EM Algorithm applied to Shim and Stephens (2015)

Brendan Law

11 January, 2019

A document showing my formal(-ish) derivation for the Expectation Maximisation (EM) algorithm as applied to Shim and Stephens (2015). The task here is to parameterise the probability that our wavelet coefficients come from either one of two latent states, each with their own conditional densities. A lot of quantities are divided by a constant to convert them into Bayes Factors, for convenience, as the closed form of the underlying conditional densities are less convenient to work with.

1 Preamble

We are working with data which has been transformed into wavelet coefficients (WC)'s across various states and scales, as well as dependent variables (eg: SNPs of interest) which we are trying to regress onto the WCs to perform association analysis across individuals, and eventually, across groups.

- Individuals, $i \in \{1, \dots, n\}$
- Scales, $s \in \{1, \dots, J\}$
- Locations, $l \in \{1, \dots, L_s\}$ (as each scale has a different number of locations, ranging from 1 at the coarsest scale to many at the finest scale?)
- For each scale and location, s, l, a vector of WCs: $\boldsymbol{y}_{sl} = (y_{sl}^1, \dots, y_{sl}^n)$, and $\boldsymbol{Y} = (\boldsymbol{y}_{11}, \dots, \boldsymbol{y}_{JL_J})$
- A vector of binary indicator variables used to indicate whether $\boldsymbol{y}_{sl}$ is associated with g (1 for association): $\boldsymbol{\gamma} = (\gamma_{11}, \dots, \gamma_{JL_J})$, where $\gamma_{sl} \in \{0, 1\}$
  - This is the latent state variable in this setting
- Parameter set, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_J)$
  - These are the free parameters we will be parameterising through the EM algorithm
- Dependent variables, $\boldsymbol{g} = (g^1, \dots, g^n)$, which represents a vector of data points, one from each individual - in this paper, the genotype data (number of copies of the minor allele) for individual i at a single SNP of interest.
- Hierarchical model:
  - $y_{sl}^i = \mu_{sl} + \gamma_{sl}\beta_{sl}g^i + \epsilon_{sl}^i$ with $\epsilon_{sl}^i \sim \mathcal{N}(0, \sigma_{sl}^2)$, implying:
    - $P(y_{sl}^i \mid \mu_{sl}, \gamma_{sl} = 0, \beta_{sl}, g^i, \sigma_{sl}^2) \sim \mathcal{N}(\mu_{sl}, \sigma_{sl}^2)$, and
    - $P(y_{sl}^i \mid \mu_{sl}, \gamma_{sl} = 1, \beta_{sl}, g^i, \sigma_{sl}^2) \sim \mathcal{N}(\mu_{sl} + \beta_{sl}g^i, \sigma_{sl}^2)$
  - $P(\gamma_{sl} = 1 \mid \boldsymbol{\pi}) = \pi_s$ for each scale s, across all locations, l

Some extra notation and assumptions regarding the model:

- Note that given their state, γ_{sl}, WCs are conditionally independent across scales and locations
- Due to the Bayesian setting of this model, the π_s's are hyperparameters, not random variables
- $P(\gamma_{sl} = 1 \mid \boldsymbol{\pi}) = \pi_s$ and $P(\gamma_{sl} = 0 \mid \boldsymbol{\pi}) = 1 - \pi_s$
- $\pi_s = 0 \Rightarrow \gamma_{sl} = 0$ and consequently $\boldsymbol{\pi} \equiv 0 \Rightarrow \boldsymbol{\gamma} \equiv 0$

- The Bayes Factor is used extensively in the paper to measure the support for $\gamma_{sl} = 1$, for a specific s, l, across all individuals i. It is easier to compute (has a closed form) thanks to the models and priors from Servin and Stephens (2007). It is denoted as such:

$$\text{BF}_{sl}(y, g) := \frac{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 1, \boldsymbol{g})}{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g})}$$

2 Complete log likelihood derivation

$$\begin{aligned}
P(\boldsymbol{Y}, \gamma \mid \boldsymbol{g}, \boldsymbol{\pi}) &= P(\boldsymbol{Y} \mid \gamma, \boldsymbol{g}, \boldsymbol{\pi}) P(\gamma \mid \boldsymbol{g}, \boldsymbol{\pi}) \\
&= P(\boldsymbol{Y} \mid \gamma, \boldsymbol{g}, \boldsymbol{\pi}) \prod_{s,l} P(\gamma_{sl} \mid \boldsymbol{\pi}) && \text{(independence of } \pi_s \text{ across scales, } \gamma \text{ independent of } \boldsymbol{g}) \\
&= P((\boldsymbol{y}_{11}, \ldots, \boldsymbol{y}_{JL}) \mid \gamma, \boldsymbol{g}) \prod_{s,l} P(\gamma_{sl} \mid \boldsymbol{\pi}) && \text{(independence of } \boldsymbol{y} \text{ of } \boldsymbol{\pi}) \\
&= \prod_{s,l} \left[P(\boldsymbol{y}_{sl} \mid \gamma_{sl}, \boldsymbol{g}) P(\gamma_{sl} \mid \pi_s) \right] && \text{(independence of } \boldsymbol{y}_{sl}\text{'s conditional on own state)} \\
&= \prod_{s,l} \prod_{k=0}^{1} \left[P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k, \boldsymbol{g}) P(\gamma_{sl} = k \mid \pi_s) \right]^{\mathbb{1}\{\gamma_{sl} = k\}} \\
&= P(\boldsymbol{Y} \mid \boldsymbol{\gamma} \equiv 0, \boldsymbol{g}) \prod_{s,l} \prod_{k=0}^{1} \frac{\left[P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k, \boldsymbol{g}) P(\gamma_{sl} = k \mid \pi_s) \right]^{\mathbb{1}\{\gamma_{sl} = k\}}}{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g})}
\end{aligned}$$

