Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation

application to cancer molecular classification

Dingming Wu, Dongfang Wang, Michael Q. Zhang and Jin Gu

BMC Genomics(2015) 16:1022

Table of contents

Motivation

Motivation

One of the major goal of cancer multi-omics study

- **discover possible cancer subtypes**
- **accurate cancer diagnoses and treatments**

Motivation

Challenges

- handle **different data types** of **different platforms** at the same time
  - **count based** data of sequencing
  - **continuous** data of microarray
  - **binary** data of genetic variations
- the data **dimension** is much **higher** than the sample number
- the **big data volumes** require **efficient** and **robust** computational algorithms

Motivation

Public perspective

- the high-dimensional cancer genomic data can be **reduced** to **a low-dimensional subspace** associated to a few major biological processes
- e.g. sustainable proliferation, apoptosis resistance, activated invasion and immune avoidance

Motivation

Public perspective

- the high-dimensional cancer genomic data can be **reduced** to **a low-dimensional subspace** associated to a few major biological processes
- e.g. sustainable proliferation, apoptosis resistance, activated invasion and immune avoidance

How to find the low-dimensional subspace?

Motivation

iCluster*, iCluster+[✳]

- based on **probabilistic principal component analysis**
- used **generalized linear models** to transform continuous, discretized and count variables as a sparse linear regression on a set of latent driving factors
- cancer subtyping can be done in the **reduced subspace** consisting of the latent driving factors

---

* Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinforma Oxf Engl. 2009;25:290612.

✳ Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A. 2013;110:424550.

Motivation

Bayesian consensus clustering, BCC*

- proposed another **Bayesian latent model**
- simultaneously **find the latent low-dimension subspaces** and **assign samples into different clusters**
- the **low computational efficiency** limits its applications on large-scale cancer omics dataset

* Lock EF, Dunson DB. Bayesian consensus clustering. Bioinforma Oxf Engl. 2013;29:26106.

Motivation

Low-rank approximation, LRA

- one kind of promising **dimension reduction** methods;
- In most cases, LRA is **convex** and can be solved using fast algorithm;
- A few studies show the advantages of LRA for single data type analysis, such as cancer copy number variations.*,⁂

* Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7(1):52342.

⁂ Zhou X, Liu J, Wan X, Yu W. Piecewise-constant and low-rank approximation for identification of recurrent copy number variations. Bioinforma Oxf Engl. 2014;30:19439.

Motivation

LRAcluster

- a novel low-rank approximation based **integrative probabilistic model**;
- deal with **different data types** with **high computational efficiency** and **stability**;
- **assumptions**:
  - a few major biological factors determine a set of high-dimensional but low-rank systems parameters;
  - the observed cancer omics data are generated based on these parameters.

Methods

Methods: LRAcluster overview

LRAcluster

- an **unsupervised** method;
- find the **principal low-dimension subspace** of **large-scale** and **high-dimensional** multi-omics data for **molecular classification**.

Methods: LRAcluster overview

Molecular Feature 1

Somatic Mutation			
	Sample 1	Sample 2	Sample 3
Gene A	1	0	0
Gene B	1	0	1
Gene C	0	1	1

Observed Data Matrix 1

Molecular Feature 2

RNA sequencing			
	Sample 1	Sample 2	Sample 3
Gene D	103	96	132
Gene E	27	42	35

Observed Data Matrix 2

Molecular Feature 3

Copy Number Variation			
	Sample 1	Sample 2	Sample 3
Gene F	-0.87	1.02	0
Gene G	-0.87	0	-0.34
Gene H	0	0.45	-0.34

Observed Data Matrix 3

The **molecular features** (such as somatic mutations, copy number variations, DNA methylations and gene expressions) are expressed as multiple **observed data matrices**.

Methods: LRAcluster overview

Somatic Mutation			
	Sample 1	Sample 2	Sample 3
Gene A	1	0	0
Gene B	1	0	1
Gene C	0	1	1
RNA sequencing			
	Sample 1	Sample 2	Sample 3
Gene D	103	96	132
Gene E	27	42	35
Copy Number Variation			
	Sample 1	Sample 2	Sample 3
Gene F	-0.87	1.02	0
Gene G	-0.87	0	-0.34
Gene H	0	0.45	-0.34

Probabilistic model

$$X_{ij} \sim \Pr(X_{ij} | \Theta_{ij})$$

The probabilistic assumption: each **observed molecular feature** of each sample is a **random variable conditional on a hidden parameter**.

Methods: LRAcluster overview

Somatic Mutation			
	Sample 1	Sample 2	Sample 3
Gene A	1	0	0
Gene B	1	0	1
Gene C	0	1	1
RNA sequencing			
	Sample 1	Sample 2	Sample 3
Gene D	103	96	132
Gene E	27	42	35
Copy Number Variation			
	Sample 1	Sample 2	Sample 3
Gene F	-0.87	1.02	0
Gene G	-0.87	0	-0.34
Gene H	0	0.45	-0.34

Probabilistic model

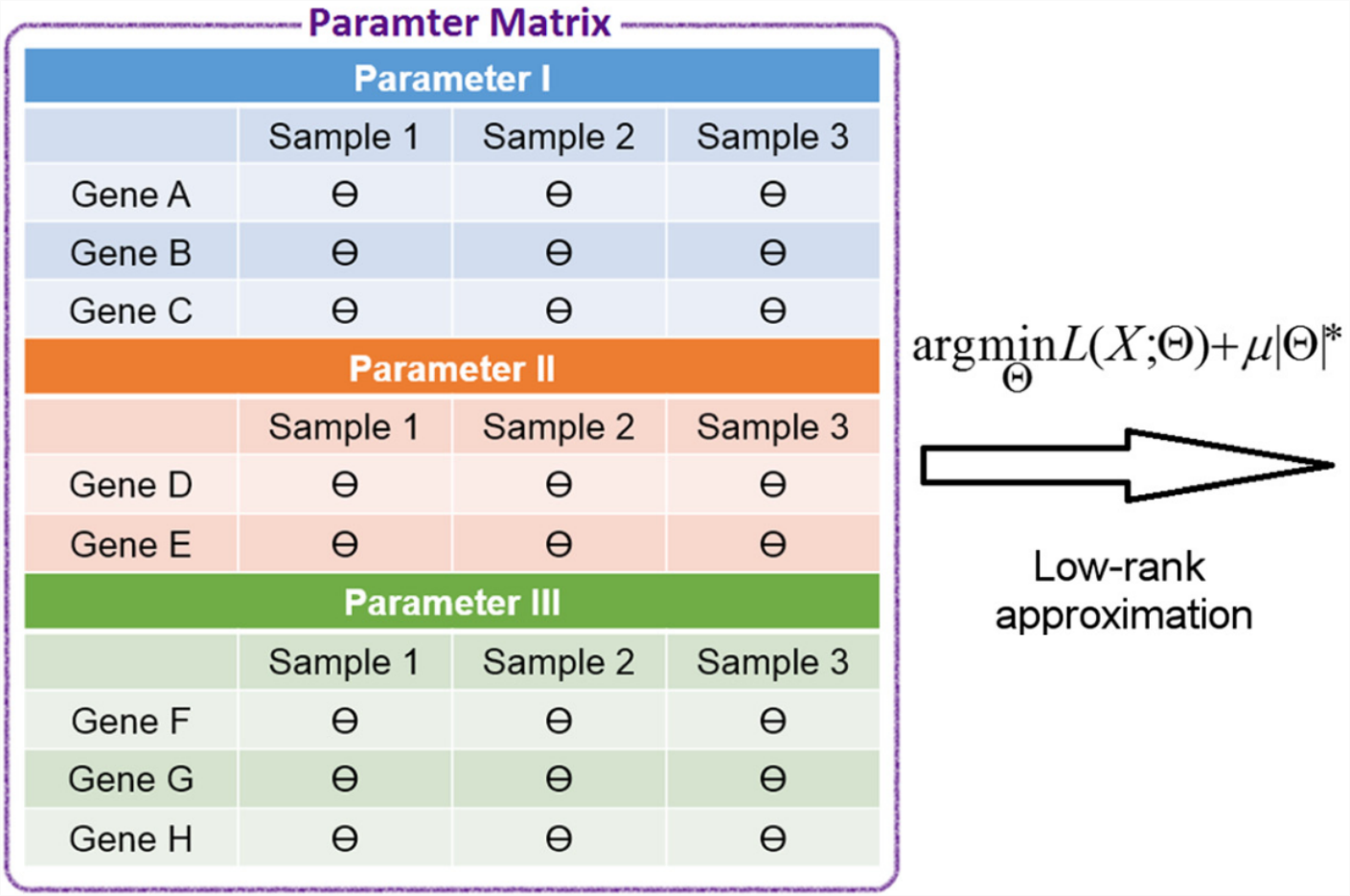
$$X_{ij} \sim \Pr(X_{ij} | \Theta_{ij})$$

Paramter Matrix

Parameter I			
	Sample 1	Sample 2	Sample 3
Gene A	Θ	Θ	Θ
Gene B	Θ	Θ	Θ
Gene C	Θ	Θ	Θ
Parameter II			
	Sample 1	Sample 2	Sample 3
Gene D	Θ	Θ	Θ
Gene E	Θ	Θ	Θ
Parameter III			
	Sample 1	Sample 2	Sample 3
Gene F	Θ	Θ	Θ
Gene G	Θ	Θ	Θ
Gene H	Θ	Θ	Θ

