# EM algorithm for a mixture of binomial logit normal distributions

Robert Vogel

August 2022 & Feb. 17, 2023

This document contains the calculations performed for maximum likelihood parameter inference under the mixture of binomial logit normal (bln) distributions with equal dispersion $\sigma^2$. The algorithm was originally derived in August of 2022, but this digitized version was prepared in February 2023.

In Mohammadi *et al.* 2019 [3], Pejman showed that allele specific expression data is effectively modeled by the overdispersed bln distribution. Let us review allele specific expression and, at a high level, the bln distribution.

Consider a gene that has at least one biallelic variant in its coding region. Let us focus our attention on any one of these variants. As humans are diploid, the genotype of any one individual at this variant is defined as the pair (allele from mother, allele from father). More generally, we may consider the genotype as the alleles from haplotypes 1 and 2, as we don't always know whether haplotype 1 is from the mother or father. Given that the variant is biallelic, the possible genotypes are (ref, ref), (alt, ref), (ref, alt), and (alt, alt). In addition, as the variant is in the coding region of a gene, the base corresponding to the ref and alt allele will be encoded in transcribed RNA. The result is that each transcirpt will contain an allele of the variant and consequently a measurable biochemical marker that correspondes to the haplotype that the transcript was transcribed. The allele specific quantification by RNA sequencing is denoted as allele specific expression (ASE) data.

The transcription process is noisy, resulting in alt and ref read counts to vary across samples. Intuitively, as there is only two outcomes, we assume that the distribution of alt counts to follow the binomial distribution. Here, the number of trials is the total count, defined as the sum of alt and ref alleles, and the probability of success is determined by the relative transcription rates. For example, if the two haplotypes have identical gene regulation sequences, we would expect that each haplotype produce the same number of transcripts. Consequently, if we were to select a transcript at random, we would have $p = 1/2$ chance selecting the alt allele. However, if haplotype 1 carries the alternative allele and so happens to also have a gene regulatory sequence that increase transcription 3 fold, the probability of selecting a read with the alternative allele is $p = 3/4$. While intuitive, application of this model was insufficienct to describe the observed variablity, enter overdispersed models.

In [3] Pejman built an intuitive overdispersed model of ASE data by a compound distribution consisting of the binomial and logit-normal distributions. The binomial distribution models the alt read counts obsered, while logit-normal models the variability in a population of the probability of selecting a single alt read. The motivation of the logit-normal distributoin is from the his logit-linear model of $p$. Specifically, the logit $(p)$ is assumed to be a linear combination of biallelic *cis* regulatory variant effects and Gaussian distributed errors. The "errors" in this case do not model true error, but the contribution of unmeasured genetic and environmental factors that regulate gene expression. Consequently, if we could measure every regulatory factor with perfect precision then variance of the error would be vanishingly small, perhaps zero, and the binomial distribution would sufficiently model the data. The resulting compound distribution results in a pricipled and overdispersed binomial distribution.

In the text that follows, I provide specific definitions of variables, distributions, etc. for the binomial-logit-normal distribution. Then I present an EM algorithm for maximum likelihood estimation for a mixture of bln distributions. I have implemented this algorithm in a software package named `blnm` using the Python programming language. Lastly, I present inferences from simulation data using the `blnm` package.

# 1 Problem statement

Consider an ASE data set for a single gene that consists of $N \in \mathbb{Z}_{\geq 0}$ independent and identically distributed samples. For each sample $i$ there exists alt count $x_i \in \mathbb{Z}_{\geq 0}$ and total count $n_i \in \mathbb{Z}_{\geq 0}$ measurements s.t. $x_i \leq n_i$, and consequently the ASE data set is defined as $\mathcal{D} = \{(x_i, n_i)\}_{i=1}^{N}$. The likelihood of the $i^{th}$ sample with respect to parameter set $\boldsymbol{\theta} = [\mu_1, \sigma_1^2, w_1, \ldots, \mu_k, \sigma_k^2, w_k]^T$ is modeled as a mixture of $K$ bln distributions,

