Kernel Principal Component Analysis

Alex Beeny (abeeny@siue.edu)

Draft: February 14, 2024

Contents

Abstract

Principal component analysis (PCA) and kernel methods are tools often used in data science. The underlying theory of these tools depend on the properties of a special type of Hilbert space called a reproducing kernel Hilbert space (RKHS). This paper explores the essence of RKHSs using data science examples, in particular, PCA and kernel PCA. When kernel methods are applied to PCA, we can analyze nonlinear data in a high-dimensional feature space with some nice properties.

1 Introduction

In linear regression, the equation of a line $y = a_0 + a_1 x$ is used to model observations based on training data. Here, the input variable x is used to predict the response variable y. The parameters a_0 and a_1 are chosen such that

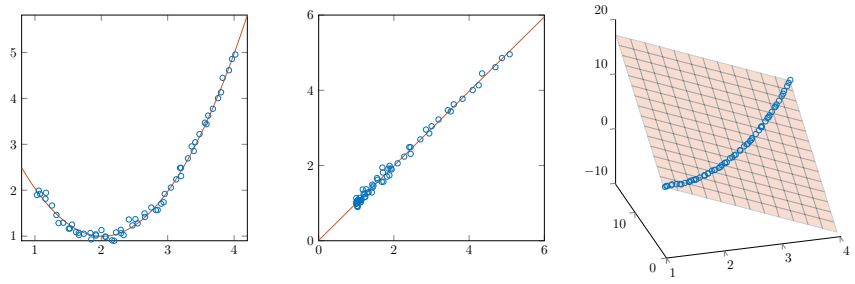

Figure 1: Plotting y against x (left) and the derived feature $z = \phi(x)$ (middle). The 3D plot (right) graphs the output y against x and x^2

the residual error[1] is minimized. It seems natural to model data using polyomial equations in a similar way, that is, determine $a_0, a_1, \ldots, a_n$ such that

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{1}$$

minimizes the residual error.

In multiple linear regression, the equation of a hyperplane

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \tag{2}$$

is used to predict y using inputs $x_1, x_2, \ldots x_n$. It follows that these observations are points in $n + 1$ dimensions.

Example 1.1. Suppose we have a set of points $\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{k}$ in $\mathbb{R}^2$. Using polynomial regression, we fit the model $y \sim 1 + x + x^2$. We can derive a new variable from x to get

$$z = \phi(x) = a_0 + a_1 x + a_2 x^2.$$

By transforming our original data, the quadradic relationship between x and y can be viewed as a linear relationship between z and y. This shows that polynomial regression is just a special case of multiple regression where each power of x is treated as a separate dimension. See Figure 1

Here, we make the distinction between different kinds of input variables. We say that x is an **attribute** and that z is a **feature**.

2 Principal Component Analysis

Principal component analysis (PCA) is a coordinate transform that minimizes projection residuals along each of the major axes $e_1, e_2, \ldots, e_m$ of the transformed space. These axes are referred to as **principal components**. After the

[1]The residual for a given observation $\left(x^{(i)}, y^{(i)}\right)$ is $\left|y^{(i)} - \left(a_0 + a_1 x^{(i)}\right)\right|$.

PCA transform is applied to a set of observations, the smallest projection error is along the first principal component. If we project down to two dimensions, the first two principal components have the smallest projection error. In general, projecting down to $k \leq m$ dimensions is most accurate along $e_1, e_2, \ldots, e_k$.

Conversely, projecting along $e_2, \ldots, e_m$ will result in the largest residual error in $m-1$ dimensions. It follows that e_1 must be the direction with the highest variance. Then e_2 is the direction with the next highest variance, and so on.

