PCA: batch preprocessing and online-PCA

Exercise T2.1: Batch vs. online-PCA (tutorial)

(a) Identify disadvantages of batch PCA that can be solved using an online PCA method.

(b) How do you find the first PC using Hebbian Learning?

(c) Why do we need Oja's rule?

(d) How do we find all other PCs?

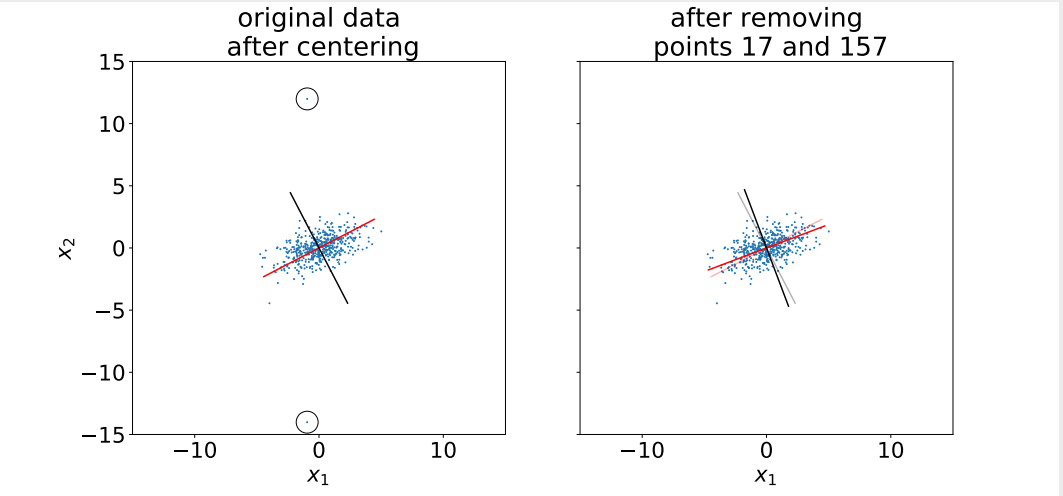
(e) Identify different ways for constructing a novelty filter?

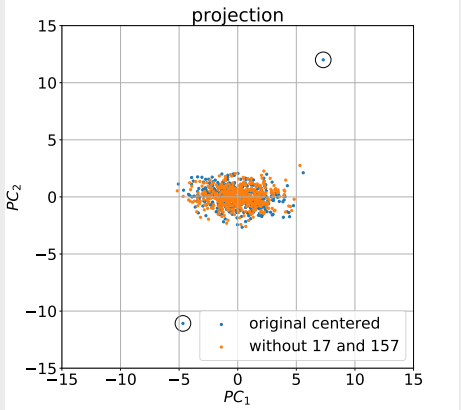
Exercise H2.1: Preprocessing (homework, 1 point)

(a) Load the dataset `pca2.csv`. Compute the Principal Components PC1 and PC2 and plot the data in the coordinate system PC2 vs. PC1. – What do you observe?

(b) Remove observations 17 and 157 (subtract 1 for zero-based indexing) and redo the above. – What is the difference?

Solution





Exercise H2.2: Sphering/Whitening (homework, 3 points)

(a) Load the dataset `pca4.csv` and visually inspect for outliers in the individual variables. Come up with a simple heuristic to remove the outliers from the data and proceed with using the "cleaned" data in all the next steps.

(b) Perform PCA on a reasonable[1] subset of this data. Use a scree plot to determine how many PCs represent the data well.

(c) "Sphere"/"Whiten" the data, i.e. transform the data into 4 new *uncorrelated-unit-variance* variables. Each new variable should have *mean 0* and a *standard deviation equal to 1*. This can be done e.g. using the transformation

$$\underline{\mathbf{V}} = \underline{\mathbf{\Lambda}}^{-1/2}\underline{\mathbf{M}}^{\top}\underline{\mathbf{X}},$$

where,

the new variables v_i form the rows of $\underline{\mathbf{V}}$,

$\underline{\mathbf{M}}$ is a matrix containing in its columns the normalized eigenvectors of the covariance matrix $\underline{\mathbf{C}}$ of the centered data $\underline{\mathbf{X}} \in \mathbb{R}^{N\times p}$

and $\underline{\mathbf{\Lambda}}$ is a diagonal matrix containing the corresponding eigenvalues[2].

