CopyMix: Mixture Model Based Single-Cell Clustering and Copy Number Profiling using Variational Inference

Negar Safinianaini[1*], Camila P. E. de Souza[2], Jens Lagergren[1,3]

[1]School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON, Canada
[3]Science for Life Laboratory, Solna, Sweden

*Correspondence: negars@kth.se

Abstract

Single-cell sequencing technologies are becoming increasingly more established, in particular, in the study of tumor heterogeneity. Investigating tumor heterogeneity is imperative to better understand how tumors evolve since each cell subpopulation harbors a unique set of genomic features that yields a unique phenotype, an issue that is bound to have clinical relevance. Clustering of cells based on copy number data, obtained from single-cell DNA sequencing, provides an opportunity to assess different tumor cell subpopulations. Accordingly, computational methods have emerged for single-cell copy number profiling as well as clustering; however, these two tasks have up to now been handled sequentially, and, included various ad-hoc preprocessing steps; hence, prone to introduce clustering artifacts. We propose CopyMix, a mixture model-based method using Variational Inference, to simultaneously cluster cells and infer a single copy number profile for each cluster, revealing the copy number variation pattern behind each cluster. CopyMix is evaluated using simulated data as well as published biological data from metastatic colorectal cancer. The results reveal high clustering performance and low errors in copy number profiling. These favorable results indicate a considerable potential to obtain clinical impact by using CopyMix in studies of cancer tumor heterogeneity.
Keywords: Mixture Models; Variational Inference; Copy number profiling; Single-cell

1 Introduction

A tumor typically consists of a collection of heterogeneous cell populations, each having distinct genetic and phenotypic properties, in particular, concerning the capacity to promote cancer progression, metastasis, and therapy resistance Eirew et al. (2015); Nowell (1976). Single-cell sequencing technologies Gawad et al. (2016); Navin et al. (2011); Shapiro et al. (2013); Zahn et al. (2017) provide a new opportunity to investigate the genomic profile of individual cells both regarding single nucleotide variation (SNV) and copy number variation (CNV). CNVs and SNVs have been shown to be important contributors to phenotypic variation relating to health and disease Baslan et al. (2012); Lawson et al. (2018). Moreover, although single-cell SNV profiling is hampered by experimental imperfections such as drop-outs, copy number profiling (detecting single-cell copy number variations) is feasible at least at coarser resolutions.

Clustering of cells based on their individual copy number profiles provides the opportunity for SNV profiling of the clusters and, in general, opportunities to better understand tumor subpopulations and tumor heterogeneity, issues that are bound to have clinical relevance.

Current single-cell datasets pose a wealth of computational challenges. As answers to some of those, methods have emerged for single-cell copy number profiling as well as clustering of cells based on their already inferred copy number profiles, e.g., Garvin et al. (2015); Zahn et al. (2017); Leung et al. (2017); Vitak et al. (2017); however, these two tasks have up to now been handled sequentially with various ad-hoc preprocessing steps. The sequential approach is vulnerable to clustering artifacts since preprocessing decisions constrain all later analyses and are typically irreversible. Even if each task has maximized its performance, the final result may still fall short of the best possible in the traditional single-phase inference framework Blocker and Meng (2013). As a solution to this issue, We propose a novel framework, referred to as CopyMix, performing the two tasks simultaneously. As to the best of our knowledge, no earlier work has simultaneously performed the two tasks, and, the need is listed in Zhang and Campbell (2020) as a future direction. Similar to CopyMix, simultaneous inference have been considered but for single-cell DNA methylation Kapourani and Sanguinetti (2019); de Souza et al. (2020), single-cell SNV data Roth et al. (2016), and bulk chIP-seq data from several replicates Zuo et al. (2016).

CopyMix is a model-based approach enjoying the advantages of a fully probabilistic framework such as transparency, uncertainty measurements, and modeling flexibility. CopyMix uses a mixture model with components, corresponding to clusters, each having a specific copy number profile modeled by a sequence of latent variables which follows a Markov structure similar to the state-of-the-art methods using Hidden Markov Models (HMMs) Shah et al. (2006); Vitak et al. (2017); Zahn et al. (2017), revealing the copy number variation pattern behind each cluster. CopyMix combines statistical strength from data across all cells and neighboring sites. We deploy a Bayesian treatment and infer all quantities of interest using Variational Inference (VI) Jordan et al. (1999), which typically yields faster than inference methods like Markov chain Monte Carlo sampling Blei et al. (2017). Compared to Expectation-Maximization (EM), VI has multiple advantages, e.g., it estimates the posterior distributions rather than point estimates, protects against over-fitting, and allows for performing principled determination of the optimal number of mixture components Bishop (2006).

We perform experiments on biologically inspired simulated data, the published biological data from Leung et al. (2017), and evaluate CopyMix results w.r.t. the work by Leung et al. (2017). However, it is not equitable to compare CopyMix to the state-of-the-art methods in copy number profiling and single-cell clustering since various ad-hoc processes (e.g., outlier removal, curve smoothing, etc.) are performed within these approaches, which add vital contributions to these sequential approaches. A sequential version of CopyMix, resembling Shah et al. (2006); Vitak et al. (2017) but excluding their ad-hoc processes, is compared to CopyMix, confirming the hypothesis that single-phase inference framework is superior over a sequential version of the framework.

2 CopyMix

CopyMix is a probabilistic clustering method based on a mixture model with components corresponding to clusters, each having a specific copy number profile modeled by a sequence of latent variables. Similarly, as in Shah et al. (2006); Vitak et al. (2017); Zahn et al. (2017), we assume that each latent copy number sequence is governed by a Markov chain (describing a sequence of possible copy numbers in which the

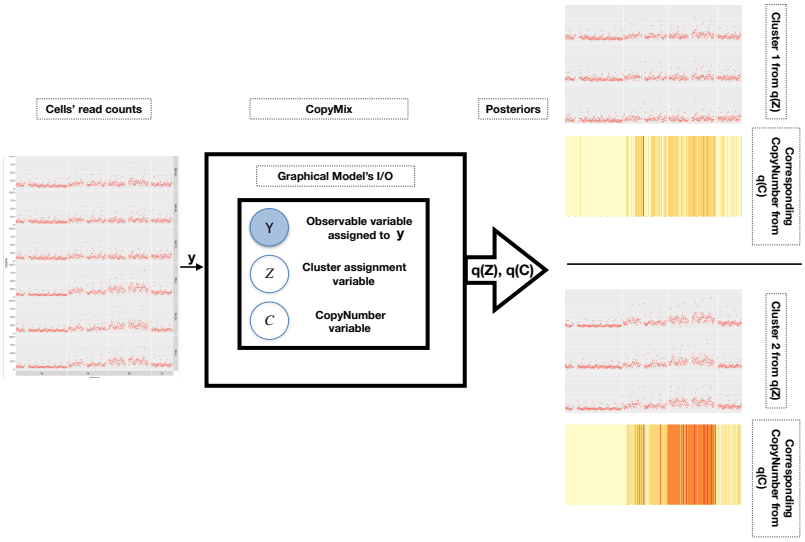

Figure 1: The overview of CopyMix with: input y, cells' reads; outputs $q(Z)$ and $q(C)$, posterior distributions resulting in clusters within y with their corresponding copy number profiles illustrated by heatmaps (this is an example of binary clustering).

probability of each copy number value, defined as a state, depends only on the previous state in the sequence). In CopyMix, we assume that the data (read counts) follow a Poisson distribution as in Witten (2011); we consider read counts per fixed equal size genomic bin as in Leung et al. (2017). Moreover, we assume the read counts are emitted from a latent sequence of copy number states. Fig. 1 illustrates an overview of CopyMix; it analyzes the input cell sequences, containing read counts per predefined genomic bin, and produces clusters of cells with their corresponding copy number sequences (profiles) explaining the clusters. Fig. 2 illustrates CopyMix graphical model where Y denotes the observable variables, read counts per predefined bins, C the latent copy number states forming a Markov chain, and Z the latent cell-specific cluster assignment variables. We notice in Fig. 2 that each Y has two levels of dependencies, which are reflections of the assumption that the Poisson distribution over the read counts depends on a latent cell-specific cluster assignment, Z, and a corresponding latent copy number state, C. Intuitively, a higher copy number should correspond to a higher read count. We incorporate this assumption in our model by defining the rate of Poisson distribution as the product of copy number state and cell-specific rate, θ_n, see Eq. (1); this also implies that θ_n corresponds to the average sequencing coverage for a haploid genome. Due to this multiplicative structure for the Poisson rate, a copy number state of zero results in the lowest rate implying copy number deletion events; however, since the rate of Poisson cannot be zero, our implementation handles this by assigning a tiny number instead of zero. Our model is described in more detail in what follows, using conjugacy regarding the priors.

