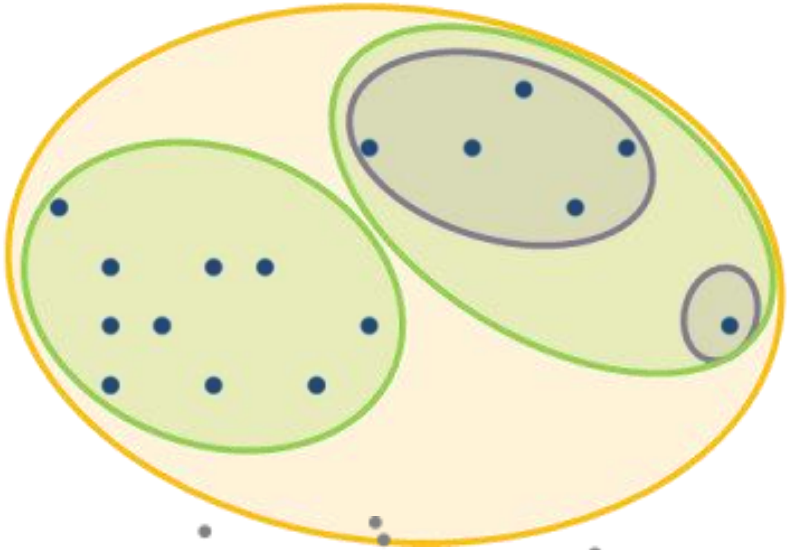
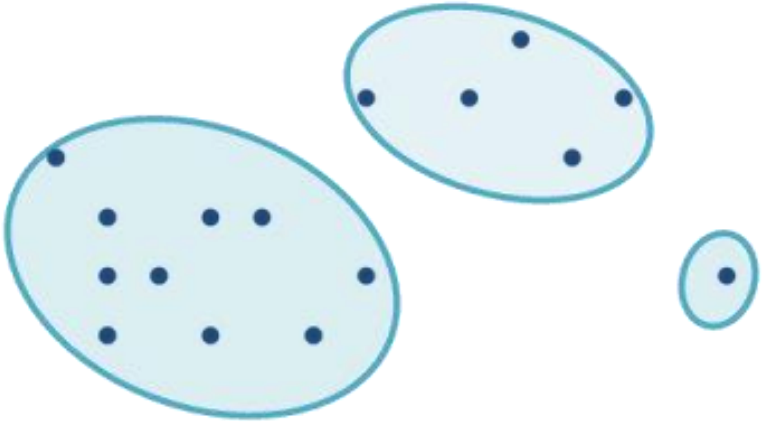
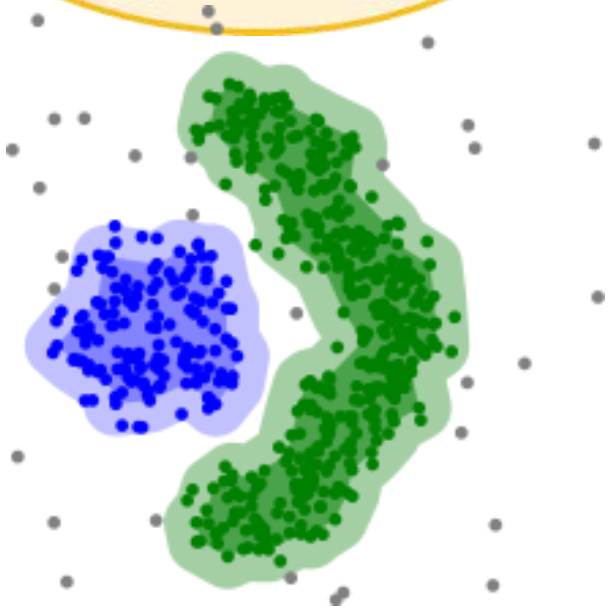
Fuzzy C-means clustering (Dembele, 2003)

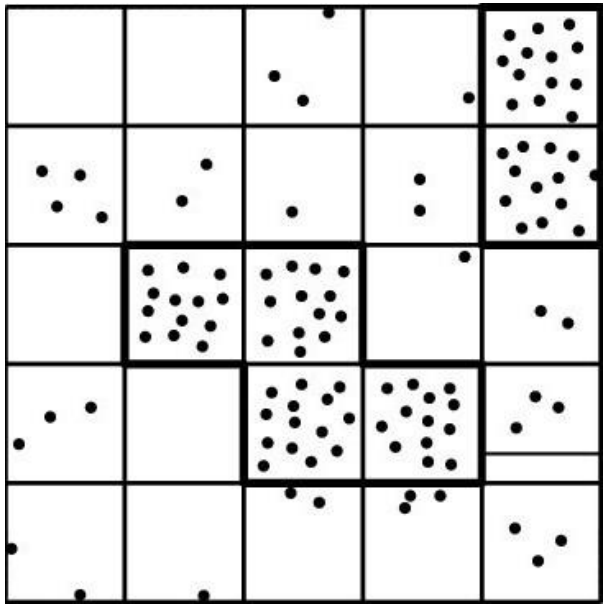
Alexander Glötzl
Algorithmische Bioinformatik
FAKULTÄT FÜR PHYSIK

Universität Regensburg









$$J(K,m) = \sum_{k=1}^{K} \sum_{i=1}^{N} (u_{ki})^m d^2(\mathbf{x}_i, \mathbf{c}_k)$$

U	Cluster 1	Cluster 2	Cluster 3
Gene 1	0.10	**0.70**	0.20
Gene 2	**0.82**	0.06	0.12
Gene 3	0.12	0.23	**0.65**
Gene 4	…		
Gene 5			

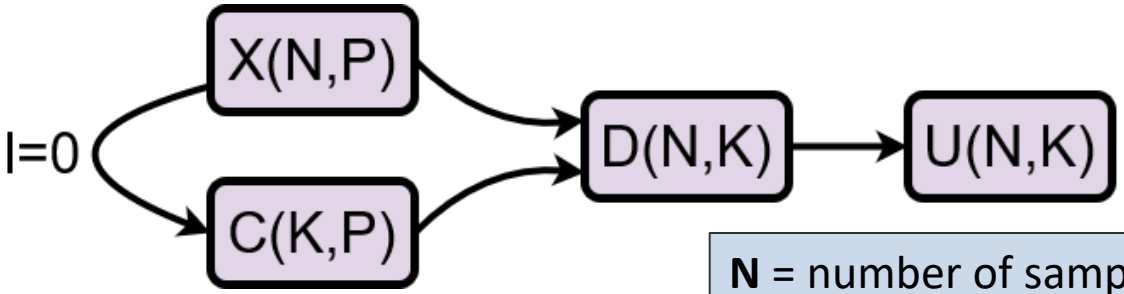
$$d^2(\mathbf{x}_i, \mathbf{c}_k) = (\mathbf{x}_i - \mathbf{c}_k)^T (\mathbf{x}_i - \mathbf{c}_k)$$

d	ΔCluster1	ΔCluster2	ΔCluster3
Gene 1	3.54	**1.10**	3.09
Gene 2	**0.98**	4.51	3.40
Gene 3	3.39	2.80	**1.22**
Gene 4	…		
Gene 5			

$$J(K,m) = \sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ki})^m d^2(\mathbf{x}_i,\mathbf{c}_k) \qquad d^2(\mathbf{x}_i,\mathbf{c}_k) = (\mathbf{x}_i - \mathbf{c}_k)^T(\mathbf{x}_i - \mathbf{c}_k)$$

$$\mathbf{c}_k^l = \frac{\sum_{i=1}^{N}(u_{ki}^{(l-1)})^m \mathbf{x}_i}{\sum_{i=1}^{N}(u_{ki}^{(l-1)})^m}; k = 1,2,\ldots,K$$

$$u_{ki}^{(l)} = \sum_{s=1}^{K}\left[\frac{d^2(\mathbf{x}_i,\mathbf{c}_k^{(l)})}{d^2(\mathbf{x}_i,\mathbf{c}_s^{(l)})}\right]^{\frac{-1}{(m-1)}}$$

