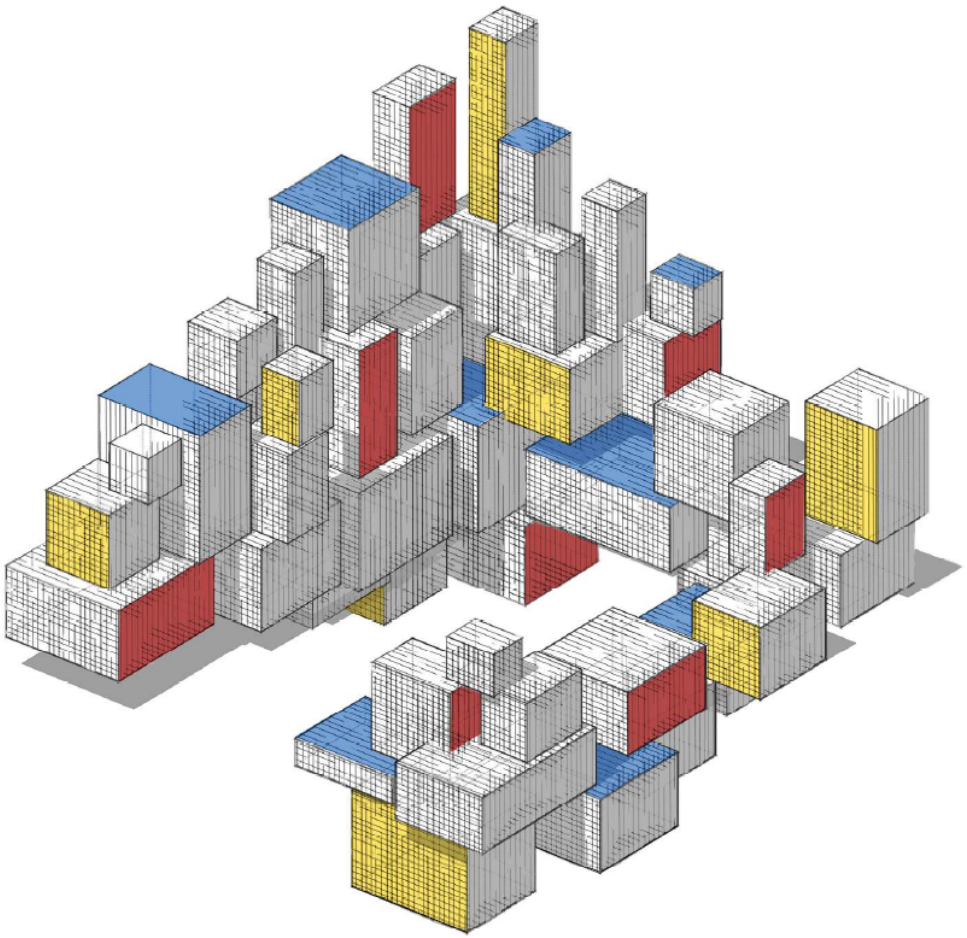
Gaussian Process Regression



Purpose

In this lecture we discuss:

- Gaussian Process
- Gaussian Process Regression
- Kernel PCA

Gaussian Process

A Gaussian process (GP) on a space $\mathcal{X}$ is a stochastic process $\{Z_{\boldsymbol{x}}, \boldsymbol{x} \in \mathcal{X}\}$ where, for any choice of indices $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the vector $[Z_{\boldsymbol{x}_1}, \ldots Z_{\boldsymbol{x}_n}]^\top$ has a multivariate Gaussian distribution.

As such, the distribution of a GP is completely specified by

- its mean function $\mu : \mathcal{X} \to \mathbb{R}$ and
- covariance function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

The covariance function is a finite positive semidefinite function, and hence, in view of the Moore–Aronzajn theorem, can be viewed as a reproducing kernel on $\mathcal{X}$.

Gaussian Process (GP) Regression

The objective of GP regression is to learn a regression function g that predicts a response $y = g(\boldsymbol{x})$ for each feature vector $\boldsymbol{x}$.

This is done in a Bayesian fashion, by establishing:

1. The prior pdf for g as the distribution of a GP with mean 0 and covariance function (that is, kernel) κ, e.g., a Gaussian kernel.
2. The likelihood of the data, for a given g. Specifically, given g, we model the $\{Y_i\}$ as

$$Y_i = g(\boldsymbol{x}_i) + \varepsilon_i\,, \quad i = 1, \ldots, n, \tag{1}$$

   where $\{\varepsilon_i\} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ (for simplicity assume σ^2 is known).

