ETH zürich

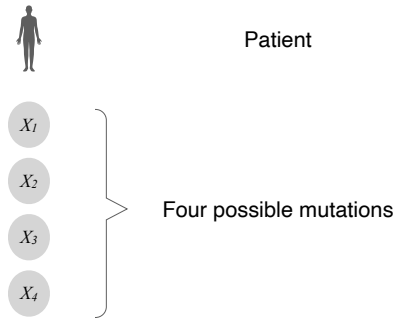
Network-Based Clustering of Pan-Cancer Data Accounting for Clinical Covariates

Fritz Bayer

09.12.2022

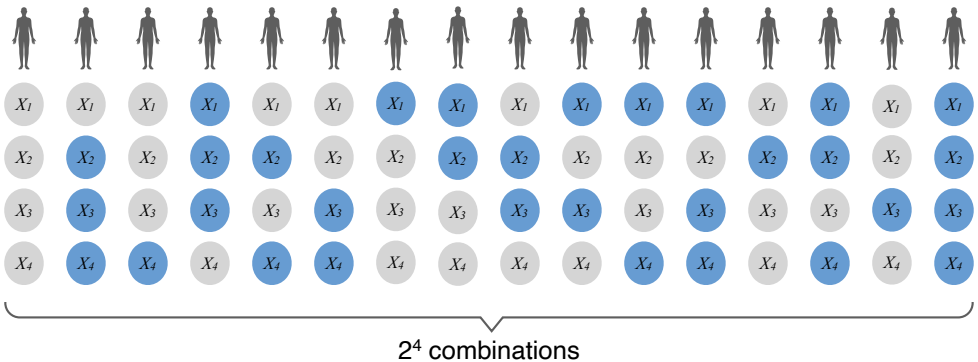
Heterogeneity of Cancer
Motivation

- Cancer occurs when cells are struck by mutations in their genomes
- Goal: Identifying subgroups of patients with shared biological properties based on mutations



