In [1]:
```
## settings
import matplotlib.pyplot as plt
import numpy as np
%matplotlib qt5
import scipy, scipy.stats
plt.rcParams['figure.figsize'] = (8.0, 4.0)
from ipywidgets import interact, fixed
%matplotlib inline
```

\usepackage*amssymb*

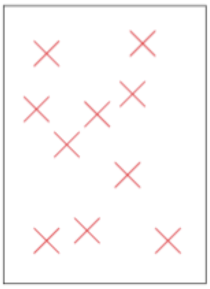
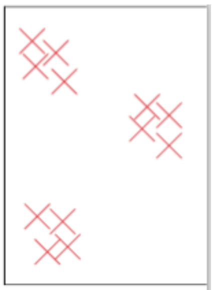
5.4. Further Projection Methods: Projection Pursuit

5.4.1. Projection Pursuit

- A simple way to find a low-dimensional projection is PCA.
- The key disadvantage is that PCA is only variance based: no other structural property is respected.

Generalization of the principle 'to maximize a criterion for well-structuredness' yields other (maybe) more meaningful projections.

- central element is the measure for structuring ('Strukturbewertungsmaß')
- Many structures manifest in the occurrence of local clustering

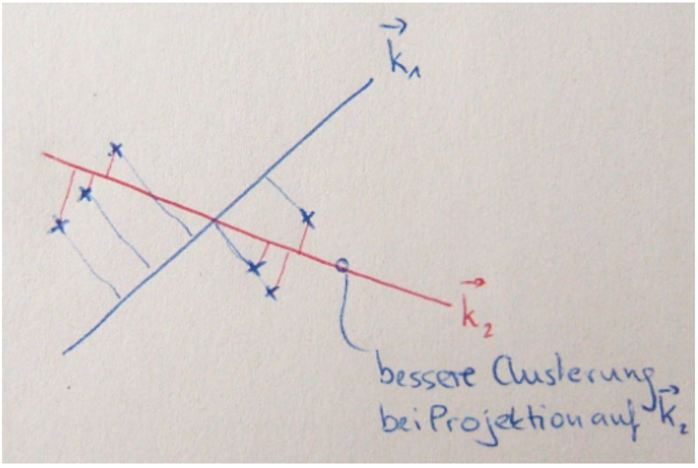



- the seemingly 'more structured' plot is characterized by
  - more short data point pair distances
  - while having the same global variance as the other plot.
- a method that is only sensitive to variance can't assess any difference between these plots.

- The preference of projections that exhibit local clustering can be expressed via a specially-crafted structure-quality measure.
- Here we first implement the idea for a projection on a 1D-axis $\hat{k}$

$$E(\hat{k}) = \underbrace{s(\hat{k})}_{\text{global dispersion along } \hat{k}} \cdot \underbrace{d(\hat{k})}_{\text{mean vicinity along } \hat{k}}$$

- by Friedman & Tukey 1974, IEEE Trans. Comp. C-23, 881ff.



- Illustration of the different structure quality of 1D-projections of a 2D dataset

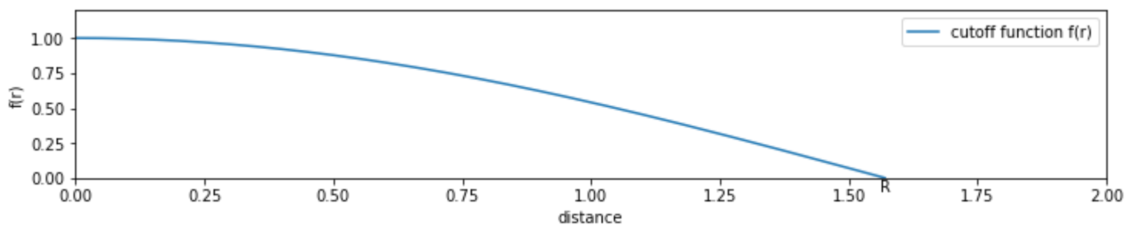
Here it is

$$s(\hat{k}) = \sqrt{\frac{\sum_{i=p\cdot N}^{(1-p)N} (\vec{r}_i \cdot \hat{k} - \langle \vec{r} \rangle \cdot \hat{k})^2}{(1-2p)N}}$$

where

$$\begin{aligned} \vec{r}_i &= i\text{-th data point} \in \mathbb{R}^d \\ &\quad \text{sorted ascending in } \vec{r}_i\hat{k} \\ N &= \text{nr. of data points} \\ p &\ll 1 \text{ for removed outliers} \\ \langle r \rangle &= \sum_{i=p\cdot N}^{(1-p)N} \frac{\vec{r}_i}{(1-2p)N} \\ d(\hat{k}) &= \sum_{\substack{i,j=1 \\ i<j}}^{N} f(r_{ij}) \cdot \Theta(R - r_{ij}) \\ r_{ij} &= |\vec{r}_i \cdot \hat{k} - \vec{r}_j \cdot \hat{k}| \end{aligned}$$

In [10]:
```
# plot f(r)↔
```



- shape of cutoff function $f(r)$ in the above equation
- The cutoff value R defines the length up to which local clusterings shall be registered.
- R is fix so that in the double sum above only $O(N \log N)$ terms contribute
  - this is done to reduce the computational effort.
- The maximization of $E(\hat{k})$ is done on the sphere surface S^{D-1}
  - generally there are many local maxima (very narrow, difficult to find)
  - good starting directions are given by PCA axes
  - because of the cut-off function $E(\hat{k})$ is relatively smooth.

