Introduction to Chemometrics

Chemometrics for Spectroscopists

Intensive Course Kraków
2021-11-29 – 12-03

Claudia Beleites
Chemometric Consulting Claudia Beleites

What is Chemometrics?

Definition

Chemometrics means applying statistics (mathematics) to solve chemical problems:

- Design of Experiments (DoE)
- Data analysis
- Quantitative Structure-Activity Relations (QSAR)

Very similar disciplines (statistics for other sciences):

- Biometrics
- Psychometrics
- ...

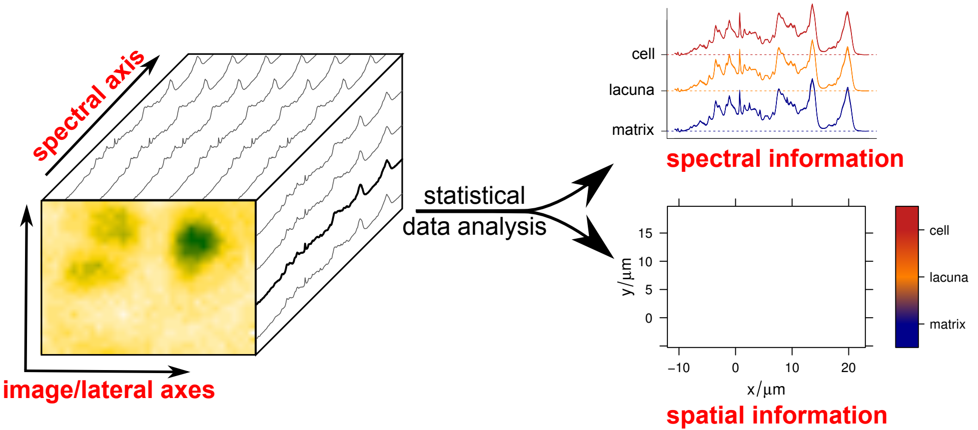
Why bother about statistics?

- Data and/or problem is too complex for “simple” analysis:
  - too many factors influencing the problem
    many sources of variance
  - information is spread out over many variables;
    no single variable carries enough information
  - information is hidden between lots of noise
- even a univariate linear calibration is chemometrics
- practical considerations
  - extract information with lowest possible number of experiments
  - “rescue” experiments that could not be performed according to sampling plan
  - good practice for data analysis

Analytical Process

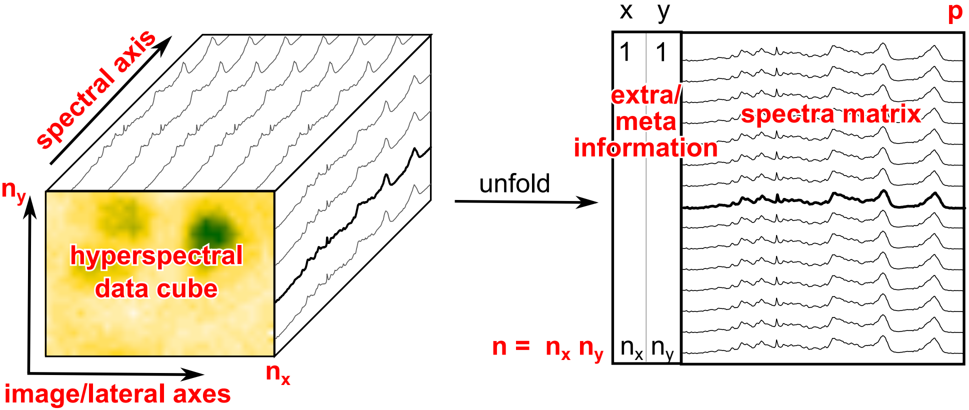
1. Design of experiment
   sampling theory
2. Measurements
3. Modeling
   - choice of model type
   - pre-processing
   - choice of model hyper-parameters
   - model fitting
4. Interpretation of the model
5. Validation: Assessment of model quality
6. Use the model

Hyperspectral Data

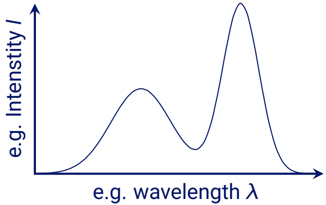


- imaging data acquired with high spectral resolution
- p spectal bands (wavelengths, wavenumbers, frequencies)
- image domain: $n = n_x \times n_y$ pixels
- analyse image and/or spectral domain

(Hyper)spectral Data Matrix



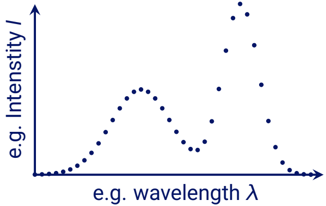
Different views on Spectra



A spectrum is a ...

Spectroscopist: continuous function of wavelength $I(\lambda)$

Different views on Spectra

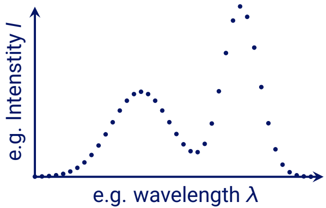


A spectrum is a ...

Spectroscopist: continuous function of wavelength $I(\lambda)$

Computer: vector of I_i at discrete wavelengths $I_i(\lambda_i)$

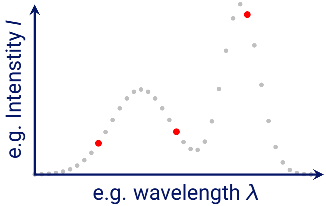
Different views on Spectra



A spectrum is a ...

