# CN-SBM: Categorical Block Modelling For Primary and Residual Copy Number Variation

**Kevin Lam**
Department of Statistics
University of British Columbia
kevin.lam@stat.ubc.ca

**William Daniels, J Maxwell Douglas, Daniel Lai, Samuel Aparicio**
Department of Molecular Oncology
BC Cancer Research Centre
saparicio@bccrc.ca

**Benjamin Bloem-Reddy, Yongjin Park**
Department of Statistics
University of British Columbia
benbr, ypp@stat.ubc.ca

## Abstract

Cancer is a genetic disorder whose clonal evolution can be monitored by tracking noisy genome-wide copy number variants. We introduce the Copy Number Stochastic Block Model (CN-SBM), a probabilistic framework that jointly clusters samples and genomic regions based on discrete copy number states using a bipartite categorical block model. Unlike models relying on Gaussian or Poisson assumptions, CN-SBM respects the discrete nature of CNV calls and captures subpopulation-specific patterns through block-wise structure. Using a two-stage approach, CN-SBM decomposes CNV data into primary and residual components, enabling detection of both large-scale chromosomal alterations and finer aberrations. We derive a scalable variational inference algorithm for application to large cohorts and high-resolution data. Benchmarks on simulated and real datasets show improved model fit over existing methods. Applied to TCGA low-grade glioma data, CN-SBM reveals clinically relevant subtypes and structured residual variation, aiding patient stratification in survival analysis. These results establish CN-SBM as an interpretable, scalable framework for CNV analysis with direct relevance for tumor heterogeneity and prognosis.

## 1 Introduction

Heterogeneity in cancer poses a great challenge in cancer genome analysis. Structural alterations, common in tumors, drive key aspects of tumor evolution such as progression, therapy resistance, and subtype diversity. Copy number variants (CNV) in the single cell DNA sequence (Laks et al., 2019) and tissue-level whole genome sequencing data (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) provide evidence that cancer genomes are rearranged over the course of clonal evolution. Genome-wide CNV profiles offer a broad overview of these large-scale structural variations, such as aneuploidies, large deletions, and duplications. One way this is accomplished is by partitioning the genome into fixed-size bins (e.g., 500 kb or 1 Mb) and estimating integer copy numbers through read counting and normalization.

Copy number calling pipelines, such as HMMcopy (Shah et al., 2006; Lai et al., 2012) and scAbsolute (Schneider et al., 2024), reduce potential biases (GC and mappability) and generate regularized copy number profiles for downstream analysis. At the single-cell level, it is then possible to investigate cell-to-cell variation that gives rise to subpopulation structures, while sample-to-sample heterogeneity can be characterized using bulk sequencing data, such as that provided by the Cancer Genome Atlas project (Weinstein et al., 2013) and even larger data from the ICGC consortium (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).

CNV profiles show us a stationary view of structural variation, exhibiting characteristic block structures. Here, we explore the possibility that CNV profiles contain richer information, which can be decomposed into primary (main) and

Proceedings of the 20th Conference on Machine Learning in Computational Biology (MLCB 2025).

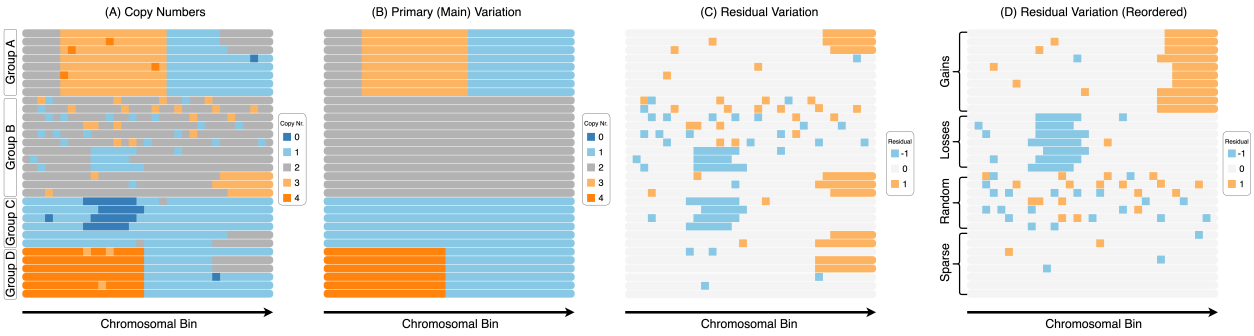


Figure 1: **Example Separation of Main and Residual Copy Number Variation Across Chromosomal Bins. (A)** Raw copy number profiles, visualized per chromosomal bin. **(B)** Main variation pattern capturing the primary copy number trends across samples. **(C)** Residual variation highlighting deviations from the main pattern. **(D)** Residual variation reordered into categories (gains, losses, random, sparse) for clearer interpretation.

residual variation (Fig. 1). Main variation captures large-scale, recurrent alterations that characterize subpopulations, while residual variation reflects finer, cell- and/or sample-specific deviations from these dominant patterns, which may arise from biologically meaningful events such as focal amplifications or deletions. Separating these components enables clearer identification of subpopulation structures while preserving residual signals at the same time (Funnell et al., 2022). This decomposition requires statistical models that can help distinguish structured biological variation from residual variation.

**Contributions.** An ad-hoc approach to modelling integer copy number states as continuous or count-valued variables using Gaussian or Poisson distributions may be computationally convenient, however, such assumptions can misrepresent the inherently discrete and multimodal nature of CNV data. To address this, we introduce the *Copy Number Stochastic Block Model (CN-SBM)*, which directly models copy number states using categorical block models (Keribin et al., 2015). This avoids restrictive distributional assumptions and captures structured, block-wise variation across subpopulations.

We develop a scalable variational inference scheme for CN-SBM, enabling efficient application to large cohorts and high-resolution genomic data. In contrast to the approach in Keribin et al. (2015), our method updates the variational distributions directly and allows for empty clusters, implicitly enabling automatic model selection for the number of clusters. CN-SBM also lends itself naturally to a two-stage analysis: the first stage captures primary variation, while the second identifies finer-scale residual variations that emerge after accounting for first-stage primary variations. For datasets with approximately 1,000 samples at 500 kb resolution, CN-SBM is computationally efficient, often converging within minutes when accelerated by GPU computing (Fig. 5).

We made our code available at: https://github.com/lamke07/cnsbm2025.

## 2 Methods

**Model Setup** To facilitate modelling, we consider scWGS data in a matrix format, where each row corresponds to a cell, and each column represents the copy number within a fixed-size chromosomal bin (e.g. 500 kilobases (kb)), ensuring consistency in data representation. Let $\mathbf{C} = (c_{ij})$ denote the $N \times M$ copy number data matrix, where $c_{ij} \in \mathbb{N}_0$ corresponds to the copy number of cell $i$ at bin $j$ across the genome. In the *(bipartite) copy number stochastic block model (CN-SBM)* we propose the use of categorical block models (Keribin et al., 2015): in this setup, each cell $i$ belongs to a **cell cluster** with up to $K$ components, indexed by $g_i \in \{1, \dots, K\}$, and each bin $j$ belongs to a **bin cluster** with up to $L$ components, indexed by $h_j \in \{1, \dots, L\}$.

The observed copy number $c_{ij}$ is modelled as a categorical variable whose distribution depends only on the latent cluster pair $(g_i, h_j)$ (Fig. 2). Copy number values are assumed to lie in a fixed, discrete set $\mathcal{C} = \{0, 1, \dots, \geq 11\}$ with $|\mathcal{C}| = n_{\text{cat}}$, where highly amplified values are grouped into a final category. The generative process is:

1. For each cell $i \in \{1, \dots, N\}$, draw cell cluster assignment $g_i \sim \text{Cat}(\boldsymbol{\pi}^g)$.
2. For each bin $j \in \{1, \dots, M\}$, draw bin cluster assignment $h_j \sim \text{Cat}(\boldsymbol{\pi}^h)$ .
3. For each pair $(i, j)$, draw copy number $c_{ij} \sim \text{Cat}\left(\boldsymbol{\pi}^{(g_i, h_j)}\right)$, where $\boldsymbol{\pi}_{k,l}$ is the categorical distribution over copy number values for the cluster pair $(k, l)$.

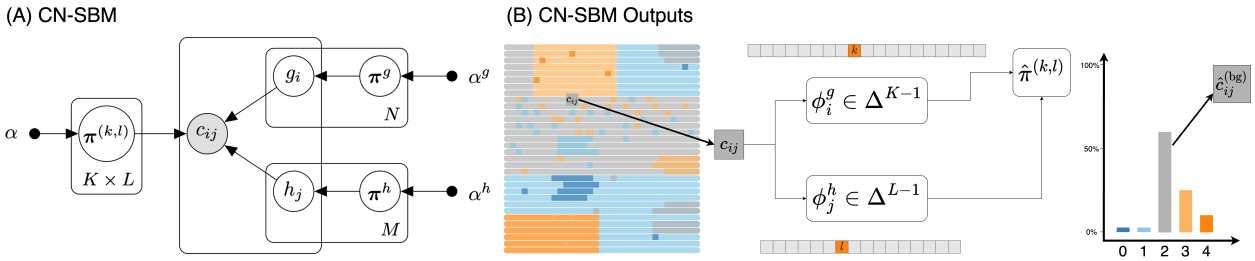


Figure 2: **Overview of the CN-SBM model for primary variation. (A)** Graphical model of the CN-SBM. Observed copy number states $c_{ij}$ are generated based on latent cell ($g_i$) and bin ($h_j$) cluster assignments, with cluster pairs governed by a categorical distribution $\boldsymbol{\pi}^{(g_i,h_j)}$. **(B)** Outputs of CN-SBM: for each observed value $c_{ij}$, the **primary** copy number state $\hat{c}_{ij}^{(bg)}$ is inferred from the most probable cluster pair $(\hat{g}_i, \hat{h}_j)$ under the soft cluster assignments $\phi_i^g, \phi_j^h$ and the corresponding estimated categorical distribution $\hat{\boldsymbol{\pi}}^{(\hat{g}_i,\hat{h}_j)}$ over the copy numbers.