$$\mathfrak{L}\left(\boldsymbol{\theta}; x_i, n_i\right) = \sum_{\forall \mathbf{z}_i \in \Omega_Z} \Pr\left(X_i = x_i | n_i, \mu_k, \sigma_k^2, \mathbf{z}_i\right) \Pr\left(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{w}\right). \tag{1}$$

with sample space $\Omega_Z$ being the set of all $K$ possible "one-hot" encoded vectors $\mathbf{z}_i \in \{0,1\}^{K\times 1}$ and $\mathbf{w} \in \mathbb{R}^{K\times 1}$ s.t. $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$. The probability of alt counts given $\mathbf{z}_i$ is modeled

$$\Pr\left(X_i = x_i | n_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{z}_i\right) = \prod_{k=1}^{K} p_{X_i}\left(x_i; n_i, \mu_k, \sigma_k^2\right)^{z_{ik}} \tag{2}$$

with $p_{X_i}\left(x_i; n_i, \mu_k, \sigma_k^2\right)$ being the $k^{th}$ bln probability mass function, what I call the location $\boldsymbol{\mu} \in \mathbb{R}^{K\times 1}$, and dispersion $\boldsymbol{\Sigma} \in \mathbb{R}_{>0}^{K\times 1}$ with the $k^{th}$ element denoted $\sigma_k^2$. Recall the definition of the bln probability mass function.

**Definition 1** (bln probability mass function)**.** *Let $X \in \mathbb{Z}_{\geq 0}$ be a bln distributed random variable with $n$ representing the number of independent trials, location* *parameter $\mu$, and dispersion parameter $\sigma^2$. The bln distribution is defined as the compound distribution*

$$p_X(x; n, \mu, \sigma^2) = \int_{s=-\infty}^{\infty} p_{X|S}(x; n, g(s)) f_S(s; \mu, \sigma^2) ds$$

*of the binomial $p_{X|S}(x; n, g(s))$ and normal $f_S(s, \mu, \sigma^2)$ distributions. The function $g : \mathbb{R} \to [0, 1]$ is the logistic function, $(1 + e^{-s})^{-1}$.*

The "one-hot" encoded component membership of sample $i$ is modeled by the multinomial distribution

$$p_{\mathbf{Z}_i}(\mathbf{z}_i|\mathbf{w}) = \prod_{k=1}^{K} w_k^{z_{ik}}. \tag{3}$$

where the number of samples drawn is one.

The task at hand is to derive the maximum likelihood estimators of model parameters in (1). Given data set $\mathcal{D}$, distributions (2) and (3) the log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathcal{D}) = \sum_{i=1}^{N} \log \left( \sum_{\forall \mathbf{z}_i \in \Omega_Z} \prod_{k=1}^{K} p_{X_i}\left(x_i; n_i, \mu_k, \sigma_k^2\right)^{z_{ik}} w_k^{z_{ik}} \right). \tag{4}$$

The log of the sum of terms precludes an analytical solution. As such, we move our aim of analytical estimators to a computational procedure for maximum likelihood estimates of parameter values. The standard procedure for latent variable models is the expectation-maximization algorithm, otherwise denoted EM [1, 2]. In what follows we derive such an algorithm for our problem.

## 2 An EM algorithm

In this section, we provide a short and high level description of EM, and apply that understanding to derive equations for our model. The log-likelihood of (4) has two latent variables for each sample $i$: the discrete class membership $\mathbf{Z}_i$ and continuous $S_i$ from the definition of the binomial-logit-normal pmf in Def. 1. Consequently, there will be many equations containing sums, integrals, and functions. This makes any exposition of the method messy, we've tried to make points more clear by adopting sensible abbreviations.

