École Polytechnique Fédérale de Lausanne

MATH-498 - Mathematics projects

Kernel methods in Machine Leaning and Statistics

Student : Yerkin Yesbay
Professor : Prof. Victor Panaretos
Supervisor : Kartik Waghmare

Fall semester 2022

1 Introduction

In the last two decades learning algorithms utilizing positive definite kernels have become very popular in machine learning and statistics. Kernel methods show spectacular efficiency in nonlinear pattern recognition comparable to neural networks, while the former has a strong mathematical foundation and the latter is still more of a “black box”. The mathematical theory behind kernel methods is truly fascinating, gathering together statistics, linear algebra, functional analysis, Gaussian processes and potentially any other field where positive definite kernels arise.

Kernel methods in machine learning can be viewed as the most elegant way to generalize linear models via *kernel trick*, that allows to implicitly map the data from its original domain into a higher- and possibly infinite-dimensional Hilbert spaces without excessive computations. At the same time linear methods extended into reproducing kernel Hilbert spaces, as it turns out, maintain the applicability of classical results of optimization in the finite dimensional Euclidean setting.

In statistics reproducing kernels are mainly known for their powerful applications in two-sample and independence testing. The kernel trick was spread from mapping single points to embedding probability distributions into the RKHS.

The main goal of this project was to study the theory of several machine learning methods employing kernel approach, such as ridge regression, support vector machines and principal component analysis, as well as kernel two-sample and independence tests to some extent. The secondary goal was to test their performance in simulations and practical use.

The structure of the report is as follows. In section 1 we present the mathematical theory of ridge regression, SVM and PCA in the finite-dimensional setting. In section 2 we first present the basic elements of RKHS theory and prove a short form of Kimmeldorf-Wahba theorem, then we show how the optimization problems of the methods above is solved in RKHS. In section 3 we outline some results of kernel two-sample and independence testing. In section 4 we visually demonstrate the use of kernel methods on simulated data. In section 5 we outline the result of application of kernel SVM and PCA on real data. In section 6 we provide a brief conclusion.

2 Linear methods

2.1 Ridge regression

In this subsection we are going to discuss a common approach to regression that can be well modified using reproducing kernels.

2.1.1 Reminder on least squares

Suppose we are given a data set $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, are drawn from an unknown distribution $\mathbf{X}, \mathbf{Y} \sim P_{XY}$. We wish to predict the value of $\mathbf{Y}$ based on the value of $\mathbf{X}$.

To evaluate the quality of a predictor $f(x)$ we introduce the notion of *risk* R_l associated with a specific *loss-function* l.

$$R_l(f) = \mathbb{E}l(f(\mathbf{X}), \mathbf{Y}),$$

where the expectation is taken with respect to the joint distribution of (X, Y). So the best predictor for the specific loss function is the one that minimizes the associated risk. We will consider the quadratic loss function $l(y, a) = (y - a)^2$, for which the associated risk is the (population) mean squared error

$$R_l(f) = \mathbb{E}(f(\mathbf{X}) - \mathbf{Y})^2.$$

The function f^* that minimizes the risk is called the *target function.* Recall that among square-integrable Borel functions of $\mathbf{X}$ the function f^* that minimizes the mean squared error is the conditional expectation $\mathbb{E}(\mathbf{Y}|\mathbf{X})$.

$$f^*(\mathbf{X}) = \mathbb{E}(\mathbf{Y}|\mathbf{X}) = \arg\min_f \mathbb{E}(f(\mathbf{X}) - \mathbf{Y})^2. \tag{1}$$

To obtain a good accuracy of a non-parametric estimation for the conditional distribution (e.g. a histogram type) it takes a data set of large size, and it grows drastically with growth of dimensionality of $\mathbf{X}$. Hence, we approximate the best predictor within some parametric space of functions S. A simple option is to consider linear functions of $\mathbf{X}$, i.e. of the following form

$$f(x) = w^T x, \;\; w \in \mathbb{R}^p.$$

Note that affine functions, i.e. of the form $f(x) = w^T x + b$, can be included by adding a constant as a component of x.

Since we don't have a direct access to P_{XY}, the distribution of $(\mathbf{X}, \mathbf{Y})$, we estimate the best predictor in S

$$f_S^* = \arg\min_{f \in S} \int_{\mathbb{R}^{p+1}} l(f(x), y) dP_{XY}(x, y),$$

by minimizing the empirical risk $\widehat{R}_n$.

$$\widehat{f}_S = \arg\min_{f \in S} \widehat{R}_n(f) = \arg\min_{f \in S} \int_{\mathbb{R}^{p+1}} l(f(x), y) dP_n(x, y) = \arg\min_{f \in S} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i).$$

Minimization of empirical risk associated with the square loss over linear functions leads to *ordinary least squares.* Given the design matrix X and vector $y = (y_1, \ldots, y_n)^T$, the solution has a closed form

$$\widehat{f}(x) = \widehat{w}^T x, \;\; \widehat{w} = (X^T X)^{-1} X^T y. \tag{2}$$

2.1.2 Bias-variance tradeoff

Note that the quadratic risk of an estimator $\widehat{f}(x) = \widehat{f}_n(x; D_n)$ can be decomposed as

$$\begin{aligned} \mathbb{E}(\widehat{f}(\mathbf{X}) - \mathbf{Y})^2 &= \mathbb{E}(\widehat{f}(\mathbf{X}) - f^*(\mathbf{X}) + f^*(\mathbf{X}) - \mathbf{Y})^2 = \\ &= \mathbb{E}(\widehat{f}(\mathbf{X}) - f^*(\mathbf{X}))^2 + \mathbb{E}(f^*(\mathbf{X}) - \mathbf{Y})^2 + 2\mathbb{E}(\widehat{f}(\mathbf{X}) - f^*(\mathbf{X}))(f^*(\mathbf{X}) - \mathbf{Y}). \end{aligned} \tag{3}$$

To calculate the last term in (3) we use total probability rule and recall that $f^*(\mathbf{X}) = \mathbb{E}(\mathbf{Y}|\mathbf{X})$

$$\mathbb{E}(\widehat{f}(\mathbf{X}) - f^*(\mathbf{X}))(f^*(\mathbf{X}) - \mathbf{Y}) = \mathbb{E}(\widehat{f}(\mathbf{X}) - f^*(\mathbf{X}))\mathbb{E}[f^*(\mathbf{X}) - \mathbf{Y}|X] = 0. \tag{4}$$

Note that the sample D_n on which $\widehat{f}$ is computed is independent of $\mathbf{X}$ and therefore can be treated as a deterministic function w.r.t. to $\mathbf{X}$. Hence, the quadratic risk of $\widehat{f}$ admits the decomposition

$$R(\widehat{f}) = \mathbb{E}(\widehat{f}(\mathbf{X}) - f^*(\mathbf{X}))^2 + \mathbb{E}(f^*(\mathbf{X}) - \mathbf{Y})^2, \tag{5}$$

where the second term is constant.

Substituting $\mathbf{X}$ by a fixed point x_0 and taking expectation w.r.t. to the training sample D_n one obtains the bias-variance decomposition of the first term of (5)

$$\mathbb{E}(\widehat{f}(x_0) - f^*(x_0))^2 = \mathbb{E}[\widehat{f}(x_0)) - \mathbb{E}\widehat{f}(x_0)]^2 + (\mathbb{E}\widehat{f}(x_0) - f^*(x_0))^2 = \mathrm{Var}\{\widehat{f}(x_0)\} + [\mathrm{bias}\{\widehat{f}(x_0)\}]^2. \quad (6)$$

Recall the following classical result of least squares regression.

Theorem 1 (Gauss-Markov). *For a design matrix X and response variables vector y of i.i.d. observations suppose that X has full rank and*

$$\mathbb{E}[y|X] = Xw, \;\; Cov(y|X) = \sigma^2 I_n.$$

Then $\widehat{w} = (X^T X)^{-1} X^T y$ is the best linear unbiased estimator (BLUE) of w, i.e. $\widehat{w}$ has the smallest variance among all linear unbiased estimators.

However, a biased estimator (e.g. constant $c\widehat{w}$ with $|c| < 1$) of w may have a smaller variance. Keeping in mind (6), it might be the case that by adding bias to the predictor and reducing the variance, we may as well reduce the total risk. One way of doing that is by introducing *regularization.*

2.1.3 Regularized least squares

We will consider a special case of Tikhonov regularization, namely the l_2 regularization. We modify the problem of empirical risk minimization by adding a regularization term to the empirical risk, thus obtaining the *ridge regression.*

$$\min_{f \in S} \widehat{R}_n(f) + \lambda ||f||_2^2, \quad (7)$$

where the norm of $f_w(x) = w^T x$ is defined as the Euclidean norm of w, i.e. $||f_w||^2 = ||w||_2^2 = w^T w$. Note, that this definition is consistent with the operator norm

$$||f|| = \inf_{x \in \mathbb{R}^p} \frac{||f(x)||_{\mathbb{R}^p}}{||x||_{\mathbb{R}^p}}.$$

Then the regularized empirical risk associated with the square loss can be written as

$$\widehat{R}_n(f) + \lambda ||f||_2^2 = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda w^2 = ||Xw - y||^2 + \lambda ||w||^2.$$

One can see that the minimized function is convex w.r.t. to w, so to minimize it we simply find the stationary point $\widehat{w}$.

$$\frac{\partial}{\partial w}\left(\frac{1}{n}||Xw - y||^2 + \lambda ||w||^2\right) = 0 \iff \frac{1}{n}(X^T X w - X^T y) + \lambda w = 0, \quad (8)$$

which yields the solution

$$\widehat{w}^{\lambda} = (X^T X + n\lambda I_p)^{-1} X^T y. \quad (9)$$

Note that $X^T X$ has non-negative eigenvalues, so the eigenvalues of $X^T X + n\lambda I_p$ are strictly positive with lower bound $n\lambda$ and thus $X^T X + n\lambda I_p$ is invertible.
The motivation behind penalizing $||w||^2$ is to shrink the optimal w towards zero and thus reduce the variance of $\widehat{w}$. The parameter λ is fixed and is to define how much we want to shrink w.

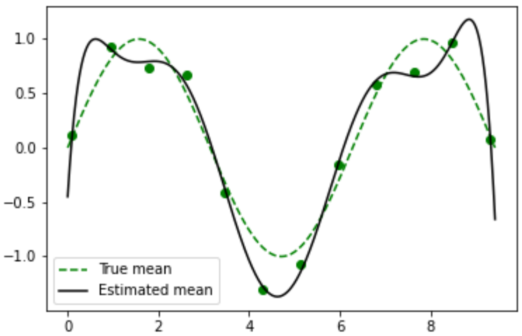

Figure 1: Ordinary least squares

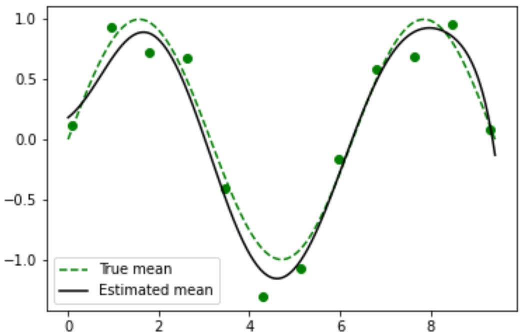

Figure 2: Ridge regression with $\lambda = 1$

In the default setting of linear regression we usually assume that the number of observation by far exceeds the number of variables. When this is not the case the OLS model may easily overfit, that is it will fit too well to single observations, but will deviate a lot form the target function

Consider polynomial regression with $n = 12$ observations and $p = 9$ explanatory variables that are powers $1, x, \dots, x^{p-1}$ of $x \in \mathbb{R}$. Figures 1 and 2 illustrate the behaviour of the estimators for OLS regression and ridge regression with $\lambda = 1$. One can notice that OLS allows for a more complex model in the sense of variability and fluctuations. Penalizing the norm of a predictor, which is a linear combination of some basis function, can be a good alternative to reducing the dimensionality of the basis. The optimal value for the parameter λ is usually searched using cross-validation.

Let $rank(X) = p$. Consider the reduced singular value decomposition of $n \times p$ matrix X (derived in the subsection 2.4.2).

$$X = U\Sigma V^T, \;\; U^TU = UU^T = V^TV = VV^T = I_p, \tag{10}$$

where U is $n \times p$, V is $p \times p$ matrices and Σ is an $p \times p$ diagonal matrix with singular values $\sigma_1, \dots, \sigma_p$ on the diagonal. The matrix U consists of the orthogonal column-vectors $u_1, \dots, u_p$.