We place Dirichlet priors over all categorical distributions, that is, $\boldsymbol{\pi}^g \sim \text{Dir}_{[K]}(\alpha^g)$ and $\boldsymbol{\pi}^h \sim \text{Dir}_{[L]}(\alpha^h)$ for the possible cell and bin cluster assignments, respectively, and $\boldsymbol{\pi}^{(k,l)} \sim \text{Dir}_{[n_{\text{cat}}]}(\alpha)$, for each of the $K \cdot L$ individual block (cluster) distributions, $1 \le k \le K, 1 \le l \le L$. Assuming independence among cluster assignments and copy number observations, the joint distribution factorizes as:

$$p(\mathbf{C}, \mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\}) = \underbrace{\prod_{i=1}^{N}\prod_{j=1}^{M} \text{Cat}(c_{ij}; \boldsymbol{\pi}^{(g_i,h_j)})}_{\text{data model}} \cdot \underbrace{\prod_{i=1}^{N} \text{Cat}(g_i; \boldsymbol{\pi}^g)}_{\text{cell clusters}} \cdot \underbrace{\prod_{j=1}^{M} \text{Cat}(h_j; \boldsymbol{\pi}^h)}_{\text{bin clusters}}$$
$$\cdot \underbrace{\prod_{k=1}^{K}\prod_{l=1}^{L} \left(\text{Dir}_{[n_{\text{cat}}]}(\boldsymbol{\pi}^{(k,l)}; \alpha)\right)}_{\text{block distribution prior}} \cdot \underbrace{\text{Dir}_{[K]}(\boldsymbol{\pi}^g; \alpha^g)}_{\text{cell cluster prior}} \cdot \underbrace{\text{Dir}_{[L]}(\boldsymbol{\pi}^h; \alpha^h)}_{\text{bin cluster prior}}.$$

Here, $\{\boldsymbol{\pi}^{(k,l)}\} := \{\boldsymbol{\pi}^{(k,l)}\}_{k\in[K],l\in[L]}$ denotes the collection of all $K \cdot L$ block-specific probability vectors. Using this factorization, the complete conditional distributions for the latent variables are:

$$\begin{aligned}
p(g_i = k|\mathbf{h}, \boldsymbol{\pi}^g, \{\boldsymbol{\pi}^{(k,l)}\}, \mathbf{c}_{i,:}) &\propto \exp\left(\sum_{j=1}^{M}\sum_{l=1}^{L} \mathbb{1}(h_j = l) \log \pi_{c_{ij}}^{(k,l)} + \log \pi_k^g\right) \\
p(h_j = l|\mathbf{g}, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\}, \mathbf{c}_{:,j}) &\propto \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}(g_i = k) \log \pi_{c_{ij}}^{(k,l)} + \log \pi_l^h\right) \\
p(\boldsymbol{\pi}^g|\mathbf{g}) &= \text{Dir}_{[K]}\left(\boldsymbol{\pi}^g; \alpha^g + \sum_{i=1}^{N} \mathbb{1}(g_i = \cdot)\right) \\
p(\boldsymbol{\pi}^h|\mathbf{h}) &= \text{Dir}_{[L]}\left(\boldsymbol{\pi}^h; \alpha^h + \sum_{j=1}^{M} \mathbb{1}(h_j = \cdot)\right) \\
p(\boldsymbol{\pi}^{(k,l)}|\mathbf{g}, \mathbf{h}, \mathbf{C}) &= \text{Dir}_{[n_{\text{cat}}]}\left(\boldsymbol{\pi}^{(k,l)}; \alpha + \sum_{i=1}^{N}\sum_{j=1}^{M} \mathbb{1}(g_i = k)\mathbb{1}(h_j = l)\mathbb{1}(c_{ij} = \cdot)\right),
\end{aligned} \tag{2.1}$$

where we have leveraged the conjugacy between the Dirichlet and categorical distributions for the latent probability vectors $\boldsymbol{\pi}^g$, $\boldsymbol{\pi}^h$, $\{\boldsymbol{\pi}^{(k,l)}\}$. The '$\cdot$' symbol in the indicator functions denotes the appropriate index over the relevant latent variable dimension, allowing for a concise representation of cluster-specific counts.

**Variational Inference** The CN-SBM assumes a latent variable model setup of the form $p(\mathbf{C}, \mathbf{z}) = p(\mathbf{C}|\mathbf{z})p(\mathbf{z})$, where $\mathbf{z} = (\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})$ contains all latent variables. We adopt a mean-field variational approximation (Ghahramani

and Beal, 2000) of the posterior that factorizes over the latent variables as:

$$q(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\}) = \prod_{i=1}^{N} q(g_i|\boldsymbol{\phi}^g) \cdot \prod_{j=1}^{M} q(h_j|\boldsymbol{\phi}^h) \cdot \prod_{k=1}^{K}\prod_{l=1}^{L} \left[q(\boldsymbol{\pi}^{(k,l)}|\boldsymbol{\gamma}^{(k,l)})\right] \cdot q(\boldsymbol{\pi}^g|\boldsymbol{\gamma}^g) \cdot q(\boldsymbol{\pi}^h|\boldsymbol{\gamma}^h),$$

where $q(g_i = k|\boldsymbol{\phi}^g) = \phi^g_{ik}$ and $q(h_j = l|\boldsymbol{\phi}^h) = \phi^h_{jl}$ are variational distributions corresponding to soft cluster assignments for the cells and bins. The variational distributions over the Dirichlet parameters are modeled as $\boldsymbol{\pi}^g \sim \text{Dir}_{[K]}(\boldsymbol{\gamma}^g)$, $\boldsymbol{\pi}^h \sim \text{Dir}_{[L]}(\boldsymbol{\gamma}^h)$, and $\boldsymbol{\pi}^{(k,l)} \sim \text{Dir}_{[n_{\text{cat}}]}(\boldsymbol{\gamma}^{(k,l)})$ for each block $(k, l)$. The evidence lower bound (ELBO) on the log-marginal likelihood $\log p(\mathbf{C})$ is:

$$\log p(\mathbf{C}) \geq \mathbb{E}_q[\log p(\mathbf{C}|\mathbf{g}, \mathbf{h}, \{\boldsymbol{\pi}^{(k,l)}\})] - D_{KL}(q(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})||p(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})),$$

which can be computed analytically in the CN-SBM setup (Appendix A.1). We aim to maximize the ELBO using **Coordinate Ascent Variational Inference (CAVI)** (Blei and Jordan, 2006; Bishop, 2006; Blei et al., 2017), an iterative optimization algorithm for Bayesian inference. For a factorized variational distribution $q(\mathbf{z}) = \prod_{i=1}^{m} q_i(\mathbf{z}_i)$, the optimal update for each $q_j(\mathbf{z}_j)$ while holding the remaining variables $\mathbf{z}_i, i \neq j$, fixed is

$$q_j(\mathbf{z}_j) \propto \exp\left(\mathbb{E}_{i\neq j}[\log p(\mathbf{z}_j|\mathbf{z}_{-j}, \mathbf{x})]\right), \tag{2.2}$$

where the expectation is taken with respect to the variational distributions over all other latent variables $\mathbf{z}_{-j}$. This update has a closed-form expression requiring only natural parameter updates when the complete conditionals $p(\mathbf{z}_j|\mathbf{z}_{-j}, \mathbf{x})$ belong to an exponential family, which is the case in the CN-SBM framework. In particular, the variational update steps are analytically derived by computing expectations over the conditionals presented in eq. (2.1). The full variational inference scheme is outlined in algorithm 1. In this context, $\psi$ is the digamma function, which provides the expectation of a log-transformed Dirichlet variable, given by the form $\mathbb{E}[\log \pi_{k'}] = \psi(\alpha_{k'}) - \psi\left(\sum_k \alpha_k\right)$. This expectation quantifies the relative importance of category $k'$ within the total concentration of the Dirichlet distribution. These updates are guaranteed to converge to a local optimum of the ELBO (Boyd and Vandenberghe, 2004; Bishop, 2006).

**Algorithm 1** Variational inference for the CN-SBM

**Require:** Copy number matrix $\mathbf{C} = (c_{ij})$.

Initialize local parameters $\boldsymbol{\phi}^g, \boldsymbol{\phi}^h$ and global parameters $\boldsymbol{\gamma}^g, \boldsymbol{\gamma}^h, \{\boldsymbol{\gamma}^{(k,l)}\}$.

**while** ELBO not converged **do**

**Local updates for each row and column**

- for row $i = 1, \dots N$ and row clusters $k = 1, \dots, K$
  - $q(g_i = k) \propto \exp\left(\sum_{j=1}^{M}\sum_{l=1}^{L} \phi^h_{jl} \cdot \left(\psi(\gamma^{(k,l)}_{c_{ij}}) - \psi\left(\sum_{c\in\mathcal{C}} \gamma^{(k,l)}_c\right)\right) + \psi(\gamma^g_k) - \psi\left(\sum_{k=1}^{K} \gamma^g_k\right)\right)$
- for column $j = 1, \dots M$ and column clusters $l = 1, \dots, L$
  - $q(h_j = l) \propto \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K} \phi^g_{ik} \cdot \left(\psi(\gamma^{(k,l)}_{c_{ij}}) - \psi\left(\sum_{c\in\mathcal{C}} \gamma^{(k,l)}_c\right)\right) + \psi(\gamma^h_l) - \psi\left(\sum_{l=1}^{L} \gamma^h_l\right)\right)$

**Global updates for cluster and cluster proportions**

- $\boldsymbol{\gamma}^g = \alpha^g + \sum_{i=1}^{N} \phi^g_{i,\cdot}$
- $\boldsymbol{\gamma}^h = \alpha^h + \sum_{j=1}^{N} \phi^h_{j,\cdot}$
- $\boldsymbol{\gamma}^{(k,l)} = \alpha + \sum_{i=1}^{N}\sum_{j=1}^{M} \phi^g_{ik}\phi^h_{jl}\mathbb{1}(c_{ij} = \cdot), \qquad 1 \leq k \leq K, 1 \leq l \leq L.$

**end while**

The local updates for cell cluster assignments $g_i$ and bin cluster assignments $h_j$ can be viewed as soft assignments over $K$ and $L$ clusters, respectively, where each cluster's relevance is determined by expected log-probabilities of the observed data categories. These expectations are computed as weighted averages over the columns (for cells) or rows (for bins), using the current corresponding soft responsibilities $\phi^h_{jl}$ and $\phi^g_{ik}$. The prior expectations $\mathbb{E}_q[\log \pi^g_k]$ and $\mathbb{E}_q[\log \pi^h_l]$ act as regularizers, reflecting the influence of Dirichlet priors. Global parameters $\boldsymbol{\gamma}^g$, $\boldsymbol{\gamma}^h$, and $\boldsymbol{\gamma}^{(k,l)}$ summarize the local assignments by aggregating responsibilities across cells, bins, and block interactions, with each update incorporating the corresponding prior parameters $\alpha^g, \alpha^h$ and $\alpha$, respectively.

