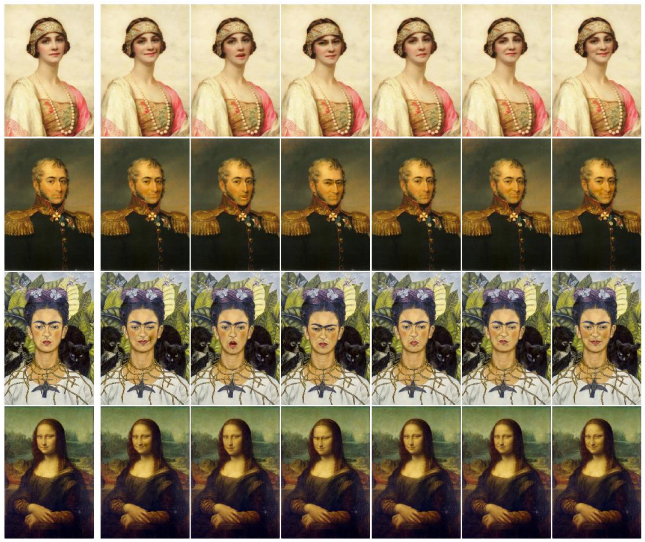
RJ-Clust: A Fast Clustering Algorithm for High-Dimensional Data

joint work with
Dr. Valen E. Johnson

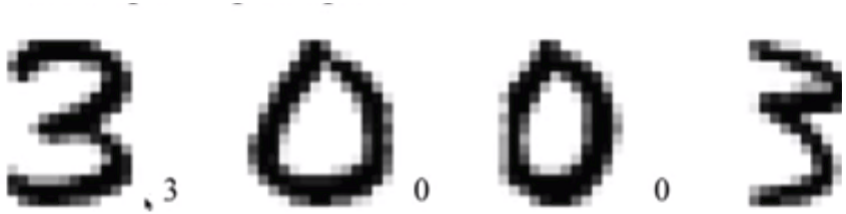
Department of Statistics
Texas A&M University

October 12, 2018

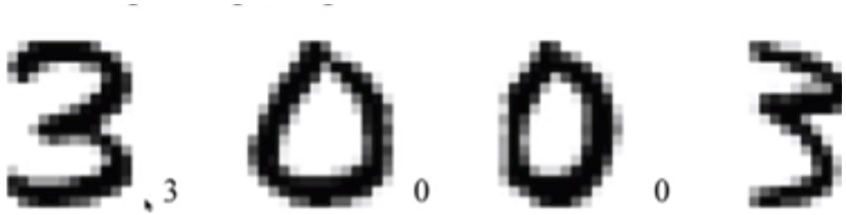
Clusters are everywhere



Human vs Machine

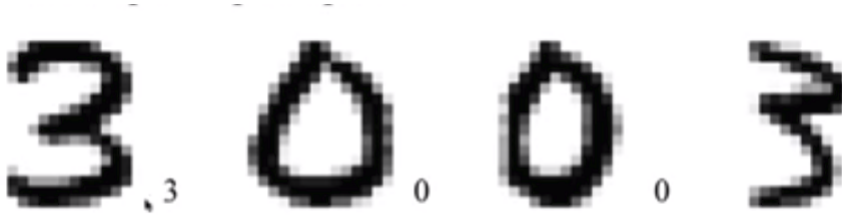


Human vs Machine



Machine can detect

Human vs Machine



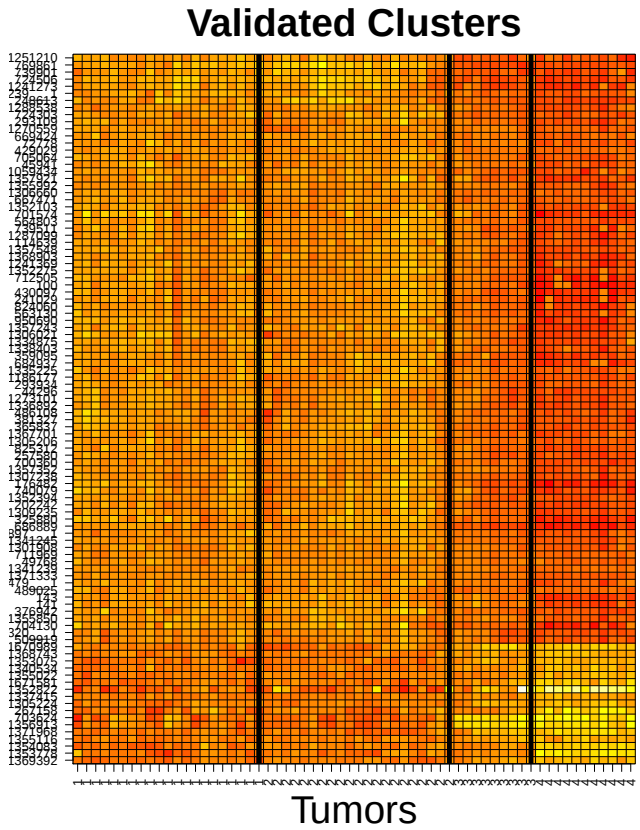
Machine can detect

"*The Brain is nothing but a statistical decision organ*" - H. Barlow

Clustering which humans cannot observe

Massive Galaxy Clusters
(P.Courtsey- Hubble)

Gene-expression Blood Tumor Clusters
(P.Courtsey- **RJClust**)



Clustering definition

"Discover the **natural groups** of a set of objects **automatically** in the **unlabeled** data".

