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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Identifying subgroups of shared biological properties based on mutational features is a key step towards precision treatment of cancer patients. However, clustering patients based on their mutational profile is challenging due to considerable heterogeneity within and across cancer types. Here, we approach the heterogeneity of cancer by learning probabilistic relationships within pan-cancer data. We present a network-based clustering method, that integrates mutational and clinical covariate data in distinct networks of their probabilistic relationships. To avoid learning the clusters based on covariates such as age and stage, we remove their effect on the cluster assignment, by exploiting causal relationships among the variables. In simulations, we demonstrate that our method outperforms standard clustering methods. We apply our method to a large-scale genomic dataset of 8085 cancer patients, where we identify novel clusters that are predictive of survival beyond clinical information and could serve as biomarkers for targeted treatment. [1]

1 Introduction

Cancer progresses in diverse ways leading to a heterogeneous landscape of mutations within and across cancer types. This heterogeneity is a considerable challenge for the task of using genomic data to learn clinically and biologically meaningful clusters [Greenman et al., 2007, McDermott et al., 2011, Ciriello et al., 2013].

Modelling interactions among the mutations via networks of their probabilistic relationships has shown to be promising for mutation-based stratification [Hofree et al., 2013, Hou et al., 2016, Kuipers et al., 2018] and the identification of drug targets [Cheng et al., 2019]. Since clinical covariates such as age and sex are highly correlated to the mutational profiles [Milholland et al., 2015, Lopes-Ramos et al., 2020], they act as hidden variables and hence limit the accuracy of models that are learned from purely mutational data.

In this work, we integrate mutational and clinical covariate data in networks of their probabilistic relationships as outlined in Figure 1. In particular, we model the following two biological aspects: (1) cluster-specific probabilistic relationships among the mutations and clinical covariates, and (2) causal effects from clinical covariates on the mutations.

[1] Code implementing our method and reproducible benchmarks and data analysis are available at `https://anonymous.4open.science/r/LMRL-NeurIPS`.

Submitted to the NeurIPS Learning Meaningful Representations of Life Workshop 2022. Do not distribute.

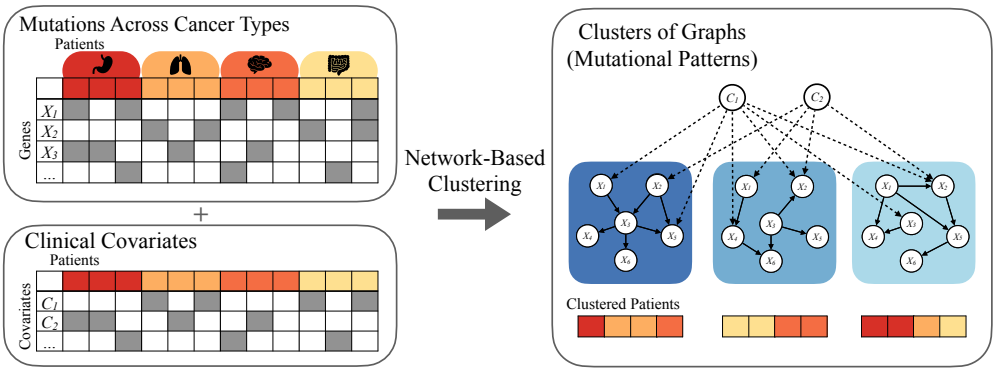

Figure 1: Network-based clustering of mutational profiles and clinical covaritates learns graphs that represent distinct probabilistic relationships across different cancer types.

2 Methods

2.1 Covariate-Adjusted Clustering

We want to the cluster mutational profiles of tumours into K different groups. To account for cluster-specific probabilistic relationships, we model each cluster by a Bayesian network. Each Bayesian network $(\mathcal{G}, \theta)$ comprises a directed acyclic graph $\mathcal{G}$ and associated local probability distributions θ. The nodes in the Bayesian networks represent mutations and clinical covariates. We differentiate between two different types of covariates: (1) *cluster-independent covariates* that have only outgoing edges into the mutations, and (2) *cluster-dependent covariates* that have incoming edges from the mutations.

