INAUGURAL - DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich-Mathematischen Gesamtfakultät

der

Ruprecht - Karls - Universität

Heidelberg

vorgelegt von

Diplom-Mathematiker Dominic Edelmann

aus

Heidelberg / Deutschland

Tag der mündlichen Prüfung:

Structures of Multivariate Dependence

Betreuer: **Prof. Dr. Rainer Dahlhaus**
Prof. Dr. Christoph Schnörr

Zusammenfassung

Die Untersuchung von Abhängigkeitsstrukturen spielt in der heutigen Statistik eine wichtige Rolle. Innerhalb der letzten Jahrzehnte wurden zahlreiche Abhängigkeitsmaße eingeführt, sowohl für univariate als auch für multivariate Zufallsvektoren. In dieser Thesis betrachten wir den Distanzkorrelationskoeffizienten, ein neues Abhängigkeitsmaß für Zufallsvariablen beliebiger Dimension, welches von Székely, Rizzo und Bakirov [102] and Székely und Rizzo [100] eingeführt wurde. Insbesondere definieren wir eine affin invariante Version der Distanzkorrelation und berechnen diesen Koeffizienten für zahlreiche Verteilungen: für die bivariate und die multivariate Normalverteilung, für die multivariate Laplaceverteilung und für bestimmte bivariate Gamma- und Poissonverteilungen. Darüber hinaus zeigen wir eine nützliche Reihenentwicklung der Distanzkovarianz für die Klasse der Lancasterverteilungen auf und leiten eine Verallgemeinerung eines Integrals her, welches in der Theorie der Distanzkorrelation eine fundamentale Rolle spielt.

Ferner untersuchen wir eine Problemstellung zum Clustern von Zufallsvariablen, welches in Gaußschen graphischen Modellen mit niederem Rang auftritt. Im Falle fester Stichprobengrößen stellen wir fest, dass dieses Problem mathematisch equivalent zum Problem des Clustern von Daten in unabhängige Unterräume ist. In der asymptotischen Situation leiten wir einen Schätzer her, welcher im Falle verrauschter Daten konsistent die Clusterstruktur erfasst.

Abstract

The investigation of dependence structures plays a major role in contemporary statistics. During the last decades, numerous dependence measures for both univariate and multivariate random variables have been established. In this thesis, we study the distance correlation coefficient, a novel measure of dependence for random vectors of arbitrary dimension, which has been introduced by Székely, Rizzo and Bakirov [102] and Székely and Rizzo [100]. In particular, we define an affinely invariant version of distance correlation and calculate this coefficient for numerous distributions: for the bivariate and the multivariate normal distribution, for the multivariate Laplace and for certain bivariate gamma and Poisson distributions. Moreover, we present a useful series representation of distance covariance for the class of Lancaster distributions and derive a generalization of an integral, which plays a fundamental role in the theory of distance correlation.

We further investigate a variable clustering problem, which arises in low rank Gaussian graphical models. In the case of fixed sample size, we discover that this problem is mathematically equivalent to the subspace clustering problem of data for independent subspaces. In the asymptotic setting, we derive an estimator, which consistently recovers the cluster structure in the case of noisy data.

Acknowledgments

First of all, I want to thank Rainer Dahlhaus for his support throughout the last years. I got to know him not only as an excellent researcher and teacher, but also as a circumspect group leader who always backs his students and staff. Second, I would like to thank Christoph Schnörr for fruitful discussions and for giving me a different view on my work. Amongst other things, he called my attention to literature outside of the field of mathematical statistics; this literature considerably affected this thesis.

The investigation of the distance correlation coefficient plays an important role in this work. I want to thank Tilmann Gneiting for introducing me into this topic and being an amazing coauthor. I further thank Donald Richards and Mercedes Richards for great conversations and for teaching me a lot about mathematics, Jamaica and the financial world. I hereby promise to read "The intelligent investor" as soon as possible. Johannes Dueck has been my office mate for the last six(!) years. Although we often have lively discussion, sometimes even severe disputes about research, soccer and everything else, he is always there when you need him. Thank you for being an awesome friend.

There are a lot of great people at the mathematical institutes and beyond, who supported me since I entered the university. I would like to thank Marion Münster, Dagmar Neubauer, Elke Carlow, Evelyn Wilhelm, Barbara Werner, Sebastian Schweer, Peter Büermann, Roman Safreider, Stefan Richter, Jochen Fiedler, Fabian Rathke, Fabian Bachl, Bernhard Schmitzer, Markus Speth, Jörg Kappes, Florian Becker, Bernhard Kausler, Eva-Maria Didden, Sophon Tunyavetchakit, Cornelia Wichelhaus, Matthias Katzfuß, Stephan Hemri, Kira Feldmann, Mark Podolskij, Michael Scheuerer, Jens Kastendieck, Maike Frank, Torsten Schweiger, Robert Dalitz, Matthias Klinger, Matthias Maier, Jan Mewes, Manuel Kudruss, Katharina Beuke, Carolin Margraf, Munir Hiabu, Axel Bücher and many other colleagues in Heidelberg.

I had the great good fortune to grow up in such a large family that mentioning all the names of my siblings, cousins and other relatives would definitely go beyond the scope of these acknowledgments. Still, I would like to particularly thank my parents Alfred and Barbara for the best education and excellent medical support.

Funding by the *Deutsche Forschungsgemeinschaft* (German Research Foundation) within the programme "Spatio/Temporal Graphical Models and Applications in Image Analysis," grant GRK 1653, is also gratefully acknowledged.

Contents

Chapter 1

Introduction

One of the essential problems in statistics is the investigation of the dependence structure between random variables. The mathematical study of these dependencies goes back at least to Gauss' *Theoria combinationis observationum erroribus minimis obnoxiae* [28] and the theoretical literature on this topic, which has emerged since then, is immense. The profound statistical analysis of the dependencies between random quantities of any kind is indispensable in all contemporary nature and social sciences; even outside the world of research, some of the ideas originated from this field are ubiquitous. The undoubtedly most celebrated concept in daily life is the notion of *correlation*, which alone brings up nearly one hundred million search results on Google.

The word correlation, which is colloquially often used as a synonym for *Pearson correlation*, is actually used to describe various normalized dependence measures. These coefficients attempt to quantify the strength of dependence between two random variables via one single number (usually in the interval $([-1, 1])$. While, in general, this single number is naturally not sufficient to express the potentially elaborate dependence structure between two random variables, it is mostly easy to interpret and estimate, which predestine those coefficients for the use in practice. During the last decades, a vast amount for dependence measures between random variables and random vectors have been proposed. Let us mention the Pearson correlation coefficient, Spearman's rank correlation coefficient [94], Goodman's and Kruskal's gamma [32] and the maximum information coefficient [77] as examples for measures in bivariate analysis. For measures between multivariate distributions, we adduce the canonical correlation coefficient [39] and the total correlation [111].

Each of the above mentioned coefficients has its advantages and disadvantages and it is not trivial to decide which measure to choose in a certain application. In his 1959 paper, Rényi [76] postulates a set of seven properties, which - according to him - a "natural" measure of dependence should fulfill.

(i) $\delta(X,Y)$ *is defined for any pair of random variables* X *and* Y*, neither of them being constant with probability* 1*.*

(ii) $\delta(X,Y) = \delta(Y,X)$*.*

(iii) $0 \leq \delta(X,Y) \leq 1$*.*

(iv) $\delta(X,Y) = 0$ *if and only if* X *and* Y *are independent.*

(v) $\delta(X,Y) = 1$ *if either* $X = g(Y)$ *or* $Y = f(X)$ *where* g *and* f *are Borel-measurable functions.*

(vi) If $f : \mathbb{R} \to \mathbb{R}$ *and* $g : \mathbb{R} \to \mathbb{R}$ *are bijective functions,* $\delta(f(X), g(Y)) = \delta(X,Y)$*.*

(vii) If the joint distribution of X *and* Y *is normal, then* $\delta(X,Y) = |\rho(X,Y)|$*.*

He further shows that the maximal correlation coefficient

$$m(X,Y) = \sup_{f,g \text{ Borel-measurable}} \rho_{Pearson}(f(X), g(Y)),$$

introduced by Gebelein in 1941 [29] satisfies all of these postulates. Yet, while every item in the preceding list obviously represent a desirable property of a dependence measure, the maximal correlation coefficient suffers from drawbacks in other respects. In particular, both sample and population measure are nontrivial to determine, moreover the calculation of the sample measures in practice is computationally hard.

Merely half a decade ago, Székely, Rizzo and Bakirov [102] and Székely and Rizzo [100] introduced the *distance correlation* as a new measure of dependence. $\mathcal{V}(X,Y)$, the *distance covariance* between X and Y is defined to be the positive square-root of $\mathcal{V}^2(X,Y)$ with

$$\mathcal{V}^2(X,Y) \propto \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(s,t) - f_X(s) f_Y(t)|^2}{\|s\|^{p+1} \, \|t\|^{q+1}} \, \mathrm{d}s \, \mathrm{d}t, \tag{1.0.1}$$

where $f_{X,Y}$ is the joint characteristic function of (X,Y), and $f_X(s) = f_{X,Y}(s,0)$ and $f_Y(t) = f_{X,Y}(0,t)$ are the corresponding marginal characteristic functions. Then the *distance correlation coefficient* between X and Y is given by

$$\mathcal{R}(X,Y) = \frac{\mathcal{V}(X,Y)}{\sqrt{\mathcal{V}(X,X)\mathcal{V}(Y,Y)}}. \tag{1.0.2}$$

Since the characteristic function $f_{X,Y}$ factorizes only in the case of independence, it can be easily seen, that - as the maximal correlation coefficient - *the distance correlation*

is 0 *if and only if* X *and* Y *are independent*. Moreover, different to the Pearson correlation or the maximal correlation coefficient, it applies to random vectors of arbitrary dimensions, rather than to univariate quantities only. Finally, and most importantly, its sample measure is astonishingly simple to define and can be computed in reasonable time. To evaluate it, we find the pairwise distances between the sample values for the first variable, and center the resulting distance matrix; then do the same for the second variable. The square of the sample distance covariance equals the average entry in the componentwise or Schur product of the two centered distance matrices. Given the theoretical appeal of the population quantity, and the striking simplicity of the sample version, it is not surprising that the distance covariance is experiencing a wealth of applications, despite having been introduced only a few years ago. As examples of the ubiquity of distance correlation methods in practice, we note the results on large astrophysical databases [79], on familial relationships and mortality [53] and long-range concerted motion in proteins [82].

While its sample measure is both easy to explain and compute, the calculation of the population distance correlation coefficients remains an intractable problem generally. For half a decade, Székely's result on the distance correlation for the bivariate normal distribution [102] has been the only success in that direction. Hence, the state of distance correlation theory until then that the empirical coefficients could be calculated readily but their population counterparts were unknown, generally. On being given random vectors X and Y, the fundamental obstacle in calculating the population distance correlation coefficient (1.0.2) is the computation of the singular integral (1.0.1). In particular, the singular nature of the integrand precludes evaluation of the integral by expanding the numerator, $|f_{X,Y}(s,t) - f_X(s)\, f_Y(t)|^2$, and subsequent term-by-term integration of each of the resulting three terms. The first part (Chapters 3-5) of this work is dedicated to novel approaches to tackle these analytical problems. In particular, we will derive the distance correlation coefficients for several multivariate distributions.

The second part of this work (Chapter 6) deals with dependence structures in a different way. In particular, we will investigate systems, where groups of random variables are *linearly dependent*, i.e. any of these random variables can be exactly or approximately expressed by a linear combination of the other variables in that group. This phenomenon, which is denoted by the terms *collinearity* or *multicollinearity*, leads to difficulties in many statistical problems. It classically arises as a problem in multiple linear regression [113]. Aside other issues, it leads to an ill-conditioned (or even non-defined) inverse covariance matrix of the predictor variables and hence often leads to a poor estimate of the regression coefficients.

The motivation for the clustering task, we will consider in Chapter 6, originates from

the field of Gaussian graphical models (GGMs). Gaussian graphical models (GGMs) [52, 62], also referred to as covariance selection models, provide a helpful framework to explore the dependence structure of multivariate Gaussian data. First developed by Dempster in 1972 [15], they recently became increasingly popular due to their importance for the analysis of high-dimensional data. In a GGM, the dependence structure of a p-variate normally distributed random vector $\mathbf{X} = (X_1, \ldots, X_p)^t$ is represented by a graph $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, p\}$ corresponds to the p the univariate random variables $X_1, \ldots, X_p$ and $\mathcal{E}$ is a set of edges between these random variables expressing the partial correlation structure. In particular, the set $\mathcal{E}$ will contain the edge (i, j) if and only if the partial correlation between the corresponding random variables X_i and X_j is not zero. In applications, these edges are often interpreted as "direct" connections, since the dependency between two variables connected by an edge cannot be fully explained by the other variables in the model.

In particular in highdimensional graphical models, it seems unavoidable to find groups of highly correlated genes or exact linear dependencies between random variables [88, 104]. But also in applications of low dimensional graphical models, collinearity can be frequently observed. In their paper *Graphical models for multivariate time series from intensive care monitoring* [27], Gather, et al. investigate the dependencies between ten time series representing physiological variables from intensive care monitoring. For their analysis, they apply graphical interaction models for multivariate time series [13], which represents an extension of the concept of GGMs to multivariate time series. Already for these ten time series, Gather, et al. find highly associated groups and need to deal with the problem of collinearity.

The issues of collinear groups of random variables in graphical models are twofold. On one hand, collinearity reduces the accuracy of the estimation in the model as noted by Bühlmann et al. [6] and Reid et al. [74]. On the other hand, the interpretation of the edges in Graphical Model as direct associations is questionable, since the partial correlations between groups of collinear groups of random variables are virtually 0, even if the corresponding variables are highly correlated. For instance, Gather et al. [27] denote , that they often *"...cannot identify known associations when a group of variables is included which are only slightly different representations of the same physiological process..."*. In Chapter 6, we attempt to make a first step towards solving this problem.

1.1 Related Literature

1.1.1 Distance Correlation

Székely, Rizzo and Bakirov [102] and Székely and Rizzo [100], in two seminal papers, introduced the distance covariance and distance correlation as powerful measures of dependence. In later papers, Rizzo and Székely [80, 81] and Székely and Rizzo [98, 96, 97] gave applications of the distance correlation concept to several problems in mathematical statistics. In recent years, there have appeared an enormous number of papers in which the distance correlation coefficient has been applied to many fields. In particular, the concept of distance covariance has been extended to abstract metric spaces [64] and has been related to machine learning [90]; and there have been applications to detecting associations in large astrophysical databases [66, 79] and to measuring nonlinear dependence in time series data [116]. We refer to the introduction of this thesis for further literature and details on the history of distance correlation. For a mathematical review of the central theorems of distance correlation, see section 2.4.

1.1.2 Variable Clustering

Clustering refers to the task of partitioning given objects (such as data points or random variables) into groups (or clusters), such that the objects in a group share certain similarities. While there is a vast literature for data clustering, i.e. the clustering of data points (see [43] for a review), the literature for variable clustering is comparatively small. However, there are many problems in statistics, where the clustering of random variables can be beneficial.

Particularly in applications, where dimension reduction is needed, variable clustering techniques possess certain advantages compared to classical dimension reduction techniques such as principal component analysis (PCA). While PCA is known to achieve effective dimension reduction, the interpretation of the obtained factors can often be difficult in practice, since these factors are typically functions of all random variables under consideration (see e.g. [107, 83]). Variable Clustering, on the other hand, divides the random variables into disjoint groups. Principal component analysis of these disjoint clusters then yields factors with disjoint loadings enabling more facile interpretation [61]. A similar objective was recently achieved by the celebrated sparse PCA approaches [14, 91]. Moreover variable clustering techniques can be useful to detect structural characteristics of the random variables under consideration, e.g. to find groups which are highly related or contribute to the same functional system. Examples include gene pathway analysis [22, 117] or detection of functional regions of the brain using fMRI data [106]. Finally, variable clustering has been applied to tackle the problem of multicollinearity in regression. In particular, it has been recently proposed

to combine variable clustering of the predictor variables and subsequent estimation of the regression coefficients via group lasso [6] or sparse regression using cluster prototypes [74]. Despite the substantial demand for variable clustering in applications, the methodological literature on techniques for this task is astonishingly small. Maybe, this is best reflected by the fact that many applied scientists refer to the VARCLUS procedure [84] contained in the software SAS, an ad-hoc method, for which (to the best knowledge of the author) no theoretical guarantees are known.

Most variable clustering techniques considered in literature are hierarchical approaches based on some kind of similarity matrix, such as correlations [22, 42], partial correlations [109] or mutual information [51] between the random variables. While these methods lump together random variables which are similar in some bivariate sense, these methods do not consider relations involving more than two covariates. In particular, hierarchical clustering based on the standard or squared correlation coefficient is not effective for attacking the problem of collinearity. Recently, methods taking more complex dependencies into account have been suggested. Bühlmann et al. [6] perform hierarchical clustering using the canonical correlation coefficient, while Ferenci and Kovács [25] exploit total correlation which represents a generalization of the mutual information coefficient to random vectors.

The matroid approach is a particularly interesting procedure, which goes back to an idea by Greene [35, 36] and has lately been considered by Woolston [114] in his PhD thesis. He first determines the intrinsic rank of every possible subset of covariates using some dependency criterion such as the variance inflation factor (VIF) or the minimal eigenvalue of the covariance matrix. A subset of random variables with intrinsic dimension j is then named a rank-j flat, if we are unable to add another covariate to the subsets without increasing its rank. Hence, the rank-j flats are the maximal subsets which can be represented by a j-dimensional projection. The rank-j flats yield several possibilities to cluster the data into dependent subsets. For example every random variable which is assigned to more than one flat could be clustered into the flat with maximum or minimum rank. As Woolston remarks, the advantage of the matroid approach compared to hierarchical procedures is, that it *"seeks not only to identify 1-dimensional clusters of mutually correlated variable, but also higher dimensional near dependencies in which collections of the observed variables are identified as falling close to lower dimensional subspaces"*. The drawback lies of course in its combinatorial nature; for a p-dimensional data set, it requires to determine the intrinsic rank of $2^p - 1$ subsets, which are already more than a million possibilities for $p = 20$.

1.1.3 Subspace Clustering

The approach, which we are going to pursue in Chapter 6 is highly related to subspace clustering, a method to categorize data being intensely investigated during the past two decades. In the following, we introduce the reader to the problem of subspace clustering and present two state-of-the-art methods. For a detailed overview over methods from the machine learning and computer vision communities, see [108], for methods from the data mining community, see [71].

Although many of the data available in e.g. computer vision are high-dimensional, it can often be observed that the data lies in lower-dimensional structures. While traditional dimension reduction techniques such as PCA aim at finding *one low-dimensional subspace* to fit the data, the intrinsic assumption of subspace clustering is, that we observe data points which are drawn from from a *union of subspaces* of lower dimension. More specifically, consider we have given k linear subspaces of $\mathbb{R}^p$ $S_1, \dots, S_k$ of respective dimensions $d_1, \dots, d_k$ as well as n samples $Y_1, \dots, Y_n \in \mathbb{R}^p$ which can be organized in a data matrix $\boldsymbol{Y} = [Y_1, \dots, Y_n] \in \mathbb{R}^{p \times n}$. The underlying assumption of subspace clustering is, that for each $i \in \{1, \dots, n\}$, there is a fraction of the data points, which lie in S_i. Hence, there is a subset of indices $C_i \subset \{1, \dots, n\}$, such that for each $l \in C_i$, it holds $Y_l \in S_i$. The aim of subspace segmentation is now, given the data $\boldsymbol{Y}$, find the number of subspaces k as well as their dimension d_i and recover the affiliation of the data points to their respective subspaces (i.e. the index sets C_i). To enable the solution of this problem, one naturally has to impose some restrictions on the subspaces, usually either disjointness of the subspaces (i.e. $S_i \cap S_j = \emptyset$ for $i \neq j$) or independence of the subspaces, i.e. $S_i \cap \bigoplus_{j \neq i} S_j = \emptyset$ for all i. It is apparent that independence of the subspaces implies disjointness of the subspaces, hence the latter assumption is stronger than the first one.

A strategy to solve subspace clustering, which is particularly interesting for our purposes has been proposed by Costeira and Kanade [12]. They consider a rank r skinny SVD of $\boldsymbol{Y} = U \Lambda V^t$, where $r = \sum_{i=1}^{k} d_i$. They then suggest composing the orthogonal projection matrix on the rows (the so-called SIM or shape iteration matrix) of $\boldsymbol{Y}$, i.e.

$$Q = VV^t \in \mathbb{R}^{n \times n}.$$

It can now be shown [48, 63], that if the subspaces $S_1, \dots, S_k$ are independent, it holds

$$Q_{lm} = 0 \text{ if } Y_l \text{ and } Y_m \text{ are in different subspaces,}$$

where Q_{lm} denotes the (l, m)-th entry of Q. Hence, in the absence of noise Q can be directly used to obtain the segmentation of the data into their respective subspaces.

For a data matrix which is contaminated by errors, such as e.g. noise or outliers, Liu et al. [63] recently proposed a method closely related to [12]. In particular, they assume that the observed data is given by

$$\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{E},$$

where $\boldsymbol{Y}$ is as described above and $\boldsymbol{E}$ is an error matrix of some kind. As before they aim at solving the subspace segmentation problem or equivalently finding the true SIM Q. For this purpose, they consider the rank minimization problem

$$\min_{P,E} rank(P) + \lambda \|\boldsymbol{E}\|_{2,1}, \quad \text{s.t.} \quad \boldsymbol{Z} = \boldsymbol{Z}P + \boldsymbol{E},$$

where $\|E\|_{2,1} = \sum_{l=1}^{n} \sqrt{\sum_{m=1}^{n} |E_{lm}|^2}$ is the $\ell_{2,1}$-norm. Since this problem is NP-hard, they replace the rank function by the nuclear norm, resulting in the following convex optimization problem:

$$\min_{P,E} \|P\|_\star + \lambda \|\boldsymbol{E}\|_{2,1}, \quad \text{s.t.} \quad \boldsymbol{Z} = \boldsymbol{Z}P + \boldsymbol{E}.$$

It can be proven, that in the case of noncontaminated data (e.g. $\boldsymbol{E} = 0$), solving this problem exactly recovers the SIM, i.e. $P^\star = Q$. Moreover it can handle a fair amount of outliers and sample-specific corruptions. For data corrupted by Gaussian noise, their experiments suggests, that construction of an affinity matrix via $P^\star$ and using this affinity matrix for subsequent spectral clustering can often deliver satisfactory results.

Another method, which solves subspace clustering via a convex optimization program is sparse subspace clustering (SSC) [23]. It relies on the fact, that (under certain conditions) every data point in a union of subspace-model can be reconstructed by a combination of other points in the dataset, i.e. for any data point Y_l in a subspace S_i, there is a vector $c_l = (c_{l1}, c_{l2}, \ldots, c_{ln})^t$ satisfying

$$Y_l = \boldsymbol{Y} c_l, \quad c_{ll} = 0. \tag{1.1.1}$$

The representation (1.1.1) is not unique in general, since possibly $n_i > d_i + 1$; there could even be nonzero elements referring to points not in the subspace S_i for the case of nonindendepent subspaces. However, as long as the number n_i of data points in S_i exceeds the dimension d_i, there clearly exists a representation of the type (1.1.1) such that all nonzero elements refer to elements of the same subspace S_i (i.e. $c_{lm} \neq 0 \Rightarrow Y_m \in S_i$). The key observation of [23] is, that there are also sparse representations of that kind (ideally involving exactly d_i nonzero element). [23] refer to such a representation as a a *subspace sparse* representation. They show that for independent subspaces as well as under mild conditions for disjoint subspaces, such a subspace sparse representation

can be efficiently recovered by solving the convex optimization program

$$\min \|c_l\|_1 \quad \text{s.t.} \quad Y_l = \boldsymbol{Y} c_l, \quad c_{ll} = 0.$$

The segmentation can then be inferred via solving the optimization problem for all data points:

$$\min \|\boldsymbol{C}\|_1 \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{Y}\,\boldsymbol{C}, \quad \operatorname{diag}(\boldsymbol{C}) = 0,$$

since C_{lm} is clearly 0, when Y_l and Y_m belong to different subspaces. In the case of noise and/or outliers, they suggest the approach

$$\begin{aligned} &\min \|\boldsymbol{C}\|_1 + \lambda_e \|E\|_1 + \frac{\lambda_z}{2} \|Z\|_F^2, \\ &\text{s.t. } \boldsymbol{Y} = \boldsymbol{Y}\,\boldsymbol{C} + \boldsymbol{E} + \boldsymbol{Z}, \quad diag(\boldsymbol{C}) = 0. \end{aligned}$$

Similarly to [63], they then proceed by constructing an affinity matrix (e.g. $A = |C| + |C^t|$) and subsequent spectral clustering.

1.2 Outline and Contribution

Let us shortly sketch the outline of the remainder of this thesis and adduce the main contributions. In Chapter 2, we state the mathematical foundations required in the course of this work. In particular, we recapitulate some well-known facts about invariant measures and the gamma function. We will further give a brief introduction into the theory of zonal polynomials and into the main theorems of distance correlation.

We proceed in Chapter 3 by introducing an alternative version of distance correlation, termed the *affinely invariant distance correlation*. We compute the population version of the affinely invariant distance covariance for the multivariate normal and derive several limit theorems, for the cases where either one or both of the dimensions of the random vectors under consideration go to infinity. The chapter is concluded by an application of our results on wind vector data.

Chapter 4 deals with an integral which is fundamental for the theory of distance correlation. We derive an extension of this integral, which may potentially be used to generalize the class of α-distance dependence measures to α outside the range $(0, 2)$.

Subsequently, Chapter 5 deals with the computation of the distance correlation coefficients for random vectors, whose joint distributions are in the class of *Lancaster distributions*. After giving several examples for Lancaster distributions, we state a theorem, which facilitates the computation of the distance covariance immensely, for distributions being in that class. We point out the significance of this results by calculating the distance covariance explicitly for the examples given in the beginning of this chapter.

Chapter 6 is dedicated to a novel approach for the clustering of random variables. After motivating a specific clustering task by an application in low-rank Gaussian Graphical models, we remark that this problem is highly related to the problem of subspace clustering for data. It is further proven that the clustering can be exactly recovered in the noiseless case; when noise is included, we receive an asymptotic guarantee to retain the clusters, for the setting where the sample size goes to infinity. Finally, Chapter 7 summarizes the work and gives an outlook into possible future work.

The main contributions are:

- Definition of the affinely invariant distance correlation and proof of the consistency of its sample measure (section 3.1). Computation of the affinely invariant distance correlation for the multivariate normal (section 3.2). Derivation of several limit theorems for the multivariate normal (section 3.3).

- Development of an formula for the distance covariance for the class of Lancaster distributions. Explicit calculation of the affinely invariant distance correlation coefficient for several distributions in that class, namely the bivariate and multivariate normal distributions, and for bivariate gamma and Poisson distributions (Chapter 5).

- Computation of the regular distance correlation coefficient for the multivariate normal (Appendix A.1) and the affinely invariant distance correlation for the multivariate Laplace (Appendix A.2).

- Generalization of an fundamental integral appearing in the theory of distance correlation (Chapter 4).

- Motivation of a novel approach to the clustering of random variables and derivation of a consistent procedure to recover the clustering in the case of noisy data for the probabilistic PPCA model (Chapter 6).

1.3 Notation

To conclude this chapter, we give an overview over the notation throughout this thesis. This list is by no means complete; due to the diverse problems considered in this thesis, it will be unavoidable to introduce additional notation in the respective chapters.

For a complex value z, the complex conjugate will be denoted by $\bar{z}$ and $|z| = z\bar{z}$. The real part of z will be denoted by $\Re(z)$, the imaginary part by $\Im(z)$. For a column vector $s \in \mathbb{R}^p$, where p is positive integer, we will denote by $|s|_p$ the standard Euclidean norm

of s, i.e. if $s = (s_1, \ldots, s_p)'$ then

$$|s|_p = (s_1^2 + \cdots + s_p^2)^{1/2}.$$

Moreover, for vectors u and v of the same dimension, p, we let $\langle u, v \rangle_p$ be the standard Euclidean scalar product of u and v. For a matrix $M \in \mathbb{R}^{m \times n}$, M^t will denote its transpose and $\mathrm{tr}\,(M)$ its trace; the spectral norm of M will be denoted by $\|M\|$, its Frobenius norm by $\|M\|_F$. Moreover, denote by $S(p)$ the space of symmetric $p \times p$-matrices and by $O(p)$ the orthogonal group of matrices in $\mathbb{R}^{p \times p}$.

For jointly distributed random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, let Σ_X and Σ_Y denote their respective covariance matrices, further let

$$f_{X,Y}(s,t) = \mathbb{E} \exp\big[i \langle s, X \rangle_p + i \langle t, Y \rangle_q\big]$$

be the joint characteristic function of (X, Y), and let $f_X(s) = f_{X,Y}(s, 0)$ and $f_Y(t) = f_{X,Y}(0, t)$ be the marginal characteristic functions of X and Y, respectively.

Analogously, given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from jointly distributed random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, we denote by $S_{\boldsymbol{X}}$ and $S_{\boldsymbol{Y}}$ the usual sample covariance matrices, further let

$$f^n_{\boldsymbol{X},\boldsymbol{Y}}(s,t) = \frac{1}{n} \sum_{j=1}^{n} \exp\big[i \langle s, X_j \rangle_p + i \langle t, Y_j \rangle_q\big].$$

be the sample characteristic function. Finally, we write $f^n_{\boldsymbol{X}}(s) = f^n_{\boldsymbol{X},\boldsymbol{Y}}(s, 0)$ and $f^n_{\boldsymbol{Y}}(t) = f^n_{\boldsymbol{X},\boldsymbol{Y}}(0, t)$ for the respective empirical characteristic functions of the marginals.

Chapter 2

Preliminaries

2.1 Invariant Measures

In the course of this thesis, we will require some measures, which are invariant under certain transformations. The best-known measure of this type is of course the Lebesgue measure λ, which is invariant under translation $\lambda(A) \mapsto \lambda(A+r)$, $r \in \mathbb{R}$. An extension of the Lebesgue measure is the Haar measure, which can be defined on Lie groups (or even more general, on locally compact topological groups). For an introduction into Lie groups as well as for the following definition, see [49].

Definition 2.1.1. *Let G be a Lie group. A nonzero Borel measure μ on G is called a (left) Haar measure if it is invariant under left translation, i.e.*

$$\int_G f(gx)d\mu = \int_G f(x)d\mu \tag{2.1.1}$$

for integrable functions $f : G \mapsto \mathbb{R}$ and elements $g \in G$.

Further, by Theorem 8.21 and 8.23 in [49]:

Theorem 2.1.2. *Let G be a Lie group. Then there exists a Haar measure on G and for any two Haar measures μ and ν on G, there is a $c > 0$, such that*

$$c\mu = \nu,$$

hence the Haar measure is unique up to multiplication with a constant factor.

Since for $p \in \mathbb{N}$, the orthogonal group $O(p)$ is well-known to be a compact Lie group [38], there exists a Haar measure μ on $O(p)$. Due to the fact that μ is a Borel measure and $O(p)$ is compact, μ is finite. Hence, there exists a unique Haar measure μ^*, such that

$$\int_{O(p)} d\mu^* = 1.$$

We will refer to that measure as the *normalized Haar measure on* $O(p)$.

By the transition to *polar coordinates* for an $x \in \mathbb{R}^p$, we will refer to the decomposition

$$x = r\,\theta,$$

where $r \in \mathbb{R}_+$ and $\theta = (\theta_1, \ldots, \theta_p)' \in S^{p-1}$.

In particular, we will make use of the following theorem [24, Proposition XVI.2.1].

Theorem 2.1.3. *Let f be an integrable function on $\mathbb{R}^p$, then*

$$\int_{\mathbb{R}^p} f(x)dx = \int_0^\infty \int_{S^{p-1}} f(r\theta)\, r^{p-1}\, \mathrm{d}\theta\, \mathrm{d}r,$$

where $S^{p-1} = \{x \in \mathbb{R}^p | x'x = 1\}$ denotes the unit sphere and $\mathrm{d}\theta$ refers to the unnormalized surface measure on S^{p-1}.

A detailed treatment of the unnormalized surface measure on S^{p-1} and its generalizations can be found in section 2.1.4 of [68]. We just state the following well-known facts.

Remark 2.1.4. *Let $\mathrm{d}\theta$ refer to the unnormalized surface measure on S^{p-1}. Then*

(i)

$$\int_{S^{p-1}} f(O\theta)\, \mathrm{d}\theta = \int_{S^{p-1}} f(\theta)\, \mathrm{d}\theta \tag{2.1.2}$$

for all integrable functions $f : S^{p-1} \mapsto \mathbb{R}$ and orthogonal matrices $O \in O(p)$.

(ii) The surface of the sphere in $\mathbb{R}^p$ is given by

$$\int_{S^{p-1}} \mathrm{d}\theta = \frac{2\pi^{p/2}}{\Gamma(p/2)}.$$

2.2 The Gamma Function

The gamma function naturally arises in the evaluation of certain integrals and power series, that we will encounter in this work. Although we assume the reader to be well-acquainted with the definition and the properties of the gamma function, we state some basic facts for easy reference (see e.g. [26]).

Theorem 2.2.1. *The gamma integral*

$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt \tag{2.2.1}$$

converges absolutely for $\Re(z) > 0$ and represents an analytical function in this area.

Theorem 2.2.2. *The gamma function Γ is analytically continuable to the whole complex plane with exception of the points*

$$z \in S := \{0, -1, -2, -3, \ldots\}.$$

For this area, it satisfies the functional equation

$$\Gamma(z+1) = z\,\Gamma(z). \tag{2.2.2}$$

In particular, it holds, for $n \in \mathbb{N}_0$

$$\Gamma(n+1) = n!.$$

Definition 2.2.3. *Let k be a nonnegative integer. For $\alpha \in \mathbb{C}$, the rising factorial $(\alpha)_k$ is defined by*

$$(\alpha)_k = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} = \alpha\,(\alpha+1)\,(\alpha+2)\,\cdots\,(\alpha+k-1). \tag{2.2.3}$$

Finally, we will need Stirling's formula for positive real values [26].

Theorem 2.2.4 (Stirling's formula)**.** *For $x \in \mathbb{R}_+$, it holds*

$$\Gamma(x) = \sqrt{2\pi}x^{x-\frac{1}{2}}e^{-x}(1 + O(x^{-1})). \tag{2.2.4}$$

2.3 Zonal Polynomials and Hypergeometric Functions of Matrix Argument

Many integrals arising in the theory of multivariate statistics cannot be evaluated in closed form expressions of elementary functions. For the computation of these integrals and the formulation of the outcome in an efficient and reasonable way, we will need the theory of zonal polynomials. To the best knowledge of the author, zonal polynomials first show up in the well-known paper by James [44], where they are employed to express a certain integral over the group of orthogonal matrices $O(p)$. The theory was then developed by James [45, 46, 47] and Constantine [10, 11]. In this work, we will mainly follow the book of Muirhead [68, Chapter 7].

Definition 2.3.1. *A partition κ is a vector of nonnegative integers $(k_1, \ldots, k_p)$ such that $k_1 \geq \cdots \geq k_p$. Moreover we will denote by $|\kappa|$ the sum of the entries of κ, i.e. $|\kappa| = k_1 + \cdots + k_p$. The length of κ is the largest integer j such that $k_j > 0$ and will be denoted by $\ell(\kappa)$.*

Definition 2.3.2. *For two partitions $\kappa = (k_1, \ldots, k_p)$ and $\iota = (j_1, \ldots, j_p)$ with $|\kappa| = |\iota| = k$, we write $\kappa > \iota$ if $k_i > j_i$ for the first index i for which they are different. If $\kappa > \iota$, we will further say, that the monomial $\lambda_1^{k_1} \cdots \lambda_p^{k_p}$ is of higher weight than the monomial $\lambda_1^{j_1} \cdots \lambda_p^{j_p}$.*

Example 2.3.3. *Let $\kappa = (3, 2, 2, 0)$ and $\iota = (3, 2, 1, 1)$. Then $\kappa > \iota$ and the monomial $\lambda_1^3 \lambda_2^2 \lambda_3^2$ is of higher weight than the monomial $\lambda_1^3 \lambda_2^2 \lambda_3 \lambda_4$.*

Definition 2.3.4. *Let $\kappa = (k_1, \ldots, k_p)$ be a partition with $|\kappa| = k$. The zonal polynomial $C_\kappa(\Lambda)$ is the unique function $C_\kappa : S(p) \mapsto \mathbb{R}$, such that:*

(i) $C_\kappa(\Lambda)$ is a symmetric polynomial in the eigenvalues $\lambda_1, \ldots, \lambda_p$ of Λ.

(ii) $C_\kappa(\Lambda)$ is homogeneous of degree $|\kappa|$ in Λ: For any $\delta \in \mathbb{R}$,

$$C_\kappa(\delta\Lambda) = \delta^{|\kappa|} C_\kappa(\Lambda). \tag{2.3.1}$$

(iii) The term of highest weight in $C_\kappa(\Lambda)$ is $\lambda_1^{k_1} \cdots \lambda_p^{k_p}$, i.e.

$$C_\kappa(\Lambda) = d_k \lambda_1^{k_1} \cdots \lambda_p^{k_p} + \text{ terms of lower weight},$$

where d_k is a constant.

(iv) $C_\kappa(\Lambda)$ is an eigenfunction of the differential operator Δ_Λ given by

$$\Delta_\Lambda = \sum_{i=1}^{p} \lambda_i^2 \frac{\partial^2}{\partial \lambda_i^2} + \sum_{i=1}^{p} \sum_{j \neq i} \frac{\lambda_i^2}{\lambda_i - \lambda_j} \frac{\partial}{\partial \lambda_i},$$

i.e.

$$\Delta_\Lambda C_\kappa(\Lambda) = \alpha C_\kappa(\Lambda),$$

where $\alpha \in \mathbb{R}$ is a constant which does not depend on Λ.

(v) For any nonnegative integer k,

$$\sum_{|\kappa|=k} C_\kappa(\Lambda) = (\operatorname{tr} \Lambda)^k. \tag{2.3.2}$$

Note, that by (i) of the above definition, we have

$$C_\kappa(K'\Lambda K) = C_\kappa(\Lambda) \tag{2.3.3}$$

for all $K \in O(p)$, since the eigenvalues are not affected by the transition from Λ to $K'\Lambda K$.