From these two we then derive, via Bayes' formula, the posterior distribution of g given the data.

Vector of Regression Values

Let $\boldsymbol{g} = [g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_n)]^\top$ be the vector of regression values.

Placing a GP prior on the function g is equivalent to placing a multivariate Gaussian prior on the vector $\boldsymbol{g}$:

$$\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{2}$$

where the covariance matrix $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]$ of $\boldsymbol{g}$ is a Gram matrix (implicitly associated with a feature map through the kernel κ).

The likelihood of the data $\boldsymbol{y}$ given $\boldsymbol{g}$, denoted $p(\boldsymbol{y} \mid \boldsymbol{g})$, is obtained directly from the model (1):

$$(\boldsymbol{Y} \mid \boldsymbol{g}) \sim \mathcal{N}(\boldsymbol{g}, \sigma^2 \mathbf{I}_n). \tag{3}$$

Posterior Distribution of $\boldsymbol{g}$

To derive the posterior distribution of $(\boldsymbol{g} \mid \boldsymbol{Y})$, we first note that the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{g}$ is again normal, with mean $\mathbf{0}$ and covariance matrix:

$$\mathbf{K}_{\boldsymbol{y},\boldsymbol{g}} = \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \mathbf{K} \\ \mathbf{K} & \mathbf{K} \end{bmatrix}. \tag{4}$$

The posterior can then be found by conditioning on $\boldsymbol{Y} = \boldsymbol{y}$:

$$(\boldsymbol{g} \mid \boldsymbol{y}) \sim \mathcal{N}\left(\mathbf{K}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \boldsymbol{y},\ \mathbf{K} - \mathbf{K}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{K}\right).$$

This only gives information about g at the observed points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

What about $g(\widetilde{\boldsymbol{x}})$ at a new input $\widetilde{\boldsymbol{x}}$?

Posterior Predictive Distribution

Define $\widetilde{g} := g(\widetilde{\boldsymbol{x}})$ for a new input $\widetilde{\boldsymbol{x}}$. We wish to find the posterior predictive distribution of $\widetilde{g}$, i.e., the distribution of $\widetilde{g}$ given the data $\boldsymbol{y}$.

The joint distribution of $[\boldsymbol{y}^\top, \widetilde{g}]^\top$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix

$$\widetilde{\mathbf{K}} = \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \boldsymbol{\kappa} \\ \boldsymbol{\kappa}^\top & \kappa(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}) \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\kappa} = [\kappa(\widetilde{\boldsymbol{x}}, \boldsymbol{x}_1), \ldots, \kappa(\widetilde{\boldsymbol{x}}, \boldsymbol{x}_n)]^\top$. It follows, from the standard conditioning formulas for the multivariate distribution, that $(\widetilde{g} \,|\, \boldsymbol{y})$ has a normal distribution with mean and variance given respectively by

$$\mu(\widetilde{\boldsymbol{x}}) = \boldsymbol{\kappa}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \boldsymbol{y} \tag{6}$$

and

$$\sigma^2(\widetilde{\boldsymbol{x}}) = \kappa(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}) - \boldsymbol{\kappa}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \boldsymbol{\kappa}. \tag{7}$$

Example: GP Regression

Suppose the regression function is $g(x) = 2\sin(2\pi x), \quad x \in [0, 1]$. We use GP regression to estimate g, using a Gaussian kernel with bandwidth parameter 0.2. The responses were obtained from (1), with noise level $\sigma = 0.5$.

