

Introduction to U-Statistics with Applications in Machine Learning

Fabian Baier

Thesis for the attainment of the academic degree

Bachelor of Science

at the TUM School of Computation, Information and Technology of the Technical University of Munich

Supervisor:

Prof. Dr. Aleksey Min

Advisors:

Prof. Dr. Aleksey Min

Submitted:

Munich, 11. December 2025

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

A handwritten signature in black ink, appearing to read "Fabian Baier".

Munich, 11. December 2025

Fabian Baier

Zusammenfassung

In dieser Bachelorarbeit wird eine Einführung in die Theorie der U-Statistiken gegeben, sowie eine explizite Anwendung im Rahmen von Maximum Mean Discrepancy vorgestellt. U-Statistiken sind von besonderem Interesse, da sie unverzerrte Schätzer sind und unter allen unverzerrten Schätzern die minimale Varianz besitzen. Viele Statistiken lassen sich als U-Statistiken schreiben, weshalb das Untersuchen der allgemeinen Eigenschaften von U-Statistiken, wie zum Beispiel der zentrale Grenzwertsatz oder das Gesetz der großen Zahlen, von großer Wichtigkeit ist. Die Theorie der U-Statistik wird ohne Vorwissen aufgebaut, und unter anderem werden die Martingale Eigenschaft und die Reverse Martingale Eigenschaft bewiesen. Ebenso werden die Hoeffding Repräsentation, sowie auch die komplette Degeneriertheit der kanonischen Funktionen, hergeleitet. Anschließend wird ein explizites Beispiel von einer U-Statistik gegeben, die Maximum Mean Discrepancy (MMD), welche viele wichtige Anwendungen, unter anderem im Machine Learning, hat. MMDs messen die Distanz zwischen zwei Wahrscheinlichkeitsverteilungen und sind effizient zu berechnen für höhere Dimensionen. Zur Veranschaulichung werden Zeitreihen aus dem Hochfrequenzbereich von unterschiedlichen Aktien unter Verwendung von MMDs verglichen. Abschließend wird ein MMD-basierter statistischer Test vorgestellt und dessen Annahmen durch eine Monte Carlo Simulationsstudie überprüft. Dieser Test ist hilfreich für Modellselektion und Modellspezifikation und ist für hochdimensionale Deep Learning Modelle geeignet. Der Test wurde im Rahmen eines Forschungspraktikums am Lehrstuhl für Finanzmathematik der TUM implementiert.

Abstract

This bachelor's thesis is about giving an introduction to the U-statistics theory from the ground up, without assuming any prior knowledge. An important application of this theory is the maximum mean discrepancy (MMD) measure. U-statistics are of particular interest, due to their property that they are unbiased estimators and have the minimal variance under all unbiased estimators. In recent years, this theory has become increasingly important due to the rise of machine learning and especially generative models. It can also be used for model specification and model selection tests based on the MMD, an important subclass of a U-statistic. For the theoretical part of U-statistics, we will talk about their martingale and reverse martingale properties, the Hoeffding representation, as well as the complete degeneracy of canonical functions. The martingale properties and the complete degeneracy will be proved, and the Hoeffding representation will be derived in detail. We then introduce the notion of an MMD and its connection to U-statistics. MMDs measure the distance between probability distributions and are efficient to estimate, even for high dimensions. As an application of MMDs, we will discuss first a simple interpretation of MMD using time series of high-frequency data from different companies and second, examine an MMD-based statistical test for model specification and selection. For this test, the code was implemented during a research internship at the chair of mathematical finance of TUM.

Contents

| | | |
|---------------------|--|-----------|
| 1 | Introduction | 1 |
| 2 | Prerequisites | 3 |
| 2.1 | Basics of probability theory | 3 |
| 2.2 | Conditional expectation | 5 |
| 2.3 | Martingales | 7 |
| 3 | U-statistics | 11 |
| 3.1 | Motivation | 11 |
| 3.2 | Definition and examples | 11 |
| 3.3 | Canonical functions and degeneracy of U-statistics | 15 |
| 3.4 | The Hoeffding representation | 18 |
| 3.5 | The martingale structure of U-statistics | 19 |
| 3.6 | V-statistics | 22 |
| 4 | Maximum Mean Discrepancy (MMD) | 25 |
| 4.1 | The notion of MMD and its connection to U-Statistics | 25 |
| 4.2 | MMD estimators and important properties | 25 |
| 5 | Case study: A high-frequency time series | 29 |
| 5.1 | The multi-level microprice and its applications | 29 |
| 5.2 | Practical considerations | 31 |
| 5.3 | MMD comparison | 32 |
| 6 | Application: An MMD-based statistical test | 35 |
| 6.1 | Motivation | 35 |
| 6.2 | The BFM test statistic | 35 |
| 6.3 | Central limit theorem | 37 |
| 6.4 | Estimation of variance | 38 |
| 6.5 | Simulation Study | 39 |
| 6.5.1 | Monte Carlo Setup for Figure 1 | 39 |
| 6.5.2 | Monte Carlo Setup for Figure 2 | 40 |
| 6.5.3 | Simulation for Figure 3: Model Selection (Degenerate Case) | 41 |
| 6.5.4 | Simulation for Figure 4: Model Selection (Non-degenerate Case) | 42 |
| 7 | Conclusion | 45 |
| A | Appendix | 47 |
| A.1 | Supporting Data | 47 |
| A.2 | Code sources | 47 |
| Bibliography | | 53 |

1 Introduction

The purpose of this thesis is to provide a clear and self-contained introduction to some fundamental ideas around *U-statistics* and their connection to the *Maximum Mean Discrepancy (MMD)*, which has a wide range of applications.

In 1948, Hoeffding, in his article [Hoe48], introduced a new class of statistics, the U-statistics, where the U stands for unbiased. It opened a new branch of probability theory, where he proved the central limit theorem for U-statistics. U-statistics are a generalization of sums of independent random variables, thus proving results such as the law of large numbers, asymptotics, convergence properties, and the law of iterated logarithm were of interest and presented in a coherent framework by [KB13]. It was shown by [Hoe48] that, if the kernel of a U-statistic has a rank of one, in which case we call the U-statistic non-degenerate, it is asymptotically equivalent to the sum of independent random variables. In this case, the entire probability theory developed for sums of independent random variables can be applied to the non-degenerate U-statistics.

For the theory of U-statistics, we follow the book by [KB13]. We will discuss how the parametric functional arises naturally in statistics, and that the U-statistic is an unbiased estimator of it. All the theory developed is for symmetric kernels, i.e. they are symmetric in their arguments. This is not a limitation, as all non-symmetric kernels can be transformed into symmetric kernels, and after the transformation their expected values are the same. Additionally, we restrict the discussion to one-sample U-statistics and where the target space of the kernel is real-valued. The theory has been extended to multi-sample U-statistics, where the target space of the kernel can also be a Banach space (see e.g. [KB13]). In this thesis, we will focus on an intuitive understanding of the martingale structure of the U-statistics, the Hoeffding representation, and canonical functions.

The main advantage of the U-statistics is that it was shown by [Hoe48] that any statistic that can be written as a U-statistic has the minimum variance among all unbiased estimators. Thus, the study of U-statistics provides a useful framework for deriving important properties of estimators, e.g. [BFM25], showing the asymptotic normality of their test statistic.

The theory would not be useful if there were no important statistics, which can be written as U-statistics. Thus, as an inspiration, the sample mean, the sample variance and covariance, as well as some modern statistics, such as the maximum mean discrepancy (MMD) and the kernel Stein discrepancy (KSD), can be written as U-statistics. The KSD measure was introduced in [LLJ16] and has an important application in goodness-of-fit testing. For example, the first theoretically valid composite goodness-of-fit test by [BRB25], where the derivation of the test is heavily based on U-statistics theory.

In this thesis, we will focus on the MMD, a measure between probability distributions, which was introduced by [Smo+07]. The empirical MMD can be expressed as a U-statistic. We will compare different MMD estimators and derive the results we need for Chapter 6, where, as an application of the theory, we consider the MMD-based statistical test by [BFM25] and perform the Monte Carlo simulation study, as in their paper.

The MMD measure and its estimators have several important applications, such as change-point detection ([ACH19]), conditional independence testing ([Zha+11]), adaptive MCMC methods ([Sej+14]) and causal inference ([LP+15]). In recent years, due to the rise in popularity of machine learning, the MMD has also been used in generative machine learning ([Zho+20]) for an MMD-based Generative Adversarial Network (GAN). One of the main advantages of MMD is that it can be used for high dimensions, and there exists a linear version of it ([Gre+12]), making it computationally highly efficient.

The remainder of this thesis is organized as follows. Chapter 2 recalls measure-theoretic probability theory preliminaries, including conditional expectation and martingales. It contains examples that will be needed for a better understanding of later proofs and concepts of U-statistics.

Chapter 3 introduces U-statistics, focusing on Hoeffding's representation and canonical functions. The martingale and reverse martingale property of U-statistics will be proved. For the property of complete degeneracy, a proof idea will be given.

Chapter 4 presents the maximum mean discrepancy and shows its link to U-statistics. The most common estimators of MMD will be discussed and for the application in Chapter 6, some essential results, which are based on U-statistics theory, will be proved.

For an intuitive understanding of the concept of MMD, a small example using high-frequency financial data will be given in Chapter 5.

Chapter 6 concludes with an application of the learned theory. A statistical test for model selection and specification will be presented, which was introduced by [BFM25]. Additionally, an empirical validation of the test will be given using a Monte Carlo simulation study.

Lastly, in Chapter 7, we discuss the main results and give an outlook into possible further research areas for statistical tests, which are based on U-statistics as estimators.

Declaration on the usage of AI tools:

I hereby declare that I used Grammarly to improve my writing style and ChatGPT to generate the code for all figures and tables in this thesis, as well as to reformat the LaTeX code. The entire code in the repository <https://github.com/fabianbaiertum/bachelor-thesis> is generated with the help of ChatGPT, besides the multi-level microprice function. Note that the code generated by ChatGPT was reviewed by me and went through several iterations before reaching its final form. For the BFM test in Chapter 6, only the code to generate the figures is written with the help of ChatGPT.

2 Prerequisites

2.1 Basics of probability theory

For this chapter, we are referring to the lecture notes [Kra23] of the Probability Theory graduate course taught by Professor Felix Krahmer at TUM in the winter semester 2023/24.

Let us start by recalling the foundational objects of probability theory: probability measures, random variables, and expectations. To formally introduce probability measures, we need the concept of a σ -algebra, which is defined as follows.

Definition 2.1.1 (σ -Algebra) *Let Ω be a sample space. A σ -algebra \mathcal{F} on Ω is a collection of subsets of Ω such that:*

1. $\Omega \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$.
3. If $(A_n)_{n \in \mathbb{N}}$ is a sequence of sets in \mathcal{F} , then

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}.$$

If \mathcal{F} is a σ -algebra on Ω , then (Ω, \mathcal{F}) is called a measurable space.

The following example considers a simple σ -algebra, which will be used throughout this chapter to illustrate the concepts intuitively.

Example 2.1.2 Let $\Omega = \{1, 2, 3\}$. Define $A = \{1, 2\}$ and $B = \{3\}$, a partition of Ω .

One σ -algebra on Ω is

$$\mathcal{F} = \{\emptyset, A, B, \Omega\}.$$

Indeed:

- $\Omega \in \mathcal{F}$.
- The complement of A in Ω is B , which is in \mathcal{F} , and vice versa.
- Any countable union of these sets is again one of the sets in \mathcal{F} .

Next, we define probability spaces, which have a nice interpretation. The sample space Ω represents all possible outcomes of a random experiment, the σ -algebra \mathcal{F} specifies the collection of events we can assign probabilities to, and the probability measure \mathbb{P} quantifies the likelihood of these events.

Definition 2.1.3 (Probability Measure and Probability Space) Let $\Omega \neq \emptyset$ be a non-empty set (the sample space), and let \mathcal{F} be a σ -algebra on Ω . A probability measure is a function

$$\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$$

satisfying:

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

2. For any countable collection of pairwise disjoint sets $(A_n)_{n \in \mathbb{N}} \subseteq \mathcal{F}$,

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

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

We want to define random variables, one of the main objects in probability theory. For this, we need the notion of measurable functions.

Definition 2.1.4 (Measurable function) Let (X, \mathcal{X}) and (S, \mathcal{S}) be measurable spaces. A map $f : X \rightarrow S$ is called measurable (or $X - \mathcal{S}$ -measurable) if

$$f^{-1}(B) \in \mathcal{X} \quad \text{for every } B \in \mathcal{S}.$$

A function is measurable if, for every measurable subset of the target space, its preimage is a measurable subset of the domain.

Given this, we can define random elements, or more specifically, for our needs, random variables and their distribution.

Definition 2.1.5 (Random Element, Random Variable and Distribution) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let (S, \mathcal{S}) be a measurable space. A measurable function

$$X : \Omega \rightarrow S$$

is called a random element of S .

The distribution of X is the probability measure μ_X on (S, \mathcal{S}) defined by

$$\mu_X(A) := \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A), \quad \text{for } A \in \mathcal{S}.$$

In the special case $S = \mathbb{R}$ and $\mathcal{S} = \mathcal{B}(\mathbb{R})$ (the Borel σ -algebra on \mathbb{R} , i.e. the smallest σ -algebra containing all open subsets of \mathbb{R}), the random element X is called a random variable.

Given a random variable, it is natural to ask what its average or expected value is. Thus, we define the expected value of a random variable as follows.

Definition 2.1.6 (Expected Value) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \rightarrow \mathbb{R}$ be a random variable. If X is integrable, i.e.

$$\int_{\Omega} |X(\omega)| d\mathbb{P}(\omega) < \infty,$$

which is the definition of $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

Then the expected value of X is defined as

$$\mathbb{E}[X] := \int_{\Omega} X(\omega) d\mathbb{P}(\omega) = \mathbb{E}_{X \sim \mathbb{P}}[X].$$

Equivalently, if μ_X denotes the distribution of X , then

$$\mathbb{E}[X] = \int_{\mathbb{R}} x d\mu_X(x).$$

2.2 Conditional expectation

We will need the conditional expectation and its properties to define martingales. It is also an incredibly useful tool for proving U-statistics results.

Definition 2.2.1 (Conditional Expectation) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Let $X : \Omega \rightarrow \mathbb{R}$ be an $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$ -measurable random variable with $\mathbb{E}[|X|] < \infty$

or $X \geq 0$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ -algebra.

The conditional expectation $\mathbb{E}[X | \mathcal{G}]$ of X given \mathcal{G} is a random variable $Y : \Omega \rightarrow \mathbb{R}$ with the following properties:

(C1) Y is $(\mathcal{G}, \mathcal{B}(\mathbb{R}))$ -measurable.

(C2) For all $A \in \mathcal{G}$,

$$\int_A X d\mathbb{P} = \int_A Y d\mathbb{P}.$$

If $\mathbb{E}[|X|] < \infty$, then $\mathbb{E}[X | \mathcal{G}]$ is almost surely finite. Every random variable fulfilling (C1) and (C2) is called a version of $\mathbb{E}[X | \mathcal{G}]$.

It can also be shown that the conditional expectation exists and is almost surely unique. So from now on, we will talk about the conditional expectation, not just a version of it. To get a better understanding of conditional expectation, consider the following two examples.

Example 2.2.2 Let $\Omega = \{1, 2, 3\}$ and consider the full σ -algebra

$$\mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \Omega\}.$$

Define the coarser σ -algebra as in the σ -algebra example

$$\mathcal{G} = \{\emptyset, \{1, 2\}, \{3\}, \Omega\}.$$

Then clearly $\mathcal{G} \subseteq \mathcal{F}$. Intuitively, \mathcal{G} represents a situation where we can only distinguish whether the outcome lies in $A = \{1, 2\}$ or in $B = \{3\}$, while \mathcal{F} contains all possible information about Ω .

If $X : \Omega \rightarrow \mathbb{R}$ is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$, then the conditional expectation $\mathbb{E}[X | \mathcal{G}]$ is a random variable that is \mathcal{G} -measurable, i.e. it only depends on whether the outcome is in A or in B .

Example 2.2.3 (Conditional Expectation on a Finite Space) Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with

$$\Omega = \{1, 2, 3\}, \quad \mathbb{P}(\{1\}) = \frac{1}{2}, \quad \mathbb{P}(\{2\}) = \frac{1}{4}, \quad \mathbb{P}(\{3\}) = \frac{1}{4},$$

and $\mathcal{F} = \mathcal{P}(\Omega)$. Define a random variable

$$X(1) = 2, \quad X(2) = 0, \quad X(3) = 1.$$

Now take the coarser σ -algebra

$$\mathcal{G} = \{\emptyset, \{1, 2\}, \{3\}, \Omega\}.$$

Step 1: Form of the conditional expectation.

Since \mathcal{G} has atoms $A = \{1, 2\}$ and $B = \{3\}$, the conditional expectation $Y := \mathbb{E}[X | \mathcal{G}]$ must be constant on each atom. Hence we look for Y of the form

$$Y(\omega) = y_A \text{ for } \omega \in A, \quad Y(\omega) = y_B \text{ for } \omega \in B.$$

2 Prerequisites

Step 2: Determining y_A and y_B .

By the (C2) property of conditional expectation, for all $C \in \mathcal{G}$, $\int_C X d\mathbb{P} = \int_C Y d\mathbb{P}$. In particular:

$$\int_A X d\mathbb{P} = X(1) \mathbb{P}(\{1\}) + X(2) \mathbb{P}(\{2\}) = 1, \mathbb{P}(A) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) = \frac{3}{4}, y_A = \frac{\int_A X d\mathbb{P}}{\mathbb{P}(A)} = \frac{1}{3/4} = \frac{4}{3},$$

$$\int_B X d\mathbb{P} = X(3) \mathbb{P}(\{3\}) = \frac{1}{4}, \mathbb{P}(B) = \mathbb{P}(\{3\}) = \frac{1}{4}, y_B = \frac{\int_B X d\mathbb{P}}{\mathbb{P}(B)} = \frac{1/4}{1/4} = 1.$$

Because Y is constant on A , its integral over A is just that constant times the probability of event A . So, solving for that constant, we get the conditional expectation of X over the set A . Analogously, the same holds for the set B .

Thus

$$\mathbb{E}[X | \mathcal{G}](\omega) = \begin{cases} \frac{4}{3}, & \omega \in \{1, 2\}, \\ 1, & \omega = 3. \end{cases}$$

Step 3: Verification of the properties for conditional expectation.

Finally we check (C2) property for all $C \in \mathcal{G}$:

- $C = \emptyset$: both integrals are 0.
- $C = \{1, 2\}$: $\int_C X d\mathbb{P} = 1, \int_C Y d\mathbb{P} = 1$.
- $C = \{3\}$: $\int_C X d\mathbb{P} = \frac{1}{4}, \int_C Y d\mathbb{P} = \frac{1}{4}$.
- $C = \Omega$: $\int_\Omega X d\mathbb{P} = \frac{5}{4}, \int_\Omega Y d\mathbb{P} = \frac{5}{4}$.

Hence Y satisfies (C2) and indeed $Y = \mathbb{E}[X | \mathcal{G}]$.

As we will often use expectations conditioned on random variables, we define what this means, as right now, we have only defined what it means to be an expectation conditioned on a σ -algebra. For this, we need the definition of the σ -algebra generated by a random variable X .

