NNT : 2017SACLC002

Thèse de Doctorat de l'Université Paris-Saclay & Doctor of Engineering Science, KU Leuven

préparée à CentraleSupélec

École doctorale N°580
Sciences et Technologies de l'Information et de la Communication
Specialité: Mathématiques et Informatique

Arenberg Doctoral School
Faculty of Engineering Science

par Wacha Bounliphone

Tests d'hypothèses statistiquement et algorithmiquement efficaces de similarité et de dépendance

Thèse présentée et soutenue à Gif-sur-Yvette, le 30 janvier 2017.

Composition du Jury:

M.	**Nikos Paragios**	Professeur Université Paris-Saclay & Inria	(Président du Jury)
M.	**Jean-Philippe Vert**	Directeur de Recherche MINES ParisTech	(Rapporteur)
M.	**Dominik Janzing**	Professeur Max Planck Institute for Intelligent Systems	(Rapporteur)
M.	**Arthur Gretton**	Professeur University College London	(Examinateur)
M.	**Jesse Davis**	Professeur KU Leuven	(Examinateur)
M.	**Arthur Tenenhaus**	Professor KU Leuven	(Directeur de thèse)
M.	**Matthew Blaschko**	Professeur Université Paris-Saclay	(Co-Directeur de thèse)

Acknowledgements

This thesis is the achievement of efforts of many individuals. May these words acknowledge their investment and contributions.

First of all, I would like to express my sincere and deepest gratitude towards my thesis advisors Prof. Artur Tenenhaus and Prof. Matthew Blaschko for their continuous support and advice during my whole PhD. Your passion and expertise for machine learning and statistics, your ambition and rigor at work has been a source of admiration and motivation. Thanks for investing a lot of your time to guide me and to improve my work during the last three years. Je vous remercie également pour tout vos conseils, c'est un honneur de vous connaître: toujours de bonne humeur, intègre et très humain.

Second, I would also like to thank my reviewers Prof. Jean-Philippe Vert and Prof. Daminik Janzing for spending their valuable time for reviewing my thesis and providing useful comments and suggestions. I would also like to thank Prof. Arthur Gretton (it is my pleasure and honor to having collaborated with you during these years), Prof. Jesse Davis and Prof. Nikos Paragios for their role as an examiner. I am honored that such brilliant people would spend time on my work.

Ma reconnaissance va également à Dr. Guillaume Saint-Pierre, Prof. Philippe Saint-Pierre, Prof. Michel Broniatowski et Prof. Paul Deheuvels pour m'avoir fortement encouragé et poussé à faire une thèse. Sans eux, je n'aurai pas eu la chance de faire cette incroyable aventure.

Je tiens égalemenet à remercier à tous les membres du CVN, les anciens et les présents pour l'ambiance sympathique qui y règne. Tout d'abord, je remercie Nikos Paragios pour m'avoir accueilli au Center for Visual Computing (CVN). Grâce à vous, on se sent parfaitement accueilli, motivé et encouragé pour travailler dans un excellent environnement. Puis je remercie Natalia Leclerc pour m'avoir aidé dans les procédures administratives mais surtout pour sa gentillesse, tu joues un rôle majeur, comme une mère merci de prendre soin de nous, de nous rassurer et encourager durant cette aventure. Puis je remercie Jiaqian Yu et Eugene Belilovsky pour leur aide, leur encouragement et surtout leur bonne humeur, je suis contente d'avoir partagé d'excellents moments avec vous en conférence ou durant les deadline. Également un grand merci pour leur humour et gentillesse à mes collègues du CVN: Erwan Zerhouni, Puneet Dokania Kumar, Enzo Ferrante, Evgenios Kornaropoulos, Siddhartha Chandra, Hariprasad Kannan, Mihir Sahasrabudhe, Maxim Berman, Stefan Kinauer, Stavros Tsogkas and Khue Le-Huu.

Également, sincère remerciements aux meilleures filles et collègues de Supélec, Gisela Lechuga et Wenmeng Xiong. Merci aussi à Luc et Anne Batalie pour avoir facilité toutes les procédures administratives.

I would like to express my sincere gratitude to all my colleagues in the PSI-VISICS lab

for providing a great work atmosphere, for chilling out together in Leuven and beach-volley games in the summer: Xuanli Chen, Yu-Hui Huang, José Oramas, Xu Jia, Amal Rannen, Rahaf Aljundi, Gina Stavropoulou, Amir Ghodrati, Bert de Brabandere, Davy Neven, Klaas Kelchtermans, Jan-Pieter d'Anvers, Jay Chakravarty and Vivek Sharma.

Il y aussi des amis que je considère comme membres de la famille que je souhaite remercier. Eric Souvanna, Samson Tat et Clément Trieu, même si vous ne le savez pas, les moments passés avec vous m'ont aidé durant cette thèse. Nos soirées, notre séjour à Londres en passant par le Millennium Bridge et surtout notre voyage en Australie de Melbourne à Townsville sont des souvenirs mémorables.

Plus personnellement, je remercie ma famille qui, si elle n'est pas bien grande n'en est pas moins le plus incroyable des soutiens. Tout d'abord mes parents Paul et Sophie mais également mon frère Jacques et ma soeur Wela qui ont toujours été présent avant et pendant ma thèse. Leur soutien et leurs encouragements m'ont beaucoup aidé. Ensuite, je remercie la famille Soukhavongsa et Sok, merci d'avoir toujours pris soin de moi. Plus particulièrement je remercie ma cousine Sophie Soukhavongsa pour tout les plats qu'elle m'a préparé et tout les moments passé avec ses enfants. Merci également à ma belle-famille, les beaux-parents Marcel et Sandrine Tong, Tonton Soukaya et Tata Ty Sayatham ainsi que Evelyne, William et Ethan Robert et les petites Malisa et Athina pour tout les moments en famille ensemble.

Enfin mes derniers remerciements sont pour Christian Tong, mon meilleur ami et surtout mon amour à qui cette thèse est dédiée. Sans lui cette thèse ne se serait jamais réalisée. Merci pour avoir supporté 24h/24 les moments de joie et de stress d'une doctorante. Merci d'avoir voyager partout dans le monde avec moi, avec toi j'ai pu m'évader et me ressourcer. Merci d'avoir toujours été présent pour moi, pour ta patience, pour tout les mots pour me rassurer et m'encourager durant cette aventure.

Abstracts

Abstract — The dissertation presents novel statistically and computationally efficient hypothesis tests for relative similarity and dependency, and precision matrix estimation. The key methodology adopted in this thesis is the class of U-statistic estimators. The class of U-statistics results in a minimum-variance unbiased estimation of a parameter. We make use of asymptotic distributions and strong consistency of U-statistic estimators to develop novel non-parametric statistical hypothesis tests.

The first part of the thesis focuses on relative similarity tests applied to the problem of model selection. Probabilistic generative models provide a powerful framework for representing data. Model selection in this generative setting can be challenging, particularly when likelihoods are not easily accessible. To address this issue, we provide a novel non-parametric hypothesis test of relative similarity and test whether a first candidate model generates a data sample significantly closer to a reference validation set. Our model selection criterion is based on the Maximum Mean Discrepancy (MMD) and measures the distance of the generated samples to some reference target set.

Subsequently, the second part of the thesis focuses on developing a novel non-parametric statistical hypothesis test for relative dependency. Tests of dependence are important tools in statistical analysis, and several canonical tests for the existence of dependence have been developed in the literature. For many problems in data analysis, however, the question of whether there exist dependencies is secondary. The determination of whether one dependence is stronger than another is frequently necessary for decision making in real-world scenarios. We present a statistical test which determine whether one variables is significantly more dependent on a first target variable or a second. Dependence is measured via the Hilbert-Schmidt Independence Criterion (HSIC). The resulting tests of dependence and relative similarity are consistent and unbiased (being based on U-statistics) and can be computed in $\mathcal{O}(n^2)$, where n is the sample size.

Finally, a novel method for structure discovery in a graphical model is proposed. Making use of a result that zeros of a precision matrix can encode conditional independencies, we develop a test that estimates and bounds an entry of the precision matrix. Methods for structure discovery in the literature typically make restrictive distributional (e.g. Gaussian) or sparsity assumptions that may not apply to a data sample of interest, and direct estimation of the uncertainty of an estimate of the precision matrix for general distributions remains challenging. Consequently, we derive a new test that makes use of results for U-statistics and applies them to the covariance matrix, which then implies a bound on the precision matrix. The resulting test enables one to answer with statistical significance whether an entry in the precision matrix is non-zero, and $\mathcal{O}(n^{-1/2})$ convergence results are known for a wide range of distributions. The computational complexity is linear in the sample size.

Keywords: U-statistics, hypothesis testing, dependency, similarity, kernel methods.

Titre — Tests d'hypothèses statistiquement et algorithmiquement efficaces de similarité et de dépendance

Résumé — Cette thèse présente de nouveaux tests d'hypothèses statistiques efficaces pour la relative similarité et dépendance, et l'estimation de la matrice de précision. La principale méthodologie adoptée dans cette thèse est la classe des estimateurs U-statistiques. La classe des U-statistiques donne lieu à une estimation sans biais de variance minimale d'un paramètre. L'utilisation d'estimateurs U-statistics à distributions asymptotiques fortement consistantes permet le développement de nouveaux tests d'hypothèses statistiques non paramétriques.

La première partie de la thèse porte sur les tests de relative similarité appliqués au problème de la sélection de modèles. Les modèles génératifs probabilistes fournissent un cadre puissant pour représenter les données. La sélection de modèles dans ce contexte génératif peut être difficile, en particulier lorsque les probabilités ne sont pas facilement accessibles. Pour résoudre ce problème, nous proposons un nouveau test d'hypothèse non paramétrique de relative similarité et testons si un premier modèle candidat génère un échantillon de données significativement plus proche d'un ensemble de validation de référence. Notre critère de sélection de modèle est basé sur le Maximum Mean Discrepancy (MMD) et mesure la distance entre des échantillons générés et une cible de référence fixée.

La deuxième partie de la thèse développe un test d'hypothèse statistique non paramétrique pour la relative dépendance. Plusieurs tests de dépendances statistiques ont déjà été développés dans la littérature. Toutefois, en présence de dépendances multiples, les méthodes existantes ne répondent qu'indirectement à la question de la relative dépendance. Or, savoir si une dépendance est plus forte qu'une autre est important pour la prise de décision dans des scénarios réels. Nous présentons un test statistique qui détermine si une variable dépend beaucoup plus d'une première variable cible ou d'une seconde variable. La dépendance est mesurée au moyen du Hilbert-Schmidt Independence Criterion (HSIC). Les tests de relatifs similarité et de dépendance résultants sont cohérents et non biaisés (étant basés sur les U-statistiques) et peuvent être calculés en $\mathcal{O}(n^2)$, où n est la taille de l'échantillon.

Enfin, une nouvelle méthode de découverte de structure dans un modèle graphique est proposée. En partant du fait que les zéros d'une matrice de précision représentent les indépendances conditionnelles, nous développons un nouveau test statistique qui estime une borne pour une entrée de la matrice de précision. Les méthodes existantes de découverte de structure font généralement des hypothèses restrictives de distributions gaussiennes ou parcimonieuses qui ne correspondent pas forcément à l'étude de données réelles. Ainsi, l'estimation directe de l'incertitude d'une estimation de la matrice de précision pour les distributions générales demeure difficile. Nous introduisons ici un nouveau test utilisant les propriétés des U-statistics appliqués à la matrice de covariance, et en déduisons une borne sur la matrice de précision. Sans faire d'hypothèse sur la distribution, ce test permet de déterminer significativement si un coefficient de la matrice de précision est non nul. Nous démontrons une convergence du test en $\mathcal{O}(n^{-1/2})$ pour une large gamme de distributions. La complexité algorithmique est linéaire en la taille de l'échantillon.

Titel — Statistisch en computationeel efficënte hypothesetoetsen voor similariteit en afhankelijkheid

Abstract — Dit werk presenteert nieuwe reken- en statistisch efficiënte hypothesetesten op relatieve similariteit en afhankelijkheid, en op precisiematrixschatting. De belangrijkste methode voorgesteld in dit werk, is de klasse van U-statistic schatters. De klasse van U-statistics resulteert in een zuivere schatting met minimum-variantie van een parameter. We maken gebruik van asymptotische distributies en sterke consistentie van U-statisticschatters om nieuwe niet-parametrische statistische hypothesetoetsen te ontwikkelen.

Het eerste deel van dit werk richt zich op de relatieve similariteitstesten toegepast op het probleem van modelselectie. Probabilistische generatieve modellen bieden een waardevol kader voor het weergeven van data. Modelselectie bij generatieve modellen kan echter een uitdaging zijn, vooral als observaties van de aannemelijkheidsfunctie moeilijk te verkrijgen zijn. Om dit probleem aan te pakken, stellen we een nieuwe niet-parametrische hypothesetest op relatieve similariteit voor en testen we of een monster gegenereerd door een eerstekandidaatsmodel aanzienlijk dichter ligt bij de validatieset. Ons criterium voor modelselectie is gebaseerd op de Maximum Mean Discrepancy (MMD) en meet de afstand tussen de gegenereerde monsters en een referentieset.

Vervolgens richt het tweede deel van dit werk zich op het ontwikkelen van een nieuwe niet-parametrische statistische hypothesetest op relatieve afhankelijkheid. Afhankelijkheidstesten zijn belangrijk voor statistische analyse. Er zijn verschillende canonieke testen ontwikkeld op het bestaan van afhankelijkheid in de literatuur. Voor talrijke problemen in gegevensanalyse is echter de vraag of afhankelijkheid bestaat van secundair belang. Het kunnen vaststellen of een afhankelijkheid sterker is dan een andere, is vaak noodzakelijk voor de besluitvorming in real-world scenario's. We presenteren een statistische test die kan bepalen of een variabele significant sterker of minder sterk afhankelijk is van een doelvariabele. Afhankelijkheid wordt gemeten via de Hilbert-Schmidt Independence Criterium (HSIC). De resulterende testen van afhankelijkheid en relatieve similariteit zijn consistent en zuiver (omdat ze gebaseerd zijn op U-statistics) en kunnen worden berekend in $\mathcal{O}(n^2)$, waarbij n de grootte is van de steekproef.

Tenslotte wordt een nieuwe werkwijze voor structuurontdekking in grafische modellen voorgesteld. Gebruikmakend van het feit dat nullen in een precisiematrix voorwaardelijke onhankelijkheden kunnen coderen, ontwikkelen we een test die de waarden in een precisiematrix schat en begrenst. Methoden voor structuurontdekking in de literatuur maken doorgaans veronderstellingen over de distributie (bijvoorbeeld een Gaussische verdeling) of de spaarsheid die vaak niet van toepassing zijn op een bepaalde steekproef. Het verkrijgen van een directe schatting voor de onzekerheid van een waarde uit de precisiematrix blijft uitdagend. Daarom leiden we een nieuwe test af die gebruik maakt van de resultaten van U-statistics en deze toepast op de covariantiematrix, wat een begrenzing op de waarde van de precisiematrix impliceert. De resulterende test laat ons toe om met statische significantie the bepalen of een waarde in de precisiematrix verschillend is van nul. $\mathcal{O}(n^{-1/2})$ convergentieresultaten zijn gekend voor een breed scale aan verdelingen. De computationele complexiteit groeit lineair met de grootte van de steekproef.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hypothesis testing is one of the most important procedure in statistics as a method of making decisions using data. In order to undertake hypothesis testing, we need to express a research hypothesis as a null when nothing happened, or there were no differences, or no cause and effect and alternative hypothesis when we are correct in a theory [Casella and Berger, 2002]. In this thesis, we consider the problem of hypothesis tests for similarity and dependency which are of fundamental importance in statistics.

First, the concept of distances between distributions is important. Relative similarity tests determine whether the distribution of a source sample is closer to the distribution of one target sample or another. An example application is to test if a new process or treatment is superior to a current process or treatment.

Second, the concept of dependence relations between random variables plays a very important role in many fields of mathematics and is one of the most widely studied subjects in probability and statistics. The investigation of such inter-relationships is of great scientific importance. For example, in studies of complex diseases, the exploration of the inter-relationship among the responsible genes is crucial for the understanding of the disease pathologies. Moreover, the concept of conditional independence, describes the relationship between two variables while conditioning on another variable. A well known theorem by Hammersley and Clifford [1971] relates such conditional dependencies to an underlying graph topology. The study of the topology of a graphical model reflects the conditional independence assertions between the observable variables.

This thesis focuses on the exploration of methods that exploit such information.

Contents

1.1 Motivation and Tasks of Interest

There are a variety of problems for testing similarity and dependence addressed in the machine learning and statistics literature. In this thesis, we focus our attention on the study of relative similarity, relative dependency and the concept of conditional independence. A more detailed look at related work is provided in the sequel.

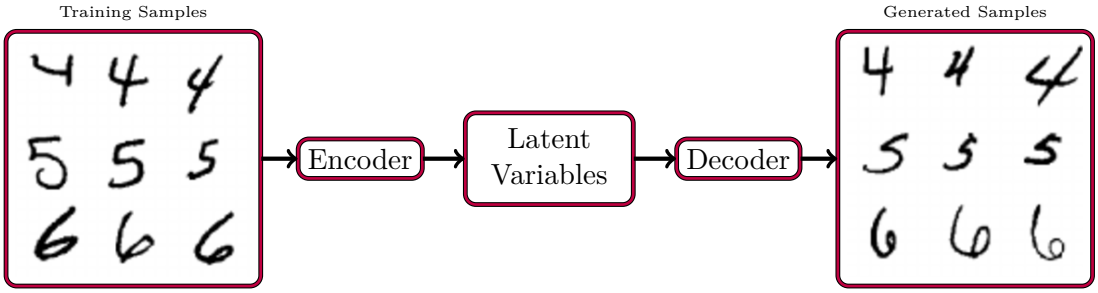
1.1.1 The Study of Relative Similarity

Many problems in testing and learning require evaluating the similarity of distributions. Examples of these applications include lexical semantic similarity, comparing the diversities of two communities or measuring a delay between two signals (e.g. stock market fluctuations in financial data or electrocardiograms in medical data). Another potential example can be found in model selection. It is important to be able to evaluate the quality of samples from a generator by measuring their similarity to a sample of reference data. E.g. when a complex generative model based on deep learning techniques is learned, it is necessary to provide feedback on the quality of the samples produced. The two-sample test is a task that deals with testing whether distributions $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{y}}$ are different on the basis of samples drawn from each of them. Formally, given independent and identically distributed (i.i.d.) samples $\mathbf{x}$ and $\mathbf{y}$ drawn from $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{y}}$, respectively, we want to test if $\mathbb{P}_{\mathbf{x}} = \mathbb{P}_{\mathbf{y}}$.

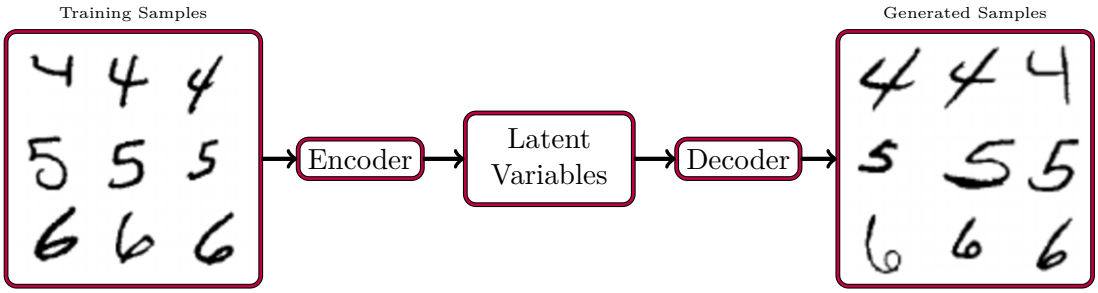
The past decade saw significant advances in the task of testing distribution similarity on complex structured data. One of the main driving forces of these advances is the use of kernel methods. Kernel methods are a class of non-parametric learning techniques relying on the Reproducing Kernel Hilbert Space (RKHS) construction, and utilize positive definite kernels taken over sufficiently rich function classes. In the last few decades kernel methods have been employed in the design of new methods to tackle several machine learning problems to compare objects which have much more complex structure [Schölkopf and Smola, 2001]. Key to the success of any kernel method is the definition of an appropriate kernel for the data at hand.

Gretton et al. [2012a] propose three simple multivariate tests for comparing samples from two distributions $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{y}}$. The test statistic is the maximum mean discrepancy (MMD), defined as the maximum deviation in the expectation of a function evaluated on each of the random variables $\mathbf{x}$ and $\mathbf{y}$, taken over a sufficiently rich function class, the RKHS. The resulting test has $\mathcal{O}(n^2)$ computational complexity, where n is the sample size, and can be applied to a variety of problems.

The ability of the kernel two-sample test to give excellent performance on complex structured data yields a promising approach for the task of testing relative similarity. The problem setting is to determine whether a target distribution is closer to one of two candidate distributions. An important potential application of the proposed test can be found in recent work with deep neural networks. The problem is to determined which of two models generates samples that are closer to a real-world reference sample of interest by testing the models' samples



(a) Samples generated from a first model.



(b) Samples generated from a second model.

Figure 1.1 – Illustration of the Variational Auto-encoder reference model from Kingma and Welling [2014]. A first goal of this thesis is to test for relative similarity for model selection in generative models. We may have several different models under consideration, and we want to know which is the best match to a data sample.

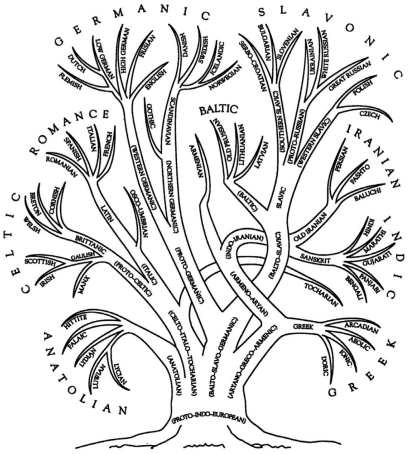
against the reference data set (Figure 1.1).

1.1.2 The Study of Relative Dependency

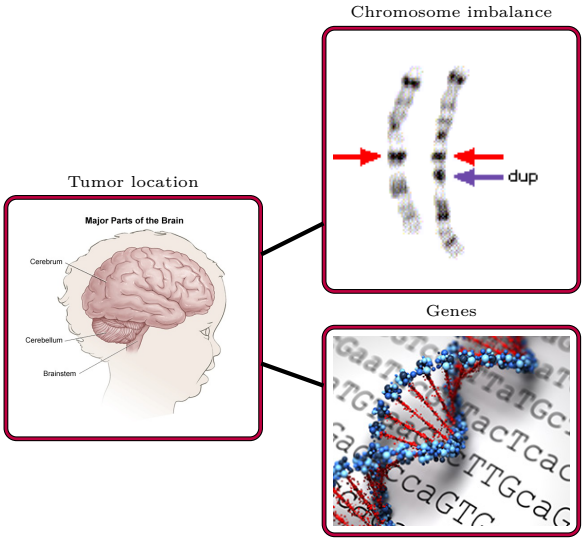
Parallel to the advances on evaluating distribution similarity, testing dependency has also achieved promising results in several applications. It is very important to understand the relationship between variables to draw the right conclusion from a statistical analysis. For example, analyzing relationships between financial variables is useful in many ways. Such an analysis can be helpful in identify the factors that are most responsible for profits for instance. A first step in the analysis is to quantify the relationship between the two sets of variables using coefficients of association and then decide if the association is significant by using a statistical test.

Many different coefficients and tests have been published as measures of association between two data samples. Also, kernel methods have been successfully used for capturing dependence of variables and the resulting methods can be applied to a variety of problems related to the estimation of dependency in structured domains, such as text, images, graphs and captions (Figure 1.2).

However, it is important to note that, despite these advances, the question of relative depen-



(a) Illustration of the Indo-European family tree from Schaller-Schwaner [2015].



(b) Illustrations from Chromosome Disorder Outreach [2016], National Human Genome Research Institute [2016], Nationwide Children's Hospital [2016].

Figure 1.2 – The secondary goal of this thesis is to find significant relative statistical dependencies between different sets of variables. (a) The identification of language groups from a multilingual parallel corpus. (b) Brain tumors have different genetics origins depending on the location of the tumor in the brain. The goal is to identify the mechanisms responsible for the tumor.

dencies has not been previously addressed in the literature.

1.1.3 The Study of Structure Discovery

Finally, conditional independence (CI) tests have received special attention lately in the machine learning literature as an important indicator of the relationship between variables in a model. Statistical dependence between observed variables can have a conditional relation, for instance conditional on the time or location of a study. A more explicit form of conditioning may result by design or by statistical analysis of several variables. In that case the distinction between conditional and marginal (in)dependence becomes relevant.

Indeed, conditional independence tests are especially important and are challenging for the task of learning the structure of a probabilistic graphical model from data.

For Gaussian graphical models, originating with the seminal work of Dempster [1972], a considerable body of literature has been developed on the identification of non-zero entries in the inverse of the covariance matrix of a random vector, called the *precision matrix*. For a large class of distributions, a non-zero entry in the precision matrix corresponds to two variables that have a non-zero partial correlation, i.e. they are correlated conditioning on all

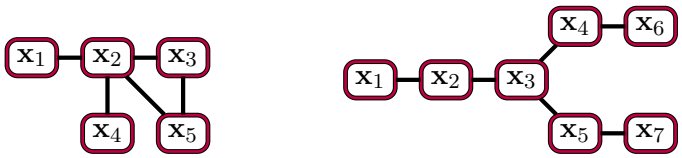


Figure 1.3 – A third goal of the thesis is discovering structure in graphical model. A graphical model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables. For a large class of distributions, the presence of an edge in the graph corresponds to a non-zero entry in the precision matrix.

other variables (Figure 1.3). A large fraction of this literature focuses on a specific setting in which *(i)* data are assumed to be Gaussian distributed, and *(ii)* the precision matrix is assumed to obey a sparsity assumption.

However, while there exist many methods in the literature using strong assumptions of data being generated by discrete or Gaussian multivariate distributions, other distributions pose new challenges in statistical modeling for real-world data.

1.2 Research Questions

From the previous section it is evident that the potential of new statistical hypothesis testing for dependence and similarity has not been sufficiently explored in prior work. We therefore advance the theory and practice of hypothesis testing in three related settings. First, the family of two-sample tests that make use of distances between distributions is one of the most commonly used strategies to quantify the similarity between two data samples. We propose to compare samples from two probability distributions which have different distance to a reference distribution. Second, when there exists classical criteria for test of dependence, the question of whether dependence exists is secondary; there may be multiple dependencies, and the question becomes which dependency is the strongest. Furthermore, for conditional dependency, the methods for structure discovery in the literature typically make restrictive Gaussian assumptions or sparsity assumptions that may not apply to a data sample of interest. So direct estimation of the uncertainty of an estimate of the precision matrix for general distributions remains challenging.

Considering these points, the objective of this thesis is to exploit new statistically and efficiency hypothesis testing for similarity and dependence. For this reason and for clarity of presentation, we split the thesis into three questions to address specifically some of these factors. For the formal setup of this section, suppose that we have random variables $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$, $\mathbf{y} \sim \mathbb{P}_{\mathbf{y}}$ and $\mathbf{z} \sim \mathbb{P}_{\mathbf{z}}$, that take values on $(\mathcal{X}, \mathcal{B}_{\mathbf{x}})$, $(\mathcal{Y}, \mathcal{B}_{\mathbf{y}})$, and $(\mathcal{Z}, \mathcal{B}_{\mathbf{z}})$, respectively, where here $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ are separable metrics and $\mathcal{B}_{\mathbf{x}}$, $\mathcal{B}_{\mathbf{y}}$, and $\mathcal{B}_{\mathbf{z}}$ are Borel σ-algebras.

Research Questions:

1. Is the probability measure $\mathbb{P}_{\mathbf{x}}$ significantly closer to $\mathbb{P}_{\mathbf{y}}$ or to $\mathbb{P}_{\mathbf{z}}$?

2. Is the dependency between $\mathbf{x}$ and $\mathbf{y}$ significantly stronger than the dependency between $\mathbf{x}$ and $\mathbf{z}$?

3. Can we develop a statistically and computationally efficient estimator of the topology of graphical models for non-Gaussian distributions?

Answering these research questions resulted in the following contributions.

1.3 Overview and Contributions of the Thesis

The line of work presented in this thesis focused on reasoning about statistically and computationally efficient hypothesis tests for dependency and similarity. The work covered in this thesis has been collected in several papers. The content and contributions of these papers are presented in Chapters 3, 4 and 5. As a whole, the contents of these papers address the research questions introduced earlier with each chapter having specific contributions.

In Chapter 2, we present fundamental principles and tools used in the different methods presented in the thesis. The objective of this chapter is to lay the foundations for the rest of the thesis.

Chapter 3 presents a novel non-parametric of relative similarity. The contributions of this part is based on classic theory of U-statistics to analyze the asymptotic distribution of the kernel statistic, the Maximum Mean Discrepancy (MMD), which is our notion of similarity. This answers Research Question 1, which analyzes the relative similarity between probability distributions in different settings. Our main contribution is a test of relative similarity with $\mathcal{O}(n^{-1/2})$ convergence and $\mathcal{O}(n^2)$ computation that is consistent under mild conditions on separable topological domains enriched with kernels. The content of this work is based on the following publication

- W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *The 4th International Conference on Learning Representations*, 2016a.

In Chapter 4 of this thesis, we present a novel non-parametric test of relative dependence. This chapter directly applies the U-statistic asymptotic distribution theorem on our notion of dependence, the Hilbert-Schmidt Independent Criterion. This answers Research Question 2, which analyzes the relative dependence for different outputs in case of multiple dependencies. Again, our main contribution prove that this scheme is consistent in the kernel independent test setup under mild conditions on separable topological domains enriched with kernels. The content of this work is based on the following publication

- W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. A low variance consistent test of relative dependency. In F. Bach and D. Blei, editors, *Proceedings of*

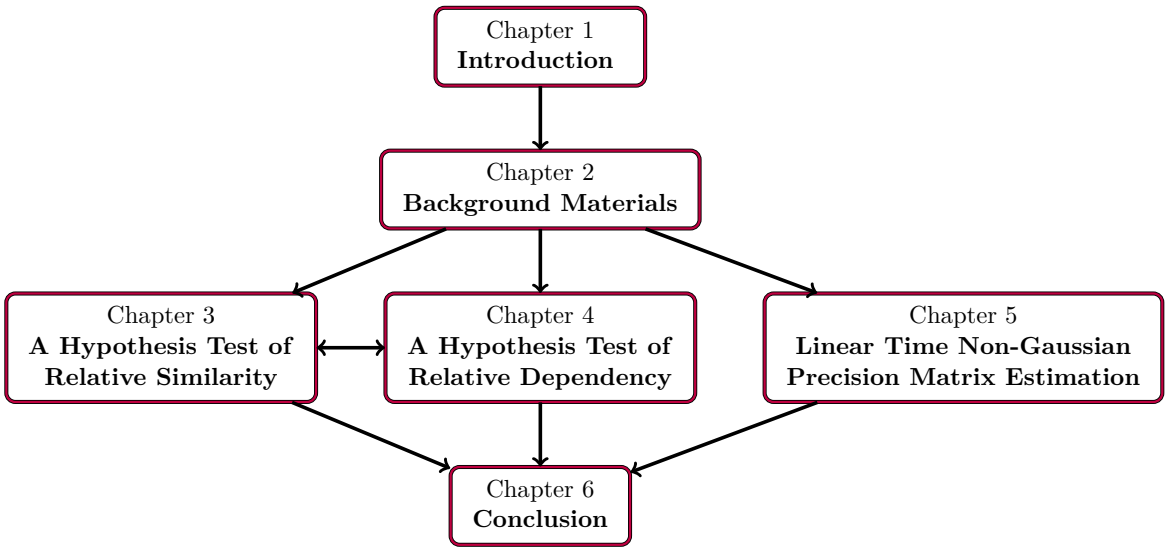


Figure 1.4 – The outline of this thesis.

Setting	Statistical Convergence Rate	Computation Time	Chapter
$\widehat{\mathrm{MMD}}_u^2 [\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}] - \widehat{\mathrm{MMD}}_u^2 [\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^2)$	3
$\widehat{\mathrm{HSIC}}_u [\mathcal{H}_\mathbf{x}, \mathcal{H}_\mathbf{y}, (\mathbf{X_n}, \mathbf{Y_n})] - \widehat{\mathrm{HSIC}}_u [\mathcal{H}_\mathbf{x}, \mathcal{H}_\mathbf{z}, (\mathbf{X_n}, \mathbf{Z_n})]$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^2)$	4
$\mathrm{Cov}\left(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}\right)$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n)$	5

Table 1.1 – Key theoretical contributions.

The 32nd International Conference on Machine Learning, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 20–29, 2015a.

Chapter 5 presents a study of edge detection in an undirected graphical model. We construct a new non-parametric linear time test for precision matrix estimation. The contributions of this part is a direct application of a U-statistic asymptotic distribution theorem on the covariance matrix. This chapter gives a positive answer to Research Question 3 by showing the effectiveness of the statistical test with millions of observations from non-Gaussian distributions. The content of this work is based on the following publication

- W. Bounliphone and M. B. Blaschko. Linear time non-Gaussian precision matrix estimation. 2016. arXiv:1604.01733 – under submission.

Chapter 6 concludes the thesis and revisits the Research Questions devised in Section 1.2. An outline of the thesis with dependencies between chapters is illustrated in Figure 1.4. Key theoretical contributions developed in this thesis are summarized in Table 1.1.

Chapter 2

Background Materials

The work presented in this thesis is related to a variety problems in statistical estimation and computational efficiency. In this chapter we introduce essential background knowledge necessary for the development of our later theory. We start by introducing the class of U-statistic, which allows a minimum-variance unbiased estimation of a parameter. Then, we introduce the kernels methods and the concept of reproducing kernel Hilbert space (RKHS). Afterward, we present the two statistics, the kernel Maximum Mean Discrepancy (MMD) and the Hilbert-Schmidt Independence Criterion (HSIC), which are used in the two novel non-parametric statistical hypothesis tests in Chapters 3 and 4. Subsequently, we introduce basic results for the estimation of the precision matrix.

Contents

2.1 U-statistics Estimator

When we assume the existence of a simple random sample $\mathbf{X_r} = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$, U-statistics generalize common notions of unbiased estimation. In fact, the letter "U" in U-statistics stands for "unbiased." In this section, we study their properties: unbiased, minimum variance, concentration, asymptotic variance, asymptotic covariance and asymptotic distribution. The basic theory of U-statistics can be found in Hoeffding [1948], Lee [1990], Lehmann [1999, Ch. 6] and Serfling [2009, Ch. 5].

Suppose that we have a sample $\mathbf{X_r} = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ of size r drawn independently and identically distributed (i.i.d.) from $\mathbb{P}_{\mathbf{x}}$. U-statistics concern an unbiased and minimum variance estimation of a parameter θ of $\mathbb{P}_{\mathbf{x}}$ using $\mathbf{X}_r$. That is, θ may be represented as

$$\theta = \mathbb{E}_{\mathbf{x}} \left[h \left(\mathbf{x}_1, \ldots, \mathbf{x}_r \right) \right], \tag{2.1.1}$$

for some function h, called a kernel[1] of the estimator. The smallest integer r for which Equation (2.1.1) holds is called the degree of θ. Without loss of generality we may assume that h is symmetric. If not, h also satisfies Equation (2.1.1) for any permutation $(i_1, \ldots, i_r)$ of the set $\{1, \ldots r\}$ and therefore so does the symmetric kernel

$$\frac{1}{r!} \sum_{\pi^r} h \left(\mathbf{x}_{i_1}, \ldots \mathbf{x}_{i_r} \right), \tag{2.1.2}$$

where the summation extends over all $r!$ permutations π^r of the set $\{1, \ldots r\}$.

We now turn to the estimation of θ by means of the n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where we shall assume that $n \geq r$. Clearly, $h \left(\mathbf{x}_1, \ldots, \mathbf{x}_r \right)$ is an unbiased estimator of θ and so is $h \left(\mathbf{x}_{i_1} \ldots \mathbf{x}_{i_r} \right)$ for any r-tuples drawn without replacement from the set $\{1, \ldots n\}$. Then, for any kernel, the corresponding U-statistic for estimation of θ using a sample $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of size n larger than r is constructed in the following way:

Definition 2.1. (U-statistic, [Serfling, 2009, Section 5.1.1]). *Given a kernel h of degree r and a sample $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of size $n \geq r$, the corresponding U-statistic for estimation θ is obtained by averaging the kernel h symmetrically over the observations*

$$U_n := \frac{1}{(n)_r} \sum_{i_r^n} h \left(\mathbf{x}_{i_1} \ldots \mathbf{x}_{i_r} \right), \tag{2.1.3}$$

where $(n)_r = \frac{n!}{(n-r)!} = r! \binom{n}{r}$ is the Pochhammer symbol and where the summation is over the set i_r^n of all $\frac{n!}{(n-r)!}$ r-tuples drawn without replacement from the set $\{1, \ldots n\}$.

Note that in this definition, we do not require the kernel h to be symmetric in its arguments. If the kernel h is symmetric in its arguments, we may without loss of generality restrict attention to the cases in which $1 \leq i_1 < \cdots < i_r \leq n$. In that case, we can drop the normalization by

[1] Note that the term kernel in U-statistics has a different meaning than in kernel methods in the machine learning community.

$r!$ due to the symmetrization and then, U_n has the equivalent form

$$U_n = \frac{1}{\binom{n}{r}} \sum h\left(\mathbf{x}_{i_1} \dots \mathbf{x}_{i_r}\right), \tag{2.1.4}$$

where the summation is over all combinations $\binom{n}{r}$ of r distinct elements, $1 \leq i_1 < \cdots < i_r \leq n$, drawn without replacement from the set $\{1, \dots n\}$.

We now introduce the variance and covariance of U-statistics and with this characterization, we can examine the asymptotic distribution of U-statistics.

Theorem 2.1. (Variance of a U-statistic, [Serfling, 2009, Section 5.2.1]). *The variance of the U-statistic given in Equation (2.1.3) is equal to*

$$\operatorname{Var} U_n = \binom{n}{r}^{-1} \sum_{k=1}^{r} \binom{r}{k} \binom{n-r}{r-k} \zeta_i, \tag{2.1.5}$$

where

$$\zeta_i = \operatorname{Var}\left(\mathbb{E}_{\mathbf{x}_{i+1},\dots,\mathbf{x}_r}\left[h\left(\mathbf{x}_1, \dots, \mathbf{x}_i, \mathbf{x}_{i+1}, \dots, \mathbf{x}_r\right)\right]\right), \tag{2.1.6}$$

with $\mathbb{E}_{x_{i+1},\dots,x_r}$ denotes the integral of the function $h\left(\mathbf{x}_1, \dots, \mathbf{x}_i, \mathbf{x}_{i+1}, \dots, \mathbf{x}_r\right)$ with respect to the variables of integration, $\mathbf{x}_{i+1}, \dots, \mathbf{x}_r$. If the distribution $\mathbb{P}_{\mathbf{x}}$ has a density f then

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_{i+1},\dots,\mathbf{x}_r}\left[h\left(\mathbf{x}_1, \dots, \mathbf{x}_i, \mathbf{x}_{i+1}, \dots, \mathbf{x}_r\right)\right] \\ &\quad = \int_{\mathbb{R}^{r-i}} h\left(\mathbf{x}_1, \dots, \mathbf{x}_i, \mathbf{x}_{i+1}, \dots, \mathbf{x}_r\right) f\left(\mathbf{x}_{i+1}, \dots, \mathbf{x}_r\right) \mathrm{d}\mathbf{x}_{i+1} \dots \mathrm{d}\mathbf{x}_r. \end{aligned} \tag{2.1.7}$$

Remark 2.1. *Depending on whether ζ_1 is vanishing or not, the asymptotic distribution is different. For $\zeta_1 > 0$, the asymptotic distribution is Gaussian, while for $\zeta_1 = 0$, the asymptotic distribution is an infinite sum of $\chi^2(1)$ random variables. In this thesis, we only have the case where $\zeta_1 > 0$.*

Theorem 2.2. (Asymptotic distribution of a U-statistic, [Serfling, 2009, Theorem. A p. 192]). *If $\mathbb{E}\left[h^2\right] < \infty$ and $\zeta_1 > 0$, then as $n \to \infty$, U_n is asymptotically normal with mean θ and variance $\frac{r^2}{n}\zeta_1$*

$$n^{1/2}\left(U_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, r^2\zeta_1\right). \tag{2.1.8}$$

Theorem 2.3. (Covariance of two U-statistics, [Serfling, 2009, Section 6]). *Let $U_n^{(1)}$ and $U_n^{(2)}$ be two U-statistics, both based on a common sample $\mathbf{X_n} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ but having different kernels h and g of degrees r_1 and r_2 respectively, with $r_1 \leq r_2$. Then the covariance of two U-statistics is equal to*

$$\operatorname{Cov}\left(U_n^{(1)}, U_n^{(2)}\right) = \binom{n}{r_1}^{-1} \sum_{k=1}^{r_1} \binom{r_2}{k} \binom{n-r_2}{r_1-k} \zeta_i, \tag{2.1.9}$$

where

$$\zeta_i = \mathrm{Cov}\left(\mathbb{E}_{\mathbf{x}_{i+1},\ldots,\mathbf{x}_{r_1}}\left[h\left(\mathbf{x}_1,\ldots,\mathbf{x}_i,\mathbf{x}_{i+1},\ldots,\mathbf{x}_{r_1}\right)\right];\right. \\ \left.\mathbb{E}_{\mathbf{x}_{j+1},\ldots,\mathbf{x}_{r_2}}\left[g\left(\mathbf{x}_1,\ldots,\mathbf{x}_j,\mathbf{x}_{j+1},\ldots,\mathbf{x}_{r_2}\right)\right]\right). \tag{2.1.10}$$

Theorem 2.4. (Joint asymptotic distribution of U-statistics, [Hoeffding, 1948, Theorem. 7.1]). *Let $U_n^{(j)}$, $j = 1,\ldots,m$, be U-statistics having expectations θ_j and kernels $h^{(j)}$ of degrees r_j and let denote by $\mathbf{U_n}$ and $\boldsymbol{\theta}$ the m-vector $\left(U_n^{(1)},\ldots,U_n^{(m)}\right)^T$ and $(\theta_1,\ldots,\theta_m)^T$, respectively. If $\mathrm{Var}\, h^{(j)}\left(\mathbf{x}_1,\ldots,\mathbf{x}_{r_j}\right) < \infty$ for all $j = 1,\ldots,m$, then $\mathbf{U}_n$ converges asymptotically in distribution to a multivariate normal distribution with mean vector zero and $\boldsymbol{\Sigma}$ the limiting covariance matrix of $n^{1/2}(U_n^{(j)} - \theta_j)$, the entries of which are given by Equations (2.1.5) and (2.1.9), provided $\boldsymbol{\Sigma}$ is positive definite:*

$$n^{1/2}\left(\mathbf{U_n} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}_m\left(\mathbf{0}, \boldsymbol{\Sigma}\right). \tag{2.1.11}$$

Applications of U-statistics have been used in many domains in high dimensional statistical inference and estimation, in many different hypothesis tests, feature selection, and the estimation of high dimensional graphical models [Callaert and Janssen, 1978, Chang et al., 2016, Chen and Shao, 2007].