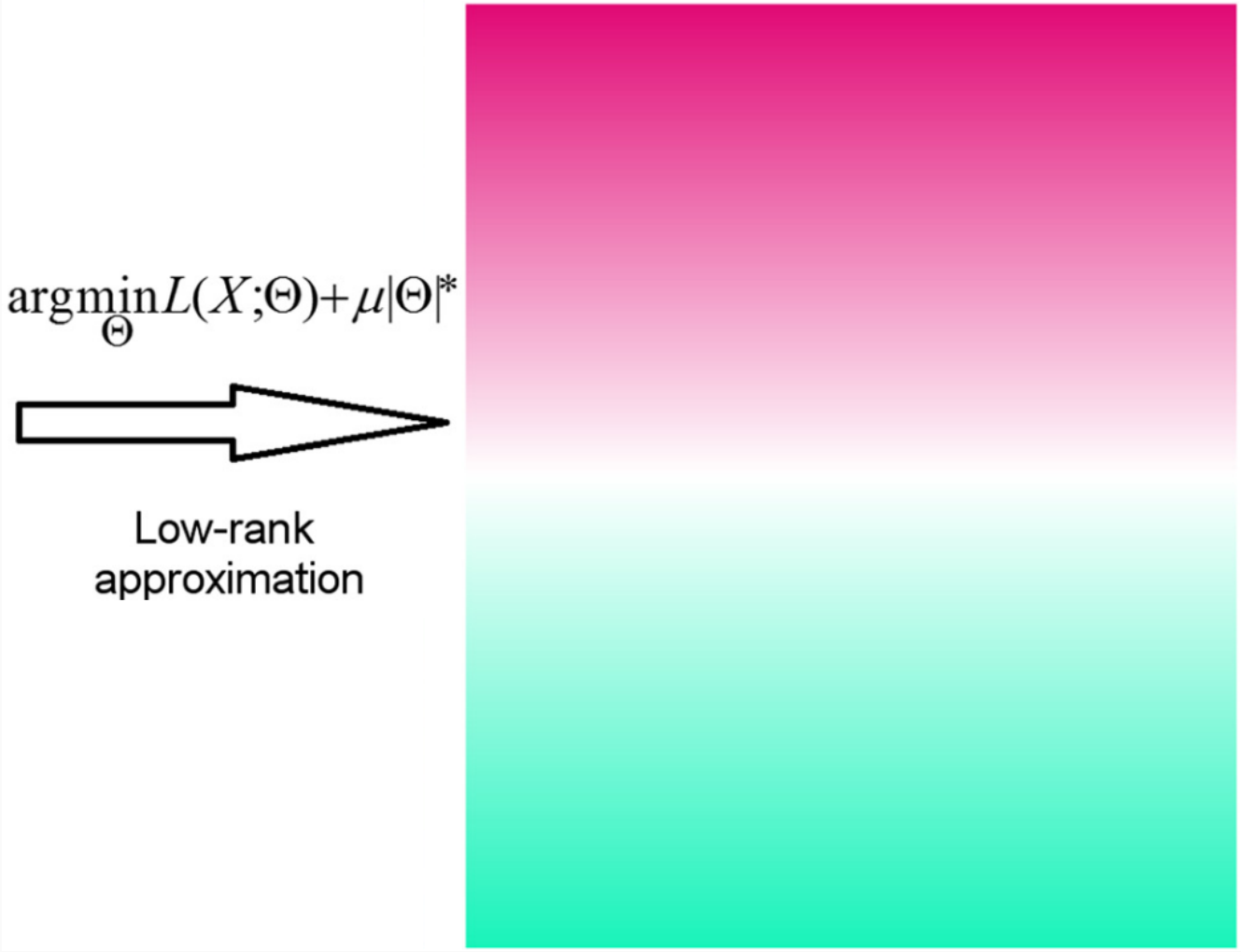
Each observed data matrix is conditional on a **size-matched parameter matrix** and different types of data follow **different probabilistic models**.

Methods: LRAcluster overview

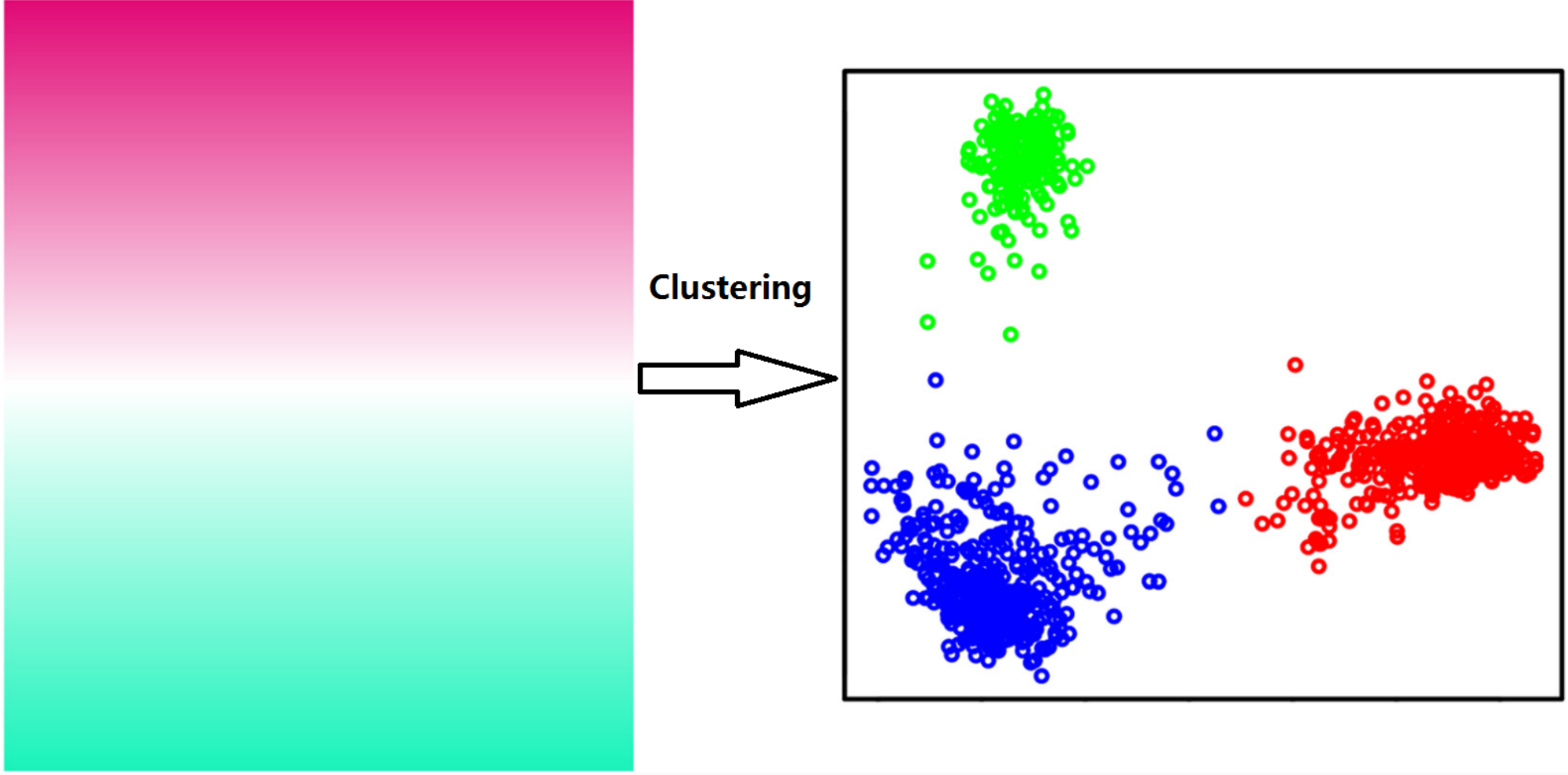


The **low-rank assumption** of the parameter matrix leads to a **penalty function** corresponding to a **structural complexity constraint** of the model.

Methods: LRAcluster overview



Then, the **low-rank parameter matrix** can be decomposed into a **low-dimensional representation of the original data**.


Clustering

Methods: LRAcluster overview

Somatic Mutation			
	Sample 1	Sample 2	Sample 3
Gene A	1	0	0
Gene B	1	0	1
Gene C	0	1	1
RNA sequencing			
	Sample 1	Sample 2	Sample 3
Gene D	103	96	132
Gene E	27	42	35
Copy Number Variation			
	Sample 1	Sample 2	Sample 3
Gene F	-0.87	1.02	0
Gene G	-0.87	0	-0.34
Gene H	0	0.45	-0.34

Probabilistic model

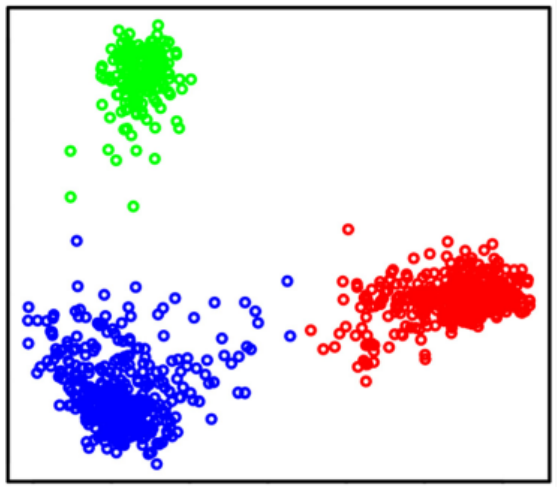
$$X_{ij} \sim \Pr(X_{ij} | \Theta_{ij})$$