**Missing Data** To handle missing values, we extended the CAVI algorithm to optimize the ELBO over the full data-generating process, bounding $\log p(\mathbf{C}_{\text{obs}}, \mathbf{C}_{\text{mis}})$, rather than just $\log p(\mathbf{C}_{\text{obs}})$. This encourages the latent structure to reflect the entire dataset, rather than only the observed portion. Let $\mathbf{B} = (b_{ij})$ denote the binary missingness matrix, where $b_{ij} = 1$ if $c_{ij}$ is observed. We introduce importance weights $w_{ij}$ to adjust each observation's influence, applying them as multiplicative factors in the variational updates in Algorithm 1 and ELBO computations (Appendix A.3). For instance, setting $w_{ij} = b_{ij}$ ignores missing entries, while $w_{ij} = b_{ij}/\hat{\zeta}_{ij}$ incorporates inverse-propensity weighting via an estimated missingness model $\hat{\zeta}_{ij}$, e.g. from logistic regression or frequency-based heuristics.

**Initialization and Model Refinement** Since CAVI guarantees only local convergence, the choice of initialization of variational parameters affects the quality of the final solution, with the converged ELBO often varying substantially across runs (Blei et al., 2017). To improve robustness and convergence, we support several initialization schemes for the variational parameters. These include random sampling from the prior and clustering-informed initialization such as k-means (MacQueen, 1967; Lloyd, 1982) and spectral clustering (Ng et al., 2001) applied independently to rows and columns. For joint initialization, we also incorporate spectral biclustering (Kluger et al., 2003), which empirically yielded the best performance in our experiments.

Given a fitted model, local exploration strategies can be applied that search for improved optima by splitting or merging clusters based on the posterior MAP hard assignments $\hat{\mathbf{g}}_{\text{MAP}}, \hat{\mathbf{h}}_{\text{MAP}}$. These splits can be performed by reassigning labels within existing sub-clusters either via standard clustering methods or by assigning new labels according to the categorical mode. To correct for over-segmentation, clusters can be merged when redundancy is detected, such as nearly identical dominant categories or when cluster sizes fall below a certain threshold, based on the number of cells (rows) or bins (columns). Throughout this process, model selection can be guided by improvements in the ELBO or by penalized criteria such as the integrated completed likelihood (ICL) (Biernacki et al., 2000; Côme and Latouche, 2015; Bar-Hen et al., 2022), which account for both model fit and complexity.

## 3 Experiments

**Datasets** We consider three datasets for our experiments, all requiring a copy number matrix with rows as cells or patient samples and columns as genomic bins. Since the CN-SBM accepts categorical inputs, it can accommodate both single-cell and bulk copy number data. First, we generated 2,500 single-cell copy number profiles across eight clones using the `CNAsim` simulator (Weiner and Bansal, 2023). This simulation includes chromosome-arm and whole-chromosome-level copy number alterations, along with configurable noise to control segment lengths and per-bin copy number jitter. The resulting profiles were aggregated into non-overlapping 500 kb genomic bins. Next, we incorporated single-cell data from the study by Funnell et al. (2022), which includes DLP+ sequencing of patient-derived xenografts (PDX) from various tumours collected over multiple time points spanning several years. Copy number calling was performed using HMMCopy (Shah et al., 2006; Lai et al., 2012) at 500 kb resolution, yielding total copy number profiles. Specifically, we considered the OV2295 ($N = 1084$), SA1096 ($N = 802$) and SA535 ($N = 1801$) cell lines.

Lastly, we also evaluated our method on bulk sequencing data from The Cancer Genome Atlas (TCGA) project (Weinstein et al., 2013), where each matrix row corresponds to a patient tumour sample. Copy number states were previously inferred using the ASCAT algorithm (Van Loo et al., 2010), which estimates allele-specific copy numbers. We rebinned the segmented data into 500 kb bins to match the resolution used in our single-cell analyses and computed the total copy number as the sum of major and minor alleles. Unlike single-cell data, this dataset reflects inter-patient population-level copy number diversity rather than intra-tumour heterogeneity. For benchmarking, we selected breast cancer ($N = 998$), ovarian cancer ($N = 532$), and low-grade glioma ($N = 490$). Some datasets have missing values in the copy number matrix, which can be due to low sequencing coverage, high-variability regions (e.g., mutation hotspots), or technical artifacts like alignment errors and GC-content bias. Since standard clustering algorithms, including those in our study, do not handle missing data, we impute missing entries by sampling from the empirical marginal distribution of observed copy number states across five seeds.

**Experiment Settings** We evaluate two experimental settings. In the first, models are trained on the fully imputed dataset to assess clustering performance under complete-data assumptions. In the second, we simulate additional missingness by randomly withholding approximately 1% of observed entries, which are excluded during training and used as a held-out set to evaluate predictive accuracy. Specifically, we assess how well the fitted categorical stochastic block model infers the unseen values, based on the posterior over latent assignments and block parameters. We benchmarked our method against several co-clustering approaches, including the Poisson bipartite stochastic block model (PoissonSBM) (Bar-Hen et al., 2022; Chiquet et al., 2024), the categorical latent block model (Blockcluster) (Keribin et al., 2015; Bhatia et al., 2017), k-means clustering (KMeans), and spectral biclustering with log-transformed or bi-stochastic input (SpecBi). All methods can produce discrete partitions over the rows (cells/samples) and columns (genomic bins), allowing direct comparison. We excluded biclustering methods that allow overlapping cluster memberships, such as the FABIA (Hochreiter et al., 2010) and Cheng and Church algorithm (Cheng and Church, 2000).

We fixed the maximum number of row and column clusters to $K = 10, L = 30$ for the CNAsim and TCGA datasets, and $K = 15, L = 30$ for the Funnell datasets, based on prior cluster expectations and practical run-time constraints. For example, PoissonSBM uses top-down hierarchical agglomeration with repeated model fitting to optimize the ICL, making model selection computationally prohibitive. KMeans and spectral biclustering do not provide probabilistic outputs. As such, we estimated empirical categorical distributions within each block (row-column cluster pair) using

| | Method | CNAsim | TCGA | | | Funnell | | |
|---|---|---|---|---|---|---|---|---|
| | | Simulated ($N = 2500$) | BRCA ($N = 998$) | OV ($N = 532$) | LGG ($N = 490$) | OV2295 ($N = 1084$) | SA1096 ($N = 802$) | SA535 ($N = 1801$) |
| **LL** $\times 10^3$ (↑) | SpecBi (bist) | -89.1 ± 0.3 | -75.3 ± 0.1 | -45.5 ± 0.3 | -19.0 ± 0.2 | -54.7 ± 1.0 | -49.0 ± 0.4 | -107.2 ± 1.7 |
| | SpecBi (log) | -88.1 ± 0.6 | -74.7 ± 0.2 | -44.4 ± 0.1 | -18.8 ± 0.2 | -55.5 ± 1.0 | -48.2 ± 0.8 | -97.5 ± 1.1 |
| | KMeans | -41.0 ± 0.6 | -60.1 ± 0.2 | -37.6 ± 0.1 | **-11.0 ± 0.1** | -37.7 ± 0.6 | -27.3 ± 0.4 | -43.9 ± 1.0 |
| | PoissonSBM | -41.6 ± 3.0 | -115.8 ± 18.0 | -59.7 ± 2.9 | -125.5 ± 32.8 | -60.4 ± 8.2 | -46.0 ± 8.0 | -44.3 ± 3.3 |
| | Blockcluster | -38.1 ± 0.4 | **-57.0 ± 0.2** | -37.0 ± 0.3 | -11.5 ± 0.3 | -39.0 ± 0.8 | -27.7 ± 0.4 | -46.9 ± 1.7 |
| | CN-SBM (ref.) | **-37.6 ± 0.3** | -57.1 ± 0.1 | **-36.8 ± 0.1** | -11.3 ± 0.2 | **-36.3 ± 1.2** | **-26.5 ± 0.5** | **-42.5 ± 0.4** |
| **ICL** $\times 10^6$ (↑) | SpecBi (bist) | -8.89 ± 0.02 | -7.52 ± 0.03 | -4.58 ± 0.02 | -1.92 ± 0.02 | -5.62 ± 0.07 | -4.99 ± 0.10 | -11.27 ± 0.61 |
| | SpecBi (log) | -8.82 ± 0.07 | -7.52 ± 0.05 | -4.47 ± 0.01 | -1.93 ± 0.02 | -5.58 ± 0.13 | -4.95 ± 0.06 | -10.08 ± 0.40 |
| | KMeans | -4.04 ± 0.09 | -6.02 ± 0.03 | -3.81 ± 0.01 | -1.13 ± 0.01 | -3.72 ± 0.03 | -2.76 ± 0.02 | -4.42 ± 0.05 |
| | PoissonSBM | -5.31 ± 0.00 | -12.77 ± 1.26 | -6.06 ± 0.21 | -15.97 ± 1.12 | -5.08 ± 0.00 | -5.28 ± 0.00 | -4.35 ± 0.00 |
| | Blockcluster | -3.78 ± 0.06 | -5.72 ± 0.01 | -3.74 ± 0.01 | -1.21 ± 0.01 | -4.11 ± 0.11 | – | -4.68 ± 0.04 |
| | CN-SBM (ref.) | **-3.58 ± 0.02** | **-5.62 ± 0.01** | **-3.70 ± 0.01** | **-1.12 ± 0.01** | **-3.54 ± 0.10** | **-2.67 ± 0.02** | **-4.20 ± 0.09** |

Table 1: Held-out log-likelihood (top) and ICL (bottom) across simulated and real datasets. Values were scaled by $10^3$ for log-likelihood and $10^6$ for ICL. Dashes indicate missing values due to failed runs.

observed frequency counts. The block distributions were estimated as $\hat{\pi}_c^{(k,l)} = (n_c^{(k,l)})/(\sum_{c'} n_{c'}^{(k,l)})$, where $n_c^{(k,l)}$ is the count of copy number state $c$ in block $(k, l)$. These distributions were used to compute ICL and held-out log-likelihood scores. For PoissonSBM and Blockcluster, we also used empirical block distributions, as the inferred probability distributions from their implementations generally yielded poorer results. For PoissonSBM, the assumed data distribution struggles to assign high probability to individual counts, while the block probabilities inferred by Blockcluster sometimes produced extreme ICL and log-likelihood values, distorting average performance metrics (Appendix B.2 and Table 3).