Definition 2.2.4 Let $X : \Omega \rightarrow S$ be a random variable, where (S, \mathcal{S}) is a measurable space. The σ -algebra generated by X is defined as

$$\sigma(X) := \{X^{-1}(B) : B \in \mathcal{S}\}.$$

Equivalently, $\sigma(X)$ is the smallest σ -algebra on Ω that makes X measurable. It represents the information contained in the random variable X : if you only know the value of X , then the events you can distinguish are exactly those in $\sigma(X)$.

Finally, we get that for random variables X and Y , the conditional expectation of X given Y is defined as

$$\mathbb{E}[X | Y] := \mathbb{E}[X | \sigma(Y)].$$

We give a short example for the σ -algebra generated by a random variable.

Example 2.2.5 Suppose $\Omega = \{1, 2, 3, 4\}$ and define

$$X(1) = X(2) = 0, \quad X(3) = X(4) = 1.$$

Then the atoms of $\sigma(X)$ are the sets where X is constant:

$$\{1, 2\}, \quad \{3, 4\}.$$

Hence

$$\sigma(X) = \{\emptyset, \{1, 2\}, \{3, 4\}, \Omega\}.$$

This shows explicitly how $\sigma(X)$ partitions Ω according to the values of X .

Interpretation of $\mathbb{E}[X | Y]$. $\mathbb{E}[X | Y]$, the expectation of the r.v. X conditioned on the r.v. Y , is the *best guess for X if you know Y* . There are two extreme cases:

- If X is a function of Y , then knowing Y means knowing X . In this case

$$\mathbb{E}[X | Y] = X.$$

- If X and Y are independent, then Y gives no information about X . Our best guess for X is simply its expectation:

$$\mathbb{E}[X | Y] = \mathbb{E}[X].$$

This interpretation is confirmed by the following theorem.

Theorem 2.2.6 (a) If X is \mathcal{F} -measurable, then

$$\mathbb{E}[X | \mathcal{F}] = X \quad \text{almost surely.}$$

(b) If $\sigma(X)$ and \mathcal{F} are independent, then

$$\mathbb{E}[X | \mathcal{F}] = \mathbb{E}[X] \quad \text{almost surely.}$$

Note that:

- $\mathbb{E}[\mathbb{E}[X | \mathcal{F}]] = \mathbb{E}[X].$
- If X is \mathcal{F} -measurable, $\mathbb{E}[|XY|] < \infty$ and $\mathbb{E}[|Y|] < \infty$, then

$$\mathbb{E}[XY | \mathcal{F}] = X\mathbb{E}[Y | \mathcal{F}] \quad \text{almost surely.}$$

This also holds if $X \geq 0$ and $Y \geq 0$.

Theorem 2.2.7 (Tower Property) For $\mathcal{F}_1 \subseteq \mathcal{F}_2$, one has

$$\mathbb{E}[\mathbb{E}[X | \mathcal{F}_1] | \mathcal{F}_2] = \mathbb{E}[\mathbb{E}[X | \mathcal{F}_2] | \mathcal{F}_1] = \mathbb{E}[X | \mathcal{F}_1] \quad \text{almost surely.}$$

We will refer to this as: “The smaller σ -algebra wins.”

The theorems about the tower property, independence and measurability of the conditional expectation will be extensively used throughout the U-statistics chapter.

2.3 Martingales

We want to describe how a random variable evolves with time and study some of the properties of this process. For this, we introduce the concept of a stochastic process and its connection to a filtration, which models the currently available information.

Definition 2.3.1 (Stochastic Process and Filtration) Let $I \subseteq \mathbb{R}$.

A family of random variables $(X_t)_{t \in I}$ with

$$X_t : \Omega \rightarrow \mathbb{R}, \quad t \in I,$$

all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is called a real-valued stochastic process.

A family of σ -algebras $(\mathcal{F}_t)_{t \in I}$ with $\mathcal{F}_t \subseteq \mathcal{F}$ is called a filtration if

$$\mathcal{F}_s \subseteq \mathcal{F}_t \quad \text{for all } s, t \in I \text{ with } s < t.$$

2 Prerequisites

A stochastic process $(X_t)_{t \in I}$ is said to be adapted to the filtration $(\mathcal{F}_t)_{t \in I}$ if

$$X_t \text{ is } \mathcal{F}_t\text{-measurable for all } t \in I.$$

The filtration generated by $(X_t)_{t \in I}$ is defined as

$$\mathcal{F}_t := \sigma(X_s : s \in I, s \leq t), \quad t \in I,$$

and is called the natural or canonical filtration of X . Clearly, $(X_t)_{t \in I}$ is adapted to its natural filtration.

Given this definition of a stochastic process, it is natural to name the property that the conditional expectation of the process at a later time, given the information up to an earlier time, equals the random variable of that earlier time. This property we will call a martingale.

Definition 2.3.2 (Martingales in continuous time) A stochastic process $(M_t)_{t \in I}$ is called a martingale, supermartingale, or submartingale with respect to a filtration $(\mathcal{F}_t)_{t \in I}$ if the following conditions hold:

(M1) $(M_t)_{t \in I}$ is adapted to $(\mathcal{F}_t)_{t \in I}$ (i.e., M_t is \mathcal{F}_t -measurable for all $t \in I$).

(M2) $\mathbb{E}[|M_t|] < \infty \quad \forall t \in I$.

(M3) For all $s, t \in I$ with $s < t$,

$$\mathbb{E}[M_t | \mathcal{F}_s] \begin{cases} = M_s, & \text{martingale}, \\ \leq M_s, & \text{supermartingale}, \\ \geq M_s, & \text{submartingale}. \end{cases}$$

Remark. If $(M_t)_{t \in I}$ is a martingale, then $t \mapsto \mathbb{E}[M_t]$ is constant (for a supermartingale it would be decreasing and a submartingale increasing). That means a martingale can be thought of as a fair game.

Later, we will show that U-statistics have a martingale structure. As a simple example to see why a martingale can be thought of as a fair game, consider the symmetric random walk. The simple symmetric random walk got its name from the fact that, if at each time step we take either a step forward or backwards with the same probability, and independent of our previous steps, we are walking randomly and in a symmetric fashion.

Example 2.3.3 (Martingales from a Simple Symmetric Random Walk) Let $(\xi_k)_{k \geq 1}$ be i.i.d. with $\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = \frac{1}{2}$ and define the random walk

$$S_0 := 0, \quad S_t := \sum_{k=1}^t \xi_k \quad (t \geq 1).$$

Let $\mathcal{F}_t := \sigma(\xi_1, \dots, \xi_t)$ be the natural filtration.

Claim: $(S_t)_{t \geq 0}$ is a martingale w.r.t. $(\mathcal{F}_t)_{t \geq 0}$.

Verification. S_t is \mathcal{F}_t -measurable and $\mathbb{E}|S_t| < \infty$. For $s < t$,

$$\mathbb{E}[S_t | \mathcal{F}_s] = \mathbb{E}[S_s + (\xi_{s+1} + \dots + \xi_t) | \mathcal{F}_s] = S_s + \sum_{k=s+1}^t \mathbb{E}[\xi_k | \mathcal{F}_s] = S_s,$$

since ξ_{s+1}, \dots, ξ_t are independent of \mathcal{F}_s and have mean 0.

We now extend the notion of a martingale to discrete-time processes and introduce two related and very useful concepts: *martingale-difference sequences* and *reverse martingales*. These will play an important role later when analyzing U-statistics.

For a shorter notation, we will write

$$Z \in L^1 \iff \mathbb{E}[|Z|] < \infty.$$

For the following definitions let $n \in \mathbb{N}_0$.

Definition 2.3.4 (Martingales in discrete-time) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathbb{F} = (\mathcal{F}_n)_{n \geq 0}$ a filtration with $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$. A process $M = (M_n)_{n \geq 0}$ is called a (discrete-time) martingale with respect to \mathbb{F} if

$$(i) M_n \in L^1, \quad (ii) M_n \text{ is } \mathcal{F}_n\text{-measurable (adapted)}, \quad (iii) \mathbb{E}[M_{n+1} | \mathcal{F}_n] = M_n \text{ a.s. } (n \geq 0).$$

For this discrete-time version, notice the change of indices compared to the other definition of a martingale. We now only have to check the martingale property for each previous time step, instead of all possible previous time steps.

Next, we define a property for stochastic processes whose conditional expectation, given the current information, is equal to zero. Processes with this property are referred to as martingale differences.

Definition 2.3.5 (Martingale Difference) A sequence $(D_{n+1})_{n \geq 0}$ is called a martingale difference (with respect to \mathbb{F}) if $D_{n+1} \in L^1$, D_{n+1} is \mathcal{F}_{n+1} -measurable, and $\mathbb{E}[D_{n+1} | \mathcal{F}_n] = 0$ a.s.

The last property of stochastic processes, which we will need for our analysis of U-statistics is the reverse martingale. It is also known as a backwards martingale, since the indices can be thought of as moving backwards in time. Thus, we need a decreasing filtration, where we lose information over time.

Definition 2.3.6 (Reverse Martingale) Given a decreasing filtration $\mathbb{G} = (\mathcal{G}_n)_{n \geq 0}$ with $\mathcal{G}_{n+1} \subseteq \mathcal{G}_n$, a process $R = (R_n)_{n \geq 0}$ is called a reverse martingale if $R_n \in L^1$, R_n is \mathcal{G}_n -measurable, and $\mathbb{E}[R_n | \mathcal{G}_{n+1}] = R_{n+1}$ a.s.

As an example of a reverse martingale, consider the following.

Example 2.3.7 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G}_n = \sigma(X_k : k \geq n)$ the tail σ -algebra generated by an i.i.d. sequence $(X_k)_{k \geq 1}$ of integrable random variables. Define

$$R_n = \mathbb{E}[X_1 | \mathcal{G}_n].$$

Then $(R_n)_{n \geq 0}$ is a reverse martingale with respect to $\mathbb{G} = (\mathcal{G}_n)_{n \geq 0}$, because by the tower property of conditional expectations,

$$\mathbb{E}[R_n | \mathcal{G}_{n+1}] = \mathbb{E}[\mathbb{E}[X_1 | \mathcal{G}_n] | \mathcal{G}_{n+1}] = \mathbb{E}[X_1 | \mathcal{G}_{n+1}] = R_{n+1}.$$

Discrete-time simplification for martingales. In continuous time (with index set \mathbb{R}_+), the martingale property reads $\mathbb{E}[M_t | \mathcal{F}_s] = M_s$ for all $0 \leq s < t$. In discrete time (with index set \mathbb{N}), it is equivalent to require only the one-step condition $\mathbb{E}[M_{n+1} | \mathcal{F}_n] = M_n$ for all n . Indeed, if (i)–(iii) hold, then for any $m > n$,

$$\mathbb{E}[M_m | \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[M_m | \mathcal{F}_{m-1}] | \mathcal{F}_n] = \mathbb{E}[M_{m-1} | \mathcal{F}_n] = \cdots = \mathbb{E}[M_{n+1} | \mathcal{F}_n] = M_n,$$

by repeated application of the tower property. Hence, checking the one-step condition is sufficient.

Connection between martingales and martingale differences. Let $\mathbb{F} = (\mathcal{F}_n)_{n \geq 0}$ be a filtration.

(Martingale \Rightarrow Martingale Difference). Assume $M = (M_n)_{n \geq 0}$ is a martingale: $M_n \in L^1$, M_n is \mathcal{F}_n -measurable, and $\mathbb{E}[M_{n+1} | \mathcal{F}_n] = M_n$ a.s. Define

$$D_{n+1} := M_{n+1} - M_n, \quad n \geq 0.$$

Then $D_{n+1} \in L^1$ and D_{n+1} is \mathcal{F}_{n+1} -measurable. Moreover,

$$\mathbb{E}[D_{n+1} | \mathcal{F}_n] = \mathbb{E}[M_{n+1} - M_n | \mathcal{F}_n] = \mathbb{E}[M_{n+1} | \mathcal{F}_n] - M_n = M_n - M_n = 0,$$

so $(D_{n+1}, \mathcal{F}_{n+1})$ is a martingale difference.

(Martingale Difference \Rightarrow Martingale). Conversely, suppose $(D_{n+1})_{n \geq 0}$ is a martingale difference: $D_{n+1} \in L^1$, D_{n+1} is \mathcal{F}_{n+1} -measurable, and $\mathbb{E}[D_{n+1} | \mathcal{F}_n] = 0$ a.s. Let $M_0 \in L^1$ be \mathcal{F}_0 -measurable and define

$$M_n := M_0 + \sum_{k=0}^{n-1} D_{k+1}, \quad n \geq 0.$$

Then $M_n \in L^1$ and M_n is \mathcal{F}_n -measurable. Furthermore,

$$\mathbb{E}[M_{n+1} | \mathcal{F}_n] = \mathbb{E}[M_n + D_{n+1} | \mathcal{F}_n] = M_n + \mathbb{E}[D_{n+1} | \mathcal{F}_n] = M_n,$$

so $M = (M_n, \mathcal{F}_n)$ is a martingale.

We now present an important example of martingale differences that we will later apply to U-statistics.

Example 2.3.8 (Martingale differences using conditional expectations) Let $\mathbb{F} = (\mathcal{F}_k)_{k \geq 0}$ be a filtration and $n \in \mathbb{N}$. Assume Y_n is \mathcal{F}_n -measurable with $\mathbb{E}|Y_n| < \infty$. Define, for $k = 1, \dots, n$,

$$\xi_{n,k} := \mathbb{E}[Y_n | \mathcal{F}_k] - \mathbb{E}[Y_n | \mathcal{F}_{k-1}].$$

Then $(\xi_{n,k}, \mathcal{F}_k)_{k=1}^n$ is a martingale difference and

$$Y_n - \mathbb{E}Y_n = \sum_{k=1}^n \xi_{n,k}.$$

Proof. Set $M_k := \mathbb{E}[Y_n | \mathcal{F}_k]$ for $k = 0, \dots, n$. Then $\xi_{n,k} = M_k - M_{k-1}$, so

$$\sum_{k=1}^n \xi_{n,k} = \sum_{k=1}^n (M_k - M_{k-1}) = M_n - M_0 = \mathbb{E}[Y_n | \mathcal{F}_n] - \mathbb{E}[Y_n | \mathcal{F}_0] = Y_n - \mathbb{E}Y_n,$$

because Y_n is \mathcal{F}_n -measurable and \mathcal{F}_0 is trivial. Moreover,

$$\mathbb{E}[\xi_{n,k} | \mathcal{F}_{k-1}] = \mathbb{E}[M_k - M_{k-1} | \mathcal{F}_{k-1}] = \mathbb{E}[M_k | \mathcal{F}_{k-1}] - M_{k-1} = M_{k-1} - M_{k-1} = 0,$$

hence $(\xi_{n,k})$ is a martingale difference.

3 U-statistics

3.1 Motivation

U-statistics provide the natural mathematical framework for unbiased and efficient estimation of population functionals. Many statistics of interest can be written as averages of a symmetric kernel over tuples of observations, such as the sample mean and variance. In this chapter, we will cover the notion of a U-statistic and important examples, canonical functions and their property of complete degeneracy, which will play a crucial role in the MMD-based statistical test in Chapter 6. We then derive the Hoeffding representation, i.e. we can write every U-statistic as a linear combination of complete degenerate U-statistics. Afterwards, we show the martingale and reverse martingale structure of the U-statistics. At the end of the chapter, we will give a short introduction to the V-statistic, a biased estimator, and compare it to the U-statistic. The theory is based on [KB13], where interested readers can find the extensions to Banach spaces as well as the asymptotic theory of U-statistics. Whenever it is possible, an intuitive example will be given and explained in detail.

3.2 Definition and examples

Now, we can formally introduce the concept of the parametric functional and its estimator of the U-statistic. Parametric functionals arise naturally in statistics as quantities that describe certain characteristics of an underlying probability distribution P . Examples include the mean, variance, covariance, correlation, and higher-order moments. Each such functional can be written as the expected value of a measurable function Φ of a finite number of i.i.d. random variables drawn from P . The function Φ , called the *kernel*, determines how the sample values interact to estimate the corresponding population parameter.

Definition 3.2.1 (Parametric (regular) functional) Let $\mathcal{P} = \{P\}$ be a class of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A mapping $\Theta : \mathcal{P} \rightarrow \mathbb{R}$ (or \mathbb{R}^k) is called a parametric (regular) functional if there exist an integer $m \geq 1$ and a measurable function $\Phi : \mathbb{R}^m \rightarrow \mathbb{R}$ such that, for $X_1, \dots, X_m \stackrel{i.i.d.}{\sim} P$, the functional can be written as

$$\Theta(P) = \mathbb{E}[\Phi(X_1, \dots, X_m)] = \int \cdots \int \Phi(x_1, \dots, x_m) P(dx_1) \cdots P(dx_m). \quad (3.1)$$

The function Φ is called the *kernel* of $\Theta(P)$ and m is the *degree* of $\Theta(P)$. For sake of simplicity, we sometimes just call it *parametric functional* and often write it just as Θ , omitting the probability distribution.

Thus, U-statistics provide a unified framework for constructing unbiased estimators of a wide class of such functionals by replacing the population expectation in (3.1) with the corresponding average over all combinations of sample observations.

Example 3.2.2 Consider the following examples of different parametric functionals with their respective kernels. The derivation of the mean and variance will be given in Examples 3.2.6 and 3.2.7.

- **Mean:**

$$\Theta(P) = \mathbb{E}_P[X_1], \quad \Phi(x_1) = x_1, \quad m = 1.$$

The mean is a functional of degree one, with Φ being the identity function. The U-statistic estimator of it will be the sample mean.

- **Variance:**

$$\Theta(P) = \frac{1}{2} \mathbb{E}[(X_1 - X_2)^2], \quad \Phi(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2, \quad m = 2.$$

The corresponding U-statistic is the unbiased sample variance.

- **Third central moment (skewness):**

$$\Theta(P) = \frac{1}{6} \mathbb{E}[(X_1 + X_2 + X_3 - 3\mathbb{E}[X])^3],$$

which can be expressed as a third-degree functional with

$$\Phi(x_1, x_2, x_3) = \frac{1}{6}(x_1 + x_2 + x_3)^3.$$

This functional captures asymmetry in the distribution.

Without loss of generality, Φ may be taken symmetric in its arguments. Otherwise we replace it by its symmetrization

$$\Phi_0(x_1, \dots, x_m) := \frac{1}{m!} \sum_{\sigma \in S_m} \Phi(x_{\sigma(1)}, \dots, x_{\sigma(m)}),$$

where S_m is the permutation group on $\{1, \dots, m\}$.

Properties of Φ_0 .

1. Φ_0 is symmetric.

2. For all $P \in \mathcal{P}$ and $X_1, \dots, X_m \stackrel{\text{i.i.d.}}{\sim} P$,

$$\Theta(P) = \mathbb{E}_{P^{\otimes m}}[\Phi(X_1, \dots, X_m)] = \mathbb{E}_{P^{\otimes m}}[\Phi_0(X_1, \dots, X_m)],$$

so the target functional is unchanged by symmetrization.

Example 3.2.3 (symmetrizing $\Phi(x_1, x_2)$) Consider the (non-symmetric) kernel

$$\Phi(x_1, x_2) = x_1^2 - x_2, \quad (x_1, x_2) \in \mathbb{R}^2.$$

Clearly, $\Phi(x_1, x_2) \neq \Phi(x_2, x_1)$ in general.

Step 1: Symmetrization.