The genome considered is partitioned into M equally sized segments of consecutive positions called bins. Let $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nM})$, where Y_{nm} is the random but observable number of reads aligned to bin m for cell n taking values in $\{0, 1, 2, \ldots\}$ for $m \in [M] = \{1, \ldots, M\}$ and $n \in [N] = \{1, \ldots, N\}$. We

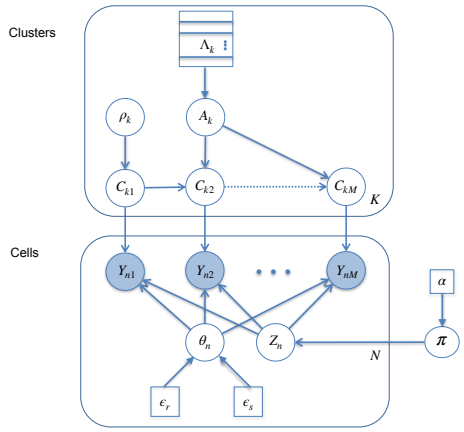

Figure 2: Probabilistic graphical model representing our proposed model. Shaded nodes represent observed values, the unshaded ones are the latent variables and the squares are the hyperparameters of the model; a posterior distribution over the values of the unshaded nodes is approximated using Variational Inference. Y_{nm}, observed read counts from cell n and bin m; C_{km}, corresponding latent copy number state forming a Markov chain; θ_n, cell-specific rate; Z_n, latent cell-specific cluster assignment variable. π and A_k, the priors over Z_n and C_{km} respectively; ρ_k, prior over the starting copy number state.

assume that $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$ are independent and given a vector of true copy number states, $Y_{n1},\ldots,Y_{nM}$ are also independent with the distribution of Y_{nm} depending on the true copy number state at bin m. The cluster membership of cell n is indicated by the hidden variable Z_n that takes values in $[K]=\{1,\ldots,K\}$. We assume there are $K << N$ vectors of true hidden copy number states, i.e., one for each cluster. The variables $Z_1,\ldots,Z_N$ are independent following a categorical distribution with $P(Z_n=k)=\pi_k$ and $\sum_{k=1}^{K}\pi_k=1$. If $Z_n=k$ then the distribution of $\mathbf{Y}_n$ depends on the k-th vector of true hidden copy number states, defined as $\mathbf{C}_k=(C_{k1},\ldots,C_{kM})$, with each C_{km} taking values in $[J]=\{1,\ldots,J\}$. We assume that $C_{k1},\ldots,C_{kM}$ follow a discrete-time homogeneous Markov chain with initial probabilities $\rho_{kj}=P(C_{k1}=j)$, $j\in[J]$ and transition probabilities $a_{ij}^k=P(C_{km}=j|\,C_{km-1}=i)$, $i,j\in[J]$. Consequently, given the cluster assignment and the corresponding true hidden vector of copy number states, $Y_{n1},\ldots,Y_{nM}$ are independent with Y_{nm} following a distribution with parameters depending on the hidden true state at bin m for cluster k, that is, $Y_{nm}|Z_n=k,C_{km}=j\sim F_{j\theta_n}$. As mentioned earlier, we assume $F_{j\theta_n}$ to be a Poisson distribution with rate $j\times\theta_n$, that is,

$$F_{j\theta_n}(y)\equiv\frac{(j\times\theta_n)^y e^{-(j\times\theta_n)}}{y!}\text{ with } j\theta_n>0. \tag{1}$$

We, w.l.o.g., consider the initial probabilities, ρ_{kj}'s, to be fixed and known. We let $\mathbf{\Psi}$ be the set containing all the unknown model parameters, i.e., $\mathbf{\Psi}=\{\mathbf{A},\boldsymbol{\theta},\boldsymbol{\pi}\}$, where $\mathbf{A}=\{\mathbf{A}_k:k\in[K]\}$ with $\mathbf{A}_k=\{a_{ij}^k:i,j\in[J]\}$; $\boldsymbol{\theta}=\{\theta_n:n\in[N]\}$; and $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_K)$. In order to infer $\mathbf{\Psi}$, the hidden states $\mathbf{C}=\{\mathbf{C}_1,\ldots,\mathbf{C}_K\}$, and $\mathbf{Z}=(Z_1,\ldots,Z_n)$ we apply VI; that is, we derive an algorithm that, for given data, approximates the posterior distribution of the parameters by finding the Variational Distribution (VD), $q(\mathbf{Z},\mathbf{C},\mathbf{\Psi})$, with smallest Kullback-Leibler divergence to the posterior distribution $P(\mathbf{Z},\mathbf{C},\mathbf{\Psi}|\mathbf{Y})$, which is equivalent to

maximizing the evidence lower bound (ELBO) Blei et al. (2017) given by

$$\text{ELBO}(q) = \text{E}[\log P(\mathbf{Y}, \mathbf{Z}, C, \mathbf{\Psi})] - \text{E}[\log q(\mathbf{Z}, \mathbf{C}, \mathbf{\Psi})]. \tag{2}$$

We consider the following prior distributions for the parameters in $\mathbf{\Psi}$.

- $\mathbf{a}_i^k = (a_{i1}^k, \ldots, a_{iJ}^k) \sim \text{Dirichlet}(\mathbf{\Lambda}_i^{k0})$.
- $\theta_n \sim \text{Gamma}(\epsilon_{sn}^0, \epsilon_{rn}^0)$, with ϵ_{sn}^0, shape, and ϵ_{rn}^0, rate, hyperparameters.
- $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_1^0, \ldots, \alpha_K^0)$

The main steps of our VI approach, for inferring $\mathbf{Z}, \mathbf{C}$, and $\mathbf{\Psi}$, are described below.

Step 1. VD factorization We assume the following factorization of the VD:

$$q(\mathbf{Z}, \mathbf{C}, \mathbf{\Psi}) = q(\boldsymbol{\pi}) \prod_{n=1}^{N} q(Z_n) \prod_{k=1}^{K} q(\mathbf{C}_k) \prod_{k=1}^{K} \prod_{i=1}^{J} q(\mathbf{a}_i^k) \prod_{n=1}^{N} q(\theta_n) \ . \tag{3}$$

Step 2. Joint distribution

The logarithm of the joint distribution satisfies: $\log P(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \mathbf{\Psi}) = \log P(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \mathbf{\Psi}) + \log P(\mathbf{C}|\mathbf{\Psi}) + \log P(\mathbf{Z}|\mathbf{\Psi}) + \log P(\mathbf{\Psi})$. For details of the calculations see Appendix A.1.

Step 3. VD computation

We now derive a coordinate ascent algorithm for the VD. That is, we derive an update equation for each term in the factorization, Eq. (3), by calculating the expectation of $\log P(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \mathbf{\Psi})$ over the VD of all random variables except the one currently being updated Bishop (2006). For example, we obtain the update equation for $\boldsymbol{\pi}$, $q(\boldsymbol{\pi})$, by calculating $E_{-\boldsymbol{\pi}}(\log P(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \mathbf{\Psi}))$, where $-\boldsymbol{\pi}$ indicates that the expectation is taken with respect to the VD of all other random variables than $\boldsymbol{\pi}$, i.e., $\mathbf{Z}, \mathbf{C}$ and $\{\mathbf{A}, \boldsymbol{\theta}\}$ except $\boldsymbol{\pi}$. See A.2 and A.3 for the update equation of each term in Eq. (3).