Exploiting our knowledge about the covariates, we can correspondingly make assumptions about the direction of the causal pathways. As an example, the covariate sex may make specific mutations more likely to occur, but it is not the consequence of specific mutations. Thus, the covariate sex has a downstream causal effect on the mutations and is therefore a cluster-independent covariate. In contrast, in acute myeloid leukemia the covariate cancer type is determined based on the mutational profile of the patients [Gerstung et al., 2017], making it a cluster-dependent covariate.

Since we are interested in mutational patterns, we do not want to cluster patients based on cluster-independent covariates such as age and sex, as these can increase the variance or bias the cluster assignment. We therefore need to correct for the effect of cluster-independent covariates on the clustering. In contrast, cluster-dependent covariates can carry information about the mutational patterns. Hence they can be modelled analogously to the mutations without further adjustment.

With this setup, we assume the generating probability distribution of the cluster-dependent variables X_V to be conditioned on the cluster-independent covariates X_C:

$$p(X_V) = \sum_{k=1}^{K} \pi_k P(X_V \mid X_C, \mathcal{G}_k, \theta_k) \tag{1}$$

where $\sum_{k=1}^{K} \pi_k = 1$ is the weight of each cluster and the probability of X_V given cluster k is

$$P(X_V \mid X_C, \mathcal{G}_k, \theta_k) = \prod_{i \in V} P\big(X_i \mid X_C, X_{pa(i)_k}, \theta_k\big) \tag{2}$$

Knowledge of the generating distribution allows us to define a mixture model to cluster the mutational profiles. In our mixture model, we learn the membership probabilities $\tilde{\phi}(X_V \mid k)$ of the mutations for each cluster using the EM-algorithm similar to [Ko et al., 2009, Kuipers et al., 2018]. To adjust for the effects of clinical covariates on the clustering, we employ the following adjusted membership probability function

$$\tilde{\phi}(X_V \mid k) = \frac{\gamma_k \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k\Big)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\Big)} \tag{3}$$

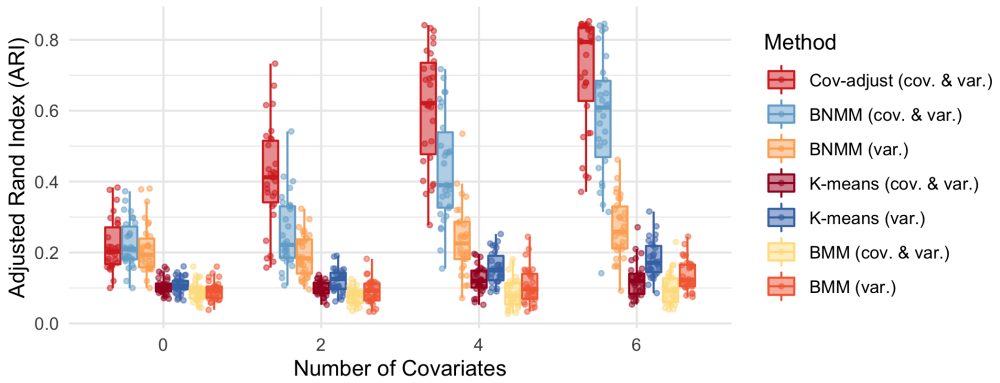

Figure 2: Simulation benchmark including our proposed cluster algorithm (cov-adjust), a Bayesian network mixture model (BNMM), k-means, and a Bernoulli mixture model (BMM).

where $\gamma_k = \frac{\sum_{1=1}^{N} \tilde{\phi}(X_V|k)}{N}$ is the weight of each cluster and $(\hat{\mathcal{G}}_k, \hat{\theta}_k)$ are the learned Bayesian network parameters. In this adjusted function, the probability of the mutations is conditioned on the covariates. This has two major advantages: first, it corrects for confounding that might be induced when neglecting the covariates, and second, it adjusts for the effects the covariates might have on the clustering. We prove that this correction leads to a lower variance in the membership probability function in Appendix A and outline the individual steps of our algorithm in Appendix B.