2.2 Kernel Methods

In this section, we introduce concepts and notation required to understand reproducing kernel Hilbert spaces. There is a significant interest in these methods from the statistics and mathematics communities. A more detailed account of this topic can be found in Aronszajn [1950], Schölkopf and Smola [2001, Ch. 1] and Berlinet and Thomas-Agnan [2011, Ch. 1], for example. For basic definitions of Hilbert spaces and their applications, we refer to Dieudonné [1960, Ch. 6] and Rudin [1987, Ch. 4].

2.2.1 Definitions and Properties of Kernels

Many machine learning algorithms can be expressed in terms of inner products between observations, $\langle \mathbf{x}, \mathbf{x}' \rangle$, or inner products between matrix structured observations, $\mathbf{X}\mathbf{X}'^T$. An inner product can be interpreted as a measure of similarity between $\mathbf{x}$ and $\mathbf{x}'$. However, the class of linear functions induced by this inner product may be too restrictive for many real-world problems. Kernel methods aim to build more flexible and powerful tools by replacing $\langle \mathbf{x}, \mathbf{x}' \rangle$ with some other, non-linear similarity measures. This is the so-called *kernel trick*: wherever inner product are used, they are replaced by kernel functions. We now introduce formally kernel functions and their properties.

Definition 2.2. (Kernel function) *Given a general set $\mathcal{X}$ and two observations $\mathbf{x}$ and $\mathbf{x}'$,*

a kernel function (or kernel) is an inner product $\langle ., . \rangle_{\mathcal{H}}$ *in a feature space* $\mathcal{H}$

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (\mathbf{x}, \mathbf{x}') &\longmapsto k(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}, \end{aligned} \tag{2.2.1}$$

where $\phi(\mathbf{x}) : \mathcal{X} \longrightarrow \mathcal{H}$*, called the feature map, is a nonlinear mapping from an observation to its feature space representation.*

Likewise, we can interpret $k(\mathbf{x}, \mathbf{x}')$ as a non-linear similarity measure between $\mathbf{x}$ and $\mathbf{x}'$ by substituting $\langle \mathbf{x}, \mathbf{x}' \rangle$ with $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$. The machine learning algorithms remain the same, we only change the space in which these algorithms operate. Note that the feature map $\phi(\mathbf{x})$ is not unique and need not be known explicitly.

The natural question is to find a full characterization of functions that are kernels. We start with basic definitions and results.

Definition 2.3 (Gram Matrix)**.** *Given a symmetric function* $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ *and observations* $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq \mathcal{X}$*, then the real symmetric* $n \times n$ *matrix* $\mathbf{K}$ *with elements*

$$[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{2.2.2}$$

is called the Gram matrix of k *with respect to* $\mathbf{x}$*.*

Definition 2.4 (Positive Semi-Definite Matrix)**.** *A symmetric function* $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ *is a positive definite kernel (or a kernel for simplicity), if for any finite set of observations it gives rise to a positive definite Gram matrix.*

Furthermore, kernels have a set of closure properties, meaning that certain operations on kernels still yield kernels. These closure operations allow us to create new kernels. Primitive closure operations are:

1. Positive linear combination: If k_1 and k_2 are kernels, and $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 k_1(\mathbf{x}, \mathbf{x}') + \alpha_2 k_2(\mathbf{x}, \mathbf{x}')$ is a kernel.
2. Non-negativity: If k is a kernel, and $b \geq 0$, then $k(\mathbf{x}, \mathbf{x}') + b$ is a kernel.
3. Tensor product: If k_1 and k_2 are kernels, then $k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}')$ is a kernel.

Example 2.1. *Based on these primitives, we can derived more complicated closure operations. The following functions are well-known examples of kernels for* $\mathcal{X} = \mathbb{R}$

- *Linear kernel:* $\langle \mathbf{x}, \mathbf{x}' \rangle$*.*
- *Gaussian kernel with bandwidth* $\sigma > 0$*:* $\exp(-\sigma \|\mathbf{x} - \mathbf{x}'\|^2)$*.*
- *Laplace kernel with bandwidth* $\sigma > 0$*:* $\exp(-\sigma \|\mathbf{x} - \mathbf{x}'\|)$*.*
- *polynomial kernel with* $c \geq 0$ *and degree* $p \in \mathbb{N}$*:* $(\langle \mathbf{x}, \mathbf{x}' + c \rangle)^p$*.*

All these examples are kernels on $\mathbb{R}^p$, which are the most common in the literature, but there exist kernels on many other domains, too, for instance graphs, strings, images, etc. [Shawe-Taylor and Cristianini, 2004].

2.2.2 Reproducing Kernel Hilbert Spaces

We now develop a characterization of kernel functions by constructing a specific Hilbert space, specifying both its topology and inner product. A Hilbert space $\mathcal{H}$ is a real or complex inner product space that is also a complete metric space with respect to the distance function induced by the inner product. We now give the definition of a reproducing kernel Hilbert space.

Definition 2.5 (Reproducing Kernel Hilbert Spaces, RHKS)**.** *Given a set $\mathcal{X}$ and a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \longrightarrow \mathbb{B}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space induced by the inner product $\langle \cdot, \cdot \rangle$ if there exists a function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ with the following properties*

1. *k has the reproducing property*

$$\langle f, k(\mathbf{x}, \cdot) \rangle = f(\mathbf{x}) \forall f \in \mathcal{H}; \tag{2.2.3}$$

   in particular,

$$\langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot) \rangle = k(\mathbf{x}, \mathbf{x}'). \tag{2.2.4}$$

2. *The RKHS is simply the Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \longrightarrow \mathbb{R}$ spanned by the kernel $k(\mathbf{x}, \cdot)$, i.e. $\mathcal{H} = \overline{\text{span}\{k(\mathbf{x}, \cdot) \,|\mathbf{x} \in \mathcal{X}\}}$.*

The RKHS $\mathcal{H}$ is fully characterized by the reproducing kernel k and vice versa, as stated in the following theorem:

Theorem 2.5. *For every positive definite function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$, there exists a unique RKHS with k as its reproducing kernel.*

Kernel methods are widely applied in the problem of supervised learning. Early examples include kernel PCA [Schölkopf et al., 1997], kernel ICA [Bach and Jordan, 2002], and kernel dimensionality reduction [Fukumizu et al., 2009]. They are also well established in the areas of estimation, analysis of probability distributions, and hypothesis testing. In the next section, we present two statistics based on kernel methods that we will use in Chapters 3 and 4.

2.3 The Maximum Mean Discrepancy

In this section, we introduce a kernel method for comparing samples from two probability distributions. A statistical test for the *two-sample problem* determines with significance if

two samples are drawn from different distributions. The test we consider in this thesis consists in maximizing the mean discrepancy between probabilistic distributions. We explain a framework for the distribution analysis via the kernel mean embedding: each distribution is mapped into a RKHS via an expectation operation. Then, we define the measure of discrepancy between distribution functions in term of their kernel mean embeddings. Most of the theory follows from Berlinet and Thomas-Agnan [2011, Ch. 4], Gretton et al. [2006], Smola et al. [2007] and Gretton et al. [2012a].

For the formal setup of this section, suppose that we have random variables $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$ and $\mathbf{y} \sim \mathbb{P}_{\mathbf{y}}$ that take values in $(\mathcal{X}, \mathcal{B}_{\mathbf{x}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathbf{y}})$, respectively, here $\mathcal{X}$ and $\mathcal{Y}$ are two separable metrics and $\mathcal{B}_{\mathbf{x}}$ and $\mathcal{B}_{\mathbf{y}}$ are Borel σ-algebras. We want to determine whether $\mathbb{P}_{\mathbf{X}} \neq \mathbb{P}_{\mathbf{Y}}$. Furthermore, we define the positive definite kernel $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ associated with a RKHS $\mathcal{H}$.

2.3.1 Hilbert Space Embedding of Distributions

We start by extending the notion of feature maps to the embedding of a probability distribution. We first provide the conditions under which the embedding $\mu_{\mathbf{x}}$ exists and belongs to $\mathcal{H}$. The existence and uniqueness of the mean embedding in the RKHS is guaranteed by Riesz representation theorem [Fréchet, 1907, Riesz, 1907]. Then, we use the following lemma to define the embedding.

Lemma 2.1. *If the kernel k is measurable and $\mathbb{E}_{\mathbf{x}}\sqrt{k(\mathbf{x},\mathbf{x})} < \infty$, where $\mathbf{x}$ is a random variable with distribution $\mathbb{P}_{\mathbf{x}}$, then there exists $\mu_{\mathbf{x}} \in \mathcal{H}$ such that*

$$\mathbb{E}_{\mathbf{x}} f(\mathbf{x}) = \langle f, \mu_{\mathbf{x}} \rangle, \forall f \in \mathcal{H}. \tag{2.3.1}$$

Definition 2.6 (Kernel Mean Embedding)**.** *Suppose that the space $\mathcal{M}_+(\mathcal{X})$ consists of all probability measures $\mathbb{P}_{\mathbf{x}}$ on a measurable space $(\Omega, \mathcal{X})$. The kernel mean embedding of probability measures in $\mathcal{M}_+(\mathcal{X})$ into an RKHS $\mathcal{H}$ induced by the kernel $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$, is defined by a mapping*

$$\begin{aligned} \mu_{\mathbf{x}} : \mathcal{M}_+ &\longrightarrow \mathcal{H} \\ \mathbb{P}_{\mathbf{x}} &\longmapsto \int k(\mathbf{x}, \cdot)\, d\mathbb{P}_{\mathbf{x}}(\mathbf{x}) = \mu_{\mathbf{x}} = \mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \cdot). \end{aligned} \tag{2.3.2}$$

Then, in the next lemma, we provide the existence condition of the mean embedding $\mu_{\mathbf{x}}$ into an RKHS $\mathcal{H}$.

Lemma 2.2 (Smola et al. [2007])**.** *If $\mathbb{E}_{\mathbf{x}}\sqrt{k(\mathbf{x},\mathbf{x})} < \infty$, $\mu_{\mathbf{x}} \in \mathcal{H}$ and $\mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \cdot) = \langle f, \mu_{\mathbf{x}} \rangle_{\mathcal{H}}$.*

The mean embedding is a very powerful tool as it gives us a non-parametric representation of a distribution as an element in a Hilbert space, which we can use in statistical methods. A schematic view of the kernel embedding and the mean element is illustrated in Figure 2.1. In practice, we do not have access to the true distribution $\mathbb{P}_{\mathbf{x}}$. Instead, we use the sample from

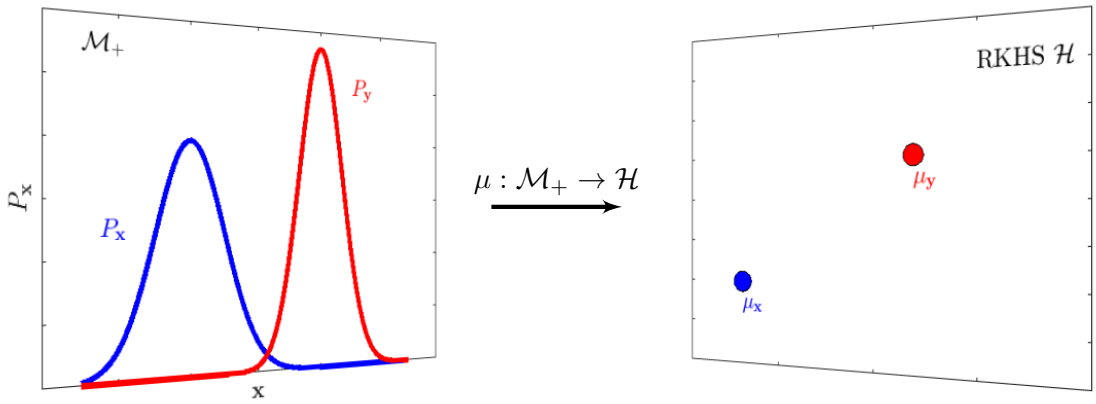


Figure 2.1 – Illustration of kernel embedding and the mean elements of two distributions.

this distribution. Given an i.i.d. sample $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the empirical estimate $\hat{\mu}_\mathbf{x}$ of the mean embedding $\mu_\mathbf{x}$ is

$$\hat{\mu}_\mathbf{x} = \frac{1}{n} \sum_{i=1}^{n} k\left(\mathbf{x}_i, .\right). \tag{2.3.3}$$

Clearly, $\hat{\mu}_\mathbf{x}$ is unbiased and $\hat{\mu}_\mathbf{x} \xrightarrow{d} \mu_\mathbf{x}$.

2.3.2 Universal and Characteristic Kernels

The notions of universal and characteristic kernels play crucial roles in the analysis of kernel mean embeddings. We now formally introduce the notion of universal and characteristic kernels. For more detail, we refer to Fukumizu et al. [2007], Gretton et al. [2006], Steinwart [2001] and Sriperumbudur et al. [2011].

Definition 2.7 (Universal kernel, [Steinwart, 2001])**.** *The kernel k on $\mathcal{X}$ is called universal if $k(\mathbf{x}, \cdot)$ is continuous for all $\mathbf{x}$ and the corresponding RKHS $\mathcal{H}$ induced by k is dense in $\mathcal{C}(\mathcal{X})$, a space of bounded continuous functions on $\mathcal{X}$.*

Definition 2.8 (Characteristic kernel, [Fukumizu et al., 2007])**.** *The kernel k is called characteristic if the map $\mathbb{P}_\mathbf{x} \mapsto \mathbb{E}_\mathbf{x} k(\mathbf{x}, \cdot)$ is injective and the RKHS $\mathcal{H}$ is called characteristic if its reproducing kernel is characteristic.*

For kernels which are universal / characteristic (the notions are identical for finite signed Borel measures [Sriperumbudur et al., 2011]), such as the Gaussian and Laplace kernels, the mean embedding operator is injective, so if we have samples from two distributions $\mathbf{x} \sim \mathbb{P}_\mathbf{x}$ and $\mathbf{y} \sim \mathbb{P}_\mathbf{y}$ then $\mu_\mathbf{x} = \mu_\mathbf{y} \iff \mathbb{P}_\mathbf{x} = \mathbb{P}_\mathbf{y}$. This is the basis for the Maximum Mean Discrepancy (MMD) test statistic [Gretton et al., 2012a], which measures $\|\mu_\mathbf{x} - \mu_\mathbf{y}\|$. We present the MMD in the next section.

2.3.3 Definition and Properties of MMD

As we have seen, kernel embeddings of Borel probability measures in $\mathcal{M}_+$ exist, and we can introduce the notion of distance between Borel probability measures in this set using the Hilbert space distance between their embeddings. Let $\mathbf{x}$ and $\mathbf{y}$ be random variables defined on a topological space $\mathcal{X}$, with respective Borel probability measures $\mathbb{P}_\mathbf{x}$ and $\mathbb{P}_\mathbf{y}$. Given observations $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{Y_m} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ of these variables, obtained i.i.d. from $\mathbb{P}_\mathbf{x}$ and $\mathbb{P}_\mathbf{y}$, we want to determine whether $\mathbb{P}_\mathbf{x} \neq \mathbb{P}_\mathbf{y}$.

Lemma 2.3. (Dudley [2002, Lemma 9.3.2]). *The Borel probability measures $\mathbb{P}_\mathbf{x}$ and $\mathbb{P}_\mathbf{y}$ are equal if and only if $\mathbb{E}_\mathbf{x} f(\mathbf{x}) = \mathbb{E}_\mathbf{y} f(\mathbf{y})$, $\forall f \in \mathcal{C}(\mathcal{X})$.*

Definition 2.9. Maximun Mean Discrepancy – MMD, [Gretton et al., 2012a, Lemma 4]. *Assume that the mean embeddings $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ exist then*

$$\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}] = \|\mu_\mathbf{x} - \mu_\mathbf{y}\|^2_\mathcal{H}. \tag{2.3.4}$$

The following alternative representation of the squared MMD will be useful.

Lemma 2.4. (Gretton et al. [2012a, Lemma 6]). *Let $\mathcal{H}$ be an RKHS, with the continuous feature mapping $\phi(\mathbf{x}) \in \mathcal{H}$ from each $\mathbf{x} \in \mathcal{X}$, such that the inner product between the features is given by the kernel function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$. Then the squared population* MMD *is given by*

$$\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{x}'}\left[k(\mathbf{x}, \mathbf{x}'))\right] - 2\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[k(\mathbf{x}, \mathbf{y})\right] + \mathbb{E}_{\mathbf{y},\mathbf{y}'}\left[k(\mathbf{y}, \mathbf{y}')\right], \tag{2.3.5}$$

where $\mathbf{x}$, $\mathbf{x}'$ and $\mathbf{y}$, $\mathbf{y}'$ are obtained i.i.d. from $\mathbb{P}_\mathbf{x}$ and $\mathbb{P}_\mathbf{y}$, respectively.

A schematic view of the kernel embedding and the mean element is illustrated in Figure 2.2. This quantity can, in general, be estimated using empirical expectations.

Lemma 2.5. (Gretton et al. [2012a, Lemma 6]). *Given i.i.d. samples $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{Y_m} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ from $\mathbb{P}_\mathbf{x}$ and $\mathbb{P}_\mathbf{y}$ respectively, an unbiased empirical estimate of $\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}]$ can be written in terms of k as*

$$\begin{aligned}\widehat{\mathrm{MMD}^2_u}[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_m}] = &\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(\mathbf{x}_i, \mathbf{x}_j) \\ &+ \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{n} k(\mathbf{y}_i, \mathbf{y}_j) \\ &- \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(\mathbf{x}_i, \mathbf{y}_j).\end{aligned} \tag{2.3.6}$$

Furthermore, let $m = n$ and $\mathbf{v_n} = (\mathbf{v}_1, ..., \mathbf{v}_n)$ be n i.i.d. random variables where $\mathbf{v} := (\mathbf{x}, \mathbf{y}) \sim \mathbb{P}_\mathbf{x} \times \mathbb{P}_\mathbf{y}$, then Equation (2.3.6) is an unbiased estimate which is a sum of two U-statistics and

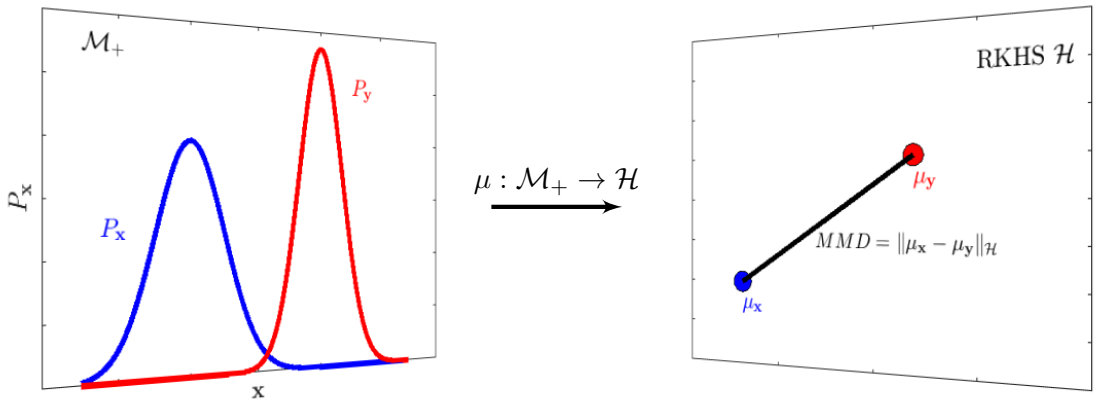


Figure 2.2 – Illustration of the Maximum Mean Discrepancy in the RKHS $\mathcal{H}$.

a sample average. That is

$$\widehat{\mathrm{MMD}}_u^2 \left[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}\right] = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} h\left(\mathbf{v}_i, \mathbf{v}_j\right), \tag{2.3.7}$$

where h is the U-statistic kernel of degree 2 such that

$$h\left(\mathbf{v}_i, \mathbf{v}_j\right) = k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{y}_i, \mathbf{y}_j) - k(\mathbf{x}_i, \mathbf{y}_j) - k(\mathbf{x}_j, \mathbf{y}_i), \tag{2.3.8}$$

and we assume that $\mathbb{E}\left[h\right] < \infty$.

2.3.4 Application to the Two-Sample Test

We have described a metric on probability distributions based on the difference of their Hilbert-Schmidt embeddings and its empirical estimate. In this section, we introduce a statistical two-sample test. The goal of a two-sample test is to decide whether a distribution $\mathbb{P}_{\mathbf{x}}$ is different from $\mathbb{P}_{\mathbf{y}}$ with statistical significance on the basis of the samples. The task is formulated as a statistical hypothesis test. Following Casella and Berger [2002, Ch. 8], we briefly introduce the framework of statistical hypothesis testing applied in the two-sample context. Given i.i.d. samples $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{Y_n} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ of size n drawn from $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{y}}$, respectively, the statistical test, $\mathcal{T} : \mathcal{X}^n \times \mathcal{Y}^n \mapsto \{0, 1\}$ is used to distinguished between the null hypothesis $\mathcal{H}_0 : \mathbb{P}_{\mathbf{x}} = \mathbb{P}_{\mathbf{y}}$ and the alternative hypothesis $\mathcal{H}_1 : \mathbb{P}_{\mathbf{x}} \neq \mathbb{P}_{\mathbf{y}}$. This is achieved by comparing the test statistic $\widehat{\mathrm{MMD}}_u^2 \left[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}\right]$ with a particular threshold: if the threshold is exceeded, then the test rejects the null hypothesis. The acceptance region of the test is defined as any real number below the threshold. Since the test is based on a

finite sample, the notion of statistical error is an integral part of hypothesis testing. A Type I error occurs when the null hypothesis is true, but is rejected. Conversely, a Type II error occurs when the null hypothesis is false, but erroneously fails to be rejected. The level α of a test is the probability of rejecting the null hypothesis when it is true. This α is a design parameter of the test and is an upper bound on the Type I error, used to set the threshold. Gretton et al. [2012a] proposed two statistical approaches for the two-sample problem. The first is based the McDiarmid [1989] or the Hoeffding [1963] concentration inequalities on the $\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]$ statistic and the second is based on the asymptotic distribution of the unbiased estimate of MMD^2. The approach we adopt here is the second: we show that MMD is asymptotically normally distributed, and we derive its variance to formulate statistics for a significant test.

Theorem 2.6. (Asymptotic distribution of the MMD [Gretton et al., 2012a, Theorem 19])**.** *If $\mathbb{E}[h] < \infty$, given i.i.d. samples $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{Y_n} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ drawn from $\mathbb{P}_\mathbf{x}$ and $\mathbb{P}_\mathbf{y}$. When $\mathbb{P}_\mathbf{x} \neq \mathbb{P}_\mathbf{y}$, an unbiased estimate of $\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}]$ given in Equation (2.4) converges asymptotically in distribution to a Gaussian distribution with mean zero and variance $\sigma^2_{\mathrm{MMD}^2_\mathbf{xy}}$*

$$n^{1/2}\left(\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}] - \widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathrm{MMD}^2_\mathbf{xy}}\right), \tag{2.3.9}$$

where

$$\sigma^2_{\mathrm{MMD}^2_\mathbf{XY}} = 4\left(\mathbb{E}_{\mathbf{v}_1}\left[\left(\mathbb{E}_{\mathbf{v}_2} h(\mathbf{v}_1, \mathbf{v}_2)\right)^2\right] - \left[\mathbb{E}_{\mathbf{v}_1, \mathbf{v}_2} h(\mathbf{v}_1, \mathbf{v}_2)\right]^2\right), \tag{2.3.10}$$

where $\mathbf{v}_i = (\mathbf{x}_i, \mathbf{y}_i)$.

When $\mathbb{P}_\mathbf{x} = \mathbb{P}_\mathbf{y}$, the U-statistic is degenerate (cf. Remark 2.1), we have $\mathbb{E}_{\mathbf{v}_2} h(\mathbf{v}_1, \mathbf{v}_2) = 0$. In that case, $\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}]$ converges in distribution according to

$$n\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}] \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l \{z_l^2 - 2\}, \tag{2.3.11}$$

where $z_l \sim \mathcal{N}(0, 2)$ i.i.d. and λ_l are the solution to the eigenvalue equation,

$$\int_\mathcal{X} \tilde{k}(\mathbf{x}, \mathbf{x}')\psi_i(\mathbf{x}) d\mathbb{P}_\mathbf{x} = \psi_i(\mathbf{x}'), \tag{2.3.12}$$

with $\tilde{k}(\mathbf{x}, \mathbf{x}') := k(\mathbf{x}, \mathbf{x}') - \mathbb{E}_\mathbf{x} k(\mathbf{x}, \mathbf{x}_i) - \mathbb{E}_\mathbf{x} k(\mathbf{x}, \mathbf{x}_j) + \mathbb{E}_{\mathbf{x},\mathbf{x}} k(\mathbf{x}_i, \mathbf{x}_j)$ is the centered RKHS kernel.

We then use the $1-\alpha$ quantile of this distribution as the test threshold. Since the distribution under $\mathcal{H}_0$ is complicated, we need to accurately approximate its quantile. There are two ways to estimate this quantile. The first is to use the bootstrap on the aggregated data and the second is to approximate the null distribution by fitting the Pearson curve to its first four moments.

Furthermore, the MMD has been applied in many applications such as hypothesis testing

[Chwialkowski et al., 2015, Zaremba et al., 2013], clustering, density estimation, covariate shift [Zhang et al., 2013], and generative models [Bounliphone et al., 2016a].

2.4 The Hilbert-Schmidt Independence Criterion

There are different methods to measure the dependence or association between random variables. A random variable is said to be dependent on another random variable if its variations can be (partially) explained by those of another. Measuring statistical dependence is an important tool in statistical analysis, and is widely applied in many data analysis contexts. For instance, dependence measures can help answer questions such as whether the price of one stock is linked to another, whether students examination results are connected to one traditional teaching method or to another; or whether the one micro-climate is synchronized with another. A large variety of dependence concepts have been studied by a number of authors, an overview of which can be found in Joe [1997]. Classical criteria include Pearson's linear correlation, Spearman's ρ, Kendall's τ, the RV coefficient and the mutual information [Casella and Berger, 2002, Escoufier, 1973]. More recent research on dependence measures has devoted considerable interest to non-parametric measures of dependence using criteria based on functions in RKHSs. This has been applied even when the dependence is non-linear, or the variables are non-euclidean (for instance images, graphs and strings). This was first accomplished by Bach and Jordan [2002], who introduced a kernel dependence functional using a regularized estimate of the spectral norm of the correlation operator between two RKHSs. Then, Gretton et al. [2005b] employed the covariance operator instead of the correlation operator. Others statistics for test of dependence are diverse and include kernel measures of covariance [Gretton et al., 2008, Zhang et al., 2011] and correlation [Dauxois and Nkiet, 1998, Gretton et al., 2008], distance covariances (which are instances of kernel tests) [Sejdinovic et al., 2013, Székely et al., 2007], kernel regression tests [Cortes et al., 2009, Gunn and Kandola, 2002], rankings [Heller et al., 2013], and space partitioning approaches [Gretton and Gyorfi, 2010, Kinney and Atwal, 2014, Reshef et al., 2011]. Specialization of such methods to univariate linear dependence can yield similar tests to classical approaches such as Bring [1996], Darlington [1968].

In our work, we choose the Hilbert-Schmidt Independence Criterion (HSIC) as a measure of dependency. The HSIC is a kernel based method to detect dependence between random variables: both the joint probability measures and the product of the marginal distribution are mapped into a infinite-dimensional feature space. We introduce HSIC in two different ways: first, we define it as the Hilbert-Schmidt norm of the cross-covariance operator [Gretton et al., 2005b] and then as a special cased of the Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a].

For the formal setup of this section, suppose that we have random variables $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$ and $\mathbf{y} \sim \mathbb{P}_{\mathbf{y}}$, that take values on $(\mathcal{X}, \mathcal{B}_{\mathbf{x}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathbf{y}})$, respectively. Here $\mathcal{X}$, $\mathcal{Y}$ are two separable metrics and $\mathcal{B}_{\mathbf{x}}, \mathcal{B}_{\mathbf{y}}$ are Borel σ-algebras. Then, $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{y}})$ is again measurable and the joint distribution is $\mathbb{P}_{\mathbf{xy}}$, which assigns values to the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{y}})$.

Empirical samples from $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{y}})$ are assumed to be of size n. We have that $\mathbf{x}$ and $\mathbf{y}$ are independent if and only if $\mathbb{P}_{\mathbf{xy}} = \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$.

Furthermore, we define kernels $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ on the space $\mathcal{X}$ and $\mathcal{Y}$, with the corresponding feature maps ϕ and φ and denote the corresponding RKHSs by $\mathcal{H}_{\mathbf{x}}$ and $\mathcal{H}_{\mathbf{y}}$, respectively. Throughout, we assume the integrability conditions $\mathbb{E}_{\mathbf{x}}[k] < \infty$ and $\mathbb{E}_{\mathbf{y}}[h] < \infty$.

2.4.1 HSIC using the Cross-Covariance Operator

We first introduce HSIC as the Hilbert-Schmidt norm of the cross-covariance operator which follows from Baker [1973], Fukumizu et al. [2004]. Similarly to the definition of the mean embedding (Definition 2.6), the cross-covariance operator in the RKHS is an important concept.

Definition 2.10. (Cross-Covariance operator, [Baker, 1973]). *The cross-covariance operator associated with the joint measure in $\mathbb{P}_{\mathbf{xy}}$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{y}})$ is a linear operator $\mathcal{C}_{\mathbf{xy}} : \mathcal{H}_{\mathbf{y}} \longrightarrow \mathcal{H}_{\mathbf{x}}$ defined as*

$$\begin{aligned} \mathcal{C}_{\mathbf{xy}} :&= \mathbb{E}_{\mathbf{xy}}\left[(k(\mathbf{x},.) - \mu_{\mathbf{x}}) \otimes (l(\mathbf{y},.) - \mu_{\mathbf{y}})\right] \\ &= \mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x},.) \otimes l(\mathbf{y},.)\right] - \mu_{\mathbf{x}} \otimes \mu_{\mathbf{y}}, \end{aligned} \tag{2.4.1}$$

where $\mu_{\mathbf{x}} = \mathbb{E}_{\mathbf{x}} k(\mathbf{x},.)$ and $\mu_{\mathbf{y}} = \mathbb{E}_{\mathbf{y}} l(\mathbf{y},.)$ and $\otimes$ denotes the tensor product operator formally defined as

$$\begin{aligned} &f \otimes g : \mathcal{F} \longrightarrow \mathcal{G} \\ &(f \otimes g)h \mapsto f\langle g, h\rangle_{\mathcal{G}}, \forall h \in \mathcal{G}. \end{aligned} \tag{2.4.2}$$

Next, our goal is to derive the Hilbert-Schmidt norm of this cross-covariance operator $\mathcal{C}_{\mathbf{xy}}$ as the basis of our measure of dependence, called the Hilbert-Schmidt Independent Criterion.

The Hilbert-Schmidt (HS) norm of a linear operator $\mathcal{C} : \mathcal{G} \longrightarrow \mathcal{F}$ is defined as

$$\|\mathcal{C}\|_{HS} = \sum_{ij} \langle \mathcal{C}\mathbf{v}_i, \mathbf{u}_j\rangle^2_{\mathcal{F}}, \tag{2.4.3}$$

where $\mathbf{v}_i$ and $\mathbf{u}_j$ are orthonormal bases of $\mathcal{F}$ and $\mathcal{G}$ respectively. Furthermore, a linear operator is called Hilbert-Schmidt operator if its Hilbert-Schmidt norm exists. The set of Hilbert-Schmidt operators $\mathcal{C} : \mathcal{G} \longrightarrow \mathcal{F}$ also form a Hilbert space $\mathcal{H}$ with inner product defined as

$$\langle \mathcal{C}, \mathcal{D}\rangle_{\mathcal{H}} := \sum_{ij} \langle \mathcal{C}\mathbf{v}_i, \mathbf{u}_j\rangle_{\mathcal{F}} \langle \mathcal{D}\mathbf{v}_i, \mathbf{u}_j\rangle_{\mathcal{F}}. \tag{2.4.4}$$

Then, we can compute the Hilbert-Schmidt norm of a tensor product operator as

$$\begin{aligned} \|f \otimes g\|_{HS} &= \langle f, (f \otimes g)g \rangle_{HS} = f \otimes g \\ &= \langle f, f \rangle_{\mathcal{F}} \langle g, g \rangle_{\mathcal{G}} = \|f\|_{\mathcal{F}}^2 \|g\|_{\mathcal{G}}^2. \end{aligned} \tag{2.4.5}$$

We now define the Hilbert-Schmidt Independence Criterion (HSIC) as the HS norm of the cross-covariance operator $\mathcal{C}_{\mathbf{xy}}$ in the following definition.

Definition 2.11. (Hilbert-Schmidt Independence Criterion – HSIC), [Gretton et al., 2005a, Definition 1]). *Given separable RKHSs, $\mathcal{F}, \mathcal{G}$ and a joint measure $\mathbb{P}_{\mathbf{xy}}$ over $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{y}})$, we define the Hilbert-Schmidt Independence Criterion (HSIC) as the squared HS-norm of the associated cross-covariance operator $\mathcal{C}_{\mathbf{xy}}$*

$$\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] := \|\mathcal{C}_{\mathbf{xy}}\|_{HS}^2. \tag{2.4.6}$$

In terms of kernels, HSIC can be expressed as following

$$\begin{aligned} \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] :&= \mathbb{E}_{\mathbf{xy}}\left[\langle k(\mathbf{x}, .) \otimes l(\mathbf{y}, .), k(\mathbf{x}, .) \otimes l(\mathbf{y}, .) \rangle\right] \\ &\quad - 2\mathbb{E}_{\mathbf{xy}}\left[\langle k(\mathbf{x}, .) \otimes l(\mathbf{y}, .), \mu_{\mathbf{x}} \otimes \mu_{\mathbf{y}} \rangle\right] + \langle \mu_{\mathbf{x}} \otimes \mu_{\mathbf{y}}, \mu_{\mathbf{x}} \otimes \mu_{\mathbf{y}} \rangle_{\mathcal{F}} \\ &= \mathbb{E}_{\mathbf{xx'yy'}}\left[k(\mathbf{x}, \mathbf{x}') l(\mathbf{y}, \mathbf{y}')\right] \\ &\quad - 2\mathbb{E}_{\mathbf{xy}}\left[\mathbb{E}_{\mathbf{x}'}[k(\mathbf{x}, \mathbf{x}')] \mathbb{E}_{\mathbf{y}'}[l(\mathbf{y}, \mathbf{y}')]\right] + \mathbb{E}_{\mathbf{xx}'}\left[k(\mathbf{x}, \mathbf{x}')\right] \mathbb{E}_{\mathbf{yy}'}\left[l(\mathbf{y}, \mathbf{y}')\right]. \end{aligned} \tag{2.4.7}$$

A nice property of HSIC is that with universal kernels HSIC $= 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent. This as been proved in the following theorem.

Theorem 2.7. (Independence and HSIC, [Gretton et al., 2005a, Theorem. 4]). *Let $\mathcal{F}$ and $\mathcal{G}$ be separable RKHSs with universal kernels k, l on respective compact domains $\mathcal{X}$ and $\mathcal{Y}$, then HSIC $= 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent.*

Furthermore, the population HSIC can be estimated by estimating each term in Equation (2.4.7) using the kernel matrices $\mathbf{K}$ and $\mathbf{L}$.

Theorem 2.8. (Unbiased estimator for HSIC $\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right]$, [Song et al., 2012, Theorem. 1]). *We denote by $\mathbf{S_n}$ the set of observations $(\mathbf{X_n}, \mathbf{Y_n}) = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots (\mathbf{x}_n, \mathbf{y}_n)\}$ of size n drawn i.i.d. from $P_{\mathbf{xy}}$. The unbiased estimator HSIC $\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right]$ is given by*

$$\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] = \frac{1}{n(n-3)}\left[\mathrm{Tr}\left(\tilde{\mathbf{K}}\tilde{\mathbf{L}}\right) + \frac{\mathbf{1}'\tilde{\mathbf{K}}\mathbf{1}\mathbf{1}'\tilde{\mathbf{L}}\mathbf{1}}{(n-1)(n-2)} - \frac{2}{n-2}\tilde{\mathbf{K}}\mathbf{1}'\tilde{\mathbf{L}}\mathbf{1}\right], \tag{2.4.8}$$

where $\mathbf{1}$ is a vector 1s of size n and $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ are related to $\mathbf{K}$ and $\mathbf{L}$ by $\tilde{\mathbf{K}}_{ij} = (1 - \delta_{ij})\mathbf{K}_i$ and $\tilde{\mathbf{L}}_{ij} = (1 - \delta_{ij})\mathbf{L}_{ij}$.

The HSIC unbiased estimator in Equation (2.4.8) can be alternatively formulated using U-statistics.

Theorem 2.9. *U-statistic of estimator of HSIC, [Song et al., 2012, Theorem. 3]. The finite sample unbiased estimator of $\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]$ given in Equation (2.4.8) can be written as a U-statistic,*

$$\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] = (n)_4^{-1} \sum_{(i,j,q,r)\in i_4^n} h_{ijqr}, \tag{2.4.9}$$

where $(n)_4 = \dfrac{n!}{(n-4)!}$, the index set i_4^n denotes the set of all $4-$tuples drawn without replacement from the set $\{1, \dots n\}$, and where h is the U-statistic kernel of degree 4 such that

$$h_{ijqr} = \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(l_{st} + l_{uv} - 2l_{su}), \tag{2.4.10}$$

where the summation is over all $4! = 24$ quadruples (s,t,u,v) selected without replacement from (i,j,q,r) [Song et al., 2012, Equation (11)], and the kernels k and l are associated uniquely with respective reproducing kernel Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$.

2.4.2 HSIC using the Maximum Mean discrepancy

In Section 2.3, we have derived a kernel based method for the two-sample test problem based on the Maximum Mean Discrepancy (MMD). We show here that the HSIC can be seen as a special case of the MMD [Gretton et al., 2012a, Section 7.4].

The kernel product in an RKHS $\mathcal{F}$ over $\mathcal{X} \times \mathcal{Y}$ is defined by

$$\begin{aligned} k \cdot l : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) &\longrightarrow \mathbb{R} \\ \left((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')\right) &\mapsto k\left(\mathbf{x}, \mathbf{x}'\right) \cdot l\left(\mathbf{y}, \mathbf{y}'\right), \end{aligned} \tag{2.4.11}$$

where $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ are kernels on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then the MMD for the joint probability measures $\mathbb{P}_{\mathbf{xy}}$ and the product of the marginals $\mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$ can be expressed as

$$\begin{aligned} \mathrm{MMD}^2\left[\mathcal{F}, \mathbb{P}_{\mathbf{xy}}, \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}\right] &= \left\|\mu_{\mathbf{xy}} - \mu_{\mathbf{x}}\mu_{\mathbf{y}}\right\|_{\mathcal{F}}^2 \\ &= \left\langle \mu_{\mathbf{xy}} - \mu_{\mathbf{x}}\mu_{\mathbf{y}}, \mu_{\mathbf{xy}} - \mu_{\mathbf{x}}\mu_{\mathbf{y}} \right\rangle \\ &= \left\langle \mu_{\mathbf{xy}}, \mu_{\mathbf{xy}} \right\rangle - 2\left\langle \mu_{\mathbf{xy}}, \mu_{\mathbf{x}}\mu_{\mathbf{y}} \right\rangle + \left\langle \mu_{\mathbf{x}}\mu_{\mathbf{y}}, \mu_{\mathbf{x}}\mu_{\mathbf{y}} \right\rangle \\ &= \mathbb{E}_{\mathbf{xx'yy'}}\left[k(\mathbf{x},\mathbf{x}')l(\mathbf{y},\mathbf{y}')\right] - 2\mathbb{E}_{\mathbf{xy}}\left[\mathbb{E}_{\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')]\mathbb{E}_{\mathbf{y}'}[l(\mathbf{y},\mathbf{y}')]\right] \\ &\quad + \mathbb{E}_{\mathbf{xx}'}\left[k(\mathbf{x},\mathbf{x}')\right]\mathbb{E}_{\mathbf{yy}'}\left[l(\mathbf{y},\mathbf{y}')\right], \end{aligned} \tag{2.4.12}$$

which is exactly the expression of HSIC given in Equation (2.4.7).

2.4.3 Application to the Independence Test

We have introduce a notion of independence and we now describe how to use HSIC as a basis of an independence test. Similarly to Section 2.3.4, we describe a statistical hypothesis test for dependence. Given i.i.d. samples $\mathbf{X_n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{Y_n} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ of size n drawn from $\mathbb{P}_{\mathbf{xy}}$, the statistical test, $\mathcal{T} : \mathcal{X}^n \times \mathcal{Y}^n \mapsto \{0, 1\}$ is used to distinguished between the null hypothesis $\mathcal{H}_0 : \mathbb{P}_{\mathbf{xy}} = \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$ and the alternative hypothesis $\mathcal{H}_1 : \mathbb{P}_{\mathbf{xy}} \neq \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$. This is achieved by comparing the test statistic $\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]$ with a particular threshold.

Theorem 2.10. (Asymptotic distribution of MMD, [Gretton et al., 2012a, Theorem. 19])**.** *If $\mathbb{E}[h] < \infty$, given the set of observations $\mathbf{S_n} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots (\mathbf{x}_n, \mathbf{y}_n)\}$ of size n drawn i.i.d. from P_{xy}. When $\mathbb{P}_{\mathbf{xy}} \neq \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$, an unbiased estimate of $\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right]$ given in Equation (2.4.7) converges asymptotically in distribution to a Gaussian distribution with mean zero and variance $\sigma^2_{\mathrm{HSIC}_{\mathbf{xy}}}$*

$$n^{1/2}\left(\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathrm{HSIC}_{\mathbf{xy}}}\right), \tag{2.4.13}$$

and where

$$\sigma^2_{\mathrm{HSIC}_{\mathbf{xy}}} = 16\left(\mathbb{E}_{\mathbf{x}_i}\left(\mathbb{E}_{\mathbf{x}_j, \mathbf{x}_q, \mathbf{x}_r} h_{ijqr}\right)^2 - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right]\right). \tag{2.4.14}$$

Its empirical estimate is $\hat{\sigma}^2_{\mathrm{HSIC}_{\mathbf{xy}}} = 16\{R_{\mathbf{xy}} - (\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right])^2\}$ where

$$R_{\mathbf{xy}} = \frac{1}{n}\sum_{i=1}^{n}\left((n-1)_3^{-1}\sum_{(j,q,r)\in i_3^n \setminus \{i\}} h_{ijqr}\right)^2, \tag{2.4.15}$$

and the index set $i_3^n \setminus \{i\}$ denotes the set of all $3-$tuples drawn without replacement from the set $\{1, \ldots n\} \setminus \{i\}$.

