Copula-based Poisson mixture models for RNAseq data

Dimitris Karlis Gildas Mazo Andrea Rau*

1 Reminder about the EM algorithm

Let $\mathbf{y}$ be the observed data and $(\mathbf{y}, z)$ be the complete data. The heuristic of the EM algorithm consists of maximizing

$$\mathrm{E}\left(\log f(\mathbf{y}, Z)|\mathbf{y}\right).$$

The law of $Z|\mathbf{y}$ is

$$P(Z = k|\mathbf{y}) \equiv f(k|\mathbf{y}) = \frac{f(\mathbf{y}|k)\pi_k}{\sum_{k=1}^K \pi_k f(\mathbf{y}|k)} = \frac{f_k(\mathbf{y})\pi_k}{\sum_{k=1}^K \pi_k f_k(\mathbf{y})}.$$

The EM algorithm works on $b = 0, 1, \ldots$ as follows.

E step. For $k = 1, \ldots, K$, $i = 1, \ldots, n$, compute

$$f(k|\mathbf{y}_i; \boldsymbol{\theta}^{(b)}) = \frac{f_k(\mathbf{y}_i; \boldsymbol{\theta}^{(b)})\pi_k^{(b)}}{\sum_{k=1}^K \pi_k^{(b)} f_k(\mathbf{y}_i; \boldsymbol{\theta}^{(b)})}.$$

M step. Maximize over $(\boldsymbol{\pi}, \boldsymbol{\theta})$:

$$\begin{aligned}
&\sum_{i=1}^n \sum_{k=1}^K f(k|\mathbf{y}_i; \boldsymbol{\theta}^{(b)}) \log f(\mathbf{y}_i, k; \boldsymbol{\theta}) \\
=&\sum_{i=1}^n \sum_{k=1}^K f(k|\mathbf{y}_i; \boldsymbol{\theta}^{(b)}) \log f_k(\mathbf{y}_i; \boldsymbol{\theta}) + \sum_{i=1}^n \sum_{k=1}^K f(k|\mathbf{y}_i; \boldsymbol{\theta}^{(b)}) \log \pi_k
\end{aligned}$$

2 Andrea's model

2.1 The Poisson mixture model for RNAseq data

The following description closely follows that in [1]. Let Y_{ijl} be the random variable corresponding to the digital gene expression measure (DGE) for biological entity i ($i = 1, \ldots, n$) of condition j ($j = 1, \ldots, d$) in biological replicate

*alphabetical order

l ($l = 1, \ldots, r_j$), with y_{ijl} being the corresponding observed value of Y_{ijl}. Let $q = \sum_{j=1}^{d} r_j$ be the total number of variables (all replicates in all conditions) in the data, such that $\mathbf{y} = (y_{ijl})$ is the $n \times q$ matrix of the DGE for all observations and variables, and $\mathbf{y}_i$ is the q-dimensional vector of DGE for all variables of observation i. We use dot notation to indicate summations in various directions, e.g., $y_{\cdot jl} = \sum_i y_{ijl}$, $y_{i\cdot\cdot} = \sum_j \sum_l y_{ijl}$, and so on.

To cluster RNA-seq data, we consider a model-based clustering procedure based on mixture of Poisson distributions. The data $\mathbf{y}$ are assumed to come from K distinct subpopulations (clusters), each of which is modeled separately:

$$f(\mathbf{y}; K, \boldsymbol{\Psi}_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_{ik})$$

where $\boldsymbol{\Psi}_K = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\theta}')'$, $\boldsymbol{\theta}'$ contains all of the parameters in $\{\boldsymbol{\theta}_{ik}\}_{i,k}$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ are the mixing proportions, with $\pi_k \in (0, 1)$ for all k and $\sum_{k=1}^{K} \pi_k = 1$. Samples are assumed to be independent conditionally on the components:

$$f_k(\mathbf{y}_i; \boldsymbol{\theta}_{ik}) = \prod_{j=1}^{d} \prod_{l=1}^{r_j} \mathcal{P}(y_{ijl}; \mu_{ijlk}),$$

where $\mathcal{P}(\cdot; \mu_{ijlk})$ denotes the standard Poisson probability mass function with mean μ_{ijlk}.