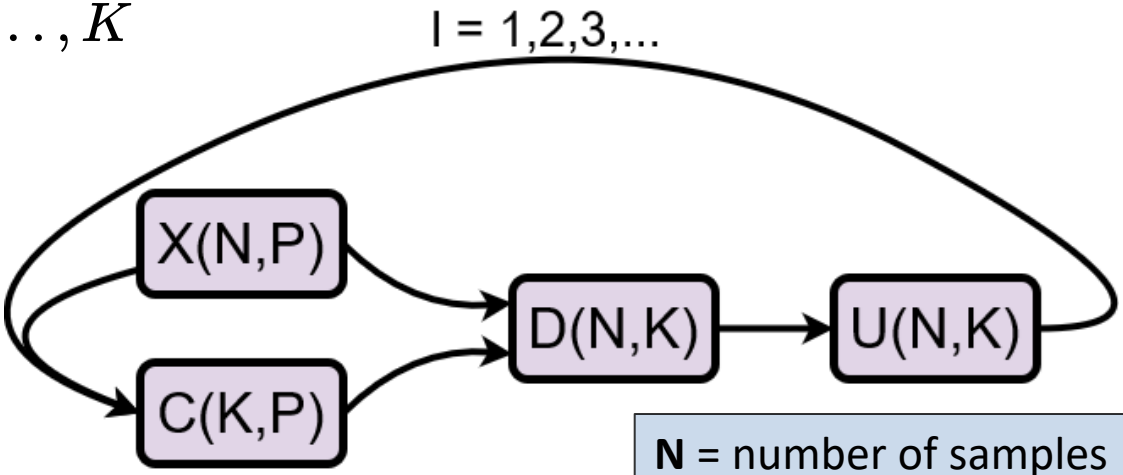


N = number of samples
P = number of features
K = number of clusters

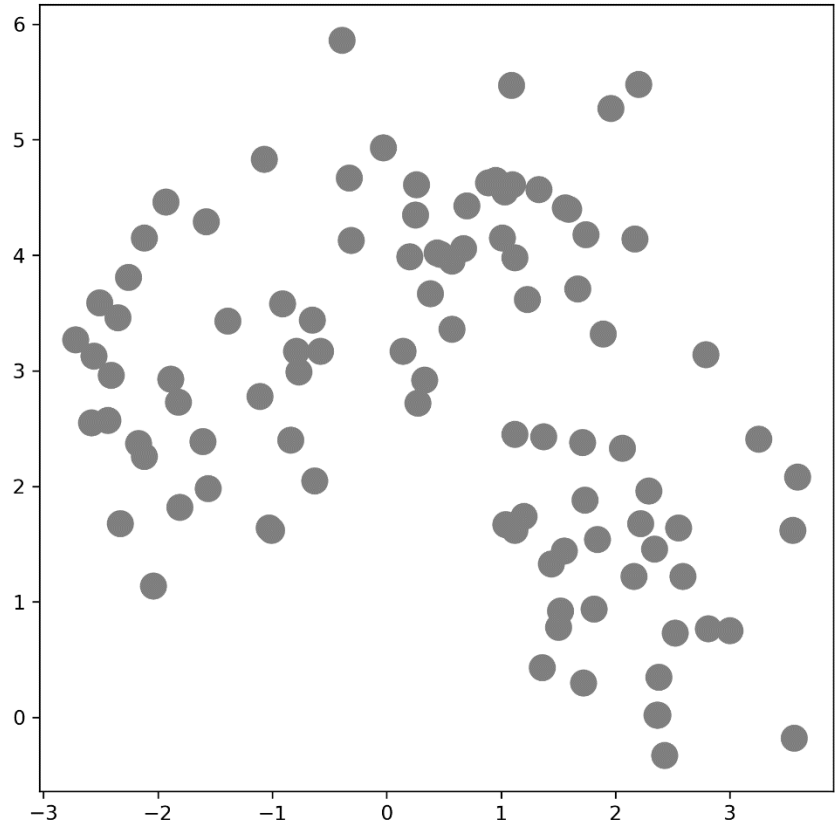
$$J(K,m) = \sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ki})^m d^2(\mathbf{x}_i,\mathbf{c}_k) \qquad d^2(\mathbf{x}_i,\mathbf{c}_k) = (\mathbf{x}_i - \mathbf{c}_k)^T(\mathbf{x}_i - \mathbf{c}_k)$$

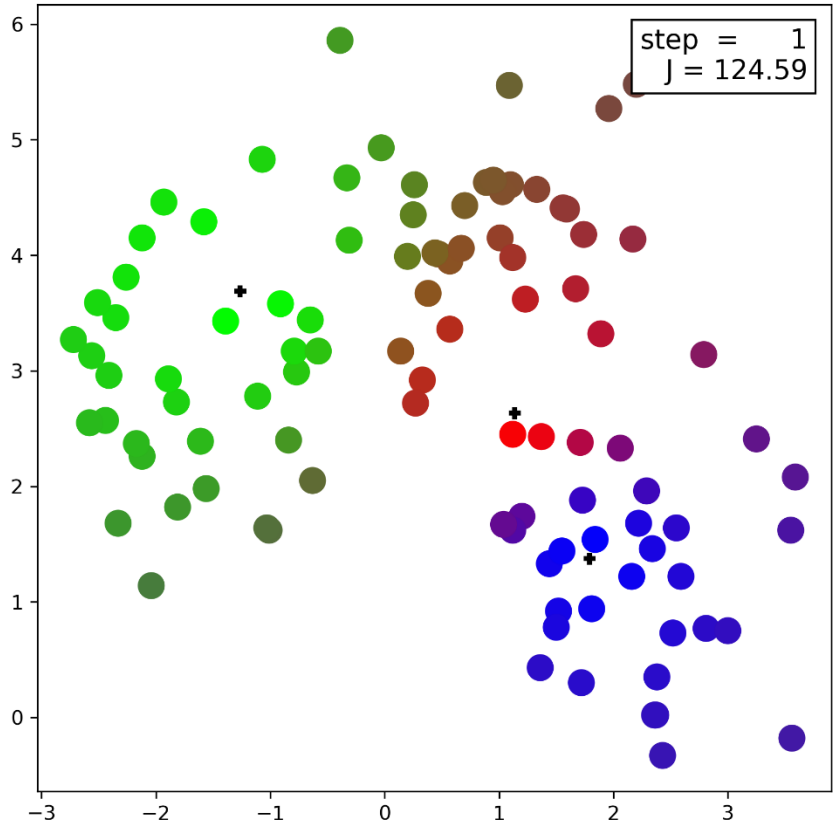
$$\mathbf{c}_k^l = \frac{\sum_{i=1}^{N}(u_{ki}^{(l-1)})^m \mathbf{x}_i}{\sum_{i=1}^{N}(u_{ki}^{(l-1)})^m}; k = 1,2,\ldots,K$$

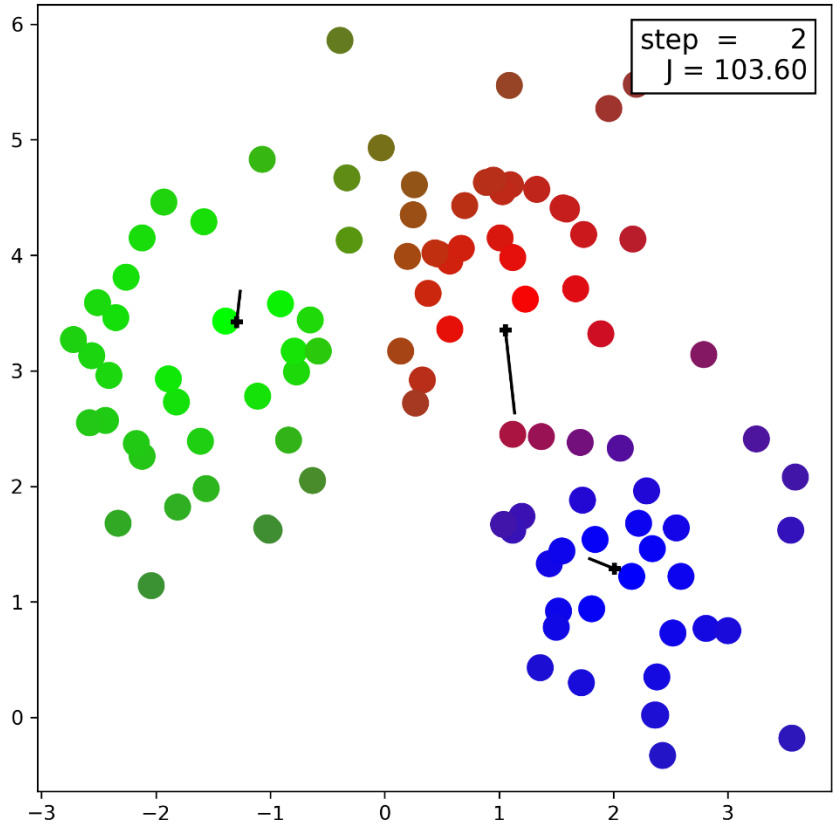
$$u_{ki}^{(l)} = \sum_{s=1}^{K}\left[\frac{d^2(\mathbf{x}_i,\mathbf{c}_k^{(l)})}{d^2(\mathbf{x}_i,\mathbf{c}_s^{(l)})}\right]^{\frac{-1}{(m-1)}}$$