Bishop [1] gives a fantastic discussion of EM, which the following explanation follows, see [1] chapter 9 section 4 for a more thorough treatment of the subject. The EM algorithm performs maximum likelihood estimation by a two step iterative procedure. In Eq. 9.70, Bishop [1] expresses the log-likelihood of the data as a sum of two quantities

$$\ell(\boldsymbol{\theta}; \mathcal{D}) = \sum_{i=1}^{N} \mathcal{L}\left(q_{S_i,\mathbf{Z}_i}, \boldsymbol{\theta}\right) + \mathrm{KL}\left(q_{S_i,\mathbf{Z}_i} || p_{S_i,\mathbf{Z}_i}\right) \tag{5}$$

where

$$\mathcal{L}\left(q_{S_i,\mathbf{Z}_i},\boldsymbol{\theta}\right) := \mathbb{E}_{\underline{S_i},\underline{\mathbf{Z}_i}\sim q_i}\left[\log\left(\frac{\Pr\left(X_i,S_i,\mathbf{Z}_i\middle|n_i,\boldsymbol{\theta}\right)}{q_{S_i,\mathbf{Z}_i}\left(s_i,\mathbf{z}_i\middle|\boldsymbol{\theta}\right)}\right).\right] \tag{6}$$

which will subsequently be referred to as $\mathcal{L}$ and KL is the Kullback-Leibler divergence between the unknown distribution $q_{S_i,\mathbf{Z}_i}$ and the posterior distribution of our latent variables $p_{S_i,\mathbf{Z}_i}$. To make (6) clear, we used a shorthand notation of the expectation over $q_{S_i,\mathbf{Z}_i}$ defined as

$$\mathbb{E}_{S_i,\mathbf{Z}_i\sim q_i}\left[h(S_i,\mathbf{Z}_i)\right] = \sum_{\forall\mathbf{z}_i\in\Omega_Z}\int_{s_i=-\infty}^{\infty} h(S_i,\mathbf{Z}_i)\; q_{S_i,\mathbf{Z}_i}\left(S_i,\mathbf{Z}_i\middle|\boldsymbol{\theta}\right)\; ds_i \tag{7}$$

for any arbitrary function $h$. At first reading the decomposition (5) appears unhelpful as we still do not have estimates of parameters $\boldsymbol{\theta}$ and now have an unknown distribution function $q_{S_i,\mathbf{Z}_i}$. However, the log in (5) now operates on distribution functions, instead of the sum of distribution functions as in the data log-likelihood (4). This simplifies the inference problem.

The EM algorithm performs iterative updates of parameter values from $\boldsymbol{\theta}^{(j)}$ to $\boldsymbol{\theta}^{(j+1)}$. The E step is simply the calculation of $\mathcal{L}$ given a suitable $q_{S_i,\mathbf{Z}_i}$. The terms of $\mathcal{L}$ that depend on $\boldsymbol{\theta}$ are kept, and the resulting function is often designated $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(j)})$. The M step maximizes $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(j)})$ with respect to $\boldsymbol{\theta}$ resulting in the update parameters $\boldsymbol{\theta}^{(j+1)}$. This procedure repeats itself until convergence, defined as the difference in the log-likelihood between iterations $j+1$ and $j$ as being below an established threshold.

To gain some intuition to why this works in practice a more careful analysis of (5) is helpful. Here we see that when the KL divergence between $q_{S_i,\mathbf{Z}_i}$ and $p^{(j)}_{S_i,\mathbf{Z}_i}$ is zero that $\mathcal{L}$ is equal to the data log-likelihood. When the KL divergence is greater than zero $\mathcal{L} < \ell(\boldsymbol{\theta},\mathcal{D}_i)$, which togheter makes $\mathcal{L}$ a lower bound on the data log-likelihood.

Now consider a parameter set $\boldsymbol{\theta}^{(j)}$, when $q_{S_i,\mathbf{Z}_i} = p^{(j)}_{S_i,\mathbf{Z}_i}$ the KL divergence is zero and the $\mathcal{L}$ is equal to data log-likelihood at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(j)}$. Any deviation in the parameters will result in a non-zero KL divergence, and $\mathcal{L} < \ell(\boldsymbol{\theta},\mathcal{D})$. Together, the lower bound property of $\mathcal{L}$ and equality with $\ell(\boldsymbol{\theta},\mathcal{D})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(j)}$ implies that

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}} = \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathcal{D})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}}.$$

Consequently, any change in parameters that results in an increase in $\mathcal{L}$ would also increase the data log likelihood. Hence, each iteration of the the E and M steps increases the log-likelihood, and convergence is achieved when the gradient of both $\mathcal{L}$ and $\ell$ are zero.

