SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

Bayesian Biclustering for Patient Stratification

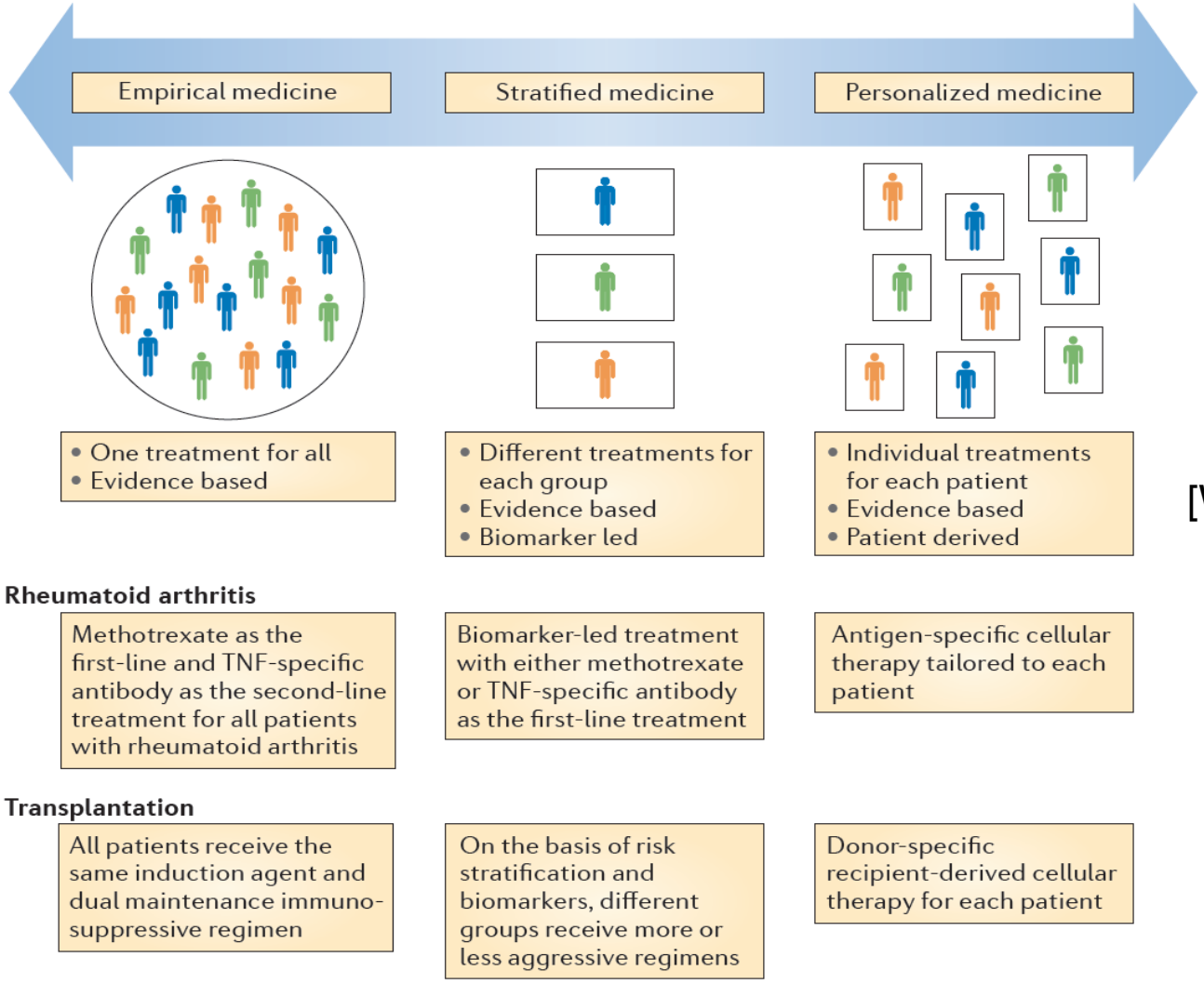
Martin Ester

CMPT 884 Spring 2018

Outline

- Introduction
- Problem Definition
- Related Work
- The Bayesian Biclustering Model

