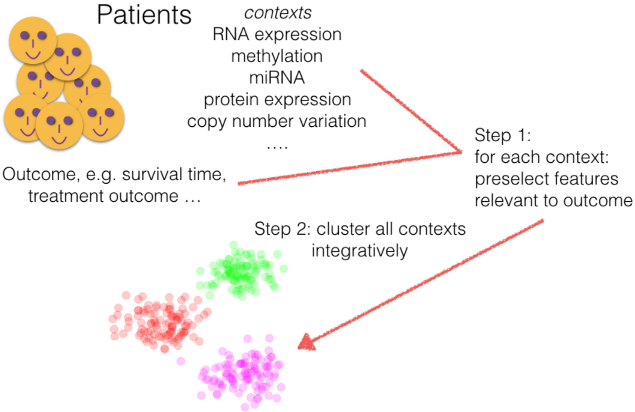
Bayesian hierarchical context-dependent clustering for multi-omics (pan-)cancer data

Magdalena Strauss, John Reid, Lorenz Wernisch

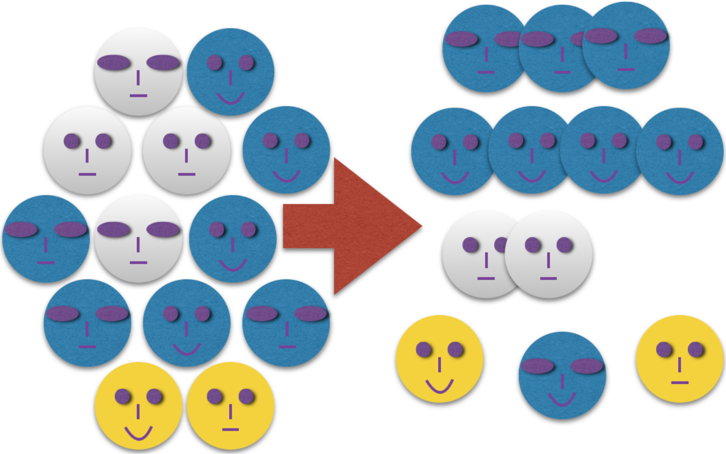
MRC Biostatistics Unit, University of Cambridge
Alan Turing Institute
12th July 2018

Motivation



Purpose: cluster patients in terms of multiple contexts integratively, determine relevance of context to outcome

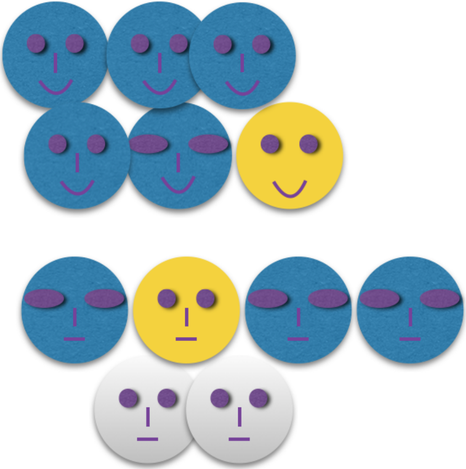
Illustrating context-dependent clustering



Context clusters

Context clusters

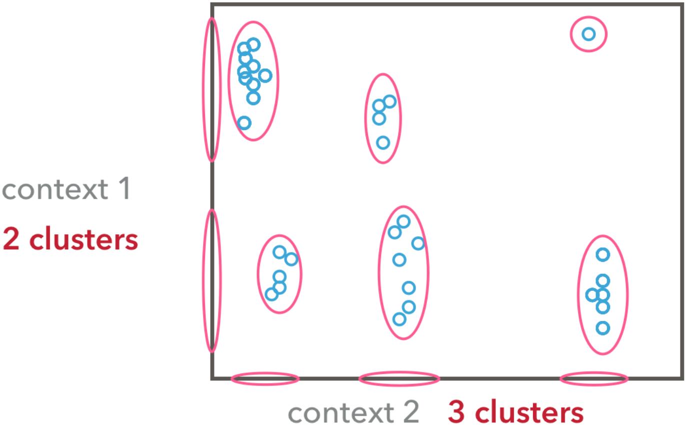
Context clusters



Summary

- Global clusters with context coordinates, i.e. model parameters are based on the context and not the global cluster
- Information sharing between the different contexts (coordinates) via the global clusters: patients in the same global cluster are in the same context cluster for each context