(d) Generate 3 heatmap plots. Specifically a heatmap of:

(i) the 4×4 covariance matrix $\underline{\mathbf{C}}$,

(ii) the covariance matrix of the data projected onto all 4 PCs, i.e. $\underline{\mathbf{M}}^{\top}\underline{\mathbf{X}}$, and

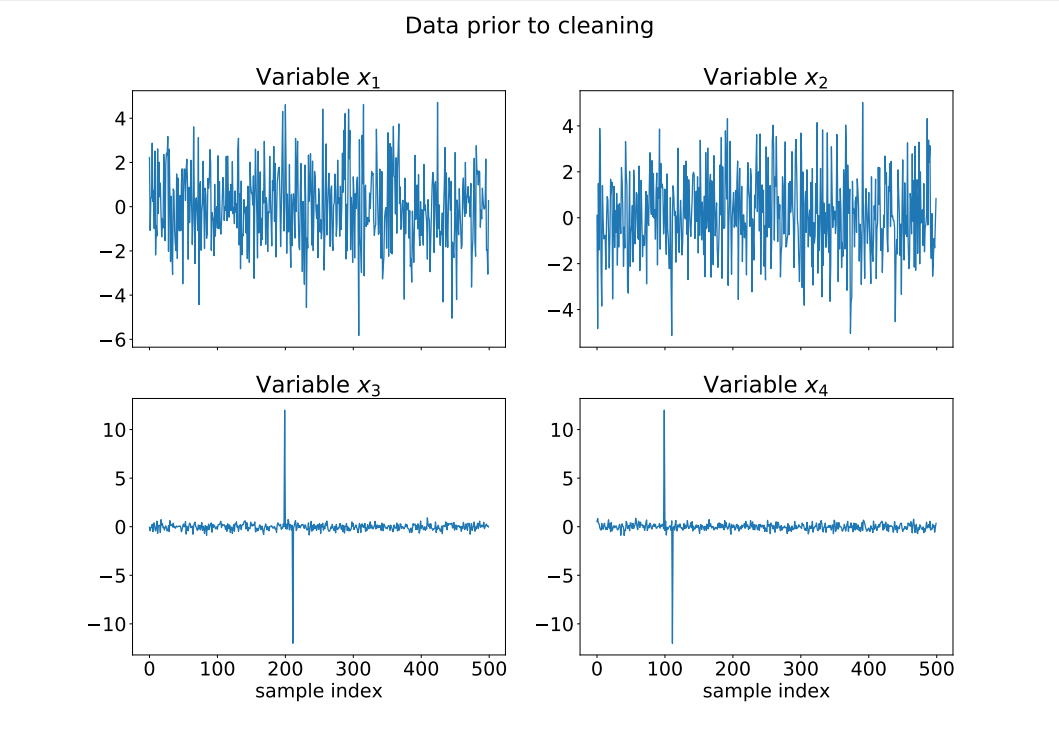
(iii) the covariance matrix of the transformed variables v_i.

[1] The reason for leaving this choice up to you is for you to find a good tradeoff between speed (small subset) vs. robust/stable results (large subset but slow). Hint: Start small then double the size of the subset, if results don't change, i.e. are stable, the subset is reasonably large.