  Khakabi and Ester, PSB 2016
- Evaluation
- Future Work

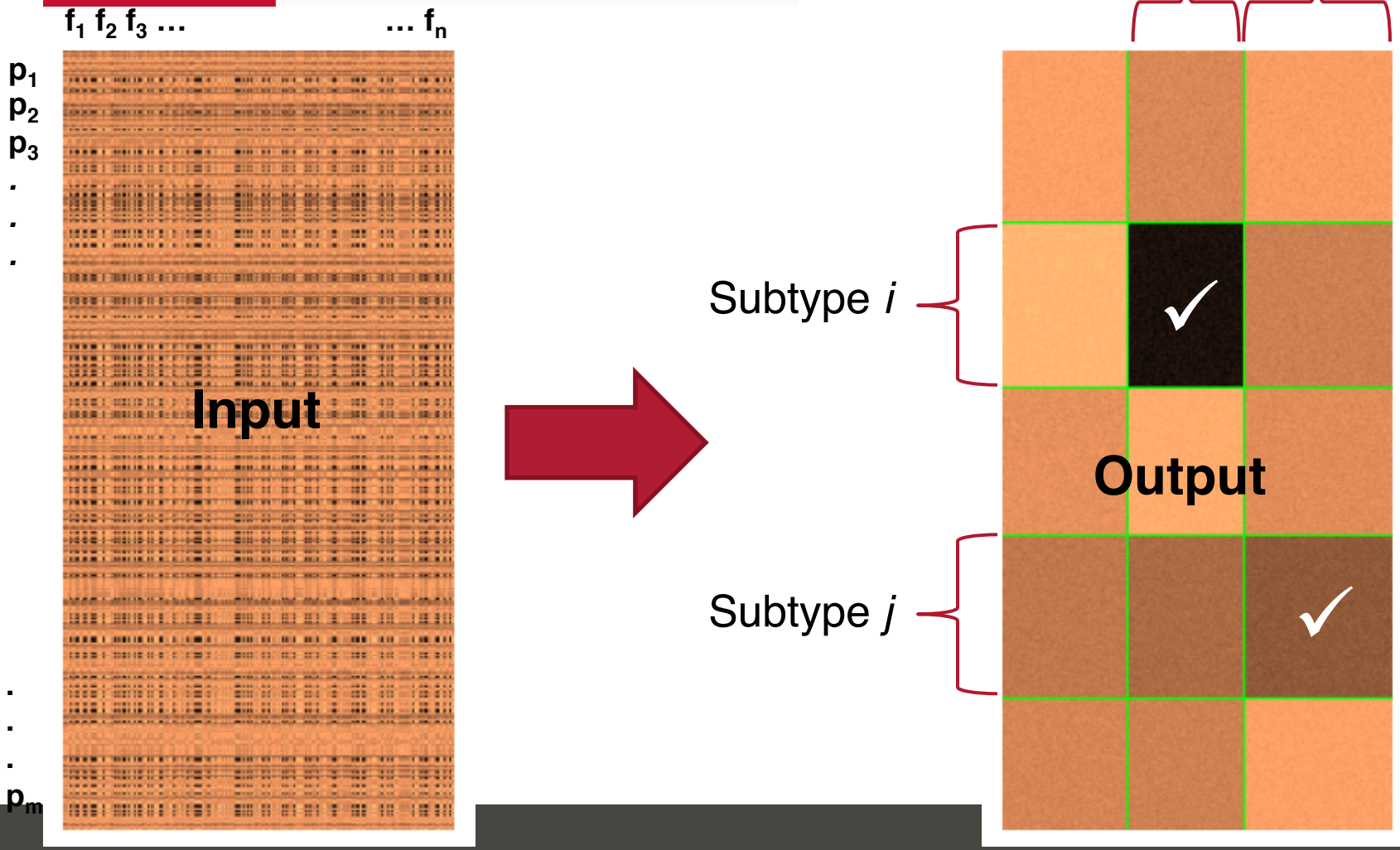


[Willis & Lord 2015]

Introduction

- How to identify the (molecular) subtypes of a disease?
- What are the subtype-specific biomarkers?
- How to identify subtypes that are distinct and detectable with biomarkers?

