is equivalent to the following: If for all $y \in S$ it holds

$$\pi(y) = \sum_{x\in S} \mathbb{P}(\theta^{(n+1)} = y|\theta^{(n)} = x)\pi(x), \tag{6.2}$$

then π is the stationary distribution of $(\theta^{(n)})_{n\in\mathbb{N}}$. If $\pi = (\pi(x))_{x\in S}$ is written as a row vector and $\Pi = (\Pi(x, y))_{x,y\in S}$ is matrix containing the transition probabilities $\Pi(x, y) = \mathbb{P}(\theta^{(n+1)} = y|\theta^{(n)} = x)$, condition (6.2), can be written as

$$\pi = \pi\Pi.$$

Why is the stationary distribution important for us? As we mentioned Markov Chains shall help us to sample from a complicated posterior distribution $\pi(\theta|x)$. The idea behind Markov Chain Monte Carlo algorithms is to construct a Markov Chain which has $\pi(\theta|x)$

as stationary distribution. Then, under certain conditions the distribution of the Markov Chain at time n converges against the stationary distribution. Hence, we can sample from a distribution close to $\pi(\theta|x)$ if we run the Markov Chain long enough provided it converges.

There is theory on the general convergence of Markov Chains, e.g. see (Meyn and Tweedie, 2009), particularly for countable or finite state spaces. As we will consider $\mathbb{R}^d$ as a state space, we will not state the theory here. Instead, we provide a simple example of a convergence to a stationary distribution in order to give some intuition.

Example 6.1.6. Suppose a frog is jumping back and forth between two leafs. Thereby, it stays on its current leaf with probability $\frac{3}{4}$ and jumps to the other with probability $\frac{1}{2}$. Then, this Markov Chain has transition matrix

$$\Pi = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

It is easy to show that the stationary distribution of this Markov Chain is given by the probability measure $\mu = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. However, suppose the Markov Chain starts with an arbitrary initial distribution μ_0. Also, let us denote the state space by $S = \{1, 2\}$. We then have for $x \in S$

$$\begin{aligned} & \mathbb{P}(\theta^{(n+1)} = x) \\ = \sum_{x_1,\dots,x_n \in S} & \mathbb{P}(\theta^{(n+1)} = x, \theta^{(n)} = x_n, \dots, \theta^{(1)} = x_1) \\ = \sum_{x_1,\dots,x_n \in S} & \mathbb{P}(\theta^{(n+1)} = x | \theta^{(n)} = x_n, \dots, \theta^{(1)} = x_1) \\ & \cdot \mathbb{P}(\theta^{(n)} = x_n | \theta^{(n-1)} = x_{n-1}, \dots, \theta^{(1)} = x_1) \cdots \mathbb{P}(\theta^{(1)} = x_1) \\ = \sum_{x_1,\dots,x_n \in S} & \mathbb{P}(\theta^{(n+1)} = x | \theta^{(n)} = x_n) \mathbb{P}(\theta^{(n)} = x_n | \theta^{(n-1)} = x_{n-1}) \cdots \mathbb{P}(\theta^{(1)} = x_1) \\ = \sum_{x_1,\dots,x_n \in S} & \Pi(x_n, x)\Pi(x_{n-1}, x_n) \cdots \Pi(x_1, x_2)\mu_0(x_1) = \mu_0 \Pi^{n-1} \Pi(\cdot, x), \end{aligned}$$

where $\Pi(\cdot, x)$ is the column of Π corresponding to x. Hence, the distribution of the Markov Chain at time n, denoted by μ_n, is given by

$$\begin{aligned} \mu_n = \mu_0 \Pi^n = \mu_0 \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}^n &= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2^n} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} \mu_0(1) & \mu_0(2) \end{pmatrix} \begin{pmatrix} 1 + \frac{1}{2^n} & 1 - \frac{1}{2^n} \\ 1 - \frac{1}{2^n} & 1 + \frac{1}{2^n} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2} + \frac{1}{2^n}(\mu_0(1) - \frac{1}{2}) & \frac{1}{2} + \frac{1}{2^n}(\mu_0(2) - \frac{1}{2}) \end{pmatrix}. \end{aligned}$$

Thereby, we used the fact that $\mu_0(1) + \mu_0(2) = 1$. In fact, we have that μ_n converges against the stationary distribution:

$$\lim_{n\to\infty} \mu_n = \lim_{n\to\infty} \begin{pmatrix} \frac{1}{2} + \frac{1}{2^n}(\mu_0(1) - \frac{1}{2}) & \frac{1}{2} + \frac{1}{2^n}(\mu_0(2) - \frac{1}{2}) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

In this example we observed that although the Markov Chain can be initialized with any distribution, after a sufficiently long time its distribution becomes close to the stationary one. Now with this intuition, the motivation and procedure behind the algorithms we will present can be summarized as follows.

1. Suppose one observes a sample $x = (x_1, \dots, x_n)$ whose distribution depends on a unknown parameter θ, and is interested in a statistic $t(\theta)$.
2. Choose a suitable prior $\pi(\theta)$ and identify the posterior distribution $\pi(\theta|x)$.
3. Construct a Markov Chain whose stationary distribution is the posterior $\pi(\theta|x)$.
4. Sample m independent trials $\hat{\theta}_1, \dots, \hat{\theta}_m$ from the Markov Chain after a sufficiently long running time and compute the posterior mean
$$\widehat{t(\theta)} = \sum_{k=1}^{m} t(\hat{\theta}_k).$$

We finish this section by introducing Gibbs Sampling which is the type of Markov Chain Monte Carlo algorithm we will encounter later. Beside the Metropolis algorithm it is one of the most commonly used MCMC methods. For more detailed information we refer to (Gelman et al., 2014).

Gibbs sampling is mostly used when there are multiple parameters involved, so that it is difficult to draw from the joint distribution, but possible from conditional distributions. Particularly, Bayesian Models with conjugate priors are suitable for Gibbs Sampling as the conditional distribution before and after resampling will be of the same type. When we introduce the algorithm, we will use the notation $\theta = (\theta_1, \dots, \theta_d)$ for the parameters. These can denote anything, e.g. numbers, vectors, matrices, etc.

Algorithm 4 Gibbs Sampling

```
Input: number of iterations $m$, data $x$, initial values $\theta_0$
Output: sample $\theta = \theta^{(m)}$
Set initial values $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})$.
for $i = 1, \dots, m$ do
    for $j = 1, \dots, d$ do
        Sample $\theta_j^{(i)} \sim \pi(\theta_j|\theta_1^{(i)}, \dots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \dots, \theta_d^{(i-1)}, x)$
Return $\theta^{(m)}$.
```

Before we finish this section, let us provide a simple example for a Gibbs sampling algorithm from (Gelman et al., 2014, p.277).

Example 6.1.7 (Gibbs Sampling)**.** Suppose we observe one single sample from a bivariate normal distribution
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$
where $\rho \in (-1, 1)$ is known and we would like to infer θ. The next step is to choose a suitable prior distribution for θ. We will do this by picking a improper prior distribution

by setting $\pi(\theta) \sim 1$. This prior is called improper as it integrates to ∞ and therefore, is no probability distribution. However, improper priors can yield proper posterior distributions, as we will observe next. We have

$$\pi(\theta|x) \propto l(\theta)\pi(\theta) \propto \exp\left(-\frac{1}{2}\begin{pmatrix} x_1 - \theta_1 & x_2 - \theta_2 \end{pmatrix}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}\begin{pmatrix} x_1 - \theta_1 \\ x_2 - \theta_2 \end{pmatrix}\right).$$

Hence, the posterior distribution is proper and given by

$$\pi(\theta|x) \sim \mathcal{N}_2\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Of course this is an easy distribution to sample from, so that one usually would not apply Gibbs Sampling. But for understanding the algorithm, this example can be helpful. We need conditional distributions which can be determined from Proposition 5.1.6. Hence, we have

$$\begin{aligned} \pi(\theta_1|\theta_2, x) &\sim \mathcal{N}(x_1 + \rho(\theta_2 - x_2), 1 - \rho^2) \\ \pi(\theta_2|\theta_1, x) &\sim \mathcal{N}(x_2 + \rho(\theta_1 - x_1), 1 - \rho^2). \end{aligned}$$

Therefore, we have the following Gibbs Sampling algorithm:

Algorithm 5 Gibbs Sampling Example

```
Input: number of iterations $m$, data $x$, initial values $\theta_0$
Output: sample $\theta = \theta^{(m)}$
Set initial values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$.
for $i = 1, \dots, m$ do
    Resample $\theta_1$:    $\theta_1^{(i)} \sim \mathcal{N}(x_1 + \rho(\theta_2 - x_2), 1 - \rho^2)$.
    Resample $\theta_2$:    $\theta_2^{(i)} \sim \mathcal{N}(x_2 + \rho(\theta_1 - x_1), 1 - \rho^2)$.
Return $\theta^{(m)}$.
```

This concludes our introduction to Bayesian Models and Markov Chain Monte Carlo algorithms. In the next chapter, we will learn how these yield new estimation methods for CCA.

6.2 The Extended Rank Likelihood for GCCCA

In this section we introduce a Bayesian method for the Gaussian Copula CCA. We begin by defining the notation and the setup throughout the section. Then, we continue by defining the likelihood in this model which is a nontrivial question. Further, we address the choice of priors and finally, present the Gibbs Sampling Algorithm which is based on (Hoff, 2007b).

Notation 6.2.1. Suppose we observe $Y^{(1)}, \dots, Y^{(n)} \overset{iid}{\sim} \mathrm{NPN}_p(C, f)$, so that $Z^{(1)}, \dots, Z^{(n)} \overset{iid}{\sim} \mathcal{N}_p(0, C)$ are joint normal and we have

$$Y_k^{(i)} = F_k^{-1}(\Phi(Z_k^{(i)}))$$

for all $i \in [n]$ and $k \in [p]$, where we used the representation from (5.1). Thereby, C is a correlation matrix and $F_1, \dots, F_p$ are continuous distribution functions. We summarize the $Y^{(i)}$'s and $Z^{(i)}$'s in the data matrices

$$\boldsymbol{Z} = \begin{pmatrix} Z^{(1)} \\ \vdots \\ Z^{(n)} \end{pmatrix} = \begin{pmatrix} Z_1^{(1)} & \dots & Z_p^{(1)} \\ \vdots & \ddots & \vdots \\ Z_1^{(n)} & \dots & Z_p^{(n)} \end{pmatrix} \qquad \boldsymbol{Y} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{pmatrix} = \begin{pmatrix} Y_1^{(1)} & \dots & Y_p^{(1)} \\ \vdots & \ddots & \vdots \\ Y_1^{(n)} & \dots & Y_p^{(n)} \end{pmatrix}.$$

Our goal is to estimate the correlation matrix C. As in Section 5.3, the problem is that the normal $Z^{(i)}$'s are unobserved. There, we were able to tackle this issue by estimating the entries of C with Spearman's ρ. This time we will go in a different direction and infer C with a Bayesian Method which has been presented in Hoff (2007b). The procedure can generally be applied for estimating correlation matrices of Gaussian Copulas and is therefore, not specific to our application in CCA estimation.