Graphical illustration



Description

- Global clusters, and clusters for each data type ('context')
- Integrates several data sets, e.g. several cancer types or subtypes
- Prior selection of genes, miRNAs expression levels ... associated with a particular outcome, e.g. survival (Cox's model)
- Known outcome only used for this prior feature selection, not for the clustering (prediction)
- Bayesian model
- Hierarchical structure using tensor products of finite Dirichlet priors

Bayesian methods

- Parameters as stochastic quantities with a probability distribution $\longrightarrow$ posterior distribution
- There may be prior information concerning the parameter $\longrightarrow$ prior distribution
- Bayes' theorem: $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- That is if $H(\theta)$ is the prior distribution of a parameter θ, and $p(x \mid \theta)$ is the distribution of the data, then for the posterior distribution $q(\theta \mid x)$: $q(\theta \mid x) = \frac{p(x|\theta)H(\theta)}{\int p(x|\theta)d\theta}$
- Sample from the posterior distributions of the parameters using, e.g, MCMC or importance sampling methods
- Suited to the modelling of complicated hierarchical structures
- Methods to obtain summary clustering as well

Modelling a single context: Mixture models

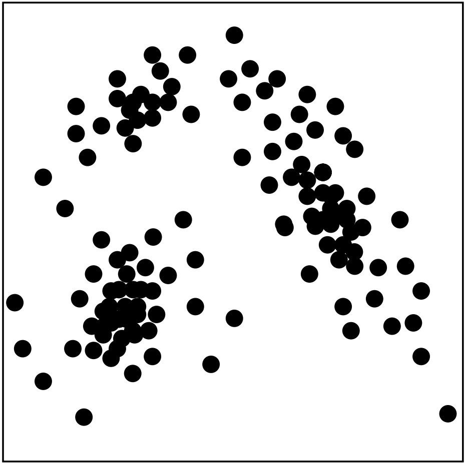
Mixture models

We model each context using a mixture model:

$$p(x) = \sum_{j=1}^{K} \pi_j F(x|\theta_j). \tag{1}$$

- K is the number of mixture components
- π_j are the mixture proportions
- F is a parametric density (such as a Gaussian)
- θ_j are the parameters associated with the j-th component

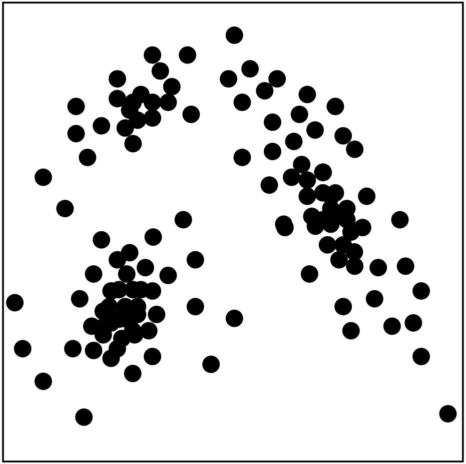
Mixture models: quick illustration



Mixture models: quick illustration

We model as a mixture of 3 Gaussians:

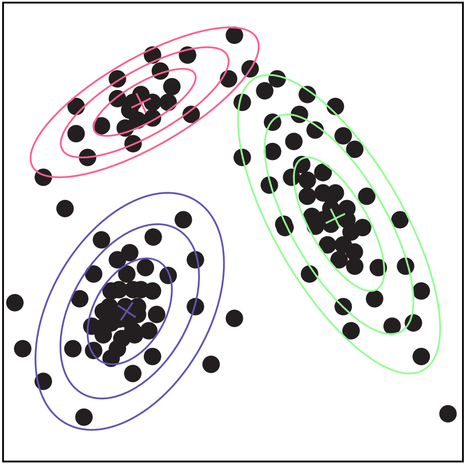
$$p(\mathbf{x}) = \pi_1 f(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1) + \pi_2 f(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2) + \pi_3 f(\mathbf{x}|\boldsymbol{\mu}_3, \Sigma_3)$$



Mixture models: quick illustration

We model as a mixture of 3 Gaussians:

$$p(\mathbf{x}) = \pi_1 f(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1) + \pi_2 f(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2) + \pi_3 f(\mathbf{x}|\boldsymbol{\mu}_3, \Sigma_3)$$



Mixture models: quick illustration

We model as a mixture of 3 Gaussians:

$$p(\mathbf{x}) = \pi_1 f(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1) + \pi_2 f(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2) + \pi_3 f(\mathbf{x}|\boldsymbol{\mu}_3, \Sigma_3)$$

