Clusteranalyse von POPRDI4Y

Gunter Ritter

Fakultät für Informatik and Mathematik
Universität Passau

ritter@fim.uni-passau.de

6. Oktober 2012

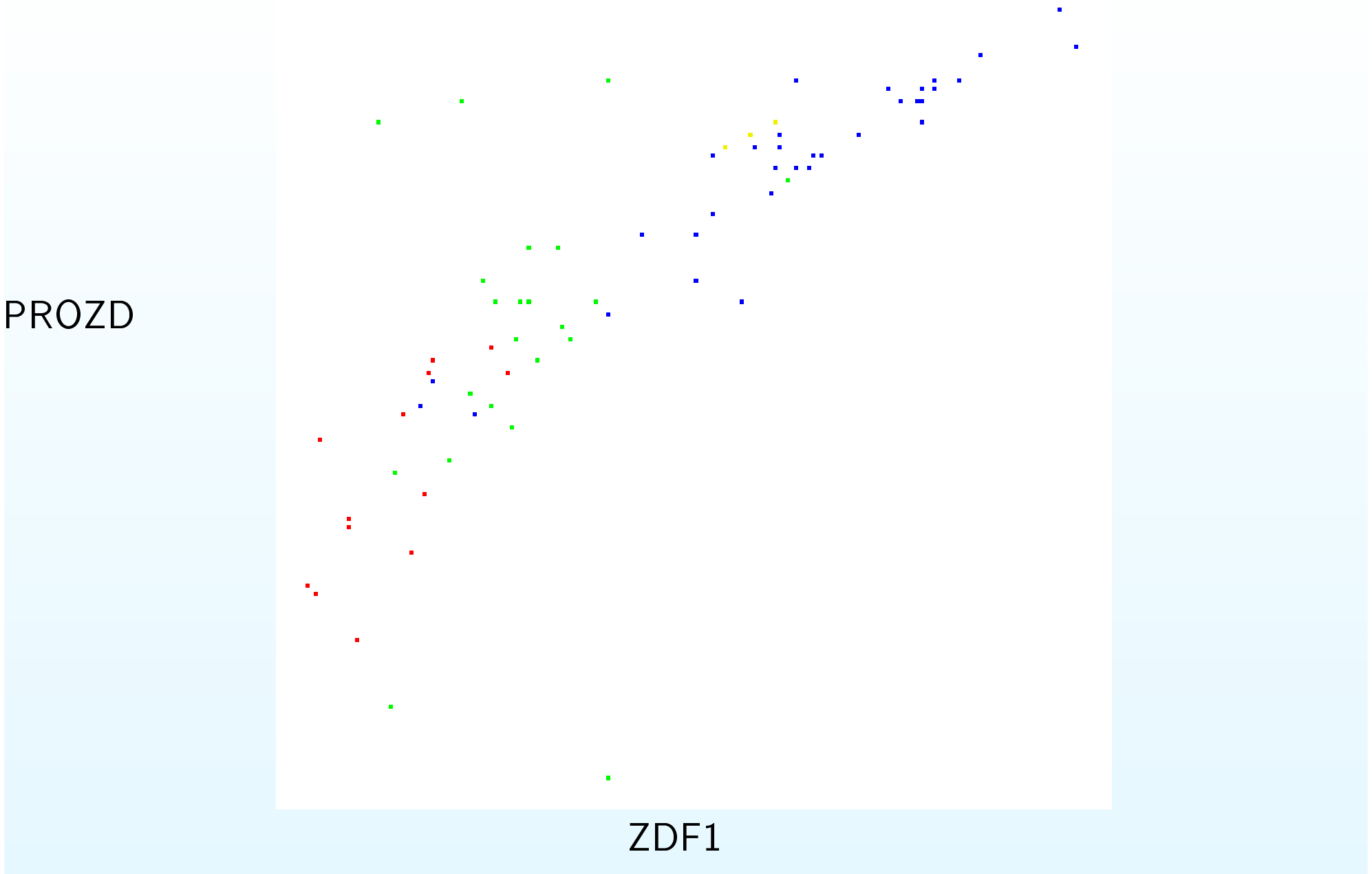
Der Datensatz

REG-MAT-KIE-SIC-DAT2-NAME	DAT	N	LBI	RDI	WTI	SCH LAG	PRIM SFR	FAC SFR	ZDF1	PRO ZD
ddd-2-0-geo-group3-ar	-120	34	NIL	35.3	2.6	NIL	42.4	24.2	47.1	69
mit-1-1-typ-group2-arn	-200	5	1.23	27.0	3.59	122	0	40	40	30
mit-1-1-typ-group2-be	-200	331	1.24	26.5	2.9	121	16	20.7	29.7	72
mit-1-0-geo-group1-bi1	-300	4111	1.07	29.1	3.1	114	44	2.6	26.3	68
mit-2-0-geo-group1-bi2	-300	77	1.08	43.7	2.4	105	32.6	5.8	10.7	42
mit-1-1-geo-group2-bie	-200	8	1.39	29.5	2.78	126	14	0	50	78
mit-1-1-typ-group2-bn	-200	25	1.31	26.3	2.1	119	15.7	15.7	30.4	72
ddd-1-1-geo-group2-bo	-200	211	1.27	27.6	3.5	116	16.8	23	35.2	69
mit-1-1-typ-group2-by	-200	8	1.11	32.6	2.9	113	15.8	15.8	15	57
mit-1-1-geo-group3-c	-80	50	1.32	29.5	2.57	121	22	2	18	63
eur-1-1-geo-group1-cl	-300	134	1.16	33.4	2.3	131	7.5	14.9	6	60

81 Objekte;
10 numerische Variablen, 8 MMe, 4 Geometriemaße, 4 rel. Hfgk.;
7 mit fehlenden Werten, davon 3 mit ≥ 3

N und ZDF1 wurden von mir logarithmiert (nach Visualisierung)

Visualisierung der angegebenen Gruppen



Die verwendeten Programmpakete (C++)

Iterative Relokations-Algorithmen, basierend auf Gaußschem Likelihood, Misch-, Klassifikationsmodell