For $m = 2$, the permutation group $S_2 = \{\text{id}, (12)\}$. The symmetrized kernel is

$$\Phi_0(x_1, x_2) := \frac{1}{2} \sum_{\sigma \in S_2} \Phi(x_{\sigma(1)}, x_{\sigma(2)}) = \frac{1}{2} (\Phi(x_1, x_2) + \Phi(x_2, x_1)).$$

Substitute Φ :

$$\Phi(x_1, x_2) = x_1^2 - x_2, \quad \Phi(x_2, x_1) = x_2^2 - x_1,$$

hence

$$\boxed{\Phi_0(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2 - x_1 - x_2).}$$

By construction, $\Phi_0(x_1, x_2) = \Phi_0(x_2, x_1)$, so Φ_0 is symmetric.

Step 2: Target functional is unchanged.

Let $X, X_1, X_2 \stackrel{\text{i.i.d.}}{\sim} P$. Then

$$\mathbb{E}[\Phi(X_1, X_2)] = \mathbb{E}[X_1^2] - \mathbb{E}[X_2] = \mathbb{E}[X^2] - \mathbb{E}[X],$$

while

$$\mathbb{E}[\Phi_0(X_1, X_2)] = \frac{1}{2} (\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] - \mathbb{E}[X_1] - \mathbb{E}[X_2]) = \mathbb{E}[X^2] - \mathbb{E}[X].$$

Thus

$$\boxed{\mathbb{E}_{P^{\otimes 2}}[\Phi] = \mathbb{E}_{P^{\otimes 2}}[\Phi_0] = \Theta(P) := \mathbb{E}[X^2] - \mathbb{E}[X].}$$

Thus, we can w.l.o.g. always assume that the kernel is symmetric in its arguments.

To estimate a parametric functional on a finite sample, we will introduce the U-statistic, an unbiased estimator of the parametric functional, where the U in U-statistics stands for unbiased. This leads us to the following definition:

Definition 3.2.4 (U-statistic (unbiased estimator of $\Theta(P)$)) Let the kernel Φ be symmetric. Given $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P$, the U-statistic of degree m is defined as:

$$U_n := \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} \Phi(X_{i_1}, \dots, X_{i_m})$$

and satisfies

$$\mathbb{E}[U_n] = \Theta(P),$$

so it is an unbiased estimator for the functional $\Theta(P)$. Here, $\binom{n}{m}$ denotes the binomial coefficient, defined as

$$\binom{n}{m} = \frac{n!}{m!(n-m)!},$$

which represents the number of distinct ways to choose m elements from a set of n elements.

In general, the kernel can also belong to a Banach space, thus for completeness the following definition:

Definition 3.2.5 (Banach space valued U-statistic, UR-statistic) Let B be a separable real Banach space and let $\Phi : \mathbb{R}^m \rightarrow B$ be a symmetric (measurable) kernel. For $n \geq m$, the U-statistic is called a UB-statistic. Clearly, $U_n \in B$. If $B = H$ is a Hilbert space, then it is called UH-statistic; If $B = \mathbb{R}$ it is a UR-statistic (i.e., a real-valued U-statistic).

In the following, we will only cover UR-statistic theory and for simplicity of notation, we refer to it as a U -statistic. We also focus on one-sample U-statistics here, while the theory can be extended to multi-sample U-statistics.

In the following examples, we will see that the sample mean and sample variance can be written as U-statistics.

Example 3.2.6 (Sample mean) Let $X, X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P$ and $\Phi(x) = x$. The functional is

$$\Theta(P) = \mathbb{E}_P[X] = \mu(P) = \int x dP(x).$$

The associated U -statistic ($m = 1$) is the sample mean:

$$U_n = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}, \quad \mathbb{E}[U_n] = \Theta(P).$$

Example 3.2.7 (Sample variance) Let $X, X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P$ and take degree $m = 2$ with

$$\Phi(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2, \quad \Theta(P) = \frac{1}{2} \mathbb{E}[(X_1 - X_2)^2] = \text{Var}(X).$$

The associated U -statistic is

$$U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \frac{1}{2} (X_i - X_j)^2.$$

Using the identity

$$\sum_{1 \leq i < j \leq n} (X_i - X_j)^2 = n \sum_{i=1}^n (X_i - \bar{X})^2,$$

we obtain

$$U_n = \frac{1}{\binom{n}{2}} \frac{1}{2} n \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Hence U_n equals the unbiased sample variance and satisfies $\mathbb{E}[U_n] = \Theta(P) = \text{Var}(X)$.

The identity holds by the following arguments.

$$\begin{aligned} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 &= \sum_{i < j} (X_i^2 + X_j^2 - 2X_i X_j) \\ &= (n-1) \sum_{i=1}^n X_i^2 - 2 \sum_{i < j} X_i X_j \\ &= (n-1) \sum_{i=1}^n X_i^2 - \left[\left(\sum_{i=1}^n X_i \right)^2 - \sum_{i=1}^n X_i^2 \right] \\ &= n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2. \end{aligned} \tag{I}$$

On the other hand,

$$\begin{aligned} n \sum_{i=1}^n (X_i - \bar{X})^2 &= n \sum_{i=1}^n (X_i^2 - 2X_i \bar{X} + \bar{X}^2) \\ &= n \sum_{i=1}^n X_i^2 - 2n\bar{X} \sum_{i=1}^n X_i + n^2\bar{X}^2 \\ &= n \sum_{i=1}^n X_i^2 - 2n^2\bar{X}^2 + n^2\bar{X}^2 \\ &= n \sum_{i=1}^n X_i^2 - n^2\bar{X}^2. \end{aligned} \tag{II}$$

Since $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, we have

$$n^2\bar{X}^2 = \left(\sum_{i=1}^n X_i \right)^2.$$

Substituting this into (II) yields

$$n \sum_{i=1}^n (X_i - \bar{X})^2 = n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2,$$

which is identical to the expression in (I). Therefore,

$$\sum_{1 \leq i < j \leq n} (X_i - X_j)^2 = n \sum_{i=1}^n (X_i - \bar{X})^2.$$

Example 3.2.8 (Specific kernel with m=2) Let $\Phi(x_1, x_2) = x_1 x_2$. For i.i.d. $X, X_1, X_2 \sim P$,

$$\Theta(P) = \mathbb{E}[\Phi(X_1, X_2)] = \mathbb{E}[X_1 X_2] = (\mathbb{E}[X])^2 = \mu^2(P) = \left(\int x dP(x) \right)^2.$$

The degree-2 U-statistic is

$$U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} X_i X_j = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} X_i X_j, \quad \mathbb{E}[U_n] = \Theta(P).$$

(Equivalently, in ordered form: $U_n = \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j$.)

Example 3.2.9 (U-statistic for $m = 2, n = 3$) To get a better understanding, we see that a U-statistic is just a generalisation of the mean. For a general symmetric kernel $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $n = 3$,

$$U_3 = \binom{3}{2}^{-1} \sum_{1 \leq i < j \leq 3} \Phi(X_i, X_j) = \frac{1}{3} (\Phi(X_1, X_2) + \Phi(X_1, X_3) + \Phi(X_2, X_3)).$$

3.3 Canonical functions and degeneracy of U-statistics

We will now make use of the probability theory chapter, especially the conditional expectation properties. We introduce canonical functions and their property of complete degeneracy, which is crucial for U-statistics theory, as it is a major tool to prove statements, such as the martingale structure.

Let us introduce some notations to simplify the definition of canonical functions. Let $X_1, \dots, X_m \stackrel{\text{i.i.d.}}{\sim} P$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, let $\Phi : \mathbb{R}^m \rightarrow \mathbb{R}$ be measurable with $\mathbb{E}|\Phi(X_1, \dots, X_m)| < \infty$, and set

$$\Theta(P) := \mathbb{E}\Phi(X_1, \dots, X_m).$$

For $c = 0, 1, \dots, m$ define

$$\Phi_c(x_1, \dots, x_c) := \mathbb{E}[\Phi(x_1, \dots, x_c, X_{c+1}, \dots, X_m)] = \mathbb{E}[\Phi(X_1, \dots, X_m) \mid X_1 = x_1, \dots, X_c = x_c], \quad (3.2)$$

so that $\Phi_0 = \Theta(P) = \mathbb{E}_{P^{\otimes m}}[\Phi(X_1, \dots, X_m)]$ and $\Phi_m = \Phi$. For $c = 0, 1, \dots, m-1$ one has

$$\Phi_c(x_1, \dots, x_c) = \mathbb{E}[\Phi_{c+1}(x_1, \dots, x_c, X_{c+1})], \quad (3.3)$$

To see why the recursion in Equation (3.3) holds, fix $c \in \{0, \dots, m-1\}$ and set the σ -algebras $\mathcal{G} := \sigma(X_1, \dots, X_c) \subset \mathcal{H} := \sigma(X_1, \dots, X_c, X_{c+1})$. By the *tower property* (here smaller σ -algebra wins),

$$\mathbb{E}[Z \mid \mathcal{G}] = \mathbb{E}[\mathbb{E}[Z \mid \mathcal{H}] \mid \mathcal{G}] \quad \text{for all integrable } Z.$$

Apply this with $Z = \Phi(X_1, \dots, X_m)$. Conditioning on the event $X_1 = x_1, \dots, X_c = x_c$ gives

$$\begin{aligned} \Phi_c(x_1, \dots, x_c) &= \mathbb{E}[\Phi(X_1, \dots, X_m) \mid X_1 = x_1, \dots, X_c = x_c] \\ &= \mathbb{E}[\mathbb{E}[\Phi(X_1, \dots, X_m) \mid X_1 = x_1, \dots, X_c = x_c, X_{c+1}] \mid X_1 = x_1, \dots, X_c = x_c] \\ &= \mathbb{E}[\Phi_{c+1}(x_1, \dots, x_c, X_{c+1}) \mid X_1 = x_1, \dots, X_c = x_c]. \end{aligned}$$

If X_{c+1} is independent of (X_1, \dots, X_c) (as in the i.i.d. setting), the conditional distribution of X_{c+1} given $X_1 = x_1, \dots, X_c = x_c$ is just P , so we may drop the outer conditioning and write

$$\Phi_c(x_1, \dots, x_c) = \mathbb{E}_{X_{c+1} \sim P}[\Phi_{c+1}(x_1, \dots, x_c, X_{c+1})], \quad c = 0, 1, \dots, m-1.$$

Introduce the centered versions

$$\tilde{\Phi} := \Phi - \Theta(P), \quad \tilde{\Phi}_c := \Phi_c - \Theta(P), \quad 1 \leq c \leq m.$$

Definition 3.3.1 (Canonical functions) Define g_1, \dots, g_m recursively by

$$\begin{aligned} g_1(x_1) &:= \tilde{\Phi}_1(x_1), \\ g_2(x_1, x_2) &:= \tilde{\Phi}_2(x_1, x_2) - g_1(x_1) - g_1(x_2), \\ g_3(x_1, x_2, x_3) &:= \tilde{\Phi}_3(x_1, x_2, x_3) - \sum_{i=1}^3 g_1(x_i) - \sum_{1 \leq i < j \leq 3} g_2(x_i, x_j), \\ &\vdots \\ g_m(x_1, \dots, x_m) &:= \tilde{\Phi}_m(x_1, \dots, x_m) - \sum_{i=1}^m g_1(x_i) - \sum_{1 \leq i_1 < i_2 \leq m} g_2(x_{i_1}, x_{i_2}) - \cdots \\ &\quad - \sum_{1 \leq i_1 < \dots < i_{m-1} \leq m} g_{m-1}(x_{i_1}, \dots, x_{i_{m-1}}). \end{aligned}$$

By using the definition of canonical functions, we get for $c = 1, \dots, m$,

$$\tilde{\Phi}_c(x_1, \dots, x_c) = \sum_{d=1}^c \sum_{1 \leq i_1 < \dots < i_d \leq c} g_d(x_{i_1}, \dots, x_{i_d}). \quad (3.4)$$

Definition 3.3.2 (Property of complete degeneracy) For $d = 1, \dots, m$ the functions g_d are symmetric in their d arguments and satisfy the property of complete degeneracy:

$$\begin{aligned} \mathbb{E} g_1(X_1) &= 0, \\ \mathbb{E} g_2(x_1, X_2) &= 0, \\ &\vdots \\ \mathbb{E} g_m(x_1, \dots, x_{m-1}, X_m) &= 0. \end{aligned}$$

As the notation suggests, the canonical functions possess the property of complete degeneracy, which we show for $m \leq 3$.

Complete degeneracy of the canonical functions for $m \leq 3$. Let $X_1, X_2, X_3 \stackrel{\text{i.i.d.}}{\sim} P$ and let

$$\Phi_c(x_1, \dots, x_c) = \mathbb{E}[\Phi(X_1, \dots, X_m) \mid X_1 = x_1, \dots, X_c = x_c], \quad c = 0, 1, \dots, m.$$

Set $\Phi_0 = \mathbb{E}\Phi(X_1, \dots, X_m)$ and the centered versions $\tilde{\Phi}_c = \Phi_c - \Phi_0$. For $m = 1, 2, 3$, we get the canonical functions

$$\begin{aligned} g_1(x_1) &= \tilde{\Phi}_1(x_1), & g_2(x_1, x_2) &= \tilde{\Phi}_2(x_1, x_2) - g_1(x_1) - g_1(x_2), \\ g_3(x_1, x_2, x_3) &= \tilde{\Phi}_3(x_1, x_2, x_3) - \sum_{i=1}^3 g_1(x_i) - \sum_{1 \leq i < j \leq 3} g_2(x_i, x_j). \end{aligned}$$

Key recursion (tower property). For $c = 0, 1, \dots, m-1$,

$$\Phi_c(x_1, \dots, x_c) = \mathbb{E}[\Phi_{c+1}(x_1, \dots, x_c, X_{c+1}) \mid X_1 = x_1, \dots, X_c = x_c],$$

hence, in particular for $m \leq 3$,

$$\mathbb{E}_{X_1} \Phi_1(X_1) = \Phi_0, \quad \mathbb{E}_{X_2} \Phi_2(x_1, X_2) = \Phi_1(x_1), \quad \mathbb{E}_{X_3} \Phi_3(x_1, x_2, X_3) = \Phi_2(x_1, x_2).$$

Subtracting Φ_0 yields the same relations with tildes: $\mathbb{E}\tilde{\Phi}_1(X_1) = 0$, $\mathbb{E}_{X_2} \tilde{\Phi}_2(x_1, X_2) = \tilde{\Phi}_1(x_1)$, $\mathbb{E}_{X_3} \tilde{\Phi}_3(x_1, x_2, X_3) = \tilde{\Phi}_2(x_1, x_2)$.

Degeneracy for g_1 (case $m \geq 1$).

$$\mathbb{E} g_1(X_1) = \mathbb{E} \tilde{\Phi}_1(X_1) = \mathbb{E} \Phi_1(X_1) - \Phi_0 = \Phi_0 - \Phi_0 = 0.$$

Degeneracy for g_2 (available when $m \geq 2$). For P -a.e. fixed x_1 ,

$$\mathbb{E}_{X_2} g_2(x_1, X_2) = \mathbb{E}_{X_2} \tilde{\Phi}_2(x_1, X_2) - g_1(x_1) - \mathbb{E}_{X_2} g_1(X_2) = \tilde{\Phi}_2(x_1) - g_1(x_1) - 0 = 0.$$

Equivalently, $\mathbb{E}[g_2(X_1, X_2) | X_1] = 0$ a.s.; by symmetry also $\mathbb{E}[g_2(X_1, X_2) | X_2] = 0$.

Degeneracy for g_3 (available when $m = 3$). For $P^{\otimes 2}$ -a.e. fixed (x_1, x_2) ,

$$\begin{aligned} \mathbb{E}_{X_3} g_3(x_1, x_2, X_3) &= \underbrace{\mathbb{E}_{X_3} \tilde{\Phi}_3(x_1, x_2, X_3)}_{= \tilde{\Phi}_2(x_1, x_2)} - g_1(x_1) - g_1(x_2) - \underbrace{\mathbb{E}_{X_3} g_2(x_1, X_3)}_{= 0} - \underbrace{\mathbb{E}_{X_3} g_2(x_2, X_3)}_{= 0} - g_2(x_1, x_2) \\ &= (\tilde{\Phi}_2(x_1, x_2) - g_1(x_1) - g_1(x_2)) - g_2(x_1, x_2) = 0, \end{aligned}$$

since $g_2(x_1, x_2) = \tilde{\Phi}_2(x_1, x_2) - g_1(x_1) - g_1(x_2)$. Equivalently, $\mathbb{E}[g_3(X_1, X_2, X_3) | X_j] = 0$ a.s. for $j = 1, 2, 3$.

Conclusion. For $m \leq 3$ the components g_1, g_2, g_3 are symmetric and satisfy the complete degeneracy conditions:

$$\mathbb{E} g_1(X_1) = 0, \quad \mathbb{E} g_2(x_1, X_2) = 0 \quad (P\text{-a.e. } x_1), \quad \mathbb{E} g_3(x_1, x_2, X_3) = 0 \quad (P^{\otimes 2}\text{-a.e. } (x_1, x_2)).$$

□

To illustrate the workings of canonical functions, consider the following example.

Example 3.3.3 ($m = 2$, projections Φ_c for $\Phi(x_1, x_2) = x_1 x_2$) Let $X_1, X_2 \stackrel{i.i.d.}{\sim} P$ with $\mu := \mathbb{E}_P[X]$ and kernel $\Phi(x_1, x_2) = x_1 x_2$. The projections are

$$\Phi_0 = \mathbb{E}[\Phi(X_1, X_2)] = \mu^2, \quad \Phi_1(x_1) = \mathbb{E}[\Phi(x_1, X_2)] = \mathbb{E}[x_1 X_2] = \mu x_1, \quad \Phi_2(x_1, x_2) = \Phi(x_1, x_2) = x_1 x_2.$$

We get the centered versions:

$$\tilde{\Phi}(x_1, x_2) = x_1 x_2 - \mu^2, \quad \tilde{\Phi}_1(x_1) = \mu(x_1 - \mu), \quad \tilde{\Phi}_2(x_1, x_2) = x_1 x_2 - \mu^2.$$

The corresponding canonical functions are:

$$g_1(x) = \tilde{\Phi}_1(x_1) = \mu(x_1 - \mu), \quad g_2(x_1, x_2) = \tilde{\Phi}_2(x_1, x_2) - g_1(x_1) - g_1(x_2) = x_1 x_2 - \mu x_1 - \mu x_2 + \mu^2.$$

Next, we check for complete degeneracy:

$$\mathbb{E} g_1(X_1) = \mu \mathbb{E}[X_1 - \mu] = 0, \quad \mathbb{E} g_2(x_1, X_2) = x_1 \mu - \mu x_1 - \mu \mathbb{E}[X_2] + \mu^2 = x_1 \mu - \mu x_1 - \mu^2 + \mu^2 = 0,$$

and symmetrically $\mathbb{E} g_2(X_1, x_2) = 0$. Hence g_1 and g_2 are canonical.

Lastly, we express g_c explicitly in terms of the projections Φ_c . We omit the derivation, as it adds little additional insight; see [KB13] on page 21 for a full proof.