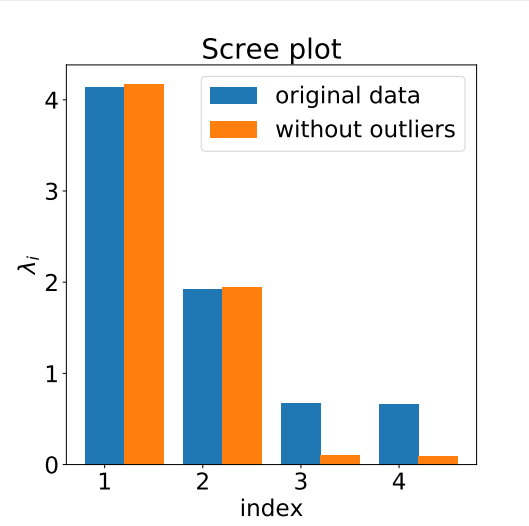
[2] If your data is stored in a $p \times N$ matrix, you can use $\underline{\mathbf{V}} = \underline{\mathbf{M}}^{\top}\underline{\mathbf{X}}\,\underline{\mathbf{M}}\,\underline{\mathbf{\Lambda}}^{-1/2}$, which will give you a $p \times 4$ matrix for $\underline{\mathbf{V}}$.

Solution

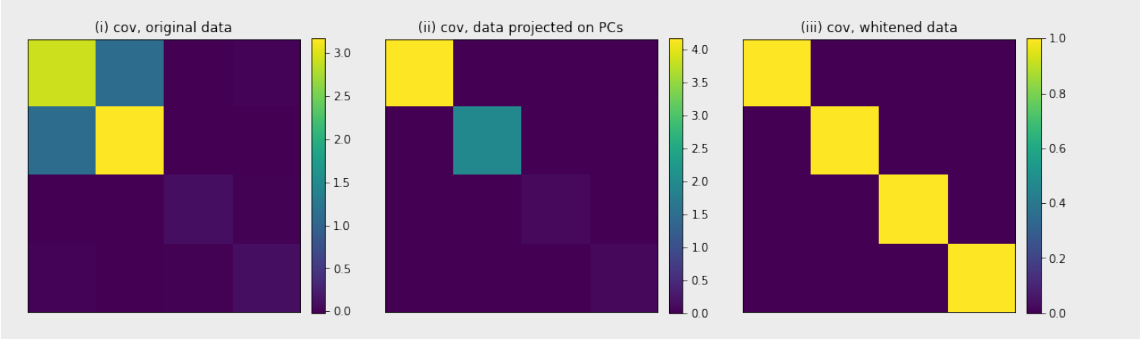
(a)



(b)



(d)


(i) cov, original data
3.0
2.5
2.0
1.5
1.0
0.5
0.0
(ii) cov, data projected on PCs
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
(iii) cov, whitened data
1.0
0.8
0.6
0.4
0.2
0.0

Exercise H2.3: Oja's Rule: Derivation (homework, 2 points)

Consider a linear connectionist neuron whose output $y = y(t)$ at time t is an inner product of the N-dim input vector $\underline{\mathbf{x}} = \underline{\mathbf{x}}(t)$ with the N-dim weight vector $\underline{\mathbf{w}}$:

$$y = \underline{\mathbf{w}}^\top \underline{\mathbf{x}}.$$

The Hebbian update rule for learning the weights can be written as

$$w_i(t+1) = w_i(t) + \varepsilon y(t) x_i(t), \quad i = 1, 2, ..., N$$

where ε is the learning rate and t is the iteration step. As was shown in the lecture, the Hebbian learning rule leads to a divergence of the length of the weight vector. Therefore, the following *explicit normalization* was introduced by Oja:

$$w_i(t+1) = \frac{w_i(t) + \varepsilon y(t) x_i(t)}{\left(\sum_{j=1}^{N} \left[w_j(t) + \varepsilon y(t) x_j(t) \right]^2 \right)^{\frac{1}{2}}} \qquad (*)$$

Task: Derive an approximation to this update rule for a small value of the learning rate parameter ε by Taylor-expanding the right hand side of the equation (*) with respect to ε around $\varepsilon = 0$. Show that neglecting terms of second or higher order in ε gives *Oja's rule*:

$$w_i(t+1) = w_i(t) + \varepsilon y(t) \left[x_i(t) - y(t) w_i(t) \right].$$

Solution

In matrix notation, this reads

Let

$$f(\varepsilon) := w_i(t+1) = \frac{w_i(t) + \varepsilon\, y(t)\, x_i(t)}{\sqrt{\sum_{j=1}^{N} \left[w_j(t) + \varepsilon\, y(t)\, x_j(t) \right]^2}}$$