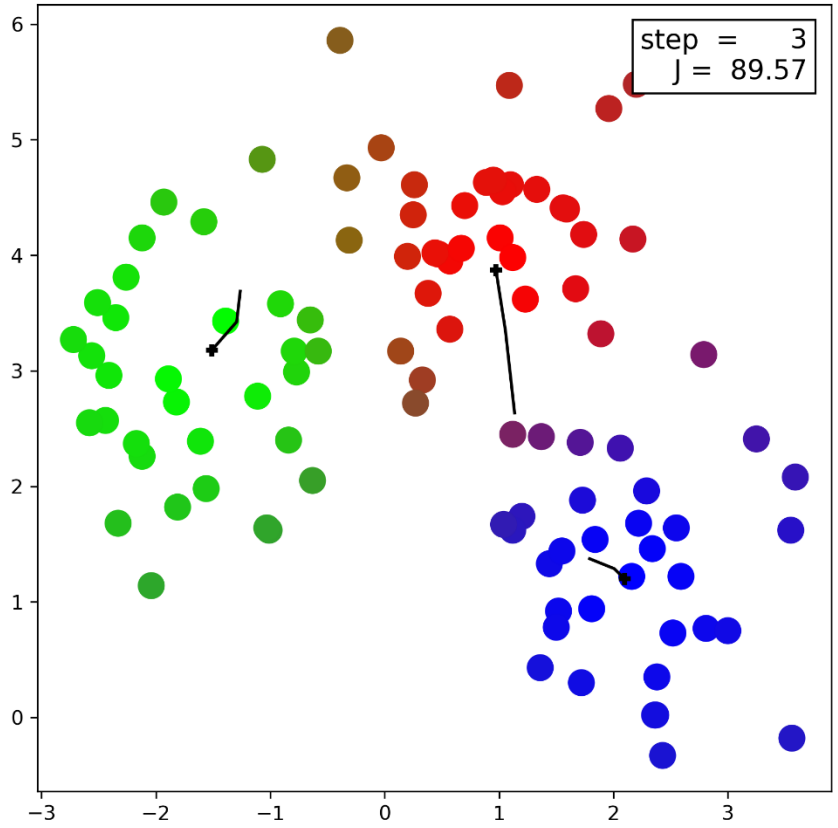


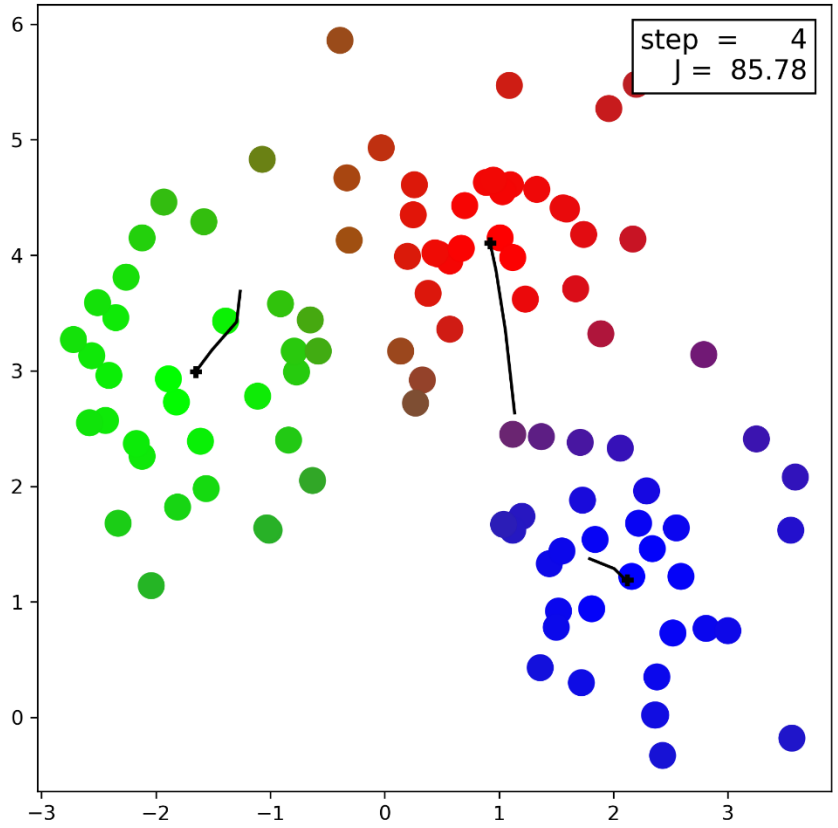
N = number of samples
P = number of features
K = number of clusters