When $\mathbb{P}_{\mathbf{xy}} = \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$, the U-statistic is degenerate, we have $\mathbb{E}h = 0$. In this case, $\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right]$ converges in distribution according to

$$\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] \xrightarrow{d} \frac{1}{n}\sum_{l=1}^{\infty}\lambda_l\{z_l^2 - 2\}, \tag{2.4.16}$$

where $z_l^2 \sim \chi^2(1)$ i.i.d. and λ_l are the solutions to the eigenvalue equation

$$\int h_{ijqr}\psi_l(\mathbf{S}_j)d\mathbb{P}_{\mathbf{S}_i, \mathbf{S}_q, \mathcal{S}_r} = \lambda_l\psi_l(\mathbf{S}_j). \tag{2.4.17}$$

We then use the $1-\alpha$ quantile of this distribution as the test threshold. Since the distribution under $\mathcal{H}_0$ is complicated, we need to accurately approximate its quantile. There are several possible strategies to do so analogous to the discussion for the MMD in Section 2.3.4.

The HSIC has been applied in many applications [Gretton et al., 2005a], including conditional

independence [Fukumizu et al., 2007] and causal discovery [Peters et al., 2014, Zhang et al., 2011].

2.5 Estimation of the Structure of the Graphical Models

In this section, we give a brief overview of the estimation of precision matrices and testing conditional independence on undirected graphical models. More details account on this topic can be found in Dawid [1979], Dempster [1972], Lauritzen [1996] and Whittaker [2009].

The importance of estimating covariance matrices and their inverses, called precision matrices, is fundamental in modern multivariate analysis and in a wide array of scientific applications. The covariance matrix reveals marginal correlations between variables, while the precision matrix encodes conditional correlations between pairs of variables given the remaining variables.

Generally, graphical models blend probability theory and graph theory together. They are powerful tools for analyzing relationships between a large number of random variables. A *graph* is a set of vertices $\mathrm{V} = \{1, \ldots, p\}$ and a set of edges $\mathrm{E} \subseteq \mathrm{V} \times \mathrm{V}$. An *undirected graphical model* is a joint probability distribution, $\mathbb{P}_{\mathbf{x}}$, defined on an undirected graph G, where the vertices V in the graph index a collection of random variables $\mathbf{x} = \{\mathbf{x}_1, \ldots \mathbf{x}_p\}$ and the edges encode conditional independence relationships among random variables

$$\mathbb{P}_{\mathbf{x}} \propto \prod_{c \in \mathcal{C}} \Psi_c(\mathbf{x}_c), \tag{2.5.1}$$

where $\mathcal{C}$ is the set of maximal cliques in the graph and $\{\Psi_c\}_{c \in \mathcal{C}}$ are non-negative potential functions.

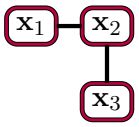
It is well known that recovering the structure of an undirected Gaussian graph is equivalent to the recovery of the support of the precision matrix (Figure 2.3). Formally, suppose we have a sample $\mathbf{X}_p = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ of dimension p and size n with the mean of each $\mathbf{x}_i$ equal to zero, and a covariance matrix of size $p \times p$ is $\Sigma_{ij} = \mathbb{E}(\mathbf{x}_i \mathbf{x}_j^T)$ such that $\mathbf{x} \sim \mathcal{N}_p(0, \mathbf{\Sigma})$ then

$$(i, j) \notin \mathrm{E} \iff \mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j | V \setminus \{i, j\} \iff \Sigma_{ij}^{-1} = 0. \tag{2.5.2}$$

Testing conditional independence is an important concept in statistics, artificial intelligence, and related fields [Dawid, 1979]. A common measure for the testing of independence of two variables conditioned on a set of variables is the *partial correlation* $\rho_{\mathbf{x}_1 \mathbf{x}_2 \cdot \mathbf{x}_3}$. With the assumption that all variables are multivariate Gaussian, the partial correlation is zero if and only if $\mathbf{x}_1$ is conditionally independent from $\mathbf{x}_2$ given a set of variables, $\mathbf{x}_3$:

$$\mathcal{H}_0 : \rho_{\mathbf{x}_1 \mathbf{x}_2 \cdot \mathbf{x}_3} = 0 \qquad \text{vs} \qquad \mathcal{H}_1 : \rho_{\mathbf{x}_1 \mathbf{x}_2 \cdot \mathbf{x}_3} \neq 0. \tag{2.5.3}$$

The distribution of the sample partial correlation for a Gaussian distribution was described



$$\boldsymbol{\Sigma}^{-1} = \begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{array} \begin{array}{c} \begin{array}{ccc} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{array} \\ \begin{pmatrix} * & * & 0 \\ * & * & * \\ 0 & * & * \end{pmatrix} \end{array}$$

Figure 2.3 – For a given random variables $\mathbf{x} \sim \mathcal{N}_3(0, \boldsymbol{\Sigma})$ with the covariance matrix $\boldsymbol{\Sigma}$, we consider the problem of non-parametric testing of the hypothesis of conditional independence of In this case, we have $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent conditioned on $\mathbf{x}_3$, this hypothesis is denoted as $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_3 | \mathbf{x}_2$.

by Fisher [1924] and we would reject $\mathbf{H}_0$ if the absolute value of a transformed test statistic exceeded the critical value from the Student table evaluated at $\delta/2$. The computational complexity of the partial correlation is $\mathcal{O}(np^2 + p^3)$ which simplifies to $\mathcal{O}(np^2)$ as $n \geq p$. However, as mentioned in [Kendall, 1946, Ch. 26 & 27], this hypothesis test makes a strong assumption that the data are Gaussian distributed, and in particular that the fourth-order moment is constrained.

Furthermore, tests of conditional independence can be made without any assumption of normality in the distribution, using for instance the permutation distribution of $\rho_{XY.Z}$ or bootstrap techniques, but this becomes too computationally expensive in practice when n tends to be large.

Chapter 3

A Hypothesis Test of Relative Similarity

Probabilistic generative models provide a powerful framework for representing data that avoids the expense of manual annotation typically needed by discriminative approaches. Model selection in this generative setting can be challenging, however, particularly when likelihoods are not easily accessible. In that regard, this chapter addresses Research Question 1 by proposing a statistical test of relative similarity, which is used to determine which of two models generates samples that are significantly closer to a real-world reference dataset of interest. We use as our test statistic the difference in maximum mean discrepancies (MMDs) between the reference dataset and each model dataset, and derive a powerful, low-variance test based on the joint asymptotic distribution of the MMDs between each reference-model pair. In experiments on deep generative models, including the variational auto-encoder and generative moment matching network, the tests provide a meaningful ranking of model performance as a function of parameter and training settings.

Work corresponding to this chapter is published in:

- W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *The 4th International Conference on Learning Representations*, 2016a.

This publication has received the Université Paris-Saclay STIC Doctoral School Best Scientific Contribution Award.

Project: https://github.com/wbounliphone/relative_similarity_test & https://github.com/eugenium/MMD.

Contents

3.1 Introduction

Generative models based on deep learning techniques aim to provide sophisticated and accurate models of data, without expensive manual annotation [Bengio, 2009, Kingma et al., 2014]. This is especially of interest as deep networks tend to require comparatively large training samples to achieve a good result [Krizhevsky et al., 2012]. Model selection within this class of techniques can be a challenge, however.

First, likelihoods can be difficult to compute for some families of recently proposed models based on deep learning [Goodfellow et al., 2014, Li et al., 2015b]. The current best method to evaluate such models is based on Parzen-window estimates of the log likelihood [Goodfellow et al., 2014, Section 5]. Second, if we are given two models with similar likelihoods, we typically do not have a computationally inexpensive hypothesis test to determine whether one likelihood is significantly higher than the other. Permutation testing or other generic strategies are often computationally prohibitive, bearing in mind the relatively high computational requirements of deep networks [Krizhevsky et al., 2012].

In this work, we provide an alternative strategy for model selection, based on a novel, non-parametric hypothesis test of relative similarity. We treat the two trained networks being compared as generative models [Goodfellow et al., 2014, Hinton et al., 2006, Salakhutdinov and Hinton, 2009], and test whether the first candidate model generates samples significantly closer to a reference validation set. The null hypothesis is that the ordering is reversed, and

the second candidate model is closer to the reference (further, both samples are assumed to remain distinct from the reference, as will be the case for any sufficiently complex modeling problem).

Our model selection criterion is based on the maximum mean discrepancy (MMD) [Gretton et al., 2006, 2012a], which represents the distance between embeddings of empirical distributions in a reproducing kernel Hilbert space (RKHS). The maximum mean discrepancy is a metric on the space of probability distributions when a characteristic kernel is used [Fukumizu et al., 2007, Gretton et al., 2006, Sriperumbudur et al., 2011], meaning that the distribution embeddings are unique for each probability measure. Recently, the MMD has been used in training generative models adversarially, [Dziugaite et al., 2015, Li et al., 2015b], where the MMD measures the distance of the generated samples to some reference target set; it has been used for statistical model criticism [Lloyd and Ghahramani, 2015]; and to minimize the effect of nuisance variables on learned representations [Louizos et al., 2016].

Rather than *train* a single model using the MMD distance to a reference distribution, our goal in this work is to *evaluate the relative performance* of two models, by testing whether one generates samples significantly closer to the reference distribution than the other. This extends the applicability of the MMD to problems of model selection and evaluation. Key to this result is a novel expression for the joint asymptotic distribution of two correlated MMDs (between samples generated from each model, and samples from the reference distribution). Li et al. [2015a] have derived the joint distribution of a specific MMD estimator under the assumption that the distributions are equal. By contrast, we derive the case in which the distributions are unequal, as is expected due to irreducible model error.

We derive the joint asymptotic distribution of the MMDs in Section 3.2.1 and we formulate a hypothesis test of *relative similarity*, to determine whether the difference in MMDs is statistically significant. Our first test benchmark is on a synthetic data for which the ground truth is known (Section 3.3), where we verify that the test performs correctly under the null and the alternative.

Finally, in Section 3.4, we demonstrate the performance of our test over a broad selection of model comparison problems in the deep learning setting, by evaluating relative similarity of pairs of model outputs to a validation set over a range of training regimes and settings. Our benchmark models include the variational auto-encoder [Kingma and Welling, 2014] and the generative moment matching network [Li et al., 2015b]. We first demonstrate that the test performs as expected in scenarios where the same model is trained with different training set sizes, and the relative ordering of model performance is known. We then fix the training set size and change various architectural parameters of these networks, showing which models are significantly preferred with our test. We validate the rankings returned by the test using a separate set of data for which we compute alternate metrics for assessing the models, such as classification accuracy and likelihood.

For the formal setup of this whole chapter, suppose that we have random variables $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$, $\mathbf{y} \sim \mathbb{P}_{\mathbf{y}}$ and $\mathbf{z} \sim \mathbb{P}_{\mathbf{z}}$, that take values on $(\mathcal{X}, \mathcal{B}_{\mathbf{x}})$, $(\mathcal{Y}, \mathcal{B}_{\mathbf{y}})$ and $(\mathcal{Z}, \mathcal{B}_{\mathbf{z}})$ respectively, here $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are three separable metric and $\mathcal{B}_{\mathbf{x}}, \mathcal{B}_{\mathbf{y}}, \mathcal{B}_{\mathbf{z}}$ are Borel σ-algebras. Furthermore, we

define the kernel $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$, with the corresponding feature map ϕ such that $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ and denote the corresponding RKHSs with $\mathcal{H}$.

Let now reformulate our Research Question 1 into a mathematical setting. In Section 2.3, we have shown that MMD is a metric on two probability distribution $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{y}}$. With this choice, the problem we would like to solve is described as follows

Problem 3.1. *Given separable RKHSs $\mathcal{H}$ with $\mathbb{P}_{\mathbf{x}} \neq \mathbb{P}_{\mathbf{y}}$ and $\mathbb{P}_{\mathbf{x}} \neq \mathbb{P}_{\mathbf{z}}$, the statistical relative similarity test $\mathcal{T}_{MMD} : \mathcal{X}^n \times \mathcal{X}^n \times \mathcal{X}^n \mapsto \{0, 1\}$ is used to test the null hypothesis*

$$\mathcal{H}_0^{MMD} : \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right] \leq \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right], \tag{3.1.1}$$

versus the alternative hypothesis

$$\mathcal{H}_1^{MMD} : \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right] > \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right], \tag{3.1.2}$$

at a given significance level α.

3.2 A Test of Relative Similarity

In this section, we calculate two dependent MMD statistics and derive the joint asymptotics distribution of this dependent quantities, which is used to construct a consistent test for the Problem 3.1. This is a direct application of Theorem 2.4 using the definition of MMD written as a U-statistic of degree 2.

3.2.1 Joint Asymptotic Distribution of Two Correlated MMD

In this section, we derive our statistical test for relative similarity as measured by MMD. In order to maximize the statistical efficiency of the test, we will reuse samples from the reference distribution, denoted by $P_{\mathbf{x}}$, to compute the MMD estimates with two candidate distributions $P_{\mathbf{y}}$ and $P_{\mathbf{z}}$. We consider two MMD estimates $\widehat{\mathrm{MMD}}_u^2\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}\right]$ and $\widehat{\mathrm{MMD}}_u^2\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}\right]$, and as the data sample $\mathbf{X_n}$ is identical between them, these estimates will be correlated. We therefore first derive the joint asymptotic distribution of these two metrics and use this to construct a statistical test.

Theorem 3.1 (Joint asymptotic distribution of two correlated MMD)**.** *We assume that $\mathbb{P}_{\mathbf{x}} \neq \mathbb{P}_{\mathbf{y}}$, $\mathbb{P}_{\mathbf{x}} \neq \mathbb{P}_{\mathbf{z}}$, $\mathbb{E}(k(\mathbf{x}_i, \mathbf{x}_j)) < \infty$, $\mathbb{E}(k(\mathbf{y}_i, \mathbf{y}_j)) < \infty$ and $\mathbb{E}(k(\mathbf{x}_i, \mathbf{y}_j)) < \infty$, then*

$$\begin{aligned} n^{1/2} \left(\begin{pmatrix} \widehat{\mathrm{MMD}}_u^2\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}\right] \\ \widehat{\mathrm{MMD}}_u^2\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}\right] \end{pmatrix} - \begin{pmatrix} \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right] \\ \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right] \end{pmatrix} \right) \\ \xrightarrow{d} \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\mathrm{MMD}^2_{\mathbf{xy}}} & \sigma_{\mathrm{MMD}^2_{\mathbf{xy,xz}}} \\ \sigma_{\mathrm{MMD}^2_{\mathbf{xy,xz}}} & \sigma^2_{\mathrm{MMD}^2_{\mathbf{XZ}}} \end{pmatrix} \right), \end{aligned} \tag{3.2.1}$$

where $\sigma^2_{\mathrm{MMD}^2_{\mathbf{xy}}}, \sigma^2_{\mathrm{MMD}^2_{\mathbf{xz}}}$ *are respectively the variances of* $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right]$ *and* $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right]$ *are as in Equation (2.3.10) and the covariance term* $\sigma_{\mathrm{MMD}^2_{\mathbf{xy},\mathbf{xz}}}$ *of* $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right]$ *and* $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right]$ *is as in Equation (2.1.9).*

Proof: Equation (3.2.1) is a direct application of [Hoeffding, 1963, Theorem 7.1], described in Theorem 2.4, which gives the joint asymptotic distribution of U-statistics, which here is the MMD statistics. □

Theorem 3.2 (Empirical estimate of the variance/covariance of MMD)**.** *We note* $[\tilde{\mathbf{K}}_{\mathbf{xx}}]_{ij} = [\mathbf{K}_{\mathbf{xx}}]_{ij}$ *for all* $i \neq j$ *and* $[\mathbf{K}_{\mathbf{xx}}]_{ij} = 0$ *for* $j = i$*. Same for* $\mathbf{K}_{\mathbf{yy}}$ *and* $\mathbf{K}_{\mathbf{zz}}$*. The empirical estimate of the variance term* $\sigma^2_{\mathrm{MMD}^2_{\mathbf{xy}}}$ *of* $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right]$ *in Equation (3.2.1), neglecting higher order terms, can be computed in* $\mathcal{O}(n^2)$ *and is*

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{MMD}^2_{\mathbf{xy}}} = \frac{4(n-2)}{n(n-1)}\Bigg\{ & \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\right)^2 \\ & - 2\left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right) \\ & + \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\right)^2 \\ & - 2\left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right) \\ & + \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - 2\left(\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right)^2 + \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1}\Bigg\},\end{aligned} \tag{3.2.2}$$

where $\mathbf{1}$ *is a vector of 1s of size* $n \times 1$*. The empirical estimate of the variance term* $\sigma^2_{\mathrm{MMD}^2_{\mathbf{xz}}}$ *is similar to Equation (3.2.2) by substituting* $\mathbf{K}_{\mathbf{xy}}$ *and* $\tilde{\mathbf{K}}_{\mathbf{yy}}$ *by* $\mathbf{K}_{\mathbf{xz}}$ *and* $\tilde{\mathbf{K}}_{\mathbf{zz}}$*, respectively. Moreover, the empirical estimate of covariance* $\sigma_{\mathrm{MMD}^2_{\mathbf{xy},\mathbf{xz}}}$ *in Equation (3.2.1), neglecting higher order terms, can be computed in* $\mathcal{O}(n^2)$ *and is*

$$\begin{aligned}\hat{\sigma}_{\mathrm{MMD}^2_{\mathbf{xy},\mathbf{xz}}} = \frac{4(n-2)}{n(n-1)}\Bigg\{ & \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{K}_{\mathbf{xx}}\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{K}_{\mathbf{xx}}\mathbf{1}\right)^2 \\ & - \left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{K}_{\mathbf{xx}}K_{\mathbf{xz}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{K}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^TK_{\mathbf{xz}}\mathbf{1}\right) \\ & - \left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{K}_{\mathbf{xx}}K_{\mathbf{xy}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{K}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^TK_{\mathbf{xz}}\mathbf{1}\right) \\ & + \left(\frac{1}{n^3}\mathbf{1}^TK_{\mathbf{yx}}K_{\mathbf{xz}}\mathbf{1} - \frac{1}{n^4}\mathbf{1}^TK_{\mathbf{xy}}\mathbf{1}\mathbf{1}^TK_{\mathbf{xz}}\mathbf{1}\right)\Bigg\}.\end{aligned} \tag{3.2.3}$$

Proof: A complete proof is given in Section 3.6.1. □

3.2.2 A Statistical Test via Two Uncorrelated MMD Statistics

Based on the empirical distribution from Equation (3.2.1), we now describe a statistical test to solve the following problem:

Problem 3.2 (Relative similarity test)**.** *Given observations* $\mathbf{X_n} := \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, $\mathbf{Y_n} := \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ *and* $\mathbf{Z_n} := \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ *i.i.d. from* $\mathbb{P}_\mathbf{x}$, $\mathbb{P}_\mathbf{y}$ *and* $\mathbb{P}_\mathbf{z}$, *respectively, such that* $\mathbb{P}_\mathbf{x} \neq \mathbb{P}_\mathbf{y}$, $\mathbb{P}_\mathbf{x} \neq \mathbb{P}_\mathbf{z}$, *we test the hypothesis that* $\mathbb{P}_\mathbf{x}$ *is closer to* $\mathbb{P}_\mathbf{z}$ *than* $\mathbb{P}_\mathbf{y}$, *i.e. we test the null hypothesis*

$$\mathcal{H}_0 : \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}\right] \leq \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{z}\right] \tag{3.2.4}$$

versus the alternative hypothesis

$$\mathcal{H}_1 : \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}\right] > \mathrm{MMD}\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{z}\right] \tag{3.2.5}$$

at a given significance level α

The test statistic $\widehat{\mathrm{MMD}}^2_u\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}\right] - \widehat{\mathrm{MMD}}^2_u\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}\right]$ is used to compute the p-value for the standard normal distribution. The test statistic is obtained by rotating the joint distribution (cf. Equation 3.2.1) by $\pi/4$ about the origin, and integrating the resulting projection on the first axis. Denote the asymptotically normal distribution of $n^{1/2}\left[\widehat{\mathrm{MMD}}^2_u\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}\right]; \widehat{\mathrm{MMD}}^2_u\left[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}\right]\right]^T$ as $\mathcal{N}_2\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. The resulting distribution from rotating by $\pi/4$ and projecting onto the primary axis is $\mathcal{N}\left([\mathbf{Q}\boldsymbol{\mu}]_1, [\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^T]_{11}\right)$ where $\mathbf{Q} = \frac{\sqrt{2}}{2}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$ is the rotation matrix by $\pi/4$ and

$$[\mathbf{Q}\boldsymbol{\mu}]_1 = \left[\frac{\sqrt{2}}{2}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix}\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}\right]\\ \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{z}\right]\end{pmatrix}\right]_1 \tag{3.2.6}$$

$$= \frac{\sqrt{2}}{2}\left[\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}\right] - \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{z}\right]\right]; \tag{3.2.7}$$

$$[\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^T]_{11} = \frac{1}{2}\left[\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix}\sigma^2_{\mathrm{MMD}^2_{\mathbf{xy}}} & \sigma_{\mathrm{MMD}^2_{\mathbf{xy},\mathbf{xz}}}\\ \sigma_{\mathrm{MMD}^2_{\mathbf{xy},\mathbf{xz}}} & \sigma^2_{\mathrm{MMD}^2_{\mathbf{xz}}}\end{pmatrix}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}\right]_{11} \tag{3.2.8}$$

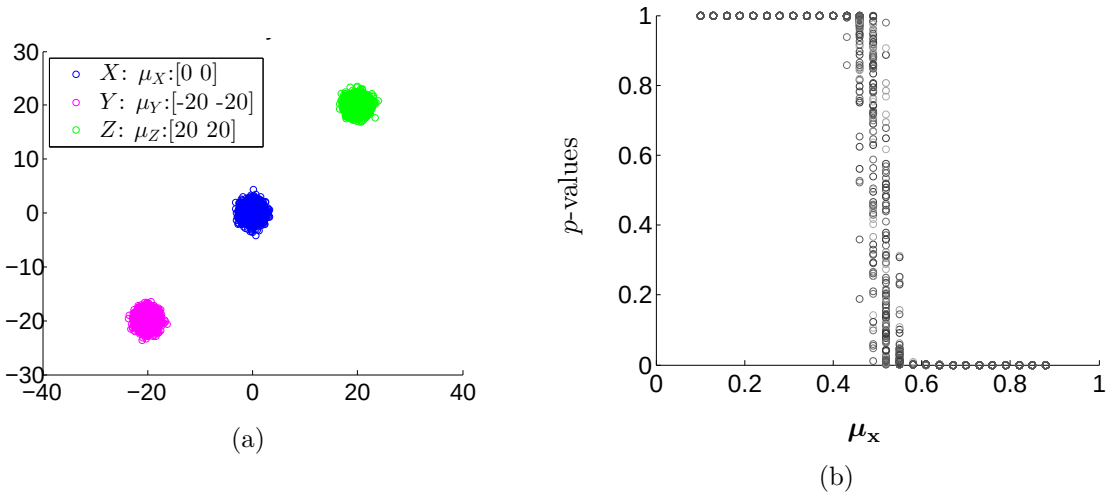
$$= \frac{1}{2}\left(\sigma^2_{\mathrm{MMD}^2_{\mathbf{xy}}} + \sigma^2_{\mathrm{MMD}^2_{\mathbf{xz}}} - 2\sigma_{\mathrm{MMD}^2_{\mathbf{xy},\mathbf{xz}}}\right), \tag{3.2.9}$$

with $\boldsymbol{Q}$ is the rotation matrix by $\pi/4$. Then, the p-values for testing $\mathcal{H}_0$ versus $\mathcal{H}_1$ are

$$p \leq \Phi\left(-\frac{\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}\right] - \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{z}\right]}{\sqrt{\sigma^2_{\mathrm{MMD}^2_{\mathbf{xy}}} + \sigma^2_{\mathrm{MMD}^2_{\mathbf{xz}}} - 2\sigma_{\mathrm{MMD}^2_{\mathbf{xy},\mathbf{xz}}}}}\right), \tag{3.2.10}$$

where Φ is the cumulative distribution function of a standard normal distribution.

In practice, the terms in Equation (3.2.10) are replaced by their empirical expectations, which are given in Lemma 2.5 and Theorem 3.2.



(a) (b)

Figure 3.1 – (a) Illustration of the synthetic dataset where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are, respectively, Gaussian distributed with means $\boldsymbol{\mu}_{\mathbf{x}} = [0, 0]^T$, $\boldsymbol{\mu}_{\mathbf{y}} = [-20, -20]^T$, $\boldsymbol{\mu}_{\mathbf{z}} = [20, 20]^T$ and with variance $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. (b) For $n = 1000$, we fixed $\boldsymbol{\mu}_{\mathbf{y}} = [-5, -5]$, $\boldsymbol{\mu}_{\mathbf{z}} = [5, 5]$ and varied $\boldsymbol{\mu}_{\mathbf{x}}$ such that $\boldsymbol{\mu}_{\mathbf{x}} = (1-\gamma)\boldsymbol{\mu}_{\mathbf{y}} + \gamma\boldsymbol{\mu}_{\mathbf{z}}$, for 41 regularly spaced values of $\gamma \in [0.1,\ 0.9]$ versus p-values for 100 repeated tests.

In Section 3.6.3, we prove that Equation (3.2.9), obtained by first performing a rotation followed by integration into the first axis, is equivalent to calculating the variance of the difference $\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}] - \widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}]$, which is a U-statistic.

3.3 Experimental Validation

We verify the validity of the hypothesis test described above using a synthetic data set in which we can directly control the relative similarity between distributions.

We constructed three Gaussian distributions as illustrated in Figure. 3.1a. These Gaussian distributions are specified with different means so that we can control the degree of relative similarity between them. The question is whether the similarity between $\mathbf{x}$ and $\mathbf{z}$ is greater than the similarity between $\mathbf{x}$ and $\mathbf{y}$. In these experiments, we used a Gaussian kernel with bandwidth selected as the median pairwise distance between data points, and we fixed $\boldsymbol{\mu}_{\mathbf{y}} = [-20, -20]$, $\boldsymbol{\mu}_{\mathbf{z}} = [20, 20]$ and varied $\boldsymbol{\mu}_{\mathbf{x}}$ such that $\boldsymbol{\mu}_{\mathbf{x}} = (1-\gamma)\boldsymbol{\mu}_{\mathbf{y}} + \gamma\boldsymbol{\mu}_{\mathbf{z}}$, for 41 regularly spaced values of $\gamma \in [0.1,\ 0.9]$ (avoiding the degenerate cases $\mathbb{P}_{\mathbf{x}} = \mathbb{P}_{\mathbf{y}}$ or $\mathbb{P}_{\mathbf{x}} = \mathbb{P}_{\mathbf{z}}$).

Figure 3.1b shows the p-values of the relative similarity test for different distribution. When γ is varying around 0.5, i.e., when $\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{Z_n}, \mathbf{Y_n}]$ is almost equal to $\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}]$, the p-values quickly transition from 1 to 0, indicating strong discrimination of the test. Figure 3.2 shows an empirical scatter plot of the pairs of MMD statistics along with a 2σ iso-curve of the estimated distribution, demonstrating that the parametric Gaussian distribution is well calibrated to the empirical values.

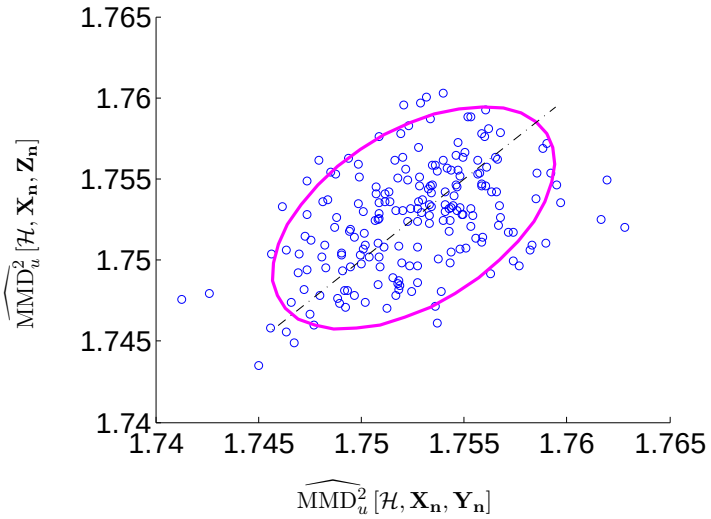


Figure 3.2 – The empirical scatter plot of the joint MMD statistics with $n = 1000$ for 200 repeated tests, along with the 2σ iso-curve of the analytical Gaussian distribution estimated by Equation (3.2.1). The analytical distribution closely matches the empirical scatter plot, verifying the correctness of the variances.

3.4 Model Selection for Deep Unsupervised Neural Networks

An important application of the problem of relative similarity(Problem 3.2) can be found in recent work on unsupervised learning with deep neural networks [Bengio et al., 2014, Goodfellow et al., 2014, Kingma and Welling, 2014, Larochelle and Murray, 2011, Li et al., 2015b, Salakhutdinov and Hinton, 2009]. As noted by several authors, the evaluation of generative models is a challenging open problem [Goodfellow et al., 2014, Li et al., 2015b], and the distributions of samples from these models are very complex and difficult to evaluate. The performance of the relative test of similarity can be used to compare different model settings, or even model families, in a statistically valid framework. To compare two models using our test, we generate samples from both, and compare these to a set of real target data samples that were not used to train either model.

In the experiments, in the sequel, we focus on the recently introduced variational auto-encoder (VAE) [Kingma and Welling, 2014] and the generative moment matching networks (GMMN) [Li et al., 2015b]. The former trains an encoder and decoder network jointly minimizing a regularized variational lower bound [Kingma and Welling, 2014]. While the latter class of models is purely generative minimizing an MMD based objective, this model works best when coupled with a separate auto-encoder which reduces the dimensionality of the data. An architectural schematic for both classes of models is provided in Figure 3.3. Both these models can be trained using standard backpropagation [Rumelhart et al., 1988]. Using the latent variable prior we can directly sample the data distribution of these models without using MCMC procedures [Hinton et al., 2006, Salakhutdinov and Hinton, 2009].

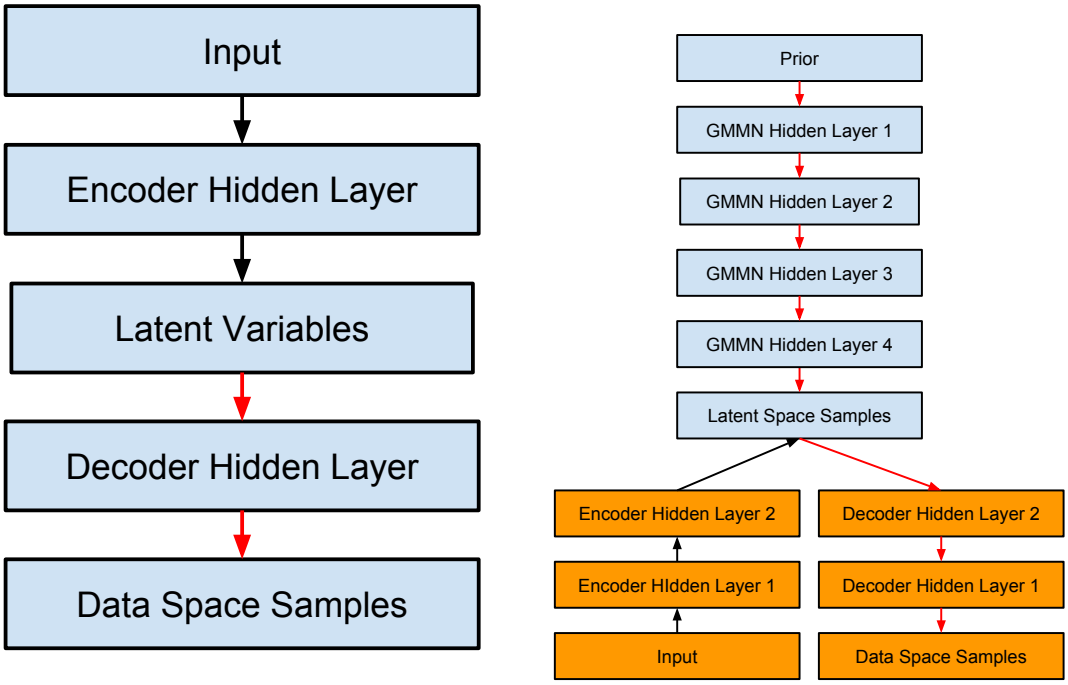
We use the MNIST and FreyFace datasets for our analysis [Goodfellow et al., 2014, Kingma

and Welling, 2014, LeCun et al., 1998]. We first demonstrate the effectiveness of our test in a setting where we have a theoretical basis for expecting superiority of one unsupervised model versus another. Specifically, we use a setup where more training samples were used to create one model versus the other. We find that the relative test of similarity agrees with the expected results (models trained with more data generalize better). We then demonstrate how the relative test of similarity can be used in evaluating network architecture choices, and we show that our test strongly agrees with other established metrics, but in contrast can provide significance results using just the validation data while other methods may require an additional test set.

Several practical matters must be considered when applying the relative similarity test. The selection of kernel can affect the quality of results, particularly more suitable kernels can give a faster convergence. In this work we extend the logic of the median heuristic [Gretton et al., 2012b] for bandwidth selection by computing the median pairwise distance between samples from $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{z}}$ and averaging that with the median pairwise distance between samples from $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{y}}$, which helps to maximize the difference between the two MMD statistics. Although the derivations for the variance of our statistic hold for all cases, the estimates require asymptotic arguments and thus a sufficiently large n. Selecting the kernel bandwidth in an appropriate range can therefore substantially increase the power of the test at a fixed sample size. While we observed the median heuristic to work well in our experiments, there are cases where alternative choices of kernel can provide greater power: for instance, the kernel can be chosen to maximize the expected test power on a held-out dataset [Gretton et al., 2012b].

3.4.1 Variational Auto-Encoder Sample Size and Architecture Experiments

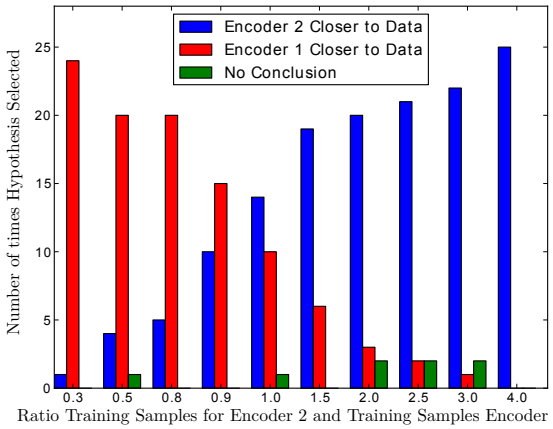
We use the architecture from Kingma and Welling [2014] with a hidden layer at both the encoder and decoder and a latent variable layer as shown in Figure 3.3a. We use sigmoidal activation for the hidden layers of encoder and decoder. For the FreyFace data, we use a Gaussian prior on the latent space and data space. For MNIST, we used a Bernoulli prior for the data space. We fix the training set size of the second auto-encoder to 300 images for the FreyFace data and 1500 images for the MNIST data. We vary the number of training samples for the first auto-encoder. We then generate samples from both auto-encoders and compare them using the relative test of similarity to a held out set of data. We use 1500 FreyFace samples as the target in the relative test of similarity and 15000 images from MNIST. Since a single sample of the data might lead to better generalization performance by chance, we repeat this experiment multiple times and record whether the relative similarity test indicated a network is preferred or if it failed to reject the null hypothesis. The results are shown in Figure 3.4 which demonstrates that we are closely following the expected model preferences. Additionally for MNIST we use another separate set of supervised training and test data. We encode this data using both auto-encoders and use logistic regression to obtain a classification accuracy. The indicated accuracies closely match the results of the relative similarity test, further validating the test. We consider model selection between networks using different architectures. We train two encoders, one a fixed reference model (400 hidden units and 20



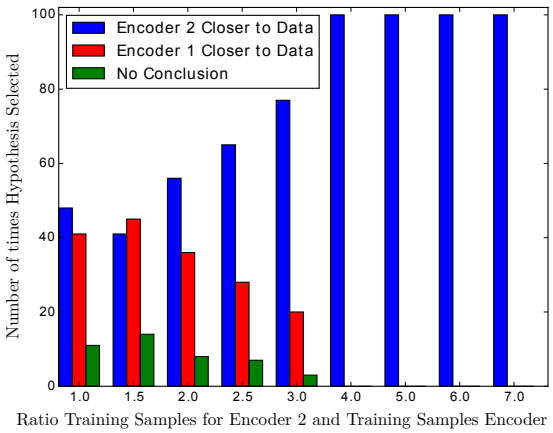
(a) Variational auto-encoder reference model (b) Auto-Encoder + GMMN reference model

Figure 3.3 – In Figure 3.3a, we have 400 hidden nodes (both encoder and decoder) and 20 latent variables in the reference model for our experiments. In Figure 3.3b, we illustate that the auto-encoder (indicated in orange) is trained separately and has 1024 and 32 hidden nodes in decode and encode hidden layers. The GMMN has 10 variables generated by the prior, and the hidden layers have 64, 256, 256, 1024 nodes in each layer respectively. In both networks red arrows indicate the data flow during sampling

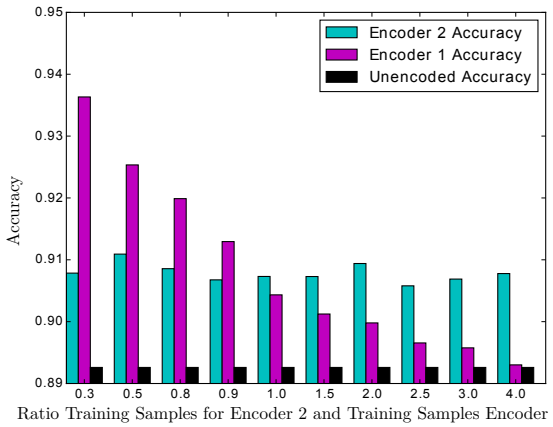
.



(a) Decision of the relative test of similarity for 25 trials



(b) Selected Hypothesis (100 runs)



(c) Average Accuracy Using Autoencoder Features (25 trials)

Figure 3.4 – In Figure 3.4a, we show the effect of varying the training set size of one auto-encoder trained on MNIST data. In Figure 3.4c As a secondary validation we compute the classification accuracy of MNIST on a separate train/test set encoded using encoder 1 and encoder 2. In Figure 3.4b We then show the effect of varying the training set size of one auto-encoder using the FreyFace data. We note that due to the size of the FreyFace dataset, we limit the range of ratios used. From this figure we see that the results of the relative similarity test match our expectation: more data produces models which more closely match the true distribution.

Hidden VAE 1	Latent VAE 1	Result RelativeMMD	Accuracy (%) VAE 1	Accuracy (%) VAE 2	Lower Bound VAE 1	Lower Bound VAE 2
200	5	Favor VAE 2	92.8 ± 0.3	**94.7 ± 0.2**	-126	**-97**
200	20	Favor VAE 2	92.6± 0.3	**94.5 ± 0.2**	-115	**-105**
400	50	Favor VAE 1	**94.6 ± 0.2**	94.0 ± 0.2	**-99.6**	-123.44
800	20	Favor VAE 1	**94.8 ± 0.2**	93.9 ± 0.2	**-111**	-115
800	50	Favor VAE 1	94.2 ± 0.3	94.5 ± 0.2	**-101**	-103

Table 3.1 – We compare several variational auto encoder (VAE) architectural choices for the number of hidden units in both decoder and encoder and the number of latent variables for the VAE. The reference encoder, denoted encoder 2, has 400 hidden units and 20 latent variables. We denote the competing architectural models as encoder 1. We vary the number of hidden nodes in both the decoder and encoder and the number of latent variables. Our test closely follows the performance difference of the auto-encoder on a supervised task (MNIST digit classification) as well as the variational lower bound on a withheld set of data. The data used for evaluating the Accuracy and Lower Bound is separate from that used to train the auto-encoders and for the hypothesis test.

latent variables), and the other varying as specified in Table 3.1. 25000 images from the MNIST data set were used for training. We use another 20000 images as the target data in the relative test of similarity. Finally, we use a set of 10000 training and 10000 test images for a supervised task experiment. We use the labels in the MNIST data and perform training and classification using an ℓ_2-regularized logistic regression on the encoded features. In addition we use the supervised task test data to evaluate the variational lower bound of the data under the two models [Kingma and Welling, 2014]. We show the result of this experiment in Table 3.1. For each comparison we take a different subset of training data which helps demonstrate the variation in lower bound and accuracy when re-training the reference architecture. We use a significance value of 5% and indicate when the test favors one auto-encoder over another or fails to reject the null hypothesis. We find that the evaluation of the relative test of similarity for the models closely matches performance on the supervised task and the test set variational lower bound.

3.4.2 Generative Moment Matching Networks Architecture Experiments

We demonstrate our hypothesis test on a different class of deep generative models called Generative Moment Matching Networks (GMMN) [Li et al., 2015b]. This recently introduced model has shown competitive performance in terms of test set likelihood on the MNIST data. Furthermore the training of this model is based on the MMD criterion. Li et al. [2015b] proposes to use that model along with an auto-encoder, which is the setup we employ in this work. Here a standard auto-encoder model is trained on the data to obtain a low dimensional representation, then a GMMN network is trained on the latent representations (Figure 3.3).

We use the relative similarity test to evaluate various architectural choices in this new class of models. We start from the baseline model specified in Li et al. [2015b] and associated software. The details of the reference model are specified in Figure 3.3.

	RelativeMMD Preference				
Experimental Condition (A/B)	A	Inconclusive	B	Avg Likelihood A	Avg Likelihood B
Dropout/No Dropout	199	17	360	-9.01 ± 55.43	76.76 ± 42.83
More/Fewer GMMN Layers	105	14	393	-73.99 ± 40.96	**249.6 ± 8.07**
More/Fewer Nodes	450	13	113	**125.2 ± 43.4**	-57 ± 49.57
More/Fewer AE layers	231	21	324	41.78 ± 44.07	25.96 ± 55.85

Table 3.2 – For each experimental condition (e.g. dropout or no dropout) we show the number of times when the relative test of similarity prefers models in group 1 or 2 and number of inconclusive tests. We use the validation set as the target data for Relative MMD. An average likelihood for the MNIST test set for each group is shown with error bars. We can see that the MMD choices are in agreement with likelihood evaluations. Particularly we identify that models with fewer GMMN layers and models with more nodes have more favorable samples, which is confirmed by the likelihood results.

We vary the number of auto-encoder hidden layers (1 to 4), generative model layers (1, 4, or 5), the number of network nodes (all or 50% of the reference model), and use of drop-out on the auto-encoder. We use the same training set of 55000, validation set of 5000 and test set of 10000 as in Goodfellow et al. [2014], Li et al. [2015b]. In total we train 48 models. We use these to compare 4 simplified binary network architecture choices using the relative test of similarity: using dropout on the auto-encoder, few (1) or more (4 or 5) GMMN layers, few (1 or 2) or more (3 or 4) auto-encoder layers, and the number of network nodes. We use our test to compare these model settings using the *validation set* as the target in the relative similarity test, and samples from the models as the two sources. To validate our results we compare it to likelihoods computed on the test set. The results are shown in Table 3.2. We see that the likelihood results computed on a separate test set follow the conclusions obtained from MMD on the validation set. Particularly, we find that using fewer hidden layers for the GMMN and more hidden nodes generally produces better models.