Using (9) and (10) the solution vector $\widehat{w}^\lambda$ can be written as

$$\widehat{w}^\lambda = V(\Sigma^T\Sigma + n\lambda I_p)^{-1}\Sigma U^T y, \tag{11}$$

where $(\Sigma^T\Sigma + n\lambda I_p)^{-1}\Sigma$ is a diagonal matrix with values $\sigma_j/(\sigma_j^2 + n\lambda), \;\; j = 1, \dots, p$. Hence, the following identity holds

$$\widehat{w}^\lambda = \sum_{j=1}^{p} \frac{\sigma_j}{\sigma_j^2 + n\lambda} v_j u_j^T y. \tag{12}$$

From this expression one can see that for ordinary least squares ($\lambda = 0$) small singular values of X add large variance to the solution. But adding $n\lambda$ in the denominator stabilizes the solution. It also adds numerical stability to the computation of inverse matrix by imporving the condition number of matrix X associated with the spectral norm from $\max \sigma_j / \min \sigma_j$ to

$$\frac{(\max \sigma_j^2 + n\lambda)^{1/2}}{(\min \sigma_j^2 + n\lambda)^{1/2}}.$$

2.2 Support Vector Machines

2.2.1 Motivation and preliminary statements

We shall begin by introducing the linear Support Vector Machines for binary classification. Given a data set $(x_1, y_1), \ldots, (x_n, y_n)$ where each point $x_i \in \mathbb{R}^p$ is assigned a class label $y_i = \pm 1$, for every new point x our goal is to choose which class in belongs to.
Assuming that the classes are linearly separable (figure 3), that is there exists a hyper plane $L = \{x \in \mathbb{R}^p : \langle x, w \rangle - b = 0\}$ such that

$$y_i(\langle x_i, w \rangle - b) > 0 \quad i = 1, \ldots, n, \tag{13}$$

we would like to find a good linear boundary. If the boundary is given by L, then a point x is assigned label $y = 1$ if $\langle x, w \rangle - b > 0$, and it is assigned label $y = -1$ otherwise.

We will stick to the classical approach to SVM using the concept of *margin*. A hyper plain L separates the classes with a margin of size d if the distance from L to any point x_i from the data set is no smaller than d (figure 4). The optimal linear boundary would be the one with the largest margin.

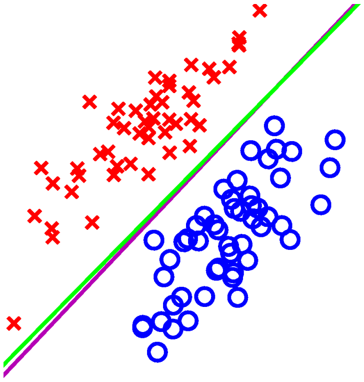

Figure 3

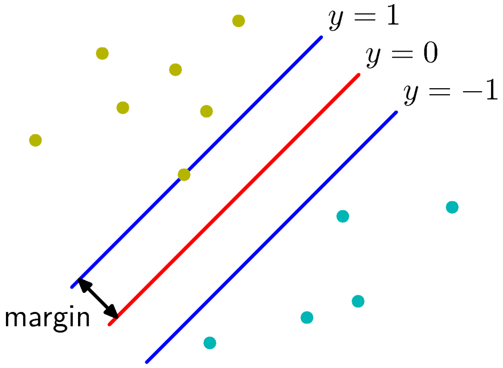

Figure 4

Before we formulate the optimization problem, we would like to make some helpful remarks. Let the hyper plain L be given by the equation $\langle x, w \rangle - b = 0$. If $x_0 \in L$ and $\langle x', w \rangle - b > 0$ one can calculate the distance between L and x as

$$dist(x', L) = \langle x' - x_0, \frac{w}{||w||} \rangle = \frac{\langle x', w \rangle - b}{||w||} \tag{14}$$

Similarly if $\langle x', w \rangle - b < 0$ then $dist(x', L) = -(\langle x', w \rangle - b)/||w||$.
Consider hyper plains $L_1 = \{\langle x, w \rangle - b = 1\}$ and $L_{-1} = \{\langle x, w \rangle - b = -1\}$. To calculate the distance between L and L_1 one should find the the dot product between $w/||w||$ and $x' - x_0$ for arbitrary $x' \in L_1, x_0 \in L$

$$dist(L_1, L) = \langle x' - x_0, \frac{w}{||w||} \rangle = \frac{b + 1 - b}{||w||} = \frac{1}{||w||}. \tag{15}$$

Similarly if $x' \in L_{-1}$ then

$$dist(L_{-1}, L) = |\langle x' - x_0, \frac{w}{||w||} \rangle| = \frac{|b - 1 - b|}{||w||} = \frac{1}{||w||}. \tag{16}$$

Let the hyper plain L with a normal vector w and an intercept b satisfy (13). Since the data set is finite there exists $\delta > 0$ such that

$$y_i(\langle x_i, w\rangle - b) \geq \delta, \;\; i = 1, \ldots, n.$$

Obviously $\widetilde{w} = w/\delta$ and $\widetilde{b} = b/\delta$ define the same hyper plain and the following inequalities hold

$$y_i(\langle x_i, \widetilde{w}\rangle - \widetilde{b}) \geq 1, \;\; i = 1, \ldots, n.$$

Now we can finally pose the optimization problem of SVM.

2.2.2 Hard margin SVM

Assume that the data set is linearly separable. Then without loss of generality, for any separating hyper plain L defined by w and b the following holds

$$y_i(\langle x_i, w\rangle - b) \geq 1, \;\; i = 1, \ldots, n. \tag{17}$$

This guarantees that no point x_i from the data set lies between the hyper plains L_{-1} and L_1. Since our goal is to find the separating hyper plain with the largest possible margin, (15) implies that we should maximize $1/||w||$. We rewrite it as the following constrained minimization problem.

$$\begin{aligned} &\min_{w,b} ||w||^2, \\ \text{s.t.} \;\; &y_i(\langle x_i, w\rangle - b) \geq 1 \end{aligned} \tag{18}$$

The constraints (17) can be rewritten as follows.

$$\min_{1\leq i\leq n} y_i(\langle x_i, w\rangle - b) \geq 1.$$

Remark 1. *Note that for the optimal solution w, b the an equality will be reached by some points from both classes, i.e.*

$$\begin{aligned} \min\nolimits_{y_i=1}(\langle x_i, w\rangle - b) = 1, \\ \min\nolimits_{y_i=-1}(-\langle x_i, w\rangle + b) = 1. \end{aligned} \tag{19}$$

Proof. Indeed, if none of the equalities are met, then for some $\varepsilon > 0$ the latter holds.

$$y_i(\langle x_i, w\rangle - b) \geq 1 + \varepsilon.$$

Then for $\widetilde{w} = w/(1+\varepsilon)$ and $\widetilde{b} = b/(1+\varepsilon)$ the constraints (17) are met and $||\widetilde{w}|| < ||w||$, so w turns out to be suboptimal.
If just one of the equalities (19) is not met, say upper one, then for some $\varepsilon > 0$

$$\langle x_i, w\rangle - b \geq 1 + \varepsilon.$$

In that case suppose $\widetilde{w} = \dfrac{w}{1+\varepsilon/2}$ and $\widetilde{b} = \dfrac{b+\varepsilon/2}{1+\varepsilon/2}$ for which the constraints are satisfied and w is suboptimal. □

The points x_i touching the margin, i.e. $y_i(\langle x_i, w\rangle - b) = 1$, are called the *support vectors* and they define the optimal parameters w, b.

Also note that the constraints (17) can be rewritten as identity

$$\min_{1\le i\le n} \frac{y_i(\langle x_i, w\rangle - b)}{||w||} = \frac{1}{||w||}.$$

Hence, instead of the the constrained optimization problem, maximization of the margin size can be posed as

$$\max_{w,b} \min_{1\le i\le n} \frac{y_i(\langle x_i, w\rangle - b)}{||w||}. \tag{20}$$

However, we find it more convenient to stick to the constrained formulation (18) which in fact gives us a quadratic programming problem. To find the necessary conditions for a local minimum we employ the Lagrange principle. The Lagrangian with multipliers $\lambda_i \ge 0$ is given as

$$\Lambda(w, b, \lambda) = \frac{1}{2}||w||^2 + \sum_{i=1}^{n} \lambda_i[1 - y_i(\langle x_i, w\rangle - b)].$$

If w, b is a local minimum then there exist $\lambda_i \ge 0$ such that the following holds.

$$\frac{\partial\Lambda}{\partial w} = 0 \iff w - \sum_i \lambda_i y_i x_i = 0, \tag{21}$$

$$\frac{\partial\Lambda}{\partial b} = 0 \iff \sum_i \lambda_i y_i = 0, \tag{22}$$

$$y_i(\langle x_i, w\rangle - b) - 1 \ge 0, \;\; i = 1, \dots, n, \tag{23}$$

$$\lambda_i(y_i(\langle x_i, w\rangle - b) - 1) = 0, \;\; i = 1, \dots, n. \tag{24}$$

From (21) it follows that the optimal w is a linear combination of x_i's. Also (11) implies that if $y_i(\langle x_i, w\rangle - b) > 1$ then $\lambda_i = 0$, which means that only support vectors contribute to w.
We use stationarity conditions (21) and (22) to write the dual problem functional as a function of λ.

$$\widetilde{\Lambda}(\lambda) = \inf_{w,b}[\Lambda(w, b, \lambda)] =$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle + \sum_{i=1}^{n} \lambda_i(1 - y_i b) =$$

$$= \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle \tag{25}$$

Problem (18) satisfies strong duality conditions, so it is equivalent to maximizing $\widetilde{\Lambda}$ over λ with respect to the following constraints.

$$\lambda_i \ge 0,$$

$$\lambda_i(y_i(\sum_j \lambda_j y_j \langle x_j, x_i\rangle - b) - 1) = 0,$$

$$y_i(\sum_j \lambda_j y_j \langle x_j, x_i\rangle - b) - 1 \geq 0.$$

Hence, to find the solution it is sufficient to know the dot products $\langle x_i, x_j\rangle$ and not exact values of x's. Also to estimate the class for a new point x it is sufficient to calculate the dot products.

$$\langle w, x\rangle - b = \sum_i \lambda_i y_i \langle x_i, x\rangle - b \tag{26}$$

The model we described in this subsection is called the *hard margin* SVM, because it does not allow the classes to cross their corresponding boundaries L_1 and L_{-1}. It has an obvious disadvantage that it necessarily requires linear separability of the classes which is very often not the case. In addition it is very sensitive to the configuration of the support vectors.

In the following subsection we discuss a less conservative and more robust modification of support vector machines.

2.2.3 Soft margin SVM

The idea of the *soft margin* SVM is to relax the linear constraints of the optimization problem allowing some points of the data set to be misclassified or to lie closer to the classification boundary. Instead of (17) we will require the following constraints.

$$\begin{array}{c} y_i(\langle x_i, w\rangle - b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \quad i = 1, \dots, n. \end{array} \tag{27}$$

Here ξ_i is the quantity to which we allow x_i to violate the constraints (17) and is not some fixed constant but rather a parameter for which we seek the optimal value. To not let the violations be arbitrarily large we add penalization for nonzero ξ_i to the minimized functional.

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i. \tag{28}$$

The constant C is a hyperparameter and defines how strong we penalize the violations, i.e. how "hard" margin we want.

Note that the constraints (27) may be rewritten as

$$\xi_i \geq \max\{0, 1 - y_i(\langle x_i, w\rangle - b)\}.$$

And similar to remark 1 here one can show that if $\xi_i > \max\{0, 1 - y_i(\langle x_i, w\rangle - b)\}$ for some i then w, b, ξ are suboptimal values, as the penalization term can be lessened with (27) being held.
Thus, in the constrained optimization problem it is sufficient to consider only those ξ for which the following identities hold.

$$\xi_i = \max\{0, 1 - y_i(\langle x_i, w\rangle - b)\}, \;\; i = 1, \dots, n. \tag{29}$$

And it allows us to reformulate the original problem as the following unconstrained optimization problem

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i(\langle x_i, w\rangle - b)\}. \tag{30}$$

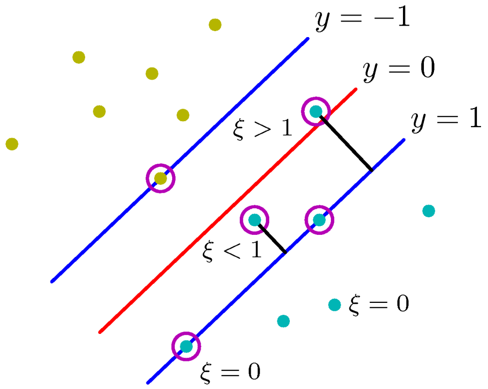

Figure 5: If x_i does not violate its respective class boundary given by L_{y_i} then the corresponding slack variable $\xi_i = 0$. Otherwise it is proportional to the distance from x_i to L_{y_i}

In the literature the variables ξ_i defined in (29) are called the *slack variables.*

The sum of slack variables can be represented as empirical risk (without averaging) associated with the *hinge loss* function.

$$l(y, a) = \max\{0, 1 - ya\},$$

$$\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} l(y_i, \langle x_i, w\rangle - b).$$

To minimize (28) under constraints (27) we write the corresponding Lagrangian

$$\Lambda(w, b, \lambda, \nu) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i(y_i(\langle x_i, w\rangle - b) + \xi_i - 1) - \sum_{i=1}^{n} \nu_i \xi_i. \tag{31}$$

The KKT conditions are given as follows.

$$\frac{\partial \Lambda}{\partial w} = w - \sum_{i=1}^{n} \lambda_i y_i x_i = 0, \tag{32}$$

$$\frac{\partial \Lambda}{\partial b} = \sum_{i=1}^{n} \lambda_i y_i = 0, \tag{33}$$

$$\frac{\partial \Lambda}{\partial \xi_i} = C - \lambda_i - \nu_i = 0. \tag{34}$$

$$\lambda_i \geq 0, \ \ \nu_i \geq 0$$

$$\lambda_i(y_i(\langle x, w\rangle - b) + \xi_i - 1) = 0 \tag{35}$$

$$\nu_i \xi_i = 0$$

From (32) and (35) it follows that x_i contributes to w only if $y_i(\langle x, w\rangle - b) = 1 - \xi_i$, that is only if x_i touches the margin or violates its class boundary. Such points are called support vectors for soft margin SVM.

