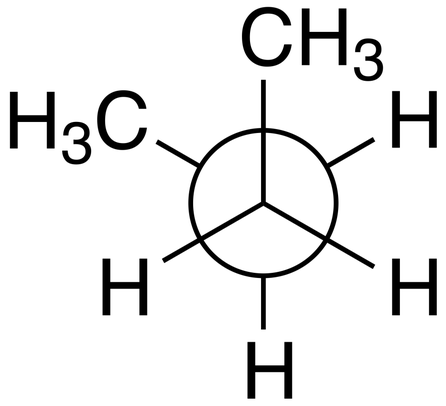
3.2.2 Dimensionality reduction

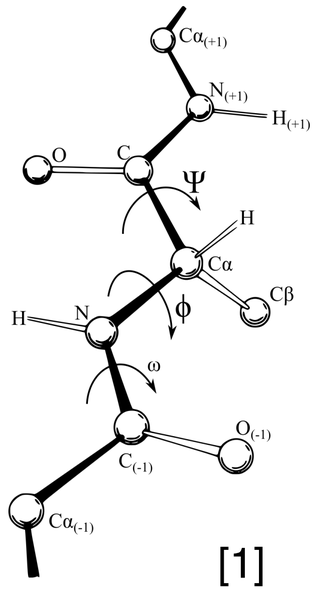
- This module will consist of a mini-lecture and exercises on dimensionality reduction
- At the end of this module, you should be able to answer the following questions:
  - What are the benefits of dimensionality reduction?
  - In principal components analysis, what are the meanings of the eigenvectors and eigenvalues?
  - What does it mean to project a configuration onto an eigenvector?

Dimensionality reduction with principal component analysis

Dimensionality reduction

- Biomolecular simulations have 3N dimensions, but in practice, functionally relevant macromolecular motion can be described with a lot fewer.
- Cartesian coordinates (x, y, and z) of each atom are the most common description but molecular systems can be described by other coordinate systems
- You have probably already seen
  - the free energy of butane as a function of the dihedral angle
  - the Ramachandran diagram, where protein backbones are described by the ϕ and ψ angles



