Shim and Stephens (2015) with HMT - Beta propreties

Brendan Law

April 2, 2019

Just some properties of the β estimator under the HMT version of [**?**]. Refer to the supporting documents (the ones titled 'EM_algo_Shim_Stephens_...', etc) for some of the workings. Also, a lot of the closed form workings and derivations were done in the supplementary material to [**?**].

1 Properties of the base case - no HMT

1.1 β Coefficient transform

This section is all about the β coefficient transform from the wavelet space to the data space. This is important, as the β coefficient is estimated in the wavelet space, and then transformed into the data space in order to provide interpretable estimates of effect sizes at different base locations along the genome.

It's worth noting that there are also some practicalities involved with its estimation:

- Even though quantile regression is performed (to make the model less sensitive to extreme values in the dependent variable), this is not performed for the purposes of estimating effect sizes.
- Otherwise, the data $\rightarrow$ wavelet $\rightarrow$ inverse wavelet $\rightarrow$ data space transform would be complicated by the fact that we would need to go data $\rightarrow$ wavelet $\rightarrow$ quantile transform $\rightarrow$ inverse quantile transform $\rightarrow$ inverse wavelet $\rightarrow$ data, which is infeasible.
- Hence, estimation here occurs without a quantile transforms, and the translation of estimates from the wavelet back to the data space is a simple process due to the fact that the wavelet transform is *linear* and *one-to-one*.

1.2 Sketching out the wavelet linear transform

To represent the model in the wavelet space, we have that:

$$\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{\beta} g + \boldsymbol{E}$$

So we have that:

$$\begin{bmatrix} y^1_{1,1} & \dots & y^n_{1,1} \\ \vdots & \ddots & \vdots \\ y^1_{J,L_J} & \dots & y^n_{J,L_J} \end{bmatrix} = \begin{bmatrix} \mu^1_{1,1} & \dots & \mu^n_{1,1} \\ \vdots & \ddots & \vdots \\ \mu^1_{J,L_J} & \dots & \mu^n_{J,L_J} \end{bmatrix} + \begin{bmatrix} \beta_{1,1} \\ \vdots \\ \beta_{J,L_J} \end{bmatrix} \begin{bmatrix} g^1 & \dots & g^n \end{bmatrix} + \begin{bmatrix} \epsilon^1_{1,1} & \dots & \epsilon^n_{1,1} \\ \vdots & \ddots & \vdots \\ \epsilon^1_{J,L_J} & \dots & \epsilon^n_{J,L_J} \end{bmatrix}$$

Where the $\boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{E}$ matrices are all of dimension $B \times n$. This represents the B base pair locations (which also correspond to the number of wavelet coefficients the transform creates – $B/2$ at the lowest/finest level, all the way to 1 at the highest/coarsest level), and the n individuals in the dataset. The $\boldsymbol{\beta}$ vector is $B \times 1$, and g is $1 \times n$, with $\boldsymbol{\beta} g$ becoming a $B \times n$ matrix:

$$\begin{bmatrix} \beta_{1,1} g^1 & \dots & \beta_{1,1} g^n \\ \vdots & \ddots & \vdots \\ \beta_{J,L_J} g^1 & \dots & \beta_{J,L_J} g^n \end{bmatrix}$$

Now, to represent the linear transform from the data space to the wavelet space, we note that:

$$\begin{aligned} y &= Wd \\ \Rightarrow d &= W^{-1}y \end{aligned}$$

$$\begin{aligned} \boldsymbol{D} &= \boldsymbol{W}^{-1}\boldsymbol{Y} \\ \Rightarrow \boldsymbol{D} &= \boldsymbol{W}^{-1}\boldsymbol{M} + \boldsymbol{W}^{-1}\boldsymbol{\beta}g + \boldsymbol{W}^{-1}\boldsymbol{E}, \\ \text{where } \boldsymbol{W}^{-1}\boldsymbol{\beta} &:= \boldsymbol{\alpha} \end{aligned}$$