3.4.3 Discussion

In these experiments we have seen that the relative similarity test can be used to compare deep generative models obtaining judgments aligned with other metrics. Comparisons to other metrics are important for verifying our test is sensible, but it can occlude the fact that MMD is a valid evaluation technique on its own. When evaluating only sample generating models where likelihood computation is not possible, MMD is an appropriate and tractable metric to consider in addition to Parzen-Window log likelihoods and visual appearance of the samples. In several ways it is potentially more appropriate than Parzen-windows as it allows one to consider directly the discrepancy between the test data samples and the model samples while allowing for significance results. In such a situation, comparing the performance of several models using the MMD against a single set of test samples, the RelativeMMD test can provide an automatic significance value without expensive cross-validation procedures.

Gaussian kernels are closely related to Parzen-window estimates, thus computing an MMD in this case can be considered related to comparing Parzen window log-likelihoods. The MMD

gives several advantages, however. First, the asymptotics of MMD are quite different to Parzen-windows, since the Parzen-window bandwidth shrinks as m grows. Asymptotics of relative tests with shrinking bandwidth are unknown: even for two samples this is challenging [Krishnamurthy et al., 2015]. Other two sample tests are not easily extendable to relative tests [Friedman and Rafsky, 1979, Hall and Tajvidi, 2002, Rosenbaum, 2005]. This is because the tests above rely on graph edge counting or nearest neighbor-type statistics, and null distributions are obtained via combinatorial arguments which are not easily extended from two to three samples. MMD is a U-statistic, hence its asymptotic behavior is much more easily generalized to multiple dependent statistics.

There are two primary advantages of the MMD over the variational lower bound, where it is known [Kingma and Welling, 2014]: first, we have a characterization of the asymptotic behavior, which allows us to determine when the difference in performance is significant; second, comparing two lower bounds produced from two different models is unreliable, as we do not know how conservative either lower bound is.

3.5 Conclusion

In this chapter, we have presented a study of a hypothesis test of relative similarity of distributions, and its application to model selection. We have described a novel non-parametric statistical hypothesis test for relative similarity based on the Maximum Mean Discrepancy. The test is consistent, and the computation time is quadratic. Our proposed test statistic is theoretically justified for the task of comparing samples from arbitrary distributions as it can be shown to converge to a quantity which compares all moments of the two pairs of distributions.

We evaluate test performance on synthetic data, where the degree of similarity can be controlled. Our experimental results on model selection for deep generative networks show that the relative test of similarity can be a useful approach to comparing such models. There is a strong correspondence between the test results and the expected likelihood, prediction accuracy, and variational lower bounds on the models tested. Moreover, our test has the advantage over these alternatives of providing guarantees of statistical significance to its conclusions. This suggests that the relative similarity test will be useful in evaluating hypotheses about network architectures, for example that AE-GMMN models may generalize better when fewer layers are used in the generative model.

3.6 Detailed Proofs

3.6.1 Proof of Theorem 3.2

Proof: We have that Equations (3.2.2) and (3.2.3) are a direct application of Theorems 2.1 and 2.3, respectively. We remind that the corresponding U-statistic kernels of degree 2 for

$\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]$ and $\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}]$ are, respectively (cf. Equation (2.3.8)),

$$h(\mathbf{u}_i, \mathbf{u}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{y}_i, \mathbf{y}_j) - k(\mathbf{x}_i, \mathbf{y}_j) - k(\mathbf{x}_j, \mathbf{y}_i), \text{ with } \mathbf{u} = (\mathbf{x}, \mathbf{y}) \sim \mathbb{P}_\mathbf{x} \times \mathbb{P}_\mathbf{y}, \quad (3.6.1)$$

$$g(\mathbf{v}_i, \mathbf{v}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{z}_i, \mathbf{z}_j) - k(\mathbf{x}_i, \mathbf{z}_j) - k(\mathbf{x}_j, \mathbf{z}_i), \text{ with } \mathbf{v} = (\mathbf{x}, \mathbf{z}) \sim \mathbb{P}_\mathbf{x} \times \mathbb{P}_\mathbf{z}. \quad (3.6.2)$$

Then the variance/covariance for a U-statistic with a kernel of order 2 is given by

$$\mathrm{Var}\left(\widehat{\mathrm{MMD}}_u^2\right) = \frac{4(n-2)}{n(n-1)}\zeta_1 + \frac{2}{n(n-1)}\zeta_2, \quad (3.6.3)$$

where ζ_1 and ζ_2 are given as in Equation (2.1.7). Moreover, Equation (3.6.3) with neglecting higher terms can be written as

$$Var\left(\widehat{\mathrm{MMD}}_u^2\right) = \frac{4(n-2)}{n(n-1)}\zeta_1 + \mathcal{O}(n^{-2}). \quad (3.6.4)$$

In Sutherland et al. [2016] and Sutherland [2016], they develop our approach including an estimate of ζ_2.

We now give an explicit expression for ζ_1 for the variance of $\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]$ and the covariance of $\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}]$ and $\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{z}]$ which follow from Equation (2.1.7). We note $[\tilde{\mathbf{K}}_{\mathbf{xx}}]_{ij} = [\mathbf{K}_{\mathbf{xx}}]_{ij}$ for all $i \neq j$ and $[\mathbf{K}_{\mathbf{xx}}]_{ij} = 0$ for $j = i$. Same for $\mathbf{K}_{\mathbf{yy}}$ and $\mathbf{K}_{\mathbf{zz}}$. We will also make use of the fact that $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ for an appropriately chosen inner product, and function ϕ.

Variance of $\widehat{\mathrm{MMD}}_u^2[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]$

$$\zeta_1 = \mathrm{Var}\left[\mathrm{E}_{\mathbf{u_2}}\left[h(\mathbf{u}_1, \mathbf{u}_2)\right]\right] \quad (3.6.5)$$

$$= \mathrm{E}_{\mathbf{u_1}}\left[\mathrm{E}_{\mathbf{u_2}}\left[\left\{h(\mathbf{u}_1, \mathbf{u}_2)\right\}^2\right]\right] - \left(\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{x}, \mathbb{P}_\mathbf{y}]\right)^2 \quad (3.6.6)$$

$$= \mathrm{E}_{\mathbf{x_1}, \mathbf{y_1}}\left[\left\{\mathrm{E}_{\mathbf{x_2}, \mathbf{y_2}}\left[k(\mathbf{x}_1, \mathbf{x}_2) + k(\mathbf{y}_1, \mathbf{y}_2) - k(\mathbf{x}_1, \mathbf{y}_2) - k(\mathbf{x}_2, \mathbf{y}_1)\right]\right\}^2\right] \quad (3.6.7)$$

$$- \left(\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{X}, \mathbb{P}_\mathbf{Y}]\right)^2$$

$$= \mathrm{E}_{\mathbf{x_1}, \mathbf{y_1}}\left[\left\{\mathrm{E}_{\mathbf{x_2}}\left[k(\mathbf{x}_1, \mathbf{x}_2)\right] + \mathrm{E}_{\mathbf{y_2}}\left[k(\mathbf{y}_1, \mathbf{y}_2)\right] - \mathrm{E}_{\mathbf{y_2}}\left[k(\mathbf{x}_1, \mathbf{y}_2)\right] - \mathrm{E}_{\mathbf{x_2}}\left[k(\mathbf{x}_2, \mathbf{y}_1)\right]\right\}^2\right] \quad (3.6.8)$$

$$- \left(\mathrm{MMD}^2[\mathcal{H}, \mathbb{P}_\mathbf{X}, \mathbb{P}_\mathbf{Y}]\right)^2.$$

Let first calculate the term $\mathrm{E}_{\mathbf{x_2}}[k(\mathbf{x}_1, \mathbf{x}_2)]$ in Equation (3.6.8). If the distribution from $\mathbb{P}_\mathbf{x}$

has a density f, then

$$\begin{aligned} \mathrm{E}_{\mathbf{x_2}}\left[k(\mathbf{x}_1, \mathbf{x}_2)\right] &= \mathrm{E}_{\mathbf{x_2}}\left[\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2)\rangle\right] \\ &= \int_{\mathbb{R}} \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) f(\mathbf{x}_2) \mathrm{d}\mathbf{x}_2 \rangle \\ &= \langle \phi(\mathbf{x}_1), \int_{\mathbb{R}} \phi(\mathbf{x}_2) f(\mathbf{x}_2) \mathrm{d}\mathbf{x}_2 \rangle \\ &= \langle \phi(\mathbf{x}_1), \overline{\mathbf{x}} \rangle, \end{aligned} \tag{3.6.9}$$

where $\overline{\mathbf{x}}$ is by definition the mean of the sample $\mathbf{x}$. Similarly, we have that the remaining terms in Equation (3.6.8) are

$$\mathrm{E}_{\mathbf{y_2}}\left[k(\mathbf{y}_1, \mathbf{y}_2)\right] = \langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle \tag{3.6.10}$$

$$\mathrm{E}_{\mathbf{y_2}}\left[k(\mathbf{x}_1, \mathbf{y}_2)\right] = \langle \phi(\mathbf{x}_1), \overline{\mathbf{y}} \rangle \tag{3.6.11}$$

$$\mathrm{E}_{\mathbf{x_2}}\left[k(\mathbf{x}_2, \mathbf{y}_1)\right] = \langle \phi(\mathbf{y}_1), \overline{\mathbf{x}} \rangle. \tag{3.6.12}$$

Finally, we have that

$$\zeta_1 = \mathrm{Var}\left[\mathrm{E}_{\mathbf{u_2}}\left[h(\mathbf{u}_1, \mathbf{u}_2)\right]\right] \tag{3.6.13}$$

$$= \mathrm{E}_{\mathbf{x_1}, \mathbf{y_1}}\left[\left\{ \langle \phi(\mathbf{x}_1), \overline{\mathbf{x}} \rangle + \langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle - \langle \phi(\mathbf{x}_1), \overline{\mathbf{y}} \rangle - \langle \overline{\mathbf{x}}, \phi(\mathbf{y}_1) \rangle \right\}^2\right] \tag{3.6.14}$$

$$- \left(\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right]\right)^2. \tag{3.6.15}$$

We note many terms in expansion of the squares in Equation (3.6.15) above cancel out due to independence. For example $\mathrm{E}_{\mathbf{x_1}, \mathbf{y_1}}\left[\langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle \langle \phi(\mathbf{x}_1), \overline{\mathbf{y}} \rangle\right] - \mathbb{E}_{\mathbf{y}_1}\left[\langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle\right] \mathbb{E}_{\mathbf{x}_1}\left[\langle \phi(\mathbf{x}_1), \overline{\mathbf{y}} \rangle\right] = 0$.

We can thus simplify to the following expression for ζ_1.

$$\zeta_1 = \mathrm{E}_{\mathbf{x_1},\mathbf{y_1}}\left[\left\{\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle + \langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle - \langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle - \langle\overline{\mathbf{x}},\phi(\mathbf{y}_1)\rangle\right\}^2\right] - \left(\mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{x}},\mathbb{P}_{\mathbf{y}}\right]\right)^2 \tag{3.6.16}$$

$$\begin{aligned}
&= \mathrm{E}_{\mathbf{x_1},\mathbf{y_1}}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle^2 + 2\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle - 2\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\right. \qquad (3.6.17)\\
&\quad - 2\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle + \langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle^2\\
&\quad - 2\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle - 2\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle\\
&\quad + \langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle^2 + 2\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle\\
&\quad \left. + \langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle^2\right] - \left(\mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{X}},\mathbb{P}_{\mathbf{Y}}\right]\right)^2\\
&= \mathrm{E}_{\mathbf{x_1},\mathbf{y_1}}\left[\langle x_1,\overline{\mathbf{x}}\rangle^2 - 2\langle\mathbf{x}_1,\overline{\mathbf{x}}\rangle\langle\mathbf{x}_1,\overline{\mathbf{y}}\rangle + \langle\mathbf{y}_1,\overline{\mathbf{y}}\rangle^2\right. \qquad (3.6.18)\\
&\quad \left. - 2\langle\mathbf{y}_1,\overline{\mathbf{y}}\rangle\langle y_1,\overline{\mathbf{x}}\rangle + \langle\mathbf{x}_1,\overline{\mathbf{y}}\rangle^2 + \langle\mathbf{y}_1,\overline{\mathbf{x}}\rangle^2\right] - \text{centering terms}\\
&= \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle^2] - \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle]^2 \qquad (3.6.19)\\
&\quad - 2(\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle] - \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle]\,\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle])\\
&\quad + \mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle^2] - \mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle]^2\\
&\quad - 2(\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle] - \mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle]\,\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle])\\
&\quad + \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle^2] - \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle]^2\\
&\quad + \mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle^2] - \mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle]^2.
\end{aligned}$$

We next substitute the kernel MMD definition from Equation (2.3.6), expand the terms in the expectation, and determine their empirical estimates in order to compute the variances in practice. Now, let derive the twelve terms of Equation (3.6.19)

1.

$$\begin{aligned}
\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle^2] &\approx \frac{1}{n}\sum_{i=1}^{n}\langle\phi(\mathbf{x}_i),\frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n}\phi(\mathbf{x}_j)\rangle\langle\phi(\mathbf{x}_i),\frac{1}{n-1}\sum_{\substack{k=1\\k\neq i}}^{n}\phi(\mathbf{x}_k)\rangle \qquad (3.6.20)\\
&= \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq i}}^{n}k(\mathbf{x}_i,\mathbf{x}_j)k(\mathbf{x}_i,\mathbf{x}_k)\\
&= \frac{1}{n(n-1)^2}\mathbf{1}\tilde{\mathbf{K}}_{\mathbf{xx}}\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}.
\end{aligned}$$

2.

$$\begin{aligned}
\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle]^2 &\approx \frac{1}{n^2(n-1)^2}\left(\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle\right)^2 \qquad (3.6.21)\\
&= \frac{1}{n^2(n-1)^2}\left(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\right)^2.
\end{aligned}$$

3.

$$\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle] \approx \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{k=1}^{n}\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle\langle\phi(\mathbf{x}_i),\phi(\mathbf{y}_k)\rangle \tag{3.6.22}$$
$$= \frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{K}_{\mathbf{xy}}\mathbf{1}.$$

4.

$$\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle]\,\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle] \approx \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle \times \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\phi(\mathbf{x}_i),\phi(\mathbf{y}_j)\rangle \tag{3.6.23}$$
$$= \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}.$$

5.

$$\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle^2] \approx \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{k=1\\k\neq i}}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_j)\rangle\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_k)\rangle \tag{3.6.24}$$
$$= \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}.$$

6.

$$\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle]^2 \approx \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_j)\rangle \times \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_j)\rangle \tag{3.6.25}$$
$$= \frac{1}{n^2(n-1)^2}\left(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\right)^2.$$

7.

$$\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle] \approx \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{k=1}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_j)\rangle\langle\phi(\mathbf{y}_i),\phi(\mathbf{x}_k)\rangle \tag{3.6.26}$$
$$= \frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1}.$$

8.

$$\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle]\,\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle] \approx \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_j)\rangle \times \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{x}_j)\rangle \tag{3.6.27}$$

$$= \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}.$$

9.

$$\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle^2] \approx \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\langle\phi(\mathbf{x}_i),\phi(\mathbf{y}_j)\rangle\langle\phi(\mathbf{x}_i),\phi(\mathbf{y}_k)\rangle \tag{3.6.28}$$

$$= \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1}.$$

10.

$$\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle]^2 \approx \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\phi(\mathbf{x}_i),\phi(\mathbf{y}_j)\rangle \times \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\phi(\mathbf{x}_i),\phi(\mathbf{y}_j)\rangle \tag{3.6.29}$$

$$= \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right)^2.$$

11.

$$\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle^2] \approx \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{x}_j)\rangle\langle\phi(\mathbf{y}_i),\phi(\mathbf{x}_k)\rangle \tag{3.6.30}$$

$$= \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1}.$$

12.

$$\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle]^2 \approx \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{x}_j)\rangle \times \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\phi(\mathbf{y}_i),\phi(\mathbf{x}_j)\rangle \tag{3.6.31}$$

$$= \frac{1}{n^4}\left(\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right)^2.$$

Finally, substituting empirical expectations over the data sample for the population expecta-

tions in Equation (3.6.19) gives

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{MMD}^2_{\mathbf{XY}}} = \frac{4(n-2)}{n(n-1)}\Bigg\{ & \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\right)^2 \\ & - 2\left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right) \\ & + \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\right)^2 \\ & - 2\left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right) \\ & + \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - 2\left(\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right)^2 + \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1}\Bigg\}. \end{aligned} \tag{3.6.32}$$

Using the order of operations implied by the parentheses in the following Equation (3.6.33), the computational cost of the empirical variance of is $\mathcal{O}(n^2)$.

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{MMD}^2_{\mathbf{XY}}} = \frac{4(n-2)}{n(n-1)}\Bigg\{ & \frac{1}{n(n-1)^2}(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})(\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}) - \left(\frac{1}{n(n-1)}\left((\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})\mathbf{1}\right)\right)^2 \\ & - 2\left(\frac{1}{n^2(n-1)}(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})(\mathbf{K}_{\mathbf{xy}}\mathbf{1}) - \frac{1}{n^3(n-1)}\left((\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})\mathbf{1}\right)\left(\mathbf{1}^T(\mathbf{K}_{\mathbf{xy}}\mathbf{1})\right)\right) \\ & + \frac{1}{n(n-1)^2}(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}})(\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}) - \left(\frac{1}{n(n-1)}\left((\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}})\mathbf{1}\right)\right)^2 \\ & - 2\left(\frac{1}{n^2(n-1)}(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}})(\mathbf{K}_{\mathbf{xy}}\mathbf{1}) - \frac{1}{n^3(n-1)}\left((\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}})\mathbf{1}\right)\left(\mathbf{1}^T(\mathbf{K}_{\mathbf{xy}}\mathbf{1})\right)\right) \\ & + \frac{1}{n^3}(\mathbf{1}^T\mathbf{K}_{\mathbf{xy}})(\mathbf{K}_{\mathbf{xy}}\mathbf{1}) - 2\left(\frac{1}{n^2}\left((\mathbf{1}^T\mathbf{K}_{\mathbf{xy}})\mathbf{1}\right)\right)^2 + \frac{1}{n^3}(\mathbf{1}^T\mathbf{K}_{\mathbf{xy}})(\mathbf{K}_{\mathbf{xy}}\mathbf{1})\Bigg\}. \end{aligned} \tag{3.6.33}$$

Covariance of $\widehat{\mathrm{MMD}}^2_u[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]$ and $\widehat{\mathrm{MMD}}^2_u[\mathcal{H}, \mathbf{X_n}, \mathbf{Z_n}]$ Using the same derivation than for the variance of $\widehat{\mathrm{MMD}}^2_u[\mathcal{H}, \mathbf{X_n}, \mathbf{Y_n}]$ above, we have

$$\zeta_1 = \mathrm{E}_{\mathbf{x_1},\mathbf{y_1},\mathbf{z_1}}\left[\mathrm{E}_{\mathbf{x_2},\mathbf{y_2},\mathbf{z_2}}\left[h(\mathbf{x}_1,\mathbf{y}_1)g(\mathbf{x}_1,\mathbf{z}_1)\right]\right] - \left(\mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{x}},\mathbb{P}_{\mathbf{y}}\right]\mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{x}},\mathbb{P}_{\mathbf{z}}\right]\right) \tag{3.6.34}$$

$$\begin{aligned} &= \mathrm{E}_{\mathbf{x_1},\mathbf{y_1},\mathbf{z_1}}[(\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle + \langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle - \langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle - \langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle) \\ &\qquad (\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle) + \langle\phi(\mathbf{z}_1),\overline{\mathbf{z}}\rangle - \langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle - \langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle)] \\ &\quad - \mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{x}},\mathbb{P}_{\mathbf{y}}\right]\mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{x}},\mathbb{P}_{\mathbf{z}}\right] \\ &= \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle^2\right] - \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\right]^2 \\ &\quad - \left(\mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle\right] - \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\right]\mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle\right]\right) \\ &\quad - \left(\mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\right] - \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{x}}\rangle\right]\mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\right]\right) \\ &\quad + \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle\right] - \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\right]\mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle\right] \end{aligned} \tag{3.6.35}$$

Using the similar derivations than above, we obtain the following approximation for ζ_1

$$\begin{aligned}\zeta_1 \approx & \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\right)^2 \\ & - \left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{K}_{\mathbf{xz}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right) \\ & - \left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right) \\ & + \left(\frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xz}}\mathbf{1} - \frac{1}{n^4}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right).\end{aligned} \tag{3.6.36}$$

Finally, the empirical estimate in order to compute the covariance term $\sigma_{\text{MMD}^2_{\mathbf{xy},\mathbf{xz}}}$ in Equation (3.6.36), neglecting higher order terms is

$$\begin{aligned}\hat{\sigma}_{\text{MMD}^2_{\mathbf{xy},\mathbf{xz}}} = & \frac{4(n-2)}{n(n-1)}\Bigg\{\frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\right)^2 \\ & - \left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{K}_{\mathbf{xz}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right) \\ & - \left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right) \\ & + \left(\frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xz}}\mathbf{1} - \frac{1}{n^4}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right)\Bigg\}.\end{aligned} \tag{3.6.37}$$

And again, using the order of operations implied by the parentheses in Equation (3.6.38), the computational cost of the empirical variance is $\mathcal{O}(n^2)$.

$$\begin{aligned}\hat{\sigma}_{\text{MMD}^2_{\mathbf{xy},\mathbf{xz}}} = & \frac{4(n-2)}{n(n-1)}\Bigg\{\frac{1}{n(n-1)^2}(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})(\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}) - \left(\frac{1}{n(n-1)}\left(\mathbf{1}^T(\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1})\right)\right)^2 \\ & - \left(\frac{1}{n^2(n-1)}(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})(\mathbf{K}_{\mathbf{xz}}\mathbf{1}) - \frac{1}{n^3(n-1)}\left((\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})\mathbf{1}\right)\left(\mathbf{1}^T(\mathbf{K}_{\mathbf{xz}}\mathbf{1})\right)\right) \\ & - \left(\frac{1}{n^2(n-1)}(\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}})(\mathbf{K}_{\mathbf{xy}}\mathbf{1}) - \frac{1}{n^3(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{xx}}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right) \\ & + \left(\frac{1}{n^3}(\mathbf{1}^T\mathbf{K}_{\mathbf{xy}})(\mathbf{K}_{\mathbf{xz}}\mathbf{1}) - \frac{1}{n^4}\left((\mathbf{1}^T\mathbf{K}_{\mathbf{xy}})\mathbf{1}\right)\left(\mathbf{1}^T(\mathbf{K}_{\mathbf{xz}}\mathbf{1})\right)\right)\Bigg\}.\end{aligned} \tag{3.6.38}$$

□

3.6.2 Derivation of the variance of the difference of two MMD statistics

In this section we propose an alternate strategy of deriving directly the variance of a u-statistic of the difference of MMDs with a joint variable. This formulation agrees with the derivation of the covariance matrix and subsequent projection, and provides extra insights.

Let $\mathbf{d_n} = (\mathbf{d}_1, ..., \mathbf{d}_n)$ be n i.i.d. random variables where $\mathbf{d} := (\mathbf{x}, \mathbf{y}, \mathbf{z}) \sim \mathbb{P}_\mathbf{x} \times \mathbb{P}_\mathbf{y} \times \mathbb{P}_\mathbf{z}$. Then

the difference of the unbiased estimators of $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right]$ and $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right]$ is given by

$$\widehat{\mathrm{MMD}_u^2}\left[\mathcal{H}, \mathbf{x}, \mathbf{y}\right] - \widehat{\mathrm{MMD}_u^2}\left[\mathcal{H}, \mathbf{x}, \mathbf{z}\right] = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} f(\mathbf{d}_i, \mathbf{d}_j) \tag{3.6.39}$$

with f, the kernel of $\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right] - \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right]$ of order 2 as follows

$$\begin{aligned} f(\mathbf{d}_1, \mathbf{d}_2) &= (k(\mathbf{x}_1, \mathbf{x}_2) + k(\mathbf{y}_1, \mathbf{y}_2) - k(\mathbf{x}_1, \mathbf{y}_2) - k(\mathbf{x}_2, \mathbf{y}_1)) \qquad (3.6.40) \\ &\quad - (k(\mathbf{x}_1, \mathbf{x}_2) + k(\mathbf{z}_1, \mathbf{z}_2) - k(\mathbf{x}_1, \mathbf{z}_2) - k(\mathbf{x}_2, \mathbf{z}_1)) \\ &= (k(\mathbf{y}_1, \mathbf{y}_2) - k(\mathbf{x}_1, \mathbf{y}_2) - k(\mathbf{x}_2, \mathbf{y}_1)) - (k(\mathbf{z}_1, \mathbf{z}_2) - k(\mathbf{x}_1, \mathbf{z}_2) - k(\mathbf{x}_2, \mathbf{z}_1)) \qquad (3.6.41) \end{aligned}$$

Equation (3.6.39) is a U-statistic and thus we can apply Equation (3.2.2) to obtain its variance. We denote $\mathrm{Var}\left(\widehat{\mathrm{MMD}_u^2}\left[\mathcal{H}, \mathbf{x}, \mathbf{y}\right] - \widehat{\mathrm{MMD}_u^2}\left[\mathcal{H}, \mathbf{x}, \mathbf{z}\right]\right)$ by $\hat{\sigma}_{\mathrm{MMD}^2_{\mathbf{xy}} - \mathrm{MMD}^2_{\mathbf{xz}}}$

$$\hat{\sigma}_{\mathrm{MMD}^2_{\mathbf{xy}} - \mathrm{MMD}^2_{\mathbf{xz}}} = \frac{4(n-2)}{n(n-1)} \zeta_1 + \mathcal{O}(n^{-2}). \tag{3.6.42}$$

We first note

$$\begin{aligned} \mathbb{E}_{\mathbf{d}_1}(f(\mathbf{d}_1, \mathbf{d}_2)) =& \langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle - \langle \phi(\mathbf{x}_1), \overline{\mathbf{y}} \rangle - \langle \overline{\mathbf{x}}, \phi(\mathbf{y}_1) \rangle \qquad (3.6.43) \\ & - (\langle \phi(\mathbf{z}_1), \overline{\mathbf{z}} \rangle - \langle \phi(\mathbf{x}_1), \overline{\mathbf{z}} \rangle - \langle \overline{\mathbf{x}}, \phi(\mathbf{z}_1) \rangle) \\ \mathbb{E}_{\mathbf{d}_1, \mathbf{d}_2}(f(\mathbf{d}_1, \mathbf{d}_2)) =& \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right] - \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right]. \qquad (3.6.44) \end{aligned}$$

We are now ready to derive the dominant leading term ζ_1, in the variance expression of Equation (3.6.42).

$$\begin{aligned} \zeta_1 &= \mathrm{Var}(\mathbb{E}_{\mathbf{d}_1}(f(\mathbf{d}_1, \mathbf{d}_2))) \qquad (3.6.45) \\ &= \mathrm{E}_{\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1}[(\langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle - \langle \phi(\mathbf{x}_1), \overline{\mathbf{y}} \rangle - \langle \overline{\mathbf{x}}, \phi(\mathbf{y}_1) \rangle - (\langle \phi(\mathbf{z}_1), \overline{\mathbf{z}} \rangle - \langle \phi(\mathbf{x}_1), \overline{\mathbf{z}} \rangle - \langle \overline{\mathbf{x}}, \phi(\mathbf{z}_1) \rangle)^2] \\ &\quad - \left(\mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{y}}\right] - \mathrm{MMD}^2\left[\mathcal{H}, \mathbb{P}_{\mathbf{x}}, \mathbb{P}_{\mathbf{z}}\right]\right)^2 \qquad (3.6.46) \end{aligned}$$

We note many terms in expansion of the squares above cancel out due to independence. For example $\mathbb{E}_{\mathbf{y}_1, \mathbf{z}_1}[\langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle \langle \phi(\mathbf{z}_1), \overline{\mathbf{z}} \rangle] - \mathrm{E}_{\mathbf{y}_1}[\langle \phi(\mathbf{y}_1), \overline{\mathbf{y}} \rangle] \, \mathrm{E}_{\mathbf{z}_1}[\langle \phi(\mathbf{z}_1), \overline{\mathbf{z}} \rangle] = 0$.

We can thus simplify to the following expression for ζ_1

$$\begin{aligned}\zeta_1 = &\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle^2] - \mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle]^2 \\ &+ \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle^2] - \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle]^2 \\ &+ \mathrm{E}_{\mathbf{y}_1}[\langle\overline{\mathbf{x}},\phi(\mathbf{y}_1)\rangle^2] - \mathrm{E}_{\mathbf{y}_1}[\langle\overline{\mathbf{x}},\phi(\mathbf{y}_1)\rangle]^2 \\ &+ \mathrm{E}_{\mathbf{z}_1}[\langle\phi(\mathbf{z}_1),\overline{\mathbf{z}}\rangle^2] - \mathrm{E}_{\mathbf{z}_1}[\langle\phi(\mathbf{z}_1),\overline{\mathbf{z}}\rangle]^2 \\ &+ \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle^2] - \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle]^2 \\ &+ \mathrm{E}_{\mathbf{z}_1}[\langle\overline{\mathbf{x}},\phi(\mathbf{z}_1)\rangle^2] - \mathrm{E}_{\mathbf{z}_1}[\langle\overline{\mathbf{x}},\phi(\mathbf{z}_1)\rangle]^2 \\ &- 2(\mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\langle\overline{\mathbf{x}},\phi(\mathbf{y}_1)\rangle] - \mathrm{E}_{\mathbf{y}_1}[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle]\,\mathrm{E}_{\mathbf{y}_1}[\langle\overline{\mathbf{x}},\phi(\mathbf{y}_1)\rangle]) \\ &- 2(\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle] - \mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle]\,\mathrm{E}_{\mathbf{x}_1}[\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle]) \\ &- 2(\mathrm{E}_{\mathbf{z}_1}[\langle\phi(\mathbf{z}_1),\overline{\mathbf{z}}\rangle\langle\overline{\mathbf{x}},\phi(\mathbf{z}_1)\rangle] - \mathrm{E}_{\mathbf{z}_1}[\langle\phi(\mathbf{z}_1),\overline{\mathbf{z}}\rangle]\,\mathrm{E}_{\mathbf{z}_1}[\langle\overline{\mathbf{x}},\phi(\mathbf{z}_1)\rangle]).\end{aligned} \tag{3.6.47}$$

We can empirically approximate these terms as follows and we have

$$\begin{aligned}\hat{\sigma}_{\mathrm{MMD}^2_{\mathbf{xy}}-\mathrm{MMD}^2_{\mathbf{xz}}} = \frac{4(n-2)}{n(n-1)}\Bigg\{&\frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\right)^2 \\ &+ \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}{}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \left(\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right)^2 \\ &+ \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}{}^T\mathbf{1} - \left(\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right)^2 \\ &+ \frac{1}{n(n-1)^2}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{zz}}\tilde{\mathbf{K}}_{\mathbf{zz}}\mathbf{1} - \left(\frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{zz}}\mathbf{1}\right)^2 \\ &+ \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{K}_{\mathbf{xz}}{}^T\mathbf{1} - \left(\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right)^2 \\ &+ \frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}{}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1} - \left(\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right)^2 \\ &- 2\left(\frac{1}{n^2(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{K}_{\mathbf{xy}}\mathbf{1} - \frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\times\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right) \\ &- 2\left(\frac{1}{n^3}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{K}_{\mathbf{xy}}{}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1} - \frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\times\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xz}}\mathbf{1}\right) \\ &- 2\left(\frac{1}{n^2(r-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{zz}}\mathbf{K}_{\mathbf{xz}}{}^T\mathbf{1} - \frac{1}{n(n-1)}\mathbf{1}^T\tilde{\mathbf{K}}_{\mathbf{yy}}\mathbf{1}\times\frac{1}{n^2}\mathbf{1}^T\mathbf{K}_{\mathbf{xy}}\mathbf{1}\right)\Bigg\}.\end{aligned} \tag{3.6.48}$$

3.6.3 Equality

In this section, we prove that Equation (3.2.9) is equal to the variance of the difference of $\mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{x}},\mathbb{P}_{\mathbf{y}}\right]$ and $\mathrm{MMD}^2\left[\mathcal{H},\mathbb{P}_{\mathbf{x}},\mathbb{P}_{\mathbf{z}}\right]$.

$$\begin{aligned}\sigma^2_{XY} + \sigma_{XZ} - 2\sigma_{XYXZ} = &\mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle^2\right] - \mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\right]^2 \\ &+ \mathrm{E}_{\mathbf{z}_1}\left[\langle\phi(\mathbf{z}_1),\overline{\mathbf{y}}\rangle^2\right] - \mathrm{E}_{\mathbf{z}_1}\left[\langle\phi(\mathbf{z}_1),\overline{\mathbf{y}}\rangle\right]^2 \\ &- 2\left(\mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle\right] - \mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{y}}\rangle\right]\mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle\right]\right) \\ &- 2\left(\mathrm{E}_{\mathbf{z}_1}\left[\langle\phi(\mathbf{z}_1),\overline{\mathbf{z}}\rangle\langle\phi(\mathbf{z}_1),\overline{\mathbf{x}}\rangle\right] - \mathrm{E}_{\mathbf{z}_1}\left[\langle\phi(\mathbf{z}_1),\overline{\mathbf{z}}\rangle\right]\mathrm{E}_{\mathbf{z}_1}\left[\langle\phi(\mathbf{z}_1),\overline{\mathbf{x}}\rangle\right]\right) \\ &+ \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle^2\right] - \mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\right]^2 \\ &+ \mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{z}}\rangle^2\right] - \mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{z}}\rangle\right]^2 \\ &+ \mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle^2\right] - \mathrm{E}_{\mathbf{y}_1}\left[\langle\phi(\mathbf{y}_1),\overline{\mathbf{x}}\rangle\right]^2 \\ &+ \mathrm{E}_{\mathbf{z}_1}\left[\langle\phi(\mathbf{z}_1),\overline{\mathbf{x}}\rangle^2\right] - \mathrm{E}_{\mathbf{z}_1}\left[\langle\phi(\mathbf{z}_1),\overline{\mathbf{x}}\rangle\right]^2 \\ &- 2\left(\mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{y}}\rangle\right]\mathrm{E}_{\mathbf{x}_1}\left[\langle\phi(\mathbf{x}_1),\overline{\mathbf{z}}\rangle\right]\right).\end{aligned} \tag{3.6.49}$$

We have shown that Equation (3.2.9) is equal to Equation (3.6.49).

3.6.4 Calibration of the test

We show here that our derived test is well calibrated. A calibrated test should output a uniform distribution of p-values when the two MMD distances are equal. The empirical distributions of p-values for various sets of $\mathbb{P}_{\mathbf{x}}$, $\mathbb{P}_{\mathbf{y}}$ and $\mathbb{P}_{\mathbf{z}}$ are given in Figure 3.6. Similarly, for a given significance level α, the false positive rate should be equal to α. The empirical false positive rates for varying α are shown in Figure 3.5 further demonstrating the proper calibration of the test.

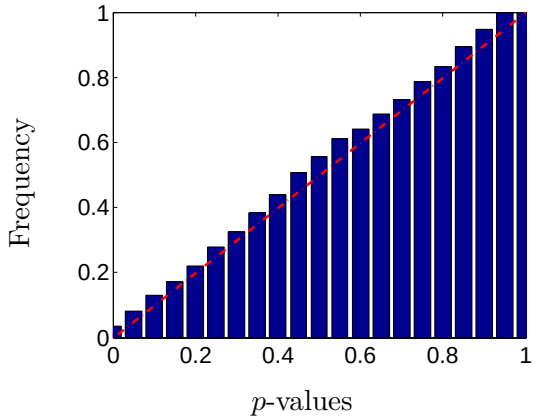
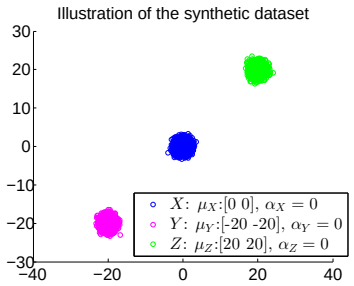
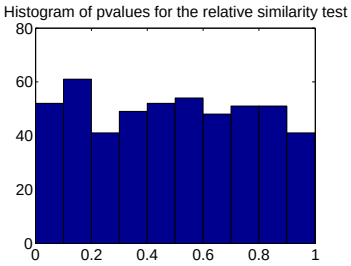


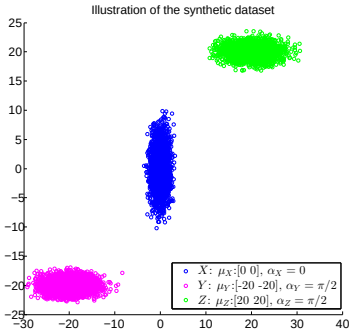
Figure 3.5 – Verification of the calibration of the relative similarity test. Here we demonstrate that the empirical frequency of p-values equals the significance level α.



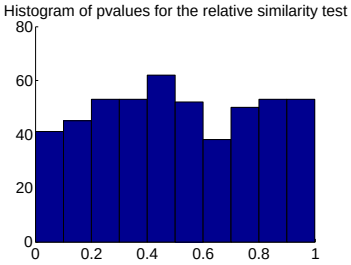
(a) Illustration of the synthetic data with different means for $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.



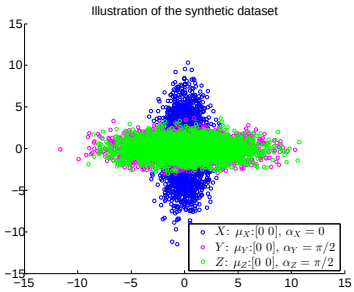
(b) Uniform histogram of p-values



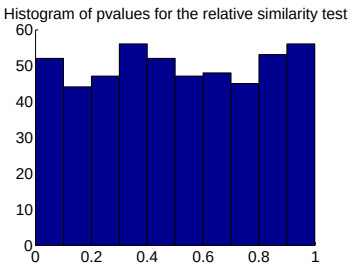
(c) Illustration of the synthetic data with different means and orientations for $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.



(d) Uniform histogram of p-values



(e) Illustration of the synthetic data with different orientations for $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.



(f) Uniform histogram of p-values

Figure 3.6 – Verification of the calibration of the relative similarity test. In all cases, the two target distributions are constructed to be equally distant from the source distribution. A well calibrated test should consequently produce a uniform distribution of p-values.

Chapter 4

A Hypothesis Test of Relative Dependency

In the previous chapter, we investigated the problem of relative similarity. The main result was a novel non-parametric statistical hypothesis test of relative similarity using the asymptotic distribution of two MMD statistics. By using the HSIC statistics to measure the dependency between two probabilistic distributions, it is natural to derive a novel-non parametric statistical hypothesis test of relative dependency. In this chapter, we present our approach to tackle the second Research Question 2: Is the dependency between $\mathbf{x}$ and $\mathbf{y}$ stronger than the dependency between $\mathbf{x}$ and $\mathbf{z}$?

We describe a novel non-parametric statistical hypothesis test of relative dependence between a source variable and two candidate target variables. Such a test enables us to determine whether one source variable is significantly more dependent on a first target variable or a second. Dependence is measured via the Hilbert-Schmidt Independence Criterion (HSIC), resulting in a pair of empirical dependence measures (source-target 1, source-target 2). We test whether the first dependence measure is significantly larger than the second. Modeling the covariance between these HSIC statistics leads to a provably more powerful test than the construction of independent HSIC statistics by sub-sampling. The resulting test is consistent and unbiased, and being based on U-statistics has favorable convergence properties. The test can be computed in quadratic time, matching the computational complexity of standard empirical HSIC estimators. The effectiveness of the test is demonstrated on several real-world problems: we identify language groups from a multilingual corpus, and we prove that tumor location is more dependent on gene expression than chromosomal imbalances.

The work covered in this chapter is based on:

- W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. A low variance consistent test of relative dependency. In F. Bach and D. Blei, editors, *Proceedings of The 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 20–29, 2015a.

Project: https://github.com/wbounliphone/reldep.

Contents

4.1 Introduction

Tests of dependence are important tools in statistical analysis, and are widely applied in many data analysis contexts as described in the introduction of Section 2.4.

For many problems in data analysis, however, the question of whether dependence exists is secondary: there may be multiple dependencies, and the question becomes which dependence is the strongest. For instance, in neuroscience, multiple stimuli may be present (e.g. visual and audio), and it is of interest to determine which of the two has a stronger influence on brain activity [Trommershauser et al., 2011]. In automated translation [Peters et al., 2012], it is of interest to determine whether documents in a source language are a significantly better match to those in one target language than to another target language, either as a measure of difficulty of the respective learning tasks, or as a basic tool for comparative linguistics.

We present a statistical test which determines whether two target variables have a significant difference in their dependence on a third, source variable. The dependence between each of the target variables and the source is computed using the Hilbert-Schmidt Independence Criterion (HSIC). Care must be taken in analyzing the asymptotic behavior of the test statistics, since the two measures of dependence will themselves be correlated: they are both computed with respect to the same source. Thus, we derive the *joint* asymptotic distribution of both dependencies. The derivation of our test utilizes classical results of U-statistics [Arcones and Gine, 1993, Hoeffding, 1963, Serfling, 2009]. In particular, we make use of results by Hoeffding [1963] and Serfling [2009] to determine the asymptotic joint distributions of the statistics (see Theorem 2.4) as described in Section 2.1. Consequently, we derive the *lowest* variance unbiased estimator of the test statistic.

We prove our approach to have greater statistical power than constructing two uncorrelated statistics on the same data by subsampling, and testing on these. In experiments, we are able to successfully test which of two variables is most strongly related to a third, in synthetic examples, in a language group identification task, and in a task for identifying the relative strength of factors for Glioma type in a pediatric patient population.

For the formal setup of this whole chapter, suppose that we have random variables $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$, $\mathbf{y} \sim \mathbb{P}_{\mathbf{y}}$ and $\mathbf{z} \sim \mathbb{P}_{\mathbf{z}}$, that take values on $(\mathcal{X}, \mathcal{B}_{\mathbf{x}})$, $(\mathcal{Y}, \mathcal{B}_{\mathbf{y}})$ and $(\mathcal{Z}, \mathcal{B}_{\mathbf{z}})$ respectively, here $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are three separable metric and $\mathcal{B}_{\mathbf{x}}, \mathcal{B}_{\mathbf{y}}, \mathcal{B}_{\mathbf{z}}$ are Borel σ-algebras. Then, $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{y}})$ and $(\mathcal{X} \times \mathcal{Z}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{z}})$ are again measurable and the joint distribution is $\mathbb{P}_{\mathbf{xy}}$ and $\mathbb{P}_{\mathbf{xz}}$ is taking values in the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{y}})$ and $(\mathcal{X} \times \mathcal{Z}, \mathcal{B}_{\mathbf{x}} \times \mathcal{B}_{\mathbf{z}})$ respectively. We denote the observations $\mathbf{X_n} := \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, $\mathbf{Y_n} := \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ and $\mathbf{Z_n} := \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ of size n, drawn i.i.d. from $\mathbb{P}_{\mathbf{x}}$, $\mathbb{P}_{\mathbf{y}}$ and $\mathbb{P}_{\mathbf{z}}$ respectively and $\mathbf{S_n} = (\mathbf{X_n}, \mathbf{Y}, \mathbf{Z_n})$ the joint sample which are drawn i.i.d from $\mathbb{P}_{\mathbf{xyz}}$. Furthermore, we define kernels $k(.,.)$, $l(.,.)$ and $m(.,.)$ on the space $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, and denote the corresponding RKHSs with $\mathcal{H}_{\mathbf{x}}$, $\mathcal{H}_{\mathbf{y}}$ and $\mathcal{H}_{\mathbf{z}}$ respectively and $\mathbf{K}$, $\mathbf{L}$ and $\mathbf{M} \in \mathbb{R}^{n \times n}$ are kernel matrices containing $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $l_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$ and $m_{ij} = m(\mathbf{z}_i, \mathbf{z}_j)$. Throughout, we assume the integrability $\mathbb{E}_{\mathbf{x}}[k] < \infty$ and $\mathbb{E}_{\mathbf{y}}[h] < \infty$.