Clustering definition

"Discover the **natural groups** of a set of objects **automatically** in the **unlabeled** data".

- objects in the same group should be similar as possible.
- objects in different groups should be different as possible.

Clustering definition

"Discover the **natural groups** of a set of objects **automatically** in the **unlabeled** data".

- objects in the same group should be similar as possible.
- objects in different groups should be different as possible.

Similar in shape, size and density.

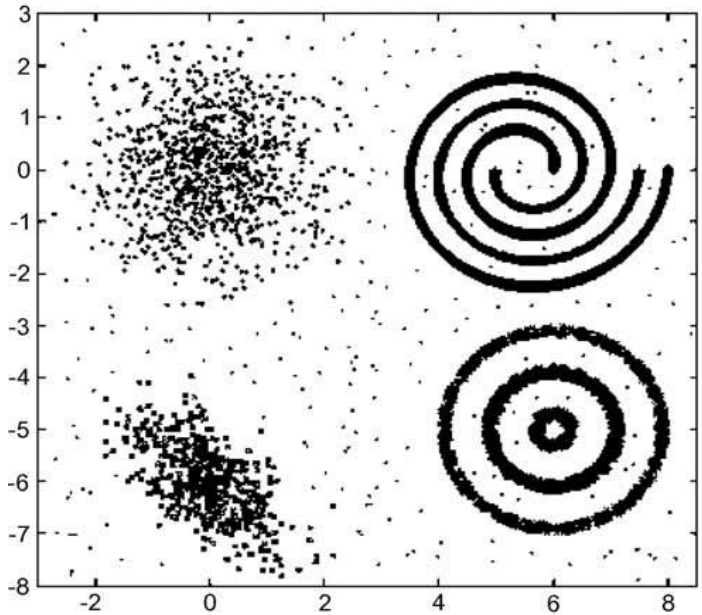
Useful overview : Tabachnik and Fidell, 2007; Duda et al. 2001.

Clustering methods differ in the choice of the objective function.

Popular Clustering Algorithms: Many....

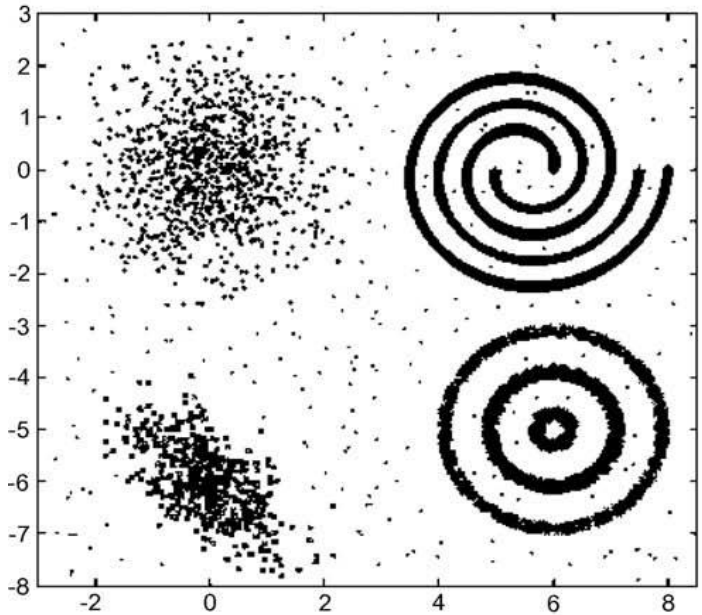
- **Hierarchical Methods** - (Aggolomerative/Divisive) Single, Average, Complete link, Ward,..
- **Partition Methods** - (K-means and variants)Mahalanobis distance, pam, L_1 distance, Fuzzy C- means, Kernel K-means, K-medoid, bisecting K-means, Penalized K-means,...
- **Density based Methods** - GMM, JarvisPatrick algorithm, DBSCAN
- **Subspace Clustering** - CLIQUE, minimum cut algorithm, NCut, Modified NCut, spectral clustering, Laplacian Eigen Map, normalized eigen vectors of kernel matrix etc.....
- **Information thoery** - Minimum Entropy method, maximizing mutual information (MI), etc
- **Cluster Ensembles** - Co-occurrence Matrix, Consensus matrix (CC) ....