Further by Corollary 7.2.4 in Muirhead [68], we obtain:

Remark 2.3.5. *If Λ is of rank r then $C_\kappa(\Lambda) = 0$ whenever $\ell(\kappa) > r$.*

There is a natural extension of the zonal polynomials to nonsymmetric matrices [68, p. 237]. When X is positive definite and Y is symmetric, the eigenvalues of $X^{1/2}YX^{1/2}$ and XY obviously coincide. Hence, we may then define

$$C_\kappa(XY) := C_\kappa(X^{1/2}YX^{1/2}). \tag{2.3.4}$$

Obviously, properties $(i)-(v)$ also hold for this extended definition of zonal polynomials. Moreover, by Theorem 7.2.5 in [68]:

Theorem 2.3.6. *For any symmetric matrices $\Lambda_1, \Lambda_2 \in \mathbb{R}^{p\times p}$,*

$$\int_{O(p)} C_\kappa(K'\Lambda_1 K\Lambda_2)\,\mathrm{d}K = \frac{C_\kappa(\Lambda_1)\,C_\kappa(\Lambda_2)}{C_\kappa(I_p)}, \tag{2.3.5}$$

where $I_p = \mathrm{diag}(1,\ldots,1) \in \mathbb{R}^{p\times p}$ denotes the identity matrix and the integral is with respect to the normalized Haar measure on $O(p)$.

For a partition κ with one part (i.e. $\kappa = (k)$), the value of the zonal polynomials can be explicitly stated using the notion of the rising factorial (2.2.3) (see [37, Lemma 6.8]).

Theorem 2.3.7. *Let $\lambda_1,\ldots,\lambda_p$ be the eigenvalues of Λ. Then, for a partition (k) with one part,*

$$C_{(k)}(\Lambda) = \frac{k!}{(\frac{1}{2})_k} \sum_{i_1+\cdots+i_p=k} \prod_{j=1}^{p} \frac{(\frac{1}{2})_{i_j}\,\lambda_j^{i_j}}{i_j!}\,, \tag{2.3.6}$$

where the sum is over all nonnegative integers $i_1,\ldots,i_p$ such that $i_1+\cdots+i_p = k$. In particular, on setting $\lambda_j = 1$, $j = 1,\ldots,p$, we obtain from (2.3.6)

$$C_{(k)}(I_p) = \frac{(\frac{1}{2}p)_k}{(\frac{1}{2})_k}, \tag{2.3.7}$$

Definition 2.3.8. *For any $\alpha \in \mathbb{C}$ and any partition $\kappa = (k_1,\ldots,k_p)$, the partitional rising factorial is defined as*

$$(\alpha)_\kappa = \prod_{j=1}^{p} \big(\alpha - \tfrac{1}{2}(j-1)\big)_{k_j}. \tag{2.3.8}$$

While the concept of the zonal polynomials is already interesting for itself and yields a wealth of applications, one leading motivation to define zonal polynomials is to extend the usual generalized hypergeometric functions to functions of matrix arguments [46, 68, 37]:

Definition 2.3.9. *Let $\alpha_1, \ldots, \alpha_l, \beta_1, \ldots, \beta_m \in \mathbb{C}$ where $-\beta_i + \frac{1}{2}(j-1)$ is not a non-negative integer, for all $i = 1, \ldots, m$ and $j = 1, \ldots, p$. Then the ${}_lF_m$ generalized hypergeometric function of matrix argument is defined as*

$$ {}_lF_m(\alpha_1, \ldots, \alpha_l; \beta_1, \ldots, \beta_m; S) = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{|\kappa|=k} \frac{(\alpha_1)_\kappa \cdots (\alpha_l)_\kappa}{(\beta_1)_\kappa \cdots (\beta_m)_\kappa} C_\kappa(S), \tag{2.3.9} $$

where $S \in S(p)$ is a $p \times p$-symmetric matrix.

It is well-known [68], that the hypergeometric series (2.3.9) converges for all S if $l \leq m$ and for $\|S\| < 1$ if $l = m+1$. A complete analysis of the convergence properties of this series was derived by Gross and Richards [37], and we refer the reader to that paper for the details.

For the case $n = 1, S = s$, we maintain the regular generalized hypergeometric functions (of scalar argument), see [70], p.404 for reference:

$$ {}_lF_m(\alpha_1, \ldots, \alpha_l; \beta_1, \ldots, \beta_m; s) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{(\alpha_1)_k \cdots (\alpha_l)_k}{(\beta_1)_k \cdots (\beta_m)_k} s^k. \tag{2.3.10} $$

As already mentioned before, the convergence of hypergeometric functions has been elaborately studied. In the course of this work, the use of Gauss' Theorem for hypergeometric functions [70, 3] will be sufficient.

Theorem 2.3.10 (Gauss' Theorem for hypergeometric functions)**.** *If $\Re(c-a-b) > 0$, the series ${}_2F_1(a, b; c; s)$ as defined in (2.3.10) also converges for the special value $s = 1$ and*

$$ {}_2F_1(a, b; c; 1) = \frac{\Gamma(c)\,\Gamma(c-a-b)}{\Gamma(c-a)\,\Gamma(c-b)}. $$

It is clear from the definition of the hypergeometric functions ${}_lF_m$, that a multitude of connections between the ${}_lF_m$ with different parameters $\alpha_1, \ldots, \alpha_l, \beta_1, \ldots, \beta_m$ can be established. To obtain explicit expressions in terms of elementary functions, we need to state some of these connections for the ${}_2F_1$. It is easy to see, that

$$ {}_2F_1(a+1, b+1; c+1; x) = \frac{c}{ab} \frac{d}{dx} {}_2F_1(a, b; c). \tag{2.3.11} $$

Moreover, there is a set of contiguous relations holding for the ${}_2F_1$. In particular, we have [2, p.94]:

$$ \begin{aligned} {}_2F_1(a, b; c; x) = x(1-x) \frac{(a+1)(b+1)}{c(c+1)} {}_2F_1(a+2, b+2; c+2; x) \\ + \frac{(c-(a+b+1)x)}{c} {}_2F_1(a+1, b+1; c+1; x) \end{aligned} \tag{2.3.12} $$

and

$$c(1-z)\frac{d}{dx}{}_2F_1(a,b;c;x) = (c-a)(c-b){}_2F_1(a,b;c+1;x) + c(a+b-c){}_2F_1(a,b;c;x). \tag{2.3.13}$$

For many hypergeometric functions, explicit representations in terms of elementary functions have been established; see [70] for a survey. Let us note that [2, 70]

$${}_2F_1(\tfrac{1}{2},\tfrac{1}{2};\tfrac{3}{2};s^2) = s^{-1}\sin^{-1}(s). \tag{2.3.14}$$

By (2.3.11) and (2.3.14), we obtain

$${}_2F_1(\tfrac{3}{2},\tfrac{3}{2};\tfrac{5}{2};s) = \tfrac{1}{2}\big(s^{-3/2}\sin^{-1}(\sqrt{s}) + s^{-1}(1-s)^{-\frac{1}{2}}\big).$$

Inserting the latter equation and (2.3.14) into (2.3.12) gives us

$${}_2F_1(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2};s^2) = s\sin^{-1}s + (1-s^2)^{1/2}. \tag{2.3.15}$$

Further, exploiting (2.3.13) yields

$${}_2F_1(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{3}{2};\rho^2) = \frac{3(1-\rho^2)^{1/2}}{4} + \frac{(1+2\rho^2)\sin^{-1}\rho}{4\rho}. \tag{2.3.16}$$

Finally, by repeated application of (2.3.13), it can be shown that for $k = 2, 3, 4, \ldots$,

$${}_2F_1(-\tfrac{1}{2},-\tfrac{1}{2};k+\tfrac{1}{2};\rho^2) = \rho^{-2(k-1)}(1-\rho^2)^{1/2}P_{k-1}(\rho^2) + \rho^{-(2k-1)}Q_k(\rho^2)\sin^{-1}\rho, \tag{2.3.17}$$

where P_k and Q_k are polynomials of degree k.

2.4 Distance Correlation

The goal of this section will be mainly to introduce the reader to the concept of distance correlation, a novel measure of independence introduced by Székely, et al. [102, 100]. We will further state some important recent results concerning the application of distance correlation to time series [116]. The distance covariance is defined for random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ of arbitrary dimension p and q and measures any kind of dependencies between X and Y. It can be formulated via an integral involving characteristic functions of these vectors.

Definition 2.4.1. *The distance covariance between random vectors $X \in \mathbb{R}^p$ and $Y \in$*

$\mathbb{R}^q$ *with finite first moments is the nonnegative number* $\mathcal{V}(X,Y)$ *defined by*

$$\mathcal{V}^2(X,Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(s,t) - f_X(s) f_Y(t)|^2}{|s|_p^{1+p}\, |t|_q^{1+q}} \,\mathrm{d}s\,\mathrm{d}t, \tag{2.4.1}$$

where $|z|$ *denotes the modulus of* $z \in \mathbb{C}$ *and*

$$c_p = \frac{\pi^{\frac{1}{2}(p+1)}}{\Gamma\left(\frac{1}{2}(p+1)\right)} = \tfrac{1}{2} \int_{S^{p-1}} \mathrm{d}\theta, \tag{2.4.2}$$

where $\mathrm{d}\theta$ *denotes the unnormalized surface measure on* S^{p-1}.

The distance correlation $\mathcal{R}$ is then just a normalized version of the distance covariance, in the way that $\mathcal{R}(X,Y) = 1$ if $X = Y$.

Definition 2.4.2. *The distance correlation between* X *and* Y *is the nonnegative number defined by*

$$\mathcal{R}(X,Y) = \frac{\mathcal{V}(X,Y)}{\sqrt{\mathcal{V}(X,X)\mathcal{V}(Y,Y)}} \tag{2.4.3}$$

if both $\mathcal{V}(X,X)$ *and* $\mathcal{V}(Y,Y)$ *are strictly positive, and defined to be zero otherwise.*

One can further specify sample measures for the distance covariance and the distance correlation in analogous fashion. Particularly, given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from jointly distributed random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ and setting

$$\boldsymbol{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{p \times n} \qquad \text{and} \qquad \boldsymbol{Y} = [Y_1, \ldots, Y_n] \in \mathbb{R}^{q \times n},$$

one can define:

Definition 2.4.3. *The sample distance covariance is the nonnegative number* $\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y})$ *defined by*

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f^n_{\boldsymbol{X},\boldsymbol{Y}}(s,t) - f^n_{\boldsymbol{X}}(s) f^n_{\boldsymbol{Y}}(t)|^2}{|s|_p^{1+p}\, |t|_q^{1+q}} \,\mathrm{d}s\,\mathrm{d}t,$$

where c_p *is the constant given in (2.4.2).*

Definition 2.4.4. *The sample distance correlation then is defined by*

$$\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{X})\mathcal{V}_n(\boldsymbol{Y}, \boldsymbol{Y})}} \tag{2.4.4}$$

if both $\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{X})$ *and* $\mathcal{V}_n(\boldsymbol{Y}, \boldsymbol{Y})$ *are strictly positive, and defined to be zero otherwise.*

Both the sample distance covariance and the sample distance correlation can be proven to be consistent.

Theorem 2.4.5. *If X and Y possess finite first moments, then, for $n \to \infty$*

$$\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y}) \xrightarrow{a.s.} \mathcal{V}(X, Y), \quad \mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) \xrightarrow{a.s.} \mathcal{R}(X, Y).$$

Certainly the most prominent feature of distance correlation, is that it defines independence, i.e. $\mathcal{R}(X, Y) = 0$ if and only if X and Y are independent. This property is immediately clear from (2.4.1) along with the fact, that the characteristic function of (X, Y) coincides with the product of the marginal characteristic functions merely in the case of independence.

However, it has to be noted that distance correlation is by far not the only measure satisfying this property and there may exist measures with more tempting theoretical characteristics, e.g. the maximal correlation coefficient [29, 76]. What makes distance correlation stand out from the others is the striking simplicity of its sample measure, which can be expressed as the Schur product of the centralized distance matrices [102].

Theorem 2.4.6. *Let*

$$a_{kl} = |X_k - X_l|_p, \quad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl},$$

and

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot},$$

similarly define $b_{kl} = |Y_k - Y_l|_q$, $\bar{b}_{k\cdot}$, $\bar{b}_{\cdot l}$, $\bar{b}_{\cdot\cdot}$, and B_{kl}, where $k, l = 1, \ldots, n$. Then

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl} B_{kl}. \tag{2.4.5}$$

This intriguingly simple version of the sample measure can be stated in an alternative form that will prove useful in the following.

Corollary 2.4.7. *Let*

$$S_1 = \frac{1}{n^2}\sum_{k=1,l=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q,$$

$$S_2 = \frac{1}{n^2}\sum_{k=1,l=1}^{n} |X_k - X_l|_p \, \frac{1}{n^2}\sum_{k=1,l=1}^{n} |Y_k - Y_l|_q,$$

$$S_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q.$$

Then

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = S_1 + S_2 - S_3. \tag{2.4.6}$$

Having defined distance covariance as well as its sample version, we now state the most relevant properties of these measures. Property (i) and (ii) of the following have already been mentioned before and are easy to prove, property (iii) is shown in [102, p. 2779].

Theorem 2.4.8. *(i) If X and Y possess finite first moments, then $0 \leq \mathcal{R}(X, Y) \leq 1$ and $\mathcal{R}(X, Y) = 0$ if and only if X and Y are independent.*

(ii) $0 \leq \mathcal{R}_n \leq 1$.

(iii) If $\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) = 1$, then $p = q$ and there exist a vector a, a nonzero real number b and an orthogonal matrix C, such that $\boldsymbol{Y} = a + bC\boldsymbol{X}$.

The proof of Theorem 2.4.6 and thereby the simplicity of the sample measures essentially rely on the fact, that a certain multidimensional singular integral involving a parameter x can be evaluated to be a constant multiple of the euclidean norm of x. The outcome of this integral is well-known and appears in many different fields of probability and statistics see e.g. the books of Chilès and Delfiner [9] and Rachev et al. [73]. A proof of a general form of this result, which is stated in the following lemma, can be found in [99]:

Lemma 2.4.9. *Suppose that $\alpha \in \mathbb{C}$ satisfies $0 < \Re(\alpha) < 2$. Then, for all $x \in \mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} \frac{1 - \cos(\langle t, x \rangle)}{|t|_d^{d+\alpha}} \, \mathrm{d}t = C(d, \alpha) \, |x|_d^{\alpha}, \tag{2.4.7}$$

where

$$C(d, \alpha) = \frac{2\pi^{d/2} \, \Gamma(1 - \alpha/2)}{\alpha \, 2^{\alpha} \, \Gamma\big((d + \alpha)/2\big)}. \tag{2.4.8}$$

The integrals at 0 and ∞ are meant in the principal value sense: $\lim_{\varepsilon \to 0} \int_{\mathbb{R}^d \setminus \{\varepsilon B + \varepsilon^{-1} B^C\}}$, where B is the unit ball (centered at 0) in $\mathbb{R}^d$ and B^C is the complement of B.

As we have explained above, the preceding Lemma secures the simplicity of the sample version of distance correlation and is hence fundamental for the idea of distance correlation itself. This suggests, that studying the underlying Lemma 2.4.9 might lead to both a better understanding of distance correlation and possible ways of generalizing distance correlation. In the course of this thesis, we will present several extensions of this integral.

Though the computation of the sample distance correlation is easy to perform for a given data sample, the calculation of distance covariance and distance correlation for

certain multivariate distributions represents a challenging problem since the integral (2.4.1) is difficult to analytically evaluate even for simple bivariate distributions. This means, however, that the physical interpretation of distance correlation is not clear and one does not really know what exactly one estimates when determining the distance correlation for a given data sample. For a better understanding of the concept of distance correlation, the knowledge of its exact value for a wide class of multivariate distributions is crucial. In their groundbreaking paper [102], Székely et al. state the important result for the bivariate normal.

Theorem 2.4.10. *If X and Y are standard normal with $cor(X, Y) = \rho$, then*

(i) $\mathcal{R}(X, Y) \leq |\rho|$,

(ii) $\mathcal{R}^2(X, Y) = \frac{\rho \sin^{-1} \rho + \sqrt{1-\rho^2} - \rho \sin^{-1} \rho/2 - \sqrt{4-\rho^2} + 1}{1+\pi/3-\sqrt{3}}$,

(iii) $\inf_{\rho \neq 0} \frac{\mathcal{R}(X,Y)}{\rho} = \lim_{\rho \to 0} \frac{\mathcal{R}(X,Y)}{\rho} = \frac{1}{2(1+\pi/3-\sqrt{3})^{1/2}}$.

Besides the theoretical investigation of distance correlation, Székely et al. [102] show the potential of this concept for applications. Most importantly, they propose a test for independence, which has been shown to outperform the celebrated MIC [77] in various settings, see [92] and [33] for reference.

Theorem 2.4.11. *Suppose $T(X, Y, \alpha, n)$ is the test that rejects independence if*

$$\frac{n\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y})}{S_2} > (\Phi^{-1}(1 - \alpha/2))^2, \tag{2.4.9}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and S_2 is defined as in Corollary 2.4.7. Further let $\alpha(X, Y, n)$ denote the achieved significance level of $T(X, Y, \alpha, n)$. For random vectors X and Y with finite first moments and all $0 < \alpha \leq 0.215$, it holds

(i) $\lim_{n\to\infty} \alpha(X, Y, n) \leq \alpha$.

(ii) $\sup_{X,Y}\{\lim_{n\to\infty} \alpha(X, Y, n) | \mathcal{V}(X, Y) = 0\} = \alpha$.

Let us note, that there are other possibilities to define measures of independence, which satisfy the pleasant properties given in Theorem 2.4.8. Székely et al. [102] suggest to define a generalized distance correlation for random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ via

$$\mathcal{V}^2(X, Y; \omega, \zeta) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(s, t) - f_X(s) f_Y(t)|^2 \omega(s)\zeta(t) \, \mathrm{d}s \, \mathrm{d}t, \tag{2.4.10}$$

where ω and ζ are suitable weight functions. A correlation measure $\mathcal{R}_{\omega,\zeta}$ can then be obtained by

$$\mathcal{R}_{\omega,\zeta}(X, Y) = \frac{\mathcal{V}(X, Y; \omega, \zeta)}{\sqrt{\mathcal{V}(X, X, \omega, \omega)\, \mathcal{V}(Y, Y, \zeta, \zeta)}}.$$

It should be noted, that ω and ζ should be chosen non-integrable, since,

$$\lim_{\varepsilon \to 0} \mathcal{R}_{\omega,\zeta}(\varepsilon X, \varepsilon Y) = \rho^2(X, Y)$$

for real-valued X, Y and integrable weight functions ω and ζ. Hence, for uncorrelated random variables X and Y, $\mathcal{R}_{\omega,\zeta}(X, Y)$ can be arbitrarily close to 0, even if X and Y are dependent.

Though Székely and Rizzo prove in [101], that distance correlation is the unique measure of the type $\mathcal{V}^2(X, Y; \omega, \zeta)$, which is scale-equivariant and invariant to all shift and orthogonal transformations on X and Y, we will show in section 3.1 that there exist other weight functions ω and ζ which lead to alternative dependence measures with very interesting properties.

In the discussion of [100], Rémillard [75] proposes the use of the distance correlation to explore nonlinear dependencies in time series data. Zhou [116] pursued this approach recently and defined the auto distance covariance function and the auto distance correlation function, along with natural sample versions, for a strongly stationary vector-valued time series:

Definition 2.4.12. *Let $X = (X_j)_{j=-\infty}^{\infty}$ be a strictly stationary multivariate time series of dimension p. Then the auto distance covariance function $\mathcal{V}_X$ is, for $k \geq 0$ defined as*

$$\mathcal{V}_X(k) = \frac{1}{c_p^2} \int_{\mathbb{R}^{2p}} \frac{|f_{X_0,X_k}(s,t) - f_{X_0}(s) f_{X_k}(t)|^2}{|s|_p^{p+1} |t|_p^{p+1}} dsdt,$$

moreover, the auto distance correlation function $\mathcal{R}_X$ is, for $k \geq 1$, defined as

$$\mathcal{R}_X(k) = \sqrt{\frac{\mathcal{V}_X(k)}{\mathcal{V}_X(0)}}$$

if $\mathcal{V}_X(0)$ is strictly positive, 0 otherwise.

Definition 2.4.13. *Let $\boldsymbol{X} = (X_j)_{j=1}^n$ be an observation of a strictly stationary multivariate time series of dimension p. Then the sample auto distance covariance function $\mathcal{V}_X^n$ is, for $k \geq 0$, defined as*

$$\mathcal{V}_{\boldsymbol{X}}^n(k) = \frac{1}{c_p^2} \int_{\mathbb{R}^{2p}} \frac{|f_k^n(s,t) - f^n(s) f^{n,k}(t)|^2}{|s|_p^{p+1} |t|_p^{p+1}} dsdt,$$

where

$$f_k^n(s,t) = \frac{1}{n-k} \sum_{j=1}^{n-k} \exp\big[i\langle s, X_j\rangle_p + i\langle t, X_{j+k}\rangle_q\big],$$

is the empirical characteristic function of $((X_j, X_{j+k}))_{j=1}^{n-k}$ *and*

$$f^n(s) = f_k^n(s,0), \quad f^{n,k}(t) = f_k^n(0,t)$$

are the respective marginal characteristic functions. With $\boldsymbol{Y} = (X_j)_{j=1}^{n-k}$ *and* $\boldsymbol{Z} = (X_j)_{j=k+1}^{n}$*, the sample auto distance correlation function* $\mathcal{R}_{\boldsymbol{X}}$ *is, for* $k \geq 1$ *defined as*

$$[\mathcal{R}_{\boldsymbol{X}}^n(k)]^2 = \frac{\mathcal{V}_{\boldsymbol{X}}^n(k)}{\sqrt{\mathcal{V}_{\boldsymbol{Y}}^n(0)\,\mathcal{V}_{\boldsymbol{Z}}^n(0)}},$$

whenever the denominator is strictly positive, 0 *otherwise.*

Furthermore, Zhou [116] is able to show the consistency of the sample auto distance covariance functions under moderate assumptions that involve the physical dependence measures $\delta(\cdot,\cdot)$. For details on the physical dependence measures, the reader is referred to [116] and [115].

Theorem 2.4.14. *Suppose* $(\mathbb{E}[|X|_p^{1+r_0}])^{\frac{1}{1+r_0}} < \infty$ *for some* $r_0 > 0$ *and* $\sum_{k=0}^{\infty} \delta(k, 1+r_0) < \infty$*. Then, for all* $k \geq 0$

$$\mathcal{V}_{\boldsymbol{X}}^n(k) \xrightarrow{\mathbb{P}} \mathcal{V}_X(k) \text{ as } n \to \infty.$$

Chapter 3

The Affinely Invariant Distance Correlation

After having introduced the notion of distance correlation, we now extend the results given in Section 2.4. In particular, we define an alternative version of distance correlation, which does not only feature the desirable properties stated in Theorem 2.4.8, but is also invariant under the group of affine transformations (Section 3.1). This measure is called the *affinely invariant distance correlation*. In the following, we derive the population version of the affinely invariant distance correlation for the multivariate normal, thereby widely generalizing the result of Székely et al. [102] stated in Theorem 2.4.10. In Section 3.3, we derive several limit theorems for the multivariate normal, which have relevance for the application of distance correlation to high-dimensional data. We close this chapter with an illustration of our results in Section 3.4, where we apply the concept of affinely invariant distance correlation to a time series of wind vector data. While being purely exploratory, the methods and the output presented in the latter section indicate the potential of distance correlation for the investigation of vector-valued time series. Finally, we mention that the statements given in this chapter represent a slightly extended version of the paper [18] by Dueck, Edelmann, Gneiting and Richards.

3.1 Definition and Properties

The goal of this section will be to define an alternative version of distance correlation which satisfies a crucial group invariance property while retaining all important features of standard distance correlation. To begin with, we point out an invariance property of distance correlation which was already stated in [102] and [100].

Theorem 3.1.1. *Let p and q be positive integers and $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random*

vectors. Moreover let

$$\boldsymbol{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{p\times n} \qquad \textit{and} \qquad \boldsymbol{Y} = [Y_1, \ldots, Y_n] \in \mathbb{R}^{q\times n}.$$

be a random sample from X and Y, respectively. Then for arbitrary constant vectors $a_1 \in \mathbb{R}^p, a_2 \in \mathbb{R}^q$, nonzero number $b_1, b_2 \in \mathbb{R}$ and orthonormal matrices $C_1 \in \mathbb{R}^{p\times p}$, $C_2 \in \mathbb{R}^{q\times q}$:

$$\mathcal{R}(a_1 + b_1 C_1 X, a_2 + b_2 C_2 Y) = \mathcal{R}(X, Y)$$

and

$$\mathcal{R}_n(a_1 + b_1 C_1 \boldsymbol{X}, a_2 + b_2 C_2 \boldsymbol{Y}) = \mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}).$$

Proof. By invariance of the Euclidean norm under orthogonal transformations, we obtain

$$a_{kl} := |(a_1 + b_1 C_1 X_k) - (a_1 + b_1 C_1 X_l)|_p = |b_1 C_1 (X_k - X_l)|_p = |b_1||X_k - X_l|_p$$

and similarly

$$b_{kl} := |(a_2 + b_2 C_2 Y_k) - (a_2 + b_2 C_2 Y_l)|_q = |b_2||Y_k - Y_l|_q.$$

Hence, we can conclude by Theorem 2.4.6, that

$$\mathcal{V}_n^2(a_1 + b_1 C_1 \boldsymbol{X}, a_2 + b_2 C_2 \boldsymbol{Y}) = |b_1|\,|b_2| \mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}).$$

It obviously follows, that

$$\mathcal{R}_n(a_1 + b_1 C_1 \boldsymbol{X}, a_2 + b_2 C_2 \boldsymbol{Y}) = \mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}).$$

The respective equality for the population version is clear by the consistency of the sample version and the uniqueness of the limit. □

The above theorem states that distance correlation is invariant under certain orthogonal transformations of (X, Y). However, the distance correlation fails to be invariant under the group of all invertible affine transformations of (X, Y). This led Székely, et al.[102] and Székely and Rizzo [100] to propose an affinely invariant sample version of the distance covariance and distance correlation.

Definition 3.1.2. *The sample affinely invariant distance covariance is the nonnegative number $\widetilde{\mathcal{V}}_n(\boldsymbol{X}, \boldsymbol{Y})$ defined by*

$$\widetilde{\mathcal{V}}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{V}_n^2(S_{\boldsymbol{X}}^{-1/2}\boldsymbol{X}, S_{\boldsymbol{Y}}^{-1/2}\boldsymbol{Y}) \tag{3.1.1}$$

if $S_{\boldsymbol{X}}$ and $S_{\boldsymbol{Y}}$ are positive definite. otherwise.

Definition 3.1.3. *The sample affinely invariant distance correlation is defined by*

$$\widetilde{\mathcal{R}}_n(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\widetilde{\mathcal{V}}_n(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\widetilde{\mathcal{V}}_n(\boldsymbol{X}, \boldsymbol{X})\widetilde{\mathcal{V}}_n(\boldsymbol{Y}, \boldsymbol{Y})}}, \tag{3.1.2}$$

if the quantities in the denominator are strictly positive, and defined to be zero otherwise.

We now adapt this proposal by introducing an affinely invariant population version of distance correlation.

Definition 3.1.4. *The affinely invariant distance covariance between random variables X and Y with finite second moments is the nonnegative number $\widetilde{\mathcal{V}}(X, Y)$ defined by*

$$\widetilde{\mathcal{V}}^2(X, Y) = \mathcal{V}^2(\Sigma_X^{-1/2}X, \Sigma_Y^{-1/2}Y). \tag{3.1.3}$$

Definition 3.1.5. *The affinely invariant distance correlation between X and Y is the nonnegative number defined by*

$$\widetilde{\mathcal{R}}(X, Y) = \frac{\widetilde{\mathcal{V}}(X, Y)}{\sqrt{\widetilde{\mathcal{V}}(X, X)\widetilde{\mathcal{V}}(Y, Y)}}, \tag{3.1.4}$$

if both $\widetilde{\mathcal{V}}(X, X)$ and $\widetilde{\mathcal{V}}(Y, Y)$ are strictly positive, and defined to be zero otherwise.

Clearly, the population affinely invariant distance correlation and its sample version are invariant under the group of invertible affine transformations, and in addition to satisfying this often-desirable group invariance property [21], they inherit the properties of the standard distance dependence measures. In particular:

Theorem 3.1.6. *(i) $0 \leq \widetilde{\mathcal{R}}(X, Y) \leq 1$ and, for populations with finite second moments and positive definite covariance matrices, $\widetilde{\mathcal{R}}(X, Y) = 0$ if and only if X and Y are independent.*

(ii) $0 \leq \widetilde{\mathcal{R}}_n(\boldsymbol{X}, \boldsymbol{Y}) \leq 1$.

(iii) $\widetilde{\mathcal{R}}_n(\boldsymbol{X}, \boldsymbol{Y}) = 1$ implies that $p = q$, that the linear spaces spanned by $\boldsymbol{X}$ and $\boldsymbol{Y}$ have full rank, and that there exist a vector $a \in \mathbb{R}^p$, a nonzero number $b \in \mathbb{R}$, and an orthogonal matrix $C \in \mathbb{R}^{p \times p}$ such that $S_{\boldsymbol{Y}}^{-1/2}\boldsymbol{Y} = a + b\, C S_{\boldsymbol{X}}^{-1/2}\boldsymbol{X}$.

Our next result shows that the sample affinely invariant distance correlation is a consistent estimator of the respective population quantity.

Theorem 3.1.7. *Let $(X, Y) \in \mathbb{R}^{p+q}$ be jointly distributed random vectors with positive definite marginal covariance matrices $\Sigma_X \in \mathbb{R}^{p \times p}$ and $\Sigma_Y \in \mathbb{R}^{q \times q}$, respectively. Suppose*

that $(X_1, Y_1), \ldots, (X_n, Y_n)$ *is a random sample from* (X, Y)*, and let* $\boldsymbol{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{p \times n}$ *and* $\boldsymbol{Y} = [Y_1, \ldots, Y_n] \in \mathbb{R}^{q \times n}$*. Also, let* $\widehat{\Sigma}_{\boldsymbol{X}}$ *and* $\widehat{\Sigma}_{\boldsymbol{Y}}$ *be strongly consistent estimators for* Σ_X *and* Σ_Y*, respectively. Then*

$$\mathcal{V}_n^2(\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2}\boldsymbol{X}, \widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2}\boldsymbol{Y}) \to \widetilde{\mathcal{V}}^2(X, Y),$$

almost surely, as $n \to \infty$*. In particular, the sample affinely invariant distance correlation satisfies*

$$\widetilde{\mathcal{R}}_n(\boldsymbol{X}, \boldsymbol{Y}) \to \widetilde{\mathcal{R}}(X, Y), \tag{3.1.5}$$

almost surely.

Proof. As the covariance matrices Σ_X and Σ_Y are positive definite, we may assume that the strongly consistent estimators $\widehat{\Sigma}_{\boldsymbol{X}}$ and $\widehat{\Sigma}_{\boldsymbol{Y}}$ also are positive definite. Therefore, in order to prove the first statement it suffices to show that

$$\mathcal{V}_n^2(\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2}\boldsymbol{X}, \widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2}\boldsymbol{Y}) - \mathcal{V}_n^2(\Sigma_X^{-1/2}\boldsymbol{X}, \Sigma_Y^{-1/2}\boldsymbol{Y}) \to 0, \tag{3.1.6}$$

almost surely. By the decomposition (2.4.6), the left-hand side of (3.1.6) can be written as an average of terms of the form

$$\left|\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2}(X_k - X_l)\right|_p \left|\widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2}(Y_k - Y_m)\right|_q - \left|\Sigma_X^{-1/2}(X_k - X_l)\right|_p \left|\Sigma_Y^{-1/2}(Y_k - Y_m)\right|_q.$$

Using the identity

$$\begin{aligned}
&\left|\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2}(X_k - X_l)\right|_p \left|\widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2}(Y_k - Y_m)\right|_q \\
&= \left|(\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2} - \Sigma_X^{-1/2} + \Sigma_X^{-1/2})(X_k - X_l)\right|_p \left|(\widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2} - \Sigma_Y^{-1/2} + \Sigma_Y^{-1/2})(Y_k - Y_m)\right|_q,
\end{aligned}$$

we obtain

$$\begin{aligned}
&\left|\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2}(X_k - X_l)\right|_p \left|\widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2}(Y_k - Y_m)\right|_q - \left|\Sigma_X^{-1/2}(X_k - X_l)\right|_p \left|\Sigma_Y^{-1/2}(Y_k - Y_m)\right|_q \\
&\qquad \le \|\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2} - \Sigma_X^{-1/2}\| \, \|\widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2} - \Sigma_Y^{-1/2}\| \, |X_k - X_l|_p \, |Y_k - Y_m|_q \\
&\qquad\quad + \|\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2} - \Sigma_X^{-1/2}\| \, |X_k - X_l|_p \left|\Sigma_Y^{-1/2}(Y_k - Y_m)\right|_q \\
&\qquad\quad + \|\widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2} - \Sigma_Y^{-1/2}\| \left|\Sigma_X^{-1/2}(X_k - X_l)\right|_p |Y_k - Y_m|_q,
\end{aligned}$$

where the matrix norm $\|\Lambda\|$ is the largest eigenvalue of Λ in absolute value. Now we can separate the three sums in (2.4.6) and place the factors like $\|\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2} - \Sigma_X^{-1/2}\|$ in front of the sums, since they appear in every summand. Then, $\|\widehat{\Sigma}_{\boldsymbol{X}}^{-1/2} - \Sigma_X^{-1/2}\|$ and $\|\widehat{\Sigma}_{\boldsymbol{Y}}^{-1/2} - \Sigma_Y^{-1/2}\|$ tend to zero and the remaining averages converge to constants (representing some distance correlation components) almost surely as $n \to \infty$, and this completes the proof of the first statement. Finally, the property (3.1.5) of strong

consistency of $\widetilde{\mathcal{R}}_n(\boldsymbol{X}, \boldsymbol{Y})$ is obtained immediately upon setting $\widehat{\Sigma}_{\boldsymbol{X}} = S_{\boldsymbol{X}}$ and $\widehat{\Sigma}_{\boldsymbol{Y}} = S_{\boldsymbol{Y}}$. □

Affinely invariant distance covariance can also be viewed as a generalized distance covariance in the sense of (2.4.10). To see this, notice the following extension of Lemma 2.4.9.

Lemma 3.1.8. *If $0 < \alpha < 2$, then for all $x \in \mathbb{R}^d$*

$$\int_{\mathbb{R}^d} \frac{1 - \cos(\langle t, x\rangle)}{\sqrt{t'At}^{d+\alpha}} dt = \frac{1}{\sqrt{\det(A)}} C(d,\alpha) \big(\sqrt{x'A^{-1}x}\big)^{\alpha},$$

where

$$C(d,\alpha) = \frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((d+\alpha)/2)}$$

and $\Gamma(\cdot)$ is the complete gamma function.

Proof. Since A is symmetric and positive definite, we can find an orthogonal matrix P and a diagonal matrix D such that $A = P^{-1}DP$. Hence, we have $A^{1/2} = P^{-1}D^{1/2}P$ and $A^{1/2}$, $A^{-1/2}$ are symmetric. Therefore by substituting $A^{1/2}t \to t$, we obtain

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{1 - \cos(\langle t, x\rangle)}{\sqrt{t'At}^{d+\alpha}} dt &= \frac{1}{\sqrt{\det(A)}} \int_{\mathbb{R}^d} \frac{1 - \cos(\langle A^{-1/2}t, x\rangle)}{|t|_d^{d+\alpha}} dt \\
&= \frac{1}{\sqrt{\det(A)}} \int_{\mathbb{R}^d} \frac{1 - \cos(\langle t, A^{-1/2}x\rangle)}{|t|_d^{d+\alpha}} dt \\
&= \frac{1}{\sqrt{\det(A)}} C(d,\alpha) |A^{-1/2}x|_d^{\alpha}.
\end{aligned}$$

where the last line follows by Lemma 2.4.9. □

Motivated by this lemma, we define weight functions ω_{M_1} and ω_{M_2} via

$$\omega_{M_1}(s) = \frac{\sqrt{\det M_1}}{c_p\, |M_1^{1/2}s|_p^{1+p}}, \quad \omega_{M_2}(t) = \frac{\sqrt{\det M_2}}{c_q\, |M_2^{1/2}t|_q^{1+q}}.$$

Then, making use of the notion of (2.4.10), we obtain

$$\mathcal{V}^2(X, Y; \omega_{M_1}, \omega_{M_2}) = \frac{\sqrt{\det M_1 \det M_2}}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(s,t) - f_X(s) f_Y(t)|^2}{|M_1^{1/2}s|_p^{1+p}\, |M_2^{1/2}t|_q^{1+q}} \mathrm{d}s\, \mathrm{d}t. \quad (3.1.7)$$

Exploiting Lemma 3.1.8 and the proof of Székely [102, Theorem 1, pp. 7], we see that a respective sample measure can be defined similarly to (2.4.5):

Definition 3.1.9. *Let*

$$a_{kl} = |M_1^{-1/2}(X_k - X_l)|_p, \quad \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl},$$

and

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot},$$

similarly $b_{kl} = |M_2^{-1/2}(Y_k - Y_l)|_q$, $\bar{b}_{k\cdot}$, $\bar{b}_{\cdot l}$, $\bar{b}_{\cdot\cdot}$, *and* B_{kl}, *where* $k, l = 1, \ldots, n$.
Now define $\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}, \omega_{M_1}, \omega_{M_2})$ *via*

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}, \omega_{M_1}, \omega_{M_2}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}. \tag{3.1.8}$$

By the same arguments as in [102] and [100], we conclude that the sample version $\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}, \omega_{M_1}, \omega_{M_2})$ is consistent. Furthermore, it is clear by the proof of Theorem 3.1.7 that M_1 and M_2 in the sample version can be replaced by strongly consistent estimators $\widehat{M}_{1n}$ and $\widehat{M}_{2n}$ of M_1 and M_2, respectively.

There are two important consequences of the preceding comments. First, note that the affinely invariant distance covariance can be regarded as measure of the type defined in (2.4.10), since

$$\widetilde{\mathcal{V}}(X, Y) = \mathcal{V}(X, Y, \omega_{\Sigma_X}, \omega_{\Sigma_Y}).$$

Secondly, any choice of positive definite matrices M_1 and M_2 and strongly consistent estimators $\widehat{M}_{1n}$ and $\widehat{M}_{2n}$ yields a measure of dependence

$$\mathcal{R}(X, Y, \omega_{M_1}, \omega_{M_2}) = \frac{\mathcal{V}(X, Y; \omega_{M_1}, \omega_{M_2})}{\sqrt{\mathcal{V}(X, X; \omega_{M_1}, \omega_{M_1})\mathcal{V}(Y, Y; \omega_{M_2}, \omega_{M_2})}},$$

and a respective consistent sample measure

$$\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}, \omega_{\widehat{M}_{1n}}, \omega_{\widehat{M}_{2n}}) = \frac{\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y}, \omega_{\widehat{M}_{1n}}, \omega_{\widehat{M}_{2n}})}{\sqrt{\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{X}, \omega_{\widehat{M}_{1n}}, \omega_{\widehat{M}_{1n}})\mathcal{V}_n(\boldsymbol{Y}, \boldsymbol{Y}, \omega_{\widehat{M}_{2n}}, \omega_{\widehat{M}_{2n}})}},$$

which satisfy the crucial properties of distance correlation stated in Theorem 2.4.8.

