In [9]:
```
import numpy, sklearn, sklearn.datasets, utils, time
%matplotlib inline
```

Principal Component Analysis

In this exercise, we will experiment with two different techniques to compute the PCA components of a dataset:

- **Singular Value Decomposition (SVD)**
- **Power Iteration**: A technique that iteratively optimizes the PCA objective.

We consider a random subset of the Labeled Faces in the Wild (LFW) dataset, readily accessible from sklearn, and we apply some basic preprocessing to discount strong variations of luminosity and contrast.

In [10]:
```
D = sklearn.datasets.fetch_lfw_people(resize=0.5)['images']
D = D[numpy.random.mtrand.RandomState(1).permutation(len(D))[:2000]]*1.0
D = D - D.mean(axis=(1,2),keepdims=True)
# Xc = D
D = D / D.std(axis=(1,2),keepdims=True)
print(D.shape)
```

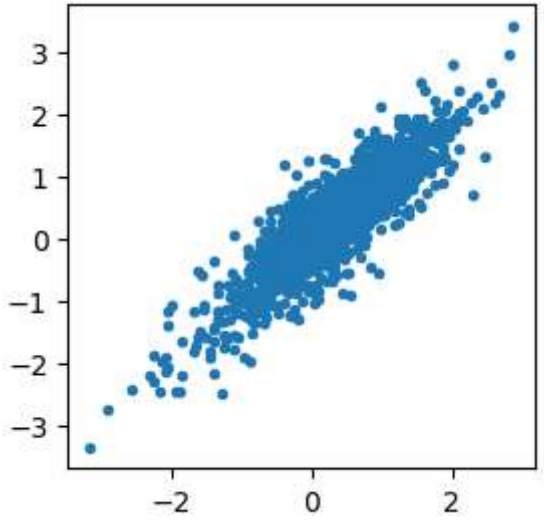
```
(2000, 62, 47)
```

Two functions are provided for your convenience and are available in `utils.py` that is included in the zip archive. The functions are the following:

- **`utils.scatterplot`** produces a scatter plot from a two-dimensional data set.
- **`utils.render`** takes an array of data points or objects of similar shape, and renders them in the IPython notebook.

Some demo code that makes use of these functions is given below.

In [11]:
```
utils.scatterplot(D[:,32,20],D[:,32,21]) # Plot relation between adjacent pixels
utils.render(D[:30],15,2,vmax=5)         # Display first 10 examples in the data
```



PCA with Singular Value Decomposition (15 P)

Principal components can be found computing a singular value decomposition. Specifically, we assume a matrix $\bar{X}$ whose columns contain the data points represented as vectors, and where the data points have been centered (i.e. we have substracted to each of them the mean of the dataset). The matrix $\bar{X}$ is of size $d \times N$ where d is the number of input features and N is the number of data points. This matrix, more specifically, the rescaled matrix $Z = \frac{1}{\sqrt{N}}\bar{X}$ is then decomposed using singular value decomposition:

$$U\Lambda V = Z$$

The k principal components can then be found in the first k columns of the matrix U.

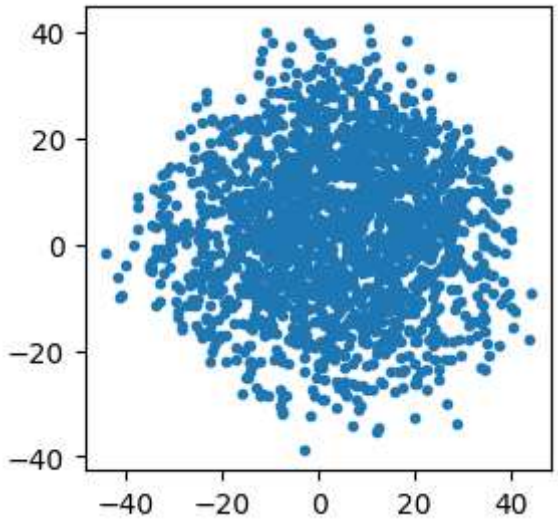
Tasks:

- **Compute the principal components of the data using the function `numpy.linalg.svd`.**
- **Measure the computational time required to find the principal components. Use the function `time.time()` for that purpose. Do *not* include in your estimate the computation overhead caused by loading the data, plotting and rendering.**
- **Plot the projection of the dataset on the first two principal components using the function `utils.scatterplot`.**
- **Visualize the 60 leading principal components using the function `utils.render`.**

Note that if the algorithm runs for more than 3 minutes, there may be some error in your implementation.