Step 4. Summary of updates

We update the parameters of each posterior distribution presented in Step 3; for the summary of the update equations, see Alg. 2, in Appendix A.2. A shorter version of Alg. 2, providing an overview of CopyMix, is shown in Alg. 1. As shown in lines 8 and 9 of Alg. 1, we aim to find the posterior distributions of cluster assignments and copy number states. Note that the update equations in line 5 of Alg. 1 can be run in a parallel fashion, keeping the algorithm's computational complexity low. The implementation of CopyMix is available at **CopyMix**.

Algorithm 1 CopyMix: a Variational Inference Algorithm

procedure ESTIMATE-POSTERIORS(Y): ▷ Y input is N sequences of length M
Assign priors hyperparameters
Initialise variational distributions parameters
    repeat
        Update variational parameters (using **Step 4. Summary of updates**)
    until convergence of the ELBO in Eq. 2
Output all posterior distributions including:
$q(Z_n = k) \quad \forall n, k;$ ▷ cluster assignment probabilities
$q(C_{km} = j) \quad \forall k, m, j;$ ▷ copy number probabilities
return posterior distribution

3 Empirical investigation

3.1 Experimental setup

3.1.1 Dataset

We perform experiments on both simulated data and a published biological dataset Leung et al. (2017) (available from the NCBI Sequence Read Archive (SRA) under accession number SRP074289). The biological data stem from single-cell whole-genome sequencing from 18 primary colon cancer tumor cells and 18 metastatic cells from matched liver samples for one patient referred to as CRC2 in Leung et al. (2017). In this work, we mainly analyze CRC2 patient data (more complex than CRC1 according to Leung et al. (2017)) considering chromosomes 18, 19, 20, and 21 comprising a total of 904 genomic bins of an approximate size of 200Kb each. Each data point corresponds to the number of reads aligned per bin per cell after GC correction Leung et al. (2017). These counts stemming from the tumor cell's genome reflect the hidden copy numbers—copy number integers are conventionally ranged from 0 to 6, as in Leung et al. (2017). The higher the counts are, the higher the copy numbers are expected to be. Regarding clustering in Leung et al. (2017), it is reasonable that the primary tumor cells from the colon and the metastatic cells from the liver should cluster separately, as inter-tumor copy number differences are expected to dominate over minor intra-tumor copy number variations; therefore, we use the terms primary and metastatic clusters.

For the simulated data, we randomly generate 30 datasets for each test case, following our graphical model and inspired by biological datasets from Leung et al. (2017); for data generation details, see README.txt in the **CopyMix**. We examine different test cases, orthogonal to each other, to show our approach's robustness; each scenario, for both clustering and copy number state inference, is reproducible.

3.1.2 Experimental Protocol

We run VI for different numbers of clusters considering various random and informative initialization methods, e.g., k-means and Baum-Welch, as recommended by Blei et al. (2017). For each given number of clusters, we choose the best VI run based on the highest log-likelihood. Next, we select the number of clusters by an elbow selection method Johnson and Wichern (2007) based on the Deviance Information Criterion Spiegelhalter et al. (2002) over different numbers of clusters. We assess the clustering performance of CopyMix on 30 simulated datasets for each test case via V-measure Rosenberg and Hirschberg

(2007). The V-measure, a value between 0 and 1, quantifies to which extent a cluster only contains data points that are members of a single class (the higher V-measure, the better clustering). Regarding inference of copy number profiles, we evaluate the performance of CopyMix on simulated data by calculating the proportion of discordant (unequal) position values between true and inferred vectors, as in de Souza et al. (2020), of copy number states. For biological data evaluation, V-measure is used to assess the clustering performance where the true clusters are assumed to correspond to the primary and metastatic cells; note that these may not be ground truth, but we believe there is certain concordance with the unknown ground truth. Regarding state inference, we compare the copy number profiles between the results from CopyMix and the original paper Leung et al. (2017).

3.2 Experimental Results

We perform experiments on simulated data and biological data, and we evaluate a sequential version of CopyMix to verify the superiority of single-phase inference over the sequential framework.

Simulated data Here, we evaluate CopyMix, and compare it to the initialization methods used by VI on 30 simulated datasets for each test scenario. The comparison is to illustrate, firstly, that VI results are not already achieved by the initialization methods, and secondly, that VI can recover the poor initialization performance and achieve optimal performance. We select the number of cells, sequence lengths, rate parameters, transition probabilities based on Leung et al. (2017). We assume equal amplification on the read counts across bins since the bin size is high enough (200kb as in Leung et al. (2017)), meaning that the effect of unequal amplification is negligible. Figures 3, 4 and 5 provide summaries of V-measures and proportions of discordant positions; note that the median values in the boxplots, highlighted in red for initialisation method and in blue for CopyMix, aim to facilitate the comparisons. The figures illustrate the performance comparison of CopyMix and the chosen initialization method, by CopyMix, for different settings: copy number transition matrices and number of clusters (Fig. 3); sequence length (Fig.4); and number of cells (Fig. 5). For all the tests, CopyMix results in V-measure of 1 except for one case (Fig. 3) where V-measure is in an interval of [.85, 1] with the median being 1; proportions of discordant positions are mostly 0 and otherwise are in the interval of [0, .2]. These results are optimal as the copy number profiling errors are near 0 and clustering near 1 (or 100%).

The results in Fig. 3 are obtained considering 45 cells—each with a sequence length of 800 bins and the mean value around 100—clustered into two and three clusters, wherein each clustering scenario, each cluster bears a different transition pattern than other clusters. The copy number transition patterns (forming the transition matrix of a Markov chain) vary across configurations A to E, which are orthogonal to each other; Configurations A to E are formed by combining the following state patterns.

- **single state**, a copy number state is inclined to stay at one certain state;
- **self-transition state**, a copy number state is inclined to stay at its current state for awhile;
- **oscillating state**, a copy number state changes shortly after being at a state;
- **spike state**, a sudden change of copy number state arises for a specific short range of positions in the copy number sequence;
- **scaled state**, a copy number changes by a multiplicative or additive constant.

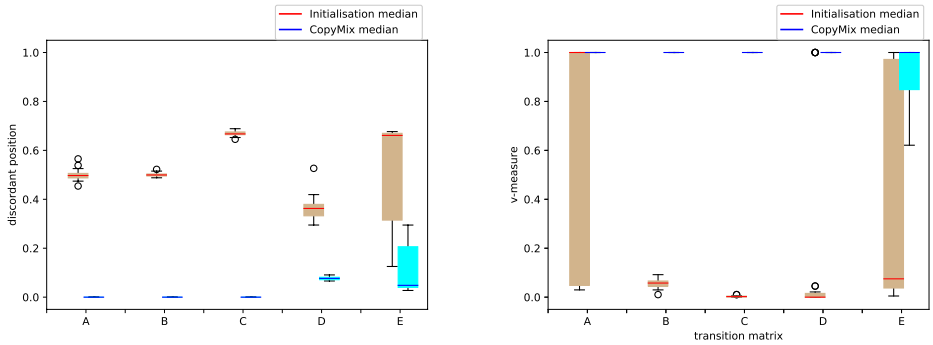

Figure 3: Performance of initialization methods (brown bars with red median values) and CopyMix (cyan bars with blue median values) across different sets of transition matrices. LHS: discordant positions between true and inferred vectors of copy number states. RHS: V-measure between true and inferred cell clustering assignments.