Even more important, the asymptotic properties of the test statistic (2.4.9) are not affected by the transition from $\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y})$ to $\mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y}, \omega_{\widehat{M}_{1n}}, \omega_{\widehat{M}_{2n}})$. Hence, a completely analogous but different test can be stated for any pair of strongly consistent estimators $\widehat{M}_{1n}$ and $\widehat{M}_{2n}$. In particular, the affinely invariant distance correlation features a test

analogous to Theorem 2.4.11. Noting the results of Kosorok [55], we raise the possibility that $\widehat{M}_{1n}$ and $\widehat{M}_{2n}$ can be chosen in a judicious, data-dependent way so that the power of the test for independence increases. In any case, tests for independence based on distances are increasingly considered in the recent literature, see for examples the papers by Heller et al. [40] and Székely and Rizzo [102].

3.2 The Affinely Invariant Distance Correlation for Multivariate Normal Populations

We now consider the problem of calculating the affinely invariant distance correlation between the random vectors X and Y where $(X, Y) \sim \mathcal{N}_{p+q}(\mu, \Sigma)$, a multivariate normal distribution with mean vector $\mu \in \mathbb{R}^{p+q}$ and covariance matrix $\Sigma \in \mathbb{R}^{(p+q)\times(p+q)}$. Naturally, we have to assume, that Σ_X and Σ_Y are nonsingular since otherwise the affinely invariant distance covariance (3.1.3) does not exist. For the case in which $p = q = 1$, i.e., the bivariate normal distribution, the problem was solved by Székely, et al. in [102] and is stated in Theorem 2.4.10. In that case, the formula for the affinely invariant distance correlation depends only on ρ, the correlation coefficient, and appears in terms of the functions $\sin^{-1}\rho$ and $(1-\rho^2)^{1/2}$. Using equation (2.3.15), this result can be expressed as

$$\widetilde{\mathcal{R}}^2(X, Y) = \frac{{}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; \frac{1}{2}; \rho^2\right) - 2\,{}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; \frac{1}{2}; \frac{1}{4}\rho^2\right) + 1}{1 + \pi/3 - \sqrt{3}},$$

where ${}_2F_1$ is the generalized ${}_2F_1$-hypergeometric function.

We will see in Corollary 3.2.6, that the general case can be stated in terms of generalized hypergeometric functions of matrix arguments (see Definition 2.3.9), representing natural generalizations of these functions. Furthermore, while the bivariate result is just a function of the correlation coefficient ρ, we will see that the general result is a symmetric function of the canonical correlation coefficients $\lambda_1, \ldots, \lambda_p$ between the random vectors X and Y. The proof of the theorems of these section will make heavy use of the theory of zonal polynomials. We refer the reader to Section 2.3 and the chapter 7 of Muirhead [68] for further details.

It will prove useful to define Loewner's partial ordering for symmetric matrices.

Definition 3.2.1. *Let $S(p)$ denote the space of symmetric $p \times p$-matrices. Then Loewner's partial ordering on $S(p)$ is defined by*

$$A \geq_\ell B \Leftrightarrow A - B \text{ is positive semi-definite.}$$

The following Lemma is well-known (see for example [93]).

Lemma 3.2.2. *For symmetric matrices $A, B \in S(p)$, $A \geq_\ell B$ implies $\alpha_i \geq \beta_i$, $i = 1, \ldots, p$, where $\alpha_1 \geq \ldots \geq \alpha_p$ and $\beta_1 \geq \ldots \geq \beta_p$ are the eigenvalues of A and B respectively.*

As an immediate consequence of 3.2.2, we get

Lemma 3.2.3. *For symmetric matrices $A, B \in S(p)$, $A \geq_\ell B$ implies*

(i) $\det(A) \geq \det(B)$

(ii) $\operatorname{tr}(A) \geq \operatorname{tr}(B)$

(iii) $\|A\| \geq \|B\|$.

Theorem 3.2.4 states the key result of this section which obtains an explicit formula for the affinely invariant distance covariance in the case of a Gaussian population of arbitrary dimension and arbitrary covariance matrix with positive definite marginal covariance matrices.

Theorem 3.2.4. *Suppose that $(X, Y) \sim \mathcal{N}_{p+q}(\mu, \Sigma)$, where*

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$$

with $\Sigma_X \in \mathbb{R}^{p\times p}$, $\Sigma_Y \in \mathbb{R}^{q\times q}$, and $\Sigma_{XY} \in \mathbb{R}^{p\times q}$. Then

$$\widetilde{\mathcal{V}}^2(X, Y) = 4\pi \, \frac{c_{p-1}}{c_p} \frac{c_{q-1}}{c_q} \sum_{k=1}^{\infty} \frac{2^{2k} - 2}{k!\, 2^{2k}} \, \frac{(\frac{1}{2})_k \, (-\frac{1}{2})_k \, (-\frac{1}{2})_k}{(\frac{1}{2}p)_k \, (\frac{1}{2}q)_k} \, C_{(k)}(\Lambda), \tag{3.2.1}$$

where

$$\Lambda = \Sigma_Y^{-1/2} \, \Sigma_{YX} \, \Sigma_X^{-1} \, \Sigma_{XY} \, \Sigma_Y^{-1/2} \in \mathbb{R}^{q\times q}. \tag{3.2.2}$$

Proof. We may assume, with no loss of generality, that μ is the zero vector. Since Σ_X and Σ_Y both are positive definite the inverse square-roots, $\Sigma_X^{-1/2}$ and $\Sigma_Y^{-1/2}$, exist. By considering the standardized variables $\widetilde{X} = \Sigma_X^{-1/2} X$ and $\widetilde{Y} = \Sigma_Y^{-1/2} Y$, we may replace the covariance matrix Σ by

$$\widetilde{\Sigma} = \begin{pmatrix} I_p & \Lambda_{XY} \\ \Lambda_{XY}{}' & I_q \end{pmatrix},$$

where

$$\Lambda_{XY} = \Sigma_X^{-1/2} \, \Sigma_{XY} \, \Sigma_Y^{-1/2}. \tag{3.2.3}$$

Once we have made these reductions, it follows that the matrix Λ in (3.2.2) can be written as $\Lambda = {\Lambda_{XY}}'\Lambda_{XY}$ and that it has norm less than 1. Indeed, by the partial Iwasawa decomposition [54, 8] of $\widetilde{\Sigma}$, viz., the identity,

$$\widetilde{\Sigma} = \begin{pmatrix} I_p & 0 \\ {\Lambda_{XY}}' & I_q \end{pmatrix} \begin{pmatrix} I_p & 0 \\ 0 & I_q - {\Lambda_{XY}}'\Lambda_{XY} \end{pmatrix} \begin{pmatrix} I_p & \Lambda_{XY} \\ 0 & I_q \end{pmatrix},$$

where the zero matrix of any dimension is denoted by 0, we see that the matrix $\widetilde{\Sigma}$ is positive semidefinite if and only if $I_q - \Lambda$ is positive semidefinite. Hence, $\Lambda \leq_\ell I_q$ in the Loewner ordering and therefore $\|\Lambda\| \leq 1$ by Lemma 3.2.3.

We proceed to calculate the distance covariance $\widetilde{\mathcal{V}}(X,Y) = \mathcal{V}(\widetilde{X},\widetilde{Y})$. It is well-known [1] that the characteristic function of $(\widetilde{X},\widetilde{Y})$ is

$$f_{\widetilde{X},\widetilde{Y}}(s,t) = \exp\Big[-\tfrac{1}{2}\begin{pmatrix} s \\ t \end{pmatrix}'\widetilde{\Sigma}\begin{pmatrix} s \\ t \end{pmatrix}\Big] = \exp\big[-\tfrac{1}{2}(|s|_p^2 + |t|_q^2 + 2s'\Lambda_{XY}t)\big],$$

where $s \in \mathbb{R}^p$ and $t \in \mathbb{R}^q$. Therefore,

$$\big|f_{\widetilde{X},\widetilde{Y}}(s,t) - f_{\widetilde{X}}(s)f_{\widetilde{Y}}(t)\big|^2 = \big(1 - \exp(-s'\Lambda_{XY}t)\big)^2 \exp(-|s|_p^2 - |t|_q^2),$$

and hence

$$\begin{aligned} c_p c_q\, \mathcal{V}^2(\widetilde{X},\widetilde{Y}) &= \int_{\mathbb{R}^{p+q}} \big(1 - \exp(-s'\Lambda_{XY}t)\big)^2 \exp(-|s|_p^2 - |t|_q^2)\, \frac{\mathrm{d}s}{|s|_p^{p+1}}\, \frac{\mathrm{d}t}{|t|_q^{q+1}} \\ &= \int_{\mathbb{R}^{p+q}} \big(1 - \exp(s'\Lambda_{XY}t)\big)^2 \exp(-|s|_p^2 - |t|_q^2)\, \frac{\mathrm{d}s}{|s|_p^{p+1}}\, \frac{\mathrm{d}t}{|t|_q^{q+1}}, \end{aligned} \tag{3.2.4}$$

where the latter integral is obtained by making the change of variables $s \mapsto -s$ within the former integral.

By a Taylor series expansion, we obtain

$$\begin{aligned} \big(1 - \exp(s'\Lambda_{XY}t)\big)^2 &= 1 - 2\exp(s'\Lambda_{XY}t) + \exp(2s'\Lambda_{XY}t) \\ &= \sum_{k=2}^{\infty} \frac{2^k - 2}{k!} (s'\Lambda_{XY}t)^k. \end{aligned}$$

Substituting this series into (3.2.4) and interchanging summation and integration, a procedure which is straightforward to verify by means of Fubini's theorem, and noting that the odd-order terms integrate to zero, we obtain

$$c_p c_q\, \mathcal{V}^2(\widetilde{X},\widetilde{Y}) = \sum_{k=1}^{\infty} \frac{2^{2k} - 2}{(2k)!} \int_{\mathbb{R}^{p+q}} (s'\Lambda_{XY}t)^{2k} \exp(-|s|_p^2 - |t|_q^2)\, \frac{\mathrm{d}s}{|s|_p^{p+1}}\, \frac{\mathrm{d}t}{|t|_q^{q+1}}. \tag{3.2.5}$$

To calculate, for $k \geq 1$, the integral

$$\int_{\mathbb{R}^{p+q}} (s'\Lambda_{XY}t)^{2k} \exp(-|s|_p^2 - |t|_q^2) \frac{\mathrm{d}s}{|s|_p^{p+1}} \frac{\mathrm{d}t}{|t|_q^{q+1}}, \tag{3.2.6}$$

we change variables to polar coordinates, putting $s = r_x\theta$ and $t = r_y\phi$ where $r_x, r_y > 0$, $\theta = (\theta_1, \ldots, \theta_p)' \in S^{p-1}$, and $\phi = (\phi_1, \ldots, \phi_q)' \in S^{q-1}$. By Theorem 2.1.3, the integral (3.2.6) reads:

$$\int_0^\infty \int_0^\infty \int_{S^{q-1}} \int_{S^{p-1}} r_x^{2k-2} r_y^{2k-2} \exp(-r_x^2 - r_y^2)(\theta'\Lambda_{XY}\phi)^{2k} \mathrm{d}\theta \mathrm{d}\phi \mathrm{d}r_x \mathrm{d}r_y,$$

which obviously separates into a product of multiple integrals over (r_x, r_y), and over (θ, ϕ), respectively. The integrals over r_x and r_y are standard gamma integrals,

$$\begin{aligned}\int_0^\infty \int_0^\infty r_x^{2k-2} r_y^{2k-2} \exp(-r_x^2 - r_y^2)\, \mathrm{d}r_x \mathrm{d}r_y &= \tfrac{1}{4} \int_0^\infty \int_0^\infty u_x^{k-3/2} u_y^{k-3/2} \exp(-u_x - u_y)\, \mathrm{d}u_x \mathrm{d}u_y \\ &= \tfrac{1}{4} \left[\Gamma(k - \tfrac{1}{2})\right]^2 = \left[(-\tfrac{1}{2})_k\right]^2 \pi,\end{aligned}$$

where the first transformation follows by substituting $u_x = r_x^2$, $u_y = r_y^2$.

The remaining factor is the integral

$$\int_{S^{q-1}} \int_{S^{p-1}} (\theta'\Lambda_{XY}\phi)^{2k} \mathrm{d}\theta \mathrm{d}\phi, \tag{3.2.7}$$

where $\mathrm{d}\theta$ and $\mathrm{d}\phi$ are unnormalized surface measures on S^{p-1} and S^{q-1}, respectively. By a standard invariance argument,

$$\int_{S^{p-1}} (\theta'v)^{2k} \mathrm{d}\theta = |v|_p^{2k} \int_{S^{p-1}} \theta_1^{2k} \mathrm{d}\theta,$$

$v \in \mathbb{R}^p$. Indeed, denoting this integral by $g(v)$, it follows by (2.1.2) that $g(v) = g(Hv)$ for all $H \in O(p)$. By choosing H to be a specific orthogonal matrix such that $Hv = (|v|_p, 0, \ldots, 0)'$ we obtain

$$\begin{aligned} g(v) &= g\big((|v|_p, 0, \ldots, 0)'\big) \\ &= \int_{S^{p-1}} (\theta_1 |v|_p)^{2k} \mathrm{d}\theta \\ &= |v|_p^{2k} \int_{S^{p-1}} \theta_1^{2k} \mathrm{d}\theta. \end{aligned}$$

Setting $v = \Lambda_{XY}\phi$, we obtain

$$\int_{S^{q-1}} \int_{S^{p-1}} (\theta' \Lambda_{XY} \phi)^{2k} \, \mathrm{d}\theta \, \mathrm{d}\phi \quad = \int_{S^{q-1}} |\Lambda_{XY}\phi|_p^{2k} \int_{S^{p-1}} \theta_1^{2k} \, \mathrm{d}\theta \, \mathrm{d}\phi. \tag{3.2.8}$$

$$= \int_{S^{q-1}} |\Lambda_{XY}\phi|_p^{2k} \gamma_{p,k} \, \mathrm{d}\phi, \tag{3.2.9}$$

with

$$\gamma_{p,k} = \int_{S^{p-1}} \theta_1^{2k} \, \mathrm{d}\theta.$$

To evaluate $\gamma_{p,k}$, we make the following considerations. Let $V = (V_1, \ldots, V_p)' \sim \mathcal{N}_p(0, I_p)$; by Muirhead [68, Theorem 1.5.7], the random vector $V/(V'V)^{1/2}$ is uniformly distributed on the sphere S^{p-1}. Let $U = V_1/(V'V)^{1/2}$; writing

$$U^2 = \frac{V_1^2}{V'V} \equiv \frac{V_1^2}{V_1^2 + (V_2^2 + \cdots + V_p^2)},$$

and noting that $V_2^2 + \cdots + V_p^2 \sim \chi_{p-1}^2$ independently of $V_1^2 \sim \chi_1^2$, it follows that $U^2 \sim \mathrm{Beta}(\frac{1}{2}, \frac{1}{2}(p-1))$, a beta distribution. Hence,

$$\gamma_{p,k} = \int_{S^{p-1}} \mathrm{d}\theta \, E(U^{2k}) = 2c_{p-1} \frac{\Gamma(k + \frac{1}{2})\Gamma(\frac{1}{2}p)}{\Gamma(k + \frac{1}{2}p)\Gamma(\frac{1}{2})} = 2c_{p-1} \frac{(\frac{1}{2})_k}{(\frac{1}{2}p)_k}, \tag{3.2.10}$$

since $2c_{p-1} = \frac{2\pi^{p/2}}{\Gamma(p/2)}$ is the surface area of S^{p-1} (see Remark 2.1.4), the remaining factor follows from the well-known moments of the beta distribution.

Therefore, in order to evaluate (3.2.9), it remains to evaluate

$$J_k(\Lambda) = \int_{S^{q-1}} |\Lambda_{XY}\phi|_p^{2k} \, \mathrm{d}\phi = \int_{S^{q-1}} (\phi' \Lambda \phi)^k \, \mathrm{d}\phi.$$

By the invariance of the surface measure under orthogonal transformations (see 2.1.2), it follows that $J_k(\Lambda) = J_k(K'\Lambda K)$ for all $K \in O(q)$. Integrating with respect to the normalized Haar measure on the orthogonal group, we conclude by (2.1.1) that

$$J_k(\Lambda) = \int_{O(q)} J_k(K'\Lambda K) \, \mathrm{d}K = \int_{S^{q-1}} \int_{O(q)} (\phi' K' \Lambda K \phi)^k \, \mathrm{d}K \, \mathrm{d}\phi. \tag{3.2.11}$$

We now make use of some properties of the zonal polynomials introduced in Section 2.3. By (2.3.2),

$$(\phi' K' \Lambda K \phi)^k = (\operatorname{tr} K'\Lambda K \phi\phi')^k = \sum_{|\kappa|=k} C_\kappa(K'\Lambda K \phi\phi'),$$

where $C_\kappa(K'\Lambda K \phi\phi')$ is meant to be understood in the sense of (2.3.4). Therefore, by

(2.3.5),

$$\int_{O(q)} (\phi' K' \Lambda K \phi)^k \, \mathrm{d}K = \sum_{|\kappa|=k} \int_{O(q)} C_\kappa(K' \Lambda K \phi\phi') \, \mathrm{d}K = \sum_{|\kappa|=k} \frac{C_\kappa(\Lambda)\, C_\kappa(\phi\phi')}{C_\kappa(I_q)}.$$

Since $\phi\phi'$ is of rank 1 then, by Remark 2.3.5, $C_\kappa(\phi\phi') = 0$ if $\ell(\kappa) > 1$; it now follows, by (2.3.2) and the fact that $\phi \in S^{q-1}$, that

$$C_{(k)}(\phi\phi') = \sum_{|\kappa|=k} C_\kappa(\phi\phi') = (\operatorname{tr} \phi\phi')^k = (\phi'\phi)^k = |\phi|_q^{2k} = 1.$$

Therefore,

$$\int_{O(q)} (\phi' K' \Lambda K \phi)^k \, \mathrm{d}K = \frac{C_{(k)}(\Lambda)}{C_{(k)}(I_q)} = \frac{(\frac{1}{2})_k}{(\frac{1}{2}q)_k} \, C_{(k)}(\Lambda),$$

where the last equality follows by (2.3.7). Substituting this result at (3.2.11), we obtain

$$J_k(\Lambda) = 2c_{q-1} \frac{(\frac{1}{2})_k}{(\frac{1}{2}q)_k} \, C_{(k)}(\Lambda).$$

Collecting together these results, we obtain

$$\begin{aligned} \widetilde{\mathcal{V}}^2(X,Y) &= \frac{1}{c_p\, c_q} \sum_{k=1}^{\infty} \frac{2^{2k}-2}{(2k)!} \big(\big[(-\tfrac{1}{2})_k \big]^2 \pi \big) \, \gamma_{p,k} \, J_k(\Lambda) \\ &= 4\pi \, \frac{c_{p-1}}{c_p} \frac{c_{q-1}}{c_q} \sum_{k=1}^{\infty} \frac{2^{2k}-2}{(2k)!} \, \frac{(\frac{1}{2})_k \, (\frac{1}{2})_k \, (-\frac{1}{2})_k \, (-\frac{1}{2})_k}{(\frac{1}{2}p)_k \, (\frac{1}{2}q)_k} \, C_{(k)}(\Lambda). \end{aligned}$$

By using the identity $(2k)! = k!\, 2^{2k}\, (\frac{1}{2})_k$, we obtain the representation (3.2.1), as desired. □

Remark 3.2.5. *By Theorem 3.2.4, we see, that $\tilde{\mathcal{V}}(X,Y)$ is just a function depending only on the dimensions p and q and the eigenvalues of the matrix Λ, i.e. the squared canonical correlation coefficients of the subvectors X and Y. For fixed dimensions this implies $\widetilde{\mathcal{R}}(X,Y) = g(\lambda_1, \ldots, \lambda_r)$, where $r = \min(p,q)$ and $\lambda_1, \ldots, \lambda_r$ are the canonical correlation coefficients of X and Y. Due to the functional invariance the maximum likelihood estimator (MLE) for affinely invariant distance correlation in the Gaussian setting is hence defined by $g(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)$, where $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r$ are the MLEs of the canonical correlation coefficients.*

Let us note, that by interchanging the roles of X and Y in Theorem 3.2.4, we would obtain (3.2.1) with Λ in (3.2.2) replaced by

$$\Lambda_0 = \Sigma_X^{-1/2} \, \Sigma_{XY} \, \Sigma_Y^{-1} \, \Sigma_{YX} \, \Sigma_X^{-1/2} \in \mathbb{R}^{p \times p}.$$

Since Λ and Λ_0 have the same characteristic polynomial and hence the same set of nonzero eigenvalues, and noting that $C_\kappa(\Lambda)$ depends only on the eigenvalues of Λ, it follows that $C_{(k)}(\Lambda) = C_{(k)}(\Lambda_0)$. Therefore, the series representation (3.2.1) for $\widetilde{\mathcal{V}}^2(X,Y)$ remains unchanged if the roles of X and Y are interchanged.

Corollary 3.2.6. *In the setting of Theorem 3.2.4, we have*

$$\widetilde{\mathcal{V}}^2(X,Y) = 4\pi \frac{c_{p-1}}{c_p}\frac{c_{q-1}}{c_q} \times \Big({}_3F_2\big(\tfrac{1}{2},-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p,\tfrac{1}{2}q;\Lambda\big) - 2\,{}_3F_2\big(\tfrac{1}{2},-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p,\tfrac{1}{2}q;\tfrac{1}{4}\Lambda\big) + 1\Big). \quad (3.2.12)$$

Proof. It is evident that the partitional rising factorial introduced in (2.3.8) satisfies

$$(\tfrac{1}{2})_\kappa = \begin{cases} (\tfrac{1}{2})_{k_1}, & \text{if } \ell(\kappa) \le 1, \\ 0, & \text{if } \ell(\kappa) > 1. \end{cases}$$

Therefore, we now can write the series in (3.2.1), up to a multiplicative constant, in terms of a generalized hypergeometric function of matrix argument, in that

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{2^{2k}-2}{k!\,2^{2k}} \frac{(\frac{1}{2})_k(-\frac{1}{2})_k(-\frac{1}{2})_k}{(\frac{1}{2}p)_k(\frac{1}{2}q)_k} C_{(k)}(\Lambda) \\
&\quad = \sum_{k=1}^{\infty} \frac{2^{2k}-2}{k!\,2^{2k}} \sum_{|\kappa|=k} \frac{(\frac{1}{2})_\kappa(-\frac{1}{2})_\kappa(-\frac{1}{2})_\kappa}{(\frac{1}{2}p)_\kappa(\frac{1}{2}q)_\kappa} C_\kappa(\Lambda) \\
&\quad = \sum_{k=1}^{\infty} \frac{1}{k!} \sum_{|\kappa|=k} \frac{(\frac{1}{2})_\kappa(-\frac{1}{2})_\kappa(-\frac{1}{2})_\kappa}{(\frac{1}{2}p)_\kappa(\frac{1}{2}q)_\kappa} C_\kappa(\Lambda) - 2\sum_{k=1}^{\infty} \frac{1}{k!\,2^{2k}} \sum_{|\kappa|=k} \frac{(\frac{1}{2})_\kappa(-\frac{1}{2})_\kappa(-\frac{1}{2})_\kappa}{(\frac{1}{2}p)_\kappa(\frac{1}{2}q)_\kappa} C_\kappa(\Lambda) \\
&\quad = \left[{}_3F_2\left(\tfrac{1}{2},-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p,\tfrac{1}{2}q;\Lambda\right) - 1\right] - 2\left[{}_3F_2\left(\tfrac{1}{2},-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p,\tfrac{1}{2}q;\tfrac{1}{4}\Lambda\right) - 1\right].
\end{aligned}$$

Due to property (2.3.1) it remains to show that the zonal polynomial series expansion for the ${}_3F_2\left(\frac{1}{2},-\frac{1}{2},-\frac{1}{2};\frac{1}{2}p,\frac{1}{2}q;\Lambda\right)$ generalized hypergeometric function of matrix argument converges absolutely for all Λ with $\Lambda \le_\ell I_q$ in the Loewner ordering. By (2.3.7)

$$\begin{aligned}
{}_3F_2\left(\tfrac{1}{2},-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p,\tfrac{1}{2}q;\Lambda\right) &\le \sum_{k=0}^{\infty} \frac{2^{2k}}{k!\,2^{2k}} \frac{(-\frac{1}{2})_k(-\frac{1}{2})_k}{(\frac{1}{2}p)_k} \\
&= {}_2F_1\left(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p;1\right).
\end{aligned}$$

The latter series converges due to Theorem 2.3.10 and so we have absolute convergence at (3.2.12) for all Σ with positive definite marginal covariance matrices. □

Corollary 3.2.7. *Let us consider the setting of Theorem 3.2.4 and set $q = 1$. Then we have*

$$\widetilde{\mathcal{V}}^2(X,Y) = 4\frac{c_{p-1}}{c_p}\Big({}_2F_1\left(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p;\lambda\right) - 2\,{}_2F_1\left(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p;\tfrac{1}{4}\lambda\right) + 1\Big), \tag{3.2.13}$$

where $\lambda := \Lambda \in \mathbb{R}$.

PROOF. First note that $c_0{c_1}^{-1} = \pi^{-1}$. It is further evident, that

$${}_3F_2\big(\tfrac{1}{2},-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p,\tfrac{1}{2};s\big) = {}_2F_1\big(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p,;s\big)$$

□

For cases in which $q = 1$ and p is odd, we can obtain explicit expressions for $\widetilde{\mathcal{V}}^2(X,Y)$. In such cases, the affinely invariant distance covariance in (3.2.13) can be expressed as hypergeometric functions of the form ${}_2F_1(-\frac{1}{2},-\frac{1}{2};k+\frac{1}{2};\rho^2)$, $k \in \mathbb{N}$, and we have shown in Section 2.3, that these latter functions are expressible in closed form in terms of elementary functions.
Let us consider again the case in which $p = q = 1$. Then $\lambda = \rho^2$, where ρ is the Pearson correlation coefficient and (3.2.13) and (2.3.15) yield

$$\begin{aligned}\widetilde{\mathcal{V}}^2(X,Y) &= \frac{4}{\pi}\Big({}_2F_1\left(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2};\rho^2\right) - 2\,{}_2F_1\left(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2};\tfrac{1}{4}\rho^2\right) + 1\Big)\\ &= \frac{4}{\pi}\Big(\rho\sin^{-1}\rho + \sqrt{1-\rho^2} - \rho\sin^{-1}\rho/2 - \sqrt{4-\rho^2} + 1\Big).\end{aligned} \tag{3.2.14}$$

For $p = 3$, we see by (2.3.16)

$$\begin{aligned}\widetilde{\mathcal{V}}^2(X,Y) = \frac{8}{\pi}\Bigg(&\frac{3(1-\rho^2)^{1/2}}{4} + \frac{(1+2\rho^2)\sin^{-1}\rho}{4\rho}\\ &- \frac{3(1-(\rho^2/4))^{1/2}}{4} + \frac{(1+\rho^2)\sin^{-1}\rho/2}{2\rho} + 1\Bigg),\end{aligned} \tag{3.2.15}$$

where we set $\lambda = \rho^2$. Further, it is clear by (2.3.17), that, for $q = 1$ and p odd, the affinely invariant distance covariance $\widetilde{\mathcal{V}}^2(X,Y)$ can be expressed in closed form in terms of elementary functions and the $\sin^{-1}(\cdot)$ function.

The appearance of the generalized hypergeometric functions of matrix argument also yields a useful expression for the affinely invariant distance variance. In order to state this result, we shall define for each positive integer p the quantity

$$A(p) = \frac{\Gamma(\frac{1}{2}p)\,\Gamma(\frac{1}{2}p+1)}{\big[\Gamma\big(\frac{1}{2}(p+1)\big)\big]^2} - 2\,{}_2F_1\big(-\tfrac{1}{2},-\tfrac{1}{2};\tfrac{1}{2}p;\tfrac{1}{4}\big) + 1. \tag{3.2.16}$$

Corollary 3.2.8. *In the setting of Theorem 3.2.4, we have*

$$\widetilde{\mathcal{V}}^2(X,X) = 4\pi \frac{c_{p-1}^2}{c_p^2}\, A(p). \tag{3.2.17}$$

PROOF. We are in the special case of Theorem 3.2.4 for which $X = Y$, so that $p = q$ and $\Lambda = I_p$. By applying (2.3.7) we can write the series in (3.2.1) as

$$\begin{aligned} 4\pi \frac{c_{p-1}^2}{c_p^2} \sum_{k=1}^{\infty} \frac{2^{2k}-2}{k!\,2^{2k}} \frac{(\frac{1}{2})_k (-\frac{1}{2})_k (-\frac{1}{2})_k}{(\frac{1}{2}p)_k (\frac{1}{2}p)_k}\, C_{(k)}(I_p) \\ = 4\pi \frac{c_{p-1}^2}{c_p^2} \sum_{k=1}^{\infty} \frac{2^{2k}-2}{k!\,2^{2k}} \frac{(-\frac{1}{2})_k (-\frac{1}{2})_k}{(\frac{1}{2}p)_k} \\ = 4\pi \frac{c_{p-1}^2}{c_p^2} \Big(\big[{}_2F_1\left(-\tfrac{1}{2}, -\tfrac{1}{2}; \tfrac{1}{2}p; 1\right) - 1 \big] - 2 \big[{}_2F_1\left(-\tfrac{1}{2}, -\tfrac{1}{2}; \tfrac{1}{2}p; \tfrac{1}{4}\right) - 1 \big] \Big). \end{aligned}$$

By Theorem 2.3.10, the series ${}_2F_1(-\frac{1}{2}, -\frac{1}{2}; \frac{1}{2}p; z)$ also converges for the special value $z = 1$, and then

$${}_2F_1(-\tfrac{1}{2}, -\tfrac{1}{2}; \tfrac{1}{2}p; 1) = \frac{\Gamma(\frac{1}{2}p)\,\Gamma(\frac{1}{2}p+1)}{\big[\Gamma\big(\frac{1}{2}(p+1)\big)\big]^2},$$

thereby completing the proof. □

For cases in which p is odd, we can again obtain explicit values by using (2.3.15) and the contiguous relation (2.3.13). This leads in such cases to explicit expressions for the exact value of $\widetilde{\mathcal{V}}^2(X,X)$. In particular, if $p = 1$ then it follows from (2.4.2) and (2.3.15) that

$$\widetilde{\mathcal{V}}^2(X,X) = \frac{4}{3} - \frac{4(\sqrt{3}-1)}{\pi}; \tag{3.2.18}$$

and for $p = 3$, we deduce from (2.4.2) and (2.3.16) that

$$\widetilde{\mathcal{V}}^2(X,X) = 2 - \frac{4(3\sqrt{3}-4)}{\pi}.$$

Hence, for $q = 1$ and p odd, (3.2.13) and the latter observation allows us to state the affinely invariant distance correlation $\tilde{\mathcal{R}}(X,Y)$ in terms of elementary functions. In particular, combining (3.2.14) and (3.2.18) yields the result (2.4.10) stated by Székely. But even for cases where $\tilde{\mathcal{R}}(X,Y)$ cannot be explicitly obtained, the representation of the affinely invariant distance covariance and variance in Corollaries 3.2.6 and 3.2.8 enable the explicit and efficient calculation of the affinely invariant distance correlation (3.1.4). For calculating the graphs in our illustrations, we use the algorithm of Koev and Edelman [50] to evaluate the generalized hypergeometric function of matrix argument,

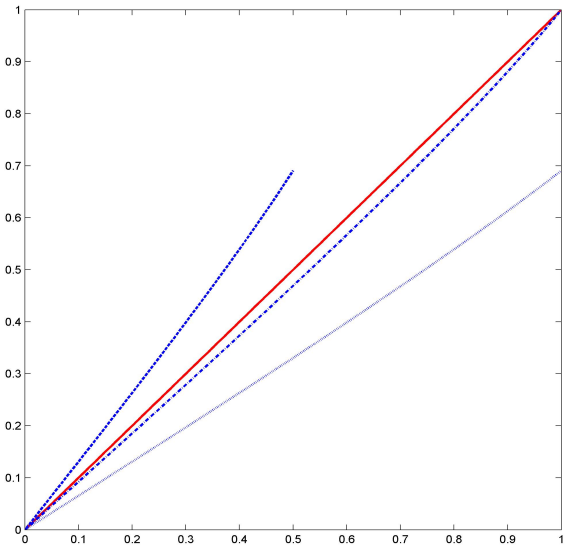

Figure 3.1: The affinely invariant distance correlation for subvectors of a multivariate normal population, where $p = q = 2$, as a function of the parameter r in three distinct settings. The solid diagonal line is the identity function and is provided to serve as a reference for the three distance correlation functions. See the text for details.

with C and Matlab code being available at these authors' websites.

Figure 3.1 concerns the case $p = q = 2$ in various settings, in which the matrix Λ_{XY} depends on a single parameter r only. The dotted line shows the affinely invariant distance correlation when

$$\Lambda_{XY} = \begin{pmatrix} 0 & 0 \\ 0 & r \end{pmatrix};$$

this is the case with the weakest dependence considered here. The dash-dotted line applies when

$$\Lambda_{XY} = \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}.$$

The strongest dependence corresponds to the dashed line, which shows the affinely invariant distance correlation when

$$\Lambda_{XY} = \begin{pmatrix} r & r \\ r & r \end{pmatrix};$$

in this case we need to assume that $0 \leq r \leq \frac{1}{2}$ in order to retain positive definiteness. In Figure 3.2, panel (a) shows the affinely invariant distance correlation when $p = q = 2$ and

$$\Lambda_{XY} = \begin{pmatrix} r & 0 \\ 0 & s \end{pmatrix},$$

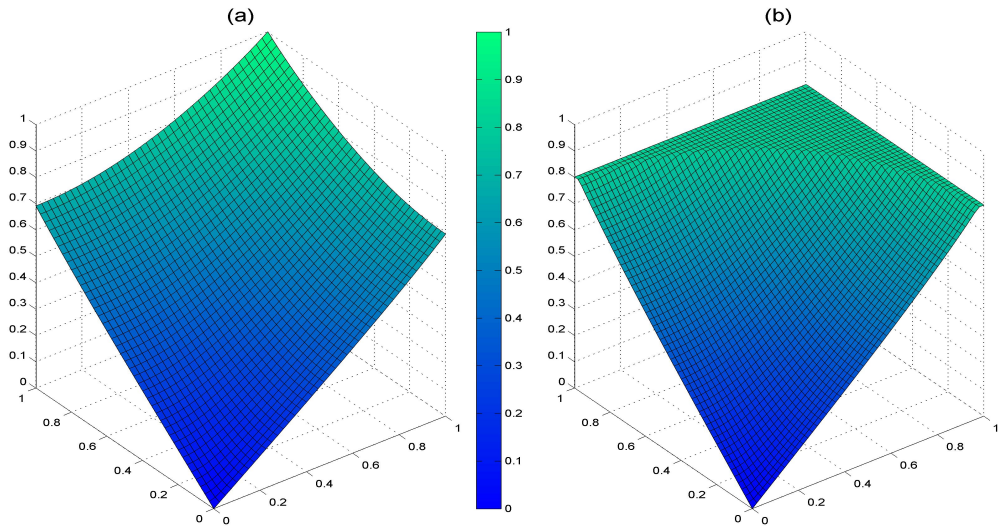

Figure 3.2: The affinely invariant distance correlation between the p- and q-dimensional subvectors of a $(p+q)$-dimensional multivariate normal population, where (a) $p = q = 2$ and $\Lambda_{XY} = \mathrm{diag}(r, s)$, and (b) $p = 2$, $q = 1$ and $\Lambda_{XY} = (r, s)'$.

where $0 \leq r, s \leq 1$. With reference to Figure 3.1, the margins correspond to the dotted line and the diagonal corresponds to the dash-dotted line.
Panel (b) of Figure 3.2 concerns the case in which $p = 2$, $q = 1$ and $\Lambda_{XY} = (r, s)'$, where $r^2 + s^2 \leq 1$. Here, the affinely invariant distance correlation attains an upper limit as $r^2 + s^2 \uparrow 1$, and we have evaluated that limit numerically as 0.8252.

Note, that the exact formula given for the affinely invariant distance covariance in Theorem 3.2.4 opens up the possibility to analytically study this measure in the case of a multivariate normal population. In the following, we will use this result to investigate the asymptotic behavior of the affinely invariant distance covariance and the affinely invariant distance correlation, in particular in high-dimensional settings, where the dimension p and q go to infinity.

3.3 Limit Theorems

In the course of this section, we study the limiting behavior of the affinely invariant distance correlation measures for subvectors of multivariate normal populations.

Our first result quantifies the asymptotic decay of the affinely invariant distance correlation in the case in which the cross-covariance matrix converges to the zero matrix, in

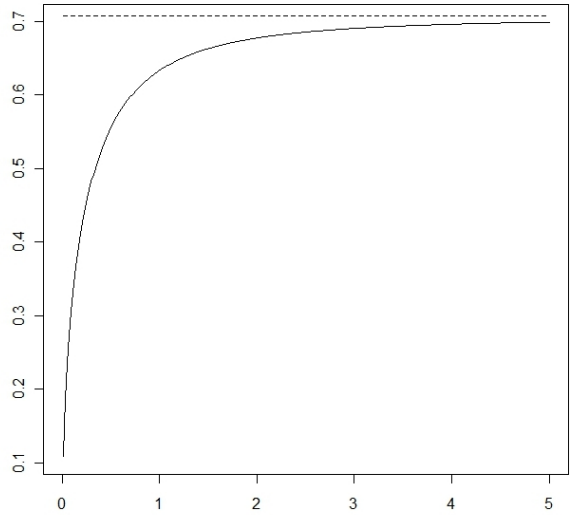

Figure 3.3: The affinely invariant distance variance $\mathcal{V}(X, X)$ subject to the dimension p. The horizontal line represents the level $\sqrt{2}$.

that

$$\mathrm{tr}\,(\Lambda) = \|\Lambda_{XY}\|_F^2 \longrightarrow 0,$$

where the matrices $\Lambda = \Lambda_{XY}{}'\Lambda_{XY}$ and Λ_{XY} are defined in (3.2.2) and (3.2.3), respectively.