Paramter Matrix

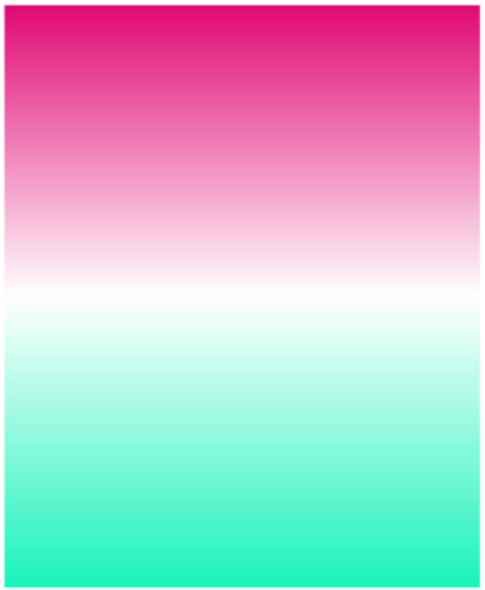
Parameter I			
	Sample 1	Sample 2	Sample 3
Gene A	Θ	Θ	Θ
Gene B	Θ	Θ	Θ
Gene C	Θ	Θ	Θ
Parameter II			
	Sample 1	Sample 2	Sample 3
Gene D	Θ	Θ	Θ
Gene E	Θ	Θ	Θ
Parameter III			
	Sample 1	Sample 2	Sample 3
Gene F	Θ	Θ	Θ
Gene G	Θ	Θ	Θ
Gene H	Θ	Θ	Θ

$$\arg\min_{\Theta} L(X;\Theta) + \mu|\Theta|^*$$

Low-rank approximation



Clustering



Methods: LRAcluster overview

Somatic Mutation			
	Sample 1	Sample 2	Sample 3
Gene A	1	0	0
Gene B	1	0	1
Gene C	0	1	1
RNA sequencing			
	Sample 1	Sample 2	Sample 3
Gene D	103	96	132
Gene E	27	42	35
Copy Number Variation			
	Sample 1	Sample 2	Sample 3
Gene F	-0.87	1.02	0
Gene G	-0.87	0	-0.34
Gene H	0	0.45	-0.34

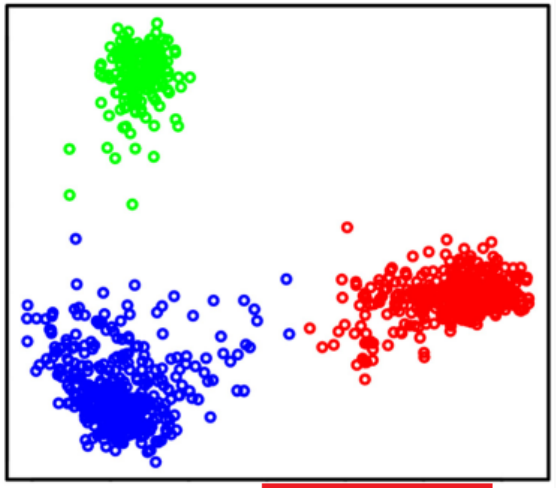
Probabilistic model $X_{ij} \sim \Pr(X_{ij}|\Theta_{ij})$

Parameter Matrix

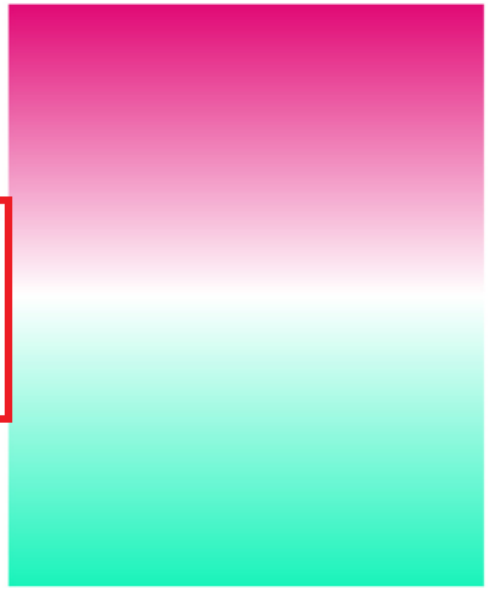
Parameter I			
	Sample 1	Sample 2	Sample 3
Gene A	Θ	Θ	Θ
Gene B	Θ	Θ	Θ
Gene C	Θ	Θ	Θ
Parameter II			
	Sample 1	Sample 2	Sample 3
Gene D	Θ	Θ	Θ
Gene E	Θ	Θ	Θ
Parameter III			
	Sample 1	Sample 2	Sample 3
Gene F	Θ	Θ	Θ
Gene G	Θ	Θ	Θ
Gene H	Θ	Θ	Θ

$\arg\min_{\Theta} L(X;\Theta)+\mu|\Theta|^*$

Low-rank approximation



Clustering



Methods: Probabilistic model

Donation

- The k-th type of omics data are denoted as $X^{(k)}$.
- x_{ij}: the **row** index represents the i-th **molecular feature** and the **column** index represent the j-th **sample**.
- $\Theta^{(k)}$ denotes the size-matched parameter matrix of $X^{(k)}$.

Methods: Probabilistic model

The probabilistic model specifies **the probability density (mass) function** of the observations given the parameters for each data type:

- $\Pr\left(X_{ij}^{(k)} \mid \Theta_{ij}^{(k)}\right) \propto \exp\left(-\frac{1}{2}\left(X_{ij}^{(k)} - \Theta_{ij}^{(k)}\right)^2\right)$ for **real-type data**, **Gaussian distribution** (CNV and DNA methylation data here);
- $$\Pr\left(X_{ij}^{(k)} \mid \Theta_{ij}^{(k)}\right) = \frac{e^{\Theta_{ij}^{(k)}}}{1+e^{\Theta_{ij}^{(k)}}} I\left(X_{ij}^{(k)} = 1\right) + \frac{1}{1+e^{\Theta_{ij}^{(k)}}} I\left(X_{ij}^{(k)} = 0\right)$$ for **binary data**, **Bernoulli distribution** (Somatic mutation data);
- $\Pr\left(X_{ij}^{(k)} \mid \Theta_{ij}^{(k)}\right) \propto \left(\lambda_{ij}^{(k)}\right)^{X_{ij}^{(k)}} e^{\left(-\lambda_{ij}^{(k)}\right)}, \lambda_{ij}^{(k)} = e^{\Theta_{ij}^{(k)}}$ for **count data**, **Poisson distribution** (RNAseq normalized count data here).
- **Categorical data** can be transformed using dummy code and thus can be treated as **binary variables**.

Methods: Probabilistic model

The **likelihood function** of above probabilistic model is written as the minus log of the probability density (mass) function:

$$L\left(\Theta^{(k)},;X^{(k)}\right) = -\sum_{ij} \ln\left(\Pr\left(X_{ij}^{(k)} \mid \Theta_{ij}^{(k)}\right)\right) \tag{1}$$

For integrative analysis

- two or more observed data matrices $X^{(k)}(k = 1, 2, \ldots, K)$.
- the overall parameter matrix Θ stacks all the parameter matrices $\Theta^{(k)}$ used for each observed data matrix.
- The overall likelihood function is the sum of the likelihood functions of different data types:

$$L(\Theta) = \sum_{k} L\left(\Theta^{(k)},;X^{(k)}\right) \tag{2}$$

Methods: Probabilistic model

Low-rank assumptions

- The probabilistic model assumes that the observations X_{ij} are independently distributed conditional on the ultrahigh dimensional parameter matrix Θ.
- The prior assumption of the model is that Θ has low-rank structure.
- The low-rank assumption is used to penalize the freedom of the model and eventually leads to the following optimization problem:

$$\arg\min_{\Theta} L(\Theta) + \mu \mid \Theta \mid^{*} \tag{3}$$

where μ is a tuning parameter and $\mid \cdot \mid^{*}$ denotes the nuclear norm of the matrix.

Methods: Fast low-rank approximation

Fast low-rank approximation

- The solution of the optimization problem (3) mimics a **singular value thresholding (SVT)** method* which suggests a general framework to solve the optimization problem $\arg\min_{\Theta} f(\Theta) + \mu \mid \Theta \mid^{*}$ where f is a convex function.

* Cai JF, Cands EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20:195682

Methods: Fast low-rank approximation

The iterative solution framework steps:

(1) initialize Θ^0 and iterate the following two steps until convergence;

(2) $\Theta^{2n+1} = \Theta^{2n} - \delta_n \nabla f$;

(3) $\Theta^{2n+2} = D_\mu\left(\Theta^{2n+1}\right)$.

- ∇f is the gradient of the un-regularized likelihood function $L(\Theta) = \sum_k L\left(\Theta^{(k)},; X^{(k)}\right)$ and δ_n is the step length.
- D_μ represents the "singular value shrinkage operator": denote the SVD of a matrix Θ as $\Theta = U\Sigma V^T$, then $D_\mu\left(\Theta\right) = U D_\mu\left(\Sigma\right) V^T$.
- $D_\mu\left(\Sigma\right)$ is a diagonal matrix with the same size as Σ and each diagonal element is the shrinkage of the singular value of Σ.
- For a positive singular value λ, the shrinkage result is $(\lambda - \mu)$ when $\lambda > \mu$ and 0 when $\lambda \leq \mu$.

Methods: Fast low-rank approximation

Initializations

- The objective function of LRAcluster is **convex**, so any initial value of the iteration will converge to the global minimum.
- LRAcluster simply initializes Θ as a zero matrix.
- The user defined parameter μ is hard to choose in practical use.
- Instead of μ, LRAcluster receives the **rank** r (also **the target dimension**) as the user defined constraint parameter , and μ is automatically chosen as the **rank** $r+1$ **largest singular value** in each iteration, which is to guarantee that Θ has rank r and the shrinkage has minimal effect on Θ.
- The algorithm converges definitely when step length $\delta \in [0.5, 2]$*, so we set $\delta = 0.5$ which ensures convergence for real applications.