Now, when setting up a Bayesian model as in Definition 6.1.1, we have the data $\boldsymbol{Y}$ and unknown parameters $C, F_1, \dots, F_p$. Hence, our "likelihood" would look somewhat like this

$$l(C) = \pi(\boldsymbol{Y}|C, F_1, \dots, F_p),$$

which is obviously inconvenient to work with. The idea from Hoff (2007b) is to replace the data $\boldsymbol{Y}$ by some event which yields information about $\boldsymbol{Z}$, so that the resulting likelihood is independent of the nuisance parameters $F_1, \dots, F_p$.

Due to the model assumptions we have some information on $\boldsymbol{Z}$ when observing $\boldsymbol{Y}$, namely that the univariate marginals of the $Y^{(i)}$'s are increasing transformations of the $Z^{(i)}$'s. Hence, if we observe $Y_j^{(i_1)} \leq Y_j^{(i_2)}$, we must have $Z_j^{(i_1)} \leq Z_j^{(i_2)}$. More generally, observing $\boldsymbol{Y}$ means that $\boldsymbol{Z}$ must lie in the set

$$D := \left\{ \mathbf{Z} \in \mathbb{R}^{n \times p} | \max\left\{ Z_j^{(k)} : Y_j^{(k)} \leq Y_j^{(i)} \right\} \leq Z_j^{(i)} \leq \min\left\{ Z_j^{(k)} : Y_j^{(k)} \leq Y_j^{(k)} \right\} \right\}.$$

Now, we can replace our data $\boldsymbol{Y}$ by the occurrence of this event. Then, the likelihood, also called "extended rank likelihood", is given by

$$l(C) = \pi(\boldsymbol{Z} \in D|C, F_1, \dots, F_p) = \mathbb{P}(\boldsymbol{Z} \in D|C).$$

We observe that the last equality holds since the distribution of the $Z^{(i)}$'s is independent of $F_1, \dots, F_p$. Hence, our likelihood does not depend on the nuisance parameters which is very convenient. As we work with a Bayesian model, the full posterior distribution is given by

$$\pi(C|\boldsymbol{Z} \in D) \propto \mathbb{P}(\boldsymbol{Z} \in D|C)\pi(C),$$

where $\pi(C)$ is a prior distribution of Z. We address the choice of priors next.

Later we will resample C and $\boldsymbol{Z}$ with a Gibbs Sampling algorithm for which we need a suitable prior for C. We learned in the previous section that suitable choices of prior distributions for Gibbs Sampling are conjugate priors. One such distribution is the Inverse-Wishart which is conjugate to the multivariate normal distribution. In order to present it we need to define Wishart distribution first.

Definition 6.2.2 (Wishart Distribution)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}_d(0, \Sigma)$, where Σ is positive definite. Then,

$$\sum_{i=1}^{n} X_i X_i^T \sim \mathcal{W}_d(n, \Sigma)$$

follows a *Wishart Distribution* with n degrees of freedom and covariance matrix Σ. Hence, $\mathcal{W}_d(n, \Sigma)$ is a distribution on the set of symmetric and positive semidefinite matrices in $\mathbb{R}^{d \times d}$.

Proposition 6.2.3 (Corollary 7.2 in (Bilodeau and Brenner, 1999))**.** *Let $\Sigma \in \mathbb{R}^{d \times d}$ be symmetric and positive definite and let $n \in \mathbb{N}$. Further, let $W \sim \mathcal{W}_d(n, \Sigma)$. If $n > d - 1$, then $\mathbb{P}(W \text{ is invertible}) = 1$.*

The preceding proposition justifies the following definition:

Definition 6.2.4 (Inverse Wishart Distribution)**.** Let $W \sim \mathcal{W}_d(n, \Sigma)$ follow a Wishart Distribution with symmetric and positive definite covariance matrix Σ with n degrees of freedom where $n > d - 1$. Then, its inverse W^{-1} follows an *Inverse Wishart Distribution.* With $\Psi = \Sigma^{-1}$ we write

$$W^{-1} \sim \mathcal{IW}_d(n, \Psi).$$

Some important properties of the Inverse-Wishart distribution are the following:

Proposition 6.2.5. *Let $W \sim \mathcal{W}_d(n, \Sigma)$ with $n > d + 1$ and $\Psi = \Sigma^{-1}$. We have*

1. $\mathbb{E}[W] = n\Sigma$
2. $\mathbb{E}[W^{-1}] = \dfrac{\Psi}{n - d - 1}$.

Proof. For 1. see page 66 of (Mardia et al., 1979). For 2. we refer to Theorem 3.1 in (von Rosen, 1988). □

Theorem 6.2.6. *Let $\Psi \in \mathbb{R}^{d \times d}$ be symmetric and positive definite and let $\Sigma \sim \mathcal{IW}_d(n, \Psi)$. Further, let $X_1, \dots, X_m \overset{iid}{\sim} \mathcal{N}_d(0, \Sigma)$ and summarize the samples into the matrix $X = (X_1 | \dots | X_m) \in \mathbb{R}^{d \times m}$. Then, we have for the posterior distribution of Σ*

$$\Sigma | X \sim \mathcal{IW}_d(n + m, \Psi + XX^T).$$

Proof. We refer to (Zhang, 2021). □

Hence, the Inverse-Wishart distribution is conjugate to the multivariate normal distribution, and therefore perfectly suitable for our Gibbs Sampling algorithm later. However, there is one drawback: The theorem is for sampling from normals with general covariance matrices while the GCCCA is parametrized with a correlation matrix. Therefore, we will resample over a covariance matrix V instead of C and C will be equal to its corresponding correlation matrix.

For the prior distribution of V, we let $V = (v_{ij}) \sim \mathcal{IW}_p(n_0, n_0 V_0)$ have inverse Wishart Distribution, so that $\mathbb{E}[V^{-1}] = V_0^{-1}$, where V_0 is the initial positive definite covariance matrix and $n_0 \geq p$ is some prior weight. Thereby, the initial V_0 will be chosen to be the covariance matrix of the initial $\boldsymbol{Z}_0$ which will be based on the observed $\boldsymbol{Y}$. We state the algorithm first and then explain how starting values can be chosen and the intuition behind all the steps. The Gibbs Sampling scheme is given as follows:

Algorithm 6 Gaussian Copula: Correlation Matrix Estimation

```
Input: Data Matrix $\boldsymbol{Y}$, prior weight $n_0$, iterations $m$
Output: Posterior Sample of Correlation Matrix $C = (c_{ij}) \sim \pi(C|\boldsymbol{Z} \in D)$
 1: Set starting values $\boldsymbol{Z} = \boldsymbol{Z}_0$ and $V_0$ as described after.
 2: Generate $V = (v_{ij}) \sim \mathcal{IW}_p(n_0, n_0 V_0)$
 3: for $k = 1, \dots, m$ do
 4:     Resample $\boldsymbol{Z}$:
 5:     for $j = 1, \dots, p$ do
 6:         for each unique $Y \in \{Y_j^{(n)}, \dots, Y_j^{(n)}\}$ do
 7:             Set $Z_l = \max\{Z_j^{(i)} : Y_j^{(i)} \leq Y, i \in [n]\}$, $Z_u = \min\{Z_j^{(i)} : Y \leq Y_j^{(i)}, i \in [n]\}$
 8:             for each $i$ such that $Y_j^{(i)} = Y$ do
 9:                 Set $\sigma_j^2 = v_{jj} - V_{[j,-j]} V_{[-j,-j]}^{-1} V_{[-j,j]}$.
10:                 Compute $\mu_{i,j} = \boldsymbol{Z}_{[i,-j]} (V_{[j,-j]} V_{[-j,-j]}^{-1})^T$
11:                 Sample $u_{i,j} \sim \text{Unif}\left(\Phi\left(\frac{Z_l - \mu_{i,j}}{\sigma_j}\right), \Phi\left(\frac{Z_u - \mu_{i,j}}{\sigma_j}\right)\right)$
12:                 Set $Z_j^{(i)} = \mu_{i,j} + \sigma_j \cdot \Phi^{-1}(u_{i,j})$.
13:     Resample $V \sim \mathcal{IW}_p(n_0 + n, n_0 V_0 + \boldsymbol{Z}^T \boldsymbol{Z})$
14: Compute C: For all $i, j$, set $c_{ij} = v_{ij} / \sqrt{v_{ii} v_{jj}}$
```

Thereby, $\boldsymbol{Z}_{[i,-j]}$ denotes the i-th row of the matrix $\boldsymbol{Z}$ where the j-th column is removed. Further, $V_{[-j,-j]}$ is the matrix V with its j-th column and j-th row removed, and $V_{[-j,-j]}^{-1}$ is its inverse. Finally, $V_{[j,-j]}$ denotes the j-th row of V where the j-th entry is removed.

Before addressing the choice of initial values let us explain the steps of this algorithm. As we mentioned before, the Inverse-Wishart distribution is supported over the space of general covariance matrices, so that we resample over $\boldsymbol{Z}$ and V instead of $\boldsymbol{Z}$ and C. This is done in the lines 5-13 of the algorithm Thereby, we sample from the distributions

(i) $Z_j^{(i)} \sim p(Z_j^{(i)} | \boldsymbol{Z} \in D, \boldsymbol{Z} \setminus \{Z_j^{(i)}\}, V)$

(ii) $V \sim \mathcal{IW}_p(n_0 + n, n_0 V_0 + \boldsymbol{Z}^T \boldsymbol{Z})$

As the sampling of V is clear, let us explain how we do it for the $Z_j^{(i)}$'s. When determining a new value for $Z_j^{(i)}$, we need to make sure that the new $\boldsymbol{Z}$ remains in D. Hence, we calculate and admissible interval $[Z_u, Z_l]$ in line 7 based on the structure of $Y_j^{(1)}, \dots, Y_j^{(n)}$. Then, we determine the conditional normal distribution parameters of $Z_j^{(i)}$ given $Z_1^{(i)}, \dots, Z_{j-1}^{(i)}, Z_{j+1}^{(i)}, \dots, Z_p^{(i)}$ in line 9-10. This is done with the formula from Proposition 5.1.6. As this distribution is truncated to the set $[Z_u, Z_l]$ we apply the generalized inverse transform method to sample from this conditional distribution.