Calculating $P(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k)$ is a marginalization problem, which is NP-hard in general for Bayesian networks [Cooper, 1990] and typically requires approximations in high-dimensional settings [Bayer et al., 2022]. However, $P(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k)$ simplifies under the condition that the covariates have only outgoing edges to the mutations due to the factorization property of Bayesian networks [Maathuis et al., 2018]. This significantly reduces the computational cost and allows for an exact calculation of the membership probability for each sample.

3 Experiments

3.1 Simulations

In order to evaluate the performance of our covariate-adjusted clustering method, we benchmarked it against standard clustering methods over a broad range of different simulations. To account for our biological assumptions, we simulated data with cluster-specific probabilistic relationships and causal effects from covariates on the clustered variables. We simulated a Bayesian network for each cluster, including covariates as nodes with outgoing edges and constant conditional probability tables.

We compared our method against the following clustering algorithms: a Bayesian network mixture model (BNMM), k-means and a Bernoulli Mixture Model (BMM). Since these clustering algorithms do not adjust for covariates, we applied them once by adding the covariates to the clustered variables and once by excluding the covariates from the analysis. Code implementing our method and reproducible benchmarks are available at `https://anonymous.4open.science/r/LMRL-NeurIPS`.

Figure 2 shows the benchmark results over different numbers of covariates, where clustering accuracy was assessed via the adjusted rand index (ARI). A more detailed analysis including the benchmarks for varied cluster-specific parameters can be found in Appendix C. In the presence of covariates, the most accurate clustering performance is reached by our covariate-adjusted method. This is in line with our theoretical expectations, which removed the effect of the covariates on the cluster membership probability while adjusting for their other effects. In general, the network-based clustering methods have a higher clustering accuracy, which reflects that these structures allow one to model probabilistic relationships among the variables. Figure 2 shows that k-means and BMM perform better when the covariates are excluded from the analysis. This can be explained by the fact that these models assume independent probability distributions and hence including the covariates only increases the variance in the cluster assignment. In contrast, the performance of the BNMM improves when including the covariates, since their induced probabilistic relationships allow one to model the data more accurately.

3.2 Application to Pan-Cancer Data

We applied our clustering algorithm to a large-scale genomic dataset from the Cancer Genome Atlas (TCGA) [TCGA-Research-Network, 2008], which includes the mutational profiles and clinical information of 8085 patients from 22 different cancer types. For each primary cancer type, we considered the 16 most significantly mutated genes analogous to Kuipers et al. [2018], adding up to a total of 201 genes across all cancer types. In addition, we included the clinical covariates age, sex, and cancer type in the analysis. Since the variables age and sex have a downstream causal effect on the mutations, we adjusted for them in our covariate-adjusted clustering framework. The number of clusters was determined by calculating the Akaike information criterion (AIC) for a range of different sizes.

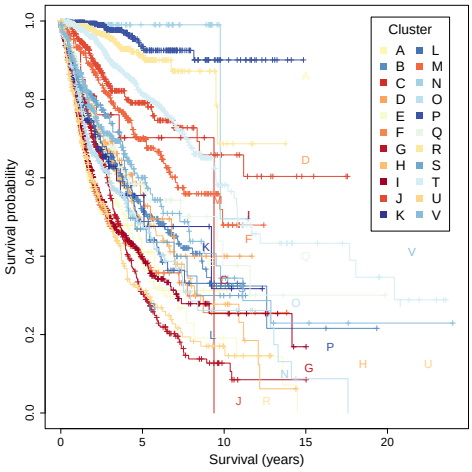

Figure 3: Survival probabilities of the novel cancer subgroups, labelled A-V.

While there is no ground truth cluster assignment in pan-cancer data, similarity in survival outcome within the clusters is a strong indicator of shared biological properties. We applied the Cox proportional hazards regression model and corrected for the effects of age, stage, and cancer type on the results since these are strong predictors for survival.