King-Sun Fu Prize Lecture: A.K. Jain, 2009



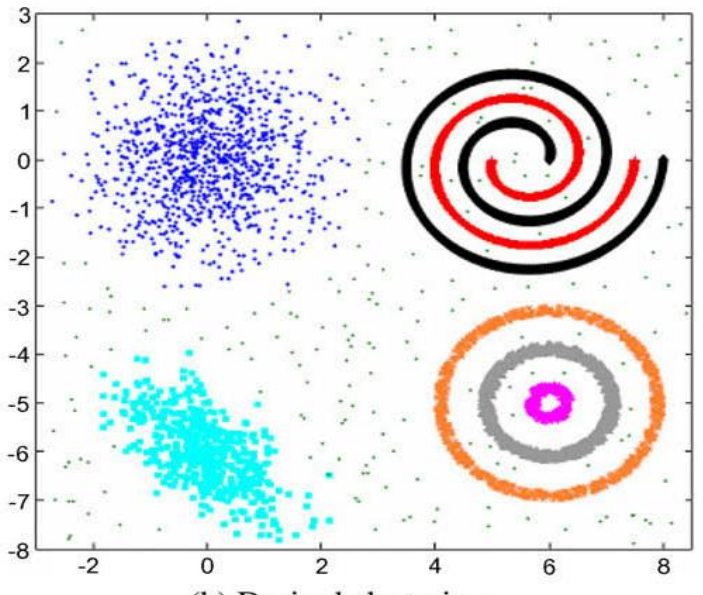
(a) Input data

King-Sun Fu Prize Lecture: A.K. Jain, 2009



(a) Input data

None of the available clustering algorithms can find all 7 clusters.



(b) Desired clustering

50 years and beyond K-means

"*Despite these developments, no single algorithm has emerged to displace the k-means scheme and its variants*"

— Shah and Koltun, PNAS, 2017

50 years and beyond K-means

"*Despite these developments, no single algorithm has emerged to displace the k-means scheme and its variants*"

— Shah and Koltun, PNAS, 2017

"*The endurance of these methods is in part due to their* **simplicity** *and in part due to difficulties associated with some of the new techniques, such as* **additional hyperparameters** *that need to be tuned,* **high computational cost**, *and .....*"

Effective and Fast Clustering in High Dimension

New Clustering Method: RJClust

Goal

Interested in clustering N rows on the basis of P features when $P \gg N$.

$$\boldsymbol{X} = \begin{array}{|cccccc|c|} \hline X_{1,1} & X_{1,2} & \cdots & X_{1,j} & \cdots & X_{1,P} & 1 \\ X_{2,1} & X_{2,2} & \cdots & X_{2,j} & \cdots & X_{2,P} & 1 \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 2 \\ X_{k,1} & X_{k,2} & \cdots & X_{k,j} & \cdots & X_{k,P} & 2 \\ X_{k+1,1} & X_{k+1,2} & \cdots & X_{k+1,j} & \cdots & X_{k+1,P} & 2 \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 3 \\ X_{N,1} & X_{N,2} & \cdots & X_{N,j} & \cdots & X_{N,P} & 3 \\ \hline \end{array}$$

Basic Idea of our Algorithm

If $\mathbf{X}$ is $N \times P$ matrix of features, we base clustering on

$$\mathbf{R} = \mathbf{X}\mathbf{X}^T,$$

an $N \times N$ matrix

Basic Idea of our Algorithm

If $\mathbf{X}$ is $N \times P$ matrix of features, we base clustering on

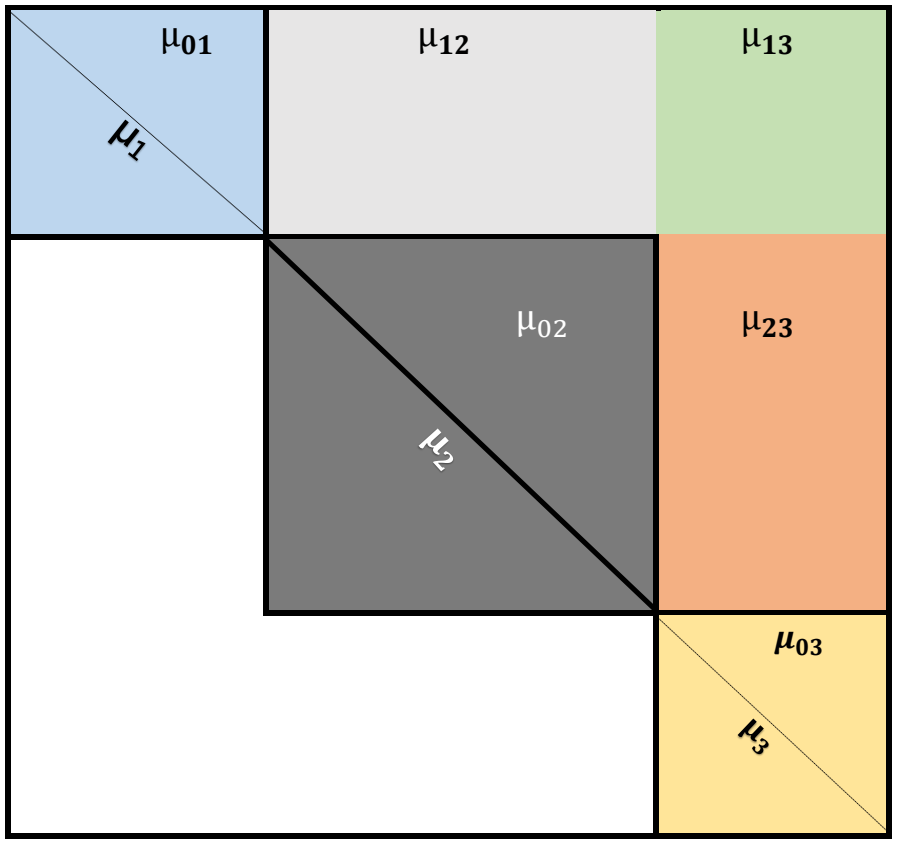
$$\mathbf{R} = \mathbf{X}\mathbf{X}^T,$$

an $N \times N$ matrix

Is the R-transformation **invariant** to the **number of Clusters** in the rows of the X ?