After the resampling is run long enough, we terminate with a covariance matrix V and calculate its corresponding correlation matrix in line 14.

Finally, let us address the question of starting values. As V is the covariance matrix of $\boldsymbol{Z}$ in the algorithm, we chose the initial V_0 as the covariance matrix of $\boldsymbol{Z}_0$, i.e. $V_0 = \frac{1}{n-1}(\tilde{\boldsymbol{Z}}_0^T\tilde{\boldsymbol{Z}}_0)$, where $\tilde{\boldsymbol{Z}}_0 = \boldsymbol{Z}_0 - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{Z}_0$. So how do we choose $\boldsymbol{Z}_0$? As we need to make sure that $\boldsymbol{Z}_0 \in D$, we need to keep the ordering of the columns of $\boldsymbol{Y}$ while transforming the data to a normal distribution. We do this the following way.

Let the columns of $\boldsymbol{Y}$ be denoted by $\boldsymbol{Y} = (\gamma_1|\cdots|\gamma_p)$. Then, we set $\boldsymbol{Z}_0 = (\tilde{\zeta}_1|\cdots|\tilde{\zeta}_p)$, where the $\tilde{\zeta}_j$'s are transformations of the γ_j's which we explain next. For $j \in [p]$, γ_j is the vector of j-th marginals of the sample, i.e. $\gamma_j = \begin{pmatrix} Y_j^{(1)} & \cdots & Y_j^{(n)} \end{pmatrix}^T$. Now, let $F_j^{(n)}$ be the empirical distribution function of this sample. Then, define

$$\zeta_j = \left(\Phi^{-1}\left(\tfrac{n}{n+1}F_j^{(n)}(Y_j^{(1)})\right) \quad \cdots \quad \Phi^{-1}\left(\tfrac{n}{n+1}F_j^{(n)}(Y_j^{(n)})\right)\right)^T, \tag{6.3}$$

and its mean and variance

$$\begin{aligned} m_j &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\Phi^{-1}\left(\frac{n}{n+1}F_j^{(n)}(Y_j^{(i)})\right) \\ s_j^2 &\coloneqq \frac{1}{n-1}\sum_{i=1}^{n}\left(\Phi^{-1}\left(\frac{n}{n+1}F_j^{(n)}(Y_j^{(i)})\right) - m_j\right)^2 \end{aligned}$$

Now, define the standardized $\tilde{\zeta}_j = (\zeta_j - m_j\mathbf{1}_n)/s_j$ and $\boldsymbol{Z}_0 = (\tilde{\zeta}_1|\cdots|\tilde{\zeta}_p)$. In summary, we "transport" the γ_j's to normal variables in (6.3) while maintaining. Thereby, the $\frac{n}{n+1}$ factor is the to ensure that finite values are generated. Finally, we standardize the columns. During the whole process the ordering of the γ_j was maintained so that we have $\boldsymbol{Z}_0 \in D$.

Finally, one can compute C which is the corresponding correlation matrix to V. Then, the generated Markov Chain has the stationary distribution $\pi(C|\boldsymbol{Z} \in D)$ by (Hoff, 2007b). Now, the remaining part is to estimate the GCCCA parameters of Y which can be done with the posterior mean.

Algorithm 7 GCCCA Estimation

```
Input: As Algorithm 6
Output: GCCCA Parameters $\hat{W}$ and $\hat{\Lambda}$
Generate $k$ posterior samples $\hat{C}^{(1)}, \dots, \hat{C}^{(k)} \sim \pi(C|\boldsymbol{Z} \in D)$ with Algorithm 6.
for $j = 1, \dots, k$ do
    Let $\hat{C}^{(j)}$ have block structure $\hat{C}^{(j)} = \begin{pmatrix} \hat{C}_{11}^{(j)} & \hat{C}_{12}^{(j)} \\ \hat{C}_{21}^{(j)} & \hat{C}_{22}^{(j)} \end{pmatrix}$.
    Set $\hat{W}^{(j)} = \hat{C}_{11}^{(j)^{-1/2}}\hat{C}_{12}^{(j)}\hat{C}_{22}^{(j)^{-1/2}}$.
Compute the posterior mean $\hat{W} = \frac{1}{k}\sum_{j=1}^{k}\hat{W}^{(j)}$.
Compute $\hat{\Lambda}$ with the singular value decomposition of $\hat{W} = \hat{Q}_1\hat{\Lambda}\hat{Q}_2^T$.
```

This concludes our subsection presenting this Bayesian GCCCA method. In the next section we present a similar procedure for the CMCCA model.

6.3 Outlook: The Multirank Likelihood for CMCCA

In this final section we will briefly discuss a Bayesian method for the CMCCA model which is also presented in (Bryan et al., 2024). We begin by the defining the notation and setup and then present the so-called multirank likelihood on which this method is based on.

Notation 6.3.1. In this section we assume the same setup as in Notation 5.4.7. Hence, we observe $Y^{(1)}, \dots, Y^{(n)}$ i.i.d. satisfying the cyclically monotone CCA, so that we have

$$\boldsymbol{Y} = \begin{pmatrix} Y_1^{(1)} & Y_2^{(1)} \\ \vdots & \vdots \\ Y_1^{(n)} & Y_2^{(n)} \end{pmatrix} = \begin{pmatrix} G_1(Z_1^{(1)}) & G_2(Z_2^{(1)}) \\ \vdots & \vdots \\ G_1(Z_1^{(n)}) & G_2(Z_2^{(n)}) \end{pmatrix}, \qquad \boldsymbol{Z} = \begin{pmatrix} Z_1^{(1)} & Z_2^{(1)} \\ \vdots & \vdots \\ Z_1^{(n)} & Z_2^{(n)} \end{pmatrix}$$

for cyclically monotone transformations G_1 and G_2. Further, we have $Z^{(1)}, \dots, Z^{(n)} \overset{iid}{\sim} \mathcal{N}_p(0, C)$, so that

$$C = \begin{pmatrix} I_{p_1} & W \\ W^T & I_{p_2} \end{pmatrix}$$

is positive definite and W has singular value decomposition $W = Q_1 \Lambda Q_2^T$.

As in Section 5.4, our goal is to estimate the CMCCA parameter W or equivalently Q_1, Q_2 and Λ. In particular, we do not need to know the cyclically monotone functions G_1 and G_2. Like in the previous section the estimation would be straightforward if we knew the latent $Z^{(i)}$'s, but we do not. In Section 5.4 we took on the problem by sampling from a grid and then use optimal transport, but this time we follow a Bayesian approach.

For stating the likelihood function, note that we have data $\boldsymbol{Y}$ and unknown parameters $Q_1, Q_2, \Lambda, G_1, G_2$, so that our likelihood would be given by

$$l(Q_1, Q_2, \Lambda) = \pi(\boldsymbol{Y} | Q_1, Q_2, \Lambda, G_1, G_2),$$

which is difficult to work with. Hence, we follow a similar approach as in the previous section by replacing the data $\boldsymbol{Y}$ by the occurrence of an event making the resulting likelihood independent of the nuisance parameters G_1 and G_2.

Again when we observe data $\boldsymbol{Y}$ we can learn some information on the structure of the $Z^{(i)}$'s. The starting point is the following: As $Y^{(1)}, \dots, Y^{(n)}$ is a sample from the CMCCA model, we must have $Y_j^{(i)} = G_j(Z_j^{(i)})$, where G_j is cyclically monotone. By the characterizations in Definition 4.1.6 we have

$$\sum_{i=1}^{n} \left\| Z_j^{(i)} - Y_j^{(i)} \right\|^2 = \min_{\pi \in S_n} \left\| Z_j^{(i)} - Y_{\pi(j)}^{(i)} \right\|^2 \qquad j = 1, 2.$$

More generally, $\boldsymbol{Z}$ must lie in the subset of $\mathbb{R}^{n \times p}$,

$$D \coloneqq \left\{ \mathbf{Z} : \{ (Z_j^{(1)}, Y_j^{(1)}), (Z_j^{(2)}, Y_j^{(2)}), \dots, (Z_j^{(n)}, Y_j^{(n)}) \} \text{ is cyclically monotone for } j = 1, 2 \right\}.$$

Now, we can take the occurrence of the event $\{\boldsymbol{Z} \in D\}$ as our data and obtain the so-called multirank likelihood

$$l(Q_1, Q_2, \Lambda) = \mathbb{P}(\boldsymbol{Z} \in D | Q_1, Q_2, \Lambda, G_1, G_2) = \mathbb{P}(\boldsymbol{Z} \in D | Q_1, Q_2, \Lambda).$$

We observe that the last equality holds since the distribution of $\boldsymbol{Z}$ is independent of G_1 and G_2. Hence, our likelihood function is independent of the nuisance parameters G_1 and G_2, as desired. Finally, one can specify a prior distribution for Q_1, Q_2 and Λ and obtain the posterior distribution

$$\pi(Q_1, Q_2, \Lambda | \boldsymbol{Z} \in D) \propto \mathbb{P}(\boldsymbol{Z} \in D | Q_1, Q_2, \Lambda) \pi(Q_1, Q_2, \Lambda).$$

We end the section at this point, as the next steps go beyond the scope of this thesis. The choice of priors on Q_1, Q_2 and Λ involves highly complex domains. Further, the resulting Gibbs Sampling scheme involves distributions which it is difficult to sample from, e.g. a matrix normal distribution truncated to a set with a cyclically monotone constraint. For interested readers we refer to (Bryan et al., 2024) for more information.

This finalizes the chapter on Bayesian methods for Canonical Correlation Analysis. In the next chapter we will conduct simulations to compare all five introduced methods.