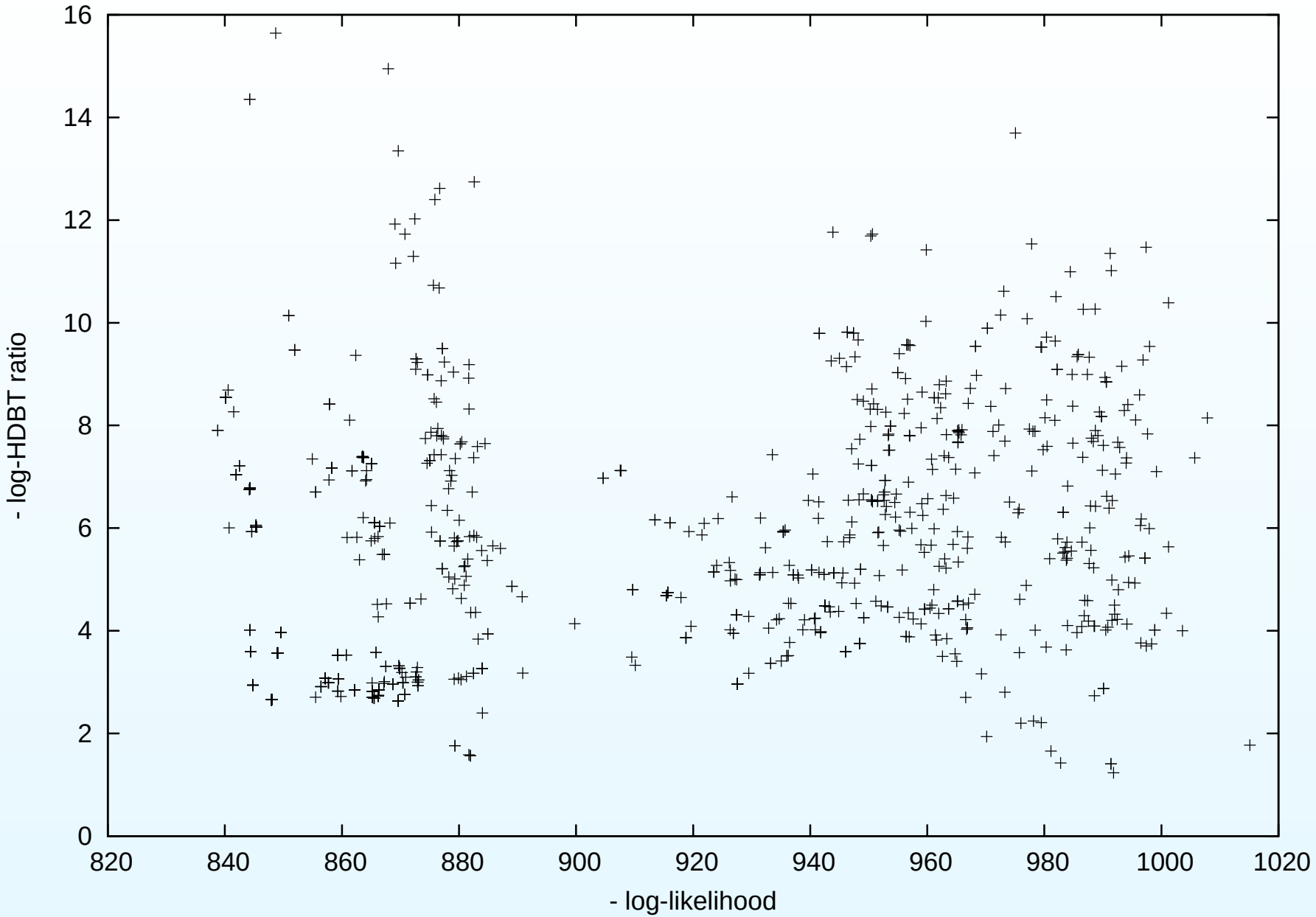
Funktionalität: # Komponenten, # Ausreißer, Minimalgröße der Cluster, Skala voll/diagonal/sphärisch, homo- heteroskedastisch, MAP/ML, Varianten $\Rightarrow$ hier für fehlende Werte
Optimierung mit Skalenbalance $\Rightarrow$ Strafterm mit log-HDBT ratio