In [13]:
```
### REPLACE BY YOUR CODE

k = 60

N = D.shape[0]
D = D.reshape(N, -1)

Z = D.T

Z = Z - numpy.mean(Z, axis = 1)[:, numpy.newaxis]
Z = Z / numpy.std(Z)

Z = (Z / N**.5)

t_start = time.time()
U, S, Vt = numpy.linalg.svd(Z, full_matrices=False)
t_end = time.time()

print('Time: %.3f seconds'%(t_end - t_start))

print(U[:, 0].shape)
utils.scatterplot(U[:, 0]@D.T, U[:, 1]@D.T)
utils.render(U[:, :k].T,15,4)

###
```

```
Time: 1.232 seconds
(2914,)
```



When looking at the scatter plot, we observe that much more variance is expressed in the first two principal components than in individual dimensions as it was plotted before. When looking at the principal components themselves which we render as images, we can see that the first principal components correspond to low-frequency filters that select for coarse features, and the following principal components capture progressively higher-frequency information and are also becoming more noisy.

PCA with Power Iteration (15 P)

The first PCA algorithm based on singular value decomposition is quite expensive to compute. Instead, the power iteration algorithm looks only for the first component and finds it using an iterative procedure. It starts with an initial weight vector $\boldsymbol{w} \in \mathbb{R}^d$, and repeatedly applies the update rule

$$\boldsymbol{w} \leftarrow S\boldsymbol{w} \,/\, ||S\boldsymbol{w}||.$$

where S is the covariance matrix defined as $S = \frac{1}{N}\bar{X}\bar{X}^\top$. Like for standard PCA, the objective that iterative PCA optimizes is $J(\boldsymbol{w}) = \boldsymbol{w}^\top S\boldsymbol{w}$ subject to the unit norm constraint for $\boldsymbol{w}$. We can therefore keep track of the progress of the algorithm after each iteration.

Tasks:

- **Implement the power iteration algorithm. Use as a stopping criterion the value of $J(w)$ between two iterations increasing by less than 0.01.**
- **Print the value of the objective function $J(w)$ at each iteration.**
- **Measure the time taken to find the principal component.**
- **Visualize the the eigenvector w obtained after convergence using the function `utils.render`.**

Note that if the algorithm runs for more than 1 minute, there may be some error in your implementation.

In [29]:
```
### REPLACE BY YOUR CODE

def J_w(w, S_X):
    return w.T @ S_X @ w

# find the largest principal component
def largest_principal_component(X, num_iteration=1000, tolerence=1e-3):
    N = X.shape[1]
    b = numpy.random.rand(X.shape[0])
    b /= numpy.linalg.norm(b)
    S = ((1/N) * X @ X.T)

    cost = J_w(b, S)
    print('iteration %.2d  J(w)= %.3f' %(0, cost))
```

```
    for i in range(num_iteration):
        b_new = S @ b
        b_new /= numpy.linalg.norm(b_new)

        cost = J_w(b_new, S)
        print('iteration %.2d  J(w)= %.3f' %(i+1, cost))

        if numpy.linalg.norm(b - b_new) < tolerence:
            print('stop criterion satisfied')
            break

        b = b_new

    return b


Z = D.T
Z = Z - numpy.mean(Z, axis = 1)[:, numpy.newaxis]
# Z = Z / numpy.std(Z)
print(Z.shape)
t_start = time.time()
w1 = largest_principal_component(Z)
t_end = time.time()

print('Time: %.3f seconds'%(t_end - t_start))
utils.render(w1.T,1,1)
###
```

```
(2914, 2000)
iteration 00  J(w)= 0.221
iteration 01  J(w)= 171.771
iteration 02  J(w)= 206.325
iteration 03  J(w)= 219.384
iteration 04  J(w)= 230.323
iteration 05  J(w)= 240.306
iteration 06  J(w)= 248.660
iteration 07  J(w)= 254.954
iteration 08  J(w)= 259.300
iteration 09  J(w)= 262.119
iteration 10  J(w)= 263.872
iteration 11  J(w)= 264.935
iteration 12  J(w)= 265.569
iteration 13  J(w)= 265.944
iteration 14  J(w)= 266.165
iteration 15  J(w)= 266.294
iteration 16  J(w)= 266.370
iteration 17  J(w)= 266.414
iteration 18  J(w)= 266.440
iteration 19  J(w)= 266.455
iteration 20  J(w)= 266.464
iteration 21  J(w)= 266.469
iteration 22  J(w)= 266.472
iteration 23  J(w)= 266.474
iteration 24  J(w)= 266.475
iteration 25  J(w)= 266.475
iteration 26  J(w)= 266.476
Time: 0.898 seconds
```

