Diploma Thesis in Mathematics

Asymmetries of Time Series under Inverting their Direction

Jonas Peters
Ruprecht-Karls-Universität Heidelberg
Tübingen, August 2008

Prof. Dr. Rainer Dahlhaus

Ruprecht-Karls-Universität Heidelberg

PD Dr. Dominik Janzing

Universität Karlsruhe (TH)
MPI für biologische Kybernetik Tübingen

BIOLOGISCHE KYBERNETIK

Declaration:
Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, keine anderen als die angegebenen Hilfsmittel benutzt und alle Stellen, die dem Wortlaut oder Sinne nach anderen Werken entnommen sind, durch Angabe der Quellen als Entlehnungen kenntlich gemacht habe.

Title of the diploma thesis:
Asymmetries of Time Series under Inverting their Direction

Involved institutes:
Ruprecht-Karls-Universität Heidelberg
Max-Planck-Institut für biologische Kybernetik Tübingen

Full name and address:
Jonas Peters
Konrad-Adenauer-Str. 52 / 64.1
72072 Tübingen
Germany

jonas.peters@gmx.de

Signed:

Acknowledgement:

I thank PD Dr. Dominik Janzing for the mentoring at the MPI, many helpful discussions and that he always had time to answer my questions, Prof. Dr. Rainer Dahlhaus for the unproblematic mentoring in Heidelberg, which allowed me to work in Tübingen, Prof. Dr. Bernhard Schölkopf and Dr. Arthur Gretton for the discussions and Prof. Dr. Bernhard Schölkopf for giving me the opportunity to write my diploma thesis at the MPI Tübingen.

I thank my parents for supporting me during my whole studies.

Calvin: *Ah! I got the letter I wrote to myself!*
Hobbes: *What did you write?*
Calvin: *“Dear Calvin, Hi! I’m writing this on Monday. What day is it now?*
How are things going? Your pal, Calvin.”
Calvin: *My past self is corresponding with my future self.*
Hobbes: *Too bad you can’t write back.*

Bill Watterson, 19 April 1995

Contents

1 Introduction

1.1 Problem and Motivation

Consider the following problem: We are given m ordered values $X_1, \ldots, X_m$ from a time series, but we do not know if the sample has been reversed. Our task is to find out whether $X_1, \ldots, X_m$ or $X_m, \ldots, X_1$ represents the true time direction.

This problem regards the general difference in backward and forward going time and thus we cannot expect it to be easily solvable. There is a lot of literature about the asymmetries regarding the direction of time, whereas we will only mention shortly how the problem can be related to physics, causality and everyday life. The following ideas should rather be taken as thought impulses than as complete overviews of the fields.

Physics The question of the direction of time can be related in particular to the second law of thermodynamics. One possible formulation of the latter states that the entropy of a closed physical system can only increase but never decrease.[1] This may suggest to use the entropy criteria in the following way: for every time t, compute the entropy of the system, and propose the direction for which the entropy increases as the correct one. If you consider real-world time series from stock values, EEG data or geophysical data, for example, this method is not applicable. This is mainly due to two different reasons: firstly, for these time series the entropy is often very hard to recognize in the data and secondly, most of the observed time series can hardly be considered as closed systems.

Causality It is a basic principle that *every cause precedes its effect* or equivalently that the future cannot influence the past. There is a lot of philosophical work about the meaning of causality and its relation to time. Here we rather address the understanding of time and causality in everyday life; it is common sense that you cannot alter past events and that a car engine never starts *before* the key is turned.

Using the idea of the temporal ordering of cause and effect we can solve the time direction problem in the following way: if we identify one cause and its effect in the data $X_1, \ldots, X_m$, we can order the whole series. Say, we found that X_5 is causing X_2, then the true time ordering must be $X_m, \ldots, X_1$. See Figure 1.1 for an example[2]. In the following subsection we will introduce causality in a more formal way and make use of this approach in the ARMA method (see below).

[1] From a microphysical perspective the entropy is actually constant in time but only increases after appropriate *coarse-graining* the physical state space [1].

[2] Many thanks to my sister Mira for providing this sketch.

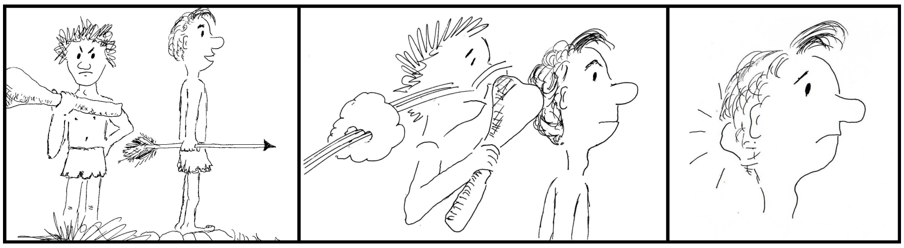


Figure 1.1: Since we can identify the blow with the club as the cause and the small bump as the effect, we can deduce that the pictures are in the correct ordering.

Everyday Life One of the time asymmetries in everyday life is that we seem to care more about the future than the past. We often regard something that will happen to us as more important than something similar that has already happened to us: We prefer a lot of unsatisfactory work to be done and the vacations to be ahead to the other way around. And as [2] mentions at first glance it is astonishing that we dread death, whereas we do not care too much about the non-existence before our birth.

Further, if we make decisions about our actions, we usually take into account (possible) events in the future rather than those in the past. It does not seem to be rational if someone acts in a special way in order to ensure the occurrence of a past event. This is surely related to the causal point of view stating that we can only change *future* events or cause them to occur.

Opposing to decision making our knowledge is focused on the past and the memory even contains *only* past events. We know much more about the past than about the future: It is easier to tell who won the last European Championship in soccer than to predict who will win the next one. Admittedly, there are a lot of periods and events from the past we do not know anything about and we are quite good in predicting the next total eclipse of the sun, for example. In general, however, our knowledge is biased towards the past.

1.2 Causality

We now formalize the concept of causality and relate it to the language of statistics. (Some of the following ideas can also be found in [3].) In many cases correlations or even dependencies beyond the second order do not give sufficient information about the relationship between two random variables. We are often interested in a deeper understanding of this relationship, namely we want to identify cause and effect. We can use standard statistical tools in order to detect a dependence between smoking habits and lung cancer, but it is more difficult to infer the causal structure. We suspect that smoking causes lung cancer, but how can we disprove that people with a higher chance of getting lung cancer (driven by a specific genotype, for example) feel a higher urge for smoking, too? For a long time, there was no formalism in classical statistics to deal with this sort of causal problems until Pearl [4] established formal notations and computation rules for the field of causality.

One possible definition of causality is the following: we say that a random variables X is causing a second variable Y if and only if Y can be written as a linear deterministic function $f(X)$ plus some noise, which is independent of X. This can be seen as a constraint on the joint distribution

$\mathbf{P}^{(X,Y)}$ of X and Y: Let ϵ be the residuum after computing a linear regression of Y on X. If X and ϵ are statistically independent (note that they are *uncorrelated* by construction), the joint distribution $P(X,Y)$ admits a linear model from X to Y. If X and Y are not independent it turns out (see Theorem 2.10) that the only case admitting a linear model in *both* directions is when $P(X,Y)$ is a bivariate Gaussian distribution.

Thus in this case causal inference (i.e. identifying cause and effect) can be done in the following way: Assume the linear model to be true and the noise to be non-Gaussian. Then consider the direction as causal that can better be fit by a linear model. This is the rationale behind LiNGAM [5] and also applies to causal inference with n variables $X_1, \ldots, X_n$ that linearly influence each other.

The problem of identifying cause and effect in a time series, is different from usual causal inference problems because of the following reasons (and therefore the conventional methods [4, 6] are not easily applicable): (1) The standard framework requires iid data of the joint distribution over all involved random variables, but not only single time instances. (2) For interesting classes of time series like MA and ARMA models (introduced in Section 4.2.1), the observed variables (X_t) are not causally sufficient since the (hidden) noise variables influence more than one of the observed variables. (3) Finite windows of observed variables are typically confounded by observable ancestors, which further complicates the problem. (4) as opposed to many real-world problems in causal inference we have at least *partial* knowledge of the ground truth [7]. This is an advantage because it makes it easier to evaluate our methods.

In [5] the authors applied their causal discovery algorithm LiNGAM to this problem. Their approach was able to propose a hypothetical time direction for 14 out of 22 time series (for the other cases their algorithm did not give a consistent result); however, only 5 out of these 14 directions turned out to be correct. The reasons for this could be the problems described below.

We have already seen how causality can help us to solve the problem of identifying the true time direction. And because of these differences from usual causal inference problems we conversely hope that studying the asymmetries of time series can also provide new insights for causal inference, too.

1.3 Proposed Methods

In this work we propose the following two methods for identifying the true time direction:

The SVM Method Consider a strictly stationary time series (that means the w-dimensional distribution of $(X_{t+h}, X_{t+1+h} \ldots, X_{t+w+h})$ does not depend on h for any choice of $w \in \mathbb{N}$). We assume the difference between the two different time directions to be a difference in the finite-dimensional distributions $(X_t, X_{t+1} \ldots, X_{t+w})$ and $(X_{t+w}, X_{t+w-1} \ldots, X_t)$. We try to learn the nature of this difference without further specifying it. For many time series we represent both distributions in a Hilbert Space (more specifically in an RKHS, see Section 3.1.2) and investigate if there are similarities between the difference of the forward and backward distributions. If this is the case, we learn these similarities using Support Vector Machines (SVMs) within this Hilbert Space.

The ARMA Method This method is based on the causality approach mentioned above. We assume the data to be an autoregressive moving average process (ARMA) with noise independent of the last values of the time series; together with this additional condition these time series are called *causal* ARMA processes.

As a main result of this paper we will show that the identifiability result from linear causal models extends to ARMA processes: if we assume that the time series is generated by a causal ARMA model with a non-vanishing AR part and with non-Gaussian noise, the process is not invertible; that means if the true direction follows a causal ARMA model with non-Gaussian noise, the other direction is not.

This result can be used in the following way: We fit the observed data to an ARMA model and test whether the regression residuals are statistically independent of the past values. Whenever the dependence in one direction is significantly weaker than in the other we infer the former to be the true one. To this end, we need a good dependence measure that is applicable to continuous data and finds dependencies beyond second order. In this work we use the Hilbert-Schmidt Independence Criterion (HSIC) [8] (see Section 3.3).

1.4 Outline

In Section 2 we formally introduce the concept of causality used here and show how linear causal models can be identified.

We define Reproducing Kernel Hilbert Spaces and Support Vector Machines in Section 3. As already mentioned before, we use the Hilbert-Schmidt Independence Criterion for independence testing, which we describe in this section, too.

In Section 4 we prove one of the main theorems — that non-Gaussian ARMA processes admit an ARMA model at most in one time direction— and we further explain the two different methods we employ for identifying the true time direction of time series data in more detail.

Section 5 contains results of our methods on both simulated and real data.

2 Causal Inference on Linear Models

2.1 Causal Models

As mentioned before the concept of statistical dependence between two random variables is often not sufficient. In many cases we assume a causal relationship between the two variables and need to know, which one is the cause and which one the effect. We know that the position of the earth relatively to the sun (which can be expressed in terms of an angle) and the temperature on the earth are strongly dependent. Without having defined the term causality yet, the position should cause the temperature and not vice versa. The goal of causal inference is to identify these sorts of causal relationship between random variables. For a long time, however, there was no formal mathematical framework for these questions. The following definition of causal models and their corresponding graphs is from Judea Pearl [4].

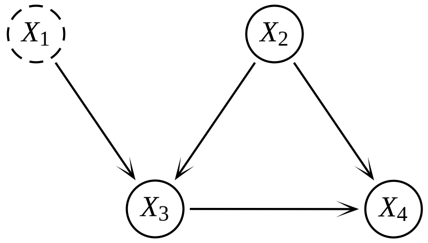
2.1.1 Definitions

Definition 2.1 [of a causal model] Let $X = \{X_1, \dots, X_n\}$ be a set of random variables and let $\mathrm{PA}_i \subset X$ be a subset of X (the *parents* of X_i). Assume that we only observe some of the variables X_i. The error variables ϵ_i represent errors due to omitted factors; they are always unobserved and independent of the variables PA_i. A set of equations

$$X_i = f_i(\mathrm{PA}_i, \epsilon_i) \quad \forall i = 1, \dots n, \qquad f_i \text{ belonging to some function class } \mathcal{F} \tag{2.1}$$

is called a *causal model* if they describe the process of generating the data. It is important to point out that this is not only a statement about probability distributions and conditional independencies, but also about the way, the process is generated in reality. We assume that the value of X_i is produced in the specific way of the equation above.

The corresponding *causal graph* is constructed by drawing a node for every random variable X_i and directed arrows from all its parents into this node. Nodes from unobserved variables are often drawn with a dashed line. The following is an example for a causal graph with $\mathrm{PA}_1 = \emptyset, \mathrm{PA}_2 = \emptyset, \mathrm{PA}_3 = \{X_1, X_2\}, \mathrm{PA}_4 = \{X_2, X_3\}$:



Definition 2.2 A causal model is called *Markovian* if its corresponding graph is a Directed Acyclic Graph (DAG) (i.e. it contains no cycles) and if all noise variables are jointly independent.

It can be shown ([4], Theorem 1.4.1) that a Markovian causal model satisfies the *Markov condition* meaning that every variable X_i is independent of all its non-descendants given its parents in the graph.
The Markov condition also exists in Bayesian Networks; these do not treat causality and therefore the Markov condition in causal models is sometimes called the *causal Markov condition*.

If further the joint distribution $\mathbf{P}^{(X_1,\dots,X_n)}$ is absolutely continuous with respect to a product measure, a Markovian causal model satisfies the factorization[1] ([9], Theorem 3.27):

$$p(x_1,\dots,x_n) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}_i)\,.$$

As said before a causal model satisfies more than just constraints on the probability distribution: If we do a hypothetical intervention on a parent on X_i, the probability distribution of X_i will change. We say that a parent effects its children. This is a condition, which cannot be written in terms of (conditional) probabilities, but deals with the generating process in reality.

2.1.2 What are causal models good for?

We want to give an idea, what causal graphs can be used for and how they can help to understand the data. We will demonstrate this by introducing Judea Pearl's *do*-notation. Out of reasons for simplicity we will only consider discrete random variables. Causal graphs provide a context in which a lot of causal problems can be formulated and solved. One of the most famous examples is the old debate about the causal relationship between smoking and lung cancer. The tobacco industry tried to explain the observed correlation between smoking and lung cancer by a genotype, which increases both the risk for getting cancer and the inborn craving for nicotine.
A powerful method to deal with these questions is the *do*-notation. When we $do(X_j = \tilde{x}_j)$ we set the variable X_j to $\tilde{x}_j$ while leaving all the other variables unchanged. Then we investigate how much this changes the distribution of another variable X_i. More formally:

Definition 2.3 Define the *causal effect* of X_j on $X_1,\dots,X_n$ to be the following distribution over $x_1,\dots,x_n$:

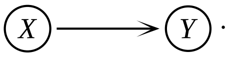
$$p(x_1,\dots,x_n \mid do(X_j = \tilde{x}_j)) := \prod_{i\neq j}^{n} p(x_i|\mathrm{PA}_i)\cdot\delta_{x_j,\tilde{x}_j}\,,$$

where $\delta_{x_j,\tilde{x}_j} = 1$ if $x_j = \tilde{x}_j$ and $\delta_{x_j,\tilde{x}_j} = 0$ otherwise.

As we see in the following example, this differs from the usual conditional distribution
$p(x_1,\dots,x_n \mid X_j = \tilde{x}_j) = p(x_1,\dots,x_n \mid \tilde{x}_j)$:

[1] $p(x_1,\dots,x_n)$ denotes the density with respect to the product measure mentioned above. This can be, for example, the probability density function or the probability mass function (or a combination of both) evaluated at $(x_1,\dots,x_n)$. We adapt to this notation because it is widely used in the domain of causal inference.

Example 2.4 Assume we have the following process:



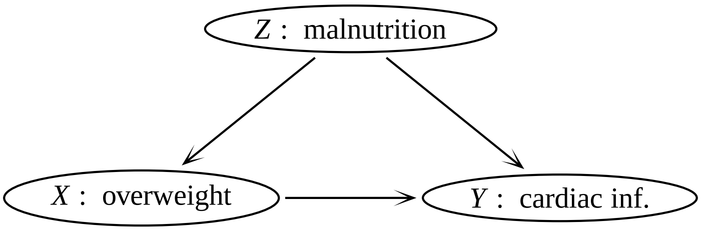
Then

$$p(y \mid do(X = \tilde{x})) = \sum_x p(x, y \mid do(X = \tilde{x})) = p(y \mid \tilde{x}),$$

$$\text{but} \qquad p(x \mid do(Y = \tilde{y})) = \sum_y p(x, y \mid do(Y = \tilde{y})) = p(x) \neq p(x \mid \tilde{y}).$$

It is important to realize, that the *do*-procedure is just a hypothetical intervention:

Example 2.5 It is widely believed that both malnutrition and overweight are risk factors for cardiac infarction. And certainly, malnutrition can also be a cause for overbalance. Therefore we can assume the following causal graph



We have $p(y \mid do(X = \tilde{x})) = \sum_z p(y \mid \tilde{x}, z)p(z)$. This shows that the *do* procedure should be interpreted as setting the variable X to $\tilde{x}$ hypothetically. We can hardly change the weight of a person by adding a few pounds in a surgery, for example. Instead we use the new distribution in order to determine the strength of the effect that overbalance has on the probability of getting a heart attack.

There are many other examples, where a change of variables is not only ethically irresponsible, but also physically impossible.

Notice that the equation for $p(y|do(X = \tilde{x}))$ in the last example shows that $do(X = \tilde{x})$ is equivalent to removing the arc between Z and X and setting X to the constant $\tilde{X}$. Instead of using Definition 2.3 we can generalise this idea and define the $(doX = \tilde{x})$-notation by removing every arc between the node X and its parents in the corresponding causal graph.[2]

Pearl provides a comprehensive theory for the *do*-calculus. He gives several rules for calculations within the *do*-framework, for example. Also, he develops a criterion in order to decide if a set of observed variables is sufficient in order to compute the causal effect of X on Y.

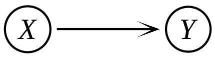
[2] This can be formalized by adding an additional parent node to X in the causal graph, which controls the function f in equation (2.1) and sets it to a constant $f \equiv \tilde{x}$ in the case of $do(X = \tilde{x})$. See Section 3.2.2 in [4] for more details.

2.2 Inferring causal graphs

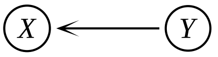
Usually causal graphs are a good possibility for including prior knowledge into data analysis. Note that this is conceptually different from putting prior distributions on parameters, which is done in Bayesian Statistics. Sometimes, however, we do not have this prior knowledge and then we want to infer the causal graph itself based on iid samples of the joint distribution $p(x_1, \dots, x_n)$. This is surely a hard problem and depending on the real generating process not even always possible. To make life easier we require the functions in equation (2.1) to be additive in the noise argument and linear in the parents of the node:

$$X_i = f_i(\mathrm{PA}_i, \epsilon_i) = g_i(\mathrm{PA}_i) + \epsilon_i \quad \forall i = 1, \dots n, \qquad g_i \text{ linear}.$$

The question arises, under which conditions we can distinguish between



and



.

Before we answer this question theoretically, we give two examples: in the first one both directions are possible, whereas in the second one $X \to Y$ is the true direction and the process cannot be reversed.

Example 2.6 (i) Assume

$$Y = \phi X + \epsilon, \quad \epsilon \perp\!\!\!\perp X\,, \tag{2.2}$$

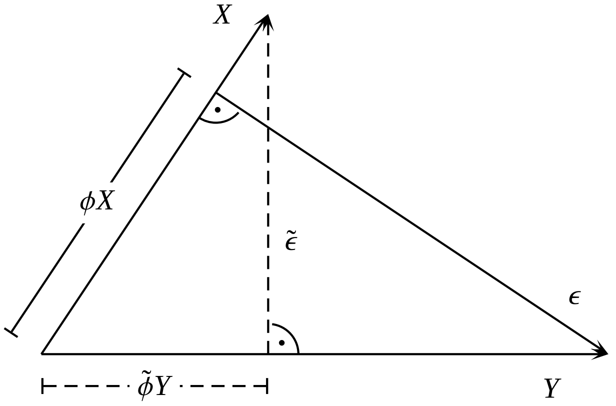
where X and ϵ are normally distributed. $\epsilon \perp\!\!\!\perp X$ means that ϵ and X are independent. Let σ^2 be the variance of the noise. Now we want to construct a new noise $\tilde{\epsilon}$, such that

$$X = \tilde{\phi} Y + \tilde{\epsilon}, \quad \tilde{\epsilon} \perp\!\!\!\perp Y\,. \tag{2.3}$$

Define $\mathbf{L}^2$ from the usual space $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbf{P})$ of square integrable random variables by identifying all random variables U, V, for which $U - \mathbf{E}U = V - \mathbf{E}V$ holds $\mathbf{P}$-almost surely. Then it is easy to see that

$$\langle U, V \rangle := \mathrm{cov}(U, V)$$

defines a dot product on $\mathbf{L}^2$. Therefore we can interpret (2.2) and (2.3) geometrically:



Now starting from X, Y and ϵ we construct $\tilde\epsilon$ by projecting X on the one dimensional subspace spanned by Y:

$$\tilde\epsilon = X - \frac{\langle X, Y\rangle}{||Y||^2} Y.$$

Then, by construction

$$X = \tilde\phi Y + \tilde\epsilon,$$

where $\tilde\phi := \frac{\langle X,Y\rangle}{||Y||^2}$. Because we used the projection, it is clear that $\tilde\epsilon$ and Y are uncorrelated:

$$\mathrm{cov}(\tilde\epsilon, Y) = \langle X, Y\rangle - \frac{\langle X, Y\rangle}{||Y||^2}\langle Y, Y\rangle = 0\,.$$

Since all distributions are Gaussian, $\tilde\epsilon$ is Gaussian, too and thus $\tilde\epsilon \perp\!\!\!\perp Y$. Furthermore we can determine $\tilde\phi$:

$$\tilde\phi = \frac{\phi \mathrm{var}(X)}{\phi^2 \mathrm{var}(X) + \sigma^2} = \frac{\phi}{\phi^2 + \sigma^2/\mathrm{var}(X)} \neq \frac{1}{\phi}\,.$$

(ii) Let X and ϵ be two iid random variables with distribution

$$\mathbf{P}(X = -0.5) = \mathbf{P}(X = 0.5) = 0.5\,,$$
$$\mathbf{P}(\epsilon = -0.5) = \mathbf{P}(\epsilon = 0.5) = 0.5\,.$$

Thus the variable

$$Y = X + \epsilon$$

has the distribution $\mathbf{P}(Y = -1) = \mathbf{P}(Y = 1) = 0.25$, $\mathbf{P}(Y = 0) = 0.5$. Assume that

$$X = bY + \tilde\epsilon\,,$$

for some $b \in \mathbb{R}$ and $\tilde\epsilon \perp\!\!\!\perp Y$. It is clear that $Y = 1$ implies $X = 0.5$ and $\tilde\epsilon = 0.5 - b$. On the other hand it follows from $Y = -1$ that $X = -0.5$ and $\tilde\epsilon = -(0.5 - b)$. Therefore Y and $\tilde\epsilon$ are not independent. More formally we have

$$\mathbf{P}(Y = 1) = \mathbf{P}(Y = 1) \cdot \mathbf{P}(\tilde\epsilon = 0.5 - b) \qquad \text{and}$$
$$\mathbf{P}(Y = -1) = \mathbf{P}(Y = -1) \cdot \mathbf{P}(\tilde\epsilon = -(0.5 - b))\,.$$

Thus $\mathbf{P}(\tilde{\epsilon} = -(0.5 - b)) = \mathbf{P}(\tilde{\epsilon} = 0.5 - b) = 1$, which implies $b = 0.5$ and $\tilde{\epsilon} \equiv 0$. Then $X = 0.5Y$, which is obviously a contradiction.

The intermediate result $b = 0.5$ in the second example is not surprising: the argumentation from the first example holds for the second example, too (projections are unique) and we get

$$b = \tilde{\phi} = \frac{\phi \cdot \mathrm{var}(X)}{\phi \cdot \mathrm{var}(X) + \mathrm{var}(\epsilon)} = \frac{1 \cdot 1/4}{1 \cdot 1/4 + 1/4} = \frac{1}{2},$$

which leads to uncorrelated noise. Because we do not have a Gaussian distribution we cannot deduce independence.

The normality assumption in the first example is not a coincidence, either. We will see that the Gaussian distribution is the only distribution, for which a linear causal relation (2.2) between X and Y can be reversed.

In the rest of the section we want to further investigate this special role of the Gaussian distribution. Therefore we need some auxiliary results. We first prove the following lemma, which is intuitively clear. The proof, however, has to be done carefully.

Lemma 2.7 *Let X and ϵ be two independent variables and assume ϵ to be non-deterministic. Then*

$$\epsilon \not\perp\!\!\!\perp (X + \epsilon)\,.$$

Proof Of course the proof becomes trivial if the variables have finite variance. Then $\mathrm{cov}(X, X + \epsilon) = \mathrm{var}(X) > 0$. For the general case, however, the argumentation is a bit more complex. Assume $\epsilon \perp\!\!\!\perp (X + \epsilon)$. Then for every $u, v \in \mathbb{R}$:

$$\begin{aligned}
\varphi_{(\epsilon, X+\epsilon)}(u, v) &= \mathbf{E}\left[\exp(iu\epsilon + iv\epsilon + ivX)\right] \\
&= \mathbf{E}\left[\exp(iu\epsilon + iv\epsilon) \cdot \exp(ivX)\right] \\
&= \mathbf{E}\left[\exp(iu\epsilon + iv\epsilon)\right] \cdot \mathbf{E}\left[\exp(ivX)\right] \\
&= \varphi_\epsilon(u + v) \cdot \varphi_X(v)\,.
\end{aligned}$$

We also have

$$\begin{aligned}
\varphi_{(\epsilon, X+\epsilon)}(u, v) &= \mathbf{E}\left[\exp(iu\epsilon + iv\epsilon + ivX)\right] \\
&= \mathbf{E}\left[\exp(iu\epsilon) \cdot \exp(iv\epsilon + ivX)\right] \\
&= \mathbf{E}\left[\exp(iu\epsilon)\right] \cdot \mathbf{E}\left[\exp(iv\epsilon + ivX)\right] \\
&= \varphi_\epsilon(u) \cdot \varphi_{(\epsilon+X)}(v) \\
&= \varphi_\epsilon(u) \cdot \varphi_\epsilon(v) \cdot \varphi_X(v)\,.
\end{aligned}$$

We know that $\varphi_X(0) = 1$ and that characteristic functions are continuous. Thus there exists a non-empty open interval $V = (-r, r) \subset \mathbb{R}$, such that $|\varphi_X(v)| > 0\ \forall v \in V$. Thus we have for all $u \in \mathbb{R}$ and $v \in V$:

$$\varphi_\epsilon(u + v) = \varphi_\epsilon(u) \cdot \varphi_\epsilon(v)\,.$$

Note that this is still true for an arbitrary $v \in \mathbb{R}$: Choose $n \in \mathbb{N}$, such that $\|v/n\| \leq r$. It follows

$$\begin{aligned}\varphi_\epsilon(u+v) &= \varphi_\epsilon\left(u + (n-1)\frac{v}{n} + \frac{v}{n}\right)\\ &= \varphi_\epsilon\left(u + (n-1)\frac{v}{n}\right) \cdot \varphi_\epsilon\left(\frac{v}{n}\right)\\ &\vdots\\ &= \varphi_\epsilon(u) \cdot \varphi_\epsilon\left(\frac{v}{n}\right)^n = \varphi_\epsilon(u) \cdot \varphi_\epsilon(v)\end{aligned}$$

Then we know

$$\varphi_\epsilon(u) = z^u \qquad \text{for some } z \in \mathbb{C}\backslash\{c \in \mathbb{C} : \operatorname{Im} c = 0, \operatorname{Re} c < 0\}\,.$$

We can write $z = \exp(a + ib)$ and since $\|\varphi_\epsilon\|_\infty \leq 1$ we deduce that $a = 0$. It follows

$$\varphi_\epsilon(u) = \exp(ib \cdot u)\,.$$

Because of the uniqueness of characteristic functions this implies $\mathbf{P}(\epsilon = b) = 1$ and ϵ is degenerate. □

Furthermore we will use the following result, which was proved by Skitovich and Darmois independently [10] [11] [12]:

Theorem 2.8 *[Darmois-Skitovich] Let $X_1, \ldots, X_n$ be independent, non-degenerate random variables. If the two linear combinations*

$$\begin{aligned} l_1 &= a_1X_1 + \ldots + a_nX_n, \qquad a_i \neq 0\\ l_2 &= b_1X_1 + \ldots + b_nX_n, \qquad b_i \neq 0\end{aligned}$$

are independent, each X_i is normally distributed.