Structure of $E(\mathbf{R})$

R preserves the number of Clusters of **X**.



R preserves the number of Clusters of **X**.

$$R_{ij} = \sum_{k=1}^{P} x_{ik} x_{jk} / P$$

R preserves the number of Clusters of **X**.

$$R_{ij} = \sum_{k=1}^{P} x_{ik} x_{jk} / P$$

Theorem 1: For P$\to \infty$, under a block covariance structure among the features, if $\sum_{s=1}^{L_P} |B_{s,P}|^2 \tau_s^2 = o(p)$, then

$$\sqrt{P}(R_{ij} - \mu_{\gamma_i,\gamma_j}) \xrightarrow{d} N(0, \sigma^2_{\gamma_i,\gamma_j}), \qquad i \neq j$$

$$\sqrt{P}(R_{ii} - \mu_{\gamma_i}) \xrightarrow{d} N(0, \sigma^2_{\gamma_i}), \qquad i = j$$

where,

γ_i = cluster identifier of subject i.

$|B_{s,P}|$ = cardinality of the features with same covariance in s^{th} block

τ_s^2 = covariance of the s^{th} block.

CLT on the rows of **R**

Ignoring the small correlations within rows/columns of **R**,

Theorem 2:For fix N and for large P, the rows of **R**, $\boldsymbol{R}_i = (R_{i1}, R_{i2}, \cdots, R_{iN})$ in the same cluster. Let $\mathcal{C}$ be the class of all convex subsets, we can have a Lyapunov-type bound for

$$sup_{A \in \mathcal{C}} |P(\boldsymbol{R}_i \in A) - P(\boldsymbol{Z} \in A)| \leq cN^{1/4}/\sqrt{P}$$

where, c is a constant.

3 Steps of the Algorithm

1. R-step: $\mathbf{R} = \mathbf{X}\mathbf{X}^T/P$.
2. J-step:

$$\mathbf{R} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \Longrightarrow \begin{bmatrix} \frac{R_{21}+R_{31}}{2} & R_{12} & R_{13} & R_{11} \\ R_{21} & \frac{R_{12}+R_{32}}{2} & R_{23} & R_{22} \\ R_{31} & R_{32} & \frac{R_{13}+R_{23}}{2} & R_{33} \end{bmatrix} = \mathbf{J}$$

$$E(\boldsymbol{J}_{k\cdot}). = E(\boldsymbol{J}_{l\cdot}).$$

   when k^{th} and l^{th} row belong to the same cluster

3. Treat rows of $\mathbf{J}$ as $(N+1) \times 1$ vectors from Gaussian Mixture Model.

Structure of J for 2 Clusters

E(J)

	1	2	3	4	5	6	7
J_1	μ_{11}	μ_{11}	μ_{11}	μ_{11}	μ_{12}	μ_{12}	$\boldsymbol{\mu_1}$
J_2	μ_{11}	μ_{11}	μ_{11}	μ_{11}	μ_{12}	μ_{12}	$\boldsymbol{\mu_1}$
J_3	μ_{11}	μ_{11}	μ_{11}	μ_{11}	μ_{12}	μ_{12}	$\boldsymbol{\mu_1}$
J_4	μ_{11}	μ_{11}	μ_{11}	μ_{11}	μ_{12}	μ_{12}	$\boldsymbol{\mu_1}$
J_5	μ_{12}	μ_{12}	μ_{12}	μ_{12}	μ_{22}	μ_{22}	$\boldsymbol{\mu_2}$
J_6	μ_{12}	μ_{12}	μ_{12}	μ_{12}	μ_{22}	μ_{22}	$\boldsymbol{\mu_2}$

Σ_1

	J_{11}	J_{12}	J_{13}	J_{14}	J_{15}	J_{16}	J_{17}
J_{11}	.	.	.	.	.	.	.
J_{12}	.	$\sigma^2_{1,1}$	$\sigma^2_{1,1,1}$	$\sigma^2_{1,1,1}$	$\sigma^2_{1,1,2}$	$\sigma^2_{1,1,2}$	$\eta_{1,1}$
J_{13}	.		$\sigma^2_{1,1}$	$\sigma^2_{1,1,1}$	$\sigma^2_{1,1,2}$	$\sigma^2_{1,1,2}$	$\eta_{1,1}$
J_{14}	.			$\sigma^2_{1,1}$	$\sigma^2_{1,1,2}$	$\sigma^2_{1,1,2}$	$\eta_{1,1}$
J_{15}	.				$\sigma^2_{1,2}$	$\sigma^2_{1,2,2}$	$\eta_{1,2}$
J_{16}	.					$\sigma^2_{1,2}$	$\eta_{1,2}$
J_{17}	.						σ^2_{10}