* Cai JF, Cands EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20:195682

Methods: Fast low-rank approximation

How to determine r

- The target rank (or dimension) r is the **only user-defined parameter** in dimension reduction step.
- The log likelihood $L(\theta; X)$ corresponding to the optimized solution θ^* (denoted as L_r^*) is used for guiding the choice of parameter r.
- For the same dataset, larger r means weaker penalization of the model freedom and leads to better data fitting (larger likelihood L_r^*).
- Thus, $L_{r=0}^*$is the minimum and $L_{r=+\infty}^*$ is the maximum among all the L_r^*.
- The quantity L_r^* describes to what extend the model fits the data.
- As LRAcluster mainly deals with large dataset, L_r^* is usually a big value.
- So, instead of L_r^*, LRAcluster uses the normalized quantity $\frac{L_{r=+\infty}^* - L_r^*}{L_{r=+\infty}^* - L_{r=0}^*}$ (between 0 and 1) as explained variation for choosing a desirable rank r.

Methods: Dimension reduction and clustering

Dimension reduction

- The dimension reduction is straightforward after getting the low-rank matrix Θ.
- As the rank of Θ is no more than r, the SVD of that matrix $\Theta = U\Sigma V^T$ has Σ with no more than r non-zero singular values.
- So the first r columns of ΣV^T are the dimension reduction result of the original data matrix X with the target dimension (rank) r.

Methods: Dimension reduction and clustering

Clustering

- LRAcluster uses k-means to identify the candidate molecular subtypes in the reduced low-dimensional subspace.
- Silhouette values* is used to determine the cluster number k.
- Any other unsupervised clustering algorithm can be used instead of k-means.

* Rousseeuw P. silhouettes - A graphical aid to the integration of clusteranalysis. J Comput Appl Math. 1987;20:5365.

Case Study

Case Study: Datasets

Datasets

- The datasets were downloaded from publicly released TCGA level 3 data (processed data from UCSC Cancer Genome Browser*).
- The whole dataset consists of 11 types of cancer (BRCA, COAD, GBM, HNSC, KIRC, LGG, LUAD, LUSC, PRAD, STAD, and THCA) with **somatic mutations**, **copy number variations**, **DNA methylations** and **gene expressions**.

* Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, et al. The UCSC Cancer Genomics Browser: update 2015. Nucleic Acids Res. 2015; 43(Database issue):D812817

Case Study: Datasets

Somatic mutation and copy number variation data

- Our preliminary studies indicate that the massive passenger variations of the complete datasets deteriorated the clustering stability.
- Thus, only the somatic mutations and copy number variations of the ∼500 genes reported as causally implicated in cancer in COSMIC* were included in this study.

* Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43(Database issue):D805811.

Case Study: Datasets

DNA methylation data

- Using Illumina HumanMethylation450 BeadChip (450 k array), probes annotated as promoter-associated (based on the annotations of IlluminaHumanMethylation450k.db*) were selected (if a gene has multiple promoter associated probes, only one of them was chosen).
- Overall, ∼8,000 probes were used.

Gene expression data

- The normalized count-based data from RNA-Seq were all included with ∼20,000 genes.

* Triche T, Jr. IlluminaHumanMethylation450k.db: Illumina Human Methylation 450k annotation data.

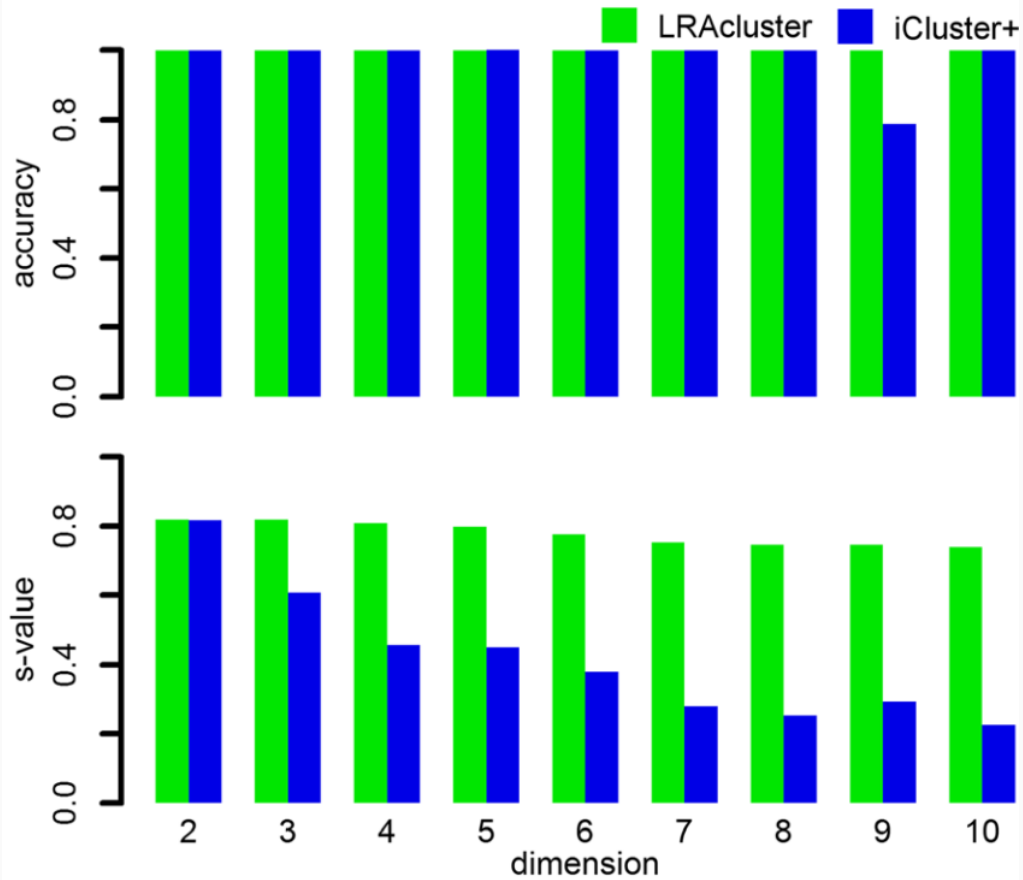
Case Study: The computational performances of LRAcluster

The computational performances of LRAcluster

- The three cancer-type testing dataset consists of **BRCA, COAD, LUAD** cancer types with **RNA-Seq** and **DNA methylation data**, which was used to compare the **clustering performances** and **time consumption** between **LRAcluster** and **iCluster+**.
- The molecular features (genes for expression data and probes for DNA methylation data) with **largest variances** across all samples are selected to construct datasets of different sizes.
- The smallest dataset containing **top 100 molecular features** of each data type is used to test LRAcluster and iCluster+'s clustering performances with different target dimension.
- Time consumption of the two methods was recorded for datasets with **different feature sizes**.

Case Study: The computational performances of LRAcluster

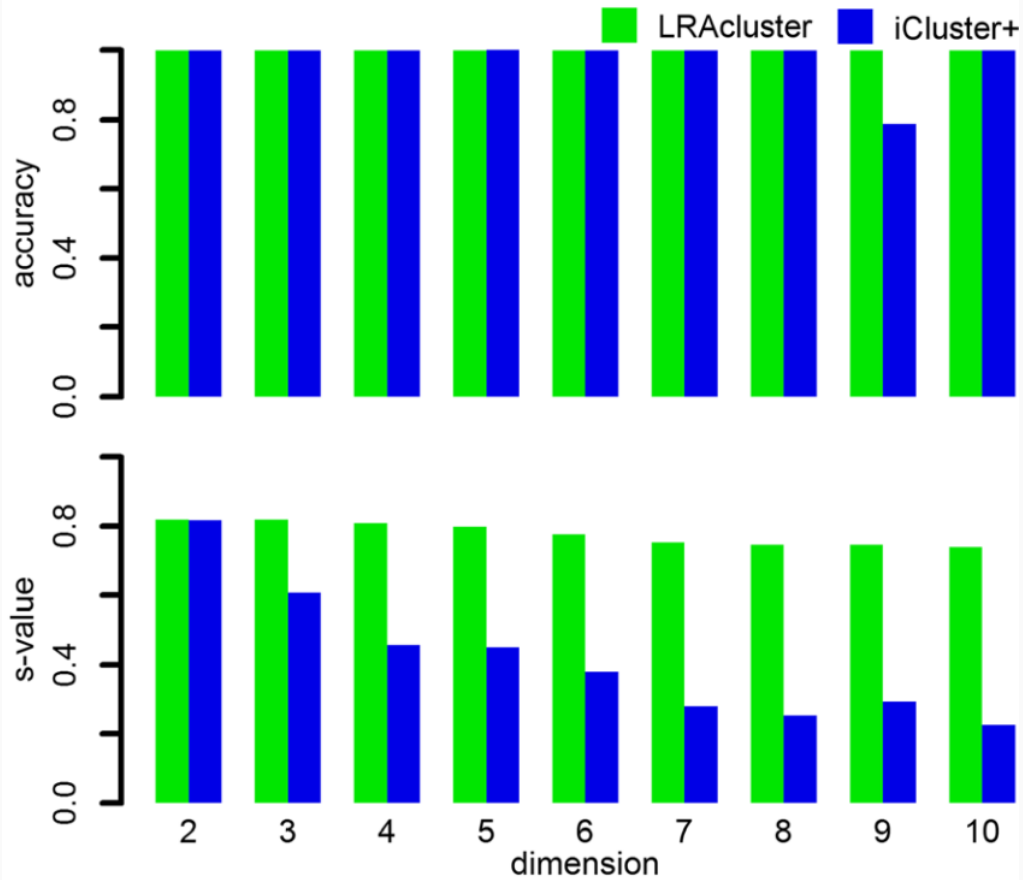
Clustering performances



Both LRAcluster and iCluster+ got high classification **accuracy** for the three cancer types in the reduced low-dimension subspaces.

Case Study: The computational performances of LRAcluster

Clustering performances



But, the **silhouette values** show that LRAcluster is superior to iCluster+, especially when the target dimension is large.