There exist different proofs of this theorem, all using characteristic functions. We will sketch one proof (see e.g. Chapter 8 in [13]), which has the advantage that it can be generalized to the case of an infinite sum of random variables. This proof, however, requires the following powerful theorem from Linnik [14] and Zinger [15]:

Theorem 2.9 *Let $f_1, \ldots, f_n$ be characteristic functions, which satisfy*

$$\prod_{i=1}^{n} f_i^{\alpha_i}(t) = f(t)\,,$$

for some $\alpha_i > 0$ and for all t in a neighbourhood of zero. Here, f is the characteristic function of a normal distribution. Then every f_i itself is a characteristic function of a normal distribution.

This theorem is a generalization of Cramér's theorem [16], that only covers the case $\alpha_i = 1$.

Proof [of the Darmois-Skitovich theorem] We give the main steps of this proof, leaving some details to the reader.

By a linear transformation it can be shown that without loss of generality (wlog) we can set $a_i = 1$ for all i. We have

$$\prod_{i=1}^{n} \varphi_i(u + b_i v) = \prod_{i=1}^{n} \varphi_i(u) \prod_{i=1}^{n} \varphi_i(b_i v)\,, \tag{2.4}$$

where φ_i denotes the characteristic function of X_i. We prove by contradiction that none of the φ_i vanishes on the real line: If any of them did, there would be a root u_0 of φ_j with smallest absolute value (φ_i is continuous and $\varphi_i(0) = 1$). Then for all $v \in \mathbb{R}$

$$\prod_{i=1}^{n} \varphi_i(u_0 + b_i v) = 0\,.$$

Choosing v, such that $|b_i v| < |u_0/2|$ for all i yields

$$\prod_{i=1}^{n} \varphi_i\left(\frac{u_0}{2}\right) \prod_{i=1}^{n} \varphi_i\left(\frac{u_0}{2} + b_i v\right) = 0$$

and either $\frac{u_0}{2}$ or $\frac{u_0}{2} + b_i v$ is a root, which leads to a contradiction. Thus we can take logarithms in (2.4) and obtain

$$\sum_{i=1}^{n} \psi_i(u + b_i v) = \sum_{i=1}^{n} \psi_i(u) + \sum_{i=1}^{n} \psi_i(b_i v) =: A(u) + B(v)\,,$$

where $\psi_i = \ln \varphi_i$. Skitovich now considers finite differences and concludes that ψ_1 is a polynomial of second degree. We take a different approach by integrating over u:

$$\begin{aligned}
& \sum_{i=1}^{n} \int_0^x \psi_i(u + b_i v)(x - u)du = \int_0^x A(u)(x - u)du + B(v)\frac{x^2}{2} \\
\Rightarrow \quad & \sum_{i=1}^{n} \int_0^{x + b_i v} \psi_i(t)(x - t + b_i v)dt - \int_0^{b_i v} \psi_i(t)(x - t + b_i v)dt = C(x) + B(v)\frac{x^2}{2} \\
\Rightarrow \quad & \sum_{i=1}^{n} \int_0^{x + b_i v} \psi_i(t)(x - t + b_i v)dt = C(x) + B_1(v)\frac{x^2}{2} + B_2(v)x + B_3(v)
\end{aligned}$$

Here, $B_1(v), B_2(v), B_3(v)$ and $C(x)$ are chosen such that the equations are satisfied. Differentiating both sides twice with respect to v and setting $v = 0$ afterwards yields

$$\sum_{i=1}^{n} b_i^2 \psi_i(x) = R(x)$$

and thus

$$\prod_{i=1}^{n} \varphi_i(x)^{c_i^2} = \exp(R(x))\,, \tag{2.5}$$

where R is a polynomial of second degree with complex coefficients. Using $R(0) = 0$ and $R(-x) = \overline{R(x)}$, we see that the right hand side of (2.5) is the characteristic function of a normal distribution and thus we can apply Theorem 2.9. □

Now we are able to prove the following key statement of this section.

Theorem 2.10 *Let X and Y be two random variables, for which*

$$Y = \phi X + \epsilon, \quad \epsilon \perp\!\!\!\perp X,\ \phi \neq 0$$

holds.
Then we can reverse the process, i.e. there exists $\psi \in \mathbb{R}$ and a noise $\tilde{\epsilon}$, such that

$$X = \psi Y + \tilde{\epsilon}, \quad \tilde{\epsilon} \perp\!\!\!\perp Y,$$

if and only if $X, Y, \epsilon, \tilde{\epsilon}$ are Gaussian distributed.

Later we generalize this theorem, but to make the proof better understandable we prove the simple case first.

Proof If X and ϵ are Gaussian distributed, the statement follows from Example 2.6. Conversely, we assume that

$$\begin{array}{rrcrcr} & Y = & \phi X & + & \epsilon \\ \text{and} & \tilde{\epsilon} = & (1-\phi\psi)X & - & \psi\epsilon \end{array}$$

are independent. Distinguish between the following cases:

1. $(1-\phi\psi) \neq 0$ and $\psi \neq 0$
   Here, Theorem 2.8 implies that X, ϵ and thus also $Y, \tilde{\epsilon}$ are normally distributed.
2. $\psi = 0$
   We have $(1-\phi\psi)X \perp\!\!\!\perp \phi X + \epsilon$. $\psi = 0$ implies

   $$X \perp\!\!\!\perp \phi X + \epsilon,$$

   which is a contradiction to Lemma 2.7.
3. $(1-\phi\psi) = 0$
   It follows $-\psi\epsilon \perp\!\!\!\perp \phi X + \epsilon$. Thus

   $$\epsilon \perp\!\!\!\perp \phi X + \epsilon$$

   and we can apply Lemma 2.7 again.

□

This result is already known. The LiNGAM algorithm [5] we mentioned before actually makes use of this fact, which can be seen as a special case of Independent Component Analysis (see Theorem 11 in [17]). Although our proof and the one given in [17] are not the same, both are based on the Darmois-Skitovich Theorem 2.8. Our proof, however, can be generalized in different ways (see Theorem 2.11 and Section 4.2).

Theorem 2.11 *Let $X_1, \dots, X_n$ and Y be random variables, for which*

$$Y = \sum_{i=1}^{n} \phi_i X_i + \epsilon, \qquad \epsilon \perp\!\!\!\perp (X_1, \dots, X_n),\ \phi_i \neq 0$$

holds. Then we can reverse the process, i.e. there exists $\psi_i \in \mathbb{R}$ and a noise $\tilde{\epsilon}$, such that

$$X_1 = \sum_{i=2}^{n} \psi_i X_i + \psi Y + \tilde{\epsilon}, \qquad \tilde{\epsilon} \perp\!\!\!\perp (Y, X_2, \dots, X_n)$$

if and only if $X_1, \dots, X_n, Y, \epsilon, \tilde{\epsilon}$ are Gaussian distributed.

Proof The proof is analogue to the one from Theorem 2.10: If all variables are Gaussian, we can define $\hat{X}_1$ as the projection of X_1 on $\text{span}(Y, X_2, \dots, X_n)$. Then we can define the new noise $\tilde{\epsilon} := X_1 - \hat{X}_1$ and by construction

$$X_1 = \sum_{i=2}^{n} \psi_i X_i + \psi Y + \tilde{\epsilon}\,,$$

where $\tilde{\epsilon}$ is independent of Y and of all $X_i,\ i = 2, \dots, n$.
Conversely, we assume

$$\begin{aligned} Y &= \sum_{i=1}^{n} \phi_i X_i + \epsilon \\ \text{and} \quad \tilde{\epsilon} &= (1 - \psi\phi_1) X_1 - \sum_{i=2}^{n} (\psi_i + \psi\phi_i) X_i - \psi\epsilon \end{aligned}$$

are independent. Again, it is straightforward (mainly by using Lemma 2.7) to argue why the coefficients cannot vanish. We can apply Theorem 2.8 and it follows that all involved variables are Gaussian distributed. □

3 Theory of Statistical Methods

In this section we introduce Reproducing Kernel Hilbert Spaces, Support Vector Machines and the Hilbert-Schmidt Independence Criterion. We present these concepts in a strict mathematical context.

3.1 Kernels

3.1.1 Definition of Kernels

In the following subsection let $\mathcal{X}$ be a separable metric space with Borel σ-algebra Γ. We think of $\mathcal{X}$ as being an input space, in which we receive the data; that means single data points are treated as $(\mathcal{X}, \Gamma)$ -valued random variables. Note that we implicitly assume the existence of a probability space $(\Omega, \mathcal{A}, \mu)$.

Definition 3.1 Consider $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $x_1, \ldots, x_m \in \mathcal{X}$. Then the matrix K with

$$K_{ij} := k(x_i, x_j)$$

is called the *Gram matrix* of a kernel k.

Definition 3.2 A symmetric $m \times m$ matrix K satisfying

$$\langle Kc, c\rangle = \sum_{i,j} c_i c_j K_{ij} \geq 0 \qquad \forall c_i \in \mathbb{R}$$

is called *positive definite*[1]. Obviously this is equivalent to all eigenvalues of K being non-negative.

Remark 3.3 If a matrix K comes from a dot product of a dot product space $(V, \ (.,.))$, i.e.

$$K_{ij} := (v_i, v_j), \qquad v_1, \ldots, v_m \in V\,,$$

we have that

$$\sum_{i,j=1}^{n} c_i c_j K_{ij} = \sum_{i,j=1}^{n} c_i c_j (v_i, v_j) = \Big(\sum_{i=1}^{n} c_i v_i, \sum_{j=1}^{n} c_j v_j\Big) \geq 0 \qquad \forall c_i \in \mathbb{K}\,,$$

and therefore K is positive definite.

[1] Strictly speaking such a matrix should be called *positive semi-definite*. Nevertheless we adapt to the commonly used notation and omit the prefix *semi-*. Conversely we say a matrix is *strictly positive definite* if it additionally satisfies $\sum_{i,j} c_i c_j K_{ij} = 0 \Leftrightarrow c_i = 0 \; \forall i$.

Definition 3.4 We call $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ a *positive definite kernel* (or just *kernel*) if its corresponding Gram matrix is positive definite for every choice of $x_1, \dots, x_m \in \mathcal{X}$.

Example 3.5 The following functions are well-known examples of kernels for $\mathcal{X} = \mathbb{R}^n$:

- Gaussian kernel with bandwidth $\sigma > 0$

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

- polynomial kernel of degree $d \in \mathbb{N}$

$$k(x, y) = \langle x, y\rangle^d$$

- inhomogeneous polynomial kernel with $c \geq 0$ and degree $d \in \mathbb{N}$

$$k(x, y) = (\langle x, y\rangle + c)^d$$

- sigmoid kernel with $\kappa > 0$ and $\theta < 0$

$$k(x, y) = \tanh(\kappa\langle x, y\rangle + \theta)$$

- B_n splines of odd order n:

$$k(x, y) = B_n(\|x - y\|) \qquad \text{where } B_n = \underbrace{1_{[-\frac{1}{2},\frac{1}{2}]} * \dots * 1_{[-\frac{1}{2},\frac{1}{2}]}}_{n\text{-times}}$$

  using the convolution $(f * g)(t) = \int f(z)g(t - z)\, dz$.

All these examples are kernels on $\mathbb{R}^d$, which is the most important space for practical purposes, but there exist kernels on many other domains, too (e.g. graphs, sets of strings [18], etc.).

3.1.2 Reproducing Kernel Hilbert Spaces

Now we consider Hilbert spaces whose dot products are related to such kernels. These spaces turn out to be very useful. They consist of real-valued functions $f : \mathcal{X} \to \mathbb{R}$, for which -as usual- summation and multiplication by a scalar is defined pointwise:

$$\begin{aligned} (\lambda \cdot f)(x) &:= \lambda \cdot f(x) && \forall\lambda \in \mathbb{R}, \forall f \in \mathcal{H} \text{ and } \forall x \in \mathcal{X} \\ (f + g)(x) &:= f(x) + g(x) && \forall f \in \mathcal{H}, \forall g \in \mathcal{H} \text{ and } \forall x \in \mathcal{X} \end{aligned}$$

The follwing definitions can be found in [19], for example.

Definition 3.6 Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is called a *Reproducing Kernel Hilbert Space (RKHS)* if there is a kernel k such that

- $k(x, .) \in \mathcal{H} \quad \forall x \in \mathcal{X}$
- $\langle f, k(x, .)\rangle = f(x) \quad \forall f \in \mathcal{H}$

For $f = k(x', .)$ the second condition yields $\langle k(x', .), k(x, .)\rangle = k(x', x)$. This explains the term *Reproducing* Kernel Hilbert Space. Notice that the two conditions together imply that

$$\mathcal{H} = \overline{\text{span}\{k(x, .) \mid x \in \mathcal{X}\}}\,.$$

This can be seen as follows: Consider an element $g \in \mathcal{H}$ with $g \perp k(x, .)\ \forall x \in \mathcal{H}$. The reproducing property implies $g(x) = \langle g, k(x, .)\rangle = 0\ \forall x \in \mathcal{H}$. Thus $g \equiv 0$.

There is a natural way to represent the data in an RKHS using the following definition:

Definition 3.7 The *feature map* is defined as

$$\Phi : \begin{array}{ccc} \mathcal{X} & \to & \mathcal{H} \\ x & \mapsto & k(x, .) \end{array} .$$

An RKHS should be thought of a high-dimensional (or even infinite-dimensional) feature space. Mapping the data into the RKHS corresponds to extracting relevant features. This sometimes makes it easier to work with the data (see Example 3.11). Usually in a high-dimensional feature space computations, especially evaluations of the dot product, are quite expensive. For an RKHS, however, we have

$$\langle \Phi(x_i), \Phi(x_j)\rangle = k(x_i, x_j)\,,$$

which can be computed very efficiently.

Remark 3.8 Assume $\mathcal{H}$ is an RKHS with kernel k and $\tilde{k}$ is another kernel of $\mathcal{H}$ satisfying the conditions of Definition 3.6. Then

$$\tilde{k}(x, x') = \tilde{k}(x', x) = \langle \tilde{k}(x', .), k(x, .)\rangle = \langle k(x, .), \tilde{k}(x', .)\rangle = k(x, x')$$

and we can see that the kern of an RKHS is unique.

There is also a more abstract characterization of an RKHS [19]:

Proposition 3.9 *Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{H}$ is an RKHS if and only if for every $x \in \mathcal{X}$ the point evaluation operator*

$$\delta_x : \begin{array}{ccc} \mathcal{H} & \to & \mathbb{R} \\ f & \mapsto & f(x) \end{array}$$

is a bounded linear functional.

Proof This proposition is a consequence of Riesz' representation theorem (e.g. [20]).

- $\Rightarrow$: Since $\mathcal{H}$ is an RKHS there is a kernel $k(., .)$, such that

$$\delta_x(f) = f(x) = \langle f, k(x, .)\rangle \quad \forall f \in \mathcal{H}, \forall x \in \mathcal{X}\,.$$

Let $x \in \mathcal{X}$ be fixed. The linearity of δ_x is clear and its operator norm is bounded because of the Cauchy-Schwartz inequality:

$$\|\delta_x\| = \sup_{\|f\|=1} \left|\langle f, k(x, .)\rangle\right| = \left|\langle \frac{k(x, .)}{\|k(x, .)\|}, k(x, .)\rangle\right| = \|k(x, .)\| = \sqrt{k(x, x)}$$

(In fact, this is true in general: In every Hilbert space $\mathcal{H}$ the functional $f \mapsto \langle f, g\rangle$ for a fixed $g \in \mathcal{H}$ is linear and bounded.)

- $\Leftarrow$: It follows from Riesz that $\forall x \in \mathcal{X}\, \exists g_x \in \mathcal{H}$ such that $\delta_x = \langle ., g_x \rangle$. Thus

  $$g_{x'}(x) = \delta_x(g_{x'}) = \langle g_{x'}, g_x \rangle = \langle g_x, g_{x'} \rangle = \delta_{x'}(g_x) = g_x(x')\,.$$

  We can define the symmetric function $k(x, x') = g_x(x')$. Remark 3.3 guarantees that k is positive definite and thus a kernel.

□

We now proof the existence of such an RKHS by an explicit construction.

Proposition 3.10 *For any given kernel k there exists an RKHS with corresponding kernel k.*

Proof Define the space

$$\mathcal{H}^0 := \left\{ f : \mathcal{X} \to \mathbb{R} \mid f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i) \quad \text{for some } m \text{ and some } \alpha_i \in \mathbb{R} \right\}$$

Define further for $f = \sum_{i=1}^{m} \alpha_i k(., x_i)$ and $g = \sum_{j=1}^{m'} \beta_j k(., y_j)$

$$\begin{aligned} \langle f, g \rangle &= \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, y_j) \\ &= \sum_{j=1}^{m'} \beta_j f(y_j) \\ &= \sum_{i=1}^{m} \alpha_i g(x_i) \end{aligned}$$

The last two identities show that the expression does not depend on the expansion of f or g. Therefore this form is well-defined. It is easy to check that this form is bilinear and symmetric. It is positive semi-definite, since

$$\langle f, f \rangle = \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0\,.$$

It can be proved [21] that the Cauchy-Schwarz inequality

$$|\langle v, w \rangle| \leq ||v|| \cdot ||w||$$

holds not only for dot products, but also for symmetric positive semi-definite bilinear forms. It follows that

$$f(x)^2 = \langle f, k(x, .) \rangle^2 \overset{\text{C-S}}{\leq} \langle f, f \rangle \cdot k(x, x) \quad \forall x \in \mathcal{X}$$

and therefore

$$\langle f, f \rangle = 0 \qquad \Rightarrow \qquad f \equiv 0\,.$$

Thus we have shown that $\langle ., . \rangle$ is a dot product.
As the final step of the construction we define $\mathcal{H}$ to be the completion of $\mathcal{H}^0$. This standard procedure creates a new space, in which the initial space can be embedded and in which all Cauchy sequences converge: $\mathcal{H}$ is a Hilbert space with $\mathcal{H}^0$ as a dense subspace.
Convergence in $\mathcal{H}$ implies pointwise convergence in $\mathbb{R}$: For each $x \in \mathcal{X}$ we have

$$|f_i(x) - f_j(x)| = \left|\langle k(x, .), f_i - f_j \rangle\right| \leq \sqrt{k(x, x)} \cdot \|f_i - f_j\|_{\mathcal{H}} .$$

Thus the reproducing property holds in the completion $\mathcal{H}$, too:

$$\begin{aligned} \langle f, k(x, .) \rangle &= \langle \lim_{i \to \infty} f_i, k(x, .) \rangle \\ &= \lim_{i \to \infty} \langle f_i, k(x, .) \rangle \\ &= \lim_{i \to \infty} f_i(x) \\ &= f(x) \end{aligned}$$

Since all conditions are met, $\mathcal{H}$ is an RKHS with kernel k. □

We have seen that every function $f \in \mathcal{H}$ can be written as the limit of a sequence $(\sum_{i=1}^n \alpha_{n,i} k(x_{n,i}, .))_n$ in the RKHS norm. It is clear that this class of functions strongly depends on the kernel k. We will see later (in Section 3.3.3), how to choose k in order to make the class very rich, but still handable.
With the following example we try to give an intuition, why the concept of an RKHS can be useful.

Example 3.11 Let $\mathcal{X} = \mathbb{R}^2$ be the input space and consider a polynomial kernel of degree 2:

$$\begin{aligned} k((a, b), (x, y)) &= \langle (a, b), (x, y) \rangle^2 \\ &= (ax + by)^2 \\ &= a^2 x^2 + 2abxy + b^2 y^2 \end{aligned}$$

Define

$$\mathcal{H}^0 := \operatorname{span}\left\{ k((a, b), .) \mid (a, b) \in \mathbb{R}^2 \right\}.$$

Using the map

$$\psi : \begin{array}{ccc} \mathcal{H}^0 & \to & \mathbb{R}^3 \\ ((x, y) \mapsto cx^2 + dxy + ey^2) & \mapsto & (c, \frac{d}{\sqrt{2}}, e) \end{array}$$

we can see easily that $\mathcal{H}^0$ is isometric isomorphic to $\mathbb{R}^3$: ψ is surely linear and injective. Furthermore we have

$$\begin{aligned} (1, 0, 0) &= \psi\Big(k((1, 0), .)\Big) \\ (0, 1, 0) &= \psi\Big(k((1, \frac{1}{\sqrt{2}}), .) - k((0, \frac{1}{\sqrt{2}}), .) - k((1, 0), .)\Big) \\ (0, 0, 1) &= \psi\Big(k((0, 1), .)\Big) \end{aligned}$$

Additionally this map turns out to be isometric in the sense that it preserves dot products:

$$\begin{aligned}
\Big\langle \sum_{i=1}^{n} \alpha_i k((a_i,b_i),.), \sum_{j=1}^{m} \beta_j k((\tilde a_j,\tilde b_j),.)\Big\rangle_{\mathcal{H}^0} &= \sum_{i,j} \alpha_i\beta_j k((a_i,b_i),(\tilde a_j,\tilde b_j)) \\
&= \sum_{i,j} \alpha_i\beta_j (a_i^2\tilde a_j^2 + 2a_ib_i\tilde a_j\tilde b_j + b_j^2\tilde b_j^2) \\
&= \Big\langle \Big(\sum_i \alpha_i a_i^2, \sum_i \alpha_i \sqrt{2}a_ib_i, \sum_i \alpha_i b_i^2\Big), \\
&\qquad \Big(\sum_j \beta_j\tilde a_j^2, \sum_j \beta_j \sqrt{2}\tilde a_j\tilde b_j, \sum_j \beta_j\tilde b_j^2\Big)\Big\rangle_{\mathbb{R}^3} \\
&= \Big\langle \psi\Big(\sum_{i=1}^{n} \alpha_i k((a_i,b_i),.)\Big), \psi\Big(\sum_{j=1}^{m} \beta_j k((\tilde a_j,\tilde b_j),.)\Big)\Big\rangle_{\mathbb{R}^3}
\end{aligned}$$

Since $\mathbb{R}^3$ is complete it follows that $\mathcal{H}^0$ is, too. Ergo $\mathcal{H}^0 = \mathcal{H}$.
For this polynomial kernel we showed that working in the RKHS is equivalent to mapping the data into $\mathbb{R}^3$ via the mapping

$$\tilde\Phi : \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R}^3 \\ (x,y) & \mapsto & (x^2, y^2, \sqrt{2}xy) \end{array} .$$

and working there using the usual Euclidean dot product. If we now receive data belonging to two different classes (o and $*$), it may happen that it is difficult to separate the data in the input space (e.g. we have to use a circle), whereas in the feature space the data can be separated easily (e.g. using a hyperplane), compare Figure 3.1. This example relates to the concept of Kernel Support Vector Machines (Section 3.2.4) and shows, why it can be an advantage to work in a feature space.

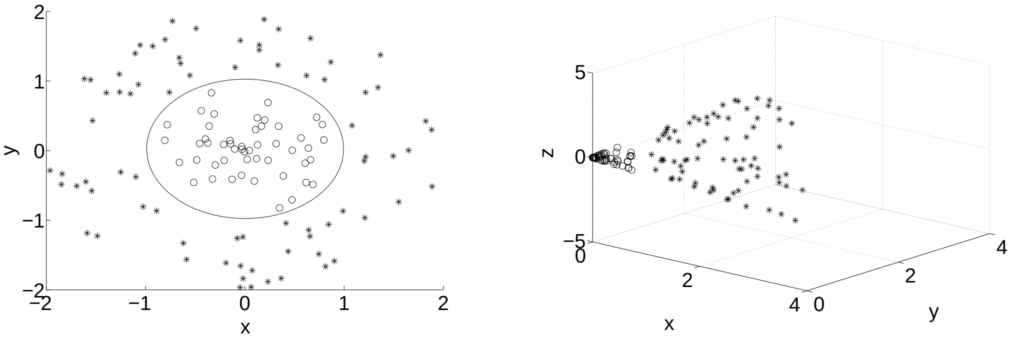


Figure 3.1: The left picture shows the data in input space, the right picture its representation in the RKHS

Remark 3.12 In this work we make the following assumptions on the kernel. (Notice that all of them are satisfied by the Gaussian kernel.)

- k is bounded, i.e. $\exists c \in \mathbb{R} : \ k(x,\tilde x) < c \ \forall x, \tilde x \in \mathcal{X}$. It follows that all functions in the RKHS are bounded

$$|f(x)| \le \|f\|_{\mathcal{H}} \cdot \|k(x,.)\|_{\mathcal{H}} \le \|f\|_{\mathcal{H}} \cdot \sqrt{k(x,x)} \le \|f\|_{\mathcal{H}} \cdot \sqrt{c}$$

- k is continuous

These conditions together with the fact that $\mathcal{X}$ is separable guarantee that the corresponding RKHS is separable, too; this is not hard to see and is shown for example in [22]. The separability of the RKHS is needed for the definition of the Hilbert-Schmidt norm.

3.1.3 A Hilbert Space Embedding of Distributions

We already saw, how we can represent single data points in an RKHS. In this section we learn a way to represent probability distributions $\mathbf{P}$ on $\mathcal{X}$ in an RKHS.
Therefore consider the mapping[2]

$$\begin{array}{ccc} \mathcal{H} & \to & \mathbb{R} \\ f & \mapsto & \mathbf{E}[f(X)] \end{array}.$$

This is well-defined because f is continuous and bounded. Furthermore this is obviously a linear function in f and it is continuous since

$$\begin{aligned} \left|\mathbf{E}[f(X)]\right| &= \left|\mathbf{E}\langle f, k(X,.)\rangle\right| \\ &\leq \mathbf{E}\left|\langle f, k(X,.)\rangle\right| \\ &\leq \mathbf{E}\,\|f\|_{\mathcal{H}}\,\|k(X,.)\|_{\mathcal{H}} \\ &= \|f\|_{\mathcal{H}}\,\mathbf{E}\,\sqrt{k(X,X)}\,, \end{aligned}$$

so

$$\sup_{\|f\|_{\mathcal{H}}=1} |\mathbf{E} f(X)| < \infty\,.$$

By Riesz' representation theorem there is an element $\mu[\mathbf{P}] \in \mathcal{H}$ such that

$$\mathbf{E}[f(X)] = \langle f, \mu[\mathbf{P}]\rangle \quad \forall f \in \mathcal{H}\,.$$

Therefore we can represent any probability measure in an RKHS using

$$\mathbf{P} \longmapsto \mu[\mathbf{P}]\,.$$

We will refer to $\mu[\mathbf{P}]$ as being the *mean element*.