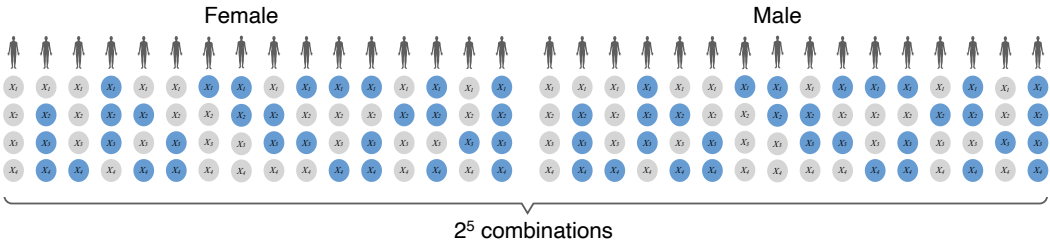
Heterogeneity of Cancer

Motivation



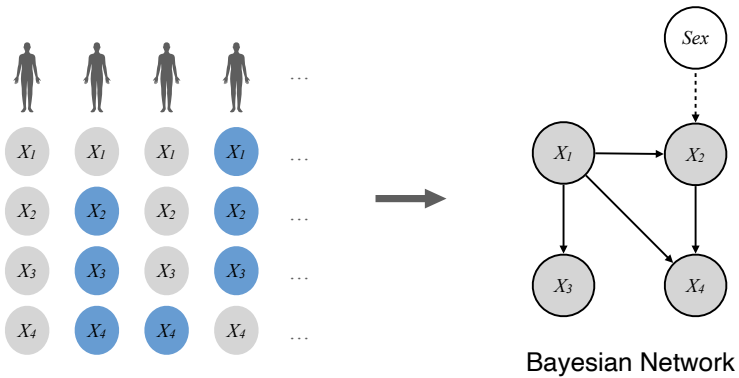
Heterogeneity of Cancer

Motivation



Network-Based Clustering

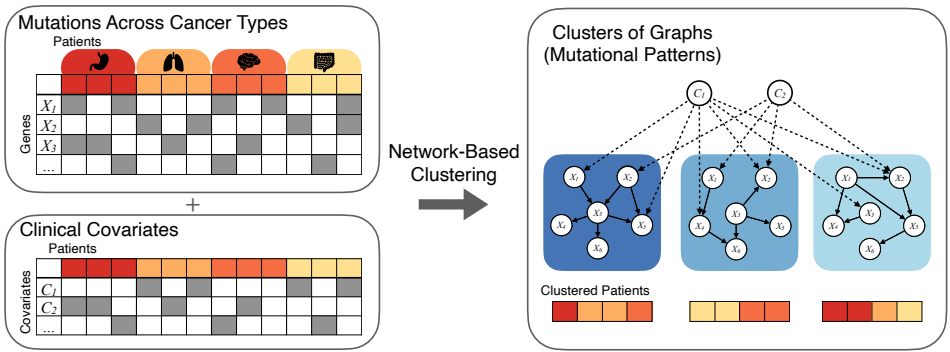
Methods



Network-Based Clustering

Methods

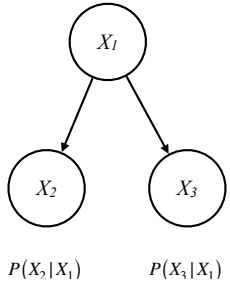
- Goal: Identifying subgroups of shared biological properties
- Method: Cluster mutational and clinical covariate data based on distinct probabilistic relationships



Introduction to Bayesian Networks
Methods

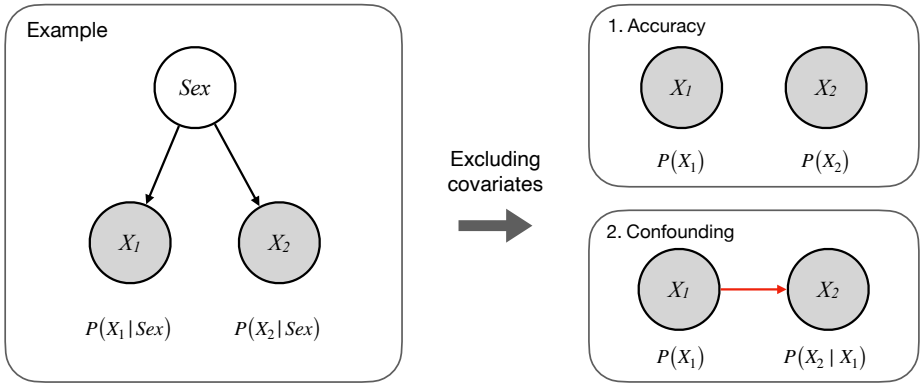
- DAG $\mathcal{G} = (V, E)$ with nodes V and edges E
- Nodes V are associated with variables X_V with probability distribution $P(X_V)$
- Factorization (Markov conditions)

$$P(X_V|\mathcal{G}, \theta) = \prod_{i \in V} P\big(X_i \mid X_{pa(i)}\big)$$



Network-Based Clustering

Methods



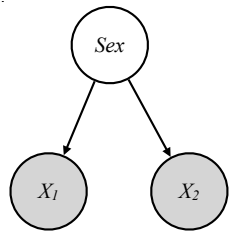
Covariate-Adjusted Clustering
Methods

- Covariate-Adjusted membership probability

$$\tilde{\phi}(X_V \mid k) = \frac{\gamma_k \cdot P\left(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k\right)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\left(X_V \mid X_C, \hat{\mathcal{G}}_{k'}, \hat{\theta}_{k'}\right)}$$