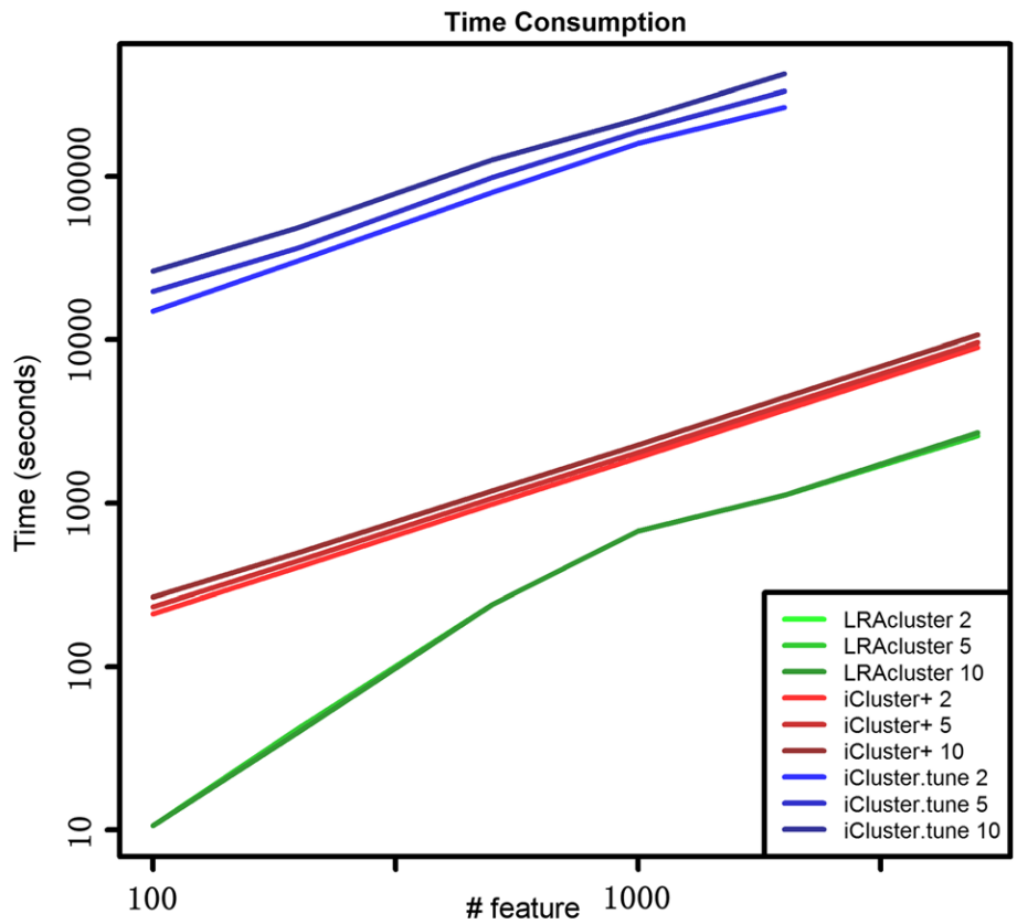
Case Study: The computational performances of LRAcluster

Clustering performances

- These results indicate that iCluster+ will encounter **local optimal problems** when the model becomes complex.
- The **convexity** of LRAcluster model ensures stable model fitting.

Case Study: The computational performances of LRAcluster

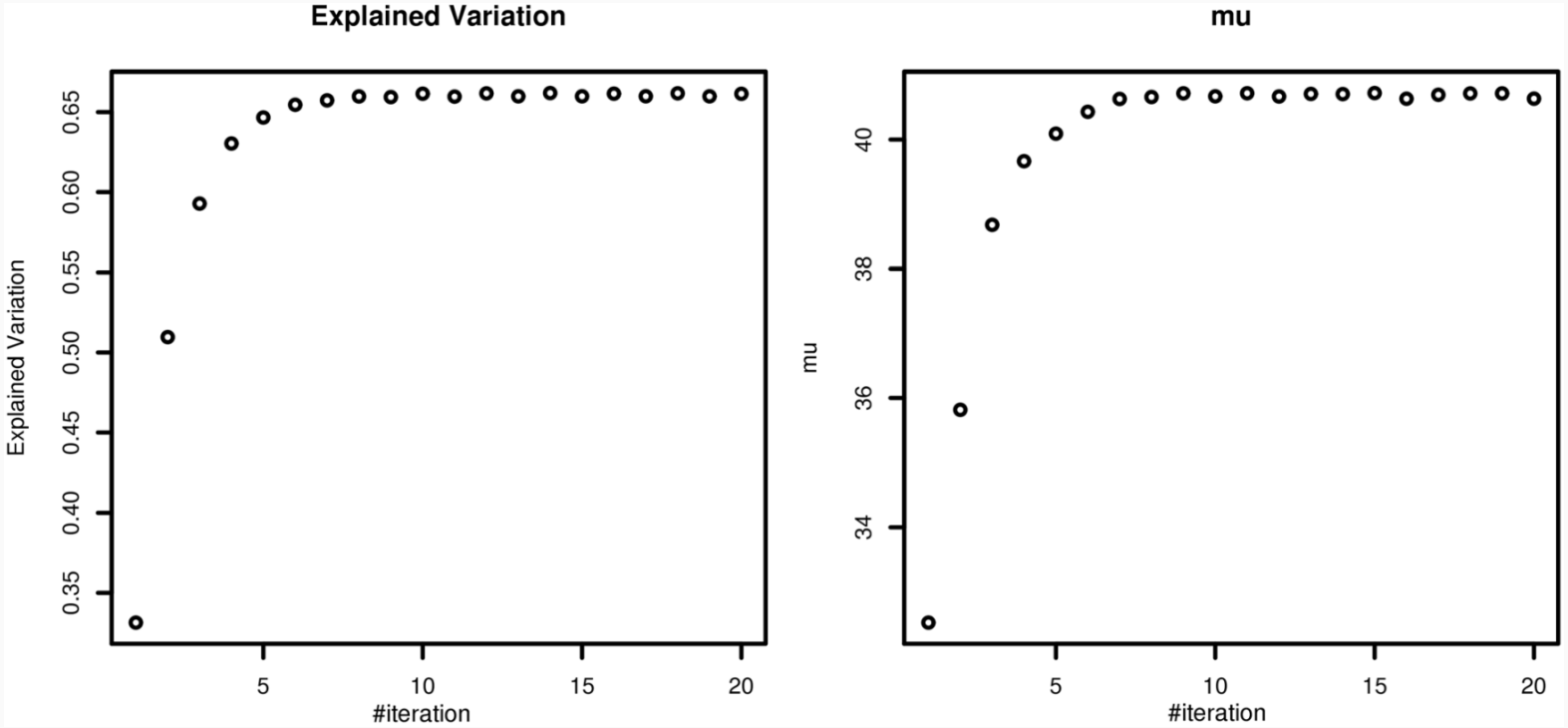
Time consumption



LRAcluster runs $\sim$ 5 fold faster than iCluster+ with fixed penalty parameter and much faster ($\sim$300 fold) if that parameter is optimized.

Case Study: The computational performances of LRAcluster

Convergence



The dynamic changes of the "explained variance" and the penalty parameter μ demonstrated that LRAcluster can quickly converge within only a few iterations.

Case Study: The computational performances of LRAcluster

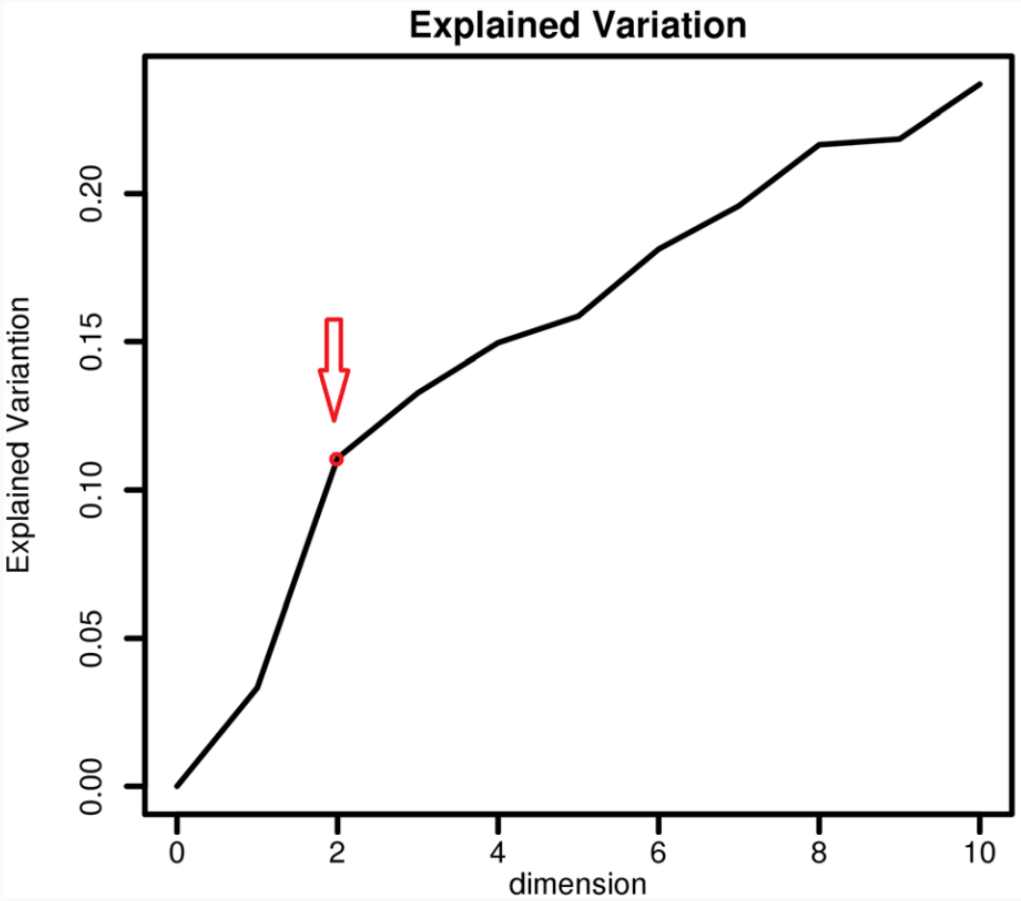
Parameters

- **Rank** (or **dimension**) of the reduced subspace r:
  can be chosed according to the **curve** of **explained variance**;
- **Cluster number** c:
  can be chosen according to the **curve** of **silhouette value (s-value)**.