MIXTURE: Mixture-Likelihood

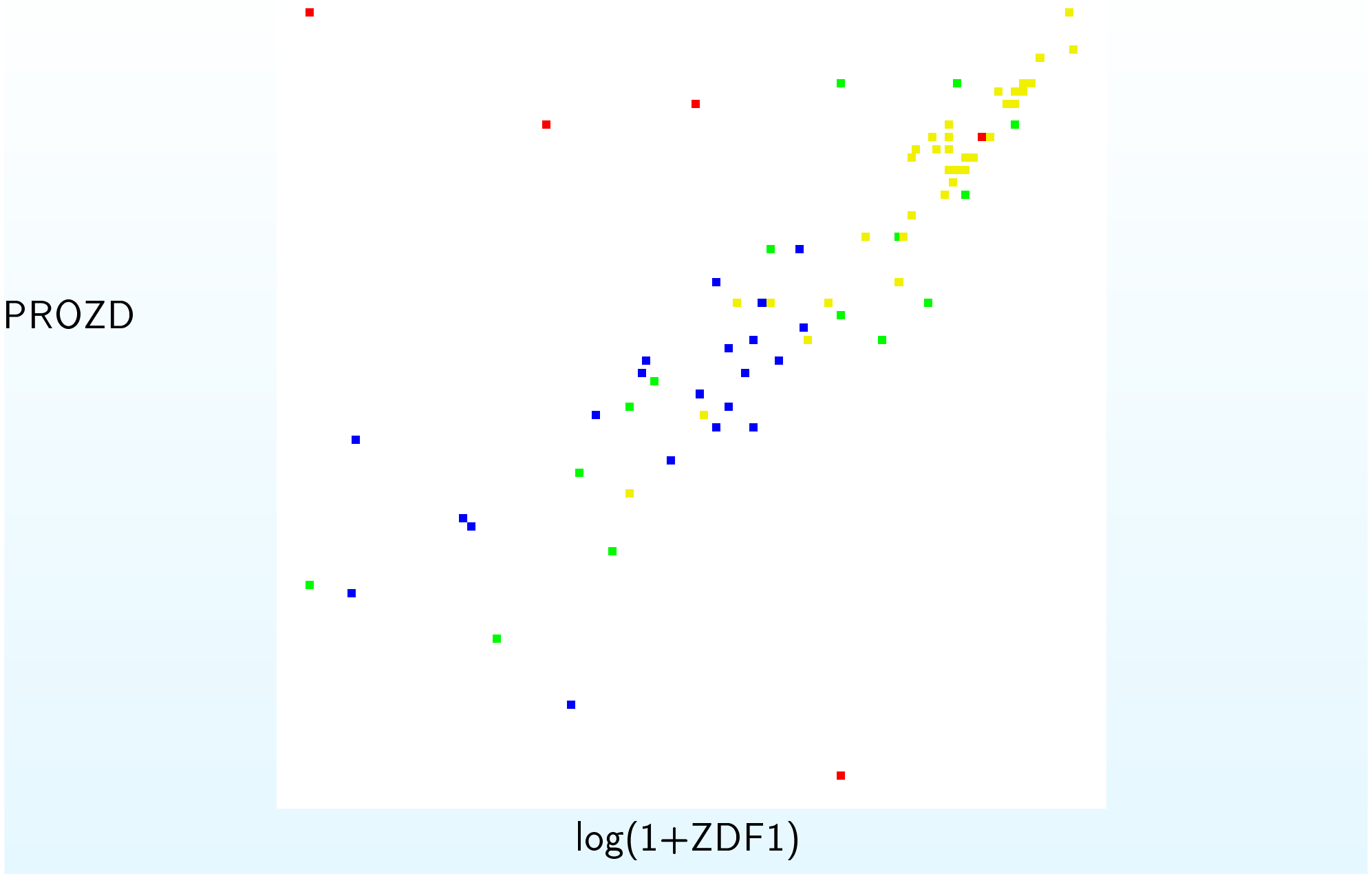
CLOUDS: Klassifikations-Likelihood

SELCLUS: Wie CLOUDS, aber mit Variablenselektion (Irrelevanz), entwickelt für Analyse von Genexpressionsdaten