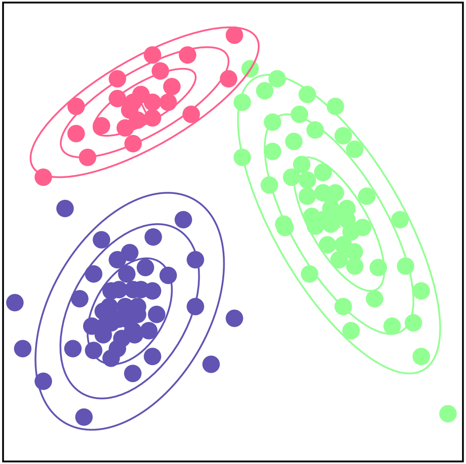


Figure by P. Kirk

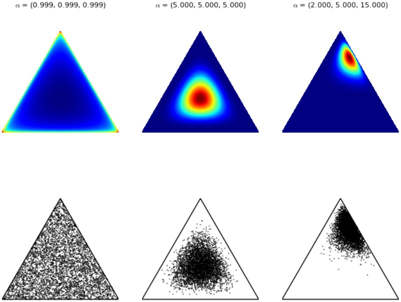
Bayesian mixture models

$$
\begin{aligned}
\boldsymbol{\pi} \mid \alpha &\sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \\
k_n \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi}, \quad n = 1, \ldots, N \\
\theta \mid H &\sim H \\
x_n &\sim F(\cdot \mid \theta))
\end{aligned}
$$

- Observed data $x_1, \ldots, x_n$
- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the collection of K mixture proportions
- α is a concentration parameter
- H is the prior for the component parameters

Prior and posterior distribution of mixture weights

Prior distribution: Dirichlet $\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$
Posterior distribution: Dirichlet $\left(\frac{\alpha}{K} + n_1, \ldots, \frac{\alpha}{K} + n_K\right)$
n_j number of samples in cluster j
Dirichlet $(\alpha_1, \ldots, \alpha_K)(\boldsymbol{\pi}) \propto \prod_{j=1}^{K} \pi_j^{\alpha_j - 1}$



Plot by Thomas Boggs. https://gist.github.com/tboggs

Context-dependent clustering model

$$\boldsymbol{\rho} \mid \gamma \sim \text{Dirichlet}\left(\frac{\gamma}{S}, \ldots, \frac{\gamma}{S}\right) \quad \text{global clusters}$$

$$\boldsymbol{\pi}^{(c)} \mid \alpha^{(c)} \sim \text{Dirichlet}\left(\frac{\alpha^{(c)}}{K^{(c)}}, \ldots, \frac{\alpha^{(c)}}{K^{(c)}}\right) \quad \text{context clusters}$$

$$s_n \mid \boldsymbol{\rho} \sim \boldsymbol{\rho}, \quad n = 1, \ldots, N \quad \text{global clusters}$$

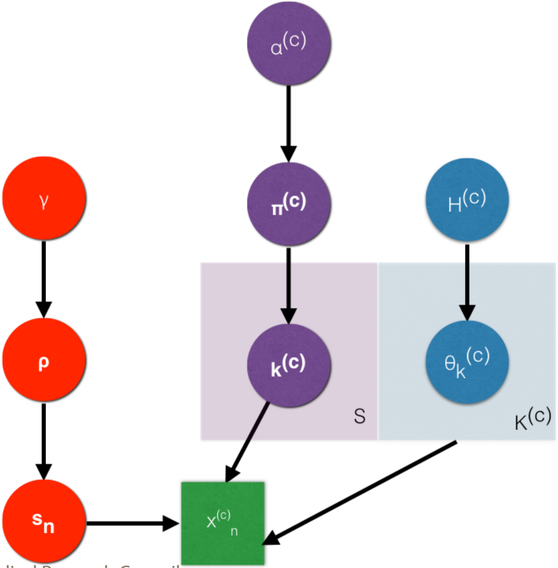
$$k^{(c)}(s) \mid \boldsymbol{\pi}^{(c)} \sim \boldsymbol{\pi}^{(c)}, \quad s = 1, \ldots, S \quad \text{context clusters}$$

$$\theta^{(c)}(k^{(c)}) \mid H^{(c)} \sim H^{(c)}, \quad k^{(c)} = 1, \ldots, K^{(c)}$$

