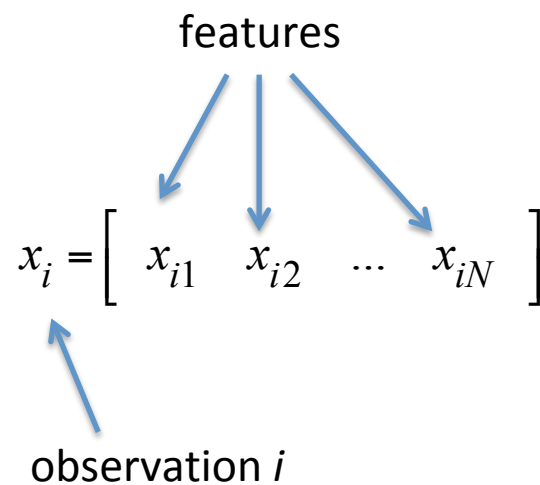


Fingerprints

Many names for the same thing:

- Fingerprint vector
- Feature vector
- Descriptor
- Representation



In materials science, the fingerprint is a mathematical representation/mapping of the positions and chemical identity of all atoms into a vector x .

Fingerprints

Discuss with your neighbor, which general criteria a good fingerprint (for materials) should satisfy

Fingerprints – general criteria

A good material fingerprint should satisfy the following:

1. **Uniqueness:** Two different materials should give different x
2. **Invariance:** x should be invariant under transformations that preserve the material (e.g. rotation, translation should yield same x). Could in principle be learned but hard and no reason!
3. **Descriptive:** materials that we expect to be similar should be represented by x that are close (in some norm – usually Euclidian).
4. **Smooth:** x should vary smoothly with the input (if smoothness is defined in input space). For materials: smooth in atomic positions.
5. **Simple:** generating the fingerprint should be computationally efficient

Fingerprints

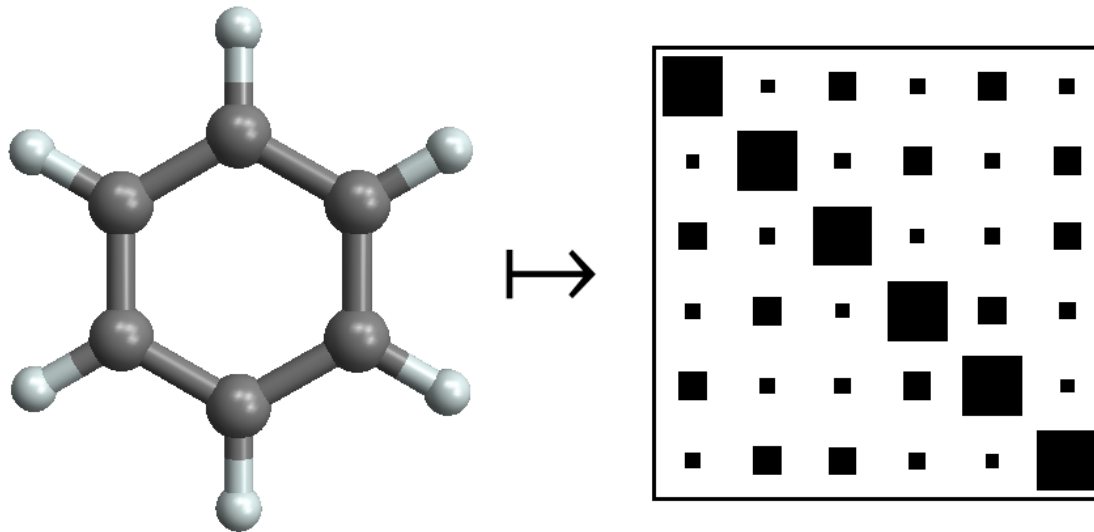
Two situations:

- (1) ML for one specific material or class of materials
(e.g. the perovskites) → local fingerprint
- (2) ML on the entire space of materials → global fingerprint

In case (1) the dimension of the fingerprint vector is unimportant (except subject to usual considerations of bias vs. overfitting).