Example 2.1. Consider the following matrix

$$A = \begin{bmatrix} 5 & 3 & 6 & 7 & 6 \\ 4 & 5 & 7 & 1 & 3 \\ 5 & 7 & 6 & 1 & 0 \\ 6 & 10 & 12 & 12 & 11 \\ 9 & 10 & 12 & 13 & 9 \end{bmatrix}.$$

The column means are $\mu = [5.8, 7, 8.6, 6.8, 5.8]$. Then the mean-centered data becomes

$$X = A - \mu = \frac{1}{5} \begin{bmatrix} -4 & -20 & -13 & 1 & 1 \\ -9 & -10 & -8 & -29 & -14 \\ -4 & 0 & -13 & -29 & -29 \\ 1 & 15 & 17 & 26 & 26 \\ 16 & 15 & 17 & 31 & 16 \end{bmatrix}.$$

The covariance matrix is

$$C = X^T X = \frac{1}{5} \begin{bmatrix} 74 & 85 & 93 & 179 & 104 \\ 85 & 190 & 170 & 225 & 150 \\ 93 & 170 & 196 & 313 & 238 \\ 179 & 225 & 313 & 664 & 484 \\ 104 & 150 & 238 & 484 & 394 \end{bmatrix}.$$

Diagonalizing C gives

$$V = \begin{bmatrix} 0.1888 & -0.2020 & -0.6366 & 0.5495 & -0.4651 \\ 0.2755 & -0.7886 & 0.1472 & -0.4502 & -0.2791 \\ 0.3606 & -0.3464 & 0.3128 & 0.5836 & 0.5582 \\ 0.6979 & 0.2522 & -0.4422 & -0.3707 & 0.3411 \\ 0.5209 & 0.3922 & 0.5288 & 0.1316 & -0.5271 \end{bmatrix},$$

$$D = \begin{bmatrix} 264.8458 & 0 & 0 & 0 & 0 \\ 0 & 27.9766 & 0 & 0 & 0 \\ 0 & 0 & 9.3198 & 0 & 0 \\ 0 & 0 & 0 & 1.4579 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

If we keep all 5 principal component vectors, then $V_5 = V$ and the projection

of X along V is

$$P = XV = \begin{bmatrix} -1.9469 & 4.3453 & -0.8756 & -0.2039 & 0 \\ -6.9742 & -0.0660 & 1.4352 & 0.7590 & 0 \\ -8.1577 & -2.6752 & -0.8063 & -0.5704 & 0 \\ 8.4282 & -0.2330 & 1.8282 & -0.4996 & 0 \\ 8.6507 & -1.3711 & -1.5815 & 0.5149 & 0 \end{bmatrix}.$$

Here, the last column of P is the zero vector because the last eigenvalue of C is zero[2]. To perfectly reconstruct A, we need $k = 4$ principal components and the row vector μ

$$A = PV^T + \mu = PV_4^T + \mu.$$

If we use $k = 3$ principal components, then the projection of X onto V_3 is

$$P = XV_3 = \begin{bmatrix} -1.9469 & 4.3453 & -0.8756 \\ -6.9742 & -0.0660 & 1.4352 \\ -8.1577 & -2.6752 & -0.8063 \\ 8.4282 & -0.2330 & 1.8282 \\ 8.6507 & -1.3711 & -1.5815 \end{bmatrix}$$

and A is approximately reconstructed by

$$A \approx PV_3^T + \mu = \begin{bmatrix} 5.1 & 2.9 & 6.1 & 6.9 & 6.0 \\ 3.6 & 5.3 & 6.6 & 1.3 & 2.9 \\ 5.3 & 6.7 & 6.3 & 0.8 & 0.1 \\ 6.3 & 9.8 & 12.3 & 11.8 & 11.1 \\ 8.7 & 10.2 & 11.7 & 13.2 & 8.9 \end{bmatrix}.$$

We can compute the reconstruction error using

$$E_k = \|A - (PV_k^T + \mu)\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm. By the SVD, we have $X = USV^T$, where $S = \sqrt{D}$. So, the projection of X onto V_k is

$$P = XV_k = US_k,$$

where S_k is the diagonal matrix of the first k singular values. Then the reconstruction error becomes

$$\begin{aligned} \|A - (PV_k^T + \mu)\|_F &= \|(A - \mu) - PV_k^T\|_F \\ &= \|X - PV_k^T\|_F \\ &= \|USV^T - US_kV^T\|_F \\ &= \|U(S - S_k)V^T\|_F \\ &= \|S - S_k\|_F \\ &= \sigma_k + \sigma_{k+1} + \cdots + \sigma_p. \end{aligned}$$

Hence,

$$E_3 = \sigma_3 + \sigma_4 = \sqrt{1.4579} + 0 = 1.2074.$$

[2]Since we subtracted the column means from a square matrix A, the dimension of the row space was reduced to 4.