Proposition 3.13 *It holds*

$$\mu[\mathbf{P}] : \begin{array}{ccc} \mathcal{X} & \to & \mathbb{R} \\ x & \mapsto & \mathbf{E}_X[k(X,x)] \end{array}.$$

Proof For every x we have

$$\mu[\mathbf{P}](x) = \langle k(.,x), \mu[\mathbf{P}]\rangle = \mathbf{E}[k(X,x)]\,.$$

□

[2] Instead of $\mathbf{E}_{X\sim\mathbf{P}} f(X)$ we use the shorthand $\mathbf{E}_X f(X)$ or even $\mathbf{E} f(X)$.

3.2 Support Vector Machines

A Support Vector Machine (SVM) is an algorithm that addresses the task of classification, which itself is a common problem in data analysis: in a data space $\mathcal{X}$ we are given a point Z, that we want to assign to one of J different classes. Obviously there is a huge variety of applications: It was estimated that between January and March 2008 92.3 per cent of email traffic is spam [23] and therefore good spam filters are needed, which can classify incoming mail reliably as being spam or not. Looking at an MRI scan a computer can do a ”pre-diagnosis” and predict if a patient has got a special disease. The post companies use computers for assigning hand written numbers on an envelope to one of the digits $0, \ldots, 9$ in order to sort the mail according to their destinations. You can also think of a company, which checks a calling customers prefix, age, income and the number of children, decides that he probably rather have complaints than the intention to open a bank account and therefore place him on hold.
In a lot of applications we have to deal with a huge amount of data and therefore efficient algorithms are indispensable.
Formally, in binary classification problems the data is given in $\mathcal{X}\times\{-1,1\}$, i.e. we receive the data in an input space $\mathcal{X}$ together with an attached label (-1 or 1). We expect there to be a function (or classification rule)

$$f: \mathcal{X} \to \{-1,1\},$$

such that the data lie on the subspace $(\mathcal{X}, f(\mathcal{X}))$. For received data the goal is to learn this function f. Obviously we have to restrict the class of function candidates by introducing some smoothness condition, for example. Otherwise there is no way of learning f. We consider classification rules that are constructed using hyperplanes. Most of the definitions and statements given in this subsection can also be found in [19].

3.2.1 Hyperplanes

Let $\mathcal{H}$ be a complete vector space with a dot product (i.e. a Hilbert space). For now you can think of $\mathcal{H}$ being $\mathbb{R}^n$, but it is important to notice that the following works for every complete dot product space. In Section 3.2.4 we use the algorithm on Reproducing Kernel Hilbert Spaces and obtain the so-called kernel SVM.

Definition 3.14 A subset $H \subset \mathcal{H}$ is called a *hyperplane* if there exist $\mathbf{w} \in \mathcal{H} \setminus \{0\}$ and $b \in \mathbb{R}$, such that

$$H = \{\mathbf{x} \in \mathcal{H} \,|\, \langle \mathbf{w}, \mathbf{x}\rangle + b = 0\}.$$

We call $(\mathbf{w}, b)$ a representation of H.[3]

These hyperplanes separate the space into two half spaces: a single point lies either on one side of it or on the other and we can define the classifier:

Definition 3.15 For given $\mathbf{w} \in \mathcal{H}$ and $b \in \mathbb{R}$ define the classification rule

$$f_{\mathbf{w},b}: \begin{array}{ccc} \mathcal{H} & \longrightarrow & \mathbb{R} \\ \mathbf{x} & \longmapsto & \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}\rangle + b) \end{array}.$$

[3] In this subsection we use the following convention: small bold letters denote vectors in vector spaces, small normal letters numbers in $\mathbb{R}$.

Thus we say a given point $\mathbf{x}$ belongs to class 1 if it lies on one side of the hyperplane and to class -1 if it lies on the other side. Note that $f_{\mathbf{w},b}$ and $f_{-\mathbf{w},-b}$ define two different classifiers.

By elementary geometry it can be shown that a representation of a hyperplane is not unique:

Remark 3.16

$$\{\mathbf{x} \in \mathcal{H} \,|\, \langle \mathbf{w}_1, \mathbf{x}\rangle + b_1 = 0\} = \{\mathbf{x} \in \mathcal{H} \,|\, \langle \mathbf{w}_2, \mathbf{x}\rangle + b_2 = 0\}$$

if and only if there exists a $k \in \mathbb{R}$ such that

$$\mathbf{w}_2 = k\mathbf{w}_1 \qquad \text{and} \qquad b_2 = kb_1\,.$$

In particular this means that $\mathbf{w}$ and b are not uniquely determined by H.

In a real situation we use the given training data in order to define a unique representation of a hyperplane:

Definition 3.17 (i) Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{H}$. We call $(\mathbf{w}, b)$ a *canonical representation* of the hyperplane H if

$$\min_{i=1,\ldots,m} |\langle \mathbf{w}, \mathbf{x}_i\rangle + b| = 1.$$

(ii) The *margin* ρ is the distance from the closest point $\mathbf{x}_i$ to the hyperplane[4]:

$$\rho := \min_{i=1,\ldots,m} \operatorname{dist}(\mathbf{x}_i, H)\,.$$

Remark 3.18 Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{H}$. Then for $(\mathbf{w}, b) \in \mathcal{H} \times \mathbb{R}$ and its generated hyperplane H the following is equivalent:

(i) $(\mathbf{w}, b)$ is the canonical form of H.

(ii) The margin of H is $\frac{1}{\|w\|}$.

(See Figure 3.2.)

Proof Let $\mathbf{z} \in H$. Notice that the projection from $\mathbf{x} - \mathbf{z}$ onto $\frac{\mathbf{w}}{\|\mathbf{w}\|}$ is exactly the distance from $\mathbf{x}$ to the hyperplane. Thus we have

$$\begin{aligned}
|\langle \mathbf{w}, \mathbf{x}\rangle + b| &= |\langle \mathbf{w}, \mathbf{x}\rangle + b - (\langle \mathbf{w}, \mathbf{z}\rangle + b)| \\
&= |\langle \mathbf{w}, \mathbf{x} - \mathbf{z}\rangle| \\
&= \|\mathbf{w}\|\, |\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, \mathbf{x} - \mathbf{z}\rangle| \\
&= \|\mathbf{w}\| \operatorname{dist}(\mathbf{x}, H)
\end{aligned}$$

Ergo $|\langle \mathbf{w}, \mathbf{x}\rangle + b|$ and $\operatorname{dist}(\mathbf{x}, H)$ are minimized by the same $\mathbf{x}_1$, say, and the statement follows. □

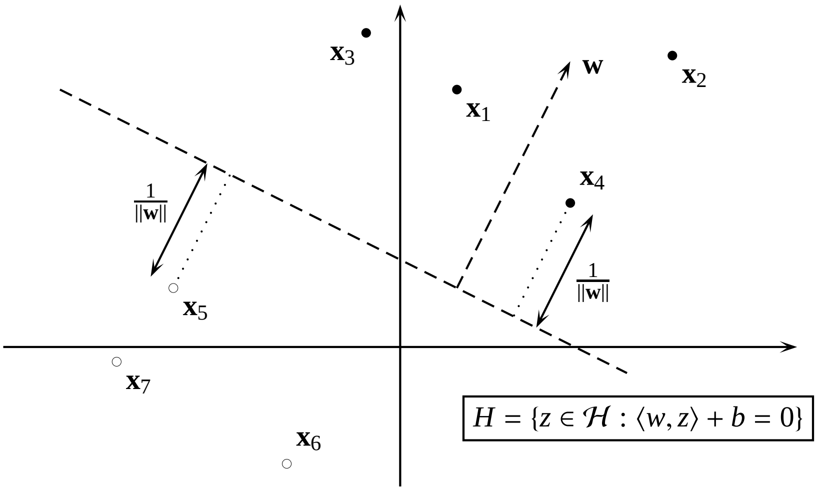


Figure 3.2: A hyperplane in its canonical form $(\mathbf{w}, b)$; its margin is $\frac{1}{||w||}$.

Remark 3.16 and Remark 3.18 together show that there are exactly two canonical representations of a hyperplane H: If $(\mathbf{w}, b)$ is one canonical representation, $(-\mathbf{w}, -b)$ is the other. As we want to construct classifiers from these hyperplanes the distinction is necessary. Although they describe the same hyperplane and therefore the same set of points in $\mathcal{H}$, the corresponding classifiers differ (see Definition 3.15).

Now we introduce the *Support Vector Machine*, which is a method that constructs such a hyperplane from the labelled data.

3.2.2 Hard Margin SVM

Given some training data we want to learn the classification rule, i.e. the form of the hyperplane (cf Definition 3.15).

Among all possible hyperplanes a hard margin SVM chooses the hyperplane, such that

1. all training data are classified correctly and
2. the margin is maximized.

The hyperplane drawn in Figure 3.2 is an example for a hard margin hyperplane (the two classes are represented by white and black points).

Remark 3.19 (i) Note that the training data does not have to be separable by a hyperplane, furthermore even if possible a separation may not be advisable as it may lead to overfitting (cf soft margin SVMs and kernel SVMs).

[4] Observe that $\langle \mathbf{x}, . \rangle$ is continuous and therefore H is a closed subspace of $\mathcal{H}$. Thus the distance $\text{dist}(\mathbf{x}, H)$ is well-defined.

(ii) Maximizing the margin is intuitively a good idea: we choose the hyperplane, such that the distance to the point closest to it is as large as possible. Assume in a survey people are asked if they like the TV show Musikantenstadl. All young people (aging 21-57) said no, the others (aging 63-103) said yes. Then we suspect the border to be at around 60 because it seems to be the best generalization.

This choice can even be justified theoretically: there are bounds on the probability of making a test error. Of course these bounds themselves only hold with a certain probability, but it can be seen that they get the tighter the larger the margin is (for details see Section 7.2. of [19]).

For canonical hyperplanes the margin is always $1//||\mathbf{w}||$ and thus maximizing the margin corresponds to minimizing $||\mathbf{w}||$ or equivalently $1/2 \cdot ||\mathbf{w}||^2$. We take the latter because then the problem turns out to be a quadratic programming problem. Further, given some data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ in $\mathcal{H} \times \{-1, 1\}$ we assume that both classes -1 and 1 occur within the y_i at least once.

Now we can summarize the hard margin SVM procedure to be the following optimization problem:

$$\begin{array}{rl} \min_{\mathbf{w},b} & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} & f_{\mathbf{w},b}(\mathbf{x}_i) = y_i \quad \forall i = 1, \ldots, m \\ \text{and} & (\mathbf{w}, b) \text{ is a canonical representation} \end{array}$$

This can be rewritten as

$$O_1 : \begin{array}{rl} \min_{\mathbf{w},b} & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} & y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1 \quad \forall i = 1, \ldots, m \end{array}$$

Assume the problem is feasible, which means that the data can be separated linearly in the space $\mathcal{H}$. If $(\hat{\mathbf{w}}, \hat{b})$ is the solution of this problem, we can check easily that for at least one of the vectors $\mathbf{x}_i$ the constraint is precisely met:

$$y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) = 1\,.$$

This means that $(\hat{\mathbf{w}}, \hat{b})$ is automatically a canonical representation of a hyperplane.

Proposition 3.20 *O_1 is equivalent to*

$$O_2 : \begin{array}{rl} \max_{\alpha\in\mathbb{R}^m} & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j=1}^m \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle \\ \textit{subject to} & \alpha_i \geq 0 \quad \forall i = 1, \ldots, m \quad \textit{and} \quad \sum_{i=1}^m \alpha_i y_i = 0 \end{array}$$

Notice that this is a quadratic programming problem (with positive definite matrix in the objective function) and can be solved very efficiently by the ellipsoid method [24].

Proof Because of the inequality constraint we introduce some slack variables η_i, such that O_1 becomes

$$\begin{array}{rl} \min_{\mathbf{w},b,\eta} & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} & y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - \eta_i - 1 = 0 \quad \forall i = 1, \ldots, m \end{array}$$

Because of the convexity of the objective function, the convexity of the constraints and the satisfied interior point condition (or Slater constraint qualification), we know that the problem is strong Lagrangian (“min of the primal=max of the dual”); for details see e.g. Chapter 5.3 in [25] ($\mathbb{R}^k$) or Chapter 10 in [26] (general Hilbert spaces). Thus solving the

primal is equivalent to solving the dual.
To compute the dual we note that the Lagrangian equals

$$L(\mathbf{w}, b, \eta, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \Big(y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - \eta_i - 1 \Big) .$$

The dual problem is defined to be

$$\begin{array}{rl} \max_{\alpha \in \mathbb{R}^m} & g(\alpha) := \inf_{\mathbf{w},b,\eta} L(\mathbf{w}, b, \eta, \alpha) \\ \text{subject to} & \alpha \in Y := \{\alpha \in \mathbb{R}^m \mid g(\alpha) > -\infty\} \end{array}$$

The constraint for α is equivalent to[5] $\alpha \geq 0$, which implies that we can set $\eta = 0$. The current optimization problem is

$$\max_{\alpha \in \mathbb{R}^m_{\geq 0}} \inf_{\mathbf{w},b} \tilde{L}(\mathbf{w}, b, \alpha) := \max_{\alpha \in \mathbb{R}^m_{\geq 0}} \inf_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \Big(y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \Big) . \tag{3.1}$$

For the optimal solution $\hat{\alpha}$ let $(\hat{\mathbf{w}}, \hat{b})$ be minimizing $\tilde{L}$. Then we know by the Lagrangian Sufficiency Theorem (or Kuhn-Tucker Saddle Point Condition) that $(\hat{\mathbf{w}}, \hat{b})$ is also a solution to the primal O_1. It follows

$$\begin{aligned} \frac{\partial}{\partial b} \tilde{L}(\hat{\mathbf{w}}, b, \hat{\alpha})_{b=\hat{b}} = 0 \ &\Rightarrow \sum_{i=1}^{m} \hat{\alpha}_i y_i = 0 \\ \frac{\partial}{\partial \mathbf{w}} \tilde{L}(\mathbf{w}, \hat{b}, \hat{\alpha})_{\mathbf{w}=\hat{\mathbf{w}}} = 0 \ &\Rightarrow \hat{\mathbf{w}} = \sum_{i=1}^{m} \hat{\alpha}_i y_i \mathbf{x}_i \end{aligned}$$

Plugging this into (3.1) yields O_2. □

Notice that the solution $\hat{\mathbf{w}}$ can be written as a linear combination of the training data. Some of the Lagrangian multipliers α_i may be zero and thus the solution $\hat{\mathbf{w}}$ is not supported by these vectors.

Definition 3.21 Assume $(\hat{\mathbf{w}}, \hat{b})$ is the solution of the optimization problem O_2 described above. The vectors $\mathbf{x}_i$ which satisfy

$$\alpha_i > 0$$

are called *support vectors*.

In Figure 3.2 the points $\mathbf{x}_4$ and $\mathbf{x}_5$ are support vectors.
It can further be shown (see Chapter 7.3. in [19]) that the data points $\mathbf{x}_i$ corresponding to Lagrangian multipliers $\alpha_i > 0$ satisfy the constraints of O_1 exactly.

[5] This is shorthand for $\alpha_i \geq 0 \ \forall i = 1, \ldots, m$.

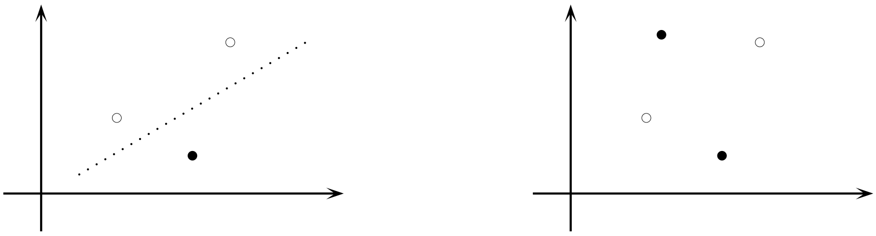


Figure 3.3: The VC dimension of $\mathbb{R}^2$ is 3: in the left picture, a separation of the points is possible for any labelling, whereas the right figure is an example, where the points are not linearly separable.

Furthermore this helps to determine the threshold b: if $\alpha_j > 0$ (i.e. x_j is a support vector), then

$$\begin{aligned} & y_j(\langle \mathbf{x}_j, \hat{\mathbf{w}} \rangle + \hat{b}) = 1 \\ \Rightarrow \quad & \langle \mathbf{x}_j, \sum_{i=1}^{m} \hat{\alpha}_i y_i \mathbf{x}_i \rangle + \hat{b} = y_j \\ \Rightarrow \quad & \sum_{i=1}^{m} \hat{\alpha}_i y_i \langle \mathbf{x}_j, \mathbf{x}_i \rangle + \hat{b} = y_j \\ \Rightarrow \quad & \hat{b} = y_j - \sum_{i=1}^{m} \hat{\alpha}_i y_i \langle \mathbf{x}_j, \mathbf{x}_i \rangle \end{aligned}$$

Ergo the threshold b can be computed by choosing only one support vector $\mathbf{x}_j$ and using this equation. In practice, however, we find small differences in the values for b if we use different support vectors. This is due to numerical problems and is dealt with by averaging over all values of b we obtain for different support vectors.

Remark 3.22 The hard margin SVM has two major drawbacks:

1. We need the training data to be linearly separable, otherwise we cannot construct the separating hyperplane. The concept of soft margin classifiers (Section 3.2.3) relax the constraint that all training data must be classified correctly.
2. We can only separate the data linearly. The Vapnik-Chervonenkis (VC) dimension [19] of the class of half-spaces in $\mathbb{R}^n$ can be shown to be $n + 1$. In the case of $n = 2$, this means that we can arrange 3 points in $\mathbb{R}^2$ in such a way that for any labelling of these 3 points, we can separate the two classes by a straight line. This is not possible for any arrangement of 4 points in $\mathbb{R}^2$. We can always attach class labels to 4 points, such that the classes cannot be separated anymore (see Figure 3.3).

   Now we can either conclude that the class of hyperplane classifiers is too small or we find a way of mapping our data in a high-dimensional space, where the class gets more powerful. The latter is exactly realised by kernel SVMs, which we explain in Section 3.2.4.

3.2.3 Soft margin SVM

In the last section we constructed the hyperplane, which separated all training data according to their classes. In order to relax this constraint, we could try to separate as many data as possible. Unfortunately, this turns out to be an $\mathcal{NP}$-hard problem (see for example [27]).
Instead, Cortes and Vapnik [28] changed the constraints of the original optimization problem O_1. They added some slack variables $\xi_i \geq 0$ and obtained

$$y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, m \, .$$

Some classification errors on the training set are now allowed, but if ξ_i gets too large, the constraints can always be satisfied. Therefore we penalize big values of ξ_i and add the term

$$\frac{C}{m} \sum_{i=1}^{m} \xi_i \, ,$$

to the objective function. Here $C > 0$ is a constant, which adjusts the strength of the regularization term [6]. This leads to the following quadratic programming problem

$$\tilde{O}_1 : \begin{array}{ll} \min_{\mathbf{w},b,\xi} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m} \xi_i \\ \text{subject to} & y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, m \end{array}$$

A computation analogue to the one before shows that this is equivalent to

$$\tilde{O}_2 : \begin{array}{ll} \max_{\alpha \in \mathbb{R}^m} & \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{subject to} & \frac{C}{m} \geq \alpha_i \geq 0 \quad \forall i = 1, \ldots, m \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \end{array}$$

As before, we can write the solution as a linear combination of the training data:

$$\hat{\mathbf{w}} = \sum_{i=1}^{m} \hat{\alpha}_i y_i \mathbf{x}_i \, ,$$

and again the support vectors are those $\mathbf{x}_i$, which support the solution, i.e. $\alpha_i > 0$. As before Lagrangian multipliers greater than zero correspond to precisely met constraints

$$y_i(\langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle + \hat{b}) = 1 - \xi_i \, ,$$

and we can conclude that for support vectors $\mathbf{x}_j$ satisfying additionally $\xi_i = 0$ (both together is equivalent to $0 < \alpha_i < C$) the following holds:

$$\hat{b} = y_j - \sum_{i=1}^{m} \hat{\alpha}_i y_i \langle \mathbf{x}_j, \mathbf{x}_i \rangle \, . \tag{3.2}$$

And again we average over all those vectors $\mathbf{x}_j$ to deal with the numerical problems.
In practice, there are different procedures to choose C: as a rule of thumb you can set $C = 10 \cdot m$ for first results; you can also choose C by cross-validation or raise C from a very low starting value until the training error (misclassifications on the training set) is lower than a certain fraction.

[6] There exist other formulations as well, where the interpretation of the adjusting constant gets somehow easier. See for example Schölkopf's ν classifier in [29].

3.2.4 Kernel SVM

As mentioned before, SVMs can get much more powerful if we map the data in a high-dimensional space, the so-called feature space. Performing an SVM classification in the high-dimensional feature space often corresponds to a non-linear classification in the input space; such a procedure, however, is usually computationally very expensive because of the evaluations of the dot products in the high-dimensional space.
We can avoid this problem if we choose the feature space to be an RKHS. Recall the definition of the feature map:

$$\Phi : \begin{array}{ccc} \mathcal{X} & \to & \mathcal{H} \\ \mathbf{x} & \mapsto & k(.,\mathbf{x}) \end{array} .$$

Then we perform the usual SVM in the RKHS. This so-called *kernelizing* of the SVM is possible because the whole original SVM algorithm only depends on the dot product of the data $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. This means, in the feature space we only have to consider dot products $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$, which can be calculated very quickly by writing it in terms of the kernel:

$$\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)\,.$$

Herein lies a lot of the power of the kernel trick: the expensive evaluation of the dot product in a suited high-dimensional feature space can be replaced by a relatively cheap evaluation of the kernel function.
As a summary the procedure of the kernel SVM is as follows:

1. Choose a kernel,
2. map the data into the (possibly infinite-dimensional) Reproducing Kernel Hilbert Space and
3. apply an SVM in this RKHS.

It seems obvious that choosing the kernel and its parameters carefully can increase the performance of an SVM a lot.
Note that the soft margin SVM now has an additional advantage: we introduced it as a possibility of creating a hyperplane, even if the training data was not separable. Using a kernel SVM it is often the case that separating all training data is possible. It may not be advisable though because this would lead to overfitting: If there is an outlier in the data, which is wrongly labelled, a classifier which tries to be correct on the whole training data depends a lot on the outlier and will not perform very well on the test data. How a bad choice of C can lead to over- or underfitting is shown in Figures 3.4-3.6.

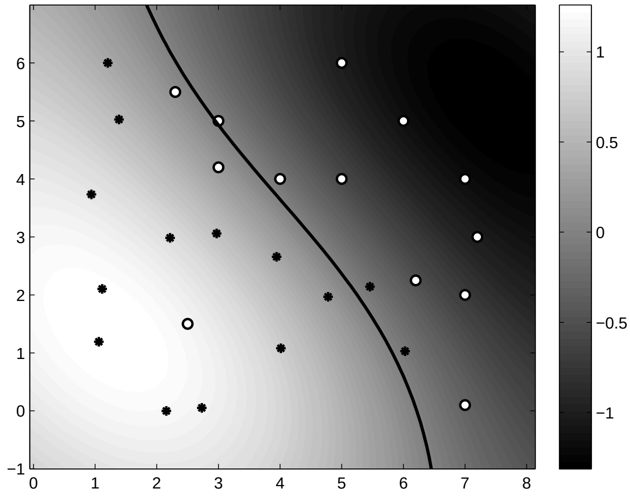


Figure 3.4: The dark line is the hyperplane in the RKHS represented in the input space and therefore our decision boundary. Choosing $C = 1 \cdot m$ leads to underfitting: 5 classification errors.

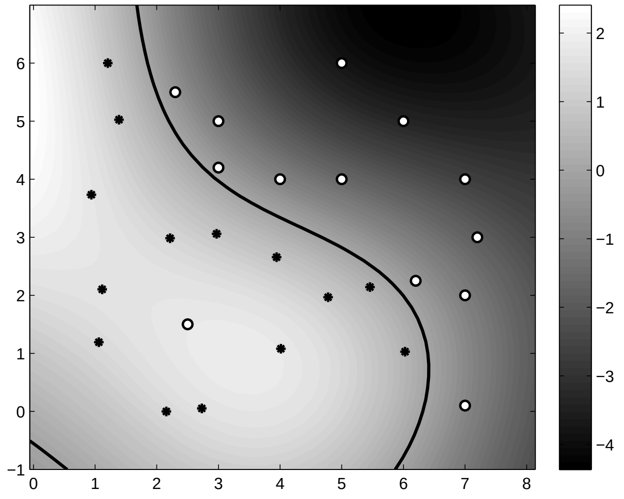


Figure 3.5: Choosing $C = 10 \cdot m$ results in only 1 classification error.

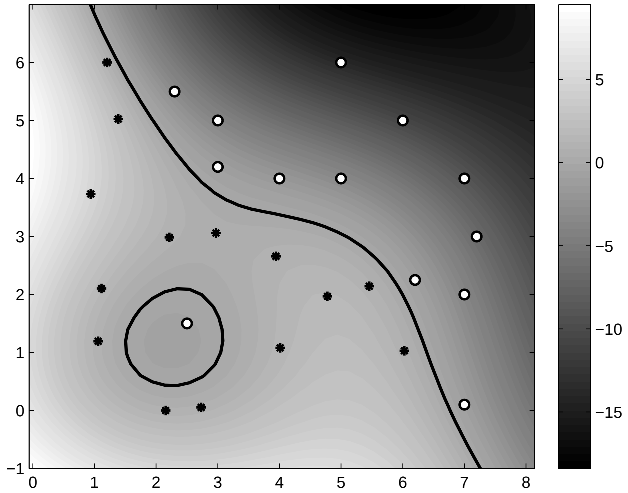


Figure 3.6: $C = 1000 \cdot m$ leads to overfitting: all points are classified correctly.

3.3 Hilbert-Schmidt Independence Criterion

There are different methods to measure the dependence or association between two random variables. Optimally, this measure should be zero if and only if the variables are independent. Some approaches, however, only take second order dependence into account, which means they cannot detect dependencies beyond correlation. Other methods (like the widely used χ^2 test) only work for discrete data, which is a drawback because there is no canonical way of discretizing continuous data. If we use too few bins for discretization, we loose information and if we use too many, we do not have enough data in each bin. In our work we chose the Hilbert-Schmidt Independence Criterion (HSIC); it does not suffer from those problems. Note that our method works for other independence tests as well. In the experiments we tried several independence tests and the HSIC performed best, probably because of the reasons mentioned above.