Let's now reformulate our Research Question 2 into a mathematical setting. In Section 2.4, we have shown that HSIC determines independence: HSIC $= 0$ if and only if $\mathbb{P}_{\mathbf{xy}} = \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$ when kernels k and l are characteristic on their respective marginal domains. With this choice, the problem we would like to solve is described as follows:

Problem 4.1. *Given separable RKHSs $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ with* $\mathrm{HSIC}[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}] > 0$ *and* $\mathrm{HSIC}[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}] > 0$*, the statistical relative independence test $\mathcal{T}_{HSIC} : \mathcal{X}^n \times \mathcal{X}^n \times \mathcal{X}^n \mapsto \{0, 1\}$ is use to test the null hypothesis*

$$\mathcal{H}_0^{HSIC} : \mathrm{HSIC}[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}] \leq \mathrm{HSIC}[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z},}, \mathbb{P}_{\mathbf{xz},}] \tag{4.1.1}$$

versus the alternative hypothesis

$$\mathcal{H}_1^{HSIC} : \mathrm{HSIC}[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}] > \mathrm{HSIC}[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}], \tag{4.1.2}$$

at a given significance level α.

4.2 A Test of Relative Dependence

In this section we calculate two dependent HSIC statistics and derive the joint asymptotic distribution of these dependent quantities, which is used to construct a consistent test for Problem 4.1. This is a direct application of Theorem 2.4 using the definition of HSIC written as a U-statisticof degree 4. We next construct a simpler consistent test, by computing two independent HSIC statistics on sample subsets. While the simpler strategy is superficially attractive and less effort to implement, we prove the dependent strategy is strictly more powerful.

4.2.1 Joint Asymptotic Distribution of HSIC

In the present section, we compute each HSIC estimate on the full dataset, and explicitly obtain the correlations between the resulting empirical dependence measurements. Let $\widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]$ and $\widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]$ be respectively the unbiased estimators of $\text{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right]$ and $\text{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]$, written as a sum of U-statistics with respective U-statistic kernels h_{ijqr} and g_{ijqr} of degree 4 as described in Equation (2.4.10),

$$h_{ijqr} = \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(l_{st} + l_{uv} - 2l_{su}), \tag{4.2.1}$$

$$g_{ijqr} = \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} k_{st}(d_{st} + d_{uv} - 2d_{su}). \tag{4.2.2}$$

Theorem 4.1 (Joint asymptotic distribution of two HSIC)**.** *If* $\mathbb{E}[h^2] < \infty$ *and* $\mathbb{E}[g^2] < \infty$*, then the multivariate vector* $\left[\text{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right]; \text{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]\right]^T$ *converges asymptotically in distribution to a multivariate normal distribution with mean vector zero and* $\mathbf{\Sigma}$ *the limiting covariance matrix as following*

$$n^{1/2}\left(\begin{pmatrix}\widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] \\ \widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]\end{pmatrix} - \begin{pmatrix}\text{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] \\ \text{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_{\text{HSIC}_{\mathbf{xy}}} & \sigma_{\text{HSIC}_{\mathbf{xy,xz}}} \\ \sigma_{\text{HSIC}_{\mathbf{xy,xz}}} & \sigma^2_{\text{HSIC}_{\mathbf{xz}}}\end{pmatrix}\right), \tag{4.2.3}$$

where $\sigma^2_{\text{HSIC}_{\mathbf{XY}}}$ *and* $\sigma^2_{\text{HSIC}_{\mathbf{XY}}}$ *are as in Equation (2.4.14). The empirical estimate of* $\sigma_{\text{HSIC}_{\mathbf{xy,xz}}}$ *is* $\hat{\sigma}_{\text{HSIC}_{\mathbf{xy,xz}}} = \frac{16}{n}\left(R_{\mathbf{XYXZ}} - \widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]\widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]\right)$*, where*

$$R_{\mathbf{XYXZ}} = \frac{1}{n}\sum_{i=1}^{n}\left((n-1)_3^{-2}\sum_{(j,q,r)\in i_3^n\setminus\{i\}} h_{ijqr}g_{ijqr}\right), \tag{4.2.4}$$

where h_{ijqr} *and* g_{ijqr} *are the U-statistic kernel of degree 4 described in Equations (4.2.1) and (4.2.2) and the index set* $i_3^n\setminus\{i\}$ *denotes the set of all* $3-$*tuples drawn without replacement from the set* $\{1,\dots n\}\setminus\{i\}$*.*

Proof: Equation (4.2.3) follows from the application of [Hoeffding, 1963, Theorem 7.1], described in Theorem 2.4, which gives the joint asymptotic distribution of U-statistics. And Equation (4.2.4) is a direct application of Theorem 2.3 and is constructed with the definition of the variance of a U-statistic as given by [Serfling, 2009, Ch. 5] where one variable is fixed. □

Based on the joint asymptotic distribution of HSIC described in Theorem 4.1, we can now describe a statistical test to solve Problem 4.1 described above. This is achieved by projecting the distribution to 1D using the statistic $\widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] - \widehat{\text{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]$,

and determining where the statistic falls relative to a conservative estimate of the $1-\alpha$ quantile of the null $\mathcal{H}_0$. We now derive this conservative estimate. A simple way of achieving this is to rotate the distribution by $\frac{\pi}{4}$ counter-clockwise about the origin, and to integrate the resulting distribution projected onto the first axis (cf. Figure 4.3). Let's denote the asymptotically normal distribution of $n^{1/2}\left[\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]; \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]\right]^T$ as $\mathcal{N}_2\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. The distribution resulting from rotation and projection is

$$\mathcal{N}\left([\mathbf{Q}\boldsymbol{\mu}]_1, [\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^T]_{11}\right), \tag{4.2.5}$$

where $\mathbf{Q} = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ is the rotation matrix by $\pi/4$ and similarly to Section 3.2.2, we have

$$[\mathbf{Q}\mu]_1 = \frac{\sqrt{2}}{2}\left(\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]\right); \tag{4.2.6}$$

$$[\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^T]_{11} = \frac{1}{2}\left(\sigma^2_{\mathrm{HSIC}_{\mathbf{xy}}} + \sigma^2_{\mathrm{HSIC}_{\mathbf{xz}}} - 2\sigma_{\mathrm{HSIC}_{\mathbf{xy},\mathbf{xz}}}\right). \tag{4.2.7}$$

Following the empirical distribution from Equation (4.2.5), a test where the test statistic is the difference of two HSIC estimator $\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]$ has p-value

$$p \leq 1 - \boldsymbol{\Phi}\left(\frac{\left(\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]\right)}{\sqrt{\sigma^2_{\mathrm{HSIC}_{\mathbf{xy}}} + \sigma^2_{\mathrm{HSIC}_{\mathbf{xz}}} - 2\sigma_{\mathrm{HSIC}_{\mathbf{xy},\mathbf{xz}}}}}\right), \tag{4.2.8}$$

where $\boldsymbol{\Phi}$ is the CDF of a standard normal distribution, and we have made the most conservative possible assumption that $\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right] = 0$ under the null (the null also allows for the difference in population dependence measures to be negative).

To implement the test in practice, the variances of $\sigma^2_{\mathrm{HSIC}_{\mathbf{xy}}}, \sigma^2_{\mathrm{HSIC}_{\mathbf{xz}}}$ and $\sigma_{\mathrm{HSIC}_{\mathbf{XY},\mathbf{XZ}}}$ may be replaced by their empirical estimates. The test will still be consistent for a large enough sample size, since the estimates will be sufficiently well converged to ensure the test is calibrated.

Equation (4.2.4) is expensive to compute naïvely, because even computing the kernels h_{ijqr} and g_{ijqr} of the U-statistic itself is a non trivial task. Following Song et al. [2012, Section 2.5], we first form a vector $\mathbf{h_{XY}}$ with entries corresponding to $\sum_{(j,q,r)\in i_3^n\setminus\{i\}} h_{ijqr}$, and a vector $\mathbf{h_{XZ}}$ with entries corresponding to $\sum_{(j,q,r)\in i_3^n\setminus\{i\}} g_{ijqr}$. Collecting terms in Equations (4.2.1) and (4.2.2) related to kernel matrices $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$, $\mathbf{h_{XY}}$ can be written as

$$\begin{aligned}\mathbf{h_{XY}} = {} & (n-2)^2\left(\tilde{\mathbf{K}} \odot \tilde{\mathbf{L}}\right)\mathbf{1} - n(\tilde{\mathbf{K}}\mathbf{1}) \odot (\tilde{\mathbf{L}}\mathbf{1}) \\ & + (n-2)\left((\mathrm{Tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}))\mathbf{1} - \tilde{\mathbf{K}}(\tilde{\mathbf{L}}\mathbf{1}) - \tilde{\mathbf{L}}(\tilde{\mathbf{K}}\mathbf{1})\right) \\ & + (\mathbf{1}^T\tilde{\mathbf{L}}\mathbf{1})\tilde{\mathbf{K}}\mathbf{1} + (\mathbf{1}^T\tilde{\mathbf{K}}\mathbf{1})\tilde{\mathbf{L}}\mathbf{1} - ((\mathbf{1}^T\tilde{\mathbf{K}})(\tilde{\mathbf{L}}\mathbf{1}))\mathbf{1},\end{aligned} \tag{4.2.9}$$

where $\odot$ denotes the Hadamard product. Then $R_{\mathbf{XYXZ}}$ in Equation (4.2.4) can be computed as $R_{\mathbf{XYXZ}} = (4n)^{-1}(n-1)_3^{-2}\mathbf{h_{XY}}^T\mathbf{h_{XZ}}$. Using the order of operations implied by the parentheses in Equation (4.2.9), the computational cost of the cross covariance term is

$\mathcal{O}(n^2)$. Combining this with the unbiased estimator of HSIC in Equation (2.4.8) leads to a final computational complexity of $\mathcal{O}(n^2)$.

In addition to the asymptotic consistency result, we provide a finite sample bound on the deviation between the difference of two population HSIC statistics and the difference of two empirical HSIC estimates.

Theorem 4.2 (Generalization bound on the difference of empirical HSIC statistics)**.** *Assume that k, l, and d are bounded almost everywhere by 1, and are non-negative. Then for $n > 1$ and all $\delta > 0$ with probability at least $1 - \delta$, for all $\mathbb{P}_{\mathbf{xy}}$ and $\mathbb{P}_{\mathbf{xz}}$, the generalization bound on the difference of empirical HSIC statistics is*

$$\Big| \{\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]\} \tag{4.2.10}$$

$$- \left\{\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]\right\} \Big|$$

$$\leq 2\left\{\sqrt{\frac{\log(6/\delta)}{\alpha^2 n}} + \frac{C}{n}\right\}, \tag{4.2.11}$$

where $\alpha > 0.24$ and C are constants.

Proof: In Gretton et al. [2005a] a finite sample bound is given for a single HSIC statistic. Equation (4.2.10) is proved by using a union bound:

$$\Big| \{\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]\} \tag{4.2.12}$$

$$- \left\{\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]\right\} \Big|$$

$$= \Big| \left\{\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]\right\}$$

$$+ \left\{\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]\right\} \Big|$$

$$\leq \left|\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right]\right|$$

$$+ \left|\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right] - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]\right|$$

$$\leq 2\left\{\sqrt{\frac{log(6/\delta)}{\alpha^2 m}} + \frac{C}{n}\right\}.$$

□

Corollary 4.1. $\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, (\mathbf{X_n}, \mathbf{Y_n})\right] - \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, (\mathbf{X_n}, \mathbf{Z_n})\right]$ *converges to the population statistic at rate $\mathcal{O}(n^{1/2})$.*

4.2.2 A Simple and Consistent Statistical Test via Two Uncorrelated HSIC Statistics

From the result of the joint asymptotic distribution of the two HSIC statistics in Equation (4.2.3), a simple, consistent test of relative dependence can be constructed as follows: split the samples from $\mathbb{P}_x$ into two equal sized sets denoted by $\mathbf{X}'_{\mathbf{n/2}}$ and $\mathbf{X}''_{\mathbf{n/2}}$, and drop the second half of the sample pairs with $\mathbf{Y_n}$ and the first half of the sample pairs with $\mathbf{Z_n}$. We will denote the remaining samples as $\mathbf{Y}'_{\mathbf{n/2}}$ and $\mathbf{Z}''_{\mathbf{n/2}}$. We can now estimate the joint distribution of $n^{1/2}\left[\widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \left(\mathbf{X}'_{\mathbf{n/2}}, \mathbf{Y}'_{\mathbf{n/2}}\right)\right]; \widehat{\mathrm{HSIC}}_u\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \left(\mathbf{X}''_{\mathbf{n/2}}, \mathbf{Z}''_{\mathbf{n/2}}\right)\right]\right]^T$ as

$$\mathcal{N}_2\left(\begin{pmatrix}\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}'}, \mathcal{H}_{\mathbf{y}'}, \mathbb{P}_{\mathbf{x}'\mathbf{y}'}\right] \\ \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}''}, \mathcal{H}_{\mathbf{z}''}, \mathbb{P}_{\mathbf{x}''\mathbf{z}''}\right]\end{pmatrix}, \begin{pmatrix}\sigma^2_{\mathrm{HSIC}_{\mathbf{x}'\mathbf{y}'}} & 0 \\ 0 & \sigma^2_{\mathrm{HSIC}_{\mathbf{x}''\mathbf{z}''}}\end{pmatrix}\right), \tag{4.2.13}$$

which we will write as $\mathcal{N}\left(\boldsymbol{\mu}', \boldsymbol{\Sigma}'\right)$. Given this joint distribution, we need to determine the distribution over the half space defined by $\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{y}}, \mathbb{P}_{\mathbf{xy}}\right] < \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}}, \mathcal{H}_{\mathbf{z}}, \mathbb{P}_{\mathbf{xz}}\right]$. As in the previous section, we achieve this by rotating the distribution by $\frac{\pi}{4}$ counter-clockwise about the origin, and integrating the resulting distribution projected onto the first axis. The resulting projection of the rotated distribution onto the primary axis is

$$\mathcal{N}\left(\left[\mathbf{Q}\boldsymbol{\mu}'\right]_1, \left[\mathbf{Q}\boldsymbol{\Sigma}'\mathbf{Q}^T\right]_{11}\right) \tag{4.2.14}$$

where

$$\left[\mathbf{Q}\boldsymbol{\mu}'\right]_1 = \frac{\sqrt{2}}{2}\left(\mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}'}, \mathcal{H}_{\mathbf{y}'}, \mathbb{P}_{\mathbf{x}'\mathbf{y}'}\right] - \mathrm{HSIC}\left[\mathcal{H}_{\mathbf{x}''}, \mathcal{H}_{\mathbf{z}''}, \mathbb{P}_{\mathbf{x}''\mathbf{z}''}\right]\right), \tag{4.2.15}$$

$$\left[\mathbf{Q}\boldsymbol{\Sigma}'\mathbf{Q}^T\right]_{11} = \frac{1}{2}\left(\sigma^2_{\mathrm{HSIC}_{\mathbf{x}'\mathbf{y}'}} + \sigma^2_{\mathrm{HSIC}_{\mathbf{x}''\mathbf{z}''}}\right). \tag{4.2.16}$$

From this empirically estimated distribution, it is straightforward to construct a consistent test (cf. Equation (4.2.8)). The power of this test varies inversely with the variance of the distribution in Equation (4.2.14).

4.2.3 The Dependent Test is More Powerful

While discarding half the samples leads to a consistent test, we might expect some loss of power over the approach in Section 4.2.1, due to the increase in variance with lower sample size. In this section, we prove the Section 4.2.1 test is more powerful than that of Section 4.2.2, regardless of $\mathbb{P}_{\mathbf{xy}}$ and $\mathbb{P}_{\mathbf{xz}}$. We call the simple and consistent approach in Section 4.2.2, the *independent approach*, and the lower variance approach in Section 4.2.1, the *dependent approach*. The following theorem compares these approaches.

Theorem 4.3. *The asymptotic relative efficiency (ARE) of the independent approach relative to the dependent approach is always greater than 1.*

Remark 4.1. *The asymptotic relative efficiency is defined in e.g. Serfling [2009, Chap. 5, Section 1.15.4]. If m_A and m_B are the sample sizes at which tests "perform equivalently" (i.e.*

have equal power), then the ratio $\frac{m_A}{m_B}$ *represents the relative efficiency. When* m_A *and* m_B *tend to* $+\infty$ *and the ratio* $\frac{m_A}{m_B} \to L$ *(at equivalent performance), then the value* L *represents the asymptotic relative efficiency of procedure* B *relative to procedure* A*. This example is relevant to our case since we are comparing two test statistics with different asymptotically normal distributions.*

The following lemma is used for the proof of Theorem. 4.3.

Lemma 4.1 (Lower Variance)**.** *The variance of the dependent test statistic is smaller than the variance of the independent test statistic.*

Proof: From the convergence of moments in the application of the central limit theorem [von Bahr, 1965], we have that $\sigma^2_{\text{HSIC}_{\mathbf{x}'\mathbf{z}'}} = 2\sigma^2_{\text{HSIC}_{\mathbf{xy}}}$. Then the variance summary in Equation (4.2.7) is $\frac{1}{2}(\sigma^2_{\text{HSIC}_{\mathbf{xy}}}+\sigma^2_{\text{HSIC}_{\mathbf{xz}}}-2\sigma_{\text{HSIC}_{\mathbf{xy},\mathbf{xz}}})$ and the variance summary in Equation (4.2.15) is $\frac{1}{2}(2\sigma^2_{\text{HSIC}_{\mathbf{xy}}}+2\sigma^2_{\text{HSIC}_{\mathbf{xz}}})$ where in both cases the statistic is scaled by $\sqrt{n}$. We have that the variance of the independent test statistic is smaller than the variance of the dependent test statistic when

$$\begin{aligned}\frac{1}{2}(\sigma^2_{\text{HSIC}_{\mathbf{xy}}}+\sigma^2_{\text{HSIC}_{\mathbf{xz}}}-2\sigma_{\text{HSIC}_{\mathbf{xy},\mathbf{xz}}}) &< \frac{1}{2}(2\sigma^2_{\text{HSIC}_{\mathbf{xy}}}+2\sigma^2_{\text{HSIC}_{\mathbf{xz}}}) \\ \iff -2\sigma_{\text{HSIC}_{\mathbf{xy},\mathbf{xz}}} &< \sigma^2_{\text{HSIC}_{\mathbf{xy}}}+\sigma^2_{\text{HSIC}_{\mathbf{xz}}}. \end{aligned} \tag{4.2.17}$$

which is implied by the positive definiteness of $\mathbf{\Sigma}$. □

Proof of Theorem 4.3.: The Type II error probability of the independent test at level α is

$$\Phi\left[\Phi^{-1}(1-\alpha)-\frac{m^{-1/2}(\,\text{HSIC}\,[\mathcal{H}_{\mathbf{x}},\mathcal{H}_{\mathbf{y}},\mathbb{P}_{\mathbf{xy}}]-\text{HSIC}\,[\mathcal{H}_{\mathbf{x}},\mathcal{H}_{\mathbf{z}},\mathbb{P}_{\mathbf{xz}}]\,)}{\sqrt{\sigma^2_{\text{HSIC}_{\mathbf{x}'\mathbf{y}'}}+\sigma^2_{\text{HSIC}_{\mathbf{x}''\mathbf{z}''}}}}\right], \tag{4.2.18}$$

where we again make the most conservative possible assumption that $\text{HSIC}\,[\mathcal{H}_{\mathbf{x}},\mathcal{H}_{\mathbf{y}},\mathbb{P}_{\mathbf{xy}}]-\text{HSIC}\,[\mathcal{H}_{\mathbf{x}},\mathcal{H}_{\mathbf{z}},\mathbb{P}_{\mathbf{xz}}]$ under the null. The Type II error probability of the dependent test at level α is

$$\Phi\left[\Phi^{-1}(1-\alpha)-\frac{m^{-1/2}(\,\text{HSIC}\,[\mathcal{H}_{\mathbf{x}},\mathcal{H}_{\mathbf{y}},\mathbb{P}_{\mathbf{xy}}]-\text{HSIC}\,[\mathcal{H}_{\mathbf{x}},\mathcal{H}_{\mathbf{z}},\mathbb{P}_{\mathbf{xz}}]\,)}{\sqrt{\sigma^2_{\text{HSIC}_{\mathbf{xy}}}+\sigma^2_{\text{HSIC}_{\mathbf{xz}}}-2\sigma_{\text{HSIC}_{\mathbf{xy},\mathbf{xz}}}}}\right], \tag{4.2.19}$$

where Φ is the CDF of the standard normal distribution. The numerator in Equation (4.2.18) is the same as the numerator in Equation (4.2.19), and the denominator in Equation (4.2.19) is smaller due to Lemma 4.1. The lower variance dependent test therefore has higher ARE, i.e., for a sufficient sample size $n > \tau$ for some distribution dependent $\tau \in \mathbb{N}_+$, the dependent test will be more powerful than the independent test. □

4.3 Generalizing to more than Two HSIC Statistics

The generalization of the dependence test to more than three random variables follows from the earlier derivation by applying successive rotations to a higher dimensional joint Gaussian distribution over multiple HSIC statistics. Given observations $\mathbf{X}_1 := \{\mathbf{x}_1^1, \dots \mathbf{x}_n^1\}, \dots, \mathbf{X}_r := \{\mathbf{x}_1^r, \dots, \mathbf{x}_n^r\}$ i.i.d. drawn, respectively, from $\mathbb{P}_{\mathbf{x}_1}, \dots, \mathbb{P}_{\mathbf{x}_r}$. We define a generalized statistical test, $\mathcal{T}_g : (\mathcal{X}^n)^r \to \{0, 1\}$ to test the null hypothesis

$$\mathcal{H}_1 : \sum_{(i,j)\in\{1,\dots,r\}^2} \mathbf{v}_{(i,j)} \operatorname{HSIC}\left[\mathcal{H}_{\mathbf{x}_i}, \mathcal{H}_{\mathbf{x}_j}, \mathbb{P}_{\mathbf{x}_i\mathbf{x}_j}\right] \leq 0 \tag{4.3.1}$$

versus the alternative hypothesis

$$\mathcal{H}_1 : \sum_{(i,j)\in\{1,\dots,r\}^2} \mathbf{v}_{(i,j)} \operatorname{HSIC}\left[\mathcal{H}_{\mathbf{x}_i}, \mathcal{H}_{\mathbf{x}_j}, \mathbb{P}_{\mathbf{x}_i\mathbf{x}_j}\right] > 0 \tag{4.3.2}$$

where $\mathbf{v}$ is a vector of weights on each HSIC statistic. We may recover the test in the previous section by setting $\mathbf{v}_{(1,2)} = +1$, $\mathbf{v}_{(1,3)} = -1$ and $\mathbf{v}_{(i,j)} = 0$ for all $(i, j) \in \{1, 2, 3\}^2 \setminus \{(1, 2), (1, 3)\}$.

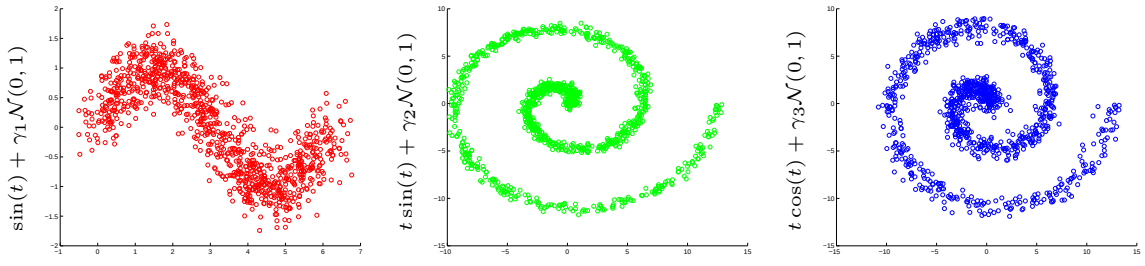
The derivation of the test follows the general strategy used in the previous section: we construct a rotation matrix so as to project the joint Gaussian distribution onto the first axis, and read the p-value from a standard normal table. To construct the rotation matrix, we simply need to rotate $\mathbf{v}$ such that it is aligned with the first axis. Such a rotation can be computed by composing n 2-dimensional rotation matrices as in Algorithm 1. We note that this same principle is applicable to any linear combination of U-statistics, including for the MMD as in Chapter 3.

Algorithm 1 Successive rotation for generalized high-dimensional relative tests of dependency (cf. Section 4.3)

Require: $\mathbf{v}$
Ensure: $[\mathbf{Qv}]_i = 0 \;\; \forall i \neq 1$, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$
  $\mathbf{Q} = I$
  for $i = 2$ **to** n **do**
    $\mathbf{Q}_i = \mathbf{I}$; $\theta = -\tan^{-1}\frac{v_i}{[\mathbf{Qv}]_1}$
    $[\mathbf{Q}_i]_{11} = \cos(\theta)$; $[\mathbf{Q}_i]_{1i} = -\sin(\theta)$; $[\mathbf{Q}_i]_{i1} = \sin(\theta)$; $[\mathbf{Q}_i]_{ii} = \cos(\theta)$
    $\mathbf{Q} = \mathbf{Q}_i\mathbf{Q}$
  end for

4.4 Experiments

In this section, we apply our estimates of statistical dependence to three challenging problems. The first is a synthetic data experiment, in which we can directly control the relative degree of



(a) $\gamma_1 = 0.3$ (b) $\gamma_2 = 0.3$ (c) $\gamma_3 = 0.6$

Figure 4.1 – Illustration of a synthetic dataset sampled from the distribution in Equation (4.4.1).

functional dependence between variates. The second experiment uses a multilingual corpus to determine the relative relations between European languages. The last experiment is a 3-block dataset which combines gene expression, comparative genomic hybridization, and a qualitative phenotype measured on a sample of Glioma patients.

4.4.1 Synthetic Experiments

We designed a synthetic problem motivated by constructing 3 distributions as defined in Equation (4.4.1) and illustrated in Figure 4.1.

$$\begin{aligned} &\text{Let } t \sim \mathcal{U}[(0, 2\pi)], \\ &(a)\ x_1 \sim t + \gamma_1\mathcal{N}(0,1) \quad y_1 \sim \sin(t) + \gamma_1\mathcal{N}(0,1); \\ &(b)\ x_2 \sim t\cos(t) + \gamma_2\mathcal{N}(0,1) \quad y_2 \sim t\sin(t) + \gamma_2\mathcal{N}(0,1); \\ &(c)\ x_3 \sim t\cos(t) + \gamma_3\mathcal{N}(0,1) \quad y_3 \sim t\sin(t) + \gamma_3\mathcal{N}(0,1). \end{aligned} \tag{4.4.1}$$

These distributions are specified so that we can control the relative degree of functional dependence between the variates by varying the relative size of noise scaling parameters γ_1, γ_2 and γ_3. The question is then whether the dependence between (a) and (b) is larger than the dependence between (a) and (c). In these experiments, we fixed $\gamma_1 = \gamma_2 = 0.3$, while we varied γ_3, and used a Gaussian kernel with bandwidth σ selected as the median pairwise distance between data points. This kernel is sufficient to obtain good performance, although others choices exist [Gretton et al., 2012b].

Figure 4.2 shows the power of the dependent and the independent tests as we vary γ_3. It is clear from these results that the dependent test is far more powerful than the independent test over the great majority of γ_3 values considered. Figure 4.3 demonstrates that this superior test power arises due to the tighter and more concentrated distribution of the dependent statistic.

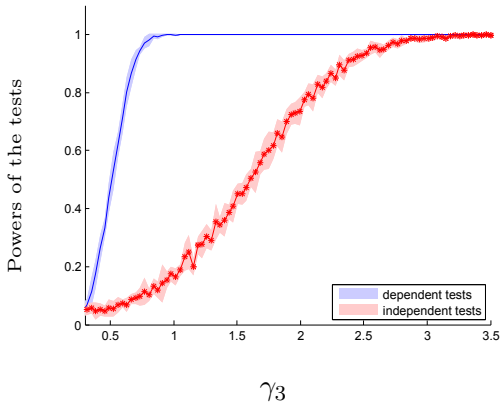
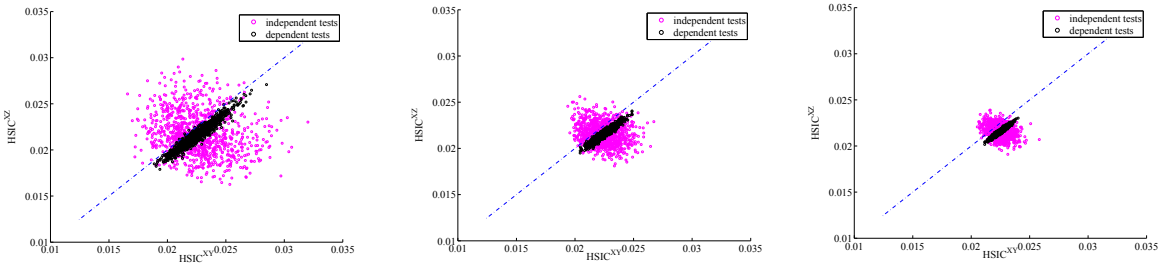


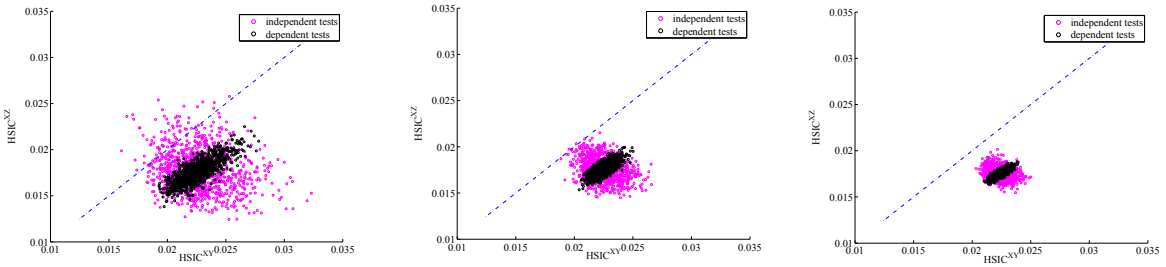
Figure 4.2 – Power of the dependent and independent test as a function of γ_3 on the synthetic data described in Section 4.4.1. For values of $\gamma_3 > 0.3$ the distribution in Figure 4.1(a) is closer to Figure 4.1(b) than to Figure 4.1(c). The problem becomes difficult as $\gamma_3 \to 0.3$. As predicted by theory, the dependent test is significantly more powerful over almost all values of γ_3 by a substantial margin.



(a) n=500, $\gamma_3 = 0.7$
$p_{dep} = 0.0189$, $p_{indep} = 0.3492$

(b) n=1000, $\gamma_3 = 0.7$
$p_{dep} = 10^{-4}$, $p_{indep} = 0.3690$

(c) n=3000, $\gamma_3 = 0.7$
$p_{dep} = 10^{-6}$, $p_{indep} = 0.2876$



(d) n=500, $\gamma_3 = 1.7$
$p_{dep} = 10^{-9}$, $p_{indep} = 0.982$

(e) n=1000, $\gamma_3 = 1.7$
$p_{dep} = 10^{-10}$, $p_{indep} = 0.0326$

(f) n=3000, $\gamma_3 = 1.7$
$p_{dep} = 10^{-13}$, $p_{indep} = 0.005$

Figure 4.3 – For the synthetic experiments described in Section 4.4.1, we plot empirical HSIC values for dependent and independent tests for 100 repeated draws with different sample sizes. Empirical p-values for each test show that the dependent distribution converges faster than the independent distribution even at low sample size, resulting in a more powerful statistical test.

Source	Target 1	Target 2	p-value
es	pt	fi	**0.0066**
fr	it	da	**0.0418**
it	es	fi	**0.0169**
pt	es	da	**0.0173**
de	nl	fi	$< \mathbf{10^{-4}}$
nl	en	es	$< \mathbf{10^{-4}}$
da	sv	fr	$< \mathbf{10^{-6}}$
sv	en	it	$< \mathbf{10^{-4}}$
en	de	es	$< \mathbf{10^{-4}}$

Table 4.1 – A selection of relative dependency tests between two pairs of HSIC statistics for the multilingual corpus data.

4.4.2 Multilingual Data

In this section, we demonstrate dependence testing to predict the relative similarity of different languages. We use a real world dataset taken from the parallel European Parliament corpus [Koehn, 2005]. We choose 3000 random documents in common written in: Finnish (fi), Italian (it), French (fr), Spanish (es), Portuguese (pt), English (en), Dutch (nl), German (de), Danish (da) and Swedish (sv). These languages can be broadly categorized into either the Romance, Germanic or Uralic groups [Gray and Atkinson, 2003]. In this dataset, we considered each language as a random variable and each document as an observation. Our first goal is to test if the statistical dependence between two languages in the same group is greater than the statistical dependence between languages in different groups.

For pre-processing, we removed stop-words (`http://www.nltk.org`) and performed stemming (`http://snowball.tartarus.org`). We applied the TF-IDF model as a feature representation and used a Gaussian kernel with the bandwidth σ set per language as the median pairwise distance between documents. For all tests, we set the significance value to $\alpha = 5\%$.

In Table 4.1, a selection of tests between language groups (Germanic, Romance, and Uralic) is given: all p-values strongly support that our relative dependence test finds the different language groups with very high significance.

Further, if we focus on the Romance family, our test enables one to answer more fine-grained questions about the relative similarity of languages within the same group. As before, we determine the ground truth similarities from the topology of the tree of European languages determined by the linguistics community [Bouckaert et al., 2012, Gray and Atkinson, 2003] as illustrated in Figure 4.4 for the Romance group. We have run the test on all triplets from the corpus for which the topology of the tree specifies a correct ordering of the dependencies. In a fraction of a second (excluding kernel computation), we are able to recover certain features of the subtree of relationships between languages present in the Romance language group (Table 4.2). The test always indicates the correct relative similarity of languages when nearby languages (pt, es) are compared with those further away (ft, it), however errors are

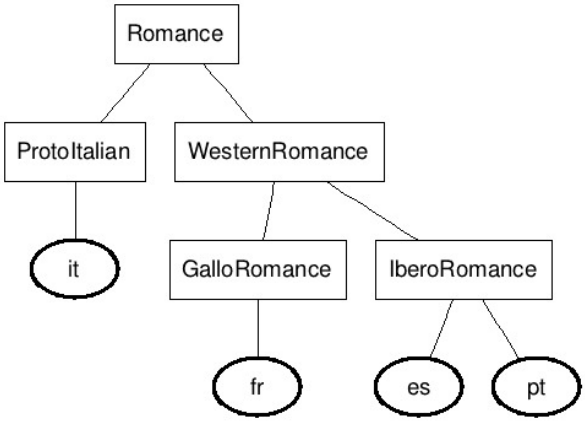


Figure 4.4 – Partial tree of Romance languages adapted from Gray and Atkinson [2003].

Source	Target 1	Target 2	p-value
fr	es	it	**0.0157**
fr	pt	it	0.1882
es	fr	it	0.2147
es	pt	it	$< \mathbf{10^{-4}}$
es	pt	fr	$< \mathbf{10^{-4}}$
pt	fr	it	0.7649
pt	es	it	**0.0011**
pt	es	fr	$< \mathbf{10^{-8}}$

Table 4.2 – Relative dependency tests between Romance languages. The tests are ordered such that a low p-value corresponds with a confirmation of the topology of the tree of Romance languages determined by the linguistics community [Bouckaert et al., 2012, Gray and Atkinson, 2003].

made when comparing triplets of languages for which the nearest common ancestor is more than one link removed.

In our next tests, we evaluate our more general framework for testing relative dependencies with more than two HSIC statistics. We chose four languages, and tested whether the average dependence between languages in the same group is higher than the dependence between groups. The results of these tests are in Table 4.3. As before, our generalized test is able to distinguish language groups with high significance.

4.4.3 Pediatric Glioma Data

Brain tumors are the most common solid tumors in children and have the highest mortality rate of all pediatric cancers. Despite advances in multimodality therapy, children with pediatric high-grade gliomas (pHGG) invariably have an overall survival of around 20% at 5 years. Depending on their location (e.g. brainstem, central nuclei, or supratentorial), pHGG present

Source	Targets	p-value
da	de sv fi	$< \mathbf{10^{-9}}$
da	sv en fr	$< \mathbf{10^{-9}}$
de	sv en it	$< \mathbf{10^{-5}}$
fr	it es sv	$< \mathbf{10^{-5}}$
es	fr pt nl	**0.0175**

Table 4.3 – Relative dependency test between four pairs of HSIC statistics for the multilingual corpus data. These tests show the ability of the relative dependence test to generalize to arbitrary numbers of HSIC statistics by constructing a rotation matrix using Algorithm 1. In all cases $\mathbf{v} = [1\ 1\ -2]$.

different characteristics in terms of radiological appearance, histology, and prognosis. The hypothesis is that pHGG have different genetic origins and oncogenic pathways depending on their location. Thus, the biological processes involved in the development of the tumor may be different from one location to another.

In order to evaluate such hypotheses, pre-treatment frozen tumor samples were obtained from 53 children with newly diagnosed pHGG from Necker Enfants Malades (Paris, France) from Puget et al. [2012]. The 53 tumors are divided into 3 locations: supratentorial (HEMI), central nuclei (MIDL), and brain stem (DIPG). The final dataset is organized in 3 blocks of variables defined for the 53 tumors: $\mathbf{x}$ is a block of indicator variables describing the location category, the second data matrix $\mathbf{y}$ provides the expression of 15 702 genes (GE). The third data matrix $\mathbf{z}$ contains the imbalances of 1229 segments (CGH) of chromosomes.

For $\mathbf{x}$, we use a linear kernel, which is characteristic for indicator variables, and for $\mathbf{y}$ and $\mathbf{z}$, the kernel was chosen to be the Gaussian kernel with σ selected as the median of pairwise distances. The p-value of our relative dependency test is $< 10^{-5}$. This shows that the tumor location in the brain is more dependent on gene expression than on chromosomal imbalances. In contrast to the independent subsampling method described in Section 4.2.2, the dependent test was also able to find the same ordering of dependence, but with a p-value that is three orders of magnitude larger ($p = 0.005$). Figure 4.5 shows iso-curves of the Gaussian distributions estimated in the independent and dependent tests. The empirical relative dependency is consistent with findings in the medical literature, and provides additional statistical support for the importance of tumor location in Gliomas [Gilbertson and Gutmann, 2007, Palm et al., 2009, Puget et al., 2012].

4.5 Conclusion

In this chapter, we have described a novel statistical test that determines whether a source random variable is more strongly dependent on one target random variable or another. This test, built on the Hilbert-Schmidt Independence Criterion, is low variance, consistent, and unbiased. We have shown that our test is strictly more powerful than a test that does not exploit the covariance between HSIC statistics, and empirically achieves p-values several orders

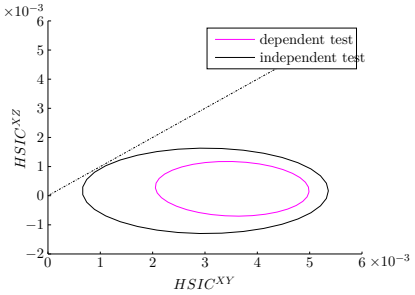


Figure 4.5 – 2σ iso-curves of the Gaussian distributions estimated from the pediatric Glioma data. As before, the dependent test has a much lower variance than the independent test. The tests support the stronger dependence on the tumor location to gene expression than chromosomal imbalances.

of magnitude smaller. We have empirically demonstrated the test performance on synthetic data, where the degree of dependence could be controlled; on the challenging problem of identifying language groups from a multilingual corpus; and for finding the most important determinant of Glioma type. The computation and memory requirements of the test are quadratic in the sample size, matching the performance of HSIC and related tests for dependence between two random variables. The test is therefore scalable to the wide range of problem instances where non-parametric dependency tests are currently applied. We have generalized the test framework to more than two HSIC statistics, and have given an algorithm to construct a consistent, low-variance, unbiased test in this setting.

Chapter 5

Linear Time Non-Gaussian Precision Matrix Estimation

In the previous chapter we investigated relative dependency using a non-parametric statistical test based on U-statistic. In this chapter, we will address conditional dependency using an analogous approach based on a U-statistic estimator of the covariance matrix and address Research Question 3.

Structure discovery in graphical models is the determination of the topology of a graph that encodes conditional independence properties of the joint distribution of all variables in the model. For some class of probability distributions, an edge between two variables is present if and only if the corresponding entry in the precision matrix is non-zero. For a finite sample estimate of the precision matrix, entries close to zero may be due to low sample effects, or due to an actual conditional independence between variables; these two cases are not readily distinguishable. Methods for structure discovery in the literature typically make restrictive (Gaussian) distributional or sparsity assumptions that may not apply to a data sample of interest, and direct estimation of the uncertainty of an estimate of the precision matrix for general distributions remains challenging. Consequently, we derive a new test that makes use of results for U-statistics and applies them to the covariance matrix. By probabilistically bounding the distortion of the covariance matrix, we can apply Weyl's theorem to bound the distortion of the precision matrix, yielding a sound test threshold for a wider class of distributions than considered in previous works. The resulting test enables one to answer with statistical significance whether an entry in the precision matrix is non-zero, and convergence results are known for a wide range of distributions. The computational complexity is linear in the sample size enabling the application of the test to large data samples for which computation time becomes a limiting factor. The soundness and effectiveness of the test is demonstrated on synthetic and real-world weather and medical datasets comprising millions of observations from non-Gaussian distributions.

Work covered this chapter is based on:

- W. Bounliphone and M. B. Blaschko. Linear time non-Gaussian precision matrix estimation. 2016. arXiv:1604.01733 – under submission.

Project: https://github.com/wbounliphone/Ustatistics_Approach_For_SD

Contents

5.1 Introduction

Graphical models are powerful tools for analyzing relationships between a set of random variables, so that key conditional independence properties can be read from a graph. Learning the structure of an underlying graphical model is of fundamental importance and has applications in a large number of domains [de Morais and Aussem, 2010, Gasse et al., 2012, Sechidis and Brown, 2015]. In many contemporary applications, a large, effectively unlimited stream of raw data with unknown multivariate distribution is to be analyzed. In such scenarios, computation becomes a fundamental limit and methods that can estimate properties of graphical models from very general distributions with computation linear in the number of observations become necessary. We can divide graphical models in two types, namely directed graphical models, e.g. Bayesian networks [Neapolitan, 2004] or undirected graphical models, e.g. Gaussian graphical models [Lauritzen, 1996, Whittaker, 2009]. Here, we focus on undirected graphical models to encode the conditional dependence structure in multivariate distributions.