7 Simulation

7.1 Setup

In chapter 5 we have introduced three different models for Canonical Correlation Analysis and presented one method for each model. The classical CCA method directly estimates the CCA parameter W by multiplying blocks of the empirical covariance matrix. In comparison, the Gaussian Copula CCA method calculates the Spearman's correlation coefficients of the sample which determine the entries of a correlation matrix in a first step. Finally, the Cyclically Monotone CCA method samples from a multivariate normal grid and transports the sample to it before estimating W. All three models are consistent provided their respective model assumptions hold.

Furthermore, we have provided two Bayesian estimation methods for the Gaussian Copula Model and the CMCCA Model, respectively, in chapter 6. In those one can sample from a Markov Chain which has the desired posterior distribution as stationary distribution if the corresponding model assumptions are satisfied. In summary, this makes five methods which have strong asymptotic results. But how do they perform in practice?

In this chapter we will make a simulation to answer this question. The main goal is to assess the performance of all methods in different scenarios which includes varying sample sizes, dimensions, and distribution of the samples. As the cyclically monotone CCA is our largest model containing both, the classical CCA and Gaussian Copula CCA, our samples will be distributed according to this model. Therefore, we will simulate an i.i.d. sample from

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} G_1(Z_1) \\ G_2(Z_2) \end{pmatrix} \qquad Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}_p \left(0, \begin{pmatrix} I_{p_1} & W \\ W^T & I_{p_2} \end{pmatrix} \right),$$

where $p_1 \leq p_2$, $p = p_1 + p_2$, W has singular value decomposition $Q_1 \Lambda Q_2^T$, and G_1 & G_2 are cyclically monotone transformations. These functions will be chosen based on Example 4.1.1. We will pick one transformation from each type, (i)-(v). Then, we will run all five presented methods for sample sizes increasing from 100 to 1000. This procedure will be repeated for the dimensions $p_1 = p_2 = 2, 3, 4$. Thereby, when we increase the dimension, we will try to change the cyclically monotone transformations as little as possible. The purpose behind this is to analyze the effect of sample size and dimension on the performance of each model.

In the simulations, we will estimate the CMCCA parameters W and Λ. Doing so we will use the loss functions

$$L_W(\hat{W}) = \frac{\|\hat{W} - W\|_F^2}{p_1 p_2}, \qquad \text{and} \qquad L_\Lambda(\hat{\Lambda}) = \frac{\|\hat{\Lambda} - \Lambda\|_F^2}{p_1},$$

where $\| \cdot \|_F$ is the Frobenius-norm of a matrix. I.e. for a matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, it holds

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}.$$

Our methods will have the following abbreviations in throughout this section:

- **CCA**: The classical CCA method from Section 5.2.
- **GCCCA**: The Gaussian Copula CCA method with Spearman's ρ from Section 5.3.
- **GCCCA-Bay**: The MCMC method for the GCCCA from Section 6.2.
- **CMCCA**: The plug-in method for the CMCCA model from Section 5.4
- **CMCCA-Bay**: The MCMC method for the CMCCA from Section 6.3.

As we did not present the estimation method for the CMCCA-Bay method, we refer to the cmcca package at https://github.com/j-g-b/cmcca. There, one can use the cmcca-mcmc() function to sample from a Markov Chain with the desired stationary distribution.

7.2 Models and Performance

7.2.1 Dimension 2

For the two dimensional case, we set

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}_4(0, \Sigma), \quad \Sigma = \begin{pmatrix} I_2 & W \\ W^T & I_2 \end{pmatrix}, \quad W = Q_1 \Lambda Q_2^T,$$

$$\Lambda = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}, \quad Q_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \quad Q_2 = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix},$$

and apply five different cyclically monotone transformations. As explained before, these will be based on Example 4.1.1. For each model we will calculate the loss for W and Λ in 100 replications for each sample size $n = 100, 250, 500, 1000$.

Model 1. The first model is a linear transformation as in Example 4.1.1 (i). Define

$$T : \mathbb{R}^2 \to \mathbb{R}^2, T(z) = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} T(Z_1) \\ T(Z_2) \end{pmatrix}.$$

Model 2. The second transformation applies univariate increasing functions to each marginal as in Example 4.1.1 (ii). Let

$$f : \mathbb{R} \to \mathbb{R}, f(x) = F_V^{-1}(\Phi(x)), \qquad G : \mathbb{R}^2 \to \mathbb{R}^2, G(z) = \begin{pmatrix} f(z_1) \\ f(z_2) \end{pmatrix}$$

where Φ is the distribution of a standard normal random variable, $V \sim \text{Exp}(1)$ is exponentially distributed, and F_V^{-1} the inverse of V's distribution function. Then, set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} G(Z_1) \\ G(Z_2) \end{pmatrix}.$$

Model 3. The next transformation is a combination of Example 4.1.1 (ii) and (v). Let G be as in Model 2. Define

$$\tilde{G} : \mathbb{R}^2 \to \mathbb{R}^2, \tilde{G}(z) = U^T G(Uz), \qquad U = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix},$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \tilde{G}(Z_1) \\ \tilde{G}(Z_2) \end{pmatrix}.$$

Model 4. In this model we transform the Z_j's by scaling them with increasing and non-negative functions of their norms as in Example 4.1.1 (iii). Let H be defined as

$$H : \mathbb{R}^2 \to \mathbb{R}^2, H(z) = \frac{\exp(0.1\|z\|)}{1 + \exp(0.1\|z\|)} \frac{z}{\|z\|}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} H(Z_1) \\ H(Z_2) \end{pmatrix}.$$

Model 5. The last transformation is as in Example 4.1.1 (iv). Define

$$J : \mathbb{R}^2 \to \mathbb{R}^2, J(x, y) = \begin{pmatrix} x^3/3 + x + y \\ y^3/3 + y + x \end{pmatrix}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} J(Z_1) \\ J(Z_2) \end{pmatrix}.$$

Performance. The following boxplots show the performance of 100 trials of each method in all models with sample sizes increasing from $n = 100$ to 1000.