The HSIC is a kernel based method to detect dependence between two random variables: both the joint probability distribution and the product of the marginal distributions are mapped in an infinite-dimensional feature space in such a way that these two points coincide if and only if the two random variables are independent.

In the first subsection we give some well-known results from functional analysis, that we need later. Afterwards we introduce HSIC in two different ways; first we define it as the Hilbert-Schmidt norm of the cross-covariance operator [8] and then as a special case of the Maximum Mean Discrepancy (MMD) [30]. We show both possibilities in order to develop a deeper understanding for the HSIC.

Remark 3.23 For the formal setup of this whole section let X and Y be two random variables, that take values on $(\mathcal{X}, \Gamma)$ and $(\mathcal{Y}, \Lambda)$, respectively; here, $\mathcal{X}$ and $\mathcal{Y}$ are two separable metric spaces, Γ and Λ are Borel σ-algebras. Then $(\mathcal{X} \times \mathcal{Y}, \Gamma \otimes \Lambda)$ is again a measurable space and X and Y are independent if and only if $\mathbf{P}^{(X,Y)} = \mathbf{P}^X \otimes \mathbf{P}^Y$.

We define kernels $k(.,.)$ and $l(.,.)$ on the spaces $\mathcal{X}$ and $\mathcal{Y}$ and denote the corresponding RKHSs with $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$, respectively.

3.3.1 Some Functional Analysis.

We want to introduce HSIC as the Hilbert-Schmidt norm of the cross-covariance operator. In order to do so we give some definitions and results from functional analysis, which include the concepts of singular value decomposition and Hilbert-Schmidt operators.

Let $\mathcal{H}$, $\mathcal{H}_1$ and $\mathcal{H}_2$ be two separable Hilbert spaces over $\mathbb{R}$. Denote the set of all continuous operators (i.e. bounded and linear functions) $T : \mathcal{H}_1 \to \mathcal{H}_2$ by $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ and set $\mathcal{L}(\mathcal{H}) := \mathcal{L}(\mathcal{H}, \mathcal{H})$. We further define

Definition 3.24 A subset $S \subset \mathcal{H}$ is called an *orthonormal system* if $\langle e_i, e_j \rangle = \delta_{ij}$ for all $e_i, e_j \in \mathcal{H}$. An orthonormal system $S \subset \mathcal{H}$ is called an *orthonormal basis* of $\mathcal{H}$ if $\mathcal{H} = \overline{\operatorname{span} S}$.

Definition 3.25

- For $T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ the adjoint of T is the unique operator $T^* \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_1)$ satisfying
$$\langle Tx, y \rangle = \langle x, T^* y \rangle \quad \forall x \in \mathcal{H}_1, y \in \mathcal{H}_2 \,.$$
- $T \in \mathcal{L}(\mathcal{H})$ is called *self-adjoint* if
$$T^* = T \,.$$
- $T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ is called *unitary* if T is invertible and
$$T^* = T^{-1} \,.$$
- $T \in \mathcal{L}(\mathcal{H})$ is called *positive* if
$$\langle Tx, x \rangle \geq 0 \quad \forall x \in \mathcal{H} \,.$$
(This implies that all eigenvalues are non-negative.)
- A linear map $T : \mathcal{H}_1 \to \mathcal{H}_2$ is called *compact* (or a compact operator) if it maps bounded subsets of $\mathcal{H}_1$ onto relatively compact subsets of $\mathcal{H}_2$. We write $\mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$ for the set of all compact operators and $\mathcal{K}(\mathcal{H}) := \mathcal{K}(\mathcal{H}, \mathcal{H})$. It is not hard to see that a compact operator is bounded and therefore continuous. Thus $\mathcal{K}(\mathcal{H}_1, \mathcal{H}_2) \subset \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$.

The following results from functional analysis are well-known. (For complete proofs see [20], for example.)

Remark 3.26

- Spectral Decomposition:
  The spectral theorem allows us to decompose a compact self-adjoint operator $T \in \mathcal{K}(\mathcal{H})$ into:
  $$Tx = \sum_{k\geq 1} \lambda_k \langle x, e_k \rangle e_k \qquad \forall x \in \mathcal{H}\,,$$
  where[7] $(\lambda_1, \lambda_2, \ldots)$ are the non-zero eigenvalues (each eigenvalue is repeated as many times as the dimension of its eigenspaces) with corresponding eigenvectors[7] $(e_1, e_2, \ldots)$ and $\lambda_k \to 0$. Furthermore we have
  $$\|T\| = \sup_{k\geq 1} |\lambda_k|\,.$$
  We can expand $(e_1, e_2, \ldots)$ to an orthonormal basis of $\mathcal{H}$ by adding an orthonormal basis of ker T:
  $$\mathcal{H} = \ker T \;\oplus\; \overline{\mathrm{lin}(e_1, e_2, \ldots)}\,.$$
  Notice, however, that opposed to $E_\lambda := \ker\,(\lambda - T)$ with $\lambda \neq 0$, the space $E_0 = \ker T$ can be infinite-dimensional and even non-separable.
  For a general compact operator $T \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$ the spectral theorem does not hold, of course, but we can apply it to the self-adjoint operator TT^*. To do so we first need the following two auxiliary results.
- Roots of operators:
  Using the spectral decomposition you can construct roots from compact operators: For every positive, self-adjoint operator $T \in \mathcal{K}(\mathcal{H})$ there is a unique positive, self-adjoint operator $S \in \mathcal{K}(\mathcal{H})$, such that
  $$S^2 = T\,.$$
  We write $S = T^{\frac{1}{2}}$. In case of the positive, self-adjoint operator TT^* we write $|T| := (TT^*)^{\frac{1}{2}}$.
- Polar Decomposition:
  For every $T \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$ there is a unique operator $U \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$, such that
  $$T = U|T|\,, \qquad U\big|_{(\ker U)^\perp} \text{is unitary} \quad \text{and} \quad \ker U = \ker T\,.$$
  (This reminds of the polar decompostion in the complex plane: $z = |z| \exp(i\phi)$.)
- Singular Value Decomposition:
  For every $T \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$ there exist orthonormal systems $(e_1, e_2, \ldots)$ of $\mathcal{H}_1$ and $(f_1, f_2, \ldots)$ of $\mathcal{H}_2$ (both possibly finite) and a non-decreasing sequence (s_k) converging to zero, such that
  $$Tx = \sum_{k=1}^{\infty} s_k \langle x, e_k \rangle f_k \qquad \forall x \in \mathcal{H}_1\,.$$
  This can be shown as follows: Write $T = U|T|$ and according to the spectral decomposition we have $|T|x = \sum_k s_k \langle x, e_k \rangle e_k$. Now define $f_k := U(e_k)$.

[7] These collections may be finite.

It follows that the s_k^2 are the eigenvalues of TT^* (repeated according to the dimension of their eigenspaces). The $s_k = s_k(T)$ are called *singular values*.

Definition 3.27 $T \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$ is called a *Hilbert-Schmidt operator* if $(s_k(T)) \in l^2$, i.e. $\sum_{k=1}^{\infty} s_k^2 < \infty$. For these operators define

$$\|T\|_{\mathrm{HS}} := \left\|(s_k(T))\right\|_{l^2} = \left(\sum_{k=1}^{\infty} s_k^2\right)^{1/2} .$$

The linear space of all Hilbert-Schmidt operators $T : \mathcal{H} \to \mathcal{H}$ is denoted by HS($\mathcal{H}$).

Proposition 3.28 *Assume $T \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$ is a Hilbert-Schmidt operator. Then the following holds for all orthonormal bases (g_m) of $\mathcal{H}_1$ and (h_n) of $\mathcal{H}_2$;*

$$\|T\|_{HS}^2 = \sum_{m,n=1}^{\infty} \langle Tg_m, h_n\rangle^2 = \sum_{m=1}^{\infty} \|Tg_m\|^2$$

Proof Let $Tx = \sum_{k=1}^{\infty} s_k\langle x, e_k\rangle f_k$ be the singular value decomposition of T. Following VI.6.2 in [20] we have

$$\begin{aligned}
\sum_{m,n=1}^{\infty} \langle Tg_m, h_n\rangle^2 &= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\langle Tg_m, h_n\rangle\langle h_n, Tg_m\rangle = \sum_{m=1}^{\infty} \|Tg_m\|^2 \\
&= \sum_{m=1}^{\infty} \left\|\sum_{k=1}^{\infty} s_k\langle g_m, e_k\rangle f_k\right\|^2 = \sum_{m=1}^{\infty}\sum_{k=1}^{\infty} s_k^2|\langle g_m, e_k\rangle|^2 \\
&= \sum_{k=1}^{\infty}\sum_{m=1}^{\infty}\langle g_m, s_k e_k\rangle^2 = \sum_{k=1}^{\infty} \|s_k e_k\|^2 \\
&= \sum_{k=1}^{\infty} s_k^2
\end{aligned}$$

Here we used Parseval's equality [20] twice. □

Proposition 3.29 $\|.\|_{HS}$ *is a norm on HS($\mathcal{H}$).*

Proof This follows from Proposition 3.28 and the fact that $\|.\|_{l^2}$ itself is a norm. Note, for example

$$\begin{aligned}
\|T\|_{\mathrm{HS}} = 0 \quad &\Rightarrow \quad \sum_{m=1}^{\infty} \|Tg_m\|^2 = 0 \\
&\Rightarrow \quad Tg_m = 0 \quad \forall m \geq 0 \\
&\Rightarrow \quad Tx = \sum_{m=1}^{\infty}\langle x, g_m\rangle Tg_m = 0 \quad \forall x \in \mathcal{H} \\
&\Rightarrow \quad T = 0
\end{aligned}$$

□

Most of the following proposition is easy to prove [20] (only the completeness requires some work), but it is still useful:

Proposition 3.30 *For $T, S \in HS(\mathcal{H})$ define*

$$\langle T, S \rangle_{HS} := \sum_{m,n=1}^{\infty} \langle Tg_m, h_n\rangle\langle Sg_m, h_n\rangle = \sum_{k=1}^{\infty}\langle Sg_k, Tg_k\rangle\,,$$

for any orthonormal basis $(g_1, g_2, \ldots)$. Then $\langle .,.\rangle_{HS}$ is a dot product and induces $||.||_{HS}$. $(HS(\mathcal{H}), \langle .,.\rangle_{HS})$ is a Hilbert space.

The tensor product between two functions is the last definition in this subsection.

Definition 3.31 For $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$ define

$$f \otimes \langle g, .\rangle : \begin{array}{ccc} \mathcal{H}_2 & \to & \mathcal{H}_1 \\ h & \mapsto & \langle g, h\rangle f \end{array}.$$

As a shorthand notation we write $f \otimes g := f \otimes \langle g, .\rangle$.

It holds

$$||f \otimes g|| = \sup_{||h||_{\mathcal{H}_2}=1} |\langle g, h\rangle| \cdot ||f||_{\mathcal{H}_1} = ||g||_{\mathcal{H}_2} \cdot ||f||_{\mathcal{H}_1}$$

and also

$$\begin{aligned} ||f \otimes g||^2_{\mathrm{HS}} &= \langle f \otimes g, f \otimes g\rangle_{\mathrm{HS}} \\ &= \sum_m \langle (f \otimes g)(h_m), (f \otimes g)(h_m)\rangle_{\mathcal{H}_1} \\ &= \sum_m \langle \langle g, h_m\rangle_{\mathcal{H}_2} f, \langle g, h_m\rangle_{\mathcal{H}_2} f\rangle_{\mathcal{H}_1} \\ &= \langle f, f\rangle_{\mathcal{H}_1} \sum_m \langle g, h_m\rangle_{\mathcal{H}_2} \langle h_m, g\rangle_{\mathcal{H}_2} = ||f||^2_{\mathcal{H}_1} \cdot ||g||^2_{\mathcal{H}_2}, \end{aligned}$$

where $(h_1, h_2, \ldots)$ is a (possibly finite) orthonormal basis of $\mathcal{H}_2$.

3.3.2 HSIC using the cross-covariance operator

Let X and Y be two random variables taking values on $(\mathcal{X}, \Gamma)$ and $(\mathcal{Y}, \Lambda)$, respectively and let $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ be the corresponding Reproducing Kernel Hilbert Spaces. We define the cross-covariance operator, which – for carefully chosen kernel – captures all sorts of dependencies between X and Y. This definition is similar to the one given by Baker [31], although he uses measures defined directly on the function spaces.

Definition 3.32 The *cross-covariance operator* $C_{\mathcal{X},\mathcal{Y}} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ is defined as being the unique linear operator satisfying

$$\langle f, C_{\mathcal{X},\mathcal{Y}} g\rangle = \mathbf{E}_{X,Y} f(X)g(Y) - \mathbf{E}_X f(X)\mathbf{E}_Y g(Y) \qquad \forall f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}\,.$$

The expectations exist since f and g are continuous and bounded functions (see Remark 3.12).

The existence of such an operator is again ensured by Riesz representation theorem: obviously the right-hand side is linear in f and additionally it is bounded with a similar argumentation as above:

$$\begin{aligned}\left|\mathbf{E}_{X,Y} f(X)g(Y) - \mathbf{E}_X f(X)\mathbf{E}_Y g(Y)\right| &\leq \left|\mathbf{E}_{X,Y} f(X)g(Y)\right| + \left|\mathbf{E}_X f(X)\mathbf{E}_Y g(Y)\right| \\ &\leq \mathbf{E}_{X,Y}|f(X)| \cdot |g(Y)| + \mathbf{E}_X|f(X)| \cdot \mathbf{E}_Y|g(Y)| \\ &\leq \mathbf{E}_{X,Y}||f||_{\mathcal{H}_X}||k(X,.)||_{\mathcal{H}_X} \cdot |g(Y)| \\ &\qquad + \mathbf{E}_X||f||_{\mathcal{H}_X}||k(X,.)||_{\mathcal{H}} \cdot \mathbf{E}_Y|g(Y)| \\ &= ||f||_{\mathcal{H}_X}\mathbf{E}_{X,Y}\sqrt{k(X,X)} \cdot |g(Y)| \\ &\qquad + ||f||_{\mathcal{H}_X}\mathbf{E}_X\sqrt{k(X,X)} \cdot \mathbf{E}_Y|g(Y)|\end{aligned}$$

so

$$\sup_{||f||_{\mathcal{H}}=1} \left|\mathbf{E}_{X,Y} f(X)g(Y) - \mathbf{E}_X f(X)\mathbf{E}_Y g(Y)\right| < \infty\,.$$

Thus the expression can be written as a dot product $\langle f, C_{\mathcal{X},\mathcal{Y}}(g)\rangle$. The right-hand side is also linear in g, which implies linearity of $C_{\mathcal{X},\mathcal{Y}}$.
Our next goal is to derive the Hilbert-Schmidt norm of this operator. If the norm is finite, $C_{\mathcal{X},\mathcal{Y}}$ is a Hilbert-Schmidt operator and we define

$$\mathrm{HSIC}(\mathbf{P}^{(X,Y)}) := ||C_{\mathcal{X},\mathcal{Y}}||^2_{\mathrm{HS}}\,.$$

Lemma 3.33

$$C_{\mathcal{X},\mathcal{Y}} = \mathbf{E}_{X,Y}[k(X,.) \otimes l(Y,.)] - \mu[\mathbf{P}^X] \otimes \mu[\mathbf{P}^Y]$$

Proof We have for every $g \in \mathcal{H}_{\mathcal{Y}}$ and every $x \in \mathcal{X}$

$$\begin{aligned}C_{\mathcal{X},\mathcal{Y}}g(x) &= \langle k(.,x), C_{\mathcal{X},\mathcal{Y}}g\rangle \\ &= \mathbf{E}_{X,Y}k(X,x)g(Y) - \mathbf{E}_X k(X,x)\mathbf{E}_Y g(Y) \\ &= \mathbf{E}_{X,Y}(k(X,.)g(Y))(x) \;-\; \mathbf{E}_X k(X,.)\mathbf{E}_Y g(Y)\,(x) \\ &= \mathbf{E}_{X,Y}(k(X,.) \otimes l(Y,.))g\,(x) \;-\; \mu[\mathbf{P}^X] \otimes \mu[\mathbf{P}^Y]g\,(x) \\ &= \left(\mathbf{E}_{X,Y}k(X,.) \otimes l(Y,.) \;-\; \mu[\mathbf{P}^X] \otimes \mu[\mathbf{P}^Y]\right) g\,(x)\end{aligned}$$

Ergo

$$C_{\mathcal{X},\mathcal{Y}} = \mathbf{E}_{X,Y}[k(X,.) \otimes l(Y,.)] - \mu[\mathbf{P}^X] \otimes \mu[\mathbf{P}^Y]$$

□

The lemma helps us to express the HSIC in terms of kernels:

$$\mathrm{HSIC}(\mathbf{P}^{(X,Y)}) \;=\; \mathbf{E}_{X,Y}\mathbf{E}_{\tilde{X},\tilde{Y}}k(X,\tilde{X})l(Y,\tilde{Y}) - 2\mathbf{E}_{X,Y}\mathbf{E}_{\tilde{X}}\mathbf{E}_{\tilde{Y}}k(X,\tilde{X})l(Y,\tilde{Y}) + \mathbf{E}_X\mathbf{E}_{\tilde{X}}\mathbf{E}_Y\mathbf{E}_{\tilde{Y}}k(X,\tilde{X})l(Y,\tilde{Y}),$$

which is computed in [8].

3.3.3 HSIC using the Maximum Mean Discrepancy

The Maximum Mean Discrepancy (MMD) provides a kernel based method for the so-called Two-Sample-Problem. For this setting let $\mathcal{X}$ be a separable metric space with Borel-σ algebra Γ and probability measures $\mathbf{P}$ and $\mathbf{Q}$. We are given two sets of iid samples $\{X_1, \dots, X_m\}$ and $\{Y_1, \dots, Y_n\}$, which are drawn from $\mathbf{P}$ and $\mathbf{Q}$ respectively. The Two-Sample-Problem asks if the two samples come from two different measures or if $\mathbf{P}$ and $\mathbf{Q}$ are actually the same. The MMD measures the difference between two probability measures and can be used to create a statistical test for testing

$$\begin{array}{lll} \mathrm{H}_0: & \mathbf{P} = \mathbf{Q} & \quad \text{against} \\ \mathrm{H}_1: & \mathbf{P} \neq \mathbf{Q} & \end{array}$$

Therefore the MMD is one solution to the Two-Sample-Problem.

Considering the distributions $\mathbf{P} = \mathbf{P}^{(X,Y)}$ and $\mathbf{Q} = \mathbf{P}^X \otimes \mathbf{P}^Y$ leads to an independence criterion, which turns out to be HSIC. This section consists of the following paragraphs:

- MMD as the Maximum Difference in Means
- MMD as the Distance of Mean Elements
- Conditions for the MMD to be a metric
- HSIC as a special Case of MMD

MMD as the Maximum Difference in Means We now investigate the difference between two probability measures $\mathbf{P}$ and $\mathbf{Q}$. The MMD measures this difference depending on a function class $\mathcal{F}$.

Definition 3.34 Let $\mathcal{X}$ be a measurable space with measures $\mathbf{P}$ and $\mathbf{Q}$ and let $\mathcal{F}$ be any class of measurable functions $f : \mathcal{X} \to \mathbb{R}$. Then define the *Maximum Mean Discrepancy (MMD)* as

$$\mathrm{MMD}(\mathcal{F}, \mathbf{P}, \mathbf{Q}) = \sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim \mathbf{P}} f(X) - \mathbf{E}_{Y \sim \mathbf{Q}} f(Y) \right| .$$

Notice that for some classes we obtain well-known concepts, e.g.:

- $\mathcal{F} = \{1_A \mid A \in \mathcal{B}\}$ leads to the total variation between the measures $\mathbf{P}$ and $\mathbf{Q}$.
- $\mathcal{F} = \{f \mid f$ continuous and bounded leads to the metrization of the weak convergence.
- $\mathcal{F} = \{f | f = \exp(i\langle t, .\rangle), t \in \mathbb{R}^d\}$ leads to the biggest difference in the characteristic functions of $\mathbf{P}$ and $\mathbf{Q}$.

Surely the MMD is zero if $\mathbf{P}$ equals $\mathbf{Q}$. And the larger the class $\mathcal{F}$, the more probability measures we are able distinguish. The following lemma (e.g. [21]) shows a sufficient condition for two measures being equal.

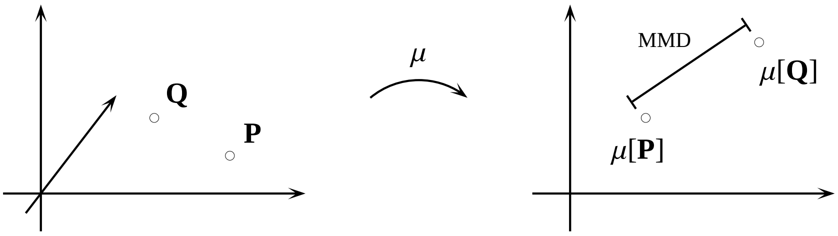


Figure 3.7: **P** and **Q** are mapped from the space of all probability measures (left) into an RKHS (right). The MMD can be shown to be their distance in the RKHS.

Lemma 3.35 *Let $\mathcal{X}$ be a metric space and* $\mathbf{P}, \mathbf{Q}$ *two Borel measures on* $(\mathcal{X}, \Gamma)$*. If* $\int f d\mathbf{P} = \int f d\mathbf{Q}$ *for all* $f \in C_b(\mathcal{X})$*, then* $\mathbf{P} = \mathbf{Q}$.

Ergo, for $\mathcal{F}$ being the class of bounded continuous functions the MMD is zero only if $\mathbf{P} = \mathbf{Q}$. This means, the MMD defines a metric on the space of all probability measures. For such a huge class, the quantity is very hard to compute, of course. The question arises, how to choose a function class $\mathcal{F}$, which satisfies the following three criteria:

1. It is big enough to guarantee that the MMD is a metric.
2. It is small enough, such that it can be computed efficiently.
3. It is chosen in a way, such that sample estimate of the MMD converges reasonably fast to the true value.

Gretton et al. [32] proposed to choose $\mathcal{F}$ as being the unit ball in an RKHS:

$$\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}.$$

Now the question for a good function class reduces to the problem of choosing a good kernel. Before we come back to this question, we first introduce a different way of interpreting the MMD:

MMD as the Distance of Mean Elements The MMD has a very nice geometric interpretation, too. Recall that we can represent probability measures as single points in an RKHS via $\mathbf{P} \mapsto \mu[\mathbf{P}] = \mathbf{E}_X k(X, .)$. If you take two measures **P** and **Q** and map them into the RKHS, the distance between these two mean elements turns out to be the MMD defined above.

Proposition 3.36 *Let* $k(.,.) \in \mathcal{L}^1$*. Then*

$$MMD(\mathbf{P}, \mathbf{Q}) = \|\mu[\mathbf{P}] - \mu[\mathbf{Q}]\|_{\mathcal{H}}$$

(see Figure 3.7).

Proof

$$\begin{aligned}
\mathrm{MMD}(\mathbf{P},\mathbf{Q}) &= \sup_{\|f\|_{\mathcal{H}}\leq 1} \left|\mathbf{E}_{\mathbf{P}} f(X) - \mathbf{E}_{\mathbf{Q}} f(Y)\right| \\
&= \sup_{\|f\|_{\mathcal{H}}\leq 1} \left|\langle \mu[\mathbf{P}], f\rangle_{\mathcal{H}} - \langle \mu[\mathbf{Q}], f\rangle_{\mathcal{H}}\right| \\
&= \sup_{\|f\|_{\mathcal{H}}\leq 1} \left|\langle \mu[\mathbf{P}] - \mu[\mathbf{Q}], f\rangle_{\mathcal{H}}\right| \\
&\overset{\text{C-S}}{=} \left\langle \mu[\mathbf{P}] - \mu[\mathbf{Q}], \frac{\mu[\mathbf{P}] - \mu[\mathbf{Q}]}{\|\mu[\mathbf{P}] - \mu[\mathbf{Q}]\|}\right\rangle_{\mathcal{H}} \\
&= \|\mu[\mathbf{P}] - \mu[\mathbf{Q}]\|_{\mathcal{H}}
\end{aligned}$$

□

This way of looking at the MMD is nice because of three different reasons:

1. It connects two quite different mathematical approaches of defining a distance for probability measures.
2. Using this formulation we can write the MMD in an easy closed-form expression:

$$\begin{aligned}
\mathrm{MMD}(\mathbf{P},\mathbf{Q}) &= \|\mu[\mathbf{P}] - \mu[\mathbf{Q}]\|_{\mathcal{H}} \\
&= \langle \mu[\mathbf{P}], \mu[\mathbf{P}]\rangle - 2\langle \mu[\mathbf{P}], \mu[\mathbf{Q}]\rangle + \langle \mu[\mathbf{Q}], \mu[\mathbf{Q}]\rangle \\
&= \mathbf{E}_X \mathbf{E}_{\tilde{X}} k(X,\tilde{X}) - 2\mathbf{E}_X \mathbf{E}_Y k(X,Y) + \mathbf{E}_Y \mathbf{E}_{\tilde{Y}} k(Y,\tilde{Y})
\end{aligned}$$

3. The problem of choosing a kernel that assures the MMD to be a metric is now equivalent to finding a kernel that makes the embedding

$$\mathbf{P} \longmapsto \mu[\mathbf{P}]$$

injective.

Conditions for the MMD to be a metric Now we answer the question, for which kernels the MMD is a metric or equivalently the embedding μ is injective by stating results from the literature. There are two different sufficient conditions on the kernel: It has to be either

- a universal kernel or
- a convolution kernel on $\mathbb{R}^d$, for which the Radon-Nikodym derivative of its inverse Fourier transform is supported almost everywhere.

Definition 3.37 Let $(\mathcal{X}, d)$ be a compact metric space. A kernel on $\mathcal{X}$ is called *universal* if the corresponding RKHS is dense in the space $C(\mathcal{X})$ of all continuous functions. Such an RKHS is also called *universal*.

The following theorem shows that this assumption is indeed sufficient (the proof is given by Gretton et al. [33]):

Theorem 3.38 *Let $\mathcal{F} := \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$ be the unit ball in a universal RKHS on a compact metric space $\mathcal{X}$. Then*

$$MMD(\mathbf{P},\mathbf{Q}) := MMD(\mathcal{F},\mathbf{P},\mathbf{Q}) = 0 \quad \Longleftrightarrow \quad \mathbf{P} = \mathbf{Q}$$

Definition 3.39 Let $\mathcal{X} = \mathbb{R}^d$. A kernel k on $\mathcal{X}$ is called a *convolution kernel* if it can be written as

$$k(x,y) = \psi(x-y)\,,$$

where ψ is a bounded continuous positive definite function.

Bochner's theorem (e.g. Theorem 6.6. in [34]) states that every positive definite function is the Fourier transform of a Borel measure:

Theorem 3.40 *Let $\psi : \mathbb{R}^d \to \mathbb{C}$ be a continuous function. It is positive definite if and only if*

$$\psi(x) = \int_{\mathbb{R}^d} \exp(-i\langle x, w\rangle) d\Lambda(w)\,,$$

where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.