Used the BRCA dataset with known ER+/ER- subtypes as an example.

Case Study: The computational performances of LRAcluster

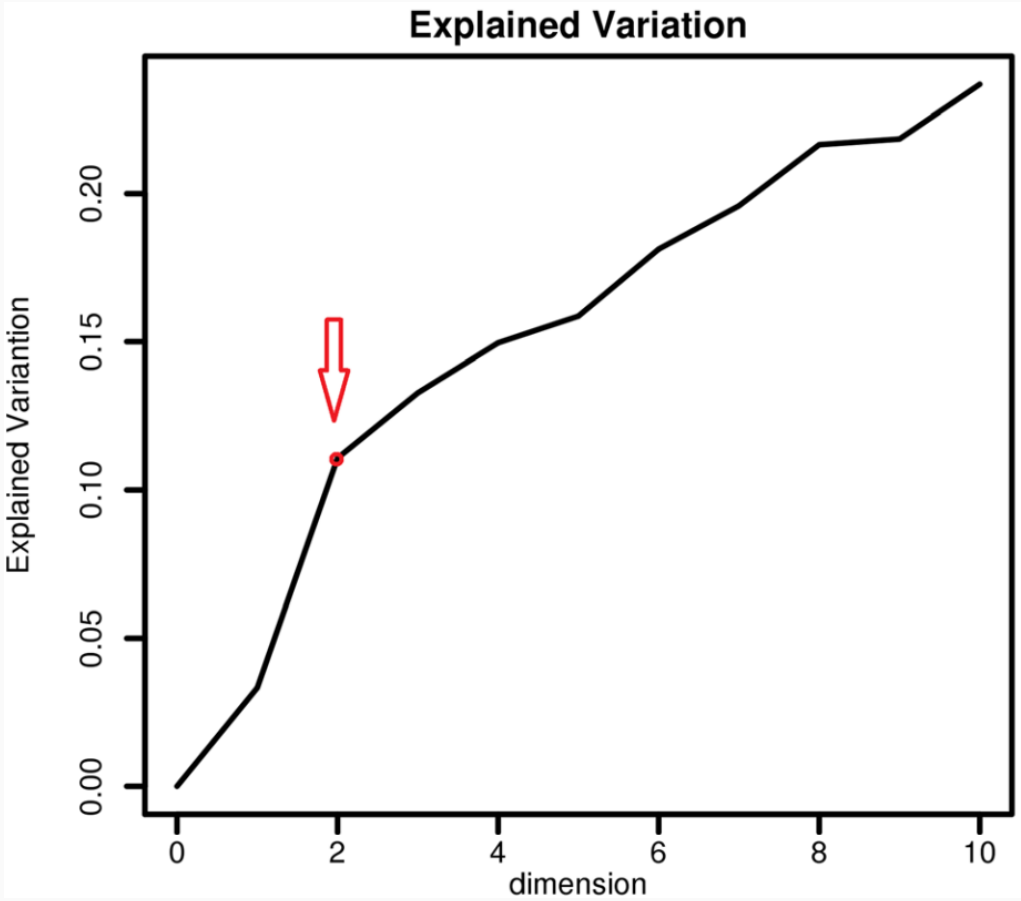
Parameters: Rank r



For the BRCA dataset, dimension r should be chosen as 2, because there was a turning point at 2 on the curve of the explained variance.

Case Study: The computational performances of LRAcluster

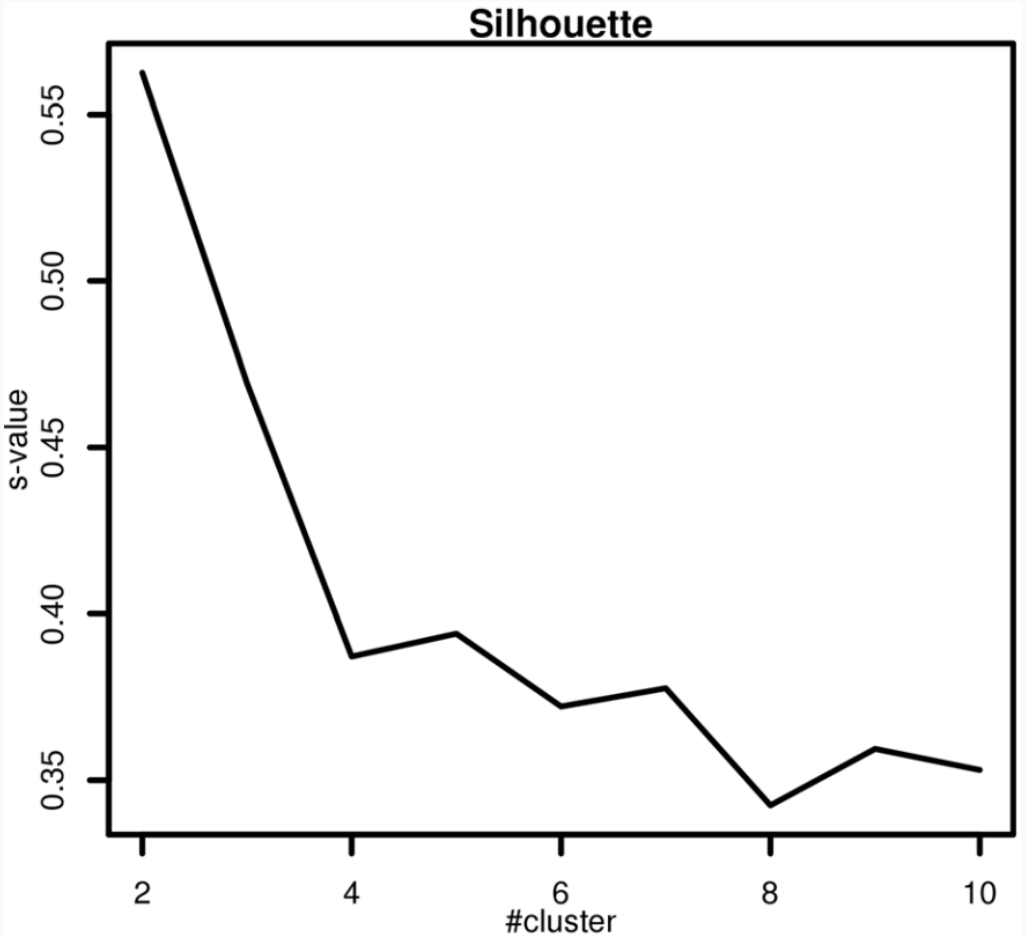
Parameters: Rank r



This empirical rule is based on the principle that the increase of model fitness is much slower after the changing point.

Case Study: The computational performances of LRAcluster

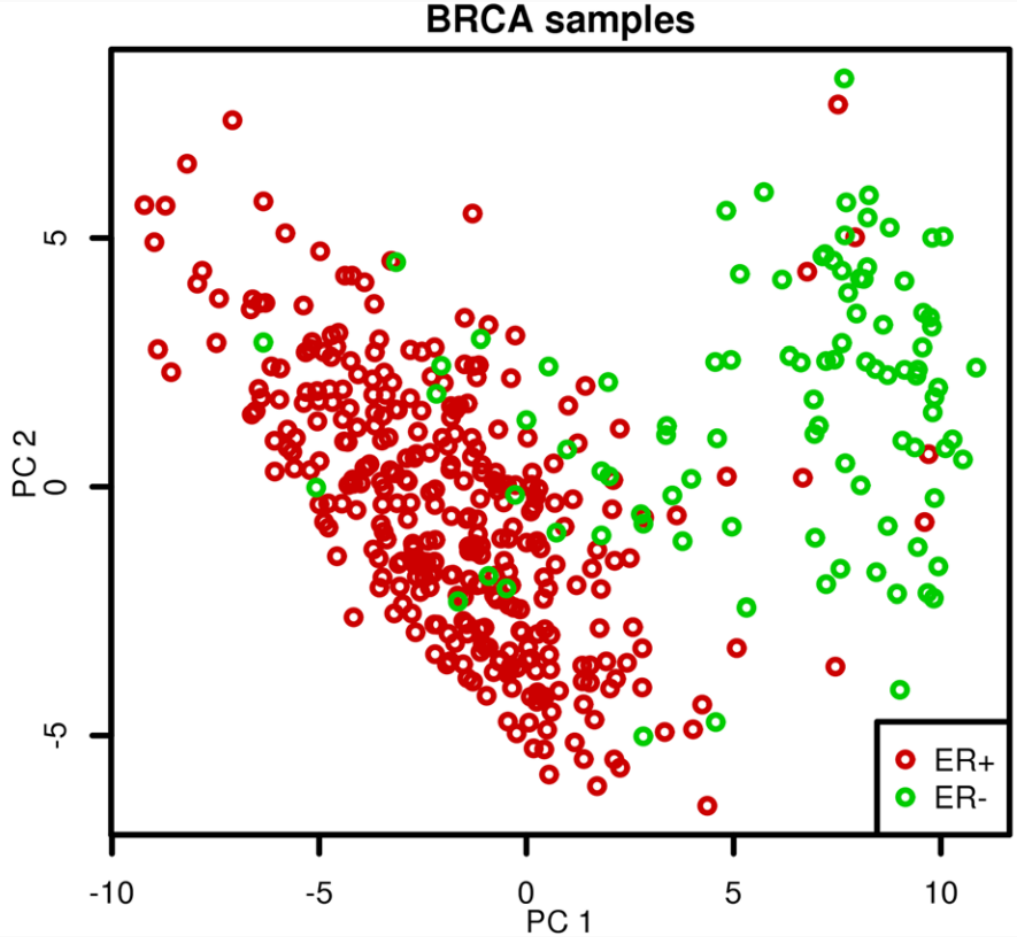
Parameters: Cluster number c



Larger s-value indicates better clustering performance. For the BRCA dataset, the largest s-value was achieved when $c = 2$.