Hypothesis testing with statistical measures of dependence is a relatively well developed field with a number of general results. Classical tests such as Spearman's ρ and Kendall's τ are widely applied [Kendall, 1946]. Recently, for multivariate non-linear dependencies, novel statistical tests were introduced, with prominent examples including the generalized variance and kernel canonical correlation analysis [Bach and Jordan, 2002], the Hilbert-Schmidt independence criterion [Gretton et al., 2005a], distance based correlation [Székely et al., 2007]

and rankings [Heller et al., 2013]. Testing the conditional dependence is even more challenging, and only few dependence measures have been generalized to the conditional case [Doran et al., 2014, Fukumizu et al., 2007, 2009, Zhang et al., 2011]. We note that their work requires the estimate of a regularization parameter with appropriate asymptotic decrease to estimate the distribution of the test statistic under the null hypothesis, as well as for kernel selection, and has quadratic space usage rendering it inapplicable to very large data sets. These results, however, do not directly extend to the test that we analyze here: that of independence between two variables *conditioned* on all the others:

$$\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j | \mathbf{x}_{V \setminus \{i,j\}}. \tag{5.1.1}$$

For Gaussian graphical models, the non-zero entry in the inverse of the covariance matrix (called the *precision* matrix), can be shown to correspond to the underlying structure of the graphical model [Dempster, 1972]. This observation has motivated a range of structure discovery techniques for estimating the precision matrix using model selection and parameter estimation methods [Drton and Perlman, 2004, Roverato and Whittaker, 1996, Yuan and Lin, 2007]. Furthermore, estimation in high-dimensional settings ($n \ll p$, where n is the sample size and p is the dimension) has been the focus on recent research [Banerjee et al., 2008, Friedman et al., 2008, Li and Gui, 2006, Liu et al., 2013, Meinshausen and Bühlmann, 2006, Ravikumar et al., 2011, Ren et al., 2015, Schäfer and Strimmer, 2005, Yuan and Lin, 2007] where methods impose a strong sparsity constraint on the entries of the precision matrix. The consequence of this method to estimate the sparse precision matrix has been the development of diverse statistical hypothesis tests [G'Sell et al., 2013, Jankova and van de Geer, 2015, Lockhart et al., 2014]. Each of these methods explicitly assumes that the data distribution is multivariate Gaussian. By contrast, we instead focus in this paper on designing a test for the $p \ll n$ case, and in particular ensure that the test has computational complexity *linear* in n, while making minimal distributional assumptions and no sparsity assumptions.

For non-Gaussian graphical models, several techniques focus on the existence of a relationship between conditional independence and the structure of the inverse covariance matrix. Loh and Wainwright [2013] have established several theoretical results by extending a number of interesting links between covariance matrices and graphical models for discrete random variables and tree-structured graphs.

While there exist many convenient methods using Gaussian multivariate distributions or discrete variables, other distributions pose new challenges in statistical modeling. In Wasserman et al. [2014], the authors consider the problem of providing non-parametric confidence guarantees that do not assume a Gaussian distribution or sparsity using finite sample Berry-Esseen bounds on the accuracy of the normal approximation and the bootstrap approach. In contrast to this article, these bootstrap estimates add a significant computational overhead to the approach.

Our contribution: Consequently, we develop a statistically and computationally efficient framework for hypothesis testing of whether an entry of the precision matrix is non-zero based on a data sample from the joint distribution $\mathbb{P}_{\mathbf{x}}$. The proposed test does not depend

on the data being Gaussian distributed or other parametric assumptions and does not require sparsity. Also, the test not only has asymptotic guarantees, but can be applied to finite samples without the need to set a regularization parameter or perform a computationally expensive bootstrap procedure.

5.2 Proposed Method

In this section, we first develop a U-statistic estimator of the covariance matrix and its uncertainty. Based on this distribution over covariance matrices, we subsequently probabilistically bound its distortion and use this bound to compute a test threshold for the empirical precision matrix. Finally, we analyze the computational and statistical properties of the resulting algorithm.

5.2.1 Structure Discovery in Undirected Graphical Models

In this section, we first give a U-statistic estimator of the covariance matrix to define a hypothesis test for discovering the structure of graphical models. We show that this estimator can be computed in time linear in the number of samples and study its asymptotic distribution. We will denote the covariance matrix by $\mathbf{\Sigma}$ with its unbiased estimator $\hat{\mathbf{\Sigma}}$ using Definition 5.1.

We focus here on U-statistic estimates of $\hat{\Sigma}$ and its asymptotic normal distribution to calculate conservative bounds on the threshold for our hypothesis test. We develop the full covariance between the elements of $\hat{\mathbf{\Sigma}}$, which we denote $\mathrm{Cov}(\hat{\mathbf{\Sigma}}) \in \mathbb{R}^{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}}$, where the size is due to the symmetry of $\hat{\mathbf{\Sigma}}$. We note $U(\boldsymbol{A})$ the function returning the upper triangular part and diagonal of a matrix $\boldsymbol{A}$.

Definition 5.1 (U-statistic estimator for the covariance matrix)**.** *Let* $\mathbf{u}_r = (\mathbf{x}_{i_r}, \mathbf{x}_{j_r})^T$*, with* $r = 1, 2$ *be ordered pairs of samples, with* $1 \leq i \leq p$*. Consider* $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$*, the covariance functional between* $\mathbf{x}_{i_1}$ *and* $\mathbf{x}_{i_2}$ *and* h*, the kernel of order 2 for the functional* $\mathbf{\Sigma}$ *such that*

$$h(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2} (\mathbf{x}_{i_1} - \mathbf{x}_{i_2}) (\mathbf{x}_{j_1} - \mathbf{x}_{j_2}) . \tag{5.2.1}$$

The corresponding U*-statistic estimator of the covariance* $\mathbf{\Sigma}$ *is*

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{i,j=1}^{n} (\mathbf{x}_{i_1} - \mathbf{x}_{i_2})(\mathbf{x}_{j_1} - \mathbf{x}_{j_2}) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}_j})(\mathbf{x}_i - \overline{\mathbf{x}_j}) \tag{5.2.2}$$

where $\overline{\mathbf{x}_i} = \frac{1}{n} \sum_{q=1}^{n} \mathbf{x}_{qi}$*.* $\hat{\mathbf{\Sigma}}$ *can be computed in* $\mathcal{O}(n)$*, with* n *the sample size.*

Theorem 5.1 (Joint asymptotic normal distribution of the covariance matrix, [Serfling, 2009])**.** *For all* (i, j, k, l) *range over each of the* p *variates in a covariance matrix* $\mathbf{\Sigma}$*, if* $\mathrm{Var}(\hat{\Sigma}_{ij}) > 0$ *and* $\mathrm{Var}(\hat{\Sigma}_{kl}) > 0$*, then* $\left[\hat{\Sigma}_{ij}; \hat{\Sigma}_{kl}\right]^T$ *converges in distribution (as* $n \longrightarrow \infty$*)*

to a Gaussian random variables

$$n^{\frac{1}{2}} \begin{pmatrix} \hat{\Sigma}_{ij} - \Sigma_{ij} \\ \hat{\Sigma}_{kl} - \Sigma_{kl} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \text{Var}(\hat{\Sigma}_{ij}) & \text{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) \\ \text{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) & \text{Var}(\hat{\Sigma}_{kl}) \end{pmatrix} \right). \tag{5.2.3}$$

where $\text{Var}(\hat{\Sigma}_{ij}), \text{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$ *and* $\text{Var}(\hat{\Sigma}_{kl})$ *are derived in Theorem 5.2.*

The preceding theorem indicates that the empirical covariance estimate has asymptotic Gaussian distribution with a $\mathcal{O}(n^{-1/2})$ convergence rate. The following two theorems show that the covariance of the empirical covariance estimates are also known and can be empirically computed in time linear in n.

Theorem 5.2 (Variance/Covariance of the U-statistic for the covariance matrix)**.** *We note respectively h and g the corresponding kernel of order 2 for the two unbiased estimates $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$, where*

$$\begin{aligned} h(\mathbf{u}_1, \mathbf{u}_2) &= \frac{1}{2} (\mathbf{x}_{i_1} - \mathbf{x}_{i_2}) (\mathbf{x}_{j_1} - \mathbf{x}_{j_2}), \text{ with } \mathbf{u}_r = (\mathbf{x}_{i_r}, \mathbf{x}_{j_r})^T, \text{for } r = 1, 2; \\ g(\mathbf{v}_1, \mathbf{v}_2) &= \frac{1}{2} (\mathbf{x}_{k_1} - \mathbf{x}_{k_2}) (\mathbf{x}_{l_1} - \mathbf{x}_{l_2}), \text{ with } \mathbf{v}_r = (\mathbf{x}_{k_r}, \mathbf{x}_{l_r})^T. \end{aligned} \tag{5.2.4}$$

The low variance, unbiased estimates of the covariance between two U-statistics estimates $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$, where $i \leq j$, $k \leq l$ range over each of the p variates in a covariance matrix $\hat{\Sigma}$ is

$$[\text{Cov}(\hat{\Sigma})]_{(ijkl)} := \text{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) = \binom{n}{2}^{-1} (2(n-2)\zeta_1) + \mathcal{O}(n^{-2}), \tag{5.2.5}$$

where $\zeta_1 = \text{Cov}\left(\mathbb{E}_{\mathbf{u}_2}[h(\mathbf{u}_1, \mathbf{u}_2)], \mathbb{E}_{\mathbf{v}_2}[g(\mathbf{v}_1, \mathbf{v}_2)]\right)$ *and* $\text{Cov}(\hat{\mathbf{\Sigma}}) \in \mathbb{R}^{\left(p^2 - \binom{p}{2}\right) \times \left(p^2 - \binom{p}{2}\right)}$.

Proof: Equation (5.2.5) is directly constructed with the definition of the covariance of a U-statistic from Hoeffding [1948]. □

Theorem 5.3. *Each entry of* $\text{Cov}(\hat{\mathbf{\Sigma}})$ *can be estimated in time linear in n. For each $1 \leq i, j, k, l \leq p$,* $\text{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$ *can be estimated using one of seven different cases through simple variable substitution.*

▶ **Case 1** : $i \neq j, k, l$; $j \neq k, l$; $k \neq l$

$$\begin{aligned} \zeta_1 = \frac{1}{4} \Big\{ & \overline{\mathbf{x}_i \mathbf{x}_j \mathbf{x}_k \mathbf{x}_l} - \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j \mathbf{x}_k \mathbf{x}_l} - \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_i \mathbf{x}_k \mathbf{x}_l} - \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_i \mathbf{x}_j \mathbf{x}_l} \\ & + \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_j \mathbf{x}_l} + \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_i \mathbf{x}_l} - \overline{\mathbf{x}_i \mathbf{x}_j \mathbf{x}_k}\ \overline{\mathbf{x}_l} + \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l}\ \overline{\mathbf{x}_j \mathbf{x}_k} \\ & + \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_l}\ \overline{\mathbf{x}_i \mathbf{x}_k} - (\overline{\mathbf{x}_i \mathbf{x}_j} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j})\ (\overline{\mathbf{x}_k \mathbf{x}_l} - 2\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_l}) \Big\}. \end{aligned} \tag{5.2.6}$$

▶ **Case 2** : $i = j;\ j \neq k, l;\ k = l$

$$\zeta_1 = \frac{1}{4}\Big\{\overline{\mathbf{x}_i^2\mathbf{x}_k^2} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_i\mathbf{x}_k^2} - 2\ \overline{\mathbf{x}_i^2\mathbf{x}_{k_1}}\ \overline{\mathbf{x}_k} + 4\overline{\mathbf{x}_i\mathbf{x}_k}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_k} - \Big(\overline{\mathbf{x}_i^2} - 2\ \overline{\mathbf{x}_i}^2\Big)\Big(\overline{\mathbf{x}_k^2} - 2\ \overline{\mathbf{x}_k}^2\Big)\Big\}. \tag{5.2.7}$$

▶ **Case 3** : $i = j;\ j \neq k, l;\ k \neq l$

$$\zeta_1 = \frac{1}{4}\Big\{\overline{\mathbf{x}_i^2\mathbf{x}_k\mathbf{x}_l} - 2\ \overline{\mathbf{x}_i\mathbf{x}_k\mathbf{x}_l}\ \overline{\mathbf{x}_i} - \overline{\mathbf{x}_i^2\mathbf{x}_l}\ \overline{\mathbf{x}_k} + 2\ \overline{\mathbf{x}_i\mathbf{x}_l}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_k} - \overline{\mathbf{x}_i^2\mathbf{x}_{k_1}}\ \overline{\mathbf{x}_l} + 2\ \overline{\mathbf{x}_i\mathbf{x}_k}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l} - \Big(\overline{\mathbf{x}_i^2} - 2\ \overline{\mathbf{x}_i}^2\Big)\left(\overline{\mathbf{x}_k\mathbf{x}_l} - 2\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_l}\right)\Big\}. \tag{5.2.8}$$

▶ **Case 4** : $i = k;\ j \neq i, k, l;\ k \neq l$

$$\zeta_1 = \frac{1}{4}\Big\{\overline{\mathbf{x}_i^2\mathbf{x}_j\mathbf{x}_l} - \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j\mathbf{x}_i\mathbf{x}_l} - \overline{\mathbf{x}_i^2\mathbf{x}_l}\ \overline{\mathbf{x}_j} - \overline{\mathbf{x}_i\mathbf{x}_j\mathbf{x}_l}\ \overline{\mathbf{x}_i} + \overline{\mathbf{x}_i}^2\ \overline{\mathbf{x}_j\mathbf{x}_l} + \overline{\mathbf{x}_i\mathbf{x}_{l_1}}\ \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_i} - \overline{\mathbf{x}_i^2\mathbf{x}_j}\ \overline{\mathbf{x}_l} + \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j\mathbf{x}_i}\ \overline{\mathbf{x}_l} + \overline{\mathbf{x}_i^2}\ \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_l} - \left(\overline{\mathbf{x}_i\mathbf{x}_j} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j}\right)\left(\overline{\mathbf{x}_i\mathbf{x}_l} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l}\right)\Big\}. \tag{5.2.9}$$

▶ **Case 5** : $i = k;\ i \neq j;\ j = l$

$$\zeta_1 = \frac{1}{4}\Big\{\overline{\mathbf{x}_i^2\mathbf{x}_j^2} - 2\overline{\mathbf{x}_i\mathbf{x}_j^2}\ \overline{\mathbf{x}_i} + \overline{\mathbf{x}_i}^2\ \overline{\mathbf{x}_j^2} - 2\overline{\mathbf{x}_i^2\mathbf{x}_j}\ \overline{\mathbf{x}_j} + 2\overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_j\mathbf{x}_i} + \overline{\mathbf{x}_i^2}\ \overline{\mathbf{x}_j}^2 - \left(\overline{\mathbf{x}_i\mathbf{x}_j} - 2(\overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j})\right)^2\Big\}. \tag{5.2.10}$$

▶ **Case 6** : $i = j = k;\ i \neq l$

$$\zeta_1 = \frac{1}{4}\Big\{\overline{\mathbf{x}_i^3\mathbf{x}_l} - 3\ \overline{\mathbf{x}_i^2\mathbf{x}_l}\ \overline{\mathbf{x}_i} + 2\ \overline{\mathbf{x}_i\mathbf{x}_l}\ \overline{\mathbf{x}_i}^2 - \overline{\mathbf{x}_i^3}\ \overline{\mathbf{x}_l} + 2\ \overline{\mathbf{x}_i^2}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l} - \Big(\overline{\mathbf{x}_i^2} - 2\ \overline{\mathbf{x}_i}^2\Big)\left(\overline{\mathbf{x}_i\mathbf{x}_l} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l}\right)\Big\}. \tag{5.2.11}$$

▶ **Case 7** : $i = j, k, l$

$$\zeta_1 = \frac{1}{4}\Big\{\overline{\mathbf{x}_i^4} - 4\overline{\mathbf{x}_i^3}\ \overline{\mathbf{x}_i} + 4\overline{\mathbf{x}_i^2}\ \overline{\mathbf{x}_i}^2 - \Big(\overline{\mathbf{x}_i^2} - 2\overline{\mathbf{x}_i}^2\Big)^2\Big\}. \tag{5.2.12}$$

A proof of Theorem 5.3 is given in Section 5.5.

We have shown that, for an extremely general class of multivariate distributions (non-pathological with finite second moment), empirical estimates of covariances and their joint distribution can be computed in linear time. By contrast, estimation of the precision matrix remains challenging, and for general, non-Gaussian distributions, the asymptotic distribution of an empirical estimate is unknown. We consequently use the distribution of the covariance estimate to probabilistically bound the distortion of the precision matrix in the next section.

5.2.2 Hypothesis Test using a U-statistic Estimator for the Covariance Matrix

This section describes the main novel theoretical contributions of this work. We first describe our statistical test for structure discovery in undirected graphical models, based on the U-statistic estimator $\hat{\mathbf{\Sigma}}$ of the covariance matrix reviewed in the previous section and discuss its performance (Section 5.2.2). We subsequently study the computational efficiency of the proposed statistical test (Section 5.2.3).

We now describe our test for structure discovery in undirected graphical model, based on the asymptotic Gaussian distribution of the empirical covariance matrix described in Theorems 5.1 and 5.2. We begin with an introduction to the terminology of statistical hypothesis testing, as it applies to edge discovery. We denote the precision matrix by $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, with $\hat{\mathbf{\Theta}}$ its empirical estimate.

Given $\mathbf{X}$ a design matrix of size $n \times p$ and for all $(i, j) \in \{1, ..., p\}$, the statistical test $(\mathcal{T}_{ij}, \hat{\Theta}_{ij}, \delta) : (\mathbf{X}, i, j) \longmapsto \{0, 1\}$, is used to distinguish between the following null hypothesis

$$\mathcal{H}_0(i, j) : \Theta_{i,j} = \nu, \tag{5.2.13}$$

and the two-sided alternative hypothesis

$$\mathcal{H}_1(i, j) : \Theta_{i,j} \neq \nu, \tag{5.2.14}$$

with $\nu \in \mathbb{R}$ at a significance level δ. This is achieved by comparing the test statistic, $|\hat{\Theta}_{ij}|$ with a particular threshold t: if the threshold is exceeded, then the test rejects the null hypothesis. The acceptance region of the test is thus defined as any real number below the threshold.

We discuss the assumptions and explain in Theorem 5.5 how the threshold is determined and show that it is a conservative bound. Theorem 5.5 is proved using Lemmas 5.1 and 5.2 and Weyl's Theorem 5.4.

Lemma 5.1. *With probability at least* $1-\delta$*, we have the two following inequalities*

$$\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\|_2 \leq \sqrt{2\lambda_{\max}}\Phi^{-1}\left(1-\delta/2\right), \tag{5.2.15}$$

$$\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\|_2 \leq \sqrt{2\,\mathrm{Tr}[\mathrm{Cov}(\hat{\boldsymbol{\Sigma}})]}\Phi^{-1}\left(1-\delta/2\right), \tag{5.2.16}$$

where $\Phi(\cdot)$ *is the CDF of* $\mathcal{N}(0,1)$ *and* $\lambda_{\max}$ *is the largest eigenvalue of* $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}})$*. Equation* (5.2.15) *is tighter than Equation* (5.2.16)*.*

Proof: As $\hat{\boldsymbol{\Sigma}}$ is a U-statistic, we have that $U(\hat{\boldsymbol{\Sigma}})$, a vector containing its upper diagonal component (including the diagonal), is Gaussian distributed with covariance $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}})$ (cf. Theorems 5.1, 5.2). Therefore, with probability at least $1-\delta$,

$$\|U(\boldsymbol{\Sigma})-U(\hat{\boldsymbol{\Sigma}})\|_2 \leq \sqrt{\lambda_{\max}}\Phi^{-1}\left(1-\delta/2\right) \tag{5.2.17}$$

and furthermore

$$\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\|_F \leq \sqrt{2}\|U(\boldsymbol{\Sigma})-U(\hat{\boldsymbol{\Sigma}})\|_2 \tag{5.2.18}$$

which combined with the fact that $\|\cdot\|_2 \leq \|\cdot\|_F$ yields the desired result. □

Lemma 5.2 (Bounding the deviation of the empirical precision matrix as a function of eigenvalues)**.** *Given* $\mathbf{x}$ *a set of random variables drawn from a distribution for which*

$$\sum_{k=1}^{p}\frac{1}{\alpha_k\hat{\alpha}_k}-2\,\mathrm{Tr}\left[\hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^{\boldsymbol{T}}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\boldsymbol{T}}\right], \tag{5.2.19}$$

converges at a rate $\mathcal{O}(n^{-1/2})$ *with a precision matrix* $\boldsymbol{\Theta}$*, and an empirical estimate of the precision matrix* $\hat{\boldsymbol{\Theta}}$ *corresponding to a covariance matrix* $\hat{\boldsymbol{\Sigma}}$ *with eigenvalues* $\hat{\alpha}_1,\ldots,\hat{\alpha}_p$*, then with high probability*

$$|\hat{\Theta}_{ij}-\Theta_{ij}| \leq \mu\sqrt{\sum_{k=1}^{p}\left(\frac{1}{\alpha_k}-\frac{1}{\hat{\alpha}_k}\right)^2} \quad \forall i,j\in\{1,...,p\}, \tag{5.2.20}$$

for a distribution dependent constant μ*.*

Proof: We denote respectively $\hat{\boldsymbol{\Sigma}}$ the perturbed matrix of $\boldsymbol{\Sigma}$, with $\alpha_1 \geq ... \geq \alpha_p$ the eigenvalues of $\boldsymbol{\Sigma}$ and $\hat{\alpha}_1 \geq ... \geq \hat{\alpha}_p$ the eigenvalues of an empirical estimate of the true covariance matrix $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{\Theta}}$ the perturbed matrix of $\boldsymbol{\Theta}$. We then have that $|\hat{\Theta}_{ij}-\Theta_{ij}| \leq \|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_F$ for all $i,j \in \{1,...,p\}$. We will use the property of the singular value decomposition that $\hat{\boldsymbol{\Sigma}}=\hat{\boldsymbol{V}}\hat{\boldsymbol{A}}\hat{\boldsymbol{V}}^{\boldsymbol{T}}$, where $\hat{V}$ is an $n\times n$ unitary matrix and a diagonal matrix $\hat{\boldsymbol{A}}$ with $\hat{A}_{ii}=\hat{\alpha}_i$ is the i-th eigenvalue of $\hat{\boldsymbol{\Sigma}}$. Furthermore, we have that $\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Theta}$ and the empirical estimate of $\boldsymbol{\Theta}$ is $\hat{\boldsymbol{\Theta}}$ such that $\hat{\boldsymbol{\Theta}}=\hat{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^{\boldsymbol{T}}$ where $\hat{\boldsymbol{U}}$ is an $n\times n$ unitary matrix and a diagonal matrix $\hat{\boldsymbol{\Lambda}}$

with $\hat{\Lambda}_{ii} = 1/\hat{\alpha}_i$.

$$\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F^2 = \operatorname{Tr}\left[\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}} + \boldsymbol{\Theta}\boldsymbol{\Theta} - 2\hat{\boldsymbol{\Theta}}\boldsymbol{\Theta}\right] \tag{5.2.21}$$

$$= \operatorname{Tr}\left[\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}} + \boldsymbol{\Lambda}\boldsymbol{\Lambda} - 2\hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^{\boldsymbol{T}}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\boldsymbol{T}}\right] \tag{5.2.22}$$

$$= \operatorname{Tr}\left[\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}} + \boldsymbol{\Lambda}\boldsymbol{\Lambda} - 2\boldsymbol{\Lambda}\hat{\boldsymbol{\Lambda}}\right] + 2\operatorname{Tr}\left[\boldsymbol{\Lambda}\hat{\boldsymbol{\Lambda}} - \hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^{\boldsymbol{T}}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\boldsymbol{T}}\right] \tag{5.2.23}$$

$$= \underbrace{\sum_{k=1}^{p}\left(\frac{1}{\alpha_k} - \frac{1}{\hat{\alpha}_k}\right)^2}_{(5.2.24)-A} + \underbrace{2\sum_{k=1}^{p}\frac{1}{\alpha_k\hat{\alpha}_k} - 2\operatorname{Tr}\left[\hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{U}}^{\boldsymbol{T}}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\boldsymbol{T}}\right]}_{(5.2.24)-B} \tag{5.2.24}$$

$$\leq \mu\left(\sum_{k=1}^{p}\left(\frac{1}{\alpha_k} - \frac{1}{\hat{\alpha}_k}\right)^2\right). \tag{5.2.25}$$

The bound in Equation (5.2.25) will hold with high probability, e.g. when the finite moment conditions of Xia et al. [2013] are satisfied, as Equation. (5.2.24) is then guaranteed to converge with rate $\mathcal{O}(n^{-1/2})$. □

We have now shown that we can compute a bound on the distortion purely from the eigenvalues of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$.

Theorem 5.4 (Weyl's Theorem, [Weyl, 1912])**.** *For two positive definite matrices Σ and $\hat{\Sigma}$ with corresponding eigenvalues α_k and $\hat{\alpha}_k$, respectively, if*

$$|\alpha_k - \hat{\alpha}_k| \leq \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \leq \varepsilon, \tag{5.2.26}$$

where $0 < \varepsilon < \alpha_k \ \forall k \in \{1, ..., p\}$, then

$$\alpha_k - \varepsilon \leq \hat{\alpha}_k \leq \alpha_k + \varepsilon \quad \forall k \in \{1, ..., p\}. \tag{5.2.27}$$

Theorem 5.5 (Conservative threshold)**.** *For all $(i, j) \in \{1, ..., p\}$, the threshold t for testing $\mathcal{H}_0 : \Theta_{i,j} = \nu$ versus $\mathcal{H}_1 : \Theta_{i,j} \neq \nu$ is given by $\mathbb{P}$ for a small probability $\delta \in (0, 1)$ such that*

$$\mathbb{P}\left(|\hat{\Theta}_{i,j}| > t|\Theta_{i,j} = 0\right) < \delta, \tag{5.2.28}$$

such that

$$t = \mu\sqrt{\sum_{k=1}^{p}\left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2}, \tag{5.2.29}$$

where t is a conservative threshold, $\hat{\alpha}_k$ is the k-th eigenvalue of the empirical covariance matrix $\hat{\Sigma}$, $\forall k, \hat{\alpha}_k > \varepsilon$, μ is a distribution dependent constant satisfying the Equation (5.2.25), and ε is an error bound such that

$$\varepsilon_{\text{Eig}} = \sqrt{2\lambda_{max}}\Phi\left(1 - \delta/2\right), \text{ or,} \tag{5.2.30}$$

$$\varepsilon_{\text{Trace}} = \sqrt{2\operatorname{Tr}[\operatorname{Cov}(\hat{\boldsymbol{\Sigma}})]}\Phi\left(1 - \delta/2\right), \tag{5.2.31}$$

where λ_{max} is the largest eigenvalue of $\operatorname{Cov}(\hat{\boldsymbol{\Sigma}})$ and $\operatorname{Tr}[\operatorname{Cov}(\hat{\boldsymbol{\Sigma}})]$ is the trace of $\operatorname{Cov}(\hat{\boldsymbol{\Sigma}})$.

Proof: We have shown that we can compute the distortion of $\hat{\mathbf{\Theta}}$ purely from the eigenvalues of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$. Therefore, we use Weyl's theorem on the covariance matrix to get error bounds for the eigenvalues of Σ. For $\varepsilon < \hat{\alpha}_k$, $\forall k$, inequality (5.2.27) gives the following bounds for the eigenvalues of the precision matrix $\mathbf{\Theta}$

$$\left(\frac{1}{\alpha_k} - \frac{1}{\hat{\alpha}_k}\right)^2 \leq \left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2 \quad \forall k \in \{1, ..., p\}. \tag{5.2.32}$$

Combining Equation (5.2.25) and Equation (5.2.32) gives with high probability

$$\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F \leq \mu \sqrt{\sum_{i=1}^{p} \left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2}, \tag{5.2.33}$$

which implies a bound on the individual entries of the precision matrix as

$$|\hat{\Theta}_{ij} - \Theta_{ij}| \leq \|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F. \tag{5.2.34}$$

□

Remark 5.1. *In the case that ε is larger than the smallest eigenvalue of $\hat{\mathbf{\Sigma}}$, the test threshold is unbounded and we can never reject the null hypothesis. In this case, additional data are necessary to decrease ε in order to have a non-trivial bound. Theorem 5.1 guarantees that ε converges to zero as a function of the sample size at a rate $\mathcal{O}(n^{-1/2})$.*

The computation of the statistical test for structure discovery in multivariate graphical models is described in detail in Algorithm 2. We now discuss the computational efficiency of the test.

5.2.3 Computational Efficiency of the Test

We now address the computational efficiency for the proposed test. Due to the fact that the computational complexity of the statistical test is linear ($\mathcal{O}\left(np^4\right)$ for the eigenvalue threshold and $\mathcal{O}\left(np^2 + p^3\right)$) for the trace threshold), we show theoretical results of the performance of the test in term of computational cost and power of a statistical test.

Theorem 5.6. *For a fixed computational budget N less than the time required to process all data points, the trace bound decreases at the same asymptotic rate as the eigenvalue bound as a function of N and p.*

Proof: We note that the bound in Equation (5.2.16) is strictly larger than that of Equation (5.2.15), but its computation $C_{\text{Trace}}(n, p) \asymp np^2$ as opposed to $C_{\text{Eig}}(n, p) \asymp np^4$, where $\asymp$ denotes that the function is asymptotically bounded above and below (see e.g. Temlyakov, *Greedy Approximation* (2011) for a formal definition of the notation). The number of samples processed is $n_{\text{Trace}}(N, p) \asymp N/p^2$ for the trace test and $n_{\text{Eig}}(N, p) \asymp N/p^4$ for the eigenvalue test.

Algorithm 2 Linear time hypothesis test for a non-zero precision matrix entry

Require: δ, the significance level of the test; μ, a constant satisfying (5.2.25); $\mathbf{X_p} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ a sample matrix variables of dimension p with sample size n.

Ensure:

1: Compute $\hat{\mathbf{\Sigma}}$, the unbiased estimator of Σ from $\mathbf{X_p}$ (cf. Definition. 5.1).

2: Compute $\hat{\mathbf{\Theta}} = \hat{\mathbf{\Sigma}}^{-1}$, the estimator of the precision matrix.

3: Compute $U([\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})])$ the upper triangular of the covariance of $U(\hat{\mathbf{\Sigma}})$ where (i, j, k, l) vary over the set of p variables (cf. Theorem 5.2).

4: Compute

- λ_{max}, the largest eigenvalue of $\mathrm{Cov}(\hat{\mathbf{\Sigma}})$, or
- $\mathrm{Tr}[\mathrm{Cov}(\hat{\mathbf{\Sigma}})]$, the trace of $\mathrm{Cov}(\hat{\mathbf{\Sigma}})$.

5: Compute one of the two error bounds ε (cf. Equations (5.2.30) and (5.2.31))

- $\varepsilon_{\mathrm{Eig}} = \sqrt{2\lambda_{max}}\Phi^{-1}(1 - \delta/2)$, or
- $\varepsilon_{\mathrm{Trace}} = \sqrt{2\,\mathrm{Tr}[\mathrm{Cov}(\hat{\mathbf{\Sigma}})]}\Phi^{-1}(1 - \delta/2)$

where Φ is the CDF of a standard normal distribution.

6: **if** ε is greater than the smallest eigenvalue of $\hat{\mathbf{\Sigma}}$ **then**

7: $t = \infty$

8: **else**

9: Compute the conservative threshold for the two error bound, $t = \mu\sqrt{\sum_{k=1}^{p}\left(\frac{-\varepsilon}{\hat{\alpha}_k(\hat{\alpha}_k - \varepsilon)}\right)^2}$, where $\hat{\alpha}_k$ is the k-th eigenvalue of the unbiased estimator $\hat{\mathbf{\Sigma}}$.

10: **end if**

11: **return** t.

For a full rank $(p^2 - \binom{p}{2}) \times (p^2 - \binom{p}{2})$ p.s.d. matrix, the trace is $\mathcal{O}(p^2\lambda_{max})$. We have when the sample sizes are equal $\varepsilon_{\text{Trace}} \in \mathcal{O}(p\varepsilon_{\text{Eig}})$. Furthermore, Equation (5.2.29) is asymptotically linear in ε as ε approaches zero from the right, and $\varepsilon_{\text{Eig}} \in \mathcal{O}(\lambda(p)n^{-1/2})$, where $\lambda(p)$ gives the dependence of ε_{Eig} on the dimensionality of the data. Therefore, at a fixed computational budget the eigenvalue threshold is $\mathcal{O}(\lambda(p)n_{\text{Eig}}(n,p)^{-1/2}) = \mathcal{O}(\lambda(p)(Np^{-4})^{-1/2}) = \mathcal{O}(\lambda(p)N^{-1/2}p^2)$, while the trace threshold is $\mathcal{O}(\lambda(p)p(n_{\text{Trace}}(n,p))^{-1/2}) = \mathcal{O}(\lambda(p)N^{-1/2}p^2)$ □

In the experiments, we set $\mu = 1$, which we have empirically validated to result in a sound test threshold for a wide range of distributions. As discussed below, for a trace threshold on a matrix with condition number $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$, the trace over-estimates Equation (5.2.24)-A by at least a factor of $1 + \frac{(p^2 - \binom{p}{2} - 1)\lambda_{\min}}{\lambda_{\max}}$, and the resulting test is therefore valid for distributions for which Equation (5.2.24)-B is asymptotically at most $\frac{p^2 - \binom{p}{2} - 1}{\kappa}$ as large as Equation (5.2.24)-A.

Theorem 5.7. *For a test with computational cost $\Omega(n^s)$ and a threshold that decreases as $\Omega(n^r)$, our test is asymptotically more powerful in the regime $n \gg p$ whenever $\frac{r}{s} > -\frac{1}{2}$.*

Proof: Our tests have computation $C_{\text{Trace}}(n) \asymp C_{\text{Eig}}(n) \asymp n$. The convergence of our test threshold is $\mathcal{O}(n^{-1/2})$ so for a fixed computational budget N, the test threshold is $\mathcal{O}(N^{-1/2})$. For a test with computational cost $\Omega(n^s)$ and a computational budget N, $\mathcal{O}(N^{1/s})$ samples will be processed. As n^r is decreasing in n for any consistent test, this implies that the test threshold is $\Omega(N^{r/s})$ which is asymptotically larger than $\mathcal{O}(N^{-1/2})$ whenever $\frac{r}{s} > -\frac{1}{2}$. □

Corollary 5.1. *Any test that is superlinear must have a threshold that converges faster than $\mathcal{O}(n^{-1/2})$ to be asymptotically more powerful at a fixed computational budget than the tests proposed here.*

In this section, we have derived two variants of a statistical hypothesis test that determines if an empirical estimate of an entry of the precision matrix significantly deviates from zero. We have shown that the two variants are asymptotically identical in the case that computation rather than data size is a limiting factor (Theorem 5.6), we have further demonstrated that our test is asymptotically more powerful than any method with superlinear computation and $\mathcal{O}(n^{-1/2})$ convergence (Corollary 5.1). In the next section, we show that these theoretical results are matched by empirical performance.

5.3 Experiments

In this section, we demonstrate the soundness and effectiveness of the proposed test for non-zero entries in a precision matrix, which enables one to answer if an edge is significantly present in a graphical model for a broad class of distributions. This is demonstrated both in terms of experiments on randomly generated graphical models with known analytic precision matrices Θ (Section 5.3.1), as well as and on real-world climate and weather data from the National Centers for Environmental Information [2016] (Section 5.3.2) and on real-world medical data

obtained from the Centers for Disease Control and Prevention [2014] (Section 5.3.3). In all experiments, we have used a significance upper bound of $\delta < 0.05$.

5.3.1 Synthetic Datasets

In these simulations, we generated the data sample $\mathbf{X}_p = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ of dimension p and size n with the mean of each $\mathbf{x}_i$ equal to zero, and a covariance matrix $\mathbf{\Sigma}$, from multivariate Gaussian or Laplace distributions with known analytic precision matrices $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, such that

$$\Sigma_{ij} = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2} \forall (i,j) \in \{1, ..., p\}. \tag{5.3.1}$$

1. Multivariate Gaussian distribution:

$$f(\mathbf{X}_p, \mathbf{\Sigma}) = \frac{1}{\sqrt{2\pi^p |\mathbf{\Sigma}|}} \exp\left\{-\frac{1}{2}\mathbf{X}_p^T \mathbf{\Theta} \mathbf{X}_p\right\}. \tag{5.3.2}$$

2. Multivariate Laplace distribution [Gómez, 1998]:

$$f(\mathbf{X}_p, \mathbf{\Sigma}) = \frac{p\Gamma\left(\frac{p}{2}\right)}{\pi^{\frac{p}{2}} \Gamma\left(1 + \frac{p}{\omega}\right) 2^{1+\frac{d}{\omega}} |\mathbf{\Sigma}|^{-\frac{1}{2}}} \exp\left\{\frac{1}{2}\left[\mathbf{X}_p^T \mathbf{\Theta} \mathbf{X}_p\right]^{\frac{\omega}{2}}\right\}, \tag{5.3.3}$$

   where $\Gamma(p)$ is the gamma function evaluated at p. For $\omega = 1$, the multivariate Laplace distribution is derived.

In Figure 5.1, we plot the sample size for 101 regularly spaced values of $n \in [10000, 1010000]$ versus the empirical threshold t_{Eig} and t_{Trace} (cf. Equation (5.2.29)) of the test. We clearly distinguish that the threshold t_{Eig} based on the eigenvalue bound in Equation (5.2.17) is less than the threshold t_{Trace} based on the trace bound in Equation (5.2.16) (see Lemma 5.1).

In Figure 5.2, we illustrate the inequality of Weyl's Theorem (Theorem 5.4). We show the boxplots of the eigenvalues of $\mathbf{\Theta}$ obtained from the simulation study. As expected, for a known precision matrix $\mathbf{\Theta}$, the eigenvalues $1/\alpha_i, i \in \{1, ...p\}$ is bounded by the two error bounds ε_{Eig} and $\varepsilon_{\text{Trace}}$. As the sample size n increases, the two bounds become tighter. We compare our hypothesis test with the eigenvalue threshold and the trace threshold (*edgeTest-eig* and *edgeTest-tr*) to the classical Fisher test (*FisherTest*) (cf. Section 2.5) for the Gaussian and the Laplace distributions. The simulations are repeated 100 times to provide statistical significance.

In Figure 5.3 we plot the significance level of the test δ against the false positive rate, which refers to the probability of falsely rejecting $\mathcal{H}_0$ for $n = 100000$ and $p = 6$. The diagonal dotted black line indicates that the significance level of different tests is equal to false positive rate. Curves above the diagonal indicate that the test does not obey the semantics of (a bound on) the false positive probability, while a curve under the diagonal indicates that the proposed test is conservative but sound. For the Gaussian distribution (Figure 5.3a), the Fisher test

(green curve) is well calibrated while the proposed test is sound. However, for the Laplace distribution (Figure 5.3b), the Fisher test is not valid. By contrast, the proposed test with the trace and the eigenvalue threshold (blue and pink curves) is sound when applied to more general families of distributions, including heavy tailed distributions such as the Laplacian.
Furthermore, we compare our method to permutation tests that are widely-applicable to non-parametric tests. The permutation procedure is a robust but computationally intensive alternative. Random shuffles of the data are used to get the correct distribution of a test statistic under a null hypothesis, and these sampling distributions are valid regardless of whether or not its distribution is known for any sample size. We calculate the test statistic for each resampling and the threshold o this procedure is choose to be the δ quantile of the distributions of the permuted test statistic. We found the error rates of permutation tests to be systematically higher than the target level (red curve) showing that the permutation test is not valid.
In addition to the comparison to the permutation tests and the Fisher tests, we have compare the proposed test to the Gaussian graphical model under a sparseness condition. At $\delta = 0.05$ we obtain an empirical false positive rate of 0.265, demonstrating that the violation of the test assumptions lead to an invalid test procedure.

In Figure 5.4, we compare the power of the tests by plotting the sample size for 101 regularly spaced values of $n \in [10000, 1010000]$ against the power of the test. We take into account an effect in the graph in the sense that we want to detect edge only when there is a non-negligible conditional dependence between two edges in the graph, i.e. when $|\Theta_{ij}| > 0.5$ for all $(i, j) \in \{1, ...p\}$ (see Lauritzen [1996]). We show that the power of the test approaches 1 after only a few thousand samples, and the linear scaling of the test means that it is applicable to millions of datapoints, enabling the discovery of subtle effects in complicated distributions.

5.3.2 Zone climate associations datasets

In the second experiment, we present our method on real-world measurements of the weather in Europe and on the east coast of the United States. We are particularly interested by modeling micro-climate dependencies as a function of the air temperature.

We have collected the datasets from the NCEI [National Centers for Environmental Information, 2016]. The NCEI acquires, historical weather datasets in the world. These data include quality controlled daily, monthly, seasonal, and yearly measurements of temperature recorded over decades, meaning that $n \gg p$. We have selected the temperature within a range of 50 miles for few cities in Europe and in the east coast of the United State and we have preprocessed the datasets by selecting common hourly measurements of temperature from 1942 to 2016 for each city. In Table 5.1, we present three collections of datasets, (I), (II) and (III) for different cities and millions of hourly measurements of temperatures.