Theorem 3.3.1. *Suppose that* $(X, Y) \sim \mathcal{N}_{p+q}(\mu, \Sigma)$*, where*

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$$

with $\Sigma_X \in \mathbb{R}^{p\times p}$ *and* $\Sigma_Y \in \mathbb{R}^{q\times q}$ *being positive definite, and suppose that the matrix* Λ *in (3.2.2) has positive trace. Then,*

$$\lim_{\mathrm{tr}\,(\Lambda)\to 0} \frac{\widetilde{\mathcal{R}}^2(X, Y)}{\mathrm{tr}\,(\Lambda)} = \frac{1}{4\,pq\sqrt{A(p)A(q)}}, \tag{3.3.1}$$

where $A(p)$ *is defined in (3.2.16).*

Proof. By Theorem 3.2.4, $\widetilde{\mathcal{V}}^2(X, Y)$ is given by:

$$\widetilde{\mathcal{V}}^2(X, Y) = 4\pi\,\frac{c_{p-1}}{c_p}\frac{c_{q-1}}{c_q}\sum_{k=1}^{\infty}\frac{2^{2k}-2}{k!\,2^{2k}}\,\frac{(\frac{1}{2})_k\,(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k\,(\frac{1}{2}q)_k}\,C_{(k)}(\Lambda). \tag{3.3.2}$$

We further note that $\widetilde{\mathcal{V}}^2(X, X)$ and $\widetilde{\mathcal{V}}^2(Y, Y)$ do not depend on Σ_{XY}, as can be seen from their explicit representations in terms of $A(p)$ and $A(q)$ given in (3.2.17). In studying

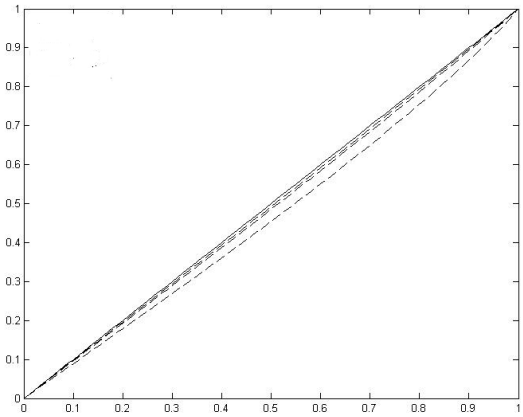

Figure 3.4: The affinely invariant distance variance between two bivariate random variables with $\Lambda_{XY} = r\, I_p$. The solid line represents the identity. The dashed lines sketch the distance correlation for $p = 1, 5, 10$.

the asymptotic behavior of $\widetilde{\mathcal{V}}^2(X, Y)$, we may interchange the limit and the summation in the series representation (3.3.2). Hence, it suffices to find the limit term-by-term. Since $C_{(1)}(\Lambda) = \mathrm{tr}\,(\Lambda)$ then the ratio of the term for $k = 1$ and $\mathrm{tr}\,(\Lambda)$ equals

$$\frac{c_{p-1}}{c_p}\frac{c_{q-1}}{c_q}\frac{\pi}{pq}.$$

For $k \geq 2$, it follows from (2.3.6) that $C_{(k)}(\Lambda)$ is a sum of monomials in the eigenvalues of Λ, with each monomial being of degree k, which is greater than the degree, viz. 1, of $\mathrm{tr}\,(\Lambda)$; therefore,

$$\lim_{\mathrm{tr}\,(\Lambda)\to 0} \frac{C_{(k)}(\Lambda)}{\mathrm{tr}\,(\Lambda)} = \lim_{\Lambda \to 0} \frac{C_{(k)}(\Lambda)}{\mathrm{tr}\,(\Lambda)} = 0. \tag{3.3.3}$$

Collecting these facts together, we find

$$\lim_{\mathrm{tr}\,(\Lambda)\to 0} \frac{\widetilde{\mathcal{R}}^2(X, Y)}{\mathrm{tr}\,(\Lambda)} = \frac{\frac{c_{p-1}}{c_p}\frac{c_{q-1}}{c_q}\frac{\pi}{pq}}{\widetilde{\mathcal{V}}(X, X)\,\widetilde{\mathcal{V}}(Y, Y)} = \frac{1}{4\,pq\sqrt{A(p)A(q)}}.$$

□

If $p = q = 1$ we are in the situation of Theorem 2.4.10 . Applying the identity (2.3.15), we obtain

$${}_2F_1(-\tfrac{1}{2}, -\tfrac{1}{2}; \tfrac{1}{2}; \tfrac{1}{4}) = \frac{\pi}{12} + \frac{\sqrt{3}}{2},$$

and $(\mathrm{tr}\,(\Lambda))^{1/2} = |\rho|$. Thus we obtain

$$\lim_{\rho \to 0} \frac{\widetilde{\mathcal{R}}(X, Y)}{|\rho|} = \frac{1}{2\left(1 + \frac{1}{3}\pi - \sqrt{3}\right)^{1/2}},$$

analogously to Theorem 2.4.10 (iii).

In the remainder of this section we consider situations in which one or both of the dimensions p and q grow without bound. We will repeatedly make use the following lemma.

Lemma 3.3.2. *Let c_p be defined as in (2.4.2), then*

$$\frac{c_{p-1}}{\sqrt{p}\, c_p} \longrightarrow \frac{1}{\sqrt{2\pi}} \tag{3.3.4}$$

as $p \to \infty$.

Proof. By the functional equation for the gamma function (2.2.2), we find that

$$\frac{c_{p-1}^2}{p\, c_p^2} = \frac{[\Gamma(\frac{p+1}{2})]^2}{2\pi\, \Gamma(\frac{p}{2})\, \Gamma(\frac{p}{2}+1)},$$

Now, by Stirling's approximation (Theorem 2.2.4)

$$\begin{aligned}
&\frac{1}{2\,\pi} \lim_{p\to\infty} \frac{\Gamma(\frac{p+1}{2})^2}{\Gamma(\frac{p}{2})\Gamma(\frac{p}{2}+1)} \\
&= \frac{1}{2\,\pi} \lim_{p\to\infty} \frac{\sqrt{(p/2+1)p/2}\Big(\frac{(p+1)/2}{e}\Big)^{p/2+1}\Big(\frac{(p+1)/2}{e}\Big)^{p/2}}{(p+1)/2\Big(\frac{p/2+1}{e}\Big)^{p/2+1}\Big(\frac{p/2}{e}\Big)^{p/2}} \\
&= \frac{1}{2\,\pi} \lim_{p\to\infty} \frac{\sqrt{(p+2)p}}{p+1} \lim_{p\to\infty} \Big(\frac{p+1}{p+2}\Big)^{p/2+1} \lim_{p\to\infty} \Big(\frac{p+1}{p}\Big)^{p/2} \\
&= \frac{1}{2\,\pi}\, e^{-1/2}\, e^{1/2} = \frac{1}{2\,\pi}.
\end{aligned}$$

□

Theorem 3.3.3. *For each positive integer p, suppose that $(X_p, Y_p) \sim \mathcal{N}_{2p}(\mu_p, \Sigma_p)$, where*

$$\Sigma_p = \begin{pmatrix} \Sigma_{X,\,p} & \Sigma_{XY,\,p} \\ \Sigma_{YX,\,p} & \Sigma_{Y,\,p} \end{pmatrix}$$

with $\Sigma_{X,\,p} \in \mathbb{R}^{p\times p}$ and $\Sigma_{Y,\,p} \in \mathbb{R}^{p\times p}$ being positive definite and such that

$$\Lambda_p = \Sigma_{Y,\,p}^{-1/2}\, \Sigma_{YX,\,p}\, \Sigma_{X,\,p}^{-1}\, \Sigma_{XY,\,p}\, \Sigma_{Y,\,p}^{-1/2} \neq 0.$$

Then

$$\lim_{p\to\infty} \frac{p}{\operatorname{tr}(\Lambda_p)}\, \widetilde{\mathcal{V}}^2(X_p, Y_p) = \frac{1}{2} \tag{3.3.5}$$

and

$$\lim_{p\to\infty} \frac{p}{\operatorname{tr}(\Lambda_p)}\, \widetilde{\mathcal{R}}^2(X_p, Y_p) = 1. \tag{3.3.6}$$

In particular, if $\Lambda_p = r^2 I_p$ for some $r \in [0,1]$, then $\operatorname{tr}(\Lambda_p) = r^2 p$, and so (3.3.5) and (3.3.6) reduce to

$$\lim_{p\to\infty} \widetilde{\mathcal{V}}^2(X_p, Y_p) = \frac{1}{2}\, r^2 \qquad \text{and} \qquad \lim_{p\to\infty} \widetilde{\mathcal{R}}(X_p, Y_p) = r,$$

respectively. The following corollary concerns the special case in which $r = 1$; we state it separately for emphasis.

Corollary 3.3.4. *For each positive integer p, suppose that $X_p \sim \mathcal{N}_p(\mu_p, \Sigma_p)$, with Σ_p being positive definite. Then*

$$\lim_{p\to\infty} \widetilde{\mathcal{V}}^2(X_p, X_p) = \frac{1}{2}. \tag{3.3.7}$$

Proof of Theorem 3.3.3 and Corollary 3.3.4. In order to prove (3.3.5) we study the limit for the terms corresponding separately to $k = 1$, $k = 2$, and $k \geq 3$ in (3.3.2).

For $k = 1$, on recalling that $C_{(1)}(\Lambda_p) = \operatorname{tr}(\Lambda_p)$, the ratio of that term to $\operatorname{tr}(\Lambda_p)/p$ is given by

$$\frac{c_{p-1}^2}{c_p^2}\frac{\pi}{p},$$

which tends to $1/2$ due to Lemma 3.3.2.

For $k = 2$, we first deduce from (2.3.2) that $C_{(2)}(\Lambda_p) \leq (\operatorname{tr}\Lambda_p)^2$. Moreover, $\operatorname{tr}(\Lambda_p) \leq p$ because $\Lambda_p \leq_\ell I_p$ and Lemma 3.2.3. Thus, the ratio of the second term in (3.3.2) to $\operatorname{tr}(\Lambda_p)/p$ is a constant multiple of

$$\frac{p}{\operatorname{tr}(\Lambda_p)}\,\frac{c_{p-1}^2}{c_p^2}\,\frac{C_{(2)}(\Lambda_p)}{(\frac{1}{2}p)_2\,(\frac{1}{2}p)_2} \leq \frac{c_{p-1}^2}{c_p^2}\,\frac{p^2}{(\frac{1}{2}p)_2\,(\frac{1}{2}p)_2} = 4\,\frac{p}{(p+1)^2}\,\frac{c_{p-1}^2}{p\,c_p^2}$$

which, by Lemma 3.3.2, converges to zero as $p \to \infty$.

Finally, suppose that $k \geq 3$. Obviously the largest eigenvalue of Λ_p is equal to the smallest eigenvalue of $\|\Lambda_p\| I_p$, and so it follows from (2.3.6) that $C_{(k)}(\Lambda_p) \leq \|\Lambda_p\|^k\, C_{(k)}(I_p)$. Further, note that $\operatorname{tr}(\Lambda_p) \geq \|\Lambda_p\|$. Then by $\lambda_p \leq_\ell I_p$ and applying (2.3.7) and Lemma 3.2.3 we obtain

$$\frac{C_{(k)}(\Lambda_p)}{\operatorname{tr}(\Lambda_p)} \leq \frac{\|\Lambda_p\|^k\, C_{(k)}(I_p)}{\|\Lambda_p\|} = \|\Lambda_p\|^{k-1} C_{(k)}(I_p) \leq C_{(k)}(I_p) = \frac{(\frac{1}{2}p)_k}{(\frac{1}{2})_k}. \tag{3.3.8}$$

Therefore,

$$4\pi\,\frac{p}{\operatorname{tr}(\Lambda_p)}\,\frac{c_{p-1}^2}{c_p^2}\sum_{k=3}^{\infty}\frac{2^{2k}-2}{k!\,2^{2k}}\,\frac{(\frac{1}{2})_k\,(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k\,(\frac{1}{2}p)_k}\,C_{(k)}(\Lambda_p)$$
$$\leq 4\pi\,p\,\frac{c_{p-1}^2}{c_p^2}\,\sum_{k=3}^{\infty}\frac{2^{2k}-2}{k!\,2^{2k}}\,\frac{(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k}.$$

By Lemma 3.3.2, each term $pc_{p-1}^2/(\frac{1}{2}p)_k c_p^2$ converges to zero as $p\to\infty$, and this proves both (3.3.5) and its special case, (3.3.7). Then, (3.3.6) follows immediately. □

Finally, we consider the situation in which q, the dimension of Y, is fixed while p, the dimension of X, grows without bound.

Theorem 3.3.5. *For each positive integer p, suppose that $(X_p, Y)\sim\mathcal{N}_{p+q}(\mu_p,\Sigma_p)$, where*

$$\Sigma_p=\begin{pmatrix}\Sigma_{X,\,p} & \Sigma_{XY,\,p}\\ \Sigma_{YX,\,p} & \Sigma_Y\end{pmatrix}$$

with $\Sigma_{X,\,p}\in\mathbb{R}^{p\times p}$ and $\Sigma_Y\in\mathbb{R}^{q\times q}$ being positive definite and such that

$$\Lambda_p=\Sigma_Y^{-1/2}\,\Sigma_{YX,\,p}\,\Sigma_{X,\,p}^{-1}\,\Sigma_{XY,\,p}\,\Sigma_Y^{-1/2}\neq 0.$$

Then

$$\lim_{p\to\infty}\frac{\sqrt{p}}{\operatorname{tr}(\Lambda_p)}\,\widetilde{\mathcal{V}}^2(X_p,Y)=\sqrt{\frac{\pi}{2}}\,\frac{c_{q-1}}{q\,c_q}\tag{3.3.9}$$

and

$$\lim_{p\to\infty}\frac{\sqrt{p}}{\operatorname{tr}(\Lambda_p)}\,\widetilde{\mathcal{R}}^2(X_p,Y)=\frac{1}{2q\sqrt{A(q)}}.\tag{3.3.10}$$

Proof. By (3.2.1),

$$\widetilde{\mathcal{V}}^2(X_p,Y)=4\pi\,\frac{c_{p-1}}{c_p}\,\frac{c_{q-1}}{c_q}\sum_{k=1}^{\infty}\frac{2^{2k}-2}{k!\,2^{2k}}\,\frac{(\frac{1}{2})_k\,(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k\,(\frac{1}{2}q)_k}\,C_{(k)}(\Lambda_p).$$

We now examine the limiting behavior, as $p\to\infty$, of the terms in this sum for $k=1$ and, separately, for $k\geq 2$.

For $k=1$, the limiting value of the ratio of the corresponding term to $\operatorname{tr}(\Lambda_p)/\sqrt{p}$ equals

$$\pi\,\frac{c_{q-1}}{q\,c_q}\,\lim_{p\to\infty}\frac{\sqrt{p}}{\operatorname{tr}(\Lambda_p)}\,\frac{c_{p-1}}{p\,c_p}\,C_{(1)}(\Lambda_p)=\sqrt{\frac{\pi}{2}}\,\frac{c_{q-1}}{q\,c_q}$$

by Lemma 3.3.2 and the fact that $C_{(1)}(\Lambda_p)=\operatorname{tr}(\Lambda_p)$.

For $k \geq 2$, the ratio of the sum to $\mathrm{tr}\,(\Lambda_p)/\sqrt{p}$ equals

$$
\begin{aligned}
4\pi\,\frac{\sqrt{p}}{\mathrm{tr}\,(\Lambda_p)}\,\frac{c_{p-1}}{c_p}\,\frac{c_{q-1}}{c_q}\sum_{k=2}^{\infty}\frac{2^{2k}-2}{k!\,2^{2k}}\,\frac{(\frac{1}{2})_k\,(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k\,(\frac{1}{2}q)_k}\,C_{(k)}(\Lambda_p) \\
\leq 4\pi\,\frac{\sqrt{p}}{\|\Lambda_p\|}\,\frac{c_{p-1}}{c_p}\,\frac{c_{q-1}}{c_q}\sum_{k=2}^{\infty}\frac{2^{2k}-2}{k!\,2^{2k}}\,\frac{(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k}\,\|\Lambda_p\|^k \\
\leq 4\pi\,\sqrt{p}\,\frac{c_{p-1}}{c_p}\,\frac{c_{q-1}}{c_q}\sum_{k=2}^{\infty}\frac{2^{2k}-2}{k!\,2^{2k}}\,\frac{(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k},
\end{aligned}
$$

where we have used (3.3.8) to obtain the last two inequalities. By applying (3.3.4), we see that the latter upper bound converges to 0 as $p \to \infty$, which proves (3.3.9), and then (3.3.10) follows immediately. □

We illustrate special cases of our limiting results in Figure 3.3 and 3.4. Figure 3.3 gives the affinely invariant distance variance for dimensions $p = 1, \ldots, 5$. For non-integer values of p the graph shows the value of the representation (3.2.17), which is continuous in p. The level of the limit $\sqrt{2}$ is marked by the horizontal dashed line.
Figure 3.4 plots the value of the affinely invariant distance correlation for random variables $X, Y \in \mathbb{R}^p$ with $\Lambda_{X,Y} = r\,I_p$. The solid lines marks the identity, while the dashed lines represent the affinely invariant distance correlations for different values of p. The line being farthest away from the identity corresponds to $p = 1$, the middle one corresponds to $p = 5$, while the nearest graph sketches the affinely invariant distance correlation for $p = 10$.

The results in this section have practical implications for affinely invariant distance correlation analysis of large-sample, high-dimensional Gaussian data. In the setting of Theorem 3.3.5, $\mathrm{tr}\,(\Lambda_p) \leq q$ is bounded, and so

$$\lim_{p\,\to\,\infty} \widetilde{\mathcal{R}}(X_p, Y) = 0.$$

As a consequence of Theorem 3.1.7 on the consistency of sample measures, it follows that the direct calculation of affinely invariant distance correlation measures for such data will return values which are virtually zero. In practice, in order to obtain values of the sample affinely invariant distance correlation measures which permit statistical inference, it will be necessary to calculate $\widehat{\Lambda}_p$, the maximum likelihood estimator of Λ_p, and then to rescale the distance correlation measures with the factor $\sqrt{p}/\mathrm{tr}\,(\widehat{\Lambda}_p)$. In the scenario of Theorem 3.3.3 the asymptotic behavior of the affinely invariant distance correlation measures depends on the ratio $p/\mathrm{tr}\,(\Lambda_p)$; and as $\mathrm{tr}\,(\Lambda_p)$ can attain any value in the interval $[0, p]$, a wide range of asymptotic rates of convergence is conceivable. In all these settings, the series representation (3.2.1) can be used to obtain complete asymptotic expansions in powers of p^{-1} or q^{-1}, of the affinely invariant distance covariance

or correlation measures, as p or q tend to infinity.

3.4 Time Series of Wind Vectors at the Stateline Wind Energy Center

Recently, Zhou suggested the use of distance correlation for time series. In [116], he defines the auto distance correlation function and shows the consistency of a respective sample measure under moderate assumptions (see 2.4.12-2.4.14). It is straightforward to extend these notions to the affinely invariant distance correlation.

Definition 3.4.1. *Let $X = \{X_j\}_{j=-\infty}^{\infty}$ be a strictly stationary multivariate time series of dimension p and let Σ_{X_0} denote the covariance matrix of X_0. Then the affinely invariant distance covariance function is, for $k \geq 0$, given by*

$$\widetilde{\mathcal{V}}_X(k) = \mathcal{V}_{\widetilde{X}}(k),$$

where $\widetilde{X} = (\Sigma_{X_0}^{-1/2} X_j)_{j=-\infty}^{\infty}$ and $\mathcal{V}_{\widetilde{X}}$ is defined as in 2.4.12. For an integer k, define the affinely invariant auto distance correlation function as

$$\widetilde{\mathcal{R}}_X(k) = \sqrt{\frac{\widetilde{\mathcal{V}}_X(k)}{\widetilde{\mathcal{V}}_X(0)}}. \tag{3.4.1}$$

We further extend the idea of Zhou to define *cross* distance covariance functions and *cross* distance correlation functions between two jointly strictly stationary, vector-valued time series, namely

Definition 3.4.2. *Let $X = (X_j)_{j=-\infty}^{\infty}$ and $Y = (Y_j)_{j=-\infty}^{\infty}$ be two strictly stationary multivariate time series of dimensions p and q, respectively. Then the cross distance covariance function $\mathcal{V}_X$ is, for $k \in \mathbb{Z}$, defined as*

$$\mathcal{V}_{X,Y}(k) = \frac{1}{c_p\, c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X_0,Y_k}(s,t) - f_{X_0}(s) f_{Y_k}(t)|^2}{|s|_p^{p+1} |t|_q^{q+1}} ds dt,$$

moreover, the cross distance correlation function $\mathcal{R}_X$ is, for $k \in \mathbb{Z}$ defined as

$$[\mathcal{R}_X(k)]^2 = \frac{\mathcal{V}_{X,Y}(k)}{\sqrt{\mathcal{V}_X(0)\, \mathcal{V}_Y(0)}}$$

if the denominator is strictly positive, 0 otherwise.

The affinely invariant cross distance covariance function and the affinely invariant cross distance correlation function can the be defined in analogous fashion.

Definition 3.4.3. *Let $X = (X_j)_{j=-\infty}^{\infty}$ and $Y = (Y_j)_{j=-\infty}^{\infty}$ be two strictly stationary multivariate time series of dimensions p and q, respectively. Moreover let Σ_{X_0} and Σ_{Y_0} denote the covariance matrices of X_0 and Y_0, respectively. Then the affinely invariant cross distance covariance function $\mathcal{V}_X$ is, for $k \in \mathbb{Z}$, defined as*

$$\widetilde{\mathcal{V}}_{X,Y}(k) = \mathcal{V}_{\widetilde{X},\widetilde{Y}}(k),$$

where $\widetilde{X} = 8\Sigma_{X_0}^{-1/2} X_j)_{j=-\infty}^{\infty}$ and $\widetilde{Y} = (\Sigma_{Y_0}^{-1/2} Y_j)_{j=-\infty}^{\infty}$. For an integer k, define the affinely invariant cross distance correlation function as

$$[\widetilde{\mathcal{R}}_{X,Y}(k)]^2 = \frac{\widetilde{\mathcal{V}}_{X,Y}(k)}{\sqrt{\widetilde{\mathcal{V}}_X(0)\widetilde{\mathcal{V}}_Y(0)}}. \tag{3.4.2}$$

The corresponding sample versions can be defined in the natural way, as in the case of the non-affine distance correlation [116].

We illustrate these concepts on time series data of wind observations at and near the Stateline wind energy center in the Pacific Northwest of the United States. Specifically, we consider time series of bivariate wind vectors at the meteorological towers at Vansycle, right at the Stateline wind farm at the border of the states of Washington and Oregon, and at Goodnoe Hills, 146 km west of Vansycle along the Columbia River Gorge. Further information can be found in the paper by Gneiting, et al. [31], who developed a regime-switching space-time (RST) technique for 2-hour-ahead forecasts of hourly average wind speed at the Stateline wind energy center, which was then the largest wind farm globally. For our purposes, we follow Hering and Genton [41] in studying the time series at the original 10-minute resolution, and we restrict our analysis to the longest continuous record, the 75-day interval from August 14 to October 28, 2002.

Thus, we consider time series of bivariate wind vectors over $10,800$ consecutive 10-minute intervals. We write V_j^{NS} and V_j^{EW} to denote the north-south and the east-west component of the wind vector at Vansycle at time j, with positive values corresponding to northerly and easterly winds. Similarly, we write G_j^{NS} and G_j^{EW} for the north-south and the east-west component of the wind vector at Goodnoe Hills at time j, respectively.

Figure 3.5 shows the classical (Pearson) sample auto and cross correlation functions for the four univariate time series. The auto correlation functions generally decay with the temporal, but do so non-monotonously, due to the presence of a diurnal component. The cross correlation functions between the wind vector components at Vansycle and Goodnoe Hills show remarkable asymmetries and peak at positive lags, due to the

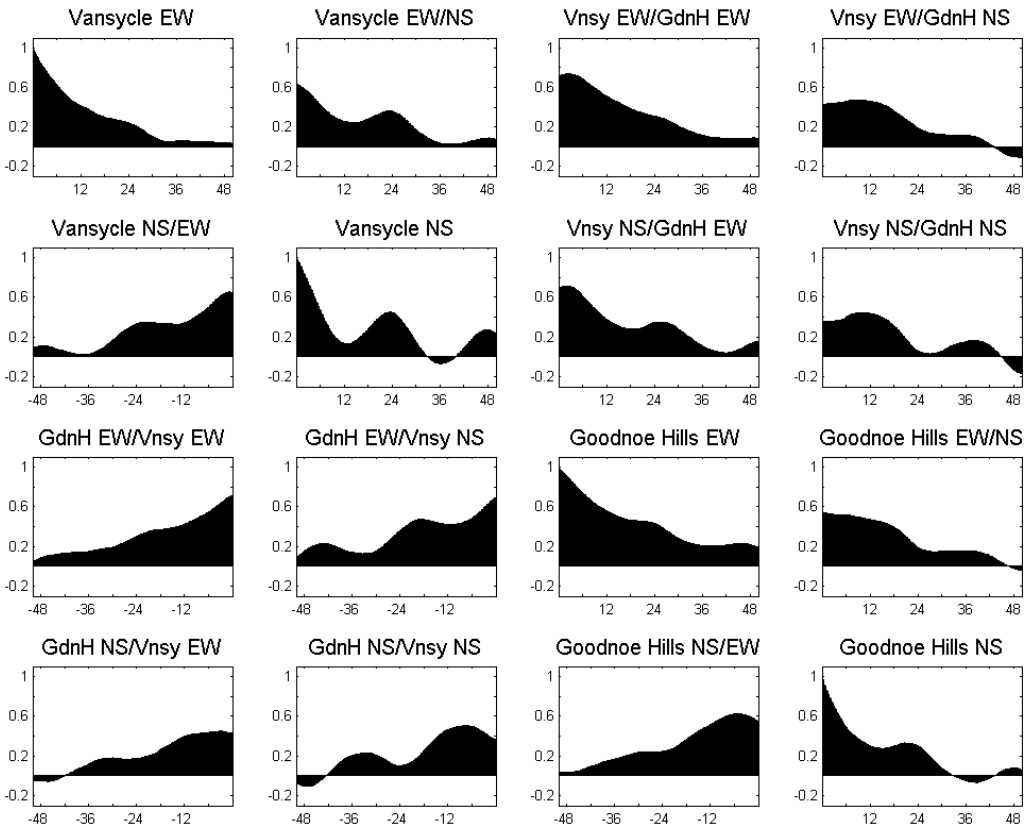

Figure 3.5: Sample auto and cross Pearson correlation functions for the univariate time series V_j^{EW}, V_j^{NS}, G_j^{EW}, and G_j^{NS}, respectively. Positive lags indicate observations at the westerly site (Goodnoe Hills) leading those at the easterly site (Vansycle), or observations of the north-south component leading those of the east-west component, in units of hours.

prevailing westerly and southwesterly wind [31]. In another interesting feature, the cross correlations between the north-south and east-west components at lag zero are strongly positive, documenting the dominance of southwesterly winds.

Figure 3.6 shows the sample auto and cross distance correlation functions for the four time series; as these variables are univariate, there is no distinction between the standard and the affinely invariant version of the distance correlation. The patterns seen resemble those in the case of the Pearson correlation. For comparison, we also display values of the distance correlation based on the sample Pearson correlations shown in Figure 3.5, and converted to distance correlation under the assumption of bivariate Gaussianity, using the results of Székely, et al. (Theorem 2.4.10) and Section 3.2; in every single case, these values are smaller than the original ones.

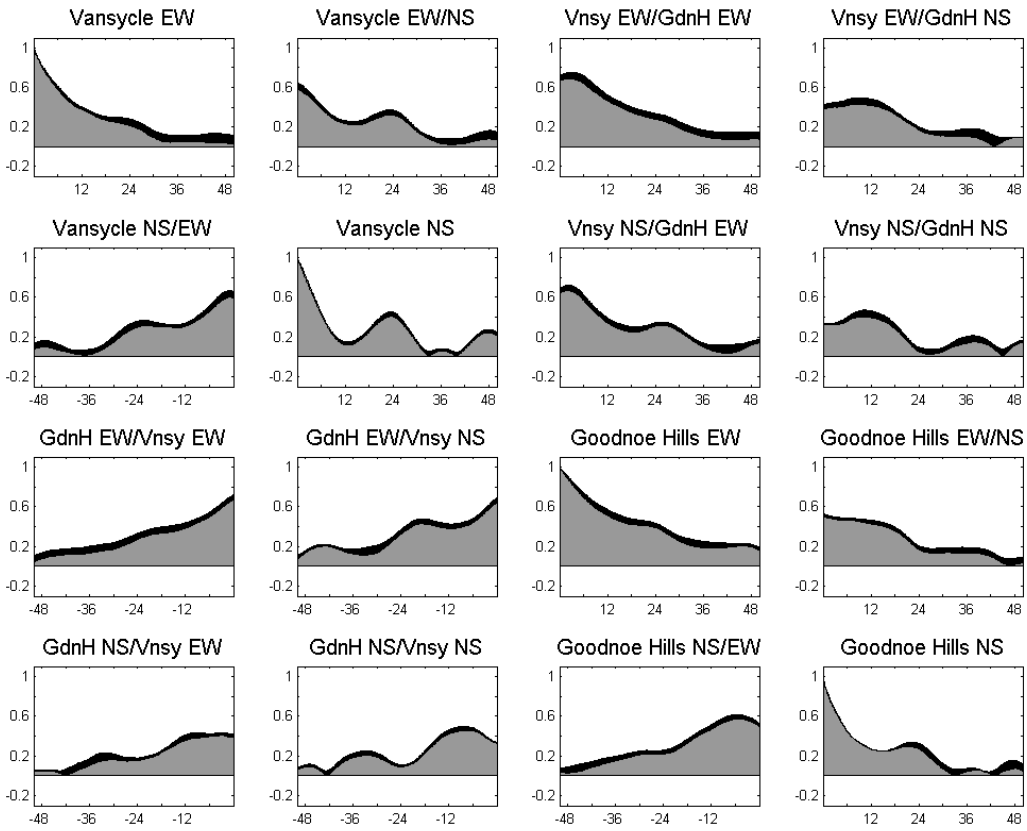

Figure 3.6: Sample auto and cross distance correlation functions for the univariate time series V_j^{EW}, V_j^{NS}, G_j^{EW}, and G_j^{NS}, respectively. For comparison, we also display, in gray, the values that arise when the sample Pearson correlations in Figure 3.5 are converted to distance correlation under the assumption of Gaussianity; these values generally are smaller than the original ones. Positive lags indicate observations at Goodnoe Hills leading those at Vansycle, or observations of the north-south component leading those of the east-west component, in units of hours.

Having considered the univariate time series setting, it is natural and complementary to look at the wind vector time series $(V_j^{\mathrm{EW}}, V_j^{\mathrm{NS}})$ at Vansycle and $(G_j^{\mathrm{EW}}, G_j^{\mathrm{NS}})$ at Goodnoe Hills from a genuinely multivariate perspective. To this end, Figure 3.7 shows the sample affinely invariant auto and cross distance correlation functions for the bivariate wind vector series at the two sites. Again, a diurnal component is visible, and there is a remarkable asymmetry in the cross-correlation functions, which peak at lags of about two to three hours.

In light of our analytical results in Section 3.2, we can compute the affinely invariant distance correlation between subvectors of a multivariate normally distributed random vector. In particular, we can compute the affinely invariant auto and cross distance

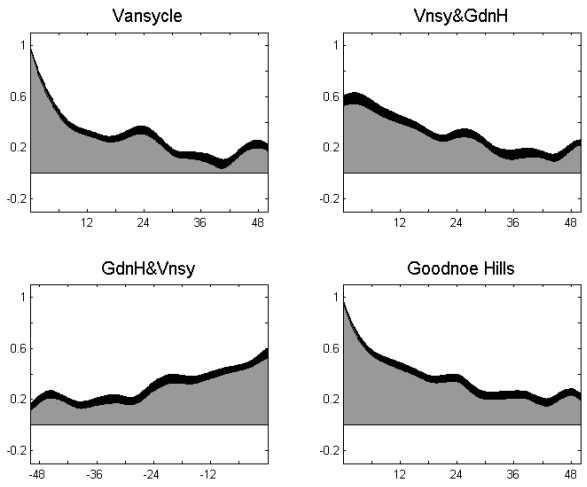

Figure 3.7: Sample auto and cross affinely invariant distance correlation functions for the bivariate time series $(V_j^{\mathrm{EW}}, V_j^{\mathrm{NS}})'$ and $(G_j^{\mathrm{EW}}, G_j^{\mathrm{NS}})'$ at Vansycle and Goodnoe Hills. For comparison, we also display, in gray, the values that are generated when the Pearson correlation in Figure 3.5 is converted to the affinely invariant distance correlation under the assumption of Gaussianity; these converted values generally are smaller than the original ones. Positive lags indicate observations at Goodnoe Hills leading those at Vansycle, in units of hours.

correlation between bivariate subvectors of a 4-variate Gaussian process with Pearson auto and cross correlations as shown in Figure 3.5. In Figure 3.7, values of the affinely invariant distance correlation that have been derived from Pearson correlations in these ways are shown in gray; the differences from those values that are computed directly from the data are substantial, with the converted values being smaller, possibly suggesting that assumptions of Gaussianity may not be appropriate for this particular data set.

We wish to emphasize that our study is purely exploratory: it is provided for illustrative purposes and to serve as a basic example. In future work, the approach hinted at here may have the potential to be developed into parametric or nonparametric bootstrap tests for Gaussianity. For this purpose recall that, in the Gaussian setting, the affinely invariant distance correlation is a function of the canonical correlation coefficients, i.e. $\widetilde{\mathcal{R}} = g(\lambda_1, \ldots, \lambda_r)$. For a parametric bootstrap test, one could generate B replicates of $g(\lambda_1^\star, \ldots, \lambda_r^\star)$, leading to a pointwise $(1-\alpha)$-confidence band. The test would now reject Gaussianity if the sample affinely invariant distance correlation function does not lie within this band. For the nonparametric bootstrap test, one could obtain ensembles $\widetilde{\mathcal{R}}_n^\star$ by resampling methods, again defining a pointwise $(1-\alpha)$-confidence band and checking if $g(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)$ is located within this band.

Following the pioneering work of Zhou [116], the distance correlation may indeed find a wealth of applications in exploratory and inferential problems for time series data.

Chapter 4

A Generalization of an Integral Arising in Distance Correlation

In this chapter, we derive an extension of Lemma 2.4.9, which is known to be fundamental for the concept of distance correlation. The following result will generalize Lemma 2.4.9 in two ways. First, we show that the regularization $\lim_{\varepsilon\to 0}\int_{\mathbb{R}^d\setminus\{\varepsilon B+\varepsilon^{-1}B^C\}}$ of the integral is not needed, since the integral converges absolutely under the stated condition on α. Second, we show how the knowledge of Lemma 2.4.9 helps to solve a more general integral, where we insert a truncated Maclaurin expansion of the function $\cos(\langle t, x\rangle)$ into the integrand. We further proof that this generalization is valid for all $\alpha \in \mathbb{C}$ such that $2(m-1) < \Re(\alpha) < 2m$, where m is any positive integer. Let us note, that the latter extension of this integral may be used to generalize the class of α-distance dependence measures [102, p. 2784] to α outside the range $(0, 2)$. The content of this chapter is adapted from the paper [20] by Dueck, Edelmann and Richards.

Throughout this chapter, we will denote the truncated Maclaurin expansion of the cosine function by

$$\cos_m(v) := \sum_{j=0}^{m-1}(-1)^j\frac{v^{2j}}{(2j)!}, \tag{4.0.1}$$

where the expansion is halted at the mth ($m \in \mathbb{N}$) summand. Further, we let

$$B_a = \{x \in \mathbb{R}^d : |x|_d < a\}$$

denote the ball which is centered at the origin and which is of radius a.

We will make frequent use of the following argument.

Lemma 4.0.4. *Let $x \in \mathbb{R}^d \backslash \{0\}$ and let $\gamma_{d,k}$ be defined as in (3.2.10). For $\alpha \in \mathbb{C}$,*

$$\int_{B_a} \frac{\langle t, x\rangle^{2k}}{|t|_d^{d+\alpha}} \mathrm{d}t = |x|_d^{2k} \gamma_{d,k} (2k-\alpha)^{-1} a^{2k-\alpha} \tag{4.0.2}$$

with absolute convergence if and only if $\Re(\alpha) < 2k$.

Proof. Transformation to polar coordinates (see Theorem 2.1.3) yields

$$\int_{B_a} \frac{\langle t, x\rangle^{2k}}{|t|_d^{d+\alpha}} \mathrm{d}t = \int_0^a \int_{S^{d-1}} r^{2k-\alpha-1} \langle \theta, x\rangle^{2k} \mathrm{d}\theta \mathrm{d}r.$$

By a standard invariance argument (see the proof of Theorem 3.2.4 for details), we see that the latter integral is equal to

$$|x|_d^{2k} \int_0^a \int_{S^{d-1}} r^{2k-\alpha-1} \theta_1^{2k} \mathrm{d}\theta \mathrm{d}r = |x|_d^{2k} \gamma_{d,k} \int_0^a r^{2k-\alpha-1} \mathrm{d}r.$$

By evaluation of the latter integral, we obtain the desired result. □

Theorem 4.0.5 states the main result of this chapter, generalizing the integral stated in Lemma 2.4.9.

Theorem 4.0.5. *Let $m \in \mathbb{N}$ and $x \in \mathbb{R}^d \backslash \{0\}$. For $\alpha \in \mathbb{C}$,*

$$\int_{\mathbb{R}^d} \frac{\cos_m(\langle t, x\rangle) - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \mathrm{d}t = C(d,\alpha) |x|_d^{\alpha}, \tag{4.0.3}$$

with absolute convergence if and only if $2(m-1) < \Re(\alpha) < 2m$, where $C(d,\alpha)$ is given in (2.4.8).

Proof. We shall establish the proof by induction on m. Consider the case in which $m = 1$. To determine the range of convergence, we split the integral into two parts:

$$\begin{aligned} &\int_{\mathbb{R}^d} \frac{\cos_1(\langle t, x\rangle) - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \mathrm{d}t = \int_{\mathbb{R}^d} \frac{1 - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \mathrm{d}t \\ &= \int_{B_a} \frac{1 - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \mathrm{d}t + \int_{\mathbb{R}^d \backslash B_a} \frac{1 - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \mathrm{d}t. \end{aligned} \tag{4.0.4}$$

By applying (4.0.1) and interchanging integral and summation by means of Fubini's theorem, we see that the first integral equals

$$\int_{B_a} \frac{1 - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \mathrm{d}t = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{(2j)!} \int_{B_a} \frac{\langle t, x\rangle^{2j}}{|t|_d^{d+\alpha}} \mathrm{d}t. \tag{4.0.5}$$

By Lemma 4.0.4, the j-th summand converges if and only if $\Re(\alpha) < 2j$. Hence, for $\Re(\alpha) < 2$, inserting (4.0.2) yields

$$\int_{B_a} \frac{1 - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \, \mathrm{d}t = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{(2j)!} |x|_d^{2j} \, \gamma_{d,j} (2j - \alpha)^{-1} a^{2j-\alpha}. \tag{4.0.6}$$

Since $\gamma_{d,j}$ is obviously decreasing in j, this series converges if and only if $\Re(\alpha) < 2$. For the second integral in (4.0.4), we apply the bound $|1 - \cos(\langle t, x\rangle)| \leq 2$ to deduce that the integrand is integrable over $\mathbb{R} \setminus B_a$ if and only if $\Re(\alpha) > 0$. Consequently, for $m = 1$, the integral converges for all $x \in \mathbb{R}^d \setminus \{0\}$ if and only if $0 < \Re(\alpha) < 2$.