Using our method to cluster the mutational profiles based on their probabilistic relationships, we identified 22 novel cancer subgroups for which the survival probabilities are shown in Figure 3. Each subgroup is associated with a distinct network, representing its subgroup-specific probabilistic relationships. Table 1 shows the corrected likelihood ratio (LR) of the Cox proportional hazards regression model for different clustering algorithms, where we only considered network-based approaches due to their superior performance in the benchmark study. High corrected LRs indicate cluster assignments that are highly predictive in survival. The lowest corrected LR is reached by the BNMM that neglects the clinical covariates, highlighting the importance of integrating these when clustering mutational profiles. In contrast, the cancer subgroups found with our covariate-adjusted clustering method were most predictive in survival beyond clinical information (LR $= 46.6$, p-value $= 1.0 \cdot 10^{-10}$), confirming our theoretical expectations and simulation results.

Table 1: Likelihood ratio for different clustering algorithms

Method	Corrected likelihood ratio	P-value
Cov-adjust	**46.6**	$1.0 \cdot 10^{-10}$
BNMM (mut. & cov.)	43.8	$4.0 \cdot 10^{-10}$
BNMM (mut.)	34.4	$5.6 \cdot 10^{-07}$

4 Conclusions

We introduced a novel network-based clustering method that clusters mutations and covariates based on their distinct probabilistic relationships. Exploiting causal relationships among the variables, we showed how to adjust for the effects of clinical covariates on the mutations.

In simulations we demonstrated that our method learns more accurate cluster assignments than standard clustering algorithms. We applied our method to a large-scale pan-cancer dataset, where we identified novel clusters which are significantly predictive of survival beyond clinical information. Adjusting for the clinical covariates increased the accuracy of the survival prediction, highlighting their impact on the learned mutational networks.

Our identified mutational networks could serve to uncover mechanistic insights within cancer subgroups and could provide biomarkers for targeted treatment.

References

Fritz M. Bayer, Giusi Moffa, Niko Beerenwinkel, and Jack Kuipers. High-Dimensional Inference in Bayesian Networks. *arXiv preprint*, arXiv:2112.09217, 2022.

Feixiong Cheng, Weiqiang Lu, Chuang Liu, Jiansong Fang, Yuan Hou, Diane E. Handy, et al. A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nature Communications*, 10(1):3476, 2019.

Giovanni Ciriello, Martin L. Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133, 2013.

Gregory F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2):393–405, 1990.

Moritz Gerstung, Elli Papaemmanuil, Inigo Martincorena, Lars Bullinger, Verena I. Gaidzik, Peter Paschka, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nature Genetics*, 49(3):332–340, 2017.

Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L. Dalgliesh, Christopher Hunter, Graham Bignell, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132): 153–158, 2007.

Matan Hofree, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115, 2013.

Jack P. Hou, Amin Emad, Gregory J. Puleo, Jian Ma, and Olgica Milenkovic. A new correlation clustering method for cancer mutation analysis. *Bioinformatics*, 32(24):3717–3728, 2016.

Younhee Ko, ChengXiang Zhai, and Sandra Rodriguez-Zas. Inference of gene pathways using mixture Bayesian networks. *BMC Systems Biology*, 3(1):54, 2009.

Jack Kuipers, Thomas Thurnherr, Giusi Moffa, Polina Suter, Jonas Behr, Ryan Goosen, et al. Mutational interactions define novel cancer subgroups. *Nature Communications*, 9(1):4353, 2018.

Camila M. Lopes-Ramos, John Quackenbush, and Dawn L. DeMeo. Genome-wide sex and gender differences in cancer. *Frontiers in Oncology*, 10, 2020.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of Graphical Models*. CRC Press, 2018.

Ultan McDermott, James R. Downing, and Michael R. Stratton. Genomics and the continuum of cancer care. *New England Journal of Medicine*, 364(4):340–350, 2011.