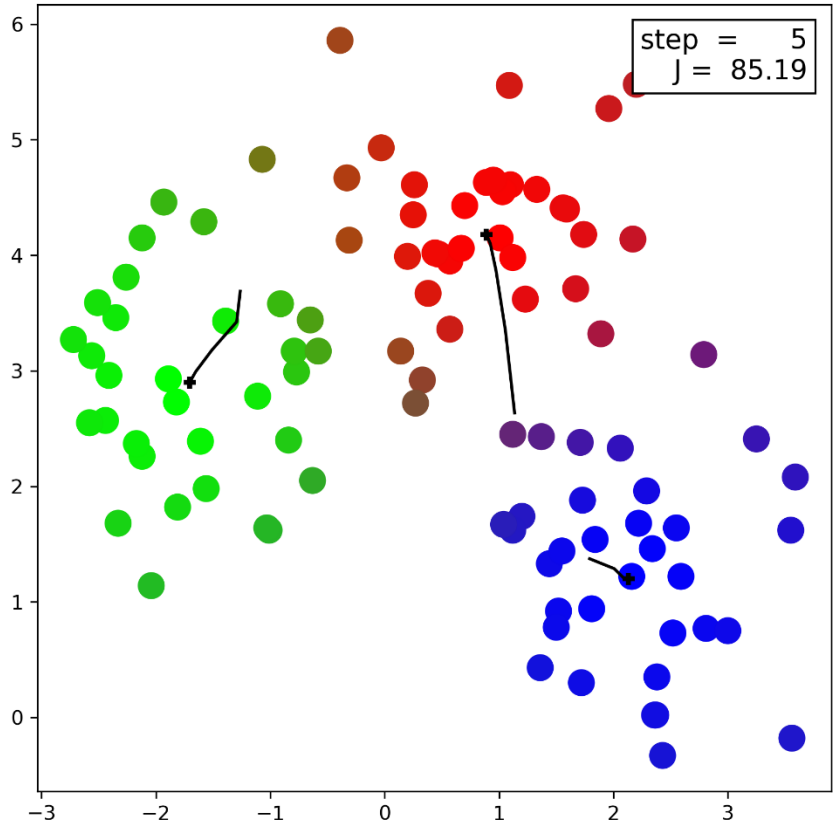

6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

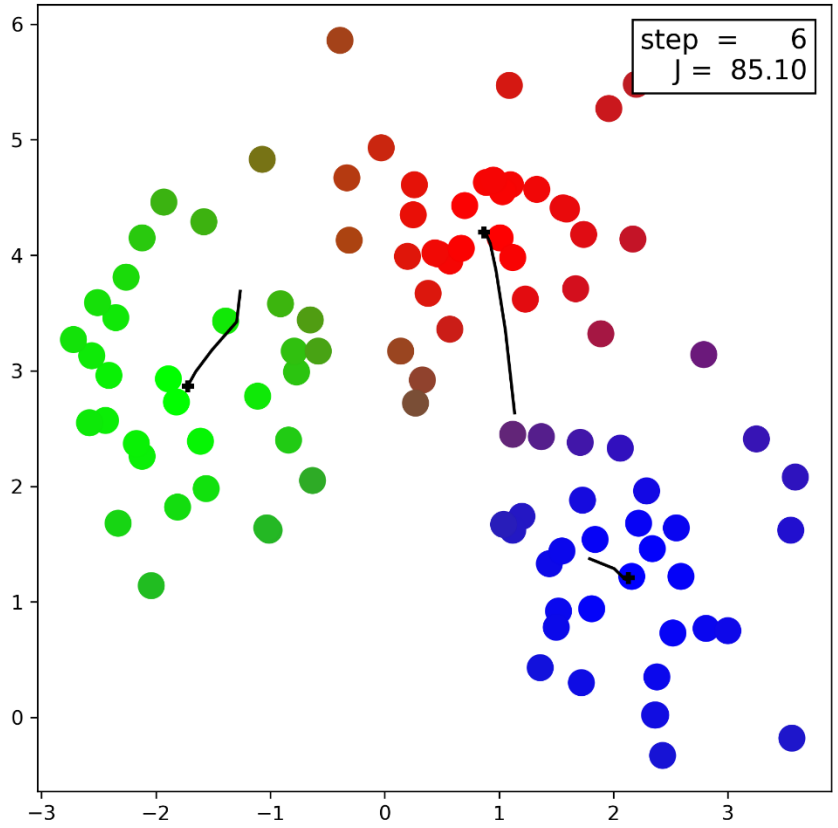

step = 1
J = 124.59
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

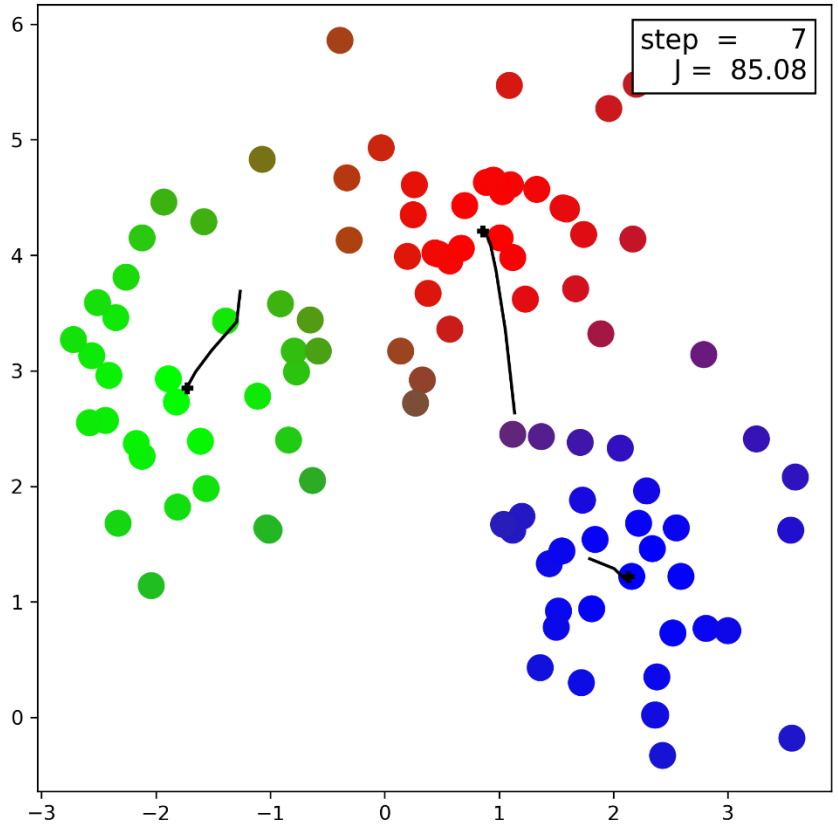

step = 2
J = 103.60
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

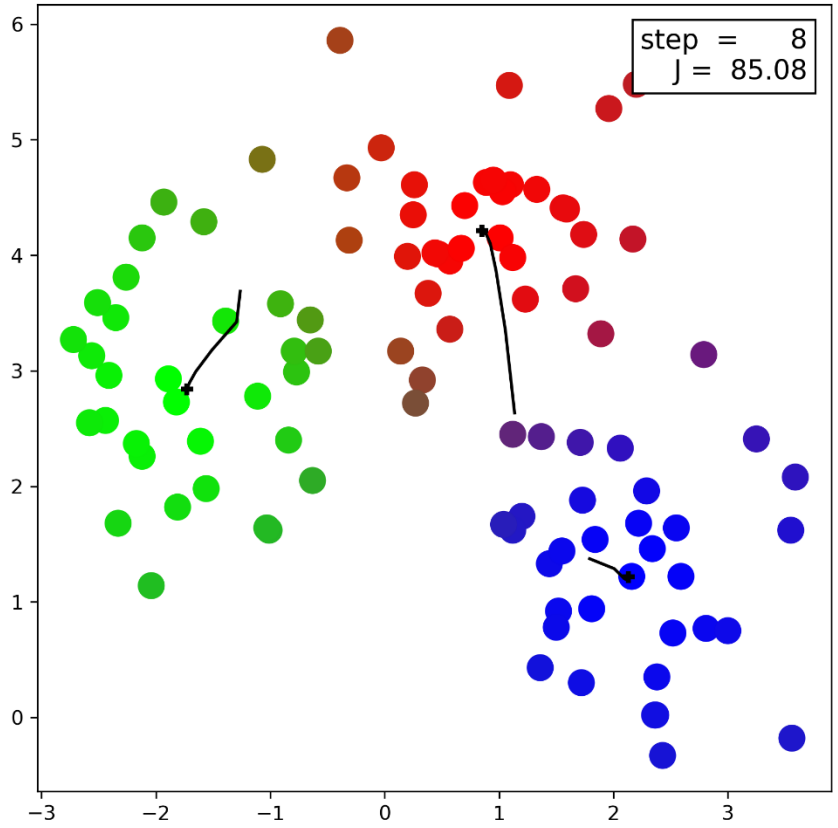

step = 3
J = 89.57
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

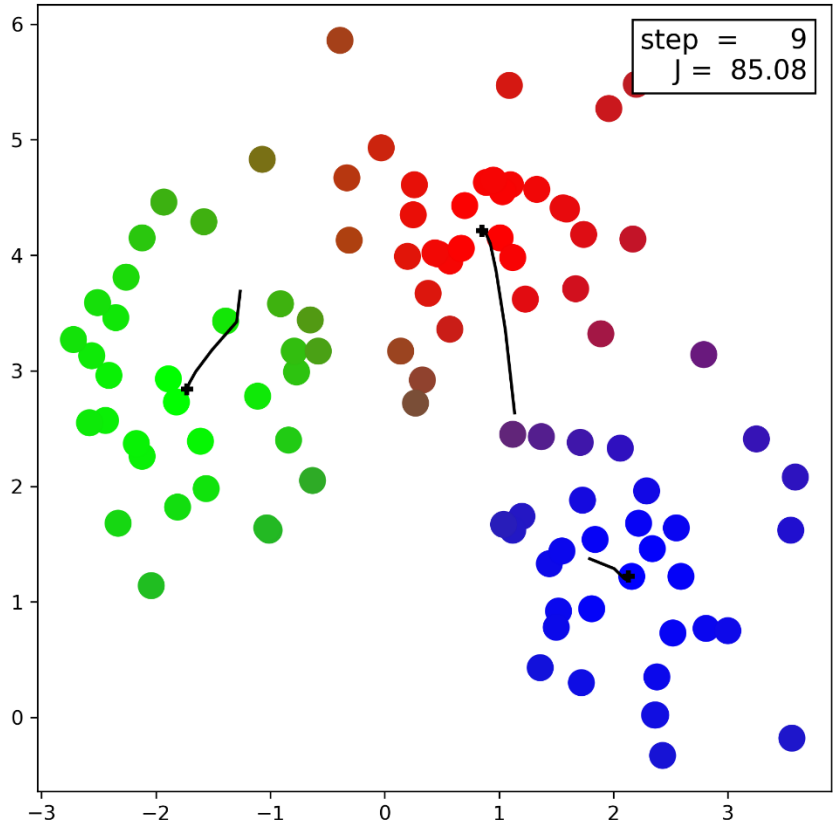

step = 4
J = 85.78
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

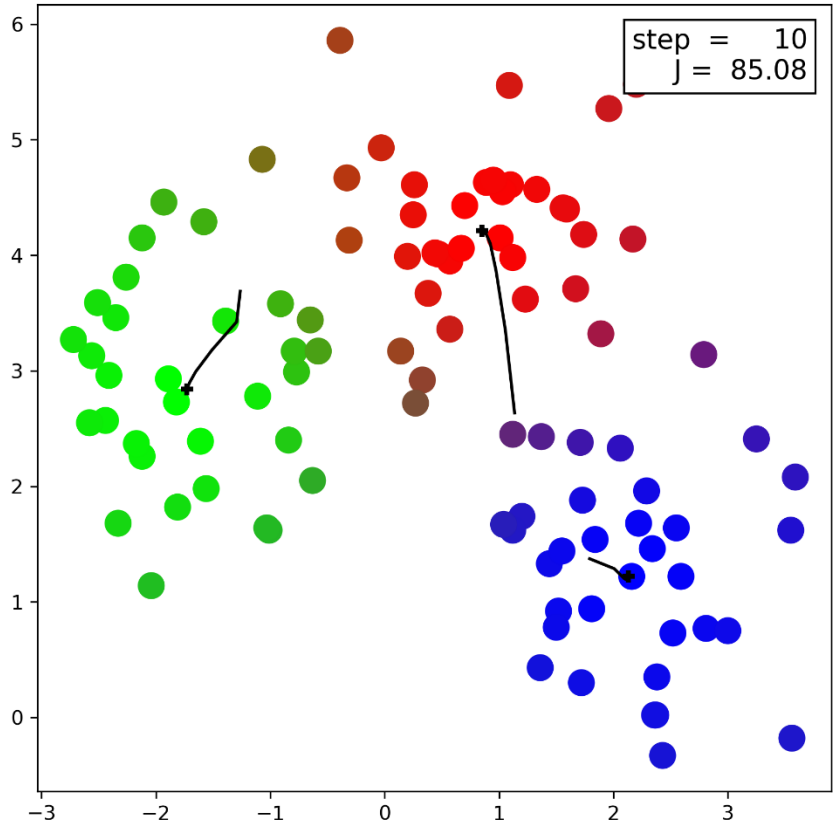

step = 5
J = 85.19
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

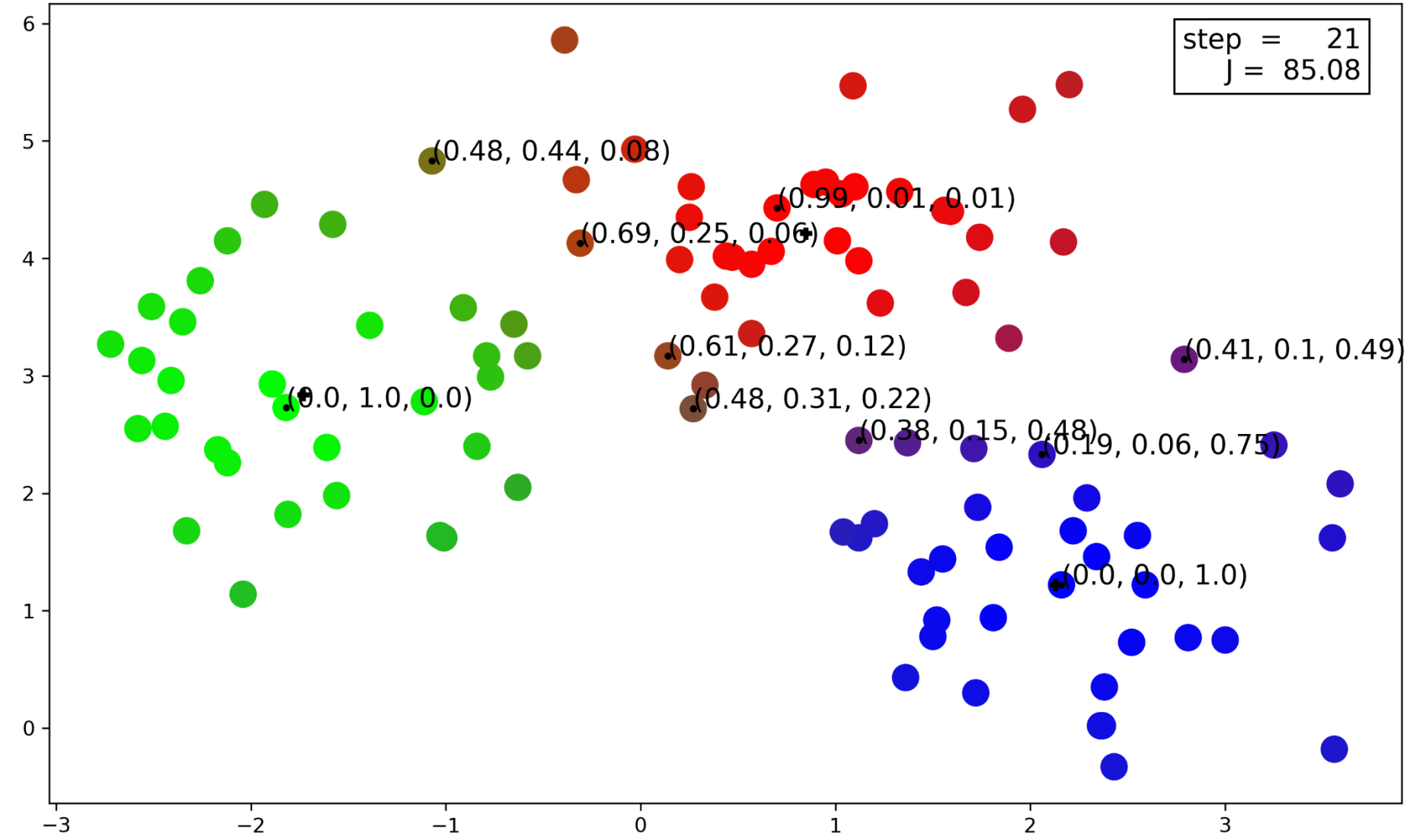

step = 6
J = 85.10
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3