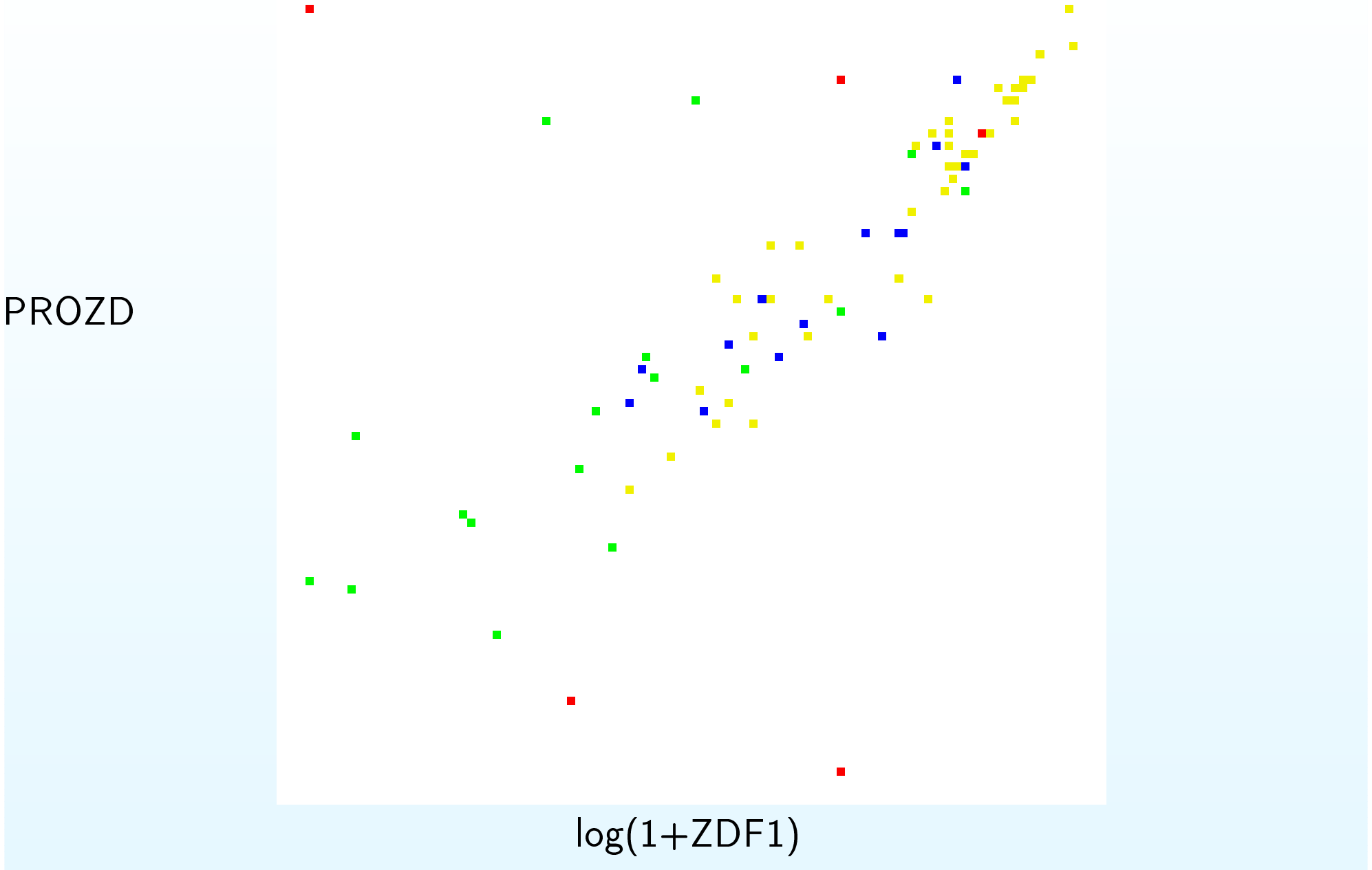
Lok. Optima: Plot -log-HDBT ratio vs. $-\log$-Likelihood