For $c = 1, \dots, m$, the canonical function can be written in terms of the projections as

$$g_c(x_1, \dots, x_c) = (-1)^c \Theta(P) + \sum_{d=1}^c (-1)^{c-d} \sum_{1 \leq j_1 < \dots < j_d \leq c} \Phi_d(x_{j_1}, \dots, x_{j_d}).$$

Equivalently, using $\tilde{\Phi}_d := \Phi_d - \Theta(P)$,

$$g_c(x_1, \dots, x_c) = \sum_{d=1}^c (-1)^{c-d} \sum_{1 \leq j_1 < \dots < j_d \leq c} \tilde{\Phi}_d(x_{j_1}, \dots, x_{j_d}).$$

3.4 The Hoeffding representation

In this section, we derive one of the most fundamental results in the theory of U-statistics – the *Hoeffding representation* (also known as the *Hoeffding decomposition*). It expresses a U-statistic as a sum of completely degenerate U-statistics. This representation not only provides valuable insight into the structure of U-statistics but also forms the basis for many asymptotic results, including variance decompositions and limit theorems.

Derivation of the Hoeffding representation. Let U_n be the degree- m U-statistic with kernel Φ and write

$$\tilde{\Phi} := \Phi - \Theta(P), \quad \Theta(P) = \mathbb{E} \Phi(X_1, \dots, X_m).$$

Let us rewrite it in centered form as follows:

$$U_n - \Theta(P) = \binom{n}{m}^{-1} S_n, \quad S_n := \sum_{1 \leq i_1 < \dots < i_m \leq n} \tilde{\Phi}(X_{i_1}, \dots, X_{i_m}). \quad (3.5)$$

For $c = 1, \dots, m$ set

$$S_{nc} := \sum_{1 \leq i_1 < \dots < i_c \leq n} g_c(X_{i_1}, \dots, X_{i_c}). \quad (3.6)$$

Using equation (3.4) with $c=m$, we get $\tilde{\Phi}(x_1, \dots, x_m) = \sum_{c=1}^m \sum_{1 \leq j_1 < \dots < j_c \leq m} g_c(x_{j_1}, \dots, x_{j_c})$, and inserting it into (3.5) yields

$$S_n = \sum_{1 \leq i_1 < \dots < i_m \leq n} \sum_{c=1}^m \sum_{1 \leq j_1 < \dots < j_c \leq m} g_c(X_{i_{j_1}}, \dots, X_{i_{j_c}}).$$

We fix c as we pull the summation over c to the front. For this fixed c , we now analyze how often each distinct term of the form $g_c(X_{r_1}, \dots, X_{r_c})$ appears as we vary all m -tuples (i_1, \dots, i_m) with $1 \leq i_1 < \dots < i_m \leq n$.

Fix a c -tuple with $1 \leq r_1 < \dots < r_c \leq n$. The term $g_c(X_{r_1}, \dots, X_{r_c})$ occurs whenever the fixed indices $\{r_1, \dots, r_c\}$ are contained in the chosen m -tuple $\{i_1, \dots, i_m\}$ from $\{1, \dots, n\}$. In other words, we must complete the fixed c -subset $\{r_1, \dots, r_c\}$ to a full m -subset by choosing the remaining $m - c$ indices among the $n - c$ elements in $\{1, \dots, n\} \setminus \{r_1, \dots, r_c\}$. The number of such completions is given by the binomial coefficient $\binom{n-c}{m-c}$. Each completion corresponds to one distinct combination in the outer sum over (i_1, \dots, i_m) , hence the same term $g_c(X_{r_1}, \dots, X_{r_c})$ appears exactly $\binom{n-c}{m-c}$ times.

Because r_1, \dots, r_c may take any values in $\{1, \dots, n\}$, we must sum over all possible c -subsets of $\{1, \dots, n\}$. This changes the index range of the inner sum from $\{1, \dots, m\}$ to $\{1, \dots, n\}$, since we are now summing over all sample indices rather than only over the kernel arguments. Formally, we replace the inner summation $\sum_{1 \leq j_1 < \dots < j_c \leq m}$ by $\sum_{1 \leq r_1 < \dots < r_c \leq n}$, yielding

$$S_n = \sum_{c=1}^m \binom{n-c}{m-c} \sum_{1 \leq r_1 < \dots < r_c \leq n} g_c(X_{r_1}, \dots, X_{r_c}) = \sum_{c=1}^m \binom{n-c}{m-c} S_{nc}.$$

The identity

$$\binom{n-c}{m-c} = \binom{n}{m} \binom{m}{c} \binom{n}{c}^{-1}$$

follows from basic properties of binomial coefficients. To see this, recall that $\binom{n}{m} = \frac{n!}{m!(n-m)!}$. Then

$$\binom{n}{m} \binom{m}{c} \binom{n}{c}^{-1} = \frac{n!}{m!(n-m)!} \cdot \frac{m!}{c!(m-c)!} \cdot \frac{c!(n-c)!}{n!} = \frac{(n-c)!}{(m-c)!(n-m)!} = \binom{n-c}{m-c}.$$

Since $\binom{n-c}{m-c} = \binom{n}{m} \binom{m}{c} \binom{n}{c}^{-1}$, we get

$$S_n = \sum_{c=1}^m \binom{n}{m} \binom{m}{c} \binom{n}{c}^{-1} S_{nc}.$$

Combining with Equation (3.5) gives the Hoeffding representation (or also known as canonical decomposition)

$$U_n - \Theta(P) = \sum_{c=1}^m \binom{m}{c} \binom{n}{c}^{-1} S_{nc}. \quad (3.7)$$

The Hoeffding representation shows that each U-statistic can be expressed as a linear combination of terms involving the canonical functions g_c . However, not all of these components necessarily contribute — some may vanish due to the degeneracy of the kernel. To formalize this idea, we introduce the concept of the *rank* of a U-statistic, which identifies the smallest order c for which the corresponding canonical function g_c is nonzero.

Definition 3.4.1 (Rank of a U-statistic / kernel) *Let $\Phi : \mathbb{R}^m \rightarrow \mathbb{R}$ be a measurable symmetric kernel with canonical functions g_1, \dots, g_m . The rank of the U-statistic with kernel Φ (equivalently, the rank of Φ) is*

$$r := \min\{c \in \{1, \dots, m\} : g_c \not\equiv 0\}.$$

Thus $g_1 = \dots = g_{r-1} \equiv 0$ and $g_r \not\equiv 0$.

Terminology:

- If $r = 1$, the U-statistic (kernel) is nondegenerate.
- If $r \geq 2$, it is degenerate and r is the order of degeneracy.
- If $r = m$, the kernel has the property of complete degeneracy.

For $c = 1, \dots, m$ define

$$U_{nc} := \binom{n}{c}^{-1} \sum_{1 \leq i_1 < \dots < i_c \leq n} g_c(X_{i_1}, \dots, X_{i_c}). \quad (3.8)$$

If the kernel has rank r (i.e. $g_1 = \dots = g_{r-1} \equiv 0$, $g_r \not\equiv 0$), then $S_{nc} = 0$ for $c < r$ (see eq. (3.10)), so the sum starts at $c = r$.

Then the Hoeffding representation can be written as

$$U_n - \Theta = \sum_{c=r}^m \binom{m}{c} U_{nc}. \quad (3.9)$$

In words, any U-statistic is a linear combination of the completely degenerate U-statistics in (3.8), with the sum starting at the rank r .

3.5 The martingale structure of U-statistics

Next, we show that each component in the Hoeffding decomposition of a U-statistic exhibits a martingale structure. Recall that the canonical functions g_1, \dots, g_m are completely degenerate, meaning that their conditional expectation with respect to any subset of their arguments is zero. Using these functions, we define partial sums $S_{n,c}$ that collect all terms of order c up to the n -th observation. We will see that, as the sample size increases, these partial sums form a martingale with respect to the natural filtration generated by the data. Intuitively, this means that adding new observations does not change the expected value of the statistic given the information available so far.

Let g_1, \dots, g_m be the canonical functions, so that $\mathbb{E}|g_c(X_1, \dots, X_c)| < \infty$ and

$$\mathbb{E}[g_c(X_1, \dots, X_c) | X_1, \dots, X_{c-1}] = 0 \quad \text{a.s. for } c = 1, \dots, m.$$

For $n \geq c$ recall that

$$S_{n,c} = \sum_{1 \leq i_1 < \dots < i_c \leq n} g_c(X_{i_1}, \dots, X_{i_c}), \quad c = 1, \dots, m,$$

and set the increasing σ -algebras

$$\mathcal{F}_k := \sigma(X_1, \dots, X_k), \quad k \in \mathbb{N}.$$

The following theorem states that this sequence forms a martingale.

Theorem 3.5.1 *For every $c \in \{1, \dots, m\}$ and integers $c \leq k \leq n$,*

$$\mathbb{E}[S_{n,c} | \mathcal{F}_k] = S_{k,c} \quad \text{a.s.}$$

Proof: By linearity of conditional expectation,

$$\mathbb{E}[S_{n,c} | \mathcal{F}_k] = \sum_{1 \leq i_1 < \dots < i_c \leq n} \mathbb{E}[g_c(X_{i_1}, \dots, X_{i_c}) | \mathcal{F}_k].$$

Split the sum into two parts: (i) indices entirely in $\{1, \dots, k\}$ and (ii) those for which at least one index exceeds k . Denote the corresponding index sets by

$$A_k := \{(i_1, \dots, i_c) : 1 \leq i_1 < \dots < i_c \leq k\}, \quad B_k := \{(i_1, \dots, i_c) : 1 \leq i_1 < \dots < i_c \leq n, \max i_j > k\}.$$

Terms with indices in A_k . If $(i_1, \dots, i_c) \in A_k$, then $g_c(X_{i_1}, \dots, X_{i_c})$ is \mathcal{F}_k -measurable, hence

$$\mathbb{E}[g_c(X_{i_1}, \dots, X_{i_c}) | \mathcal{F}_k] = g_c(X_{i_1}, \dots, X_{i_c}).$$

Summing over A_k gives

$$\sum_{(i_1, \dots, i_c) \in A_k} g_c(X_{i_1}, \dots, X_{i_c}) = S_{k,c}.$$

Terms with indices in B_k . Fix $(i_1, \dots, i_c) \in B_k$, so at least one index is $> k$. By symmetry of g_c , we may relabel the arguments so that the last index is $> k$; write

$$Y := g_c(X_{j_1}, \dots, X_{j_{c-1}}, X_{j_c}), \quad \text{with } j_c > k,$$

where $\{j_1, \dots, j_{c-1}\} \subset \{1, \dots, n\} \setminus \{j_c\}$. Apply the tower property with the larger σ -algebra $\mathcal{H} := \mathcal{F}_k \vee \sigma(X_{j_1}, \dots, X_{j_{c-1}})$:

$$\mathbb{E}[Y | \mathcal{F}_k] = \mathbb{E}[\mathbb{E}[Y | \mathcal{H}] | \mathcal{F}_k].$$

Conditionally on \mathcal{H} , the random variable X_{j_c} is independent of \mathcal{H} (because $j_c > k$ and the X_i are independent), and by the *complete degeneracy* property of g_c ,

$$\mathbb{E}[g_c(X_{j_1}, \dots, X_{j_{c-1}}, X_{j_c}) | X_{j_1}, \dots, X_{j_{c-1}}] = 0 \quad \text{a.s.}$$

Therefore $\mathbb{E}[Y | \mathcal{H}] = 0$ a.s., and hence $\mathbb{E}[Y | \mathcal{F}_k] = \mathbb{E}[0 | \mathcal{F}_k] = 0$ a.s. Summing over B_k yields

$$\sum_{(i_1, \dots, i_c) \in B_k} \mathbb{E}[g_c(X_{i_1}, \dots, X_{i_c}) | \mathcal{F}_k] = 0.$$

Combining the two parts,

$$\mathbb{E}[S_{n,c} | \mathcal{F}_k] = S_{k,c} + 0 = S_{k,c} \quad \text{a.s.}$$

This completes the proof of showing the martingale structure. \square

We now combine the previous results to express the centered U-statistic as a sum of martingale differences. This representation is particularly useful because it connects U-statistics to martingale theory, allowing us to use powerful probabilistic tools such as martingale convergence theorems and central limit results. Intuitively, each martingale difference represents the contribution of a new observation to the statistic after accounting for all information available up to the previous step. The following theorem formalizes this idea, corresponding to Example 2.3.7 for the case $Y_n = U_n$.

Theorem 3.5.2 (Centered U_n as a sum of martingale differences) Let $(X_i)_{i \geq 1}$ be i.i.d. and let $\Phi : \mathbb{R}^m \rightarrow \mathbb{R}$ be a symmetric kernel with $\mathbb{E}|\Phi(X_1, \dots, X_m)| < \infty$. Denote

$$U_n := \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} \Phi(X_{i_1}, \dots, X_{i_m}), \quad \Theta := \mathbb{E} \Phi(X_1, \dots, X_m).$$

Let g_1, \dots, g_m be the canonical functions, and for $c = 1, \dots, m$ and $n \geq c$ set

$$S_{n,c} := \sum_{1 \leq i_1 < \dots < i_c \leq n} g_c(X_{i_1}, \dots, X_{i_c}), \quad \eta_{kc} := \sum_{1 \leq i_1 < \dots < i_{c-1} \leq k-1} g_c(X_{i_1}, \dots, X_{i_{c-1}}, X_k).$$

Let $\mathcal{F}_k := \sigma(X_1, \dots, X_k)$. Define, for $k = 1, \dots, n$,

$$\xi_{nk} := \mathbb{E}[U_n | \mathcal{F}_k] - \mathbb{E}[U_n | \mathcal{F}_{k-1}].$$

Then $(\xi_{nk}, \mathcal{F}_k)_{k=1}^n$ is a martingale-difference sequence and

$$U_n - \Theta = \sum_{k=1}^n \xi_{nk}, \quad \xi_{nk} = \sum_{c=1}^m \binom{m}{c} \binom{n}{c}^{-1} \eta_{kc}. \quad (\star)$$

This is the same as in Example 2.3.8, for $Y_n = U_n$.

Proof: Set $M_k := \mathbb{E}[U_n | \mathcal{F}_k]$, $k = 0, \dots, n$. Then $\xi_{nk} = M_k - M_{k-1}$, so

$$\sum_{k=1}^n \xi_{nk} = M_n - M_0 = \mathbb{E}[U_n | \mathcal{F}_n] - \mathbb{E}[U_n | \mathcal{F}_0] = U_n - \mathbb{E}U_n = U_n - \Theta,$$

since U_n is \mathcal{F}_n -measurable and \mathcal{F}_0 is trivial. Moreover,

$$\mathbb{E}[\xi_{nk} | \mathcal{F}_{k-1}] = \mathbb{E}[M_k - M_{k-1} | \mathcal{F}_{k-1}] = \mathbb{E}[M_k | \mathcal{F}_{k-1}] - M_{k-1} = M_{k-1} - M_{k-1} = 0,$$

so $(\xi_{nk}, \mathcal{F}_k)$ is a martingale-difference sequence. It remains to compute ξ_{nk} explicitly.

By the Hoeffding representation (for expectations conditioned on the first k variables),

$$\mathbb{E}[U_n | \mathcal{F}_k] = \Theta + \sum_{c=1}^m \binom{m}{c} \binom{n}{c}^{-1} S_{k,c}, \quad \mathbb{E}[U_n | \mathcal{F}_{k-1}] = \Theta + \sum_{c=1}^{m-1} \binom{m}{c} \binom{n}{c}^{-1} S_{k-1,c}.$$

Subtracting gives

$$\xi_{nk} = \sum_{c=1}^m \binom{m}{c} \binom{n}{c}^{-1} (S_{k,c} - S_{k-1,c}).$$

By definition of $S_{k,c}$ and symmetry of g_c ,

$$S_{k,c} - S_{k-1,c} = \sum_{1 \leq i_1 < \dots < i_{c-1} \leq k-1} g_c(X_{i_1}, \dots, X_{i_{c-1}}, X_k) = \eta_{kc}.$$

Therefore

$$\xi_{nk} = \sum_{c=1}^m \binom{m}{c} \binom{n}{c}^{-1} \eta_{kc},$$

which is (\star) . \square

We will now show that a U-statistic is a reverse martingale. Intuitively, this means that as the sample size increases, the conditional expectation of the U-statistic given all future values (i.e., with less detailed information) equals the next statistic in the sequence. To formalize this idea, we introduce a decreasing filtration that captures the information contained in all U-statistics from a given index onward.

For the following theorem, set

$$\mathcal{B}_n := \sigma(U_n, U_{n+1}, U_{n+2}, \dots), \quad n \geq m.$$

Then $(\mathcal{B}_n)_{n \geq m}$ is a decreasing filtration: $\mathcal{B}_{n+1} \subseteq \mathcal{B}_n$.

Theorem 3.5.3 (U-statistics form a reverse martingale) *With the notation above, $(U_n, \mathcal{B}_n)_{n \geq m}$ is a reverse martingale:*

$$\mathbb{E}[U_n | \mathcal{B}_{n+1}] = U_{n+1} \quad \text{a.s. for all } n \geq m.$$

Proof: Integrability follows from the assumption $\mathbb{E}|\Phi(X_1, \dots, X_m)| < \infty$ and linearity of expectation; measurability $U_n \in \mathcal{B}_n$ is by definition. It remains to show the reverse martingale step.

Step 1. Fix $n \geq m$. Because Φ is symmetric and the X_i are i.i.d., for any index set $1 \leq i_1 < \dots < i_m \leq n$ we have

$$\mathbb{E}[\Phi(X_{i_1}, \dots, X_{i_m}) | \mathcal{B}_n] = \mathbb{E}[\Phi(X_1, \dots, X_m) | \mathcal{B}_n] \quad \text{a.s.} \quad (3.10)$$

Indeed, any permutation of (X_1, \dots, X_n) leaves the sequence (U_n, U_{n+1}, \dots) unchanged (U-statistics are symmetric), hence also leaves \mathcal{B}_n invariant; combined with the i.i.d. law of (X_i) , the conditional distributions agree and give (3.10).

Step 2 (Express U_n as a conditional expectation). Average (3.10) over all $\binom{n}{m}$ m -tuples $1 \leq i_1 < \dots < i_m \leq n$:

$$\mathbb{E}[U_n | \mathcal{B}_n] = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} \mathbb{E}[\Phi(X_{i_1}, \dots, X_{i_m}) | \mathcal{B}_n] = \mathbb{E}[\Phi(X_1, \dots, X_m) | \mathcal{B}_n].$$

Since U_n is \mathcal{B}_n -measurable, $\mathbb{E}[U_n | \mathcal{B}_n] = U_n$, hence

$$U_n = \mathbb{E}[\Phi(X_1, \dots, X_m) | \mathcal{B}_n] \quad \text{a.s.} \quad (3.11)$$

Step 3 (Reverse martingale step via tower property). Condition (3.11) further on \mathcal{B}_{n+1} and use $\mathcal{B}_{n+1} \subseteq \mathcal{B}_n$ (smaller σ -algebra wins) :

$$\mathbb{E}[U_n | \mathcal{B}_{n+1}] = \mathbb{E}[\mathbb{E}[\Phi(X_1, \dots, X_m) | \mathcal{B}_n] | \mathcal{B}_{n+1}] = \mathbb{E}[\Phi(X_1, \dots, X_m) | \mathcal{B}_{n+1}].$$

By the same argument as in Step 2 with n replaced by $n+1$,

$$\mathbb{E}[\Phi(X_1, \dots, X_m) | \mathcal{B}_{n+1}] = U_{n+1}.$$

Combining the last two displays yields $\mathbb{E}[U_n | \mathcal{B}_{n+1}] = U_{n+1}$ a.s., which proves the claim. \square

3.6 V-statistics

To conclude this chapter, we will see the definition of a V-statistic, named after von Mises, which will play a role in the next chapter on different MMD estimators.