with the last step due to:

$$P(\boldsymbol{Y} \mid \boldsymbol{\gamma} \equiv 0, \boldsymbol{g}) = \prod_{s,l} \prod_{k=0}^{1} P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g})$$

Therefore, the complete log likelihood:

$$\begin{aligned}
\log L(\boldsymbol{\pi}; \boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{g}) = \log P(\boldsymbol{Y} \mid \boldsymbol{\gamma} \equiv 0, \boldsymbol{g}) + \sum_{s,l} \Bigg[\mathbb{1}\{\gamma_{sl} = 0\} \Bigg(\log \frac{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g})}{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g})} + \log(1 - \pi_s) \Bigg) \ldots \\
+ \mathbb{1}\{\gamma_{sl} = 1\} \Bigg(\log \frac{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 1, \boldsymbol{g})}{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g})} + \log \pi_s \Bigg) \Bigg] \\
= \log P(\boldsymbol{Y} \mid \boldsymbol{\gamma} \equiv 0, \boldsymbol{g}) + \sum_{s,l} \Big[\mathbb{1}\{\gamma_{sl} = 0\} \big(\log(1 - \pi_s) \big) + \mathbb{1}\{\gamma_{sl} = 1\} \big(\log \text{BF}_{sl}(y, g) + \log \pi_s \big) \Big]
\end{aligned}$$

Note that, as always, the following remains a random variable representing the unknown state of the γ_{sl} variable:

$$\mathbb{1}\{\gamma_{sl} = k\} = \begin{cases} 1 & \gamma_{sl} = k \\ 0 & \gamma_{sl} \neq k \end{cases}$$

3 EM algorithm

We will now compute the MLE of the parameters in $\boldsymbol{\pi}$ by iterating through the EM algorithm and updating $\boldsymbol{\pi}$ at the end of each step.

$$\begin{aligned}
Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}^{(\text{t})}) &= \mathbb{E}_{(\boldsymbol{\gamma} \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\text{t})})} [\log L(\boldsymbol{\pi}; \mathbf{Y}, \boldsymbol{\gamma}, \mathbf{g})] && \text{(finding unknowns in this statement is the E-step)} \\
\boldsymbol{\pi}^{(\text{t}+1)} &:= \underset{\boldsymbol{\pi}}{\operatorname{argmax}}\, Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}^{(\text{t})}) && \text{(this is the M-step)}
\end{aligned}$$

4 E-step: derivation

$$
\begin{aligned}
Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}^{(\mathrm{t})}) &= \mathbb{E}_{(\boldsymbol{\gamma} \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}[\log L(\boldsymbol{\pi}; \boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{g})] \\
&= \mathbb{E}_{(\boldsymbol{\gamma} \mid \boldsymbol{Y}, \boldsymbol{g} \boldsymbol{\pi}^{(\mathrm{t})})}\Big[\log P(\boldsymbol{Y} \mid \boldsymbol{\gamma} \equiv 0, \boldsymbol{g}) + \sum_{s,l} \big[\mathbb{1}\{\gamma_{sl} = 0\}\big(\log(1-\pi_s)\big) \ldots \\
&+ \mathbb{1}\{\gamma_{sl} = 1\}\big(\log \mathrm{BF}_{sl}(y,g) + \log \pi_s\big)\big]\Big] \\
&= \log P(\boldsymbol{Y} \mid \boldsymbol{\gamma} \equiv 0, \boldsymbol{g}) + \sum_{i=1}^{n} \sum_{s,l} \big[P(\gamma_{sl} = 0 \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})\big(\log(1-\pi_s)\big) + P(\gamma_{sl} = 1 \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})\big(\log \mathrm{BF}_{sl}(y,g) + \log \pi_s\big)\big]
\end{aligned}
$$