step = 7
J = 85.08
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3


step = 8
J = 85.08
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3


step = 9
J = 85.08
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3


step = 10
J = 85.08
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

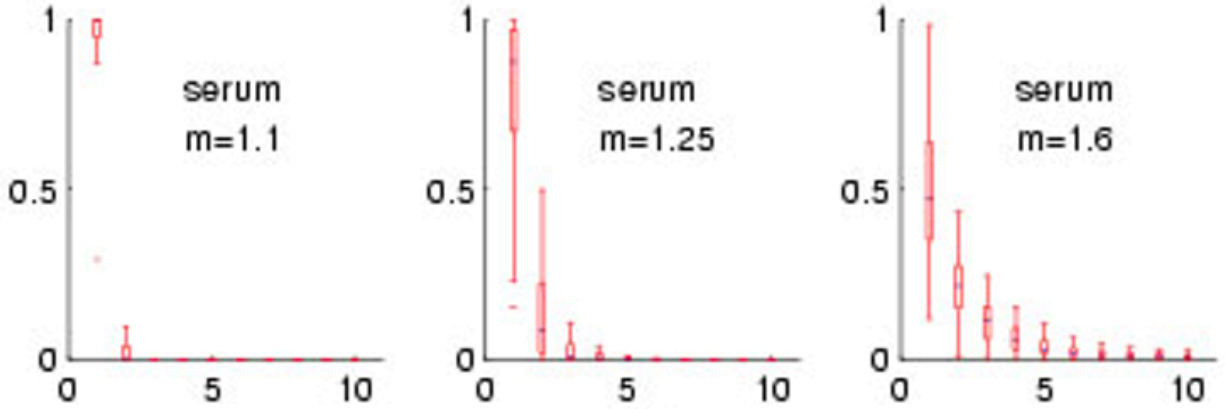

step = 21
J = 85.08
(0.48, 0.44, 0.08)
(0.99, 0.01, 0.01)
(0.69, 0.25, 0.06)
(0.61, 0.27, 0.12)
(0.41, 0.1, 0.49)
(0.0, 1.0, 0.0)
(0.48, 0.31, 0.22)
(0.38, 0.15, 0.48)
(0.19, 0.06, 0.75)
(0.0, 0.0, 1.0)
6
5
4
3
2
1
0
−3
−2
−1
0
1
2
3

Data sets:

	N	p	K used
Serum	517	13	10
Yeast	2945	16	16
Cancer	728	60	20

$$J(K,m) = \sum_{k=1}^{K} \sum_{i=1}^{N} (u_{ki})^m d^2(\mathbf{x}_i, \mathbf{c}_k)$$