From now on we assume that Λ is absolutely continuous with respect to the Lebesgue measure λ and we write

$$\frac{d\Lambda}{d\lambda} =: \Psi\,.$$

The following shows that the support of Ψ being strictly greater than zero almost everywhere, is also a sufficient condition for the injectivity of the embedding, i.e. the property that $MMD(\mathbf{P},\mathbf{Q})$ is zero only if $\mathbf{P} = \mathbf{Q}$:

Theorem 3.41 *Let k be a convolution kernel on $\mathbb{R}^d$ whose corresponding Borel measure Λ has the Radon-Nikodym derivative Ψ and let*

$$supp(\Psi) := \overline{\{x \in \mathbb{R}^d \mid \Psi(x) > 0\}} = \mathbb{R}^d\,.$$

For the unit ball $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$ in the corresponding RKHS we then have once more

$$MMD(\mathbf{P},\mathbf{Q}) = MMD(\mathcal{F},\mathbf{P},\mathbf{Q}) = 0 \quad \Longleftrightarrow \quad \mathbf{P} = \mathbf{Q}$$

This result is due to [35].

Remark 3.42 The universal property has the drawback that we require the input space $\mathcal{X}$ to be compact (that excludes $\mathbb{R}^d$), whereas the second condition needs $\mathcal{X}$ to be $\mathbb{R}^d$. Our data, however, lie in $\mathbb{R}^d$ and thus we make use of the second approach. We further note that all conditions of this section on the kernel are satisfied by the Gaussian kernel

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

It is bounded, continuous and additionally it is universal [36] and the inverse Fourier transform of a Gaussian is supported everywhere.

HSIC as a special Case of MMD We show that the HSIC can be regarded as a special case of the MMD. Consider a random vector (X, Y) taking values in the product space $(\mathcal{X}, \mathcal{Y})$ and define a product kernel on the space $(\mathcal{X}, \mathcal{Y})$ via

$$\begin{array}{ccc} \mathcal{X} \times \mathcal{Y} & \to & \mathbb{R} \\ ((x,y),(\tilde{x},\tilde{y})) & \mapsto & k(x,\tilde{x}) \cdot l(y,\tilde{y}) \end{array},$$

where $k(.,.)$ and $l(.,.)$ are kernels on $\mathcal{X}$, $\mathcal{Y}$, respectively. Then the MMD for the distributions $\mathbf{P} = \mathbf{P}^{(X,Y)}$ and $\mathbf{Q} = \mathbf{P}^X \otimes \mathbf{P}^Y$ can be expressed as

$$\mathrm{MMD}(\mathbf{P}^{(X,Y)}, \mathbf{P}^X \otimes \mathbf{P}^Y)^2 = \mathrm{HSIC}(\mathbf{P}^{(X,Y)}),$$

which can be seen as follows:

$$\begin{aligned} \mathrm{MMD}(\mathbf{P}^{(X,Y)}, \mathbf{P}^X \otimes \mathbf{P}^Y)^2 &= \langle \mu[\mathbf{P}^{(X,Y)}] - \mu[\mathbf{P}^X \otimes \mathbf{P}^Y], \mu[\mathbf{P}^{(X,Y)}] - \mu[\mathbf{P}^X \otimes \mathbf{P}^Y] \rangle \\ &= \langle \mu[\mathbf{P}^{(X,Y)}], \mu[\mathbf{P}^{(X,Y)}] \rangle - 2\langle \mu[\mathbf{P}^{(X,Y)}], \mu[\mathbf{P}^X \otimes \mathbf{P}^Y] \rangle \\ &\quad + \langle \mu[\mathbf{P}^X \otimes \mathbf{P}^Y], \mu[\mathbf{P}^X \otimes \mathbf{P}^Y] \rangle \\ &= \mathbf{E}_{X,Y} \mathbf{E}_{\tilde{X},\tilde{Y}} k(X,\tilde{X}) l(Y,\tilde{Y}) - 2\mathbf{E}_{X,Y} \mathbf{E}_{\tilde{X}} \mathbf{E}_{\tilde{Y}} k(X,\tilde{X}) l(Y,\tilde{Y}) \\ &\quad + \mathbf{E}_X \mathbf{E}_{\tilde{X}} \mathbf{E}_Y \mathbf{E}_{\tilde{Y}} k(X,\tilde{X}) l(Y,\tilde{Y}) \end{aligned}$$

which is exactly the expression for $\mathrm{HSIC}(\mathbf{P}^{(X,Y)})$ defined earlier (cf Section 3.3.2).

3.3.4 Sample Estimate of HSIC and its distribution

In any real-world situation we have to deal with a finite amount of data and thus we need a sample estimate of HSIC, which is converging reasonably fast to the true value of HSIC. If we want to create a statistical test, we further need at least an approximation of the distribution of this estimate under the null hypothesis of independence in order to bound the type one error.

Writing the HSIC in terms of kernels provides an easy way to get such a sample estimate. If we are given $(X_1, Y_1), \ldots, (X_m, Y_m)$, we can estimate $\mathrm{HSIC}(\mathbf{P}^{(X,Y)})$ by a V-statistic [37] that is denoted by $\mathrm{HS\hat{I}C}$. An unbiased estimator for $\mathrm{HSIC}(\mathbf{P}^{(X,Y)})$ is

$$k(X_1, X_2) l(Y_1, Y_2) - 2k(X_1, X_2) l(Y_1, Y_3) + k(X_1, X_2) l(Y_3, Y_4),$$

and the corresponding V-statistic has the form

$$\mathrm{HS\hat{I}C} = \frac{1}{m^2} \sum_{i,j}^{m} k(X_i, X_j) l(Y_i, Y_j) - 2\frac{1}{m^3} \sum_{i,j,f}^{m} k(X_i, X_j) l(Y_i, Y_f) + \frac{1}{m^4} \sum_{i,j,f,g}^{m} k(X_i, X_j) l(Y_f, Y_g).$$

We can think of the V-statistic as being a plug-in estimator: if F denotes the distribution function, you estimate the magnitude of interest $\theta(F)$ by $\theta(\hat{F}_m)$, where

$$\hat{F}_m(x) = \frac{1}{m} \sum_{i=1}^{m} 1_{X_i \leq x}$$

denotes the empirical (or sample) distribution function.

Writing[8] $(K)_{ij} := k_{ij} := k(X_i, X_j)$ and $(L)_{ij} := l_{ij} := l(Y_i, Y_j)$ and $H := I - \frac{1}{m}\mathbf{1}\cdot\mathbf{1}^t$ we can express this estimate for $\mathrm{HSIC}(\mathbf{P}^{(X,Y)})$ as a simple product of matrices:

$$
\begin{aligned}
\frac{1}{m^2}\mathrm{trace}(KHLH) &= \frac{1}{m^2}\sum_{f,l}(KH)_{fl}(LH)_{lf} \\
&= \frac{1}{m^2}\sum_{f,l}\left(-\frac{1}{m}\sum_{s\neq l}k_{fs} + \left(1-\frac{1}{m}\right)k_{fl}\right)\left(-\frac{1}{m}\sum_{s\neq f}l_{ls} + \left(1-\frac{1}{m}\right)l_{lf}\right) \\
&= \frac{1}{m^2}\sum_{f,l}\Bigg[\frac{1}{m^2}\sum_{s\neq l}\sum_{r\neq f}k_{fs}l_{lr} - \frac{m-1}{m^2}\sum_{s\neq f}k_{fl}l_{ls} - \frac{m-1}{m^2}\sum_{s\neq l}k_{fs}l_{lf} \\
&\qquad\qquad + \frac{(m-1)^2}{m^2}k_{fl}l_{lf}\Bigg] \\
&= \frac{1}{m^2}\sum_{f,l}\left[k_{fl}l_{lf}\right] - \frac{2}{m^3}\sum_{f,l}\left[k_{fl}l_{lf} + \sum_{s\neq f}k_{fl}l_{ls} + \sum_{s\neq l}k_{fs}l_{lf}\right] \\
&\qquad + \frac{1}{m^4}\sum_{f,l}\left[\sum_{s\neq l}\sum_{r\neq f}k_{fs}l_{lr} + k_{fl}l_{lf} + 2\sum_{s\neq l}k_{fs}l_{lf}\right] \\
&= \frac{1}{m^2}\sum_{i,j}^{m}k_{ij}l_{ij} - 2\frac{1}{m^3}\sum_{i,j,f}^{m}k_{ij}l_{if} + \frac{1}{m^4}\sum_{i,j,f,g}^{m}k_{ij}l_{fg} \\
&= \hat{\mathrm{HSIC}}
\end{aligned}
$$

Moreover, this shows that the estimate $\hat{\mathrm{HSIC}}$ can be computed in $O(m^2)$ time. [8] investigated the convergence of this estimator and computed corresponding deviation bounds.

As already mentioned, for a hypothesis test for independence we further need to know the distribution of the test statistic $\hat{\mathrm{HSIC}}$ under the assumption of independence.
One approach is using a **bootstrap estimator**: you brake the connection between X_i and Y_i, create new pairs (X_i, Y_j) by shuffling and compute a new value for $\hat{\mathrm{HSIC}}$. This is done many times and since X_i and Y_j can be regarded as being independent, you obtain an empirical distribution of $\hat{\mathrm{HSIC}}$ under the null hypothesis of independence. This takes a lot of running time, though.
A different approach is using a **Gamma approximation** for the distribution of HSIC, which is based on the following result (see [32] and Section 5.5.2. in [37]):

Theorem 3.43 *Under the assumption of independence (i.e. $HSIC(\mathbf{P}^{(X,Y)}) = 0$), we have*

$$
m \cdot \hat{HSIC} \xrightarrow{d} 6 \cdot \sum_{l=1}^{\infty}\lambda_l\, w_l^2 \,,
$$

where $w_l^2 \overset{iid}{\sim} \chi_1^2$ and λ_l solves the eigenvalue problem

$$
\lambda_l\, g_l(z_j) = \int h_{ijqr}\, g_l(z_i)\, d\mathbf{P}^{(Z_i,Z_q,Z_r)} \,,
$$

[8] Here, I is the $m \times m$ identity matrix and $\mathbf{1}$ the $m \times 1$ vector containing only ones.

where $Z_j = (X_j, Y_j)$ *and* $h_{ijqr} = \frac{1}{4!} \sum_{\{t,u,v,w\}=\{i,j,q,r\}} k_{tu} l_{tu} + k_{tu} l_{vw} - 2 k_{tu} l_{tv}$*; here, the sum represents all ordered quadruples* (t, u, v, w) *drawn without replacement from* (i, j, q, r)*. (There are* 4! *summands.)*

Although we know the asymptotic distribution of $\hat{\mathrm{HSIC}}$, it is hard to get exact quantiles, for example. This is due to the difficult eigenvalue problem and to the infinite sum of random variables. Therefore we use an approximation for this distribution, which was suggested by Kankainen [38]:

$$m\hat{\mathrm{HSIC}} \overset{d}{\approx} \Gamma(\alpha, \beta)\,,$$

where the parameters α and β are chosen, such that the first two moments of this gamma distribution are matched to the first two moments of $m \cdot \hat{\mathrm{HSIC}}$ under the independence hypothesis:

$$\alpha = \frac{(\mathbf{E}\,\hat{\mathrm{HSIC}})^2}{\mathrm{var}\,\hat{\mathrm{HSIC}}}$$

$$\beta = m \cdot \frac{\mathrm{var}\,\hat{\mathrm{HSIC}}}{\mathbf{E}\,\hat{\mathrm{HSIC}}}$$

The moments of $\hat{\mathrm{HSIC}}$ can be estimated efficiently (computable in $O(m^2)$) with a negligible bias [32]:

$$\hat{\mathbf{E}} = \frac{1}{m} + \frac{1}{m^4(m-1)^2} \sum_{i<j} k_{ij} \sum_{i<j} l_{ij} - \frac{1}{m^2(m-1)} \sum_{i<j} k_{ij} - \frac{1}{m^2(m-1)} \sum_{i<j} l_{ij}$$

$$\hat{\mathrm{var}} = \frac{2(m-4)(m-5)}{m(m-1)(m-2)(m-3)} \mathbf{1}^t \, (B - \mathrm{diag}(B))\, \mathbf{1}$$

where $B = ((HKH). \cdot (HLH)).^2$. Here, $A. \cdot B$ and $A.^2$ denote entrywise operations between matrices.

We now summarize

Theorem 3.44 *[Independence Test based on HSIC] Let* $(X_1, Y_1), \ldots, (X_m, Y_m)$ *be independent and identically distributed according to* $\mathbf{P}^{(X,Y)}$*. We can test the hypothesis*

$$\begin{aligned} H_0 &: \quad X \perp\!\!\!\perp Y \qquad \textit{against} \\ H_1 &: \quad X \not\perp\!\!\!\perp Y \end{aligned}$$

with a significance level of α *by using two kernels* k *and* l *satisfying the conditions of Theorem 3.41 (e.g. Gaussian kernels). Compute the statistic*

$$\hat{HSIC} = \frac{1}{m^2} trace(KHLH)\,,$$

where $K_{ij} = k(X_i, X_j)$*,* $L_{ij} = l(X_i, X_j)$ *and* $H = 1 - \frac{1}{m}\mathbf{1} \cdot \mathbf{1}^t$ *and define the decision function*

$$d(X_1, \ldots, X_n, Y_1, \ldots Y_m) = \begin{cases} H_0, & \hat{HSIC} \leq c \\ H_1, & \hat{HSIC} > c \end{cases} .$$

Use the fact that $m \cdot \hat{HSIC}$ is approximately $\Gamma(\alpha,\beta)$ distributed with

$$\alpha = \frac{\hat{\mathbf{E}}^2}{\hat{var}}$$
$$\beta = m \cdot \frac{\hat{var}}{\hat{\mathbf{E}}}$$

and choose c such that the type 1 error is bounded by the significance level α.

4 Causal Inference on Time Series

We now consider the problem already mentioned in the introduction: We are given some observations $X_1, \ldots, X_T$ of a (real-valued) time series, but we do not know if this sample has been reversed. That means we do not know if $X_1, \ldots, X_T$ or if $X_T, \ldots, X_1$ represents the true time direction. This can occur in the following situation: We receive the time series sample on a sheet of paper written by an Israeli. Fortunately, he used Arabic numerals, but unfortunately we do not know if he started writing on the right or on the left.
Thus we are looking for an algorithm, which can distinguish between the true and the reversed time direction based on a finite sample (see Figure 4.1).

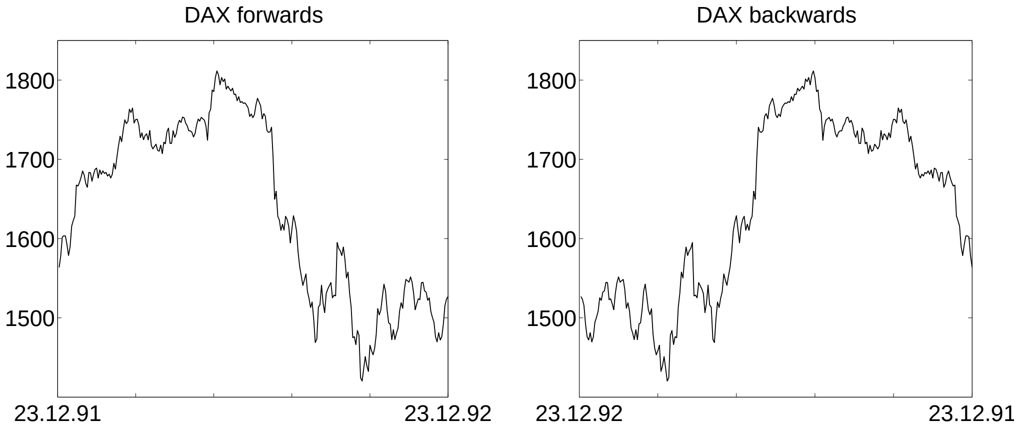


Figure 4.1: DAX values between 23.12.1991 and 23.12.1992. The left panel shows the true time direction, the right panel the reversed one. This is one of the examples, for which our ARMA method was able to identify the correct direction.

Notice that this a hard problem and we surely have to make some restrictions on the class of considered time series in order to be able to identify the true time direction.
In this section we propose two methods, one using SVMs and one using an ARMA model for time series.

4.1 Learning the Time Direction using SVMs

For this method we apply Support Vector Machines in different ways in order to distinguish between the two directions of time. Therefore assume we are given a strictly stationary time series $(X_t)_{t\in\mathbb{Z}}$. Strictly stationary means that the distribution of $(X_{t_1+h}, \ldots, X_{t_n+h})$ does not depend on h (see Definition 4.3 below). We further assume that there is a difference in the finite-dimensional distributions $(X_t, X_{t+1}, \ldots, X_{t+w})$ and $(X_{t+w}, \ldots, X_{t+1}, X_t)$. If all of these distributions were the

same, we could not detect a difference.
In the SVM approach we do not further investigate this difference and try to learn the nature of it by training an SVM on many data from time series, for which we know the true direction. As a *naive method* we just use an SVM on fixed sized windows of the time series. For the second approach (the *SVM-RKHS method*) we construct an SVM on the finite-dimensional distributions of the time series. Therefore we embed these distributions into an RKHS (see Section 4.1.1) and try to separate the points in the RKHS by a standard SVM, i.e. linearly. This is extended to a non-linear separation in the *SVM-RKHS-PCA method*, in which we choose the same embedding for the distributions, but we first apply a principal component analysis (Section 4.1.2) to the data points and perform an SVM on the new coordinates.
These methods are described in more detail in Section 4.1.3.

4.1.1 Hilbert Space Embeddings of Sample Distributions

Recall that for some kernels (like the Gaussian kernel) we have an injective embedding of probability measures into an RKHS via the mapping

$$\mu[\mathbf{P}^X] = \mathbf{E}k(X, .) \,.$$

Notice that even if we use a kernel, which does not satisfy the conditions necessary for the embedding to be injective, this embedding can still be useful. If we use a polynomial kernel of degree d on $\mathbb{R}$, for example, the embedding will be injective only on a smaller class of distributions. If two distributions coincide in the first d moments, for example, they will be indistinguishable in the RKHS.

Now assume we are given an iid sample $(X_1, \ldots, X_m)$ of the distribution $\mathbf{P}^X$. We know that we can estimate the distribution function F of X by its sample estimtate

$$\hat{F}_m(t) = \frac{1}{m} \sum_{i=1}^{m} 1_{X_i \le t} \,.$$

If $(X_1, \ldots, X_m)$ takes the value $(x_1, \ldots, x_m)$, this corresponds to a measure $\hat{\mathbf{P}}^X_m$ that has mass $\frac{1}{m}$ on each observed value x_i. Replacing $\mathbf{P}^X$ by $\hat{\mathbf{P}}^X_m$ leads to the following sample estimate of the mean element in the RKHS:

$$\hat{\mu}[\mathbf{P}^X] \;:=\; \mu[\hat{\mathbf{P}}^X_m] \;=\; \frac{1}{m} \sum_{i=1}^{m} k(x_i, .) \,. \tag{4.1}$$

It turns out that these representations are unique in the following sense[1]:

Proposition 4.1 *Let k be a strictly positive definite kernel. Then*

$$\frac{1}{m} \sum_{i=1}^{m} k(x_i, .) = \frac{1}{n} \sum_{j=1}^{n} k(\tilde{x}_j, .) \quad \Leftrightarrow \quad m = n \;\; \textit{and} \;\; x_i = \tilde{x}_{\sigma(i)} \quad \forall i = 1, \ldots, m \,,$$

for a permutation $\sigma \in S_m$.

[1] B. Schölkopf, MPI Tuebingen, told me this remark in a personal discussion.

Proof We prove the even more general case

$$\sum_{i=1}^{m} \alpha_i k(x_i, .) = \sum_{j=1}^{n} \beta_j k(\tilde{x}_j, .) \quad \Rightarrow \quad m = n \text{ and } x_i = \tilde{x}_{\sigma(i)} \quad \forall i = 1, \ldots, m$$

Wlog it is enough to show that there is a $j \in \{1, \ldots, m\}$, such that $x_1 = \tilde{x}_j$. Assume $x_1 \neq \tilde{x}_j \ \forall j$. Then we can rewrite the left hand side as

$$\sum_{i=1}^{\max(m,n)} \gamma_i k(y_i, .) = 0 \,,$$

where y_i are distinct values (either x_i or $\tilde{x}_i$) and $y_1 = x_1$, $\gamma_1 = \alpha_1$. Taking the norm yields

$$\sum_{i,j=1}^{\max(m,n)} \gamma_i \gamma_j k(y_i, y_j) = 0 \,,$$

which is contrary to the strictly positive definite kernel. □

This proposition shows that a single point in the RKHS contains all information about the whole sample. This statement, however, is not surprising, since we already know that the embedding of distributions is injective under some conditions on the measures and on the kernel. Here we showed that we do not even need these additional conditions if the kernel is strictly positive definite.

Notice that for the two SVM-RKHS approaches we want to apply an SVM (or a Principal Component Analysis (PCA), respectively) to these points in the RKHS. We have already seen that we only need the dot product matrix $\langle \phi_i, \phi_j \rangle$ of the considered points ϕ_i in order to perform an SVM. As we will see below (Section 4.1.2) the same is true for PCA. Thus we still have to compute the pairwise dot products of the points $\phi_x = \frac{1}{m} \sum_{i=1}^{m} k(x_i, .)$ in the RKHS. This is done as follows

$$\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, .), \frac{1}{n} \sum_{j=1}^{n} k(\tilde{x}_j, .) \rangle = \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} \langle k(x_i, .), k(\tilde{x}_j, .) \rangle = \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, \tilde{x}_j) \,. \tag{4.2}$$

4.1.2 PCA

Let $\mathbf{X}$ be a vector of random variables in $\mathcal{L}^2$ with mean zero and covariance matrix Σ. For m samples $\mathbf{x}_1, \ldots \mathbf{x}_m$ with sample mean zero, the sample covariance matrix is defined as $\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T$. In standard PCA we consider the eigenvalue decomposition of the sample covariance matrix (which is symmetric and therefore has only real eigenvalues). The eigenvectors are called principal components and are usually ordered according to the eigenvalues. The first principal component corresponds to the largest eigenvalue. From

$$\hat{\Sigma} v = \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{x}_i, v \rangle \mathbf{x}_i$$

it is obvious that $\hat{\Sigma}$ is positive (semi-)definite and all eigenvalues are non-negative. Further we notice that all eigenvectors v with $\lambda > 0$ lie in span$(\mathbf{x}_1, \dots, \mathbf{x}_m)$. It follows that all vectors in this span are eigenvectors of Σ to the eigenvalue λ if and only if

$$\lambda \langle \mathbf{x}_i, v \rangle = \langle \mathbf{x}_i, \hat{\Sigma} v \rangle \qquad \forall i = 1, \dots, m$$

This can be seen as follows: Construct an orthonormal basis $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{\tilde{m}}$ of the span using Gram-Schmidt, for example. We then have

$$\lambda \langle \tilde{\mathbf{x}}_i, v \rangle = \langle \tilde{\mathbf{x}}_i, \hat{\Sigma} v \rangle \qquad \forall i = 1, \dots, \tilde{m}$$

and it follows $\lambda v = \hat{\Sigma} v$.

Principal component analysis can also be done in an RKHS (see [19]). Therefore we consider m points $\phi_1, \dots, \phi_m$ in the Hilbert space and again we assume that they are centered: $\sum_{i=1}^m \phi_i = 0$. Define the covariance operator as

$$\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} \langle \phi_i, . \rangle \phi_i \, .$$

Since the Hilbert space may be infinite-dimensional, we cannot necessarily write this in terms of matrices. Again this operator is positive ($\langle \hat{\Sigma} v, v \rangle \geq 0$) and thus all eigenvalues λ are non-negative. Notice further that all non-zero eigenvalues can again be written as a linear combination of the ϕ_j:

$$v = \sum_{j=1}^{m} \alpha_j \phi_j \, . \tag{4.3}$$

Furthermore (with the same argumentation as above), the eigenvalue equation reduces to

$$\lambda \langle \phi_i, v \rangle = \langle \phi_i, \hat{\Sigma} v \rangle \qquad \forall i = 1, \dots, m \, .$$

Using (4.3) yields

$$\lambda \sum_{j=1}^{m} \alpha_j \langle \phi_i, \phi_j \rangle = \frac{1}{m} \sum_{j=1}^{m} \alpha_j \langle \phi_i, \sum_{k=1}^{m} \phi_k \langle \phi_k, \phi_j \rangle \rangle \qquad \forall i = 1, \dots, m \, ,$$

which reduces to

$$m \lambda K \alpha = K^2 \alpha \tag{4.4}$$

if we write K for the Gram matrix $K_{kl} := \langle \phi_k, \phi_l \rangle$ and α for $(\alpha_1, \dots, \alpha_m)^t$. This eigenvalue problem is equivalent to

$$m \lambda \alpha = K \alpha \tag{4.5}$$

since we can show (see e.g. the appendix of *Kernel Principal Component Analysis* in [39]) that all solutions α which satisfy (4.4) but not (4.5) are of the form $K\alpha = 0$. And for such a solution, however, we would have

$$\langle \phi_i, \sum_{j=1}^{m} \alpha_j \phi_j \rangle = (K\alpha)_i = 0 \qquad \forall i \, ,$$

which shows that we are not interested in it: It corresponds to a vector $v = \sum_{j=1}^{m} \alpha_j \phi_j$, which does not lie in span$(\phi_1, \dots, \phi_m)$. Conversely, it is obvious that all solutions of (4.5) satisfy (4.4).

Thus we have seen that performing a PCA in an RKHS basically reduces to diagonalizing the Gram matrix (4.5). Let $\alpha^{(i)}$ denote the eigenvector corresponding to the eigenvalue $m\lambda_i$. Note that we still have to normalize the solutions to ensure that the principal components have length one:

$$1 = \left\langle \sum_{j=1}^{m} \alpha_j^{(i)} \phi_j, \sum_{j=1}^{m} \alpha_j^{(i)} \phi_j \right\rangle = \langle \alpha^{(i)}, K\alpha^{(i)} \rangle = m\lambda_i \langle \alpha^{(i)}, \alpha^{(i)} \rangle \,.$$

The projection of data point ϕ_k onto the i-th principal component $\sum_{j=1}^{m} \alpha_j^{(i)} \phi_j$ is easily computed:

$$\langle \phi_k, \sum_{j=1}^{m} \alpha_j^{(i)} \phi_j \rangle = (K\alpha^{(i)})_k \,.$$

Similarly, we can compute the expansion in the principal components for a new point ψ in the RKHS:

$$\langle \psi, \sum_{j=1}^{m} \alpha_j^{(i)} \phi_j \rangle = \sum_{j=1}^{m} \alpha_j^{(i)} \langle \psi, \phi_j \rangle \,.$$

4.1.3 The SVM Method

Naive Method As a first idea we use a kernel SVM on a finite subset of the time series data with fixed length. For testing its performance consider 200 time series, for example. We first take 100 consecutive values out of each time series and then choose our training set (say 180 out of the 200 samples). Then we have 360 training points, each of which is labelled as +1 or −1 depending if they represent the true direction or if they are reversed samples. We then train the SVM and evaluate the predictions it makes on the test set consisting of the last 40 time series.

This method cannot be expected to work well: We expect that the difference between forward and backward going time can be found in the finite-dimensional distributions of the time series. The corresponding information does lie in the first 100 data points, but in a subtle way. Therefore the naive SVM is more unlikely to pick up this information than the following SVM methods, which are adjusted to the finite-dimensional distributions. In Machine Learning the way the data is presented to the machine matters.