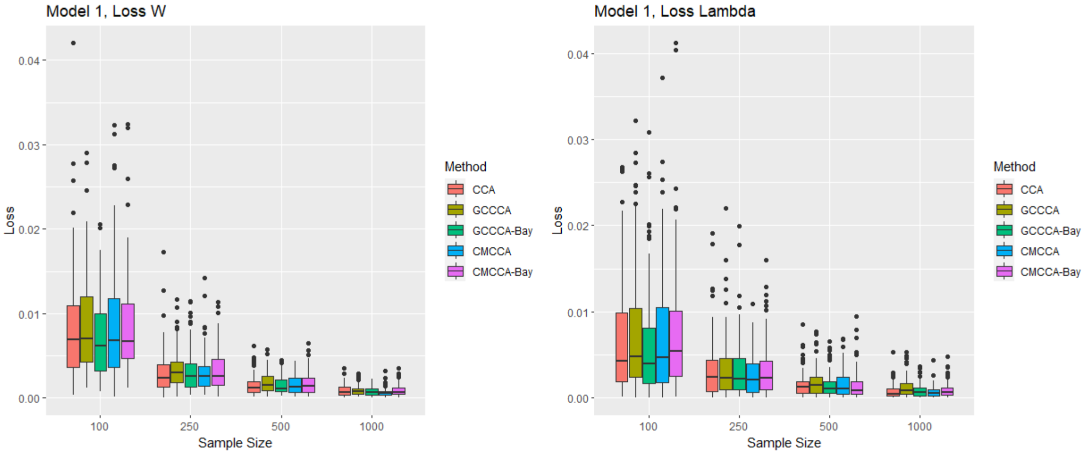

Figure 7.1: Performance in Model 1

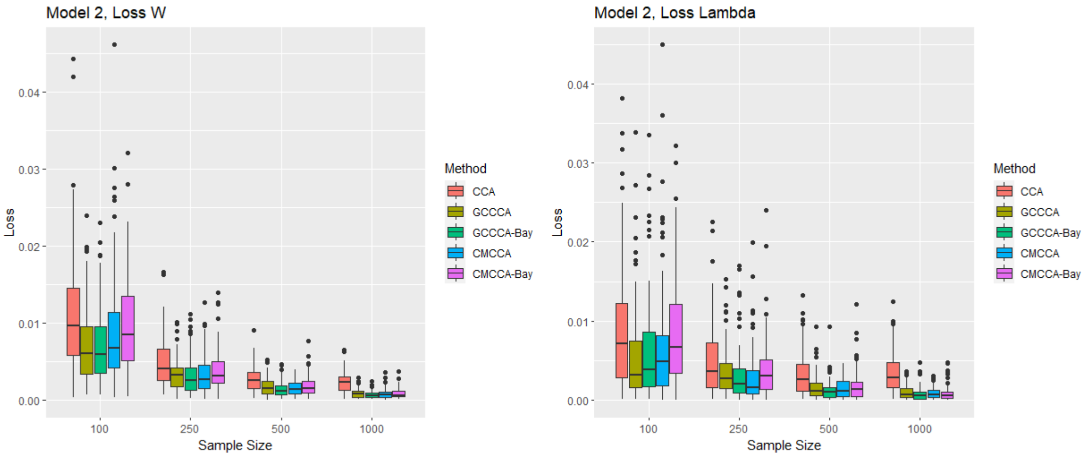

Figure 7.2: Performance in Model 2

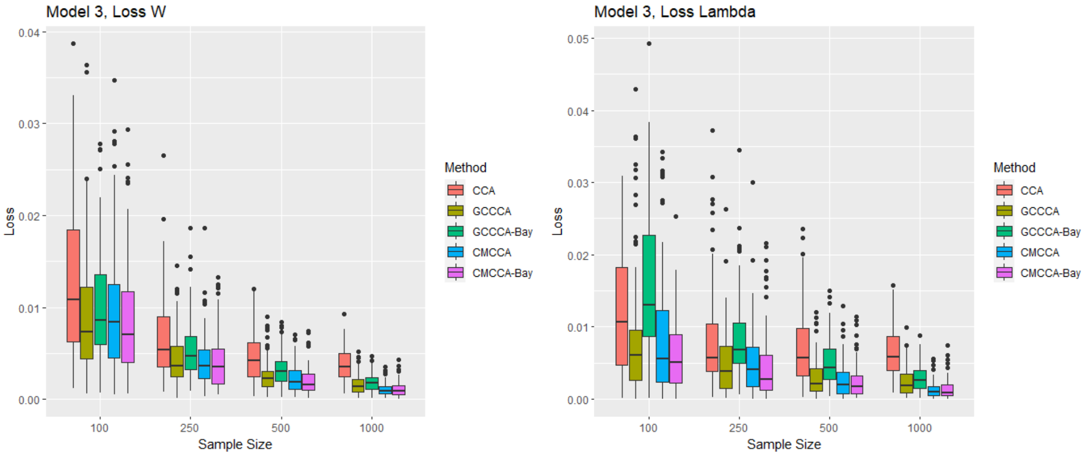

Figure 7.3: Performance in Model 3

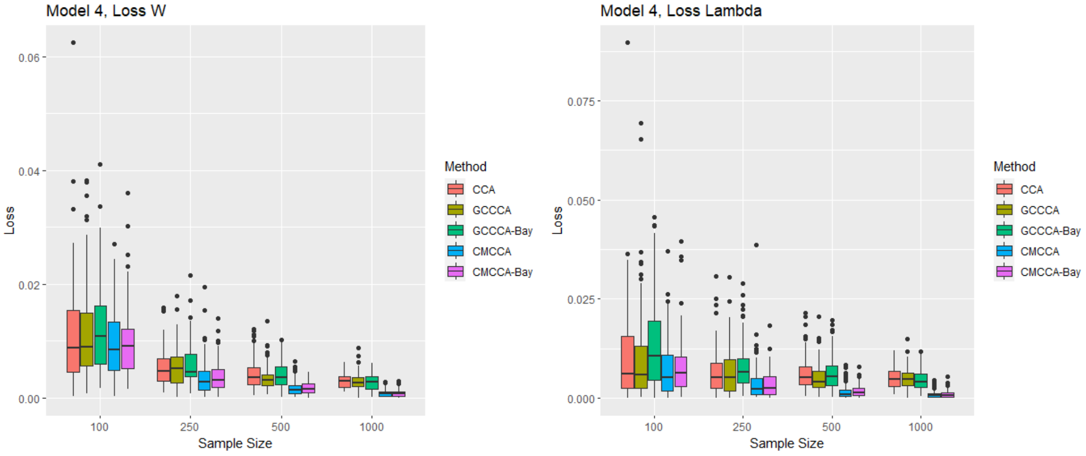

Figure 7.4: Performance in Model 4

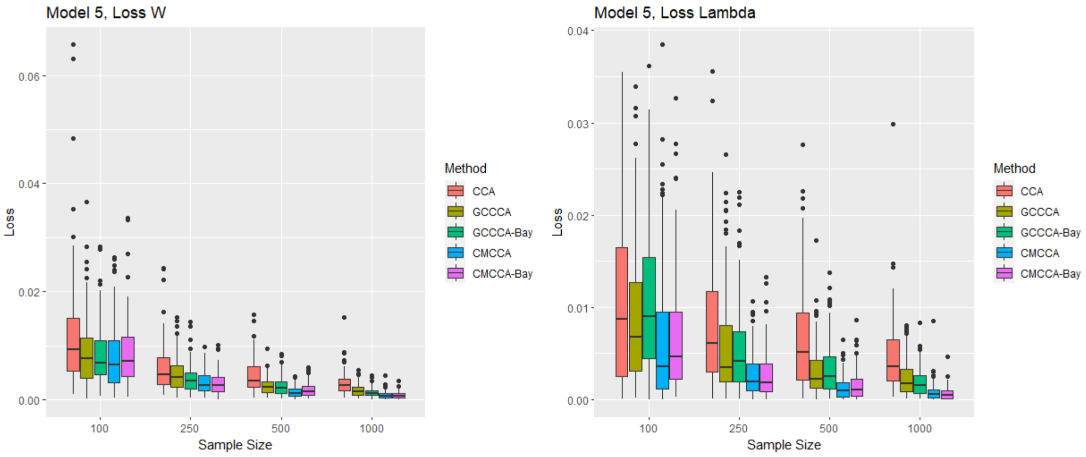

Figure 7.5: Performance in Model 5

Before we comment on the boxplots, let us address which model assumptions are satisfied in which model. As the first transformation is linear, Model 1 satisfies the CCA assumptions, and therefore, also the ones for the GCCCA. In Model 2, we do not have normally distributed data anymore, so that the classical CCA assumptions are violated. However, the type of transformation still lies in the GCCCA setup. Finally, one can check that Model 3-5 neither fulfill the classical CCA nor the GCCCA model assumptions. However, by our setup the CMCCA assumptions are satisfied in every model.

These observations are reflected in the boxplots. In figure 7.1, we see that all methods perform similarly in Model 1 with some minor fluctuations. In particular, the accuracy of each model improves with the sample size.

In figure 7.2 we observe a different result. While the other methods become more accurate with a higher sample size, the classical CCA fails to do so. This makes sense as its model assumptions are not satisfied and its consistency does not hold.

In Model 3, only the CMCCA assumptions are satisfied, so that the two corresponding methods perform best as seen in Figure 7.3. The classical CCA demonstrates the worst performance, while the GCCCA methods perform mediocrely. The distinction in accuracy of the five methods arises particularly with higher sample sizes.

In Model 4, the dominance of the CMCCA methods is clear. While the three other methods perform similar with no improvement in accuracy, the CMCCA methods perform significantly better.

Finally, Figure 7.5 is similar to Figure 7.3, with the CMCCA methods performing best, the GCCCA being mediocre and the classical CCA being the least accurate method.

In summary, the classical CCA method only performed well when its model assumptions were satisfied. The GCCCA methods were accurate in their models, but also moderate if their model assumptions were violated. Overall, the CMCCA methods performed best as they were accurate in every scenario. In particular, they beat the established methods in Model 3-5, and performed similarly in the other two.

Next, we consider the case $p_1 = p_2 = 3$.

7.2.2 Dimension 3

For the three-dimensional case, we set

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}_6(0, \Sigma), \quad \Sigma = \begin{pmatrix} I_2 & W \\ W^T & I_2 \end{pmatrix}, \quad W = Q_1 \Lambda Q_2^T,$$

$$\Lambda = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{6} \end{pmatrix}, \quad Q_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

and again apply five different cyclically monotone transformations. The functions will be the same or similar to the previous section, so that we can analyze the effect of the dimension on the performance. We will examine the accuracy for the sample sizes $n = 100, 250, 500, 1000$.

Model 1. We define

$$T : \mathbb{R}^3 \to \mathbb{R}^3, T(z) = \begin{pmatrix} 1 & 0.25 & 0 \\ 0.25 & 1 & 0.25 \\ 0 & 0.25 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} T(Z_1) \\ T(Z_2) \end{pmatrix}.$$

Model 2. Let

$$f : \mathbb{R} \to \mathbb{R}, f(x) = F_V^{-1}(\Phi(x)), \qquad G : \mathbb{R}^3 \to \mathbb{R}^3, G(z) = \begin{pmatrix} f(z_1) \\ f(z_2) \\ f(z_3) \end{pmatrix}$$

where Φ is the distribution of a standard normal random variable, $V \sim \text{Exp}(1)$ is exponentially distributed, and F_V^{-1} the inverse of V's distribution function. Then, set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} G(Z_1) \\ G(Z_2) \end{pmatrix}.$$

Model 3. Define

$$\tilde{G} : \mathbb{R}^3 \to \mathbb{R}^3, \tilde{G}(z) = U^T G(Uz), \qquad U = \begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix},$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \tilde{G}(Z_1) \\ \tilde{G}(Z_2) \end{pmatrix}.$$

Model 4. Let H be defined as

$$H : \mathbb{R}^3 \to \mathbb{R}^3, H(z) = \frac{\exp(0.1\|z\|)}{1+\exp(0.1\|z\|)} \frac{z}{\|z\|}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} H(Z_1) \\ H(Z_2) \end{pmatrix}.$$

Model 5. Define

$$J : \mathbb{R}^3 \to \mathbb{R}^3, J(x,y,z) = \begin{pmatrix} x^3/3 + 2x + y + z \\ y^3/3 + 2y + x + z \\ z^3/3 + 2z + x + y \end{pmatrix}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} J(Z_1) \\ J(Z_2) \end{pmatrix}.$$

Performance. The following boxplots show the performance of 100 trials of each method in all models with sample sizes increasing from $n = 100$ to 1000.