Case Study: The computational performances of LRAcluster

LRAcluster Clustering performance on BRCA dataset



LRAcluster find two subtypes highly consistent with known ER+/ER subtypes in the reduced 2-dimensional subspace (accuracy 92.1%).

Case Study: Application on TCGA pan-cancer dataset

Application on the large-scale TCGA pan-cancer dataset

- 11 different cancer types
- 3,319 samples
- 4 different data types including somatic mutations, copy number variations, DNA methylations, and gene expressions
- we got ten clusters in the reduced ten-dimension subspace

Case Study: Application on TCGA pan-cancer dataset

Application on the large-scale TCGA pan-cancer dataset

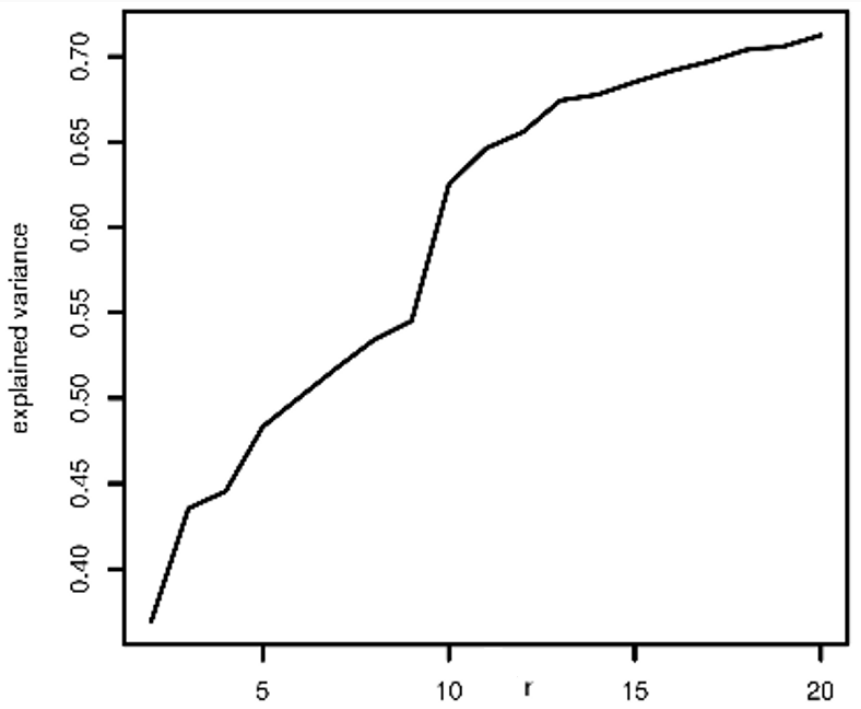


Figure S1. "Explained variance" against the parameter r.

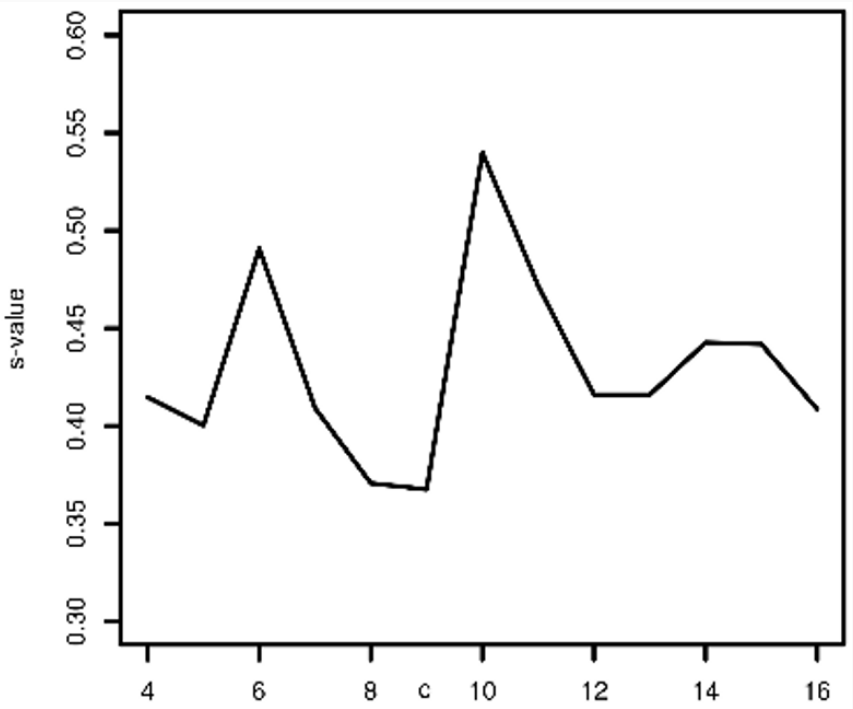


Figure S2. Silhouette value against cluster number.

The dimension and the cluster number were determined according to the curves of **explained variances** and **s-values**, respectively.

Case Study: Application on TCGA pan-cancer dataset

Application on the large-scale TCGA pan-cancer dataset

Table 1 The unsupervised clustering results of pan-cancer analysis

	BRCA	COAD	GBM	HNSC	KIRC	LGG	LUAD	LUSC	PRAD	STAD	THCA	Total
C1	1	0	0	286	0	0	0	6	0	0	0	293
C2	0	0	0	0	0	1	0	0	0	0	411	412
C3	0	0	41	0	0	451	0	0	0	0	0	492
C4	0	0	0	0	0	0	0	0	0	231	0	231
C5	0	0	0	0	0	0	0	0	293	0	0	293
C6	0	190	0	1	0	0	2	0	1	0	0	194
C7	3	17	0	0	1	0	406	7	0	0	3	437
C8	0	0	0	0	240	0	0	0	0	0	0	240
C9	448	0	1	2	1	0	4	1	0	0	0	457
C10	8	1	0	195	0	0	6	60	0	0	0	270
Total	460	208	42	484	242	452	418	74	294	231	414	3319

Results: most samples from the same cancer types are grouped as independent clusters, similar with a recent pan-cancer study*.

* Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158:92944.

Case Study: Application on TCGA pan-cancer dataset

Application on the large-scale TCGA pan-cancer dataset

Table 1 The unsupervised clustering results of pan-cancer analysis

	BRCA	COAD	GBM	HNSC	KIRC	LGG	LUAD	LUSC	PRAD	STAD	THCA	Total
C1	1	0	0	286	0	0	0	6	0	0	0	293
C2	0	0	0	0	0	1	0	0	0	0	411	412
C3	0	0	41	0	0	451	0	0	0	0	0	492
C4	0	0	0	0	0	0	0	0	0	231	0	231
C5	0	0	0	0	0	0	0	0	293	0	0	293
C6	0	190	0	1	0	0	2	0	1	0	0	194
C7	3	17	0	0	1	0	406	7	0	0	3	437
C8	0	0	0	0	240	0	0	0	0	0	0	240
C9	448	0	1	2	1	0	4	1	0	0	0	457
C10	8	1	0	195	0	0	6	60	0	0	0	270
Total	460	208	42	484	242	452	418	74	294	231	414	3319

The two brain cancers LGG and GBM are grouped together as Cluster C3.

Case Study: Application on TCGA pan-cancer dataset

Application on the large-scale TCGA pan-cancer dataset

Table 1 The unsupervised clustering results of pan-cancer analysis

	BRCA	COAD	GBM	HNSC	KIRC	LGG	LUAD	LUSC	PRAD	STAD	THCA	Total
C1	1	0	0	286	0	0	0	6	0	0	0	293
C2	0	0	0	0	0	1	0	0	0	0	411	412
C3	0	0	41	0	0	451	0	0	0	0	0	492
C4	0	0	0	0	0	0	0	0	0	231	0	231
C5	0	0	0	0	0	0	0	0	293	0	0	293
C6	0	190	0	1	0	0	2	0	1	0	0	194
C7	3	17	0	0	1	0	406	7	0	0	3	437
C8	0	0	0	0	240	0	0	0	0	0	0	240
C9	448	0	1	2	1	0	4	1	0	0	0	457
C10	8	1	0	195	0	0	6	60	0	0	0	270
Total	460	208	42	484	242	452	418	74	294	231	414	3319

HNSC are separated into Cluster C1 & C10: 40.3 % of HNSC samples in Cluster C10 are clustered together with 81.1 % of LUSC samples.

Application on the large-scale TCGA pan-cancer dataset

Table 1 The unsupervised clustering results of pan-cancer analysis

	BRCA	COAD	GBM	HNSC	KIRC	LGG	LUAD	LUSC	PRAD	STAD	THCA	Total
C1	1	0	0	286	0	0	0	6	0	0	0	293
C2	0	0	0	0	0	1	0	0	0	0	411	412
C3	0	0	41	0	0	451	0	0	0	0	0	492
C4	0	0	0	0	0	0	0	0	0	231	0	231
C5	0	0	0	0	0	0	0	0	293	0	0	293
C6	0	190	0	1	0	0	2	0	1	0	0	194
C7	3	17	0	0	1	0	406	7	0	0	3	437
C8	0	0	0	0	240	0	0	0	0	0	0	240
C9	448	0	1	2	1	0	4	1	0	0	0	457
C10	8	1	0	195	0	0	6	60	0	0	0	270
Total	460	208	42	484	242	452	418	74	294	231	414	3319

It's indicates that the squamous carcinomas of different tissue origins may share some common molecular mechanisms.