We annihilated the linear trend of the time series because we did not want the SVM adapt to this feature. It is improbable, however, that the naive SVM finds any relevant features needed for the distinction between forwards and backwards going time.

SVM-RKHS Method We learned in Section 3.3.3 that we can map a distribution of a random variable in an RKHS, such that all statistical properties are represented. If we have a finite sample of this variable, this mean element can be estimated by looking at the sample mean of the feature maps (cf (4.1)). In Section 4.1.1 we have seen that the mapping is one-to-one in the following sense: if two function values in the RKHS are the same then the samples are of the same size and consist of exactly the same points. Therefore we can say that these Hilbert space representations inherit all relevant statistical information of the finite sample in input space. Now we want to apply this idea to the finite-dimensional distributions of a time series. Since we can compute the pairwise dot products of these points we are able to perform a linear SVM in the RKHS. This approach can be summarized as follows:

1. Choose a fixed window length w and take for each time series many finite-dimensional samples

$$\begin{aligned}\mathbf{x}_{t_1} &= (X_{t_1}, X_{t_1+1}, \ldots, X_{t_1+w})\\ \mathbf{x}_{t_2} &= (X_{t_2}, X_{t_2+1}, \ldots, X_{t_2+w})\\ &\vdots\\ \mathbf{x}_{t_m} &= (X_{t_m}, X_{t_m+1}, \ldots, X_{t_m+w})\,.\end{aligned}$$

   The t_i can be chosen in a way, such that $t_{i+1} - (t_i + w) = \text{const}$, for example. The larger this gap between two samples of the time series is, the less dependent these samples will be (ideally, we would like to have iid data, which is, of course, impossible for time series). Represent the distribution of $(X_t, \ldots, X_{t+w})$ in the RKHS using the point

$$\frac{1}{m}\sum_{i=1}^{m} k(\mathbf{x}_{t_i}, .)\,.$$

2. Perform a (linear) soft margin SVM on these points (one for each time series) using (4.2).

This procedure should not be confused with the usual kernel SVM, which is fundamentally different.

SVM-PCA-RKHS Method The SVM-RKHS method just mentioned is doing an SVM on sample representations of the finite-dimensional distributions in the RKHS. Although the RKHS may be infinite-dimensional, the Support Vector Machine still performs a linear classification. It may be the case, however, that the vectors in the RKHS cannot be separated linearly. The goal of this last SVM method is to do a non-linear classification within the RKHS. This can be done using principal component analysis (PCA). Therefore we determine the principal components (in the RKHS), project all data points on the most important directions and do a usual kernel SVM classification on these coefficients. In Section 4.1.2 we have given a short review of standard PCA and showed, how it can be implemented in an RKHS. To summarize this method:

1. As above, represent each time series in the RKHS using the point

$$\frac{1}{m}\sum_{i=1}^{m} k(\mathbf{x}_{t_i}, .)\,.$$

2. Perform a PCA on these points (one for each time series) using (4.2) and expand the points with respect to the principal components: for each time series you get a vector of coefficients.

3. Discard all principal components with eigenvalue smaller than a threshold, such that you remain with shorter coefficient vectors (of length 10, say).

4. Perform a kernel SVM on these coefficient vectors.

Remark 4.2 Note that the SVM-RKHS and the SVM-RKHS-PCA method are both possibilities of performing an SVM on probability distributions and therefore are interesting concepts in itself. They combine the idea of embedding distributions into an RKHS and performing an SVM. We have not heard that this idea has been used before.

4.2 Learning the Time Direction using ARMA Models

We first introduce the concept of ARMA processes. Later we show how they can be used to distinguish between the two time directions.

4.2.1 Time Series Analysis

Time series are stochastic processes indexed over $\mathbb{Z}$, which means they are a (countable) collection of random variables. Throughout the whole section we consider only non-degenerate random variables, that means random variables, for which there is no $a \in \mathbb{R}$, such that its distribution function can be written as

$$F(x) = 1_{x \geq a}(x)\,.$$

We now give some basic definitions and important results.

Definition 4.3

- A *time series* is a family of random variables $(X_t)_{t\in\mathbb{Z}}$ over a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.
- A time series $(X_t)_{t\in\mathbb{Z}}$ is called *strictly stationary* if

$$(X_{t_1}, \dots, X_{t_k}) \overset{d}{=} (X_{t_1+h}, \dots, X_{t_k+h}) \qquad \forall k, t_1, \dots, t_k, h \in \mathbb{Z}\,.$$

- A time series $(X_t)_{t\in\mathbb{Z}}$ is called *weakly (or second-order) stationary*[2] if $X_t \in \mathcal{L}^2$ and

$$\mathbf{E}X_t = \mu \qquad \text{and} \qquad \operatorname{cov}(X_t, X_{t+h}) = \gamma_h \quad \forall t, h \in \mathbb{Z}\,,$$

  i.e., both mean and covariance do not depend on the time t, but the latter only depends on the time gap h. $h \mapsto \gamma_h$ is called the *auto-covariance function*.
- A time series $(\epsilon_t)_{t\in\mathbb{Z}}$ is called a *white noise* process if $\epsilon_t \in \mathcal{L}^2$, $\mathbf{E}\epsilon_t = 0$ and

$$\operatorname{cov}(\epsilon_t, \epsilon_{t+h}) = 0 \quad \forall h \in \mathbb{Z}$$

  .
- A time series $(\epsilon_t)_{t\in\mathbb{Z}}$ is called an *iid white noise* process if ϵ_t is iid.

Definition 4.4

- A time series $(X_t)_{t\in\mathbb{Z}}$ is called a *moving average* process of order q and we write MA(q) if it is weakly stationary and if

$$X_t = \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \qquad \forall t \in \mathbb{Z}\,,$$

  for iid white noise $\epsilon_t \in \mathcal{L}^2$.

[2] In the literature sometimes the prefix *weakly* or *strictly* is omitted; we do not adapt to this notation in order to avoid confusion.

- A time series $(X_t)_{t\in\mathbb{Z}}$ is called an *auto-regressive* process of order p and we write AR(p) if it is weakly stationary and if

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t \qquad \forall t \in \mathbb{Z}\,,$$

  for iid white noise $\epsilon_t \in \mathcal{L}^2$
- A time series $(X_t)_{t\in\mathbb{Z}}$ is called an *auto-regressive moving average* process of order (p, q) and we write ARMA(p, q) if it is weakly stationary and if

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \qquad \forall t \in \mathbb{Z}\,,$$

  for iid white noise $\epsilon_t \in \mathcal{L}^2$
- A time series $(X_t)_{t\in\mathbb{Z}}$ is called an auto-regressive *integrated* moving average process of order (p, q, d) and we write ARIMA(p,q,d) if $\Delta^d X_t$ is an ARMA(p,q), where $\Delta X_t = X_t - X_{t-1}$.

Define the backward shift operator B via $B^j X_t = X_{t-j}$ in order to simplify the notation in the definitions above. The equation for an ARMA process, for example, simplifies to

$$\phi(B) X_t = \theta(B)\epsilon_t \qquad \forall t \in \mathbb{Z}\,,$$

where $\phi(\mathrm{z}) = 1 - \phi_1 z - \ldots - \phi_p z^p$ and $\theta(\mathrm{z}) = 1 + \theta_1 z + \ldots + \theta_q z^q$.

Note that in the literature ARMA processes are sometimes defined without the iid assumption of the noise, that means they only require white noise processes.

The following remark helps us to determine the auto-covariance function of an AR process:

Remark 4.5 Assume that (X_t) is an AR(p) process. That means for all $t \in \mathbb{Z}$

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t\,.$$

Considering $\mathrm{cov}(X_t, X_{t+k}),\ k \geq 1$ yields the so-called Yule-Walker equations:

$$\gamma_k = \phi_1 \gamma_{k-1} + \ldots + \phi_p \gamma_{k-p}\,.$$

For the special case of an AR(1) process we have $\gamma_k = \phi_1 \gamma_{k-1}$. With

$$\gamma_0 = \mathrm{cov}(X_t, X_t) = \phi_1^2 \gamma_0 + \sigma^2$$

it follows

$$\gamma_0 = \frac{\sigma^2}{1 - \phi_1^2} \quad \text{and} \quad \gamma_k = \frac{\phi_1^k \cdot \sigma^2}{1 - \phi_1^2}\,.$$

Remark 4.6 Now we consider ARMA processes with additional constraints on the coefficients. The following arguments (mainly given by [40]) show why these restrictions can be regarded as natural:

- $\phi(z)$ and $\theta(z)$ do not have common zeros
  Assume there is at least one common zero. If none of the zeros lie on the unit circle, X_t is the unique weakly stationary solution of the ARMA equation, in which all common factors are cancelled. If one of the common zeros lie on the unit circle, the ARMA equation may have more than one weakly stationary solution.
- $\phi(z)$ does not have a zero on the unit circle
  If it did and additionally there are no common zeros of $\phi(z)$ and $\theta(z)$, it can be shown that the ARMA equation has no weakly stationary solution at all. A simple example for this is the equation $X_t = X_{t-1} + \epsilon_t$. Considering the variance of this process, it is clear that there is no such thing as an AR(1) process with $\phi = 1$.

Further, it is natural to consider processes for which the noise is independent of the last values of the time series; that means for every point in time there is an additive random shock, which does not depend on the last values of the time series:

Definition 4.7 An ARMA(p, q) process satisfying $\phi(B)X_t = \theta(B)\epsilon_t$ is called *causal* if

$$\epsilon_t \perp\!\!\!\perp (X_{t-1}, \ldots, X_{t-h}) \qquad \forall h \geq 1\,. \tag{4.6}$$

Proposition 4.8 *For an ARMA(p, q) process satisfying $\phi(B)X_t = \theta(B)\epsilon_t$, where $\phi(z)$ and $\theta(z)$ have no common zeros, the following is equivalent*[3]*:*

(i) The process is causal.

(ii) There exists a sequence (ψ_i), such that $\sum_{i=0}^{\infty} |\psi_i| < \infty$ and

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}\,. \tag{4.7}$$

(iii) $\phi(z)$ does not have any zeros in the unit circle $|z| \leq 1$.

If this is the case, the coefficients ψ_i of (4.7) *are determined by*

$$\psi(z) = \sum_{i=0}^{\infty} \psi_i z^i = \frac{\theta(z)}{\phi(z)} \qquad |z| \leq 1$$

and the sum (4.7) *converges absolutely with probability one (Proposition 3.1.1 in [40]). Furthermore* (4.7) *is the unique weakly stationary solution of $\phi(B)X_t = \theta(B)\epsilon_t$.*

It is important that (ii) is not a property of the process X_t alone, but rather of the relationship between X_t and ϵ_t.

Proof
- $(i) \Leftarrow (ii)$:
  Wlog let $h = 1$. Define

$$X_t^{(n)} := \sum_{k=1}^{n} \psi_i \epsilon_{t-k}\,.$$

[3] Note that in [40] causal processes are actually defined as those satisfying condition (ii).

Because cos and sin are bounded continuous functions, we know by the definition of weak convergence that the characteristic functions are converging pointwise:

$$\varphi_{\mathbf{P}^{X_t^{(n)}}}(s) \longrightarrow \varphi_{\mathbf{P}^{X_t}}(s) \qquad \forall s \in \mathbb{R}\,.$$

Further $(X_t^{(n)}, \epsilon_{t+1}) \xrightarrow{d} (X_t, \epsilon_{t+1})$ holds, since $X_t^{(n)} \xrightarrow{\mathbf{P}\text{-a.s.}} X_t$ and thus $(X_t^{(n)}, \epsilon_{t+1}) \xrightarrow{\mathbf{P}\text{-a.s.}} (X_t, \epsilon_{t+1})$. Then

$$\begin{aligned}
\varphi_{\mathbf{P}^{(\epsilon_{t+1}, X_t)}}(u, v) &= \lim_{n\to\infty} \varphi_{\mathbf{P}^{(\epsilon_{t+1}, X_t^{(n)})}}(u, v) \\
&= \lim_{n\to\infty} \varphi_{\mathbf{P}^{\epsilon_{t+1}}}(u) \cdot \varphi_{\mathbf{P}^{X_t^{(n)}}}(v) \\
&= \varphi_{\mathbf{P}^{\epsilon_{t+1}}}(u) \cdot \lim_{n\to\infty} \varphi_{\mathbf{P}^{X_t^{(n)}}}(v) \\
&= \varphi_{\mathbf{P}^{\epsilon_{t+1}}}(u) \cdot \varphi_{\mathbf{P}^{X_t}}(v) \\
&= \varphi_{\mathbf{P}^{\epsilon_{t+1}} \otimes \mathbf{P}^{X_t}}(u, v)
\end{aligned}$$

and because of the uniqueness of characteristic functions we have that ϵ_{t+1} and X_t are independent.

- $(i) \Rightarrow (ii)$:
  We know (e.g. Theorem 3.1.3 in [40], Laurent series expansion) that X_t can be written as
  $$X_t = \sum_{i\in\mathbb{Z}} \psi_i \epsilon_{t-i}\,. \tag{4.8}$$
  We have to show that for causal processes all of the $\psi_i, i < 0$ are zero. If $\psi_{i_0} \neq 0$ for $i_0 < 0$, it follows that
  $$\psi_{i_0}\epsilon_{t-i_0} + \sum_{i\in\mathbb{Z}-i_0} \psi_i \epsilon_{t-i} = X_t \;\perp\!\!\!\perp\; \psi_{i_0}\epsilon_{t-i_0}\,, \tag{4.9}$$
  where $\psi_{i_0}\epsilon_{t-i_0}$ and $\sum_{i\in\mathbb{Z}-0} \psi_i \epsilon_{t-i}$ are independent (same reasoning as above). Thus (4.9) contradicts Lemma 2.7.
- $(ii) \Leftrightarrow (iii)$:
  This is shown as Theorem 3.1.1. in [40].

□

Above we have considered processes, which have finite variance and which are weakly stationary. Of course processes with finite variance and strict stationarity are just special cases. It is possible, however, to extend the last result to strictly stationary processes, which do not require a finite variance. In order to ensure strict stationarity we consider so-called Levy skew stable (or α-stable) distributions (see Section 13.3. in [40]):

Definition 4.9 A random variable Z has a *Levy skew stable* distribution if the characteristic function of Z has the form

$$\varphi_Z(t) = \begin{cases} \exp\left(it\mu - \frac{|ct|^\alpha}{2}\,(1 - i\beta\,\mathrm{sgn}(t)\tan(\pi\alpha/2)\,)\right) & \text{for } \alpha \neq 1 \\ \exp\left(it\mu - \frac{|ct|^\alpha}{2}\,(1 + i\beta\,\mathrm{sgn}(t)\ln|t|\,)\right) & \text{for } \alpha = 1 \end{cases},$$

where the exponent α lies in $[0,2]$, the skewness parameter η in $[-1,1]$ and the scale parameter c in $\mathbb{R}$. For $\alpha = 2$ we obtain a Gaussian distribution, for $\alpha = 1$ and $\beta = 0$ a Cauchy distribution.

These distributions have the following stability property [40]:
A random variable Z is Levy stable if and only if there exist $a_n > 0$ and $b_n \in \mathbb{R}$, such that

$$Z_1 + \ldots + Z_n \quad \overset{d}{=} \quad a_n Z + b_n$$

for all $Z_1, \ldots, Z_n, Z \overset{\text{iid}}{\sim}$ Levy stable.
Now we extend the definition of ARMA processes:

Definition 4.10 A process (X_t) is called a *strictly stationary ARMA(p, q)* process if the noise ϵ_t is Levy stable distributed and the process satisfies

$$\phi(B)X_t = \theta(B)\epsilon_t \, .$$

Again, under some conditions on ϕ and θ, we can write the process as a general linear process (see Proposition 13.3.2 in [40]):

Theorem 4.11 *Let (X_t) be a strictly stationary ARMA(p, q), which satisfies $\phi(B)X_t = \theta(B)\epsilon_t$, where $\phi(z)$ and $\theta(z)$ have no common zeros. If $\phi(z)$ does not have any zeros in the unit circle $|z| \leq 1$,*

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$$

is the unique strictly stationary solution of $\phi(B)X_t = \theta(B)\epsilon_t$. The coefficients ψ_i are determined by

$$\psi(z) = \sum_{i=0}^{\infty} \psi_i z^i = \frac{\theta(z)}{\phi(z)} \quad |z| \leq 1 \, .$$

This extension to strictly stationary ARMA processes is necessary because in real data you often have noise with heavier tails than the Gaussian, which may not even have a finite variance. In our simulations we use Cauchy distributed noise as an example for a Levy stable distribution (with non-finite variance).

4.2.2 Reversibility of linear Time Series

In Section 2.2 we have already seen that linear causal relationships do not have to be reversible. In fact, the normal distribution turned out to be a necessary and sufficient condition for reversibility. One of the main theoretical results of this work is a corresponding statement for auto-regressive moving average processes.

Definition 4.12 We call a causal ARMA(p, q) process with $\phi(B)X_t = \theta(B)\epsilon_t$, *time-reversible* if it can also be written as a causal ARMA($\tilde{p}$,$\tilde{q}$) process in the different time direction, i.e. if there exist $\tilde{p}, \tilde{q}, \tilde{\phi}_1, \ldots, \tilde{\phi}_{\tilde{p}}, \tilde{\theta}_1, \ldots, \tilde{\theta}_{\tilde{q}}$ and a noise $\tilde{\epsilon}_t$, such that

$$X_t = \sum_{i=1}^{\tilde{p}} \tilde{\phi}_i X_{t+i} + \sum_{j=1}^{\tilde{q}} \tilde{\theta}_j \tilde{\epsilon}_{t+j} + \tilde{\epsilon}_t \qquad \tilde{\epsilon}_t \perp\!\!\!\perp (X_{t+1}, X_{t+2}, \ldots, X_{t+h}) \quad \forall h \, .$$

In the theoretical work [41] and [42] the authors call a strictly stationary process time-reversible if $(X_0, \dots, X_h)$ and $(X_0, \dots, X_{-h})$ are equal in distribution for all h. This notion is not appropriate for our purpose because, a priori, it could be that both forward and backward process both are ARMA processes even though they do not coincide in distribution. Nevertheless, their result that (mainly) only Gaussian ARMA processes are time-reversible is similar to the one we will prove, but as already said it is more restrictive, though.

For an AR(1) process

$$X_t = \phi_1 X_{t-1} + \epsilon_t$$

Theorem 2.10 of Section 2.2 shows that this process is only reversible for Gaussian noise. It is not straightforward to apply Theorem 2.11 to an AR(p) process

$$X_t = \sum_{i=1}^{q} \phi_1 X_{t-i} + \epsilon_t$$

because the sum does not only consist of independent random variables. In order to cope with this problem we first introduce a characterization of the normal distribution, which is a generalization of the Darmois-Skitovich theorem and then consider the MA(∞) representation of an ARMA process. Recall that the Darmois-Skitovich theorem tells us that if two different linear combinations of independent random variables are themselves independent then all summands are normally distributed. It turns out that this can be generalized to an infinite sum. This was first done by Mamai [43]:

Theorem 4.13 *Let $(X_t)_t$ be a sequence of independent random variables and assume that both $\sum_{i=1}^{\infty} a_i X_i$ and $\sum_{i=1}^{\infty} b_i X_i$ converge almost surely. Further suppose that the sequences $\{\frac{a_i}{b_i} : b_i \neq 0\}$ and $\{\frac{b_i}{a_i} : a_i \neq 0\}$ are bounded. If*

$$\sum_{i=1}^{\infty} a_i X_i \qquad \text{and} \qquad \sum_{i=1}^{\infty} b_i X_i$$

are independent, then each X_i, for which $a_i b_i \neq 0$, is normally distributed.

Before we can prove this theorem we need the following generalized version of Theorem 2.9, which was also given by Mamai [43] (see also Theorem 7.8 in [13]).

Theorem 4.14 *Let $f_1, f_2, \dots$ be a sequence of characteristic functions, which satisfy*

$$\prod_{i=1}^{\infty} f_i^{\alpha_i}(t) = f(t),$$

for some $\alpha_i > \alpha > 0$ and for all t in a neighbourhood of zero, where f is the characteristic function of a normal distribution. Then every f_i itself is the characteristic function of a normal distribution.

Proof [of Theorem 4.13] The core of the proof is the same as the one for Theorem 2.8. We have to extend all sums and products to infinity. If the series converge almost surely, it is obvious that the corresponding products of characteristic functions are converging pointwise. We even know that this convergence is uniformly in any finite interval. This can be seen, for example, in [44]. Therefore we have

$$\prod_{i=1}^{\infty} \varphi_i(a_i u + b_i v) = \prod_{i=1}^{\infty} \varphi_i(a_i u) \prod_{i=1}^{\infty} \varphi_i(b_i v) \,.$$

Now we cannot conclude as easily as in the proof of Theorem 2.8 that $\varphi_i(t) \neq 0$ for all $t \in \mathbb{R}$ (if one side vanishes, it does not imply that one of the factors does). Instead, we have to restrict ourselves to an interval around 0 and consider functions $\tilde{\varphi}_i(t) := \varphi_i(t)\varphi_i(-t) = |\varphi_i(t)^2|$, which are the characteristic functions of the random variables $Y_i := X_i - \tilde{X}_i$, where $\tilde{X}_i$ is an independent copy of X_i. These functions are always positive and bounded away from zero in an interval around the origin. Now we are able to consider logarithms (which are continuous!).

$$\sum_{i=1}^{\infty} \psi_i(a_i u + b_i v) = \sum_{i=1}^{\infty} \psi_i(a_i u) + \sum_{i=1}^{\infty} \psi_i(b_i v) =: A(u) + B(v)$$

where $\psi_i = \ln \tilde{\varphi}_i$. If Y_i turns out to be Gaussian, X_i is as well because of Cramér's theorem [16]. The rest is analogue to above if we take the uniform convergence into account, which justifies a term-by-term integration, and if we use Theorem 4.14 instead of Theorem 2.9. □

Now we are able to prove the following, central theorem:

Theorem 4.15 *Assume that (X_t) is a causal ARMA process with iid noise and non-vanishing AR part. Then the process is time-reversible if and only if the process is Gaussian distributed.*

Furthermore, if this is the case, the order of the process and the parameters stay the same: $\tilde{p} = p, \tilde{q} = q, \tilde{\phi}_i = \phi_i, \tilde{\theta}_j = \theta_j$ and even the variance of the Gaussian noise does not change.

Proof Although technically this is not necessary, we do the proof not only for the general case of an ARMA(p,q) process but also for the special cases of AR(1) and AR(p) processes with finite variance noise in order to achieve a better understanding.

- $\Leftarrow$:

1. AR(1)
Let σ^2 denote the variance of the iid white noise Gaussian process ϵ.
The reversibility is shown in Example 2.6. There we constructed a new noise $\tilde{\epsilon}$, such that

$$X_t = \frac{\phi}{\phi^2 + \sigma^2/\mathrm{var}(X_t)} X_{t+1} + \tilde{\epsilon}_t \,.$$

We have seen before that

$$\gamma_0 = \mathrm{var}(X_t) = \frac{\sigma^2}{1 - \phi_1^2} \quad \text{and} \quad \gamma_1 = \frac{\phi_1 \sigma^2}{1 - \phi_1^2} \,,$$

which implies

$$\tilde{\phi}_1 = \frac{\sigma^2/(1-\phi_1^2)}{\phi_1\sigma^2/(1-\phi_1^2)} = \phi_1 \,.$$

We know that $\tilde{\epsilon}_t \perp\!\!\!\perp X_{t+1}$, but technically we still have to check if $\tilde{\epsilon}_t$ and X_{t+k} are independent for $k \geq 2$:

$$\begin{aligned} \langle \tilde{\epsilon}_t, X_{t+k} \rangle &= \left\langle X_t - \frac{\langle X_{t+1}, X_t \rangle}{\|X_{t+1}\|^2} X_{t+1}, X_{t+k} \right\rangle \\ &= \gamma_k - \phi\gamma_{k-1} \\ &= 0 \end{aligned}$$

Now we can easily conclude

$$\begin{aligned} \langle \tilde{\epsilon}_t, \tilde{\epsilon}_{t+k} \rangle &= \langle \tilde{\epsilon}_t, X_{t+k} - \phi X_{t+k+1} \rangle \\ &= 0 \end{aligned}$$

for all $k \geq 1$. That means $(\tilde{\epsilon}_t)$ is a sequence of independent random variables. Furthermore

$$\begin{aligned} \text{var}(\tilde{\epsilon}_t) = \langle \tilde{\epsilon}_t, \tilde{\epsilon}_t \rangle &= \|X_{t+1}\|^2 - 2 \cdot \frac{\langle X_{t+1}, X_t \rangle^2}{\|X_{t+1}\|^2} + \langle X_{t+1}, X_t \rangle \\ &= \frac{\sigma^2 - 2\phi_1^2\sigma^2 + \sigma^2\phi_1^2}{1-\phi_1^2} = \sigma^2 \,, \end{aligned}$$

so the new noise $\tilde{\epsilon}_t$ has the same variance as the old one ϵ_t.

2. AR(p)
Again the reversibility was already mentioned in the proof of Theorem 2.11. We considered the projection of X_t on span$(X_{t+1}, \ldots, X_{t+p})$ and defined the new noise as being the difference between X_t and the projected vector. It remains to show, that the coefficients do not change. Therefore we use that a projection always minimises the distance between vector and projection space:

$$\begin{aligned} (\psi_1, \ldots, \psi_p) &= \text{argmin}_{\mathbf{a}} \|(a_1, \ldots, a_p)(X_{t+1}, \ldots, X_{t+p})^t - X_t\|^2 \\ &= \text{argmin}_{\mathbf{a}} \sum_{i,j=1}^{p} a_i a_j \text{cov}(X_{t+i}, X_{t+j}) - 2\sum_{i=1}^{p} a_i \text{cov}(X_t, X_{t+i}) \\ &= \text{argmin}_{\mathbf{a}} \sum_{i,j=1}^{p} a_i a_j \text{cov}(X_{t+p-i}, X_{t+p-j}) - 2\sum_{i=1}^{p} a_i \text{cov}(X_{t+p}, X_{t+p-i}) \\ &= \text{argmin}_{\mathbf{a}} \|(a_1, \ldots, a_p)(X_{t+p-1}, \ldots, X_t)^t - X_{t+p}\|^2 \\ &= (\phi_1, \ldots, \phi_p) \end{aligned}$$

The last step holds, since $\sum_{i=1}^{p} \phi_i X_{t+p-i} - X_{t+p} = -\epsilon_{t+p}$ and $\langle \epsilon_{t+p}, X_{t+p-i} \rangle = 0$, for all $i = 1, \ldots, p$.

Again, we notice that for all $k \geq p$

$$\begin{aligned}\langle \tilde{\epsilon}_t, X_{t+k} \rangle &= \langle X_t, X_{t+k} \rangle - \sum_{i=1}^{p} \phi_j \langle X_{t+i}, X_{t+k} \rangle \\ &= \gamma_k - \sum_{i=1}^{p} \phi_i \gamma_{k-i} \\ &= 0\end{aligned}$$

according to the Yule-Walker equations. As above it follows

$$\begin{aligned}\langle \tilde{\epsilon}_t, \tilde{\epsilon}_{t+k} \rangle &= \left\langle \tilde{\epsilon}_t, X_{t+k} - \sum_{i=1}^{p} X_{t+k+i} \right\rangle \\ &= 0\end{aligned}$$

It remains to check that the variance of the noise is not changing:

$$\begin{aligned}\langle \tilde{\epsilon}_t, \tilde{\epsilon}_t \rangle &= \left\| X_t - \sum_{i=1}^{p} \phi_i X_{t+i} \right\|^2 \\ &= \gamma_0 - 2 \sum_{i=1}^{p} \phi_i \gamma_i + \sum_{i,j=1}^{p} \phi_i \phi_j \gamma_{|i-j|} \\ &= \left\| X_{t+p} - \sum_{i=1}^{p} \phi_i X_{t+p+i} \right\|^2 \\ &= \sigma^2\end{aligned}$$

Notice, that in the whole proof the stationarity of the time series plays a crucial role.