**Model Fit** Table 1 provides an overview of how the CN-SBM performs against the other baselines. Three complementary criteria are reported: (i) held-out log-likelihood, (ii) integrated completed-likelihood (ICL), which penalises overly complex latent structures while rewarding good data fit, and (iii) weighted average normalized entropy to assess cluster purity (Supplementary Table 2). The model *CN-SBM (ref.)* explores the partition space using posterior parameters from a single CN-SBM run, similarly to PoissonSBM.

Model quality is evaluated using held-out log-likelihood, measuring how well the predicted distribution aligns with unseen data, and the ICL, which assesses overall model fit on the full training data. The CN-SBM achieves the highest held-out log-likelihood across most datasets, and performance margins generally increased with the size of the data. The large performance gap between PoissonSBM and other methods on the TCGA datasets and OV2295 is due to how PoissonSBM handles cluster specification. When given $K$ row and $L$ column clusters, it searches over a total of $K + L$ clusters without fixing the row and column counts. As a result, the final model often assigns more row or column clusters than specified, for example, averaging 25 row clusters in BRCA despite $K = 10$, and similarly for OV and LGG. This may be due to the wide range of copy number variation that cannot be reflected by a unimodal count distribution such as the Poisson distribution. The ICL includes a complexity penalty and thus favours simpler models when performance is comparable. The CN-SBM yields the highest ICL across all datasets (Table 1), indicating that its inferred block structure and block distributions effectively capture underlying data patterns best.

As a second order assessment, we also computed the average normalized entropy weighted by subcluster size to analyze the within-cluster purity of the inferred sub-clusters, weighted by sub-cluster sizes. Low entropy indicates clear separation between primary copy number variation and residual variation, enabling the identification of distinct groups without confounding noise and reflecting the model's ability to capture major sources of variation. We report the results in Supplementary Table 2: the CN-SBM achieves the lowest weighted entropy across most datasets, significantly outperforming spectral biclustering and showing similar performance to PoissonSBM and Blockcluster.

Standard deviations across five runs remain comparable with the baseline methods for the CN-SBM, highlighting its robustness despite variation in cell/sample counts and copy number patterns. Blockcluster often failed to converge (see SA1096 and Table 4). For the PoissonSBM, the Funnell datasets (OV2295, SA1096, SA535) and CNAsim are fully observed, so no imputation was needed, which eliminates a potential source of variability. As such, PoissonSBM showed almost no variance on these datasets due to its agglomerative fitting process, which consistently converged to the same cluster partitioning across random initializations.

## 4 Primary and Residual Variation in Low-Grade Glioma

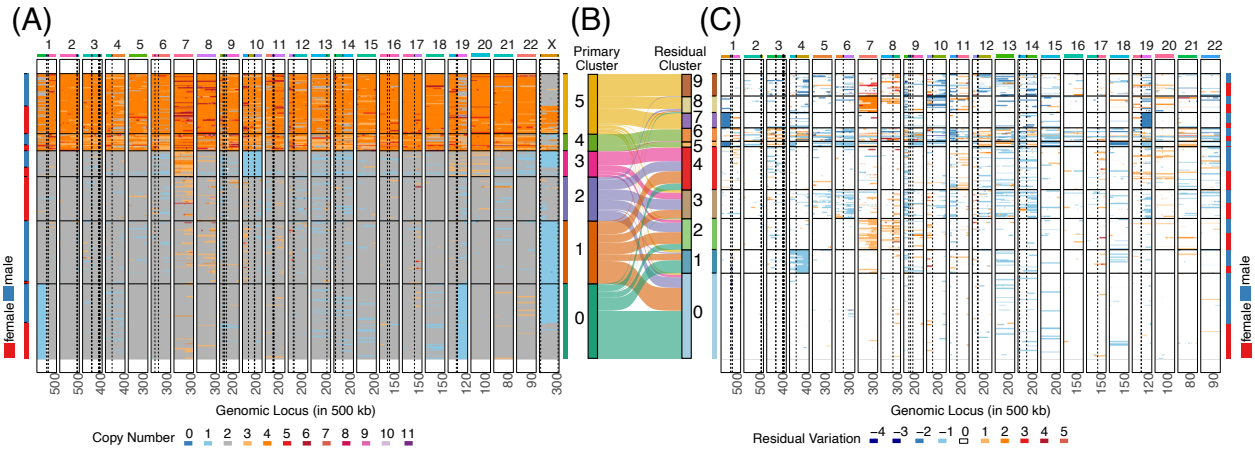


Figure 3: **Main and residual variation in TCGA low-grade glioma** ($N = 484$). **(A)** CN-SBM fitted to the LGG copy number profiles in 500 kb resolution, with samples and bins coloured according to their cluster. **(B)** Alluvial diagram linking sample cluster assignment in main variation (A) vs. residual variation (C). **(C)** Residual copy-number landscape after subtraction of main variation obtained from the CN-SBM. Samples have been reordered according to the clustering obtained from a second categorical SBM fitted to this matrix, identifying ten finer *residual* clusters.

We analyze somatic copy number alterations from the TCGA low-grade glioma (LGG) dataset, comprising $N = 490$ samples. Copy number states are thresholded at 11, with high or extreme amplifications categorized as 11 to reduce sparsity. To capture the main variation, we fit the CN-SBM and assign hard clusters. Within each co-cluster (block), we summarize copy number patterns by selecting the copy number with the highest posterior Dirichlet-mean (Fig. 2B), forming a *main variation* matrix of representative values per sample-bin pair. Subtracting this matrix from the original copy number profiles yields a residual matrix, with negative values indicating underrepresentation. We re-encode these deviations categorically and apply the CN-SBM again to capture finer structure. As before, hard cluster assignments are obtained by computing the posterior argmax assignment. This two-stage approach decomposes copy number variation into main and residual components using categorical block models. The `X` chromosome is excluded in the second stage to avoid sex-related confounding.

Application of the CN-SBM identified ten *main (variation)* clusters within the samples. Four of these, comprising just six outlier samples with profiles dissimilar from the other main patterns, were excluded to focus on six dominant clusters ($N = 484$) that capture the primary structure of copy number variation (Fig. 3A). Reordering the copy number matrix by cluster assignments reveals distinct genomic signatures. Clusters 0-3 display arm-level and whole-chromosome alterations involving chromosomes 1, 7, 10, 19, and X. In contrast, clusters 4 and 5 are marked by widespread amplification, with modal states near four, consistent with whole-genome doubling events.

Having applied the CN-SBM to the residual matrix, the second-stage analysis uncovered structured patterns that were not apparent in the primary clustering. As illustrated in the Sankey diagram (Fig. 3B), samples from the same primary cluster often split into distinct residual clusters, indicating shared patterns of deviation within otherwise similar profiles (Fig. 3C). Residual cluster 0 shows minimal deviation, suggesting strong alignment with the dominant structure. In contrast, other residual clusters reveal specific alterations: cluster 1 shows chromosome 4 loss; cluster 2 exhibits gains on chromosomes 7 and 8; cluster 3 is characterized by widespread deletions affecting chromosomes 5, 6, 9, 13, 14, and 19. Additional residual clusters 5, 7, and 8 display complex patterns involving chromosomes 1, 7, and 19. These may highlight biologically relevant variation that emerges after characterizing the primary variations.

To evaluate the prognostic relevance of the CN-SBM-inferred primary (main) and residual variation, we performed Cox proportional hazards modelling using patient survival data. The cohort comprised 213 female patients (44%), with ages ranging from 14 to 87 years (median 41.0 years, IQR 32.8-52.8). Age was discretized into three categories: less than 40, 40-60, and 60+ years. In an initial model excluding CN-SBM clusters, sex was not significantly associated with survival, which is consistent with its balanced distribution across main copy number clusters (clusters 0, 3, 4, and 5; Fig. 3A). As such, the baseline model included only age group as a covariate and achieved a concordance index (C-index) of 0.741 (SE = 0.031).

Through model selection (Appendix B.3), we incorporated both the main and a part of the residual CN-SBM cluster assignments. This substantially improved model analytics, yielding a C-index of 0.855 (SE = 0.019) (Fig. 4A). This

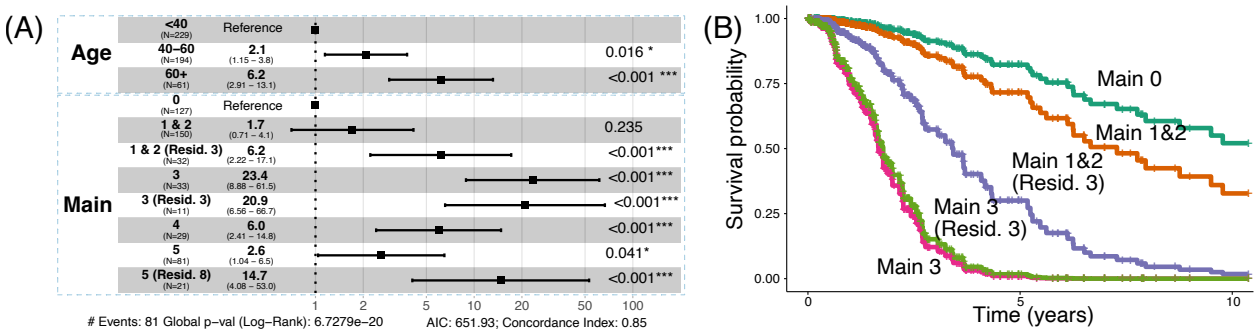


Figure 4: **(A)** Forest plot for Cox proportional hazards model evaluating impact of age groups and *main* cluster membership on patient survival. Distinct copy-number profiles segregate patients into different survival trajectories (log-rank $p < 0.001$). **(B)** Kaplan-Meier survival curves for patient profiles with 40-60 age group and varying clusters.

highlights the strong predictive and explanatory value of the discovered genomic subtypes. Most main clusters were significantly associated with survival: clusters 1 and 2 were combined (*1 & 2*) due to minimal differences, primarily limited to `X` chromosome alterations (Fig. 3A). Residual clusters 3 and 8 added further resolution: for instance, residual cluster 3 significantly modified hazard ratios within main clusters 1 and 2, while residual cluster 8 influenced outcomes among patients in main cluster 5. These results demonstrate that residual variation can encode information beyond that captured by dominant copy number profiles. As shown in Fig. 4B, predicted survival risk among patients aged 40-60 varied substantially across combinations of main and residual cluster memberships. Broad chromosomal alterations (main clusters) and finer-grained aberrations (residual clusters) provide complementary and additional stratification of LGG tumours. The two-stage CN-SBM framework thus uncovers biologically and clinically meaningful substructures with direct relevance for survival prediction.