Introduction

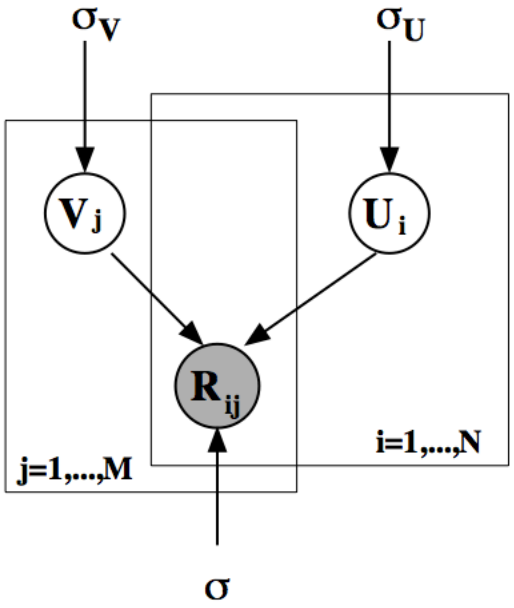


Problem Definition

Method	Probabilistic/ Deterministic	Clustering/ Biclustering	Stratification Input Datatypes
Verhaak et al. (2010)	Deterministic (HC)	Clustering	Expression
Hochreiter et al. (2010)	Deterministic (FA)	Biclustering	Expression
Hofree et al. (2013)	Deterministic (NMF)	Biclustering	Mutation
Shen et al. (2009, 2012)	Deterministic (FA)	Clustering	Multiple Datatypes
Sun et al. (2014)	Deterministic (SVD)	Clustering	Multiple Datatypes
Cho & Przytycka (2013)	Probabilistic (PGM)	Clustering	Multiple Datatypes
B2PS	**Probabilistic (PGM)**	**Biclustering**	**Multiple Datatypes**