Spectroscopist:	continuous function of wavelength $I(\lambda)$
Computer:	vector of I_i at discrete wavelengths $I_i(\lambda_i)$
Chemometric algorithm:	point in high dimensional space spanned by axes that are $I(\lambda_i)$

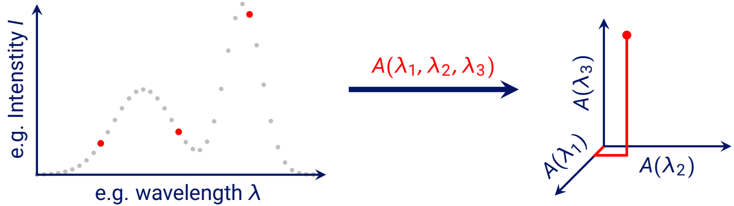
Different views on Spectra



A spectrum is a ...

Spectroscopist:	continuous function of wavelength $I(\lambda)$
Computer:	vector of I_i at discrete wavelengths $I_i(\lambda_i)$
Chemometric algorithm:	point in high dimensional space spanned by axes that are $I(\lambda_i)$

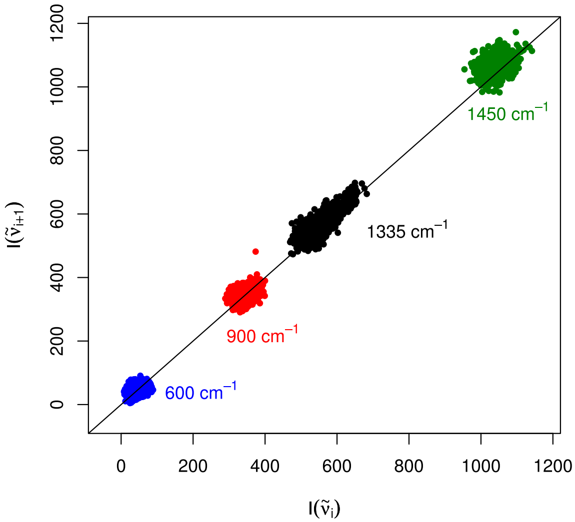
Different views on Spectra



A spectrum is a ...

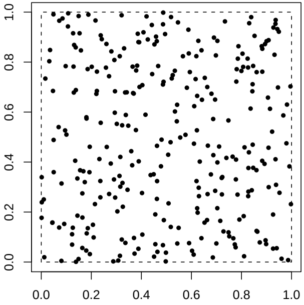
Spectroscopist:	continuous function of wavelength $I(\lambda)$
Computer:	vector of I_i at discrete wavelengths $I_i(\lambda_i)$
Chemometric algorithm:	point in high dimensional space spanned by axes that are $I(\lambda_i)$

Spectra as High-Dimensional Data



. . . are also highly correlated across wavelengths

High-dimensional Spaces

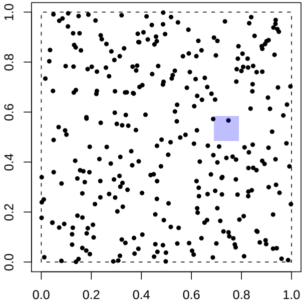


Points are far apart from each other

p-dimensional unit cube with uniformly distributed points.

p	2	10	100	1000
Distance between 2 points	0.52	1.27	4	13
edge length of cube to fill 1 % of space	0.1	0.63	0.95	0.995

High-dimensional Spaces

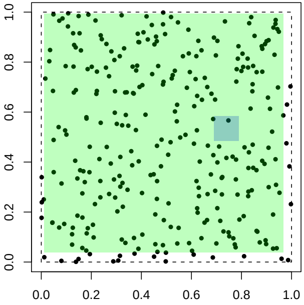


Points are far apart from each other

p-dimensional unit cube with uniformly distributed points.

p	2	10	100	1000
Distance between 2 points	0.52	1.27	4	13
edge length of cube to fill 1 % of space	0.1	0.63	0.95	0.995

High-dimensional Spaces

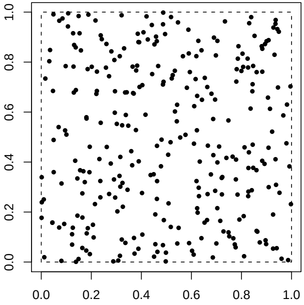


Points are far apart from each other

p-dimensional unit cube with uniformly distributed points.

p	2	10	100	1000
Distance between 2 points	0.52	1.27	4	13
edge length of cube to fill 1 % of space	0.1	0.63	0.95	0.995

High-dimensional Spaces

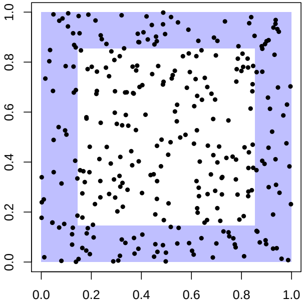


Most points are on the outside

p-dimensional unit cube with uniformly distributed points.

p	2	10	100	1000
mean distance to closest border	0.17	0.05	0.005	0.0005
outer shell width to fill 50 % of space	0.15	0.03	0.003	0.0003

High-dimensional Spaces

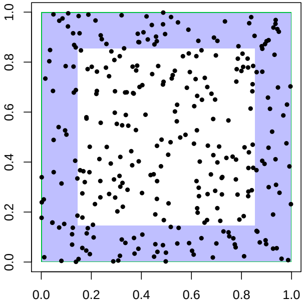


Most points are on the outside

p-dimensional unit cube with uniformly distributed points.

p	2	10	100	1000
mean distance to closest border	0.17	0.05	0.005	0.0005
outer shell width to fill 50 % of space	0.15	0.03	0.003	0.0003

High-dimensional Spaces

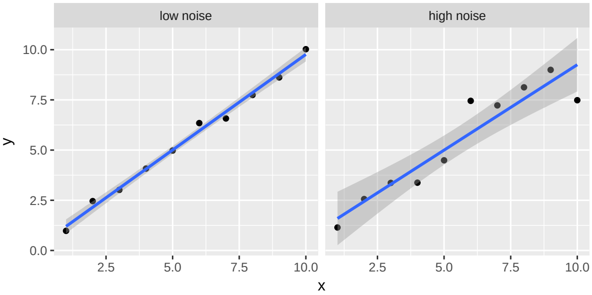


Most points are on the outside

p-dimensional unit cube with uniformly distributed points.