Σ_2

	J_{51}	J_{52}	J_{53}	J_{54}	J_{55}	J_{56}	J_{57}
J_{51}	$\sigma^2_{1,2}$	$\sigma^2_{1,1,2}$	$\sigma^2_{1,1,2}$	$\sigma^2_{1,1,2}$	.	$\sigma^2_{1,2,2}$	$\eta_{2,1}$
J_{52}		$\sigma^2_{1,2}$	$\sigma^2_{1,1,2}$	$\sigma^2_{1,1,2}$	.	$\sigma^2_{1,2,2}$	$\eta_{2,1}$
J_{53}			$\sigma^2_{1,2}$	$\sigma^2_{1,1,2}$	.	$\sigma^2_{1,2,2}$	$\eta_{2,1}$
J_{54}				$\sigma^2_{1,2}$	.	$\sigma^2_{1,2,2}$	$\eta_{2,1}$
J_{55}	.	.	.	.	.	.	.
J_{56}					.	$\sigma^2_{2,2}$	$\eta_{2,2}$
J_{57}					.		σ^2_{20}

No. of Parameters $=$
$4\binom{C}{1}+3\binom{C}{2}+\binom{C}{3}+C+C-1 \approx O(C^3)$

EM implementation on **J** with Exact Covariance Structure

For each cluster configuration, $C = 1, ..., C_{max}$,

- **Initialization :** Hierarchical agglomeration on **J** matrix.
- **Estimation of Parameters :** Apply EM Algorithm.
- **Compute BIC** for each cluster configuration,

$$2 \log [f(\mathbf{J}|C)] - K \log(N).$$

  where, $K = 4\binom{C}{1} + 3\binom{C}{2} + \binom{C}{3} + C + C - 1$

- **Choose C** corresponding to the **optimal BIC.**

Schwarz,1978; Keribin,1998; Fraley,1998, 2002; Standford,2002

AMI Comparison on 32 Genomic Data sets

Datasets	kmeans++	GMM	CC-km	CC-hc	GAP	AC-W	N-Cuts	AP	Zell	SEC	LDMGI	PIC	RCC	RCC-DR	RJ
Alizadeh-2000-v1	0.340	0.024	0.037	0.007	0.000	0.101	0.096	0.232	0.250	0.238	0.123	0.033	0.000	0.426	0.515
Alizadeh-2000-v2	0.568	0.922	0.695	0.628	1.000	0.922	0.922	0.563	0.922	0.922	0.738	0.922	1.000	1.000	1.000
Alizadeh-2000-v3	0.586	0.604	0.551	0.497	0.629	0.616	0.601	0.540	0.702	0.574	0.582	0.625	0.792	0.792	0.792
Armstrong-2002-v1	0.372	0.372	0.478	0.463	0.475	0.308	0.372	0.381	0.308	0.323	0.355	0.308	0.528	0.546	0.547
Armstrong-2002-v2	0.891	0.803	0.541	0.375	0.525	0.746	0.83	0.586	0.802	0.891	0.509	0.802	0.642	0.838	0.539
Bhattacharjee-2001	0.444	0.406	0.598	0.521	0.518	0.601	0.563	0.377	0.496	0.570	0.378	0.378	0.495	0.600	0.557
Bittner-2000	-0.012	-0.002	0.021	0.024	0.000	0.002	0.042	0.243	0.115	-0.002	0.014	0.115	-0.016	0.156	0.138
Bredel-2005	0.297	0.208	0.202	0.211	0.035	0.384	0.203	0.139	0.278	0.259	0.295	0.278	0.468	0.466	0.265
Chowdary-2006	0.764	0.808	0.499	0.298	0.000	0.859	0.859	0.443	0.859	0.859	0.859	0.859	0.360	0.393	0.585
Dyrskjot-2003	0.507	0.532	0.236	0.241	0.348	0.474	0.303	0.558	0.269	0.389	0.385	0.177	0.359	0.383	0.623
Garber-2001	0.242	0.137	0.026	0.026	0.096	0.210	0.204	0.274	0.246	0.200	0.191	0.246	0.240	0.173	0.130
Golub-1999-v1	0.688	0.583	0.688	0.418	0.044	0.831	0.650	0.430	0.615	0.615	0.615	0.615	0.527	0.490	0.420
Golub-1999-v2	0.680	0.730	0.439	0.282	0.000	0.737	0.693	0.516	0.689	0.703	0.600	0.689	0.656	0.597	0.538
Gordon-2002	0.651	0.669	0.651	0.432	0.435	0.483	0.681	0.304	-0.005	0.791	0.669	0.664	0.349	0.343	0.429
Laiho-2002	0.007	0.207	0.116	0.044	0.000	-0.007	0.030	0.061	0.073	-0.007	0.093	0.044	0.000	0.000	0.144
Lapointe-2004-v1	0.088	0.141	0.116	0.147	0.034	0.151	0.179	0.162	0.151	0.088	0.149	0.151	0.171	0.156	0.181
Lapointe-2004-v2	0.008	0.013	0.092	0.082	0.199	0.033	0.153	0.210	0.147	0.028	0.118	0.171	0.155	0.239	0.172
Liang-2005	0.301	0.301	0.236	0.261	0.243	0.301	0.301	0.481	0.301	0.301	0.301	0.301	0.401	0.419	0.481
Nutt-2003-v1	0.171	0.137	0.219	0.311	0.000	0.159	0.156	0.116	0.109	0.086	0.078	0.113	0.142	0.129	0.425
Nutt-2003-v2	-0.025	-0.025	0.138	0.131	0.035	-0.024	-0.025	-0.027	-0.031	-0.025	-0.027	-0.030	-0.030	-0.029	0.435
Nutt-2003-v3	0.063	0.259	0.163	0.169	0.000	0.004	0.080	-0.002	0.059	0.080	0.174	0.059	0.000	0.000	0.642
Pomeroy-2002-v1	0.012	-0.022	0.014	0.007	-0.007	-0.020	-0.006	0.061	-0.020	0.008	-0.026	-0.020	0.111	0.140	0.067
Pomeroy-2002-v2	0.502	0.544	0.443	0.309	0.376	0.591	0.617	0.586	0.568	0.577	0.602	0.568	0.582	0.582	0.246
Ramaswamy-2001	0.618	0.650	0.189	0.258	0.336	0.623	0.651	0.592	0.618	0.620	0.663	0.639	0.635	0.676	0.613
Risinger-2003	0.210	0.194	0.174	0.152	0.000	0.297	0.223	0.309	0.201	0.258	0.153	0.201	0.227	0.248	0.311
Shipp-2002-v1	0.264	0.149	0.079	0.087	0.079	0.208	0.132	0.113	-0.002	0.168	0.203	-0.002	0.134	0.124	0.065
Singh-2002	0.048	0.029	0.032	0.037	0.066	0.019	0.033	0.079	-0.003	0.069	-0.003	0.066	0.034	0.034	0.159
Su-2001	0.666	0.720	0.426	0.496	0.589	0.662	0.738	0.657	0.687	0.650	0.667	0.660	0.725	0.702	0.622
Tomlins	0.396	0.366	0.184	0.196	0.423	0.454	0.409	0.374	0.647	0.469	0.419	0.590	0.485	0.513	0.459
Tomlins-2006-v2	0.368	0.333	0.172	0.094	0.000	0.215	0.292	0.340	0.226	0.383	0.354	0.311	0.348	0.373	0.294
West-2001	0.489	0.413	0.358	0.337	0.00	0.489	0.442	0.258	0.515	0.489	0.442	0.515	0.391	0.391	0.308
Yeoh-2002-v2	0.385	0.343	0.021	0.018	0.000	0.383	0.479	0.405	0.530	0.550	0.337	0.442	0.496	0.465	0.127

