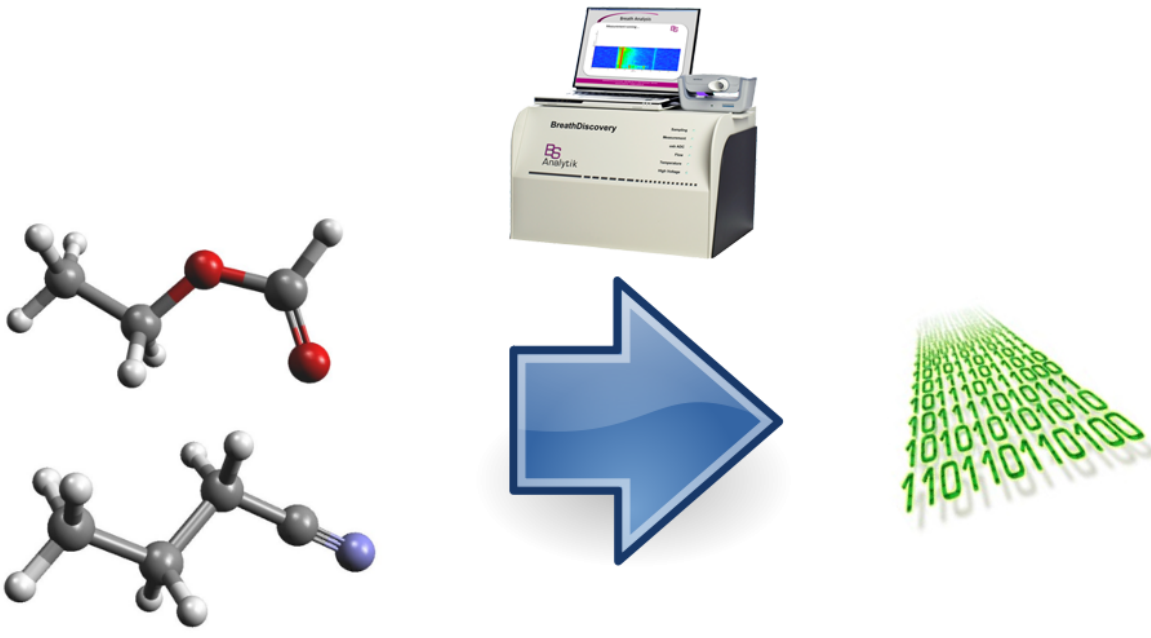
Peak Modeling for Ion Mobility Spectrometry Measurements

Dominik Kopczynski

Collaborative Research Center SFB 876,
Computer Science XI, TU Dortmund, Germany

September 12, 2012

Ion mobility spectrometry (IMS)



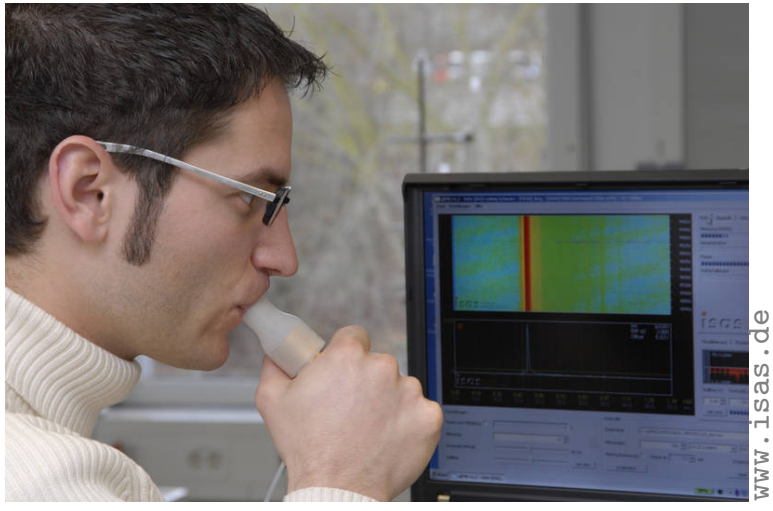
IMS measurement

IMS functionality

- Analyzes concentration of compounds in air
- Works with ambient pressure in contrast to mass spectrometry
- Measurement takes $\approx 25\,\text{ms}$