The original data, $\boldsymbol{D}$, is a $B \times n$ matrix as follows:

$$\boldsymbol{D} = \begin{bmatrix} d_1^1 & \dots & d_1^n \\ \vdots & \ddots & \vdots \\ d_B^1 & \dots & d_B^n \end{bmatrix}$$

The wavelet transform represents a linear transform from this $B \times n$ matrix, $\boldsymbol{D}$ into a $B \times n$ matrix of coefficients, $\boldsymbol{Y}$, and hence can be represented as a $B \times B$ matrix:

$$\boldsymbol{W} = \begin{bmatrix} w_{1,(1,1)} & \dots & w_{B,(1,1)} \\ \vdots & \ddots & \vdots \\ w_{1,(J,L_J)} & \dots & w_{B,(J,L_J)} \end{bmatrix}$$

where w_{ij} represents the 'amount' by which the data point at base i contributes to the wavelet coefficient indexed by j, where j represents a scale-location pair, (s, l). This is for any given individual. Furthermore, we have that $\boldsymbol{W}$ is orthogonal (from suppl mtl) hence $\boldsymbol{W}^T = \boldsymbol{W}^{-1}$

Therefore for the matrix $\boldsymbol{\alpha}$:

$$\begin{aligned} \boldsymbol{\alpha} &= \boldsymbol{W}^{-1}\boldsymbol{\beta} \\ &= \boldsymbol{W}^T\boldsymbol{\beta} \\ \Rightarrow \boldsymbol{\alpha} &= \begin{bmatrix} w_{1,(1,1)} & \dots & w_{1,(J,L_J)} \\ \vdots & \ddots & \vdots \\ w_{B,(1,1)} & \dots & w_{B,(J,L_J)} \end{bmatrix} \begin{bmatrix} \beta_{1,1} \\ \vdots \\ \beta_{J,L_J} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{s,l} w_{1,(s,l)}\beta_{s,l} \\ \vdots \\ \sum_{s,l} w_{B,(s,l)}\beta_{s,l} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_B \end{bmatrix} \end{aligned}$$

which yields us the α_i's – the estimated data space effect at each base location i.

1.3 Estimating statistical properties of α

We're also interested in some statistical properties of α, such as its mean and variance. This requires us to perform some operations on the posterior distribution of β, (i.e. $P(\beta \mid \text{Data, hyper-params})$, with all other parameters integrated out).In the no-HMT case, we have from the supplementary material, a closed form of the posterior of β, and thus we can derive its mean and variance properties, and use inverse wavelet transforms to derive means and variances of α.

1.3.1 Posterior distribution

From the no-HMT supplementary material, we have a posterior of $\beta_{s,l} = P(\beta_{s,l} \mid y_{s,l}, g)$ as follows:

$$\begin{aligned} P(\beta_{s,l} \mid y_{s,l}, g, \boldsymbol{\pi}) &= P(\beta_{s,l} \mid y_{s,l}, g, \boldsymbol{\pi}, \gamma_{s,l} = 1)P(\gamma_{s,l} = 1 \mid y_{s,l}, g, \boldsymbol{\pi}) + P(\beta_{s,l} \mid y_{s,l}, g, \boldsymbol{\pi}, \gamma_{s,l} = 0)P(\gamma_{s,l} = 0 \mid y_{s,l}, g, \boldsymbol{\pi}) \\ &= P(\beta_{s,l} \mid y_{s,l}, g, \boldsymbol{\pi}, \gamma_{s,l} = 1)\phi_{s,l} + P(\beta_{s,l} \mid y_{s,l}, g, \boldsymbol{\pi}, \gamma_{s,l} = 0)(1 - \phi_{s,l}) \end{aligned}$$

where:

- $\boldsymbol{\pi}$ is a vector of hyperparameters (in this case, the proportion of non-zero wavelet coefficients at a given scale)
- $P(\beta_{s,l} \mid y_{s,l}, g, \boldsymbol{\pi}, \gamma_{s,l} = 0)$ is a point mass at 0 (wp 1)
-

$$\begin{aligned}
\phi_{s,l} &= P(\gamma_{s,l} = 1 \mid y_{s,l}, g, \boldsymbol{\pi}) \\
&= \frac{P(y_{s,l} \mid g, \gamma_{s,l} = 1) P(\gamma_{s,l} = 1 \mid \boldsymbol{\pi})}{P(y_{s,l} \mid g, \gamma_{s,l} = 0) P(\gamma_{s,l} = 0 \mid \boldsymbol{\pi}) + P(y_{s,l} \mid g, \gamma_{s,l} = 1) P(\gamma_{s,l} = 1 \mid \boldsymbol{\pi})} \\
&= \frac{BF_{s,l}(y, g)\hat{\pi_s}}{BF_{s,l}(y, g)\hat{\pi_s} + (1 - \hat{\pi_s})}, \text{ obtaining the BFs by dividing through by the } \gamma_{s,l} = 0 \text{ case}
\end{aligned}$$

- Posterior of $\beta_{s,l}$ is a mixture (with mixing probabilities $\phi_{s,l}$) of a three-parameter t-distribution, and a point mass at 0.

1.3.2 Mean

The desired quantity is $\mathbb{E}[\alpha_b]$. Based on the 'Ash' material and 'Wavelets - Supplementary Material'. We start by finding $\mathbb{E}[\beta_{s,l}]$ and then transforming it back into the data space.

$$\begin{aligned}
\mathbb{E}[\beta_{s,l}] &= \mathbb{E}[\mathbb{E}[\beta_{s,l} \mid \gamma_{s,l}]] \\
&= \mathbb{E}[P(\gamma_{s,l} = 1)\mathbb{E}[\beta_{s,l} \mid \gamma_{s,l} = 1]] + \mathbb{E}[P(\gamma_{s,l} = 0)\mathbb{E}[\beta_{s,l} \mid \gamma_{s,l} = 0]] \\
&= \mathbb{E}[\phi_{s,l} a_{s,l} + (1 - \phi_{s,l})0] \\
&= \phi_{s,l} a_{s,l}
\end{aligned}$$

where $a_{s,l}$ represents the mean of a three-parameter t-distribution, and in this case, corresponds to $\boldsymbol{B_2^*}$, the 2nd element of $\boldsymbol{B} = (\boldsymbol{D}^{-1} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$, evaluated in the limit (of some of the hyper-parameters, κ^a, κ^b and σ_μ^2).

$$\begin{aligned}
\therefore \mathbb{E}[\alpha_b] &= \mathbb{E}[\sum_{s,l} w_{b,(s,l)} \beta_{s,l}] \\
&= \sum_{s,l} w_{b,(s,l)} \mathbb{E}[\beta_{s,l}] && \text{(linearity of expectations)} \\
&= \sum_{s,l} \phi_{s,l} a_{s,l} w_{b,(s,l)}
\end{aligned}$$

1.3.3 Variance

$$
\begin{aligned}
Var(\beta_{s,l}) &= \mathbb{E}[\beta_{s,l}^2] - \mathbb{E}[\beta_{s,l}]^2 \\
\mathbb{E}[\beta_{s,l}^2] &= \mathbb{E}[\mathbb{E}[\beta_{s,l}^2 \mid \gamma_{s,l}]] \\
\mathbb{E}[\beta_{s,l}^2 \mid \gamma_{s,l}] &= Var[\beta_{s,l} \mid \gamma_{s,l}] + \mathbb{E}[\beta_{s,l} \mid \gamma_{s,l}]^2 \\
\therefore \mathbb{E}[\mathbb{E}[\beta_{s,l}^2 \mid \gamma_{s,l}]] &= \mathbb{E}[\phi_{s,l}(Var[\beta_{s,l} \mid \gamma_{s,l} = 1] + \mathbb{E}[\beta_{s,l} \mid \gamma_{s,l} = 1]^2)] \\
&= \mathbb{E}[\phi_{s,l}(\frac{b_{s,l}v_{s,l}}{v_{s,l}-2} + a_{s,l}^2)] \\
&= \phi_{s,l}(\frac{b_{s,l}v_{s,l}}{v_{s,l}-2} + a_{s,l}^2) \\
\therefore Var(\beta_{s,l}) &= \phi_{s,l}(\frac{b_{s,l}v_{s,l}}{v_{s,l}-2} + a_{s,l}^2) - \phi_{s,l}^2 a_{s,l}^2 = \phi_{s,l}\left(\frac{b_{s,l}v_{s,l}}{v_{s,l}-2} + a_{s,l}^2 - \phi_{s,l}a_{s,l}^2\right) \\
\therefore Var(\alpha_b) &= Var(\sum_{s,l} w_{b,(s,l)}\beta_{s,l}) \\
&= \sum_{s,l} w_{b,(s,l)}^2 Var(\beta_{s,l}) \qquad \text{(linearity of variance of independent terms)} \\
&= \sum_{s,l} w_{b,(s,l)}^2 \phi_{s,l}\left(\frac{b_{s,l}v_{s,l}}{v_{s,l}-2} + a_{s,l}^2 - \phi_{s,l}a_{s,l}^2\right)
\end{aligned}
$$