We observe that the computation time has decreased significantly. The difference of performance becomes larger as the number of dimensions and data points increases. We can observe that the principal component is the same (sometimes up to a sign flip) as the one obtained by the SVD algorithm.

Exercises for the course
Machine Learning 1
Winter semester 2024/25

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 4

Exercise 1: Lagrange Multipliers (10 + 10 P)

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ be a dataset of N data points. We consider the objective function

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{N} \|\boldsymbol{\theta} - \boldsymbol{x}_k\|^2$$

to be minimized with respect to the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. In absence of constraints, the parameter $\boldsymbol{\theta}$ that minimizes this objective is given by the empirical mean $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$. However, this is generally not the case when the parameter $\boldsymbol{\theta}$ is constrained.

(a) Using the method of Lagrange multipliers, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{\theta}^\top \boldsymbol{b} = 0$, with $\boldsymbol{b}$ some unit vector in $\mathbb{R}^d$. Give a geometrical interpretation to your solution.

(b) Using the same method, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to $\|\boldsymbol{\theta} - \boldsymbol{c}\|^2 = 1$, where $\boldsymbol{c}$ is a vector in $\mathbb{R}^d$ different from $\boldsymbol{m}$. Give a geometrical interpretation to your solution.

Exercise 2: Principal Component Analysis (10 + 10 P)

We consider a dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$. Principal component analysis searches for a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that projecting the data on that vector produces a distribution with maximum variance. Such vector can be found by solving the optimization problem:

$$\arg\max_{\boldsymbol{u}} \; \frac{1}{N}\sum_{k=1}^{N} \Big[\boldsymbol{u}^\top \boldsymbol{x}_k - \frac{1}{N}\Big(\sum_{l=1}^{N} \boldsymbol{u}^\top \boldsymbol{x}_l\Big)\Big]^2 \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

(a) *Show* that the problem above can be rewritten as

$$\arg\max_{\boldsymbol{u}} \; \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

where $\boldsymbol{S} = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^\top$ is the scatter matrix, and $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$ is the empirical mean.

(b) *Show* using the method of Lagrange multipliers that the problem above can be reformulated as solving the eigenvalue problem

$$\boldsymbol{S}\boldsymbol{u} = \lambda \boldsymbol{u}$$

and retaining the eigenvector $\boldsymbol{u}$ associated to the highest eigenvalue λ.

Exercise 3: Bounds on Eigenvalues (5 + 5 + 5 + 5 P)

Let λ_1 denote the largest eigenvalue of the matrix $\boldsymbol{S}$. The eigenvalue λ_1 quantifies the variance of the data when projected on the first principal component. Because its computation can be expensive, we study how the latter can be bounded with the diagonal elements of the matrix $\boldsymbol{S}$.

(a) *Show* that $\sum_{i=1}^{d} \boldsymbol{S}_{ii}$ is an upper bound to the eigenvalue λ_1.

(b) *State* the conditions on the data for which the upper bound is tight.