2.1 Ordinary least squares and PCA

Let X and Y be random variables corresponding to observation vectors $\mathbf{x}$ and $\mathbf{y}$. Suppose we want to fit a linear model such that $Y \sim 1 + X$.

In two dimensions, the model for ordinary least squares (OLS) regression reduces to

$$y(x) = \frac{\operatorname{cov}(\mathbf{x}, \mathbf{y})}{\operatorname{var}(\mathbf{x})}(x - \overline{\mathbf{x}}) + \overline{\mathbf{y}}, \tag{3}$$

This model minimizes the sum of squared errors when used to predict $\mathbf{y}$ values. If we model X as the target instead, OLS would yield a different formula. In contrast, PCA can be used to create the regression model[3]

$$y(x) = \frac{\lambda_{\max} - \operatorname{var}(\mathbf{x})}{\operatorname{cov}(\mathbf{x}, \mathbf{y})}(x - \overline{\mathbf{x}}) + \overline{\mathbf{y}}, \tag{4}$$

where $\lambda_{\max}$ is the largest eigenvalue of the covariance matrix

$$\begin{bmatrix} \operatorname{cov}(\mathbf{x}, \mathbf{x}) & \operatorname{cov}(\mathbf{x}, \mathbf{y}) \\ \operatorname{cov}(\mathbf{y}, \mathbf{x}) & \operatorname{cov}(\mathbf{y}, \mathbf{y}) \end{bmatrix} = \begin{bmatrix} \operatorname{var}(\mathbf{x}) & \operatorname{cov}(\mathbf{x}, \mathbf{y}) \\ \operatorname{cov}(\mathbf{x}, \mathbf{y}) & \operatorname{var}(\mathbf{y}) \end{bmatrix}.$$

In particular,

$$\lambda_{\max} = \frac{1}{2}\left(\operatorname{var}(\mathbf{x}) + \operatorname{var}(\mathbf{y}) + \sqrt{(\operatorname{var}(\mathbf{x}) - \operatorname{var}(\mathbf{y}))^2 + 4\operatorname{cov}(\mathbf{x}, \mathbf{y})^2}\right).$$

The PCA model minimizes orthogonal projection residuals, that is, the line given by (4) results in the shortest total distance to the set of points $\{(x^{(i)}, y^{(i)})\}$. So, whether X or Y is treated as the target variable, the regression line is the same. See Figure 2.

3 Reproducing Kernel Hilbert Space

Recall that the PCA algorithm involves minimizing terms involving the dot product $x^\top x$. We would like to apply a feature map to our variables and replace this dot product with an inner product in the feature space. However, since the dimension of the feature space will typically be much larger than the dimension of the original space, optimization using this inner product will be expensive. This problem is solved by the so-called *kernel trick* which allows us to circumvent the computational complexity issue caused by high-dimensional spaces. In order to justify the kernel trick, we need to establish some theory of reproducing kernel Hilbert spaces.

First, we recall the definition of an inner product.

Definition 3.1. [12, p. 10] Let X be a (real) vector space. An **inner product** is a function $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ which satisfies the following properties:

[3] I think this model would be considered total least squares (TLS) regression since it uses all variables (input and output). Principal component regression (PCR) creates the regression equation using only the input variables.