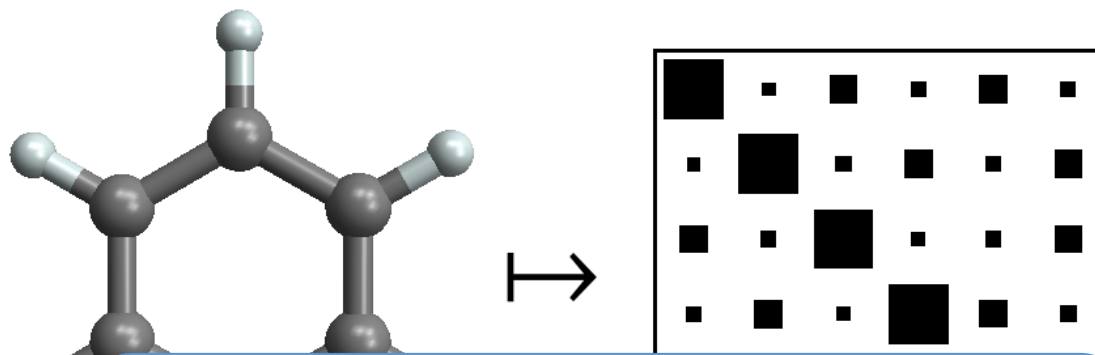
In case (2) the dimension of the fingerprint vector should ideally be the same for all materials, i.e. independent of chemical composition and structure. Much harder problem!

Coulomb matrix



$$M_{IJ} = \begin{cases} 0.5Z_I^{2.4} & \text{for } I = J \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \text{for } I \neq J \end{cases}$$

Coulomb matrix



Is it a good fingerprint?
Does it have issues, which?

$$M_{IJ} = \begin{cases} 0.5Z_I^{2.4} & \text{for } I = J \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \text{for } I \neq J \end{cases}$$

Coulomb matrix

A general problem is that Z is not very descriptive: Atoms with similar Z do not have similar chemical properties. Also Z is entangled with atom distances

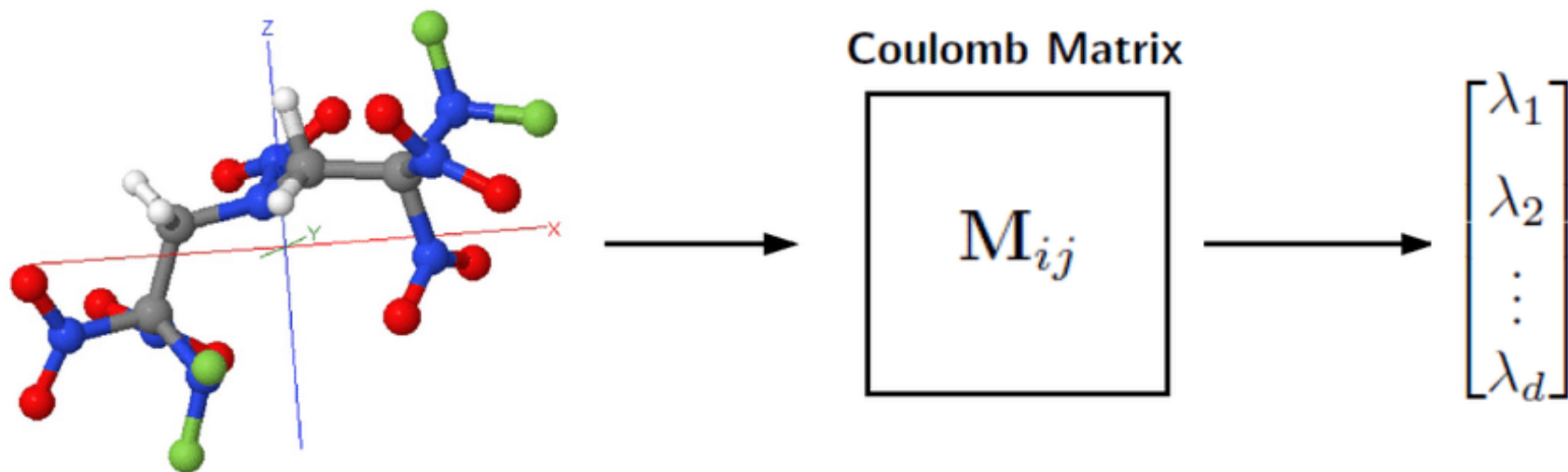
Coulomb matrix

A general problem is that Z is not very descriptive: Atoms with similar Z do not have similar chemical properties. Also Z is entangled with atom distances

Moreover, not *invariant* under ordering of atoms!

Ordering of CM \rightarrow introduces *discontinuities*.