To conclude the proof for the case in which $m = 1$, we proceed precisely as did Székely, et al. [102, p. 2771] to obtain the right-hand side of (4.0.3).

Next, we assume by inductive hypothesis that the assertion holds for a given positive integer m. Note that the right-hand side of (4.0.3), as a function of $\alpha \in \mathbb{C}$, is meromorphic with a pole at each nonnegative integer α.
By (4.0.1),

$$\cos_{m+1}(v) = \cos_m(v) + (-1)^m \frac{v^{2m}}{(2m)!}.$$

For fixed $a > 0$, we decompose the integral (4.0.3) into a sum of three terms:

$$\int_{\mathbb{R}^d} \frac{\cos_m(\langle t, x\rangle) - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \, \mathrm{d}t = T_1 + T_2 + T_3, \tag{4.0.7}$$

where

$$T_1 = \int_{B_a} \frac{\cos_{m+1}(\langle t, x\rangle) - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \, \mathrm{d}t,$$
$$T_2 = \int_{\mathbb{R}^d \setminus B_a} \frac{\cos_m(\langle t, x\rangle) - \cos(\langle t, x\rangle)}{|t|_d^{d+\alpha}} \, \mathrm{d}t,$$

and

$$T_3 = \frac{(-1)^{m-1}}{(2m)!} \int_{B_a} \frac{\langle t, x\rangle^{2m}}{|t|_d^{d+\alpha}} \, \mathrm{d}t.$$

We now determine the necessary and sufficient condition on the range of α for which the decomposition (4.0.7) entails absolute convergence of the integral. In so doing, we examine each term individually.
In the case of T_1, we proceed as in (4.0.5)-(4.0.6) to find that the series converges absolutely for all $x \in \mathbb{R}^d\{0\}$ if and only if $\Re(\alpha) < 2(m + 1)$. As regards the term T_2 we know, by inductive hypothesis, that it converges absolutely if and only if $\Re(\alpha) >$

$2(m-1)$.

By Lemma 4.0.4 we find that T_3 converges absolutely if and only if $\Re(\alpha) < 2m$ and

$$T_3 = \frac{(-1)^{m-1}}{(2m)!}\,\gamma_{d,m}\,|x|_d^{2m}\,(2m-\alpha)^{-1}\,a^{2m-\alpha}. \tag{4.0.8}$$

Moreover, the last term in (4.0.8) exists for all $\alpha \in \mathbb{C}$ such that $\Re(\alpha) \neq 2m$ and it is a meromorphic function of α.
To summarize, T_1 converges absolutely for $\Re(\alpha) < 2(m+1)$; T_2 converges absolutely for $\Re(\alpha) > 2(m-1)$; and T_3 converges absolutely for $\Re(\alpha) < 2m$. Therefore, the decomposition (4.0.7) is valid for $2(m-1) < \Re(\alpha) < 2m$, and it represents an analytic function which equals $C(d,\alpha)\,|x|_d^{\alpha}$ on the strip $\{\alpha \in \mathbb{C} : 2(m-1) < \Re(\alpha) < 2m\}$. Hence, by analytic continuation, we obtain for $2(m-1) < \Re(\alpha) < 2(m+1)$, $\Re(\alpha) \neq 2m$,

$$C(d,\alpha)\,|x|_d^{\alpha} = T_1 + T_2 + \frac{(-1)^{m-1}}{(2m)!}\,\gamma_{d,m}\,|x|_d^{2m}\,(2m-\alpha)^{-1}\,a^{2m-\alpha}. \tag{4.0.9}$$

Now fix $2m < \Re(\alpha) < 2(m+1)$ and let $a \to \infty$ in (4.0.9). It is apparent that $T_2 \to 0$ and $a^{2m-\alpha} \to 0$; therefore, for $2m < \Re(\alpha) < 2(m+1)$, we obtain

$$C(d,\alpha)\,|x|_d^{\alpha} = \lim_{a\to\infty} T_1 = \int_{\mathbb{R}^d} \frac{\cos_{m+1}(\langle t,x\rangle) - \cos(\langle t,x\rangle)}{|t|_d^{d+\alpha}}\,\mathrm{d}t,$$

which concludes the proof. □

As already stated in the introduction of this chapter Theorem 4.0.5 generalizes Lemma 2.4.9. We suppose that our extended version of this lemma may motivate the definition of α-dependent measures where α is larger than 2. We further expect, that such a theory will lead for sufficiently large $\Re(\alpha)$ to distance correlation analyses of data modeled by random vectors which do not have finite first moments, e.g., the multivariate stable distributions of index less than 2. Moreover, although the integral (4.0.3) diverges for $\Re(\alpha) = 2m$, our results raise the possibility of developing a theory of distance correlation at the poles by modifying (4.0.3) to attain convergence as $\Re(\alpha)$ converges to the poles.

Finally, we remark that our decomposition (4.0.7) was motivated by the ideas of Gelfand and Shilov [30, p. 10].

Chapter 5

Distance Correlation and Lancaster Distributions

As we have already pointed out earlier in this work, the calculation of the population version of the distance correlation or the affinely invariant distance correlation is non-trivial. On the other hand, it is certain that the evaluation of these population measures leads to a better understanding of distance correlation since it captures in which way its value depends on other parameters of the distribution. For the multivariate normal, for example, we could show that the affinely invariant distance correlation between X and Y is a symmetric function in the canonical correlations between X and Y. Moreover, as indicated in Chapter 3, the knowledge of the distance correlation for certain distribution opens up possibilities for further applications, such as high-dimensional settings (section 3.3) or bootstrap testing (section 3.4).

When aiming to find the distance covariance for multivariate distributions, the straightforward way is to calculate the occurring integrals for every single distribution separately, just as we did for the multivariate normal (section 3.2) or for the multivariate Laplace (Appendix A.2). Another approach, which we pursue in this chapter, is to calculate distance covariance for a *class of multivariate distributions*, containing various common multivariate distributions as special cases. In particular, we calculate the distance correlation coefficients for pairs (X, Y) of random vectors whose joint distributions are in the class of *Lancaster distributions*, a class of probability distributions which was made prominent by Lancaster [59, 60] and by Sarmanov [85]. The distribution functions of the Lancaster family are well-known to have attractive expansions in terms of certain orthogonal functions (Koudou [58]; Diaconis, et al. [16]). By applying those expansions, we deduce series expansions for the corresponding characteristic functions and then we obtain explicit expressions for the distance covariance and distance correlation coefficients.

Consequently we derive under mild convergence conditions a general formula for the distance covariance for the Lancaster distributions. As examples, we apply the general formula to obtain explicit expressions for the distance covariance and distance correlation for the bivariate and multivariate normal distributions, and for bivariate gamma and Poisson distributions. We remark that explicit results can also be obtained for certain negative binomial distributions and for other Lancaster-type expansions obtained by Bar-Lev, *et al.* [4]; because the formulas derived here are fully representative of the general case then we will omit the details for other cases. The content of this chapter has been extracted from the paper [19] by Johannes Dueck, Dominic Edelmann and Donald Richards.

5.1 The Lancaster Distributions

To recapitulate the class of Lancaster distributions we generally follow the standard notation in that area, as given by Koudou [57, 58]; cf., Lancaster [60], Pommeret [72], or Diaconis, et al. [16, Section 6].

Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be locally compact, separable probability spaces, such that $L^2(\mu)$ and $L^2(\nu)$ are separable. Let σ, a probability measure on $\mathcal{X} \times \mathcal{Y}$, have marginal distributions μ and ν; then there exist functions K_σ and L_σ such that

$$\sigma(\mathrm{d}x, \mathrm{d}y) = K_\sigma(x, \mathrm{d}y)\mu(\mathrm{d}x) = L_\sigma(\mathrm{d}x, y)\nu(\mathrm{d}y).$$

We note that K_σ and L_σ represent the conditional distributions of Y given $X = x$, and X given $Y = y$, respectively.

Let $\mathcal{C}$ denote a countable index set with a zero element, denoted by 0. Let $\{P_n : n \in \mathcal{C}\}$ and $\{Q_n : n \in \mathcal{C}\}$ be sequences of functions on $\mathcal{X}$ and $\mathcal{Y}$ which form orthonormal bases for the separable Hilbert spaces $L^2(\mu)$ and $L^2(\nu)$, respectively. We assume, by convention, that $P_0 \equiv 1$ and $Q_0 \equiv 1$.

Because the tensor product Hilbert space $L^2(\mu \otimes \nu) \equiv L^2(\mu) \otimes L^2(\nu)$ is separable there holds, for $\sigma \in L^2(\mu \otimes \nu)$, the expansion

$$\sigma(\mathrm{d}x, \mathrm{d}y) = \sum_{m \in \mathcal{C}} \sum_{n \in \mathcal{C}} \rho_{m,n} P_m(x) Q_n(y)\, \mu(\mathrm{d}x)\, \nu(\mathrm{d}y), \tag{5.1.1}$$

$(x, y) \in \mathcal{X} \times \mathcal{Y}$. Letting $\delta_{m,n}$ denote Kronecker's delta, the probability measure σ is

called a *Lancaster distribution* if there exists a positive sequence $\{\rho_n : n \in \mathcal{C}\}$ such that

$$\int P_m(x)\, Q_n(y)\, \sigma(\mathrm{d}x, \mathrm{d}y) = \rho_m\, \delta_{m,n}$$

for all $m, n \in \mathcal{C}$; in particular, $\rho_0 = 1$. The sequence $\{\rho_n : n \in \mathcal{C}\}$ is called a *Lancaster sequence*, and the expansion (5.1.1) reduces to

$$\sigma(\mathrm{d}x, \mathrm{d}y) = \sum_{n \in \mathcal{C}} \rho_n P_n(x) Q_n(y) \mu(\mathrm{d}x) \nu(\mathrm{d}y).$$

Koudou [57, pp. 255–256] characterized the Lancaster sequences $\{\rho_n : n \in \mathcal{C}\}$ such that the associated probability distribution σ is absolutely continuous with respect to $\mu \otimes \nu$ and has Radon-Nikodym derivative

$$\frac{\sigma(\mathrm{d}x, \mathrm{d}y)}{\mu(\mathrm{d}x)\, \nu(\mathrm{d}y)} = \sum_{n \in \mathcal{C}} \rho_n\, P_n(x)\, Q_n(y) \in L^2(\mu \otimes \nu),$$

$(x, y) \in \mathcal{X} \times \mathcal{Y}$.
In the sequel, we consider the case in which $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$ and the underlying random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ have joint distribution σ and marginal distributions μ and ν, respectively. We assume that μ, ν, and σ are absolutely continuous with respect to Lebesgue measure or counting measure on the respective sample spaces and we denote their corresponding probability density functions by ϕ_X, ϕ_Y, and $\phi_{X,Y}$, respectively. This yields the expansion,

$$\phi_{X,Y}(x, y) = \phi_X(x)\, \phi_Y(y) \sum_{n \in \mathcal{C}} \rho_n\, P_n(x)\, Q_n(y). \tag{5.1.2}$$

We will refer to (5.1.2) as the *Lancaster expansion* of the joint density function $\phi_{X,Y}$.

5.2 Examples of Lancaster Expansions

In this section, we provide examples of Lancaster expansions (5.1.2) for the bivariate and multivariate normal distributions, and for some bivariate gamma and Poisson distributions. In the sequel, we denote by $\mathbb{N}_0$ the set of nonnegative integers.

5.2.1 The Bivariate Normal Distribution

Let (X, Y) follow a bivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

denoted by $(X, Y) \sim \mathcal{N}_2(0, \Sigma)$. The joint probability density function of (X, Y) is

$$\phi_{X,Y}(x, y) = \frac{1}{2\pi}(1 - \rho^2)^{-\frac{1}{2}} \exp\left(-\frac{x^2 + y^2 - 2\rho\, x\, y}{2(1 - \rho^2)}\right),$$

$x, y \in \mathbb{R}$, and the marginal density functions are given by

$$\phi_X(x) = \phi_Y(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x^2\right).$$

In this case, the index set $\mathcal{C}$ is $\mathbb{N}_0$. For $n \in \mathbb{N}_0$, let

$$H_n(x) = (-1)^n \exp\left(\tfrac{1}{2}x^2\right)\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^n \exp\left(-\tfrac{1}{2}x^2\right),$$

$x \in \mathbb{R}$, denote the nth Hermite polynomial, $n = 0, 1, 2, \ldots$. It is well-known that the polynomials $\{H_n : n \in \mathbb{N}_0\}$ are orthogonal with respect to the standard normal distribution and form a complete orthogonal basis for the Hilbert space $L^2(X)$. Also, the Lancaster expansion of $\phi_{X,Y}$ is given by the classical formula of Mehler: For $x, y \in \mathbb{R}$,

$$\phi_{X,Y}(x, y) = \phi_X(x)\, \phi_Y(y) \sum_{n=0}^{\infty} \frac{\rho^n}{n!}\, H_n(x)\, H_n(y), \tag{5.2.1}$$

and this series converges absolutely for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$.

We remark that there are numerous extensions of Mehler's formula which represent Lancaster-type expansions for generalizations of the bivariate normal distribution; for such expansions, we refer to Srivastava and Singhal [95] and the references given there. The details in those cases are similar to the results which we derive, and we can obtain analogous formulas for the distance correlation coefficients for those distributions.

5.2.2 The Multivariate Normal Distribution

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors such that $(X, Y) \sim \mathcal{N}_{p+q}(0, \Sigma)$, a $(p + q)$-dimensional multivariate normal distribution with mean vector 0 and positive definite covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix} \tag{5.2.2}$$

where Σ_X, Σ_Y, and $\Sigma_{XY} = \Sigma_{YX}'$ are $p \times p$, $q \times q$ and $p \times q$ matrices, respectively. We denote by $\phi_{X,Y}$ the joint probability density function of (X, Y), and by ϕ_X and ϕ_Y the marginal density functions of X and Y, respectively.

We now describe the Lancaster expansion of $\phi_{X,Y}$, a result derived in [112]. In this

case, the index set $\mathcal{C}$ is $\mathbb{N}_0^{p\times q}$, the set of $p \times q$ matrices with nonnegative integer entries. For a matrix of summation indices $\boldsymbol{N} = (N_{rc}) \in \mathbb{N}_0^{p\times q}$, define $\boldsymbol{N}! = \prod_{r=1}^{p}\prod_{c=1}^{q} N_{rc}!$. For $r = 1, \ldots, p$, let

$$\boldsymbol{N}_{r\bullet} = \sum_{c=1}^{q} N_{rc}$$

and set $\boldsymbol{N}_{*\bullet} = (\boldsymbol{N}_{1\bullet}, \ldots, \boldsymbol{N}_{p\bullet})$. Similarly, for each $c = 1, \ldots, q$, define

$$\boldsymbol{N}_{\bullet c} = \sum_{r=1}^{p} N_{rc}$$

and set $\boldsymbol{N}_{\bullet *} = (\boldsymbol{N}_{\bullet 1}, \ldots, \boldsymbol{N}_{\bullet q})$. Further, we define

$$\boldsymbol{N}_{\bullet\bullet} = \sum_{r=1}^{p}\sum_{c=1}^{q} N_{rc},$$

and note that $\boldsymbol{N}_{\bullet\bullet} = \sum_{r=1}^{p} \boldsymbol{N}_{r\bullet} = \sum_{c=1}^{q} \boldsymbol{N}_{\bullet c}$.

Denoting by $(\Sigma_{XY})_{rc}$ the (r, c)th entry of Σ_{XY}, we also define

$$\Sigma_{XY}^{\boldsymbol{N}} = \prod_{r=1}^{p}\prod_{c=1}^{q}[(\Sigma_{XY})_{rc}]^{N_{rc}}.$$

We now introduce the multivariate Hermite polynomials. For any $p \in \mathbb{N}$, $\boldsymbol{k} = (k_1, \ldots, k_p) \in \mathbb{N}_0^p$, and $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, define $x^{\boldsymbol{k}} = x_1^{k_1} \cdots x_p^{k_p}$ and define the differential operator,

$$\left(-\frac{\partial}{\partial x}\right)^{\boldsymbol{k}} = \left(-\frac{\partial}{\partial x_1}\right)^{k_1} \cdots \left(-\frac{\partial}{\partial x_p}\right)^{k_p}.$$

The $\boldsymbol{k}$th *multivariate Hermite polynomial* with respect to the marginal density function ϕ_X is defined as

$$H_{\boldsymbol{k}}(x; \Sigma_X) = \frac{1}{\phi_X(x)} \left(-\frac{\partial}{\partial x}\right)^{\boldsymbol{k}} \phi_X(x). \tag{5.2.3}$$

The Lancaster expansion of the multivariate normal density function $\phi_{X,Y}$ is given by the generalized Mehler formula [112]:

$$\phi_{X,Y}(x, y) = \phi_X(x)\,\phi_Y(y) \sum_{\boldsymbol{N}\in\mathbb{N}_0^{p\times q}} \frac{\Sigma_{XY}^{\boldsymbol{N}}}{\boldsymbol{N}!} H_{\boldsymbol{N}_{*\bullet}}(x; \Sigma_X)\, H_{\boldsymbol{N}_{\bullet *}}(y; \Sigma_Y), \tag{5.2.4}$$

with absolute convergence for all $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$.

To calculate the affinely invariant distance correlation coefficient between X and Y, as

defined in equations (3.1.3) and (3.1.4), we need the Lancaster expansion of the joint density function of the standardized random vectors $\widetilde{X} = \Sigma_X^{-1/2} X$ and $\widetilde{Y} = \Sigma_Y^{-1/2} Y$. It is straightforward to verify that $(\widetilde{X}, \widetilde{Y}) \sim \mathcal{N}_{p+q}(0, \Lambda)$ where

$$\Lambda = \begin{pmatrix} I_p & \Lambda_{XY} \\ \Lambda_{XY}' & I_q \end{pmatrix}$$

with $\Lambda_{XY} = \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2}$, and then we deduce from (5.2.4) that the Lancaster expansion for $(\widetilde{X}, \widetilde{Y})$ is

$$\phi_{\widetilde{X},\widetilde{Y}}(x, y) = \phi_{\widetilde{X}}(x)\, \phi_{\widetilde{Y}}(y) \sum_{\boldsymbol{N} \in \mathbb{N}_0^{p \times q}} \frac{\Lambda_{XY}^{\boldsymbol{N}}}{\boldsymbol{N}!}\, H_{\boldsymbol{N}_{*\bullet}}(x; I_p)\, H_{\boldsymbol{N}_{\bullet *}}(y; I_q). \tag{5.2.5}$$

5.2.3 The Bivariate Gamma Distribution

The Lancaster expansion for a bivariate gamma distribution, which was derived by Sarmanov [87, 86], can be stated as follows (cf., Kotz, et al. [56, pp. 437–438]).
For $\alpha > -1$ and $n \in \mathbb{N}_0$, the classical *Laguerre polynomial* is defined by

$$\begin{aligned} L_n^{(\alpha)}(x) &= \frac{1}{n!} x^{-\alpha} \exp(x) \Big(\frac{\mathrm{d}}{\mathrm{d}x} \Big)^n x^{n+\alpha} \exp(-x) \\ &= \frac{(\alpha+1)_n}{n!} \sum_{j=0}^{n} \frac{(-n)_j}{(\alpha+1)_j} \frac{x^j}{j!}, \end{aligned} \tag{5.2.6}$$

$x > 0$, where $(\alpha)_n = \Gamma(\alpha + n)/\Gamma(\alpha)$ denotes the rising factorial.
Let $\lambda \in (0, 1)$, and let α and β satisfy $\alpha \geq \beta > 0$. Sarmanov [87, 86] derived for certain bivariate gamma random variables (X, Y) the joint probability density function,

$$\phi_{X,Y}(x, y) = \phi_X(x)\, \phi_Y(y) \sum_{n=0}^{\infty} a_n L_n^{(\alpha-1)}(x)\, L_n^{(\beta-1)}(y), \tag{5.2.7}$$

$x, y > 0$, where

$$a_n = \lambda^n \left[\frac{(\beta)_n}{(\alpha)_n} \right]^{1/2}, \tag{5.2.8}$$

$n = 0, 1, 2, \ldots$. The corresponding marginal density functions are

$$\phi_X(x) = \frac{1}{\Gamma(\alpha)}\, x^{\alpha-1}\, \exp(-x)$$

and

$$\phi_Y(y) = \frac{1}{\Gamma(\beta)}\, y^{\beta-1}\, \exp(-y),$$

which we recognize as the density functions of one-dimensional gamma random variables with index parameters α and β, respectively.

We remark that if $\alpha = \beta$ then the density function (5.2.7) reduces to the Kibble-Moran bivariate gamma density function and $\mathrm{Corr}(X, Y) = \lambda$ (Kotz, et al. [56, pp. 436–437]). Also, (5.2.7) represents the Lancaster expansion for (X, Y).

5.2.4 The Bivariate Poisson Distribution

For $a > 0$ and $x, n \in \mathbb{N}_0$, let

$$C_n(x; a) = \left(\frac{a^n}{n!}\right)^{1/2} \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{x}{k} \frac{k!}{a^k} \tag{5.2.9}$$

denote the Poisson-Charlier polynomial of degree n. For $\lambda \in [0, 1]$, Koudou [58, Section 5] (cf., Bar-Lev, et al. [4], Pommeret [72]) shows that there exists a bivariate random vector (X, Y) with probability density function

$$\phi_{X,Y}(x, y) = \phi_X(x)\, \phi_Y(y) \sum_{n=0}^{\infty} \lambda^n\, C_n(x; a)\, C_n(y; a), \tag{5.2.10}$$

$x, y \in \mathbb{N}_0$. The corresponding marginal density functions ϕ_X and ϕ_Y are given by

$$\phi_X(k) = \phi_Y(k) = \frac{a^k\, \exp(-a)}{k!},$$

$k \in \mathbb{N}_0$, so that X and Y are distributed marginally according to a Poisson distribution with parameter a. The series (5.2.10) is an expansion of Lancaster type, a special case of (5.1.2), and the resulting distribution is called a bivariate Poisson distribution.

5.3 Distance Correlation Coefficients for Lancaster Distributions

In this section, we derive a general series expression for the distance correlation coefficients for Lancaster distributions with density functions of the form (5.1.2). For a joint density function $\phi_{X,Y}$ given by (5.1.2) and $n \in \mathcal{C}$, we introduce the notation

$$\mathcal{P}_n(s) = \mathbb{E} \exp(\mathrm{i}\, \langle s, X \rangle)\, P_n(X), \tag{5.3.1}$$

$s \in \mathbb{R}^p$, and

$$\mathcal{Q}_n(t) = \mathbb{E} \exp(\mathrm{i}\, \langle t, Y \rangle)\, Q_n(Y), \tag{5.3.2}$$

$t \in \mathbb{R}^q$. To verify that each expectation $\mathcal{P}_n(s)$ converges absolutely for all $s \in \mathbb{R}^p$, we apply the Cauchy-Schwarz inequality to obtain

$$\begin{aligned} |\mathcal{P}_n(s)|^2 &= |\mathbb{E}\exp(\mathrm{i}\langle s, X\rangle)P_n(X)|^2 \\ &\le \big(\mathbb{E}|\exp(\mathrm{i}\langle s, X\rangle)|^2\big) \cdot \big(\mathbb{E}|P_n(X)|^2\big) = 1, \end{aligned}$$

because $\{P_n : n \in \mathcal{C}\}$ is an orthonormal basis for the Hilbert space $L^2(\mu)$. Similarly, $|\mathcal{Q}_n(t)| \le 1$ for all $t \in \mathbb{R}^q$.

In the following result, we will use the notation

$$\mathcal{A}_{j,k} = \int_{\mathbb{R}^p} \mathcal{P}_j(s)\, \mathcal{P}_k(-s)\, \frac{\mathrm{d}s}{|s|_p^{p+1}} \tag{5.3.3}$$

and

$$\mathcal{B}_{j,k} = \int_{\mathbb{R}^q} \mathcal{Q}_j(t)\, \mathcal{Q}_k(-t)\, \frac{\mathrm{d}t}{|t|_q^{q+1}}, \tag{5.3.4}$$

$j, k \in \mathcal{C}$, whenever these integrals converge absolutely.

We now state the main result.

Theorem 5.3.1. *Suppose that the random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ have the joint probability density function (5.1.2). Then,*

$$\mathcal{V}^2(X, Y) = \frac{1}{\gamma_p \gamma_q} \sum_{j \in \mathcal{C}, j \ne 0} \sum_{k \in \mathcal{C}, k \ne 0} \rho_j\, \overline{\rho_k}\, \mathcal{A}_{j,k}\, \mathcal{B}_{j,k}, \tag{5.3.5}$$

whenever the sum converges absolutely.

Proof. Rewriting the Lancaster expansion (5.1.2) in the form,

$$\phi_{X,Y}(x, y) - \phi_X(x)\, \phi_Y(y) = \phi_X(x)\, \phi_Y(y) \sum_{n \in \mathcal{C}, n \ne 0} \rho_n P_n(x)\, Q_n(y),$$

and taking Fourier transforms on both sides of this identity, we obtain for all $s \in \mathbb{R}^p$ and $t \in \mathbb{R}^q$ the expansion

$$f_{X,Y}(s, t) - f_X(s)\, f_Y(t) = \sum_{n \in \mathcal{C}, n \ne 0} \rho_n \mathcal{P}_n(s)\, \mathcal{Q}_n(t). \tag{5.3.6}$$

This identity is valid subject to the requirement that we may interchange summation and integration, which is justified by the assumption that the sum in the final result

converges absolutely. Using (5.3.6) we deduce that

$$|f_{X,Y}(s,t) - f_X(s)f_Y(t)|^2 = \big(\, f_{X,Y}(s,t) - f_X(s)f_Y(t)\,\big)\big(\,\overline{f_{X,Y}(s,t) - f_X(s)f_Y(t)}\,\big)$$
$$= \sum_{j\in\mathcal{C},j\neq 0}\ \sum_{k\in\mathcal{C},k\neq 0} \rho_j\,\overline{\rho_k}\,\mathcal{P}_j(s)\,\mathcal{P}_k(-s)\,\mathcal{Q}_j(t)\,\mathcal{Q}_k(-t).$$

Next, we integrate this expansion with respect to the measures $\mathrm{d}s/|s|_p^{p+1}$ and $\mathrm{d}t/|t|_q^{q+1}$; this requires that we again interchange summation and integration which, by assumption, we are able to do. On carrying through these procedures, we obtain (5.3.5). □

5.4 Examples

In this section, we demonstrate the versatility of Theorem 5.3.1 by applying it to compute the distance correlation coefficients for the bivariate normal, multivariate normal, and bivariate gamma and Poisson distributions. We verify for each example the absolute convergence of the series resulting from Theorem 5.3.1, for that convergence property cannot in general be obtained from abstract Lancaster expansions. In developing each example, we retain the corresponding notation in Section 5.2.

5.4.1 The Bivariate Bormal Distribution

In the sequel, we use the standard double-factorial notation,

$$n!! = n(n-2)(n-4)\cdots = \begin{cases} n(n-2)(n-4)\cdots 2, & \text{if } n \text{ is even} \\ n(n-2)(n-4)\cdots 1, & \text{if } n \text{ is odd} \end{cases}$$

Proposition 5.4.1. *Let $(X,Y) \sim \mathcal{N}_2(0,\Sigma)$, a bivariate normal distribution with correlation coefficient ρ. Then,*

$$\mathcal{V}^2(X,Y) = 4\pi^{-1}\sum_{l=1}^{\infty} \frac{((2l-3)!!)^2}{(2l)!}\big(1 - 2^{-(2l-1)}\big)\rho^{2l}, \tag{5.4.1}$$

and this series converges absolutely for all $\rho \in (-1,1)$.

Proof. Starting with the Lancaster expansion of the bivariate normal density function, as given in (5.2.1), and using the definitions of $\mathcal{P}_n$ and $\mathcal{Q}_n$ in (5.3.1) and

(5.3.2), respectively, we obtain by substitution and integration-by-parts,

$$\begin{aligned}\mathcal{P}_n(s) = \mathcal{Q}_n(s) &= \int_{-\infty}^{\infty} \exp(\mathrm{i}sx)\frac{1}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}x^2\right)\, H_n(x)\,\mathrm{d}x \\ &= (\mathrm{i}s)^n \exp\left(-\tfrac{1}{2}s^2\right),\end{aligned}$$

$s \in \mathbb{R}$. Therefore,

$$\begin{aligned}\mathcal{A}_{j,k} = \mathcal{B}_{j,k} &= (-1)^k\,\mathrm{i}^{j+k}\int_{-\infty}^{\infty} s^{j+k-2}\,\exp(-s^2)\,\mathrm{d}s, \\ &= \begin{cases}(-1)^k\,\mathrm{i}^{j+k}\,\pi^{1/2}\left(\frac{1}{2}\right)^{(j+k-2)/2}\,(j+k-3)!!, & \text{if } j+k \text{ is even} \\ 0, & \text{otherwise}\end{cases}\end{aligned}$$

since the latter integral is a moment of the $\mathcal{N}(0, \frac{1}{2})$ distribution. By Theorem 5.3.1, we obtain

$$\mathcal{V}^2(X,Y) = \frac{4}{\pi}\sum_{\substack{j,\,k>0 \\ j+k \text{ even}}} \frac{\rho^{j+k}}{j!\,k!}\,\left(\tfrac{1}{2}\right)^{j+k}\,\left((j+k-3)!!\right)^2.$$

Setting $j + k = 2l$ with $l \geq 1$, the double series reduces to

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \frac{4}{\pi}\sum_{l=1}^{\infty}\rho^{2l}(\tfrac{1}{2})^{2l}((2l-3)!!)^2\sum_{\substack{j,k\geq 1 \\ j+k=2l}}\frac{1}{j!\,k!} \\ &= \frac{4}{\pi}\sum_{l=1}^{\infty}\rho^{2l}(\tfrac{1}{2})^{2l}\frac{((2l-3)!!)^2}{(2l)!}\sum_{j=1}^{2l-1}\frac{(2l)!}{j!\,(2l-j)!} \\ &= \frac{4}{\pi}\sum_{l=1}^{\infty}\rho^{2l}(\tfrac{1}{2})^{2l}\frac{((2l-3)!!)^2}{(2l)!}(2^{2l}-2),\end{aligned}$$

which is the same as (5.4.1).

The absolute convergence of (5.4.1) can be verified by comparison with a geometric series. Moreover, it is straightforward to verify that the series is identical with the result obtained by Székely, et al. (see Theorem 2.4.10). □

Having obtained $\mathcal{V}(X,Y)$, we let $\rho \to 1-$ to obtain the distance variances $\mathcal{V}(X,X)$ and $\mathcal{V}(Y,Y)$; here, we are applying a well-known result that if $(X,Y) \sim \mathcal{N}_2(0,\Sigma)$ where $\mathrm{Var}(X) = \mathrm{Var}(Y)$ and $\rho = 1$ then $X = Y$, almost surely. Exactly as in (3.2.18), we obtain

$$\mathcal{V}^2(X,X) = \mathcal{V}^2(Y,Y) = \frac{4}{3} - \frac{4(\sqrt{3}-1)}{\pi}.$$

5.4.2 The Multivariate Normal Distribution

In this subsection, we will make extensive use of the notation $\boldsymbol{N}_{r\bullet}$, $\boldsymbol{N}_{\bullet c}$, $\boldsymbol{N}_{*\bullet}$, $\boldsymbol{N}_{\bullet *}$, and $\boldsymbol{N}_{\bullet\bullet}$ from Subsection 5.2.2 for the multi-index matrix $\boldsymbol{N} \in \mathbb{N}_0^{p\times q}$. We now establish the following result.

Proposition 5.4.2. *Let $(X, Y) \sim \mathcal{N}_{p+q}(0, \Sigma)$, where Σ is given in (5.2.2). Then the affinely invariant distance covariance, $\widetilde{\mathcal{V}}^2(X, Y)$, is given by*

$$\widetilde{\mathcal{V}}^2(X,Y) = \frac{1}{\gamma_p\,\gamma_q} \sum_{\boldsymbol{J}\neq\boldsymbol{0}} \sum_{\boldsymbol{K}\neq\boldsymbol{0}} A_{\boldsymbol{J},\boldsymbol{K}}\, B_{\boldsymbol{J},\boldsymbol{K}}\, \frac{\Lambda_{XY}^{\boldsymbol{J}}}{\boldsymbol{J}!} \frac{\Lambda_{XY}^{\boldsymbol{K}}}{\boldsymbol{K}!}, \tag{5.4.2}$$

where the sums are taken over all non-zero $\boldsymbol{J}, \boldsymbol{K} \in \mathbb{N}_0^{p\times q}$ such that all components of $\boldsymbol{J}_{*\bullet} + \boldsymbol{K}_{*\bullet}$ and $\boldsymbol{J}_{\bullet *} + \boldsymbol{K}_{\bullet *}$ are even,

$$A_{\boldsymbol{J},\boldsymbol{K}} = \frac{\Gamma\left(\frac{1}{2}(\boldsymbol{J}_{\bullet\bullet} + \boldsymbol{K}_{\bullet\bullet} - 1)\right)}{\Gamma\left(\frac{1}{2}(\boldsymbol{J}_{\bullet\bullet} + \boldsymbol{K}_{\bullet\bullet}) + \frac{1}{2}p\right)} \prod_{r=1}^{p} \Gamma\big(\tfrac{1}{2}(\boldsymbol{J}_{r\bullet} + \boldsymbol{K}_{r\bullet} + 1)\big) \tag{5.4.3}$$

and

$$B_{\boldsymbol{J},\boldsymbol{K}} = \frac{\Gamma\left(\frac{1}{2}(\boldsymbol{J}_{\bullet\bullet} + \boldsymbol{K}_{\bullet\bullet} - 1)\right)}{\Gamma\left(\frac{1}{2}(\boldsymbol{J}_{\bullet\bullet} + \boldsymbol{K}_{\bullet\bullet}) + \frac{1}{2}q\right)} \prod_{c=1}^{q} \Gamma\big(\tfrac{1}{2}(\boldsymbol{J}_{\bullet c} + \boldsymbol{K}_{\bullet c} + 1)\big). \tag{5.4.4}$$

Proof. In this case, the index set $\mathcal{C}$ is $\mathbb{N}_0^{p\times q}$, and we write the Lancaster expansion (5.2.5) of $(\widetilde{X}, \widetilde{Y})$ in the form

$$\phi_{\widetilde{X},\widetilde{Y}}(x,y) - \phi_{\widetilde{X}}(x)\,\phi_{\widetilde{Y}}(y) = \phi_{\widetilde{X}}(x)\,\phi_{\widetilde{Y}}(y) \sum_{\boldsymbol{N}\neq\boldsymbol{0}} \frac{\Lambda_{XY}^{\boldsymbol{N}}}{\boldsymbol{N}!}\, H_{\boldsymbol{N}_{*\bullet}}(x; I_p)\, H_{\boldsymbol{N}_{\bullet *}}(y; I_q).$$

To calculate the Fourier transform $\mathcal{P}_{\boldsymbol{N}}$ corresponding to $\widetilde{X}$, we apply the definition (5.2.3) of the multivariate Hermite polynomials and integration-by-parts to deduce that for $s \in \mathbb{R}^p$,

$$\begin{aligned}
\mathcal{P}_{\boldsymbol{N}}(s) &= \int_{\mathbb{R}^p} \exp(\mathrm{i}\langle s, x\rangle)\, \phi_{\widetilde{X}}(x)\, H_{\boldsymbol{N}_{*\bullet}}(x; I_p)\, \mathrm{d}x \\
&= (-1)^{\boldsymbol{N}_{\bullet\bullet}} \int_{\mathbb{R}^p} \exp(\mathrm{i}\langle s, x\rangle) \left(\frac{\partial}{\partial x}\right)^{\boldsymbol{N}_{*\bullet}} \phi_{\widetilde{X}}(x)\, \mathrm{d}x \\
&= \int_{\mathbb{R}^p} \phi_{\widetilde{X}}(x) \left(\frac{\partial}{\partial x}\right)^{\boldsymbol{N}_{*\bullet}} \exp(\mathrm{i}\langle s, x\rangle)\, \mathrm{d}x \\
&= (\mathrm{i}s)^{\boldsymbol{N}_{*\bullet}} \int_{\mathbb{R}^p} \phi_{\widetilde{X}}(x)\, \exp(\mathrm{i}\langle s, x\rangle)\, \mathrm{d}x \\
&= \mathrm{i}^{\boldsymbol{N}_{\bullet\bullet}}\, s^{\boldsymbol{N}_{*\bullet}}\, \exp(-\tfrac{1}{2}\langle s, s\rangle).
\end{aligned}$$

Similarly,

$$\mathcal{Q}_{\boldsymbol{N}}(t) = \mathrm{i}^{\boldsymbol{N}_{\bullet\bullet}}\, t^{\boldsymbol{N}_{\bullet *}}\, \exp(-\tfrac{1}{2}\langle t, t\rangle),$$

$t \in \mathbb{R}^q$. Therefore,

$$\int_{\mathbb{R}^p} \mathcal{P}_{\boldsymbol{J}}(s)\, \mathcal{P}_{\boldsymbol{K}}(-s) \frac{\mathrm{d}s}{|s|_p^{p+1}} = (-1)^{\boldsymbol{K}_{\bullet\bullet}}\, \mathrm{i}^{\boldsymbol{J}_{\bullet\bullet}+\boldsymbol{K}_{\bullet\bullet}} \int_{\mathbb{R}^p} s^{\boldsymbol{J}_{*\bullet}+\boldsymbol{K}_{*\bullet}}\, \exp(-\langle s, s\rangle) \frac{\mathrm{d}s}{|s|_p^{p+1}}.$$

We now change variables to polar coordinates: $s = r\omega$, where $r > 0$ and $\omega = (\omega_1, \ldots, \omega_p) \in S^{p-1}$, the unit sphere in $\mathbb{R}^p$. By 2.1.3 the latter integral reduces to

$$\int_{\mathbb{R}_+} r^{\boldsymbol{J}_{\bullet\bullet}+\boldsymbol{K}_{\bullet\bullet}-2}\, \exp(-r^2)\, \mathrm{d}r \cdot \int_{S^{p-1}} \omega^{\boldsymbol{J}_{*\bullet}+\boldsymbol{K}_{*\bullet}}\, \mathrm{d}\omega.$$

The integral over $\mathbb{R}_+$ is evaluated by replacing r by $r^{1/2}$, and we obtain its value as $\frac{1}{2}\,\Gamma\left(\frac{1}{2}(\boldsymbol{J}_{\bullet\bullet} + \boldsymbol{K}_{\bullet\bullet} - 1)\right)$.
It is easy to see that the integral over S^{p-1} equals zero if any component of $\boldsymbol{J}_{*\bullet} + \boldsymbol{K}_{*\bullet}$ is odd. For the case in which each component of $\boldsymbol{J}_{*\bullet} + \boldsymbol{K}_{*\bullet}$ is even, we obtain

$$\int_{S^{p-1}} \omega^{\boldsymbol{J}_{*\bullet}+\boldsymbol{K}_{*\bullet}}\, \mathrm{d}\omega = A(S^{p-1})\, \mathbb{E}(\omega^{\boldsymbol{J}_{*\bullet}+\boldsymbol{K}_{*\bullet}}),$$

where $A(S^{p-1}) = 2\pi^{p/2}/\Gamma(\frac{1}{2}p)$ is the surface area of S^{p-1} and ω now is a uniformly distributed random vector on S^{p-1}. It is well-known that the random vector $(\omega_1^2, \ldots, \omega_p^2) \sim D(\frac{1}{2}, \ldots, \frac{1}{2})$, a Dirichlet distribution with parameters $(\frac{1}{2}, \ldots, \frac{1}{2})$; so, by a classical formula for the moments of the Dirichlet distribution [56, p. 488],

$$\mathbb{E}(\omega^{\boldsymbol{J}_{*\bullet}+\boldsymbol{K}_{*\bullet}}) = \frac{\Gamma(\frac{1}{2}p)}{[\Gamma(\frac{1}{2})]^p} \frac{\prod_{r=1}^p \Gamma(\frac{1}{2}(\boldsymbol{J}_{r\bullet} + \boldsymbol{K}_{r\bullet} + 1))}{\Gamma(\frac{1}{2}(\boldsymbol{J}_{\bullet\bullet} + \boldsymbol{K}_{\bullet\bullet}) + \frac{1}{2}p)}.$$

Collecting together these results, we obtain

$$\int_{\mathbb{R}^p} \mathcal{P}_{\boldsymbol{J}}(s)\, \mathcal{P}_{\boldsymbol{K}}(-s) \frac{\mathrm{d}s}{|s|_p^{p+1}} = (-1)^{\boldsymbol{K}_{\bullet\bullet}}\, (-1)^{(\boldsymbol{J}_{\bullet\bullet}+\boldsymbol{K}_{\bullet\bullet})/2}\, A_{\boldsymbol{J},\boldsymbol{K}},$$

where $A_{\boldsymbol{J},\boldsymbol{K}}$ is given in (5.4.3). A similar expression can be obtained for

$$\int_{\mathbb{R}^q} \mathcal{Q}_{\boldsymbol{J}}(t)\, \mathcal{Q}_{\boldsymbol{K}}(-t) \frac{\mathrm{d}t}{|t|_q^{q+1}},$$

from which the final result (5.4.2) follows. □

Similar to the bivariate normal case, the affinely invariant distance variance $\widetilde{\mathcal{V}}^2(X, X)$ in the multivariate case can be calculated by taking $p = q$ and $\Lambda_{XY} = \rho I_p$, where

$-1 < \rho < 1$, and then letting $\rho \to 1-$ in the expression for $\widetilde{\mathcal{V}}^2(X,Y)$.
We remark also that the distance covariance and distance correlation for non-standardized jointly normal random vectors can be calculated using the arguments used earlier, and we refer to Appendix A.1 for the explicit formula.