Related Work

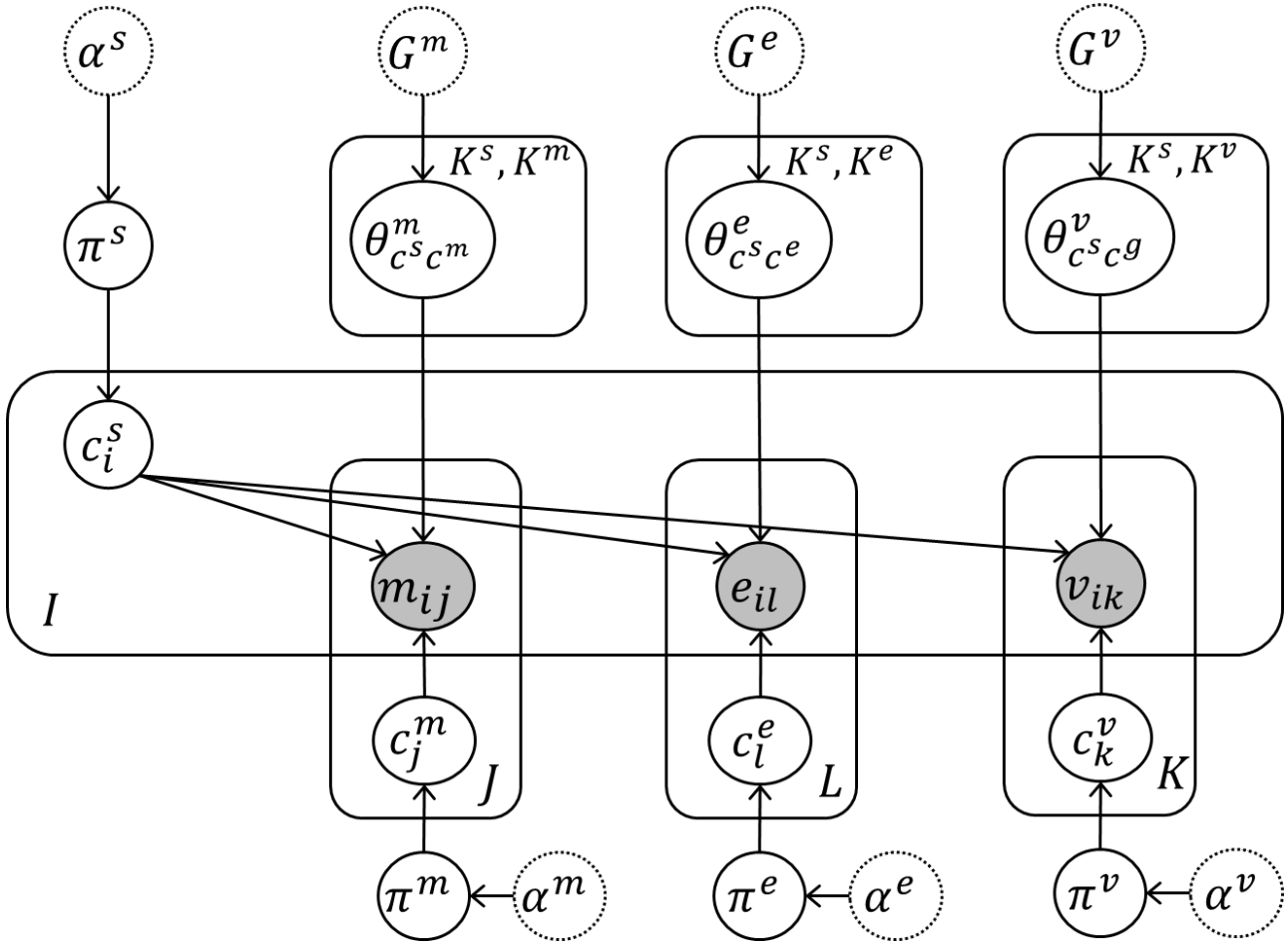
Probabilistic Matrix Factorization



Can be used for (bi-)clustering

Assign every row/column to the *i*-th cluster where *i* is the largest row/ column factor

Related Work



mij: point mutations
eij: gene expression
vij: copy number variations

B2PS (Bayesian Biclustering for Patient Stratification)

Properties

- Ability to incorporate prior knowledge
- Ability to integrate different data types
- Ability to detect the natural number of clusters
- Flexibility in the number of row and column clusters

B2PS (Bayesian Biclustering for Patient Stratification)

Pipeline

- Data preprocessing
- Prior tuning
- Parameter learning
- Consensus clustering
- Evaluation

B2PS (Bayesian Biclustering for Patient Stratification)

Prior Tuning

Priors			Num. of Sample Clusters	Num. of Feature Clusters	Log-rank p-value
Data	Sample Clustering	Gene Clustering			
weak	weak	weak	8	66	0.018
strong	**weak**	**weak**	**8**	**25**	**0.004**
strong	strong	weak	9	21	0.017
strong	weak	strong	8	73	0.019
strong	strong	strong	8	70	0.008

B2PS (Bayesian Biclustering for Patient Stratification)