Similarly to hard margin SVM, using (32) and (33) one can formulate the dual problem where x_i only appear in dot products. This again means that we only need to know the labels y_i and the dot products $\langle x_i, x_j\rangle$. On the dual problem as well as existence and uniqueness of soft-margin SVM solution we write more generally in subsection 3.5.

Figures 3, 4, and 5 are taken from Bishop and Nasrabadi 2006. Pattern recognition and machine learning. Copyright in these figures is owned by Christopher M. Bishop and they are free to use in educational purposes.

2.3 Principal Component Analysis

2.3.1 Motivation

Principal component analysis is often viewed as a way to reduce dimensionality of the input space $\mathcal{X} = \mathbb{R}^p$ based on the data set $x_1, \dots, x_n \in \mathcal{X}$. In a broader sense it is a way to analyze the covariance structure of the data set, which is also called Karhunen-Loeve decomposition.

Suppose that $x_1, \dots, x_n$ are drawn from a continuous distribution in $\mathbb{R}^p$ but which is in fact close to a degenerate distribution with a lower dimensional support, i.e. $x - \varepsilon \in L$, where $L \subset \mathcal{X}$ is a linear subspace of dimension $q < p$ and ε is noise with small variance. In other words the features are almost linearly dependent up to a small noise. This phenomenon is called multicollinearity and it often takes place in real world data. In this case it is reasonable to reduce dimensions of the input space by projecting the data onto a lower dimensional subspace.

Looking ahead we wish to note that kernel PCA also allows us to analyze data with non-linear underlying structure, that is for example when x is scattered close to a surface (manifold) $M \subset \mathcal{X}$ of dimension $q < p$.

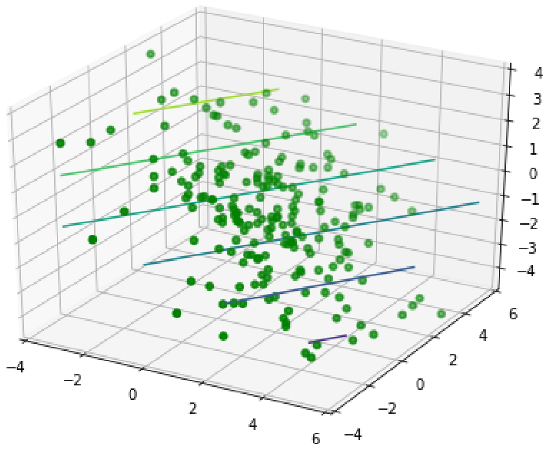

Figure 6: The cloud of points lies in a neighbourhood of a linear subspace

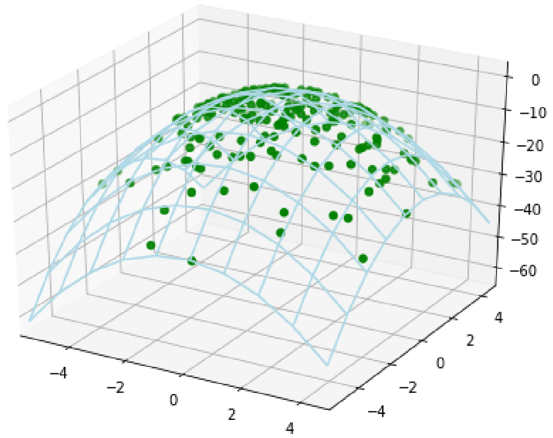

Figure 7: The cloud of points lies in a neighbourhood of a quadratic surface

2.3.2 Preliminary statements

We will make use of some facts from linear algebra. Let matrices $A \in \mathbb{R}^{n \times K}$ and $B \in \mathbb{R}^{p \times K}$ have columns $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}^p$ respectively. Then the following identity holds.

$$AB^T = \sum_{j=1}^{K} a_j b_j^T \tag{36}$$

Let $X \in \mathbb{R}^{n \times p}$ be a matrix of rank p. Then $X^T X$ is a symmetric positive definite matrix and admits the spectral decomposition

$$X^T X = V D V^T, \quad V^T V = V V^T = I_p, \tag{37}$$

where $D = diag(\sigma_1^2, \dots, \sigma_p^2)$ is the diagonal matrix consisting of eigenvalues of $X^T X$ and V consists of the column-vectors $v_1, \dots, v_p$ of orthogonal eigenvectors of $X^T X$.

One can see that the matrix U defined as $U = XVD^{-1/2}$ has orthogonal columns.

$$U^TU = D^{-1/2}V^TX^TXVD^{-1/2} = D^{-1/2}V^TVDV^TVD^{-1/2} = D^{-1/2}DD^{-1/2} = I_p.$$

Using this we obtain the reduced singular value decomposition of X

$$X = U\Sigma V^T, \quad U^TU = V^TV = I_p, \tag{38}$$

where $\Sigma = D^{1/2}$. Note that $U\Sigma$ consists of columns $\sigma_j u_j$ and from (36) it follows that

$$X = \sum_{j=1}^{p} \sigma_j u_j v_j^T. \tag{39}$$

Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$. For $q \leq p$ consider the truncated matrices $U_q = [u_1, \ldots, u_q]$,
$V_q = [v_1, \ldots, v_q]$, i.e. first q columns of U and V, and $\Sigma_{(q)} = diag(\sigma_1, \ldots, \sigma_q)$, and consider $X_{(q)}$ defined as

$$X_{(q)} = XV_{(q)}V_{(q)}^T, \tag{40}$$

that is $X_{(q)}$ consists of rows of X orthogonally projected on $span\{v_1, \ldots, v_q\}$.
Using (39) and the orthogonality of v_j we obtain

$$X_{(q)} = \sum_{j=1}^{p} \sigma_j u_j v_j^T \sum_{j=1}^{p} v_j v_j^T = \sum_{i=1}^{q} \sigma_j u_j v_j^T = U_{(q)}\Sigma_{(q)}V_{(q)}^T.$$

Now assume A and B are $n \times p$ matrices. The Frobenius dot product on the linear space of $n \times p$ is defined as

$$\langle A, B\rangle_F := tr(A^TB) = \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}b_{ij}. \tag{41}$$

It induces the Frobenius norm

$$||A||_F^2 = \langle A, A\rangle_F = tr(A^TA) = tr(AA^T) = ||A^T||_F^2. \tag{42}$$

Finally we can formulate the following theorem (Golub and Van Loan 2013).

Theorem 2 (Eckart-Young-Mirsky). *Let M be an $n \times p$ matrix of rank q, and $X_{(q)}$ defined as above. Then*

$$||X - X_{(q)}||_F \leq ||X - M||_F,$$

and

$$||X - X_{(q)}||^2 = \sum_{j=q+1}^{p} \sigma_j^2.$$

2.3.3 Problem formulation and solution

We begin with the following formulation of PCA.

Assume we are given a data set $x_1, \ldots, x_n \in \mathbb{R}^p$. We wish to project the data set onto a q-dimensional linear subspace $L \subset \mathbb{R}^p$ that minimizes the *mean squared deviation* of x_i's from their projections x_i'.

Without loss of generality, assume that the data is centered, i.e.

$$\sum_{i=1}^{n} x_i = 0, \tag{43}$$

and $rank(X) = p$ (otherwise the data belongs to a lower dimension subspace).

Then the empirical covariance matrix $\widehat{C}$ can be written as

$$\widehat{C} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = \frac{1}{n} X^T X \tag{44}$$

Let P_L be the projection operator on L. Thus, mean squared deviation of x_i's from their projections $P_L x_i$ can be written as follows.

$$\frac{1}{n}\sum_{i=1}^{n} ||x_i - P_L x_i||^2 = \frac{1}{n}\sum_{i=1}^{n}((I - P_L)x_i)^T((I - P_L)x_i) =$$

$$= \frac{1}{n} tr(\sum_{i=1}^{n} x_i^T (I - P_L)^T (I - P_L) x_i) = \frac{1}{n} tr(\sum_{i=1}^{n} x_i x_i^T (I - P_L)^T (I - P_L))$$

$$= \frac{1}{n} tr(X^T X (I - P_L)^T (I - P_L)) = \frac{1}{n} tr((X - P_L X^T)^T (X^T - P_L X^T)) =$$

$$= \frac{1}{n} ||X^T - P_L X^T||_F^2 = \frac{1}{n} ||X - X P_L^T||_F^2 \tag{45}$$

If $rank(X P_L^T) = q$ then by Eckart-Young-Mirsky theorem one has

$$\frac{1}{n}\sum_{i=1}^{n} ||x_i - P_L x_i||^2 \geq ||X - X_{(q)}||_F^2, \tag{46}$$

where $X_{(q)}$ is defined by (40). If $rank(X P_L^T) = q' < q$ then $||X - X P_L^T||_F^2 \geq ||X - X_{(q')}||^2 > ||X - X_{(q)}||$, which follows form the second proposition of the theorem.

Thus, we proved that the optimal projector of rank q is $P_L = V_{(q)} V_{(q)}^T$, i.e. the orthogonal projector onto the linear span of $v_1, \ldots, v_q$. Note that by construction of SVD, $v_1, \ldots, v_p$ are the eigenvectors of the covariance matrix $\widehat{C}$. They are called the *principal directions* and $span\{v_1, \ldots, v_q\}$ is called the q-th *principal subspace*.

One can derive another formulation of PCA. Substituting P_L by $V_{(q)} V_{(q)}^T$ we rewrite the mean squared deviation as

$$\frac{1}{n}\sum_{i=1}^{n} ||x_i - V_{(q)} V_{(q)}^T x_i||^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^T (I - V_{(q)} V_{(q)}^T)^T (I - V_{(q)} V_{(q)}^T) x_i = \frac{1}{n}\sum_{i=1}^{n} x_i^T (I - V_{(q)} V_{(q)}^T) x_i.$$

To obtain the last equality we used the fact that $V_{(q)} V_{(q)}^T$ and $I - V_{(q)} V_{(q)}^T$ have orthogonal images due to the definition of orthogonal projector. Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n} ||x_i - V_{(q)} V_{(q)}^T x_i||^2 = tr(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T - x_i^T V_{(q)} V_{(q)}^T x_i) =$$

$$= tr(\widehat{C}) - \frac{1}{n} tr(\sum_{i=1}^{n} x_i^T (V_{(q)} V_{(q)}^T)(V_{(q)} V_{(q)}^T) x_i) =$$

$$= tr(\widehat{C}) - \frac{1}{n} \sum_{i=1}^{n} ||V_{(q)} V_{(q)}^T x_i||^2$$

Note that to derive this form of mean squared deviations we only use the orthogonality of $V_{(q)}$ and therefore it is true for any orthogonal projector $\widetilde{V}\widetilde{V}^T$ of rank q. Since $tr(\widehat{C})$ is a constant, minimizing mean squared deviations with respect to $\widetilde{V}$ is equivalent to solving the following optimization problem

$$\begin{gathered} \max_{\widetilde{V}} \frac{1}{n} \sum_{i=1}^{n} ||\widetilde{V}\widetilde{V}^T x_i||^2, \\ \text{s.t. } \widetilde{V}^T \widetilde{V} = I_q. \end{gathered} \tag{47}$$

where the minimized functional will be henceforth called the *empirical variance of the projected data* (not to be confused with the empirical covariance matrix).