p	2	10	100	1000
mean distance to closest border	0.17	0.05	0.005	0.0005
outer shell width to fill 50 % of space	0.15	0.03	0.003	0.0003

Curse of Dimensionality

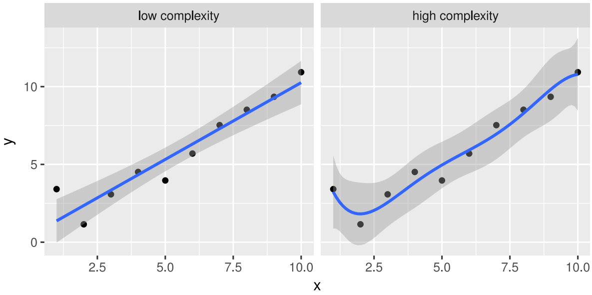


Chemometric Models are functions in *p*-dimensional space.

(Un)certainty depends on:

- noise (signal-to-noise ratio) of measurement/input variates
- degrees of freedom (no. of parameters) of the model
- sampling density

Curse of Dimensionality

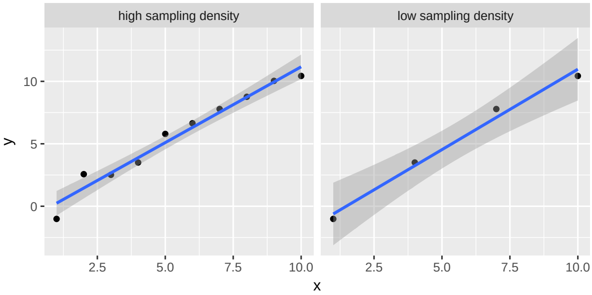


Chemometric Models are functions in *p*-dimensional space.

(Un)certainty depends on:

- noise (signal-to-noise ratio) of measurement/input variates
- degrees of freedom (no. of parameters) of the model
- sampling density

Curse of Dimensionality



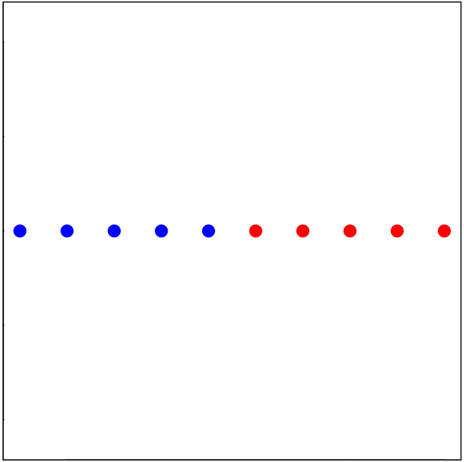
Chemometric Models are functions in *p*-dimensional space.

(Un)certainty depends on:

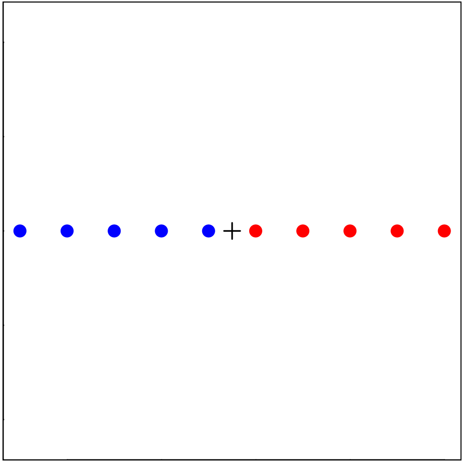
- noise (signal-to-noise ratio) of measurement/input variates
- degrees of freedom (no. of parameters) of the model
- sampling density

  constant sample density: $n \sim \left(\frac{1}{density}\right)^p$ exponential with p