- **CL**uster **I**dentification via **C**onnectivity **K**ernels (CLICK) algorithm (Sharan and Shamir, 2000)



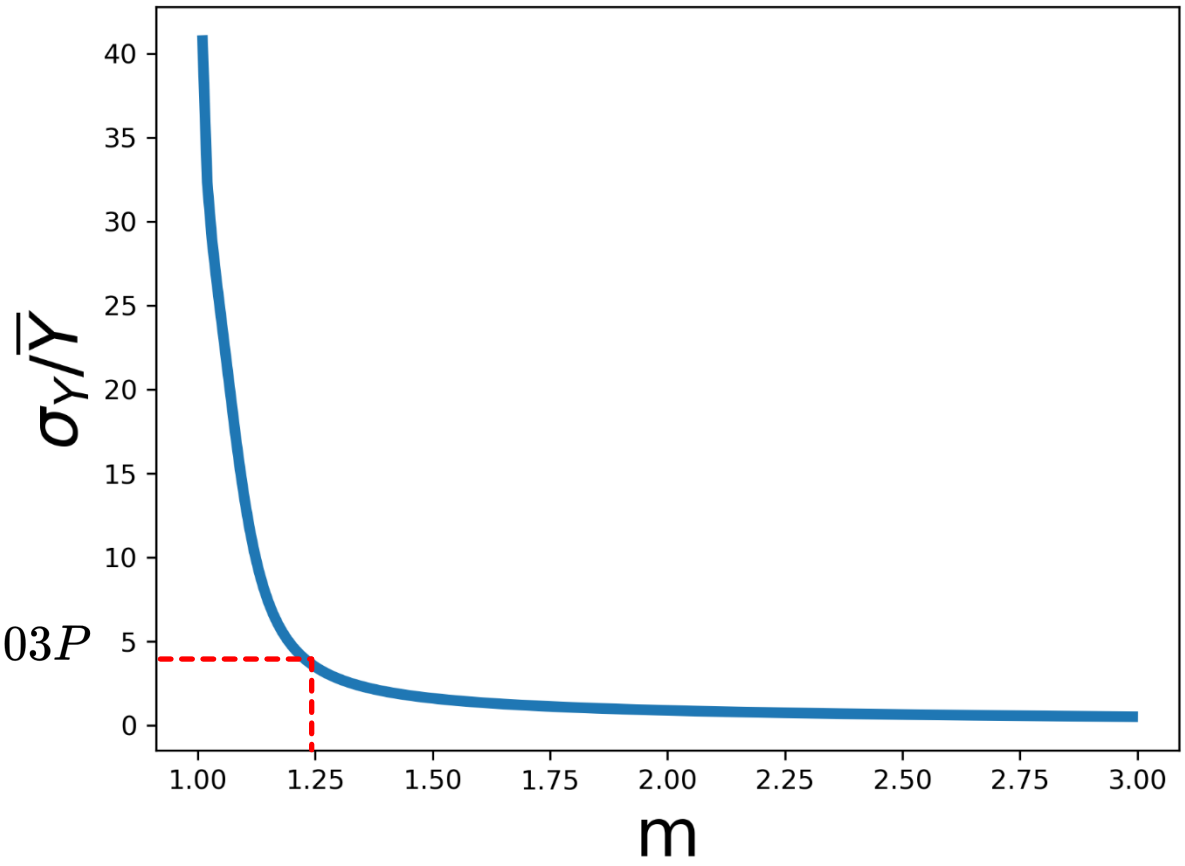
$$u_{ki}^{(l)} = \sum_{s=1}^{K} \left[\frac{d^2(\mathbf{x}_i, \mathbf{c}_k^{(l)})}{d^2(\mathbf{x}_i, \mathbf{c}_s^{(l)})} \right]^{\frac{-1}{(m-1)}}$$

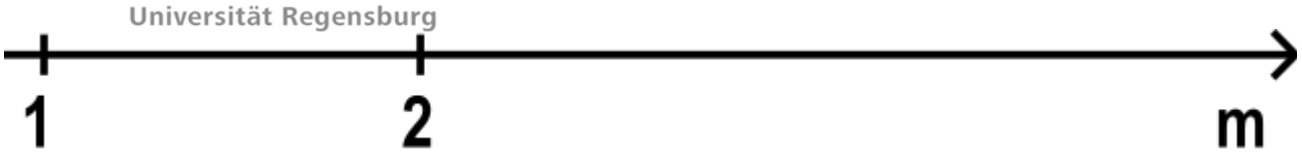
$$Y_m = d^2(\mathbf{x}_i, \mathbf{x}_k^{(l)})^{\frac{1}{(m-1)}}; \quad k \neq i = 1, 2, \ldots, N$$

$$\text{coefficient of variation} = cv(x) = \frac{\sigma_x}{\overline{x}}$$

$$cv(Y_m) = \frac{\sigma_{Y_m}}{\overline{Y_m}} \approx 0.03P$$

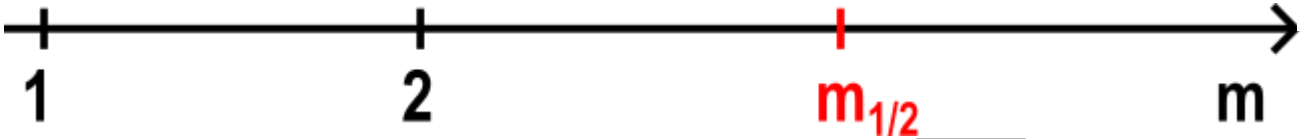
$$\frac{\sigma_{Y_m}}{\overline{Y_m}} \approx 0.03P$$





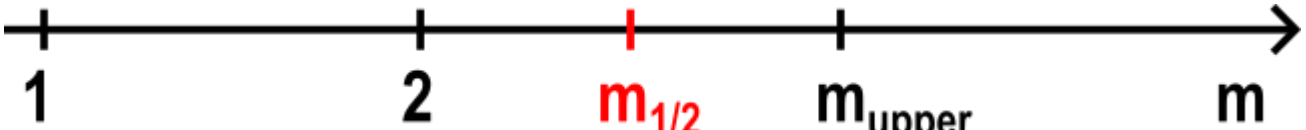
$$\sigma_{Y_2}/\overline{Y_2}$$

$$\sigma_{Y_{mub}}/\overline{Y_{mub}} \approx 0.03P$$



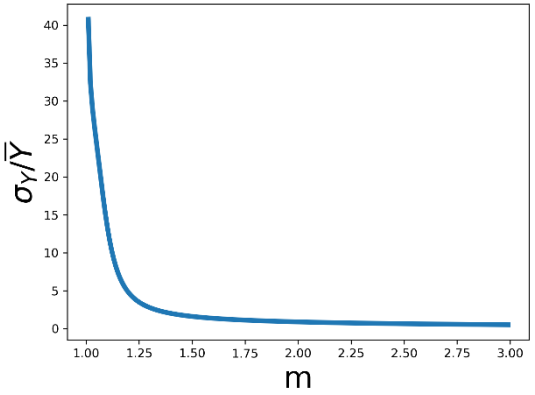
$$\sigma_{Y_{m_{1/2}}}/\overline{Y_{m_{1/2}}}$$