The configurations A to E are: **A**, similar to chromosomes 11 in CRC1 and 4 in CRC2, (one cluster follows (**scaled state**); **B**, similar to chromosome 11 in CRC1 (one cluster follows **self-transition state** and the other one **single state**); **C**, similar to chromosome 3 in CRC1 (two clusters have identical transition matrices but one arrive at **spike state**); **D**, similar to chromosomes 18 to 21 in CRC2 (clusters combine **self-transition state**, **oscillating state**, and **scaled state**); **E**, two of the clusters have the transitions mentioned in case **C**, and, the third cluster has a combination of **oscillating state** and **single state**.

We can observe that, in Fig. 3, CopyMix has a perfect performance in terms of clustering and copy number inference for easier cases and also a high performance concerning the harder cases (**D** and **E**). Next, however not shown here, we investigate decreasing the rate parameter, θ_n, for configuration **A** to see how our model handles shallow data as this can be a challenge in copy number profiling Zahn et al. (2017). Our model is robust towards lower rates up to $\theta_n = 3$, without reducing the quality of performance concerning V-measure and discordant positions. As the third set of tests, we rerun the test using configuration **A**, but we increase the size of the sequences up to 10000 (inspired by the total length of the genomic bins from biological data in Leung et al. (2017)), see Fig. 4. We can observe that our model handles these sequences without any decrease in performance. Finally, Fig. 5 illustrates the perfect performance of CopyMix when varying the number of cells from 15 to 500 when, again, using configuration **A**.

Biological data We evaluate CopyMix on the biological data from Leung et al. (2017). As aforementioned, we do not compare CopyMix to the earlier state-of-the-art methods due to their inclusion of ad-hoc processes in their framework; however, to evaluate the CopyMix results on the biological data, we use biologically verifiable results from the method in Leung et al. (2017). Using chromosomes 18 to 21, CopyMix clusters the cells into primary and metastatic groups, perfectly matching the corresponding primary and metastatic organs—V-measure is 1 w.r.t. the cells found in the two organs Leung et al. (2017). Adding chromosomes 3 and 8 to our test, metastatic tumor cells are clustered into two major subclusters as in Leung et al. (2017). As mentioned in Leung et al. (2017), amplifications on chromosomes 3 and 8, where some known genes are present, contribute to the detection of the metastatic subclusters. Note that, Leung et al. (2017) does not present a method to determine the threshold in hierarchical clustering,

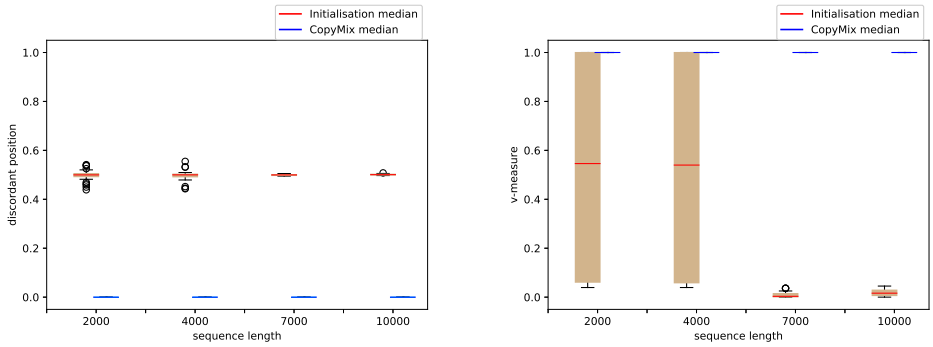

Figure 4: Performance of initialization methods and CopyMix, varying the sequence length. LHS and RHS figures, as well as color codes, are similar to Fig. 3.

which results in the reported subclusters; moreover, there is no ground truth to verify these results.

Regarding the copy number profiling, Fig. 6 illustrates the copy number profiles inferred by CopyMix and the original work by Leung et al. (2017), here referred to as Navin. The heatmaps indicate a higher copy number state by a darker color. Note that in Navin, they perform ploidy scaling, which contributes to scaling differences compared to the results of CopyMix. Moreover, the Navin heatmap, here, illustrates an average of the heatmap in the original work across the cells since there is no single copy number profile per cluster. That is, Navin, unlike CopyMix, lacks a single copy number profile per cluster, revealing the copy number variation pattern behind each cluster. CopyMix detects the breakpoints (copy number changes) similar to Navin, with an accuracy of 76%, using the breakpoint margin (of a size corresponding to 5% of the sequence length). Again, there is no ground truth to verify the Navin results from Leung et al. (2017).

A sequential CopyMix Applying a sequential version of CopyMix (1. performing copy number profiling, using EM for HMMs; 2. performing clustering over the inferred copy numbers, using EM for a mixture model), resembling Shah et al. (2006); Vitak et al. (2017) but excluding their ad-hoc processes, on the biological data, results in poor clustering results (V-measure of .24) while, as mentioned, CopyMix gains V-measure of 1. This result confirms that simultaneous detection of copy number profiles and clustering is superior to the sequential approach. In particular, we conclude that methods that use EM for HMMs, using HMMcopy Shah et al. (2006) framework, benefit from the multiple ad-hoc transformations within the HMMcopy framework, much more than the pure computational model, e.g., HMM. Adding multiple ad-hoc rules within such computational models lacks elegance, integrity, and potential model-based approach benefits.

4 Conclusion

The tasks of single-cell copy number profiling and clustering have so far been sequential, requiring multiple ad-hoc processing of data; hence prone to introduce clustering artifacts. We introduce a mixture model-based framework, CopyMix, to allow for a fully probabilistic and simultaneous single-cell copy number profiling and clustering using VI. Our approach is evaluated on both simulated and biological data

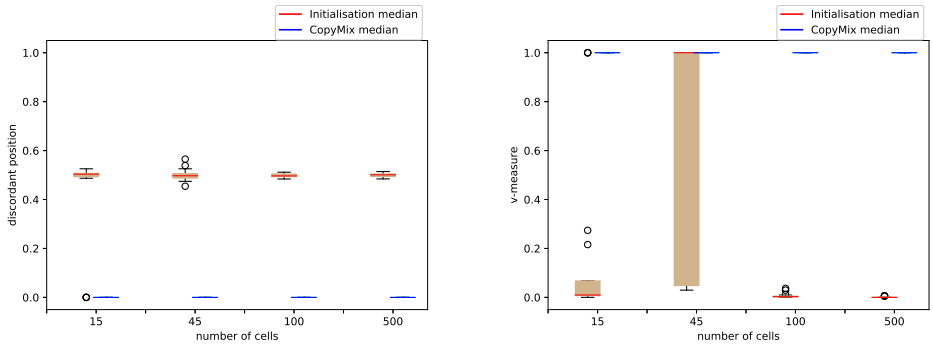

Figure 5: Performance of initialization methods and CopyMix varying the number of cells. LHS and RHS figures, as well as color codes, are similar to Fig. 3.

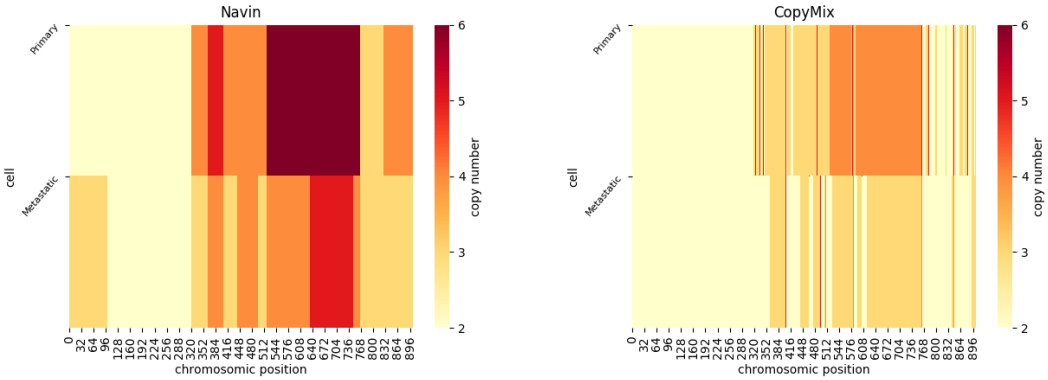

Figure 6: Copy number profiling comparison between the results from Navin Leung et al. (2017) (*LHS*) and CopyMix (*RHS*) for both primary and metastatic clusters.

resulting in good performance w.r.t. both of the tasks.