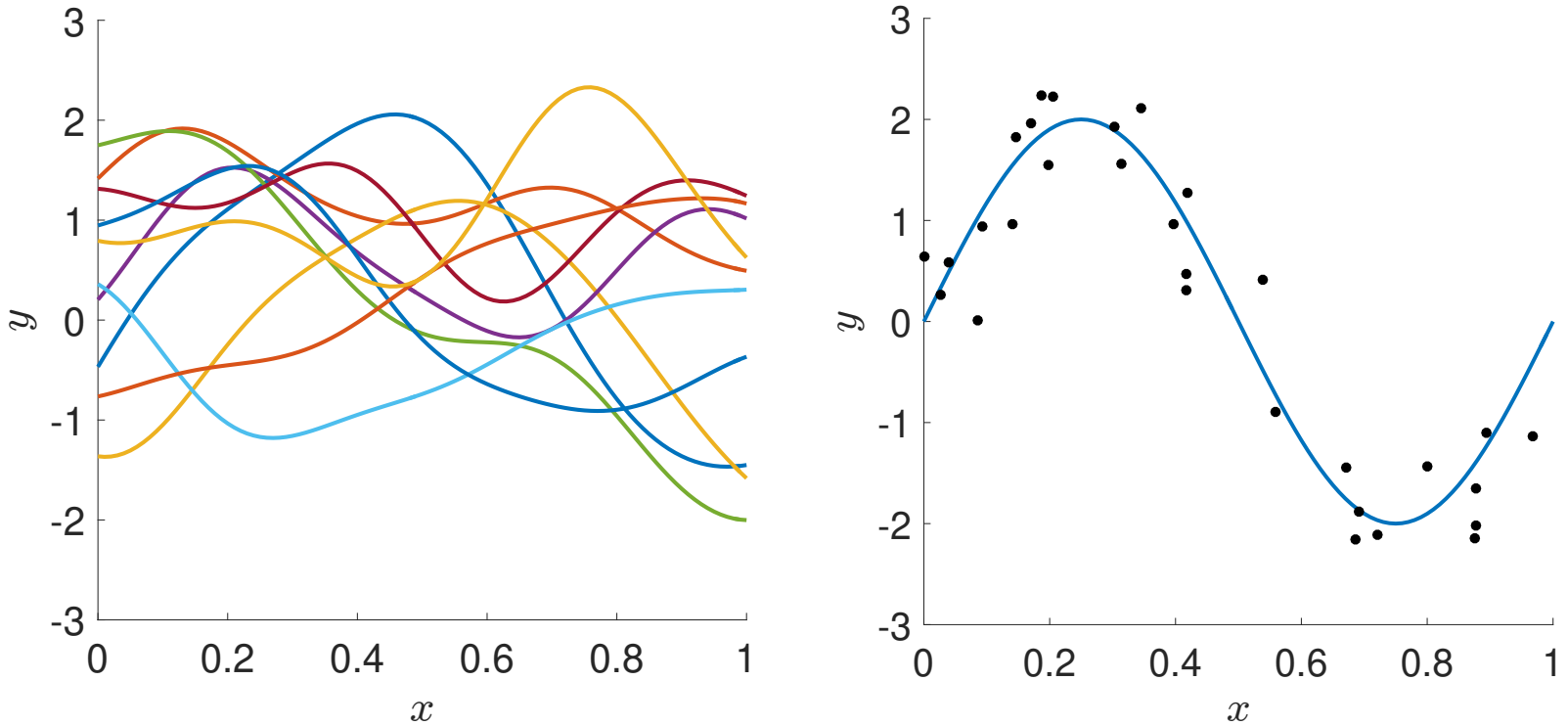


Figure: Left: samples drawn from the GP prior distribution. Right: the true regression function with the data points.