(c) *Show* that $\max_{i=1}^{d} \boldsymbol{S}_{ii}$ is a lower bound to the eigenvalue λ_1.

(d) *State* the conditions on the data for which the lower bound is tight.

Exercise 4: Iterative PCA (10 P)

When performing principal component analysis, computing the full eigendecomposition of the scatter matrix $\boldsymbol{S}$ is typically slow, and we are often only interested in the first principal components. An efficient procedure to find the first principal component is *power iteration*. It starts with a random unit vector $\boldsymbol{w}^{(0)} \in \mathbb{R}^d$, and iteratively applies the parameter update

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{S}\boldsymbol{w}^{(t)} \big/ \|\boldsymbol{S}\boldsymbol{w}^{(t)}\|$$

until some convergence criterion is met. Here, we would like to show the exponential convergence of power iteration. For this, we look at the error terms

$$\mathcal{E}_k(\boldsymbol{w}) = \left|\frac{\boldsymbol{w}^\top \boldsymbol{u}_k}{\boldsymbol{w}^\top \boldsymbol{u}_1}\right| \qquad \text{with} \quad k = 2, \ldots, d,$$

and observe that they should all converge to zero as $\boldsymbol{w}$ approaches the eigenvector $\boldsymbol{u}_1$ and becomes orthogonal to other eigenvectors.

(a) Show that $\mathcal{E}_k(\boldsymbol{w}^{(T)}) = |\lambda_k/\lambda_1|^T \cdot \mathcal{E}_k(\boldsymbol{w}^{(0)})$, i.e. the convergence of the algorithm is exponential with the number of time steps T.

Exercise 5: Programming (30 P)

Download the programming files on ISIS and follow the instructions.

Exercise 1: Lagrange Multipliers (10 + 10 P)

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ be a dataset of N data points. We consider the objective function

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{N} \|\boldsymbol{\theta} - \boldsymbol{x}_k\|^2$$

to be minimized with respect to the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. In absence of constraints, the parameter $\boldsymbol{\theta}$ that minimizes this objective is given by the empirical mean $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$. However, this is generally not the case when the parameter $\boldsymbol{\theta}$ is constrained.

(a) Using the method of Lagrange multipliers, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{\theta}^\top \boldsymbol{b} = 0$, with $\boldsymbol{b}$ some unit vector in $\mathbb{R}^d$. Give a geometrical interpretation to your solution.

Solution:

First the objective function can be writen as

$$J(\theta) = \sum_{k=1}^{N} \|\theta - x_k\|^2$$

not related to x_k

$$= \sum_{k=1}^{N} \theta^T\theta - 2\theta^T x_k + x_k^T x_k$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad \sum_{k=1}^{N} \theta^T\theta - 2\theta^T x_k$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad N\theta^T\theta - 2\theta^T \sum_{k=1}^{N} x_k$$

$$\Downarrow$$

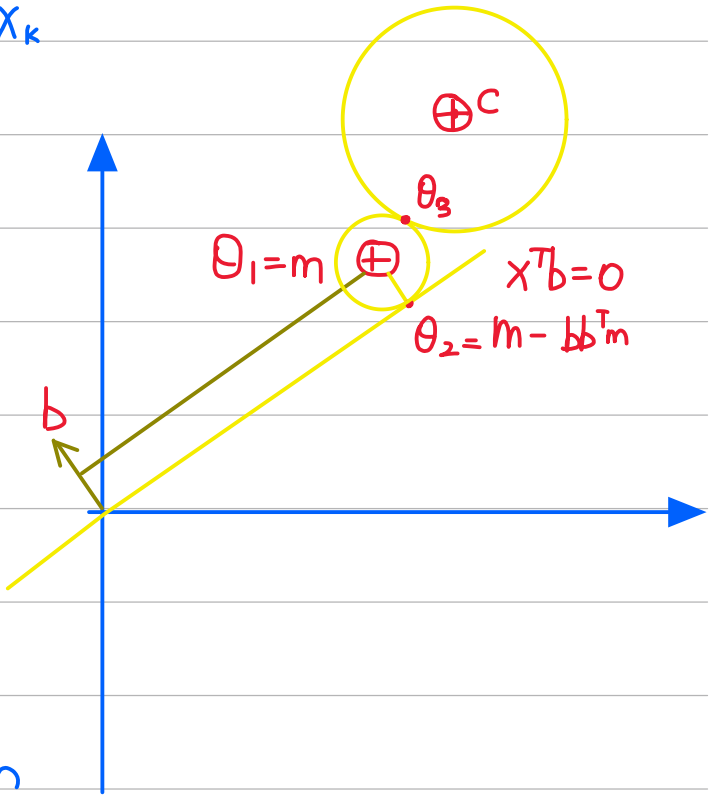
$$\underset{\theta}{\operatorname{argmin}} \quad \theta^T\theta - 2\theta^T m$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad \theta^T\theta - 2\theta^T m + m^T m$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} \quad \|\theta - m\|^2 \qquad \theta_1^* = m$$