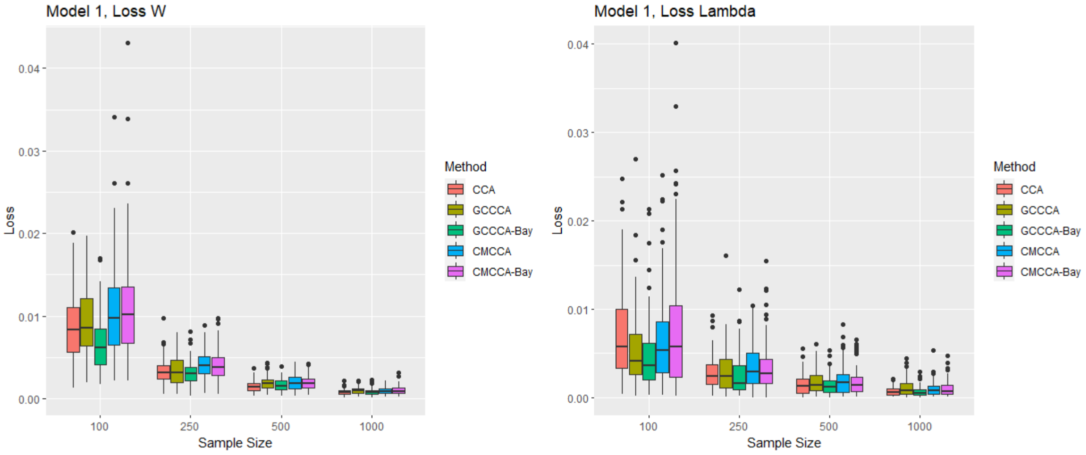

Figure 7.6: Performance in Model 1

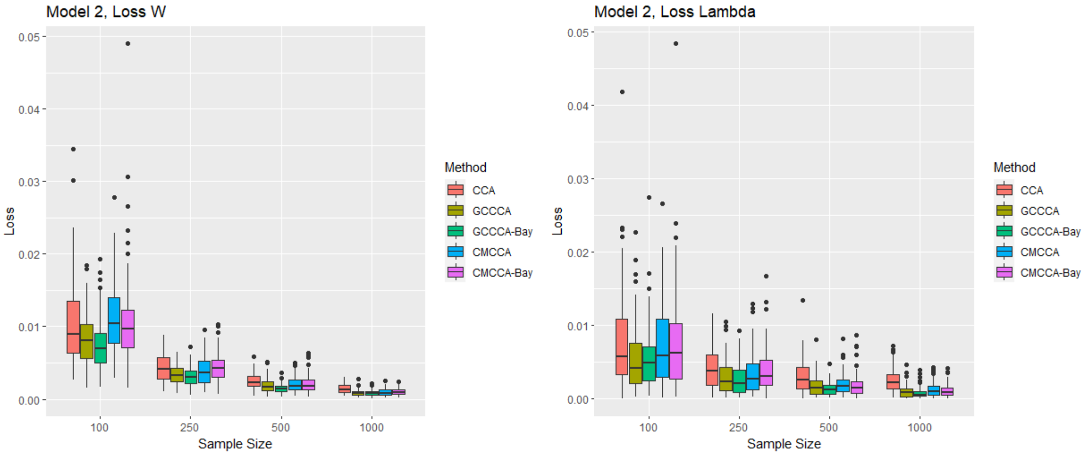

Figure 7.7: Performance in Model 2

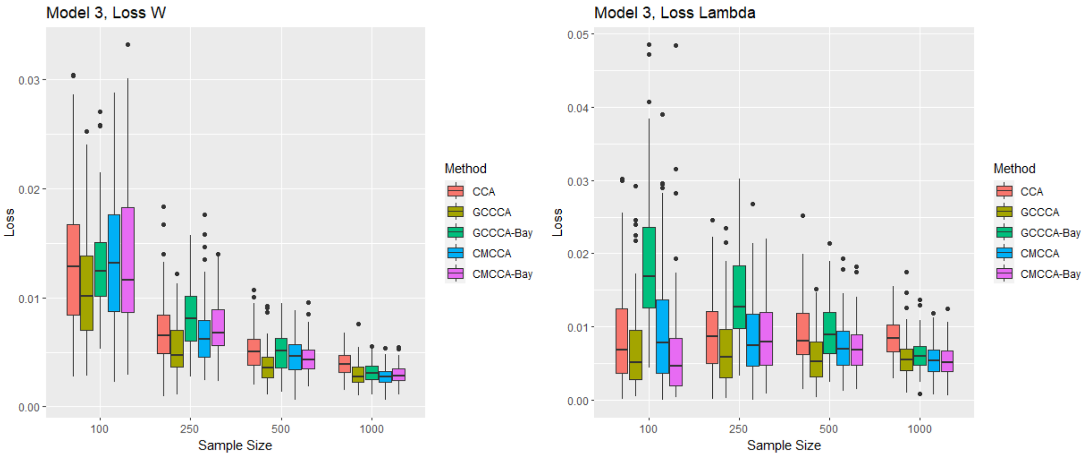

Figure 7.8: Performance in Model 3

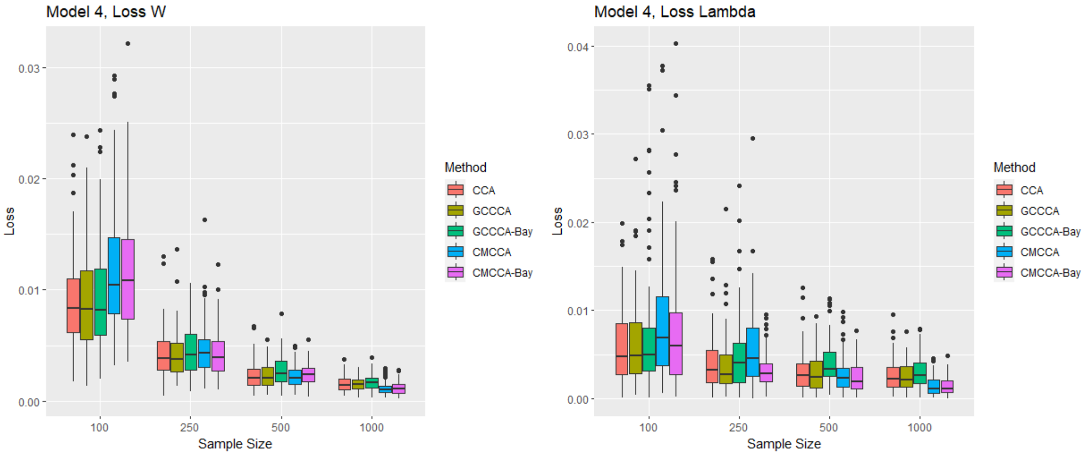

Figure 7.9: Performance in Model 4

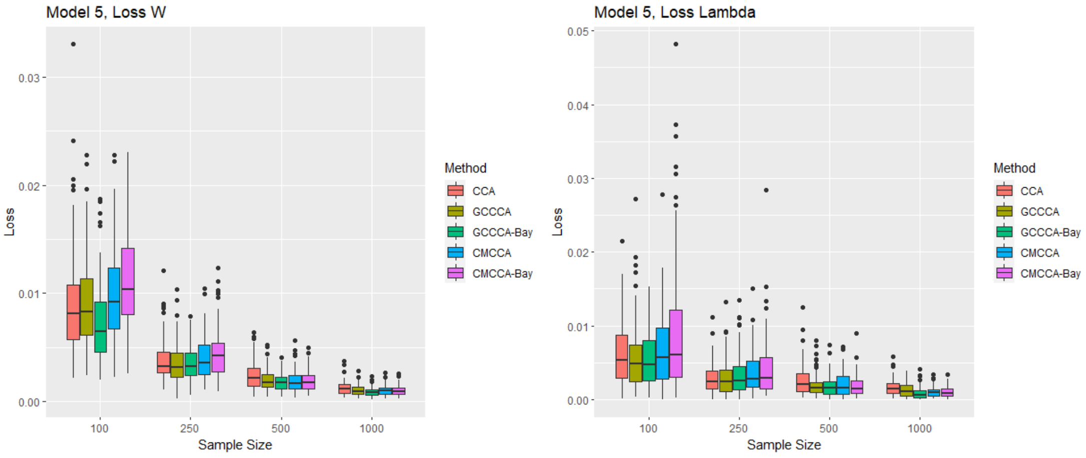

Figure 7.10: Performance in Model 5

As in the two-dimensional case, the classical CCA assumptions are satisfied only in Model

1. Further, the GCCCA assumptions are met in Model 1 and 2. Finally, the CMCCA setup is always satisfied.

In Model 1, we again observe that the accuracy for all methods improves with the sample size. In Figure 7.6 we can also see that the CMCCA estimators perform slightly worse for lower sample sizes, but are equally accurate in comparison to the others for a high n.

Figure 7.7 shows a similar result as Figure 7.2. All methods become more accurate asymptotically except the classical CCA. As in Model 1, the CMCCA methods perform slightly worse for lower sample sizes, but improve significantly asymptotically.

In Model 3, we observe that the classical CCA method performs worst while the GCCCA methods are more accurate although their model assumptions are violated, as well. Particularly, the GCCCA method using Spearman's ρ performs well in all sample sizes. However, asymptotically the CMCCA methods are slightly more accurate, particularly the plug-in method.

In Model 4, the classical CCA and GCCCA methods are misspecified, and they perform similarly. Although they become slightly more accurate for higher sample size they are not consistent asymptotically. In comparison, the CMCCA methods improve significantly for higher n and are consistent beating the other methods clearly.

Finally, Figure 7.10 shows that all methods perform similarly and become more accurate for higher sample sizes. Thereby, the GCCCA-Bay and CMCCA methods are slightly better than the other two.

Overall, the two CMCCA methods were dominant in this three-dimensional case. Although the difference in accuracy is not as big as in the two-dimensional case, the CMCCA methods performed best on average. In particular, they beat the other methods in Model 3-5. An exception is the GCCCA-Bay method in Model 5 which performed best there, but was less accurate in Model 3-4. In Model 1-2 the CMCCA methods were still able to reach a comparable performance in comparison to the correctly specified GCCCA methods. The classical CCA method performed worst as in the two-dimensional case. This is non-surprising as its assumptions were mostly misspecified.

Next, we consider the case $p_1 = p_2 = 4$.

7.2.3 Dimension 4

For the four-dimensional case, we set

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}_8(0, \Sigma), \quad \Sigma = \begin{pmatrix} I_2 & W \\ W^T & I_2 \end{pmatrix}, \quad W = Q_1 \Lambda Q_2^T,$$

$$\Lambda = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{3}{8} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{8} \end{pmatrix}, \quad Q_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix}, \quad Q_2 = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 \\ 0 & 0 & \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix},$$

and again apply five different cyclically monotone transformations. The functions will be the same or similar to the previous section. We will then examine the performance for sample sizes $n = 100, 250, 500, 1000, 2000$. Due to the higher dimension, we include the case $n = 2000$ this time.

Model 1. We define

$$T : \mathbb{R}^4 \to \mathbb{R}^4, T(z) = \begin{pmatrix} 1 & 0.25 & 0 & 0 \\ 0.25 & 1 & 0.25 & 0 \\ 0 & 0.25 & 1 & 0.25 \\ 0 & 0 & 0.25 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} T(Z_1) \\ T(Z_2) \end{pmatrix}.$$

Model 2. Let

$$f : \mathbb{R} \to \mathbb{R}, f(x) = F_V^{-1}(\Phi(x)), \qquad G : \mathbb{R}^3 \to \mathbb{R}^3, G(z) = \begin{pmatrix} f(z_1) \\ f(z_2) \\ f(z_3) \\ f(z_4) \end{pmatrix}$$

where Φ is the distribution of a standard normal random variable, $V \sim \mathrm{Exp}(1)$ is exponentially distributed, and F_V^{-1} the inverse of V's distribution function. Then, set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} G(Z_1) \\ G(Z_2) \end{pmatrix}.$$

Model 3. Define

$$\tilde{G} : \mathbb{R}^3 \to \mathbb{R}^3, \tilde{G}(z) = U^T G(Uz), \qquad U = \begin{pmatrix} 5 & 1 & 1 & 1 \\ 1 & 5 & 1 & 1 \\ 1 & 1 & 5 & 1 \\ 1 & 1 & 1 & 5 \end{pmatrix},$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \tilde{G}(Z_1) \\ \tilde{G}(Z_2) \end{pmatrix}.$$

Model 4. Let H be defined as

$$H : \mathbb{R}^4 \to \mathbb{R}^4, H(z) = \frac{\exp(0.1\|z\|)}{1 + \exp(0.1\|z\|)} \frac{z}{\|z\|}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} H(Z_1) \\ H(Z_2) \end{pmatrix}.$$

Model 5. Define

$$J : \mathbb{R}^4 \to \mathbb{R}^4, J(x, y, z, w) = \begin{pmatrix} x^3/3 + 3x + y + z + w \\ y^3/3 + 3y + x + z + w \\ z^3/3 + 3z + x + y + w \\ w^3/3 + 3w + x + y + z \end{pmatrix}$$

and set

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} J(Z_1) \\ J(Z_2) \end{pmatrix}.$$

Performance. The following boxplots show the performance of 100 trials of each method in all models with sample sizes increasing from $n = 100$ to 2000.