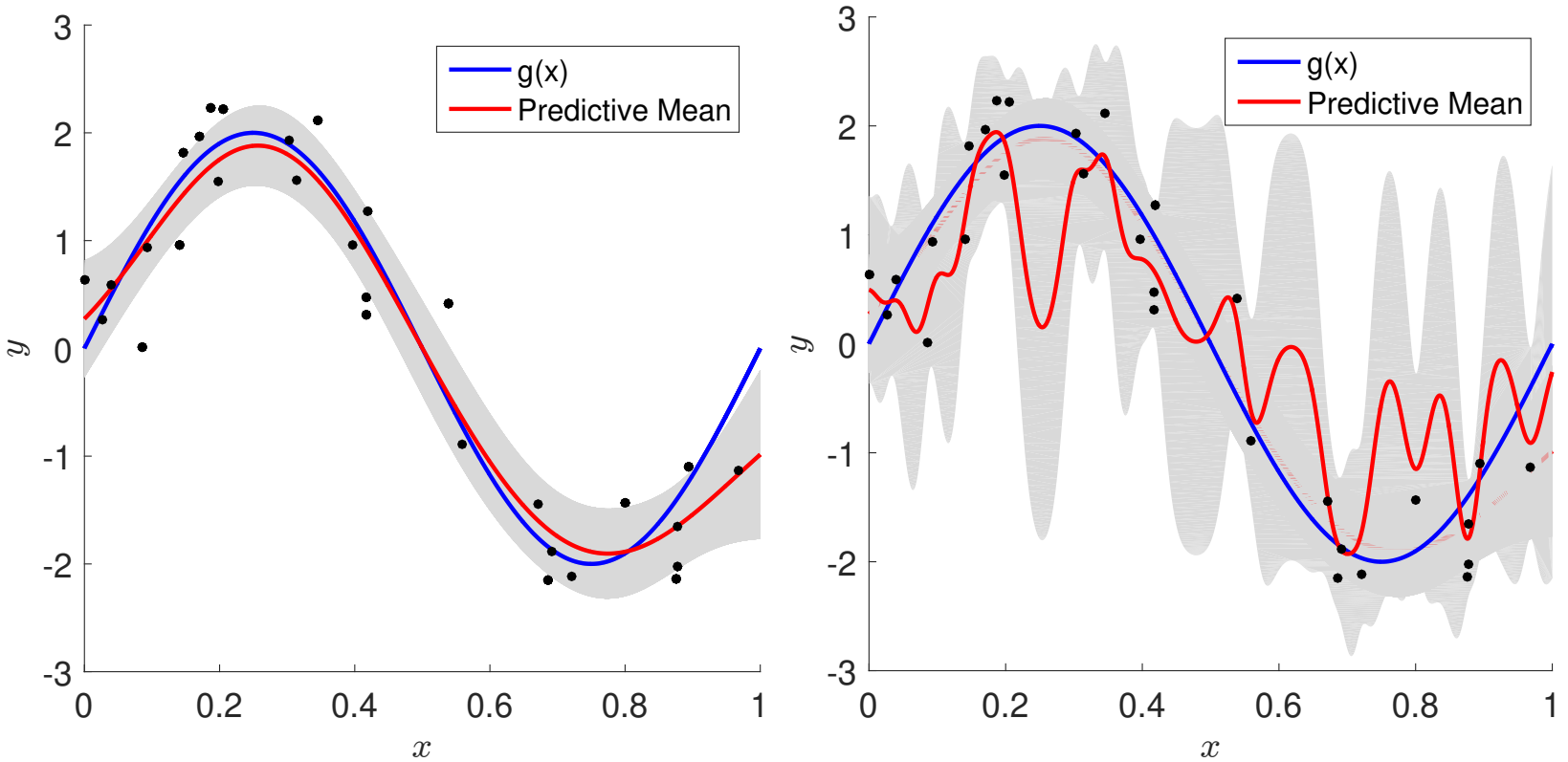


Figure: GP regression of synthetic data set with bandwidth 0.2 (left) and 0.02 (right). The black dots represent the data and the blue curve is the latent function $g(x) = 2\sin(2\pi x)$. The red curve is the mean of the GP predictive distribution given by (6), and the shaded region is the 95% confidence band, corresponding to the predictive variance given in (7).

GP Regression with Hyperparameters

Typically, the variance σ^2 appearing in (1) is not known, and the kernel κ itself depends on several parameters — for instance a Gaussian kernel with an unknown bandwidth parameter.

In the Bayesian framework, one typically specifies a hierarchical model by introducing a prior $p(\boldsymbol{\theta})$ for the vector $\boldsymbol{\theta}$ of such hyperparameters.

Now, the GP prior $(g \mid \boldsymbol{\theta})$ and the likelihood of the data $(\boldsymbol{Y} | \boldsymbol{g}, \boldsymbol{\theta})$ are both dependent on $\boldsymbol{\theta}$.

The posterior distribution of $(\boldsymbol{g} \mid \boldsymbol{y}, \boldsymbol{\theta})$ is as before.

Empirical Bayes

One approach to setting the hyperparameter $\boldsymbol{\theta}$ is to determine its posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and obtain a point estimate, for instance via its maximum a posteriori estimate. However, this can be a computationally demanding exercise.

What is frequently done in practice is to consider instead the *marginal likelihood* $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and maximize this with respect to $\boldsymbol{\theta}$. This procedure is called empirical Bayes.