Now we can evaluate each of the two conditional probability statements around γ_{sl}:

$$
\begin{aligned}
P(\gamma_{sl} = k \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) &= P(\gamma_{sl} = k \mid \boldsymbol{y}_{sl}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) \qquad \text{(independent of WCs from other scales, locations)} \\
&= \frac{P(\gamma_{sl} = k, \boldsymbol{y}_{sl} \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}{\sum_{k'} P(\gamma_{sl} = k', \boldsymbol{y}_{sl} \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})} \\
&= \frac{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) P(\gamma_{sl} = k \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}{\sum_{k'} P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k', \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) P(\gamma_{sl} = k' \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})} \\
&= \frac{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) P(\gamma_{sl} = k \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}{\sum_{k'} P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k', \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) P(\gamma_{sl} = k' \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})} \\
&= \frac{\dfrac{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) P(\gamma_{sl} = k \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}}{\dfrac{\sum_{k'} P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = k', \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) P(\gamma_{sl} = k' \mid \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}{P(\boldsymbol{y}_{sl} \mid \gamma_{sl} = 0, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})}} \\
&\qquad \text{(Divide both sides by a constant to convert into Bayes Factors)}
\end{aligned}
$$

$$
\begin{aligned}
\therefore P(\gamma_{sl} = 1 \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) &= \frac{\mathrm{BF}_{sl}(y,g)\pi_s^{(\mathrm{t})}}{BF_{sl}(y,g)\pi_s^{(\mathrm{t})} + (1-\pi_s^{(\mathrm{t})})}, \text{ and} \\
\therefore P(\gamma_{sl} = 0 \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})}) &= \frac{1-\pi_s^{(\mathrm{t})}}{BF_{sl}(y,g)\pi_s^{(\mathrm{t})} + (1-\pi_s^{(\mathrm{t})})}
\end{aligned}
$$

5 M-step: maximisation

To simplify the notation, denote:

$$A_{sl,k}^{(\mathrm{t})} := P(\gamma_{sl} = k \mid \boldsymbol{Y}, \boldsymbol{g}, \boldsymbol{\pi}^{(\mathrm{t})})$$

We have that:

$$\boldsymbol{\pi}^{(\mathrm{t}+1)} := \underset{\boldsymbol{\pi}}{\operatorname{argmax}}\, Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}^{(\mathrm{t})})$$

where:

$$Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}^{(\mathrm{t})}) \propto \sum_{s,l} \big[A_{sl,0}^{(\mathrm{t})} \log(1-\pi_s) + A_{sl,1}^{(\mathrm{t})} (\log BF_{sl}(y,g) + \log \pi_s) \big]$$

(proportionality up to the constant where $\boldsymbol{\gamma} \equiv 0$ up front)

Hence, finding the $\boldsymbol{\pi}$ which maximises the term in the sum on the right hand side above will yield an equivalent result as the one desired.

For each $s \in \{1, \ldots, J\}$,

$$
\begin{aligned}
\frac{\delta Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}^{(\mathrm{t})})}{\delta \pi_s} &= \sum_{l=1}^{L_s} \Big(- \frac{A_{sl,0}^{(\mathrm{t})}}{1-\pi_s} + \frac{A_{sl,1}^{(\mathrm{t})}}{\pi_s} \Big) \\
&= \frac{\sum_{l=1}^{L_s} (-A_{sl,0}^{(\mathrm{t})}\pi_s + A_{sl,1}^{(\mathrm{t})}(1-\pi_s))}{\pi_s(1-\pi_s)}
\end{aligned}
$$

Setting this equal to zero,

$$
\begin{aligned}
0 &= \frac{\sum_{l=1}^{L_s} (-A_{sl,0}^{(\mathrm{t})}\pi_s^{(\mathrm{t+1})} + A_{sl,1}^{(\mathrm{t})}(1-\pi_s^{(\mathrm{t+1})}))}{\pi_s^{(\mathrm{t+1})}(1-\pi_s^{(\mathrm{t+1})})} \\
\Rightarrow \sum_{l=1}^{L_s} A_{sl,0}^{(\mathrm{t})}\pi_s^{(\mathrm{t+1})} &= \sum_{l=1}^{L_s} A_{sl,1}^{(\mathrm{t})}(1-\pi_s^{(\mathrm{t+1})}) \\
\pi_s^{(\mathrm{t+1})} \sum_{l=1}^{L_s} A_{sl,0}^{(\mathrm{t})} &= (1-\pi_s^{(\mathrm{t+1})}) \sum_{l=1}^{L_s} A_{sl,1}^{(\mathrm{t})} \\
(\pi_s^{(\mathrm{t+1})}) \sum_{l=1}^{L_s} (A_{sl,0}^{(\mathrm{t})} + A_{sl,1}^{(\mathrm{t})}) &= \sum_{l=1}^{L_s} A_{sl,1}^{(\mathrm{t})} \\
\Rightarrow \pi_s^{(\mathrm{t+1})} &= \frac{\sum_{l=1}^{L_s} A_{sl,1}^{(\mathrm{t})}}{\sum_{l=1}^{L_s} (A_{sl,0}^{(\mathrm{t})} + A_{sl,1}^{(\mathrm{t})})} = \frac{\sum_{l=1}^{L_s} A_{sl,1}^{(\mathrm{t})}}{L_s}
\end{aligned}
$$

for each scale, s. The simplification in the last line follows as:

$$
\sum_{l=1}^{L_s} (A_{sl,0}^{(\mathrm{t})} + A_{sl,1}^{(\mathrm{t})}) = \sum_{l=1}^{L_s} 1 = L_s
$$

References

B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114, 2007.

H. Shim and M. Stephens. Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *The annals of applied statistics*, 9(2):655, 2015.