So when this objective is constraint by $\theta^T b = 0$ using Lagrange multiplier we can form a Lagrange function:

$$\max_{\lambda} \min_{\theta} \mathcal{L}(\theta, \lambda) = \frac{1}{2}\|\theta - m\|^2 + \lambda \theta^T b$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \theta^T} = (\theta - m) + \lambda b = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = \theta^T b = 0 \end{cases}$$

$$\theta - m + \lambda b = 0 \qquad (1)$$

$$\therefore \quad \theta^* = m - \lambda b$$

Then we need to find the expression of λ.

let (1) be left multiplied by b^T.

$$b^T\theta - b^T m + \lambda b^T b = 0$$

Since $b^T\theta = 0$ and b is a unit vector, i.e. $b^T b = 1$

$$\therefore \quad \lambda = b^T m$$

$$\therefore \quad \theta_2^* = m - b^T m b$$

$$= m - b b^T m$$

(b) Using the same method, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to $\|\boldsymbol{\theta} - \boldsymbol{c}\|^2 = 1$, where $\boldsymbol{c}$ is a vector in $\mathbb{R}^d$ different from $\boldsymbol{m}$. Give a geometrical interpretation to your solution.

Solution:

The Lagrange function is

$$\mathcal{L}(\theta, \lambda) = \frac{1}{2}\|\theta - m\|^2 + \frac{1}{2}\lambda\left(\|\theta - c\|^2 - 1\right)$$

$$\frac{\partial L}{\partial \theta^T} = (\theta - m) + \lambda(\theta - c) = 0$$

$$\therefore \quad (\theta - c) + \lambda(\theta - c) = m - c$$

$$(1+\lambda)(\theta - c) = m - c \longrightarrow \theta = c + \frac{m-c}{1+\lambda}$$

$$\therefore \quad \Downarrow \text{ take 2nd Norm}$$

$$(1+\lambda)^2 \underbrace{\|\theta - c\|^2}_{=1} = \|m - c\|^2$$

$$\Downarrow$$

$$(1+\lambda)^2 = \|m - c\|^2$$

$$\therefore \quad \lambda = -1 \pm \|m - c\|$$

$$\therefore \quad \theta^* = c \pm \frac{m-c}{\|m-c\|}$$

We can compare the objective value:

$$J\left(c + \frac{m-c}{\|m-c\|}\right) = \left\| c + \frac{m-c}{\|m-c\|} - m \right\|^2 \longrightarrow \left\| -(m-c) + \frac{m-c}{\|m-c\|} \right\|^2$$

$$J\left(c - \frac{m-c}{\|m-c\|}\right) = \left\| c - \frac{m-c}{\|m-c\|} - m \right\|^2 \longrightarrow \left\| -(m-c) - \frac{m-c}{\|m-c\|} \right\|^2$$

$$\left\| (m-c) \times \left(\frac{1}{\|m-c\|} - 1 \right) \right\|^2 \leq \left\| (m-c) \times \left(-\frac{1}{\|m-c\|} - 1 \right) \right\|^2$$

$$\Downarrow$$

$$\theta_3^* = c + \frac{m-c}{\|m-c\|}$$

Exercise 2: Principal Component Analysis (10 + 10 P)