Diagonalization of CM \rightarrow breaks the *uniqueness* (different molecules same eigenvalues). Keeping only the N largest eigenvalues ensures fixed dimension of the fingerprint.



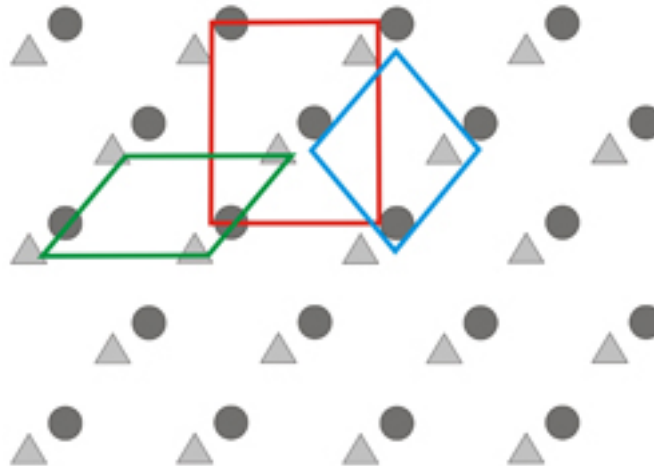
Coulomb matrix for crystals

What is the problem for periodic systems (if any)?
Can't we just use the atoms in a unit cell?

Coulomb matrix for crystals

What is the problem for periodic systems (if any)?
Can't we just use the atoms in a unit cell?

- The fingerprint x would depend on the choice of unit cell



→ Clearly, for a periodic crystal, the fingerprint must be independent on the unit cell.

Coulomb matrix for crystals

Extended Coulomb matrix:

Use the $M \times MN$ matrix obtained by extending the Coulomb matrix to pairs of atoms: M atoms in unit cell (rows) and the N closest unit cells (columns). To avoid the long range $1/r$ interaction use a faster decay function $\exp(-r)$

Coulomb matrix for crystals

Extended Coulomb matrix:

Use the $M \times MN$ matrix obtained by extending the Coulomb matrix to pairs of atoms: M atoms in unit cell (rows) and the N closest unit cells (columns). To avoid the long range $1/r$ interaction use a faster decay function $\exp(-r)$

Ewald sum matrix:

Use the $M \times M$ matrix obtained by summing for each pair of atoms (i,j) in the unit cell 0 over all other unit cells n :

$$(*) \quad x_{ij} = Z_i Z_j \sum_n \frac{1}{|r_i^n - r_j^0|}$$

Coulomb matrix for crystals

Extended Coulomb matrix:

Use the $M \times MN$ matrix obtained by extending the Coulomb matrix to pairs of atoms: M atoms in unit cell (rows) and the N closest unit cells (columns). To avoid the long range $1/r$ interaction use a faster decay function $\exp(-r)$

Ewald sum matrix:

Use the $M \times M$ matrix obtained by summing for each pair of atoms (i,j) in the unit cell 0 over all other unit cells n :

$$(*) \quad x_{ij} = Z_i Z_j \sum_n \frac{1}{|r_i^n - r_j^0|}$$

Sine matrix

The potential $(*)$ has the properties: (1) periodic in each coordinate (r_i and r_j). (2) diverges for $r_i = r_j$. A simple function that incorporates these properties:


$$x_{ij} = \left\| \mathbf{B} \cdot \sum_{k=x,y,z} \mathbf{e}_k \sin^2(\pi \mathbf{e}_k \cdot \mathbf{B}^{-1} \cdot (r_i - r_j)) \right\|^{-1}$$

Many-body tensor representation

An alternative representation that encodes similar information as the CM but avoids the invariance and uniqueness problems is:

$$f_{CM}(x, z_1, z_2) = \sum_{i,j}^{N_a} \delta(x - d_{i,j}^{-1}) \delta(z_1, z_i) \delta(z_2, z_j)$$

Different encoding of atomic number compared to CM



Many-body tensor representation

An alternative representation that encodes similar information as the CM but avoids the invariance and uniqueness problems is:

$$f_{CM}(x, z_1, z_2) = \sum_{i,j}^{N_a} \delta(x - d_{i,j}^{-1}) \delta(z_1, z_i) \delta(z_2, z_j)$$