The Taylor expansion of $f(\varepsilon)$ around $\varepsilon_0 = 0$:

$$f(\varepsilon) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\varepsilon_0)}{n!} (\varepsilon - \varepsilon_0)^n = f(0) \; + \; f'(0)\, \varepsilon \; + \; \frac{1}{2} f''(0)\, \varepsilon^2 \; + \; \frac{1}{6} f'''(0)\, \varepsilon^3 \; + \ldots$$

Neglecting terms of second order or higher in ε yields the following approximation:

$$f(\varepsilon) \; \approx \; f(0) \; + \; f'(0)\, \varepsilon \; + \; \mathcal{O}(\varepsilon^2)$$

$$f'(\varepsilon) = \frac{\partial f(\varepsilon)}{\partial \varepsilon} \overset{f(\varepsilon)=\frac{u}{v}}{=} \frac{u' \cdot v - u \cdot v'}{v^2} = \frac{u'}{v^2} \cdot v - \frac{u \cdot v'}{v^2}$$

with $u := w_i(t) + \varepsilon\, y(t)\, x_i(t), \quad$ then $\quad u' = \frac{\partial u}{\partial \varepsilon} = y(t)\, x_i(t)$
and with

$$\begin{aligned} v :&= \left(\sum_{j=1}^{N} \left[w_j(t) + \varepsilon y(t) x_j(t) \right]^2 \right)^{1/2} \\ &= \left(\sum_{j=1}^{N} \left[w_j^2(t) + 2\, \varepsilon\, y(t)\, x_j(t)\, w_j(t) + \varepsilon^2\, y^2(t)\, x_j^2(t) \right] \right)^{1/2} \end{aligned}$$

then

$$v' = \frac{1}{2}\left(\sum_{j=1}^{N}\left[w_j^2(t) + 2\,\varepsilon\,y(t)\,x_j(t)\,w_j(t) + \varepsilon^2\,y^2(t)\,x_j^2(t)\right]\right)^{-1/2} \cdot \sum_{j=1}^{N}\left[2\,y(t)\,x_j(t)\,w_j(t) + 2\,\varepsilon\,y^2(t)\,x_j^2(t)\right]$$

To evaluate $f'(\varepsilon_0) = f'(0)$:

$u(0) = w_i(t), \quad u'(0) = y(t)\,x_i(t), \quad v(0) = \left(\sum_{j=1}^{N}\left[w_j(t)\right]^2\right)^{1/2} = \underbrace{||\underline{\mathbf{w}}|| \;=\; 1}_{\substack{\text{guaranteed from} \\ \text{explicit normalization}}}$,

$$\begin{aligned} v'(0) &= \frac{1}{2}\left(\sum_{j=1}^{N}\left[w_j^2(t)\right]\right)^{-1/2} \cdot \sum_{j=1}^{N}\left[2\,y(t)\,x_j(t)\,w_j(t)\right] \\ &= \frac{y(t)\sum_{j=1}^{N}\left[\,x_j(t)\,w_j(t)\right]}{\left(\sum_{j=1}^{N}\left[w_j^2(t)\right]\right)^{1/2}} \\ &\overset{x_j\,w_j = y}{=} \frac{y^2(t)}{||\underline{\mathbf{w}}||}. \end{aligned}$$

Evaluating $f'(\varepsilon_0)$:

$$\begin{aligned} f'(\varepsilon_0) = f'(0) &= \frac{u' \cdot v}{v^2} - \frac{u \cdot v'}{v^2} \\ &= \frac{y(t)\,x_i(t) \cdot ||\underline{\mathbf{w}}||}{||\underline{\mathbf{w}}||^2} - \frac{w_i(t) \cdot \frac{y^2(t)}{||\underline{\mathbf{w}}||}}{||\underline{\mathbf{w}}||^2} \\ &= \frac{y(t)\,x_i(t)}{||\underline{\mathbf{w}}||} - \frac{w_i(t) \cdot y^2(t)}{||\underline{\mathbf{w}}||^3} \\ &\overset{||\underline{\mathbf{w}}||=1}{=} y(t)\,x_i(t) - w_i(t) \cdot y^2(t). \end{aligned}$$