We consider a dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$. Principal component analysis searches for a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that projecting the data on that vector produces a distribution with maximum variance. Such vector can be found by solving the optimization problem:

$$\arg\max_{\boldsymbol{u}} \frac{1}{N} \sum_{k=1}^{N} \Big[\boldsymbol{u}^\top \boldsymbol{x}_k - \frac{1}{N} \Big(\sum_{l=1}^{N} \boldsymbol{u}^\top \boldsymbol{x}_l \Big) \Big]^2 \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

(a) *Show* that the problem above can be rewritten as

$$\arg\max_{\boldsymbol{u}} \boldsymbol{u}^\top \boldsymbol{S} \boldsymbol{u} \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

where $\boldsymbol{S} = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^\top$ is the scatter matrix, and $\boldsymbol{m} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{x}_k$ is the empirical mean.

Solution:

We can derive the function as.

$$\frac{1}{N} \sum_{k=1}^{N} \Big[u^T x_k - \frac{1}{N} \Big(\sum_{l=1}^{N} u^T x_l \Big) \Big]^2$$

$$= \frac{1}{N} \sum_{k=1}^{N} \Big[u^T x_k - u^T m \Big]^2$$

$$= \frac{1}{N} \sum_{k=1}^{N} (u^T x_k - u^T m)^T (u^T x_k - u^T m)$$

$$= \frac{1}{N} \sum_{k=1}^{N} \big(u^T (x_k - m) \big)^T \big(u^T (x_k - m) \big)$$

$$= \frac{1}{N} \sum_{k=1}^{N} (x_k - m)^T u u^T (x_k - m)$$

$$= \frac{1}{N} \sum_{k=1}^{N} u^T (x_k - m)(x_k - m)^T u$$

$$\Downarrow \text{ delete } \frac{1}{N}$$

$$u^T \Big(\sum_{k=1}^{N} (x_k - m)(x_k - m)^T \Big) u$$

$$\Downarrow$$

$$\boxed{\arg\max_{u} u^T S u \quad \text{with} \quad \|u\|^2 = 1}$$

(b) *Show* using the method of Lagrange multipliers that the problem above can be reformulated as solving the eigenvalue problem

$$\boldsymbol{S}\boldsymbol{u} = \lambda \boldsymbol{u}$$

and retaining the eigenvector $\boldsymbol{u}$ associated to the highest eigenvalue λ.

Solution:

We can form the Lagrange function as:

$$\mathcal{L}(\lambda, u) = u^T S u - \lambda(\|u\|^2 - 1)$$

$$\frac{\partial \mathcal{L}}{\partial u^T} = Su - \lambda u = 0$$

$$\therefore \quad Su = \lambda u \qquad (1)$$

So we have proofed that it can be solved by solving an Eigenvalue problem.

Then we left multiply u^T on both sides.

$$u^T S u = \lambda u^T u = \lambda \|u\|^2 = \lambda$$

$$\therefore \quad \underset{u}{\operatorname{argmax}}\ u^T S u \iff \underset{\lambda}{\operatorname{argmax}}\ \lambda$$

$\therefore$ the optimal results is the eigenvector u^* that correspond to the largest eigenvalue λ_1.

Exercise 3: Bounds on Eigenvalues (5 + 5 + 5 + 5 P)

Let λ_1 denote the largest eigenvalue of the matrix $\boldsymbol{S}$. The eigenvalue λ_1 quantifies the variance of the data when projected on the first principal component. Because its computation can be expensive, we study how the latter can be bounded with the diagonal elements of the matrix $\boldsymbol{S}$.

(a) *Show* that $\sum_{i=1}^{d} \boldsymbol{S}_{ii}$ is an upper bound to the eigenvalue λ_1.

Solution:

$$\lambda_i = u_i^T S u_i = \sum_{k=1}^{N} (u^T(x_k - m))^2 \geq 0$$

$$\sum_i S_{ii} = \text{trace}(S)$$

From Matrix cookbook, Page 6. formular (12) we have:

$$\sum_i S_{ii} = \text{trace}(S) = \sum_i \lambda_i \geq \lambda_1$$

∴ the largest eigenvalue is upperbounded by $\sum_i S_{ii}$

(b) *State* the conditions on the data for which the upper bound is tight.