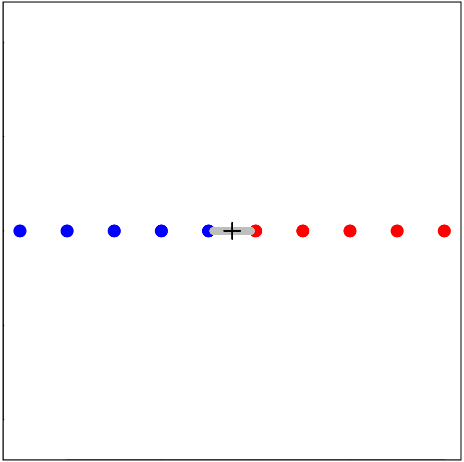
Curse of Dimensionality



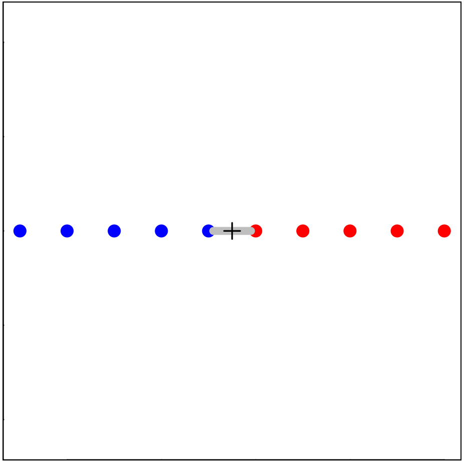
Curse of Dimensionality

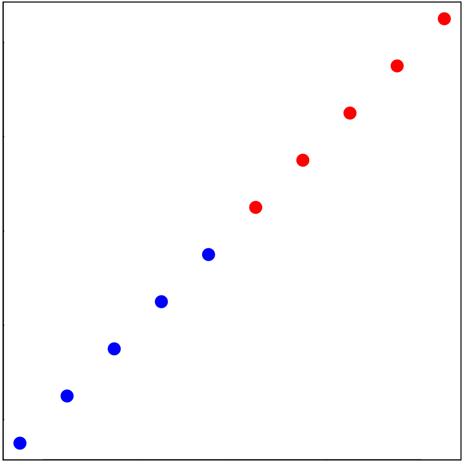


Curse of Dimensionality

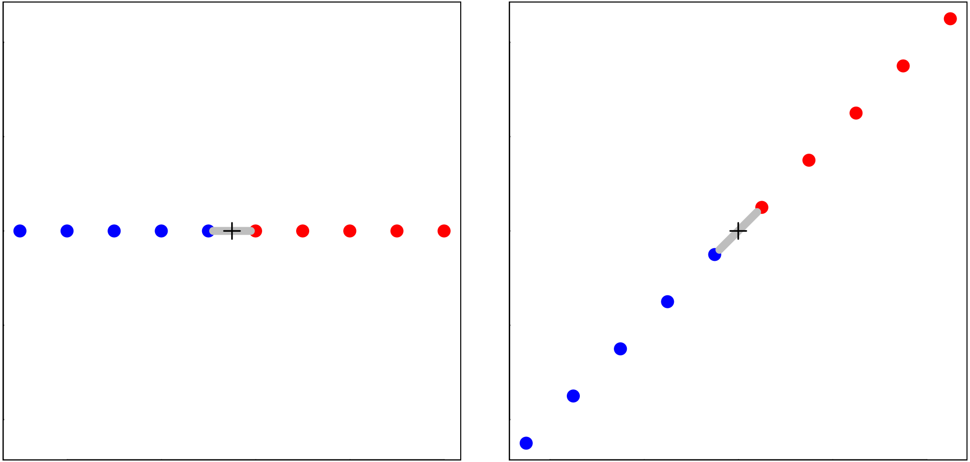


Curse of Dimensionality

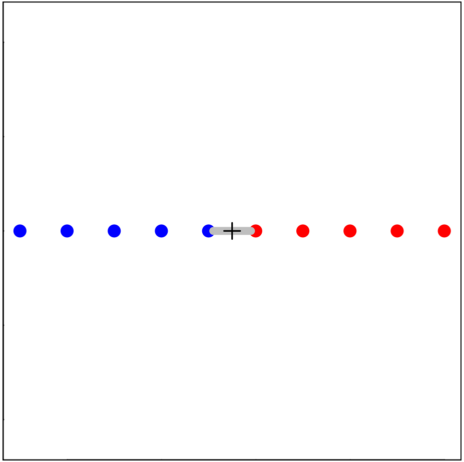


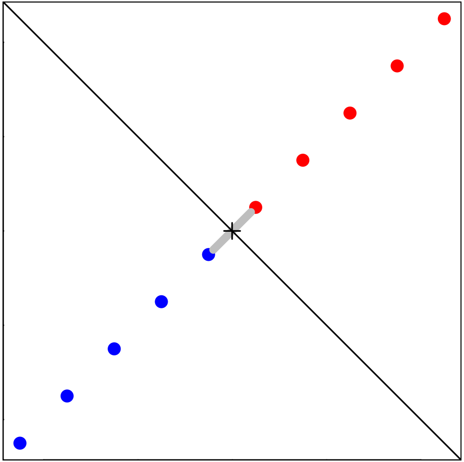


Curse of Dimensionality

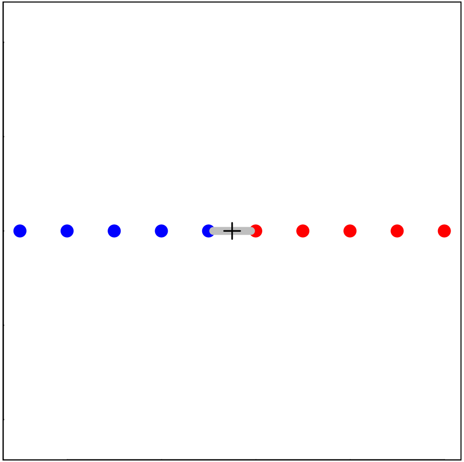


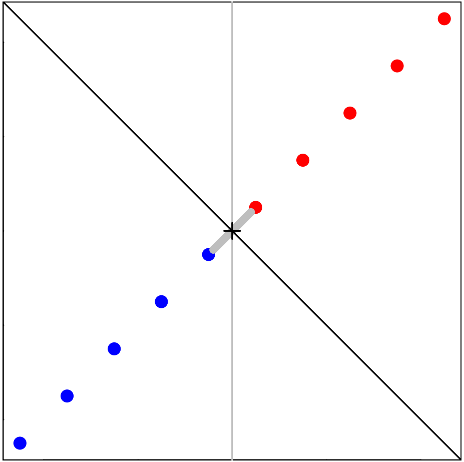