Evaluating $f(\varepsilon_0)$

$$f(0) = \frac{w_i(t)}{\left(\sum_{j=1}^{N} w_j^2(t)\right)} = \frac{w_i(t)}{||\underline{\mathbf{w}}||} = w_i(t)$$

Finally:

$$\begin{aligned} f(\varepsilon) \;&\approx\; f(0) \;+\; f'(0)\,\varepsilon \\ &= w_i(t) \;+\; \left[y(t)\,x_i(t) - w_i(t) \cdot y^2(t)\right]\;\varepsilon \\ &= w_i(t) \;+\; \varepsilon\,y(t)\,\left[\,x_i(t) - y(t) \cdot w_i(t)\right] \;\leftarrow \textit{Oja's rule} \end{aligned}$$

.

Exercise H2.4: Oja's Rule: Application (homework, 4 points)

The file `data-onlinePCA.txt` contains observations from an experiment run over an interval of time. The first datapoint was observed at $t_0 = 0$ and the last at $t_T \mathrel{\widehat{=}} 10s$.

(a) Produce a scatter plot of the data and indicate the time index by the color of the datapoints. You can break the full dataset into 10 blocks corresponding to 1 second length each and use 10 different colors, 1 color for each block.

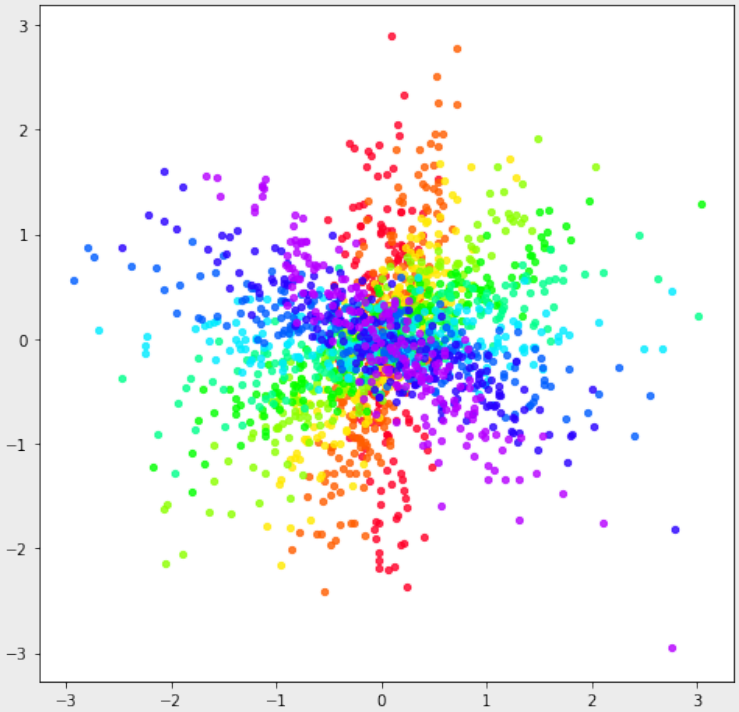
(b) Determine the principal components, using *batch PCA*, for *each of the 10 blocks separately*. Plot the direction of first PC for each block (e.g. as an arrow or the endpoint of it) together with the original data.

(c) Implement Oja's rule and apply it with a learning rate parameter $\varepsilon \in \{0.002, 0.04, 0.45\}$ to the dataset. In each iteration, take the next[3] data point and apply the learning step. Plot the weights at each timestep (as points whose x vs. y coordinates are given by the weight for x and y) in the same plot as the original data. Use the colors from (a) to indicate the time index for each plotted weight.

(d) Interpret your results.