Parameter Learning

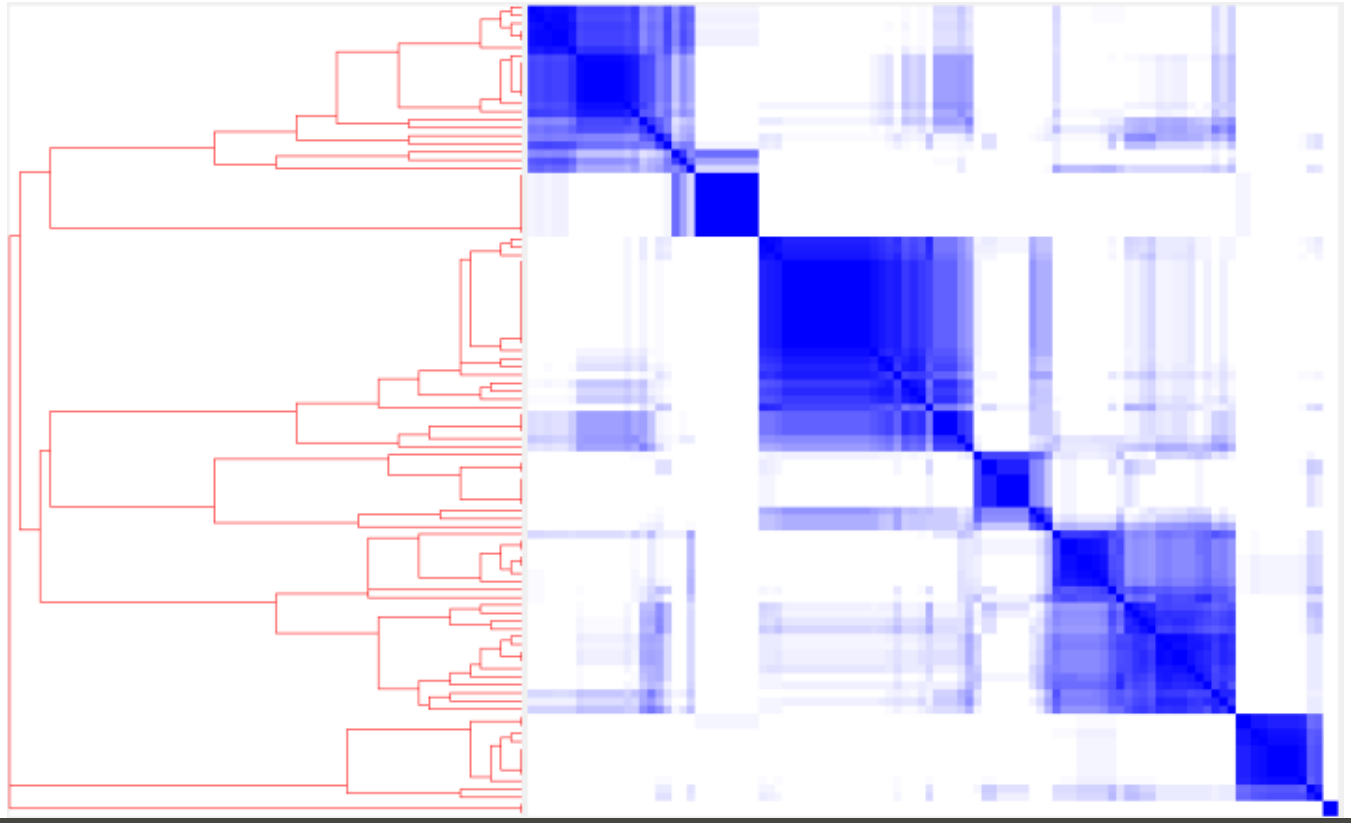
Gibbs Sampling

Column Clusters:

$$Pc\downarrow j\uparrow e = qc\downarrow -j\uparrow e, c\uparrow s, e;$$
$$\alpha\uparrow e, G\uparrow e \propto ne\downarrow q\uparrow -j +$$
$$\alpha\downarrow q\uparrow e / ne\uparrow -j + p\uparrow e$$

B2PS (Bayesian Biclustering for Patient Stratification)