Curse of Dimensionality



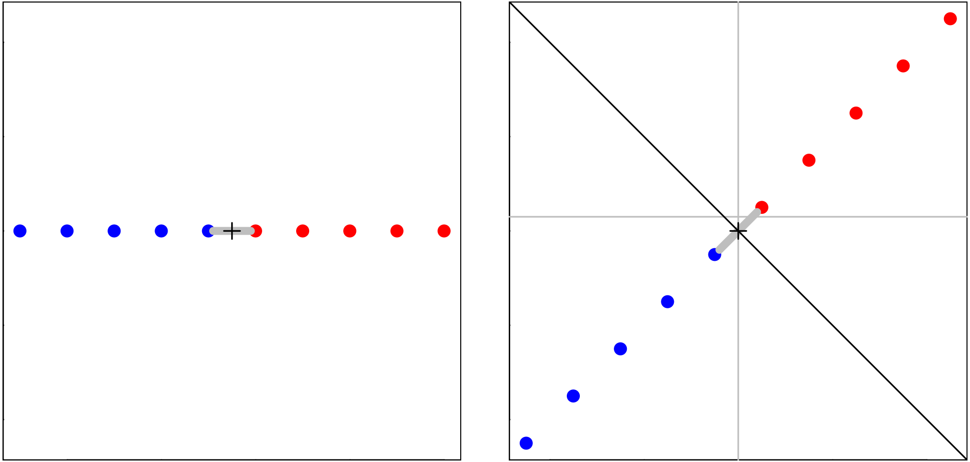


Curse of Dimensionality

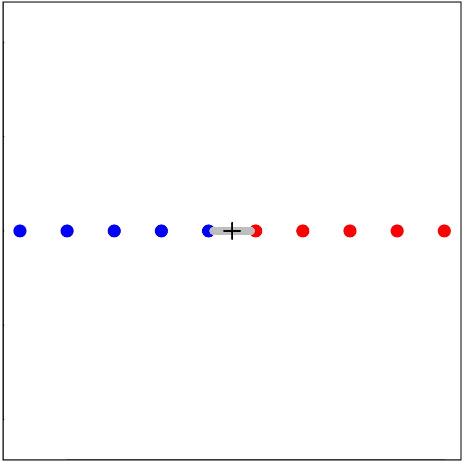


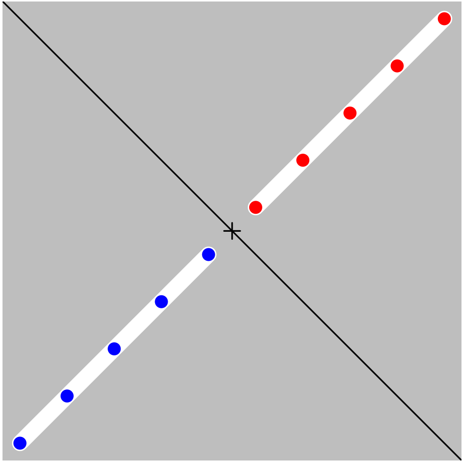


Curse of Dimensionality

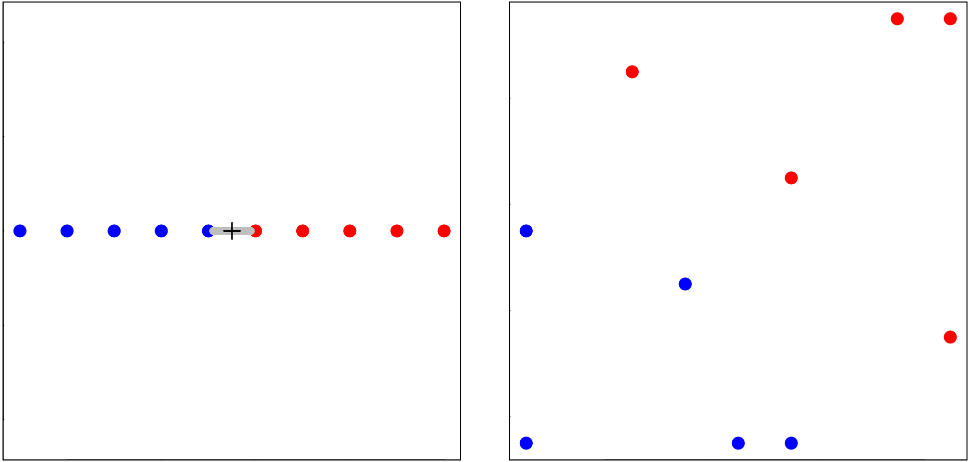


Curse of Dimensionality

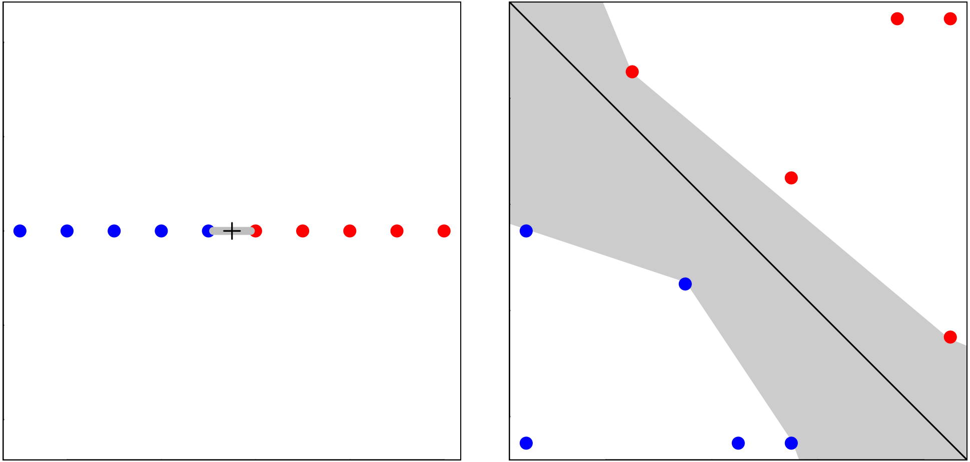




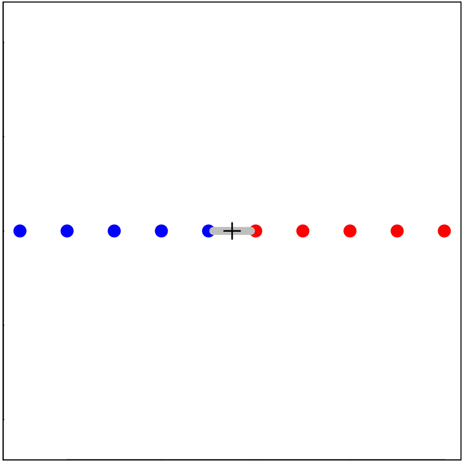
Curse of Dimensionality

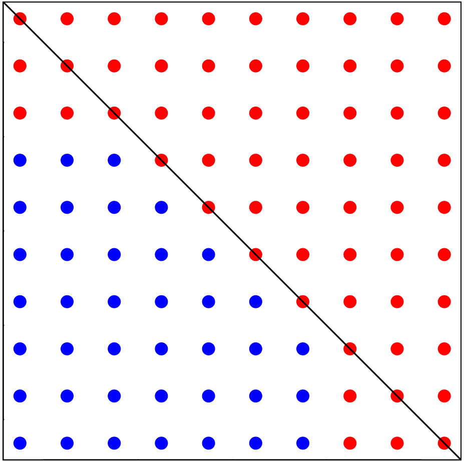