Solution:

When all other eigenvalues $\lambda_2, \lambda_3 \cdots \lambda_d$ are 0.

Then

$$\sum_i S_{ii} = \lambda_i$$

(c) *Show* that $\max_{i=1}^{d} \boldsymbol{S}_{ii}$ is a lower bound to the eigenvalue λ_1.

Solution:

$$\lambda = \max_{\|u\|=1} u^T S u \geq \max_{u \in \{e_1 \cdots e_d\}} u^T S u = \max_{i=1 \cdots d} e_i^T S e_i$$

(every element ≥ 0)

$$= \{e_1^T S e_1, e_2^T S e_2 \cdots e_d^T S e_d\}$$

[$e_1 \cdots e_d$ are canonical coordinate vectors.]

∴ $\lambda \geq \max_{i=1}^{d} S_{ii}$

(d) *State* the conditions on the data for which the lower bound is tight.

Solution:

If we want $\lambda = \max_{i=1}^{d} S_{ii} = \max_{i=1}^{d} e_i^T S e_i$

then $u \in \{e_1 \cdots e_d\}$ ✓

So that it can pick out the largest diagonal element

Exercise 4: Iterative PCA (10 P)

When performing principal component analysis, computing the full eigendecomposition of the scatter matrix $\boldsymbol{S}$ is typically slow, and we are often only interested in the first principal components. An efficient procedure to find the first principal component is *power iteration*. It starts with a random unit vector $\boldsymbol{w}^{(0)} \in \mathbb{R}^d$, and iteratively applies the parameter update

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{S}\boldsymbol{w}^{(t)} / \|\boldsymbol{S}\boldsymbol{w}^{(t)}\|$$

until some convergence criterion is met. Here, we would like to show the exponential convergence of power iteration. For this, we look at the error terms

$$\mathcal{E}_k(\boldsymbol{w}) = \left| \frac{\boldsymbol{w}^\top \boldsymbol{u}_k}{\boldsymbol{w}^\top \boldsymbol{u}_1} \right| \qquad \text{with} \quad k = 2, \ldots, d,$$

and observe that they should all converge to zero as $\boldsymbol{w}$ approaches the eigenvector $\boldsymbol{u}_1$ and becomes orthogonal to other eigenvectors.

(a) Show that $\mathcal{E}_k(\boldsymbol{w}^{(T)}) = |\lambda_k/\lambda_1|^T \cdot \mathcal{E}_k(\boldsymbol{w}^{(0)})$, i.e. the convergence of the algorithm is exponential with the number of time steps T.

Solution:

$$\mathcal{E}_k(w^{(T+1)}) = \left| \frac{w^{(T+1)^T} u_k}{w^{(T+1)^T} u_1} \right|$$

$$= \left| \frac{\frac{(Sw^{(T)})^T}{\|Sw^{(T)}\|} u_k}{\frac{(Sw^{(T)})^T}{\|Sw^{(T)}\|} u_1} \right| = \left| \frac{(Sw^{(T)})^T u_k}{(Sw^{(T)})^T u_1} \right|$$

Since $S = S^T$

$$= \left| \frac{w^{(T)^T} S u_k}{w^{(T)^T} S u_1} \right| = \left| \frac{w^{(T)^T} \lambda_k u_k}{w^{(T)^T} \lambda_1 u_1} \right|$$

$$= \left| \frac{w^{(T)^T} u_k}{w^{(T)^T} u_1} \right| \cdot \left| \frac{\lambda_k}{\lambda_1} \right|$$

$$= \mathcal{E}_k(w^{(T)}) \left| \frac{\lambda_k}{\lambda_1} \right| = \mathcal{E}_k(w^{(T-1)}) \left| \frac{\lambda_k}{\lambda_1} \right|^2$$

$$= \mathcal{E}_k(w^{(0)}) \left| \frac{\lambda_k}{\lambda_1} \right|^T$$