Different encoding of atomic number compared to CM

We can broaden the x -delta function (e.g. Gaussian) to obtain a smoother dependence on atomic coordinates. We can replace the Kronecker z -delta function by an element correlation matrix, C . We can introduce weighting function w to weigh down contributions from atoms far apart:

$$f_2(x, z_1, z_2) = \sum_{i,j}^{N_a} w_2(i, j) \underbrace{D(x, g_2(i, j))}_{D \text{ could be any function.}} C_{z_1, Z_i} C_{z_2, Z_j}$$

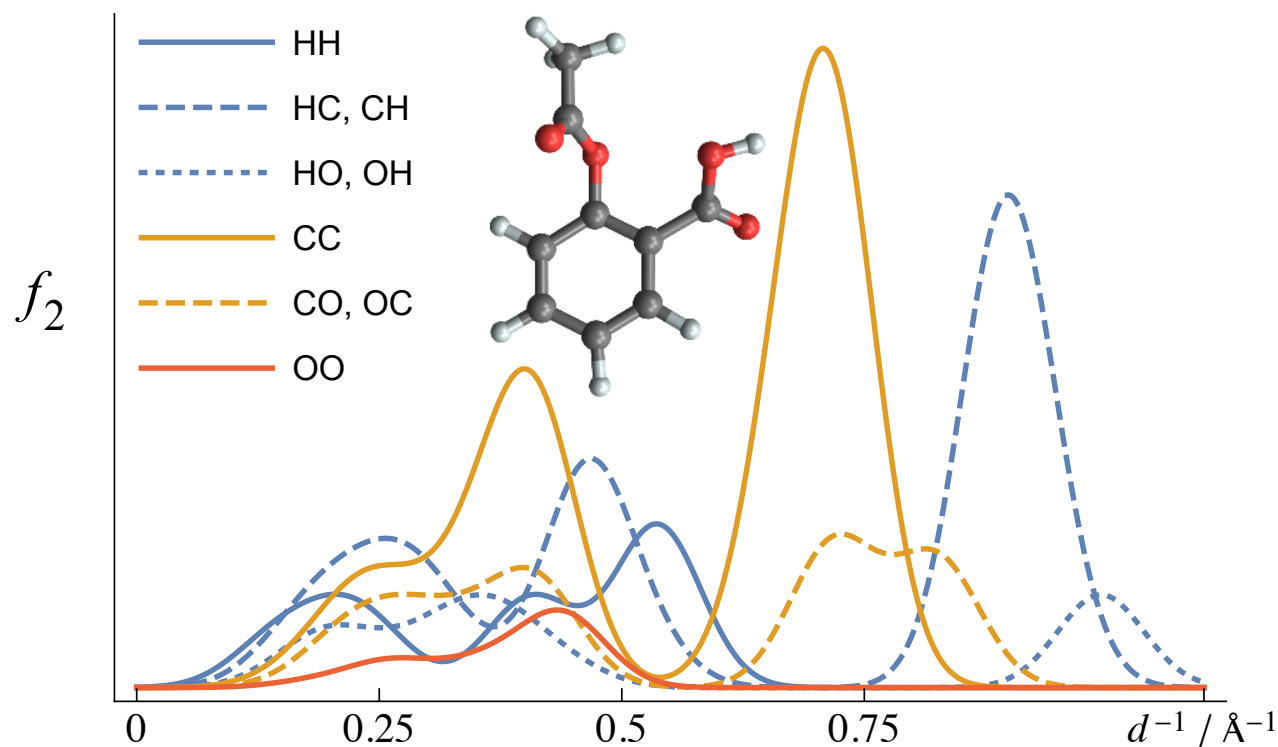
Discretizing the x -variable into N_x bins this fingerprint has dimension:

$$\dim(f_2) = N_x \times N_e \times N_e$$

N_e : Number of elements. Using all elements (global fingerprint) $N_e \approx 100$ and a very sparse matrix \rightarrow PCA

