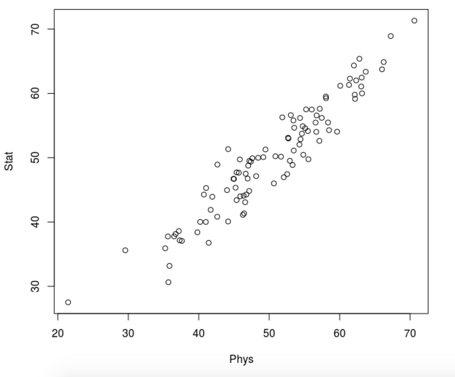
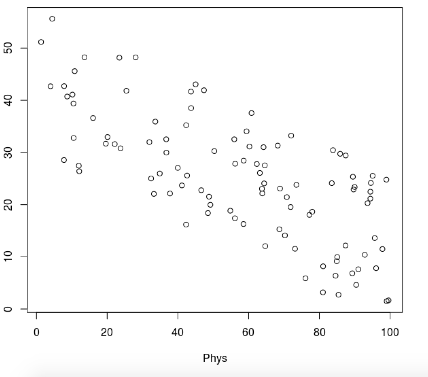
392013 Exercises Algorithmic Cheminformatics

Exercise 09. June 27, 2025
(no mandatory exercises)

1 PCA Visually

Consider 100 students with Physics and Statistics grades shown in the two diagram below.





For each of the diagrams:

1. Which is the direction (as a vector) along which the data varies the most?
2. What is (roughly) the point representing the mean value of the statistics/mathematics degree?

2 Covariance Matrix

Assume the following grades:

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

1. Define a matrix X' where the rows represent samples (students), and the columns represent features (the different tests).
2. Perform mean centering: subtract each data value from its variable's measured mean so that its empirical mean (average) is zero, call the resulting matrix X.

3. Without computing, what do you expect the covariance between a) Math and English and b) English and Art to be? Postive or negative?
4. Ignoring Bessel's correction, which formula (cmp. slide 9) should be used to compute the covariance matrix?
5. Compute the variance of each test grades and the covariance between the test grades.
6. Determine the covariance matrix a) with Bessel's corection, b) without Bessel's correction.

3 Projection onto new feature space

When analysing the Iris dataset, we stored the data in form of a 150×4 matrix where the columns are the different features, and every row represents a separate flower sample. Let's call this matrix X. Given the first and the second principal component w_1 and w_2.

1. How do we, using matrix multiplication, project the 4-dimensional feature space onto a new 2-dimensional feature space that is spanned by w_1 and w_2? How does this relate to the formula on page 7 of the slide set?
2. Give the sizes of the matrices that you used in 1.)

4 Eigenfaces

For the presentation of Eigenfaces and PCA I was using functionality from OpenCV (`https://opencv.org/`). For this exercise, we will use another set of tools, namely Google's Colabatory framework `https://colab.research.google.com` and, `scikit-learn` (`https://scikit-learn.org`), a free software machine learning library for Python for data analytics, statistics, machine learning, and `Scipy` (`https://www.scipy.org/`, the scientific Python ecosystem). All of those in itself can be quite overwhelming, however, the idea of this exercise is to show you how easy it is to use and integrate all those.

Proceed as follows:

1. Download the Jupyter notebook `plot_eigenfaces.ipynb` from here : `https://scipy-lectures.org/packages/scikit-learn/auto_examples/plot_eigenfaces.html` (at the very end of the page).
2. Upload the notebook to `https://colab.research.google.com`.

3. Run the cells subsequently (you can safely ignore everything related to Support Vector classification, and anything after "Doing the Learning: Support Vector Machines").

4. Add a cell (after cell 8) and print the vector in the 150-dimensional Eigenface-space that corresponds to the projection of the first image in the test dataset `X_test`.