$$x_n^{(c)} \sim F^{(c)}(\cdot \mid \theta^{(c)}(k^{(c)}(s_n)))$$

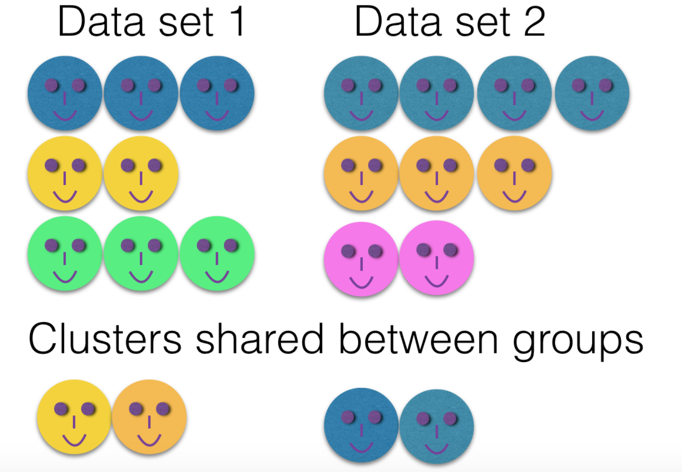
Context likelihoods weighted by inverse of context dimensions

Graphical model description

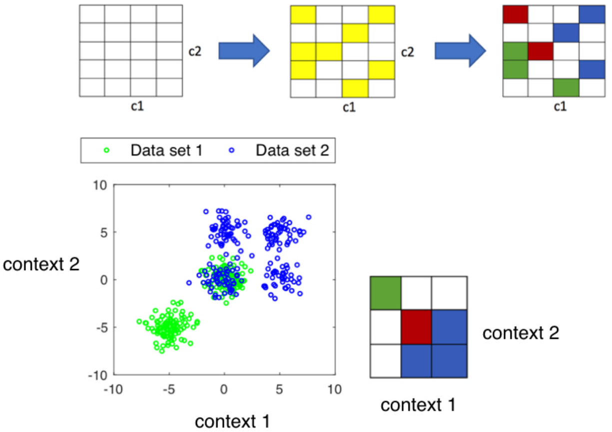


Hierarchical model

For integration of different data types, and to improve sampling properties and avoid collapse to very few context clusters



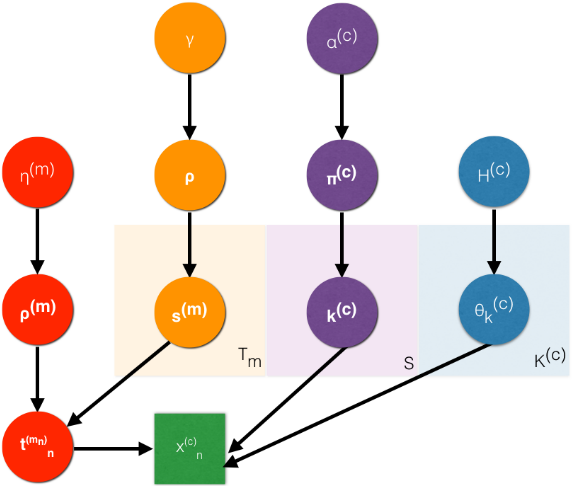
Hierarchical model



Mathematical model description

$$
\begin{aligned}
\boldsymbol{\rho}^{(\boldsymbol{m})} \mid \eta^{(m)} &\sim \text{Dirichlet}\left(\frac{\eta^{(m)}}{T_m}, \ldots, \frac{\eta^{(m)}}{T_m}\right) \\
\boldsymbol{\rho} \mid \gamma &\sim \text{Dirichlet}\left(\frac{\gamma}{S}, \ldots, \frac{\gamma}{S}\right) \\
\boldsymbol{\pi}^{(\boldsymbol{c})} \mid \alpha^{(c)} &\sim \text{Dirichlet}\left(\frac{\alpha^{(c)}}{K^{(c)}}, \ldots, \frac{\alpha^{(c)}}{K^{(c)}}\right) \\
t_n^{(m_n)} \mid \boldsymbol{\rho}^{(\boldsymbol{m})} &\sim \boldsymbol{\rho}^{(\boldsymbol{m})}, \quad n = 1, \ldots, N \\
s^{(m)}(t^{(m)}) \mid \boldsymbol{\rho} &\sim \boldsymbol{\rho}, \quad t^{(m)} = 1, \ldots, T_m \\
k^{(c)}(s) \mid \boldsymbol{\pi}^{(\boldsymbol{c})} &\sim \boldsymbol{\pi}^{(\boldsymbol{c})}, \quad s = 1, \ldots, S \\
\theta^{(c)}(k^{(c)}) \mid H^{(c)} &\sim H^{(c)}, \quad k^{(c)} = 1, \ldots, K^{(c)} \\
x_n^{(c)} &\sim F^{(c)}(\cdot \mid \theta^{(c)}(k^{(c)}(s^{(m_n)}(t_n^{(m_n)}))))
\end{aligned}
$$