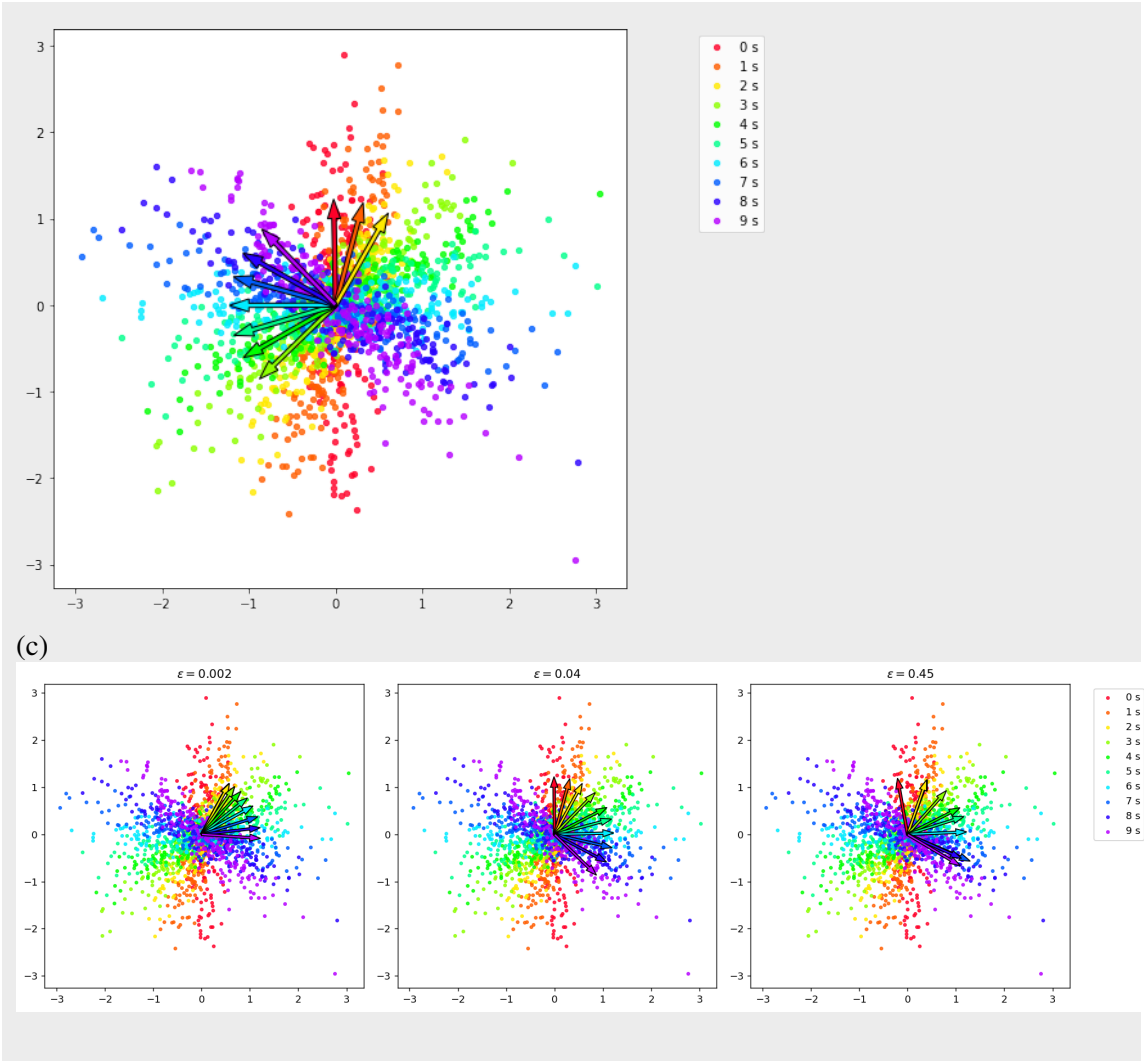
Solution

(a)



(b)

[3] In contrast to the algorithm of the lecture where datapoints are sampled randomly here we do not want to lose the temporal relation between the data points. Therefore, process the data in the original temporal order.



(c)

Total 10 points.