The matrix $X_{(q)}$ still consists of p-dimensional row-vectors even though these vectors belong to the q-dimensional principal subspace. To obtain the columns of coordinates of the points $V_{(q)} V_{(q)}^T x_i$ in the basis $v_1, \dots, v_q$ of the principal subspace one simply calculates

$$V_{(q)}^T x_i \in \mathbb{R}^q.$$

These coordinates are called the *principal components* of x_i. Being transposed into rows and stacked together they principal components of all data points constitute the matrix

$$(V_{(q)}^T X^T)^T = X V_{(q)} = U_{(q)} \Sigma_{(q)} \in \mathbb{R}^{n \times q}.$$

Alternatively one can orthogonally transform X by multiplying it with $V = V_{(p)}$ to obtain

$$XV = U\Sigma =: \widetilde{X}.$$

Recall that $U\Sigma$ consists of the columns $\sigma_j u_j$, so the first q columns of it compose the matrix $U_{(q)}\Sigma_{(q)}$. That is one can first calculate $\widetilde{X}$ and obtain any $U_{(q)}\Sigma_{(q)}$ by taking first q columns of the former.

Furthermore, the empirical covariance matrix of the transformed data is

$$\frac{1}{n} \widetilde{X}^T \widetilde{X} = \frac{1}{n} V^T X^T X V = \frac{1}{n} V^T V D V^T v = \frac{1}{n} D, \tag{48}$$

i.e. the transformed data is empirically decorrelated. Summarizing the above written, PCA produces a representation of the data as coordinates (principal components) in the basis principal directions, i.e. eigenvector of the empirical covariance matrix. For any $1 \leq q \leq p$ the q-th principal subspace carries the most empirical variance of the data among all projections of rank q, or equivalently, it gives the best approximation of the dataset in terms of mean squared deviation.

3 Kernel methods

In this section we begin the main part of the report. We will first provide a short heuristic motivation of the kernel approach to generalize linear machine learning algorithms. Then we will present the basic elements of the RKHS theory. Finally, we will show how the optimization problems we dealed with are extended from $\mathbb{R}^p$ to RKHS.

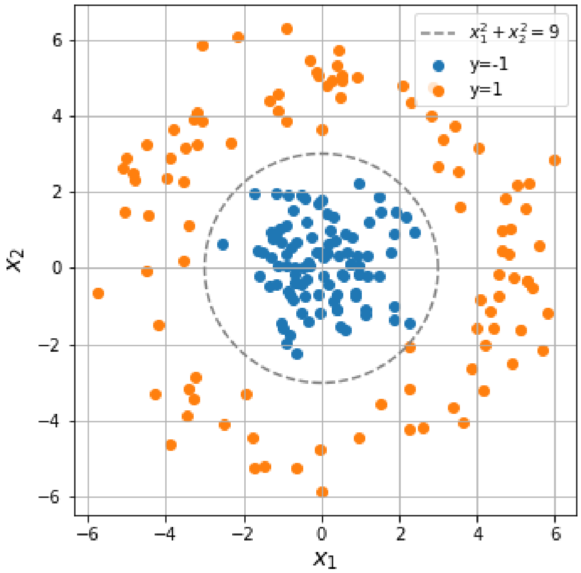

Figure 8: Linearly non-separable data

3.1 Motivation

We would like to generalize the linear machine learning algorithms to work with more complex non-linear patterns. One way doing that is by taking a non-linear transformation of the data and thus mapping it into a generally higher dimensional linear space, where we can use linear methods.

For the sake of readability in this introductory subsection we use x_j to denote the j-th entry of a vector x, whereas in general we use it to denote the j-th data point.
Consider binary classification problem with linearly non-separable data as illustrated on the figure 10. Using linear classifier like SVM does not make sense for the originally given data. However, the classes $y = 1$ and $y = -1$ can be separated by a second order polynomial curve $x_1^2 + x_2^2 = 9$. Then if we transform the data with the map $\psi : (x_1, x_2) \mapsto (x_1, x_2, x_3 = x_1^2 + x_2^2)$ into $\mathbb{R}^3$, the data can be separated in the image space by the hyperplane $x_3 = 9$.

In this toy example it is easy to visually guess the suitable transformation to make the data appropriate for linear SVM. In general we need a higher dimensional transformation that allows for a large enough variety of non-linear subsets in the original domain of the data to have linear images in the new space.

Suppose the data belongs to $\mathbb{R}^p$ and we want to include all the polynomials of $x_1, \ldots, x_p$ of order $d = 2$ to the feature space. Consider the quadratic map

$$\phi : (x_1, \ldots, x_p) \mapsto (x_1, x_2, \ldots, x_p, x_1^2, \ldots, x_p^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \ldots, \sqrt{2}x_{p-1}x_p),$$

that maps the data into a higher dimensional euclidean space. This feature map has roughly $p^2/2$ entries and may be computationally costly for large p. However, as discussed in subsection 2.2 we only need to calculate the dot products $\langle \phi(x), \phi(y) \rangle$ to fit SVM and predict class labels. It is easy to verify that the dot product can be calculated as below

$$\langle \phi(x), \phi(y) \rangle = x^T y + (x^T y)^2,$$

which makes the computation rather simple. Similarly to SVM, we will show that the machine learning models we described can be fitted and used on new data without explicit calculations of the feature maps.

We are not limited to polynomial transformations and we may search for classifiers and predictors in a large variety of functional spaces. But we also expect that if predictors f and g are close enough in such a space, their outputs $f(x)$ and $g(x)$ on test data are close as well.

3.2 Reproducing kernel Hilbert spaces

Let $\mathcal{H}$ be a Hilbert space of real-valued functions $f : \mathcal{X} \to \mathbb{R}$, i.e. $\mathcal{H}$ is equipped with a dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ that induces a norm $||f||^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}$ and $\mathcal{H}$ is complete.
Suppose that for any $x \in \mathcal{X}$ the evaluation functional $h_x : f \mapsto f(x)$ is continuous in $(\mathcal{H}, || \cdot ||)$, i.e. h_x belongs to the dual space $\mathcal{H}^*$. Then $\mathcal{H}$ is called a *reproducing kernel Hilbert space*.

By the Riesz representation theorem for any $x \in \mathcal{X}$ there exists a unique $g \in \mathcal{H}$ such that $h_x[f] = f(x) = \langle g, f \rangle_{\mathcal{H}}$. Moreover Riesz representation gives a natural isometry between $\mathcal{H}$ and $\mathcal{H}^*$, so we will hereafter abuse the notation and denote both the evaluation functional and it's representer as h_x. Using that we can calculate the value of the function h_x at any given point $y \in \mathcal{X}$ as

$$h_x(y) = \langle h_x, h_y \rangle_{\mathcal{H}}. \tag{49}$$

The function $k : \mathcal{X}^2 \to \mathbb{R}$ defined as $k(x, y) = \langle h_x, h_y \rangle_{\mathcal{H}}$ is called the *reproducing kernel* of $\mathcal{H}$. Note that by definition $k(x, y) = h_x(y)$ for any y, which means that the function $k(x, \cdot)$ is identical to h_x and thus belongs to $\mathcal{H}$. Equivalently, a reproducing kernel of $\mathcal{H}$ can be defined as a function $k : \mathcal{X}^2 \to \mathbb{R}$, such that $k(x, \cdot) \in \mathcal{H}$ and $f(x) = \langle k(x, \cdot), f \rangle$ for any $x \in \mathcal{X}, f \in \mathcal{H}$. It is easy to show that such function is unique for a given RKHS: assume that $\widetilde{k}$ is also a reproducing kernel of $\mathcal{H}$. Then

$$\langle k(x, \cdot) - \widetilde{k}(x, \cdot), f \rangle = f(x) - f(x) = 0, \forall x \in \mathcal{X}, \forall f \in \mathcal{H},$$

which means that the function $k(x, \cdot) - \widetilde{k}(x, \cdot)$ equals zero for any $x \in \mathcal{X}$, that is $k = \widetilde{k}$.

The reproducing kernel is obviously symmetric, and moreover it is a positive definite function, i.e. for any set of points $x_1, \dots, x_n \in \mathcal{X}$ the matrix $K = (k(x_i, x_j))^n_{i,j=1}$ is positive semi-definite. Indeed, let $\alpha \in \mathbb{R}^n$, then

$$\alpha^T K \alpha = \sum_{i,j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j = \langle \sum_{i=1}^{n} \alpha_i h_{x_i}, \sum_{j=1}^{n} \alpha_j h_{x_j} \rangle_{\mathcal{H}} = || \sum_{i=1}^{n} \alpha_i h_{x_i} ||^2_{\mathcal{H}} \geq 0,$$

and the identity holds iff $\sum_{i=1}^n \alpha_i h_{x_i}$=0.

We showed above explicitly that for any reproducing kernel Hilbert space there exists a unique positive definite function k, such that $f(x) = \langle k(x, \cdot), f \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$, that is a reproducing kernel. Turns out that the reverse is also true.

Theorem 3 (Aronsajn-Moore)**.** *Assume $k(x, y) : \mathcal{X} \to \mathbb{R}$ is a symmetric, positive definite function (kernel). Then there exists a unique Hilbert space of functions $\mathcal{H}$ such that for any $x \in \mathcal{X}, f \in \mathcal{H}$ the following equation is satisfied $f(x) = \langle k(x, \cdot), f \rangle_{\mathcal{H}}$, i.e. k is the reproducing kernel of $\mathcal{H}$.*

The function $\phi \in \mathcal{X} \to \mathcal{H}$ that maps $x \mapsto k(x, \cdot)$ is called the canonical *feature map*. Considering Theorem 3 we will hereafter call any symmetric positive definite function a reproducing kernel.

Examples of reproducing kernels:

- Linear kernel: let $\mathcal{X} = \mathbb{R}^p$ be a Euclidean space and consider the kernel $k(x,y) = \langle x, y\rangle = x^T y$. The associated RKHS is the space of linear functionals in $\mathbb{R}^p$ of the form $f_w(y) = w^T y$ with the norm $||f_w||_{\mathcal{H}} = ||w||$, and the corresponding feature map $\psi : w \mapsto f_w$ that can be viewed as identity. So in this case $\mathcal{X} \equiv \mathcal{H}$.

- Polynomial kernel: suppose $\mathcal{X} = \mathbb{R}^p$ and let $k(x,y) = (1 + \alpha x^T y)^d, \;\; \alpha \in \mathbb{R} \setminus \{0\}, d \in \mathbb{N}$. The corresponding RKHS is the space of polynomials of order up to d and is a finite dimensional linear space.

- Gaussian (Radial basis function) kernel: suppose $X = \mathbb{R}^p$ and $k(x,y) = \exp\{-\frac{||x-y||^2}{2a^2}\}, \;\; a > 0$. The associated RKHS is an infinite dimensional Hilbert space called Gaussian RKHS.

- Laplacian kernel: $k(x,y) = \exp\{-\frac{||x-y||}{a}\}, \;\; a > 0$ also induces an infinite dimensional RKHS.

Combinations of kernels: Suppose the k_1 and k_2 are reproducing kernels on $\mathcal{X}^2$. Then $\widetilde{k}$ defined as below is also a reproducing kernel.

- $\widetilde{k}(x,y) = \alpha_1 k_1(x,y) + \alpha_2 k_2(x,y)$ for any $\alpha_1, \alpha_2 > 0$.

- $\widetilde{k}(x,y) = k_1(x,y) k_2(x,y)$.

- $\widetilde{k}(x,y) = \sum_{z \in \mathcal{X}} k(x,z) k(z,y)$.

- $\widetilde{k}(x,y) = \dfrac{k_1(x,y)}{k(x,x)k(y,y)}$.

Moreover, if the kernels $k_x(x,x')$ and $k_y(y,y')$ are defined in $\mathcal{X}^2$ and $\mathcal{Y}^2$ respectively, then the function $k((x,x'),(y,y')) = k_x(x,x') k_y(y,y')$ is kernel defined on the product space $(\mathcal{X} \times \mathcal{Y})^2$.

3.3 Kimmeldorf and Wahba theorem

The following theorem provides a useful result for a large variety of learning problems in RKHS's.