High performance of CopyMix on the shallow-reads simulated data motivates CopyMix to be further tested for shallow data, a challenge described in Zahn et al. (2017). Another future direction is to apply a biological indicator to detect clustering artifacts, where the benefit of non-sequential methods can be illustrated. Overall, CopyMix provides various research opportunities concerning clustering and copy number profiling, such as improving computational efficiency, augmenting and refining the model, performing experiments concerning the whole genome, and, modeling biallelic deletions and aneuploidy inspired by a recently proposed CopyKat Gao et al. (2021).

Acknowledgments

We thank Dr. Andrew Roth (University of British Columbia and British Columbia Cancer Agency) for providing useful suggestions for the evaluation of CopyMix.

References

Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., ., . ., Navin, N., and Hicks, J. (2012), "Genome-wide copy number analysis of single cells," *Nature Protocols*, 7, 1024–10241.

Bishop, C. (2006), "Pattern recognition and machine learning," *Information science and statistics, New York, NY: Springer*.

Blei, D., Kucukelbir, A., and Mcauliffe, J. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877.

Blocker, A. W. and Meng, X.-L. (2013), "The potential and perils of preprocessing: Building new foundations," *Bernoulli*, 19, 1176–1211.

de Souza, C. P., Andronescu, M., Masud, T., Kabeer, F., Biele, J., Laks, E., Lai, D., Brimhall, J., Wang, B., Su, E., et al. (2020), "Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data," *PLOS Computational Biology*.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., et al. (2015), "Dynamics of genomic clones in breast cancer patient xenografts at single cell resolution," *Nature*, 518, 422.

Gao, R., Bai, S., Henderson, Y. C., ..., and Navin, N. E. (2021), "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes," *Nat Biotechnology*.

Garvin, T., Aboukhalil, R., and Kendall, J. e. a. (2015), "Interactive analysis and assessment of single-cell copy-number variations," *Nat Methods*, 1058–11060.

Gawad, C., Koh, W., and Quake, S. R. (2016), "Single-cell genome sequencing: current state of the science," *Nature Reviews Genetics*, 17, 175.

Johnson, R. A. and Wichern, D. W. (2007), *Applied multivariate statistical analysis, 6th Ed.*, Pearson Prentice Hall.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An introduction to variational methods for graphical models," *Machine learning*, 37, 183–233.

Kapourani, C.-A. and Sanguinetti, G. (2019), "Melissa: Bayesian clustering and imputation of single-cell methylomes," *Genome biology*, 20, 61.

Lawson, D., Kessenbrock, K., Davis, R., Pervolarakis, N., and Werb, Z. (2018), "Tumour heterogeneity and metastasis at single-cell resolution," *Nature Cell Biology*, 20, 1349–1360.

Leung, M., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., ., . ., and Navin, N. (2017), "Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer," *Genome Research*, 27, 1287–1299.

MacKay, D. J. (1997), "Ensemble Learning for Hidden Markov Models," Tech. rep.

McGrory, C. A. and Titterington, D. (2009), "Variational Bayesian analysis for hidden Markov models," *Australian & New Zealand Journal of Statistics*, 51, 227–244.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011), "Tumour evolution inferred by single-cell sequencing," *Nature*, 472, 90.

Nowell, P. C. (1976), "The clonal evolution of tumor cell populations," *Science*, 194, 23–28.

Rosenberg, A. and Hirschberg, J. (2007), "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

Roth, A., McPherson, A., Laks, E., Biele, J., Yap, D., Wan, A., Smith, M. A., Nielsen, C. B., McAlpine, J. N., Aparicio, S., et al. (2016), "Clonal genotype and population structure inference from single-cell tumor sequencing," *Nature methods*, 13, 573.

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, e431–e439.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013), "Single-cell sequencing-based technologies will revolutionize whole-organism science," *Nature Reviews Genetics*, 14, 618.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

Vitak, S., Torkenczy, K., Rosenkrantz, J., Fields, A., Christiansen, L., ..., and Adey, A. (2017), "Sequencing thousands of single-cell genomes with combinatorial indexing," *Nature methods*, 14.

Witten, D. (2011), "Classification and clustering of sequencing data using a Poisson model," *Ann. Appl. Stat.*, 5, 2493–2518.

Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S. P., Aparicio, S., and Hansen, C. L. (2017), "Scalable whole-genome single-cell library preparation without preamplification," *Nature methods*, 14, 167.

Zhang, A. and Campbell, K. (2020), "Computational modelling in single-cell cancer genomics: methods and future directions," *Phys. Biol.*

Zuo, C., Chen, K., Hewitt, K., Bresnick, E., and Keleş, S. (2016), "A hierarchical framework for state-space matrix inference and clustering," *Ann. Appl. Stat.*, 10, 1348–1372.

A Appendix: Derivation details

A.1 Step 2 of Section 2

In this section, we provide the calculation of each decomposed term of $\log P(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \mathbf{\Psi})$. Each term of $\log P(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \mathbf{\Psi})$ presented in Step 2 of Section 2 is calculated as follows: $\log P(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \mathbf{\Psi}) =$
$\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{j=1}^{J}\mathrm{I}(Z_n = k)\mathrm{I}(C_{km} = j)\log F_{j\theta_n}(y_{nm}) =$
$\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{j=1}^{J}\mathrm{I}(Z_n = k)\mathrm{I}(C_{km} = j)\Big[y_{nm}\log(j\theta_n) - j\theta_n - \log(y_{nm}!))\Big]$;
$\log P(\mathbf{C}|\mathbf{\Psi}) = \sum_{k=1}^{K}\big[\sum_{j=1}^{J}\mathrm{I}(C_{k1} = j)\log\rho_{kj} + \sum_{m=2}^{M}\sum_{i=1}^{J}\sum_{j=1}^{J}\mathrm{I}(C_{km-1} = i, C_{km} = j)\log a_{ij}^{k}\big]$;

$$\log P(\mathbf{Z}|\mathbf{\Psi}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{I}(Z_n = k)\log\pi_k;$$

$$\log P(\mathbf{\Psi}) = \log P(\mathbf{A}) + \log P(\boldsymbol{\theta}) + \log P(\boldsymbol{\pi}); \tag{4}$$

where,

$$\log P(\boldsymbol{\theta}) = \sum_{n=1}^{N}\big[(\epsilon_{sn}^{0} - 1)\log\theta_n - \epsilon_{rn}^{0}\theta_n\big] + C,$$

$$\log P(\mathbf{A}) \quad = \quad \sum_{k=1}^{K}\sum_{i=1}^{J}\big[\sum_{j=1}^{J}(\Lambda_{ij}^{k0}-1)\log a_{ij}^{k}\big]+C,$$

and

$$\log P(\boldsymbol{\pi}) \quad = \quad \sum_{k=1}^{K}(\alpha_k^0-1)\log\pi_k + C.$$

A.2 Step 3 and 4 of Section 2

In this section we first present the details of the step 3 of section 2, i.e., the derivation of each posterior update equation (VD computation). We then summarise them in the form of a pseudo code, covering the details of step 4 in Section 2. For convenience, we use $_+\!\approx$ to denote equality up to a constant additive factor.