Each mean μ_{ijlk} is parameterized by

$$\mu_{ijlk} = w_i s_{jl} \lambda_{jk}$$

where $w_i = y_{i\cdot\cdot}$ corresponds to the overall expression level of observation i (e.g., weakly to strongly expressed) and s_{jl} represents the normalized library size for replicate l of condition j, such that $\sum_{j,l} s_{jl} = 1$. These normalization factors take into account the fact that the number of reads expected to map to a particular gene depends not only on its expression level, but also on the library size (overall number of mapped reads) and the overall composition of the RNA population being sampled. We note that $\{s_{jl}\}_{j,l}$ are estimated from the data prior to fitting the model, and like the overall expression levels w_i, they are subsequently considered to be fixed in the Poisson mixture model. Finally, the unknown parameter vector $\boldsymbol{\lambda}_k = (\lambda_{1k}, \ldots, \lambda_{dk})$ corresponds to the clustering parameters that define the profiles of the genes in cluster k across all biological conditions.

2.2 Inference through the EM algorithm

To estimate mixture parameters $\boldsymbol{\Psi}_K = (\boldsymbol{\pi}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$ by computing the maximum likelihood estimate (MLE), an Expectation-Maximization (EM) algorithm is considered. After initializing the parameters $\boldsymbol{\Psi}_K^{(0)}$ and $\mathbf{z}^{(0)}$ by a so-called

Small-EM strategy, the E-step at iteration b corresponds to computing the conditional probability that an observation i arises from the kth component for the current value of the mixture parameters:

$$t_{ik}^{(b)} = \frac{\pi_k^{(b)} f_k(\mathbf{y}_i; \boldsymbol{\theta}_{ik}^{(b)})}{\sum_{m=1}^{K} \pi_m^{(b)} f_m(\mathbf{y}_i; \boldsymbol{\theta}_{im}^{(b)})}$$

where $\boldsymbol{\theta}_{ik}^{(b)} = \{w_i s_{jl} \lambda_{jk}^{(b)}\}_{jl}$. Then, in the M-step the mixture parameter estimates are updated to maximize the expected value of the completed likelihood, which leads to weighting the observation i for group k with the conditional probability $t_{ik}^{(b)}$. Thus,

$$\pi_k^{(b+1)} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}^{(b)}, \tag{1}$$

and

$$\lambda_{jk}^{(b+1)} = \frac{\sum_{i=1}^{n} t_{ik}^{(b)} y_{ij\cdot}}{s_{j\cdot} \sum_{i=1}^{n} t_{ik}^{(b)} y_{i\cdot\cdot}},$$

since $w_i = y_{i\cdot\cdot}$. Note that at each iteration of the EM algorithm, we obtain that $\sum_{j=1}^{d} \lambda_{jk}^{(b)} s_{j\cdot} = 1$. Thus $\lambda_{jk}^{(b)} s_{j\cdot}$ can be interpreted as the proportion of reads that are attributed to condition j in cluster k, after accounting for differences due to library size; this proportion is shared among the replicates of condition j according to their respective library sizes s_{jl}.

3 Incorporating copulas

The idea is to model the dependence between the conditions/replicates of the gene expressions with copulas.

- First we need to decide where to put the copulas. Unlike conditions, can replicates be considered independent?
- Then rewrite the model
- Finally write an EM algo as in [2, 3]. Or a pseudo-EM algo as in [4].

4 Copulas

A copula is a function C which can "couple" the marginals to model the dependence structure.

For instance, let two Poisson distributions, $f(y;\mu_j) = \mu_j^y e^{-\mu_j}/y!$, $j = 1, 2$. The (cumulative) distribution functions are given by

$$F(y;\mu_j) = \sum_{m=0}^{y} f(m;\mu_j), \quad j = 1, 2.$$

Without copulas, the joint distribution function is given by

$$F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2) = F(y_1;\mu_1)F(y_2;\mu_2)$$

(independence). We can couple the marginals to add a dependence structure. Take, for instance, the AMH copula:

$$C(u, v;\gamma) = uv + \gamma uv(1-u)(1-v), \gamma \in [-1, 1],\ u, v \in [0, 1].$$

Then

$$F(y_1, y_2;\mu_1, \mu_2, \gamma) \equiv C(F(y_1;\mu_1), F(y_2;\mu_2);\gamma),$$

is a well defined distribution function with a dependence structure.

References

[1] A. Rau et al. Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 2015.

[2] Ioannis Kosmidis and Dimitris Karlis. Model-based clustering using copulas with applications. *Statistics and computing*, 26(5):1079–1099, 2016.

[3] Aristidis K Nikoloulopoulos and Dimitris Karlis. Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation*, 39(1):172–187, 2009.

[4] Gildas Mazo. A semiparametric and location-shift copula-based mixture model. *Journal of Classification*, 34(3):444–464, 2017.