## 5 Discussion

SBMs have previously been used in various biological applications: for instance, to uncover clonal structures from binary mutation matrices (Myers et al., 2020) and to identify rare cell populations in single-cell RNA-seq data using *nested SBMs* (Morelli et al., 2021). The single-cell genotyper (Roth et al., 2016) performed clonal genotype inference via mixture modelling. We introduced the CN-SBM, a probabilistic block model for characterizing primary copy number variation and isolating residual deviations, by jointly clustering samples and genomic bins while preserving the categorical nature of copy number states. As such, CN-SBM is well-suited to the multimodal distributions typical of copy number alterations. A key strength of the model is its use in a two-stage modelling approach: the first stage captures primary chromosomal alterations, while the second isolates structured residual variation. Our analysis shows that both components provide complementary and independently prognostic information, highlighting the clinical utility of decomposing genomic variation into coarse and fine-grained layers for survival analysis.

CN-SBM can scale effectively to large cohorts and finer bin resolutions using stochastic variational inference (see Appendix A.2) and exhibits stable convergence across diverse datasets. In contrast, existing baselines such as Blockcluster frequently encounter convergence issues and require repeated initializations to produce stable results. While CN-SBM captures important genomic substructure, it currently assumes independence across genomic bins, which may limit its ability to detect spatially correlated patterns. For the characterization of primary variation, the CN-SBM benefits from pre-segmented input via existing copy number callers.

Beyond clustering, CN-SBM facilitates downstream analysis through interpretable feature summaries. By leveraging the genomic ordering of bins, contiguous regions sharing cluster labels can be merged into higher-level segments, enabling the computation of summary statistics such as event frequency and length. These derived features may support the identification of mutational signatures or focal aberrations and enhance the integration of CN-SBM outputs into broader analytical workflows.

An important avenue for future work lies in extending CN-SBM to integrate multi-omic data. Recent single-cell technologies enable simultaneous DNA and RNA profiling (e.g. Macaulay et al. (2015)), offering the opportunity to model joint structure across copy number alterations and gene expression. A multimodal extension of CN-SBM, incorporating shared and modality-specific latent variables, could improve stratification and yield deeper insight into the relationship between genomic alterations and transcriptional programs.

## Acknowledgments and Disclosure of Funding

YP and KL acknowledge the support of the CANSSI Graduate Student Enrichment Scheme (GSES). BBR acknowledges the support of NSERC: RGPIN2020-04995, RGPAS-2020-00095. YP acknowledges the support of the Canada Research Chair Tier 2 program and the NSERC Discovery Grant. We thank Charles Gadd for his insightful discussions and his assistance in accessing the TCGA datasets.

## References

Bar-Hen, A., Barbillon, P., and Donnet, S. (2022). Block models for generalized multipartite networks: Applications in ecology and ethnobiology. *Statistical Modelling*, 22(4):273–296.

Bhatia, P. S., Iovleff, S., and Govaert, G. (2017). Blockcluster: An R Package for Model-Based Co-Clustering. *Journal of Statistical Software*, 76:1–24.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, 1st ed. 2006. corr. 2nd printing 2011 edition.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 8:93–103.

Chiquet, J., Donnet, S., team, g., and Barbillon, P. (2024). Sbm: Stochastic Blockmodels.

Côme, E. and Latouche, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15(6):564–589.

Funnell, T., O'Flanagan, C. H., Williams, M. J., McPherson, A., McKinney, S., Kabeer, F., Lee, H., Salehi, S., Vázquez-García, I., Shi, H., Leventhal, E., Masud, T., Eirew, P., Yap, D., Zhang, A. W., Lim, J. L. P., Wang, B., Brimhall, J., Biele, J., Ting, J., Au, V., Van Vliet, M., Liu, Y. F., Beatty, S., Lai, D., Pham, J., Grewal, D., Abrams, D., Havasov, E., Leung, S., Bojilova, V., Moore, R. A., Rusk, N., Uhlitz, F., Ceglia, N., Weiner, A. C., Zaikova, E., Douglas, J. M., Zamarin, D., Weigelt, B., Kim, S. H., Da Cruz Paula, A., Reis-Filho, J. S., Martin, S. D., Li, Y., Xu, H., de Algara, T. R., Lee, S. R., Llanos, V. C., Huntsman, D. G., McAlpine, J. N., Shah, S. P., and Aparicio, S. (2022). Single-cell genomic variation induced by mutational processes in cancer. *Nature*, 612(7938):106–115.

Ghahramani, Z. and Beal, M. (2000). Propagation Algorithms for Variational Bayesian Learning. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., and Clevert, D.-A. (2010). FABIA: Factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(4):1303–1347.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93.

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Research*, 13(4):703–716.

Lai, D., Ha, G., and Shah, S. (2012). HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data.

Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T., Ting, J., Grewal, D., Nielsen, C., Leung, S., Bojilova, V., Smith, M., Golovko, O., Poon, S., Eirew, P., Kabeer, F., Ruiz de Algara, T., Lee, S. R., Taghiyar, M. J., Huebner, C., Ngo, J., Chan, T., Vatrt-Watts, S., Walters, P., Abrar, N., Chan, S., Wiens, M., Martin, L., Scott, R. W., Underhill, T. M., Chavez, E., Steidl, C., Da Costa, D., Ma, Y., Coope, R. J. N., Corbett, R., Pleasance, S., Moore, R., Mungall, A. J., Mar, C., Cafferty, F., Gelmon, K., Chia, S., CRUK IMAXT Grand Challenge Team, Marra, M. A., Hansen, C., Shah, S. P., and Aparicio, S. (2019). Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell*, 179(5):1207–1221.e22.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5.1, pages 281–298. University of California Press.

Morelli, L., Giansanti, V., and Cittaro, D. (2021). Nested Stochastic Block Models applied to the analysis of single cell data. *BMC Bioinformatics*, 22(1):576.

Myers, M. A., Zaccaria, S., and Raphael, B. J. (2020). Identifying tumor clones in sparse single-cell mutation data. *Bioinformatics*, 36(Supplement_1):i186–i193.

Ng, A., Jordan, M., and Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Roth, A., McPherson, A., Laks, E., Biele, J., Yap, D., Wan, A., Smith, M. A., Nielsen, C. B., McAlpine, J. N., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2016). Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature Methods*, 13(7):573–576.

Schneider, M. P., Cullen, A. E., Pangonyte, J., Skelton, J., Major, H., Van Oudenhove, E., Garcia, M. J., Chaves Urbano, B., Piskorz, A. M., Brenton, J. D., Macintyre, G., and Markowetz, F. (2024). scAbsolute: Measuring single-cell ploidy and replication status. *Genome Biology*, 25(1):62.

Shah, S. P., Xuan, X., DeLeeuw, R. J., Khojasteh, M., Lam, W. L., Ng, R., and Murphy, K. P. (2006). Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, 22(14):e431–e439.

Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L., and Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915.

Weiner, S. and Bansal, M. S. (2023). CNAsim: Improved simulation of single-cell copy number profiles and DNA-seq data from tumors. *Bioinformatics*, 39(7).

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

# A Supplementary Material: CN-SBM

## A.1 Joint Distribution and ELBO for the CN-SBM

The joint distribution of the CN-SBM expands as

$$
\begin{aligned}
p(\mathbf{C}, \mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\}) &= \underbrace{p(\mathbf{C}|\mathbf{g}, \mathbf{h}, \{\boldsymbol{\pi}^{(k,l)}\})}_{\text{model}} \cdot \underbrace{p(\mathbf{g}|\boldsymbol{\pi}^g)}_{\text{cells}} \cdot \underbrace{p(\mathbf{h}|\boldsymbol{\pi}^h)}_{\text{bins}} \cdot \underbrace{p(\boldsymbol{\pi}^g)}_{\text{prior}} \cdot \underbrace{p(\boldsymbol{\pi}^h)}_{\text{prior}} \cdot \underbrace{p(\{\boldsymbol{\pi}^{(k,l)}\})}_{\text{prior}} \\
&= \underbrace{\prod_{i=1}^{N}\prod_{j=1}^{M} \pi_{c_{ij}}^{(g_i,h_j)}}_{\text{data model}} \cdot \underbrace{\prod_{i=1}^{N} \pi_{g_i}^{g}}_{\text{cell clusters}} \cdot \underbrace{\prod_{j=1}^{M} \pi_{h_j}^{h}}_{\text{bin clusters}} \\
&\quad \underbrace{\frac{1}{B(\alpha)^{K \cdot L}} \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{c \in \mathcal{C}} (\pi_c^{(k,l)})^{\alpha-1}}_{\text{block distribution prior}} \cdot \underbrace{\frac{1}{B(\alpha^g)^K} \prod_{k=1}^{K} (\pi_k^g)^{\alpha^g-1}}_{\text{cell cluster prior}} \cdot \underbrace{\frac{1}{B(\alpha^h)^L} \prod_{l=1}^{L} (\pi_l^h)^{\alpha^h-1}}_{\text{bin cluster prior}}.
\end{aligned}
$$