Update equation for $\boldsymbol{\pi}$: $\log q(\boldsymbol{\pi})\,_+\!\approx \mathrm{E}_{-\boldsymbol{\pi}}\big(\log P(\mathbf{Y},\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})\big)$
$_+\!\approx \mathrm{E}_{-\boldsymbol{\pi}}\big(\log P(\mathbf{Z}|\boldsymbol{\Psi})\big) \;+\; \mathrm{E}_{-\boldsymbol{\pi}}\big(\log P(\boldsymbol{\pi})\big)$
$=\sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big)\log\pi_k+\log P(\boldsymbol{\pi})$
$=\sum_{k=1}^{K}\log\pi_k\Big[\sum_{n=1}^{N}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big)\Big]+\sum_{k=1}^{K}\log\pi_k(\alpha_k^0-1)$
$=\sum_{k=1}^{K}\log\pi_k\Big[\Big(\sum_{n=1}^{N}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big)+\alpha_k^0\Big)-1\Big]$. Let $\mathrm{I}(\cdot)$ be the indicator function. $q(\boldsymbol{\pi})$ is a Dirichlet distribution with parameters $\alpha=(\alpha_1,\ldots,\alpha_K)$, where

$$\alpha_k=\alpha_k^0+\sum_{n=1}^{N}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big). \tag{5}$$

Update equation for Z_n*:*

$$\begin{aligned}\log q(Z_n) \quad &_+\!\approx \quad \mathrm{E}_{-Z_n}\big(\log P(\mathbf{Y},\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})\big)\\ &_+\!\approx \quad \mathrm{E}_{-Z_n}\big(\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})\big)+\mathrm{E}_{-Z_n}\big(\log P(\mathbf{Z}|\boldsymbol{\Psi})\end{aligned}$$

Let y_{nm} be the observed number of reads aligned to bin m for cell n corresponding to the random variable Y_{nm} and let $D_{n,m,j}\equiv y_{nm}\log\theta_n+y_{nm}\log j-j\theta_n-\log(y_{nm}!)$. Note that $\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})$ and $\log P(\mathbf{Z}|\boldsymbol{\Psi})$ can be written as the sum of two terms, one that depends on Z_n and one that does not, i.e., $\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})=\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{j=1}^{J}\mathrm{I}(Z_n=k)\mathrm{I}(C_{km}=j)D_{n,m,j}\;+$
$\sum_{l\neq n}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{j=1}^{J}\mathrm{I}(Z_l=k)\mathrm{I}(C_{km}=j)D_{n,m,l}$
and

$$\log P(\mathbf{Z}|\boldsymbol{\Psi})=\sum_{k=1}^{K}\mathrm{I}(Z_n=k)\log\pi_k+\sum_{l\neq n}\sum_{k=1}^{K}\mathrm{I}(Z_l=k)\log\pi_k.$$

Consequently, $\log q(Z_n)\,_+\!\approx$
$\sum_{k=1}^{K}\mathrm{I}(Z_n=k)\big\{\mathrm{E}_{q(\boldsymbol{\pi})}(\log\pi_k)+\sum_{m=1}^{M}\sum_{j=1}^{J}\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km}=j)\big)\times\mathrm{E}_{q(\theta_n)}(D_{n,m,j})\big\}$.

The distribution $q(Z_n)$ is a Categorical distribution with parameters $\boldsymbol{\pi}_n=(\pi_{n1},\ldots,\pi_{nK})$, where

$$\pi_{nk}=\frac{\exp(\tilde{\pi}_{nk})}{\sum_{k'=1}^{K}\exp(\tilde{\pi}_{nk'})} \tag{6}$$

with

$$\tilde{\pi}_{nk} = \mathrm{E}_{q(\boldsymbol{\pi})}(\log \pi_k) + \sum_{m=1}^{M}\sum_{j=1}^{J} \mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km}=j)\big)\mathrm{E}_{q(\theta_n)}\big(D_{n,m,j}\big).$$

Update equation for θ_n:

$$\begin{aligned}\log q(\theta_n) \quad &{}_{+}\!\!\approx \quad \mathrm{E}_{-\theta_n}\big(\log P(\mathbf{Y},\mathbf{Z},\mathbf{C},\boldsymbol{\Psi}) \\ &{}_{+}\!\!\approx \quad \mathrm{E}_{-\theta_n}\big(\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})\big) + \mathrm{E}_{-\theta_n}\big(\log P(\boldsymbol{\theta})\big)\end{aligned}$$

We can write $\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})$ as in Eq. A.2 and $\log P(\boldsymbol{\theta})$ which depends on θ_n as,

$$\log P(\boldsymbol{\theta}) = (\epsilon^0_{sn}-1)\log\theta_n - \epsilon^0_{rn}\theta_n + \sum_{l\neq n}\big[(\epsilon^0_{sn}-1)\log\theta_l - \epsilon^0_{rn}\theta_l\big] + C.$$

Therefore, $\log q(\theta_n) {}_{+}\!\!\approx$
$\log\theta_n\Big[\sum_{m=1}^{M} y_{nm}\sum_{k=1}^{K}\sum_{j=1}^{J}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big)\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km}=j)\big)\Big]$
$-\,\theta_n\Big[\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{j=1}^{J}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big)\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km}=j)\big)j\Big]$
$+\log\theta_n(\epsilon^0_{sn}-1) - \theta_n\epsilon^0_{rn}$.
Note that $\sum_{k=1}^{K}\sum_{j=1}^{J}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big)\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km}=j)\big)=1$.
The distribution $q(\theta_n)$ is a Gamma distribution with shape and rate parameters:

$$\epsilon_{sn} = \epsilon^0_{sn} + \sum_{m=1}^{M} y_{nm} \tag{7}$$

$$\epsilon_{rn} = \epsilon^0_{rn} + \Big\{\sum_{m=1}^{M}\sum_{k=1}^{K}\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n=k)\big)\sum_{j=1}^{J}\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km}=j)\big)j\Big\}.$$

Update equation for $\mathbf{a}^k_i$:

$$\begin{aligned}\log q(\mathbf{a}^k_i) \quad &{}_{+}\!\!\approx \quad \mathrm{E}_{-\mathbf{a}^k_i}\big(\log P(\mathbf{Y},\mathbf{Z},\mathbf{C},\boldsymbol{\Psi})\big) \\ &{}_{+}\!\!\approx \quad \mathrm{E}_{-\mathbf{a}^k_i}\big(\log P(\mathbf{C}|\boldsymbol{\Psi})\big) + \mathrm{E}_{-\mathbf{a}^k_i}\big(\log P(\mathbf{A})\big)\end{aligned}$$

Disregarding the terms that do not depend on $\mathbf{a}^k_i$ in $\log P(\mathbf{C}|\boldsymbol{\Psi})$ and $\log P(\mathbf{A})$, we obtain:
$\log q(\mathbf{a}^k_i) {}_{+}\!\!\approx$
$\sum_{j=1}^{J}\sum_{m=2}^{M}\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km-1}=i,C_{km}=j)\big)\log a^k_{ij} + \sum_{j=1}^{J}(\Lambda^{k0}_{ij}-1)\log a^k_{ij} =$
$\sum_{j=1}^{J}\log a_{ij}\Big\{\Big[\Lambda^{k0}_{ij} + \sum_{m=2}^{M}\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km-1}=i,C_{km}=j)\big)\Big]-1\Big\}$. The distribution $q(\mathbf{a}^k_i)$ is Dirichlet with parameters $\boldsymbol{\Lambda}^k_i = (\Lambda^k_{i1},\ldots,\Lambda^k_{iJ})$, where

$$\Lambda^k_{ij} = \Lambda^{k0}_{ij} + \sum_{m=2}^{M}\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km-1}=i,C_{km}=j)\big). \tag{8}$$