Chowdary-2002: Separation of Breast and Colon Cancer

Columns are estimated clusters.
Rows are validated clusters

R-J Cluster (2.23 secs)

Tumors	C1	C2	C3
Breast	61	1	-
Colon	1	35	6

CC hc (20.42 secs)

Tumors	C1	C2	C3
Breast	46	9	7
Colon	36	6	-

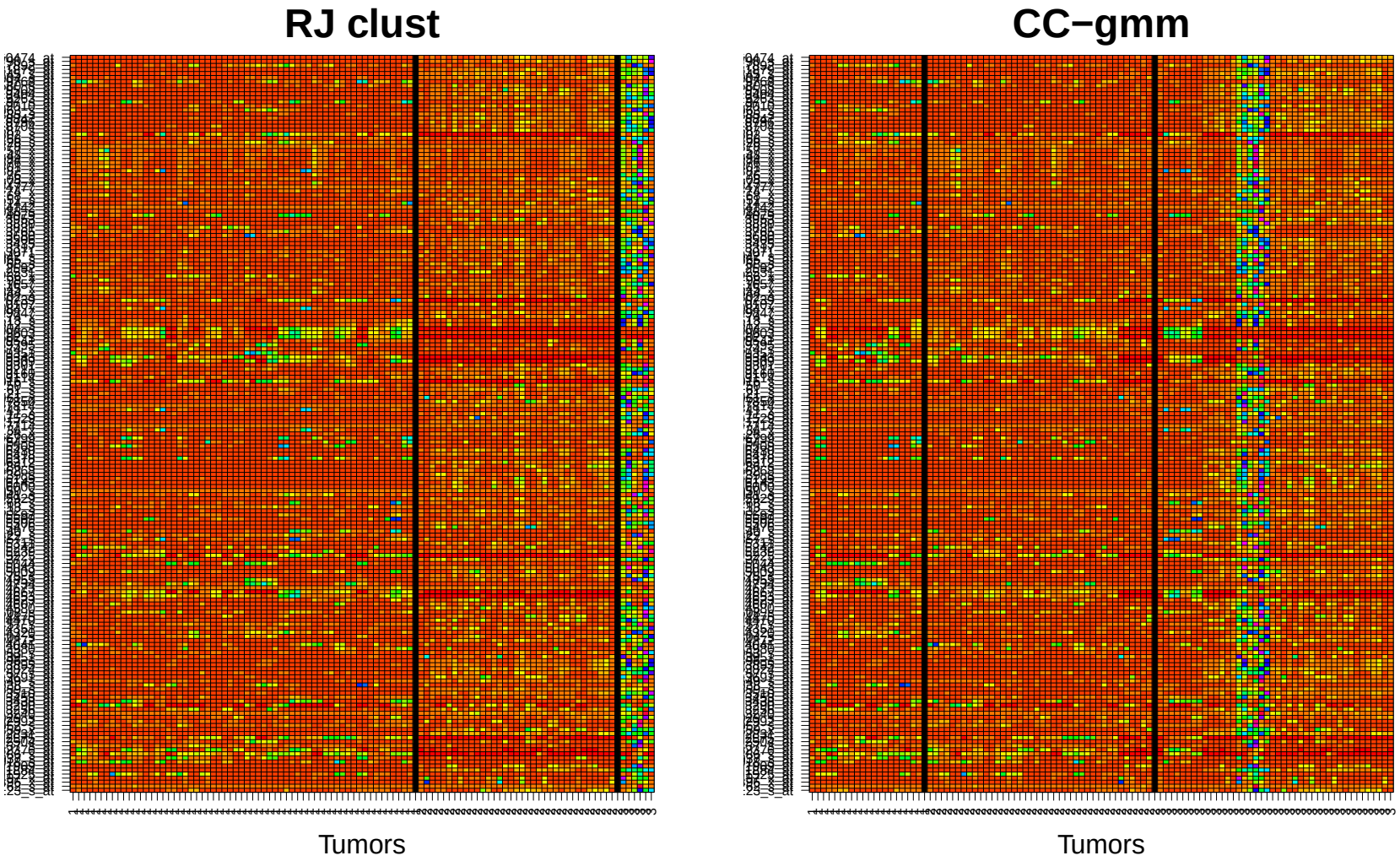
CC k-means (21.53 secs)