Theorem 4 (Kimmeldorf-Wahba)**.** *Suppose $\mathcal{H}$ is a RKHS and we are given a minimization problem*

$$\min_{f \in \mathcal{H}} L(f(x_1), \ldots, f(x_n)) + \lambda ||f||^2_{\mathcal{H}}. \tag{50}$$

Then if f is a local minimum, there exists $\alpha \in \mathbb{R}^n$ such that the following holds

$$f = \sum_{i=1}^n \alpha_i k(x_i, \cdot). \tag{51}$$

Proof. Consider the linear subspace $S = span\{k(x_1,\cdot), \ldots, k(x_n,\cdot)\}$ and it's orthogonal $S^\perp$. For an arbitrary $f \in \mathcal{H}$ there exist a unique decomposition $f = f_S + f_{S^\perp}$, such that $f_S \in S, f_{S^\perp} \in S^\perp$. By Pythagoras theorem we have

$$||f||^2_{\mathcal{H}} = ||f_S||^2_{\mathcal{H}} + ||f_{S^\perp}||^2_{\mathcal{H}},$$

and by the definition of reproducing kernel

$$f(x_i) = \langle k(x_i,\cdot), f_S + f_{S^\perp}\rangle = \langle k(x_i,\cdot), f_S\rangle = f_S(x_i).$$

Thus the minimized functional for f can be written as

$$J(f) = L(f_S(x_1), \dots, f_S(x_n)) + \lambda(||f_S||^2_{\mathcal{H}} + ||f_{S^\perp}||^2_{\mathcal{H}}).$$

If $f \notin S$, one can obtain a lower value of J within any neighbourhood of f by reducing the orthogonal component $f_{S^\perp}$, which means that f is not a local minimum. Therefore any local minimum belongs to the linear span of $k(x_1, \cdot), \dots, k(x_n, \cdot)$. □

3.4 Kernel ridge regression

Let $\mathcal{H}$ be a RKHS with the reproducing kernel k. We now want to generalized the ridge regression to solve non-linear regression problem by considering predictors $f \in \mathcal{H}$. That is we solve the following optimization problem.

$$\min_{f \in \mathcal{H}} J_\lambda(f) = \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda ||f||^2_{\mathcal{H}}. \tag{52}$$

From Kimmeldorf-Wahba theorem it follows that the optimal predictor is of the form

$$f_\alpha = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot). \tag{53}$$

Substituting $f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ and considering the $n \times n$ matrix $(K)_{ij} = k(x_i, x_j)$ we can reformulate the original problem as

$$\min_\alpha \frac{1}{n} ||y_i - K\alpha||^2 + \lambda \alpha^T K \alpha. \tag{54}$$

Setting the derivative w.r.t α to zero we obtain

$$\frac{\partial J_\lambda}{\partial \alpha} = \frac{2}{n} \left(-Ky + K^2\alpha + n\lambda K\alpha \right) = 0.$$

The matrix K is non-negative and therefore J_λ is convex w.r.t. to α, which yields that the optimal α has to satisfy the equation

$$K(y - (K + n\lambda I_n)\alpha) = 0. \tag{55}$$

The matrix $(K + n\lambda I_n)$ is invertible as its eigenvalues are no less then $n\lambda$. Consider $\widehat{\alpha} = (K + n\lambda I_n)^{-1} y$.

Lemma 1. *For any α such that $K(y - (K + n\lambda I_n)\alpha) = 0$, the following holds*

$$f_\alpha = f_{\widehat{\alpha}}. \tag{56}$$

Proof. The matrix K admits spectral decomposition $K = U\Lambda U^T$ with $U^T U = UU^T = I_n$. We then can write

$$\begin{aligned} K(K + n\lambda I_n)^{-1} &= U\Lambda U^T U(\Lambda + n\lambda I_n)^{-1} U^T = U\Lambda(\Lambda + n\lambda I_n)^{-1} U^T \\ &= U(\Lambda + n\lambda I_n)^{-1} \Lambda U^T = (K + n\lambda I_n)^{-1} K, \end{aligned}$$

where we used the fact that the matrices $(\Lambda + n\lambda I_n)^{-1}$ and Λ commute due to their diagonality. Hence, if $h \in \mathbb{R}^n$ and $Kh = 0$ then $K(K + n\lambda K)^{-1}h = (K + n\lambda K)^{-1}Kh = 0$.
Furthermore, if $Kh = 0$ then $\sum_{i=1}^n h_i k(x_i, \cdot) \equiv 0$:

$$||\sum_{i=1}^n h_i k(x_i, \cdot)||^2 = \langle\sum_{i=1}^n h_i k(x_i, \cdot), \sum_{j=1}^n h_j k(x_j, \cdot)\rangle = \sum_{i=1}^n \sum_{j=1}^n h_i k(x_i, x_j) h_j = h^T K h = 0. \tag{57}$$

If α satisfies (55), taking $h = y - (K + n\lambda I_n)\alpha$ one has $Kh = 0$ and consequently

$$0 = K(K + n\lambda I_n)h = K((K + n\lambda I_n)^{-1}y - \alpha).$$

Finally we obtain that for any $x \in \mathcal{X}$

$$f_{\widehat{\alpha}}(x) - f_\alpha(x) = \sum_{i=1}^n k(x_i, x)(\widehat{\alpha}_i - \alpha_i) = \sum_{i=1}^n k(x_i, x)((K + n\lambda I_n)^{-1}y - \alpha)_i = 0$$

□

From Lemma 1 it follows that any α for which (55) holds is a solution to (54), including $\widehat{\alpha} = (K + n\lambda I_n)^{-1}y$. Thus, we obtained a closed form solution for ridge regression in RKHS:

$$f(x)_{\widehat{\alpha}} = \sum_{i=1}^n \widehat{\alpha}_i k(x_i, x), \text{ where } \ \widehat{\alpha} = (K + n\lambda I_n)^{-1}y. \tag{58}$$

3.5 Kernel SVM

In this subsection we will extend the soft-margin SVM from $\mathbb{R}^p$ to reproducing kernel Hilbert spaces. We first consider the problem of regularized empirical risk minimization associated with the loss function $l : \{1, -1\} \times \mathbb{R} \to \mathbb{R}$

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n l(y_i, f(x_i)) + \lambda ||f||^2_{\mathcal{H}}. \tag{59}$$

By Kimmeldorf-Wahba theorem any local minimum is of the form

$$f_\alpha = \sum_{j=1}^n \alpha_j \phi(x_j) = \sum_{j=1}^n \alpha_j k(x_j, \cdot). \tag{60}$$

Following Theorem 5.5 in Steinwart and Christmann 2008, if the loss function $l(y, f(x))$ is convex w.r.t. to $f(x)$ there exists a unique solution to (59), while α is in general not unique. From (60) it follows that it is sufficient to optimize f over the linear span of $k(x_i, \cdot)$. So we can rewrite this problem as

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n l(y_i, f(x_i)) + \lambda ||f||^2_{\mathcal{H}} =$$
$$= \min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n l(y_i, \sum_{j=1}^n \alpha_j k(x_j, x_i)) + \lambda \alpha^T K \alpha. \tag{61}$$

Thus, we can apply classical results of convex finite-dimensional optimization. Now recall that in subsection 2.2.3. we formulated soft-margin SVM problem in the form of (68) with hinge loss $l(y, a) = \max\{0, 1 - ya\}$. Repeating the trick in (29) we obtain a constrained problem with slack variables equivalent to (61).

$$\min_{\alpha,\xi_i} C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\alpha^T K \alpha, \tag{62}$$

$$\xi_i \geq 1 - y_i \sum_{j=1}^{n} \alpha_j k(x_j, x_i), \;\; \xi_i \geq 0 \;\; i = 1, \ldots, n, \tag{63}$$

where $C = 1/(2n\lambda)$. The matrix K is positive semi-definite and the hinge loss is a convex function, therefore the problem (62)-(63) is convex. The corresponding Lagrangian is given as

$$\Lambda(\alpha, \xi) = C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\alpha^T K \alpha + \sum_{i=1}^{n} \lambda_i (1 - y_i \sum_{j=1}^{n} \alpha_j k(x_j, x_i) - \xi_i) - \sum_{i=1}^{n} \nu_i \xi_i, \tag{64}$$

with $\lambda_i \geq 0$ and $\nu_i \geq 0$. The stationarity conditions are given as

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{n} \alpha_i k(x_i, x_j) - \sum_{i=1}^{n} \lambda_i y_i k(x_i, x_j) = 0, \;\; j = 1, \ldots n, \tag{65}$$

$$\frac{\partial L}{\partial \xi_i} = C - \lambda_i - \nu_i = 0, \;\; i = 1, \ldots n. \tag{66}$$

Note that f is uniquely defined by n dot products $\langle f, \phi(x_j)\rangle_{\mathcal{H}}$, i.e. its values at x_j, since $f \in span\{\phi(x_1), \ldots, \phi_(x_n)\}$ (repeat (57) if unsure). Hence (65) implies that $f = \sum_{i=1}^{n} \lambda_i y_i \phi(x_i)$ and we can identify α as $\alpha_i = \lambda_i y_i$. Finally, using stationarity conditions we write the dual problem as

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j k(x_i, x_j), \tag{67}$$

$$0 \leq \lambda_i \leq C, \tag{68}$$

from which the decision function is restored as $f(x) = \sum_{j=1}^{n} \lambda_j y_j k(x_j, x)$. The constraints (68) are due to $\lambda_i \geq 0, \nu_i \geq 0$ and $C = \lambda_i + \nu_i$. The support vectors are the feature vectors $k(x_j, \cdot)$ for which $y_i \sum_{j=1}^{n} \alpha_j k(x_j, x_i) \leq 1$ and the sparcity of the solution is obtained using KKT conditions, the same way it is done in subsection 2.2.3.

We have shown that to solve the SVM optimization problem in possibly an infinite-dimensional RKHS, we actually need to solve the same finite-dimensional quadratic programming problem as in the case of $\mathbb{R}^p$, but instead of dot products $\langle x_i, x_j\rangle$ we have $k(x_i, x_j)$.

3.6 Kernel PCA

Let $\mathcal{H}$ be a RKHS with the associated reproducing kernel k and feature map $\phi : \mathcal{X} \to \mathcal{H}, \;\; \phi(x) = k(x, \cdot)$. We are given a dataset $x_1, \ldots, x_n$ that we transform to $\phi(x_1), \ldots, \phi(x_n)$. Our goal is to extract the principal components of features $\phi(x_i)$. For the sake of simplicity, we assume for now that the data is centered in the feature space, that is

$$\sum_{i=1}^{n} \phi(x_i) = 0,$$

but we will return to it later. The empirical covariance matrix in feature space is given by

$$\widehat{C} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\phi(x_i)^T.$$

If $\mathcal{H}$ is infinite-dimensional the notation $\phi(x_i)\phi(x_i)^T$ should be read as a linear operator

$$\phi(x_i)\phi(x_i)^T : u \mapsto \phi(x_i)\langle \phi(x_i), u\rangle_{\mathcal{H}}. \tag{69}$$

The image of $\widehat{C}$ belongs to the finite dimensional subspace $span\{\phi(x_1), \dots, \phi(x_n)\}$. We want to find its eigenvalues $\lambda \geq 0$ and the corresponding non-zero eigenvectors $v \in \mathcal{H} \setminus \{0\}$:

$$\lambda v = \widehat{C}v. \tag{70}$$

Since the solutions v belong to the linear span of $\phi(x_i)$ for each v there exists $\alpha = (\alpha^1, \dots, \alpha^n)^T$ such that

$$v = \sum_{i=1}^{n} \alpha^i \phi(x_i). \tag{71}$$

Plugging this into (70) and taking dot products with $\phi(x_j)$ one obtains the system of equations

$$\lambda \sum_{i=1}^{n} \langle \phi(x_j), \phi(x_i)\rangle = \frac{1}{n}\sum_{i=1}^{n} \alpha^i \langle \phi(x_j), \sum_{l=1}^{n} \phi(x_l)\langle \phi(x_l), \phi(x_i)\rangle\rangle, \;\; j = 1, \dots, n. \tag{72}$$

Considering the $n \times n$ Gram matrix $K = (\langle \phi(x_i), \phi(x_j)\rangle)_{i,j}$ one has

$$\lambda K\alpha = \frac{1}{n}K^2\alpha. \tag{73}$$

Thus to obtain the solution v one needs to solve (73) for α. For non-zero α this is equivalent to solving

$$\lambda\alpha = \frac{1}{n}K\alpha. \tag{74}$$

Let $\lambda_1 \geq \dots, \lambda_n$ be the eigenvalues of $\frac{1}{n}K$ and $\alpha_1, \dots, \alpha_n$ be the associated eigenvectors, the latter are the principal directions. Then to extract the first q principal components of a point $x \in \mathcal{X}$ with the image $\phi(x)$ one calculates

$$\langle v_j, \phi(x)\rangle = \sum_{i=1}^{n} \alpha_j^i k(x_i, x).$$

Previously we made the assumption that the data is centered in $\mathcal{H}$. We avoid explicit computation of the feature map and thus explicit centering of the images of the data in $\mathcal{H}$. Turns out that centering the images corresponds to transforming the kernel matrix

$$K \mapsto \widetilde{K} = (K - \mathbf{1}_n K - K\mathbf{1}_n + \mathbf{1}_n K \mathbf{1}_n),$$

where $\mathbf{1}_n$ denoted the $n \times n$ matrix with entries $1/n$. The following theorem states the optimality properties of kernel PCA (Schölkopf and Smola 2002).