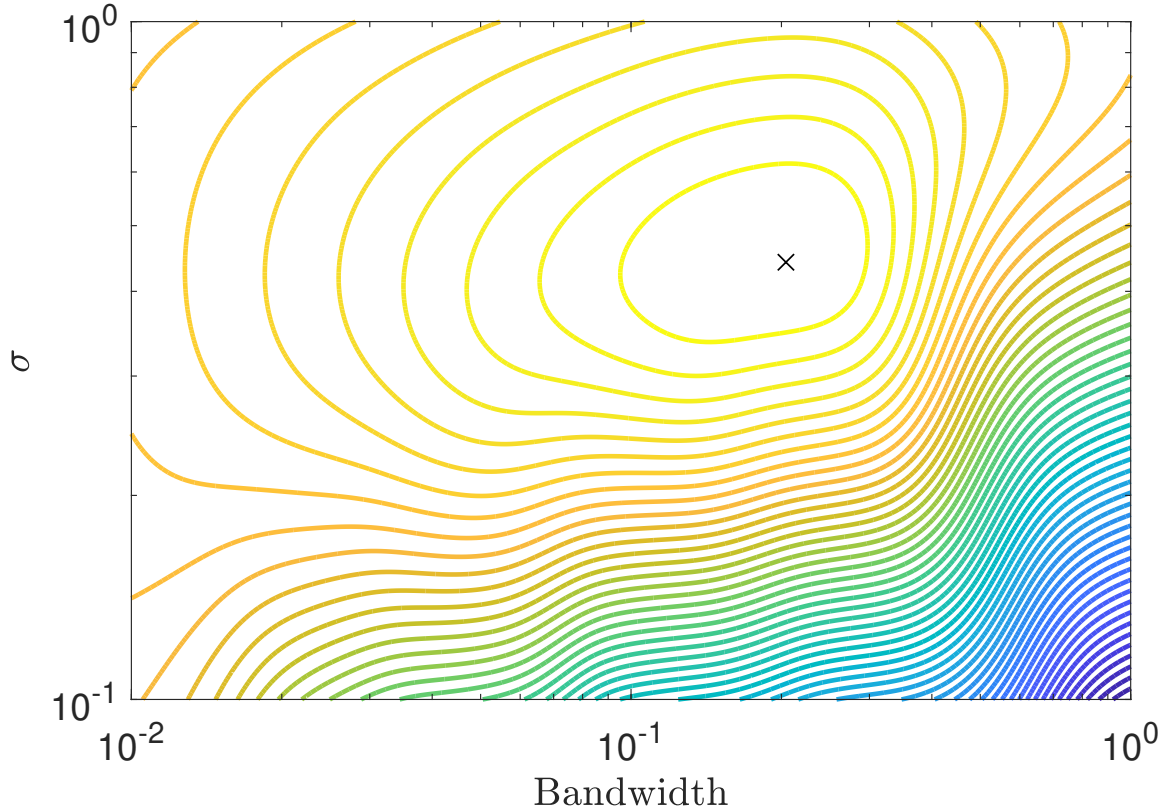
From (4), we have that $(\boldsymbol{Y} \mid \boldsymbol{\theta})$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}_{\boldsymbol{y}} = \mathbf{K} + \sigma^2 \mathbf{I}_n$, immediately giving an expression for the marginal log-likelihood:

$$\ln p(\boldsymbol{y} \mid \boldsymbol{\theta}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\det(\mathbf{K}_{\boldsymbol{y}})| - \frac{1}{2} \boldsymbol{y}^\top \mathbf{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y}. \tag{8}$$

We notice that only the second and third terms in (8) depend on $\boldsymbol{\theta}$.

GP Regression (cont.)

The figure gives the contour plot of the marginal log-likelihood as a function of the noise level σ and bandwidth parameter.



The maximum is attained for a bandwidth parameter around 0.20 and $\sigma \approx 0.44$.

Kernel Principal Component Analysis

Kernel PCA can be thought of as PCA in feature space.

We encountered PCA as a dimensionality reduction technique, based on an SVD of the matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$.

What we shall do is to first re-cast the problem in terms of the Gram matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^\top = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle]$ (note the different order of $\mathbf{X}$ and $\mathbf{X}^\top$), and subsequently replace the inner product $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ with $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ for a general reproducing kernel κ.