Tumors	C1	C2
Breast	53	9
Colon	36	6

CC gmm (41.48 secs)

Tumors	C1	C2	C3
Breast	34	8	20
Colon	7	34	1

Chowdary-2002: Separation of Breast and Colon Cancer



R-J clust found a possible subtype in the colon-cancer.

Dyrskot-2003: Bladder carcinoma Cancer Subtypes

Rows correspond to well-established tumor subtypes, T2+, T1, TA, TA-grade 3 and TA-grade 3 with CIS, and columns to the clusters indentified by each clustering method.

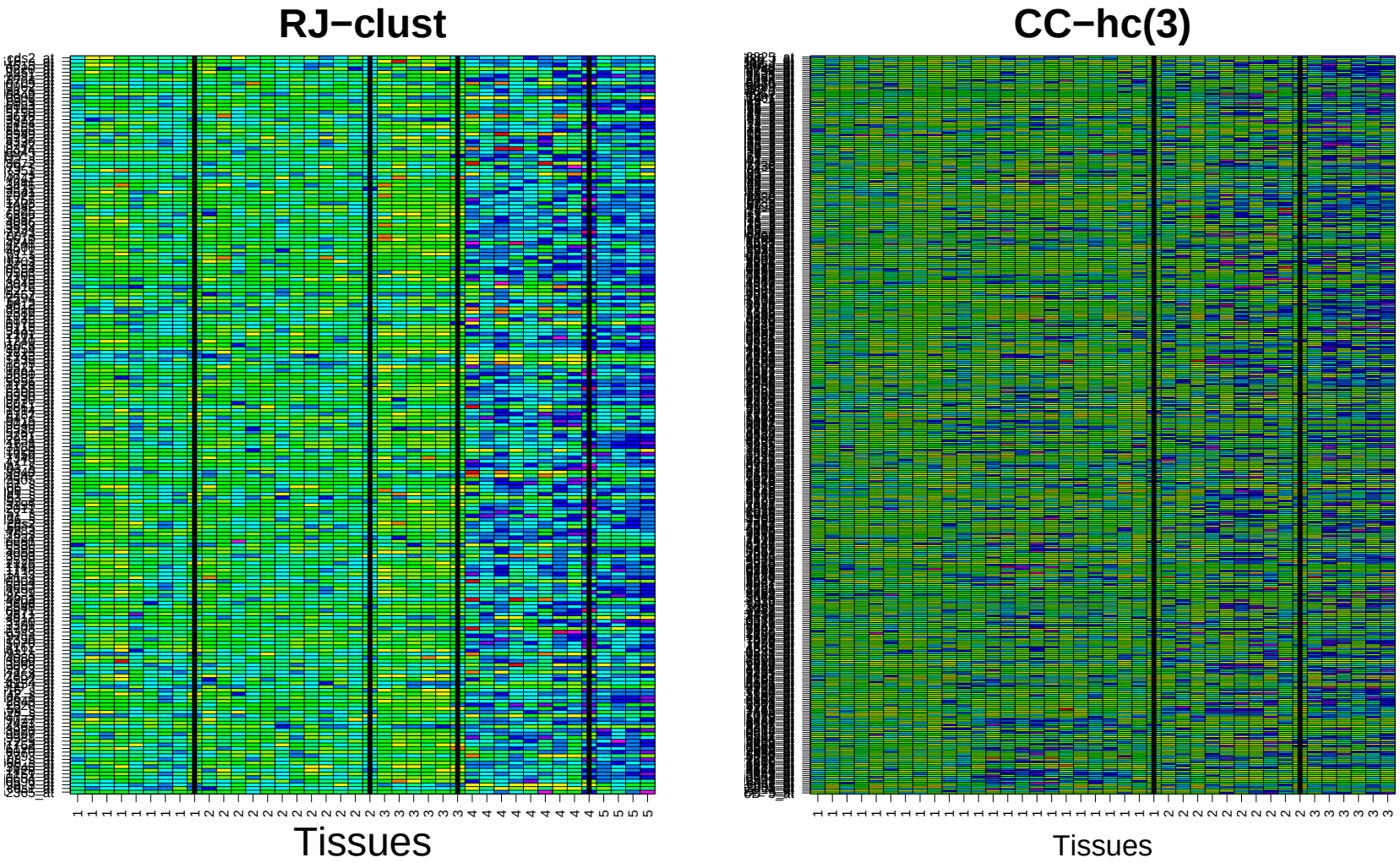
(a) R-J clustering (5.25 secs)