Theorem 5. *The first q principal components carry the most empirical variance among all projections on any q directions. Equivalently, the mean-squared deviation it the lowest for the first q principal components among all projections of the data in $\mathcal{H}$ on any q directions.*

This implies that using PCA we can extract the part of the data in $\mathcal{H}$ that lives in the q-dimensional subspace of $\mathcal{H}$, keeping as much information of our data in terms of variance.

4 Kernel two-sample and independence testing

Reproducing kernels Hilbert spaces also appear to be a powerful approach for two-sample and independence testing. Proposed by Gretton et al (2007, 2012), kernel two-sample test is a family of criterions based on statistical estimators of maximum mean discrepancies in RKHS's.

Suppose $(\mathcal{X}, d)$ and $(\mathcal{Y}, \rho)$ are metric spaces. Random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively are independent if for any two Borel subsets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$ the following holds

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B). \tag{75}$$

When $(\mathcal{X}, d) = (\mathcal{Y}, \rho)$, we say that X and Y are identically distributed if for any Borel subset $A \in \mathcal{X}$ the following holds

$$P(X \in A) = P(Y \in A) \tag{76}$$

Let P_X and P_Y be the probability distributions of X and Y in $\mathcal{X}$ and $\mathcal{Y}$, and P_{XY} be their joint distribution in $\mathcal{X} \times \mathcal{Y}$ respectively. We may rewrite conditions (75) and (76) as follows:

$$P_{XY} = P_X \times P_Y \Leftrightarrow (75),$$

$$P_X = P_Y \Leftrightarrow (76).$$

The problem of testing hypothesis on the distribution of two random variables is called two sample testing when $P_X = P_Y$ is the null hypothesis, and is called independence testing when the null hypothesis is $P_{XY} = P_X \times P_Y$.

Let X and Y be two random variables in a metric space $(\mathcal{X}, \rho)$ with Borel probability measures P_X and P_Y respectively. Let $\mathcal{F}$ be a class of Borel functions $f : \mathcal{X} \to \mathbb{R}$. The *maximum mean discrepancy* (MMD) of P_X and P_Y over the class $\mathcal{F}$ is defined as

$$MMD(P_X, P_Y, \mathcal{F}) = \sup_{f \in \mathcal{F}} |\mathbb{E}_X f - \mathbb{E}_Y f|, \tag{77}$$

where $\mathbb{E}_X f = \mathbb{E} f(X)$. Given two samples of i.i.d. random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ drawn from distributions P_X and P_Y respectively, we have the following biased (in general) empirical estimator of MMD

$$MMD_b = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{m} \sum_{i=1}^{m} f(Y_i) \right|.$$

Recall the following lemma.

Lemma 2. *Let P_X and P_Y be two probability measures on $\mathcal{X}$. Then P_X and P_Y are identical if and only if for any bounded, continuous, real-valued function $f : \mathcal{X} \to \mathbb{R}$ the following is satisfied*

$$\mathbb{E}_X f = \mathbb{E}_Y f.$$

We use $C(\mathcal{X})$ to denote the space of bounded, continuous, real-valued functions. For the sake of simplicity we say that a class $\mathcal{F}$ is dense in $C(\mathcal{X})$, if for any $f \in C(\mathcal{X})$ there exists a sequence $f_n \in \mathcal{F}$ such that $\sup_{\mathcal{X}} |f_n(x) - f(x)| \to 0$, even if $\mathcal{F}$ does not belong to $C(\mathcal{X})$. It follows from Lemma 2 that whenever $\mathcal{F}$ is dense in a neighbourhood of zero in $C(\mathcal{X})$, the MMD over $\mathcal{F}$ is a distance between distributions, that is $MMD(P_X, P_Y, \mathcal{F}) = 0$ iff $P_X = P_Y$.

Gretton, Borgwardt, et al. 2012 specified the function classes that are rich enough to identify whether $P_X = P_X$, yet restrictive enough to provide useful finite sample estimates of MMD.

Let $\mathcal{H}$ be a RKHS of functions on $\mathcal{X}$ with reproducing kernel k and let $\mathcal{F} = \{f \in \mathcal{H} : ||f||_{\mathcal{H}} \leq 1\}$ be the unit ball in $\mathcal{H}$. Assume that the linear functionals $\mathbb{E}_X : f \mapsto \mathbb{E}f(X)$ and $\mathbb{E}_Y : f \mapsto \mathbb{E}f(Y)$ are bounded. By the Riesz representations theorem, there exist $\mu_X \in \mathcal{H}$ and $\mu_Y \in \mathcal{H}$ such that $\mathbb{E}_X f = \langle f, \mu_X \rangle$ and $\mathbb{E}_Y f = \langle f, \mu_Y \rangle$. Hence one has

$$MMD(P_X, P_Y, \mathcal{F}) = \sup_{f \in \mathcal{F}} (\langle \mu_X, f \rangle - \langle \mu_Y, f \rangle) = \sup_{f \in \mathcal{F}} \langle \mu_X - \mu_Y, f \rangle = ||\mu_X - \mu_Y||_{\mathcal{H}}. \tag{78}$$

In the last equality we use the isometry of $\mathcal{H}$ and $\mathcal{H}^*$. Using the reproducing property we obtain

$$\langle \mu_X, \mu_Y \rangle = \int_{\mathcal{X}} \mu_X(y) dP_Y(y) = \int_{\mathcal{X}} \langle \mu_X, k(y, \cdot) \rangle dP_Y(y) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) dP_X(x) dP_Y(y) = \mathbb{E}_X \mathbb{E}_Y k.$$

Then using the property of Euclidean norm $||\mu_X - \mu_Y||^2_{\mathcal{H}} = ||\mu_X||^2 + ||\mu_Y||^2 - 2\langle \mu_X, \mu_Y \rangle$ one has

$$MMD(P_X, P_Y, \mathcal{F})^2 = \mathbb{E}k(X, X') + \mathbb{E}k(Y, Y') - 2\mathbb{E}k(X, Y), \tag{79}$$

where X, X' i.i.d. $\sim P_X$ and Y, Y' i.i.d. $\sim P_Y$.

The function μ_X is called the *mean embedding* of the distribution P_X in $\mathcal{H}$ and is an extension of the the feature map. The mean embedding exists if and only if the linear functional $f \mapsto \mathbb{E}_X f$ is bounded in $\mathcal{H}$. The sufficient condition for $\mathbb{E}_X(\cdot)$ to be bounded is fairly simple:

Lemma 3. *If $k(x, y)$ is measurable (w.r.t. Borel sigma-algebra $\mathcal{B}(\mathcal{X})$) and $\mathbb{E}_X k(X, X)^{1/2} < \infty$ then $\mu_X \in \mathcal{H}$.*

Proof. Using Cauchy-Schwarz inequality one has

$$|\mathbb{E}_X f| \leq \mathbb{E}_X |f| = \mathbb{E}_X |\langle k(x, \cdot), f \rangle| \leq \mathbb{E}_X (k(x, x)^{1/2} ||f||_{\mathcal{H}}) \leq ||f||_{\mathcal{H}} \mathbb{E}_X k(x, x)^{1/2}$$

□

In particular, if $k(x, x)$ is uniformly bounded on $\mathcal{X}$ the integrability of $k(x, x)^{1/2}$ holds for any probability measure.

A continuous kernel k defined on a compact metric space $\mathcal{X}$ is said to be *universal* (or c-universal) if the associated $\mathcal{H}$ is dense in $C(\mathcal{X})$. The class of universal kernels includes Gaussian (RBF) and Laplacian kernel. Fukumizu et al. 2007 introduced a more general notion: a bounded measurable kernel k is said to be *characteristic* if the mean embedding $P_X \mapsto \mu_X$ is an injective map on the space of Borel probability measures on $\mathcal{X}$ (which does not need to be compact), that is when MMD is a metric between probability measures. They also established that Gaussian and Laplacian kernels are characteristic in $\mathbb{R}^d$.As shown by Gretton, Borgwardt, et al. 2012 any universal kernel is characteristic (on a compact domain).

One can see that the elegant representation (79) of MMD over the unit ball in $\mathcal{H}$ allows to construct relatively simple natural estimators of MMD without direct computation of a supremum over an infinite class of functions.

4.1 Kernel two-sample test

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are i.i.d rv's drawn from the distributions P_X and P_Y respectively. The biased and unbiased estimators of MMD^2 are defined as

$$MMD_b^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j), \tag{80}$$

$$MMD_u^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j). \quad (81)$$

Note that MMD_u^2 is obtained by removing the diagonal terms $k(x_i, x_i)$ and $k(y_i, y_i)$ from MMD_b^2. These terms ensured non-negativity of the the statistic and therefore MMD_u^2 can take negative values.

Using the weak LLN and Chebyshev inequality one can show that these statistics are consistent estimators of $MMD^2(P_X, P_Y, \mathcal{F})$ with $\mathcal{F}$ being the unit ball in $\mathcal{H}$ with associated kernel k. When k is a characteristic kernel, one has

$$MMD_b^2 \xrightarrow{P} 0, \;\; MMD_u^2 \xrightarrow{P} 0 \text{ as } n \to \infty \quad (82)$$

if and only if $P_X = P_Y$. Here $\xrightarrow{P}$ denotes convergence in probability.

For simplicity we focus on the case when $n = m$. The following theorem allows to build two-sample test based on the estimators above (Gretton, Borgwardt, et al. 2012).

Theorem 6. *Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be defined as above.If $P_X \neq P_Y$, then $nMMD_b^2 \xrightarrow{P} \infty$. If $P_X = P_Y$ then*

$$nMMD_b^2 \xrightarrow{D} Q_1 = \sum_{i=1}^{\infty} \lambda_i \eta_i^2, \quad (83)$$

and

$$nMMD_u^2 \xrightarrow{D} Q_2 = \sum_{i=1}^{\infty} \lambda_i (2\eta_i^2 - 2) \quad (84)$$

where η_i are i.i.d. $N(0,1)$ rv's, and $\lambda_i \geq 0$ are the eigenvalues of the operator $T_k : L_2(P) \to \mathcal{H}$

$$T_k \psi(\cdot) = \int_X \widetilde{k}(x, \cdot) \psi(x) dP_X(x), \quad (85)$$

where $\widetilde{k}(x,y) = k(x,y) - \mathbb{E}_X(X, y) - \mathbb{E}_Y(x, Y) + \mathbb{E}_X \mathbb{E}_Y(X, Y)$ is the centered version of k.

If follows from Theorem 6 that if the null hypothesis $H_0 : P_X = P_Y$ holds, then for any $0 \leq \alpha \leq 1$ one has

$$\lim_{n \to \infty} P(nMMD_b^2 \leq y_{1-\alpha}) = 1 - \alpha, \quad (86)$$

$$\lim_{n \to \infty} P(nMMD_u^2 \leq z_{1-\alpha}) = 1 - \alpha, \quad (87)$$

where $y_{1-\alpha}$ and $z_{1-\alpha}$ are the $1 - \alpha$ quantiles of the random variables Q_1 and Q_2 respectively.

Gretton, Borgwardt, et al. 2012 provided several options to estimate the $1 - \alpha$ quantiles of Q_1 and Q_2. The first way is to use bootstrap resampling on the aggregated data following Arcones and Gine 1992, which is a very time expensive approach.

A more preferable way is to approximate the CDF of Q_2 with a Pearson curve fitted to the first four moments of MMD_u^2. It has $O(n^3)$ computational complexity as noted in Gretton, Fukumizu, Harchaoui, et al. 2009.

The third approach uses estimators of the first two moments of MMD_b^2 to approximate the null distribution of Q_1 with a two-parameter Gamma distribution. Specifically, we approximate

$$nMMD_b^2 \sim \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, \text{ where } \alpha = \frac{(\mathbb{E}MMD_b)^2}{Var(MMD_b)}, \;\; \beta = \frac{nVar(MMD_b)}{\mathbb{E}MMD_b}. \quad (88)$$

Although this approximations is necessarily less accurate than the Pearson curve, it performs remarkably well in the experiments and was considered to be the fastest way having $O(n^2)$ complexity. The fourth approach is to approximate the eigenvalues of T_k with the eigenvalues of the centered gram matrix HKH with $(K)_{i,j} = k(x_i, y_j)$ and $(H)_{i,j} = \delta_{i,j} - 1/n, \ \ i, j = 1, \dots, n$.

4.2 Kernel independence test

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces with reproducing kernels $k_x(x, x')$ and $k_y(y, y')$. Then $k((x, x'), (y, y')) = k_x(x, x')k_y(y, y')$ is a reproducing kernel in the product space $\mathcal{X} \times \mathcal{Y}$.
Let (X, Y) be a random vector in $\mathcal{X} \times \mathcal{Y}$ with the joint distribution P_{XY} and marginal distributions $X \sim P_X$ and $Y \sim P_Y$. The Hilbert-Schmidt (HSIC) independence criterion is defined as

$$HSIC(X, Y, \mathcal{F}) = MMD^2(P_{XY}, P_X \times P_Y, \mathcal{F}),$$

where $\mathcal{F}$ is the unit ball in the Hilbert space with the reproducing kernel k. If $\mathbb{E}_X k_x(X, X) < \infty$ and $\mathbb{E}_Y k_y(Y, Y) < \infty$, then feature embeddings of P_{XY} and $P_X \times P_Y$ are correctly defined (Sejdinovic et al. 2013).