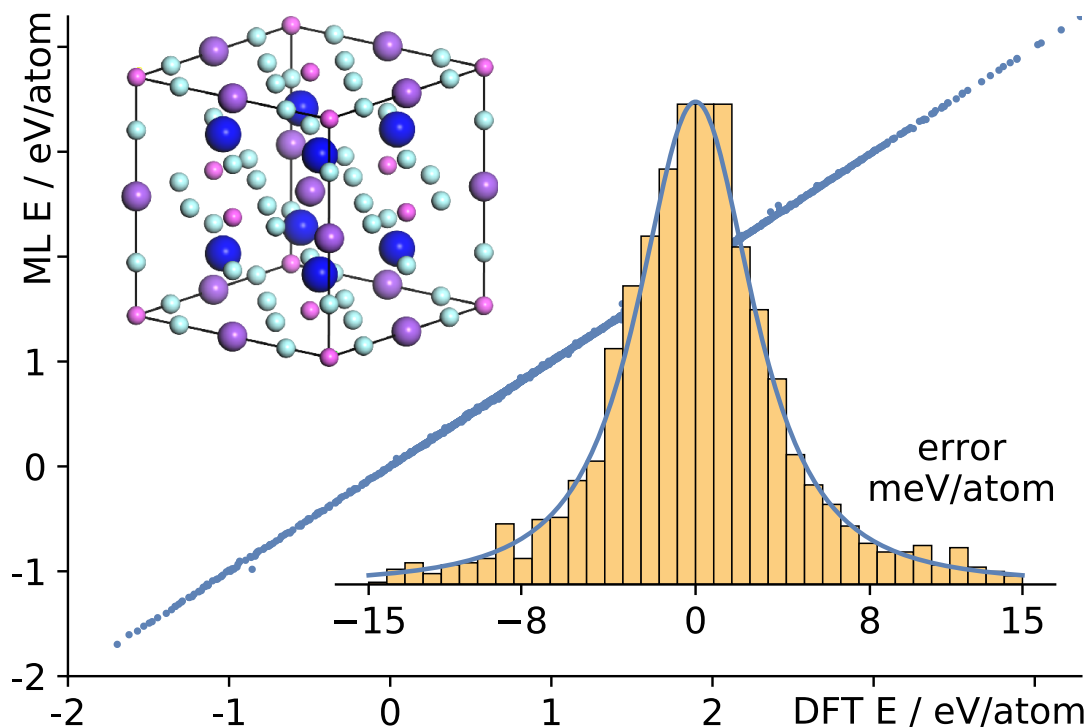
Many-body tensor representation

$$f_2(x, z_1, z_2) = \sum_{i,j}^N w_2(i,j) D(x, g_2(i,j)) C_{z_1, Z_i} C_{z_2, Z_j}$$



Many-body tensor representation

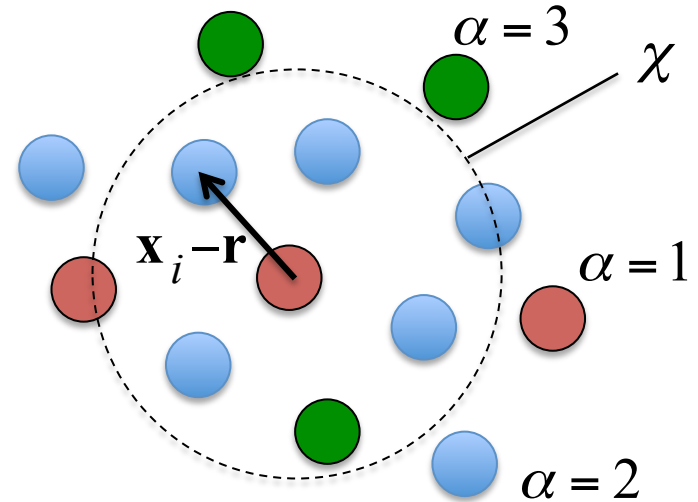
- ABC_2D_6 crystals with 12 different elements.
- Kernel ridge regression for DFT total energies for 11.000 materials.
- Training set size 9000
- Test set size 2000



Smooth overlap of atomic positions (SOAP)

The idea behind SOAP is to represent the local environment of each atom:

$$\rho_{\chi}^{\alpha}(\mathbf{r}) = \sum_{i \in \chi(\alpha)} \exp\left(-\frac{(\mathbf{x}_i - \mathbf{r})^2}{\sigma^2}\right)$$



The similarity of two local environments is defined via the kernel:

$$k(\chi, \chi') = \int d\hat{R} \left| \int \sum_{\alpha} \rho_{\chi}^{\alpha}(\mathbf{r}) \rho_{\chi'}^{\alpha}(\hat{R}\mathbf{r}) d\mathbf{r} \right|^2$$

Average over all rotations
(\rightarrow invariant)

Measure geometric similarity of identical species

Smooth overlap of atomic positions (SOAP)

A global fingerprint can be obtained by collecting all local environments of the atoms in the molecule/unit cell.