Brandon Milholland, Adam Auton, Yousin Suh, and Jan Vijg. Age-related somatic mutations in the cancer genome. *Oncotarget*, 6(28):24627–24635, 2015.

Polina Suter, Jack Kuipers, Giusi Moffa, and Niko Beerenwinkel. Bayesian structure learning and sampling of Bayesian networks with the R package BiDAG. *arXiv preprint*, arXiv:2105.00488, 2021.

TCGA-Research-Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

A Scenarios of the Covariates

We want to cluster the mutational profiles to a number of K individual clusters. We assume that (1.) the mutational profiles X_V can be represented by the probability distributions of K Bayesian networks, and (2.) the cluster-dependent clinical covariates X_C affect mutations via directed edges from covariates to mutations.

In the Bayesian network mixture model, we can consider the following three clustering variations:

- Covariate-adjusted clustering
- Jointly clustering mutational profiles and clinical covariates
- Clustering mutational profiles only

For the covariate-adjusted clustering, the membership probability function is defined as

$$\tilde{\phi}(X_V \mid k) = \frac{\gamma_k \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k\Big)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\Big)} \tag{4}$$

For the clustering of mutational profiles and clinical covariates, it is

$$\phi(X_V, X_C \mid k) = \frac{\gamma_k \cdot P\Big(X_V, X_C \mid \hat{\mathcal{G}}_k, \hat{\theta}_k\Big)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\Big(X_V, X_C \mid \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\Big)} \tag{5}$$

For the clustering of mutational profiles, it is

$$\phi(X_V \mid k) = \frac{\gamma_k \cdot P\Big(X_V \mid \hat{\mathcal{G}}_k, \hat{\theta}_k\Big)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\Big(X_V \mid \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\Big)} \tag{6}$$

If we assume that the probability distribution of the cluster-indepenent covariates X_C is constant across different clusters, i.e. $\exists\, k, k' \in \{1, ..., K\} : P(X_C \mid k) = P(X_C \mid k')$ and if the Bayesian network parameters $(\hat{\mathcal{G}}_k, \hat{\theta}_k)$ are known, then the membership probability functions defined in Equation 4 and Equation 5 are identical:

$$\begin{aligned}
\phi(X_V, X_C \mid k) &= \frac{\gamma_k \cdot P\Big(X_V, X_C \mid \hat{\mathcal{G}}_k, \hat{\theta}_k\Big)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\Big(X_V, X_C \mid \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\Big)} \\
&= \frac{\gamma_k \cdot P\Big(X_C \mid \hat{\mathcal{G}}_k, \hat{\theta}_k\Big) \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k\Big)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\Big(X_C \mid \hat{\mathcal{G}}_{k'}, \hat{\theta}_{k'}\Big) \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\Big)} \\
&= \frac{\gamma_k \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k\Big)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\Big(X_V \mid X_C, \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\Big)} = \tilde{\phi}(X_V \mid k)
\end{aligned}$$

Following Blackwell-Rao theorem, this implies that the variance of Equation 4 is lower than the variance in Equation 5, since the estimation of $P(X_C \mid \hat{\mathcal{G}}_k, \hat{\theta}_k)$ introduces further variance in the estimation. In addition, applying our causal knowledge about the direction of the causal pathways allows us to learn the DAGs $\hat{\mathcal{G}}_k$ more accurately, reducing the variance in $P(X_V \mid \hat{\mathcal{G}}_k, \hat{\theta}_k)$. If the covariates are excluded from the analysis (Equation 6), they act as hidden variables in the probability distribution of the mutations, leading to a lower accuracy in the cluster assignment.

We will now discuss two alternative scenarios to the assumption of a constant probability distribution of the covariates across the clusters. First, instead of being constant, the probability distribution of the covariates could depend on the clusters, i.e. $\exists\, k, k' \in \{1, ..., K\} : P(X_C \mid k) \neq P(X_C \mid k')$. In this case, the covariates would be cluster-dependent covariates that are informative for the cluster assignment. Thus, the covariates should be treated like cluster-dependent covariates without adjustments. Another scenario could be that the probability distribution of the covariates depends on a different grouping than the clusters $g \in \{1, ..., G\}$, i.e. $P(X_C \mid k, g) \neq P(X_C \mid k, g')$. In this case, adjusting for the covariates helps to remove the bias of the other grouping in the clustering of the mutations.