In Table 5.1 and in Figure 5.5, we emphasize the fact that for these real-world experiments, the empirical distributions are non-Gaussian. For each city, we estimate the kurtosis and the resulting values (Kurt. < 3) indicate a platykurtic distribution (flatter than a Gaussian

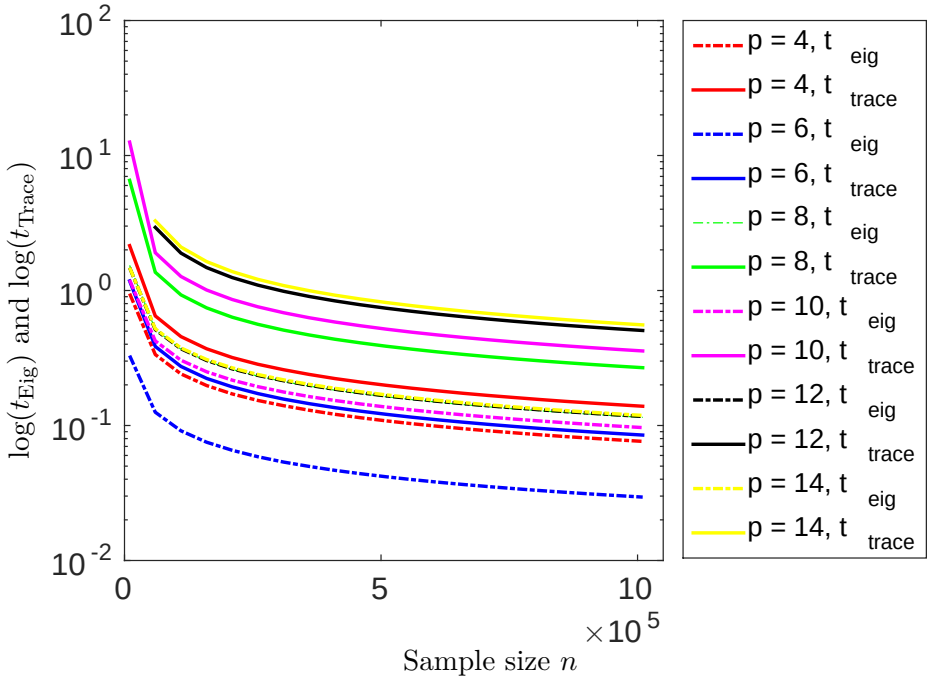


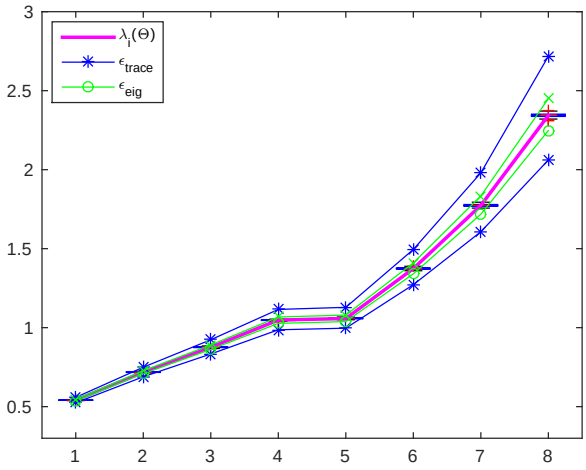
Figure 5.1 – Illustration of the sample size for 101 regularly spaces values of $n \in [10000, 1010000]$ versus the thresholds t_{Eig} and t_{Trace} (Equation (5.2.29)). We have plotted both the eigenvalue bound as well as the trace bound (cf. Lemma 5.1).

distribution with shorter tails). Additionally, we have performed one-sample Kolmogorov-Smirnov statistical tests for the normality of the empirical distributions. The resulting low p-value concludes that the empirical data are clearly non-Gaussian. Furthermore, in Figures. 5.7 and 5.8, we present the resulting undirected graphical model discovered with our statistical test for each collection. Our tests reject the null hypothesis with probability decreasing as a function of the distance between the cities, and that dependence between more distant cities can largely be explained by conditioning on cities lying in between.

Finally, we note that our test is capable of processing millions of observations with unoptimized Matlab code in a matter of seconds on a 2.80GHz CPU.

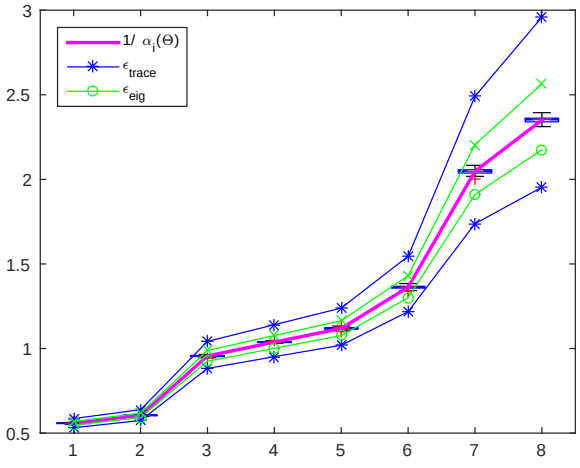
5.3.3 Risk Factors for Tuberculosis in the United States

In the third experiment, we evaluate our method on real-world data of Tuberculosis (TB) cases. TB is a potentially serious infectious disease that mainly affects the lungs. The factors responsible for TB are multiple but in the United States, the most important are problems of poverty, homelessness, and poor access to health care, which have combined to help maintain a reservoir of infected persons [Narasimhan et al., 2013, Suchindran et al., 2009]. The addition of HIV-associated immunodeficiency has had an observable impact on the incidence of TB. In the last decades, the number of new TB cases reported annually in the United States has increased and the TB morbidity has consistently demonstrated the burden of TB. Multiple



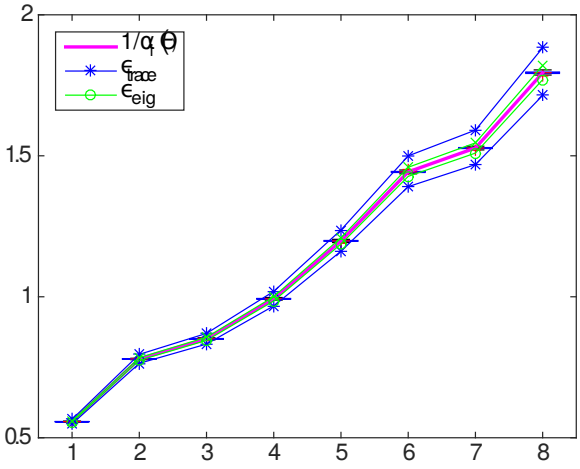
$1/\alpha_i(\mathbf{\Theta})$, for $i \in \{1, ..., 8\}$.

(a) Gaussian distribution, $n = 100000$, $p = 6$.



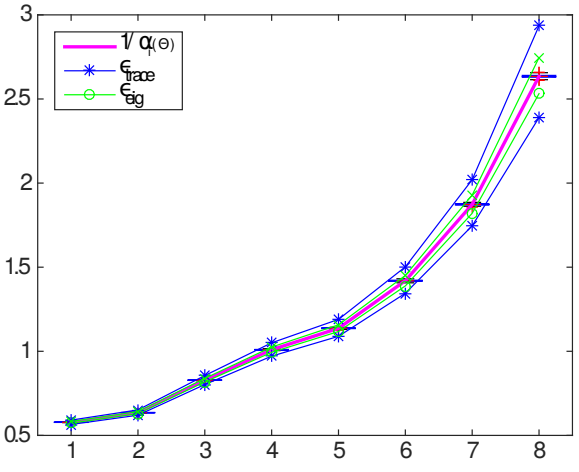
$1/\alpha_i(\mathbf{\Theta})$, for $i \in \{1, ..., 8\}$.

(b) Laplace distribution, $n = 100000$, $p = 6$.



$1/\alpha_i(\mathbf{\Theta})$, for $i \in \{1, ..., 8\}$.

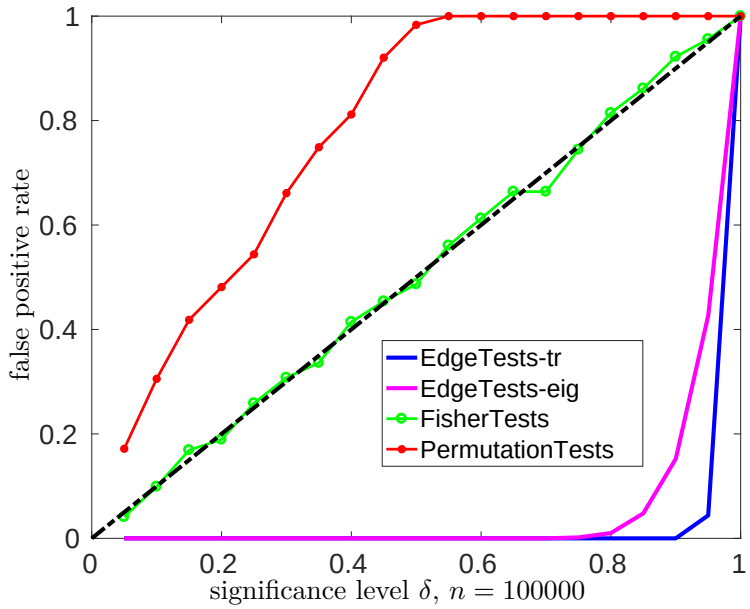
(c) Gaussian dist, $n = 500000$, $p = 6$.



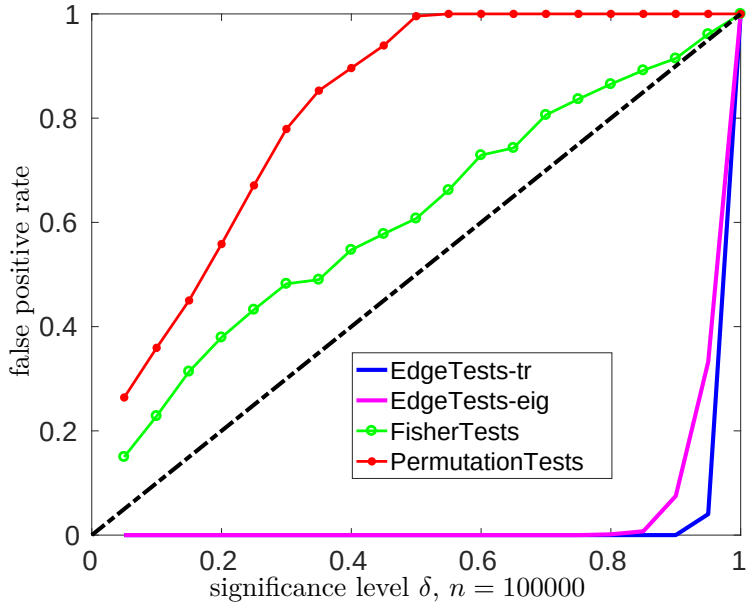
$1/\alpha_i(\mathbf{\Theta})$, for $i \in \{1, ..., 8\}$.

(d) Laplace distribution, $n = 500000$, $p = 6$.

Figure 5.2 – For a known analytic precision matrix $\mathbf{\Theta}$ of size $p = 8$ and for two different sample sizes, we show the boxplots of accuracy values of eigenvalues of 200 estimates matrices $\hat{\Theta}$ for the Gaussian (Figures 5.2a, 5.2c) and Laplace (Figures 5.2b, 5.2d) distributions with normalized data. In pink, we plot the true eigenvalue of $\mathbf{\Theta}$ and in green and blue, we plot the upper and lower bound given by Weyl's theorem. As n grows, we see that the bound more closely constrains the true eigenvalues of $\mathbf{\Theta}$.

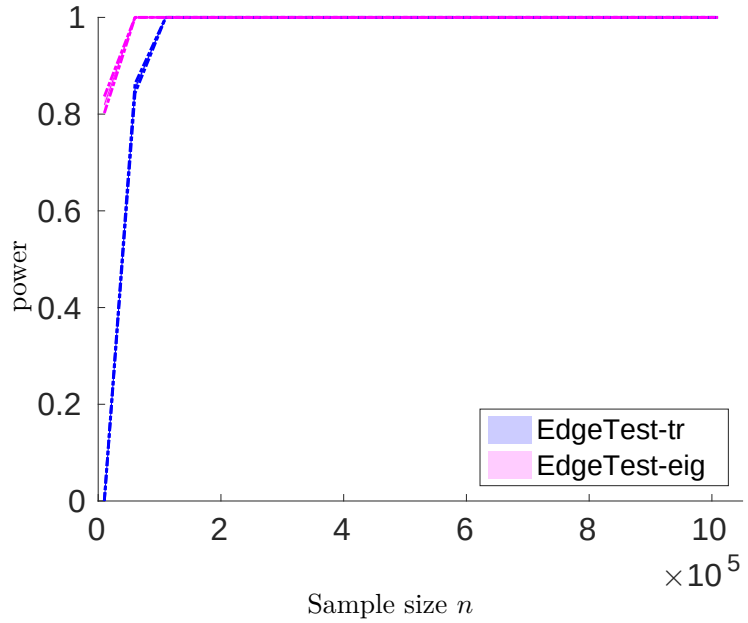


(a) Gaussian distribution

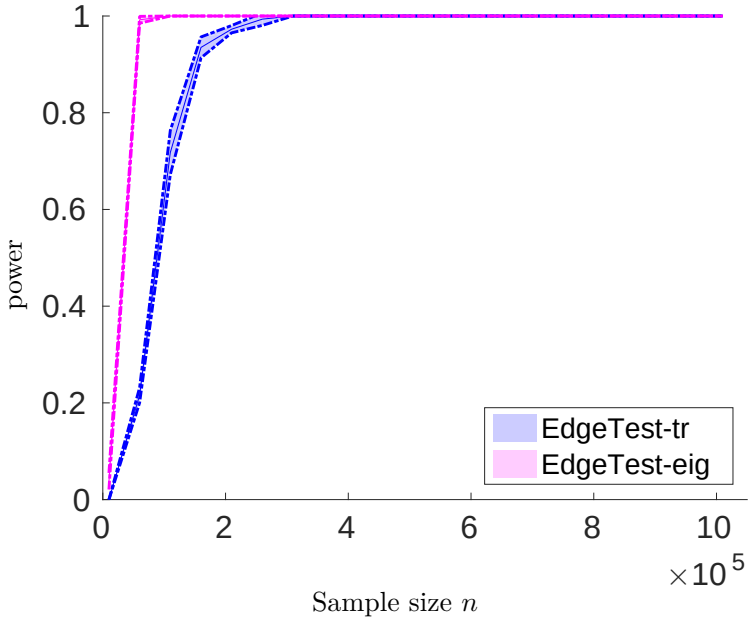


(b) Laplace distribution

Figure 5.3 – Comparison of the false positive rate for the proposed test, the Fisher test and the permutation test. For the Gaussian distribution (Figure 5.3a), the curves show that the Fisher test is well calibrated and that the proposed test is conservative (below the diagonal). For the Laplace distribution (Figure 5.3b), the Fisher test does not obey the semantics of a bound on δ (the curve is above the diagonal). By contrast, the proposed tests remains conservative and sound. In addition, the permutation test dost not obey the semantics of a bound on δ for both distributions. An explanation for the overly high false positive rate of the permutation test is that the permutations destroy the underlying edge distribution of the graph resulting in an incorrect estimate of the distribution of the statistic under the null hypothesis.

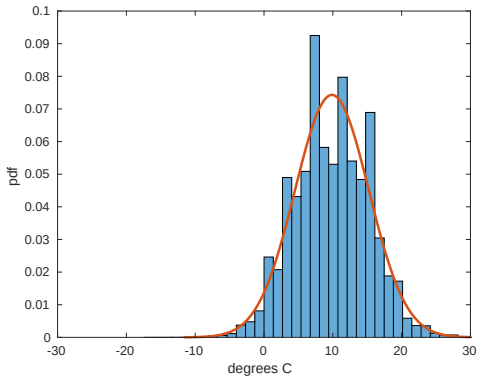


(a) Gaussiandistribution

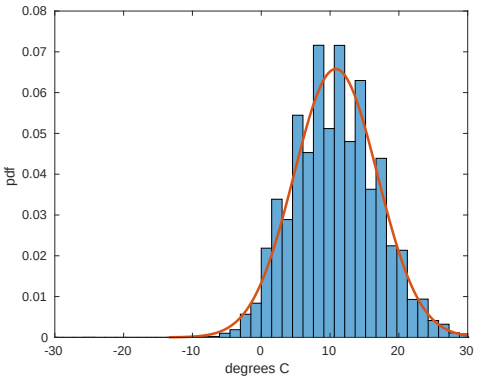


(b) Laplace distribution

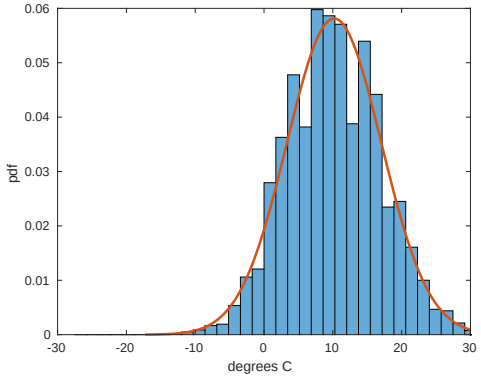
Figure 5.4 – Comparison of the powers of the proposed test using the two bounds as a function of n,when we reject the null hypothesis and when there is a large magnitude entry of $\boldsymbol{\Theta}$, here when $|\Theta_{ij}| > 0.5$.



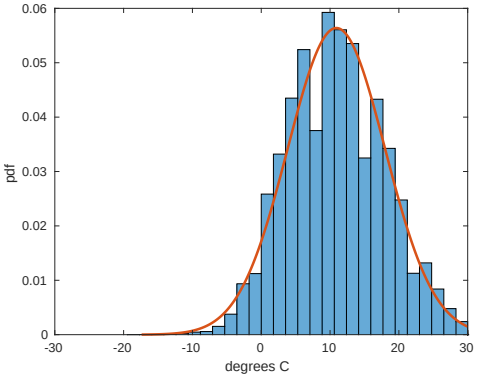
(a) Brussels



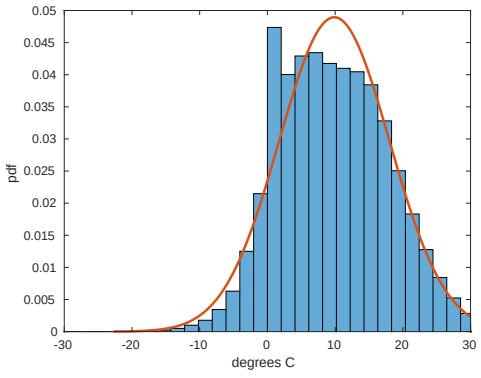
(b) London



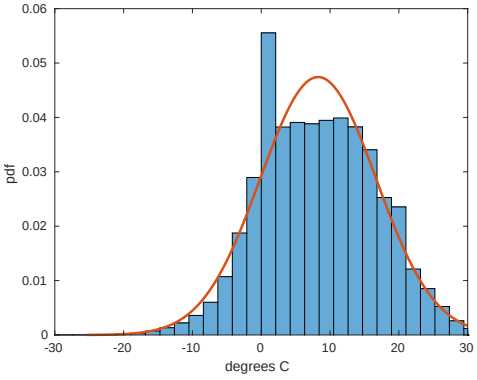
(c) Munich



(d) Paris

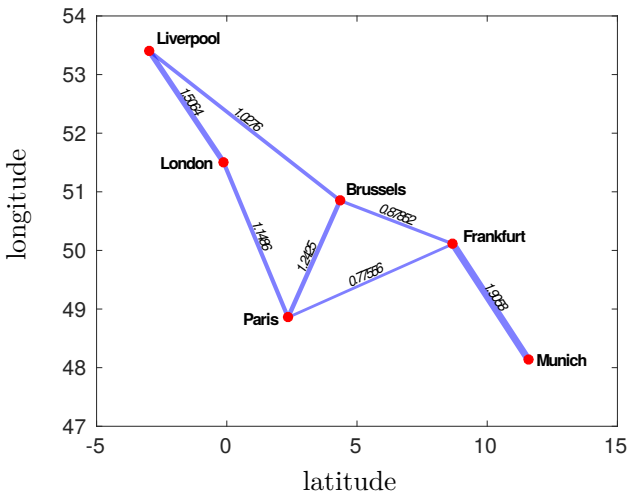


(e) Liverpool

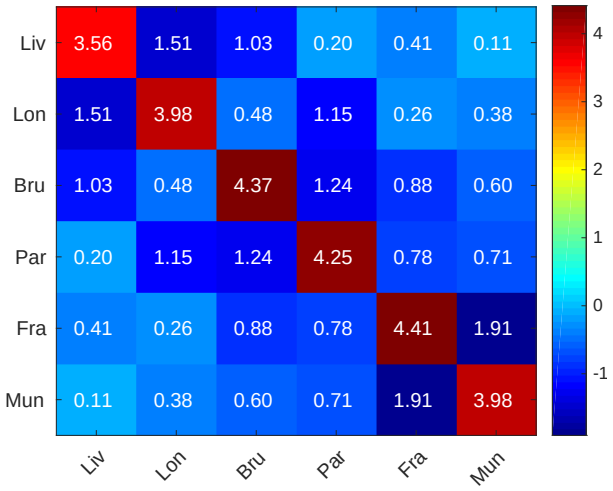


(f) Frankfurt

Figure 5.5 – Empirical distribution of the weather datasets for different cities for the collection(I).



(a)

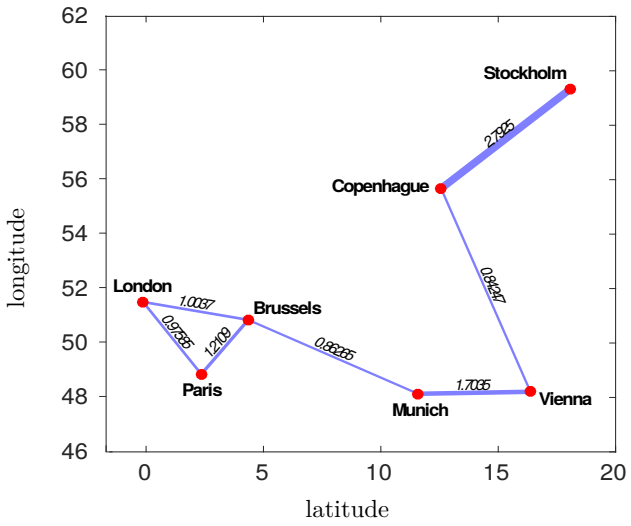


(b)

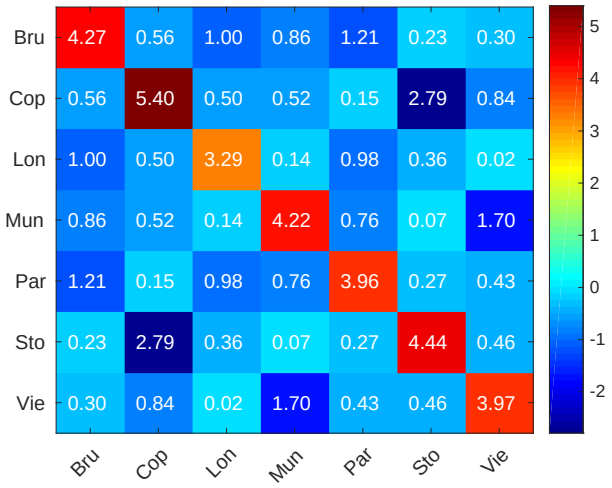
$$
\begin{array}{c} \\ Liv. \\ Lon. \\ Bru. \\ Par. \\ Fran. \\ Mun. \end{array}
\begin{array}{c}
\begin{array}{cccccc} Liv. & Lon. & Bru. & Par. & Fran. & Mun. \end{array} \\
\left(\begin{array}{cccccc}
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1
\end{array} \right)
\end{array}
$$

(c)

Figure 5.6 – Collection (I): Illustrations of the undirected graph with weight edges (Figure 5.6a), the absolute value of the test statistic matrix between the different cities (Figure 5.6b), with a threshold of $t_{eig} = 0.7279$ and the adjacency matrix showing the significant association between the cities (Figure 5.6c).



(a)

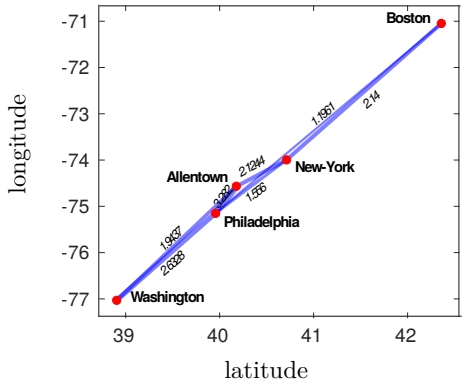


(b)

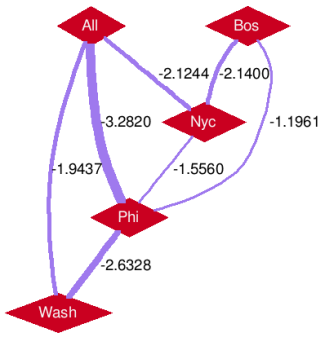
$$
\begin{array}{c}
 \\
Bru. \\
Cop. \\
Lon. \\
Mun. \\
Par. \\
Sto. \\
Vie.
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
Bru. & Cop. & Lon. & Mun. & Par. & Sto. & Vie
\end{array} \\
\left(\begin{array}{ccccccc}
1 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 1
\end{array}\right)
\end{array}
$$

(c)

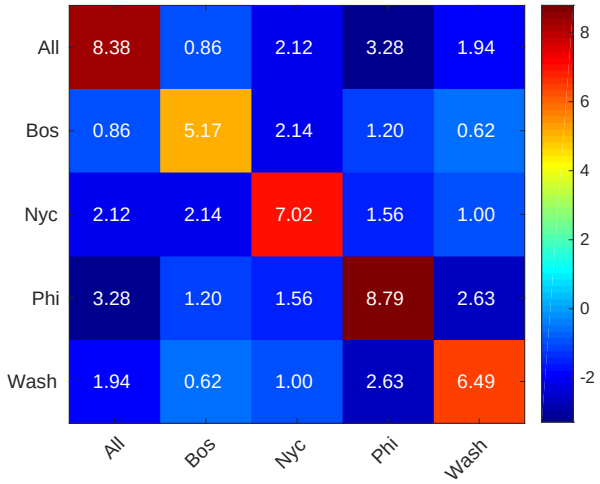
Figure 5.7 – Collection (II): Illustrations of the undirected graph with weight edges (Figure 5.7a), the absolute value of the test statistic matrix between the different cities (Figure 5.7b), with a threshold of $t_{eig} = 0.7899$ and the adjacency matrix showing the significant association between the cities (Figure 5.7c).



(a)



(b)



(c)

$$
\begin{array}{c} \\ All \\ Bos \\ Nyc \\ Phi. \\ Wash. \end{array}
\begin{array}{c}
\begin{array}{ccccc} All. & Bos. & Nyc. & Phi. & Wash. \end{array} \\
\left(\begin{array}{ccccc}
1 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 1
\end{array} \right)
\end{array}
$$

(d)

Figure 5.8 – Collection (III): Illustrations of the undirected graph with weight edges (Figures. 5.8a and 5.8b), the absolute value of the test statistic matrix between the different cities (Figure 5.8c), with a threshold of $t_{eig} = 1.0958$ and the adjacency matrix showing the significant association between the cities (Figure 5.8d).

Collection (I) $n = 1,266,463$		**Collection (II)** $n = 1,971,449$		**Collection (III)** $n = 3,463,949$	
Cities	Kurt.	Cities	Kurt.	Cities	Kurt.
Liverpool (Liv)	2.8947	Brussels (Bru)	2.8883	Allentown (All)	2.1786
London (Lon)	2.8773	Copenhagen (Cop)	2.6244	Boston (Bos)	2.4226
Brussels (Bru),	2.8954	London (Lon)	2.8350	New–York (Nyc)	2.2426
Paris (Par),	2.8146	Munich (Mun)	2.7201	Philadelphia (Phi)	2.2014
Frankfurt (Fra)	2.8561	Paris (Par)	2.7778	Washington ($Wash$)	2.1802
Munich (Mun)	2.8034	Stockholm(Sto)	2.7614		
		Vienna(Vie)	2.4138		

Table 5.1 – Description of the three collections of the datasets. For each city, the low values of the estimate kurtosis show a fatter tail. Additionally, the one-sample Kolmogorov-Smirnov statistical test of normality yields a p-value smaller than numerical precision.

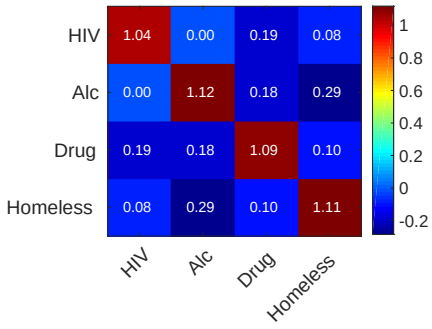
factors contributed to the recent increases in the number of TB cases. With the resurgence of tuberculosis in the United States, this is of interest to provide significant and effective information for the targeting of efforts to control tuberculosis and to reverse that trend.

We studied independently factors that increase the risk of being infected and the risk of infection leading on to active disease. The Online Tuberculosis Information System (OTIS) contains information on verified TB cases reported to the Centers for Disease Control and Prevention [2014]. The data set consist of n = 163,997 tuberculosis cases for which we have selected 4 causes of being infected: Positive to HIV (HIV); Alcohol use (Alc); Drug use ($Drug$) and Homeless in the past year ($Homeless$).

The proposed tests show a significant association between, "HIV positive - Homeless in past year" and "drug use - Homeless in past year" showing that environmental factors represent important common risk factors for TB as shown in Figure 5.9. This association that has long been highlighted in the medical literature [Narasimhan et al., 2013, Suchindran et al., 2009] and our hypothesis test produces consistent findings.

5.4 Discussion and Conclusion

We have considered the problem of structure discovery for undirected graphical models in the context of non-Gaussian multivariate distributions. By using a concentration bound from the theory of U-statistics, we have developed two sound test thresholds t_{Eig} and t_{Trace}. As a baseline, we compare to the Fisher test which is only correct under the assumption of a Gaussian distribution. As shown in the simulation studies, for non-Gaussian distributions, the Fisher test is not calibrated, while alternatively, the proposed test is sound. Among the two bounds presented here, the eigenvalue bound is preferred when the availability of data is more limited than computation, while t_{Trace} is a competitive test when we have a fixed computational budget N that is exceeded by the amount of available data. In this work,



(a)

	HIV	*Alc.*	*Drug*	*Homeless*
HIV	1	0	1	0
Alc.	0	1	1	1
Drug	1	1	1	0
Homeless	0	1	0	1

(b)

Figure 5.9 – Illustration of the absolute value of the test statistic matrix between the risk factors (Figure 5.9a), with a threshold of $t_{eig} = 0.1170$ and the adjacency matrix showing the significant association between the risk factors (Figure 5.9b).

we have constructed a conservative threshold on the absolute value of the precision matrix as a hypothesis test of the presence of an edge in a graphical model. For a wider range of distributions, we have developed a threshold based on a U-statistic empirical estimator of the covariance matrix. This is achieved by probabilistically bounding the distortion of the true covariance matrix, and then using this fixed bound in conjunction with Weyl's theorem to bound the distortion of the precision matrix. These bounds are applicable to the quantification of uncertainty in the magnitude of an effect between variables as measured by the value of the precision matrix, and can also be used to construct a hypothesis test of whether an edge is present in a graphical model by testing for significant deviations from zero. The resulting test asymptotically converges at the same $\mathcal{O}(n^{-1/2})$ rate as the U-statistic, which we have additionally verified empirically. We have shown two alternative thresholds, one based on the largest eigenvalue of $\mathrm{Cov}(\hat{\mathbf{\Sigma}})$, and a second based on the trace of $\mathrm{Cov}(\hat{\mathbf{\Sigma}})$, which strictly upper bounds the first. Simulation studies show that the test successfully recovers the structure of undirected graphical models given a sufficient number of samples. Theorem 5.7 implies that more expensive tests will eventually be dominated by our test. However, there may be situations in which subsampling and running a different test may give higher power if data are limitated. Comparison, e.g. to Wasserman et al. [2014], in such a setting is therefore an interesting direction for future work. We have empirically demonstrated the test performance on synthetic data, on the challenging problem of understanding geographic dependencies in weather, and on the problem of identifying conditionally dependent risk factors for having tuberculosis. We have demonstrated that our method is scalable to millions of data points in under a minute on a standard CPU using a direct implementation of Algorithm 2.

5.5 Proofs

In this section, we show the details of the derivation of Theorem 5.3. We first provide a short theoretical reminder of the notation in this context using the Theorem 2.3 of the covariance of two U-statistics with the corresponding U-statistic kernel of degree 2.

We derive low variance, unbiased estimates of the covariance between two U-statistics estimates $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$, where (i, j, k, l) range over each of the d variates in a covariance matrix $\hat{\boldsymbol{Sigma}}$. We note h and g the corresponding kernel of order 2 for $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$, where

$$h(u_1, u_2) = \frac{1}{2}\left(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\right)\left(\mathbf{x}_{j_1} - \mathbf{x}_{j_2}\right), \text{with } \mathbf{u}_r = (\mathbf{x}_{i_r}, \mathbf{x}_{j_r})^T \tag{5.5.1}$$

$$g(v_1, v_2) = \frac{1}{2}\left(\mathbf{x}_{k_1} - \mathbf{x}_{k_2}\right)\left(\mathbf{x}_{l_1} - \mathbf{x}_{l_2}\right), \text{with } \mathbf{v}_r = (\mathbf{x}_{k_r}, \mathbf{x}_{l_r})^T. \tag{5.5.2}$$

Then, using Theorem 2.3, the covariance $\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$ for the two U-statistics $\hat{\Sigma}_{ij}$ and $\hat{\Sigma}_{kl}$ is

$$\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl}) = \binom{n}{2}^{-1} (2(n-2)\zeta_1 + \zeta_2) \tag{5.5.3}$$

$$= \binom{n}{2}^{-1} (2(n-2)\zeta_1) + \mathcal{O}(n^{-2}), \tag{5.5.4}$$

where

$$\zeta_1 = \mathrm{Cov}\left(\mathbb{E}_{\mathbf{u}_2}[h(\mathbf{u}_1, \mathbf{u}_2)], \mathbb{E}_{\mathbf{v}_2}[g(\mathbf{v}_1, \mathbf{v}_2)]\right) \tag{5.5.5}$$

where $\mathbb{E}_{\mathbf{u}_2} = \mathbb{E}_{\mathbf{x}_{i_2}, \mathbf{x}_{j_2}}$ denotes the integral of the function $h\left(\mathbf{x}_{i_1}, \mathbf{x}_{j_1}, \mathbf{x}_{i_2}\mathbf{x}_{j_2}\right)$ with respect with respect to the variable of integration are $\mathbf{x}_{i_2}, \mathbf{x}_{j_2}$. If the distribution from $\mathbb{P}_{\mathbf{x}}$ has a density f then

$$\mathbb{E}_{\mathbf{u}_2}[h(\mathbf{u}_1, \mathbf{u}_2)] = \mathbb{E}_{\mathbf{x}_{i_2}, \mathbf{x}_{j_2}}[h(\mathbf{x}_{i_1}, \mathbf{x}_{j_1}, \mathbf{x}_{i_2}\mathbf{x}_{j_2})] \tag{5.5.6}$$

$$= \int_{\mathbb{R}} h\left(\mathbf{x}_{i_1}, \mathbf{x}_{j_1}, \mathbf{x}_{i_2}\mathbf{x}_{j_2}\right) f\left(\mathbf{x}_{i_2}, \mathbf{x}_{j_2}\right) \mathrm{d}\mathbf{x}_{i_2}\mathrm{d}\mathbf{x}_{j_2} \tag{5.5.7}$$

$$\mathbb{E}_{\mathbf{v}_2}[g(\mathbf{v}_1, \mathbf{v}_2)] = \mathbb{E}_{\mathbf{x}_{k_2}, \mathbf{x}_{l_2}}[g(\mathbf{x}_{k_1}, \mathbf{x}_{l_1}, \mathbf{x}_{k_2}\mathbf{x}_{l_2})] \tag{5.5.8}$$

$$= \int_{\mathbb{R}} g\left(\mathbf{x}_{k_1}, \mathbf{x}_{l_1}, \mathbf{x}_{k_2}\mathbf{x}_{l_2}\right) f\left(\mathbf{x}_{k_2}, \mathbf{x}_{l_2}\right) \mathrm{d}\mathbf{x}_{k_2}\mathrm{d}\mathbf{x}_{l_2}. \tag{5.5.9}$$

We now present a proof of Theorem 5.3.

5.5.1 Description of the algorithm providing the seven cases

We formally described the algorithm that provided us 7 cases for the derivation of $\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$ of Theorem 5.3, where (i, j, k, l) vary over the set of p variables.

Enumeration First, we enumerate all configurations of $\mathrm{Cov}(\hat{\Sigma}_{ij}, \hat{\Sigma}_{kl})$, which can be encoded as a non-unique assignment matrix of variables i, j, k, l to instantiated variables (a, b, c, d). For a fixed assignment of i to variable a, we can list all possible assignments of the 3 remaining variables (j, k, l) to any (a, b, c, d). Naïvely, we have 4^3 possible assignments, but many of them will be equivalent by variable substitution. To test whether two forms are equivalent, it is sufficient to test a reduced form for equality.

Reduced Form We map a variable assignment to a reduced form by re-labeling variables sorted by the number of occurrences, which reduces the number of possible matches up-to non-uniqueness of the mapping due to equal numbers of variable occurrences. This ambiguity is then resolved by testing for symmetries.

Symmetry Symmetry of the covariance operator brings the following equally that we take into consideration in testing for equivalence:

$$\begin{aligned}\mathrm{Cov}(\hat{\Sigma}_{ij},\hat{\Sigma}_{kl}) &= \mathrm{Cov}(\hat{\Sigma}_{kl},\hat{\Sigma}_{ij}) = \mathrm{Cov}(\hat{\Sigma}_{ij},\hat{\Sigma}_{lk}) = \mathrm{Cov}(\hat{\Sigma}_{lk},\hat{\Sigma}_{ij}) \\ &= \mathrm{Cov}(\hat{\Sigma}_{lk},\hat{\Sigma}_{ji}) = \mathrm{Cov}(\hat{\Sigma}_{ji},\hat{\Sigma}_{kl}) = \mathrm{Cov}(\hat{\Sigma}_{ji},\hat{\Sigma}_{lk}).\end{aligned} \tag{5.5.10}$$

The algorithm outputs each variable assignment that is not equivalent by variable substitution to any previously enumerated assignment. The resulting seven cases are given in Table 5.2.

Cases	Indices	Correspondence
1	$i \neq j,k,l;\ j \neq k,l;\ k \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ij},\hat{\Sigma}_{kl})$
2	$i = j;\ j \neq k,l;\ k = l$	$\mathrm{Cov}(\hat{\Sigma}_{ii},\hat{\Sigma}_{kk})$
3	$i = j;\ j \neq k,l;\ k \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ii},\hat{\Sigma}_{kl})$
4	$i = k;\ j \neq i,k,l;\ k \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ij},\hat{\Sigma}_{il})$
5	$i = k;\ i \neq j;\ j = l;$	$\mathrm{Var}(\hat{\Sigma}_{ij})$
6	$i = j = k;\ i \neq l$	$\mathrm{Cov}(\hat{\Sigma}_{ii},\hat{\Sigma}_{il})$
7	$i = j,k,l$	$\mathrm{Var}(\hat{\Sigma}_{ii})$

Table 5.2 – Enumeration and correspondence of the seven cases.

5.5.2 The seven exhaustive cases

We now derive linear-time finite-sample estimates of the covariance for each of the seven cases. We note $\overline{\mathbf{xyuv}} = \mathbb{E}[\mathbf{xyuv}]$, $\overline{\mathbf{xyz}} = \mathbb{E}[\mathbf{xyz}]$, $\overline{\mathbf{xy}} = \mathbb{E}[\mathbf{xy}]$, $\overline{\mathbf{x}} = \mathbb{E}[\mathbf{x}]$, and $\overline{\mathbf{xyuv}}\ \overline{\mathbf{x}} = \mathbb{E}[\mathbf{xyuv}]\mathbb{E}[\mathbf{x}]$.