Update equation for $\mathbf{C}_k$: In the calculation of $q(\mathbf{C}_k)$, the resulting distribution resembles the posterior probability of the hidden variables in an HMM and only differs in the normalization constant. Similarly to the approach adopted in MacKay (1997), we use a slightly different implementation of forward-backward algorithm. More specifically, we define a graph (one per cluster k) $G=\{V,E\}$ with vertices $V=\{C_{mj}: m\in[M], j\in[J]\}$, having weights denoted as $w_k(C_{mj})$, and edges $E=\{C_{m-1i}C_{mj}: m\in\{2,\ldots,M\}, j\in[J]\}$, having weights denoted as $w_k(C_{m-1i}C_{mj})$. Note that for the simplicity of notation we use C_{mj} instead of $C_{km}=j$ for edges and vertices. We define forward (ϕ^k_{mj}) and backward (β^k_{mj}) quantities similar to HMMs Bishop (2006). Instead of using the terminologies

of transition and emission probabilities, we formulate a corresponding algorithm in terms of the weights of the graph, $w_k(C_{m-1i}C_{mj})$ and $w_k(C_{mj})$. The initial transition probability can be defined as $w_k(C_0C_{1j})$. Notice that k indicates the cluster which the graph belongs to. We calculate forward and backward quantities using the weights, and there is no need to normalize the graph weights since a normalization is performed later when calculating the VD. Having the weights, we compute the two posterior probabilities $q(C_{km} = j)$ and $q(C_{km-1} = i, C_{km} = j)$; note that they are normalized by summing over all $j \in [J]$. McGrory and Titterington (2009) derive the exact normalization of the graph weights, for a similar but less complex graphical model; however, this is unnecessary since the normalization can be performed later, as we do.

$$\begin{aligned}\log q(\mathbf{C}_k) \quad &{}_{+}\!\approx \quad \mathrm{E}_{-\mathbf{C}_k}\big(\log P(\mathbf{Y},\mathbf{Z},\mathbf{C},\mathbf{\Psi})\big)\\ &{}_{+}\!\approx \quad \mathrm{E}_{-\mathbf{C}_k}\big(\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\mathbf{\Psi})\big) + \mathrm{E}_{-\mathbf{C}_k}\big(\log P(\mathbf{C}|\mathbf{\Psi})\big).\end{aligned}$$

We can write $\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\mathbf{\Psi})$ as the sum of two terms, one that depends on $\mathbf{C}_k$ and another that does not, that is, $\log P(\mathbf{Y}|\mathbf{Z},\mathbf{C},\mathbf{\Psi}) = \sum_{n=1}^N \sum_{m=1}^M \sum_{j=1}^J \mathrm{I}(Z_n = k)\mathrm{I}(C_{km} = j)D_{n,m,j}$
$+ \sum_{n=1}^N \sum_{m=1}^M \sum_{k'\neq k} \sum_{j=1}^J \mathrm{I}(Z_n = k)\mathrm{I}(C_{k'm} = j)D_{n,m,j}$.

Let $f(k)$ be a function of k as $\sum_{j=1}^J \mathrm{I}(C_{k1} = j)\log\rho_{kj} + \sum_{m=2}^M \sum_{i=1}^J \sum_{j=1}^J \mathrm{I}(C_{km-1} = i, C_{km} = j)\log a_{ij}^k$, We can write $\log P(\mathbf{C}|\mathbf{\Psi}) = f(k) + \sum_{k'\neq k} f(k')$.

Thus, disregarding the terms that do not depend on $\mathbf{C}_k$ we obtain: $\log q(\mathbf{C}_k) {}_{+}\!\approx$
$\sum_{m=1}^M \sum_{j=1}^J \mathrm{I}(C_{km} = j)\Big\{\sum_{n=1}^N \mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n = k)\big)\mathrm{E}_{q(\theta_n)}\big(D_{n,m,j}\big)\Big\} +$
$\sum_{j=1}^J \mathrm{I}\big(C_{k1} = j\big)\log\rho_{kj} + \sum_{m=2}^M \sum_{j=1}^J \sum_{i=1}^J \mathrm{I}\big(C_{km-1} = i, C_{km} = j\big) \times \mathrm{E}_{q(\mathbf{a}_i^k)}\big(\log a_{ij}^k\big)$. Calculations regarding directed graph are as follows: $\mathrm{u}_k(\mathbf{C}_{1:m-1}, C_m = j) = \prod_{i=1}^{m-1}\prod_{s=1}^J \exp\Big\{\sum_{n=1}^N \mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n = k)\big)\mathrm{E}_{q(\theta_n)}\big(D_{n,i,s}\big)\Big\}^{\mathrm{I}(C_{ki}=s)}$
$\exp\Big\{\sum_{n=1}^N \mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n = k)\big)\mathrm{E}_{q(\theta_n)}\big(D_{n,m,j}\big)\Big\}$
$\prod_{j=1}^J \exp\Big\{\log\rho_{kj}\Big\}^{\mathrm{I}\left(C_{k1}=j\right)}$
$\prod_{t=2}^{m-1}\prod_{s=1}^J\prod_{i=1}^J \exp\Big\{\mathrm{E}_{q(\mathbf{a}_i^k)}\big(\log a_{is}^k\big)\Big\}^{\mathrm{I}\left(C_{kt-1}=i, C_{kt}=s\right)}$
$\prod_{i=1}^J \exp\Big\{\mathrm{E}_{q(\mathbf{a}_i^k)}\big(\log a_{ij}^k\big)\Big\}^{\mathrm{I}\left(C_{km-1}=i, C_{km}=j\right)}$.
and
$\mathrm{v}_k(\mathbf{C}_{m+1:M}, C_m = j) = \prod_{i=m+1}^M\prod_{s=1}^J \exp\Big\{\sum_{n=1}^N \mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n = k)\big)\mathrm{E}_{q(\theta_n)}\big(D_{n,i,s}\big)\Big\}^{\mathrm{I}(C_{ki}=s)}$
$\prod_{r=m+1}^M\prod_{s=1}^J\prod_{i=1}^J \exp\Big\{\mathrm{E}_{q(\mathbf{a}_i^k)}\big(\log a_{is}^k\big)\Big\}^{\mathrm{I}\left(C_{kr}=i, C_{kr+1}=s\right)}$
$\prod_{i=1}^J \exp\Big\{\mathrm{E}_{q(\mathbf{a}_i^k)}\big(\log a_{ji}^k\big)\Big\}^{\mathrm{I}\left(C_{km}=j, C_{km+1}=i\right)}$.

We define $\phi_{mj}^k = \sum_{C_1=1}^J .. \sum_{C_{m-1}=1}^J u_k(\mathbf{C}_{1:m-1}, C_m = j)$ and
$\beta_{mj}^k = \sum_{C_{m+1}=1}^J .. \sum_{C_M=1}^J v_k(\mathbf{C}_{m+1:M}, C_m = j)$.

The graph weights are (we assume $w_k(C_0C_{1j})$ to be fixed):
$w_k(C_{m-1i}C_{mj}) = \exp\big\{\mathrm{E}_{q(\mathbf{a}_i^k)}(\log a_{ij}^k)\big\}$
and
$$w_k(C_{mj}) = \exp\Big\{\sum_{n=1}^N \mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n = k)\big)\mathrm{E}_{q(\theta_n)}\big(D_{n,m,j}\big)\Big\}.$$

Note that we skip writing i, j in the calculations to make them short and more readable:

$\phi_m^k = \sum_{C_1=1}^J .. \sum_{C_{m-1}=1}^J u_k(\mathbf{C}_{1:m-1}, C_m) = \sum_{C_1=1}^J .. \sum_{C_{m-1}=1}^J u_k(\mathbf{C}_{1:m-2}, C_{m-1}, C_m) =$
$\sum_{C_1=1}^J .. \sum_{C_{m-1}=1}^J u_k(\mathbf{C}_{1:m-2}, C_{m-1}) w_k(C_m) w_k(C_{m-1}C_m) = w_k(C_m) \sum_{C_{m-1}=1}^J \phi_{m-1}^k w_k(C_{m-1}C_m)$.

$\beta_m^k = \sum_{C_{m+1}=1}^J .. \sum_{C_M=1}^J v_k(\mathbf{C}_{m+1:M}, C_m) = \sum_{C_{m+1}=1}^J .. \sum_{C_M=1}^J v_k(\mathbf{C}_{m+2:M}, C_{m+1}, C_m) =$
$\sum_{C_{m+1}=1}^J .. \sum_{C_M=1}^J v_k(\mathbf{C}_{m+2:M}) w_k(C_m C_{m+1}) w_k(C_{m+1}) = \sum_{C_{m+1}=1}^J \beta_{m+1}^k w_k(C_m C_{m+1}) w_k(C_{m+1})$.