3. ARMA(p, q)
Now we consider a Gaussian ARMA(p, q) process (X_t) and define

$$\overline{X} : \begin{array}{ccc} \Omega & \to & \mathbb{R}^{\mathbb{Z}} \\ \omega & \mapsto & \{t \mapsto X_t(\omega)\} \end{array} .$$

Recall that its finite-dimensional distributions $(X_{t_1}, \ldots, X_{t_d})$ are normally distributed and therefore they are characterized only by the mean and the covariance matrix. So the distribution of the whole process only depends on the mean function

$$s \mapsto \mu_{\overline{X}}(s) := \mathbf{E} X_s$$

and its covariance function

$$(s, t) \mapsto \operatorname{cov}_{\overline{X}}(s, t) := \operatorname{cov}(X_s, X_t) .$$

We define the backward process

$$\overline{Y} : \begin{array}{ccc} \Omega & \to & \mathbb{R}^{\mathbb{Z}} \\ \omega & \mapsto & \{t \mapsto X_{-t}(\omega)\} \end{array} ,$$

which is again normally distributed and has the covariance function

$$\mathrm{cov}_{\overline{Y}}(s,t) = \mathrm{cov}(X_{-s}, X_{-t}) = \gamma_{|-s-(-t)|} = \gamma_{|s-t|} = \mathrm{cov}_{\overline{X}}(s,t)\,.$$

Ergo

$$\overline{X} \overset{d}{=} \overline{Y}\,,$$

meaning that $\mathbf{P}(X_t = Y_t \ \forall t) = 1$ (X and Y are *indistinguishable*). For the forward direction there exists a way to construct a process $\overline{\epsilon} : \Omega \to \mathbb{R}^{\mathbb{Z}}$ as a function of $\overline{X}$, such that (X_t) together with the noise (ϵ_t) is a causal ARMA process with specific coefficients; or, phrasing it differently, such that the joint distribution of $(\overline{X}, \overline{\epsilon})$ satisfies certain conditions. The explicit construction of $\overline{\epsilon}$ can be done using $\epsilon_t = \sum_{j=-\infty}^{\infty} \pi_j X_{t-j}$ [40], but is not relevant for the following. Important is that we can construct a (possibly different) noise $\tilde{\overline{\epsilon}}$ from $\overline{Y}$ in exactly the same way. Because the distribution of $(\overline{X}, \overline{\epsilon})$ only depends on the distribution of $\overline{X}$ (this is due to the fact that $\overline{\epsilon}$ is just a function of $\overline{X}$), $(\overline{X}, \overline{\epsilon})$ and $(\overline{Y}, \tilde{\overline{\epsilon}})$ have the same properties.

- ⇒:

  1. AR(1)
     This was already shown in Theorem 2.10.
  2. AR(p)
     See the general case below.
  3. ARMA(p, q)

     By assumption, we have

$$X_t = \sum_{i=1}^{\tilde{p}} \tilde{\phi}_i X_{t+i} + \sum_{j=1}^{\tilde{q}} \tilde{\theta}_j \tilde{\epsilon}_{t+j} + \tilde{\epsilon}_t \qquad \forall t \in \mathbb{Z}\,.$$

Thus using (4.7) we can write

$$\begin{aligned}
\sum_{j=1}^{\tilde{q}} \tilde{\theta}_j \tilde{\epsilon}_{t-\tilde{p}+j} + \tilde{\epsilon}_{t-\tilde{p}} &= X_{t-\tilde{p}} - \sum_{j=1}^{\tilde{p}} \tilde{\phi}_j X_{t-\tilde{p}+j} \\
&= \left(\sum_{i=0}^{\infty} \psi_i \epsilon_{t-\tilde{p}-i} \right) - \left(\sum_{j=1}^{\tilde{p}} \tilde{\phi}_j \sum_{i=0}^{\infty} \psi_i \epsilon_{t-\tilde{p}+j-i} \right) \\
&= \sum_{i=0}^{\infty} \left(\psi_{i-\tilde{p}} - \sum_{j=1}^{\tilde{p}} \tilde{\phi}_j \psi_{i+j-\tilde{p}} \right) \epsilon_{t-i}\,,
\end{aligned}$$

where $\psi_i = 0$ for all $i < 0$. Additionally we have

$$X_{t-\tilde{p}+\tilde{q}+1} = \sum_{i=0}^{\infty} \psi_i \ \epsilon_{t-\tilde{p}+\tilde{q}+1-i} = \sum_{i=\tilde{q}-\tilde{p}+1}^{\infty} \psi_{\tilde{p}-\tilde{q}-1+i} \ \epsilon_{t-i}\,.$$

Both sums are converging absolutely with probability one (see Proposition 4.8) and by assumption, the left hand sides are independent of each other. Clearly, we want to apply Theorem 4.13, but therefore we need the boundedness condition to be satisfied. This we show in Lemma 4.16 below. Given the boundedness Theorem 4.13 implies that the noise ϵ_t is Gaussian distributed. (Note that we actually have some ϵ_t occurring on both sides because of the non-vanishing AR part.) Then X_t is Gaussian distributed, too: Define again

$$X_t^{(n)} := \sum_{i=1}^{n} \psi_i \epsilon_{t-i} \,.$$

We know that $(X_t^{(n)})_n$ is converging in $\mathcal{L}^2$. Ergo

$$\|X_t\|_2 - \|X_t^{(n)}\|_2 \leq \|X_t - X_t^{(n)}\|_2 \longrightarrow 0 \,,$$

and thus

$$\sigma_n^2 := \mathrm{var} X_t^{(n)} \longrightarrow \mathrm{var} X_t =: \sigma^2 \,.$$

Furthermore $(X_t^{(n)})_n$ converges in distribution and therefore the cumulative distribution functions F_n are converging pointwise:

$$F_n(x) = \Phi_{0,\sigma_n^2}(x) \longrightarrow \Phi_{0,\sigma^2}(x) = F(x)$$

and therefore X_t is Gaussian distributed.

□

Recall the proof of Theorem 4.15. It remains to show that the boundedness condition on the coefficients is satisfied:

Lemma 4.16 *For all possible causal backward models ARMA($\tilde{p}, \tilde{q}$) both*

$$\left| \frac{\psi_{\tilde{p}-\tilde{q}-1+i}}{\sum_{j=0}^{\tilde{p}} c_j \psi_{i+j-\tilde{p}}} \right| \qquad \textit{and} \qquad \left| \frac{\sum_{j=0}^{\tilde{p}} c_j \psi_{i+j-\tilde{p}}}{\psi_{\tilde{p}-\tilde{q}-1+i}} \right| \tag{4.10}$$

are bounded in i (see (4.7) for the coefficients ψ_i).

Here, $c_1 := -\tilde{\phi}_1, \ldots, c_{\tilde{p}} := -\tilde{\phi}_{\tilde{p}} \in \mathbb{R}$ and $c_0 := 1$.

Proof First we show for an example that this Lemma holds and then we generalize the arguments for a rigorous proof: Consider an ARMA(2, 1) process with the following coefficients: $\phi_1 = 1, \phi_2 = -0.25, \theta_1 = 1$. For this process we have $\psi_i = (1 + 3i)2^{-i}$ for all i (see Chapter 3.3 in [40]). Thus the first fraction reduces to

$$\left| \frac{(1 + 3(\tilde{p} - \tilde{q} - 1 + i)) \cdot 2^{\tilde{q}+1-\tilde{p}} \cdot 2^{-i}}{\sum_{j=0}^{\tilde{p}} c_j (1 + 3(j - \tilde{p} + i)) \cdot 2^{\tilde{p}-j} \cdot 2^{-i}} \right| \,.$$

The leading terms in i determine the limit behaviour of this fraction; thus for all $\tilde{p}$ and $c_1, \ldots, c_{\tilde{p}}$ it converges against

$$\left| \frac{3 \cdot 2^{\tilde{q}+1-\tilde{p}}}{\sum_{j=0}^{\tilde{p}} 3 \cdot c_j \cdot 2^{\tilde{p}-j}} \right|$$

as $i \to \infty$ and is therefore bounded in i.

A similar reasoning can be applied to general ARMA processes since we have the following expression for ψ_i (see Chapter 3.3 in [40]):

$$\psi_i = \sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, i^t\, \xi_s^{-i},$$

where $\alpha_{s,t}$ are some coefficients, ξ_s are the distinct (possibly complex) roots of $\phi(z)$ and T_s their multiplicity. Wlog assume that $\alpha_{s,T_s-1} \neq 0\ \forall s$. We can write the left fraction of (4.10) as

$$\begin{aligned} &\frac{\sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, (\tilde{p} - \tilde{q} - 1 + i)^t\, \xi_s^{-\tilde{p}+\tilde{q}+1-i}}{\sum_{j=0}^{\tilde{p}} c_j\, \sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, (i + j - \tilde{p})^t\, \xi_s^{-i-j+\tilde{p}}} \\ = \;& \frac{\sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, \xi_s^{-\tilde{p}+\tilde{q}+1} (\tilde{p} - \tilde{q} - 1 + i)^t\, \xi_s^{-i}}{\sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, \xi_s^{\tilde{p}} \sum_{j=0}^{\tilde{p}} c_j\, \xi_s^{-j} (i + j - \tilde{p})^t\, \xi_s^{-i}}\,. \end{aligned} \tag{4.11}$$

To investigate the limit behaviour we again consider only leading terms in i. More specifically, all summands are going to zero since $|\xi_s^{-1}| < 1$. The root ξ_{s_0} with the smallest modulus converges towards zero with the slowest rate and thus the corresponding summand determines the overall convergence. We divide both numerator and denominator of (4.11) by $i^{T_{s_0}-1}\, \xi_{s_0}^{-i}$ to see that the fraction converges towards

$$\left| \frac{\alpha_{s_0,T_{s_0}-1}\, \xi_{s_0}^{-\tilde{p}+\tilde{q}+1}}{\alpha_{s_0,T_{s_0}-1}\, \xi_{s_0}^{\tilde{p}} \sum_{j=0}^{\tilde{p}} c_j\, \xi_{s_0}^{-j}} \right|$$

for $i \to \infty$. This surely implies boundedness.

Note that the coefficient

$$\alpha_{s_0,T_{s_0}-1}\, \xi_{s_0}^{\tilde{p}} \sum_{j=0}^{\tilde{p}} c_j\, \xi_{s_0}^{-j}$$

does not vanish because this implies $\sum_{j=0}^{\tilde{p}} c_j\, \xi_{s_0}^{-j} = \tilde{\phi}(\xi_{s_0}^{-1}) = 0$. That means $\xi_{s_0}^{-1}$ is a root of $\tilde{\phi}(z)$, which is contrary to the restriction of a causal backward model ($|\xi_{s_0}| > 1$, cf Proposition 4.8). □

Remark 4.17 Note that in Theorem 4.15 we excluded all pure MA processes

$$X_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i}\,.$$

This is partly necessary because for some configurations of the coefficients θ_j, the process is time-reversible even for non-Gaussian distributions. In [41], for example, the author considers MA(q) processes, whose coefficients satisfy $\theta_j = \theta_{q-j},\ j = 0, \dots, q$, where $\theta_0 := 1$. He remarks that

$$(X_{t_1}, \dots, X_{t_n}) \overset{d}{=} (X_{-t_1}, \dots, X_{-t_n}),$$

which can be seen as follows (we need the symmetry of the coefficients for the second and the fourth equality)

$$\begin{aligned}
(X_{t_1}, \dots, X_{t_n}) &= \left(\sum_{i=0}^{q} \theta_i \epsilon_{t_1-i},\ \dots\ ,\ \sum_{i=0}^{q} \theta_i \epsilon_{t_n-i} \right) \\
&\overset{d}{=} \left(\sum_{i=0}^{q} \theta_i \epsilon_{-(t_1-i)},\ \dots\ ,\ \sum_{i=0}^{q} \theta_i \epsilon_{-(t_n-i)} \right) \\
&\overset{d}{=} \left(\sum_{i=0}^{q} \theta_i \epsilon_{-t_1-q+i},\ \dots\ ,\ \sum_{i=0}^{q} \theta_i \epsilon_{-t_n-q+i} \right) \\
&= \left(\sum_{i=0}^{q} \theta_i \epsilon_{-t_1-i},\ \dots\ ,\ \sum_{i=0}^{q} \theta_i \epsilon_{-t_n-i} \right) \\
&= (X_{-t_1}, \dots, X_{-t_n})
\end{aligned}$$

With the same reasoning as in the "$\Leftarrow$"-part of the proof of Theorem 4.15, we can now conclude that the time series is time-reversible for all distributions of ϵ_t satisfying the above symmetry constraints.

This shows, why at least some of the cases of pure MA processes have to be excluded from Theorem 4.15.

4.2.3 The ARMA Method

We use these theoretical results to propose a method which is able to detect the true direction of time series. The main idea of this method is based on Theorem 4.15: We assume that the time series (X_t) is a causal stationary ARMA process with non-Gaussian iid noise. Remember that the causality assumption means that noise and past values of the time series are independent. We showed that the reversed time series cannot be expressed as a causal stationary ARMA process.

Having this result we fit an ARMA process to both directions and test for independence between noise and past values of the time series. This can be done as a significance test. If we can reject the independence assumption only in one direction, we take the other direction as the true one. If we do not reject independence in any direction, the time series may be a Gaussian process and if we reject independence in both directions, our model assumption is probably wrong. In both of the latter cases we do not decide.

We now summarize the main steps of the ARMA method:

1. ARMA Fit

   Assume that the data come from a causal ARMA process with non-vanishing AR part

and with independent, non-Gaussian noise. Fit an ARMA process to both directions $(X_1, \ldots, X_T)$ and $(X_T, \ldots, X_1)$ and compute the fitted residuals.
The ARMA coefficients are fitted using a Maximum Likelihood approach, the exact Likelihood is computed by representing the ARMA process as a State Space Model and using a Kalman Filter. We do not give further details, but refer to Chapter 12 in [40]. In the experiments we used the implementation from R (*arima* with *method*=*"ML"*) for fitting the ARMA process. Moreover, we used the Akaike Information Criterion in order to determine the order of the ARMA process.

2. Normality Test
If the residuals seem to be Gaussian, i.e. the hypothesis of a normal distribution cannot be rejected, do not make a decision. In this work we used a test based on the skewness and the kurtosis of the distribution: the so-called Jarque-Bera test uses the test statistic

$$JB = \frac{m}{6}\left(s^2 + \frac{(k-3)^2}{4}\right),$$

where m is the number of samples, s is the skewness and k the kurtosis of the sample. Under the hypothesis of a normal distribution the test statistic JB follows a Chi-Square distribution with two degrees of freedom [45].

3. Independence Test
Using HSIC and a significance level of α test if ϵ_t depends on $X_{t-1}, X_{t-2}, \ldots$ or if $\tilde{\epsilon}_t$ depends on $X_{t+1}, X_{t+2}, \ldots$ and call the p-values of both tests p_1 and p_2, respectively. According to Theorem 4.15 only one dependence should be found. If the independence is indeed rejected for only one direction, i.e. exactly one p-value is smaller than α:

$$\min(p_1, p_2) < \alpha \qquad \text{and} \qquad \max(p_1, p_2) > \alpha$$

and additionally,

$$\max(p_1, p_2) - \min(p_1, p_2) > \delta,$$

then propose the direction of $\operatorname{argmax}(p_1, p_2)$ to be the correct one. See Figure 4.2 for an example.

4. If both directions seem to lead to dependent noise, conclude that the model fit is not good enough and do not decide.

Note that there are two parameters to choose:

- the minimal difference in p-values δ and
- the significance level α.

We expect that the larger δ and the smaller α is, the less decisions our algorithms makes, but the more accurate these decisions will be.

We further point out that this is a heuristic method. Although Theorem 4.15 is a theoretical justification for our approach, we cannot bound the probability of choosing the wrong direction, for example. Furthermore we need iid data for the independence test. The noise can be assumed to be iid, but the time series values cannot. Even if we assume strict stationarity we have that the values are identically distributed, but they cannot be regarded as being independent. To come over this problem, we do not consider all consecutive values of the time series, but introduce a gap instead; that means we take only every third value of the time series for the independence test, for example. This reduces the dependence between the samples, but does not completely annihilate it.

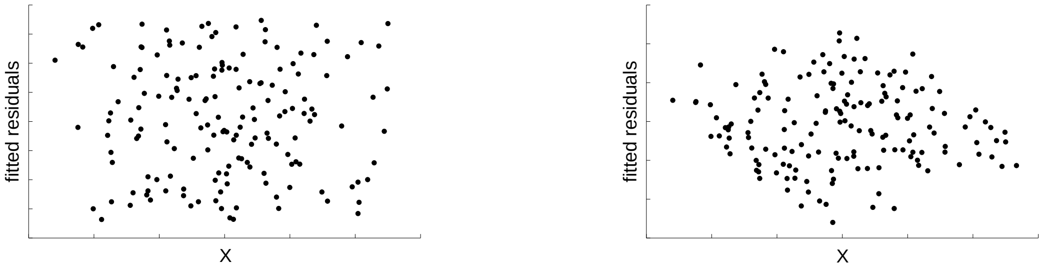


Figure 4.2: Simulated AR process with uniformly distributed noise: The fitted residuals of the forward model (left) and of the backward model (right) are plotted against past time series values. The fit in the wrong direction leads to a strong dependence between residuals and time series (p-value of 0.0008), the residuals of the forward model are regarded as independent (p-value of 0.8796).

5 Experiments

We applied the SVM and the ARMA method both to simulated ARMA processes and to real data. Mainly there are four different data sets:

- ARMA Processes (simulated)
  We simulated data from an ARMA$(2, 2)$ time series of length 500 with fixed parameters and varying kinds of noise. For different values of r we sampled the noise from

  $$\epsilon_t \sim \text{sgn}(Z) \cdot |Z|^r,$$

  where $Z \sim \mathcal{N}(0, 1)$. We then normalized it in order to obtain the same variance for all r. Only $r = 1$ corresponds to a normal distribution. For all samples the parameters were chosen to be $\phi_1 = 0.9, \phi_2 = -0.3, \theta_1 = -0.29$ and $\theta_2 = 0.5$.

- AR Processes (simulated)
  For this experiment we simulated AR(p) processes of length 500 and of different orders $(p = 1, \ldots, 5)$. Again, we used different kinds of noise (Gaussian, Laplace, Cauchy, Student-t and uniform). For each of these 25 combinations we simulated 100 time series, each of which had different parameters. These parameters were chosen randomly, but were constrained to fulfill the conditions of a causal ARMA process, of course. The noise was simulated with variance one, except for the Cauchy distribution.

- EEG Data (real)
  We used a publicly available EEG data set [46] consisting of 118 channels of a single subject. The sampling rate was 1000Hz and we considered the first 5 seconds of each channel, cut into 10 pieces. In total this gave 1180 time series of length 500.

- Mixed Collection of Time Series (real)
  We collected data consisting of 200 time series with varying length (from 100 up to 10,000 samples) from very different areas: finance, physics, transportation, crime, production of goods, demography, economy, EEG data and agriculture. Roughly two thirds of them belong to the groups economy and finance.

Once more we mention that in theory the ARMA method only works if the data follow an ARMA process with non-Gaussian noise. For stationary ARMA processes the SVM methods require non-Gaussian noise, too. If the noise were Gaussian, the backward process would be again a causal ARMA process and there would be no difference in the finite-dimensional distributions of the forward and the backward direction. Thus we expect our methods to fail for normally distributed noise. In many applications such a Gaussian distribution is assumed, but this is often done because of its nice computational properties rather than a consistency with the data. Using noise with heavier tails than the Gaussian would often be more appropriate (e.g. [47]).

5.1 SVM method

In the experiments the naive SVM method for classifying the direction of time series did not exceed chance level. As we mentioned before, we did not expect it to do so because it is more unlikely to adapt the important features relating to the finite-dimensional distributions. In the following subsection we only present the results from the SVM-RKHS method and the SVM-RKHS-PCA method. For these two methods we used a polynomial kernel of degree 4 for the Hilbert Space embedding.

Since we know the “true” time direction for all time series in the data sets we used the following procedure to test the SVM methods: We divided the data set randomly into training and test set, trained the SVM method on the training set and checked its performance on the test set. This was repeated many times in order to avoid misleading results due to particular easy test sets, for example.

ARMA Processes (simulated). For this experiment we consider an ARMA(2,2) process with coefficients $\phi_1 = 0.9, \phi_2 = -0.3, \theta_1 = -0.29$ and $\theta_2 = 0.5$. For each kind of noise (the noise is parameterized by r, only $r = 1$ corresponds to a Gaussian) we simulated 100 instances of this ARMA(2,2) process, divided these 100 time series into training set (85) and test set (15) and obtained an error rate of the SVM method on the test set.

We have seen before that the distributions of (X_t, X_{t+1}, X_{t+2}) and (X_{t+2}, X_{t+1}, X_t) coincide if and only if we consider a Gaussian distribution (this is a special case of Theorem 4.15). Thus we expect the method only to work for $r \neq 1$.

Since for each distribution of the noise all of the time series were simulated using the same coefficients we expect the finite-dimensional distributions to be similar over all of these 100 time series.

Notice, however, that the distributions used in this experiment –except for the Gaussian case– are not Levy stable and thus the ARMA process is not strictly stationary. This means that the finite-dimensional distributions of the time series vary over time. Assuming that the difference in distributions obtained by time shifts are small compared to the difference caused by a time inversion we still applied the SVM methods.

Figures 5.1 and 5.2 show that both SVMs learned indeed the true direction of this ARMA process provided the noise was sufficiently different from being a Gaussian. Although the SVM-RKHS-PCA method allows us to separate distributions in the RKHS non-linearly, it did not perform better than the linear SVM-RKHS method. This is mainly due to the following reason: Since a linear separation in the RKHS is seemingly possible, we do not gain very much from the non-linearity. Besides we have to face difficulties in regularization: even if we consider only few principal components (3 or 4), the SVM must be heavily regularized in order to avoid overfitting.

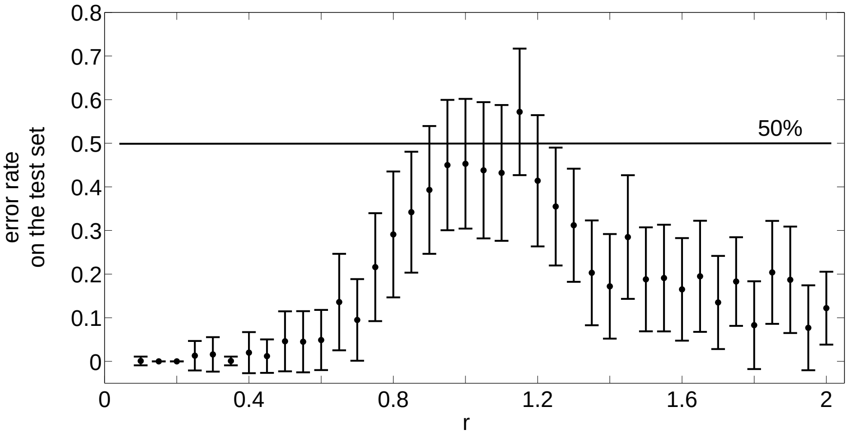


Figure 5.1: SVM-RKHS method on the ARMA processes. For each value of r (i.e. for each kind of noise) we simulated 100 instances of an ARMA(2,2) process with fixed coefficients and divided them into 85 time series for training and 15 for testing; this was done 100 times for each r. The graph shows the average classification error on the test set and the corresponding standard deviation.

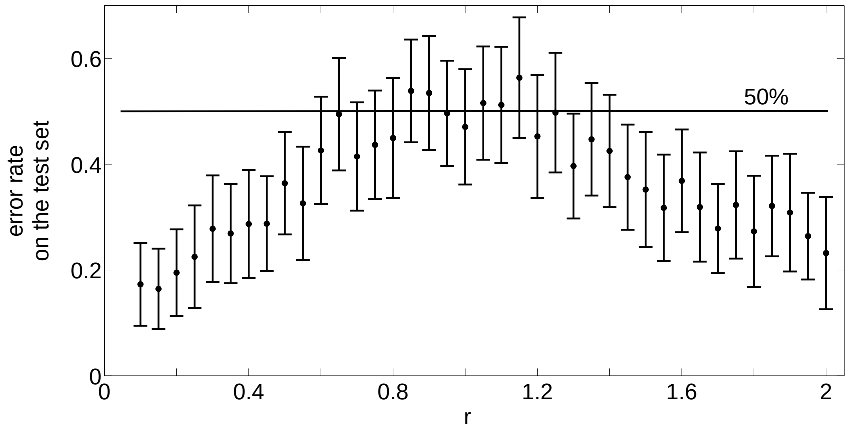


Figure 5.2: SVM-RKHS-PCA method on the ARMA processes. This time we applied the SVM-RKHS-PCA method, which allows a non-linear classification in the RKHS.

AR Processes (simulated). Recall that this experiment is different from the previous one in the following way: Again, we consider 100 time series for each order-noise combination. Each of these 100 time series was sampled with *different* (random) coefficients. In the ARMA experiment we considered 100 time series for each noise, too, but all of these 100 time series were simulated with the same *fixed* coefficients. Hence, this AR experiment is closer to the task of finding the difference between forward and backward going time, but is also much harder.
Both SVM methods did not perform better than chance on this data set. See Figures 5.3 and 5.4 for results.

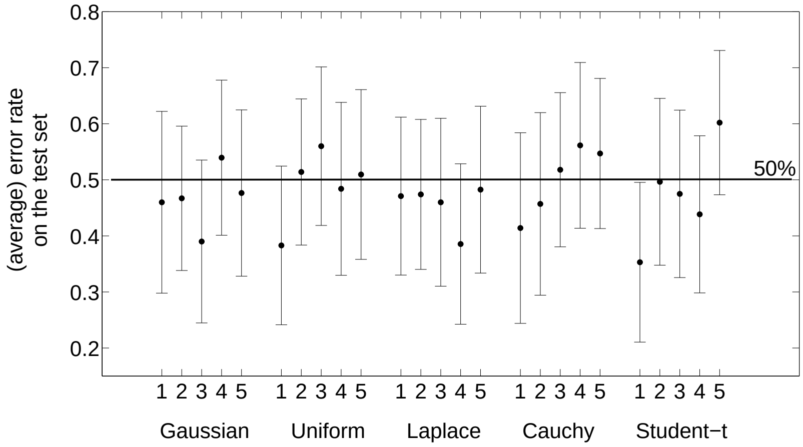


Figure 5.3: SVM-RKHS method on the AR processes. For each order-noise combination we trained the SVM on 90 time series and tested it on the remaining 10 time series; this procedure was repeated 100 times. The figure shows the average error rate on the test set together with the standard deviation. The window length was chosen to be 3 or 5 (which was decided by cross-validation) and $C = 10$. The training error was around 30%. The performance is not better than chance level.