▶ **Case 1**: $i \neq j,k,l;\ j \neq k,l;\ k \neq l$
The corresponding U-statistic kernels for this case are

$$h(\mathbf{u}_1,\mathbf{u}_2) = \frac{1}{2}\left(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\right)\left(\mathbf{x}_{j_1} - \mathbf{x}_{j_2}\right), \text{ and} \tag{5.5.11}$$

$$g(\mathbf{v}_1,\mathbf{v}_2) = \frac{1}{2}\left(\mathbf{x}_{k_1} - \mathbf{x}_{k_2}\right)\left(\mathbf{x}_{l_1} - \mathbf{x}_{l_2}\right), \tag{5.5.12}$$

then, if the distribution from $\mathbb{P}_{\mathbf{x}}$ has a density f, the expectation in Equation (5.5.7) is

$$\begin{aligned}
\mathbb{E}_{\mathbf{u}_2}[h(\mathbf{u}_1, \mathbf{u}_2)] &= \int_{\mathbb{R}} h\left(\mathbf{x}_{i_1}, \mathbf{x}_{j_1}, \mathbf{x}_{i_2}\mathbf{x}_{j_2}\right) f\left(\mathbf{x}_{i_2}, \mathbf{x}_{j_2}\right) \mathrm{d}\mathbf{x}_{i_2}\mathrm{d}\mathbf{x}_{j_2} \\
&= \int_{\mathbb{R}} \frac{1}{2}\left(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\right)\left(\mathbf{x}_{j_1} - \mathbf{x}_{j_2}\right) f\left(\mathbf{x}_{i_2}, \mathbf{x}_{j_2}\right) \mathrm{d}\mathbf{x}_{i_2}\mathrm{d}\mathbf{x}_{j_2} \\
&= \frac{1}{2}\int_{\mathbb{R}} \left(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\right)\left(\mathbf{x}_{j_1} - \mathbf{x}_{j_2}\right) f\left(\mathbf{x}_{i_2}\right) f\left(\mathbf{x}_{j_2}\right) \mathrm{d}\mathbf{x}_{i_2}\mathrm{d}\mathbf{x}_{j_2} \\
&= \frac{1}{2}\left(\int_{\mathbb{R}} \left(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\right) f\left(\mathbf{x}_{i_2}\right) \mathrm{d}\mathbf{x}_{i_2}\right)\left(\int_{\mathbb{R}} \left(\mathbf{x}_{j_1} - \mathbf{x}_{j_2}\right) f\left(\mathbf{x}_{j_2}\right) \mathrm{d}\mathbf{x}_{j_2}\right) \\
&= \frac{1}{2}\Bigg\{\left(\mathbf{x}_{i_1}\int_{\mathbb{R}} f\left(\mathbf{x}_{i_2}\right) \mathrm{d}\mathbf{x}_{i_2} - \int_{\mathbb{R}} \mathbf{x}_{i_2} f\left(\mathbf{x}_{i_2}\right) \mathrm{d}\mathbf{x}_{i_2}\right) \\
&\qquad \left(\mathbf{x}_{j_1}\int_{\mathbb{R}} f\left(\mathbf{x}_{j_2}\right) \mathrm{d}\mathbf{x}_{j_2} - \int_{\mathbb{R}} \mathbf{x}_{j_2} f\left(\mathbf{x}_{j_2}\right) \mathrm{d}\mathbf{x}_{j_2}\right)\Bigg\} \\
&= \frac{1}{2}\Big\{\left(\mathbf{x}_{i_1} - \overline{\mathbf{x}_i}\right)\left(\mathbf{x}_{j_1} - \overline{\mathbf{x}_j}\right)\Big\},
\end{aligned} \tag{5.5.13}$$

where we make use of the i.i.d. properties of $\mathbf{x}_i$ and $\mathbf{x}_j$, and we have that the probability density function f satisfy the condition

$$\int f\left(\mathbf{x}\right) d\mathbb{P}_{\mathbf{x}} = 1 \tag{5.5.14}$$

and where $\overline{\mathbf{x}_i}$ represent the mean of the sample $\mathbf{x}_{i_1}$. Similarly, using the same derivations than in Equation (5.5.13), we have that the Equation (5.5.9) is equal to

$$\mathbb{E}_{\mathbf{u}_2}[g(\mathbf{v}_1, \mathbf{v}_2)] = \frac{1}{2}\Big\{\left(\mathbf{x}_{k_1} - \overline{\mathbf{x}_k}\right)\left(\mathbf{x}_{l_1} - \overline{\mathbf{x}_l}\right)\Big\}. \tag{5.5.15}$$

Finally, we have that ζ_1 is equal to

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(\mathbf{x}_{i_1} - \overline{\mathbf{x}_i}\right)\left(\mathbf{x}_{j_1} - \overline{\mathbf{x}_j}\right), \frac{1}{2}\left(\mathbf{x}_{k_1} - \overline{\mathbf{x}_k}\right)\left(\mathbf{x}_{l_1} - \overline{\mathbf{x}_l}\right)\right] \qquad (5.5.16)\\
&= \frac{1}{4}\Big\{\mathrm{Cov}\left[\mathbf{x}_{i_1}\mathbf{x}_{j_1} - \overline{\mathbf{x}_i}\mathbf{x}_{j_1} - \mathbf{x}_{i_1}\overline{\mathbf{x}_j}; \mathbf{x}_{k_1}\mathbf{x}_{l_1} - \overline{\mathbf{x}_k}\mathbf{x}_{l_1} - \mathbf{x}_{k_1}\overline{\mathbf{x}_l}\right]\Big\}\\
&= \frac{1}{4}\Big\{\mathbb{E}_{u_1}\left[\mathbf{x}_{i_1}\mathbf{x}_{j_1}\mathbf{x}_{k_1}\mathbf{x}_{l_1} - \overline{\mathbf{x}_i}\mathbf{x}_{j_1}\mathbf{x}_{k_1}\mathbf{x}_{l_1} - \mathbf{x}_{i_1}\overline{\mathbf{x}_j}\mathbf{x}_{k_1}\mathbf{x}_{l_1}\right.\\
&\qquad\qquad - \mathbf{x}_{i_1}\mathbf{x}_{j_1}\overline{\mathbf{x}_k}\mathbf{x}_{l_1} + \overline{\mathbf{x}_i}\mathbf{x}_{j_1}\overline{\mathbf{x}_k}\mathbf{x}_{l_1} + \mathbf{x}_{i_1}\overline{\mathbf{x}_j}\ \overline{\mathbf{x}_k}\mathbf{x}_{l_1}\\
&\qquad\qquad \left. - \mathbf{x}_{i_1}\mathbf{x}_{j_1}\mathbf{x}_{k_1}\overline{\mathbf{x}_l} + \overline{\mathbf{x}_i}\mathbf{x}_{j_1}\mathbf{x}_{k_1}\overline{\mathbf{x}_l} + \mathbf{x}_{i_1}\overline{\mathbf{x}_j}\mathbf{x}_{k_1}\overline{\mathbf{x}_l}\right]\\
&\qquad - \mathbb{E}_{u_1}\left[\mathbf{x}_{i_1}\mathbf{x}_{j_1} - \overline{\mathbf{x}_i}\mathbf{x}_{j_1} - \mathbf{x}_{i_1}\overline{\mathbf{x}_j}\right]\mathbb{E}_{u_1}\left[\mathbf{x}_{k_1}\mathbf{x}_{l_1} - \overline{\mathbf{x}_k}\mathbf{x}_{l_1} - \mathbf{x}_{k_1}\overline{\mathbf{x}_l}\right]\Big\}\\
&= \frac{1}{4}\Big\{\overline{\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k\mathbf{x}_l} - \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j\mathbf{x}_k\mathbf{x}_l} - \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_i\mathbf{x}_k\mathbf{x}_l}\\
&\qquad - \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_i\mathbf{x}_j\mathbf{x}_l} + \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_j\mathbf{x}_l} + \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_i\mathbf{x}_l}\\
&\qquad - \overline{\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k}\ \overline{\mathbf{x}_l} + \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l}\ \overline{\mathbf{x}_j\mathbf{x}_k} + \overline{\mathbf{x}_j}\ \overline{\mathbf{x}_l}\ \overline{\mathbf{x}_i\mathbf{x}_k}\\
&\qquad - \left(\overline{\mathbf{x}_i\mathbf{x}_j} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_j}\right)\left(\overline{\mathbf{x}_k\mathbf{x}_l} - 2\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_l}\right)\Big\}.
\end{aligned}$$

▶ **Case 2**: $i = j;\ j \neq k, l;\ k = l$

The corresponding U-statistic kernels for this case are

$$h(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2}\left(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\right)^2, \text{ and} \qquad (5.5.17)$$

$$g(\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{2}\left(\mathbf{x}_{k_1} - \mathbf{x}_{k_2}\right)^2. \qquad (5.5.18)$$

Using the same derivations than in (5.5.13), we obtain

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(\mathbf{x}_{i_1} - \overline{\mathbf{x}_i}\right)^2; \frac{1}{2}\left(\mathbf{x}_{k_1} - \overline{\mathbf{x}_k}\right)^2\right] \qquad (5.5.19)\\
&= \frac{1}{4}\Big\{\mathrm{Cov}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}; \mathbf{x}_{k_1}^2 - 2\mathbf{x}_{k_1}\overline{\mathbf{x}_k}\right]\Big\}\\
&= \frac{1}{4}\Big\{\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}^2\mathbf{x}_{k_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}\mathbf{x}_{k_1}^2 - 2\mathbf{x}_{i_1}^2\mathbf{x}_{k_1}\overline{\mathbf{x}_k} + 4\mathbf{x}_{i_1}\overline{\mathbf{x}_i}\mathbf{x}_{k_1}\overline{\mathbf{x}_k}\right]\\
&\qquad - \mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}\right]\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{k_1}^2 - 2\mathbf{x}_{k_1}\overline{\mathbf{x}_k}\right]\Big\}\\
&= \frac{1}{4}\Big\{\overline{\mathbf{x}_i^2\mathbf{x}_k^2} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_i\mathbf{x}_k^2} - 2\ \overline{\mathbf{x}_i^2\mathbf{x}_{k_1}}\ \overline{\mathbf{x}_k} + 4\overline{\mathbf{x}_i\mathbf{x}_k}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_k}\\
&\qquad - \left(\overline{\mathbf{x}_i^2} - 2\ \overline{\mathbf{x}_i}^2\right)\left(\overline{\mathbf{x}_k^2} - 2\ \overline{\mathbf{x}_k}^2\right)\Big\}.
\end{aligned}$$

▶ **Case 3**: $i = j;\ j \neq k, l;\ k \neq l$

The corresponding U-statistic kernels for this case are

$$h(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2}\left(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\right)^2, \text{ and} \tag{5.5.20}$$

$$g(\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{2}\left(\mathbf{x}_{k_1} - \mathbf{x}_{k_2}\right)\left(\mathbf{x}_{l_1} - \mathbf{x}_{l_2}\right). \tag{5.5.21}$$

And, ζ_1 is equal to

$$\begin{aligned}
\zeta_1 &= \operatorname{Cov}\left[\frac{1}{2}\left(\mathbf{x}_{i_1} - \overline{\mathbf{x}_i}\right)^2; \frac{1}{2}\left(\mathbf{x}_{k_1} - \overline{\mathbf{x}_k}\right)\left(\mathbf{x}_{l_1} - \overline{\mathbf{x}_l}\right)\right] \\
&= \frac{1}{4}\Big\{\operatorname{Cov}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}; \mathbf{x}_{k_1}\mathbf{x}_{l_1} - \overline{\mathbf{x}_k}\mathbf{x}_{l_1} - \mathbf{x}_{k_1}\overline{\mathbf{x}_l}\right]\Big\} \\
&= \frac{1}{4}\Big\{\mathbb{E}_{u_1}\left[\mathbf{x}_{i_1}^2\mathbf{x}_{k_1}\mathbf{x}_{l_1} - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}\mathbf{x}_{k_1}\mathbf{x}_{l_1} - \mathbf{x}_{i_1}^2\overline{\mathbf{x}_k}\mathbf{x}_{l_1}\right. \\
&\qquad \left. + 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}\ \overline{\mathbf{x}_k}\mathbf{x}_{l_1} - \mathbf{x}_{i_1}^2\mathbf{x}_{k_1}\overline{\mathbf{x}_l} + 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}\mathbf{x}_{k_1}\overline{\mathbf{x}_l}\right] \\
&\quad - \mathbb{E}_{u_1}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_i}\right]\mathbb{E}_{u_1}\left[\mathbf{x}_{k_1}\mathbf{x}_{l_1} - \overline{\mathbf{x}_k}\mathbf{x}_{l_1} - \mathbf{x}_{k_1}\overline{\mathbf{x}_l}\right]\Big\} \\
&= \frac{1}{4}\Big\{\overline{\mathbf{x}_i^2\mathbf{x}_k\mathbf{x}_l} - 2\ \overline{\mathbf{x}_i\mathbf{x}_k\mathbf{x}_l}\ \overline{\mathbf{x}_i} - \overline{\mathbf{x}_i^2\mathbf{x}_l}\ \overline{\mathbf{x}_k} \\
&\qquad + 2\ \overline{\mathbf{x}_i\mathbf{x}_l}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_k} - \overline{\mathbf{x}_i^2\mathbf{x}_{k_1}}\ \overline{\mathbf{x}_l} + 2\ \overline{\mathbf{x}_i\mathbf{x}_k}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l} \\
&\quad - \left(\overline{\mathbf{x}_i^2} - 2\ \overline{\mathbf{x}_i}^2\right)\left(\overline{\mathbf{x}_k\mathbf{x}_l} - 2\ \overline{\mathbf{x}_k}\ \overline{\mathbf{x}_l}\right)\Big\}.
\end{aligned} \tag{5.5.22}$$

▶ **Case 4**: $i = k$; $j \neq i, k, l$; $k \neq l$
The corresponding U-statistic kernels for this case are

$$h(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2}\left(\mathbf{x}i_1 - \mathbf{x}i_2\right)\left(\mathbf{x}j_1 - \mathbf{x}j_2\right), \text{ and} \tag{5.5.23}$$

$$g(\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{2}\left(\mathbf{x}i_1 - \mathbf{x}i_2\right)\left(\mathbf{x}l_1 - \mathbf{x}l_2\right). \tag{5.5.24}$$

And, ζ_1 is equal to

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(\mathbf{x}_{i_1}-\overline{\mathbf{x}_i}\right)\left(\mathbf{x}_{j_1}-\overline{\mathbf{x}_j}\right);\frac{1}{2}\left(\mathbf{x}_{i_1}-\overline{\mathbf{x}_i}\right)\left(\mathbf{x}_{l_1}-\overline{\mathbf{x}_l}\right)\right] \\
&= \frac{1}{4}\Big\{\mathrm{Cov}\left[\mathbf{x}_{i_1}\mathbf{x}_{j_1}-\overline{\mathbf{x}_i}\mathbf{x}_{j_1}-\mathbf{x}_{i_1}\overline{\mathbf{x}_j};\mathbf{x}_{i_1}\mathbf{x}_{l_1}-\overline{\mathbf{x}_i}\mathbf{x}_{l_1}-\mathbf{x}_{i_1}\overline{\mathbf{x}_l}\right]\Big\} \\
&= \frac{1}{4}\Big\{\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}^2\mathbf{x}_{j_1}\mathbf{x}_{l_1}-\overline{\mathbf{x}_i}\mathbf{x}_{j_1}\mathbf{x}_{i_1}\mathbf{x}_{l_1}-\mathbf{x}_{i_1}^2\overline{\mathbf{x}_j}\mathbf{x}_{l_1}\right. \\
&\qquad -\mathbf{x}_{i_1}\mathbf{x}_{j_1}\overline{\mathbf{x}_i}\mathbf{x}_{l_1}+\overline{\mathbf{x}_i}^2\,\mathbf{x}_{j_1}\mathbf{x}_{l_1}+\mathbf{x}_{i_1}\overline{\mathbf{x}_j}\,\overline{\mathbf{x}_i}\mathbf{x}_{l_1} \\
&\qquad \left.-\mathbf{x}_{i_1}^2\mathbf{x}_{j_1}\overline{\mathbf{x}_l}+\overline{\mathbf{x}_i}\mathbf{x}_{j_1}\mathbf{x}_{i_1}\overline{\mathbf{x}_l}+\mathbf{x}_{i_1}^2\overline{\mathbf{x}_j}\overline{\mathbf{x}_l}\right] \\
&\quad -\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}\mathbf{x}_{j_1}-\overline{\mathbf{x}_i}\mathbf{x}_{j_1}-\mathbf{x}_{i_1}\overline{\mathbf{x}_j}\right]\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}\mathbf{x}_{l_1}-\overline{\mathbf{x}_i}\mathbf{x}_{l_1}-\mathbf{x}_{i_1}\overline{\mathbf{x}_l}\right]\Big\} \\
&= \frac{1}{4}\Big\{\overline{\mathbf{x}_{i_1}^2\mathbf{x}_{j_1}\mathbf{x}_{l_1}}-\overline{\mathbf{x}_i}\;\overline{\mathbf{x}_{j_1}\mathbf{x}_{i_1}\mathbf{x}_{l_1}}-\overline{\mathbf{x}_{i_1}^2\mathbf{x}_{l_1}}\;\overline{\mathbf{x}_j} \\
&\qquad -\overline{\mathbf{x}_{i_1}\mathbf{x}_{j_1}\mathbf{x}_{l_1}}\;\overline{\mathbf{x}_i}+\overline{\mathbf{x}_i}^2\;\overline{\mathbf{x}_{j_1}\mathbf{x}_{l_1}}+\overline{\mathbf{x}_{i_1}\mathbf{x}_{l_1}}\;\overline{\mathbf{x}_j}\;\overline{\mathbf{x}_i} \\
&\qquad -\overline{\mathbf{x}_{i_1}^2\mathbf{x}_{j_1}}\;\overline{\mathbf{x}_l}+\overline{\mathbf{x}_i}\;\overline{\mathbf{x}_{j_1}\mathbf{x}_{i_1}}\;\overline{\mathbf{x}_l}+\overline{\mathbf{x}_{i_1}^2}\;\overline{\mathbf{x}_j}\;\overline{\mathbf{x}_l}] \\
&\quad -\left(\overline{\mathbf{x}_i\mathbf{x}_j}-2\;\overline{\mathbf{x}_i}\;\overline{\mathbf{x}_j}\right)\left(\overline{\mathbf{x}_i\mathbf{x}_l}-2\;\overline{\mathbf{x}_i}\;\overline{\mathbf{x}_l}\right)\Big\}.
\end{aligned} \tag{5.5.25}$$

▶ **Case 5**: $i=k;\ i\neq j;\ j=l$
The corresponding U-statistic kernels for this case are

$$h(\mathbf{u}_1,\mathbf{u}_2)=\frac{1}{2}\left(\mathbf{x}_{i_1}-\mathbf{x}_{i_2}\right)\left(\mathbf{x}_{j_1}-\mathbf{x}_{j_2}\right),\text{ and} \tag{5.5.26}$$

$$g(\mathbf{v}_1,\mathbf{v}_2)=h(\mathbf{u}_1,\mathbf{u}_2). \tag{5.5.27}$$

And, ζ_1 is equal to

$$\begin{aligned}
\zeta_1 &= \mathrm{Var}\left[\frac{1}{2}\left(\mathbf{x}_{i_1}-\overline{\mathbf{x}_i}\right)\left(\mathbf{x}_{j_1}-\overline{\mathbf{x}_j}\right)\right] \\
&= \frac{1}{4}\Big\{\mathrm{Var}\left[\mathbf{x}_{i_1}\mathbf{x}_{j_1}-\overline{\mathbf{x}_i}\mathbf{x}_{j_1}-\mathbf{x}_{i_1}\overline{\mathbf{x}_j}\right]\Big\} \\
&= \frac{1}{4}\Big\{\mathrm{E}_{\mathbf{x}_1}\left[\left(\mathbf{x}_{i_1}\mathbf{x}_{j_1}-\overline{\mathbf{x}_i}\mathbf{x}_{j_1}-\mathbf{x}_{i_1}\overline{\mathbf{x}_j}\right)^2\right]-\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}\mathbf{x}_{j_1}-\overline{\mathbf{x}_i}\mathbf{x}_{j_1}-\mathbf{x}_{i_1}\overline{\mathbf{x}_j}\right]^2\Big\} \\
&= \frac{1}{4}\Big\{\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}^2\mathbf{x}_{j_1}^2-2\mathbf{x}_{i_1}\mathbf{x}_{j_1}^2\overline{\mathbf{x}_i}+\overline{\mathbf{x}_i}^2\mathbf{x}_{j_1}^2-2\mathbf{x}_{i_1}^2\mathbf{x}_{j_1}\overline{\mathbf{x}_j}+2\overline{\mathbf{x}_i}\mathbf{x}_{j_1}\mathbf{x}_{i_1}\overline{\mathbf{x}_j}+\mathbf{x}_{i_1}^2\overline{\mathbf{x}_j}^2\right] \\
&\qquad -\left(\overline{\mathbf{x}_iX_j}-2(\overline{\mathbf{x}_i}\;\overline{\mathbf{x}_j})\right)^2\Big\} \\
&= \frac{1}{4}\Big\{\overline{\mathbf{x}_i^2\mathbf{x}_j^2}-2\overline{\mathbf{x}_i\mathbf{x}_j^2}\;\overline{\mathbf{x}_i}+\overline{\mathbf{x}_i}^2\;\overline{\mathbf{x}_j^2}-2\overline{\mathbf{x}_i^2\mathbf{x}_j}\;\overline{\mathbf{x}_j}+2\overline{\mathbf{x}_i}\;\overline{\mathbf{x}_j}\;\overline{\mathbf{x}_j\mathbf{x}_i}+\overline{\mathbf{x}_i^2}\;\overline{\mathbf{x}_j}^2 \\
&\qquad -\left(\overline{\mathbf{x}_iX_j}-2(\overline{\mathbf{x}_i}\;\overline{\mathbf{x}_j})\right)^2\Big\}.
\end{aligned} \tag{5.5.28}$$

▶ **Case 6**: $i=j=k;\ i\neq l$

The corresponding U-statistic kernels for this case are

$$h(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2}\left(\mathbf{x}_{i_1} - \mathbf{x}i_2\right)^2, \text{ and} \tag{5.5.29}$$

$$g(\mathbf{v}_1, \mathbf{v}_2) = \frac{1}{2}\left(\mathbf{x}_{i_1} - \mathbf{x}i_2\right)\left(\mathbf{x}l_1 - \mathbf{x}l_2\right). \tag{5.5.30}$$

And, ζ_1 is equal to

$$\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left[\frac{1}{2}\left(\mathbf{x}_{i_1} - \overline{\mathbf{x}_i}\right)^2; \frac{1}{2}\left(\mathbf{x}_{i_1} - \overline{\mathbf{x}_i}\right)\left(\mathbf{x}l_1 - \overline{\mathbf{x}_l}\right)\right] \\
&= \frac{1}{4}\Big\{\mathrm{Cov}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}; \mathbf{x}_{i_1}\mathbf{x}_{l_1} - \overline{\mathbf{x}_{i_1}}\mathbf{x}_{l_1} - \mathbf{x}_{i_1}\overline{\mathbf{x}_{l_1}}\right]\Big\} \\
&= \frac{1}{4}\Big\{\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}^2\mathbf{x}_{i_1}\mathbf{x}_{l_1} - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}\mathbf{x}_{i_1}\mathbf{x}_{l_1} - \mathbf{x}_{i_1}^2\overline{\mathbf{x}_{i_1}}\mathbf{x}_{l_1}\right. \\
&\qquad \left. + 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}\ \overline{\mathbf{x}_{i_1}}\mathbf{x}_{l_1} - \mathbf{x}_{i_1}^2\mathbf{x}_{i_1}\overline{\mathbf{x}_{l_1}} + 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}\mathbf{x}_{i_1}\overline{\mathbf{x}_{l_1}}\right] \\
&\quad - \mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}\right]\mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}\mathbf{x}_{l_1} - \overline{\mathbf{x}_{i_1}}\mathbf{x}_{l_1} - \mathbf{x}_{i_1}\overline{\mathbf{x}_{l_1}}\right]\Big\} \\
&= \frac{1}{4}\Big\{\overline{\mathbf{x}_i^3\mathbf{x}_l} - 3\ \overline{\mathbf{x}_i^2\mathbf{x}_l}\ \overline{\mathbf{x}_i} + 2\ \overline{\mathbf{x}_i\mathbf{x}_l}\ \overline{\mathbf{x}_i}^2 - \overline{\mathbf{x}_i^3}\ \overline{\mathbf{x}_l} + 2\ \overline{\mathbf{x}_i^2}\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l} \\
&\quad - \left(\overline{\mathbf{x}_i^2} - 2\ \overline{\mathbf{x}_i}^2\right)\left(\overline{\mathbf{x}_i\mathbf{x}_l} - 2\ \overline{\mathbf{x}_i}\ \overline{\mathbf{x}_l}\right)\Big\}.
\end{aligned} \tag{5.5.31}$$

▶ **Case 7**: $i = j, k, l$
The corresponding U-statistic kernels for this case are

$$h(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2}\left(\mathbf{x}i_1 - \mathbf{x}i_2\right)^2, \text{ and} \tag{5.5.32}$$

$$g(\mathbf{v}_1, \mathbf{v}_2) = h(\mathbf{u}_1, \mathbf{u}_2). \tag{5.5.33}$$

And, ζ_1 is equal to

$$\begin{aligned}
\zeta_1 &= \mathrm{Var}\left[\frac{1}{2}\left(\mathbf{x}i_1 - \overline{\mathbf{x}_i}\right)^2\right] \\
&= \frac{1}{4}\mathrm{Var}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}\right] \\
&= \frac{1}{4}\left\{\mathrm{E}_{\mathbf{x}_1}\left[\left(\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}\right)^2\right] - \mathrm{E}_{\mathbf{x}_1}\left[\mathbf{x}_{i_1}^2 - 2\mathbf{x}_{i_1}\overline{\mathbf{x}_{i_1}}\right]^2\right\} \\
&= \frac{1}{4}\left\{\overline{\mathbf{x}_i^4} - 4\overline{\mathbf{x}_i^3}\ \overline{\mathbf{x}_i} + 4\overline{\mathbf{x}_i^2}\ \overline{\mathbf{x}_i}^2 - \left(\overline{\mathbf{x}_i^2} - 2\overline{\mathbf{x}_i}^2\right)^2\right\}.
\end{aligned} \tag{5.5.34}$$

5.5.3 Derivation in $\mathcal{O}(n)$ time for all terms

In section 5.5.2, all terms are in the form of $\mathbb{E}[\mathbf{x}]$,$\mathbb{E}[\mathbf{xy}]$,$\mathbb{E}[\mathbf{xyz}]$ and $\mathbb{E}[\mathbf{xyuy}]$ and can be computed in $\mathcal{O}(n)$ as following

$$\mathbb{E}[\mathbf{x}] = \frac{1}{n}\sum_{q=1}^{n} \mathbf{x}_q. \tag{5.5.35}$$

$$\mathbb{E}[\mathbf{xy}] = \frac{1}{n}\sum_{q=1}^{n} \mathbf{x}_q \odot \mathbf{y}_q \tag{5.5.36}$$

$$E[\mathbf{xyz}] = \frac{1}{n}\sum_{q=1}^{n} \mathbf{x}_q. \odot \mathbf{y}_q \odot \mathbf{z}_q. \tag{5.5.37}$$

$$\mathbb{E}[\mathbf{xyuy}] = \frac{1}{n}\sum_{q=1}^{n} \mathbf{x}_q \odot \mathbf{y}_q \odot \mathbf{u}_q \odot \mathbf{v}_q. \tag{5.5.38}$$

Chapter 6

Conclusion

The problem of hypothesis tests for similarity and dependency are of fundamental importance in statistics. However, there were several important open questions in the literature prior to the work done in this thesis. First, while significant progress had been achieved in modeling and using classical hypothesis tests for the task of identifying similarity between variables, or finding the dependencies among them, existing methods have mostly focused on pairwise relationships. Second, conditional independence tests are especially important and are challenging for the task of learning probabilistic graphical model structures from data. Such tests enable reasoning about interconnected nodes in networks, for example. However, while there exist many methods in the literature using strong assumptions of data being generated by discrete or Gaussian multivariate distributions, other distributions have posed new challenges in statistical modeling for real-world data.

In this thesis, we have taken as starting point these shortcomings. This chapter summarizes the contributions made in this thesis and their relationship to the three research questions posed at the beginning of the thesis.

6.1 Summary of contributions

In this thesis, we addressed the problem of novel non-parametric hypothesis tests for similarity and dependence and we explored means to achieve them. As a result, the contributions resulting from this work have developed novel statistical hypothesis tests with optimal computational complexity. A summary is given in Table 1.1 in Chapter 1.

Being based on the powerful framework of U-statistics, all resulting tests have favorable convergence properties and are consistent, low-variance, and unbiased. In the following paragraphs, we described in more detail the contributions of each component.

The first work presented in this thesis addresses the task of determining whether a target distribution is closer to one of two candidate distributions. In Chapter 3, we propose a novel non-parametric statistical hypothesis test for relative similarity based on the MMD. Based on the theory of U-statistics, we use as our test statistic the difference of MMDs between the reference dataset and each model dataset, and derive a powerful, low-variance test based on their joint asymptotic distribution.

The test is consistent, and the computation time is quadratic. Our proposed test statistic

is theoretically justified for the task of comparing samples from arbitrary distributions as it can be shown to converge to a quantity which compares all moments of the two pairs of distributions.

For the second work presented in this thesis, we describe in Chapter 4 a novel statistical test which determines whether two target variables have a significant difference in their dependence on a third, source variable. The dependence between each of the target variables and the source is computed using the Hilbert-Schmidt Independence Criterion (HSIC).

Finally, in Chapter 5, we have constructed a conservative threshold for a hypothesis test on the absolute value of the precision matrix being significantly non-zero for a wider range of distributions than has been previously considered in the literature. Previous works have primarily focused on Gaussian or discrete distributions. We have developed a threshold based on a U-statistic empirical estimator of the covariance matrix, which we use to probabilistically bound the distortion of the true covariance matrix. Using this fixed bound in conjunction with Weyl's theorem, we are able to bound the distortion of the precision matrix.

6.2 Revisiting the Research Questions

At the beginning of this thesis, we set the objective of investigating the potential of new statistical hypothesis testing for dependence and similarity that has not been explored. To this end, we formulated three research questions which were presented in introduction. We revisit and address each of them in turn.

1. Is the probability measure $\mathbb{P}_{\mathbf{x}}$ significantly closer to $\mathbb{P}_{\mathbf{y}}$ or to $\mathbb{P}_{\mathbf{z}}$?

The results obtained from Chapter 3 give a novel non-parametric statistical hypothesis test for relative similarity based on the MMD. Our proposed test statistic is theoretically justified for the task of comparing samples from arbitrary distributions as it can be shown to converge to a quantity which compares all moments of the two pairs of distributions. Experimental results on model selection for deep generative networks show that this hypothesis can be a useful approach to comparing such models. These observations were further confirmed by Sutherland et al. [2016], where the method optimizes the representation and distinguishes samples from two probability distributions by maximizing the estimated power of a statistical test based on the MMD statistics. These findings essentially show that answering such a question is of importance in the machine learning community.

2. Is the dependency between $\mathbf{x}$ and $\mathbf{y}$ significantly stronger than the dependency between $\mathbf{x}$ and $\mathbf{z}$?

This research question was explored with one main goal, which is to find significant relative dependency of different outputs. When there may be multiple dependencies, which dependence is the strongest? Dependence is measured via the HSIC resulting in a pair of empirical

dependence measures (source-target 1, source-target 2). We formulate the hypothesis test to determine whether the first dependence measure is significantly larger than the second. There do not exist competing tests and detecting the strength of dependencies between three sets of variable is of great importance in a variety of problems. For instance, to develop a treatment for cancer, we must identify the mechanisms responsible for the disease, so the goal is to determine the cause of this process. Doctors are frequently interested in genomic analysis (to locate genes, determine the function of a protein, or interpret large amounts of information generated) and chromosome imbalance (when there is extra or missing chromosomal material). In order to save time and money, the question posed by researchers at Hôpital Necker-Enfants Malades [Puget et al., 2012] was whether the dependency between the location of glioma is most associated with the expression of genes or chromosomal anomaly. This application demonstrates the real world interest in answering this research question.

3. Can we develop a statistically and computationally efficient estimator of the topology of graphical models for non-Gaussian distributions?

Aiming at answering this research question, in Chapter 5 we proposed a new framework for hypothesis testing of whether an entry of the precision matrix is non-zero based on a data sample from the joint distribution. The proposed test is sound and does not depend on the data being Gaussian distributed or other parametric assumptions and does not require sparsity. Furthermore, results from synthetic and real-world datasets show that for millions of data points and general distributions, the hypothesis test is effective at finding conditional independencies between all variables in the model. These results give a positive answer to this research question by showing that considering a very general class of distributions, and for very large sample sizes, the topology of a graphical model can be discovered.

6.3 Directions for Future Research

In the last section of this thesis, we draw some directions for future research based on the contributions and limitations of our work. We propose several research paths that can be followed.

One direction for future work lies in the extension of the relative test of dependency in Chapter 4 to random process variables. In many applications, an observation is dependent on its past values. For instance, in neuroscience, multiple stimuli may be present (e.g. visual and audio), and it is of interest to determine which of the two has a stronger influence on brain activity. For Alzeihmer's disease, we can test over long term, if drug and non-drug treatments help with both cognitive and behavioral symptoms. The question is then to determine whether a source random process variable is more strongly dependent on one target random process variable or another.

Futhermore, recent work by Chwialkowski et al. [2015] propose a class of nonparametric two-sample tests with a cost linear in the sample size. We can extend our two main Chapters 3

and 4 to have a linear time complexity.

Another possible direction is related to the Chapter 5. The idea is to select a subset of variables of interest and study whether an entry if the precision matrix is non-zero. More formally, given a sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$, we can select a subset of from $\mathbf{X}$ and then form a low-rank approximation of the covariance of $\mathbf{X}$. The most straightforward technique is to use the Nÿstrom approximation, but further research on conditions for the identifiability of partial correlations under such approximations is required.

List of publications

Papers at international scientific conferences

- W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *The 4th International Conference on Learning Representations*, 2016a.
- W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. A low variance consistent test of relative dependency. In F. Bach and D. Blei, editors, *Proceedings of The 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 20–29, 2015a.

Under submission

- W. Bounliphone, E. Belilovsky, A. Tenenhaus, I. Antonoglou, A. Gretton, and M. B. Blaschko. Fast Non-Parametric Tests of Relative Dependency and Similarity. 2016c. arXiv:1611.05740 – under submission.
- W. Bounliphone and M. B. Blaschko. Linear time non-Gaussian precision matrix estimation. 2016. arXiv:1604.01733 – under submission.

Workshops

- W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A kernel test of relative similarity. In *Women in Machine Learning Workshop*, Barcelona, Spain, Dec. 2016b.
- W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. Kernel non-parametric tests of relative dependency. International Conference of the ERCIM WG on Computational and Methodological Statistics - CMStatistics 2015, Dec. 2015b.
- W. Bounliphone, A. Gretton, and M. Blaschko. Kernel non-parametric tests of relative dependency. In *NIPS Workshop on Modern Nonparametrics 3: Automating the Learning Pipeline*, Montreal, Canada, Dec. 2014a.
- W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. A kernel test of relative dependency. In *Women in Machine Learning Workshop*, Montreal, Canada, Dec. 2014b.

Bibliography

M. A. Arcones and E. Gine. Limit theorems for U-processes. *The Annals of Probability*, pages 1494–1542, 1993.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9, 2008.

Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

Y. Bengio, E. Thibodeau-Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2011.

R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, 2012.

W. Bounliphone and M. B. Blaschko. Linear time non-Gaussian precision matrix estimation. 2016. arXiv:1604.01733 – under submission.

W. Bounliphone, A. Gretton, and M. Blaschko. Kernel non-parametric tests of relative dependency. In *NIPS Workshop on Modern Nonparametrics 3: Automating the Learning Pipeline*, Montreal, Canada, Dec. 2014a.

W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. A kernel test of relative dependency. In *Women in Machine Learning Workshop*, Montreal, Canada, Dec. 2014b.

W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. A low variance consistent test of relative dependency. In F. Bach and D. Blei, editors, *Proceedings of The 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 20–29, 2015a.

W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. Kernel non-parametric tests of relative dependency. International Conference of the ERCIM WG on Computational and Methodological Statistics - CMStatistics 2015, Dec. 2015b.

W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *The 4th International Conference on Learning Representations*, 2016a.

W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A kernel test of relative similarity. In *Women in Machine Learning Workshop*, Barcelona, Spain, Dec. 2016b.

W. Bounliphone, E. Belilovsky, A. Tenenhaus, I. Antonoglou, A. Gretton, and M. B. Blaschko. Fast Non-Parametric Tests of Relative Dependency and Similarity. 2016c. arXiv:1611.05740 – under submission.

J. Bring. A geometric approach to compare variables in a regression model. *The American Statistician*, 50(1):57–62, 1996.

H. Callaert and P. Janssen. The Berry-Esseen theorem for U-statistics. *The Annals of Statistics*, pages 417–421, 1978.

G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, 2nd edition, 2002.

Centers for Disease Control and Prevention. Online tuberculosis information system data (OTIS), 2014. http://wonder.cdc.gov/.

J. Chang, Q.-M. Shao, and W.-X. Zhou. Cramér-type moderate deviations for Studentized two-sample U-statistics with applications. *The Annals of Statistics*, 44(5):1931–1956, 2016.

L. H. Chen and Q.-M. Shao. Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli*, pages 581–599, 2007.

Chromosome Disorder Outreach. Introduction to chromosomes, 2016. http://chromodisorder.org/.

K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. In *Advances in Neural Information Processing Systems*, 2009.

R. B. Darlington. Multiple regression in psychological research and practice. *Psychological Bulletin*, 69(3):161, 1968.

J. Dauxois and G. M. Nkiet. Nonlinear canonical analysis and independence tests. *The Annals of Statistics*, 26(4):1254–1278, 1998.

A. P. Dawid. Conditional independence in statistical theory. *Royal Statistical Society*, 1979.

S. R. de Morais and A. Aussem. An efficient and scalable algorithm for local Bayesian network structure discovery. In J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases, Part III*, pages 164–179. Springer, 2010.

A. P. Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

J. Dieudonné. *Foundations of Modern Analysis*. Academic Press, Elsevier, 1960.

G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In *Conference on Uncertainty in Artificial Intelligence*, pages 132–141, 2014.

M. Drton and M. D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2nd edition, 2002.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

Y. Escoufier. Le traitement des variables vectorielles. *Biometrics*, pages 751–760, 1973.

R. A. Fisher. The distribution of the partial correlation coefficient. *Metron*, 3:329–332, 1924.

M. Fréchet. Sur les ensembles de fonctions et les opérations linéaires. *CR Acad. Sci. Paris*, 144:1414–1416, 1907.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 2008.

J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, pages 489–496, 2007.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

M. Gasse, A. Aussem, and H. Elghazel. An experimental comparison of hybrid algorithms for Bayesian network structure learning. In P. A. Flach, T. De Bie, and N. Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases, Part I*, pages 58–73. Springer, 2012.

R. J. Gilbertson and D. H. Gutmann. Tumorigenesis in the brain: Location, location, location. *Cancer Research*, 67(12):5579–5582, 2007.

E. Gómez. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory, Methods*, 27(3), 1998.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

R. D. Gray and Q. D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, 2003.

A. Gretton and L. Gyorfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pages 63–77, 2005a.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005b.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2006.

A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Neural Information Processing Systems*, pages 585–592, 2008.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012a.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1205–1213. 2012b.

M. G. G'Sell, J. Taylor, and R. Tibshirani. Adaptive testing for the graphical lasso. *arXiv:1307.4765*, 2013.

S. R. Gunn and J. S. Kandola. Structural modelling with sparse kernels. *Machine Learning*, 48(1):137–163, 2002.

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971.

R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

G. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325, 1948.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

J. Jankova and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *The Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. CRC Press, 1997.

M. G. Kendall. *The Advanced Theory of Statistics*. C. Griffin, 1946.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

J. B. Kinney and G. S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 2014.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

A. Krishnamurthy, K. Kandasamy, B. Póczos, and L. A. Wasserman. On estimating L_2^2 divergence. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

H. Larochelle and I. Murray. The neural autoregressive distribution estimator. *Journal of Machine Learning Research*, 15:29–37, 2011.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

A. J. Lee. *U-statistics: Theory and practice*. CRC Press, 1990.

E. L. Lehmann. *Elements of Large-sample Theory*. Springer, 1999.

H. Li and J. Gui. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2006.

S. Li, Y. Xie, H. Dai, and L. Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems*, pages 3348–3356, 2015a.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015b.

W. Liu et al. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, 2015.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413, 2014.

P.-L. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049, 2013.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair auto encoder. In *International Conference on Learning Representations*, 2016.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(1): 148–188, 1989.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 06 2006.

P. Narasimhan, J. Wood, C. R. MacIntyre, and D. Mathai. Risk factors for tuberculosis. *Pulmonary Medicine*, 2013.

National Centers for Environmental Information. Digital dataset of detailed hourly observational climate data for thousands of locations worldwide, 2016. https://www.ncdc.noaa.gov/.

National Human Genome Research Institute. Clinical sequencing centers aim for guidelines, 2016. https://www.genome.gov/.

Nationwide Children's Hospital. Brain tumors in children, 2016. http://www.nationwidechildrens.org/child-brain-tumor.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2004.

T. Palm, D. Figarella-Branger, F. Chapon, C. Lacroix, F. Gray, F. Scaravilli, D. W. Ellison, I. Salmon, M. Vikkula, and C. Godfraind. Expression profiling of ependymomas unravels localization and tumor grade-specific tumorigenesis. *Cancer*, 115(17):3955–3968, 2009.

C. Peters, M. Braschler, and P. Clough. *Multilingual Information Retrieval: From Research to Practice*. Springer, 2012.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, Jan. 2014.

S. Puget, C. Philippe, D. Bax, B. Job, P. Varlet, M. P. Junier, F. Andreiuolo, D. Carvalho, R. Reis, and L. Guerrini-Rousseau. Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. *PloS one*, 7(2):e30313, 2012.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *The Electronic Journal of Statistics*, 5:935–980, 2011.

Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. Detecting novel associations in large datasets. *Science*, 334(6062), 2011.

F. Riesz. Sur une espèce de géométrie analytique des systèmes de fonctions sommables. *CR Acad. Sci. Paris*, 144:1409–1411, 1907.

P. R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

A. Roverato and J. Whittaker. Standard errors for the parameters of graphical Gaussian models. *Statistics and Computing*, 6(3):297–302, 1996.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. In J. A. Anderson and E. Rosenfeld, editors, *Neurocomputing: Foundations of Research*, pages 696–699. MIT Press, 1988.

R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

I. Schaller-Schwaner. Does a picture say more than 7000 words? windows of opportunity to learn languages - an attempt at a creative reflective poster. *Language Learning in Higher Education*, 5(1):1–23, 2015.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

K. Sechidis and G. Brown. Markov blanket discovery in positive-unlabelled and semi-supervised data. In A. Appice, P. P. Rodrigues, V. Santos Costa, C. Soares, J. Gama, and A. Jorge, editors, *Machine Learning and Knowledge Discovery in Databases, Part I*, pages 351–366. Springer, 2015.

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(5): 2263–2702, 2013.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.

S. Suchindran, E. S. Brouwer, and A. Van Rie. Is HIV infection a risk factor for multi-drug resistant tuberculosis? a systematic review. *PloS one*, 4(5):e5561, 2009.

D. J. Sutherland. *Scalable, Flexible, and Active Learning on Distributions*. PhD thesis, Queensland University of Technology, 2016.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative model and models criticism via optimized maximum mean discrepancy. 2016. arXiv:1611.04488.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

J. Trommershauser, K. Kording, and M. S. Landy. *Sensory Cue Integration*. Oxford University Press, 2011.

B. von Bahr. On the convergence of moments in the central limit theorem. *The Annals of Mathematical Statistics*, 36(3):808–818, 06 1965.

L. Wasserman, M. Kolar, and A. Rinaldo. Berry-Esseen bounds for estimating undirected graphs. *The Electronic Journal of Statistics*, 8(1), 2014.

H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen. *Mathematische Annalen*, 71(4):441–479, 1912.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, 2009.

N. Xia, Y. Qin, and Z. Bai. Convergence rates of eigenvector empirical spectral distribution of large dimensional sample covariance matrix. *The Annals of Statistics*, 41(5):2572–2607, 2013.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, pages 755–763, 2013.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

Titre : Tests d'hypothèses statistiquement et algorithmiquement efficaces de similarité et de dépendance

Mots clefs : U-statistiques, tests d'hypothèses statistiques, dépendance, similarité, méthodes à noyau.

Résumé : Cette thèse présente de nouveaux tests d'hypothèses statistiques efficaces pour la relative similarité et dépendance, et l'estimation de la matrice de précision. La principale méthodologie adoptée dans cette thèse est la classe des estimateurs U-statistiques.

Le premier test statistique porte sur les tests de relative similarité appliqués au problème de la sélection de modèles. Les modèles génératifs probabilistes fournissent un cadre puissant pour représenter les données. La sélection de modèles dans ce contexte génératif peut être difficile. Pour résoudre ce problème, nous proposons un nouveau test d'hypothèse non paramétrique de relative similarité et testons si un premier modèle candidat génère un échantillon de données significativement plus proche d'un ensemble de validation de référence.

La deuxième test d'hypothèse statistique non paramétrique est pour la relative dépendance. En présence de dépendances multiples, les méthodes existantes ne répondent qu'indirectement à la question de la relative dépendance. Or, savoir si une dépendance est plus forte qu'une autre est important pour la prise de décision. Nous présentons un test statistique qui détermine si une variable dépend beaucoup plus d'une première variable cible ou d'une seconde variable.

Enfin, une nouvelle méthode de découverte de structure dans un modèle graphique est proposée. En partant du fait que les zéros d'une matrice de précision représentent les indépendances conditionnelles, nous développons un nouveau test statistique qui estime une borne pour une entrée de la matrice de précision. Les méthodes existantes de découverte de structure font généralement des hypothèses restrictives de distributions gaussiennes ou parcimonieuses qui ne correspondent pas forcément à l'étude de données réelles. Nous introduisons ici un nouveau test utilisant les propriétés des U-statistics appliqués à la matrice de covariance, et en déduisons une borne sur la matrice de précision.

Title : Statistically and computationally efficient hypothesis tests for similarity and dependency

Keywords : U-statistics, hypothesis testing, dependency, similarity, kernel methods.

Abstract : The dissertation presents novel statistically and computationally efficient hypothesis tests for relative similarity and dependency, and precision matrix estimation. The key methodology adopted in this thesis is the class of U-statistic estimators. The class of U-statistics results in a minimum-variance unbiased estimation of a parameter.

The first part of the thesis focuses on relative similarity tests applied to the problem of model selection. Probabilistic generative models provide a powerful framework for representing data. Model selection in this generative setting can be challenging. To address this issue, we provide a novel non-parametric hypothesis test of relative similarity and test whether a first candidate model generates a data sample significantly closer to a reference validation set.

Subsequently, the second part of the thesis focuses on developing a novel non-parametric statistical hypothesis test for relative dependency. Tests of dependence are important tools in statistical analysis, and several canonical tests for the existence of dependence have been developed in the literature. However, the question of whether there exist dependencies is secondary. The determination of whether one dependence is stronger than another is frequently necessary for decision making. We present a statistical test which determine whether one variables is significantly more dependent on a first target variable or a second.

Finally, a novel method for structure discovery in a graphical model is proposed. Making use of a result that zeros of a precision matrix can encode conditional independencies, we develop a test that estimates and bounds an entry of the precision matrix. Methods for structure discovery in the literature typically make restrictive distributional (e.g. Gaussian) or sparsity assumptions that may not apply to a data sample of interest. Consequently, we derive a new test that makes use of results for U-statistics and applies them to the covariance matrix, which then implies a bound on the precision matrix.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France