Given two structures A and B one defines the covariance of neighborhoods of atoms i in A and j in B:

$$C_{ij}(A, B) = k(\chi_i^A, \chi_j^B)$$

The similarity of the two structures can then be quantified by the average kernel:

$$\bar{K}(A, B) = \frac{1}{N^2} \sum_{ij} C_{ij}(A, B)$$

Other choices are possible, e.g. best match (1-1 atoms)

Performance of different fingerprints

- Predicting atomization energies of organic molecules.
- Kernel ridge regression
- Training on 5000 testing on 2000 molecules

Representation	Kernel	E / kcal mol ⁻¹		α / Å ³	
		RMSE	MAE	RMSE	MAE
CM [4]	Laplacian	4.76	3.47	0.17	0.13
BoB [15]	Laplacian	2.86	1.79	0.12	0.09
BAML [17]	Laplacian	2.54	1.15	0.12	0.07
SOAP [31]	REMatch	1.61	0.92	0.07	0.05
MBTR	Linear	1.14	0.74	0.10	0.07
MBTR	Gaussian	0.97	0.60	0.06	0.04

Performance of different fingerprints

- Predicting energies of energy for changes in geometry of organic molecules
- Kernel ridge regression + DNN
- Training on 10k random geometries from MD and testing on 2k

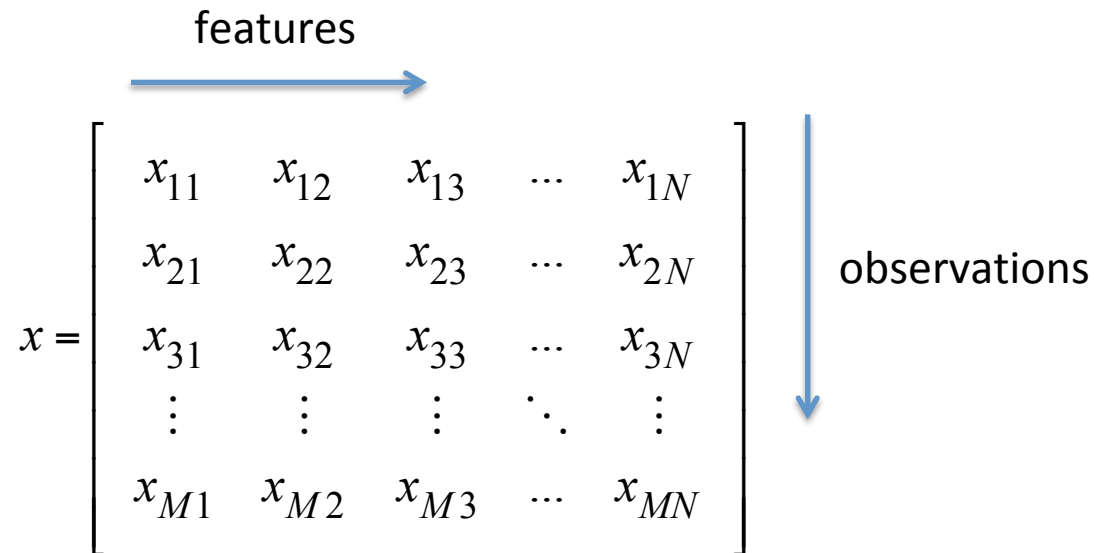
Kernel	DTNN —	GDML Matérn	CM ^{md} Gaussian		MBTR linear		MBTR Gaussian	
Molecule	MAE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
benzene	0.04	0.07	0.05	0.04	0.06	0.05	0.05	0.04
uracil	—	0.11	0.09	0.06	0.14	0.10	0.06	0.04
naphthalene	—	0.12	0.16	0.12	0.15	0.11	0.15	0.11
aspirin	—	0.27	0.40	0.27	0.26	0.18	0.32	0.22
salicylic acid	0.50	0.12	0.15	0.11	0.17	0.12	0.11	0.08
malonaldehyde	0.19	0.16	0.26	0.19	0.28	0.21	0.13	0.10
ethanol	—	0.15	0.25	0.17	0.22	0.16	0.10	0.07
toluene	0.18	0.12	0.22	0.16	0.16	0.12	0.15	0.11

Principal component analysis (PCA)

... is used to reduce the number of features in a fingerprint vector, i.e. *dimensionality reduction*.

The PCA provides a ranking of the new features (linear combinations of the original ones) according to the size of their variance over all observations.