Definition 3.6.1 (V-statistic (von Mises statistic)) *Let X_1, \dots, X_n be i.i.d. real-valued random variables on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$, and let $\Phi : \mathbb{R}^m \rightarrow \mathbb{R}$ be a symmetric, measurable kernel of degree $m \geq 1$. The V-statistic is defined as*

$$V_n(\Phi) := \frac{1}{n^m} \sum_{i_1=1}^n \cdots \sum_{i_m=1}^n \Phi(X_{i_1}, \dots, X_{i_m}). \quad (3.12)$$

Let $\Theta := \mathbb{E}\Phi(X_1, \dots, X_m)$ as before, then $V_n(\Phi)$ is a biased estimator of Θ , and $V_n(\Phi) \rightarrow \Theta$ in probability (and a.s. under some mild conditions). In contrast, the related Ustatistic averages only over distinct m -tuples.

To see that the V-statistic is biased, we look at a simple example:

Example 3.6.2 (Case $m = 2$) *Let $h(x, y) = \frac{1}{2}(x - y)^2$ be the kernel. Then*

$$\Theta := \mathbb{E}[h(X_1, X_2)] = \frac{1}{2} \mathbb{E}[(X_1 - X_2)^2] = \text{Var}(X_1),$$

as X_1 and X_2 are i.i.d..

The U- and V-statistics are

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} h(X_i, X_j), \quad V_n = \frac{1}{n^2} \sum_{i,j=1}^n h(X_i, X_j).$$

Since $h(x, x) = 0$, we have

$$V_n = \frac{1}{n^2} \sum_{i \neq j} h(X_i, X_j) = \frac{n-1}{n} U_n.$$

Taking expectations and using $\mathbb{E}[U_n] = \Theta$ (unbiasedness of U_n),

$$\mathbb{E}[V_n] = \frac{n-1}{n} \Theta = \left(1 - \frac{1}{n}\right) \text{Var}(X_1),$$

so V_n is biased downward with bias $-\text{Var}(X_1)/n$.

To understand the difference between U- and V-statistics, consider $n = 10$ and $m = 2$. We compare two equivalent U-statistic weightings and the V-statistic weighting.

U-statistics (degree 2). Let $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a *symmetric* kernel. There are two common (equivalent) summation schemes:

(i) *Ordered pairs (off-diagonal)*:

$$U_n^{\text{ord}} = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^n h(X_i, X_j). \quad (3.13)$$

(ii) *Unordered pairs (upper triangle)*:

$$U_n^{\text{unord}} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h(X_i, X_j). \quad (3.14)$$

Since h is symmetric, (3.13) and (3.14) are identical estimators: $U_n^{\text{ord}} = U_n^{\text{unord}}$. For $n = 10$, the ordered form averages over $10 \cdot 9 = 90$ off-diagonal pairs with weight $1/90$ each, while the unordered form averages over $\binom{10}{2} = 45$ pairs with weight $1/45$ each (each unordered pair counted once). So computationally, the unordered pairs version is more efficient, and thus widely used as for example, in the [BFM25] statistical test code implementation.

V-statistic (degree 2). The von Mises (V-) statistic includes the diagonal:

$$V_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h(X_i, X_j). \quad (3.15)$$

For $n = 10$, all $10^2 = 100$ pairs are included with equal weight $1/100$ (diagonal $i = j$ included). Compared to U_n , this adds diagonal terms and slightly down-weights each off-diagonal pair. The difference in weighting schemes between those three is shown in Figure 3.1. For each pair of i and j , we have a colour, which represents the weight of this pair. Teal corresponds to the equal weighting scheme. Purple means that this term is excluded from calculations, thus having a weight of zero. Yellow signals that the weight is above the equal weighting.

3 U-statistics

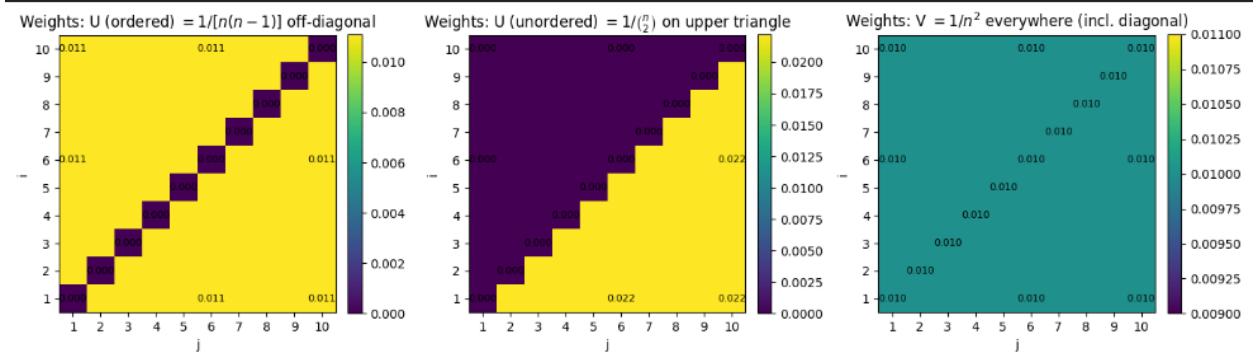


Figure 3.1 Comparison of weighting schemes between U-statistics and V-statistics

For each pair (i,j) , we now compute the kernel of x_i, x_j and sum all pairs up with their weighting. For each pair (i, j) we evaluate the kernel $h(x_i, x_j)$ and aggregate with weights w_{ij} :

$$\widehat{S} := \sum_{i=1}^n \sum_{j=1}^n w_{ij} k(x_i, x_j),$$

where the weighting scheme is

$$w_{ij} = \begin{cases} \frac{1}{n(n-1)} \mathbf{1}\{i \neq j\}, & \text{U-statistic (ordered),} \\ \binom{n}{2}^{-1} \mathbf{1}\{i < j\}, & \text{U-statistic (unordered),} \\ \frac{1}{n^2}, & \text{V-statistic.} \end{cases}$$

For the MMD, we will later see a V-statistic of degree $m=2$, which, as shown in the previous example, will be a biased MMD estimator.

4 Maximum Mean Discrepancy (MMD)

4.1 The notion of MMD and its connection to U-Statistics

A central application of U -statistics in modern statistics and machine learning is the computation of *maximum mean discrepancy* (MMD), a popular measure of discrepancy between probability distributions. MMD was first introduced by [Smo+07], and we will follow his approach. MMD is defined in terms of a *reproducing kernel Hilbert space* (RKHS) \mathcal{H} with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

Let P, Q be two Borel probability measures on \mathcal{X} . The mean embeddings of P and Q into the RKHS \mathcal{H} are defined as

$$\mu_P := \mathbb{E}_{X \sim P}[k(X, \cdot)], \quad \mu_Q := \mathbb{E}_{Y \sim Q}[k(Y, \cdot)].$$

The *maximum mean discrepancy* between P and Q is the RKHS distance of the mean embeddings:

$$\text{MMD}^2(\mathcal{H}, P, Q) := \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 = \mathbb{E}_{X, X' \sim P}[k(X, X')] + \mathbb{E}_{Y, Y' \sim Q}[k(Y, Y')] - 2 \mathbb{E}_{X \sim P, Y \sim Q}[k(X, Y)].$$

Connection to U-statistics. Given i.i.d. samples $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$, the expectations in the population MMD can be estimated unbiasedly by U -statistics. Indeed, define

$$\widehat{\text{MMD}}^2 := \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^n k(X_i, X_j) + \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^m k(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(X_i, Y_j).$$

Then each term is a (possibly two-sample) U -statistic with symmetric kernel:

- The first term is the U -statistic of degree-2 for P with kernel $k(x, x')$.
- The second term is the U -statistic of degree-2 for Q with kernel $k(y, y')$.
- The third term is a two-sample U -statistic of degree $(1, 1)$ with kernel $(x, y) \mapsto k(x, y)$.

Therefore, $\widehat{\text{MMD}}^2$ inherits the properties of U -statistics: it is unbiased, has a Hoeffding decomposition into canonical projections, and admits asymptotic normality under mild assumptions.

4.2 MMD estimators and important properties

We can now introduce different MMD estimators and derive some of their properties needed in Chapter 6. The estimators under examination here were first introduced by [Gre+12].

There are several finite-sample estimators of $\text{MMD}^2(P, Q)$. Let $\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P$ and $\{Y_j\}_{j=1}^m \stackrel{\text{i.i.d.}}{\sim} Q$, and let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be positive definite.

Let us start with the most common unbiased MMD estimator $\widehat{\text{MMD}}_u^2$.

Unbiased (U -statistic) estimator.

$$\widehat{\text{MMD}}_u^2 = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^n k(X_i, X_j) + \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^m k(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(X_i, Y_j). \quad (4.1)$$

It is unbiased, as it is written in the form of a U -statistic, thus: $\mathbb{E}[\widehat{\text{MMD}}_u^2] = \text{MMD}^2(P, Q)$.

Biased (V -statistic) estimator.

$$\widehat{\text{MMD}}_b^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k(X_i, X_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(X_i, Y_j). \quad (4.2)$$

Including the diagonal terms makes $\widehat{\text{MMD}}_b^2$ slightly biased upward (but often a bit lower variance and convenient computationally).

The z -estimator (diagonals removed everywhere). When $n = m$ and one wishes to exclude all self-similarity terms, use

$$\widehat{\text{MMD}}_z^2(P, Q) := \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^n \left\{ k(X_i, X_j) - 2k(X_i, Y_j) + k(Y_i, Y_j) \right\}. \quad (4.3)$$

Since $\{(X_i, Y_j) : i \neq j\}$ are still independent pairs, $\widehat{\text{MMD}}_z^2$ remains unbiased for $\text{MMD}^2(P, Q)$; it typically has slightly higher variance than (4.1) but excludes all diagonal contributions.

The following MMD estimator is a specific version of $\widehat{\text{MMD}}_u^2$, which is used in this form by [BFM25] and needed for Chapter 6.

Quadratic (block) MMD. Assume n is even and set $m := n/2$. Group adjacent observations into pairs $((X_{2i-1}, Y_{2i-1}), (X_{2i}, Y_{2i}))$ for $i = 1, \dots, m$. Define, for $i \neq j$,

$$q((X_{2i-1}, Y_{2i-1}), (X_{2i}, Y_{2i}); (X_{2j-1}, Y_{2j-1}), (X_{2j}, Y_{2j})) := k(X_{2i-1}, X_{2j-1}) - k(Y_{2i}, X_{2j}) - k(X_{2i}, Y_{2j}) + k(Y_{2i-1}, Y_{2j-1}).$$

Define the paired variables

$$Z_i := ((X_{2i-1}, X_{2i-1}), (X_{2i}, Y_{2i})), \quad i = 1, \dots, m.$$

Then, by the definition of q , for $i \neq j$ we have

$$q(Z_i, Z_j) = k(X_{2i-1}, X_{2j-1}) - k(Y_{2i}, X_{2j}) - k(X_{2i}, Y_{2j}) + k(Y_{2i-1}, Y_{2j-1}).$$

Then the quadratic MMD (degree-2 U-statistic over pairs) is

$$\begin{aligned} \widehat{\text{MMD}}_q^2(P, Q) &= \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^m \left\{ k(X_{2i-1}, X_{2j-1}) - k(Y_{2i}, X_{2j}) - k(X_{2i}, Y_{2j}) + k(Y_{2i-1}, Y_{2j-1}) \right\} \\ &= \frac{1}{m(m-1)} \sum_{i \neq j} q(Z_i, Z_j) \end{aligned}$$

We will now establish important properties of the MMD estimators, which will be used in Chapter 6. The proofs use the U-statistics theory discussed in Chapter 3.

Lemma 4.2.1 (Degeneracy of the $\widehat{\text{MMD}}_z^2$ estimator under H_0) Let (X_1, \dots, X_n) and (Y_1, \dots, Y_n) be i.i.d. samples from P and Q , respectively, and assume the null hypothesis $H_0 : P = Q$. Consider the MMD estimator

$$\widehat{\text{MMD}}_z^2(P, Q) = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^n \left\{ k(X_i, X_j) - 2k(X_i, Y_j) + k(Y_i, Y_j) \right\}. \quad (4.4)$$

Then $\widehat{\text{MMD}}_z^2$ is a U-statistic of degree 2 whose kernel

$$h((x, y), (x', y')) := k(x, x') - k(x, y') - k(y, x') + k(y, y')$$

is degenerate under H_0 , i.e.

$$\mathbb{E}[h((X_1, Y_1), (X_2, Y_2)) \mid (X_1, Y_1)] = 0 \quad a.s.$$

Proof: Define the paired variables

$$Z_i := (X_i, Y_i), \quad i = 1, \dots, n,$$

so that $(Z_i)_{i=1}^n$ are i.i.d. from $P \times P$.

Then we can rewrite the statistic as

$$\widehat{\text{MMD}}_z^2 = \frac{1}{n(n-1)} \sum_{i \neq j} h(Z_i, Z_j),$$

so $\widehat{\text{MMD}}_z^2$ is a U -statistic of degree 2 with kernel h and i.i.d. inputs Z_1, \dots, Z_n .

To show degeneracy under the null, we must show that the first Hoeffding projection of h vanishes almost surely, i.e.

$$\mathbb{E}[h(Z_1, Z_2) | Z_1 = z_1] = 0 \quad \text{a.s.}$$

Fix $z_1 = (x, y)$ and let $Z_2 = (X', Y')$ be an independent copy, with $X', Y' \sim P$ independent. Then

$$\mathbb{E}[h((x, y), (X', Y')) | (x, y)] = \mathbb{E}[k(x, X')] - \mathbb{E}[k(x, Y')] - \mathbb{E}[k(y, X')] + \mathbb{E}[k(y, Y')].$$

Under H_0 , $X' \sim P$ and $Y' \sim P$ have the same distribution, so for any t

$$\mathbb{E}_{U \sim P}[k(t, U)]$$

is the same whether U is written X' or Y' . Define

$$\mu_k(t) := \mathbb{E}_{U \sim P}[k(t, U)].$$

Then the conditional expectation becomes

$$\mu_k(x) - \mu_k(x) - \mu_k(y) + \mu_k(y) = 0.$$

Thus

$$\mathbb{E}[h(Z_1, Z_2) | Z_1 = z_1] = 0 \quad \text{a.s.},$$

which shows that the kernel h is *degenerate* under the null. \square

Lemma 4.2.2 (Nondegeneracy of the $\widehat{\text{MMD}}_q^2$ estimator under H_0) Assume $H_0 : P_1 = P_2 =: P$, let n be even and set $m := n/2$. Let (X_1, \dots, X_n) and (Y_1, \dots, Y_n) be i.i.d. samples from P . Group adjacent observations into pairs

$$Z_i := ((X_{2i-1}, Y_{2i-1}), (X_{2i}, Y_{2i})), \quad i = 1, \dots, m.$$

For $i \neq j$, define the symmetric kernel

$$q(Z_i, Z_j) := k(X_{2i-1}, X_{2j-1}) - k(Y_{2i}, X_{2j}) - k(X_{2i}, Y_{2j}) + k(Y_{2i-1}, Y_{2j-1}),$$

where k is a symmetric positive definite kernel such that the map $t \mapsto \mathbb{E}_{U \sim P}[k(t, U)]$ is nonconstant (e.g. k Gaussian and P non-atomic).

Then the quadratic MMD estimator

$$\widehat{\text{MMD}}_q^2 = \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^m q(Z_i, Z_j)$$

is a U -statistic of degree 2 whose kernel q is nondegenerate under H_0 , i.e.

$$\mathbb{P}(\mathbb{E}[q(Z_i, Z_j) | Z_i = z_i] = 0) < 1.$$

4 Maximum Mean Discrepancy (MMD)

Proof: By construction, under H_0 we have $X_j \sim P$, $Y_j \sim P$, all mutually independent. Since the indices in $Z_i = ((X_{2i-1}, Y_{2i-1}), (X_{2i}, Y_{2i}))$ are disjoint across i , the random elements $(Z_i)_{i=1}^m$ are i.i.d. from $(P \times P)^2$. Moreover, for $i \neq j$ we have

$$q(Z_i, Z_j) = k(X_{2i-1}, X_{2j-1}) - k(Y_{2i}, X_{2j}) - k(X_{2i}, Y_{2j}) + k(Y_{2i-1}, Y_{2j-1}),$$

so

$$\widehat{\text{MMD}}_q^2 = \frac{1}{m(m-1)} \sum_{i \neq j} q(Z_i, Z_j),$$

which shows that $\widehat{\text{MMD}}_q^2$ is a degree-2 U -statistic with kernel q and i.i.d. inputs Z_1, \dots, Z_m .

To study (non)degeneracy, we compute the first Hoeffding projection. Fix

$$z = ((x_1, y_1), (x_2, y_2))$$

and let

$$Z' = ((X_3, Y_3), (X_4, Y_4))$$

be an independent copy, with $X_3, X_4, Y_3, Y_4 \sim P$ i.i.d. Then

$$\mathbb{E}[q(Z, Z') | Z = z] = \mathbb{E}[k(x_1, X_3)] - \mathbb{E}[k(y_2, X_4)] - \mathbb{E}[k(x_2, Y_4)] + \mathbb{E}[k(y_1, Y_3)].$$

Under H_0 , X_3, X_4, Y_3, Y_4 all have the same distribution P , so for any t define

$$\mu_k(t) := \mathbb{E}_{U \sim P}[k(t, U)].$$

Then

$$\mathbb{E}[q(Z, Z') | Z = z] = \mu_k(x_1) - \mu_k(x_2) + \mu_k(y_1) - \mu_k(y_2).$$

Hence, for the random blocks Z_i ,

$$\mathbb{E}[q(Z_i, Z_j) | Z_i = z_i] = (\mu_k(X_{2i-1}) - \mu_k(X_{2i})) + (\mu_k(Y_{2i-1}) - \mu_k(Y_{2i}))$$

where

$$z_i = ((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i})).$$

By assumption on k (e.g. Gaussian kernel) and on P (non-atomic / nondegenerate), the map $t \mapsto \mu_k(t)$ is nonconstant. Therefore $\mu_k(X_{2i-1})$ and $\mu_k(X_{2i})$ are i.i.d. nonconstant random variables, so their difference

$$\mu_k(X_{2i-1}) - \mu_k(X_{2i})$$

is not almost surely zero; the same holds for the Y -part. Thus the sum of these two independent differences is also not almost surely zero, i.e.

$$\mathbb{P}(\mathbb{E}[q(Z_i, Z_j) | Z_i = z_i] = 0) < 1.$$

This is precisely the nondegeneracy condition for the kernel q . □

5 Case study: A high-frequency time series

To understand the intuition behind the notion of MMD, we will now look at an example from high-frequency financial data. In particular, we employ the LOBSTER dataset and gratefully acknowledge the permission granted to use it in this study. We use the first 5000 entries of the limit order book data from the 2nd of January 2015 for three different stocks. The limit order book data at each time step includes multiple levels of bid and ask prices with their respective volumes. For this extensive example, we want to calculate the returns of the time series from three different stocks over the same period in time. To calculate the returns, we first have to calculate an approximate fair value at each point in time from the limit order book data. The traditional approach would be using the midprice, but here we will propose our own multi-level microprice. Generally speaking, a microprice is a better approximation of the theoretical price of that asset in the current time step than the midprice, as it uses market microstructure information about order book imbalance. We now introduce our own generalisation of the microprice to multiple levels, being a more stable and robust estimate of a fair value.