B Algorithm

As outlined in Algorithm 1, we clustered the mutational profiles and clinical covariates using a mixture model. We used a Bernoulli mixture model to initialize the membership probabilities and learned the DAGs $\mathcal{G}_k$ with corresponding local probability distributions θ_k using the **BiDAG** package (GPL-3) [Suter et al., 2021]. Due to the high computational cost of the DAG structure search, we only relearned the DAGs $\mathcal{G}_k$ for every tenth update of the membership probabilities.

Our computations were performed on one CPU core of the AMD EPYC 7H12 processor (2.6 GHz nominal, 3.3 GHz peak) and 256 GB of DDR4 memory clocked at 3200 MHz. Code implementing our method is available at `https://anonymous.4open.science/r/LMRL-NeurIPS`.

A limitation of our algorithm is that computational cost is significantly higher compared to standard clustering algorithms such as k-means or a Bernoulli mixture model. In our simulations, the computation time averaged 4.18 minutes over 20 repetitions with the following parameters: 4 clusters, 20 variables, 5 covariates, and 4800 samples. However, the computational cost can increase significantly for a higher number of variables. To cluster the mutational profiles from the TCGA database (22 clusters, 201 variables, 24 covariates, 8085 samples), our method took 30.32 hours to converge.

Algorithm 1: Covariate-Adjusted Clustering

Input: A matrix of variables X_V and a matrix of covariates X_C
Output: Cluster membership probabilities $\tilde{\phi}(X_V)$ and respective Bayesian networks $(\mathcal{G}_k, \theta_k)$

Initialize membership probabilities $\tilde{\phi}(X_V \mid k)$
repeat
  $i \leftarrow 0$
  Create copy of membership probabilities $\tilde{\phi}_{\text{previous}}(X_V \mid k) \leftarrow \tilde{\phi}(X_V \mid k)$
  Learn the DAGs $\mathcal{G}_k$ given the membership probabilities $\tilde{\phi}(X_V \mid k)$
  repeat
    $i \leftarrow i + 1$
    M-Step: Learn the parameters θ_k given $\mathcal{G}_k$ and $\tilde{\phi}(X_V \mid k)$
    E-Step: Update membership probability: $\tilde{\phi}(X_V \mid k) \leftarrow \frac{\gamma_k \cdot P\left(X_V \mid X_C, \hat{\mathcal{G}}_k, \hat{\theta}_k\right)}{\sum_{k'=1}^{N} \gamma_{k'} \cdot P\left(X_V \mid X_C, \hat{\mathcal{G}}_{k'}, \hat{\theta_{k'}}\right)}$
    Update the cluster weights γ_k: $\gamma_k \leftarrow \frac{\sum_{1=1}^{N} \tilde{\phi}(X_V \mid k)}{N}$
  until $i = 10$;
  Quantify change in membership probabilities $\delta \leftarrow \sum_{X_V} \left(\tilde{\phi}(X_V \mid k) - \tilde{\phi}_{\text{previous}}(X_V \mid k)\right)^2$
until $\delta < \epsilon$;

C Additional Benchmark Information

To assess the performance of our covariate-adjusted algorithm over a range of different scenarios in our benchmark, we varied the following cluster-specific parameters: number of clusters K, number of variables N_V, number of covariates N_C, and number of samples per cluster n_k. While varying one parameter in our benchmark, all other parameters remained constant. The results of the different benchmarks are displayed in Figure 4.