The ELBO can be computed as

$$
\begin{aligned}
\log p(\mathbf{C}) &\geq \mathbb{E}_q[\log p(\mathbf{C}, \mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})] - \mathbb{E}_q[\log q(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})] \\
&= \mathbb{E}_q[\log p(\mathbf{C}|\mathbf{g}, \mathbf{h}, \{\boldsymbol{\pi}^{(k,l)}\})] - D_{KL}(q(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\}) || p(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})) \\
&= \mathbb{E}_q[\log p(\mathbf{C}|\mathbf{g}, \mathbf{h}, \{\boldsymbol{\pi}^{(k,l)}\})] \\
&\quad - D_{KL}(q(\mathbf{g}, \boldsymbol{\pi}^g) || p(\mathbf{g}, \boldsymbol{\pi}^g)) - D_{KL}(q(\mathbf{h}, \boldsymbol{\pi}^h) || p(\mathbf{h}, \boldsymbol{\pi}^h)) - D_{KL}(q(\{\boldsymbol{\pi}^{(k,l)}\}) || p(\{\boldsymbol{\pi}^{(k,l)}\})),
\end{aligned}
$$

where the individual terms are given by

$$
\begin{aligned}
\mathbb{E}_q\left[\log p(\mathbf{C}|\mathbf{g}, \mathbf{h}, \{\boldsymbol{\pi}^{(k,l)}\})\right] &= \sum_{i,j} \mathbb{E}_q\left[\log \prod_{k,l} (\pi_{c_{ij}}^{(k,l)})^{\mathbb{1}(g_i=k)\mathbb{1}(h_j=l)}\right] \\
&= \sum_{i,j}\sum_{k,l} \phi_{ik}^g \phi_{jl}^h \cdot \mathbb{E}_q\left[\log(\pi_{c_{ij}}^{(k,l)})\right] \\
&= \sum_{i,j}\sum_{k,l} \phi_{ik}^g \phi_{jl}^h \left(\psi(\gamma_{c_{ij}}^{(k,l)}) - \psi\left(\sum_{c \in \mathcal{C}} \gamma_c^{(k,l)}\right)\right) \\
D_{KL}(q(\mathbf{g}, \boldsymbol{\pi}^g) || p(\mathbf{g}, \boldsymbol{\pi}^g)) &= \sum_i D_{KL}(q(g_i, \boldsymbol{\pi}^g) || p(g_i, \boldsymbol{\pi}^g)) \\
&= D_{KL}(q(\boldsymbol{\pi}^g) || p(\boldsymbol{\pi}^g)) + \sum_i \mathbb{E}_q\left[\log \frac{q(g_i)}{p(g_i|\boldsymbol{\pi}^g)}\right] \\
D_{KL}(q(\mathbf{h}, \boldsymbol{\pi}^h) || p(\mathbf{h}, \boldsymbol{\pi}^h)) &= \sum_j D_{KL}(q(h_j, \boldsymbol{\pi}^h) || p(h_j, \boldsymbol{\pi}^h)) \\
&= D_{KL}(q(\boldsymbol{\pi}^h) || p(\boldsymbol{\pi}^h)) + \sum_j \mathbb{E}_q\left[\log \frac{q(h_j)}{p(h_j|\boldsymbol{\pi}^h)}\right] \\
D_{KL}(q(\{\boldsymbol{\pi}^{(k,l)}\}) || p(\{\boldsymbol{\pi}^{(k,l)}\})) &= \sum_{k,l} D_{KL}(q(\boldsymbol{\pi}^{(k,l)}) || p(\boldsymbol{\pi}^{(k,l)})).
\end{aligned}
$$

Here, the expected log-likelihood can be expanded as

$$
\begin{aligned}
\mathbb{E}_q[\log p(g_i|\boldsymbol{\pi}^g)] = \mathbb{E}_q[\log \pi_{g_i}^g] = \mathbb{E}_q\left[\log \prod_k (\pi_k^g)^{\mathbb{1}(g_i=k)}\right] = \sum_k \phi_{ik}^g \mathbb{E}_q[\log \pi_k^g] \\
\mathbb{E}_q[\log p(h_j|\boldsymbol{\pi}^h)] = \mathbb{E}_q[\log \pi_{h_j}^h] = \mathbb{E}_q\left[\log \prod_l (\pi_j^h)^{\mathbb{1}(h_j=l)}\right] = \sum_l \phi_{jl}^h \mathbb{E}_q[\log \pi_l^h]
\end{aligned}
$$

and $\mathbb{E}_q[\log q(g_i)] = -\sum_k \phi_{ik} \log \phi_{ik}$ is the entropy of a multinomial distribution. The KL-divergence of two Dirichlet distributions is based on the expectation of a log-transformed Dirichlet variable, i.e.

$$\begin{aligned} D_{KL}(q(\boldsymbol{\pi})||p(\boldsymbol{\pi})) &= \log \frac{\Gamma(\sum_k \alpha_k^q)}{\Gamma(\sum_k \alpha_k^p)} + \sum_k \log \frac{\Gamma(\alpha_k^p)}{\Gamma(\alpha_k^q)} + \sum_k (\alpha_k^q - \alpha_k^p)\mathbb{E}_q[\log \pi_k] \\ &= \log \frac{\Gamma(\sum_k \alpha_k^q)}{\Gamma(\sum_k \alpha_k^p)} + \sum_k \log \frac{\Gamma(\alpha_k^p)}{\Gamma(\alpha_k^q)} + \sum_k (\alpha_k^q - \alpha_k^p) \left[ \psi(\alpha_k^q) - \psi\left(\sum_j \alpha_j^q\right) \right], \end{aligned}$$

where $p(\boldsymbol{\pi}) \sim \text{Dir}_{[0,\dots,K]}(\boldsymbol{\alpha}^p)$ and $q(\boldsymbol{\pi}) \sim \text{Dir}_{[0,\dots,K]}(\boldsymbol{\alpha}^q)$.

### A.2 Stochastic Variational Inference

Stochastic Variational Inference (SVI) (Hoffman et al., 2013) extends the CAVI framework to enable scalable inference in large datasets. The core idea is that in latent variable models, each observation is typically associated with local latent variables ($g_i$, $h_j$), which influence the posterior over global variables ($\boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\}$). While standard CAVI requires full sweeps over all local variables before updating global parameters, SVI performs stochastic optimization by sampling a mini-batch of data points instead and using these intermediate values to update the global parameters, reducing the per-iteration cost. Since the underlying variational updates remain structurally identical to the ones in Algorithm 1, adapting our model to SVI is straightforward. The resulting SVI procedure is outlined in Algorithm 2.

**Algorithm 2** Stochastic variational inference for the CN-SBM (Stochastic VI)

**Require:** Copy number matrix $\mathbf{C} = (c_{ij})$, learning schedule $(\eta_t)_{t\geq 1}$.
Set $t = 1$. Initialize local parameters $\boldsymbol{\phi}^g, \boldsymbol{\phi}^h$ and global parameters $\boldsymbol{\gamma}^g, \boldsymbol{\gamma}^h, \{\boldsymbol{\gamma}^{(k,l)}\}$.
**while** ELBO not converged **do**
**Local updates:** Sample subset $\mathcal{I}_N \subseteq [N], \mathcal{I}_M \subseteq [M]$
- for row $i \in \mathcal{I}_N$ and row clusters $k = 1, \dots, K$
  - $q(g_i = k) \propto \exp\left(\sum_{j=1}^M \sum_{l=1}^L \phi_{jl}^h \cdot \left(\psi(\gamma_{c_{ij}}^{(k,l)}) - \psi\left(\sum_{c\in\mathcal{C}} \gamma_c^{(k,l)}\right)\right) + \psi(\gamma_k^g) - \psi\left(\sum_{k=1}^K \gamma_k^g\right)\right)$
- for col $j \in \mathcal{I}_M$ and col clusters $l = 1, \dots, L$
  - $q(h_j = l) \propto \exp\left(\sum_{i=1}^N \sum_{k=1}^K \phi_{ik}^g \cdot \left(\psi(\gamma_{c_{ij}}^{(k,l)}) - \psi\left(\sum_{c\in\mathcal{C}} \gamma_c^{(k,l)}\right)\right) + \psi(\gamma_l^h) - \psi\left(\sum_{l=1}^L \gamma_l^h\right)\right)$

**Intermediate global updates for cluster and cluster proportions**
- $\hat{\boldsymbol{\gamma}}^g = \alpha^g + \frac{N}{|\mathcal{I}_N|} \sum_{i\in\mathcal{I}_N} \phi_{i,\cdot}^g$
- $\hat{\boldsymbol{\gamma}}^h = \alpha^h + \frac{M}{|\mathcal{I}_M|} \sum_{j\in\mathcal{I}_M} \phi_{j,\cdot}^h$
- $\hat{\boldsymbol{\gamma}}^{(k,l)} = \alpha + \frac{NM}{|\mathcal{I}_N|\cdot|\mathcal{I}_M|} \sum_{i\in\mathcal{I}_N} \sum_{j\in\mathcal{I}_M} \phi_{ik}^g \phi_{jl}^h \mathbb{1}(c_{ij} = \cdot), \qquad 1 \leq k \leq K, 1 \leq l \leq L.$

**Update global estimates**
- $\boldsymbol{\gamma}^g \leftarrow (1-\eta_t)\boldsymbol{\gamma}^g + \eta_t \hat{\boldsymbol{\gamma}}^g$
- $\boldsymbol{\gamma}^h \leftarrow (1-\eta_t)\boldsymbol{\gamma}^h + \eta_t \hat{\boldsymbol{\gamma}}^h$
- $\boldsymbol{\gamma}^{(k,l)} \leftarrow (1-\eta_t)\boldsymbol{\gamma}^{(k,l)} + \eta_t \hat{\boldsymbol{\gamma}}^{(k,l)}$

$t \leftarrow t + 1$
**end while**

### A.3 Importance Weighting for the CN-SBM with Missing Data

In section 2, we performed model inference by maximizing the ELBO for the observed data log-likelihood, $\log p_\theta(\mathbf{C}_{\text{obs}})$, effectively ignoring the missing data. However, a more principled approach may be to optimize the ELBO for the full data-generating process, i.e., bounding $\log p(\mathbf{C}_{\text{obs}}, \mathbf{C}_{\text{mis}}) = \log p(\mathbf{C})$ instead. This ensures that the latent variables recover structure across the entire dataset, rather than just the observed portion. Letting $\mathbf{B} = (b_{ij})$ denote the binary missingness matrix as before the ELBO for the full data is given by

$$\log p(\mathbf{C}) \geq \mathbb{E}_q[\log p(\mathbf{C}|\mathbf{g}, \mathbf{h}, \{\boldsymbol{\pi}^{(k,l)}\})] - D_{KL}(q(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})||p(\mathbf{g}, \mathbf{h}, \boldsymbol{\pi}^g, \boldsymbol{\pi}^h, \{\boldsymbol{\pi}^{(k,l)}\})).$$

In the presence of missing data, we can approximate the likelihood term using inverse propensity weights (IPW):

$$\begin{aligned}
\mathbb{E}_q[\log p(\mathbf{C}|\mathbf{g},\mathbf{h},\{\boldsymbol{\pi}^{(k,l)}\})] &= \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbb{E}_{(\mathbf{z}\sim q)}[\log p(c_{ij}|\mathbf{g},\mathbf{h},\{\boldsymbol{\pi}^{(k,l)}\})] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbb{E}\left[\frac{B_{ij}}{\zeta_{ij}}\mathbb{E}_{(\mathbf{z}\sim q)}[\log p(c_{ij}|\mathbf{g},\mathbf{h},\{\boldsymbol{\pi}^{(k,l)}\})]\right] \\
&\approx \sum_{(i,j)\in\mathcal{E}^{\text{obs}}}\frac{1}{\zeta_{ij}}\mathbb{E}_{(\mathbf{z}\sim q)}[\log p(c_{ij}|\mathbf{g},\mathbf{h},\{\boldsymbol{\pi}^{(k,l)}\})] \\
&\approx \sum_{(i,j)\in\mathcal{E}^{\text{obs}}}\frac{1}{\hat{\zeta}_{ij}}\mathbb{E}_{(\mathbf{z}\sim q)}[\log p(c_{ij}|\mathbf{g},\mathbf{h},\{\boldsymbol{\pi}^{(k,l)}\})]
\end{aligned}$$

Here $\mathcal{E}^{\text{obs}}$ is the set of observed indices and $\zeta_{ij}$ denotes the probability of observing the copy number $c_{ij}$, which may depend on observed data or additional covariates. This term can be computed using only the observed data, and therefore does not require inference over the missing entries. We make two approximations: we approximate the expectation using a single observed sample by setting $b_{ij} = 1$ and estimate the observation propensity as $\hat{\zeta}_{ij}$. The variance of this estimator can become large when individual propensity scores are low. To mitigate this, it may be beneficial to exclude columns with a high proportion of missing data. We plan to perform a sensitivity analysis to understand the impact of missing data in the various missingness scenarios.