## 2.1 The E step of the mixture of bln distibutions

A parameter update to $\boldsymbol{\theta}^{(j)}$ begins by finding the distribution $q^{(j)}_{S_i,\mathbf{Z}_i}$ in which the KL divergence in (5) is zero. By definition of the KL divergence this is simply when $q^{(j)}_{S_i,\mathbf{Z}_i} = p_{S_i,\mathbf{Z}_i}(s_i, \mathbf{z}_i; \boldsymbol{\theta}^{(j)}, x_i, n_i)$, $p^{(j)}_i$ for short. Next, we substitute $p^{(j)}_i$ for $q^{(j)}_{S_i,\mathbf{Z}_i}$ into $\mathcal{L}\left(q^{(j)}_{S_i,\mathbf{Z}_i}, \boldsymbol{\theta}\right)$ of (6),

$$\begin{aligned}\mathcal{L}\left(q^{(j)}_{S_i,\mathbf{Z}_i} = p^{(j)}_i, \boldsymbol{\theta}\right) = &\overbrace{\mathbb{E}_{S_i,\mathbf{Z}_i\sim p^{(j)}_i}\left[\log\left(\Pr\left(X_i, S_i, \mathbf{Z}_i \middle| n_i, \boldsymbol{\theta}\right)\right)\right]}^{Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(j)})} \\ &- \underbrace{\mathbb{E}_{S_i,\mathbf{Z}_i\sim p^{(j)}_i}\left[\log\left(p^{(j)}_i\right)\right]}_{H:=\text{Entropy of } p^{(j)}_i \implies -H(S_i,\mathbf{Z}_i|\boldsymbol{\theta}^{(j)})}. \end{aligned} \tag{8}$$

The entropy term is not useful for computing the parameter updates, as it is independent of parameters $\boldsymbol{\theta}$. As the "E" step consists of computing the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)})$,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \mathbb{E}_{S_i,\mathbf{Z}_i\sim p^{(j)}_i}\left[\log\left(\Pr\left(X_i, S_i, \mathbf{Z}_i \middle| n_i, \boldsymbol{\theta}\right)\right)\right] \tag{9}$$

Next we apply the E step to our problem. First, let's calculate the posterior distribution of the latent variables.

**Result 1** (Posterior $j$)**.** *Under the problem setup, the posterior distribution of the $i^{th}$ sample latent variables given $\boldsymbol{\theta}^{(j)}$ is*

$$p^{(j)}_i = \frac{1}{\Omega^{(j)}_i} \prod_{k=1}^{K} \left(p_{X_i|S_i}(x_i)\ f^{(j)}_{S_{ik}}(s_i)\ w^{(j)}_k\right)^{z_{ik}}$$

*where*

$$\Omega^{(j)}_i = \sum_{k=1}^{K} \int_{s_i=-\infty}^{\infty} p_{X_i|S_i}(x_i)\ f^{(j)}_{S_{ik}}(s_i)\ w^{(j)}_k\ ds_i.$$

*Proof.* The posterior is determined by Bayes' theorem,

$$p_{S_i,\mathbf{Z}_i}(s_i, \mathbf{z}_i; \boldsymbol{\theta}^{(j)}, x_i, n_i) = \frac{\Pr\left(X_i, S_i, \mathbf{Z}_i | n_i, \boldsymbol{\theta}\right)}{\sum_{\forall \mathbf{z}_i \in \Omega_Z} \int_{s_i=-\infty}^{\infty} \Pr\left(X_i, S_i, \mathbf{Z}_i | n_i, \boldsymbol{\theta}\right)\ ds_i}$$

in which we can write the joint distribution over the observed $X_i$ and latent variables $S_i$ and $\mathbf{Z}_i$ in terms of the model defined conditional distributions. If we denote the denominator as $\Omega^{(j)}_i$ and apply the shorthand $p^{(j)}_{S_i,\mathbf{Z}_i}$ then the posterior is