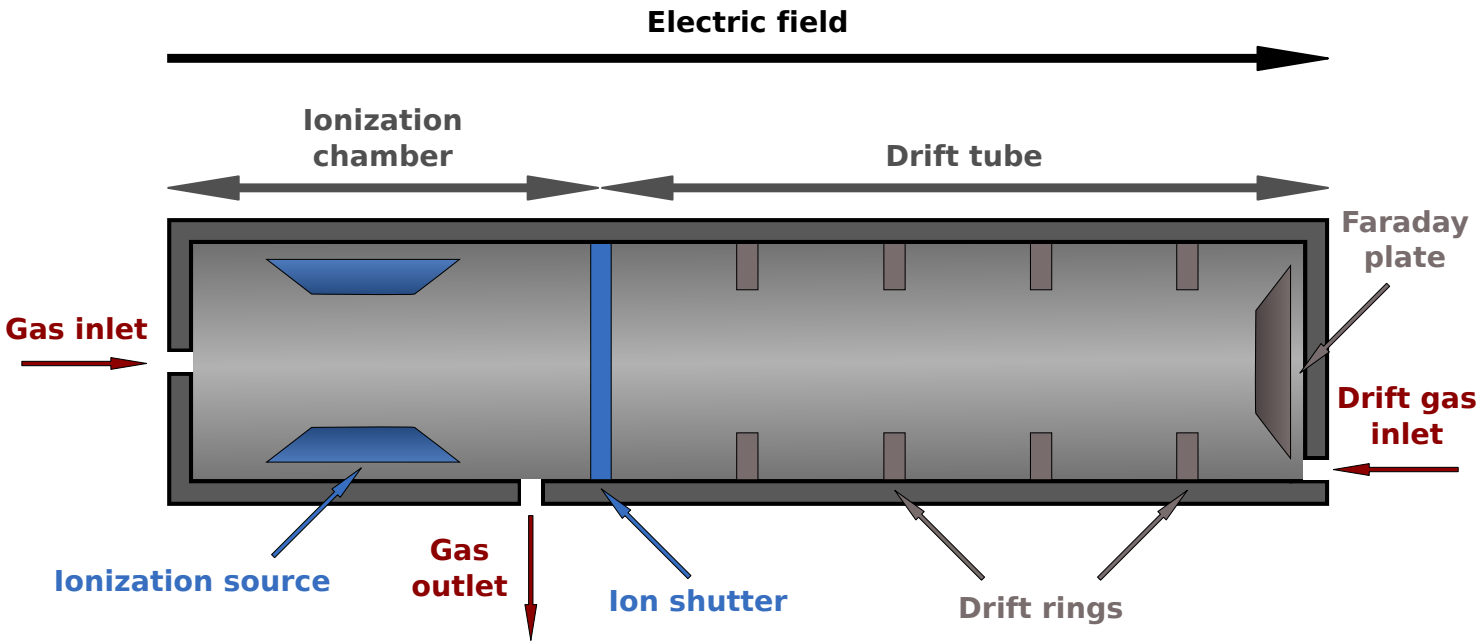
Application areas

Detecting drugs and explosives in airports



Breath gas analysis in medicine *(main reasearch)*

Cross section of IMS device



Multi-capillary column

MCC functionality

- About 1 200 columns
- Gel-coated on inside
- Mobile phase: molecules pushed by carrier gas
- Stationary phase: molecules stick to gel
- Time to cross: retention time (in s)

Coupling MCC with IMS device

After 0 sec:

- MCC: dimension in retention time (in s)
- IMS: dimension in drift time (in ms)

Coupling MCC with IMS device

After 2 sec:

- MCC: dimension in retention time (in s)
- IMS: dimension in drift time (in ms)

Coupling MCC with IMS device

After 5 sec:

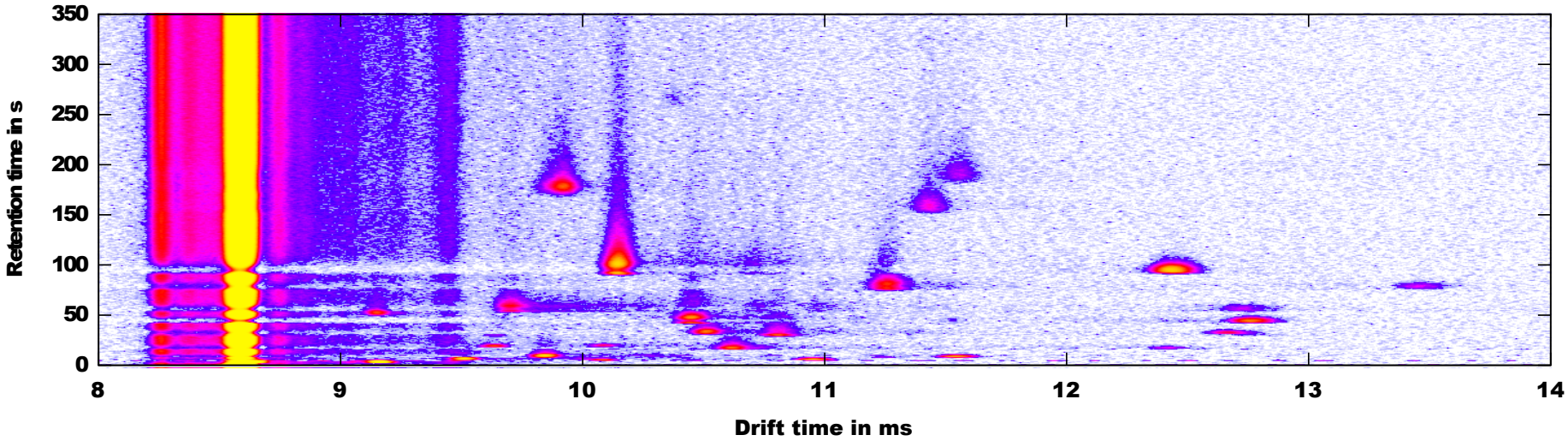
- MCC: dimension in retention time (in s)
- IMS: dimension in drift time (in ms)