Case Study: Application on TCGA pan-cancer dataset

Application on the large-scale TCGA pan-cancer dataset

- jNBS* reported similar results with LRAcluster: most of cancer types are separately clustered according to their tissue origin, and two types of squamous carcinomas, head/neck squamous carcinoma and lung squamous carcinoma are cluster together.
- But it found more cross tissue type clusters, because the jNBS analysis only used genetic (mutation & CNV) and epigenetic (DNA methylation) data, the results are hard to be directly compared.

* Liu Z, Zhang S. Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. BMC Genomics. 2015;16:503

Case Study: Application on TCGA pan-cancer dataset

Applied LRAcluster on the 11 cancer types separately

Table 2 The results of single-cancer analysis

Cancer	Dimension[a]	#Cluster[b]	Silhouette values
BRCA	2	2	0.55
COAD	4	4	0.40
GBM	8	2	0.35
HNSC	7	3	0.26
KIRC	6	2	0.36
LGG	2	3	0.44
LUAD	5	2	0.34
LUSC	5	4	0.32
PRAD	2	4	0.41
STAD	4	3	0.37
THCA	2	2	0.61

[a]The dimension of the reduced space is determined according to the curve of the explained variations of each cancer type
[b]The number of clusters is determined according to the curve of the within cluster variances

The omics data have different subtyping abilities of different cancer types.

Case Study: Application on TCGA pan-cancer dataset

Applied LRAcluster on the 11 cancer types separately

Table 2 The results of single-cancer analysis

Cancer	Dimension[a]	#Cluster[b]	Silhouette values
BRCA	2	2	0.55
COAD	4	4	0.40
GBM	8	2	0.35
HNSC	7	3	0.26
KIRC	6	2	0.36
LGG	2	3	0.44
LUAD	5	2	0.34
LUSC	5	4	0.32
PRAD	2	4	0.41
STAD	4	3	0.37
THCA	2	2	0.61

[a]The dimension of the reduced space is determined according to the curve of the explained variations of each cancer type
[b]The number of clusters is determined according to the curve of the within cluster variances

BRCA, LGG, PRAD, and THCA datasets get high silhouette values.

Case Study: Application on TCGA pan-cancer dataset

Applied LRAcluster on the 11 cancer types separately

Table 2 The results of single-cancer analysis

Cancer	Dimension[a]	#Cluster[b]	Silhouette values
BRCA	2	2	0.55
COAD	4	4	0.40
GBM	8	2	0.35
HNSC	7	3	0.26
KIRC	6	2	0.36
LGG	2	3	0.44
LUAD	5	2	0.34
LUSC	5	4	0.32
PRAD	2	4	0.41
STAD	4	3	0.37
THCA	2	2	0.61

[a]The dimension of the reduced space is determined according to the curve of the explained variations of each cancer type
[b]The number of clusters is determined according to the curve of the within cluster variances

The BRCA subtypes are significantly associated with ER status.

Case Study: Application on TCGA pan-cancer dataset

Applied LRAcluster on the 11 cancer types separately

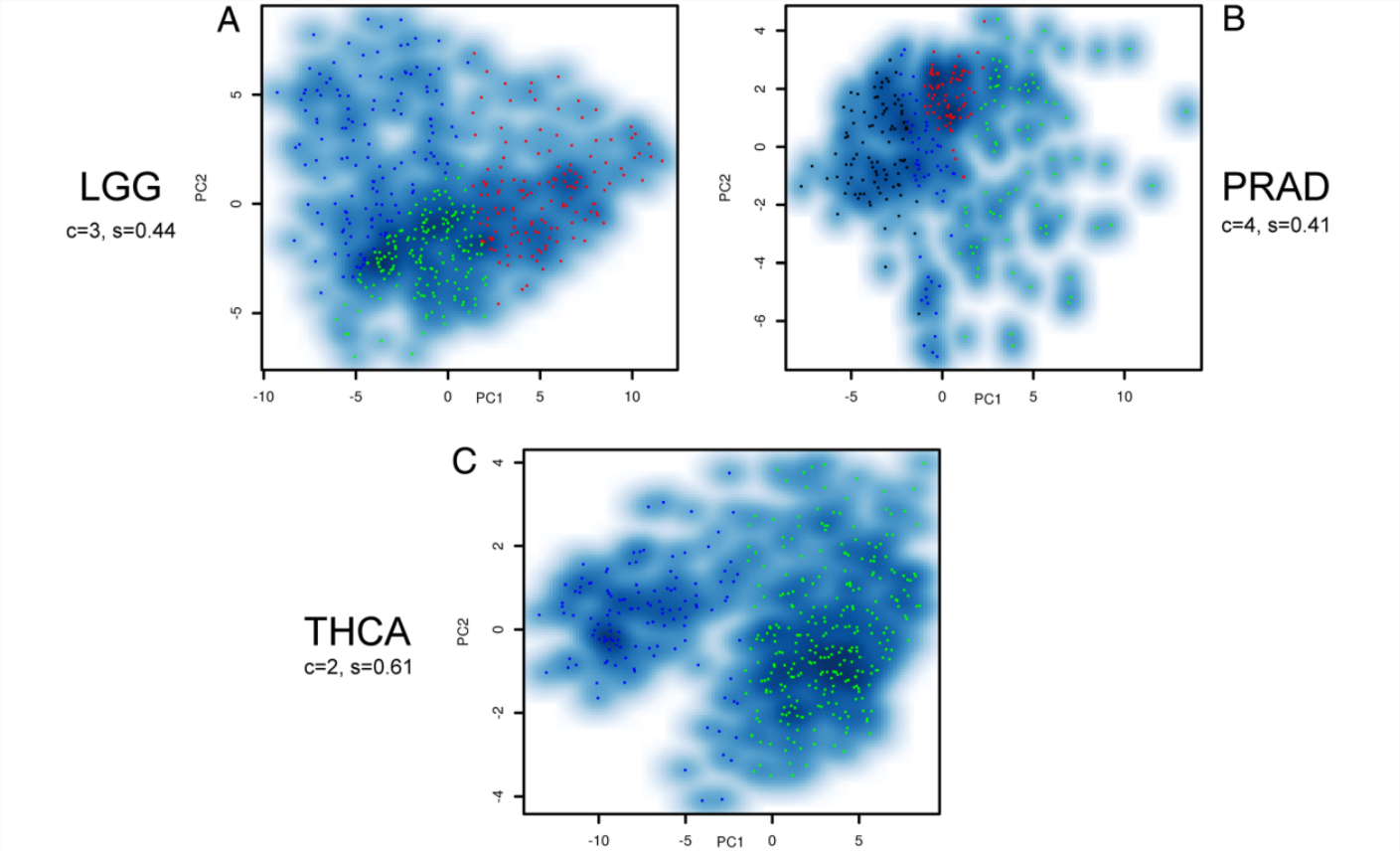


Fig. 4 The molecular subtypes identified by LRAcluster. (**a**) is for LGG, (**b**) for PRAD and (**c**) for THCA. The scatter plots show all the samples in the corresponding reduced 2-dimensional subspace. Different colors represent different molecular subtypes identified by LRAcluster, *c* indicates the number of identified clusters and *s* shows the silhouette value

But, there are no significant differences of overall survival among the identified molecular subtypes in LGG, PRAD, and THCA.

Case Study: Application on TCGA pan-cancer dataset

Applied LRAcluster on the 11 cancer types separately

Table 2 The results of single-cancer analysis

Cancer	Dimension[a]	#Cluster[b]	Silhouette values
BRCA	2	2	0.55
COAD	4	4	0.40
GBM	8	2	0.35
HNSC	7	3	0.26
KIRC	6	2	0.36
LGG	2	3	0.44
LUAD	5	2	0.34
LUSC	5	4	0.32
PRAD	2	4	0.41
STAD	4	3	0.37
THCA	2	2	0.61

[a]The dimension of the reduced space is determined according to the curve of the explained variations of each cancer type
[b]The number of clusters is determined according to the curve of the within cluster variances

For the remaining 7 cancer types, LRAcluster did not find strong molecular subtypes based on current omics data.

Conclusion

Conclusion

- **LRAcluster** probabilistically models the observed data conditional on the **size-matched parameters**.
- The **low-rank constraint** is the key to get the low-dimensional representation of the original data.
- And the **convexity** of the regularized likelihood function provides efficient gradient-descent algorithm for model fitting.
- Results show that LRAcluster runs **fast** with **high classification accuracy** and it is suitable for **large-scale** cancer multi-omics analysis.

Discussion

Strategy

- LRAcluster does not penalize the association between molecular features and the reduced subspace via sparsity assumption.
- A better strategy: find the significantly differential features between the samples in that cluster and all the other samples.
- The inter-omics regulatory information can be modeled as a separate pre-processing step to find the cancer driving factors and then only the molecular features significantly associated with these drivers are used as the input of LRAcluster.

Discussion

Joint non-negative matrix factorization (jNMF)

- jNMF is another strategy to find the shared principal subspace across multiple omics datasets*,⁕.
- Theoretically, NMF can be treated as a matrix version of latent factor analysis.
- jNMF will also encounter the optimization difficulty of non-convey loss function.
- But the advantage of jNMF is that the model can also get the molecular features (or called as modules) significantly associated each dimension.

---

* Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinforma Oxf Engl. 2011;27:i401409.

⁕ Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multidimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 2012;40:937991.

Thanks.