Curse of Dimensionality

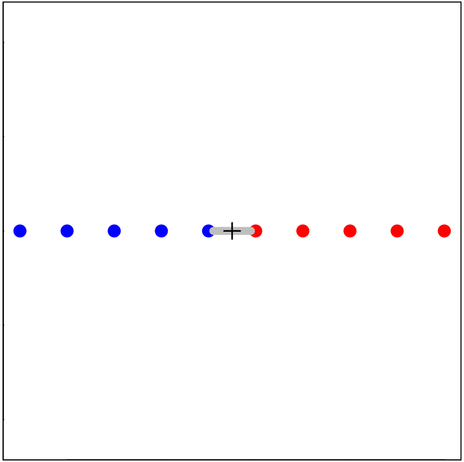


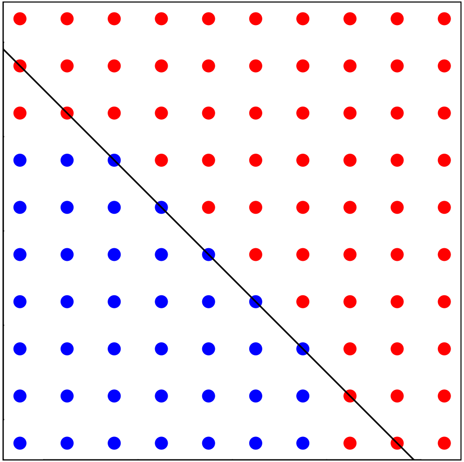
Curse of Dimensionality



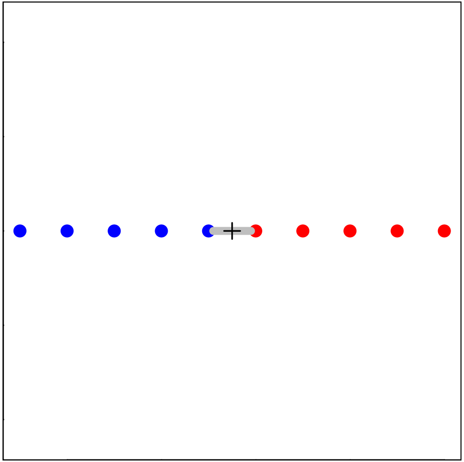


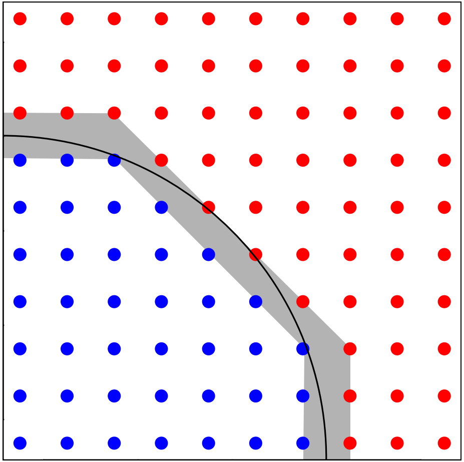
Curse of Dimensionality



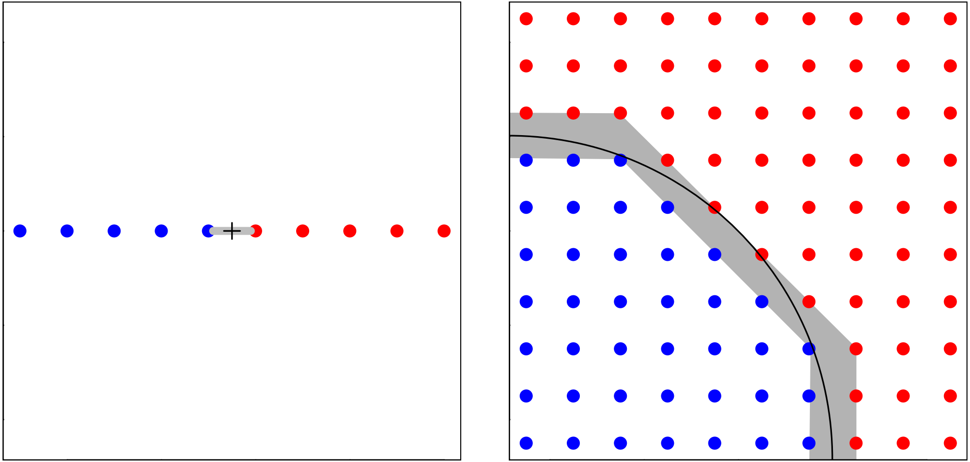


Curse of Dimensionality





Curse of Dimensionality



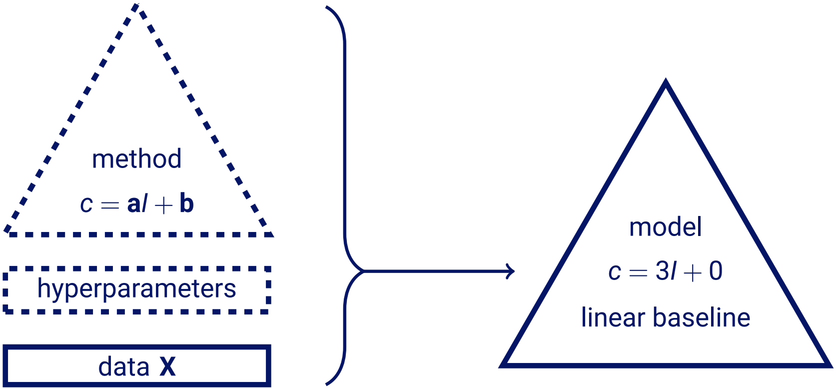
Need exponentialy increasing sample number with growing *p*!