Thus, using the reproducing property it can be written as

$$HSIC(X, Y, \mathcal{F}) = \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k_x(X, X')k_y(Y, Y') + \mathbb{E}_X\mathbb{E}_{X'}k_x(X, X')\mathbb{E}_{X'}\mathbb{E}_{Y'}k_y(Y, Y') -$$
$$-2\mathbb{E}_{X'Y'}[\mathbb{E}_X k_x(X, X')\mathbb{E}_Y k_y(Y, Y')],$$

where (X', Y') is an independent replica of (X, Y). Provided that the kernels k_x and k_y are characteristic, $HSIC(X, Y, \mathcal{F}) = 0$ if and only if $P_{XY} = P_X \times P_Y$.

Consider the sequence of i.i.d random vectors $(X_1, Y_1), \dots, (X_n, Y_n)$ drawn from the distribution P_{XY}. By substituting the expectations w.r.t. P_X, P_Y, P_{XY} and $P_X \times P_y$ to expectations w.r.t. to empirical distributions P_X^n, P_Y^n, P_{XY}^n and $P_X^n \times P_Y^n$ we obtain a biased estimator of HSIC

$$HSIC_b(X, Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} k_x(x_i, x_j)k_y(y_i, y_j) + \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} k_x(x_i, x_j)k_y(y_q, y_e) - \tag{89}$$
$$-\frac{2}{n^3}\sum_{i,j,q=1}^{n} k_x(x_i, x_j)k_y(y_i, y_q).$$

For pairs $z = (x, y)$ we can define the function $h(z_1, z_2, z_3, z_4)$

$$h(z_1, z_2, z_3, z_4) = \frac{1}{4!}\sum_{\pi \in S_4} k_x(x_{\pi(1)}, x_{\pi(2)})[k_y(y_{\pi(1)}, y_{\pi(2)}) + k_y(y_{\pi(3)}, y_{\pi(4)}) - 2k_y(y_{\pi(2)}, y_{\pi(3)})],$$

where the sum is taken over all permutations of 4 indices. Then $HSIC_b(X, Y)$ can be written as a V-statistic (Chwialkowski and Gretton 2014) associated with h

$$HSIC_b(X, Y) = \frac{1}{n}\sum_{i,j,q,r}^{n} h(w_i, w_j, w_q, w_r). \tag{90}$$

Finally $HSIC_b(X, Y)$ also allows the following matrix representation

$$HSIC_b(X, Y) = \frac{1}{n^2}Trace(K_x H K_y H), \tag{91}$$

where K_x is the $n \times n$ matrix $(K_x)_{ij} = k_x(x_i, x_j)$, K_y is the $n \times n$ matrix $(K_y)_{ij} = k_y(y_i, y_j)$, and $(H)_{i,j} = \delta_{i,j} - 1/n$ is the centering matrix.

Theorem 7. *If $P_{XY} \neq P_X \times P_Y$ then $nHSIC_b \xrightarrow{P} \infty$. If hypothesis $H_0 : P_{XY} = P_X \times P_Y$ holds one has*

$$nHSIC_b \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i \eta_i^2, \tag{92}$$

where η_i are i.i.d. rv's with distribution $N(0,1)$ and $\lambda_i \geq 0$ are the solutions of the eigenvalue problem

$$\lambda\psi(\cdot) = \int h(\cdot, z_1, z_2, z_3)\psi(w_1)dP_{z_1,z_2,z_3}.$$

Here z_i are i.i.d. $\sim P_X \times P_Y$.

To construct the test we need to approximate the $1-\alpha$ quantile of the limit distribution in Theorem 7. One way is to use permutation resampling which is computationally expensive. The second approach is to approximate the limit distribution with a two parameter Gamma distribution based on the estimators of two-first moments of $HSIC_b$. Specifically, Gretton, Fukumizu, Teo, et al. 2007 proposed the approximation

$$nHSIC_b \sim \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, \text{ where } \alpha = \frac{(\mathbb{E}HSIC_b)^2}{Var(HSIC_b)}, \ \beta = \frac{nVar(HSIC_b)}{\mathbb{E}HSIC_b}. \tag{93}$$

The estimators of these parameters can be computed in $O(n^2)$ time.

Summarizing, kernel two-sample and independence tests are powerful statistical criterions, which are relatively simple from the computational point of view. Given that the kernel is characteristic and the sample is large enough, they are guaranteed to identify with probability close to 1 if two distributions are identical and if they are independent. Some widely used statistical criterions like Pearson's R, Spearman's ρ and Kendall's τ can well recognize linear/monotonic dependence, but often fail to identify more complex types of association. More general independence tests based on domain partitioning like Pearson's Chi-square are limited to discrete and low-dimensional variables, otherwise they require extremely large sample sizes. Other MMD-based tests are not necessarily as efficient as the kernel ones, as not any functional class is dense enough in C and allows such a simple representation of MMD.

5 Simulation

In this section our goal is to demonstrate visually how kernel ridge regression, SVM and HSIC work on simple simulated datasets and recognize nonlinear patterns. Therefore, all the simulated data is two-dimensional.

The experiments we produced yield the following conclusion. Gaussian RBF kernel is a good default candidate kernel for nonlinear regression and classification as well as independence testing. It managed well enough all these tasks and it only has one hyperparameter to tune. Of course there are many kernels that may work much better for certain data, but Gaussian kernel can be used as a first approximation. Polynomials kernel (even of high degree) has rather limited power compared to Gaussian kernel.

Kernel independence test indeed does well at identifying nonlinear and non-monotonic types of dependence.

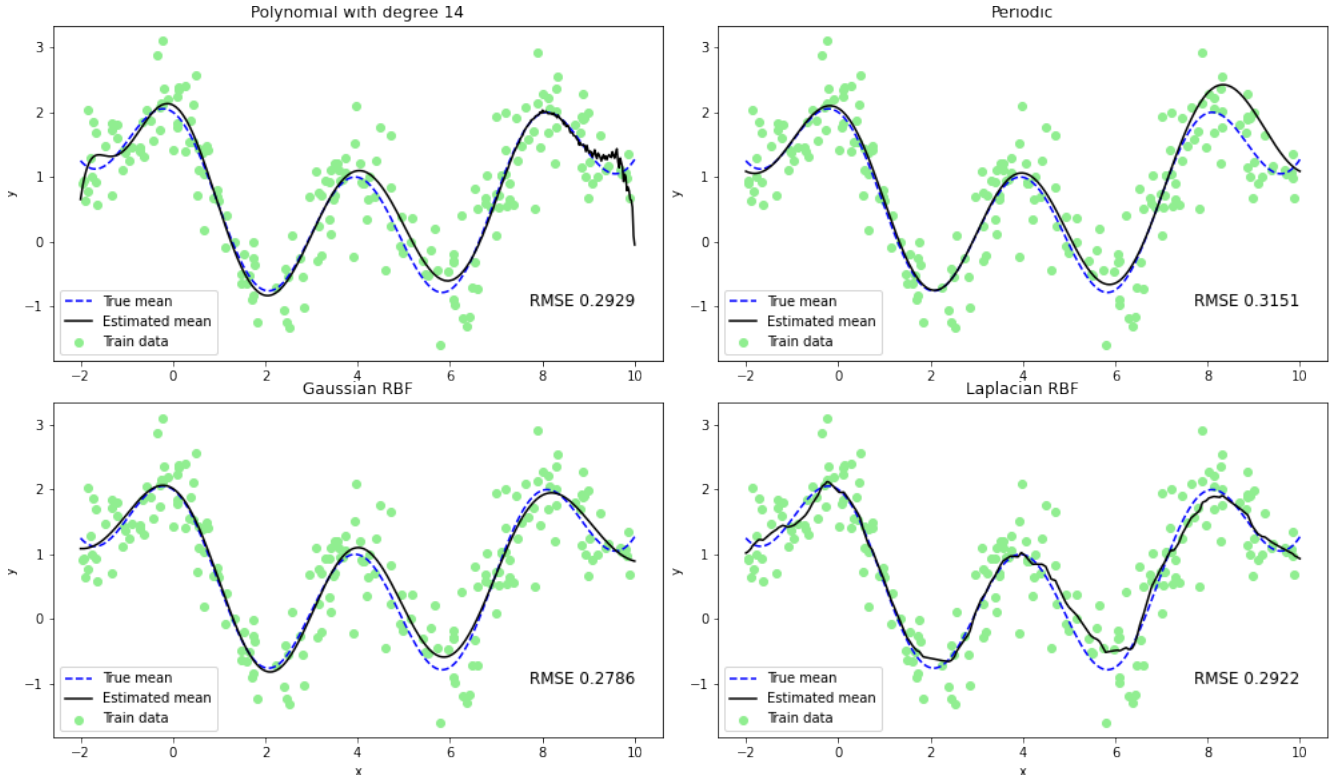

Figure 9: Ridge regression with different kernels

5.1 Kernel ridge regression

The data is simulated from the model $y_i = \mu(x_i) + \varepsilon_i$, where ε_i are i.i.d. Gaussian rv's. The function $\mu : \mathbb{R} \to \mathbb{R}$ is non-linear and deterministic. We used the Scikit-learn implementation KernelRidge with polynomial, Gaussian, Laplacian and Exp-Sine-Squared kernel (aka periodic kernel). We fitted the models on a train sample selecting the hyperparameters using grid-search and K-fold cross-validation.

On figure 9 are shown the graphs of $\widehat{\mu}(x)$ produced by the fitted models compared to $\mu(x)$, as well as the root mean squared errors on a test set generated from the same distribution. Gaussian kernel performed slightly better on the new data and gives a good smooth approximation $\widehat{\mu}$ of the true mean μ. Polynomial kernel is close to the Gaussian in terms of RMSE, but is very volatile towards the right edge of the interval. Laplacian kernel has larger RMSE and the estimated mean is the worst in terms of smoothness.

Periodic kernel has the largest RMSE although it fits very well for $x \leq 7$. We expected the periodic kernel to fit well as the true mean contains a strictly periodic component. It also requires to optimize two hyperparameters which in general takes much more computation than other kernels.

5.2 Classification with SVM

We simulated three datasets of different type and used SVM classification with 3 kernels, i.e. linear, polynomial and Gaussian RBF kernel. We used SVC with Gaussian and polynomial kernel and LinearSVC from Scikit-learn. Again, the hyperparameter $C = 1/\lambda$ (inverse of regularization strength), scale parameter of kernels and polynomial degree were selected via grid-search and K-fold cross-validation. Degree one was not included when we produced grid-search for polynomial SVM to not repeat linear SVM.

On figure 10 are shown the classification boundaries of the fitted models and the accuracy scores on a test sets generated from the same distribution.

The first dataset is the simplest one, it consists of a two-class Gaussian mixture with unit covariance matrices, different means and class probabilities $p_1 = p_2 = 0.5$. The distribution of the data implies that the optimal classification boundary should be the straight line going orthogonally though the middle of the segment joining the means. One can see on the figure, that the classification boundary of both polynomial and Gaussian RBF are almost linear. In fact, the grid-search procedure chose the smallest kernel scale for RBF and inverse regularization coefficients within the grid to make the decision boundary as "linear" as possible. That is to show that in some cases linear kernel can be the best one and is worth trying. Note that linear and polynomial SVM allow to implicitly change the intercept of the separating hyperplain so that one can adjust false-positive and false-negative rates. Other kernels such as RBF may lack this possibility.