5.1 The multi-level microprice and its applications

Microprice-based estimators play a central role in high-frequency trading and market making. The standard top-of-book microprice (as in [CJP15] on page 18 for an explicit estimator, or in [Sto17], where a function must be estimated from data) reacts only to the best bid and ask sizes, which may be susceptible to fleeting orders and spoofing. By incorporating additional depth information, a multi-level estimator yields a fair value that is less sensitive to short-lived quote updates while remaining highly responsive to genuine shifts in supply and demand.

Such fair value estimates are used in:

- **Market Making:** Centering bid and ask quotes around the microprice and adding inventory-dependent skew improves execution quality and inventory control.
- **Optimal Execution:** Deciding whether to join, improve, or step back from the touch can be guided by the position of the microprice relative to the mid-price.
- **Predictive Signals:** Deviations between microprice and mid-price often correlate with short-term price movements, producing alpha signals.

Our main goal is a generalisation of the top of the book microprice proposed by [CJP15] (page 18) to multiple levels.

Classical microprice. Let B_1 and A_1 denote the best bid and ask prices with volumes V_1^b and V_1^a . The classical microprice is

$$\text{MP}_{\text{TOB}} = \frac{V_1^a B_1 + V_1^b A_1}{V_1^a + V_1^b}. \quad (5.1)$$

This construction weights each side's price by the *opposite* side's volume, reflecting the probability that the next trade occurs at that price. When the ask side is heavy, MP_{TOB} tilts toward B_1 , anticipating a likely downward trade. However, (5.1) ignores liquidity beyond the best level. Large resting volume one or two ticks away often stabilises prices and should shift fair value in its direction.

Our extension of the top of the book microprice to multiple levels is based on the idea that the further away a quote is from the top of the book, the less important the information of this quote. There are two

approaches to constructing such an estimator: either we take the distance from the midprice of a quote, or we separate it into distance from the best bid price for quotes on the bid side and distance from the best ask price for quotes on the ask side. As our goal is to achieve a generalisation of the classical microprice, we chose the latter. After deciding which of the two approaches someone chooses, we must decide on the weighting scheme. For this, we took an exponential decay, as the further away we are from the best price on this side the less important our information is. We will now formally introduce our multi-level microprice, which is a true generalisation of the classical microprice, i.e. if we only take one level into consideration, no matter the parametrisation, we will get the same result as the classical microprice.

Multi-level extension. Consider L bid levels (B_ℓ, V_ℓ^b) and L ask levels (A_ℓ, V_ℓ^a) with tick size $\tau > 0$. We define exponential distance-based weights with a tuning parameter α

$$w_\ell^b := e^{-\alpha d_\ell^b}, \quad w_\ell^a := e^{-\alpha d_\ell^a}, \quad \alpha > 0,$$

where d_ℓ^b and d_ℓ^a denote the number of ticks between the best quote and level ℓ on the bid/ask side. The effective volumes \tilde{V}^b, \tilde{V}^a and prices \tilde{B}, \tilde{A} are then

$$\begin{aligned} \tilde{V}^b &:= \sum_{\ell=1}^L w_\ell^b V_\ell^b, & \tilde{V}^a &:= \sum_{\ell=1}^L w_\ell^a V_\ell^a, \\ \tilde{B} &:= \frac{\sum_{\ell=1}^L w_\ell^b B_\ell V_\ell^b}{\tilde{V}^b}, & \tilde{A} &:= \frac{\sum_{\ell=1}^L w_\ell^a A_\ell V_\ell^a}{\tilde{V}^a}. \end{aligned}$$

The *multi-level, distance-weighted microprice* is

$$\text{MP}_{\text{multi}} := \frac{\tilde{B}\tilde{V}^a + \tilde{A}\tilde{V}^b}{\tilde{V}^a + \tilde{V}^b}. \quad (5.2)$$

Equation (5.2) reduces to the classical form (5.1) if $L = 1$ and $w_1^b = w_1^a = 1$. Optionally, the estimator may be clamped to $[B_1, A_1]$ to guarantee it lies inside the spread.

Interpretation. Equation (5.2) maintains the probabilistic interpretation of the microprice: $\tilde{V}^a/(\tilde{V}^a + \tilde{V}^b)$ can be seen as the probability of the next trade hitting the bid, while $\tilde{V}^b/(\tilde{V}^a + \tilde{V}^b)$ is the probability of hitting the ask. The use of \tilde{B} and \tilde{A} shifts the reference prices to account for nearby liquidity.

Example comparison. Consider an order book with tick size $\tau = 0.1$ and spread $A_1 - B_1 = 0.2$:

$$\begin{array}{ll} B_1 = 100.0, & V_1^b = 200, \\ A_1 = 100.2, & V_1^a = 50, \end{array} \quad \begin{array}{ll} B_2 = 99.9, & V_2^b = 500, \\ A_2 = 100.3, & V_2^a = 50. \end{array}$$

With $\alpha = 1.0$, the distances are $d_1^b = 0, d_2^b = 1$ giving $w_1^b = 1, w_2^b = e^{-1}$, and similarly for the ask side. We obtain

$$\tilde{V}^b \approx 383.9, \quad \tilde{V}^a \approx 68.4, \quad \tilde{B} \approx 99.95, \quad \tilde{A} \approx 100.23.$$

Substituting into (5.2) yields

$$\text{MP}_{\text{multi}} \approx 100.1853.$$

For comparison, the classical microprice gives 100.16 and the mid-price is 100.10. The multi-level microprice lies closer to the ask, reflecting strong bid-side depth, as shown in Figure 5.1. The Figure shows a snapshot of a limit order book for this specific example, with the bid volumes in blue bars and ask volumes in red bars. It clearly demonstrates that the large level two bid volume should influence the fair value estimation and thus shift it towards the currently quoted best ask price.

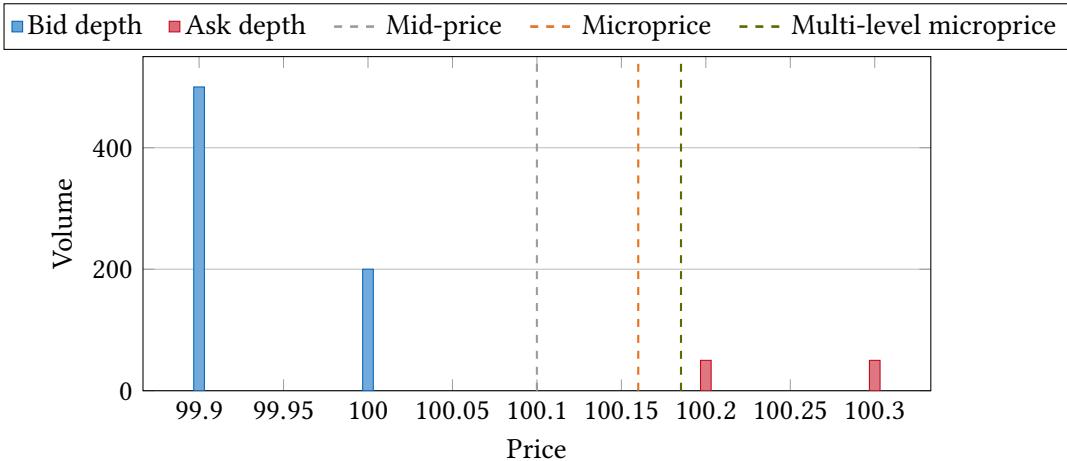


Figure 5.1 Comparison of midprice, microprice, and multi-level microprice.

In this example, if we were a market maker, we would most likely improve the bid to 100.1 or join at 100 and step back on the ask side (assuming we don't hold any inventory).

5.2 Practical considerations

The proposed multi-level microprice depends on several modelling choices that influence its behaviour in practice. In the following, we highlight the most important considerations.

Decay parameter. The tuning parameter $\alpha > 0$ determines how quickly the influence of deeper levels decays with tick distance. For small values of α , the estimator approaches an unweighted average over all visible depth, incorporating substantial information from several levels of the order book. Conversely, large values of α place almost exclusive weight on the best bid and ask, so that MP_{multi} behaves similarly to the classical top-of-book microprice. The tuning of this parameter could be based on the amount of volume quoted and its distribution or some liquidity estimate.

Truncation and stability. In empirical applications, it is neither necessary nor desirable to include all available levels. Truncating the calculation at $L = 3\text{--}5$ levels is typically sufficient, as more distant quotes contribute negligibly once exponentially discounted. For stability, levels with zero volume can be dropped, and in situations where the effective depth $\tilde{V}^a + \tilde{V}^b$ is too small, one may revert to the mid-price as a safeguard.

Reduction property. A desirable feature of the construction is its reduction to the classical microprice. When only the best levels are considered ($L = 1$) and the weights are set to one, the estimator MP_{multi} coincides exactly with MP_{TOB} . This guarantees consistency with the established definition while extending it to incorporate additional liquidity.

Scale invariance. Measuring distances in tick units ensures that the estimator is robust to changes in the nominal price level of the asset. This scale invariance is crucial for practical use, as it allows the estimator to be applied uniformly across instruments with different price ranges and tick sizes.

5.3 MMD comparison

We can now turn back to the goal of making MMD intuitive. For this, we take the LOBSTER data of three different stocks and compute their returns of their multi-level microprices. We normalise the time series by dividing it with price at time zero and multiply it by 100. We only consider the first 5000 events of the limit order book data per stock. The normalised time series of their multi-level microprices can be found in Figures 5.2-5.4, for the stocks TSLA (Tesla), INTC (Intel) and PCLN (Booking).

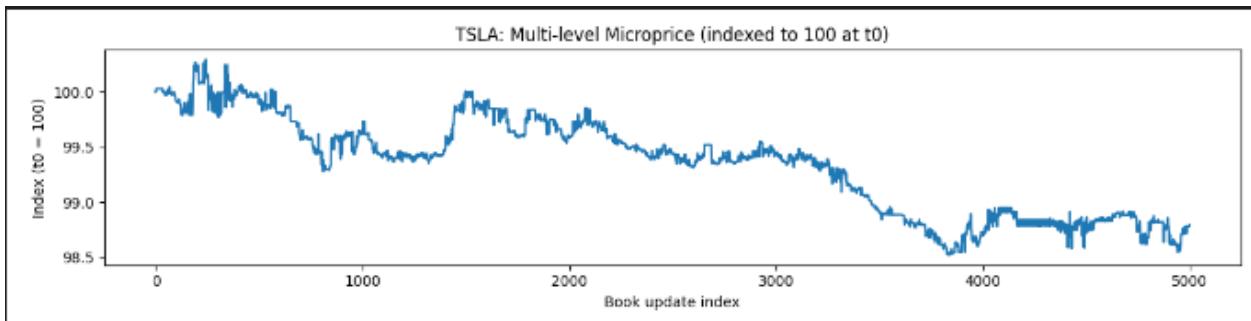


Figure 5.2 Normalised TSLA time series

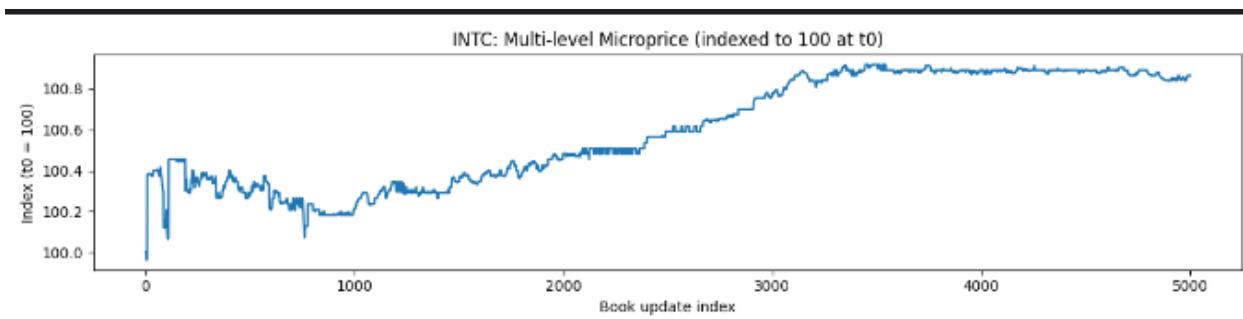


Figure 5.3 Normalised INTC time series

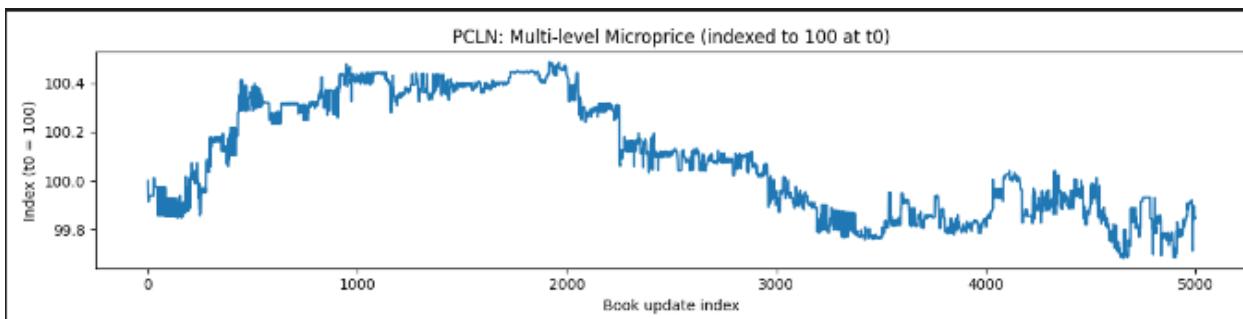


Figure 5.4 Normalised PCLN time series

Second, we plot the distributions of the returns (see Figure 5.5). From the distribution plot, we would expect that the distributions of returns of TSLA (blue) and PCLN (green) are more similar to each other than either of them to INTC. As expected, most of the returns are highly concentrated around 0 due to the usage of high frequency data.

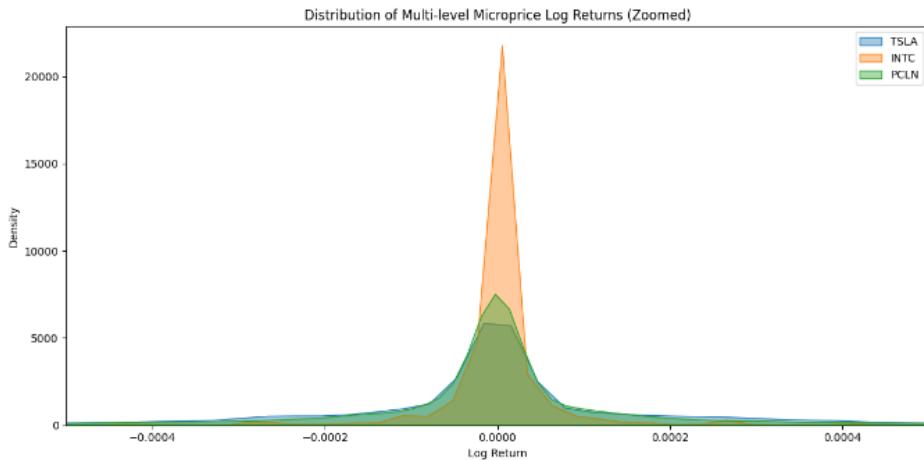


Figure 5.5 Distributions of log-returns for the three stocks

And lastly, we calculate the MMD between all pairs of stocks. Table 5.1 confirms our intuition that the distributions of TSLA and PCLN are more similar to each other. Thus, MMD measures the distance/similarity between distributions and a higher MMD value means that the distributions have a more distinct distributional divergence.

Table 5.1 MMD output of cross comparisons between stock returns

| Stock Pair | MMD [multi-level micro returns] |
|-----------------|---------------------------------|
| MMD(TSLA, INTC) | 0.319589 |
| MMD(TSLA, PCLN) | 0.0482841 |
| MMD(INTC, PCLN) | 0.275695 |

6 Application: An MMD-based statistical test

In the paper of [BFM25], they introduce an elegant MMD-based statistical test for model selection and specification. In this chapter, we will discuss the workings of their test, which is an application of U-statistics theory. We will skip the proofs, use the results discussed in Chapter 4 and focus more on intuition. At the end of the chapter, the results of several Monte-Carlo simulations will be presented. Those will verify the empirical properties of the Brueck-Fermanian-Min (BFM) test.

6.1 Motivation

The goal of the BFM test statistic is to decide, given an underlying data set X and two additional data sets Y_1 and Y_2 , which of the data sets Y_1 or Y_2 is more similar to X . Alternatively, when provided with only a single data set Y , the goal is to decide whether Y is similar to X .

The latter case is known as the *model specification* problem. Here, the null hypothesis is that X and Y are drawn from the same distribution, i.e.

$$H_{0,\mathcal{M}} : \text{MMD}(P_{\alpha_\star}, P) = 0,$$

where P is the underlying distribution of X , \mathcal{M} denotes a model (a family of probability measures), and P_{α_\star} is the distribution of Y under the fitted parameter α_\star .

This corresponds to the setting of model specification testing in the BFM framework: if $H_{0,\mathcal{M}}$ is rejected, then X and Y are deemed too different and the model is rejected. If $H_{0,\mathcal{M}}$ is not rejected, we conclude that the model provides an adequate fit.

The second case is known as the *model selection* problem. Here, two candidate models (or data sets) Y_1 and Y_2 are available, and we wish to determine which is closer to X . Let Q_{β_\star} denote the distribution of Y_2 and set $Y := Y_1$ for notational simplicity. Then the null hypothesis is

$$H_{0,\mathcal{M}_1, \mathcal{M}_2} : \text{MMD}(P_{\alpha_\star}, P) = \text{MMD}(Q_{\beta_\star}, P).$$

In other words: Y_1 and Y_2 are equally close to X in terms of the MMD, meaning that there is not enough evidence to prefer one model over the other. Rejecting $H_{0,\mathcal{M}_1, \mathcal{M}_2}$ indicates that one of the two candidate models provides a significantly better fit to the data X .

The BFM test statistic is highly relevant for applications such as generative machine learning, transfer learning, Bayesian statistics, clustering, and adaptive MCMC methods.

6.2 The BFM test statistic

To understand what the BFM test statistic tries to accomplish, we first examine its components and the issues it addresses. By standard U-statistic theory, it can be shown (see [Gre+12]) that under the null hypothesis,

$$n \widehat{\text{MMD}}^2(P_1, P_2) \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i (\chi_i^2 - 2),$$

where $\chi_i \sim \mathcal{N}(0, 2)$ and the λ_i are (possibly infinitely many) eigenvalues associated with the functional equation

$$\mathbb{E}[(k(X, y) - \mu_{P_1}(X) - \mu_{P_1}(y) + \mathbb{E}_Y[\mu_{P_1}(Y)])\psi(X)] = \lambda\psi(y), \quad y \in \mathcal{S}.$$

The problem with this asymptotic distribution is that estimating the eigenvalues λ_i is either too complex or computationally infeasible in practice. Hence, tests based directly on $n \widehat{\text{MMD}}^2(P_1, P_2)$ are difficult to implement. Another approach would be to use a pre-test to verify whether $P_{\alpha_\star} = Q_{\beta_\star} = P$, but this introduces additional layers of testing and new sources of error.

The BFM paper solves this issue by constructing a test statistic that:

- avoids pre-testing,
- is asymptotically distribution-free,
- and converges to a standard normal distribution under H_0 .