Recommended sample size

- in general (i.e. non-linear model): $n \sim \left(\frac{1}{density}\right)^p$
- univariate linear regression: 10 samples
- multivariate linear model: 3 – 5 independent samples / input variate
- linear classifiers: $5p$ independent samples / class
  IR spectrum 600 – 1800 cm^{-1}:
  p = 601 data points ⇒ 3005 samples (patients) / class
- How complex a model can we afford?

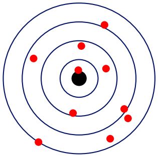
⇝ Always measure stability of the model.

Chemometric Models

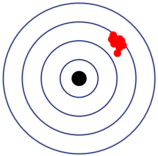


Analogy: Think of chemometric model like an analytical instrument, or the part of an instrument.

Uncertainty



Random Error
Variance type Uncertainty
low Precision



Systematic Error
Bias type Uncertainty
low Accuracy

- Spectra are subject to bias & variance
- Models are subject to bias & variance
- Predictions are subject to bias & variance

No Free Lunch Theorem

No Free Lunch in Search and Optimization

No classification method is better than any other compared across all possible problems.

For each problem that is efficiently optimized by a given heuristic, many related functions can be derived where the heuristic is bad.

- Search space = model complexity: spectra space + hyperparameters

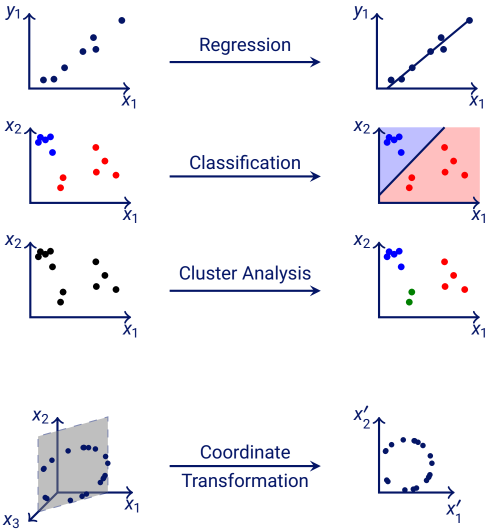
⇝ Data-driven optimization is expensive in sample size

- **Consequence:** Method needs to be adapted to problem
- External knowledge makes modeling successful
  - Domain knowledge about application/task/problem
    biological, chemical, physical, spectroscopic, . . .
  - Domain knowlege about data analysis
    judge model complexity, sample size requirements, experimental plans, . . .

Analytical Process

1. Design of experiment
   sampling theory
2. Measurements
3. Modeling
   - choice of model type
   - pre-processing
   - choice of model hyper-parameters
   - model fitting
4. Interpretation of the model
5. Validation: Assessment of model quality
6. Use the model

Data Analysis Methods

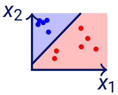


Qualitative vs. Quantitative

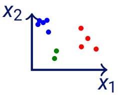
Qualitative Analysis answers a yes/no question.

- Is there any *x* in my sample?
- Does the sample express a certain property?
- Does the sample belong to a certain class/group of samples?

e.g. Classification



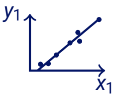
, Cluster Analysis



Quantitative Analysis quantitates properties

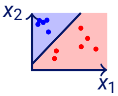
- How much *x* is in my sample?

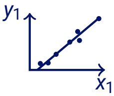
e.g. Regression



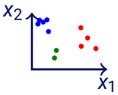
Prediction vs. Description

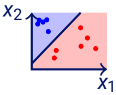
Predictive methods are used to built models that predict certain properties for new data.

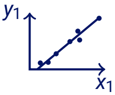




Descriptive models are used to explain a problem, its influencing factors, and so on.







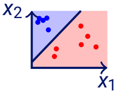
- Predictive models can also be used for description.
- Predictive models are always supervised.

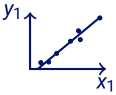
CHEMO METRIX

Supervised vs. Unsupervised

Supervised methods have a *dependent* variable and
use additional information about the particular property of interest.

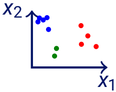
Models are built using examples with known outcome.

e.g. 

, 

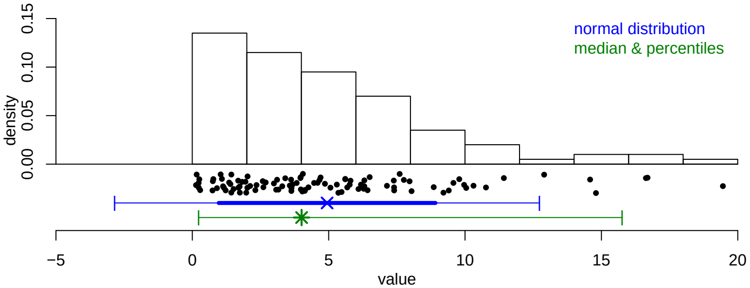
Unsupervised methods do not use additional information.

Models are built without any known true outcome.

e.g. 