Consensus Clustering

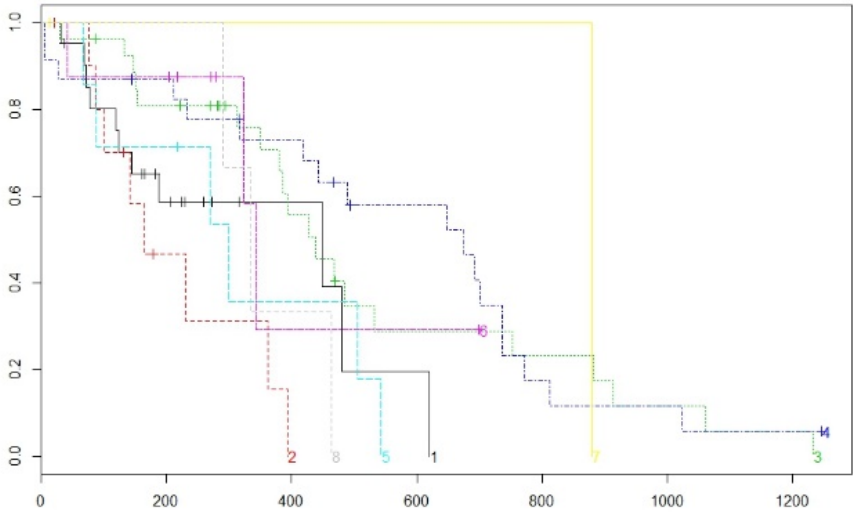


consensus row clusters from 50 executions

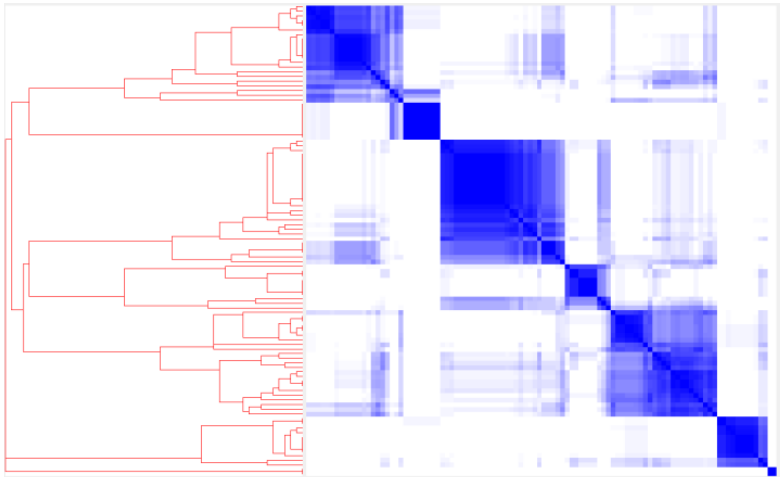
B2PS (Bayesian Biclustering for Patient Stratification)

Evaluation of Sample Clusters

Survival Analysis (log-rank test)



Robustness (Cophenetic Corr. Coeff.)



B2PS (Bayesian Biclustering for Patient Stratification)

Datasets

Disease	Samples	Features		
		Point Mutation	CNV	Expression
GBM	102	4117	23082	11874
BRCA	501	13776	23082	17814

Evaluation

Evaluation of Priors

Priors			Num. of Sample Clusters	Num. of Feature Clusters	Log-rank p-value
Data	Sample Clustering	Gene Clustering			
weak	weak	weak	8	66	0.018
strong	**weak**	**weak**	**8**	**25**	**0.004**
strong	strong	weak	9	21	0.017
strong	weak	strong	8	73	0.019
strong	strong	strong	8	70	0.008

Prior knowledge matters!

Evaluation

Evaluation of Datatypes

Data Types	Sample Clusters	Feature Clusters		Log-rank p-value	Cophenetic Corr. Coef.	GOTO	
		Exp.	CNV			Exp.	CNV
Exp.	**8**	**25**	**NA**	**0.004**	**0.958**	**3.4**	**NA**
CNV	19	NA	86	0.411	0.976	NA	1.8
Exp. and CNV	7	22	68	0.292	0.799	3.4	1.8
Mut.	No Convergence						
Mut. and Exp.	Similar to the first row						
Mut. and CNV	Similar to the second row						

Gene expression most useful,
Other datatypes may introduce noise!

Evaluation

Comparison to NMF

Method	Constraints	Sample Clusters	Feature Clusters	Log-rank p-value	Cophenetic Corr. Coef.	GOT O
B2PS	**Unrestricted**	**8**	**25**	**0.004**	**0.958**	**3.4**
NMF	# clusters = 3	3	3	0.458	0.965	2.5
B2PS	# sample clusters = 3	3	29	0.047	0.967	3.4
B2PS	# sample clusters = 3 # feature clusters = 6	3	6	0.217	0.999	3.4

Superior performance of B2PS due to

- its probabilistic nature (handles the noise)
- its flexibility in the number of column and row clusters.

Evaluation

- When does data integration help?
- How to extract biomarkers for patient clusters?
- Can we also extract driver genes?
- We often use survival data for evaluation of the clustering results. How to incorporate phenotype data such as survival data into the clustering model?

Future Work