This approach relies on specifying an appropriate model for the observation mechanism and how the missing data is generated. In the Missing at Random (MAR) setting, the observation propensity can initially be modeled using logistic regression, depending only on the row and column indices. Alternatively, a simple heuristic based on observation frequencies can be used: the propensity could be approximated as $\zeta_{ij} = \frac{|\mathcal{I}_{i,\text{obs}}|}{M}\frac{|\mathcal{J}_{j,\text{obs}}|}{N}$, where $|\mathcal{I}_{i,\text{obs}}|$ and $|\mathcal{J}_{j,\text{obs}}|$ denote the number of observed entries in row i and column j, respectively. This approach assumes that missingness is independent across rows and columns and provides a simple, data-driven estimate of the observation probability.

**Weighted CAVI Updates** When using importance weights for the CN-SBM, we can also apply approximate update steps similar to the CAVI update formulations in Algorithm 1:

$$\begin{aligned}
q(g_i = k|\boldsymbol{\phi}^g) &\propto \exp\left(\sum_{j=1}^{M}\sum_{l=1}^{L}\mathbb{E}\left[\log\left(\pi_{c_{ij}}^{(k,l)}\right)^{\mathbb{1}(h_j=l)}\right] + \mathbb{E}[\log\pi_k^g]\right) \\
&\approx \exp\left(\sum_{j:(i,j)\in\mathcal{E}^{\text{obs}}}\sum_{l=1}^{L}\frac{1}{\zeta_{ij}}\mathbb{E}[\mathbb{1}(h_j=l)]\mathbb{E}[\log\pi_{c_{ij}}^{(k,l)}] + \mathbb{E}[\log\pi_k^g]\right) \\
q(h_j = l|\boldsymbol{\phi}^h) &\propto \exp\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[\log\left(\pi_{c_{ij}}^{(k,l)}\right)^{\mathbb{1}(g_i=k)]}\right] + \mathbb{E}[\log\pi_l^h]\right) \\
&\approx \exp\left(\sum_{i:(i,j)\in\mathcal{E}^{\text{obs}}}\sum_{k=1}^{K}\frac{1}{\zeta_{ij}}\mathbb{E}[\mathbb{1}(g_i=k)]\mathbb{E}[\log\pi_{c_{ij}}^{(k,l)}] + \mathbb{E}[\log\pi_l^h]\right) \\
\boldsymbol{\gamma}^{(k,l)} &= \alpha + \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbb{E}[\mathbb{1}(g_i=k)\mathbb{1}(h_j=l)]\mathbb{1}(c_{ij}=\cdot) \\
&\approx \alpha + \sum_{(i,j)\in\mathcal{E}^{\text{obs}}}\frac{1}{\zeta_{ij}}\mathbb{E}[\mathbb{1}(g_i=k)\mathbb{1}(h_j=l)]\mathbb{1}(c_{ij}=\cdot), \qquad 1\le k\le K, 1\le l\le L.
\end{aligned} \tag{A.1}$$

The other two update steps for $\boldsymbol{\gamma}^g$ and $\boldsymbol{\gamma}^h$ remain exact as they do not depend on the observed data. However, because the importance-weighted updates are approximations, we no longer have a guarantee that each iteration will increase the ELBO. Therefore, we consider the algorithm to have converged when a moving average of the ELBO shows no significant change over time.

# B Supplementary Material: Benchmarks and Results

## B.1 Implementation and Data Sources

**Implementation Details** The CN-SBM model was implemented in `Python 3.9` using `JAX`, which enables efficient linear algebra via just-in-time (JIT) compilation. Training for the CN-SBM was performed on a virtual machine equipped with an NVIDIA RTX 3080 GPU (10 GB memory, using a single GPU core). The CAVI algorithm terminates when the ELBO improvement falls below a threshold of $10^{-4}$. For datasets at 500 kb resolution with 2,500–5,000 cells, convergence typically occurs within 20–30 minutes, with most of the time spent on JIT compilation and initialization. For higher resolutions (e.g. 100kb or 50kb) or larger cell counts, we recommend stochastic variational inference to avoid memory overruns.

Runtimes for the TCGA BRCA dataset ($N = 998$) at varying resolutions are provided in Fig. 5, where we also compared the runtimes for the PoissonSBM and the Blockcluster model on the same machine and demonstrate the scalability of the algorithm to finer resolution data using stochastic variational inference.

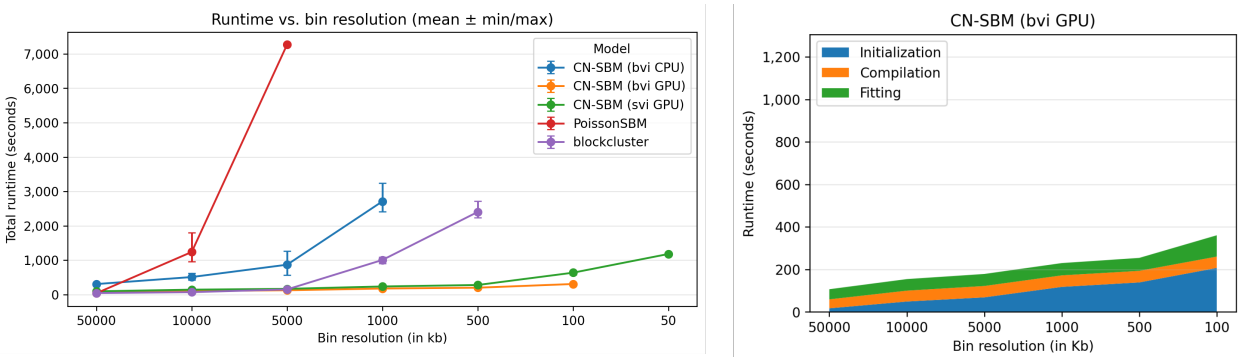


Figure 5: **Left:** Runtimes for the CN-SBM, PoissonSBM, and Blockcluster on the same machine. `bvi` refers to Algorithm 1 and `svi` to Algorithm 2. **Right:** Runtime decomposition of the CN-SBM into spectral bi-clustering initialization, JAX model compilation, and CAVI model fitting.

We use the ICL as a measure of model fit on the complete dataset. We approximate it as

$$\mathrm{ICL}(\mathbf{C}; \hat{\mathbf{g}}_{\mathrm{MAP}}, \hat{\mathbf{h}}_{\mathrm{MAP}}) \approx \log p(\mathbf{C}, \hat{\mathbf{g}}_{\mathrm{MAP}}, \hat{\mathbf{h}}_{\mathrm{MAP}} | \hat{\boldsymbol{\pi}}^g, \hat{\boldsymbol{\pi}}^h, \{\hat{\boldsymbol{\pi}}^{(k,l)}\}) - \mathrm{pen}_{\mathrm{ICL}}(\mathbf{C}, \hat{\mathbf{g}}_{\mathrm{MAP}}, \hat{\mathbf{h}}_{\mathrm{MAP}})$$

where the penalty term is defined as

$$\mathrm{pen}_{\mathrm{ICL}}(\mathbf{C}, \hat{\mathbf{g}}_{\mathrm{MAP}}, \hat{\mathbf{h}}_{\mathrm{MAP}}) = -\frac{1}{2}\left[\underbrace{(K-1)\log(N)}_{\text{Penalty on } \hat{\mathbf{g}}_{\mathrm{MAP}}} + \underbrace{(L-1)\log(M)}_{\text{Penalty on } \hat{\mathbf{h}}_{\mathrm{MAP}}} + \underbrace{(n_{\mathrm{cat}}-1)\cdot K \cdot L}_{\text{Block Complexity}} \cdot \underbrace{\log|\mathcal{E}^{\mathrm{obs}}|}_{\text{Data Size}}\right],$$

Here, $|\mathcal{E}^{\mathrm{obs}}|$ is the number of observed data points in the copy number matrix, and $\hat{\boldsymbol{\pi}}^g, \hat{\boldsymbol{\pi}}^h, \{\hat{\boldsymbol{\pi}}^{(k,l)}\}$ are maximum likelihood estimates. ICL offers a model-based assessment that balances goodness-of-fit with model complexity, by combining a likelihood term involving the maximum a posteriori (MAP) estimate with a penalty accounting for the number of effective clusters. This enables model comparison based on the final clustering configuration inferred from the MAP estimate.

**Data Sources** The CNAsim package is available at https://github.com/samsonweiner/CNAsim/tree/main. We used the command `cnasim -m 0 -n 2500 -c 7 -v -U -B 500000 -E1 0.04 -E2 0.1 -P 24` to generate 2,500 single-cell copy number profiles. TCGA data with ASCAT-inferred copy numbers were preprocessed in the repository at https://github.com/cwlgadd/TCGA using the `TCGA.data_modules.ascat.loaders.LoadASCAT()` function. Patient survival data can be extracted using the `RTCGA` package, see https://rtcga.github.io/RTCGA/reference/survivalTCGA.html for additional information. Single-cell data from Funnell et al. (2022) can be obtained online at Zenodo: https://zenodo.org/records/6998936.