5.4.3 The Bivariate Gamma Distribution

Proposition 5.4.3. *Suppose that the random vector (X,Y) is distributed according to a Sarmanov bivariate gamma distribution, as given by (5.2.7). Then,*

$$\mathcal{V}^2(X,Y) = 2^{2(1-\alpha-\beta)} \frac{\Gamma(2\alpha+1)\Gamma(2\beta+1)}{\Gamma(\alpha)\Gamma(\beta)} \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_j\, a_k\, A_{j,k}(\alpha)\, A_{j,k}(\beta), \tag{5.4.5}$$

where

$$A_{j,k}(\alpha) = \frac{(\alpha)_j\,(\alpha)_k\,(1-\alpha-j)_{j+k-2}}{j!\,k!\,(\alpha-j+2)_{j+k-2}\,\Gamma(\alpha-j+2)}\ {}_2F_1\left(-j-k+2, 2\alpha; \alpha-k+2; \tfrac{1}{2}\right).$$

Proof. By (5.2.7), there holds the expansion,

$$\phi_{X,Y}(x,y) - \phi_X(x)\,\phi_Y(y) = \phi_X(x)\,\phi_Y(y) \sum_{n=1}^{\infty} a_n L_n^{(\alpha-1)}(x)\, L_n^{(\beta-1)}(y),$$

$x, y > 0$. Then, it follows from (5.3.1) that for $s, t \in \mathbb{R}$,

$$\begin{aligned}\mathcal{P}_n(s) &= \int_0^{\infty} \exp(\mathrm{i}sx)\, L_n^{(\alpha-1)}(x)\, \phi_X(x)\, \mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \exp\big(-(1-\mathrm{i}s)x\big)\, x^{\alpha-1}\, L_n^{(\alpha-1)}(x)\, \mathrm{d}x.\end{aligned}$$

By a direct calculation using (5.2.6), we obtain

$$\begin{aligned}\mathcal{P}_n(s) &= \frac{(\alpha)_n}{n!}\,(1-\mathrm{i}s)^{-\alpha}\,\left(1-(1-\mathrm{i}s)^{-1}\right)^n \\ &= \frac{(\alpha)_n}{n!}\,(1-\mathrm{i}s)^{-(\alpha+n)}\,(-\mathrm{i}s)^n\end{aligned}$$

and, analogously,

$$\mathcal{Q}_n(t) = \frac{(\beta)_n}{n!}\,(1-\mathrm{i}t)^{-(\beta+n)}\,(-\mathrm{i}t)^n.$$

We now calculate the integral

$$\int_{\mathbb{R}} \mathcal{P}_j(s)\,\mathcal{P}_k(-s)\,\frac{\mathrm{d}s}{s^2} \equiv \frac{(\alpha)_j}{j!}\,\frac{(\alpha)_k}{k!}\,\mathrm{i}^{-j+k} \int_{\mathbb{R}} g(s)\,\mathrm{d}s, \tag{5.4.6}$$

where

$$g(s) = s^{j+k-2}\,(1-\mathrm{i}s)^{-(\alpha+j)}\,(1+\mathrm{i}s)^{-(\alpha+k)}, \tag{5.4.7}$$

$s \in \mathbb{R}$. To calculate the integral on the right-hand side of (5.4.6), we utilize Cauchy's beta integral [2, p. 48]: For $a, u, v \in \mathbb{C}$ such that $\mathrm{Re}(a) > 0$ and $\mathrm{Re}(u+v) > 1$,

$$\int_{\mathbb{R}} (1-\mathrm{i}s)^{-u}\,(1+\mathrm{i}as)^{-v}\,\mathrm{d}s = 2\pi \frac{\Gamma(u+v-1)}{\Gamma(u)\Gamma(v)}\,a^{u-1}\,(a+1)^{2-u-v}. \tag{5.4.8}$$

To differentiate the left-hand side of (5.4.8) m times with respect to a, we apply the formula,

$$\Big(\frac{\partial}{\partial a}\Big)^m (1+\mathrm{i}as)^{-v} = (-\mathrm{i})^m s^m (v)_m\,(1+\mathrm{i}as)^{-v-m};$$

by differentiating under the integral we obtain

$$\Big(\frac{\partial}{\partial a}\Big)^m \int_{\mathbb{R}} (1-\mathrm{i}s)^{-u}\,(1+\mathrm{i}as)^{-v}\mathrm{d}s = (-\mathrm{i})^m (v)_m \int_{\mathbb{R}} s^m\,(1-\mathrm{i}s)^{-u}\,(1+\mathrm{i}as)^{-v-m}\,\mathrm{d}s.$$

To differentiate the right-hand side of (5.4.8) m times with respect to a, we apply Leibniz's formula:

$$\Big(\frac{\partial}{\partial a}\Big)^m \Big[a^{u-1}\,(a+1)^{2-u-v}\Big] = \sum_{l=0}^{m} \binom{m}{l} \Big[\Big(\frac{\partial}{\partial a}\Big)^{m-l} a^{u-1}\Big] \cdot \Big[\Big(\frac{\partial}{\partial a}\Big)^{l} (a+1)^{2-u-v}\Big].$$

Noting that

$$\binom{m}{l} = \frac{(-1)^l (-m)_l}{l!},$$
$$\Big(\frac{\partial}{\partial a}\Big)^{m-l} a^{u-1} = (-1)^m\,a^{u-1-m+l}\,\frac{(1-u)_m}{(u-m)_l},$$

and

$$\Big(\frac{\partial}{\partial a}\Big)^{l} (a+1)^{2-u-v} = (-1)^l\,(a+1)^{2-u-v-l}\,(-2+u+v)_l,$$

we obtain

$$\begin{aligned}\Big(\frac{\partial}{\partial a}\Big)^m \Big[a^{u-1}\,(a+1)^{2-u-v}\Big] \\ &= (-1)^m a^{u-1-m}(a+1)^{2-u-v}(1-u)_m \sum_{l=0}^{m} \frac{(-m)_l\,(-2+u+v)_l}{l!\,(u-m)_l}\,a^l\,(a+1)^{-l} \\ &= (-1)^m a^{u-1-m}(a+1)^{2-u-v}(1-u)_m\ {}_2F_1\Big(-m,-2+u+v;u-m;\frac{a}{a+1}\Big),\end{aligned}$$

where ${}_2F_1$ denotes Gauss' hypergeometric function (see section 2.3).

Comparing the derivatives of the left- and right-hand sides of (5.4.8), we obtain

$$\begin{aligned}\int_{\mathbb{R}} s^m\,(1-\mathrm{i}s)^{-u}\,(1+\mathrm{i}as)^{-v-m}\,\mathrm{d}s = 2\pi\,(-\mathrm{i})^m\,a^{u-1-m}(a+1)^{2-u-v}\,\frac{\Gamma(u+v-1)}{\Gamma(u)\Gamma(v)} \\ \times\,\frac{(1-u)_m}{(v)_m}\,{}_2F_1\Big(-m,-2+u+v;u-m;\frac{a}{a+1}\Big).\end{aligned}$$

Substituting $a=1$, $m=j+k-2$, $u=\alpha+j$, and $v=\alpha+k-m\equiv\alpha-j+2$, the latter equation reduces to

$$\begin{aligned}\int_{\mathbb{R}} g(s)\,\mathrm{d}s = 2^{-2\alpha+1}\,\pi\,(-\mathrm{i})^{j+k-2}\,\frac{\Gamma(2\alpha+1)}{\Gamma(\alpha+j)\Gamma(\alpha-j+2)} \\ \times\,\frac{(1-\alpha-j)_{j+k-2}}{(\alpha-j+2)_{j+k-2}}\,{}_2F_1\big(-j-k+2,2\alpha;\alpha-k+2;\tfrac{1}{2}\big).\end{aligned}$$

Therefore,

$$\begin{aligned}\int_{\mathbb{R}} \mathcal{P}_j(s)\,\mathcal{P}_k(-s)\,\frac{\mathrm{d}s}{s^2} = 2^{-2\alpha+1}\pi\,(-1)^{j-1}\,\frac{(\alpha)_j\,(\alpha)_k}{j!\,k!}\,\frac{\Gamma(2\alpha+1)}{\Gamma(\alpha+j)\Gamma(\alpha-j+2)} \\ \times\,\frac{(1-\alpha-j)_{j+k-2}}{(\alpha-j+2)_{j+k-2}}\,{}_2F_1\big(-j-k+2,2\alpha;\alpha-k+2;\tfrac{1}{2}\big),\end{aligned}$$

and similarly for Y. Substituting these expressions into Theorem 5.3.1 and simplifying the outcome, we obtain the series (5.4.5) as a formal expression for $\mathcal{V}^2(X,Y)$.

Finally, we verify that (5.4.5) converges absolutely. By (5.4.7),

$$\int_{\mathbb{R}} |g(s)|\,\mathrm{d}s = \int_{\mathbb{R}} |s|^{j+k-2}\,(1+s^2)^{-(2\alpha+j+k)/2}\,\mathrm{d}s.$$

Making the change-of-variables $s^2=t/(1-t)$, the latter integral is transformed to

$$\int_0^1 t^{\frac{1}{2}(j+k-3)}\,(1-t)^{\alpha-\frac{1}{2}}\,\mathrm{d}t = B\left(\tfrac{1}{2}(j+k-1),\alpha+\tfrac{1}{2}\right), \tag{5.4.9}$$

where $B(\cdot,\cdot)$ is the classical beta function, and this integral converges absolutely because $j+k-1>0$ and $\alpha+1/2>0$ for all $j,k\in\mathbb{N}$ and $\alpha>0$. Hence, to establish that (5.4.5) converges absolutely, we need only show that the series

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_j\,a_k\,\frac{(\alpha)_j(\beta)_j}{(j!)^2}\,\frac{(\alpha)_k\,(\beta)_k}{(k!)^2} \\ \times B\left(\tfrac{1}{2}(j+k-1),\alpha+\tfrac{1}{2}\right)\,B\left(\tfrac{1}{2}(j+k-1),\beta+\tfrac{1}{2}\right) \qquad (5.4.10)$$

converges absolutely.

By (5.2.8), $0\le a_j\le\lambda^j\le 1$ for all j. Also, for $j+k\ge 3$, it follows from (5.4.9) that

$$B\left(\tfrac{1}{2}(j+k-1),\alpha+\tfrac{1}{2}\right)\le\int_0^1(1-t)^{\alpha-\frac{1}{2}}\,\mathrm{d}t=\frac{1}{\alpha+\frac{1}{2}}.$$

Therefore, (5.4.10) is bounded above by

$$\begin{aligned}
&\alpha^2\beta^2\lambda^2\left[B\left(\tfrac{1}{2},\alpha+\tfrac{1}{2}\right)\right]^2+\frac{1}{(\alpha+\frac{1}{2})^2}\sum_{\substack{j,\,k\ge 1\\ j+k\ge 3}}\frac{(\alpha)_j(\beta)_j}{(j!)^2}\,\frac{(\alpha)_k\,(\beta)_k}{(k!)^2}\lambda^{j+k}\\
&\quad\le\alpha^2\beta^2\lambda^2\left[B\left(\tfrac{1}{2},\alpha+\tfrac{1}{2}\right)\right]^2+\frac{1}{(\alpha+\frac{1}{2})^2}\Big(\sum_{j=0}^{\infty}\frac{(\alpha)_j(\beta)_j}{(j!)^2}\lambda^j\Big)\Big(\sum_{k=0}^{\infty}\frac{(\alpha)_k\,(\beta)_k}{(k!)^2}\lambda^k\Big)\\
&\quad\equiv\alpha^2\beta^2\lambda^2\big[B\big(\tfrac{1}{2},\alpha+\tfrac{1}{2}\big)\big]^2+\frac{1}{(\alpha+\frac{1}{2})^2}\,\big[{}_2F_1(\alpha,\beta;1;\lambda)\big]^2,
\end{aligned}$$

and it is well-known that this Gaussian hypergeometric series converges absolutely for all $\alpha,\beta\in\mathbb{C}$ and all $\lambda\in[0,1]$. $\square$

In calculating the distance variances $\mathcal{V}(X,X)$ and $V(Y,Y)$, only the marginal distributions are relevant. Therefore, we may assume that X and Y have any joint distribution for which the marginal distributions are gamma with parameters α and β, respectively. Letting $\beta\to\alpha$, the Sarmanov bivariate gamma distribution reduces to the Kibble-Moran distribution, and then the joint characteristic function of (X,Y) is

$$\big((1-\mathrm{i}t_1)(1-\mathrm{i}t_2)+\lambda t_1t_2\big)^{-\alpha};$$

cf. [56, p. 436]. Next, we let $\lambda\to 1-$; then this characteristic function converges to

$$\big(1-\mathrm{i}(t_1+t_2)\big)^{-\alpha}\equiv\mathbb{E}\exp\big(\mathrm{i}(t_1+t_2)X\big),$$

proving that, for $\lambda=1$, $X=Y$, almost surely. Therefore, the distance variance

$\mathcal{V}(X,X)$ is a limiting case of $\mathcal{V}(X,Y)$, viz.,

$$\begin{aligned}\mathcal{V}^2(X,X) &= \frac{1}{\gamma_1^2}\int_{\mathbb{R}^2} |f_X(s+t) - f_X(s)f_X(t)|^2 \frac{\mathrm{d}s}{s^2}\frac{\mathrm{d}t}{t^2} \\ &= \lim_{\lambda\to 1-}\lim_{\beta\to\alpha} \frac{1}{\gamma_1^2}\int_{\mathbb{R}^2} |f_{X,Y}(s,t) - f_X(s)f_Y(t)|^2 \frac{\mathrm{d}s}{s^2}\frac{\mathrm{d}t}{t^2} \\ &= \lim_{\lambda\to 1-}\lim_{\beta\to\alpha} \mathcal{V}^2(X,Y).\end{aligned}$$

Similarly,

$$\mathcal{V}^2(Y,Y) = \lim_{\lambda\to 1-}\lim_{\alpha\to\beta} \mathcal{V}^2(X,Y).$$

5.4.4 The Bivariate Poisson Distribution

Proposition 5.4.4. *Suppose that the random vector (X,Y) is distributed according to a bivariate Poisson distribution, as given by (5.2.10). Then*

$$\mathcal{V}^2(X,Y) = \frac{1}{\pi}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \lambda^{j+k}\Big(\frac{(4a)^{j+k}}{j!\,k!}\Big) A_{jk}^2, \tag{5.4.11}$$

where

$$\begin{aligned}A_{jk} = \sum_{\substack{l=0\\ l \text{ even}}}^{j-k} \binom{j-k}{l} (-1)^{l/2} \sum_{n=0}^{l/2} (-1)^n \frac{\Gamma(n+j-\frac{1}{2}l-\frac{1}{2})}{\Gamma(n+j-\frac{1}{2}l)} \\ \times {}_1F_1(n+j-\tfrac{1}{2}l-\tfrac{1}{2}; n+j-\tfrac{1}{2}l; -4a)\end{aligned} \tag{5.4.12}$$

for $j \geq k$, and $A_{jk} = A_{kj}$ for $j < k$.

Proof. By (5.2.10) and (5.3.1), we have

$$\mathcal{P}_n(s) = \mathcal{Q}_n(s) = \mathbb{E}\exp(\mathrm{i}sX)\, C_n(X;a),$$

$s \in \mathbb{R}$. Substituting the definition (5.2.9) of the Poisson-Charlier polynomials C_n into the expectation and reversing the order of summation, we obtain

$$\begin{aligned}\mathcal{P}_n(s) = \mathcal{Q}_n(s) &= \sum_{x=0}^{\infty} \exp(\mathrm{i}sx) C_n(x;a) \frac{\mathrm{e}^{-a}a^x}{x!} \\ &= \Big(\frac{a^n}{n!}\Big)^{1/2} (1-\mathrm{e}^{\mathrm{i}s})^n \exp\big(-a(1-\mathrm{e}^{\mathrm{i}s})\big).\end{aligned}$$

Therefore, for $j, k \geq 1$,

$$\begin{aligned}
&\int_{\mathbb{R}} \mathcal{P}_j(s)\mathcal{P}_k(-s)\frac{\mathrm{d}s}{s^2} \\
&\quad = \Big(\frac{a^{j+k}}{j!\,k!}\Big)^{1/2} \int_{\mathbb{R}} (1-\mathrm{e}^{\mathrm{i}s})^j\,(1-\mathrm{e}^{-\mathrm{i}s})^k\,\exp\big(-a(1-\mathrm{e}^{\mathrm{i}s}+1-\mathrm{e}^{-\mathrm{i}s})\big)\,\frac{\mathrm{d}s}{s^2} \\
&\quad = \Big(\frac{a^{j+k}}{j!\,k!}\Big)^{1/2} \int_{\mathbb{R}} (1-\mathrm{e}^{\mathrm{i}s})^j\,(1-\mathrm{e}^{-\mathrm{i}s})^k\,\exp\big(-2a(1-\cos s)\big)\frac{\mathrm{d}s}{s^2}. \qquad (5.4.13)
\end{aligned}$$

Because this integral is symmetric in j and k then we can assume, with no loss of generality, that $j \geq k$. We now write

$$\begin{aligned}
(1-e^{\mathrm{i}s})^j(1-e^{-\mathrm{i}s})^k &= (1-e^{\mathrm{i}s})^{j-k}(1-e^{\mathrm{i}s})^k(1-e^{-\mathrm{i}s})^k \\
&= (1-e^{\mathrm{i}s})^{j-k}(2(1-\cos s))^k,
\end{aligned}$$

and apply the binomial theorem in the form,

$$\begin{aligned}
(1-e^{\mathrm{i}s})^{j-k} &= (1-\cos s-\mathrm{i}\sin s)^{j-k} \\
&= \sum_{l=0}^{j-k}\binom{j-k}{l}(-\mathrm{i}\sin s)^l(1-\cos s)^{j-k-l}.
\end{aligned}$$

Then, it follows that the integral in (5.4.13) equals

$$2^k\sum_{l=0}^{j-k}\binom{j-k}{l}(-\mathrm{i})^l\int_{\mathbb{R}}(\sin s)^l(1-\cos s)^{j-l}\,\exp\big(-2a(1-\cos s)\big)\frac{\mathrm{d}s}{s^2}. \qquad (5.4.14)$$

Expanding the exponential term,

$$\exp\big(-2a(1-\cos s)\big) = \sum_{m=0}^{\infty}\frac{(-2a)^m}{m!}(1-\cos s)^m,$$

applying the half-angle identities, $\sin s = 2\sin\frac{1}{2}s\cos\frac{1}{2}s$ and $1-\cos s = 2(\sin\frac{1}{2}s)^2$, and integrating term-by-term, we deduce that (5.4.14) equals

$$\begin{aligned}
&2^k\sum_{l=0}^{j-k}\binom{j-k}{l}(-\mathrm{i})^l\sum_{m=0}^{\infty}\frac{(-2a)^m}{m!}\int_{\mathbb{R}}(2\sin\tfrac{1}{2}s\cos\tfrac{1}{2}s)^l(2(\sin\tfrac{1}{2}s)^2)^{j-l+m}\frac{\mathrm{d}s}{s^2} \\
&\quad = \sum_{l=0}^{j-k}\binom{j-k}{l}(-\mathrm{i})^l\sum_{m=0}^{\infty}\frac{(-a)^m}{m!}2^{j+k+2m}\int_{\mathbb{R}}(\cos\tfrac{1}{2}s)^l(\sin\tfrac{1}{2}s)^{2(j+m)-l}\frac{\mathrm{d}s}{s^2}. \qquad (5.4.15)
\end{aligned}$$

If l is odd then the latter integral is an odd function of s, so the integral equals 0. For the case in which l is even, we apply the identity $\sin^2 s = 1-\cos^2 s$ to write the integral

in (5.4.15) as

$$\int_{\mathbb{R}} (\cos^2 \tfrac{1}{2}s)^{l/2} (\sin \tfrac{1}{2}s)^{2(j+m)-l} \frac{\mathrm{d}s}{s^2} = \int_{\mathbb{R}} (1-\sin^2 \tfrac{1}{2}s)^{l/2} (\sin \tfrac{1}{2}s)^{2(j+m)-l} \frac{\mathrm{d}s}{s^2}. \tag{5.4.16}$$

To calculate the latter integral, we will expand the first term in the integrand by the binomial theorem and then integrate termwise. Applying the formula (Gradshteyn and Ryzhik [34, p. 483, 3.821(10)]),

$$\int_{\mathbb{R}} (\sin \tfrac{1}{2}s)^{2k} \frac{\mathrm{d}s}{s^2} = \begin{cases} \pi, & k = 1 \\ \dfrac{(2k-3)!!}{(2k-2)!!}\,\pi, & k = 2,3,4,\ldots \end{cases} \tag{5.4.17}$$

we find that (5.4.16) equals

$$\sum_{n=0}^{l/2} (-1)^n \int_{\mathbb{R}} (\sin \tfrac{1}{2}s)^{2(n+j+m)-l} \frac{\mathrm{d}s}{s^2} = \pi \sum_{n=0}^{l/2} (-1)^n \frac{(2(n+j+m)-l-3)!!}{(2(n+j+m)-l-2)!!}.$$

Substituting this result into (5.4.15), and interchanging the order of summation over m and n, we obtain

$$\begin{aligned} \int_{\mathbb{R}} \mathcal{P}_j(s)\mathcal{P}_k(-s) \frac{\mathrm{d}s}{s^2} = \pi \Big(\frac{(4a)^{j+k}}{j!\,k!} \Big)^{1/2} \sum_{\substack{l=0 \\ l \text{ even}}}^{j-k} \binom{j-k}{l} (-1)^{l/2} \\ \times \sum_{n=0}^{l/2} (-1)^n \sum_{m=0}^{\infty} \frac{(-4a)^m}{m!} \frac{(2(n+j+m)-l-3)!!}{(2(n+j+m)-l-2)!!}. \end{aligned} \tag{5.4.18}$$

Writing each double factorial in terms of rising factorials, and simplifying the resulting expressions, we find that (5.4.18) equals

$$\begin{aligned} \pi^{1/2} \Big(\frac{(4a)^{j+k}}{j!\,k!} \Big)^{1/2} \sum_{\substack{l=0 \\ l \text{ even}}}^{j-k} \binom{j-k}{l} (-1)^{l/2} \\ \times \sum_{n=0}^{l/2} (-1)^n \frac{\Gamma(n+j-\frac{1}{2}l-\frac{1}{2})}{\Gamma(n+j-\frac{1}{2}l)}\, {}_1F_1(n+j-\tfrac{1}{2}l-\tfrac{1}{2}; n+j-\tfrac{1}{2}l; -4a), \end{aligned} \tag{5.4.19}$$

where ${}_1F_1$ denotes the confluent hypergeometric function.

We remark that the individual terms in this series can be calculated in a straightforward way by differentiating a simpler hypergeometric series. Note that each confluent hypergeometric function in (5.4.19) is of the form ${}_1F_1(r-\frac{1}{2}; r; -4a)$ for $r \in \mathbb{N}$; for $r = 1$,

this latter function satisfies the well-known Kummer transformation [2, p. 191],

$$ {}_1F_1(\tfrac{1}{2};1;-4a) \equiv e^{-2a}\,{}_0F_1(1;a^2); $$

and for $r \geq 1$ we may differentiate this identity with respect to a, using the well-known formula [2, p. 94],

$$ {}_1F_1(r-\tfrac{1}{2};r;a) = \frac{(1)_{r-1}}{(\frac{1}{2})_{r-1}} \Big(\frac{\partial}{\partial a}\Big)^{r-1} {}_1F_1(\tfrac{1}{2};1;a). $$

Finally, we establish the absolute convergence of the resulting series for $\mathcal{V}^2(X,Y)$. On applying to (5.4.13) the identity

$$ |1-\mathrm{e}^{\mathrm{i}s}| = |1-\mathrm{e}^{-\mathrm{i}s}| = \big(2(1-\cos s)\big)^{1/2} = 2\big(\sin^2 \tfrac{1}{2}s\big)^{1/2} $$

and the inequality

$$ \exp\big(-2(1-\cos s)\big) \leq 1, $$

$s \in \mathbb{R}$, we obtain

$$ \begin{aligned} \left| \int_{\mathbb{R}} \mathcal{P}_j(s)\mathcal{P}_k(-s)\frac{\mathrm{d}s}{s^2} \right| &\leq \Big(\frac{a^{j+k}}{j!\,k!}\Big)^{1/2} \int_{\mathbb{R}} |1-\mathrm{e}^{\mathrm{i}s}|^j |1-\mathrm{e}^{\mathrm{i}s}|^k \exp\big(-2a(1-\cos s)\big)\frac{\mathrm{d}s}{s^2} \\ &\leq \Big(\frac{(4a)^{j+k}}{j!\,k!}\Big)^{1/2} \int_{\mathbb{R}} \big(\sin^2 \tfrac{1}{2}s\big)^{(j+k)/2} \frac{\mathrm{d}s}{s^2}. \end{aligned} $$

By the Cauchy-Schwarz inequality,

$$ \begin{aligned} \int_{\mathbb{R}} \big(\sin^2 \tfrac{1}{2}s\big)^{(j+k)/2} \frac{\mathrm{d}s}{s^2} &\equiv \int_{\mathbb{R}} \big(\sin^2 \tfrac{1}{2}s\big)^{j/2} \big(\sin^2 \tfrac{1}{2}s\big)^{k/2} \frac{\mathrm{d}s}{s^2} \\ &\leq \left(\int_{\mathbb{R}} \big(\sin^2 \tfrac{1}{2}s\big)^{j} \frac{\mathrm{d}s}{s^2} \right)^{1/2} \left(\int_{\mathbb{R}} \big(\sin^2 \tfrac{1}{2}s\big)^{k} \frac{\mathrm{d}s}{s^2} \right)^{1/2}. \end{aligned} $$

Because $(2k-3)!!/(2k-2)!! \leq 1$ for all $k \in \mathbb{N}$ then it follows from (5.4.17) that

$$ \int_{\mathbb{R}} \big(\sin^2 \tfrac{1}{2}s\big)^{j} \frac{\mathrm{d}s}{s^2} \leq \pi; $$

therefore,

$$ \left| \int_{\mathbb{R}} \mathcal{P}_j(s)\mathcal{P}_k(-s)\frac{\mathrm{d}s}{s^2} \right| \leq \Big(\frac{(4a)^{j+k}}{j!\,k!}\Big)^{1/2} \pi, $$

and the same holds for the functions $\mathcal{Q}_j$. Substituting these bounds into the general

series expansion (5.3.5), we obtain the upper bound

$$\begin{aligned}\mathcal{V}^2(X,Y) &\leq \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\frac{(4\lambda a)^{j+k}}{j!\,k!}\\ &= \big(\exp(4\lambda a)-1\big)^2 < \infty,\end{aligned}$$

for all $\lambda \in [0,1]$ and $a > 0$. Therefore, the series (5.4.12) converges absolutely. □

To calculate the distance variance, the argument given in the bivariate gamma case remains valid here. By Koudou [58, p. 103], the characteristic function of (X,Y) is

$$f_{X,Y}(s,t) = \exp\big[a(1-\lambda)(\mathrm{e}^{\mathrm{i}s}-1) + a(1-\lambda)(\mathrm{e}^{\mathrm{i}t}-1) + a\lambda(\mathrm{e}^{\mathrm{i}(s+t)}-1)\big].$$

Therefore,

$$\lim_{\lambda\to 1-} f_{X,Y}(s,t) = \exp\big[a(\mathrm{e}^{\mathrm{i}(s+t)}-1\big] \equiv f_X(s+t),$$

so we obtain

$$\mathcal{V}^2(X,X) = \mathcal{V}^2(Y,Y) = \lim_{\lambda\to 1-}\mathcal{V}^2(X,Y).$$

In this chapter, we derived a formula for the population version of distance correlation for random vectors (X,Y) lying the class of Lancaster distributions. While this formula still requires solving the nontrivial integrals (5.3.3) and (5.3.3) to obtain explicit results for certain distributions, it enormously facilitates the calculation of these results by delivering a tractable expansion of the distance correlation. The impact of this result is twofold. On one hand, it enables the efficient evaluation of distance correlation for numerous discrete and continuous distributions as we have shown in section 5.4. On the other hand, the expansion of distance correlations in terms of integrals of orthogonal polynomials and Lancaster coefficients in Theorem 5.3.1 may lead to better understanding and physical interpretation of distance correlation.

Chapter 6

Detecting Collinear Groups of Random Variables in Low-rank Models

After having considered the distance correlation and the affinely invariant distance correlation in the preceding chapters, we now take a look at a specific variable clustering problem in low rank models. In particular, our goal will be to detect groups of collinear random variables, i.e. random variables which feature an exact or approximate linear dependence. As our main result, we will show that we can indeed obtain an asymptotic guarantee to retrieve these groups in a particular Gaussian setting (the PPCA model).

We first give a motivation for our clustering task based on a specific interpretation problem in Gaussian graphical models. Subsequently, we discuss the set-up of this task and formulate an explicit problem statement. We remark, that for fixed sample size, the problem under consideration is mathematically equivalent to the problem of subspace clustering for data in the case of independent subspaces. On these grounds, we can show that the clustering can be exactly recovered in the noiseless case. When the sample size goes to infinity, the equivalence to subspace clustering is not preserved. However, for the case of known intrinsic dimension, we show that the clustering can be asymptotically retrieved under moderate assumptions. For the probabilistic PCA model, consistent estimators for the intrinsic dimension are available [89] and we can hence transfer this result to the setting of unknown intrinsic dimension. We conclude this chapter with a critical discussion of our results.

6.1 Motivation

In numerous applications, Gaussian graphical models (GGMs) are utilized to detect meaningful associations between different quantities. In particular, edges in a GGM are interpreted as "direct" connections, since the dependence betweentwo variables

connected by an edge cannot be fully explained by the other variables in the model. However, it is often neglected that this interpretation crucially depends on two assumptions:

1.) All "relevant" quantities are included in the model.

2.) There is no redundant information in the model.

If *1.)* is not satisfied, we could mistake an edge for a direct connection, when there is in fact a relevant variable not included in the model explaining the dependency between the variables connected by this edge. If *2.)* is not satisfied, direct connections could be missed out because of redundant variables. Let for example two variables X_1 and X_2 of a GGM be a slightly different representation of the same quantity. Then X_1 explains almost all the variance of X_2 (and vice-versa), hence the partial correlations between these two variables and the rest of the model are virtually zero, which induces the absence of edges even if strong direct relations are present. Particularly in large-dimensional applications, the observed dimension rarely matches the intrinsic dimension of the problem and the interpretation of the GGM is questionable. This chapter is a first step in developing methods to detect redundant variables in graphical models and to use this knowledge to define a new graphical model which fits the intrinsic dimension of the problem and hence allows for the interpretation described above.

To point out the immense problems caused by redundant information, consider the following example. Let $X = (X_1, X_2, \ldots, X_6)$ be a random vector, such that these six random variables correspond to the intrinsic dimension of the problem and the graphical model (Figure 6.1) is regular and reveals direct associations. We now add two more variables X_7 and X_8 via

$$X_7 = aX_1 + bX_2 + \epsilon, \quad X_8 = cX_3 + dX_4 + \delta,$$

where $a, b, c, d \in \mathbb{R}$. If $\epsilon = \delta = 0$, the structure of the graphical model is completely destroyed (Figure 6.2), similar for small noise ϵ, δ. This problem occurs since the three variables $\{X_7, X_1, X_2\}$ (or $\{X_8, X_3, X_4\}$ respectively) intrinsically live on a two-dimensional subspace. Hence already two of these variables explain the third one and all edges between these subsets and the rest of the model are eliminated. We suggest an approach to tackle this problem consisting of two steps. First, we aim at detecting groups of collinear random variables (in our example $\{X_7, X_1, X_2\}$ and $\{X_8, X_3, X_4\}$). Second, we propose to apply dimension reduction techniques to define a new graphical model, which better suits the intrinsic dimension of the problem. This chapter is dedicated to the first step of this approach.

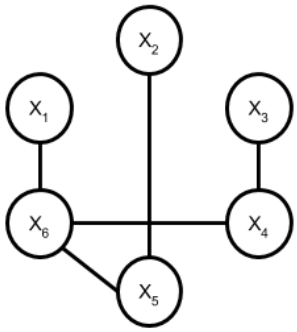

Figure 6.1: A regular Gaussian graphical model with six nodes.

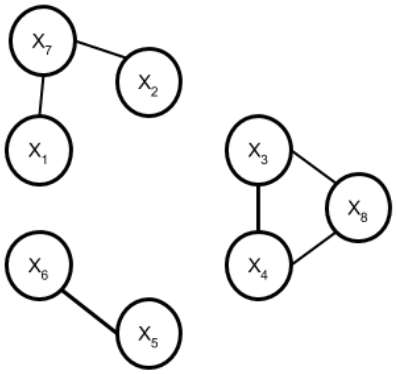

Figure 6.2: The structure of the GGM is heavily affected from collinearity.

6.2 Notation

For two partitions $\alpha = \{C_i\}_{i=1}^{k_1}$ and $\beta = \{B_i\}_{i=1}^{k_2}$, we will say that β is coarser than α if any set in α is a subset of an element in β. Analously, we will say that β is finer than α if any set in β is a subset of an element in α.

We will further use the following notation for matrices, which is due to [63]. For a matrix $M \in \mathbb{R}^{m\times n}$, $|M]_{i,j}$ denotes its (i,j)-th entry, $|M]_{i,:}$ denotes its i-th row and $|M]_{:,j}$ its j-th column. The notation $M = [M_1; M_2; \ldots; M_k]$ refers to

$$M = \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_k \end{pmatrix}$$

and analogously $M = [M_1, M_2, \ldots, M_k]$ denotes:

$$M = \begin{pmatrix} M_1 & M_2 & \cdots M_k \end{pmatrix}.$$

Moreover, it will prove useful to fix a notation for matrices which are 0 everywhere except for certain subsets of the rows and columns. In particular, for $M \in \mathbb{R}^{m\times n}$, we denote by $|M]_{S,T}$ the $m \times n$-Matrix whose entries are $|M]_{i,j}$ for $(i,j) \in S \times T$ and 0 otherwise. Moreover $|M]_{S,:} := |M]_{S,\{1,\ldots,n\}}$ and $|M]_{:,T} := |M]_{\{1,\ldots,m\},T}$. By $|M]_{l,S}$ (resp. $|M]_{S,l}$), we refer to the l-th row (resp. column) of $|M]_{:,S}$ (resp. $|M]_{S,:}$). Finally, by using the term "block-diagonality", we will refer to the property that a matrix is block-diagonal (in the usual sense) up to permutation, more precisely note the following definition.

Definition 6.2.1. *We say that a matrix M is block-diagonal with exactly k blocks, whenever there exist two permutation matrices Q_1 and Q_2, such that $Q_1\,M\,Q_2$ has the form*

$$Q_1\,M\,Q_2 = \begin{pmatrix} L_1 & 0 & 0 & 0 \\ 0 & L_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & L_k \end{pmatrix}$$

and for any i, there are no permutation matrices R_1, R_2, s.t.

$$R_1\,L_i\,R_2 = \begin{pmatrix} K_1 & 0 \\ 0 & K_2 \end{pmatrix}.$$

Obivously $rank(M) = rank(Q_1\,M\,Q_2) = \sum_{i=1}^{k} rank(L_i)$.