$$p^{(j)}_{S_i,\mathbf{Z}_i} = \frac{1}{\Omega^{(j)}_i} \prod_{k=1}^{K} \left(p_{X_i|S_i}\left(x_i; n_i, s_i\right)\ f_{S_i}\left(s_i; \mu^{(j)}_k, \sigma^{2\,(j)}_k\right)\ w^{(j)}_k\right)^{z_{ik}}$$

where

$$\begin{aligned}\Omega^{(j)}_i &= \sum_{\forall \mathbf{z}_i \in \Omega_Z} \int_{s_i=-\infty}^{\infty} \prod_{k=1}^{K} \left(p_{X_i|S_i}\left(x_i; n_i, s_i\right)\ f_{S_i}\left(s_i; \mu^{(j)}_k, \sigma^{2\,(j)}_k\right)\ w^{(j)}_k\right)^{z_{ik}}\ ds_i \\ &= \sum_{k=1}^{K} \int_{s_i=-\infty}^{\infty} p_{X_i|S_i}\left(x_i; n_i, s_i\right)\ f_{S_i}\left(s_i; \mu^{(j)}_k, \sigma^{2\,(j)}_k\right)\ w^{(j)}_k\ ds_i.\end{aligned}$$

Then by applying or shorthand notation for each distribution our result is derived. □

Next, we use Result 1 to derive the E step for the mixture of bln distributions.

**Result 2** (E step of bln mixture)**.** *Under the problem setup,* (9)*, and* (8) *the E step is*

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \sum_{i=1}^{N} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ \sum_{k=1}^{K} z_{ik} \Bigg( \log\left(p_{X_i|S_i}(x_i)\right) - \frac{1}{2}\log(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2}(x_i - \mu_k)^2 + \log(w_k) \Bigg) \right]$$

*Proof.* First, by (4) and the definition of the E step

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \sum_{i=1}^{N} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[\log\left(\Pr\left(X_i, S_i, \mathbf{Z}_i \middle| n_i, \boldsymbol{\theta}\right)\right)\right]$$

where the $\Pr\left(X_i, S_i, \mathbf{Z}_i \middle| n_i, \boldsymbol{\theta}\right)$ may be expressed in terms of the conditional distributions of our model. Writing this expression for th $i^{th}$ sample

$$Q_i(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) = \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ \sum_{k=1}^{K} z_{ik} \log\left(p_{X_i|S_i}(x_i; n_i, s_i) \;\; f_{S_i}\left(s_i; \mu_k, \sigma_k^2\right) \;\; w_k\right) \right].$$

Then the result is derived by applying the log to each term. □

## 2.2 The M step of the mixture of bln distributions

In the M step, the parameters $\boldsymbol{\theta}^{(j+1)}$ are those which maximize the constrained optimization function

$$\boldsymbol{\theta}^{(j+1)} = \underset{\boldsymbol{\theta}, \lambda}{\operatorname{argmax}} \left\{ Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) + \lambda \left(1 - \sum_{k=1}^{K} \underline{w_k}^{(j+1)}\right) \right\}. \tag{10}$$

where $\lambda$ is the Lagrange multiplier associated with the constraint that the $\sum_{k=1}^{K} w_k^{(j+1)} = 1$.

**Result 3** (M step of bln mixture)**.** *The* $(j+1)$ *update to parameters* $\boldsymbol{\theta}$ *are those that maximize the constrained optimization* (10)

$$\begin{aligned}
\hat{w}_k^{(j+1)} &= \frac{N_k}{\underline{N}} \\
\hat{\mu}_k^{(j+1)} &= \frac{1}{N_k} \sum_{i=1}^{N} \frac{w_k^{(j)}}{\Omega_i^{(j)}} \, \mathbb{E}_{S_i} \left[ s_i \; p_{X_i|S_i}(x_i) \;\middle|\; \boldsymbol{\theta}_k^{(j)} \right] \\
\hat{\sigma}_k^{2\;(j+1)} &= \left( \frac{1}{N_k} \sum_{i=1}^{N} \frac{w_k^{(j)}}{\Omega_i^{(j)}} \, \mathbb{E}_{S_i} \left[ \underline{s_i^2} \; p_{X_i|S_i}(x_i) \;\middle|\; \boldsymbol{\theta}_k^{(j)} \right] \right) - \left(\hat{\mu}_k^{(j+1)}\right)^2
\end{aligned}$$