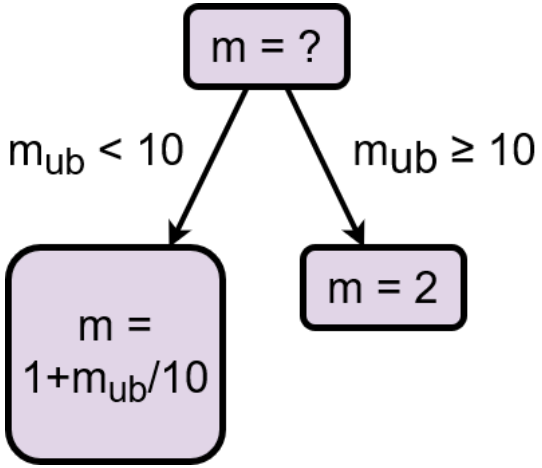
$$\sigma_{Y_{mub}}/\overline{Y_{mub}}$$



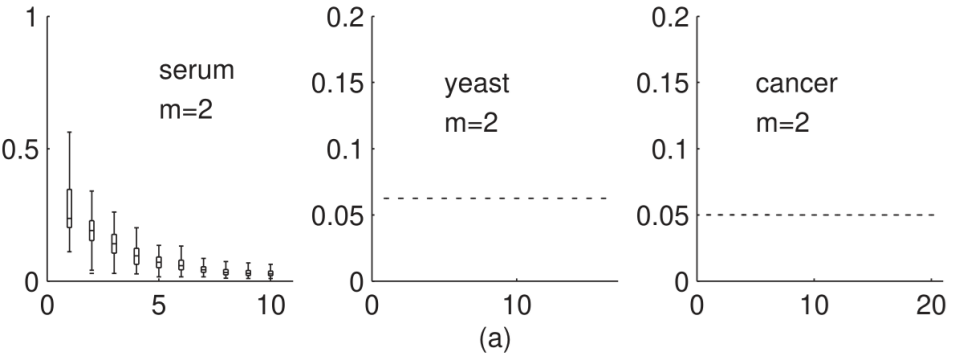
$$\sigma_{Y_{m_{1/2}}}/\overline{Y_{m_{1/2}}}$$

$$\sigma_{Y_{mub}}/\overline{Y_{mub}}$$

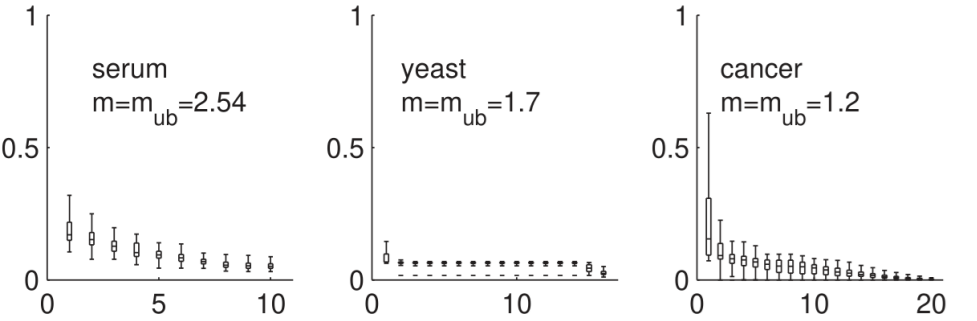


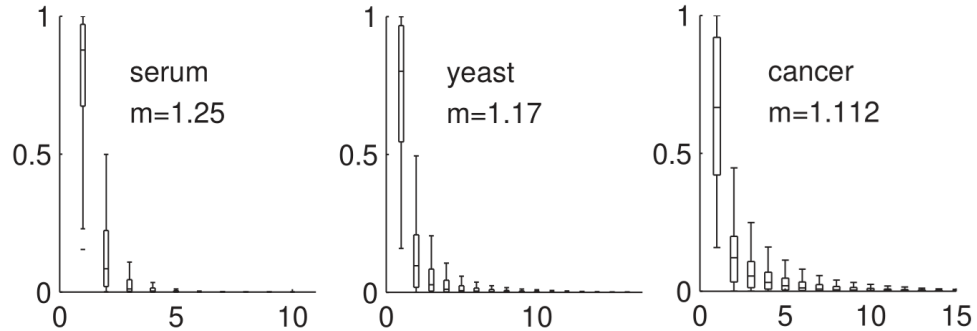


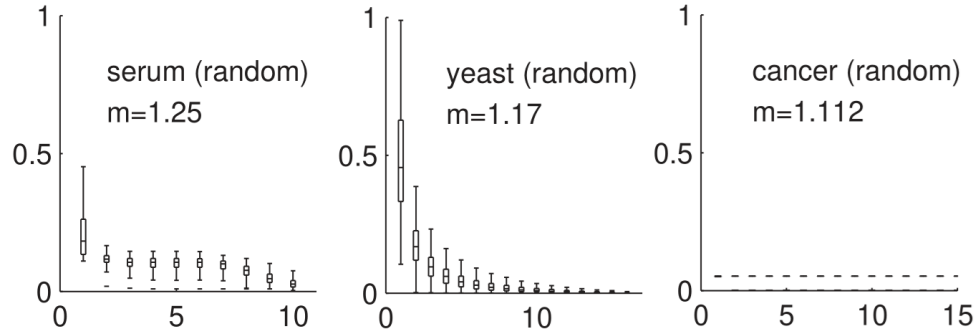
	N	p	m_{ub} exper.	m used	K used
Serum	517	13	2.548	1.25	10
Yeast	2945	16	1.71	1.17	16
Cancer	728	60	1.2	1.112	20



(a)

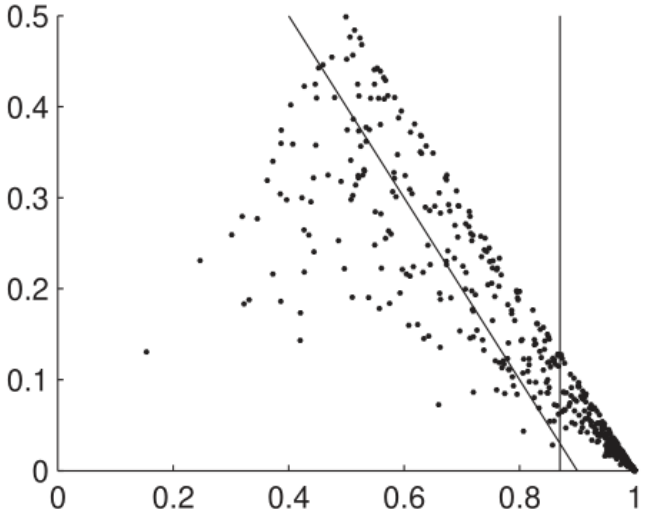




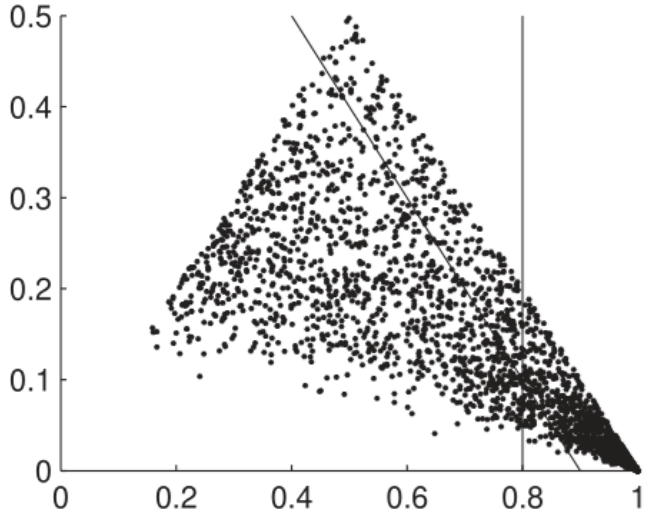


	N	p	m_{ub} exper.	m used	K used
Serum	517	13	2.548	1.25	10
Yeast	2945	16	1.71	1.17	16
Cancer	728	60	1.2	1.112	20

X(N,P)	p_1	p_2	p_3
Gene1			
Gene2			
Gene3			
...			

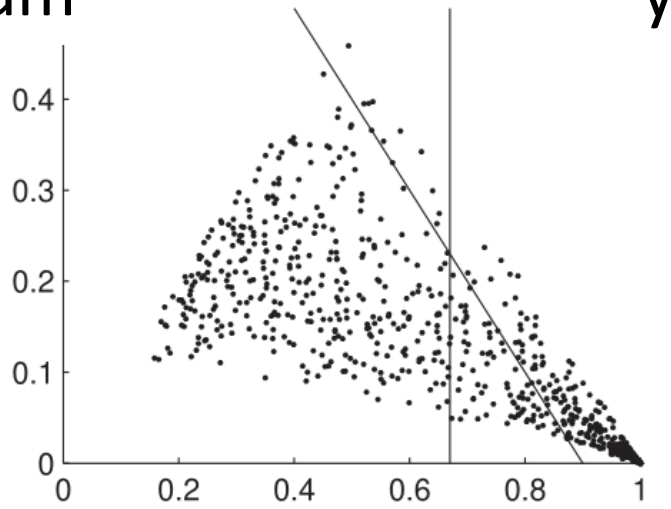


serum



yeast

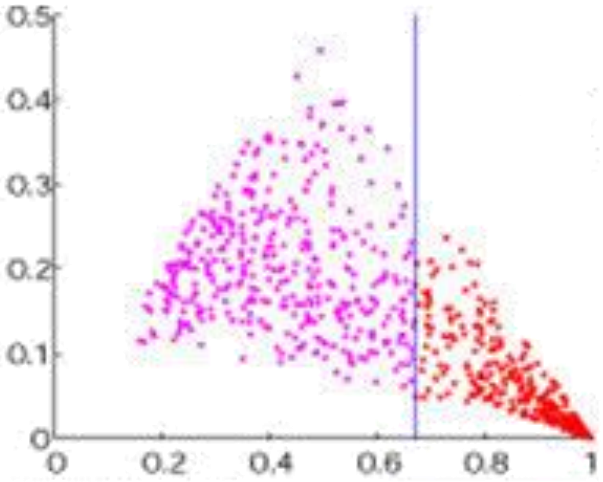
$$(\max \ of \ u_{ki}) + (2^{nd} \max \ of \ u_{ki}) \geqslant 0.9$$