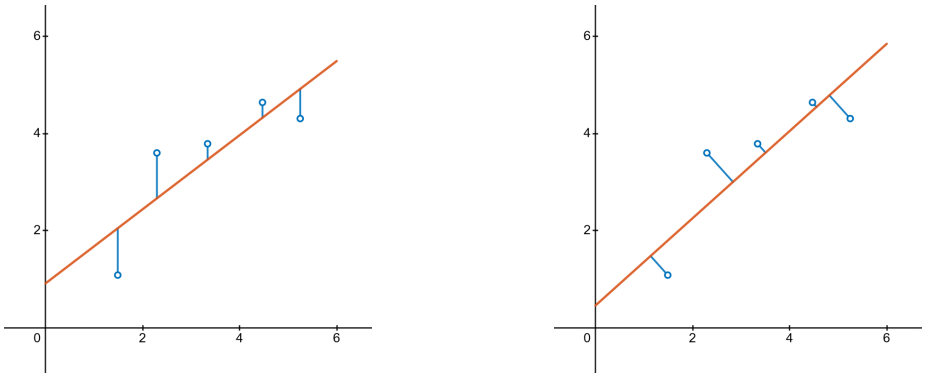

Figure 2: Ordinary least squares (left) minimizes the sum of squared errors while total least squares, i.e., the PCA model (right) minimizes the orthogonal projections.

1. Symmetry. For all $x, y \in X$,
$$\langle x, y \rangle = \langle y, x \rangle.$$
2. Linear in the first argument. For all $x, y, z \in X$, $\alpha, \beta \in \mathbb{R}$,
$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle.$$
3. Positive definite. For all $x \in X$,
$$\langle x, x \rangle \geq 0$$
and $\langle x, x \rangle = 0$ if and only if $x = 0$.

An inner product space is a vector space along with an inner product.

Since real inner products are symmetric and linear in the first argument,

$$\langle z, \alpha x + \beta y \rangle = \langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle = \alpha \langle z, x \rangle + \beta \langle z, y \rangle.$$

So, we also have linearity in the second argument.

The norm induced by an inner product is defined as

$$\|x\| = \langle x, x \rangle^{1/2}$$

and the metric induced by this norm is

$$d(x, y) = \|x - y\| = \langle x - y, x - y \rangle^{1/2}.$$

It follows that an inner product space is also a normed space and a metric space. So, the induced norm will have the following properties for all $x, y \in X$ and $\alpha \in \mathbb{R}$:

1. Triangle inequality. $\|x + y\| \leq \|x\| + \|y\|$;
2. Scalar multiplication. $\|\alpha x\| = |\alpha|\|x\|$;
3. Positivity. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.

Definition 3.2 (Hilbert space)**.** A Hilbert space is a complete inner product space. Let H be a Hilbert space. Then we denote the inner product as $\langle \cdot, \cdot \rangle_H$.

Definition 3.3 (kernel)**.** Let X be a nonempty set and $k : X \times X \to \mathbb{R}$. Then k is a (positive semi-definite) **kernel** if

1. k is symmetric: for all $x, y \in X$,
$$k(x, y) = k(y, x);$$
2. k is positive semi-definite: for all $n \in \mathbb{N}$, if $x_1, \ldots, x_n \in X$ and $c_1, \ldots, c_n \in \mathbb{R}$, then
$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$
Equivalently, the matrix $K \in \mathbb{R}^{n \times n}$ whose entries are $K_{ij} = k(x_i, x_j)$ is positive semi-definite, that is, $w^\top K w \geq 0$ for all $w \in \mathbb{R}^n$.

Definition 3.4 (feature map)**.** [8] Let H be a Hilbert space of functions $f : X \to \mathbb{R}$. A feature map is a function $\Phi : X \to H$.

Our goal is to have kernels written as inner products of feature maps:

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_H.$$

If Φ is a known feature map, then k would surely be symmetric and positive definite since inner products are. If instead we know k is a kernel, then we need to show that there is a unique Hilbert space with the desired inner product.