We now calculate the posteriors: $\mathrm{q}(\mathrm{C}_{km} = j) = \sum_{C_{k,1:m-1}}^J \sum_{C_{k,m+1:M}}^J q(C_{k,1:m-1}, C_{km} = j, C_{k,m+1:M}) =$
$\sum_{C_{1:m-1}}^J u_k(\mathbf{C}_{1:m-1}, C_m = j) \sum_{C_{m+1:M}}^J v_k(\mathbf{C}_{m+1:M}, C_m = j) = \phi_{mj}^k \beta_{mj}^k$.
$q(C_{km-1} = i, C_{km} = j) =$
$\sum_{C_{k,1:m-2}}^J \sum_{C_{k,m+1:M}}^J q(C_{k,1:m-2}, C_{k,m-1} = i, C_{km} = j, C_{k,m+1:M}) =$
$\sum_{C_{1:m-2}}^J u_k(\mathbf{C}_{1:m-2}, C_{m-1} = i) w_k(C_{m-1i} C_{mj}) w_k(C_{mj})$
$\sum_{C_{m+1:M}}^J v_k(\mathbf{C}_{m+1:M}, C_m = j) = w_k(C_{m-1i} C_{mj}) w_k(C_{mj}) \phi_{m-1i}^k \beta_{mj}^k$.

Finally, the following are the resulted update equations. Subsequently, the detailed algorithm used in CopyMix is shown in Algorithm 2.

$$\mathrm{w}_k(C_{m-1i} C_{mj}) = \exp\left\{\mathrm{E}_{q(\mathbf{a}_i^k)}(\log a_{ij}^k)\right\}$$

$$w_k(C_{mj}) = \exp\Big\{\sum_{n=1}^N \mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n = k)\big) \mathrm{E}_{q(\theta_n)}\big(D_{n,m,j}\big)\Big\}$$

$$q(C_{km} = j) = \frac{\phi_{mj}^k \beta_{mj}^k}{\sum_{j=1}^J \phi_{mj}^k \beta_{mj}^k}$$

$$q(C_{km-1} = i, C_{km} = j) = \frac{w_k(C_{m-1i} C_{mj}) w_k(C_{mj}) \phi_{m-1i}^k \beta_{mj}^k}{\sum_{j=1}^J w_k(C_{m-1i} C_{mj}) w_k(C_{mj}) \phi_{m-1i}^k \beta_{mj}^k}$$

A.3 Calculating expectations

In this section we list the calculations of the expected values that are used in the update equations in the previous Appendix. Let $\mathbf{\Phi}$ be the digamma function defined as $\mathbf{\Phi}(x) = \frac{d}{dx} \log \Gamma(x)$ which can be easily calculated via numerical approximation. The values of the expectations taken with respect to the approximated distributions are given as follows where $\mathbf{\Phi}$ is used for some of them.

- $\mathrm{E}_{q(Z_n)}\big(\mathrm{I}(Z_n = k)\big) = \pi_{nk}$
- Due to computational issues in calculating $\mathrm{E}_{q(\theta_n)}\big(D_{n,m,j}\big) = y_{nm} E_{q(\theta_n)}(\log \theta_n) + y_{nm} \log j - j E_{q(\theta_n)}(\theta_n) - \log(y_{nm}!)$, we can instead compute it using the fact that
  $\mathrm{E}_{q(\theta_n)}(\theta_n) = \epsilon_{sn}/\epsilon_{rn}$ and
  $\mathrm{E}_{q(\theta_n)}\big(\log \theta_n\big) = \mathbf{\Phi}(\epsilon_{sn}) + \log E_{q(\theta_n)}(\theta_n) - \log \epsilon_{sn}$.
  Therefore, we obtain $\mathrm{E}_{q(\theta_n)}\big(D_{n,m,j}\big) = \log F_{\tilde{\theta}}(y_{nm}) + y_{nm} \times \big[\mathbf{\Phi}(\epsilon_{sn}) - \log \epsilon_{sn}\big]$, with $F_{\tilde{\theta}}$ being a Poisson p.m.f. with rate parameter $\tilde{\theta} = j \times (\epsilon_{sn}/\epsilon_{rn})$.
- $\mathrm{E}_{q(\boldsymbol{\pi})}(\log \pi_k) = \mathbf{\Phi}(\alpha_k) - \mathbf{\Phi}\Big(\sum_{k=1}^K \alpha_k\Big)$
- $\mathrm{E}_{q(\mathbf{a}_i^k)}(\log a_{ij}^k) = \mathbf{\Phi}(\Lambda_{ij}^k) - \mathbf{\Phi}\Big(\sum_{j=1}^J \Lambda_{ij}^k\Big)$

Algorithm 2 CopyMix: a Variational Inference Algorithm

procedure ESTIMATE-POSTERIORS(Y): ▷ Y input is N sequences of length M
Initialization and assigning priors ϵ^0_{sn}, ϵ^0_{rn}, $\mathbf{\Lambda}^{k0}_i$, α^0_k, π_{nk}, $w_k(C_0C_{1j})$
Update ϵ_{sn} using y and ϵ^0_{sn} in Eq. 7
Update $w_k(C_{m-1i}C_{mj})$ using $\mathbf{\Lambda}^{k0}_i$ and π_{nk} in Eq. A.2
Update $w_k(C_{mj})$ using $\mathbf{\Lambda}^{k0}_i$, π_{nk}, ϵ^0_{rn}, ϵ_{sn} and y in Eq. A.2
  repeat
    Update $\alpha^{(c)}_k$ using $\pi^{(c-1)}_{nk}$ and α^0_k in Eq. 5
    Update $q(C_{km}=j)^{(c)}$ and $q(C_{km-1}=i, C_{km}=j)^{(c)}$ using
      $w_k(C_0C_{1j})^{(c-1)}$, $w_k(C_{m-1i}C_{mj})^{(c-1)}$ and
      $w_k(C_{mj})^{(c-1)}$ (Eq. A.2 to A.2)
    Update $\mathbf{\Lambda}^{k(c)}_i$ using $\mathbf{\Lambda}^{k0}_i$ and $q(C_{km-1}=i, C_{km}=j)^{(c)}$ in Eq. 8
    Update $\epsilon^{(c)}_{rn}$ using ϵ^0_{rn}, $\pi^{(c-1)}_{nk}$ and $q(C_{km}=j)^{(c)}$ in Eq. A.2
    Update $\pi^{(c)}_{nk}$ using $\alpha^{(c)}_k$, ϵ_{sn}, $\epsilon^{(c)}_{rn}$, $q(C_{km}=j)^{(c)}$ and y in Eq. 6
    Update $w_k(C_{mj})^{(c)}$ using $\mathbf{\Lambda}^{k(c)}_i$, $\pi^{(c)}_{kn}$, $\epsilon^{(c)}_{rn}$, ϵ_{sn} and y in Eq. A.2
    Update $w_k(C_{m-1i}C_{mj})^{(c)}$ using $\mathbf{\Lambda}^{k(c)}_i$ and $\pi^{(c)}_{nk}$ in Eq. A.2
    $c \leftarrow c+1$
  until convergence of the ELBO in Eq. 2
Output all posterior distributions including:
$q(Z_n=k) \quad \forall n,k;$ ▷ cluster assignment probabilities
$q(C_{km}=j) \quad \forall k,m,j;$ ▷ copy number probabilities
return posterior distribution

- $\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km}=j)\big) = q(C_{km}=j)$ and $\mathrm{E}_{q(\mathbf{C}_k)}\big(\mathrm{I}(C_{km-1}=i, C_{km}=j)\big) = q(C_{km-1}=i, C_{km}=j)$, which are calculated in Eq. A.2 and A.2.