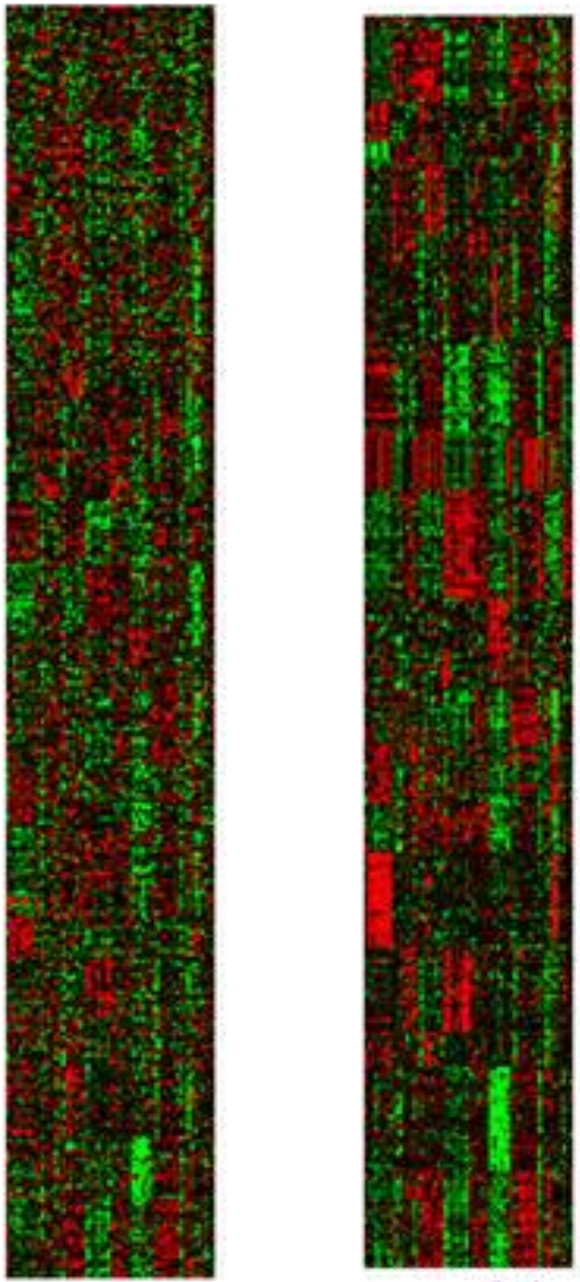


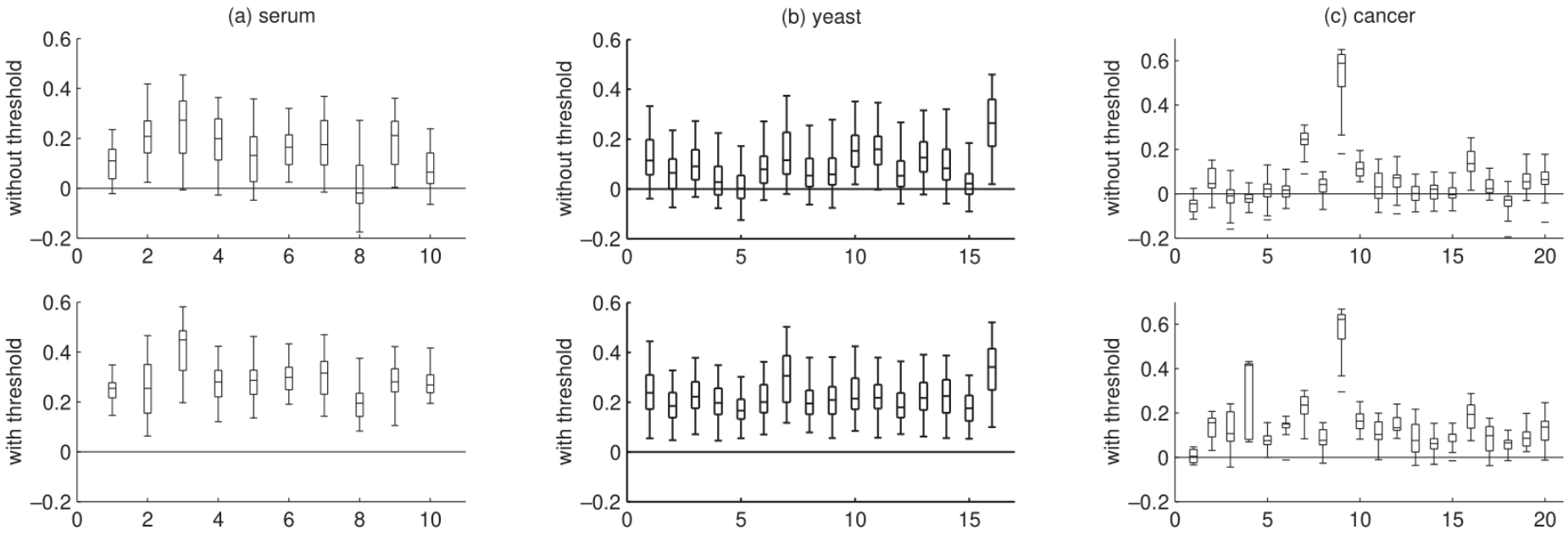
cancer

Alexander Glötzl

Cancer data set






(a) serum
(b) yeast
(c) cancer
without threshold
with threshold
0.6
0.4
0.2
0
−0.2
0
2
4
6
8
10
5
10
15
20

C_k	MIPS functional category (total ORFs)	Raw clusters N_k	M_k	Restricted clusters n_k	m_k	p-Value
11	Ribosome biogenesis (215 ORFs)	225	86	183(81%)	81(94%)	4.05E-5
	Organization of cytoplasm (554 ORFs)		102		94(92%)	9.29E-5
	Organization of nucleus chromosome (44 ORFs)		7		7(100%)	0.23
16	DNA synthesis and replication (94 ORFs)	257	27	232(90%)	27(100%)	0.0538
	Mitotic cell cycle and cycle control (352 ORFs)		41		39(95%)	0.2
	DNA recombination and DNA repair (153 ORFs)		20		20(100%)	0.1187
	Organization of nucleus (774 ORFs)		50		50(100%)	0.0033
3	Organization of mitochondrion (366 ORFs)	217	37	155(71%)	30(81%)	0.107
	Respiration (88 ORFs)		12		10(83%)	0.2814
7	Mitotic cell cycle and cycle control (352 ORFs)	163	26	116(71%)	20(77%)	0.3258
	Budding, cell polarity, filament form. (170 ORFs)		13		11(85%)	0.3687
	DNA synthesis and replication (94 ORFs)		6		6(100%)	0.125
13	TCA pathway or Krebs cycle (25 ORFs)	191	5	153(80%)	4(80%)	0.74
	C-compound, carbohydrate metabo. (415 ORFs)		26		21(81%)	0.5834
14	Nitrogen and sulfur metabolism (67 ORFs)	185	10	142(77%)	8(80%)	0.578
	Amino acid metabolism (204 ORFs)		16		14(80%)	0.2323

$$p\text{-value} = \sum_{i=m_k}^{M_k} \left[\binom{M_k}{i} \binom{N_k - M_k}{n_k - i} \right] / \binom{N_k}{n_k}$$

$$p\text{-value} = 1 - P(X < 81) =$$
$$= P(X = 81) + P(X = 82) + \ldots + P(X = 86)$$

Thank You