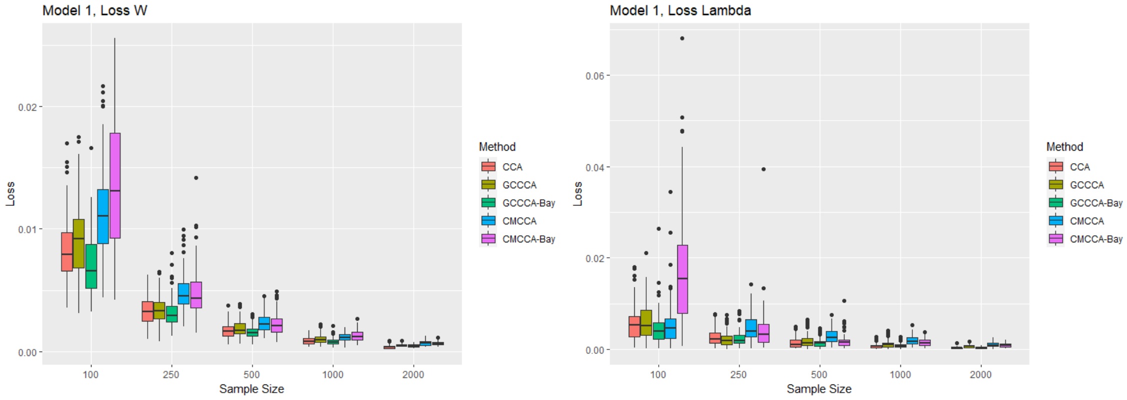

Figure 7.11: Performance in Model 1

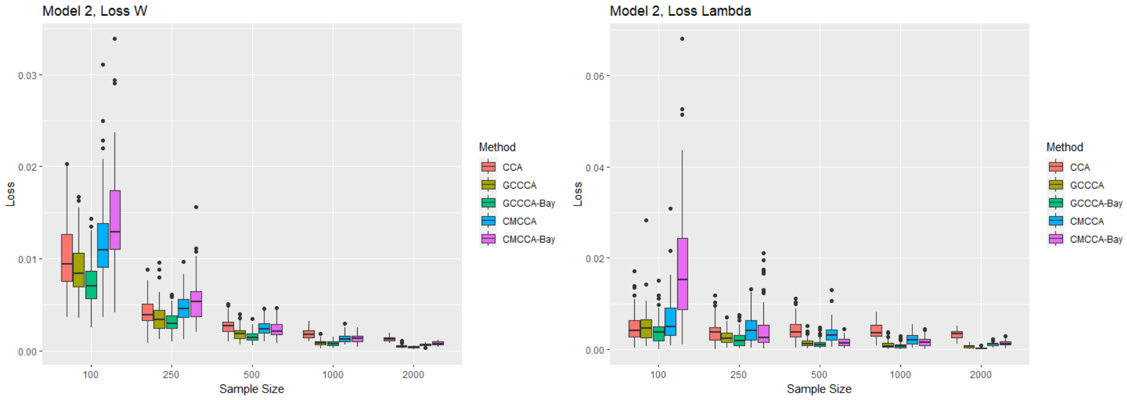

Figure 7.12: Performance in Model 2

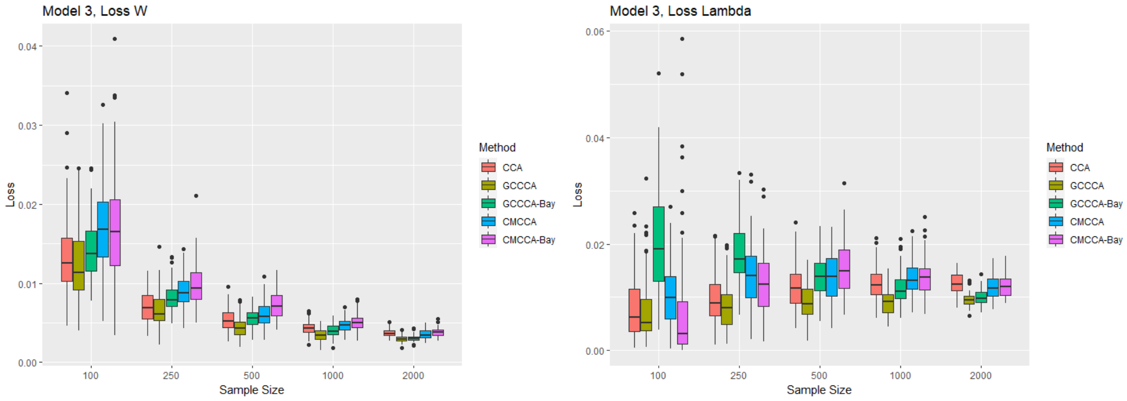

Figure 7.13: Performance in Model 3

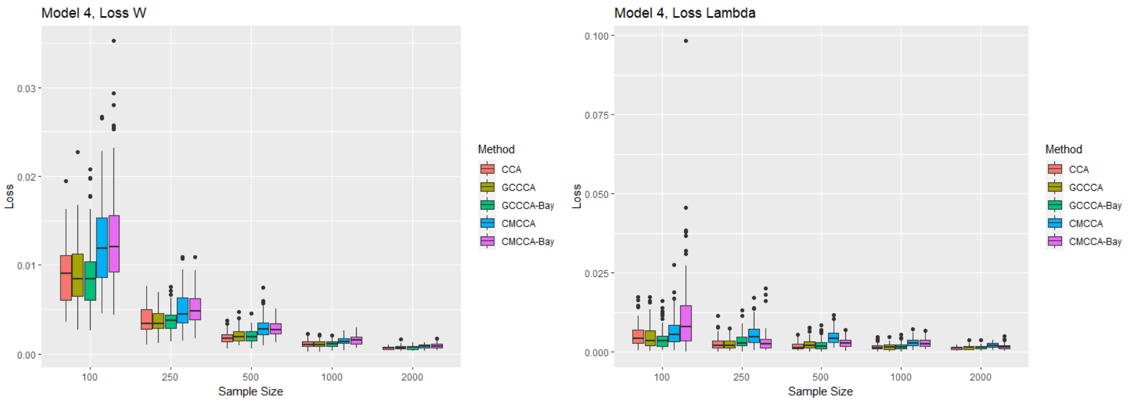

Figure 7.14: Performance in Model 4

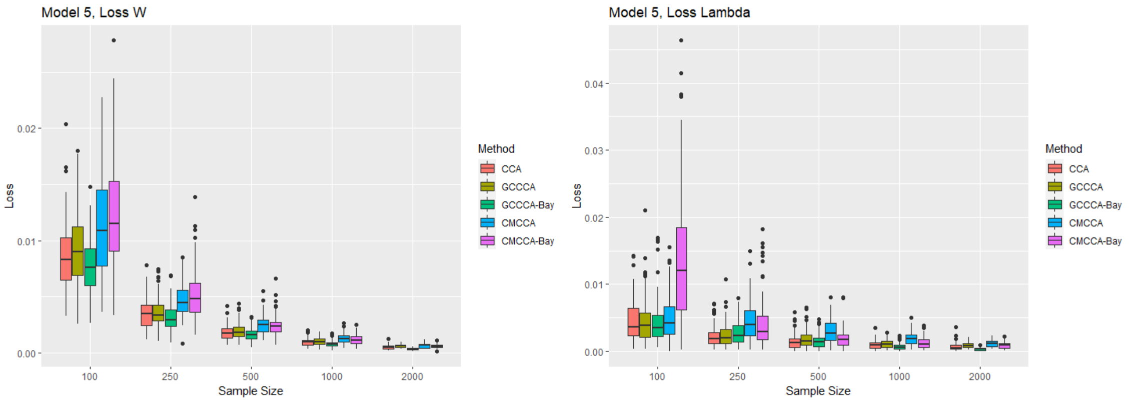

Figure 7.15: Performance in Model 5

As in the two-dimensional and three-dimensional case, the classical CCA setup is satisfied only in Model 1. Further, the GCCCA assumptions are met in Model 1 and 2. Finally, the CMCCA setup is always satisfied.

We observe in Model 1 that all methods are accurate and are consistent asymptotically. The CMCCA “take slightly longer” to converge, but still improve significantly for higher sample sizes.

In Model 2, the misspecified classical CCA estimator fails to improve asymptotically. However, the other methods are consistent with a slight edge for the GCCCA methods. The GCCCA methods outperform the CMCCA particularly for lower sample sizes while the CMCCA methods are almost as accurate as the GCCCA methods for $n = 2000$.

In Figure 7.13 we observe that the methods perform similarly asymptotically, whereby the GCCCA methods are slightly better. However, the CMCCA methods improve in accuracy for higher sample sizes as we have shown in the theory section. Hence, a $n > 2000$ simulation would be needed to observe their consistency.

Figure 7.14 and 7.15 are rather similar. In those the CMCCA methods are a bit less accurate than the other methods for lower sample sizes, but improve significantly asymptotically, so that all methods perform equally well for $n = 2000$.

7.3 Key Findings

Let us summarize the results of the simulation. As described in the setup our goal was to assess the performance of all five methods in different scenarios and analyze the effect of the sample size on the accuracy.

Classical CCA: Our simulations showed that the classical CCA method is inaccurate in many scenarios. Some examples include Figure 7.3, 7.5, 7.8 and 7.12 and other previous figures. These particularly show that the CCA may not perform well in different dimensions and sample sizes. It almost only displays accuracy if its assumptions are satisfies as in Figure 7.6 and 7.11.