Coupling MCC with IMS device

After 10 sec:

- MCC: dimension in retention time (in s)
- IMS: dimension in drift time (in ms)

Heatmap of MCC/IMS measurement



- About 3 000 000 data points
- Peak location: compound type
- Peak intensity: compound concentration
- Peaks are potential biomarkers

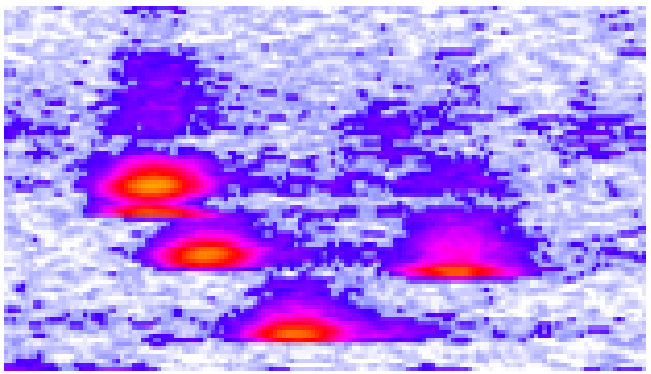
Clustering and estimating parameters

Approch

- First approach: cluster data points *(k-means)*

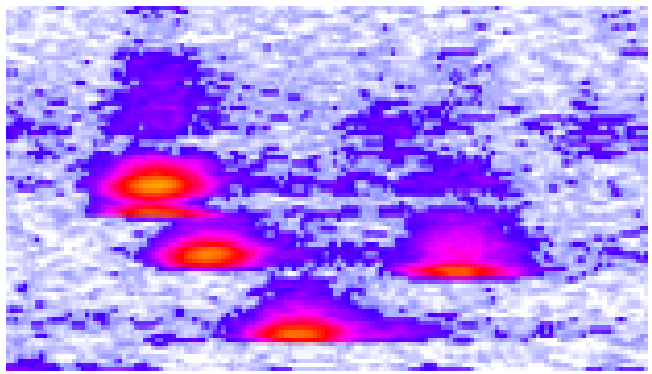
Approch

- First approach: cluster data points *(k-means)* ↯
- Problem: peaks can overlap, hard-clustering impossible



Approch

- First approach: cluster data points *(k-means)* ↯
- Problem: peaks can overlap, hard-clustering impossible



- Solution: soft-clustering with statistical models

Analysis method

Approach

- Set up likelihood function
- Use **Expectation Maximization (EM) algorithm** to maximize likelihood

Analysis method

Approach

- Set up likelihood function
- Use **Expectation Maximization (EM) algorithm** to maximize likelihood

EM algorithm

- Method to estimate parameters for statistical models
- Soft clustering $\Rightarrow$ peak decomposition
- Always finds local optimum

EM algorithm

Expectation step

- Given: $|R| \times |T|$ data points and $1 \leq j \leq c$ models
- Estimate hidden values $W_{(r,t),j}$ describing membership of every datapoint to every model