where τ_k is the weight of each cluster

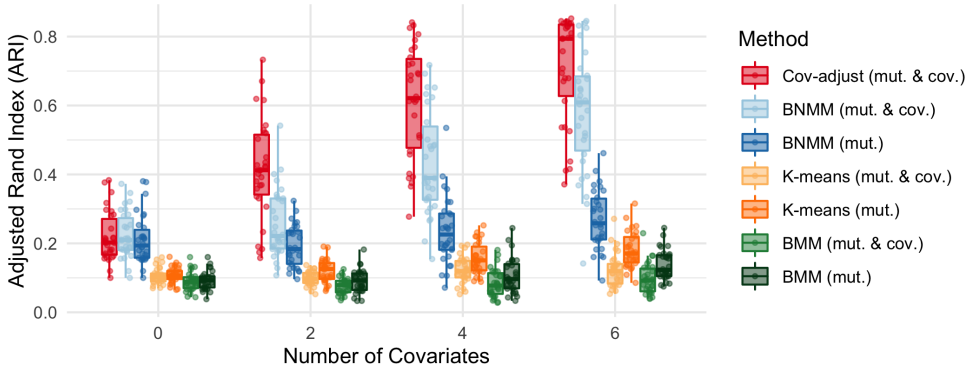
$$\gamma_k = \frac{\sum_{1=1}^{N} \tilde{\phi}(X_V \mid k)}{N}$$



Simulation Experiments

Results

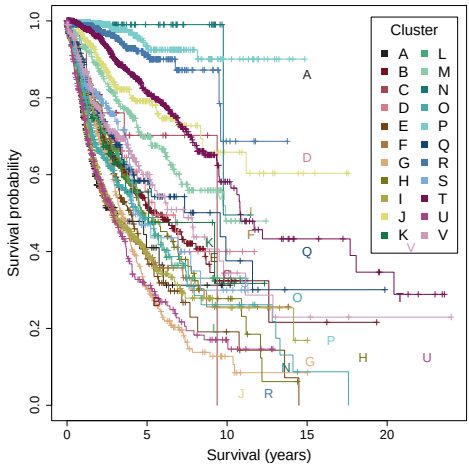
- Simulate random clusters (Bayesian networks) for a varied number of covariates
- Compared against standard clustering algorithms



Application to Pan-Cancer Data

Results

- Genomic dataset of 8085 cancer patients
- Survival analysis as check for informative clustering
- Clusters are highly significant in predicting survival ($LR = 46.6$, $p = 1 \times 10^{-10}$)

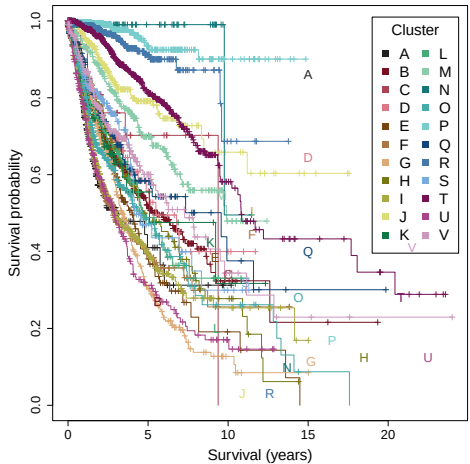


Pancancer Data Results

Results

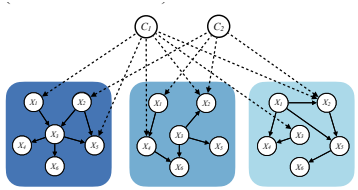
- Accounting for covariates significantly improves survival prediction
- Highlights the importance of adjusting for clinical covariates

Method	LR	P-value
Cov-adjust	**46.6**	$1 \cdot 10^{-10}$
BNMM (mut. & cov.)	43.8	$4 \cdot 10^{-10}$
BNMM (mut.)	34.4	$6 \cdot 10^{-07}$



Summary

- Method: Network-based clustering method that integrates mutational and covariate data
- Simulations: Outperforms standard clustering methods
- Pan-cancer data: Identified clusters which are predictive of survival beyond clinical information



ETH zürich

Thank you for your attention!

Code: https://github.com/cbg-ethz/graphClust_NeurIPS

Contact: frbayer@ethz.ch

Thanks to: Giusi Moffa, Niko Beerenwinkel, Jack Kuipers