Graphical model representation

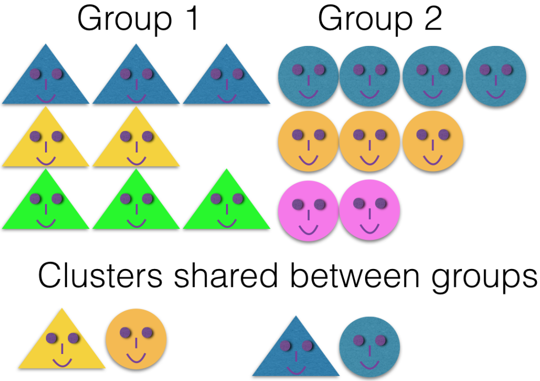


Difference to non-hierarchical model

- Integrates several data sets/cancer types
- For just one data set, we use a larger number of possibly empty global clusters and allocate them to context clusters at each iteration
- The number of group clusters is smaller
- The larger number of global clusters explores all possible combinations of context clusters more accurately
- The smaller number of group clusters leads to information sharing across the contexts

Hierarchical model: across-clustering

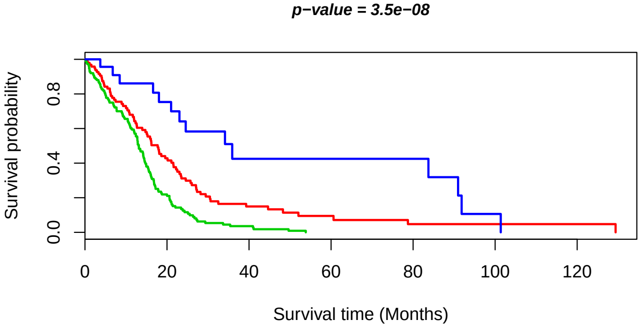
Cluster several data sets hierarchically in terms of the distances from their respective means



GBM data (1)

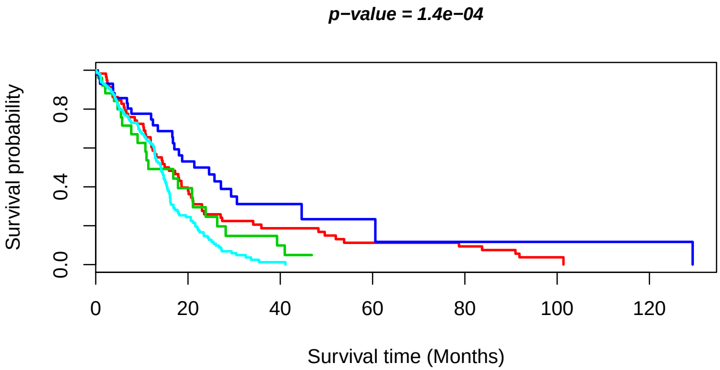
TCGA data, gene expression, DNA methylation, miRNA expression data, also analysed as part of the CancerSubtypes bioconductor package (Xu, T. et al. (2017). *Bioinformatics*). 276 patients

hyper-parameters: $\eta, \gamma, \alpha = 100, K = 4, S = 4^3, S_m = 20$



Methylation

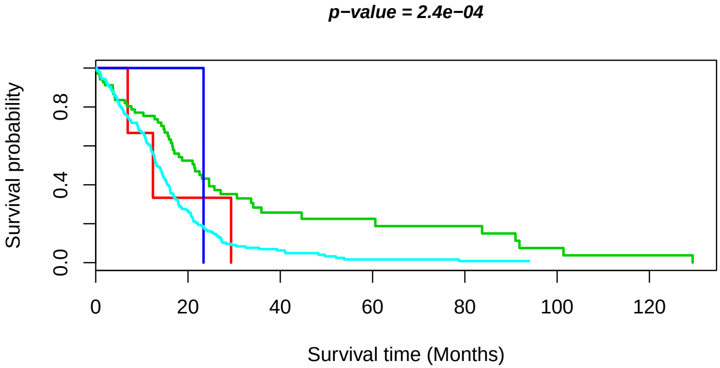
GBM data (2)



RNA expression

Strength of association with the survival outcome differs for the different contexts.