We note here that $\beta_{s,l}$'s are (conditionally) independent, given the data and hyperparameters (notation above has been very sloppy – all properties of expectations and variances are of the posterior of the beta, conditional on data and hyperparameters. It has been omitted from the above notation due to brevity. Hence $Var(\beta_{s,l}) = Var(\beta_{s,l} \mid$ Data, hyperparams), which are conditionally independent of each other). An explanation of why the β posteriors are conditionally independent can be found in 'wavelets_and_models.pdf'.

2 Properties of the base case - with HMT

2.1 Estimating statistical properties of α

All of the above still applies, except that we now calculate $\phi_{s,l}$ differently, as $\phi_{s,l}$ depends on the probabilities of the other $\gamma_{s,l}$'s as seen through the upward-downward algorithm required to solve this case.

2.1.1 Posterior distribution

$$
\begin{aligned}
\phi_{s,l} &= P(\gamma_{s,l} = 1 \mid \boldsymbol{Y}, g, \boldsymbol{\pi}) \\
&= \frac{P(\gamma_{s,l} = 1, \boldsymbol{Y} \mid \boldsymbol{\pi})}{P(\boldsymbol{Y} \mid \boldsymbol{\pi})} \\
&= \frac{\beta_{s,l}(1)\alpha_{s,l}(1)}{\sum_{m=0}^{1} \beta_{s,l}(m)\alpha_{s,l}(m)}
\end{aligned}
$$

where $\beta_{s,l}(m) := P(\mathcal{T}_{s,l} \mid \gamma_{s,l} = m, \boldsymbol{\pi})$ and $\alpha_{s,l}(m) := P(\gamma_{s,l} = m, \mathcal{T}_{1\setminus s,l} \mid \boldsymbol{\pi})$, as found in the other documentation ('EM_algo_HMT.pdf'). These are computed as part of the EM algorithm, and will drop out of it nicely (all 'scaled' to utilise the closed form of the Bayes Factor, but still with the desired result).

2.1.2 Means and variances

Can't remember how the means work. Don't believe the posterior still has the same 3-parameter t-distribution as the β priors are no longer independent and factorisable (HMT prior imposes dependencies between them). Although we do suspect that $P(\beta_{s,l} \mid \boldsymbol{Y}, g, \gamma_{s,l} = m) = P(\beta_{s,l} \mid y_{s,l}, g, \gamma_{s,l} = m)$. We'll have to calculate this one later. How did we get to this conclusion, and why is it significant? Does it mean that we postulate our joint posterior factorises into the

product of individual 3-parameter t-distributions?

A note about the variances:

- Dependence (even conditionally) between $\beta_{s,l}$'s, meaning that the linearity and additivity of variances doesn't apply as there will be covariance cross-terms to be calculated.
- Hence we'll just simulate these properties in this case.
- It's also likely we'll end up simulating the means too.

2.1.3 Simulation methodology