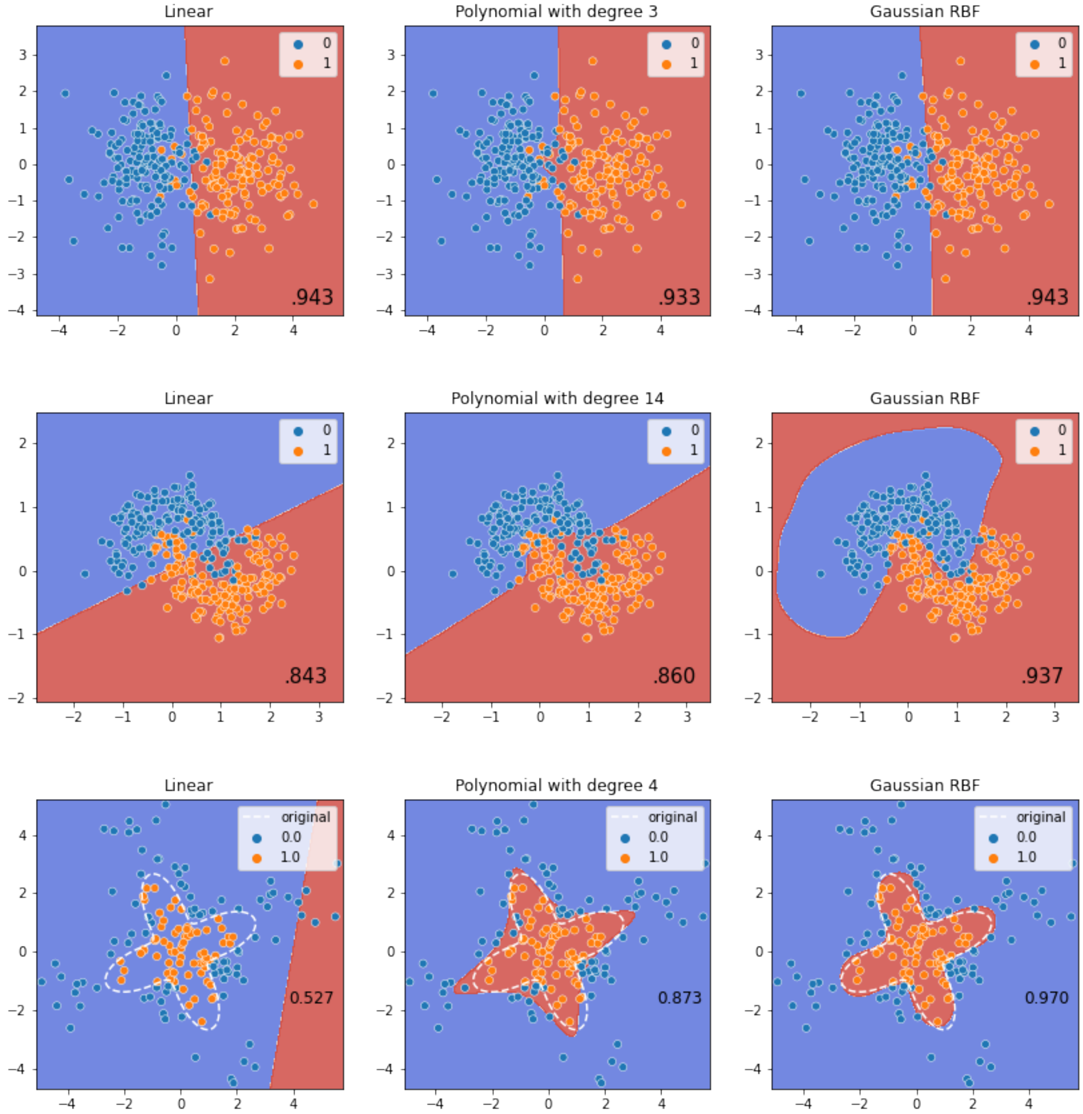

Figure 10: SVM classifiers with different kernels

The second dataset was generated as a noise scattered around two interleaving half-circles. Gaussian RBF shows the best accuracy by far, although linear and polynomial SVM have fairly good scores as well.

The third dataset was simulated specifically to have a more complex boundary and not suitable for linear classification. The white dashed line shows the original shape of the boundary. One can see that Gaussian kernel SVM fits the line almost perfectly and therefore shows the highest accuracy. Polynomial kernel shows a significantly worse result but still finds the pattern to some extent. Linear SVM just labels almost all datapoints as class zero as it dominates in number, but the penalization of slacks did not allow the line to go to far away.

5.3 Kernel independence test

We generated 3 datasets of random vectors $(X_i, Y_i) \in \mathbb{R}^2$ drawn from different distributions. In each of 3 distributions X and Y are dependent (top of figure 11). We implemented the Hilbert-Schmidt independence criterion in Python and compared it with Spearman's ρ and Kendall's τ tests in terms of p-values for association. Each dataset was sampled $N = 100$ times and each time we measured p-values of 3 tests, thus we obtained the empirical distributions of p-values for each test. The empirical CDFs of the p-values on datasets 1, 2 and 3 are shown on the bottom of figure 11 under the scatterplots of the corresponding datasets.

One can see that for the first dataset all 3 tests rejected independence at almost zero significance level for every sample. This is because the association between X and Y is monotonic.
For dataset 2 Spearman's ρ and Kendall's τ behave as if X and Y are independent, i.e. their p-values seem to be uniformly distributed. HSIC with probability ~ 1 rejects the null hypothesis on 18% level and is likely to reject in on lower significance levels as well.
For dataset 3 Spearman's ρ and Kendall's τ perform even worse, while HSIC manages to reject the

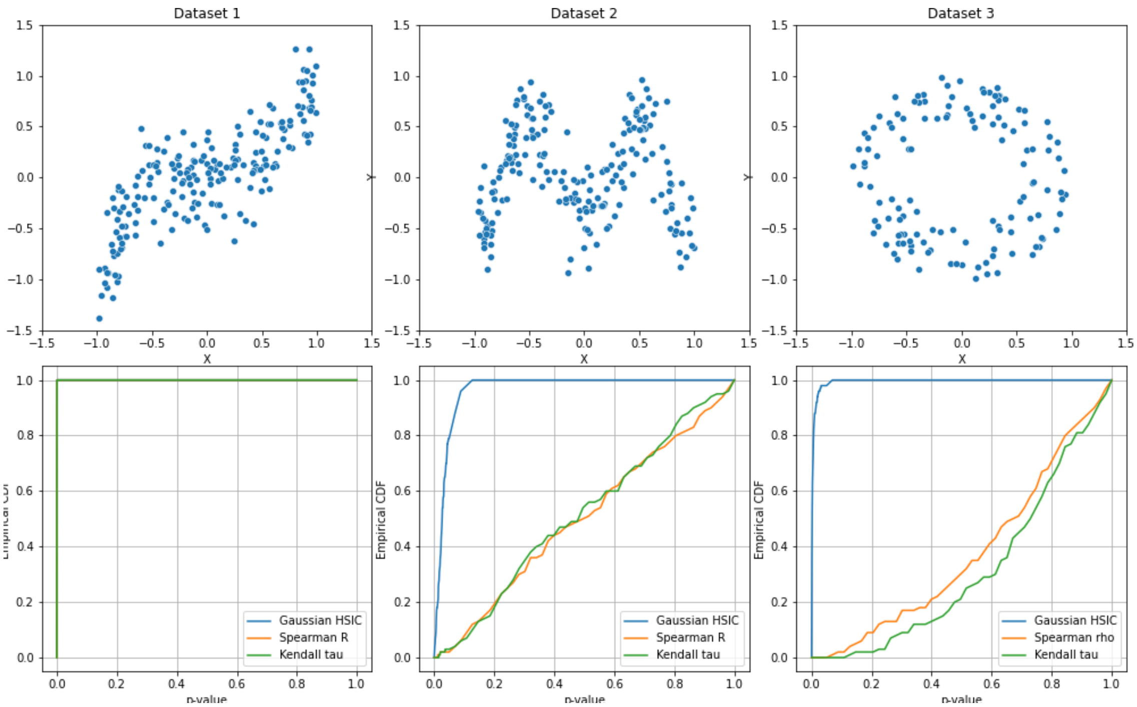

Figure 11: Data (top) and empirical CDFs of p-values' for three tests

null hypothesis on a low significance level with probability close to 1.

6 Real data application

In this section we will outline the results of applications of kernel SVM and kernel PCA for binary classification of electroencephalography (EEG) signals of Bonn University Clinic, relevant paper: Andrzejak et al. 2001. In the original dataset the response variable y takes 5 values, but we collapsed them into 2 merging all non-epileptic cases in one class.

The data consists of 11500 data points $(x_1, y_1), \dots, (x_{11500}, y_{11500})$ with 178 explanatory variables $x_i^1, \dots, x_i^{178} \in \mathbb{R}$ and response variable $y_i \in \{-1, 1\}$. The explanatory variables are values of time series , consisting of recordings of brain activity at 178 timepoints within one second interval. The target variable indicates whether an epileptic seizure activity was detected (class 1) or not (class -1) in this period of time.

On figure 12 are shown two arbitrarily picked signals where on the left an epiliptic seizure was recorded and on the right not recorded. Comparing different pairs of time series of classes 1 and -1, we noticed that in positive cases the signals have much higher magnitude and usually have visually recognizable periodic components, while negative cases tend to have higher fluctuation frequencies.

We have split the dataset into train and test sets in 70/30 ratio and fitted a kernel SVM classifier with Gaussian RBF kernel on the train data. The hyperparameters were tuned via grid-search and cross-validation. The obtained model has shown 97.2% classification accuracy on test data with 96% precision for class 1.

7 Conclusion

We provided a rather detailed presentation of the mathematical theory behind the ridge regression, support vector machines and principal component analysis and outlined the elements of the RKHS theory needed for understanding the kernel trick. Some steps of the reasoning were omitted, as they were considered redundant (like proof of convexity for quadratic or hinge loss). Some statements were given without a proof nor reference, as they are considered as well known facts (like

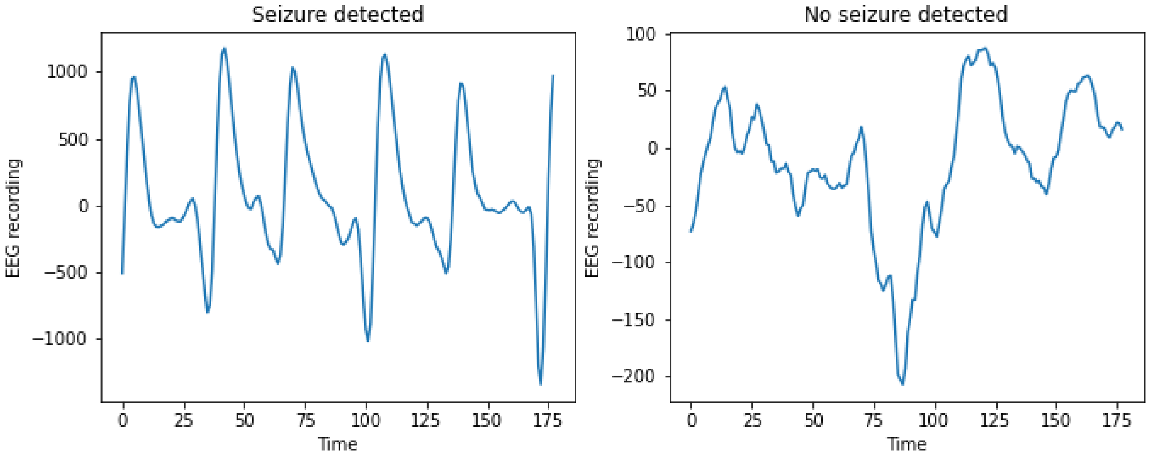

Figure 12: EEG signals in epileptic and non-epileptic cases

BLUE property of OLS solution and Riesz representation theorem).
We also described the basic results that are used to construct the kernel two-sample and independence tests.
As for the results of simulations and real data analysis, in the discussed cases kernel methods with RBF kernel have shown overall good performance and we suggest to use it as a starting point to analyse the data. However one should not always rely on a single kernel as well as neglect the use of linear kernel, since it can be optimal in certain cases, does not require hyperparameter selection and allows for a more straightforward interpretation.

Certainly, there are much more fascinating results on the kernel methods that we wanted to study in the context of this project, such as relation of kernels to Gaussian processes or Bayesian approach, but unfortunately we were limited in time and prerequisite knowledge.

References

Andrzejak, R. G., K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger (2001). "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state". In: *Physical Review E* 64.6, p. 061907.

Arcones, M. A. and E. Gine (1992). "On the bootstrap of U and V statistics". In: *The Annals of Statistics*, pp. 655–674.

Bishop, C. M. and N. M. Nasrabadi (2006). *Pattern recognition and machine learning.* Vol. 4. 4. Springer.

Chwialkowski, K. and A. Gretton (2014). "A kernel independence test for random processes". In: *International Conference on Machine Learning.* PMLR, pp. 1422–1430.

Fukumizu, K., A. Gretton, X. Sun, and B. Schölkopf (2007). "Kernel measures of conditional dependence". In: *Advances in neural information processing systems* 20.

Golub, G. H. and C. F. Van Loan (2013). *Matrix computations.* JHU press.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1, pp. 723–773.

Gretton, A., K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur (2009). "A fast, consistent kernel two-sample test". In: *Advances in neural information processing systems* 22.

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola (2007). "A kernel statistical test of independence". In: *Advances in neural information processing systems* 20.

Schölkopf, B. and A. Smola (2002). "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond". In.

Steinwart, I. and A. Christmann (2008). *Support vector machines.* Springer Science & Business Media.