Tumors	C1	C2	C3	C4	C5
T2+	9	-	-	-	-
T1	-	11	-	-	-
TA_2	-	-	5	-	-
TA_3	-	-	1	8	-
TA_{3CIS}	-	1	-	1	4

(b) CC hc (10.24 secs)

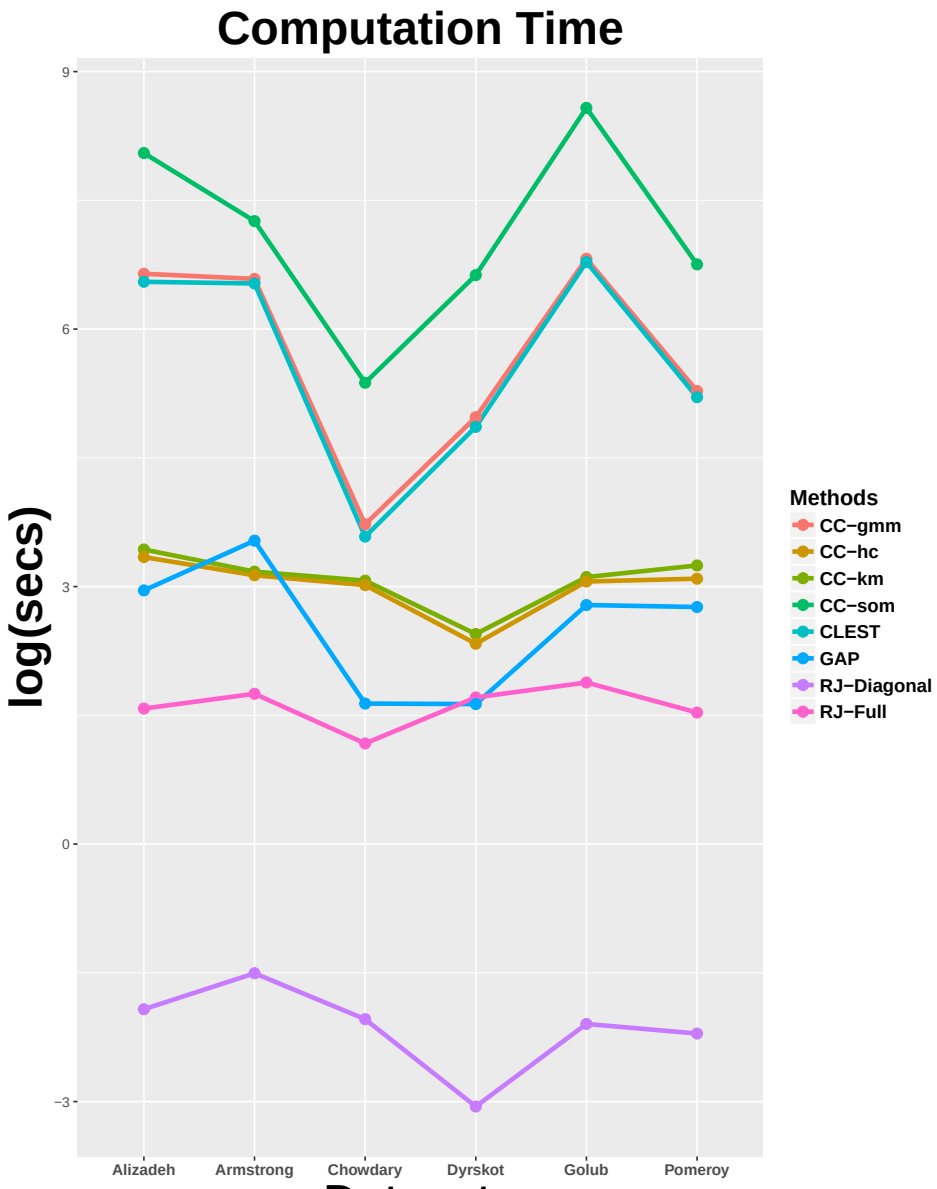
Tumors	C1	C2	C3
T2+	9	-	-
T1	9	2	-
TA_2	-	5	-
TA_3	-	-	9
TA_{3CIS}	6	-	-

Dyrskot-2003: Bladder carcinoma Cancer Subtypes



TA grade 2, grade 3 and grade 3 with CIS are accurately identified by R-J. Grage 3 with CIS is noted for frequent recurrence after treatment.

RJ-clust is exponentially faster!



Summary

- R-J clust is **fast** and **effective** in high dimension setting.
- Computation burden is mostly on N and C **but not on P**.
- R-J clust is **devoid** of tuning parameters.

Manuscript: under submission.
Rcode is available: https://github.com/srahman-24/RJClust.