EM algorithm

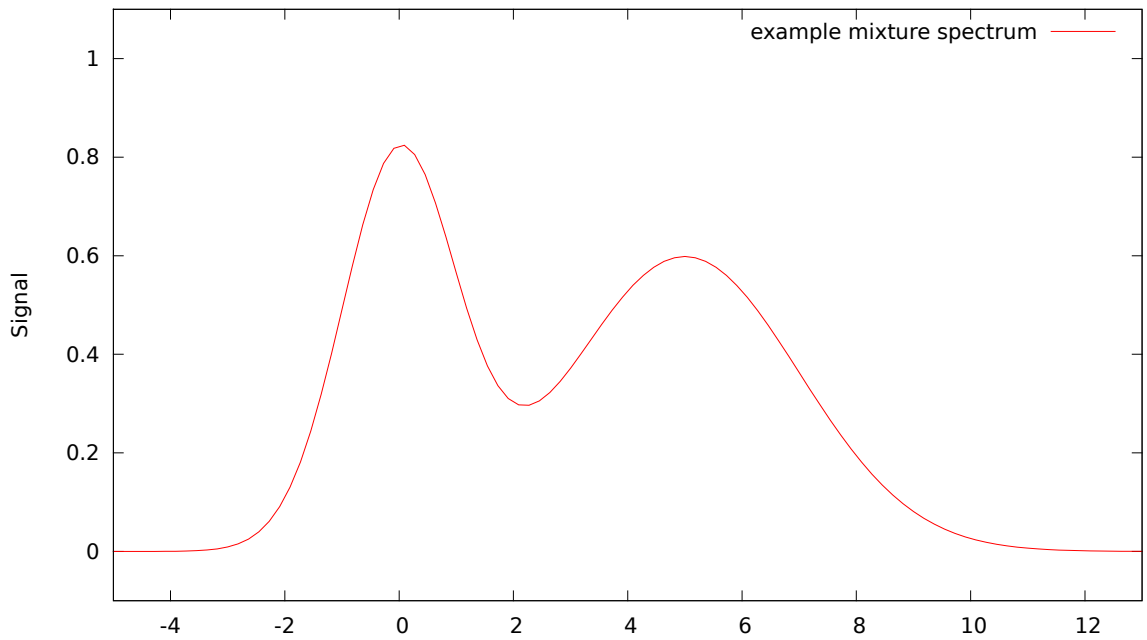
Expectation step

- Given: $|R| \times |T|$ data points and $1 \leq j \leq c$ models
- Estimate hidden values $W_{(r,t),j}$ describing membership of every datapoint to every model

Maximization step

- Estimate weight of model in relation to whole measurement
- Use maximum likelihood estimators (MLE) to estimate parameters for every model

EM example



- Given:
  - n datapoints i.e. mixture model
  - supposed statistical distribution
- Find: estimated parameters for models

Setting up ...

Notation

- s_i = signal at coordinate $x_{1 \leq i \leq n}$
- $1 \leq j \leq c$ models
- $W_{i,j}$ = membership for x_i to model j
- ω = normalized weight for model j in measurement
- Θ = parameter vector for distribution
- $P_{\Theta_j}(x_i)$ = p.d.f. of distribution

Likelihood

$$L_{x,W}(\Theta, \omega) = \prod_i^n \prod_j^c \left[\omega_j \cdot P_{\Theta_j}(x_i)\right]^{s_i \cdot W_{i,j}}$$

Setting up ... (cont'd)

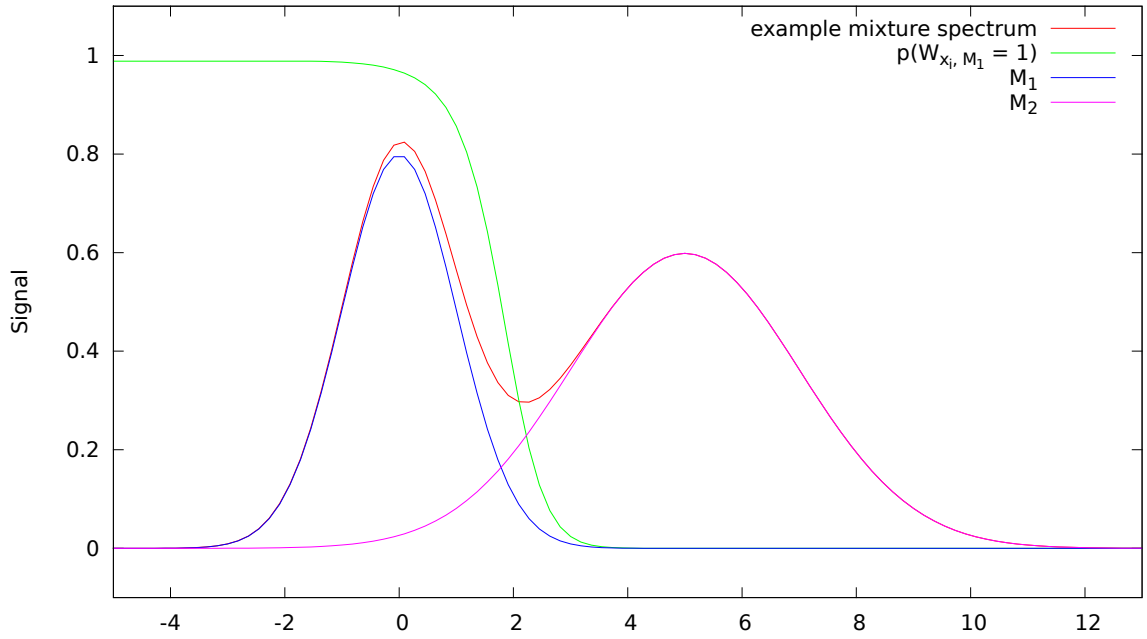
log likelihood

$$\mathcal{L}_{x,W}(\Theta,\omega) = \sum_{i}^{n} s_i \sum_{j}^{c} W_{i,j} \cdot \log\left(\omega_j \cdot P_{\Theta_j}(x_i)\right)$$