To make the link, let us start with an SVD of $\mathbf{X}^\top$:

$$\mathbf{X}^\top = \mathbf{U}\mathbf{D}\mathbf{V}^\top. \tag{9}$$

The dimensions of $\mathbf{X}^\top$, $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ are $d \times n$, $d \times d$, $d \times n$, and $n \times n$, respectively.

Singular Value Decompositions

An SVD of $\mathbf{X}^\top\mathbf{X}$ is

$$\mathbf{X}^\top\mathbf{X} = (\mathbf{U}\mathbf{D}\mathbf{V}^\top)(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top = \mathbf{U}(\mathbf{D}\mathbf{D}^\top)\mathbf{U}^\top$$

and an SVD of $\mathbf{K}$ is

$$\mathbf{K} = (\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{D}\mathbf{V}^\top) = \mathbf{V}(\mathbf{D}^\top\mathbf{D})\mathbf{V}^\top.$$

Let $\lambda_1 \geqslant \cdots \geqslant \lambda_r > 0$ denote the non-zero eigenvalues of $\mathbf{X}^\top\mathbf{X}$ (and $\mathbf{K}$) and let $\mathbf{\Lambda}$ be the corresponding $r \times r$ diagonal matrix.

We can assume that the eigenvector of $\mathbf{X}^\top\mathbf{X}$ belonging to λ_k is the k-th column of $\mathbf{U}$ and that the k-th column of $\mathbf{V}$ is an eigenvector of $\mathbf{K}$.

Let $\mathbf{U}_k$ and $\mathbf{V}_k$ contain the first k columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and let $\mathbf{\Lambda}_k$ be the corresponding $k \times k$ submatrix of $\mathbf{\Lambda}$, $k = 1, \ldots, r$.

Projections

By the SVD (9), we have $\mathbf{X}^\top \mathbf{V}_k = \mathbf{U}\mathbf{D}\mathbf{V}^\top \mathbf{V}_k = \mathbf{U}_k \mathbf{\Lambda}_k^{1/2}$.

The projection of a point $\boldsymbol{x}$ onto the k-dimensional linear space spanned by the columns of $\mathbf{U}_k$ (the first k principal components) is the linear mapping $\boldsymbol{x} \mapsto \mathbf{U}_k^\top \boldsymbol{x}$. Using the fact that $\mathbf{U}_k = \mathbf{X}^\top \mathbf{V}_k \mathbf{\Lambda}^{-1/2}$, we find that $\boldsymbol{x}$ is projected to a point $\boldsymbol{z}$ given by

$$\boldsymbol{z} = \mathbf{\Lambda}_k^{-1/2} \mathbf{V}_k^\top \mathbf{X} \boldsymbol{x} = \mathbf{\Lambda}_k^{-1/2} \mathbf{V}_k^\top \boldsymbol{\kappa}_{\boldsymbol{x}},$$

where we have (suggestively) defined $\boldsymbol{\kappa}_{\boldsymbol{x}} := [\langle \boldsymbol{x}_1, \boldsymbol{x} \rangle, \ldots, \langle \boldsymbol{x}_n, \boldsymbol{x} \rangle]^\top$.

Thus, $\boldsymbol{z}$ is completely determined by the vector of inner products $\boldsymbol{\kappa}_{\boldsymbol{x}}$ and the k principal eigenvalues and (right) eigenvectors of the Gram matrix $\mathbf{K}$. Note that each component z_m of $\boldsymbol{z}$ is of the form

$$z_m = \sum_{i=1}^{n} \alpha_{m,i}\, \kappa(\boldsymbol{x}_i, \boldsymbol{x}), \quad m = 1, \ldots, k. \tag{10}$$

Generalizations

For an uncentered data matrix $\widetilde{\mathbf{X}}$, the centered data can be written as $\mathbf{X} = \widetilde{\mathbf{X}} - \frac{1}{n}\mathbf{E}_n\widetilde{\mathbf{X}}$, where $\mathbf{E}_n$ is the $n \times n$ matrix of ones. Consequently,

$$\mathbf{X}\mathbf{X}^\top = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top - \frac{1}{n}\mathbf{E}_n\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top - \frac{1}{n}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\mathbf{E}_n + \frac{1}{n^2}\mathbf{E}_n\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\mathbf{E}_n,$$

or, more compactly, $\mathbf{X}\mathbf{X}^\top = \mathbf{H}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\mathbf{H}$, where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top, \mathbf{I}_n$.