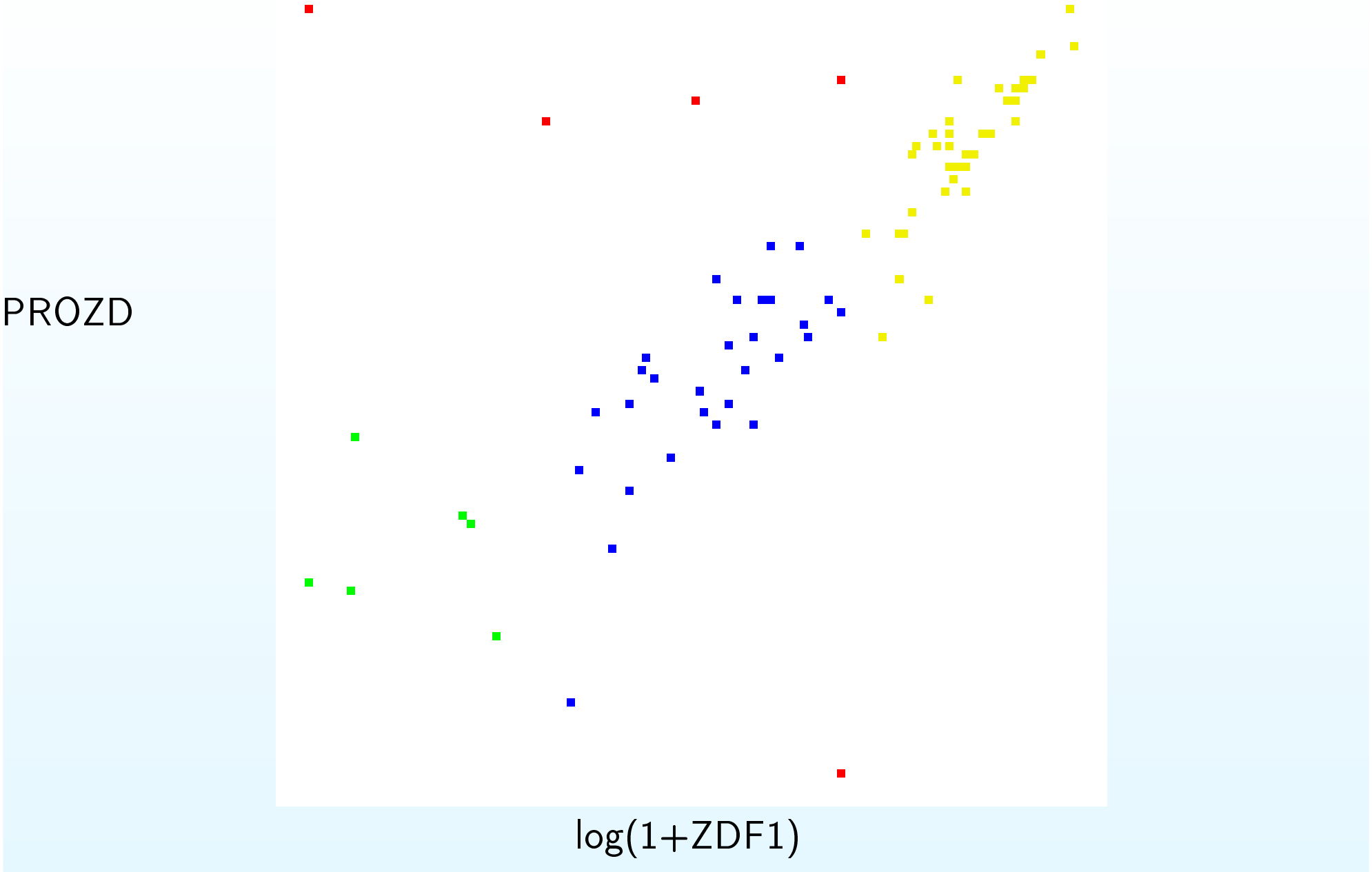
Ergebnis mit MIXTURE, 3 Komponenten, 5 Ausreißer



Ergebnis mit CLOUDS, 3 Cluster, 5 Ausr.



Ergebnis mit SELCLUS – 3 Cluster, wähle 1 MM, 5 Ausreißer



Ergebnis aus SELCLUS – 3 Cluster, wähle 1 MM, 5 Ausreißer

Gewähltes Merkmal: ZDF1

Ausreißer: l, ms, va, wl, wst

Matrix der Übereinstimmung mit den gegebenen Gruppen:

	cluster1	cluster2	cluster3	Ausreißer
group1	6	7	0	0
group2	0	18	3	4
group3	0	4	32	1
group4	0	0	3	0

Literatur

[1] J. A. Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. Trimmed k–means: An attempt to robustify quantizers. *Ann. Statist.*, 25:553–576, 1997.

[2] María Teresa Gallegos and Gunter Ritter. Trimmed ML–estimation of contaminated mixtures. *Sankhyā, Series A*, 71:164–220, 2009.

[3] María Teresa Gallegos and Gunter Ritter. Using combinatorial optimization in model-based trimmed clustering with cardinality constraints. *Computational Statistics and Data Analysis*, 54:637–654, 2010. DOI 10.1016/j.csda.2009.08.023.

[4] Christian Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, 52:258–271, 2007.

[5] Ron Kohavi and George E. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[6] Adrian E. Raftery and Nema Dean. Variable selection for model–based clustering. *J. Amer. Stat. Assoc.*, 101:168–178, 2006.

[7] A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[8] M.J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.