→ PCA essential for fingerprint containing thousands of features



The diagram shows a matrix X with M rows and N columns. A horizontal blue arrow labeled "features" points to the columns. A vertical blue arrow labeled "observations" points to the rows. The matrix is defined as:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2N} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & x_{M3} & \dots & x_{MN} \end{bmatrix}$$

Principal component analysis (PCA)

First step is to *normalize* all features so they have **zero mean** and **unit variance** over the observations.

$$\bar{x} = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} & \bar{x}_{13} & \dots & \bar{x}_{1N} \\ \bar{x}_{21} & \bar{x}_{22} & \bar{x}_{23} & \dots & \bar{x}_{2N} \\ \bar{x}_{31} & \bar{x}_{32} & \bar{x}_{33} & \dots & \bar{x}_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{x}_{M1} & \bar{x}_{M2} & \bar{x}_{M3} & \dots & \bar{x}_{MN} \end{bmatrix}$$

The PCA finds the feature combinations that maximize the variance over the observations.


$$\left\{ \begin{array}{l} \sum_i \bar{x}_{i1} = 0 \\ \sum_i \bar{x}_{i1}^2 = 1 \end{array} \right.$$

Principal component analysis (PCA)

Next step is to construct and diagonalize the co-variance matrix

$$\bar{x}^T \bar{x} = \begin{bmatrix} 1 & \dots & \bar{\sigma}_{N1}^2 \\ \vdots & \ddots & \vdots \\ \bar{\sigma}_{1N}^2 & \dots & 1 \end{bmatrix} \rightarrow V^+ \bar{x}^T \bar{x} V = \text{diag}(\{\lambda_1, \lambda_2, \dots, \lambda_N\})$$

$z = \bar{x}V$: Fingerprint in the new features

Covariance of (normalized)
feature 1 and N :

$$\sum_i \bar{x}_{iN} \bar{x}_{i1} = \bar{\sigma}_{1N}^2$$

The matrix V rotates the original features into a new set of features that are *uncorrelated* over the observations.

The variance of the new features are given by the eigenvalues squared, λ_i^2

Principal component analysis (PCA)

Next step is to construct and diagonalize the co-variance matrix

$$\bar{x}^T \bar{x} = \begin{bmatrix} 1 & \dots & \bar{\sigma}_{N1}^2 \\ \vdots & \ddots & \vdots \\ \bar{\sigma}_{1N}^2 & \dots & 1 \end{bmatrix} \rightarrow V^+ \bar{x}^T \bar{x} V = \text{diag}(\{\lambda_1, \lambda_2, \dots, \lambda_N\})$$

$z = \bar{x}V$: Fingerprint in the new features

Covariance of (normalized)
feature 1 and N :

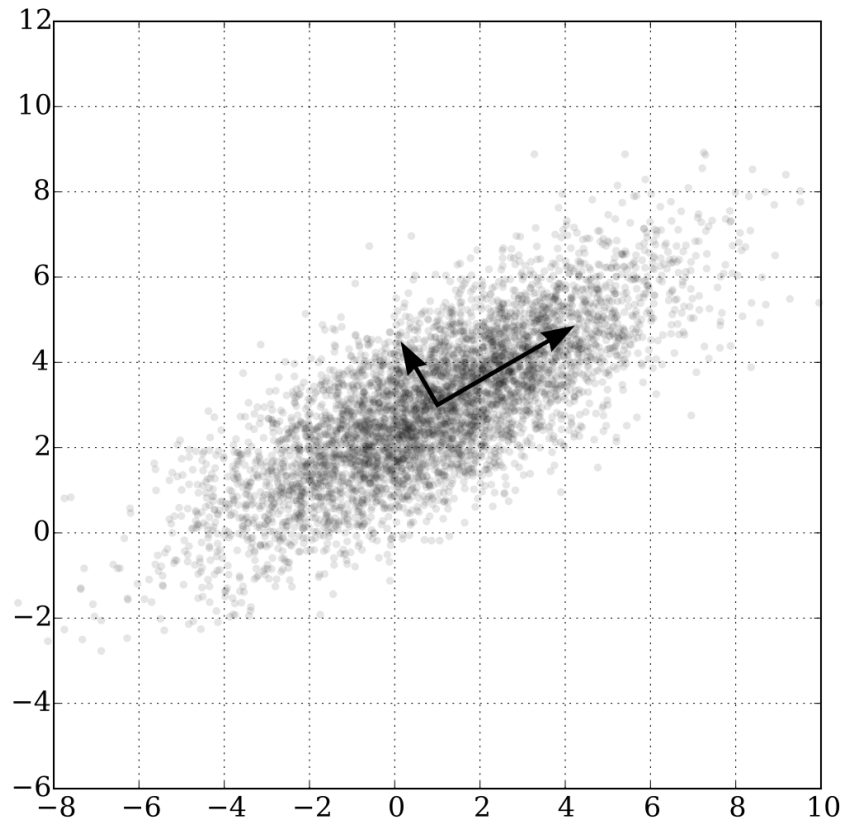
$$\sum_i \bar{x}_{iN} \bar{x}_{i1} = \bar{\sigma}_{1N}^2$$