Maximum likelihood estimators

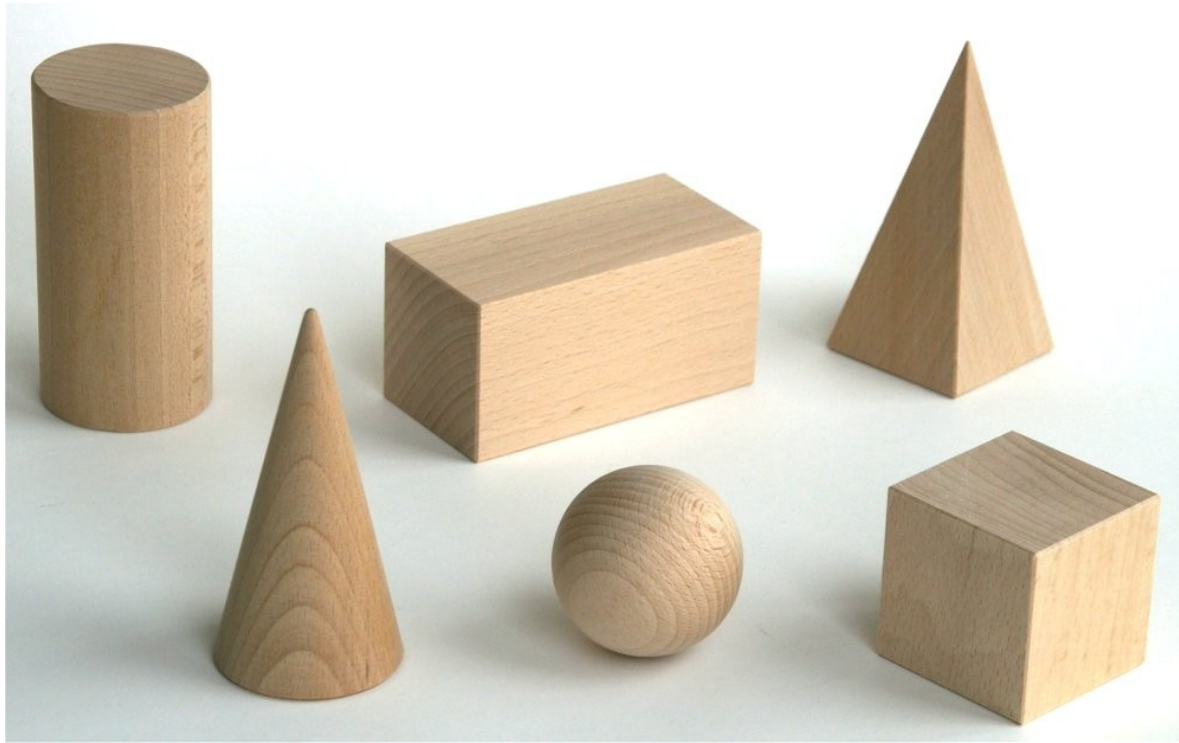
- $W^0_{i,j} = \frac{\omega_j \cdot P_{\Theta_j}(x_i)}{\sum_{k}^{c} \omega_k \cdot P_{\Theta_k}(x_i)}$
- $\omega^*_j = \frac{1}{n} \sum_{i}^{n} W^0_{i,j}$
- MLE for model parameters depend on model

EM example (cont'd)

