In [1]:
```
## settings ↔
```

\usepackage*amssymb*

4.4. Optimization methods for Clustering

- Idea: derive clustering from an optimization (minimization) of a suitable cost function for clustering into K clusters.
- Success depends on the definition of a global optimality criterion
- The usually applied criteria are derived from the covariance matrix of the data (resp. of the cluster or cluster centers).

Global data covariance matrix

$$C = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{N_i} (\vec{x}_{ij} - \vec{m})(\vec{x}_{ij} - \vec{m})^\tau$$

where $N = \sum_{i=1}^{K} N_i$ und $\vec{m} = \langle \vec{x} \rangle$ ist.

Local covariance matrix (of any cluster i)

$$C_i = \frac{1}{N_i} \sum_{j=1}^{N_i} (\vec{x}_{ij} - \vec{m}_i)(\vec{x}_{ij} - \vec{m}_i)^\tau$$

Mean covariance matrix of all clusters

$$W = \frac{1}{N} \sum_{i=1}^{K} N_i C_i$$

- also called *within*-Variance

Covariance matrix of the cluster centers

$$B = \frac{1}{K} \sum_{i=1}^{K} (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^\tau$$

also called *in between*-Variance.

- Note that the origin is here the centroid $\vec{m}$ of the data set, and not the centroid of the cluster centroid vectors: the clusters could contain differntly many data points.
- matrices B, C, W are not completely independent, but coupled via $C = B + c \cdot W$, where c is a real-valued factor.

Quality criteria (resp. cost functions)

Minimization of trace(W)

$$\text{trace}(W) = \text{trace}\left(\frac{1}{N}\sum_{i,j}^{K,N_i}(\vec{x}_{ij} - \vec{m}_i)(\vec{x}_{ij} - \vec{m}_i)^\tau\right) = \sum_{i=1}^{d}\lambda_i(W)$$

where λ_i is the i-th eigenvalue of W.

- Interpretation: trace(W) is a measure for the total variance of a typical cluster, i.e. minimization corresponds to the claim for clusters with low total variance, i.e. spatial extension.
- $\Rightarrow$ this favours compact, rather spherical clusters.

Minimization of the determinant $\det(W)$

It is

$$\det(W) = \prod_{i=1}^{d}\lambda_i(W)$$

Interpretation:

- $\sqrt{\lambda_i}$ measures the mean extension of a typical cluster along the ith-eigenvector of W.
  - The determinant is thus a measure for the squared volume of a typical cluster.
- Minimizing $\det(W)$ corresponds to the demand of clusters with least volume
  - this may favour longish slim clusters
- the quality function favours partitioning into cluster of similar form (incl. orientation!)
- Since C is const, one can as well maximize $\det(C)/\det(W)$
  - Advantage: better normalization since this is a dimensionfree number.

Maximization of $\text{trace}(BW^{-1})$

Interpretation:

- In the 1D-case, one can understand that this favours simultaneously large variance of the cluster centroids and low dispersion within the clusters.
  - W is in the denominator
- Generalization to the high-dimensional case results in the matrix form of the given quality function.

- All previous optimization methods require the search of a maximal (resp. minimal) value of the quality functions E.
- This can for instance be practically implemented by a stochastic search method of Simulated Annealing with E as cost function.

4.4.1 Simulated Annealing for Clustering

Basic Idea:

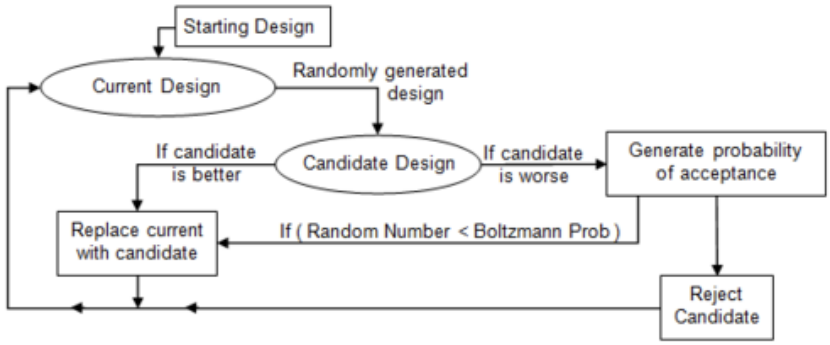

Figure from http://apmonitor.com/me575/uploads/Main/sim_annealing_flow.png (http://apmonitor.com/me575/uploads/Main/sim_annealing_flow.png)

Basic Principle:

1. random partitioning of the data set into initial clusters (e.g. heuristically)
2. compute ΔE_{ij} for each possible relabeling of an individual data element x_i into cluster C_j.
3. If $\Delta E^\star = \min_{i,j} \Delta E_{ij} < 0$, execute the according relabeling.
    - else execute the relabeling only at probability $p = \exp(-\Delta E^\star / T)$ (Boltzmann probability factor)
    - Note that T is called temperature and expresses the stochasticity of the process
4. Decrease T slightly, e.g. by setting $T = \alpha \cdot T$ with α close to but smaller than 1.
5. Interruption Condition: after a fixed number of steps (the more the better)

Remarks:

- The temperature T is slowly decreased, thus annealing.
- Alternative to the temperature, the notation $\beta = 1/T$ is used, where β is increased for annealing.
- at later time points, the acceptance is more and more limited to relabeling steps that improve the quality.
- The acceptance of steps that worsens the quality reduces the risk of getting stuck in clusterings that are only optimal under small local changes.
- The method depends on the initialization. The remedy is to repeat the algorithm with different initializations and use only the best clustering as result.

There is an incremental calculation $W^{\text{neu}} = W^{\text{alt}} + \Delta W$

- After relabeling a data point $\vec{x}_r$ from cluster o into a new cluster i requires because of $C = const \cdot W + B$ only to compute ΔW.
- Regard the change of terms of W:

$$\vec{m}_o \leftarrow \frac{n_o \vec{m}_o - \vec{x}_r}{n_o - 1} \qquad \Delta n_o = -1$$
$$\vec{m}_i \leftarrow \frac{n_i \vec{m}_i + \vec{x}_r}{n_i + 1} \qquad \Delta n_i = +1$$

  Thus we have:

$$W = \frac{1}{N} \sum_i N_i W_i$$

  with

$$W_i = \frac{1}{N_i} \sum_{j=1}^{n_i} (x_{ij} - \vec{m}_i)(x_{ij} - \vec{m}_i)^T$$

and with that:

$$\Delta W_i = (\vec{x}_r - \vec{m}_i)(\vec{x}_r - \vec{m}_i)^T \frac{n_i}{n_i + 1}$$
$$\Delta W_o = (\vec{x}_r - \vec{m}_o)(\vec{x}_r - \vec{m}_o)^T \frac{n_o}{n_o - 1}$$

- in addition, the shift of the means needs to be corrected...