To do:

- move kernel/Gram matrix to its own definition
- Define linear functionals
- Define dual space (maybe?)
- Riesz representation theorem
- Define evaluation functionals
- Define RKHS (continuous evaluation functionals)
- An RKHS defines a unique reproducing kernel (by Riesz representation theorem).

- Mercer's theorem.
- A kernel defines a feature map.
- A kernel defines a unique RKHS (by Mercer/Moore-Aronszajn).
- We finally get $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_H$.

Theorem 3.5. *[8, 11] Suppose k_1 and k_2 are kernels over $X \times X$. The following functions kernels.*

1. $k(x, y) = a_1 k_1(x, y) + a_2 k_2(x, y)$ *for all* $a_1, a_2 \geq 0$.
2. $k(x, y) = k_1(x, y) k_2(x, y)$.
3. $k(x, y) = k_1(h(x), h(y))$ *for all* $h : X \to X$.
4. $k(x, y) = g(x) g(y)$ *for all* $g : X \to \mathbb{R}$.
5. $k(x, y) = a_0 + a_1 k_1(x, y) + a_2 k_1(x, y)^2 + \cdots + a_n k_1(x, y)^n$ *for all* $n \in \mathbb{N}$ *and* $a_0, \ldots, a_n \geq 0$.
6. $k(x, y) = \exp(k_1(x, y))$.

Proof. Let $x_1, \ldots, x_n \in X$.

1. Let $a_1, a_2 \geq 0$. Define the matrices K_1 and K_2 by

$$[K_1]_{ij} = k_1(x_i, x_j), \qquad [K_2]_{ij} = k_2(x_i, x_j),$$

   for $i, j = 1, \ldots, n$. Let $K = a_1 K_1 + a_2 K_2$. Since K_1 and K_2 are symmetric, so is K. Since K_1 and K_2 are positive semi-definite and $a_1, a_2 \geq 0$,

$$w^\top K w = w^\top (a_1 K_1 + a_2 K_2) w = a_1 (w^\top K_1 w) + a_2 (w^\top K_2 w) \geq 0,$$

   for any $w \in \mathbb{R}^n$. It follows that K is symmetric positive semi-definite. Since K is the Gram matrix for k, k is a kernel.

2.
3.
4.
5.
6.

□

Theorem 3.6 (Gaussian kernel)**.** *The function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by*

$$k(x, y) = \exp\left(\frac{-\|x - y\|_2^2}{\sigma^2} \right),$$

is a kernel.

4 Kernel PCA

PCA works by computing vector projections using the dot product. This is the typical inner product for $\mathbb{R}^n$ and the resulting basis is orthogonal. The idea behind kernel PCA is to replace the dot product with a kernel function.

5 Conclusion

A Linear Algebra

B Riesz Representation Theorem

Theorem B.1 (Riesz Representation Theorem)**.** *[12, p. 22] Let $\phi : H \to \mathbb{R}$ be a continuous linear functional defined on a Hilbert space H. Then there exists a unique element $g \in H$ such that $\phi(g) = \langle f, g \rangle_H$ for all $g \in H$.*

C Mercer's Theorem

D Code

References

[1] Peter Bartlett. CS 281B / Stat 241B Statistical Learning Theory. Lecture notes, Spring 2008.

[2] Peter Bartlett. CS 281B / Stat 241B Statistical Learning Theory. Lecture notes, Spring 2014.

[3] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3), jun 2008.

[4] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[5] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 2013.

[6] The MathWorks Inc. Statistics and machine learning toolbox, 2022.

[7] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2012.

[8] Cynthia Rudin. Intuition for the Algorithms of Machine Learning. Lecture notes, 2023.

[9] Walter Rudin. *Real and Complex Analysis*. Higher Mathematics Series. McGraw-Hill Education, 1987.

[10] Cosma Rohilla Shalizi. Advanced Data Analysis from an Elementary Point of View. Draft textbook, 2021.

[11] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[12] Christopher G. Small and D.L. Mcleish. *Hilbert Space Methods in Probability and Statistical Inference*. John Wiley & Sons, Inc, 1994.