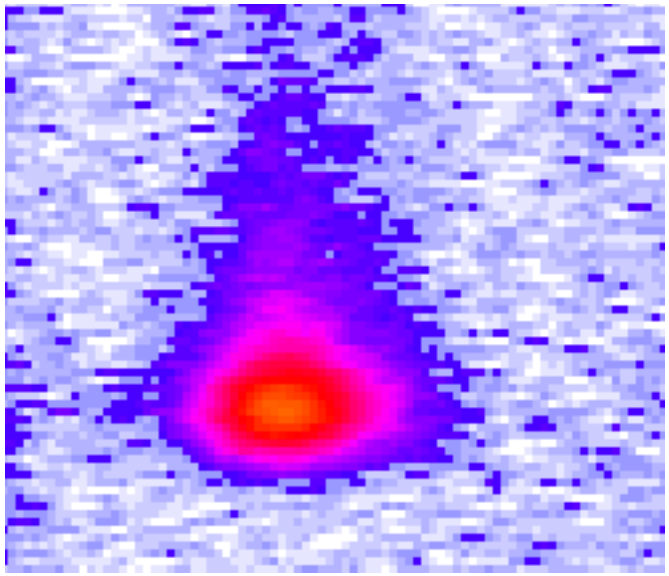


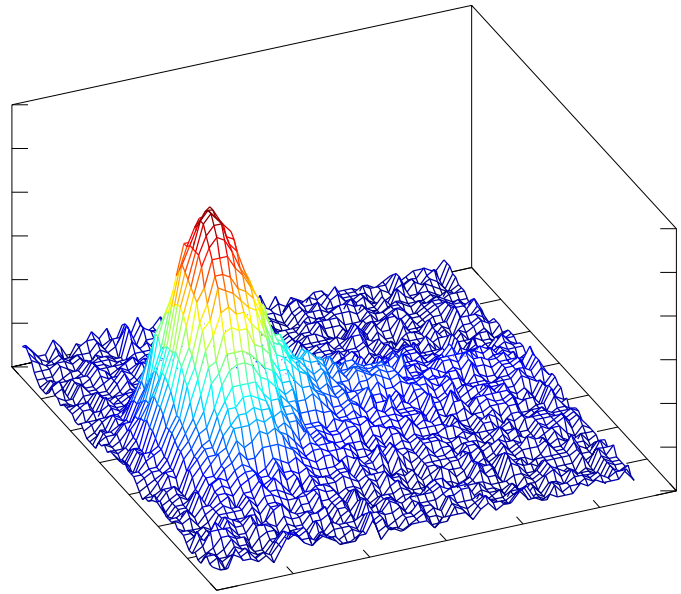
- $M_1 : \mu_1 = 0, \sigma_1 = 1, \omega_1 = 0.4$
- $M_2 : \mu_2 = 5, \sigma_2 = 2, \omega_2 = 0.6$

Describing peaks



Peaks

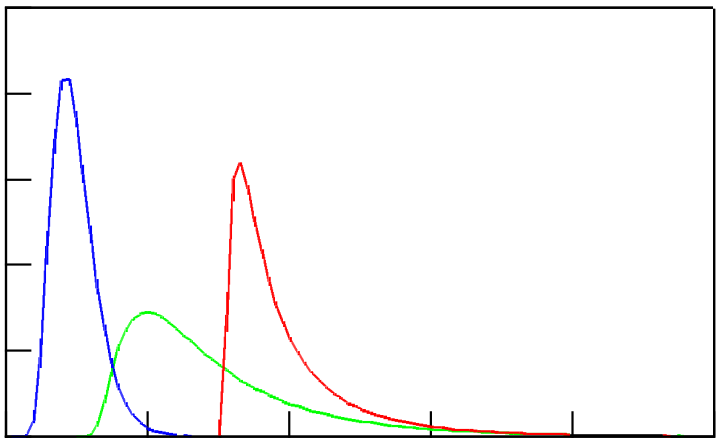




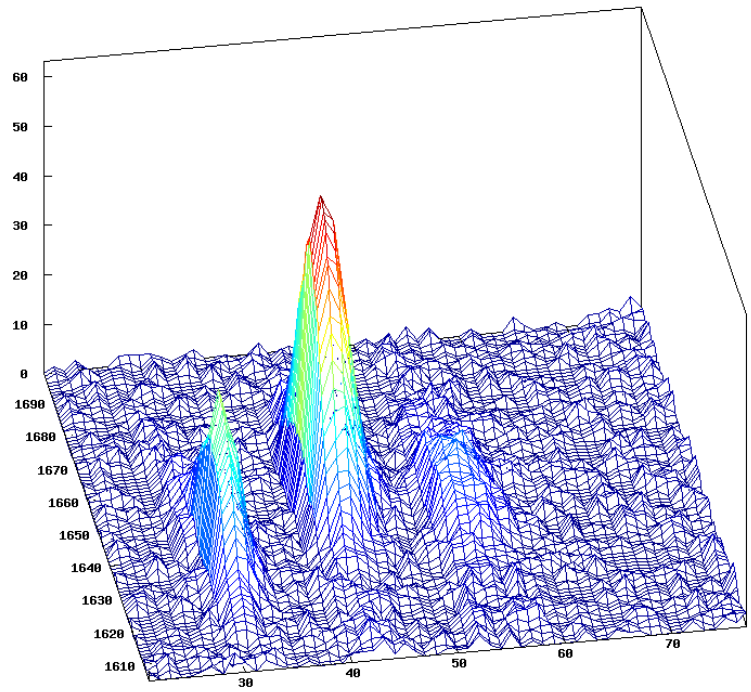
- Peaks can consist of thousands of datapoints
- Skewed in retention and drift time