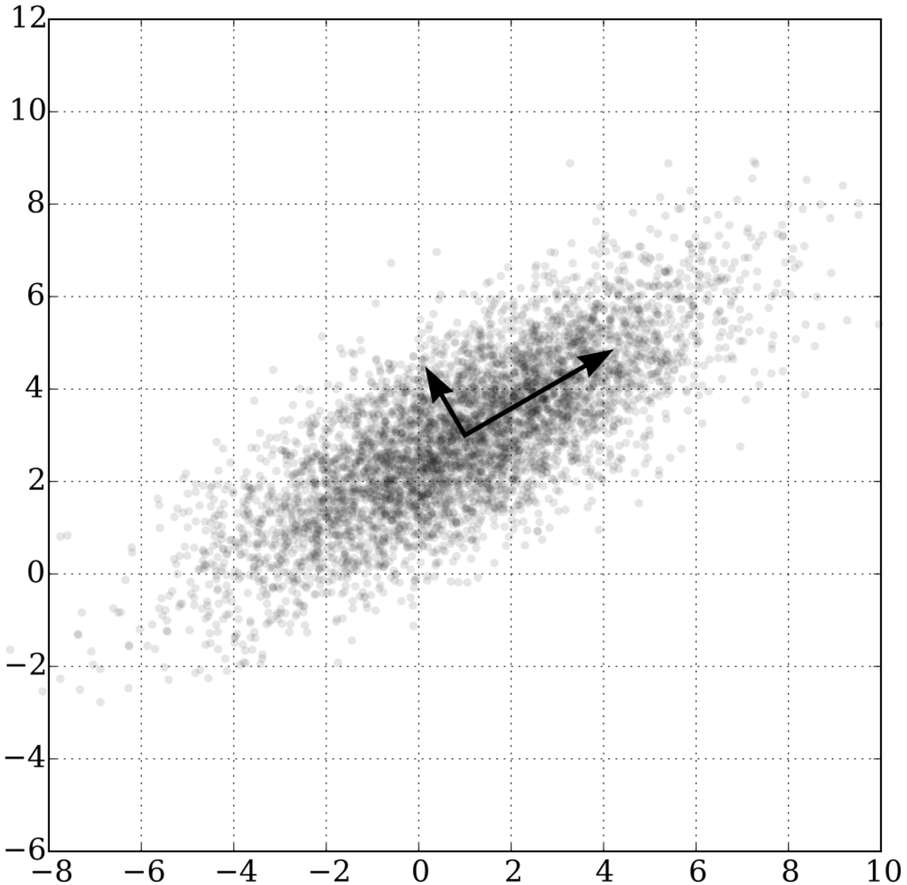


[1]

Principal component analysis (PCA)

- An *automated* way to do dimensionality reduction
- A linear transformation of coordinates in decreasing order of variance
  - First principal component has the largest variance
  - Second principal component has second largest variance
  - And so forth
- Dimensions can be reduced by keeping the highest-variance dimensions

- See https://en.wikipedia.org/wiki/Principal_component_analysis



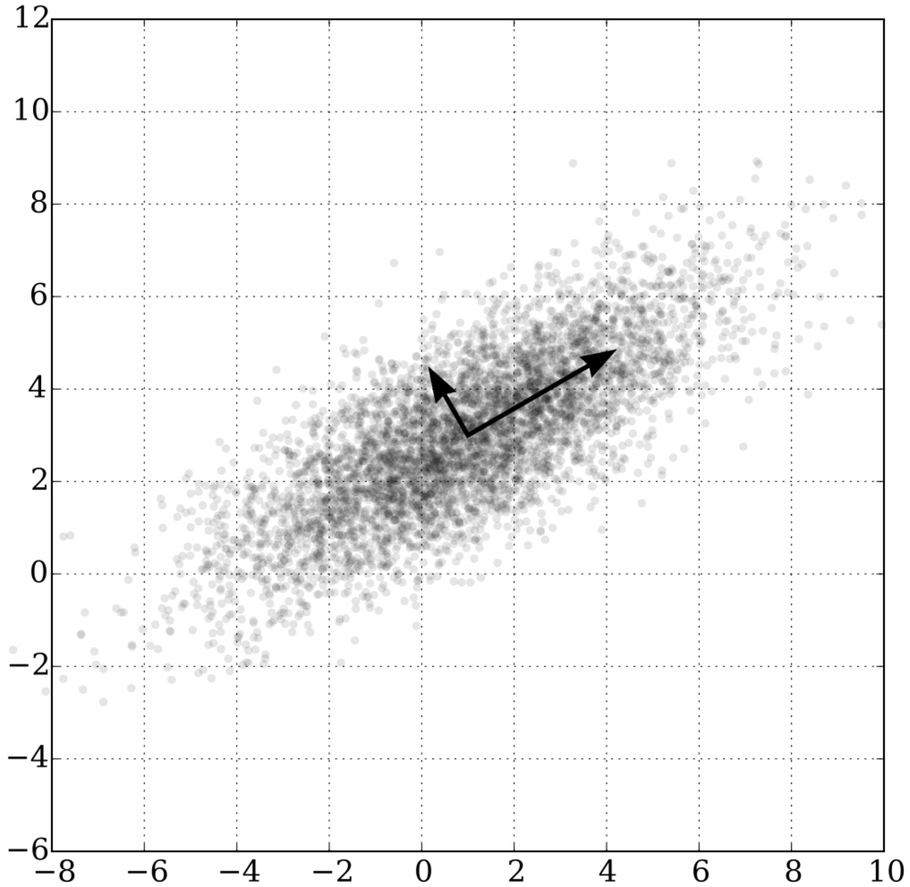
[2]

The matrices of PCA

- **CV** = **PV**
- **C**: covariance matrix
  - C_{kl} is the covariance between dimensions k and l.
  - Usually empirically estimated from data.
- **V**: matrix of column *eigenvectors*
  - V_{kl} is the
    - importance of the original coordinate k
    - in the transformed coordinate l.
    - known as the principal components
    - the columns are orthonormal vectors
- **P**: diagonal matrix of *eigenvalues*
  - $P_{kl} = \lambda_{kl}$ if k = l
  - $P_{kl} = 0$ otherwise
  - variances in transformed coordinate system
  - scaling of the eigenvectors
- All three matrices have the same size