The resulting statistic is much simpler to compute and, importantly, its asymptotic law is not influenced by parameter estimation (e.g. the choice of α_\star that minimizes some custom loss function). This was a major issue in earlier works, such as the test of [Gre+06], which requires independent samples from P_{α_n} to compute $\widehat{\text{MMD}}^2(P_{\alpha_n}, P)$. In practice, P_{α_n} is fitted using the same sample from P , making independence impossible. This means that $n \widehat{\text{MMD}}^2(P_{\alpha_n}, P)$ does *not* have the same asymptotic distribution as $n \widehat{\text{MMD}}^2(P_{\alpha_\star}, P)$, because parameter estimation introduces additional variability.

The main idea of their test statistic is taking a linear combination of the standard MMD (meaning $\widehat{\text{MMD}}_z$) and $\widehat{\text{MMD}}_q$, as introduced in chapter 4. It follows that $\widehat{\text{MMD}}_q^2(P_1, P_2)$ is an unbiased estimator of $\text{MMD}^2(P_1, P_2)$ and by U-statistic theory it follows that

$$\sqrt{n} \left\{ \widehat{\text{MMD}}_q^2(P_1, P_2) - \text{MMD}^2(P_1, P_2) \right\} \xrightarrow{d} \mathcal{N}(0, \sigma_q^2),$$

with $\sigma_q^2 > 0$, essentially if and only if P_1 or P_2 is not a Dirac probability measure.

Some might think that this implies that a test should be based only on $\widehat{\text{MMD}}_q$, but since it does not use all pairs of observations there will be a power loss. That is why a linear combination of the two is introduced – to approximately keep the power of $\text{MMD}^2(P_1, P_2)$. Thus, they define

$$\widehat{\text{MMD}}_{\epsilon_n}^2(P_1, P_2) := \widehat{\text{MMD}}^2(P_1, P_2) + \epsilon_n \widehat{\text{MMD}}_q^2(P_1, P_2),$$

with weights $\epsilon_n > 0$.

If $\epsilon_n := \epsilon > 0$ is a constant, it is obvious that

$$\sqrt{n} \widehat{\text{MMD}}_{\epsilon_n}^2(P_1, P_2) \xrightarrow{d} \epsilon \mathcal{N}(0, \sigma_q^2) \quad \text{under } H_0,$$

since $\sqrt{n} \widehat{\text{MMD}}^2(P_1, P_2) \xrightarrow{p} 0$ when $P_1 = P_2$. However, choosing a fixed $\epsilon_n = \epsilon > 0$ may lead to a power loss, similar to a test based only on $\sqrt{n} \widehat{\text{MMD}}_q^2(P_1, P_2)$. Therefore, they impose that $\epsilon_n \rightarrow 0$ in probability.

Thus, for the model specification test, they define the test statistic

$$T_n(\mathcal{M}, P) := \frac{\sqrt{n} \widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P)}{\widehat{\sigma}_n},$$

for some sequence of parameters $(\alpha_n)_{n \geq 1}$ that weakly converges to α_\star at rate $n^{-1/2}$ and $\widehat{\sigma}_n$ is the estimator of the standard deviation. In the next section, it will be stated under which conditions this test statistic will converge to a standard normal distribution under the null hypothesis.

In the case of model selection, the test for

$$\mathcal{H}_{0, \mathcal{M}_1, \mathcal{M}_2} : \text{MMD}(P_{\alpha_\star}, P) = \text{MMD}(Q_{\beta_\star}, P)$$

is based on

$$T_n(\mathcal{M}_1, \mathcal{M}_2, P) := \frac{\sqrt{n} \left(\widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_{\epsilon_n}^2(Q_{\beta_n}, P) \right)}{\widehat{\tau}_n},$$

where

$$\widehat{\tau}_n^2 = \widehat{\tau}_n^2(\epsilon_n, P_{\alpha_n}, Q_{\beta_n}, P)$$

denotes a natural estimator of the asymptotic variance of

$$\sqrt{n} \left\{ \widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_{\epsilon_n}^2(Q_{\beta_n}, P) \right\}.$$

Choice of Kernel. Throughout the BFM test, the kernel k is chosen to be the Gaussian kernel

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right),$$

where $\sigma > 0$ is the bandwidth parameter. The Gaussian kernel is characteristic, which ensures that the $\text{MMD}(P, Q)$ is zero if and only if $P = Q$, making it a suitable choice for two-sample testing.

6.3 Central limit theorem

For the asymptotic behaviour of the BFM test statistic in the case of model specification, [BFM25] show the following result.

Theorem 6.3.1 Assume $\epsilon_n \rightarrow 0$, $\epsilon_n \sqrt{n} \rightarrow \infty$ in probability, $\sqrt{n}(\alpha_n - \alpha_\star) = O_p(1)$ and that some technical assumptions as in the BFM paper hold.

1. If $P = P_{\alpha_\star}$, i.e. under $\mathcal{H}_{0,\mathcal{M}}$, we have

$$\mathcal{T}_n(\mathcal{M}, P) = \frac{\sqrt{n} \widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P)}{\widehat{\sigma}_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1),$$

where $\widehat{\sigma}_n = \widehat{\sigma}_{\alpha_n} + \epsilon_n \widehat{\sigma}_{q,\alpha_n}$.

2. If $P \neq P_{\alpha_\star}$, i.e. if $\mathcal{H}_{0,\mathcal{M}}$ is false, then $\mathcal{T}_n(\mathcal{M}, P) \rightarrow \infty$ in probability.

This result shows that the proposed test statistic converges to a standard normal distribution under the null and diverges under the alternative. Estimates for the variance will be discussed in the next section.

A similar result is obtained for the model selection case.

Theorem 6.3.2 Assume again $\epsilon_n \rightarrow 0$, $\epsilon_n \sqrt{n} \rightarrow \infty$ in probability, $\alpha_\star \in \arg \min_{\alpha \in \Theta_1} \text{MMD}^2(P_\alpha, P)$ and $\beta_\star \in \arg \min_{\beta \in \Theta_2} \text{MMD}^2(Q_\beta, P)$. Let $\sqrt{n}(\alpha_n - \alpha_\star) = O_p(1)$ and $\sqrt{n}(\beta_n - \beta_\star) = O_p(1)$, assume that the samples $(U_i)_{i \geq 1}$ and $(V_i)_{i \geq 1}$ are independent, and that the same technical assumptions as in the BFM paper hold. Then, under $\mathcal{H}_{0,\mathcal{M}_1, \mathcal{M}_2}$:

$$\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P) = \frac{\sqrt{n} \left(\widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_{\epsilon_n}^2(Q_{\beta_n}, P) \right)}{\widehat{\tau}_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1),$$

where $\widehat{\tau}_n = \widehat{\sigma}_{\alpha_n, \beta_n} + \epsilon_n \widehat{\sigma}_{q,\alpha_n, \beta_n}$.

Algorithmic implementation. The BFM paper also provides those explicit algorithms for performing these tests:

Algorithm 1: MMD-based test of $\mathcal{H}_{0,\mathcal{M}} : \text{MMD}(P_{\alpha_*}, P) = 0$

- Requirements:** I.i.d. sample $(X_i)_{1 \leq i \leq n}$ from P , generative model $F(U; \alpha) \sim P_\alpha$, estimator α_n of α_* , tuning parameter ϵ_n and confidence level γ .
- 1 Sample $(F(U_i, \alpha_n))_{1 \leq i \leq n}$, where $(U_i)_{1 \leq i \leq n} \stackrel{i.i.d.}{\sim} P_U$;
 - 2 Compute $\mathcal{T}_n(\mathcal{M}, P) = \sqrt{n} \widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P) / (\tilde{\sigma}_{\alpha_n} + \epsilon_n \tilde{\sigma}_{q, \alpha_n})$ according to (3), (10) and (11);
 - 3 Reject $\text{MMD}(P_{\alpha_*}, P) = 0$ when $|\mathcal{T}_n(\mathcal{M}, P)| > \Phi^{-1}(1 - \gamma/2)$; otherwise, accept.
-

Figure 6.1 Algorithm for model specification

Algorithm 2: MMD based test of $\mathcal{H}_{0,\mathcal{M}_1,\mathcal{M}_2} : \text{MMD}(P_{\alpha_*}, P) = \text{MMD}(Q_{\beta_*}, P)$

- Requirements:** I.i.d. sample $(X_i)_{1 \leq i \leq n}$ from P , generative models $F(U; \alpha) \sim P_\alpha$ and $G(V; \beta) \sim Q_\beta$, estimator α_n of $\arg\min_{\alpha \in \Theta_1} \text{MMD}(P_\alpha, P)$, estimator β_n of $\arg\min_{\beta \in \Theta_2} \text{MMD}(Q_\beta, P)$, tuning parameter ϵ_n and confidence level γ .
- 1 Sample $(F(U_i, \alpha_n))_{1 \leq i \leq n}$ and $(G(V_i, \beta_n))_{1 \leq i \leq n}$, where $(U_i)_{1 \leq i \leq n} \stackrel{i.i.d.}{\sim} P_U$ and $(V_i)_{1 \leq i \leq n} \stackrel{i.i.d.}{\sim} P_V$ are independent;
 - 2 Compute $\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P) = \sqrt{n} \{ \widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_{\epsilon_n}^2(Q_{\beta_n}, P) \} / (\tilde{\sigma}_{\alpha_n, \beta_n} + \epsilon_n \tilde{\sigma}_{q, \alpha_n, \beta_n})$ according to (3), (16) and (17);
 - 3 Reject $\text{MMD}(P_{\alpha_*}, P) = \text{MMD}(Q_{\beta_*}, P)$ when $|\mathcal{T}_n(\mathcal{M}_1, \mathcal{M}_2, P)| > \Phi^{-1}(1 - \gamma/2)$; otherwise, accept.
-

Figure 6.2 Algorithm for model selection

6.4 Estimation of variance

The goal is to obtain an estimator of the asymptotic variance of $\sqrt{n} \widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P)$. To achieve this, we split it into two estimations: one of $\sqrt{n} \widehat{\text{MMD}}_{\epsilon_n}^2(P_{\alpha_n}, P)$ and one of $\sqrt{n} \widehat{\text{MMD}}_q^2(P_{\alpha_n}, P)$, which are then combined.

It is well known (e.g. in [Ser80]) that the asymptotic variance of $\sqrt{n} \{ \widehat{\text{MMD}}^2(P_\alpha, P) - \text{MMD}^2(P_\alpha, P) \}$ is given by

$$\sigma_\alpha^2 := \text{Var}(2h(X, F(U; \alpha); \alpha)).$$

The empirical estimate of this variance is

$$\tilde{\sigma}_\alpha^2 := \frac{4}{n} \sum_{i=1}^n \left\{ \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n h((X_i, F(U_i; \alpha)), (X_j, F(U_j; \alpha))) - \widehat{\text{MMD}}^2(P_\alpha, P) \right\}^2.$$

Since our goal is to estimate $\sigma_{\alpha_*}^2$ but α_* is unknown, we replace α_* with α_n and define

$$\tilde{\sigma}_{\alpha_n}^2 := \frac{4}{n} \sum_{i=1}^n \left\{ \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n h((X_i, F(U_i; \alpha_n)), (X_j, F(U_j; \alpha_n))) - \widehat{\text{MMD}}^2(P_{\alpha_n}, P) \right\}^2.$$

For the other asymptotic variance of the MMD_q -part, we introduce the shorthand notation

$$q([x, u]_{1:4}; \alpha) := k(x_1, x_3) - k(x_4, F(u_2; \alpha)) - k(x_2, F(u_4; \alpha)) + k(F(u_1; \alpha), F(u_3; \alpha)).$$

Thus, the asymptotic variance of $\sqrt{n}\{\widehat{MMD}_q^2(P_\alpha, P) - MMD^2(P_\alpha, P)\}$ is given by

$$\sigma_{q,\alpha}^2 := \text{Var}(2\sqrt{2} \mathbb{E}_{[X,U]_{3,4}}[q([X, U]_{1:4}; \alpha)]).$$

An empirical estimator of $\sigma_{q,\alpha_\star}^2$ is defined as

$$\widehat{\sigma}_{q,\alpha_n}^2 := \frac{16}{n} \sum_{i=1}^{n/2} \left\{ \frac{1}{n/2-1} \sum_{\substack{j=1 \\ j \neq i}}^{n/2} q([X, U]_{2i-1, 2i; 2j-1, 2j}; \alpha_n) - \widehat{MMD}^2(P_{\alpha_n}, P) \right\}^2.$$

Notice that both estimators are always non-negative and have computational complexity $O(n^2)$. Furthermore, $\sigma_{\alpha_\star}^2 = 0$ when $P_{\alpha_\star} = P$, but $\sigma_{q,\alpha_\star}^2$ is strictly positive under the assumptions of the BFM paper. Finally, the BFM test statistic combines these two estimators as

$$\widehat{\sigma}_n = \widehat{\sigma}_{\alpha_n} + \epsilon_n \widehat{\sigma}_{q,\alpha_n}.$$

6.5 Simulation Study

In this section, we investigate the finite-sample performance of the BFM test for model specification and model selection. Since the theoretical results from the previous section are asymptotic, a Monte Carlo study is useful to verify how well they hold in practice for moderate sample sizes. We focus on the empirical level and power of the test statistic $T_n(\mathcal{M}, P)$.

We replicated the simulation study done by [BFM25] and implemented it using GPU to speed up the process. We will recreate Figures 1-4 of [BFM25].

6.5.1 Monte Carlo Setup for Figure 1

For Figure 1, we simulate from the following setting. Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P = \mathcal{N}(0, I_p)$ be p -dimensional standard normal random variables. Define the model

$$Y(\alpha) = Y + \alpha, \quad Y \sim \mathcal{N}(0, \sigma^2 I_p),$$

where $\alpha \in \mathbb{R}^p$ is an unknown shift parameter, and by construction $Y(\alpha) \sim P_\alpha$. If $\alpha = 0$ and $\sigma^2 = 1$, then $P_\alpha = P$ and the null hypothesis $H_{0,\mathcal{M}}$ is true.

Each Monte Carlo replication proceeds as follows:

1. Generate a sample X_1, \dots, X_n from P .
2. Estimate the parameter α by the sample mean

$$\alpha_n := \frac{1}{n} \sum_{i=1}^n X_i.$$

3. Generate a second sample $Y_1(\alpha_n), \dots, Y_n(\alpha_n)$ from P_{α_n} .
4. Compute the BFM test statistic $T_n(\mathcal{M}, P)$.
5. Record whether $H_{0,\mathcal{M}}$ is rejected at the chosen significance level.

This procedure is repeated 1000 times. The proportion of rejections estimates the empirical level under the null and the empirical power under the alternatives. For Figure 1, we vary the variance parameter $\sigma \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$ to generate departures from $H_{0,\mathcal{M}}$ and illustrate how the test power increases with stronger deviations.

In addition to varying σ , we also investigate:

- the sample size $n \in \{100, 250, 500, 1000\}$,
- the dimension $p \in \{2, 16\}$,
- and the tuning parameter ϵ_n , choosing rates $\epsilon_n \in \{n^{-1/2.5}, n^{-1/4.5}, n^{-1/6.5}\}$.

These choices allow us to study how the test behaves as n grows, how it scales to higher dimensions, and how sensitive it is to the tuning parameter. We compare the BFM test with the classical test based solely on $\sqrt{n} \widehat{\text{MMD}}_q^2(P_{\alpha_n}, P)$ to evaluate whether the combination of MMD_z and MMD_q improves power without distorting the empirical level.

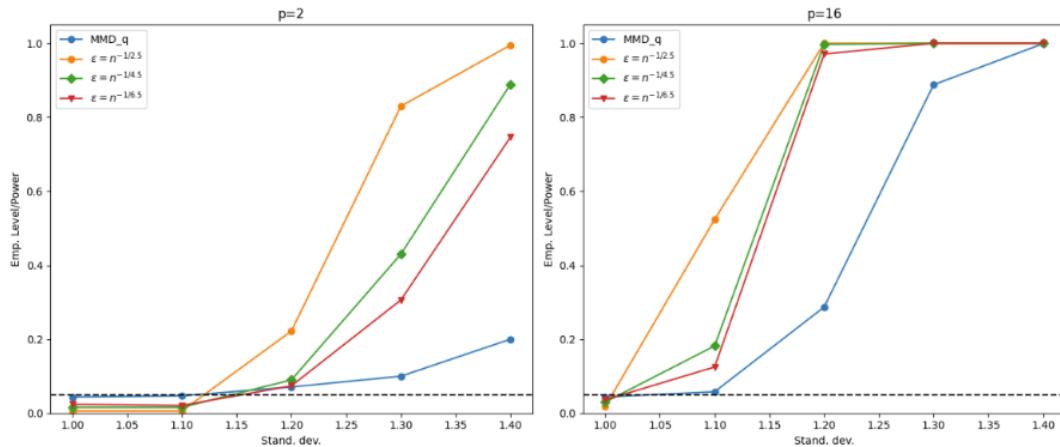


Figure 1: Empirical level and power of the tests based on $T_n(\mathcal{M}, P)$ and $\sqrt{n} \widehat{\text{MMD}}_q^2(P_{\alpha_n}, P)$ for dimensions $p = 2$ (left) and $p = 16$ (right), a sample size $n = 500$, as well as for different choices of ϵ_n (see the legend) and varying **standard deviation**. The rejection probabilities are estimated using 1000 replications of the tests based on samples of size n . The black dashed line indicates the significance level 0.05.

Figure 6.3 Replication of simulation study for BFM's Figure 1

6.5.2 Monte Carlo Setup for Figure 2

For Figure 2, we fix the tuning parameter at $\epsilon_n = n^{-1/2.5}$, as it gives the best power among the three, and vary the mean while also varying the sample size n . We consider the competing model

$$Y_1(\sigma) = \alpha_0 \mathbf{1} + \text{diag}(\sigma_1, \dots, \sigma_p) Y \sim P_\sigma, \quad Y \sim \mathcal{N}(0, I_p),$$

with pre-specified marginal mean $\alpha_0 \in \mathbb{R}$ and marginal standard deviations $\sigma = (\sigma_1, \dots, \sigma_p)$. If we fix $\alpha_0 = 0$, then the “optimal” parameters are $\sigma_1^\star = \dots = \sigma_p^\star = 1$ and $P = P_{\sigma^\star}$. We create alternatives by varying the mean

$$\alpha_0 \in \{0, 0.1, 0.2, \dots, 0.6\}.$$

For each replication we estimate the scale of each coordinate from the data by

$$\sigma_{j,n}^2 = \frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2, \quad \sigma_n := (\sigma_{1,n}, \dots, \sigma_{p,n}),$$

and then generate the model sample $(Y_1(\sigma_n), \dots, Y_n(\sigma_n))$ from P_{σ_n} . We always test the null $\mathcal{T}_n(\mathcal{M}, P)$ using the two samples (X_1, \dots, X_n) from P and $(Y_1(\sigma_n), \dots, Y_n(\sigma_n))$ from P_{σ_n} .

The number of Monte Carlo replications is 1000. We report the empirical rejection probabilities for

$$p \in \{2, 16\}, \quad n \in \{100, 250, 500, 1000\}, \quad \epsilon_n = n^{-1/2.5},$$

as a function of the mean shift α_0 . Across all settings, the BFM test based on $\mathcal{T}_n(\mathcal{M}, P)$ maintains its level and shows higher power than the test based solely on $\sqrt{n} \widehat{\text{MMD}}_q^2(P_{\sigma_n}, P)$, especially as n increases.