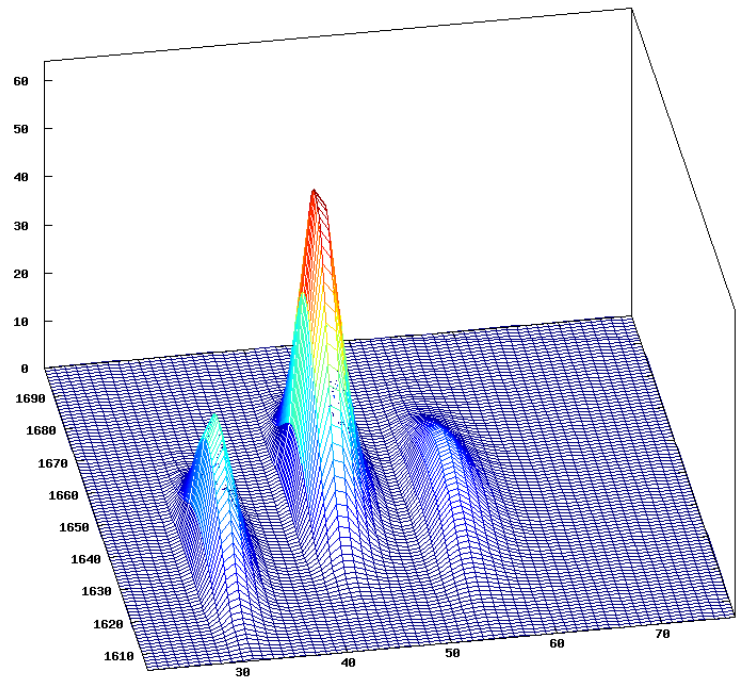
Statistical model to describe peaks

- Two-dimensional distribution model
- Inverse Gaussian distribution in retention and drift time
- $IG(x|\mu,\lambda,o) = \sqrt{\frac{\lambda}{2\pi(x-o)^3}} \exp\left(-\frac{\lambda(x-o-\mu)^2}{2\mu^2(x-o)}\right)$
- Model $M = \omega \cdot IG(\mu_r,\lambda_r,o_r) \cdot IG(\mu_t,\lambda_t,o_t)$
- Seven parameters per model
- No physical explanation, phenomenologically fits well



Example Result



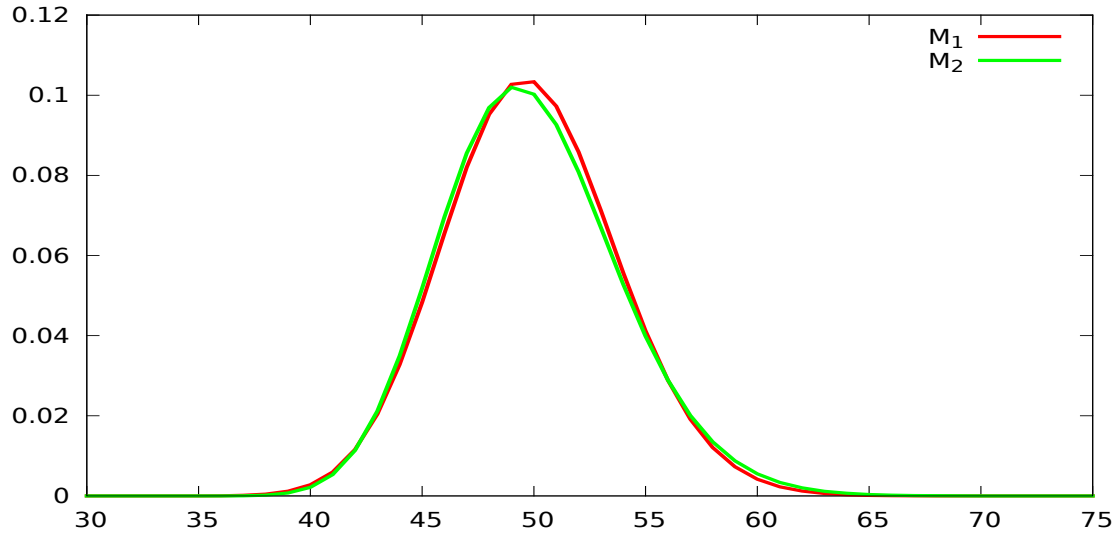


- Input: 600 datapoints
- Output: 21 parameters
- For whole measurement: data reduction factor $\approx 10\,000$

Parameter comparability

http://www.teaching-learning.eu

Model parameters not comparable



Model	μ	λ	o
M_1	71.2	24344.3	-21.3
M_2	35.2	2779.8	14.7

- Large differences caused by offset parameter o
- Search for descriptors with similar values on similar models