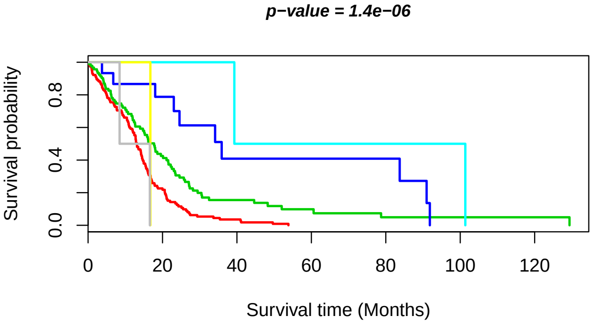
GBM data (3)



miRNA

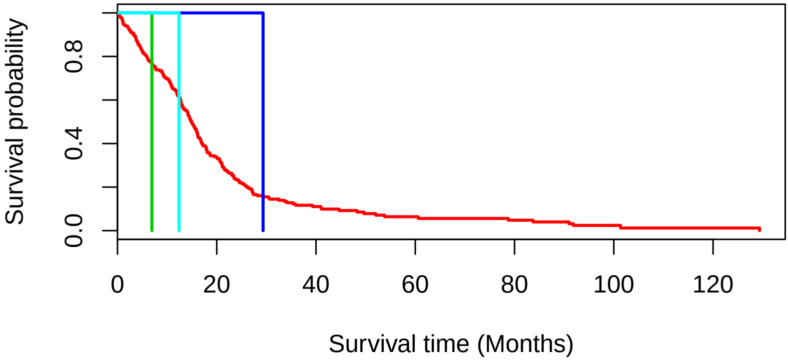
Comparison to clustering only one context on its own

Fitting using finite mixture model with optimised number of clusters (2 to 9), mclust R package



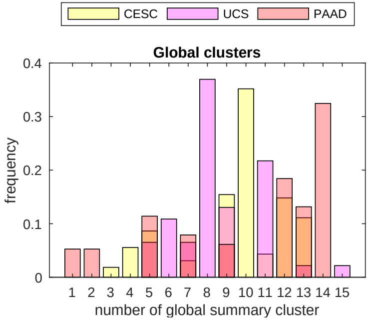
Adjusted Rand index to proposed method: 0.93

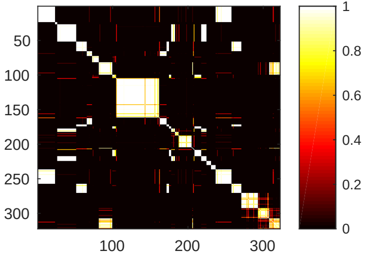
miRNA



Adjusted Rand index to proposed method: 0.08

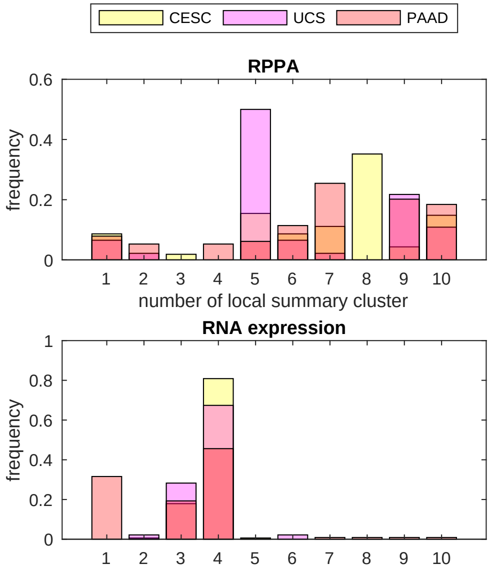
Several types





3 cancer types, 2 contexts (RNA expr, RPPA). Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma(CESC, 162 patients), Pancreatic Adenocarcinoma(PAAD, 114 patients), Uterine Carcinosarcoma(UCS, 46 patients); primary solid tumours; data from TCGA.

Several types: context clusters



Discussion and future work

- Posterior distribution of cluster allocations tends to be multimodal $\longrightarrow$ multiple runs, more effective sampling strategies
- Testing methods of selecting covariates relevant to outcome for different outcomes (sparse variable selection, deep learning)
- More systematic testing of sensitivity of context cluster structures to maximum number of global clusters
- Systematic testing of relevance of context to outcome for different contexts
- Integration of imaging data would be interesting
- Interested in Bayesian statistics, clustering, sparse regression? Talk to me