For the general kernel setting, we replace $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top$ by $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]$ and set $\boldsymbol{\kappa}_{\boldsymbol{x}} = [\kappa(\boldsymbol{x}, \boldsymbol{x}), \ldots, \kappa(\boldsymbol{x}_n, \boldsymbol{x})]^\top$, so that $\boldsymbol{\Lambda}_k$ is the diagonal matrix of the k largest eigenvalues of $\mathbf{HKH}$ and $\mathbf{V}_k$ is the corresponding matrix of eigenvectors.

Note that the "usual" PCA is recovered when we use the linear kernel $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top\boldsymbol{y}$.

Example: Kernel PCA

We simulated 200 points, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{200}$, from the uniform distribution on the set $B_1 \cup (B_4 \cap B_3^c)$, where $B_r := \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant r^2\}$.

We apply kernel PCA with $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$ and compute the functions $z_m(\boldsymbol{x}), m = 1, \ldots, 9$ in (10).

The density plots of z_m are shown in the next figure (data points are superimposed).

We see that the principal components identify the radial structure present in the data.

The final figure shows the projections $\{[z_1(\boldsymbol{x}_i), z_2(\boldsymbol{x}_i)]^\top]\}$ of the original data points onto the first two principal components.

We see that the projected points can be separated by a straight line, whereas this is not possible for the original data.