Generalization to 2D-projections:

- either achieved by two-times applying the 1D method
  - i.e. the second axis is constraint to the orthogonal space of the first axis.
- or direct minimization of an analogue defined structure quality measure for two directions $\hat{k}, \hat{l}$, e.g.

$$s(\hat{k}, \hat{l}) = \sqrt{\sum (\vec{r}_i \hat{k} - \langle \vec{r} \rangle \hat{k})^2 + (\vec{r}_i \hat{l} - \langle \vec{r} \rangle \hat{l})^2}$$

and

$$d(\hat{k}, \hat{l}) = \text{ wie zuvor, jedoch } r_{ij} = \sqrt{(r_i \hat{k} - r_j \hat{k})^2 + (r_i \hat{l} - r_j \hat{l})^2}$$

5.4.2 Further propositions for projection indices

Point of departure: statistical considerations: under all probability distributions of fixed variance σ^2, the normal distribution $N(x, \sigma)$ is the least informative

- is describes for $n \to \infty$ the resulting distribution of the superposition of n uncorrelated random vectors.
- for that reason, a random projection converges always towards the normal distribution as limes $d \to \infty$
- for the normal distribution we receive the highest value for the Shannon entropy

$$S(P) = -\int P(x) \log P(x) dx$$

Conclusion:

- The interestingness of a projected probabiltiy distribution $P(x)$ can thus be measured and judged by its deviation from a normal distribution.
- w.l.o.g. the data can always be normalized so that $P(\cdot)$ has mean 0 and variance 1
- The deviation of P from a normal distribution can be measured in different ways: (using $x = \hat{k} \cdot \vec{r}$ in the following):

$$\begin{aligned}
I_1 &:= \int \frac{(P(x) - \mathcal{N}(x, 1))^2}{2\mathcal{N}(x, 1)} dx \quad \text{(Friedman (1987))} \\
I_2 &:= \int (P(x) - \mathcal{N}(x, 1))^2 dx \quad \text{(Hall)} \\
I_3 &:= \int (P(x) - \mathcal{N}(x, 1))^2 \mathcal{N}(x, 1) dx \quad \text{(Hermite Index)}
\end{aligned}$$

Interpretation:

- I_1 favors deviations in the tails of the distribution
- I_2 offers a more balanced weighting of density differences at different x
- I_3 regards deviations within the center as particularly relevant

Practical Procedure:

- Religio of integrals on expectation values by expansion of $P(x)$ in orthogonal polynoms

$$P(x) = \sum_{\nu=0}^{\infty} a_\nu H_\nu(x) \ \ (1)$$

Example: Expansion of I_3 (Hermite Index):

- In this case we choose $H_\nu(x)$ as proportional to the so-called hermite polynomials.
- At suitable scaling we can assert the following orthogonality relations:

$$\int_{-\infty}^{\infty} H_\mu(x) H_\nu(x) N(x,1) dx = \delta_{\mu\nu}$$

- Integration of (1) after multiplying with $N(x, 1)$ results in

$$a_\mu = \int P(x) H_\mu(x) N(x,1) dx$$

  and thus

$$a_\mu = \langle H_\mu(x) N(x,1) \rangle_{\text{all data}}$$

Insertion into I_3, together with the expansion of N according to

$$N(x,1) = \sum_{\mu=0}^{\infty} b_\mu H_\mu(x)$$

in Hermite polynomials, under exploitation of the orthogonality relations

$$\langle H_\mu H_\nu \rangle = \delta_{ij}$$

gives:

$$\begin{aligned}
I_3 &= \int (P(x) - N(x))^2 N(x) dx \\
&= \int \left(\left(\sum_\nu a_\nu H_\nu \right) - \left(\sum_\mu b_\mu H_\mu \right) \right)^2 N(x) dx \\
&= \int \left[\sum_{\mu\nu} a_\mu a_\nu H_\mu(x) H_\nu(x) - 2 \sum_{\mu\nu} a_\mu b_\nu H_\mu(x) H_\nu(x) \right. \\
&\quad \left. + \sum_{\mu\nu} b_\mu b_\nu H_\mu(x) H_\nu(x) \right] N(x,1) dx \\
&= \sum_\mu \left[a_\mu^2 - 2 a_\mu b_\mu + b_\mu^2 \right] = \sum_\mu [a_\mu - b_\mu]^2
\end{aligned}$$

- Practically, this series is stopped after few terms, e.g. directly after the leading term

$$I_3^0 = (a_o - b_0)^2 = (a_0 - \frac{1}{2}\sqrt{\pi})^2$$

  and

$$a_0 = \langle H_0(x) N(x,1) \rangle_P = \langle N(x,1) \rangle_P$$

- Thus the evalution is reduced to the computation of expectation values $\langle \cdot \rangle_P$ w.r.t. the probability density of the data under projection.

```
In [ ]:
```