GCCCA: The GCCCA method with Spearman's ρ performs significantly better. If its model setup is true, then it can outperform the other methods as in Figure 7.2 and 7.12. However, if it is misspecified it may still perform moderately which is an advantage that the CCA lacks. Examples are Figure 7.3, 7.8, 7.14 and 7.15. However, there are still scenarios, where it is inaccurate, e.g. in Figure 7.4 and 7.9.

GCCCA-Bay: The Bayesian GCCCA method performs similarly as the method with Spearman's ρ. It is particularly accurate when its model assumptions are satisfied as in Figure 7.6, 7.7. and 7.12. As the Spearman method it can also perform well when it is misspecified as in Figure 7.10, 7.13 and 7.15. However, it can also be inaccurate as in Figure 7.3. 7.4 and 7.9. Another of its disadvantages is a long runtime.

We also observe that the dimensions do not influence the performance of both GCCCA models that much. Higher sample sizes improve their accuracy, but the difference in improvement is not as high as for the CMCCA models.

CMCCA: The plug-in method for CMCCA has the strongest theoretical results and also performed very well in the simulations. It was particularly dominant in dimensions 2 and 3, where it often was the most accurate as in Figure 7.3-7.5 and 7.8-7-10. Hence, it could beat the other methods in many cases. We also observed that the higher the dimension gets the higher sample size is necessary in order to improve accuracy (e.g. compare Figure 7.4 and 7.9). In dimension 4, the CMCCA method could still keep up with the other methods for high sample sizes.

CMCCA-Bay: Its performance is very similar to the CMCCA plug-in method, particularly asymptotically. However, one disadvantage is its long running time.

8 Conclusion

Let us quickly summarize the most important aspects of the thesis. In chapter 4, we learned about optimal transport and rank statistics. These allow us to transport samples to a more convenient distribution to work with and use its theory. We particularly explored this in Section 4.2 and 5.4, where we applied them for defining the center-outward distribution function and the CMCCA model, respectively. The concept of rank statistics is promising and will remain a focal research topic in the next years.

Further, we got to know three different models for Canonical Correlation Analysis, one requiring normality, one assuming a Gaussian copula dependency structure and one allowing arbitrary joint marginals, but requiring their true rank statistics to be jointly normal. For all three models we have presented a consistent estimation method and two additional Bayesian methods. In the simulations we could compare their performance and observed the the CMCCA estimation methods which are rather new, were highly accurate particularly in lower dimensions and high sample sizes.

References

Denis Agniel and Tianxi Cai. Analysis of multiple diverse phenotypes via semiparametric canonical correlation analysis. *Biometrics*, 73(4):1254–1265, 2017. ISSN 0006-341X. doi: 10.1111/biom.12690. URL `https://doi.org/10.1111/biom.12690`.

A. D. Alexandroff. Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. *Leningrad State Univ. Annals [Uchenye Zapiski] Math. Ser.*, 6:3–35, 1939.

T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003. ISBN 0-471-36091-0.

Heinz Bauer. *Measure and integration theory*, volume 26 of *De Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 2001. ISBN 3-11-016719-0. doi: 10.1515/9783110866209. URL `https://doi.org/10.1515/9783110866209`. Translated from the German by Robert B. Burckel.

Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-19745-9. doi: 10.1002/9780470316962. URL `https://doi.org/10.1002/9780470316962`. A Wiley-Interscience Publication.

Martin Bilodeau and David Brenner. *Theory of multivariate statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 1999. ISBN 0-387-98739-8.

Jordan G. Bryan, Jonathan Niles-Weed, and Peter D. Hoff. The multirank likelihood for semiparametric canonical correlation analysis, 2024.

Yasuko Chikuse. Concentrated matrix Langevin distributions. *J. Multivariate Anal.*, 85(2):375–394, 2003. ISSN 0047-259X. doi: 10.1016/S0047-259X(02)00065-9. URL `https://doi.org/10.1016/S0047-259X(02)00065-9`.

Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation, 2019. URL `https://arxiv.org/abs/1909.08733`.

Nabarun Deb, Bhaswar B. Bhattacharya, and Bodhisattva Sen. Pitman efficiency lower bounds for multivariate distribution-free tests based on optimal transport, 2023.

Rick Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. ISBN 978-1-108-47368-2. doi: 10.1017/9781108591034. URL `https://doi.org/10.1017/9781108591034`. Fifth edition of [MR1068527].

William Feller. *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons, Inc., New York-London-Sydney, second edition, 1971.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2014. ISBN 978-1-4398-4095-5.

Hans-Otto Georgii. *Stochastics*. De Gruyter Textbook. Walter de Gruyter & Co., Berlin, extended edition, 2013. ISBN 978-3-11-029254-1; 978-3-11-029360-9. Introduction to probability and statistics, Translated from the German original [MR2397455] by Marcel Ortgiese, Ellen Baake and Georgii.

A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*. Chapman & Hall/CRC, Boca Raton, FL, 2000. ISBN 1-58488-046-5.

Marc Hallin et al. On distribution and quantile functions, ranks and signs in rd. *ECARES WP*, 2017.

Peter Hoff. Simulation of the matrix bingham-von mises-fisher distribution, with applications to multivariate and relational data, 2007a.

Peter D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.*, 1(1):265–283, 2007b. ISSN 1932-6157. doi: 10.1214/07-AOAS107. URL `https://doi.org/10.1214/07-AOAS107`.

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013. ISBN 978-0-521-54823-6.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN 00063444. URL `http://www.jstor.org/stable/2333955`.

Richard A. Johnson and Dean W. Wichern. *Applied multivariate statistical analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, sixth edition, 2007. ISBN 978-0-13-187715-3; 0-13-187715-1.

IOSR Journals. The correlation of personality traits and academic performance: A canonical correlation analysis. *IOSR Journal of Humanities and Social Science*, 26:141–149, 2021. URL `https://www.iosrjournals.org/iosr-jhss/papers/Vol.26-Issue10/Ser-11/E2610114149.pdf`.

C. G. Khatri and K. V. Mardia. The von Mises-Fisher matrix distribution in orientation statistics. *J. Roy. Statist. Soc. Ser. B*, 39(1):95–106, 1977. ISSN 0035-9246. URL `http://links.jstor.org/sici?sici=0035-9246(1977)39:1<95:TVMMDI>2.0.CO;2-1&origin=MSN`.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1037. URL `https://doi.org/10.1214/12-AOS1037`.

Kantilal Varichand Mardia, John T. Kent, and John M. Bibby. *Multivariate analysis*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, 1979. ISBN 0-12-471250-9.

Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323, 1995. ISSN 0012-7094. doi: 10.1215/S0012-7094-95-08013-2. URL `https://doi.org/10.1215/S0012-7094-95-08013-2`.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management.* Princeton Series in Finance. Princeton University Press, Princeton, NJ, 2005. ISBN 0-691-12255-5. Concepts, techniques and tools.

Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability.* Cambridge University Press, Cambridge, second edition, 2009. ISBN 978-0-521-73182-9. doi: 10.1017/CBO9780511626630. URL https://doi.org/10.1017/CBO9780511626630. With a prologue by Peter W. Glynn.

Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais.* De l'Imprimerie Royale, 1781.

Melvyn B. Nathanson and David A. Ross. Continuity of the roots of a polynomial, 2023.

R. T. Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific J. Math.*, 17:497–510, 1966. ISSN 0030-8730. URL http://projecteuclid.org/euclid.pjm/1102994514.

Hongjian Shi, Mathias Drton, and Fang Han. Distribution-free consistent independence tests via center-outward ranks and signs. *J. Amer. Statist. Assoc.*, 117(537):395–410, 2022. ISSN 0162-1459. doi: 10.1080/01621459.2020.1782223. URL https://doi.org/10.1080/01621459.2020.1782223.

Hongjian Shi, Mathias Drton, Marc Hallin, and Fang Han. Distribution-free tests of multivariate independence based on center-outward quadrant, spearman, kendall, and van der waerden statistics, 2024. URL https://arxiv.org/abs/2111.15567.

Galen R. Shorack. *Probability for statisticians.* Springer Texts in Statistics. Springer, Cham, second edition, 2017. ISBN 978-3-319-52206-7; 978-3-319-52207-4. doi: 10.1007/978-3-319-52207-4. URL https://doi.org/10.1007/978-3-319-52207-4.

Y. L. Tong. *The multivariate normal distribution.* Springer Series in Statistics. Springer-Verlag, New York, 1990. ISBN 0-387-97062-2. doi: 10.1007/978-1-4613-9655-0. URL https://doi.org/10.1007/978-1-4613-9655-0.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 1998. ISBN 0-521-49603-9; 0-521-78450-6. doi: 10.1017/CBO9780511802256. URL https://doi.org/10.1017/CBO9780511802256.

Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Old and new.

Dietrich von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, 15(2):97–109, 1988. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616090.

Thomas P. Wihler. On the Hölder continuity of matrix functions for normal matrices. *JIPAM. J. Inequal. Pure Appl. Math.*, 10(4):Article 91, 5, 2009. ISSN 1443-5756.

Alex Williams. A short introduction to optimal transport and wasserstein distance, 2020. URL `http://alexhwilliams.info/itsneuronalblog/`. Accessed: 2024-05-31.

David Williams. *Probability with martingales.* Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. ISBN 0-521-40455-X; 0-521-40605-6. doi: 10.1017/CBO9780511813658. URL `https://doi.org/10.1017/CBO9780511813658`.

Grace Yoon, Raymond J. Carroll, and Irina Gaynanova. Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 107(3):609–625, 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa007. URL `https://doi.org/10.1093/biomet/asaa007`.

Zhiyong Zhang. A note on wishart and inverse wishart priors for covariance matrix. *Journal of Behavioral Data Science*, 1, 01 2021. doi: 10.35566/jbds/v1n2/p2.

Roger S. Zoh, Bani Mallick, Ivan Ivanov, Veera Baladandayuthapani, Ganiraju Manyam, Robert S. Chapkin, Johanna W. Lampe, and Raymond J. Carroll. PCAN: probabilistic correlation analysis of two non-normal data sets. *Biometrics*, 72(4):1358–1368, 2016. ISSN 0006-341X. doi: 10.1111/biom.12516. URL `https://doi.org/10.1111/biom.12516`.