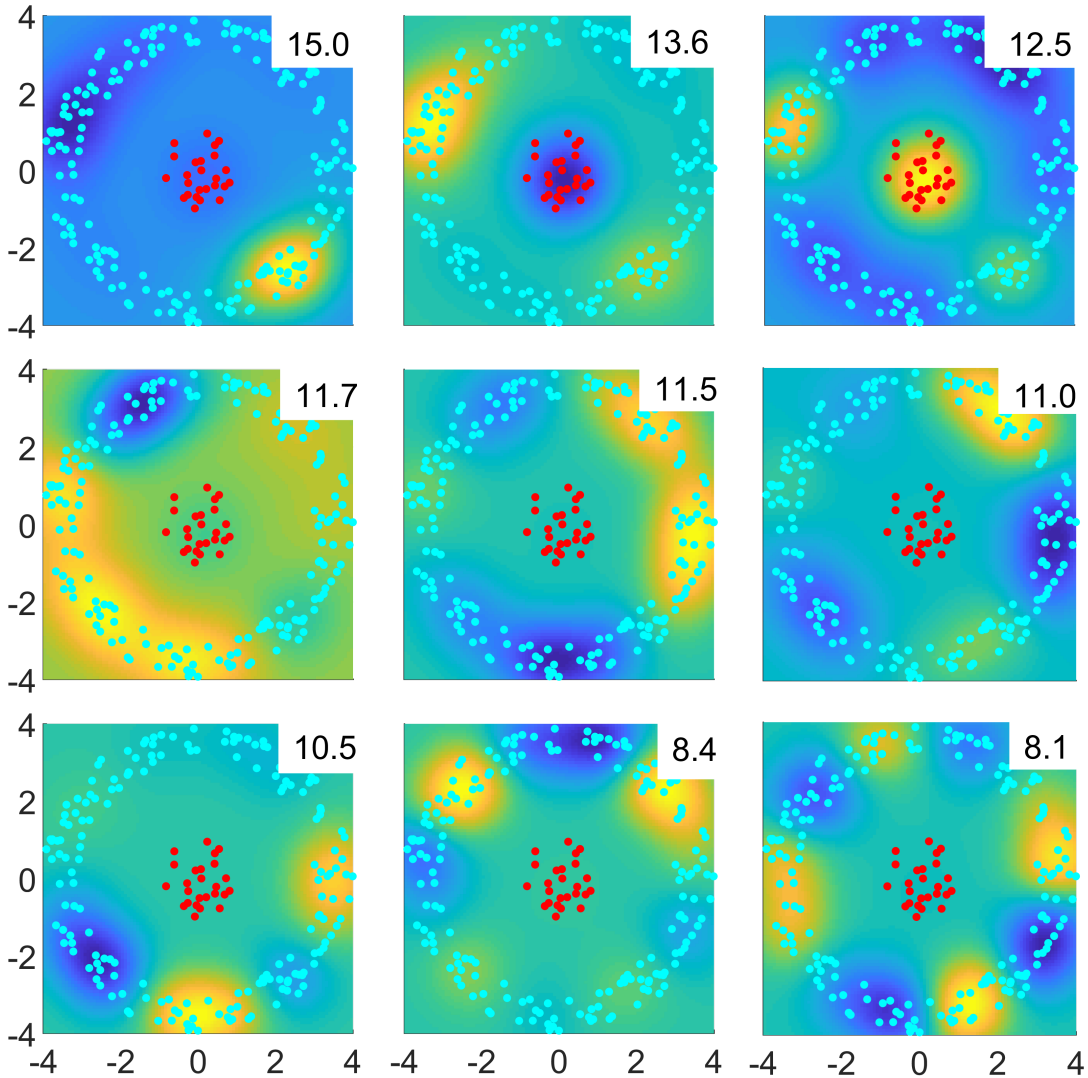


Figure: First nine eigenfunctions using a Gaussian kernel for the two-dimensional data set formed by the red and cyan points.

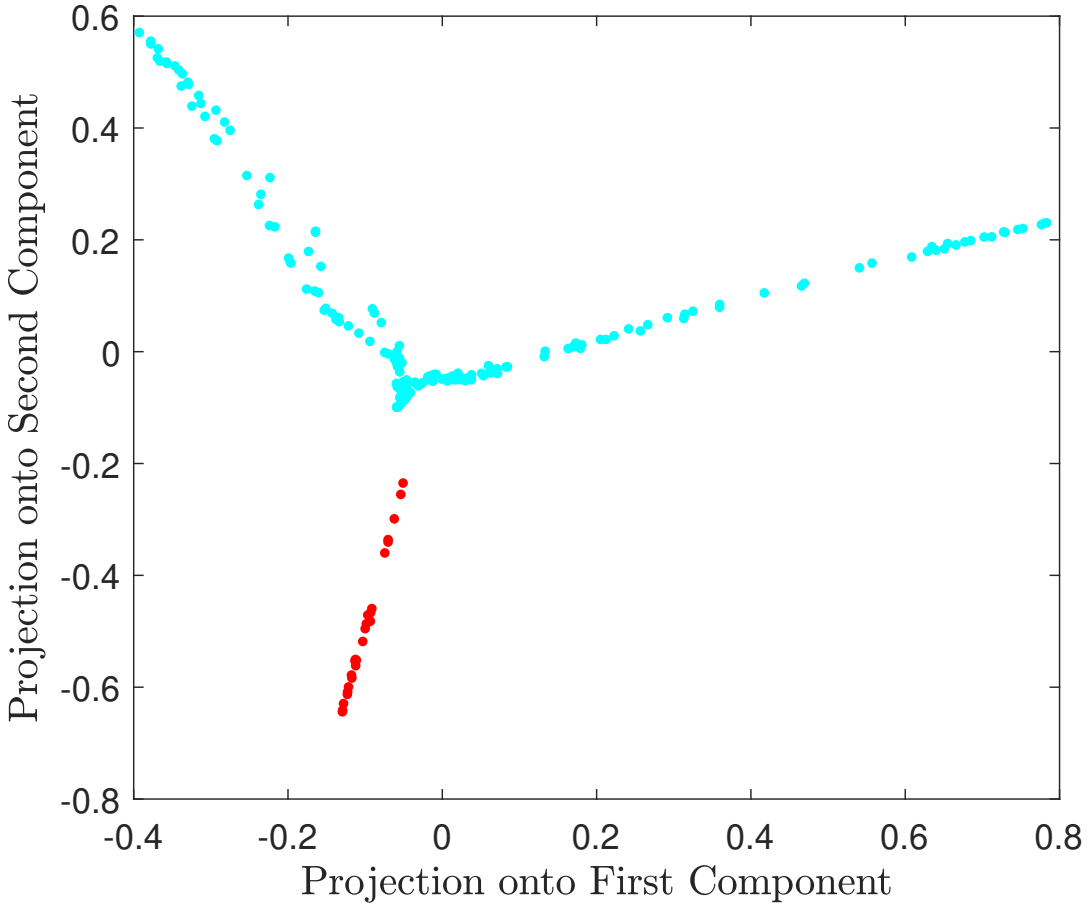


Figure: Projection of the data onto the first two principal components. Observe that already the projections of the inner and outer points are well separated.