6.3 Problem Statement

As already stated in the introduction to this chapter, our goal will be to detect groups of collinear random variables. For this purpose, let $Y = (Y_1, \ldots, Y_p)^t$ be a random vector distributed according to some distribution with mean $(0, \ldots, 0)^t$ and covariance matrix Γ. Moreover, suppose that there are m groups (or clusters) of collinear random variables. The clusters will be denoted by $C_1, \ldots, C_m$, where for $i \in \{1, \ldots, m\}$, C_i is a a subset of $\{1, \ldots, p\}$, such that l lies in C_i, if the random variable Y_l belongs to the i-th group of collinear random variables. Apparently it holds $C_1 \cup C_2 \cup \ldots C_m \subset \{1, \ldots, p\}$ and $C_i \cap C_j = \emptyset$ for $i \neq j$. As mentioned before, the random variables affiliated to a cluster C_i are assumed to be collinear, i.e. there are constants λ_l ($l \in C_i$) such that

$$\sum_{l \in C_i} \lambda_l Y_l = 0$$

and hence the covariance matrix of the random variables affiliated to C_i does not have full rank (i.e. $\text{rank}(|\Gamma]_{C_i,C_i}) < |C_i|$). Moreover the sets $C_1, \ldots, C_m$ should account for all the redundant information in the data, hence there is no collinearity among the variables expressed by the index set $D = \{1, \ldots, p\} \setminus \bigcup_{i=1}^{m} C_i$ (i.e. $\text{rank}(|\Gamma]_{D,D}) = |D|$). For technical reasons, we will treat the random variables affiliated to D as single clusters of size 1. Hence $D = C_{m+1} \cup C_{m+2} \cup \ldots \cup C_k$, with $|C_i| = 1$ for $i \in \{m+1, \ldots, k\}$ and $k - m = |D|$. This means, that for $i \in \{1, \ldots, k\}$, C_i either refers to a low-rank cluster (i.e. $\text{rank}(|\Gamma]_{C_i,C_i}) < |C_i|$) or to a single random variable being part of the set D. To simplify the notation, we will denote, for $i \in \{1, \ldots, k\}$, $p_i := |C_i|$ and $d_i = \text{rank}(|\Gamma]_{C_i,C_i})$.

Our goal will be in the following, given an *i.i.d.* sample $\boldsymbol{Z} = (Z^{(1)}, Z^{(2)}, \ldots, Z^{(n)})$ drawn from $Y+E$ (where E is some noise variable), find the number of clusters k, as well as the segmentation of the random variables, i.e. the index sets $C_1, \ldots, C_k$. It is apparent, that this problem is ill-posed in general and we need additional assumptions on the clusters to make this problem identifiable. To find assumptions, which suit the nature of our problem, we reconsider the motivation for our clustering task. Since our goal is to remove the collinearity, we should prefer a model, which accounts for all redundant information. Hence the missing dimensionality in the clusters should add up to the missing dimensionality of the model, which is expressed by $\sum_{i=1}^{k}(p_i - d_i) = (p - d)$ or $\sum_{i=1}^{k} d_i = d$, equivalently. The second goal is to identify the structure of the collinearity. Hence, among all clusterings satisfying $\sum_{i=1}^{k} d_i = d$, we will assume that $\{C_i\}_{i=1}^{k}$ is the *finest*, i.e. any other partition $\{B_i\}_{i=1}^{m}$ satisfying $\sum_{i=1}^{k} d_i = d$ is coarser than $\{C_i\}_{i=1}^{k}$. The existence of a partition satisfying $\sum_{i=1}^{k} d_i = d$ is clear since for the trivial partition $C_1 = \{1, 2, \ldots, p\}$, $|\Gamma]_{C_1,C_1} = \Gamma$ and hence $d_1 = d$. We will now show that there is a unique finest clustering with that property. Interestingly, the preceding two conditions will be sufficient to define the clustering, the resulting clusters will automatically either satisfy $\operatorname{rank}(|\Gamma]_{C_i,C_i}) < |C_i|$ or $|C_i| = 1$.

Lemma 6.3.1. *Let $Y = (Y_1, \ldots, Y_p)$ be a random vector with mean $(0, \ldots, 0)^t$, such that the corresponding covariance matrix Γ satisfies $\operatorname{rank}(\Gamma) = d$. Then there is a unique finest partition $\{C_i\}_{i=1}^{k}$, such that $\sum_{i=1}^{k} d_i = d$, where $d_i = \operatorname{rank}(|\Gamma]_{C_i,C_i})$.*

Proof. As already pointed out above, the existence of a clustering satisfying $\sum_{i=1}^{k} d_i = d$ is trivial. If a finest clustering with $\sum_{i=1}^{k} d_i = d$ exists, it is clear that this partition is unique, since any other feasible partition is coarser. It remains to show that there exists a finest clustering satisfying the desired property. Consider that the statement is false; then there are two different partitions $\{C_i\}_{i=1}^{k_1}$ and $\{B_i\}_{i=1}^{k_2}$, such that there exist no feasible partitions, which are finer than $\{C_i\}_{i=1}^{k_1}$ or $\{B_i\}_{i=1}^{k_2}$. Hence there are two sets C_l and B_m, such that

$$\{0\} \subsetneq (C_l \cap B_m) \subsetneq C_l.$$

Let us define the subsets D_1 and D_2 by

$$D_1 = (C_l \cap B_m), \quad D_2 = C_l \backslash B_m.$$

It is now straightforward to show that

$$\operatorname{rank}(|\Gamma]_{D_1,D_1}) + \operatorname{rank}(|\Gamma]_{D_2,D_2}) = \operatorname{rank}(|\Gamma]_{C_l,C_l})$$

and hence the partition $\{(C_i)_{i\neq l}, D_1, D_2\}$ is a partition into $k_1 + 1$ subsets such that

the ranks of the respective blocks in Γ add up to d. This is an contradiction to the assumption that there exists no finer feasible partition than $\{C_i\}_{i=1}^{k_1}$. □

We are now ready to formulate our problem statement.

Problem 6.3.2. *Let $Y = (Y_1, \ldots, Y_p)$ be a random vector with mean $(0, \ldots, 0)^t$, such that the corresponding covariance matrix Γ satisfies $\operatorname{rank}(\Gamma) = d$. Consider observations of the form*

$$\mathbf{Z} = \mathbf{Y} + \mathbf{E},$$

where $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}) \in \mathbb{R}^{p \times n}$ is an i.i.d. sample drawn from Y and $\boldsymbol{E}$ is a noise matrix (which will be specified in the respective subsections). Our goal is to recover the finest partition $\{C_i\}_{i=1}^{k}$, such that $\sum_{i=1}^{k} d_i = d$, where $d_i = \operatorname{rank}(|\Gamma]_{C_i,C_i})$.

Let us denote that - to the best knowledge of the author - there is only one existing variable clustering method, which is suited to this problem. While other methods intrinsically assume that the rank of all clusters in one, the matroid approach [114] allows for clusters of arbitrary (low) rank j. Indeed, the clusters in our description containing just one variable are exactly those rank-1-flats, which are not part of a rank-j-flat with $j > 1$. A cluster C_i of low rank $d_i < |C_i|$ is obviously a d_i-flat. Moreover, one can show that the condition $\sum_{i=1}^{k} d_i = d$ ensures, that this cluster is not part of a rank-j-flat with $j > d_i$. In conclusion, our clustering task may be attacked by the matroid approach, assigning each covariate to the flat with maximum rank. However, the matroid approach is - due to its combinatorial nature - computationally highly expensive and there are no known theoretical guarantees for this method. We will now show a link of the problem under consideration to subspace clustering, which enables a much more efficient solution in the case of clean data (i.e. $\boldsymbol{E} = 0$.)

6.4 Inference

6.4.1 Inference for Clean Data

Consider for now that $\boldsymbol{E} = 0$ and further assume the mild technical condition

$$\operatorname{rank}(\boldsymbol{Y}) = \operatorname{rank}(\Gamma), \tag{6.4.1}$$

which holds true almost surely for distributions with a certain degree of regularity, as long as $n > d$. Under this assumption, we immediately obtain $\operatorname{rank}(|\boldsymbol{Y}]_{C_i:}) = \operatorname{rank}(|\Gamma]_{C_i,C_i}) = d_i$ for clusters C_i with $i \in \{1, \ldots, k\}$. When we denote the linear subspace spanned by the rows of $\operatorname{rank}(|\boldsymbol{Y}]_{C_i:})$ by S_i, we realize that our problem is analogue to the problem of subspace clustering as described in subsection 1.1.3, only that the roles of the rows and the columns, i.e. of the component of the random vector and

the samples are interchanged. In addition, we can show that our clustering constraints guarantee the independence of the respective subspaces. For that purpose, we recapitulate the definition of the independence of subspaces. To isolate this concept from the concept of stochastic independence, we will refer to it as subspace-independence or short $\mathcal{S}$-independence.

Definition 6.4.1. *For subspaces $\{U_i\}_{i=1}^k$ of a vector space V, we say that $\{U_i\}_{i=1}^k$ are $\mathcal{S}$-independent (subspace-independent), if for any $i \in \{1, 2, \dots, k\}$*

$$U_i \cap \sum_{j \neq i} U_j = \{0\}.$$

Lemma 6.4.2. *Consider Problem 6.3.2 and further assume the condition (6.4.1). Then the subspaces $\{S_i\}_{i=1}^k$ spanned by the rows of $|Y]_{C_i,:}$ are $\mathcal{S}$-independent.*

Proof. Assume, that $\{S_i\}_{i=1}^k$ are not $\mathcal{S}$-independent. Then there is some $j \in \{1, 2, \dots, k\}$, such that $dim(S_j \cap \sum_{i \neq j} S_i) > 0$. Hence,

$$\begin{aligned} d = dim\left(\sum_{i=1}^{k} S_i\right) &= dim\left(\sum_{i \neq j} S_i\right) + dim(S_j) - dim\left(S_j \cap \sum_{i \neq j} S_i\right) \\ &< dim\left(\sum_{i \neq j} S_i\right) + dim(S_j) \leq \sum_{i \neq j} dim(S_i) + dim(S_j) = \sum_{i=1}^{k} dim(S_i). \end{aligned}$$

This is a contradiction to the assumption that $d = \sum_{i=1}^k d_i = \sum_{i=1}^k dim(S_i)$.

□

The preceding lemma implies that, for fixed sample size n and clean data $\boldsymbol{Y}$, our method is equivalent to the subspace clustering problem described in 1.1.3 assuming the independence of the respective subspaces $\{S_i\}_{i=1}^k$. In particular, this ensures, that the method by Costeira and Kanade [12] works in our setting, i.e. the clustering can be recovered by the orthogonal projection on the column space of $\boldsymbol{Y}$. For the sake of completeness, we work out the details of this method. The following lemma is due to [63, 12].

Lemma 6.4.3. *Consider the setting of Problem 6.3.2 and further assume the condition 6.4.1. Let $U\Lambda V^t$ denote the skinny SVD of $\boldsymbol{Y}$ and $P = UU^t$ the orthogonal projection on the column space of $\boldsymbol{Y}$. Then $|P]_{l,m} = 0$ if l and m do not belong to the same cluster (i.e. $|P]_{C_i,C_j} = 0$ for $i \neq j$).*

Proof. Let us define an auxiliary matrix W by

$$|W]_{l,m} = \begin{cases} |P]_{l,m} & \text{if } l \text{ and } m \text{ belong to the same cluster,} \\ 0 & \text{else.} \end{cases}$$

Further define $R = P - W$. It is apparent that it suffices to show that $R = 0$.
W.l.o.g., we now assume that $l \in C_i$ where $l \in \{1, \dots, p\}$ and $i \in \{1, \dots, k\}$ arbitrary. It holds:

$$|P\,\boldsymbol{Y}]_{l,:} = |\boldsymbol{Y}|_{l,:} \in S_i.$$

Moreover, with $C_{-i} := \bigcup_{j \neq i} C_j$ and noting that $|W]_{C_i,C_j} = 0$ for $i \neq j$:

$$|W\,\boldsymbol{Y}]_{l,:} = |W]_{l,C_i}\,|\boldsymbol{Y}]_{C_i,:} + |W]_{l,C_{-i}}\,|\boldsymbol{Y}]_{C_{-i},:} = |W]_{l,C_i}\,|\boldsymbol{Y}]_{C_i,:} \in S_i$$

and since $|R]_{C_i,C_j} = 0$ for $i = j$

$$|R\,\boldsymbol{Y}]_{l,:} = |R]_{l,C_i}\,|\boldsymbol{Y}]_{C_i,:} + |R]_{l,C_{-i}}\,|\boldsymbol{Y}]_{C_{-i},:} = |R]_{l,C_{-i}}\,|\boldsymbol{Y}]_{C_{-i},:} \in \sum_{j \neq i} S_j.$$

Finally it also holds that $|R\,\boldsymbol{Y}]_{l,:} = |P\,\boldsymbol{Y}]_{l,:} - |W\,\boldsymbol{Y}]_{l,:} \in S_i$. Lemma 6.4.2 yields that $S_i \cap \sum_{j \neq i} S_j = \{0\}$, hence $R\,\boldsymbol{Y} = 0$. Since the columns of P can be written as linear combinations of $\boldsymbol{Y}$, this implies that $RP = 0$ as well. Moreover $R\,R = R\,(P - W) = -R\,W$. Now, for any $i \in \{1, \dots, k\}$:

$$-|R\,R]_{C_i,C_i} = |R\,W]_{C_i,C_i} = |R]_{C_i,C_i}\,|W]_{C_i,C_i} + |R]_{C_i,C_{-i}}\,|W]_{C_{-i},C_i} = 0.$$

Hence $0 = |R\,R]_{C_i,C_i} = |R]_{C_i,;}\,(|R]_{C_i,;})^t$. It follows, that $|R]_{C_i,;} = 0$ for all $i \in \{1, \dots, k\}$ which implies $R = 0$. □

Obviously, the theorem above states that our clusters $\{C_i\}_{i=1}^k$ are separated by the zero-entries of the projection matrix P. However, it makes no statement about the connectivity of the intra-cluster entries of that matrix. The following lemma addresses this question.

Lemma 6.4.4. *Consider the setting of Problem 6.3.2. For $i \in \{1, \dots, k\}$, consider an arbitrary partition of C_i into non-empty sets D_1 and D_2 (i.e. $D_1 \cup D_2 = C_i$, $D_1 \cap D_2 = \emptyset$). Then $|P]_{D_1,D_2} \neq 0$.*

Proof. Assume that $|P]_{D_1,D_2} = 0$. It is easy to check, that

$$rank(|\boldsymbol{Y}]_{C_i,:}) = rank(|P]_{C_i,C_i}) = rank(|P]_{D_1,D_1}) + rank(|P]_{D_2,D_2}).$$

Moreover, since $|P]_{D_1,:} = |P]_{D_1,D_1}$,

$$|\boldsymbol{Y}]_{D_1,:} = |P]_{D_1,D_1}|\boldsymbol{Y}]_{D_1,:}.$$

and $|\boldsymbol{Y}]_{D_2,:} = |P]_{D_2,D_2}|\boldsymbol{Y}]_{D_2,:}$. It follows that

$$rank(|\boldsymbol{Y}]_{C_i,:}) = rank(|\boldsymbol{Y}]_{D_1,:}) + rank(|\boldsymbol{Y}]_{D_2,:}),$$

Denoting the spaces spanned by the rows $|\boldsymbol{Y}]_{D_1,:}$ and $|\boldsymbol{Y}]_{D_2,:}$ by T_1 and T_2 respectively, this implies $dim(T_1+T_2) = dim(S_i) = dim(T_1)+dim(T_2)$. Hence, $dim(T_1)+dim(T_2)+\sum_{j\neq i} dim(S_j) = d$. This is a contradiction to the assumptions, that $\{C_i\}_{i=1}^k$ is the finest partition with that property. □

Combining Lemma 6.4.3 and Lemma 6.4.4, we have that:

(i) The orthogonal projection P on the columns of $\boldsymbol{Y}$ is block-diagonal with exactly k blocks.

(ii) $|P]_{C_i,C_j} = 0$ for $i \neq j$.

Let us now consider the adjacency matrix A, whose (l,m)-th element $a_{l,m}$ is given by

$$a_{l,m} = \mathbb{1}_{\{p_{l,m}\neq 0\}},$$

where $p_{l,m}$ denotes the (l,m)-th element of P.

Obviously, (i) implies that the graph induced by A has exactly k connected components and (ii) states that the clusters are not connected in the graph. Hence the connected components of A coincide with the clusters $\{C_i\}_{i=1}^k$.

We are now ready to state the central theorem of this section.

Theorem 6.4.5. *Consider the setting of Problem 6.3.2 and further assume the condition 6.4.1. Let $U\Lambda V^t$ denote the skinny SVD of $\boldsymbol{Y}$ and $P = UU^t$ the orthogonal projection on the column space of $\boldsymbol{Y}$. The clusters $\{C_i\}_{i=1}^k$ can then be exactly recovered by finding the adjacency matrix A, whose (l,m)-th element $a_{l,m}$ is given by*

$$a_{l,m} = \mathbb{1}_{\{p_{l,m}\neq 0\}},$$

where $p_{l,m}$ denotes the (l,m)-th element of P.

For the case of clean data ($\boldsymbol{E} = 0$) considered in the current subsection, it is clear that the orthogonal projection P and hence A, as well, can be exactly recovered. For noisy data, the issue is much more involved, since P can naturally only be estimated. Furthermore, even when P can be consistently estimated, we have no guarantee to consistently retrieve the cluster structure, since the elements of A are obviously no continuous functions in the elements of P.

6.4.2 Inference in the Case of Homogeneous Noise

In the following we will consider the case of noisy data. In particular, we are going to assume that the error term $\boldsymbol{E}$ represents homogeneous noise, i.e. its columns are

independently drawn from a random vector $E = (E_1, E_2, \ldots, E_p)^t$ with mean 0 and covariance matrix $\mathbb{E}[E\,E^t] = \sigma^2\,I_p$, with some $\sigma^2 > 0$. Moreover E and Y are assumed to be independent. For convenience, we will additionally assume (6.4.1) for fixed n, yet note that this condition is automatically fulfilled when $n \to \infty$.

Our aim will be to retrieve the clustering via Theorem 6.4.5, i.e. by constructing a consistent estimator for the matrix A. First we will derive an estimator for the projection matrix P without assuming any further condition. We will then show the asymptotic normality of this estimator given the asymptotic normality of the respective sample covariance matrix. Under this assumption, we will succeed to find a consistent estimator for A, as long as the intrinsic dimension d of Y is known. Finally, in the case of a Gaussian probabilistic PCA model, we can apply known estimators for the intrinsic dimension. By combining these with the derived estimator for A, we derive an asymptotic guarantee to recover our cluster structure in the setting of unknown intrinsic dimension d.

The following notation will prove useful (see [105]):

Definition 6.4.6. *Let M be a symmetric $q \times q$ matrix. Then for some eigenvalue λ of M, we call the unique orthogonal projection Q_λ on the λ-eigenspace $\mathcal{E}_\lambda$ the eigenprojection for M associated with λ. For a set of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$, we call the orthogonal projection $Q_{\lambda_1,\lambda_2,\ldots,\lambda_d}$ on the sum of the eigenspaces $\sum_{i=1}^d \mathcal{E}_{\lambda_i}$ the total eigenprojection for M associated with $\lambda_1, \lambda_2, \ldots, \lambda_d$. If $\lambda_1, \lambda_2, \ldots, \lambda_d$ are pairwise different, we have*

$$Q_{\lambda_1,\lambda_2,\ldots,\lambda_d} = \sum_{i=1}^{d} Q_{\lambda_i}.$$

Under assumption (6.4.1), the orthogonal projection $P = UU^t$ on the column space of $\boldsymbol{Y}$ obviously coincides with the orthogonal projection on the column space of Γ. Note, that Γ can be decomposed as follows:

$$\Gamma = \lambda_1 v_1 v_1^t + \ldots \lambda_d v_d v_d^t,$$

where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_d$ are the d non-zero eigenvalues of Γ (if we count with multiplicity) and $v_1, \ldots, v_d$ are respective orthonormal eigenvectors. Then the covariance matrix $\Sigma = \Gamma + \sigma^2\,I_p$ of $Z = X + E$ can be expressed as follows:

$$\Sigma = (\lambda_1 + \sigma^2) v_1 v_1^t + \ldots (\lambda_d + \sigma^2) v_d v_d^t + \sigma^2 v_{d+1} v_{d+1}^t + \ldots \sigma^2 v_p v_p^t,$$

where $\{v_{d+1}, \ldots, v_p\}$ is an orthonormal basis of $\ker(\Gamma)$.

The total eigenprojection P for Γ associated with its non-zero eigenvalues is given by

$$P = v_1 v_1^t + \dots v_d v_d^t,$$

which is apparently the same as the total eigenprojection for Σ associated with its d largest eigenvalues. In the following, we will show that the total eigenprojection for the sample covariance matrix $\hat{\Sigma}_n$ associated with its d largest eigenvalues $a.s.$-converges to the total eigenprojection for Σ associated with its d largest eigenvalues . So, if the dimension d of Γ is known, we obtain a consistent estimator for P. For $i \in \{1, \dots, p\}$, let us denote by $\hat{\mu}_n^i$ the i-th largest eigenvalue of $\hat{\Sigma}_n$ and by $\hat{v}_n^1, \dots, \hat{v}_n^p$ corresponding orthonormal eigenvectors.
For the eigenvalues and eigenprojections of symmetric matrices, it holds an important continuity property [105, Lemma 2.1]:

Lemma 6.4.7. *Let M_n be a $q \times q$ symmetric matrix with eigenvalues $\lambda_1(M_n) \geq \lambda_2(M_n) \geq \dots \geq \lambda_q(M_n)$. Let $P_{j,t}(M_n)$ represent the total eigenprojection for M_n associated with $\lambda_j(M_n), \dots, \lambda_t(M_n)$ for $t \geq j$. If $M_n \to M$ as $n \to \infty$, then*

(i) $\lambda_j(M_n) \to \lambda_j(M)$, and

(ii) $P_{j,t}(M_n) \to P_{j,t}(M)$ provided $\lambda_{j-1}(M) \neq \lambda_j(M)$ and $\lambda_t(M) \neq \lambda_{t+1}(M)$.

It is straightforward to transfer this result to $a.s.$-convergence:

Lemma 6.4.8. *Let M_n be a random $q \times q$ symmetric matrix with eigenvalues $\lambda_1(M_n) \geq \lambda_2(M_n) \geq \dots \geq \lambda_q(M_n)$. Let $P_{j,t}(M_n)$ represent the total eigenprojection for M_n associated with $\lambda_j(M_n), \dots, \lambda_t(M_n)$ for $t \geq j$. If $M_n \overset{a.s.}{\to} M$ as $n \to \infty$, then*

(i) $\lambda_j(M_n) \overset{a.s.}{\to} \lambda_j(M)$, and

(ii) $P_{j,t}(M_n) \overset{a.s.}{\to} P_{j,t}(M)$ provided $\lambda_{j-1}(M) \neq \lambda_j(M)$ and $\lambda_t(M) \neq \lambda_{t+1}(M)$.

This immediately implies the following theorem.

Theorem 6.4.9. *Consider Problem* 6.3.2*, where $\boldsymbol{E}$ represents homogeneous noise, i.e. its columns are independently drawn from a random vector $E = (E_1, E_2, \dots, E_p)^t$ with mean 0 and covariance matrix $\mathbb{E}[E\,E^t] = \sigma^2\, I_p$, with some $\sigma^2 > 0$ where E and Y are independent. Further assume, that the intrinsic dimension d is known. Define $\hat{P}_n$ be defined as the total eigenprojection of the sample covariance matrix $\hat{\Sigma}_n$ associated with its d-largest eigenvalues $\hat{\mu}_n^1, \dots, \hat{\mu}_n^d$, i.e.*

$$\hat{P}_n = \sum_{j=1}^{d} \hat{v}_n^j (\hat{v}_n^j)^t, \tag{6.4.2}$$

where $\hat{v}_n^1, \ldots, \hat{v}_n^d$ denote the corresponding orthonormal eigenvectors. Then, for $n \to \infty$

$$\hat{P}_n \xrightarrow{a.s.} P.$$

We can further show that the asymptotic normality of the sample covariance matrix $\hat{\Sigma}_n$ leads to the asymptotic normality of the estimator $\hat{P}_n$. For this purpose, we assume that the sample covariance matrix $\hat{\Sigma}_n$ is asymptotically normal in the following sense:

$$vec(\sqrt{n}\,(\Sigma - \hat{\Sigma}_n)) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\, K) \quad \text{as} \quad n \to \infty. \tag{6.4.3}$$

where K is a $p^2 \times p^2$-matrix.

As an example, assume that $\boldsymbol{Z}$ is generated by a (Gaussian) probabilistic PCA model [103], which induces that $E \sim \mathcal{N}(0, \sigma^2 I_p)$ and $Y = WX$ where $X \sim \mathcal{N}(0, I_d)$ and $W \in \mathbb{R}^{p \times d}$ is an arbitrary rank-d matrix. Then $\boldsymbol{Z}$ is drawn from

$$Z = WX + E,$$

and $Z \sim \mathcal{N}(0, WW^t + \sigma^2 I_p)$. Since Z is normally distributed, it holds (see e.g. [105]):

$$vec(\sqrt{n}\,(\Sigma - \hat{\Sigma}_n)) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\, (I_{p^2} + I_{(p,p)})\,(\Sigma \otimes \Sigma)) \quad \text{as} \quad n \to \infty, \tag{6.4.4}$$

where $\Sigma = WW^t + \sigma^2 I_p$ and $\hat{\Sigma}_n$ denotes the respective sample covariance matrix.

We will need the following notation. The set of positive eigenvalues of Γ will be denoted by

$$L := \{\lambda_1, \ldots, \lambda_d\}.$$

For an eigenvalue λ of Γ, we denote the eigenprojection for Γ associated with this eigenvalue by P_λ. The spectral norm of a matrix B will be denoted by $\|B\|$. $A \otimes B$ will denote the Kronecker product of A and B, $vec(\cdot)$ will denote the *vec*-operator and $I_{(p,p)}$ the permuted identity matrix (also referred to as commutation matrix).
We will make use of the following rules (see [65, 69] for reference and further properties):

(i) Let A, B, C be $p \times p$-matrices. Then,

$$vec(ABC) = (C^t \otimes A)\, vec(B). \tag{6.4.5}$$

(ii) Let A, B, C, D be $p \times p$-matrices. Then,

$$(A \otimes B)\,(C \otimes D) = (AC \otimes BD). \tag{6.4.6}$$

(iii) Let A and B be $p \times p$-matrices. Then,

$$I_{(p,p)}\,(A \otimes B) = (B \otimes A)\,I_{(p,p)}. \tag{6.4.7}$$

By applying Lemma 4.1 in [105], we get

Lemma 6.4.10. *If* $\|\hat{\Sigma}_n - \Sigma\| \leq \lambda_d/2$, *then*

$$\hat{P}_n = P - \sum_{\lambda \in L}[P_\lambda\,(\hat{\Sigma}_n - \Sigma)\,(\Sigma - (\lambda + \sigma^2)I_p)^+ + (\Sigma - (\lambda + \sigma^2)I_p)^+\,(\hat{\Sigma}_n - \Sigma)\,P_\lambda] + E_n,$$

where $\|E_n\| \leq \frac{\lambda_1}{\lambda_d}\left(\frac{2\|\hat{\Sigma}_n - \Sigma\|}{\lambda_d}\right)^2\left(1 - \frac{2\|\hat{\Sigma}_n - \Sigma\|}{\lambda_d}\right)^{-1}$.

Now we are ready to prove the asymptotic normality of $\hat{P}_n$

Theorem 6.4.11. *Consider Problem* 6.3.2, *where* $\boldsymbol{E}$ *represents homogeneous noise, i.e. its columns are independently drawn from a random vector* $E = (E_1, E_2, \ldots, E_p)^t$ *with mean* 0 *and covariance matrix* $\mathbb{E}[E\,E^t] = \sigma^2\,I_p$, *with some* $\sigma^2 > 0$ *where* E *and* Y *are independent. Further assume (6.4.3) and that the intrinsic dimension* d *is known. Let* $\hat{P}_n$ *be the estimator defined in Theorem 6.4.9. Then:*

$$\sqrt{n}\,vec(\hat{P}_n - P) \xrightarrow{\mathcal{D}} \mathcal{N}\,(0,\,C\,K\,C)\,,$$

where

$$C := \sum_{\lambda \in L} \lambda^{-1}\,(P_0 \otimes P_\lambda + P_\lambda \otimes P_0).$$

Proof. From assumption (6.4.3), we know that

$$\sqrt{n}\,vec(\hat{\Sigma}_n - \Sigma) \xrightarrow{\mathcal{D}} M,$$

where $vec(M) \sim \mathcal{N}(0,\,K)$. Clearly, by Lemma 6.4.10

$$\sqrt{n}(\hat{P}_n - P) \xrightarrow{\mathcal{D}} N = -\sum_{\lambda \in L}[P_\lambda M\,(\Sigma - (\lambda + \sigma^2)\,I_p)^+ + (\Sigma - (\lambda + \sigma^2)\,I_p)^+ M\,P_\lambda].$$

Exploiting property (6.4.5) yields:

$$vec(N) = -\Big(\sum_{\lambda \in L}(\Sigma - (\lambda + \sigma^2)\,I_p)^+ \otimes P_\lambda + P_\lambda \otimes (\Sigma - (\lambda + \sigma^2)\,I_p)^+\Big)\,vec(M).$$

Noting that

$$\Sigma - (\lambda + \sigma^2)\,I_p = \sum_{\lambda' \in L}(\lambda' - \lambda)\,P_{\lambda'} - \lambda\,P_0$$

we obtain

$$\begin{aligned} vec(N) &= -\Big(\sum_{\lambda \in L} \big\{ \sum_{\lambda'' \in L} (\lambda' - \lambda)^{-1} P_{\lambda'} \otimes P_\lambda - \lambda^{-1} P_0 \otimes P_\lambda \big\} \\ &\qquad + \big\{ \sum_{\lambda'' \in L} (\lambda'' - \lambda)^{-1} P_\lambda \otimes P_{\lambda''} - \lambda^{-1} P_\lambda \otimes P_0 \big\} \Big) \, vec(M) \\ &= -\Big(\sum_{\lambda \in L} \lambda^{-1} \left(P_0 \otimes P_\lambda + P_\lambda \otimes P_0 \right) \Big) \, vec(M). \end{aligned}$$

Hence $vec(N)$ is multivariate normal with covariance matrix

$$Cov(vec(N)) = C \, K \, C.$$

□

Corollary 6.4.12. *Let us consider, we are in the setting of Theorem 6.4.11 and further assume that* $\boldsymbol{Z}$ *is generated by a probabilistic PCA model. Hence* $E \sim \mathcal{N}(0, \sigma^2 I_p)$ *and* $Y = WX$ *where* $X \sim \mathcal{N}(0, I_d)$ *and* $W \in \mathbb{R}^{p \times d}$ *is an arbitrary rank-d matrix. Then*

$$\sqrt{n} \, vec(\hat{P}_n - P) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{\lambda \in L} \lambda^{-2} \, \sigma^2 \, (\lambda + \sigma^2) \, (P_0 \otimes P_\lambda + P_\lambda \otimes P_0) \right).$$

Proof. By (6.4.4) and Theorem 6.4.11

$$\sqrt{n} \, vec(\hat{P}_n - P) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \, C \, (I_{p^2} + I_{(p,p)}) \, (\Sigma \otimes \Sigma) \, C \right),$$

where

$$C := \sum_{\lambda \in S} \lambda^{-1} \, (P_0 \otimes P_\lambda + P_\lambda \otimes P_0).$$

Exploiting property (6.4.7) yields

$$C \, I_{(p,p)} = I_{(p,p)} \, C.$$

We get

$$C \, (I_{p^2} + I_{(p,p)}) \, (\Sigma \otimes \Sigma) \, C = (I_{p^2} + I_{(p,p)}) \, C (\Sigma \otimes \Sigma) \, C.$$

Now,by property (6.4.6):

$$
\begin{aligned}
&C\,(\Sigma \otimes \Sigma)\,C \\
&= \Big(\sum_{\lambda\in L} \lambda^{-1}\,(P_0 \otimes P_\lambda + P_\lambda \otimes P_0)\Big)\,\Big(\Sigma \otimes \Sigma\Big)\,\Big(\sum_{\lambda'\in L} (\lambda')^{-1}\,(P_0 \otimes P_{\lambda'} + P_{\lambda'} \otimes P_0)\Big) \\
&= \Big(\sum_{\lambda\in L} \lambda^{-1}\,\sigma^2\,(\lambda+\sigma^2)\,(P_0 \otimes P_\lambda + P_\lambda \otimes P_0)\Big)\,\Big(\sum_{\lambda'\in L} (\lambda')^{-1}(P_0 \otimes P_{\lambda'} + P_{\lambda'} \otimes P_0)\Big) \\
&= \sum_{\lambda\in L} \lambda^{-2}\,\sigma^2\,(\lambda+\sigma^2)\,(P_0 \otimes P_\lambda + P_\lambda \otimes P_0),
\end{aligned}
$$

which completes the proof. □

In the current chapter, we have constructed a consistent estimator for the orthogonal projection $P = U\,U^t$, when our data is corrupted by homogeneous noise. Moreover we were able to show the asymptotic normality of this estimator under moderate assumptions (6.4.3), in particular for Gaussian data generated by a probabilistic PCA model. To find the adjacency matrix A, the problem is now to identify the structures of the zeros in P via the consistent estimate $\hat{P}_n$. A possible approach could e.g. be to construct a multiple hypothesis test for a certain dependence structure using the asymptotic normality result 6.4.11.

We now pursue a different approach, namely identifying the adjacency matrix A via hard thresholding of $\hat{P}_n$. While it is not clear, how to choose the threshold for fixed n, we know from the asymptotic normal result that $|p_{ij} - \hat{p}_{ij}^{(n)}|$ is of magnitude $n^{-\frac{1}{2}}$, where p_{ij} and $\hat{p}_{ij}^{(n)}$ denote the (i,j)-th element of P and $\hat{P}_n$, respectively. Hence, for fixed d, we can asymptotically find the zero-entries of P by setting all entries of $\hat{P}_n$ with absolute value smaller than b_n to 0, where $b_n^{-1} = o(\sqrt{n})$ (e.g. $b_n = c_1 n^{-\frac{1}{4}}$ with arbitrary real constant c_1). More precisely, we obtain

Corollary 6.4.13. *Consider Problem* 6.3.2*, where* $\boldsymbol{E}$ *represents homogeneous noise, i.e. its columns are independently drawn from a random vector* $E = (E_1, E_2, \ldots, E_p)^t$ *with mean* 0 *and covariance matrix* $\mathbb{E}[E\,E^t] = \sigma^2\,I_p$*, with some* $\sigma^2 > 0$ *where* E *and* Y *are independent. Further assume (6.4.3) and that the intrinsic dimension* d *is known. Let* $\hat{P}_n$ *be the estimator defined in Theorem 6.4.9.*
Now, for the (i,j)*-th element of* $\hat{P}_n$*, denoted by* $\hat{p}_{ij}^{(n)}$*, define the matrix* $\hat{A}_n$ *by*

$$\hat{a}_{ij}^n = \mathbb{1}_{\{|\hat{p}_{ij}^{(n)}| > b_n\}}$$

where b_n is some positive real null sequence satisfying $b_n^{-1} = o(\sqrt{n})$. Then

$$\mathbb{P}(\hat{A}_n = A) \overset{n\to\infty}{\longrightarrow} 1,$$

where A is defined in Theorem 6.4.5. Hence, for $n \to \infty$, the clusters $\{C_i\}_{i=1}^k$ are asymptotically recovered.

PROOF. A is obviously a binary matrix. First consider an element a_{ij} $(i, j \in \{1, \dots, p\})$ of A, such that $a_{ij} = 1$. Then

$$\mathbb{P}(\hat{a}_{ij}^n = a_{ij}) = \mathbb{P}(\mathbb{1}_{\{|\hat{p}_{ij}^{(n)}|>b_n\}} = 1) = \mathbb{P}(|\hat{p}_{ij}^{(n)}| > b_n) \leq \mathbb{P}(|\hat{p}_{ij}^{(n)} - p_{ij}| < |p_{ij}| - b_n).$$

The latter expression converges to 1, since p_{ij} is greater than 0, b_n is a null sequence and $\hat{p}_{ij}$ is a consistent estimate for p_{ij}. Now consider $a_{ij} = 0$ and let c denote the ij-th element of the diagonal of the covariance matrix in Corollary 6.4.12. Then

$$\begin{aligned}\mathbb{P}(\hat{a}_{ij}^n = a_{ij}) &= \mathbb{P}(\mathbb{1}_{\{|\hat{p}_{ij}^{(n)}|>b_n\}} = 0) = \mathbb{P}(|\hat{p}_{ij}^{(n)}| \leq b_n) \\ &\leq \mathbb{P}(|\hat{p}_{ij}^{(n)} - p_{ij}| \leq b_n) = \mathbb{P}(\sqrt{c^{-1}n}\,|\hat{p}_{ij}^{(n)} - p_{ij}| \leq \sqrt{c^{-1}n}\,b_n).\end{aligned}$$

The latter expression converges to 1 since $\sqrt{c^{-1}n}\,b_n \to \infty$ and by Corollary 6.4.12

$$\mathbb{P}(\sqrt{c^{-1}n}\,|\hat{p}_{ij}^{(n)} - p_{ij}| \leq x) \overset{n\to\infty}{\longrightarrow} 1 - 2\,\Phi(-x),$$

where Φ denotes the cumulative distribution function of the Gaussian. □

In applications, the intrinsic dimension d is naturally rarely known. However, there exist numerous methods to choose an appropriate d in practice (see [7] for a survey). However, for most of these methods, no statistical properties are known and we do not know if these methods consistently estimate the intrinsic dimension. The probabilistic PCA model represents an exception and dimensionality estimation is well studied in this setting [5, 67, 89].

A consistent estimate $\hat{d}_n$ for the intrinsic dimension is derived in [89]:

$$\hat{d}_n = \arg\min_{k\in\{1,\dots,p\}} \log\left(\Big(\prod_{i=1}^{k} u_i\Big) \times \Big(\frac{1}{p-k}\sum_{i=k+1}^{p} u_i\Big)^{p-k}\right) + \frac{k}{n}\log n, \tag{6.4.8}$$

where $u_i := \frac{1}{n}\sum_{j=1}^{n}(z_{ij}^2)$ and z_{ij} represents the (i, j)-th element of the data matrix $\boldsymbol{Z}$.

