Continuous Buhmbox - Update 11/27

Alexandre Lamy - all2187@columbia.edu

November 27, 2017

Notation

As a quick recall, remember that

$$S_{BUHMBOX} = \frac{\sum_{i<j} w_{ij} y_{ij}}{\sqrt{\sum_{i<j} w_{ij}^2}}$$

And,

$$Y = \alpha(R - R')$$

Where w_{ij} are weights and, in the standard BB case, R is the sample correlation matrix in cases and R' in controls, and in the continuous BB case, R is the weighted sample correlation matrix (using some weights ω) and R' is the unweighted sample correlation matrix (or equivalently the weighted sample correlation matrix with all equal weights). α is simply a normalization factor to assure that all y_{ij} are $\mathcal{N}(0,1)$ in the independent case.

We denote r_ω to be the weighted sample correlation between two independent random variables (X and Y) calculated on N observations(x_i and y_i's), each with weight ω_i and with $\sum_i \omega_i = 1$ (and $\forall i, \omega_i \geq 0$). And r to be r_ω where ω just gives equal weights to all observations (equivalently the unweighted sample correlation), we sometimes also denote this as $r_{1/N}$ for obvious reasons.

Then note that in the independent population case, each y_{ij} is distributed like an $r_\omega - r$ for the continuous BB case.

Theorems and empirically verified conjectures

I claim to have a proof for everything I label as a "Theorem". I have empirically verified conjecture but do not yet have proofs.

Theorem 1. *Suppose that all w_{ij} are independent and prefixed weights, and assume all y_{ij} are $\mathcal{N}(0,1)$. Then $S_{BUHMBOX}$ is distributed as a $\mathcal{N}(0,1)$.*

Remark 1. This theorem explains why proving that the y_{ij}'s are $\mathcal{N}(0,1)$ is important. However, the strong (and possibly wrong) assumption of the independence of the w_{ij}'s might be the cause of the main current challenge (see remaining problems/challenges section). The rest of the theorems involve trying to prove that the y_{ij}'s (or equivalently r_ω for arbitrary ω's) are distributed as $\mathcal{N}(0,1)$ under independence.

Theorem 2. $\mathbb{E}[r_\omega] = 0$ *for any valid ω (by valid we mean $\sum_i \omega_i = 1$ and $\forall i, \omega_i \geq 0$).*

Remark 2. This along with the linearity of expectation and the fact that r is just a special case of r_ω, proves that $\mathbb{E}[y_{ij}] = 0$ in both the standard and continuous versions of Buhmbox.

Theorem 3. $\mathrm{Var}[r] = \mathbb{E}[r^2] \approx \frac{1}{N}$ *for large* N.

Remark 3. This comes trivially from the fact that, for normal variables X and Y, the sample correlation coefficient has a variance of approximately

$$\frac{1-\rho^2}{N-2}$$

which is approximately $\frac{1}{N}$ in our case. That expression from the variance comes from long papers that use the Fischer Transformation. However, we successfully proved this from first principles, without even relying on the assumption of normality. When doing this we proved that the variance was exactly $\frac{1}{N-1}$

Conjecture 1. $\mathrm{Var}[r_\omega] = \mathbb{E}[r_\omega^2] \approx \sum_i \omega_i^2$

Remark 4. Note that this is a generalization of theorem 3. However, although being very close, we have not been able to prove this result using similar techniques (I get stuck with a really long expression, I can't simplify). This may be due to the above result being an approximation rather than an exact result. However, it is empirically an **excellent** approximation.

Conjecture 2. $\mathrm{Var}[r_\alpha - r_\beta] = \mathbb{E}[(r_\alpha - r_\beta)^2] \approx \sum_i (\alpha_i - \beta_i)^2$.
Or, equivalently $\mathbb{E}[r_\alpha r_\beta] \approx \sum_i \alpha_i \beta_i$.

Remark 5. First note that this is again a generalization of both theorem 3 and conjecture 1. Furthermore it was verified to be empirically extremely accurate and matches exactly with the formulation of standard Buhmbox. Indeed note that the standard Buhmbox just calculates y_{ij} as $r_\alpha - r_\beta$ where α gives weights of $\frac{1}{N}$ for all cases and weights of 0 for all controls, and β gives weights of $\frac{1}{N'}$ to all controls and weights of 0 to all cases. In which case this result gives us that:

$$\mathrm{Var}[y_{ij}] = \mathbb{E}[(r_\alpha - r_\beta)^2] = \sum_i (\alpha_i - \beta_i)^2 = \sum_{cases} \left(\frac{1}{N} - 0\right)^2 + \sum_{controls} \left(0 - \frac{1}{N'}\right)^2 = N\frac{1}{N^2} + N'\frac{1}{N'^2} = \frac{1}{N} + \frac{1}{N'}$$

Which exacty matches their normalization factor of $\sqrt{\frac{NN'}{N+N'}} = \sqrt{\frac{1}{\frac{1}{N}+\frac{1}{N'}}}$.

As another quick note, equivalence of the two results simply comes from the fact that

$$\mathrm{Var}[r_\alpha - r_\beta] = \mathrm{Var}[r_\alpha] + \mathrm{Var}[r_\beta] - 2\mathrm{Cov}[r_\alpha, r_\beta] = \mathrm{Var}[r_\alpha] + \mathrm{Var}[r_\beta] - 2\,\mathbb{E}[r_\alpha r_\beta]$$

And a teeny bit of algebra.

Proof wise, the first result seems almost impossible to prove directly. The second seems easier, but proving conjecture 1 would be a first step. There is also an added complexity in that the denominator has a square root that I can't get rid of, and taking the expectation of a square root seems particularly difficult.

Proofs

I have a proof for all three theorems and a good long attempt for Conjecture 1 in my notebook. I have not had time to type them up but will do so in the coming days (the proof for theorem 3 is quite long...).

Remaining Problems and Challenges

I obviously still need proofs for conjectures 1 and 2. However a few other problems also present themselves.

(1) To prove any of the theorems (or even start proving them or do anything), I had to use the following equality:

$$\mathbb{E}\left[\frac{A}{B}\right] = \frac{\mathbb{E}[A]}{\mathbb{E}[B]}$$

Where A is more or less the squared covariance and B is the product of the two variances (since $r = \frac{\operatorname{Cov}[X,Y]}{\sqrt{\operatorname{Var} X \operatorname{Var} Y}}$ and so $r^2 = \frac{\operatorname{Cov}[X,Y]^2}{\operatorname{Var} X \operatorname{Var} Y}$).

Professor Pe'er suggested doing this at some point I think, and it seems to give correct results. However, it is definitely not true in general. Even if A and B are completely independent we have $\mathbb{E}\left[\frac{A}{B}\right] = \mathbb{E}[A]\,\mathbb{E}[\frac{1}{B}]$, but I can see no reason why we would have $\mathbb{E}[\frac{1}{B}] = \frac{1}{\mathbb{E}[B]}$ (since this is not true for almost any RV).

(2) Even though I do not yet have all the proofs, I implemented continuous BB based on the conjecture 2 result. This correctly gave me a variance of 1 for the BB score under independence. However, the expected value of the BB score was clearly not 0. I looked at the values of various y_{ij}'s and these were all $\mathcal{N}(0,1)$ as expected. So the error could only come from some sort of positive covariance/correlation between the weights w_{ij} and the values of y_{ij} (rendering the result of theorem 1 inapplicable). This seems strongly problematic. I have not spent much time thinking about the problem but am unsure how to solve it.