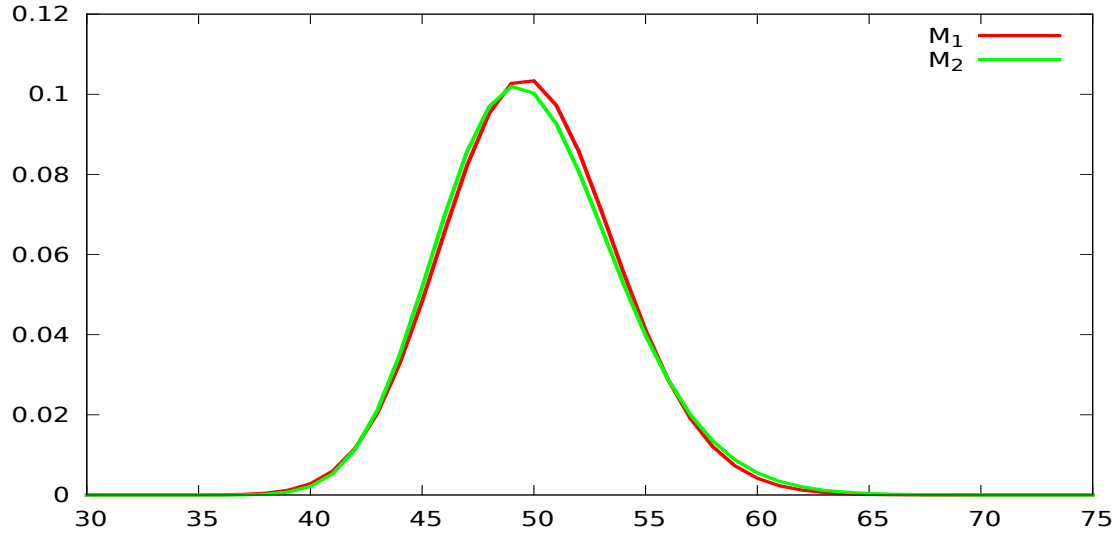
Find appropriate descriptors

Use similar descriptors instead of model parameters:

- Uncorrected mean $\mu' = \mu + o$
- Standard deviation $\sigma = \sqrt{\frac{\mu^3}{\lambda}}$
- Mode $m = \mu \left(\sqrt{1 + \frac{9\mu^2}{4\lambda^2}} - \frac{3\mu}{2\lambda} \right) + o$

Model parameters simply recomputable from descriptors

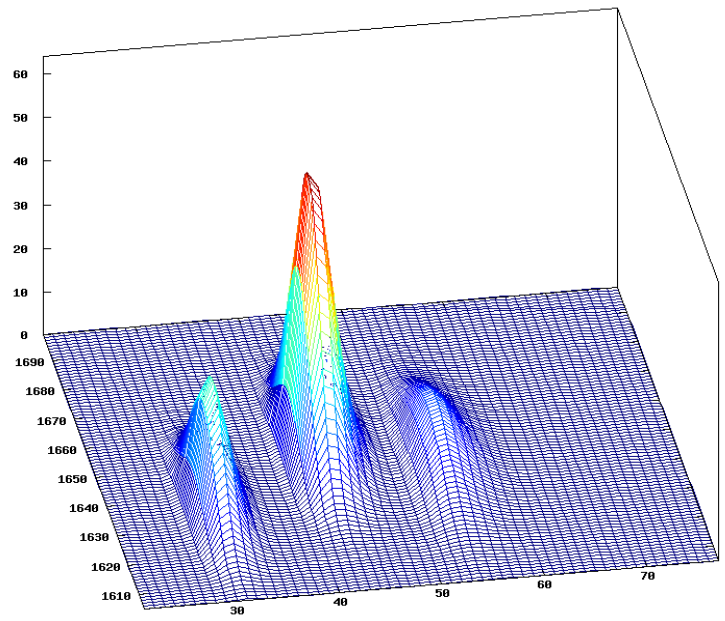
Descriptors

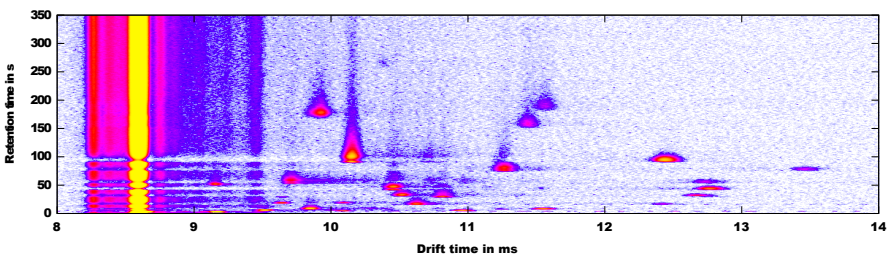


Model	μ	λ	o	μ'	σ	m
M_1	71.2	24344.3	-21.3	49.90	3.85	49.58
M_2	35.2	2779.8	14.7	49.90	3.96	49.23

Summary

- Describe peaks with statistical models
- EM algorithm for parameter estimation
- Reduction factor: $\approx 10\,000$





- Descriptors:
  - More intuitive
  - Make peaks comparable

Thanks to SFB 876 - TB1 project members

- Project leaders:
  - Sven Rahmann (Uniklinikum Essen / TU Dortmund)
  - Jörg Ingo Baumbach (KIST)
- Marianna D'Addario (TU Dortmund)
- Anne-Christin Hauschild (MPII)
- Kathrin Rupp (KIST)
- Gabriele Sprave (B&S Analytik)
- Marijan Kasunic (B&S Analytik)
- Susanne Krois (B&S Analytik)