, PCA, spectral unmixing

Parametric vs. Non-Parametric



Parametric models assume a particular distribution for the variates

e.g. Confidence interval calculation using $x = \bar{x} \pm t\left(1 - \frac{\alpha}{2}; n-1\right) s$
LDA: Multivariate Gaussian with equal COV for all classes

Non-Parametric models do not need any particular distribution

e.g. Confidence interval calculation using bootstrap
prediction interval using percentiles of test sample predictions
k-nearest Neighbour Classification

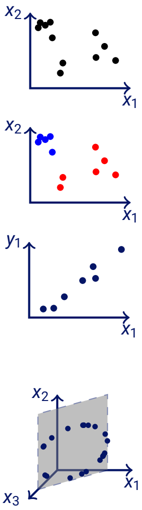
Hard vs. Soft models

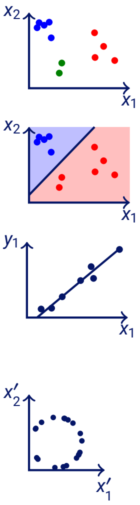
- **Hard models** are based on physical, chemical, physico-chemical, biological model
  e.g. fit reaction kinetics of pre-specified order with given pure-component spectra
- work well if assumptions met,
- but can fail badly if assumtions are violated

- **Soft models** do not use such an application background
  e.g. PLS regression
- application background enters via model interpretation
- less dependent on assumptions
- typically more uncertainty

- no sharp boundary: e.g. MCR-ALS

Data Analysis Methods

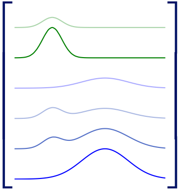




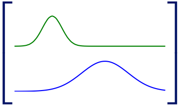
	quantitative	qualitative	predictive	descriptive	supervised
Cluster Analysis	✘	✔	✘	✔	✘
Classification	✘	✔	✔	✔	✔
Regression	✔	(✔)	✔	✔	✔
Coordinate Transformation	✔	(✔)	✔	✔	(✔)

Bilinear Models

$$A_j(\lambda) = \sum c_{ij} \quad A_i(\lambda)$$



$$= \begin{bmatrix} 0.33 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 0.33 \\ 0.33 & 0.33 \\ 0.33 & 0.67 \\ 0.00 & 1.00 \end{bmatrix}$$

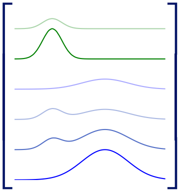


Physical background spectroscopy:

- Independence: superposition of electromagnetic waves (*linear* optics)
- Proportionality with concentration:
  - Absorption
  - Raman scattering
  - Fluorescence emission
  - Atomic emission

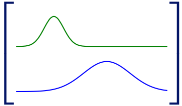
Bilinear Models

$$\mathbf{X}^{(n \times p)} = \mathbf{C}^{(n \times m)} \quad \mathbf{S}^{(m \times p)}$$



=

$$\begin{bmatrix} 0.33 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 0.33 \\ 0.33 & 0.33 \\ 0.33 & 0.67 \\ 0.00 & 1.00 \end{bmatrix}$$

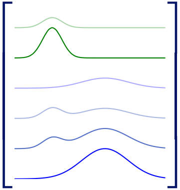


concentrations
scores

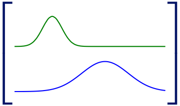
pure component spectra
loadings
latent variables/spectra

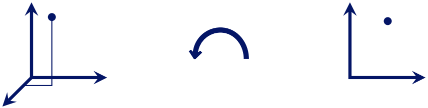
Bilinear Models

$$\mathbf{X}^{(n\times p)} = \mathbf{C}^{(n\times m)} \quad \mathbf{S}^{(m\times p)}$$



$$= \begin{bmatrix} 0.33 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 0.33 \\ 0.33 & 0.33 \\ 0.33 & 0.67 \\ 0.00 & 1.00 \end{bmatrix}$$

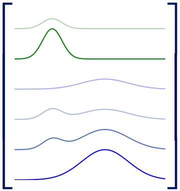




p dimensions $\gg$ m dimensions

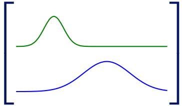
Bilinear Models

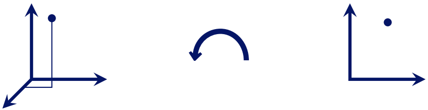
$$\mathbf{X}^{(n \times p)} = \mathbf{C}^{(n \times m)} \mathbf{S}^{(m \times p)}$$



=

$$\begin{bmatrix} 0.33 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 0.33 \\ 0.33 & 0.33 \\ 0.33 & 0.67 \\ 0.00 & 1.00 \end{bmatrix}$$





p dimensions $\gg$ m dimensions

- $p > 1$: multivariate model
- $m_{pred} > 1$: multianalyte model

Vocabulary: (Bi)linear Models

consider $\mathbf{Y} = \mathbf{B} \times \mathbf{X}$

$A = \varepsilon \cdot c$

with concentation $\mathbf{X}$

coefficient $\mathbf{B}$

signal intensity $\mathbf{Y}$

Chemist: linear in concentration c ($\mathbf{X}$) $\rightsquigarrow$ *linear model*

Statistician: linear in coefficients $\mathbf{B}$ (ε) $\rightsquigarrow$ *linear model*

Chemometrician: linear in both concentration c ($\mathbf{X}$) and coefficients $\mathbf{B}$ (ε)

$\rightsquigarrow$ *bilinear model*

Summary

- Spectra matrix + meta-data
- Spectra as points in high-dimensional space:
  Beware of curse of dimensionality
- No free lunch: data-driven optimization is costly
  Use external knowledge
- Method classes: Cluster analysis, Classification, Regression, . . .
- Bilinear models: strong physical background (linear spectroscopy)
- Bias- & variance-type uncertainty lurk everywhere
- Chemometrics on whole spectra: Variance is the danger
  Easy to overlook, often the main contributor to the total error
- Chemometrics on few wavelengths: much better situation
  Elephant in the room: wavelength selection