*with*

$$N_k^{(j)} = \sum_{i=1}^{N} \frac{w_k^{(j)}}{\Omega_i^{(j)}} \, \mathbb{E}_{S_i} \left[ \, p_{X_i|S_i}(x_i) \mid \boldsymbol{\theta}_k^{(j)} \right]$$

*Proof.* By using the partial derivatives to find a stationary point of (10) and some simple algebraic operations we have

$$\partial_\lambda : \qquad \sum_{k=1}^{K} w_k = 1 \tag{11}$$

$$\partial_{w_k} : \qquad \lambda w_k = \sum_{i=1}^{N} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ z_{ik} \right] \tag{12}$$

$$\partial_{\mu_k} : \qquad \sum_{i=1}^{N} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ z_{ik} s_i \right] = \mu_k \sum_{i=1}^{N} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ z_{ik} \right] \tag{13}$$

$$\partial_{\sigma_k^2} : \qquad \sum_{i=1}^{N} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ z_{ik} (s_i - \mu_k)^2 \right] = \sigma_k^2 \sum_{i=1}^{N} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ z_{ik} \right] \tag{14}$$

There is a set pattern in the expression above. First, we repeatedly see $\mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ z_{ik} \right]$, which by Result 1 the expectation is

$$\begin{aligned} \mathbb{E}_{S_i, \mathbf{Z}_i \sim p_i^{(j)}} \left[ z_{ik} \right] &= \frac{1}{\Omega_i^{(j)}} \sum_{\forall \mathbf{z}_i \in \Omega_Z} z_{ik} \int_{s_i=-\infty}^{\infty} p_{X_i|S_i}(x_i) \, f_{S_{ik}}^{(j)}(s_i) \, w_k^{(j)} \, ds_i \\ &= \frac{w_k^{(j)}}{\Omega_i^{(j)}} \int_{s_i=-\infty}^{\infty} p_{X_i|S_i}(x_i) \, f_{S_i} \left( s_i; \mu_k^{(j)}, \sigma_k^{2\,(j)} \right) \, ds_i \\ &= \frac{w_k^{(j)}}{\Omega_i^{(j)}} \, \mathbb{E}_{S_i} \left[ \, p_{X_i|S_i}(x_i) \mid \boldsymbol{\theta}_k^{(j)} \right]. \end{aligned} \tag{15}$$

This is the $i^{th}$ term in the definition $N_k^{(j)}$. The intuition being that this is the average number of samples attributed to the $k^{th}$ mixture component under our model with parameters $\boldsymbol{\theta}^{(j)}$.

Given $N_k$ the Lagrange multiplier and $\hat{w}_k^{(j+1)}$ are found by solving equations (11) and (12),

$$\partial_{w_k} : \qquad w_k = \frac{N_k}{\lambda} \tag{16}$$

$$\partial_\lambda : \qquad \sum_{k=1}^{K} w_k = 1 \implies \frac{1}{\lambda} \sum_{k=1}^{K} w_k = 1 \implies \lambda = N.$$

substitution of $\lambda = N$ into (16) completes the derivation.

Next we notice that the LHS of (13) and (14) are the respective moments of $s_i$ of the $k^{th}$ component. By making the respective substitutions and rearranging the equations we derive the expression for $\hat{\mu}_k^{(j+1)}$ and $\hat{\sigma}_k^{2\,(j+1)}$.

□

Results 1, 2, and 3 are all the steps of the EM algorithm required. In what follows we define the algorithm in more detail.

### 2.3 An algorithm

## 3 Application

## References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 2. Springer Series in Statistics New York, 2009.

[3] P. Mohammadi, S. E. Castel, B. B. Cummings, J. Einson, C. Sousa, P. Hoffman, S. Donkervoort, Z. Jiang, P. Mohassel, A. R. Foley, H. E. Wheeler, H. K. Im, C. G. Bonnemann, D. G. MacArthur, and T. Lappalainen. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science*, 366(6463):351–356, 2019.