Combining this consistent estimate with Theorem 6.4.13, we immediately obtain an asymptotic guarantee to recover the cluster structure for the probabilistic PCA model in the case of unknown dimension d:

Corollary 6.4.14. *Consider Problem 6.3.2 and further assume that* $\boldsymbol{Z}$ *is generated by a probabilistic PCA model. Hence* $E \sim \mathcal{N}(0, \sigma^2 I_p)$ *and* $Y = WX$ *where* $X \sim \mathcal{N}(0, I_d)$ *and* $W \in \mathbb{R}^{p \times d}$ *is an arbitrary rank-d matrix. Assume that the intrinsic dimension* d *is not known. Let* $\hat{d}_n$ *be the dimension estimator given in (6.4.8) and let* $\tilde{P}_n$ *be defined as the total eigenprojection of the sample covariance matrix* $\hat{\Sigma}_n$ *associated with its* $\hat{d}_n$*-largest eigenvalues* $\hat{\mu}_n^1, \dots, \hat{\mu}_n^{\hat{d}_n}$*, i.e.*

$$\tilde{P}_n = \sum_{j=1}^{\hat{d}_n} \hat{v}_n^j (\hat{v}_n^j)^t, \tag{6.4.9}$$

where $\hat{v}_n^1, \dots, \hat{v}_n^{\hat{d}_n}$ *denote the corresponding orthonormal eigenvectors. Now, for the* (i,j)*-th element of* $\tilde{P}_n$*, denoted by* $\tilde{p}_{ij}^{(n)}$*, define the matrix* $\tilde{A}_n$

$$\tilde{a}_{ij}^n = \mathbb{1}_{\{|\tilde{p}_{ij}^{(n)}| > b_n\}}$$

where b_n *is some positive real null sequence satisfying* $b_n^{-1} = o(\sqrt{n})$*. Then*

$$\mathbb{P}(\tilde{A}_n = A) \overset{n \to \infty}{\longrightarrow} 1,$$

where A *is defined in Theorem 6.4.5.*

6.5 Discussion and Outlook

In the preceding chapter, we have studied the problem of detecting groups of collinear variables in low-rank models. Even though we were able to construct consistent estimators to recover the cluster structure in certain settings, the obtained results are not completely satisfactory. Consequently, we close this chapter with a critical discussion of our investigations and an outlook to possible future work on this topic.

While the estimators in 6.4.13 and 6.4.14 are consistent, we did not achieve to show any further statistical properties. In particular, we do not know anything more about the quality of our estimation. In practice, the precision of the estimation may heavily depend on the choice of the threshold parameter b_n and the identification of the dimension d. Concerning b_n, methods need to be derived which select an appropriate parameter for fixed sample size n. Moreover, the asymptotic normality result suggests that choosing the same b_n for every element of the matrix is not optimal. It is further questionable if hard thresholding is the best way to identify the zeros in the projection matrix P and other approaches need to be studied in the future. On the other hand, statistical properties of estimators for the intrinsic dimension d are hardly known and

most practical researchers use ad-hoc rules to select d. Yet, the consistency of our technique relies on the exact identification of the intrinsic dimension. We suggest two possible options to improve this situation. The first point is obviously the development of better and more general estimators for the intrinsic dimension. Second, one could attempt to investigate how heavily our approach depends on the right choice of d. It would be important to know if we can still approximately recover the cluster structure, when our choice for d is only slightly different from the actual dimension. Finally, note that our approach identifies the dimension d and the zero-pattern of P successively. Techniques to identify both quantities at a time may be considered in the future, e.g. making use of estimation methods for simultaneously sparse and low rank matrices, which attracted considerable interest recently [78, 110, 17].

A promising starting point to tackle the problem for fixed sample size n is certainly the link to subspace clustering. Even more, since most of the subspace clustering methods do not consider any generative model, but only assume, that we have given some data $\boldsymbol{Y} \in \mathbb{R}^{p\times n}$, those methods may be directly applied by transposing the data matrix, hence swapping the roles of dimension p and sample size n. To put it another way, when no generative model for the data matrix $\boldsymbol{Y}$ is assumed, the terms "dimension" and "sample size" are somehow arbitrary definitions referring to the number of rows and columns, respectively of the data matrix. Although subspace clustering may naturally be a great help, one has to be careful when applying these methods. First of all, statistical methods which assume some kind of generative model, such as MPPCA may naturally not be applied. Moreover, one has to be attentive concerning the conditions under which a method works, often there are implicit conditions which are natural to assume for subspace clustering of data points, but which do not hold in our setting. For example, the sparse subspace clustering algorithm (SSC) intrinsically assumes that the number of samples n_i in each subspace S_i exceeds the dimension d_i without explicitly mentioning this fact. Finally, while results for the quality of the estimation for fixed sample size may be transferred, the link to subspace clustering can naturally not help to derive asymptotic properties.

To close this discussion, let us briefly discuss the second and still open point of our introductory motivation, the construction of a new graphical model, which does not suffer from the specified interpretability deficits. For this purpose, let us reconsider the example illustrated in Figures 6.1 and 6.2 of the introduction and choose $\epsilon = \delta = 0$. Remind that our goal was to recover Figure 6.1 from Figure 6.2, i.e. to recover the Graphical model consisting $X_1, \dots, X_6$ from the graphical model consisting of $X_1, \dots, X_8$. Applying the methods derived in this chapter reveals the clustering

$$C_1 = \{X_1, X_2, X_7\}, \quad C_2 = \{X_3, X_4, X_8\}, \quad C_3 = \{X_5\}, \quad C_4 = \{X_6\}$$

and we know both the intrinsic dimension of the whole model and of each single cluster. A possible way to proceed would certainly be to select a number of variables in each cluster that corresponds to its intrinsic dimension. Yet, note that this approach does not yield a unique solution. In particular, discarding any pair of random variables from C_1 and C_2 yields a valid model and there are thus $3 \cdot 3 = 9$ possible choices for a graphical model representing the right dimension and clustering. Unfortunately, in general, each of this model features different edges both within the clusters and in between the clusters. It is not obvious at all which additional assumption on our model are required to obtain a unique solution.

Chapter 7

Conclusion

In this thesis, we have dealt with complex dependence structures in two different ways. On the one hand, we studied the distance covariance and the distance correlation, two powerful dependence coefficients, which measure any kind of dependencies between random variables. On the other hand, we investigated the task of clustering collinear random variables in low rank systems.

With Chapters 3, 4 and 5, we hope to have contributed to the effective development of the theory of distance correlation. Beside confirming the result for the bivariate normal, which has already been shown by Székely, Rizzo and Bakirov [102], we succeeded in calculating the distance correlation for various other bivariate distributions, namely the Laplace and certain types of Poisson and Gamma distributions. For the setting of multivariate random variables, we introduced an affinely invariant version of the distance correlation as an alternative measure of dependence. In addition to the desirable properties of distance correlation, this coefficient is invariant under all invertible affine transformations. Yet, both the regular distance correlation and the affinely invariant distance correlation have its benefits (e.g. the regular distance covariance is scale-equivariant, which makes it possible to view it as a scalar product), and it may depend on the specific situation which measure to apply. An advantage of the affinely invariant distance correlation above the regular distance correlation is certainly, that its population version appears to be better interpretable. While we were able to get an explicit and readily computable expression for the affinely distance covariance of the multivariate normal distribution (which was employed both to obtain interesting limits results and for an application on wind vector data), the respective result for the regular distance covariance is much harder to handle. We were further able to give a useful series representation for the distance covariance of Lancaster distributions, which simplifies the computation of the population coefficients considerably. Finally, we derived a generalization of an integral which is at the core of the theory of distance correlation.

In Chapter 6, we attempted to make a first step towards resolving specific interpretation problems in low rank Gaussian graphical models. In particular, we defined a model considering multiple groups of collinear random variables and investigated the task of recovering these groups from both noiseless and noisy data. For fixed sample size, we find that the model is mathematically equivalent to the widely noticed model of subspace clustering of data in the case of independent subspaces. This opens up the possibility to apply methods from the vast literature of subspace clustering to help sorting out this problem; we suggest that further investigation of this link may be rewarding. Yet, since the role of sample size and dimension in the subspace clustering model and our model are swapped, the two models are not equivalent in the asymptotic setting. In the situation, where the sample size goes to infinity, we derive a consistent estimator, which asymptotically recovers the cluster structure for noisy data.

Our results on distance correlation offer several possibilities for further research. The application on wind vector data in section 3.4 is purely exploratory and for illustrative purposes. Yet, it introduces new concepts as the cross distance correlation function; a sound mathematical investigation of the convergence properties of this function could possibly lead to a better understanding of the distance correlation for dependent data. Moreover the analysis in this section may have the potential to be developed into parametric or nonparametric bootstrap tests for Gaussianity. Similarly, the extension of the integral given in Chapter 4 is purely theoretical. However, we raise the possibility, that this integral may be used to generalize the class of α-distance dependence measures to α outside the range $(0, 2)$. Finally and most importantly, we hope that further research based on the explicit formulas for the distance covariance in both finite-dimensional and asymptotic settings, together with the series representation for the class of Lancaster distributions will lead to a better physical interpretation of this coefficient.

The analysis of the clustering task in Chapter 6 is naturally by no means complete. The given estimator for the case of noisy data undoubtedly requires further investigation. In particular, the asymptotic normality of the preliminary estimator could possibly induce a test for a particular dependence structure. However, it has to be noted that this is a multiple testing problem and may be hard to tackle. Moreover, it is likely that the consistency of the given estimator holds true for a large class of distributions beyond the probabilistic PPCA setting. Finally, new approaches to solve this clustering problems may be considered in the future; the link to subspace clustering is certainly a promising starting point.

Appendix A

Appendix

A.1 The Standard Distance Correlation for the Multivariate Normal Population

In Theorem 3.2.4 and Corollary 3.2.6 we calculated the affinely invariant distance covariance for multivariate normal populations. Here, we consider the problem of deriving a formula for the standard distance covariance and distance correlation. We remark, that the following result is included in the preprint of the paper [18] by Dueck, Edelmann, Gneiting and Richards which is available on the *arXiv* (http://arxiv.org/abs/1210.2482).

We first consider the case in which Σ_X and Σ_Y are scalar matrices, say, $\Sigma_X = \sigma_x^2\, I_p$ and $\Sigma_Y = \sigma_y^2\, I_q$ with $\sigma_x, \sigma_y > 0$. Thus, suppose that $(X, Y) \sim \mathcal{N}_{p+q}(\mu, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix} = \begin{pmatrix} \sigma_x^2\, I_p & \Sigma_{XY} \\ \Sigma_{YX} & \sigma_y^2\, I_q \end{pmatrix}.$$

Putting $\Lambda = \Sigma_{YX}\Sigma_{XY}$, we follow the proofs of Theorem 3.2.4 and Corollary 3.2.6 to obtain

$$\begin{aligned} \mathcal{V}^2(X,Y) &= 4\pi \frac{c_{p-1}}{c_p}\frac{c_{q-1}}{c_q} \sum_{k=1}^{\infty} \frac{2^{2k}-2}{k!\,2^{2k}} \frac{(\frac{1}{2})_k\,(-\frac{1}{2})_k\,(-\frac{1}{2})_k}{(\frac{1}{2}p)_k(\frac{1}{2}q)_k} \frac{1}{(\sigma_x\sigma_y)^{2k-1}}\, C_{(k)}(\Lambda) \\ &= 4\pi\sigma_x\sigma_y \frac{c_{p-1}}{c_p}\frac{c_{q-1}}{c_q} \Big(\big[{}_3F_2\big(\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}; \tfrac{1}{2}p, \tfrac{1}{2}q; \Lambda/\sigma_x^2\sigma_y^2\big) - 1\big] \\ &\qquad -2\,\big[{}_3F_2\big(\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}; \tfrac{1}{2}p, \tfrac{1}{2}q; \Lambda/4\sigma_x^2\sigma_y^2\big) - 1\big] \Big). \end{aligned}$$

Next we reduce the general case to the scalar case above. By Theorem 3.1.1, we see

that we may assume, without loss of generality, that Σ_X and Σ_Y are diagonal matrices. Now denote by σ_x^2 and σ_y^2 the smallest eigenvalues of Σ_X and Σ_Y, respectively. Also, let $\Lambda_X = \Sigma_X - \sigma_x^2 I_p$ and $\Lambda_Y = \Sigma_Y - \sigma_y^2 I_q$; then, $\Sigma_X = \Lambda_X + \sigma_x^2 I_p$ and $\Sigma_Y = \Lambda_Y + \sigma_y^2 I_q$. Substituting these decompositions into the integral which defines $\mathcal{V}^2(X,Y)$, we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^{p+q}} \left(1 - \exp(s'\Sigma_{XY}t)\right)^2 \exp(-s'\Sigma_X s - t'\Sigma_Y t)\, \frac{\mathrm{d}s}{|s|_p^{p+1}}\, \frac{\mathrm{d}t}{|t|_q^{q+1}} \\
&= \int_{\mathbb{R}^{p+q}} \left(1 - \exp(s'\Sigma_{XY}t)\right)^2 \exp(-s'\Lambda_X s - t'\Lambda_Y t) \exp(-\sigma_x^2|s|_p^2 - \sigma_y^2|t|_q^2)\, \frac{\mathrm{d}s}{|s|_p^{p+1}}\, \frac{\mathrm{d}t}{|t|_q^{q+1}}.
\end{aligned}$$

Next, we apply a Taylor expansion,

$$\left(1 - \exp(s'\Sigma_{XY}t)\right)^2 = \sum_{k=2}^{\infty} \frac{2^k - 2}{k!} (s'\Sigma_{XY}t)^k$$

and, writing $\Lambda_X = \mathrm{diag}(\lambda_{x1}, \ldots, \lambda_{xp})$, we have

$$\begin{aligned}
\exp(-s'\Lambda_X s) &= \sum_{l=0}^{\infty} \frac{(-1)^l}{l!} (s'\Lambda_X s)^l \\
&= \sum_{l=0}^{\infty} \frac{(-1)^l}{l!} (\lambda_{x1} s_1^2 + \cdots + \lambda_{xp} s_p^2)^l \\
&= \sum_{l=0}^{\infty} \frac{(-1)^l}{l!} \sum_{l_1 + \cdots + l_p = l} \binom{l}{l_1, \ldots, l_p} \prod_{i=1}^{p} \lambda_{xi}^{l_i} s_i^{2l_i}.
\end{aligned}$$

Similarly, on writing $\Lambda_Y = \mathrm{diag}(\lambda_{y1}, \ldots, \lambda_{yq})$, we obtain

$$\exp(-t'\Lambda_Y t) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \sum_{m_1 + \cdots + m_q = m} \binom{m}{m_1, \ldots, m_q} \prod_{j=1}^{q} \lambda_{yj}^{m_j} t_j^{2m_j}.$$

Integrating these series term-by-term, we find that the typical integral to be evaluated is

$$\int_{\mathbb{R}^{p+q}} (s'\Sigma_{XY}t)^k \prod_{i=1}^{p} s_i^{2l_i} \prod_{j=1}^{q} t_j^{2m_j} \exp(-\sigma_x^2|s|_p^2 - \sigma_y^2|t|_q^2)\, \frac{\mathrm{d}s}{|s|_p^{p+1}}\, \frac{\mathrm{d}t}{|t|_q^{q+1}}.$$

By the substitution $t \mapsto -t$, we find that this integral vanishes if k is odd, and so we need to calculate

$$\int_{\mathbb{R}^{p+q}} (s'\Sigma_{XY}t)^{2k} \prod_{i=1}^{p} s_i^{2l_i} \prod_{j=1}^{q} t_j^{2m_j} \exp(-\sigma_x^2|s|_p^2 - \sigma_y^2|t|_q^2)\, \frac{\mathrm{d}s}{|s|_p^{p+1}}\, \frac{\mathrm{d}t}{|t|_q^{q+1}}.$$

By transformation to polar coordinates $s = r_x\theta$ and $t = r_y\phi$, where $r_x, r_y > 0$, $\theta \in S^{p-1}$,

and $\phi \in S^{q-1}$, the integral separates into a product of multiple integrals over (r_x, r_y), and over (θ, ϕ), respectively.

The integrals over r_x and r_y are standard gamma integrals:

$$\int_0^\infty \int_0^\infty r_x^{2k+2l_\bullet-2} r_y^{2k+2m_\bullet-2} \exp(-\sigma_x^2 r_x^2 - \sigma_y^2 r_y^2)\,\mathrm{d}r_x \mathrm{d}r_y = \frac{\Gamma(k+l_\bullet-\frac{1}{2})\,\Gamma(k+m_\bullet-\frac{1}{2})}{4\,\sigma_x^{2k+2l_\bullet-1}\,\sigma_y^{2k+2m_\bullet-1}},$$

where $l_\bullet = l_1 + \cdots + l_p$ and $m_\bullet = m_1 + \cdots + m_q$. As for the integrals over θ and ϕ, they are

$$\int_{S^{q-1}} \int_{S^{p-1}} (\theta' \Sigma_{XY} \phi)^{2k} \prod_{i=1}^{p} \theta_i^{2l_i} \prod_{j=1}^{q} \phi_j^{2m_j}\,\mathrm{d}\theta\,\mathrm{d}\phi.$$

To evaluate these integrals, we expand $(\theta' \Sigma_{XY} \phi)^{2k}$ using the multinomial theorem, obtaining a sum of terms, each of which is homogeneous in θ and ϕ. Then we integrate term-by-term by transforming the surface measures $\mathrm{d}\theta$ and $\mathrm{d}\phi$ to Euler angles [1, pp. 285–286]. The outcome is a multiple series expansion for the distance covariance. It does not appear to be a series that can be made simple in the general case, but it does provide an explicit expression in terms of Σ, p, and q.

A.2 The Affinely Invariant Distance Correlation for the Multivariate Laplace Distribution

Let $(X, Y) \sim L_{p+q}(\Sigma)$, i.e.

$$f_{X,Y}(s,t) = \Big(1 + \frac{1}{2} \begin{pmatrix} s \\ t \end{pmatrix}' \Sigma \begin{pmatrix} s \\ t \end{pmatrix} \Big)^{-1},$$

where $f_{X,Y}$ is the characteristic function of (X, Y). Hence, the characteristic functions of the marginals are

$$f_X(s) = \Big(1 + \frac{1}{2} s' \Sigma_{11} s\Big)^{-1} \quad \text{and} \quad f_Y(t) = \Big(1 + \frac{1}{2} t' \Sigma_{22} t\Big)^{-1},$$

respectively. Therefore, the affinely invariant distance covariance between X and Y can be computed as

$$c_p c_q \tilde{\mathcal{V}}(X,Y) = \int_{\mathbb{R}^{p+q}} \Big| \Big(1 + \frac{1}{2} \begin{pmatrix} s \\ t \end{pmatrix}' \Sigma \begin{pmatrix} s \\ t \end{pmatrix} \Big)^{-1} - \big(1 + \frac{1}{2} s' \Sigma_{11} s\big)^{-1} \big(1 + \frac{1}{2} t' \Sigma_{22} t\big)^{-1} \Big|^2 \times \frac{\sqrt{|\Sigma_{11}|}\,ds\,\sqrt{|\Sigma_{22}|}\,dt}{(s' \Sigma_{11} s)^{(p+1)/2} (t' \Sigma_{22} t)^{(q+1)/2}}.$$

By substituting $u = \sqrt{1/2}\,\Sigma_{11}^{1/2} s$ and $v = \sqrt{1/2}\,\Sigma_{22}^{1/2} t$, we obtain for the latter integral

$$2\int_{\mathbb{R}^{p+q}} \Big| \big(1 + u'u + v'v + 2u'\,\Sigma_{11}^{-\frac{1}{2}}\,\Sigma_{12}\,\Sigma_{22}^{-\frac{1}{2}}\,v\big)^{-1} - \big(1+u'u\big)^{-1}\big(1+v'v\big)^{-1}\Big|^2 \frac{du\,dv}{(u'u)^{(p+1)/2}\big(v'v\big)^{(q+1)/2}}.$$

Now we change variables to polar coordinates, putting $u = r_1\theta$ and $v = r_2\phi$ where $r_1, r_2 > 0$, $\theta = (\theta_1, \ldots, \theta_p)' \in S^{p-1}$, and $\phi = (\phi_1, \ldots, \phi_q)' \in S^{q-1}$. With $\Lambda := \Sigma_{11}^{-\frac{1}{2}}\,\Sigma_{12}\,\Sigma_{22}^{-\frac{1}{2}}$ the integral is equal to

$$2\int_{S^{p-1}\times S^{q-1}}\int_{\mathbb{R}_+\times\mathbb{R}_+} \Big| \big(1 + r_1^2 + r_2^2 + 2r_1r_2\,\theta'\,\Lambda\,\phi\big)^{-1} - \big(1+r_1^2\big)^{-1}\big(1+r_2^2\big)^{-1}\Big|^2 \frac{dr_1\,dr_2\,d\theta\,d\phi}{r_1^2 r_2^2}.$$

Again substituting $u = r_1^2$ and $v = r_2^2$ the latter integral equals

$$\frac{1}{2}\int_{S^{p-1}\times S^{q-1}}\int_{\mathbb{R}_+\times\mathbb{R}_+} \Big| \big(1 + u + v + 2\sqrt{uv}\,\theta'\,\Lambda\,\phi\big)^{-1} - \big(1+u\big)^{-1}\big(1+v\big)^{-1}\Big|^2 \frac{du\,dv\,d\theta\,d\phi}{u^{3/2}\,v^{3/2}}.$$

Furthermore, we change coordinates to $s = \frac{u}{1+u}$ and $t = \frac{v}{1+v}$. Observing that $1 + u = \frac{1}{1-s}$, $1 + v = \frac{1}{1-t}$ and

$$1 + u + v + 2\sqrt{uv}\,\theta'\,\Lambda\,\phi = \frac{1 - st + 2\,\theta'\Lambda\phi\sqrt{st}\,\sqrt{(1-s)(1-t)}}{(1-s)(1-t)}$$

the inner integral transforms to

$$\int_{[0,1]\times[0,1]} \Big| \big(1 - st + 2\theta'\,\Lambda\,\phi\,\sqrt{st}\;\sqrt{(1-s)(1-t)}\big)^{-1} - 1\Big|^2 \left(\frac{(1-s)\,(1-t)}{s\,t}\right)^{3/2} ds\,dt.$$

By expanding into negative binomial series, we obtain

$$\begin{aligned}
&\Big| \big(1 - st + 2\,\theta'\,\Lambda\,\phi\,\sqrt{st}\;\sqrt{(1-s)(1-t)}\big)^{-1} - 1\Big|^2 \\
&= \big(1 - st + 2\,\theta'\,\Lambda\,\phi\,\sqrt{st}\;\sqrt{(1-s)(1-t)}\big)^{-2} - 2\big(1 - st + 2\,\theta'\,\Lambda\,\phi\,\sqrt{st}\;\sqrt{(1-s)(1-t)}\big)^{-1} + 1 \\
&= \sum_{k=2}^{\infty} (k-1)\,\big(st - 2\,\theta'\Lambda\phi\sqrt{st}\sqrt{(1-s)(1-t)}\big)^k.
\end{aligned}$$

Moreover, by expanding into binomial series, the latter term reads

$$\sum_{k=2}^{\infty}(k-1)\sum_{i=0}^{k}\binom{k}{i}(st)^{k-i}(-1)^i(2\,\theta'\Lambda\phi\sqrt{st}\sqrt{(1-s)(1-t)})^i.$$

Hence

$$\begin{aligned}\tilde{\mathcal{V}}(X,Y) =&\frac{1}{2c_pc_q}\sum_{k=2}^{\infty}(k-1)\sum_{i=0}^{k}\binom{k}{i}(-1)^i\Big(\int_0^1 s^{k-i-3/2}(1-s)^{(i+3)/2}ds\Big)^2\\ &\times\int_{S^{p-1}\times S^{q-1}}(2\,\theta'\Lambda\phi)^i d\theta\,d\phi.\end{aligned}$$

Since $\int_{S^{p-1}\times S^{q-1}}(2\,\theta'\Lambda\phi)^i d\theta\,d\phi$ vanishes for i odd, this can be written as

$$\begin{aligned}\tilde{\mathcal{V}}(X,Y) =&\frac{1}{2c_pc_q}\sum_{k=2}^{\infty}(k-1)\sum_{j=0}^{\lfloor\frac{k}{2}\rfloor}\binom{k}{2j}\Big(\int_0^1 s^{k-j-3/2}(1-s)^{j+3/2}ds\Big)^2\\ &\times\int_{S^{p-1}\times S^{q-1}}(2\,\theta'\Lambda\phi)^{2j}d\theta\,d\phi.\end{aligned}$$

The integral with respect to s is a standard beta integral

$$\int_0^1 s^{k-j-3/2}(1-s)^{j+3/2}ds = B\left(k-j-\frac{1}{2},j+\frac{5}{2}\right),$$

where B is the beta function. Moreover the integral with respect to the spheres is well known to be

$$4c_{p-1}c_{q-1}\frac{(\frac{1}{2})_j\,(\frac{1}{2})_j}{(\frac{1}{2}p)_j(\frac{1}{2}q)_j}C_{(j)}(\Lambda),$$

where $(\alpha)_j$ denotes the rising factorial and $C_{(j)}(\cdot)$ is the top order zonal polynomial with weight j. As a result, we finally find

$$\tilde{\mathcal{V}}(X,Y) = 2\,\frac{c_{p-1}\,c_{q-1}}{c_p\,c_q}\sum_{k=2}^{\infty}(k-1)\sum_{j=0}^{\lfloor\frac{k}{2}\rfloor}2^{2j}\binom{k}{2j}B\left(k-j-\frac{1}{2},j+\frac{5}{2}\right)^2\frac{(\frac{1}{2})_j\,(\frac{1}{2})_j}{(\frac{1}{2}p)_j(\frac{1}{2}q)_j}C_{(j)}(\Lambda).$$

In the special case $\Sigma = I_{p+q}$, the affinely invariant distance covariance between X and Y reduces to

$$2\,\frac{c_{p-1}\,c_{q-1}}{c_p\,c_q}\sum_{k=2}^{\infty}(k-1)\,B\left(k-\frac{1}{2},\frac{5}{2}\right)^2>0,$$

which is a strictly positive constant.

Bibliography

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. third edition. Wiley, New York, 2003.

[2] G.E. Andrews, R.A. Askey, and R. Roy. *Special Functions*. Cambridge University Press, Cambridge, 2000.

[3] W.N. Bailey. *Generalized hypergeometric series*. Stechert-Hafner, 1964.

[4] S.K. Bar-Lev, D. Bshouty, G. Letac, I. Lu, and D.St.P. Richards. The diagonal multivariate natural exponential families and their classification. *Journal of Theoretical Probability*, 7:883–929, 1993.

[5] C. Bouveyron and S. Celeux, G.and Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32:1706–1713, 2011.

[6] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143:1835–1858, 2013.

[7] F. Camastra. Data dimensionality estimation methods: a survey. *Pattern recognition*, 36:2945–2954, 2003.

[8] W. Chang and D.St.P. Richards. Finite-sample inference with monotone incomplete multivariate normal data, I. *Journal of Multivariate Analysis*, 100:1883–1899, 2009.

[9] J.-P. Chilès and P. Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons, 2009.

[10] A.G. Constantine. Some non-central distribution problems in multivariate analysis. *The Annals of Mathematical Statistics*, 34:1270–1285, 1963.

[11] A.G. Constantine. The distribution of Hotelling's generalised T_0^2. *The Annals of Mathematical Statistics*, 38:215–225, 1966.

[12] J.P. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29:159–179, 1998.

[13] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51:157–172, 2000.

[14] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49:434–448, 2007.

[15] A.P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[16] Khare K. Diaconis, P. and L. Saloff-Coste. The diagonal multivariate natural exponential families and their classification. *Statistical Science*, 23:151–178, 2008.

[17] D. Drusvyatskiy, S.A. Vavasis, and H. Wolkowicz. Extreme point inequalities and geometry of the rank sparsity ball. *Mathematical Programming*, pages 1–24, 2014.

[18] J. Dueck, D. Edelmann, T. Gneiting, and D. Richards. The affinely invariant distance correlation. *Bernoulli*, 20:2305–2330, 2014.

[19] J. Dueck, D. Edelmann, and D. Richards. Distance correlation coefficients for lancaster distributions. *arXiv preprint http://arxiv.org/abs/1502.01413*, 2015.

[20] J. Dueck, D. Edelmann, and D. Richards. A generalization of an integral arising in the theory of distance correlation. *Statistics & Probability Letters*, 97:116–119, 2015.

[21] M.L. Eaton. *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, Hayward, California, 1989.

[22] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863–14868, 1998.

[23] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797, 2009.

[24] J. Faraut and A. Korányi. *Analysis on Symmetric Cones*. Oxford University Press, New York, 1994.

[25] T. Ferenci and L. Kovács. Using total correlation to discover related clusters of clinical chemistry parameters. In *Intelligent Systems and Informatics (SISY), 2014 IEEE 12th International Symposium on*, pages 49–54, 2014.

[26] E. Freitag and R. Busam. *Funktionentheorie 1*. Springer-Verlag Berlin Heidelberg, 2006.

[27] U. Gather, M. Imhoff, and R. Fried. Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, 21:2685–2701, 2002.

[28] C.F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae*. H. Dieterich, 1823.

[29] H. Gebelein. Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21:364–379, 1941.

[30] I.M. Gelfand and G.E. Shilov. *Generalized Functions*, volume 1. Academic Press, New York, 1964.

[31] T. Gneiting, K. Larson, K. Westrick, M. Genton, and E. Aldrich. Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time (RST) technique. *Journal of the American Statistical Association*, 101:968–979, 2006.

[32] L. A Goodman and W.H. Kruskal. Measures of association for cross classifications*. *Journal of the American Statistical Association*, 49:732–764, 1954.

[33] M. Gorfine, R. Heller, and Y. Heller. Comment on "Detecting novel associations in large data sets". *Unpublished manuscript*, 2012. `http://iew3.technion.ac.il/~gorfinm/files/science6.pdf`.

[34] I. S. Gradshteyn and I. M. Ryzhik. *Tables of Integrals, Series, and Products, fifth edition (A. Jeffrey, Editor)*. Academic Press, New York, 1994.

[35] T. Greene. The depiction of linear association by matroids. *Computational Statistics & Data Analysis*, 9:251–269, 1990.

[36] T. Greene. Descriptively sufficient subcollections of flats in matroids. *Discrete Mathematics*, 87:149–161, 1991.

[37] K.I. Gross and D.St.P. Richards. Special functions of matrix argument. I: Algebraic induction, zonal polynomials, and hypergeometric functions. *Transactions of the American Mathematical Society*, 301:781–811, 1987.

[38] B. Hall. *Lie groups, Lie algebras, and Representations: an Elementary Introduction*. Springer, 2003.

[39] W. Härdle and L. Simar. Canonical correlation analysis. *Applied Multivariate Statistical Analysis*, pages 321–330, 2007.

[40] R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100:503–510, 2013.

[41] A.S. Hering and M.G. Genton. Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, 105:92–104, 2010.

[42] K. Horimoto and H. Toh. Statistical estimation of cluster boundaries in gene expression profile data. *Bioinformatics*, 17:1143–1151, 2001.

[43] A.K. Jain, M. N. Murty, and P.J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31:264–323, 1999.

[44] A.T. James. The distribution of the latent roots of the covariance matrix. *The Annals of Mathematical Statistics*, 31:151–158, 1960.

[45] A.T. James. Zonal polynomials of the real positive definite symmetric matrices. *Annals of Mathematics*, 74:456–469, 1961.

[46] A.T. James. Distributions of matrix variates and latent roots derived from normal samples. *Annals of Mathematical Statistics*, 35:475–501, 1964.

[47] A.T. James. Calculation of zonal polynomial coefficients by use of the Laplace-Beltrami operator. *The Annals of Mathematical Statistics*, 74:1711–1718, 1968.

[48] K. Kanatani. Motion segmentation by subspace separation and model selection. 2:586–691, 2001.

[49] A.W. Knapp. *Lie groups beyond an introduction*. Springer, 2002.

[50] P. Koev and A. Edelman. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, 75:833–846, 2006.

[51] I. Kojadinovic. Agglomerative hierarchical clustering of continuous variables based on mutual information. *Computational statistics & data analysis*, 46:269–294, 2004.

[52] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[53] J. Kong, B.E.K. Klein, R. Klein, and G. Wahba. Using distance correlation and SS-ANOVA to access associations of familial relationships, lifestyle factors, diseases, and mortality. *Proceedings of the National Academy of Sciences*, 109:20352–20357, 2012.

[54] Y. Konno. Estimation of a normal covariance matrix with incomplete data under Stein's loss. *Journal of Multivariate analysis*, 52:308–324, 1995.

[55] M.R. Kosorok. Discussion of: Brownian distance covariance. *Annals of Applied Statistics*, 3:1270–1278, 2009.

[56] S. Kotz, N. Balakrishnan, and N.L. Johnson. *Continuous Multivariate Distributions: Models and Applications*. Wiley, New York, 2000.

[57] A.E. Koudou. Probabilités de lancaster. *Expositiones Mathematicae*, 14:247–275, 1996.

[58] A.E. Koudou. Lancaster bivariate probability distributions with poisson, negative binomial and gamma margins. *Test*, 7:95–110, 1998.

[59] H.O. Lancaster. The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29:719–736, 1958.

[60] H.O. Lancaster. *The Chi-Squared Distribution*. Wiley, New York, 1969.

[61] Y. M. Latha and K. Chavali. Two-stage variable clustering for data sets in distributed systems. *Advances in Digital Multimedia*, 1:124–130, 2012.

[62] S.L. Lauritzen. *Graphical models*, volume 17. Oxford University Press, 1996.

[63] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35:171–184, 2013.

[64] R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41:3284–3305, 2013.

[65] J.R. Magnus and H. Neudecker. The commutation matrix: Some properties and applications. *The Annals of Statistics*, 7:381–394, 1979.

[66] E. Martínez-Gómez, M.T. Richards, and D.St.P. Richards. Distance correlation methods for discovering associations in large astrophysical databases. *The Astrophysical Journal*, 781:39 (11 pp.), 2014.

[67] T.P. Minka. Automatic choice of dimensionality for PCA. In *NIPS*, volume 13, pages 598–604, 2000.

[68] R.J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, New York, 1982.

[69] H. Neudecker. The Kronecker matrix product and some of its applications in econometrics. *Statistica Neerlandica*, 22:69–82, 1968.

[70] F.W.J. Olver. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[71] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter - - Special issue on learning from imbalanced datasets*, 6:90–105, 2004.

[72] D. Pommeret. A characterization of lancaster probabilities with margins in a multivariate additive class. *Sankhyā*, 66:1–19, 2004.

[73] S.T. Rachev, L. Klebanov, S.V. Stoyanov, and F. Fabozzi. *The methods of distances in the theory of probability and statistics*. Springer, 2013.

[74] S. Reid and R. Tibshirani. Sparse regression and marginal testing using cluster prototypes. *arXiv preprint arXiv:1503.00334*, 2015.

[75] B. Rémillard. Discussion of: Brownian distance covariance. *Annals of Applied Statistics*, 3:1295–1298, 2009.

[76] A. Rényi. On measures of dependence. *Acta mathematica hungarica*, 10:441–451, 1959.

[77] D.N. Reshef, J.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. *Science*, 334:1518–1524, 2011.

[78] E. Richard, P.A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012.

[79] M.T. Richards, D.St.P. Richards, and E. Martínez-Gómez. Interpreting the distance correlation results for the combo-17 survey. *The Astrophysical Journal Letters*, 784:L34 (5 pp.), 2014.

[80] M. L. Rizzo and G. J. Székely. Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4:1034–1055, 2010.

[81] M.L. Rizzo and G.J. Székely. Energy: E-statistics (energystatistics). r package, version 1.4-0. 2011. available at `http://cran.us.r-project.org/web/packages/energy/index.html`.

[82] A. Roy and C.B. Post. Detection of long-range concerted motions in protein by a distance covariance. *Journal of chemical theory and computation*, 8:3009–3014, 2012.

[83] R. Sanche and K. Lonergan. Variable reduction for predictive modeling with clustering. In *Casualty Actuarial Society Forum*, pages 89–100, 2006.

[84] W.S. Sarle. The VARCLUS procedure. *SAS/STAT User's Guide,*, 1990.

[85] O.V. Sarmanov. Generalized normal correlation and two-dimensional fréchet classes. *Soviet Mathetmatics Doklady*, 7:596–599, 1966.

[86] O.V. Sarmanov. An approximate calculation of correlation coefficients between functions of dependent random variables. *Math. Notes Acad. Sciences USSR*, 7:373–377, 1970.

[87] O.V. Sarmanov. Gamma correlation process and its properties. *Doklady Akademii Nauk*, 191:30–32, 1970.

[88] J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21:754–764, 2005.

[89] A. Seghouane and A. Cichocki. Bayesian estimation of the number of principal components. *Signal Processing*, 87:562–568, 2007.

[90] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41:2263–2291, 2013.

[91] H. Shen and J.Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99:1015–1034, 2008.

[92] N. Simon and R. Tibshirani. Comment on "Detecting novel associations in large data sets," by Reshef, et al. Unpublished manuscript, available at `http://www-stat.stanford.edu/~tibs/reshef/comment.pdf`.

[93] M. Siotani. Some applications of Loewner's ordering on symmetric matrices. *Annals of the Institute of Statistical Mathematics*, 19:245–259, 1967.

[94] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15:72–101, 1904.

[95] H.M. Srivastava and J.P. Singhal. Some extensions of the Mehler formula. *Proceedings of the American Mathematical Society*, 31:135–141, 1972.

[96] G. J. Székely and M. Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013.

[97] G. J. Székely and M. Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42:2382–2412, 2014.

[98] G.J. Székely and M. Rizzo. On the uniqueness of distance correlation. *Statistics & Probability Letters*, 82:2278–2282, 2012.

[99] G.J. Székely and M.L. Rizzo. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, 22:151–183, 2005.

[100] G.J. Székely and M.L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 3:1236–1265, 2012.

[101] G.J. Székely and M.L. Rizzo. On the uniqueness of distance covariance. *Statistics & Probability Letters*, 82:2278–2282, 2012.

[102] G.J. Székely, M.L. Rizzo, and N.K. Bakirov. Measuring and testing independence by correlation of distances. *Annals of Statistics*, 35:2769–2794, 2007.

[103] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61:611–622, 1999.

[104] H. Toh and K. Horimoto. Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics*, 18:287–297, 2002.

[105] D.E. Tyler. Asymptotic inference for eigenvectors. *The Annals of Statistics*, 9:725–736, 1981.

[106] G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. *arXiv preprint arXiv:1206.6447*, 2012.

[107] I.P. Vaughan and S.J. Ormerod. Methodological insights: Increasing the value of principal components analysis for simplifying ecological data: a case study with rivers and river birds. *Journal of Applied Ecology*, 42:487–497, 2005.

[108] R. Vidal. A tutorial on subspace clustering. *Signal Processing Magazine*, 28:52–68, 2010.

[109] P.J. Waddell and H. Kishino. Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Informatics*, 11:129–140, 2000.

[110] Y. Wang, Huan Xu, and C. Leng. Provable subspace clustering: When lrr meets ssc. In *Advances in Neural Information Processing Systems (NIPS)*, pages 64–72, 2013.

[111] S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4:66–82, 1960.

[112] C. S. Withers and S. Nadarajah. Expansions for the multivariate normal. *Journal of Multivariate Analysis*, 101:1311–1311, 2010.

[113] S. Wold, A. Ruhe, H. Wold, and W.J. Dunn III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5:735–743, 1984.

[114] A.S. Woolston. *Working with collinearity in epidemiology: development of collinearity diagnostics, identifying latent constructs in exploratory research and dealing with perfectly collinear variables in regression.* University of Leeds, 2012.

[115] W.B. Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102:14150–14154, 2005.

[116] Z. Zhou. Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, 33:438–457, 2012.

[117] D. Zhu, A.O. Hero, H. Cheng, R. Khanna, and A. Swaroop. Network constrained clustering for gene microarray data. *Bioinformatics*, 21:4014–4020, 2005.