Question

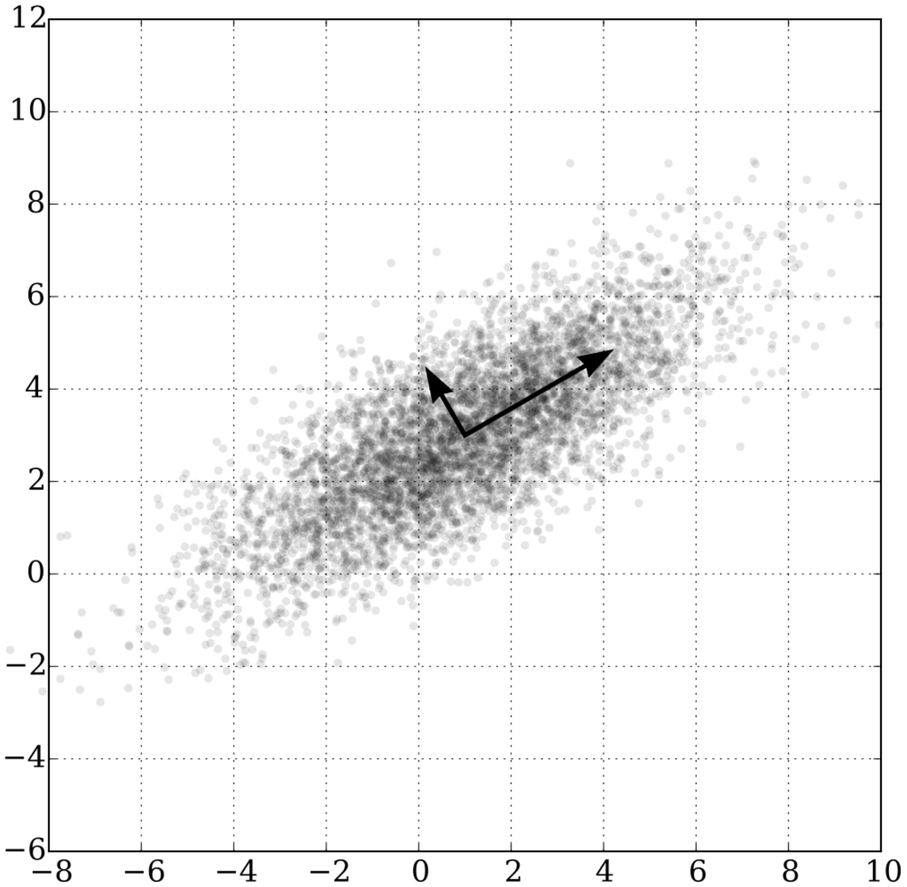
- Approximately, what is the first principal component/eigenvector?
- The vector is ~ (3.5, 2). The magnitude is ~ 4. The normalized eigenvector is ~ (0.875, 0.5).



[2]