## B.2 Additional Benchmark Details

We initially included spectral co-clustering (SpecCo) as a baseline method, however, it consistently performed worse than all other models, which is why we excluded this method from the model comparisons. Table 2 presents supplementary

performance metrics: classification accuracy on held-out copy numbers and the average normalized entropy of clusters for models trained on the complete dataset. We considered accuracy as a secondary metric as it involves two approximations: converting soft clusters to hard assignments and selecting a MAP category from the categorical block distributions. Models are not explicitly optimized for low entropy, so this serves as a sanity check on cluster quality.

| | Method | CNAsim | TCGA | | | Funnell | | |
|---|---|---|---|---|---|---|---|---|
| | | Simulated ($N = 2500$) | BRCA ($N = 998$) | OV ($N = 532$) | LGG ($N = 490$) | OV2295 ($N = 1084$) | SA1096 ($N = 802$) | SA535 ($N = 1801$) |
| **Acc.** $\times 10^2$ (↑) | SpecBi (bist) | 82.26 ± 0.05 | 52.33 ± 0.03 | 42.71 ± 0.42 | 79.60 ± 0.56 | 70.05 ± 0.94 | 60.89 ± 0.67 | 62.97 ± 1.14 |
| | SpecBi (log) | 82.48 ± 0.06 | 52.77 ± 0.19 | 43.85 ± 0.24 | 80.38 ± 0.32 | 69.22 ± 1.15 | 61.83 ± 0.77 | 66.75 ± 0.87 |
| | KMeans | 93.36 ± 0.20 | 62.19 ± 0.24 | 51.05 ± 0.15 | **87.77 ± 0.16** | **81.52 ± 0.39** | **81.32 ± 0.45** | **88.41 ± 0.43** |
| | PoissonSBM | 89.57 ± 3.77 | 14.63 ± 0.41 | 13.31 ± 0.79 | 21.32 ± 0.63 | 60.91 ± 1.51 | 73.81 ± 0.70 | 88.19 ± 0.76 |
| | Blockcluster | 93.29 ± 0.22 | 62.66 ± 0.24 | 52.35 ± 0.52 | 86.38 ± 0.74 | 79.69 ± 0.60 | 80.36 ± 0.40 | 85.43 ± 1.08 |
| | CN-SBM (ref.) | **93.81 ± 0.05** | **63.12 ± 0.09** | **53.01 ± 0.45** | 87.42 ± 0.27 | 80.99 ± 0.89 | 81.23 ± 0.62 | 87.43 ± 0.31 |
| **Entr.** $\times 10^{-2}$ (↓) | SpecBi (bist) | 24.6 ± 0.0 | 51.4 ± 0.2 | 58.5 ± 0.3 | 26.2 ± 0.3 | 33.9 ± 0.4 | 40.7 ± 0.8 | 41.2 ± 2.3 |
| | SpecBi (log) | 24.5 ± 0.2 | 51.4 ± 0.3 | 57.1 ± 0.2 | 26.3 ± 0.3 | 33.7 ± 0.8 | 40.3 ± 0.5 | 36.8 ± 1.5 |
| | KMeans | 11.1 ± 0.3 | 41.1 ± 0.2 | 48.6 ± 0.1 | 15.2 ± 0.2 | 22.3 ± 0.2 | 22.3 ± 0.2 | 16.0 ± 0.2 |
| | PoissonSBM | 10.5 ± 0.0 | 38.7 ± 0.0 | **47.3 ± 0.0** | **14.7 ± 0.0** | 22.4 ± 0.0 | **21.3 ± 0.0** | 15.7 ± 0.0 |
| | Blockcluster | 10.4 ± 0.2 | 39.0 ± 0.1 | 47.6 ± 0.1 | 16.3 ± 0.1 | 24.7 ± 0.7 | – | 17.0 ± 0.1 |
| | CN-SBM (ref.) | **9.9 ± 0.0** | **38.5 ± 0.1** | 47.4 ± 0.1 | 15.3 ± 0.1 | **21.3 ± 0.6** | 21.7 ± 0.2 | **15.3 ± 0.3** |

Table 2: Accuracy (top) on held-out data, normalized entropy (bottom) (mean ± std) across simulated and real datasets. Values were scaled by $10^{-2}$ for accuracy and entropy. Dashes for the entropy indicate missing values due to failed runs

Table 3 reports performance metrics (held-out log-likelihood, accuracy) based on the block probability distributions inferred from the PoissonSBM and Blockcluster implementations, rather than empirical block distributions. Since PoissonSBM uses a Poisson distribution, likelihood values are expected to be much lower as it cannot capture dominant copy number categories, e.g. $\text{Poisson}(x = 2; \lambda = 2) \approx 0.27$. Similarly to the results in Table 1, PoissonSBM also shows low ICL variance, as it is trained on the full dataset using exhaustive search for cluster partitions.

For the Blockcluster implementation (Bhatia et al., 2017), we initially considered using the semi-supervised implementation of Blockcluster (`R-blockcluster` package) by using spectral clustering informed inputs, however, the model runs failed multiple times. As the method already employs a smart initialization, we used the default settings. Due to issues computing log-likelihoods from its probability outputs, which led to extreme values, we relied on empirical block distributions in the main results and report package-specific metrics in Table 3.

| | Method | CNAsim | TCGA | | | Funnell | | |
|---|---|---|---|---|---|---|---|---|
| | | Simulated ($N = 2500$) | BRCA ($N = 998$) | OV ($N = 532$) | LGG ($N = 490$) | OV2295 ($N = 1084$) | SA1096 ($N = 802$) | SA535 ($N = 1801$) |
| **LL** | PoissonSBM | -198.4 ± 0.1 | -209.7 ± 0.6 | -90.0 ± 0.2 | -45.3 ± 0.3 | -134.6 ± 0.8 | -163.5 ± 1.6 | -186.6 ± 0.1 |
| | Blockcluster | -38.2 ± 0.4 | -101.5 ± 53.5 | -66.1 ± 41.3 | -118.2 ± 92.3 | -193.0 ± 308.3 | -161.6 ± 231.9 | -181.0 ± 231.3 |
| **ICL** | PoissonSBM | -18.88 ± 0.00 | -9.06 ± 0.00 | -4.93 ± 0.00 | -4.05 ± 0.00 | -9.93 ± 0.00 | -7.46 ± 0.00 | -16.11 ± 0.00 |
| | Blockcluster | -21.98 ± 21.04 | -14.14 ± 7.91 | -3.74 ± 0.01 | -1.21 ± 0.01 | -25.95 ± 37.83 | – | -68.42 ± 55.37 |
| **Acc.** | PoissonSBM | 93.62 ± 0.09 | 60.6 ± 0.21 | 50.53 ± 0.05 | 89.51 ± 0.28 | 80.39 ± 0.33 | 81.65 ± 0.22 | 88.59 ± 0.24 |
| | Blockcluster | 93.29 ± 0.22 | 46.4 ± 20.11 | 37.29 ± 21.75 | 48.75 ± 37.57 | 65.81 ± 28.27 | 58.87 ± 37.46 | 67.16 ± 31.6 |

Table 3: Held-out log-likelihood (top), ICL (middle), and accuracy (bottom) across simulated and real datasets using the probability distributions provided by each model implementation. Values were scaled by $10^3$ for log-likelihood, $10^6$ for ICL, and $10^{-2}$ for the accuracy. Dashes indicate missing values due to failed runs.

Table 4 reports the number of successful model runs across five random seeds for each experiment. PoissonSBM completed only 3 runs on the TCGA-OV dataset due to time constraints, as it explores a large number of cluster combinations. Blockcluster failed on multiple runs, likely due to convergence issues in its package implementation.

| Method | Missing | CNAsim | TCGA | | | Funnell | | |
|---|---|---|---|---|---|---|---|---|
| | | Simulated ($N = 2500$) | BRCA ($N = 998$) | OV ($N = 532$) | LGG ($N = 490$) | OV2295 ($N = 1084$) | SA1096 ($N = 802$) | SA535 ($N = 1801$) |
| PoissonSBM | False | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| | True | 5 | 5 | 3 | 5 | 5 | 5 | 5 |
| Blockcluster | False | 4 | 4 | 4 | 2 | 3 | 0 | 3 |
| | True | 3 | 4 | 2 | 3 | 4 | 3 | 3 |

Table 4: Number of successful model runs across 5 seeds for PoissonSBM and Blockcluster.

### B.3 TCGA-LGG Survival Modelling

**Model Selection Details** We observed that age, when treated as a continuous covariate, was generally not statistically significant, as indicated by its associated $p$-values. In contrast, the discretization of age into age groups more effectively captured the risk patterns present in the survival data, leading to the baseline model to include age groups.

We began by analyzing a Cox model incorporating the main variation clusters (Fig. 3A). All clusters 0-5 except 1 were found to be significant, where the $p$-value was $0.16$ for the latter. Combining cluster 1 with any of the remaining clusters would yield an overall significant model, however, merging clusters 1 and 2 was the most appropriate: these two clusters displayed highly similar genomic profiles, only differing in the copy numbers in the `X` chromosome, likely attributable to patient sex. To validate this, we refitted the CN-SBM model on the TCGA-LGG dataset excluding the `X` chromosome. In this output, samples from clusters 1 and 2 were grouped into a single, larger cluster, supporting the decision to merge them. As a result, we arrived at the model involving age groups and the main variation clusters, where clusters 1 and 2 were merged.

To assess the inclusion of residual clusters, we extended the baseline model to incorporate them (without main clusters). Residual clusters 2, 3, 4, 6, 8, and 9 were found to be significant, with clusters 3 and 8 exhibiting the largest hazard ratios, namely 6.2 (95% CI: 2.7-14.0) and 5.9 (95% CI: 2.0-17.6), respectively. We then explored individually stratifying the main clusters based on their residual cluster groupings, guided by the Sankey diagram (Fig. 3C). This analysis led to the final model shown in Fig. 3D, where we found that residual cluster 3 modified the effect of the merged main clusters 1 & 2, and residual cluster 8 similarly influenced main cluster 5. In this model, we also included the interaction between main cluster 3 and residual cluster 3 (where there was no significant effect) to illustrate how the impact of residual variation can depend on the main cluster.

This analysis highlights that residual variation does not necessarily translate to significant survival effects and may interact with other unmodeled covariates. Our model selection process was designed to evaluate the contributions of both main and residual variation clusters to patient survival. We acknowledge that more comprehensive models that include additional covariates beyond age groups and copy number variation may further improve predictive and explanatory performance.