The matrix V rotates the original features into a new set of features that are *uncorrelated* over the observations.

The variance of the new features are given by the eigenvalues squared, λ_i^2

Principal component analysis (PCA)

Consider two features and show graphically that variance maximization goes hand-in-hand with feature independence (zero covariance)



PCA - example

For representing persons we pick the features:

- Weight
- Height
- Glasses (binary)
- Gender (binary)
- Number of eyes

Three persons (observations) yield:

$$x = \begin{bmatrix} 71 & 184 & 1 & 0 & 2 \\ 55 & 165 & 0 & 0 & 2 \\ 65 & 170 & 1 & 1 & 2 \end{bmatrix}$$

Do you expect correlations?

→ Perform a PCA!



PCA - example

Normalization:

$$x = \begin{bmatrix} 71 & 184 & 1 & 0 & 2 \\ 55 & 165 & 0 & 0 & 2 \\ 65 & 170 & 1 & 1 & 2 \end{bmatrix} \rightarrow \bar{x} = \begin{bmatrix} 0.9 & 1.0 & 0.5 & -0.5 & 0 \\ -1.0 & -0.7 & -1.0 & -0.5 & 0 \\ 0.2 & -0.3 & 0.5 & 1.0 & 0 \end{bmatrix}$$

Result of PCA: $(\bar{x}V)^+ (\bar{x}V) = \text{diag}(\{\lambda_1, \lambda_2, \dots, \lambda_N\})$

$$V = \begin{bmatrix} -0.6 & 0.07 & -0.3 & -0.3 & 0 \\ -0.5 & 0.4 & -0.4 & -0.4 & 0 \\ -0.6 & -0.3 & 0.8 & 0.8 & 0 \\ -0.1 & -0.9 & -0.4 & -0.4 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \lambda = \begin{bmatrix} 2.7 \\ 1.27 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \bar{x}V = \begin{bmatrix} -1.0 & 0.6 & 0 & 0 & 0 \\ 1.3 & 0.3 & 0 & 0 & 0 \\ -0.3 & -0.9 & 0 & 0 & 0 \end{bmatrix}$$

Not to be trusted!

PCA – it's all about the data

It is important to realize that the PCA works only on data (fingerprint) and is completely uncoupled from the model and nature of target.

It might be that features with large variance have no significance for the model/target property.

Example:

You want to predict the ability of a material to absorb solar light.

Your training data shows a large variation in the mass density of the materials.

BUT, light absorption is not correlated with mass density.

Conclusion:

Data (features) with large variance are not necessarily important.

However, if data shows no variation in a feature then it can just as well be removed. Thus large variance is a necessary but not sufficient condition for importance.

(Approximate) time-plan for 10316 (3rd week)

Friday 17/1

8:30 – 10:00 Fingerprints (Kristian)

BREAK

10.30 - 12:00 Pandas presentation with exercises (Mark)

LUNCH

12:00 – 13:00 Machine learning best practices (Nikolaj)

13:00 – 15:00 Fitting with fingerprints (Nikolaj + Mark)

15:00- 16:00 Intro to SKlearn library (Nikolaj + Mark)

Monday 20/1

8:30 – 10:00 Introduction to neural networks (Jakob)

BREAK

10:30 – 11:30 Introduction to a secret materials data set (Kristian)

11:30- 12:00 Introduction to dScribe + Kaggle competition (Mark + Nikolaj)

Monday afternoon – Friday noon

Work on project (groups of 4)

13 – 15: Consultancy hours (Mark, Nikolaj, Morten, Sami: in bld. 309, 2nd floor)

Friday 24/1

13:00- 16:00 Poster session and prize ceremoni