Question

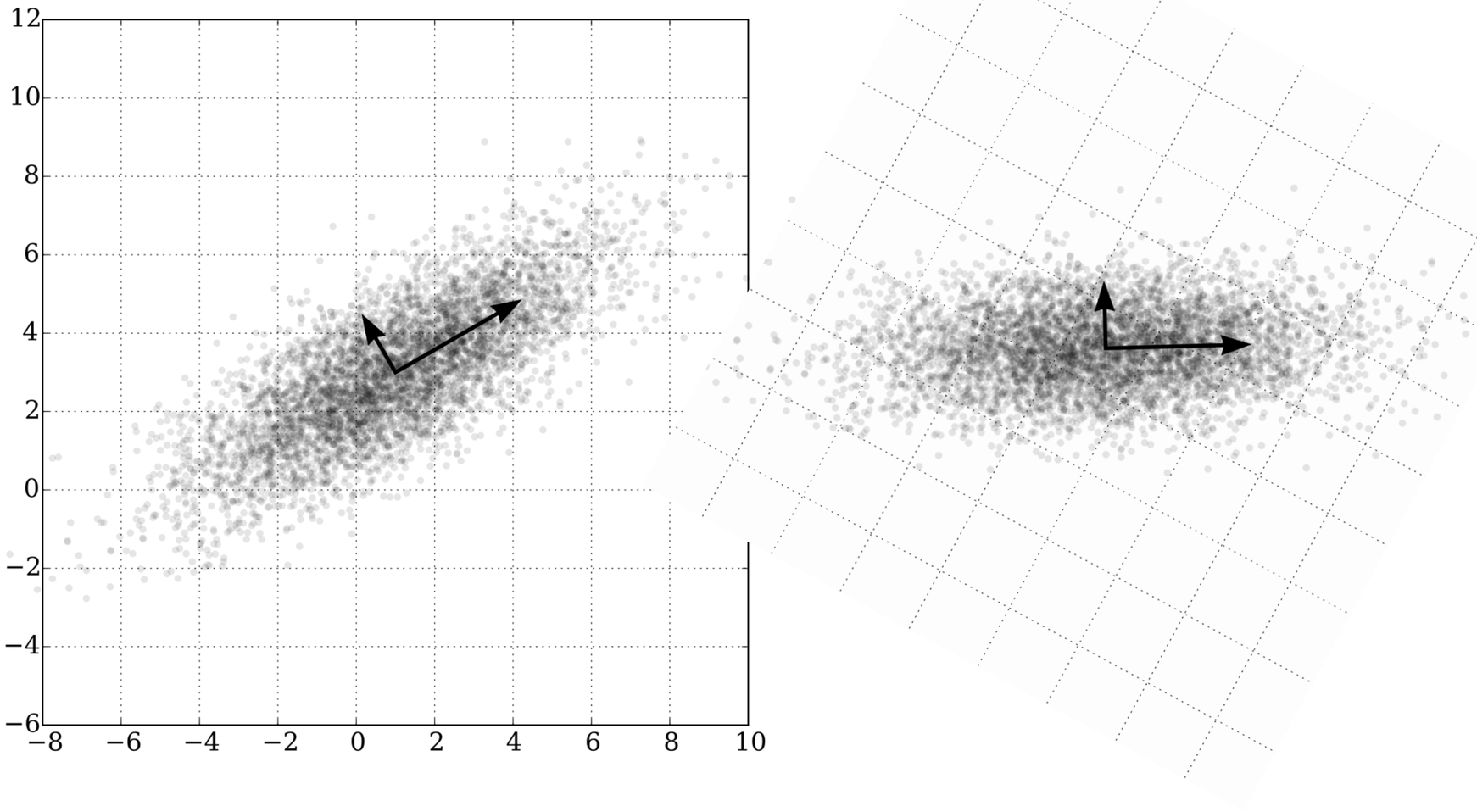
- How do the first and second eigenvalue compare?
- The first eigenvalue is approximately three times larger than the second



[2]

Question

- What would this plot look like if the x axis were projections onto the first PC and the y axis were projections on the second PC?



PCA analysis of molecular simulation

- Load and run the jupyter notebook https://github.com/CCBatIIT/modelingworkshop/exercises/mpro/RIKEN_trajectory/analysis/DimerPCA.ipynb, which demonstrates PCA analysis for a simulation of mpro

- What part of the enzyme has the strongest weight in the first three principal components?
- What is the dot product between a principal component and itself? What is the dot product between a principal component and another principal component?
- What fraction of the variance is accounted for in the first three principal components?
- At approximately what point in the trajectory does the largest conformational shift occur?

- Exercise: repeat the PCA calculation without the C terminus by setting nresidues = 300. How do the results differ?

Time-lagged independent component analysis (TICA)

- Like PCA, based on eigenvector and eigenvalues of a matrix
  - autocorrelation instead of covariance matrix
  - isolates slow motions into coordinates
- Often used in Markov state models (MSMs)
- Slow is often but not always functionally relevant

Review

- What are the benefits of dimensionality reduction?
- In principal components analysis, what are the meanings of the eigenvectors and eigenvalues?
- What does it mean to project a configuration onto an eigenvector?