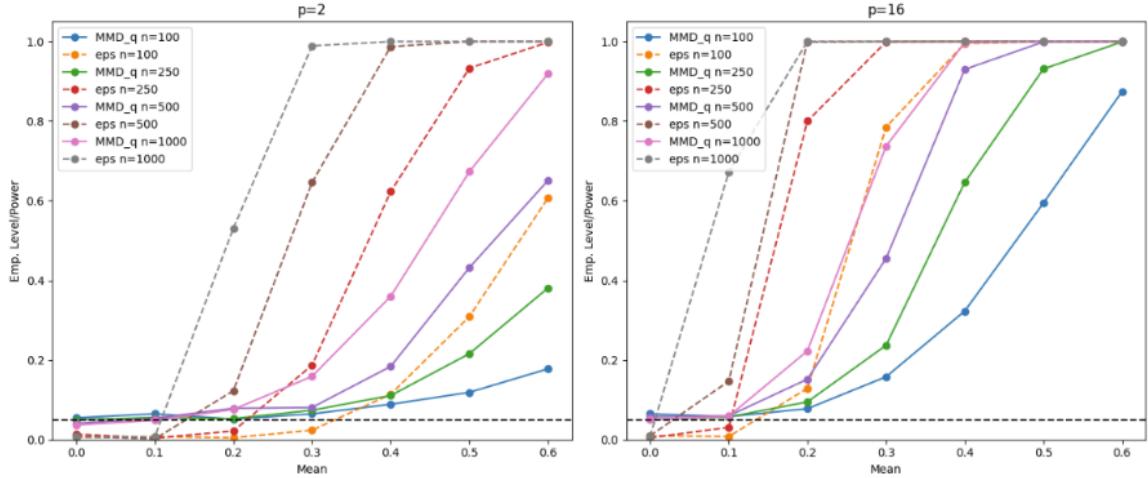


Figure 2: Empirical level and power of the tests based on $\mathcal{T}_n(\mathcal{M}, P)$ and $\sqrt{n} \widehat{\text{MMD}}_q^2(P_{\sigma_n}, P)$ for dimensions $p = 2$ (left) and $p = 16$ (right) as well as for sample sizes $n = 100, 250, 500, 1000$ (see the legend), $\epsilon_n = n^{-1/2.5}$ and varying mean . The rejection probabilities are estimated using 1000 samples of size n . The black dashed line indicates the significance level 0.05.

Figure 6.4 Replication of simulation study for BFM's Figure 2

6.5.3 Simulation for Figure 3: Model Selection (Degenerate Case)

For Figure 3, we investigate the finite-sample performance of the BFM test in a model comparison setting. We consider two competing parametric models \mathcal{M}_1 and \mathcal{M}_2 for the distribution of X . Assume p is even and let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P = \mathcal{N}(0, I_p)$.

The first model \mathcal{M}_1 is defined as

$$Y(\alpha) = Y + \alpha \sim P_\alpha, \quad Y \sim \mathcal{N}(0, \text{diag}(1, \dots, 1, \sigma^2, \dots, \sigma^2)),$$

where the first $p/2$ coordinates have variance 1 and the remaining $p/2$ coordinates have variance σ^2 . The “optimal” parameter α_* is zero, and thus \mathcal{M}_1 coincides with the true model when $\sigma^2 = 1$.

The second model \mathcal{M}_2 is given by

$$Z(\beta) = Z + \beta \sim Q_\beta, \quad Z \sim \mathcal{N}(0, I_p),$$

where $\beta \in \mathbb{R}^p$ is the parameter vector. In the degenerate setting, we set $\beta = 0$, so that \mathcal{M}_2 also coincides with the true model. Hence, both competing models are equally valid under the data-generating process.

In the simulation, we:

1. Generate n i.i.d. samples from P .
2. Estimate $\alpha_n = n^{-1} \sum_{i=1}^n X_i$ and β_n analogously.

3. Generate independent samples $Y_1(\alpha_n), \dots, Y_n(\alpha_n)$ from P_{α_n} and $Z_1(\beta_n), \dots, Z_n(\beta_n)$ from Q_{β_n} .
4. Compute the BFM model selection test statistic $T_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and record whether $H_{0, \mathcal{M}_1, \mathcal{M}_2}$ is rejected.

This procedure is repeated 1000 times to approximate the rejection probability. We vary the standard deviation $\sigma \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$, the dimension $p \in \{2, 16\}$, and the sample size $n \in \{100, 250, 500, 1000\}$. The tuning parameter is fixed to $\epsilon_n = n^{-1/2.5}$ as in the previous experiment.

The results, shown in Figure 3, confirm that all tests maintain the correct empirical level under $\sigma^2 = 1$ and show increasing power as σ^2 increases. As expected, the BFM test outperforms the benchmark based only on $\sqrt{n} \widehat{\text{MMD}}_q^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_q^2(Q_{\beta_n}, P)$, especially for larger sample sizes and higher dimensions.

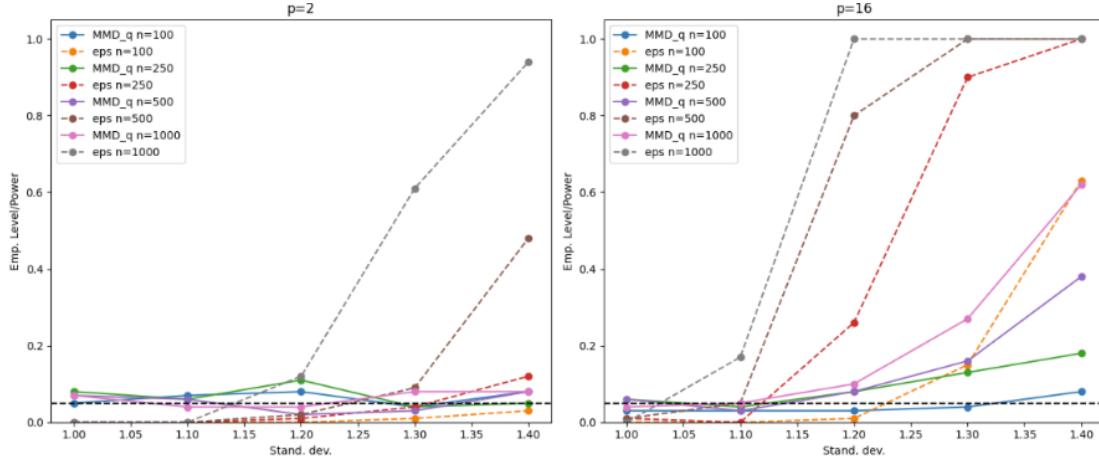


Figure 3: Degenerate case for comparison of two models: Empirical level and power of the tests based on $T_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and $\sqrt{n}(\widehat{\text{MMD}}_q^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_q^2(Q_{\beta_n}, P))$ for dimensions $p = 2$ (left) and $p = 16$ (right) as well as for sample sizes $n = 100, 250, 500, 1000$ (see the legend), $\epsilon_n = n^{-1/2.5}$ and varying **standard deviation** σ in Model \mathcal{M}_1 . Model \mathcal{M}_2 coincides with the true model ($\beta = 0$). The rejection probabilities are estimated using 1000 replications of the tests based on samples of size n . The black dashed line indicates the significance level 0.05.

Figure 6.5 Replication of simulation study for BFM's Figure 3

6.5.4 Simulation for Figure 4: Model Selection (Non-degenerate Case)

For Figure 4, we modify the model comparison setup to avoid the degenerate case. The competing models are now defined as

$$Y(\alpha) = Y + \alpha \sim P_\alpha, \quad Y \sim \mathcal{N}(0, \text{diag}(1.2^2, \dots, 1.2^2, \sigma^2, \dots, \sigma^2)),$$

and

$$Z(\beta) = Z + \beta \sim Q_\beta, \quad Z \sim \mathcal{N}(0, 1.2^2 I_p).$$

Thus, both \mathcal{M}_1 and \mathcal{M}_2 are equally far from the true distribution, except when $\sigma = 1.2$, where they coincide.

In the simulation, we:

1. Generate $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$.
2. Estimate $\alpha_n = n^{-1} \sum_{i=1}^n X_i$ and β_n analogously.
3. Generate independent samples $Y_1(\alpha_n), \dots, Y_n(\alpha_n) \stackrel{\text{i.i.d.}}{\sim} P_{\alpha_n}$ and $Z_1(\beta_n), \dots, Z_n(\beta_n) \stackrel{\text{i.i.d.}}{\sim} Q_{\beta_n}$.

4. Compute the BFM model selection test statistic $T_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and record whether $H_{0, \mathcal{M}_1, \mathcal{M}_2}$ is rejected.

We repeat the procedure 1000 times for each setting, varying

$$\sigma \in \{1.2, 1.3, 1.4, 1.5, 1.6\}, \quad p \in \{2, 16\}, \quad n \in \{100, 250, 500, 1000\}, \quad \epsilon_n = n^{-1/2.5}.$$

Figure 4 shows that, in this non-degenerate case, the BFM test again outperforms the competitor based solely on $\sqrt{n}(\widehat{\text{MMD}}_q^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_q^2(Q_{\beta_n}, P))$, particularly for large n and higher p .

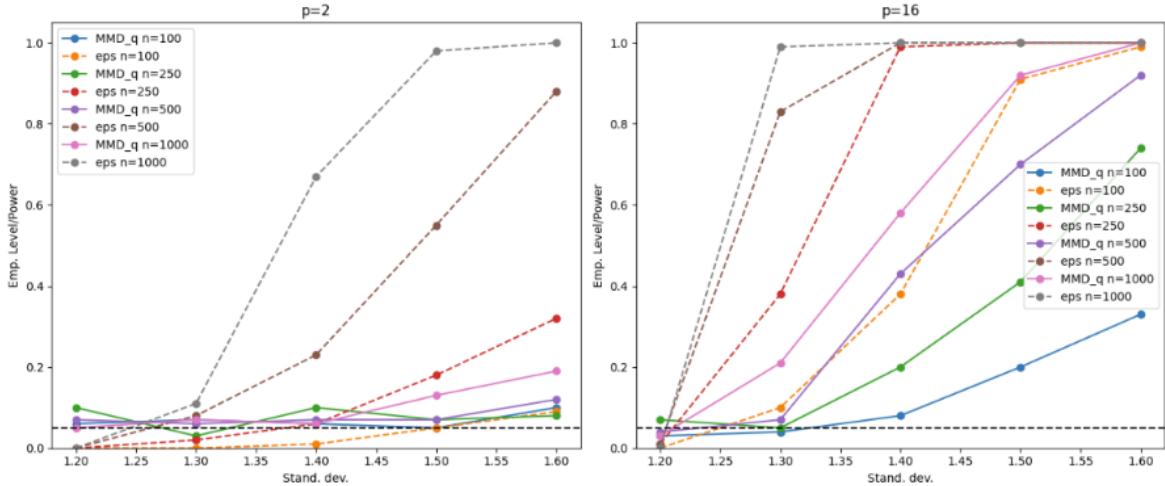


Figure 4: Non-degenerate case for comparison of two models: Empirical level and power of the tests based on $T_n(\mathcal{M}_1, \mathcal{M}_2, P)$ and $\sqrt{n}(\widehat{\text{MMD}}_q^2(P_{\alpha_n}, P) - \widehat{\text{MMD}}_q^2(Q_{\beta_n}, P))$ for dimensions $p = 2$ (left), $p = 16$ (right) as well as for sample sizes $n = 100, 250, 500, 1000$ (see the legend), $\epsilon_n = n^{-1/2.5}$ and varying **standard deviation** in Model \mathcal{M}_1 . Both models do not coincide with the true model. The rejection probabilities are estimated using 1000 replications of the tests based on samples of size n . The black dashed line indicates the significance level 0.05.

Figure 6.6 Replication of simulation study for BFM's Figure 4

7 Conclusion

This thesis provided a comprehensive and self-contained introduction to the theory of U-statistics, emphasizing their role as unbiased and minimum-variance estimators and illustrating their practical significance through the maximum mean discrepancy (MMD). We gave an intuition-based introduction to the fundamental principles of measure-theoretic probability theory, such as probability spaces, random variables and expected values. We introduced conditional expectations and its properties and martingales, which are crucial for developing the theoretical foundation required to analyze the properties and structure of U-statistics.

In Chapter 3, we deep dived into parametric functionals and their estimator, the U-statistic. We saw that U-statistics arise naturally in statistics, e.g. when we want to estimate the mean, the sample mean can be written as such. The theory is developed for kernels with symmetric arguments, but this is not a limitation of the theory, as the expected value of a non-symmetric kernel is the same value as the symmetrized kernel. Additionally, we restricted ourselves to U-statistics of one sample and the real numbers being the target space of the kernel. We introduced the canonical functions and the property of complete degeneracy, which were important tools to prove the martingale structure of U-statistics. We derived the Hoeffding representation to be able to write every U-statistic as a linear combination of degenerate U-statistics.

In Chapter 4, we connected the estimator of the maximum mean discrepancy (MMD) with the U-statistics, thus all of the theory is applicable to prove asymptotic results. We looked at several MMD estimators, including one biased estimator, and proved the needed results for Chapter 6. The MMD is a modern and heavily used estimator in machine learning and several applications as for example, the model specification test by [BFM25], which was introduced in Chapter 6.

In Chapter 5, we gave an intuitive example to understand the MMD measure in the context of financial data. For this example, we proposed our own weighting scheme for a multi-level microprice, which is a generalisation of the classical microprice. We applied it to the LOBSTER data set to compare the returns of three different stocks and measure the distance between the return distributions.

In Chapter 6, we discussed a modern goodness-of-fit test by [BFM25], which is based on a linear combination of two MMD estimators and is an application of U-statistics theory. It provides a hypothesis test for model specification and model selection, which works for independent and identically distributed samples. For the BFM test, we also replicated their Monte Carlo simulation study using GPU-accelerated code to achieve a performance improvement of over 4200 times and confirm that the asymptotic results of the test also hold for moderate sample sizes.

We now give an outlook into further research areas in U-statistics-based hypothesis tests. Here, we only presented the MMD-based hypothesis goodness-of-fit test, but there are multiple estimators for goodness-of-fit tests, which can be written as a U-statistic. For example, the kernel Stein discrepancy (KSD) is also used for goodness-of-fit tests. The KSD can also be estimated by a U-statistic, thus the U-statistics theory will be heavily used to prove results for KSD-based statistical tests. To our knowledge, the first theoretically valid composite goodness-of-fit test based on KSD is [BRB25]; Also, further research in the area for non-i.i.d. samples is needed, as right now, those tests can not handle financial time-series data. If those tests could be extended to time-series data, quantitative researchers could use them for model validation or selection and regime detection for example in order-flow, volatility or returns.

A Appendix

A.1 Supporting Data

For Chapter 5, we employed the LOBSTER dataset and thankfully acknowledge the permission to use it in this thesis. The dataset contains event-by-event data of the limit order book for three different stocks.

A.2 Code sources

The code to generate the figures for the BFM test described in Chapter 6 can be found at <https://github.com/fabianbaiertum/BFM-test>. For all other figures and the table, the code is available at <https://github.com/fabianbaiertum/bachelor-thesis>.

List of Figures

| | | |
|-----|---|----|
| 3.1 | Comparison of weighting schemes between U-statistics and V-statistics | 24 |
| 5.1 | Comparison of midprice, microprice, and multi-level microprice. | 31 |
| 5.2 | Normalised TSLA time series | 32 |
| 5.3 | Normalised INTC time series | 32 |
| 5.4 | Normalised PCLN time series | 32 |
| 5.5 | Distributions of log-returns for the three stocks | 33 |
| 6.1 | Algorithm for model specification | 38 |
| 6.2 | Algorithm for model selection | 38 |
| 6.3 | Replication of simulation study for BFM's Figure 1 | 40 |
| 6.4 | Replication of simulation study for BFM's Figure 2 | 41 |
| 6.5 | Replication of simulation study for BFM's Figure 3 | 42 |
| 6.6 | Replication of simulation study for BFM's Figure 4 | 43 |

List of Tables

| | |
|---|----|
| 5.1 MMD output of cross comparisons between stock returns | 33 |
|---|----|

Bibliography

- [ACH19] S. Arlot, A. Celisse, and Z. Harchaoui. “A Kernel Multiple Change-Point Algorithm via Model Selection”. In: *Journal of Machine Learning Research* 20 (2019), pp. 1–56.
- [BRB25] F. Brueck, V. Reimoser, and F. Baier. *Composite goodness-of-fit test with the Kernel Stein Discrepancy and a bootstrap for degenerate U-statistics with estimated parameters*. 2025. arXiv: 2510.22792 [math.ST].
- [BFM25] F. Brück, J. Fermanian, and A. Min. “Distribution Free Tests for Model Selection Based on Maximum Mean Discrepancy with Estimated Parameters”. In: *Journal of Machine Learning Research* (2025).
- [CJP15] Á. Cartea, S. Jaimungal, and J. Penalva. *Algorithmic and High-Frequency Trading*. Cambridge University Press, 2015.
- [Gre+06] A. Gretton et al. “A Kernel Method for the Two-Sample Problem”. In: *Advances in Neural Information Processing Systems*. Vol. 19. Curran Associates, Inc., 2006.
- [Gre+12] A. Gretton et al. “A Kernel Two-Sample Test”. In: *Journal of Machine Learning Research* 13 (2012), pp. 723–773.
- [Hoe48] W. Hoeffding. “A class of statistics with asymptotically normal distribution”. In: *Annals of Mathematical Statistics* 19.3 (1948), pp. 293–325.
- [KB13] V. Korolyuk and Y. Borovskich. *Theory of U-Statistics*. Springer Netherlands, 2013.
- [Kra23] F. Krahmer. *Probability Theory – Course Script (Winter Semester 2023/24)*. Lecture notes for the graduate course Probability Theory at TUM (WS 2023/24); access via TUM-Moodle. Technical University of Munich. 2023. URL: https://www.moodle.tum.de/pluginfile.php/4830986/mod_resource/content/0/skript-proba.pdf.
- [LLJ16] Q. Liu, J. Lee, and M. I. Jordan. “A Kernelized Stein Discrepancy for Goodness-of-Fit Tests”. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*. Vol. 48. Proceedings of Machine Learning Research. PMLR, 2016, pp. 276–284.
- [LP+15] D. Lopez-Paz et al. “Towards a Learning Theory of Cause-Effect Inference”. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*. PMLR, 2015, pp. 1452–1461.
- [Sej+14] D. Sejdinovic et al. “Kernel Adaptive Metropolis-Hastings”. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*. PMLR, 2014, pp. 1665–1673.
- [Ser80] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.
- [Smo+07] A. Smola et al. “A Hilbert Space Embedding for Distributions”. In: *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT 2007)*. Springer, 2007, pp. 13–31.
- [Sto17] S. Stoikov. *The Micro-Price: A High Frequency Estimator of Future Prices*. Available at SSRN. SSRN: <https://ssrn.com/abstract=2970694>, <http://dx.doi.org/10.2139/ssrn.2970694>. 2017.
- [Zha+11] K. Zhang et al. “Kernel-Based Conditional Independence Test and Application in Causal Discovery”. In: *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. AUAI Press, 2011, pp. 804–813.
- [Zho+20] Z. Zhou et al. “DC-MMDGAN: A New Maximum Mean Discrepancy Generative Adversarial Network Using Divide and Conquer”. In: *Applied Sciences* 10.9 (2020), p. 3107.