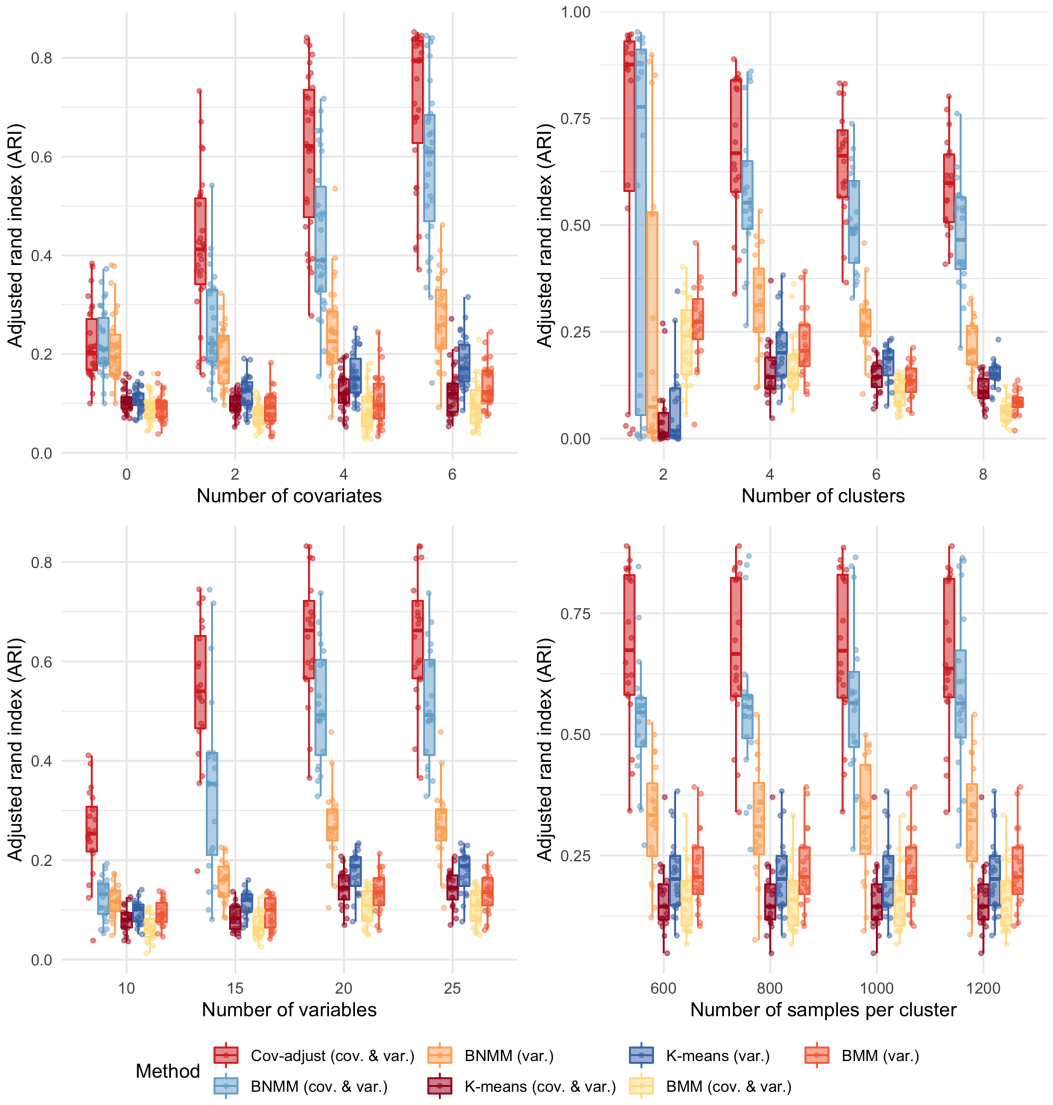

Figure 4: Simulation benchmark over a range of varied cluster-specific parameters including our proposed cluster algorithm (cov-adjust), a Bayesian network mixture model (BNMM), k-means, and a Bernoulli mixture model (BMM).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix B.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2 and Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 3.1.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.2.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] See Appendix B.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 3.1.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data is public and anonymised.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]