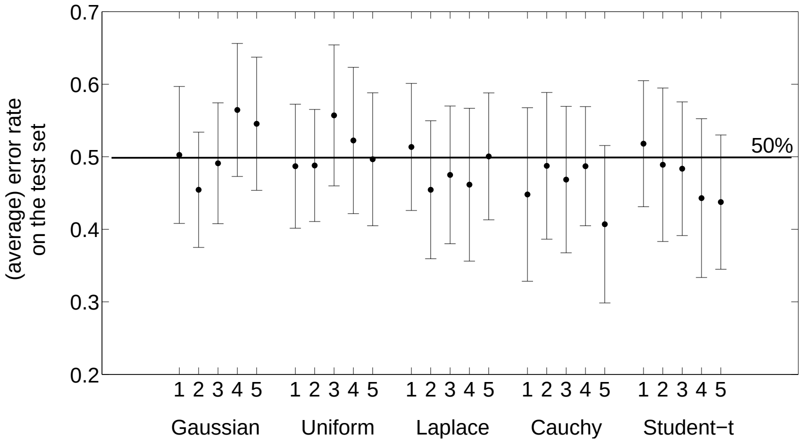


Figure 5.4: SVM-RKHS-PCA method on the AR processes. Here, we used the first 5 principal components in the RKHS for classification.

EEG Data (real). The SVM methods performed very well on this data set. We separated the data set into a training set (103 time series) and a test set (15 time series) 300 times. Because we always used the forward and backward direction of the samples, the actual sizes of training and test set were twice as large. Both methods were on average able to classify more than 95% of all time series in the test set correctly: The SVM-RKHS method achieved an average error rate of 2.9% ± 3.9%, the SVM-RKHS-PCA method of 4.8% ± 4.5%. In both cases the training error was even less (2.6% ± 0.5% and 0.9% ± 0.5%, respectively). A histogram of the achieved error rates is shown in Figure 5.5.

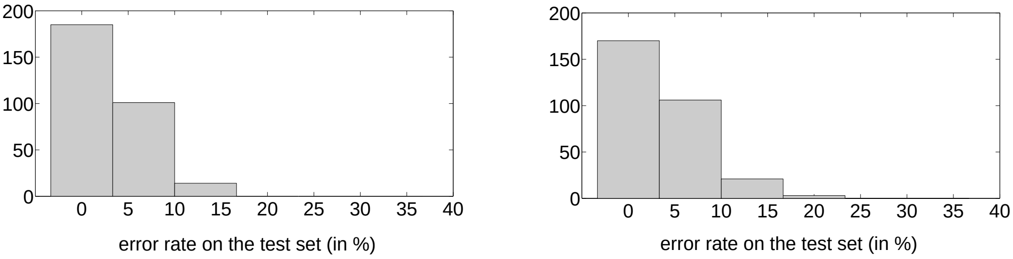


Figure 5.5: SVM-RKHS method and SVM-RKHS-PCA method on the EEG data. This figure shows the performance of the SVM methods on the EEG data set for 300 divisions into training and test set. In most cases there was no false classification on the test set at all. The SVM-RKHS method (left) and the SVM-RKHS-PCA method (right) perform almost equally well.

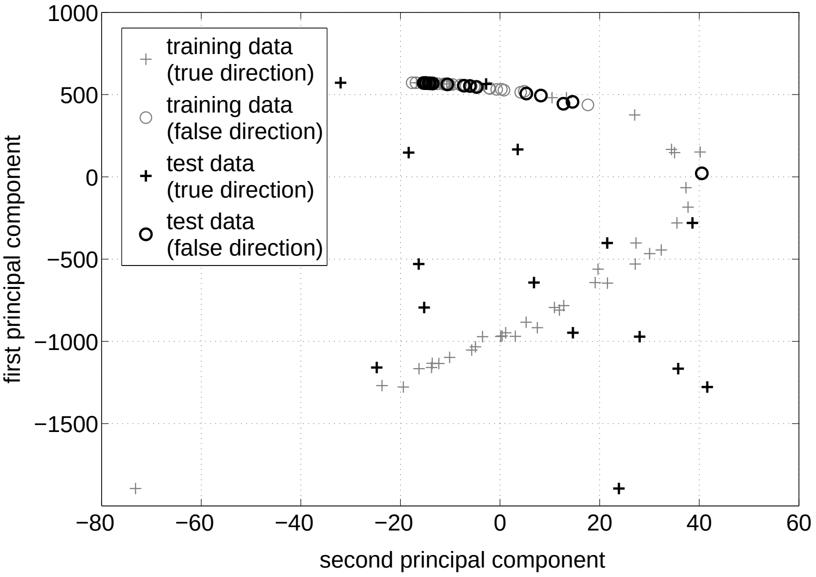


Figure 5.6: This figure explains why there is almost no difference in the performance of the SVM-RKHS and the SVM-RKHS-PCA method on the EEG data set. The plot shows 40 training points and all 15 test points in the RKHS with respect two the first two principal components corresponding to the two largest eigenvalues. It seems that a separation based only on the first principal component is already possible. Thus a linear classification in the RKHS suffices and there is not much to gain from a non-linear classifier.

As in the ARMA experiment (see above) we only used one specific kind of time series, namely EEG data. Even more, some samples belonged to the same time series since we used 10 cuts of each channel. The good performance of the SVM methods shows that the asymmetries between past and future are sufficiently significant.
Figure 5.6 shows why the SVM-RKHS-PCA method does not perform better than the SVM-RKHS method. Here, we computed the projections of the data points in the RKHS onto the first two principal components (that means the two components with the highest sample variation). From this plot it can be seen that the variation in the first principal component is already big enough to separate the data. Ergo a linear classification in the RKHS (by a hyperplane, perpendicular to the first principal component, for example) performs already very well and we do not gain much from a non-linear classification as it is done by SVM-RKHS-PCA method.

Mixed Collection of Time Series (real). The time series in this collection are very different from each other in nature and distribution. Thus, presumably, it is a more difficult problem to solve than the EEG data set. Both SVM methods performed significantly better than chance (see Figures 5.7 and 5.8, but not as good as for the EEG data.

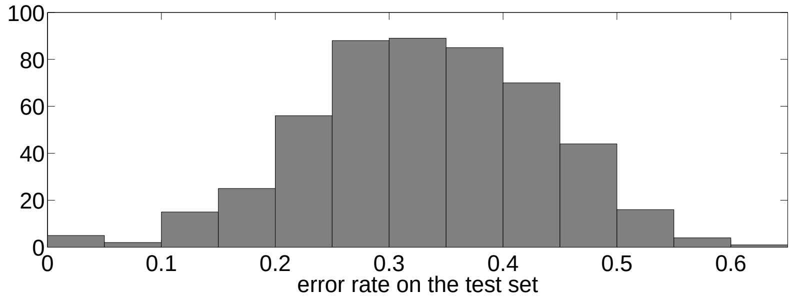


Figure 5.7: SVM-RKHS on the time series collection. 500 times we chose randomly a test set of size 20, trained the method on the remaining 180 time and looked at the performance on the test set. For the C parameter we chose $C = 10$, which resulted in a training error of 29.8% ± 1.8% and a test error of 35.7 ± 10.5%. We reached the same performance, however, for values of C, which were several magnitudes lower or higher.

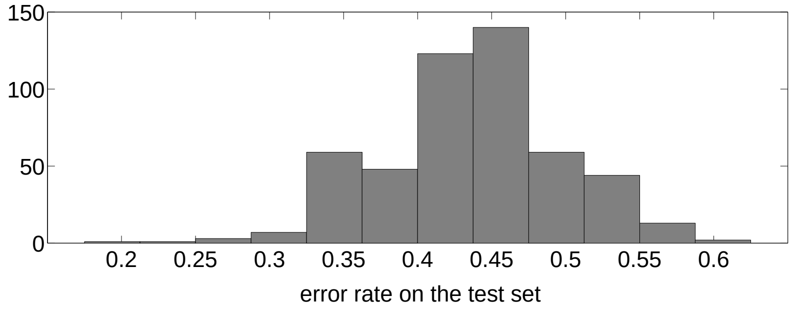


Figure 5.8: SVM-RKHS-PCA on the time series collection. For the C parameter we again chose $C = 10$, which resulted in a training error of 37.3% ± 0.8% and a test error of 43.7% ± 6.7%. Again the performance did not vary for changing C in orders of magnitude 2 or 3.

5.2 ARMA method

Recall that apart from the parameters δ (minimal difference in p-values) and α (significance level) for the HSIC we have to choose a kernel and its parameters. In the experiments we chose the Gaussian kernel and the bandwidth was chosen by the rule of thumb saying that the median of $(||x - y||^2)/(2\sigma^2)$ should be 1.

ARMA Processes (simulated). Recall that according to Theorem 4.15 the ARMA method only works if the ARMA processes are simulated with non-Gaussian noise. This experiment shows that the assumption of non-Gaussian noise is essential. We simulated ARMA(2,2) processes with different noise distributions. These are parameterized by a value r, which ranges between 0.1 and 2. Only $r = 1$ corresponds to a normal distribution.
We then fit an ARMA model to the data without making use of the fact that we already know the order of the process; instead we used the Akaike Information Criterion which penalizes the order of the model. When we detected a dependence between residuals and past values of the time series, we rejected this direction, otherwise we accepted it. (We only wanted to show the necessity of non-Gaussian noise and thus did not perform the whole ARMA method). For the true direction we obviously expect the independence to be rejected in very few cases (depending on the significance level). Theorem 4.15 states that only for $r = 1$, the residuals of the reversed direction will be independent. Since we are dealing with a finite amount of data, the noise cannot be distinguished from a Gaussian distribution if r is close to 1; in these cases we will still be able to fit a backward model reasonably well. For the independence test we used a significance level of $\alpha = 0.01$. See Figure 5.9 for details.

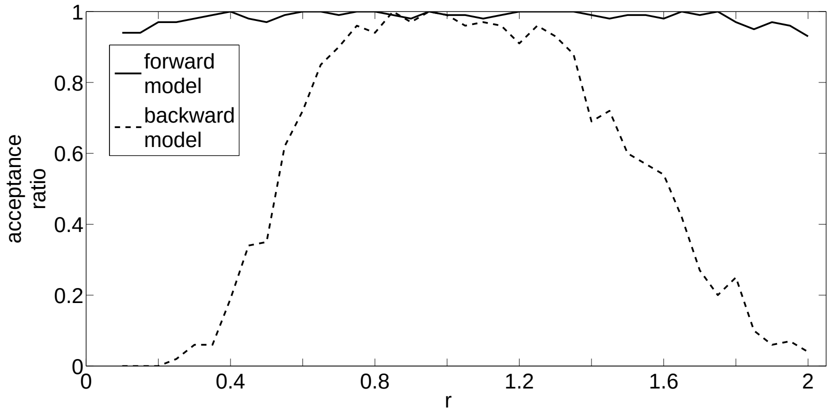


Figure 5.9: ARMA method on the ARMA processes. For each value of r (expressing the non-Gaussanity of the noise) we simulated 100 instances of an ARMA(2,2) process with 500 time points and show the acceptance ratio for the forward model (solid line) and for the backward model (dashed line). When the noise is significantly different from Gaussian noise ($r = 1$), the correct direction can be identified.

As a comparison we also did the same experiment for an MA process with coefficients $\theta_1 = -0.3, \theta_2 = -0.3$ and $\theta_3 = 1$. For this special arrangement, Theorem 4.15 does not hold and as we have seen (Remark 4.17) the process is time-reversible for all distributions of ϵ_t. Thus we expect both forward and backward model to be accepted most of the times. See Figure 5.10 for details.

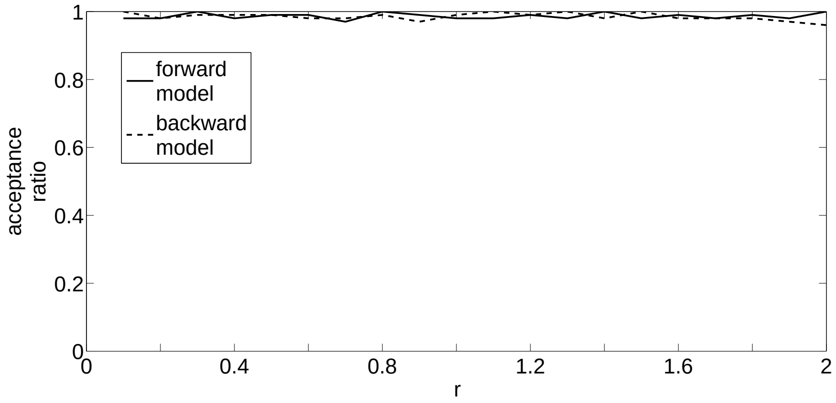


Figure 5.10: Here we simulated for each value of r 100 instances of an MA(3) process with 300 time points and show the acceptance ratio for the forward model (solid line) and for the backward model (dashed line). Since the AR part vanishes and the MA coefficients are carefully chosen according to Remark 4.17, the process is time-reversible for all considered distributions.

AR Processes (simulated). For each order-noise combination we received two numbers: the number of classified time series (out of 100), and the proportion of correctly classified time series (out of those classified), which are shown in Figure 5.11. In order to make the results for the different kinds of noise more distinguishable, we used very unconservative parameters: the minimal difference in p-values δ was chosen to be 0.05 and the significance level α to be 0.1. This ensures that we have some false decisions and can observe a difference in the performance for the different noise distributions. The method works for all distributions except for the Gaussian (as expected). Further it works best for the Cauchy distribution, and it is slightly better in the uniform case than in the Student-t case, for example. This seems reasonable since it is harder to distinguish between a Student-t and a Gaussian distribution than between a uniform and a Gaussian distribution.

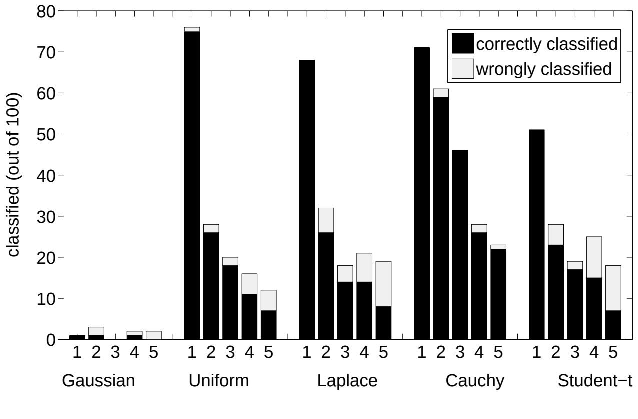


Figure 5.11: ARMA method on the AR processes. The histogram shows the number of classified time series (out of 100) and the proportion of correctly classified time series. The parameters (minimal difference in p-values $\delta = 5\%$, significance level $\alpha = 10\%$) were chosen such that many time series were classified, albeit with some resulting loss of accuracy. Still in most cases (except the Gaussian) the correct classification rate significantly exceeds 50%.

EEG Data (real). The results of the ARMA method on the EEG data set are shown for different values of α and δ in Figure 5.12. As α shrinks and δ grows, the algorithm makes fewer mistakes, but also classifies fewer time series. That said, classification accuracy consistently exceeds 68%.

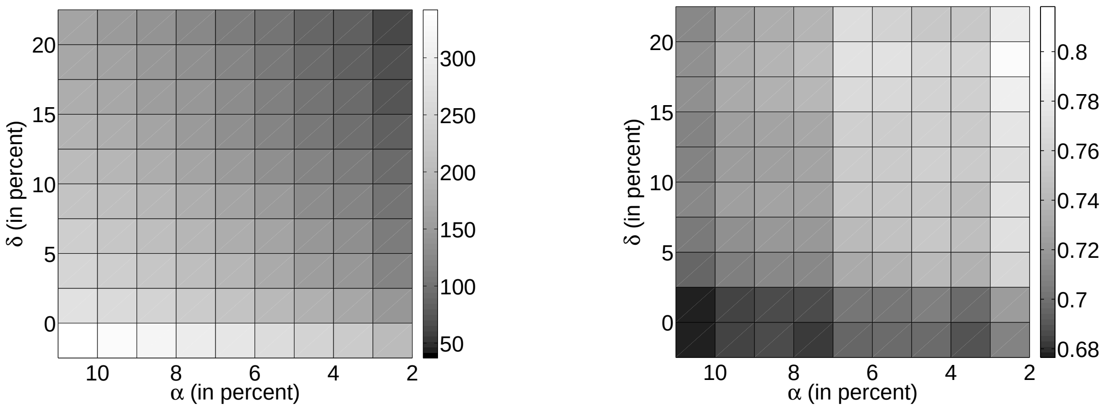


Figure 5.12: ARMA method on EEG data. The left panel shows the number of classified time series (out of 1180), and the right panel the proportion of correctly classified time series, depending on the parameters α and δ. The results are consistently better than chance, reaching a correct classification rate of up to 82%.

Mixed Collection of Time Series (real). In order to obtain a larger data set, we cut the long time series into pieces of length 400. This way we could use 576 instead of 200 time series. Since the performance depends strongly on the chosen parameters, we give the results for different values. The classification consistently exceeds 50% and the more conservative the parameters are chosen, the larger the proportion of correctly classified time series is. See Figure 5.13 for details.

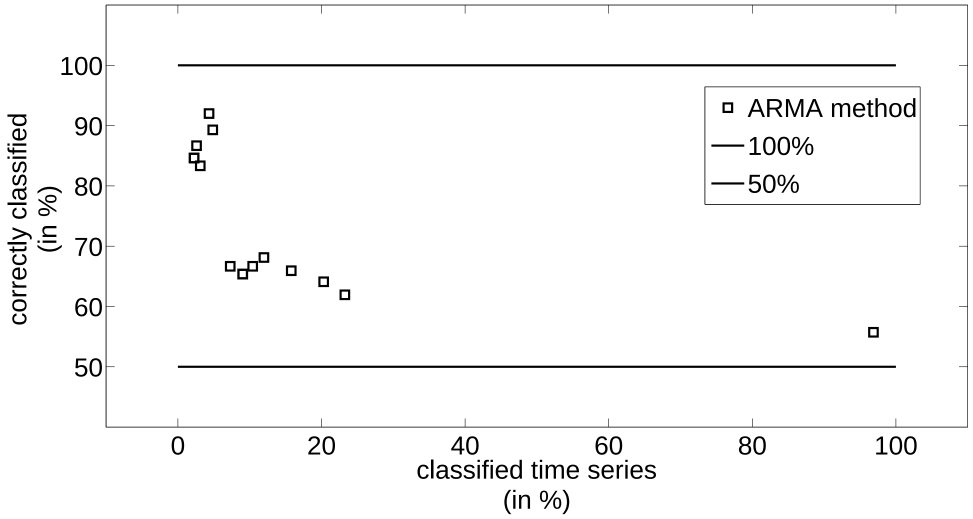


Figure 5.13: ARMA method on the time series collection. We cut the longer time series into smaller pieces of length 400 and obtained 576 time series. We show the results for different values of the parameters: The minimal difference in p-values δ ranges between 0% and 20%, the significance level α between 10% and 0.1%. The point with the highest classification rate corresponds to the highest value of δ and the lowest value of α.

6 Conclusion

We have proposed two methods to detect the time direction of time series. Our methods were based on the theory of Reproducing Kernel Hilbert Spaces, Support Vector Machines, Principal Component Analysis, the Hilbert-Schmidt Independence Criterion and Time Series analysis. Therefore we explained these concepts and gave the main results and proofs.

For the *SVM method* we combined the concept of Hilbert Space embeddings of distributions and Support Vector Machines in a new way: It is possible to represent probability distributions uniquely in an RKHS. We showed how to perform a linear SVM in the RKHS and we could even extended this to a non-linear classifier by applying a PCA to the data in the RKHS and an SVM on the coefficients in the direction of the principal components afterwards.

The *ARMA method* is based on a theoretical result we proved as Theorem 4.15: Causal ARMA processes with non-vanishing AR part can be reversed in time if and only if they are normally distributed. Based on this result we fit an ARMA model to both time directions and check whether the residuals are independent of the former values of the time series. For non-Gaussian distributions this should be the case only for the true time direction. In order to detect the dependence between noise and time series for the wrong time direction we needed a powerful independence test. In this work we used the kernel-based Hilbert-Schmidt Independence Criterion.

Using different kinds of experiments we showed that the SVM methods are able to learn the difference in the finite-dimensional distributions between forward and backward going time series if time series from test and training set are sufficiently similar. When we simulated all time series as instances of an ARMA process with fixed coefficients, for example, we were able to detect the true time direction for all noise distributions except the Gaussian. When we trained the SVM on a set of mixed ARMA processes, which means each time series in the training and test set had different coefficients, we did not achieve a performance better than chance. Probably the size of the training set would have to be much larger; with small training sets it is likely that the SVM adapts to differences in the finite-dimensional distributions, which are dominant for this specific kind of time series, but which cannot be generalized to other time series. Therefore these features should not be regarded as essential differences between forward and backward going time.
Since there were 10 samples of each channel of the EEG data, we again have similar time series in training and test set. Here, the SVM methods performed well. On the collection of time series, however, we did not cut the time series into several pieces and received worse results for this method.
It is interesting to further investigate the reasons why we did not achieve better results for the SVM methods. For an AR(1) process, for example, the dependence between noise and time series implies conditions on the distribution of two adjacent random variables of the time series. These constraints on the distributions can be expressed in terms of their moments. Using a poly-

nomial kernel for the Hilbert Space embedding we should at least in principal be able to detect this difference even if we train the SVM on a set of AR(1) processes with different coefficients and even different kinds of noise.

The experiments with simulated data sets show that the ARMA method is able to identify the true direction in most cases unless the ARMA processes were Gaussian distributed (and thus time-reversible). For real world time series (EEG and the time series collection) we found that in many cases the data did not admit an ARMA model in either direction, or the distributions were close to Gaussian. For a considerable fraction, however, the residuals were significantly less dependent for one direction than for the other. For these cases, we mostly recovered the true direction.

Classification accuracies were not on par with the classification problems commonly considered in Machine Learning, but we believe that this is owed to the difficulty of the task; it is remarkable that we could at all identify the true time direction in time series (even in real data) and thus we consider our results rather to be encouraging.

It is possible to think of an extension of the ARMA method to non-linear time series models. As we found out [48], the result that a linear model with independent additive noise can be reversed if and only if the noise is normally distributed can be extended under some technical conditions to non-linear models: if we can write $Y = f(X) + \epsilon$, where ϵ and X are independent, then a representation $X = g(Y) + \tilde{\epsilon}$, where $\tilde{\epsilon}$ and Y independent is possible if and only if f is linear and all involved variables are Gaussian. It may be possible to prove a similar result in non-linear time series analysis.
We can also think of different, more subtle asymmetries between past and future in time series that are similar to this approach, i.e. if there is a simple generative model in the forward but not the backward direction in a more general sense. Since every cause precedes its effect, finding the true time direction of time series would shed further light on the statistical asymmetries between cause and effect.

Bibliography

[1] R. Balian. *From microphysics to macrophysics*. Springer, 1992.

[2] P. Horwich. *Asymmetries in Time: Problems in the Philosophy of Science*. MIT, 1987.

[3] J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Detecting the direction of time series. *Neural Information Processing Systems*, 2008 (submitted).

[4] J. Pearl. *Causality*. Cambridge University Press, 2002.

[5] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[6] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, 1993. (2nd ed. MIT Press 2000).

[7] M. Eichler and V. Didelez. Causal reasoning in graphical time series models. *In: Proceedings of the 23nd Annual Conference on Uncertainty in Artifical Intelligence*, pages 109–116, 2007.

[8] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pages 63–78. Springer-Verlag, 2005.

[9] S. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, New York, oxford statistical science series edition edition, 1996.

[10] V. P. Skitovic. Linear forms in independent random variables and the normal distribution law (in Russian). *Izvestiia AN SSSR, Ser. Matem.*, 18:185–200, 1954.

[11] V. P. Skitovic. Linear combinations of independent random variables and the normal distribution law. *Select. Transl. Math. Stat. Probab.*, 2:211–228, 1962.

[12] G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst.Internationale Statist.*, 21:2–8, 1953.

[13] B. Ramachandran. *Advanced Theory of Characteristic Functions*. Statistical Publishing Society, Calcutta, 1967.

[14] Yu. V. Linnik. A problem concerning characteristic functions of probability distribution (in Russian). *Usp. Matem. Nauk*, 10(1):137–138, 1955.

[15] A. A. Zinger and Yu. V. Linnik. An analytic extension of Cramér's theorem and its application (in Russian). *Vestnik Leningrad. Univ.*, 11:51–56, 1955.

[16] H. Cramér. Über eine Eigenschaft der normalen Verteilungsfunktion. *Math. Z.*, 41:405–414, 1936.

[17] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[18] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

[19] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Massachusetts, 2002.

[20] D. Werner. *Funktionalanalysis*. Springer, Berlin, sixth edition, 2007.

[21] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, USA, 2002.

[22] M. N. Lukic and J. H. Beder. Stochastic processes with sample paths in Reproducing Kernel Hilbert Spaces. *Trans. Am. Math. Soc.*, 353:3945–3969, 2001.

[23] Netzzeitung. Website, 17.4.2008, 2:55pm. `http://www.netzeitung.de/internet/976597.html`.

[24] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan. Polynomial solvability of convex quadratic programming. *Sov. Math.*, 20:1108–1111, 1979.

[25] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, second edition, 1999.

[26] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, New York, 2006.

[27] S. Ben-David and H. Simon. Efficient learning of linear perceptrons. *Proceedings of Neural Information Processing Systems*, pages 189–195, 2000.

[28] C. Cortez and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[29] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[30] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, 2007.

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

[32] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. *Proceedings of Neural Information Processing Systems*, 20:1–8, 2007.

[33] A. Gretton, K. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. *Technical Report* **157** *MPI for biological Cybernetics, Tübingen*, 157:1–8, 2008.

[34] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, 2005.

[35] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Conference on Learning Theory*, 2008. to appear.

[36] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

[37] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons Inc, USA, 2002.

[38] A. Kankainen. Consistent testing of total independence based on the empirical characteristic function. *PhD Thesis, University of Jyväskylä*, 1995.

[39] B. Schölkopf, C. J. C. Burges, and A. J. Smola (editors). *Advances in Kernel Methods*. MIT Press, Massachusetts, 1999.

[40] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, second edition, 1991.

[41] G. Weiss. Time-reversibility of linear stochastic processes. *J. Appl. Prob.*, 12:831–836, 1975.

[42] F. J. Breidt and R. A. Davis. Time-reversibility, identifiability and independence of innovations for stationary time series. *Journal of Time Series Analysis*, 13(5):379–390, 1991.

[43] L. V. Mamai. On the theory of characteristic functions. *Select. Transl. Math. Stat. Probab.*, 4:153–170, 1963.

[44] M. Loeve. *Probability Theory*. Van Nostrand, Princeton, second edition, 1960.

[45] C. M. Jarque and A. K. Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163–172, 1987.

[46] This data set (exp. iva, subj. 3) can be downloaded after registration. Website, 3.6.2008, 3:51pm. `http://ida.first.fraunhofer.de/projects/bci/competition_iii/#datasets`.

[47] B. Mandelbrot. On the distribution of stock price differences. *Operations Research*, 15(6):1057–1062, 1967.

[48] P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Neural Information Processing Systems*, 2008 (submitted).