Supplementary materials for "False discovery rate control for grouped or discretely supported p-values with application to a neuroimaging study"

Hien D. Nguyen*, Yohan Yee, Geoffrey J. McLachlan
and Jason P. Lerch

December 2019

The material contained herein is supplementary to the article named in the title and published in SORT-Statistics and Operations Research Transactions Volume 43(2).

* HDN is at the Department of Mathematics and Statistics, La Trobe University, Bundoora 3086, Victoria Australia (Corresponding author; email: h.nguyen5@latrobe.edu.au). GJM is at the School of Mathematics and Physics and Centre for Innovation in Biomedical Imaging Technology, University of Queensland, St. Lucia 4072, Queensland Australia. YY and JPL are at the Mouse Imaging Centre, Hospital for Sick Children, M5T 3H7 Toronto, Ontario Canada.

1 Notes regarding the binned estimation of the empirical Bayes model for grouped or discretely supported p-values

1.1 An EM algorithm for MML estimation

In order to compute $\hat{\boldsymbol{\theta}}_n$, we can utilize the EM algorithm of McLachlan and Jones (1988) for truncated and binned data. Suppose that we observe a realization x_i for each datum X_i. Further, let $n_j = \sum_{i=1}^{n} x_{ij}$, for each $j \in [m]$. Define $\boldsymbol{\theta}^{(0)}$ to be some initial value of the EM algorithm and denote the value of $\boldsymbol{\theta}^{(r)}$ after the rth iteration by $\boldsymbol{\theta}^{(r)\top} = \left(\pi_0^{(r)}, \mu_0^{(r)}, \sigma_0^{2(r)}, \mu_1^{(r)}, \sigma_1^{2(r)}\right)$. Without going into the details of its derivation, the EM algorithm proceeds as follows. At each iteration of the EM algorithm, perform an E-step, followed by an M-step.

On the $(r+1)$th E-step (expectation-step), compute $\alpha_{jk}^{(r+1)}$, $\beta_{jk}^{(r+1)}$, and $\gamma_{jk}^{(r+1)}$ for each $j \in [m]$ and $k \in \{0,1\}$, where

$$\alpha_{jk}^{(r+1)} = \frac{\pi_k^{(r)} \int_{B_j} \phi\left(z; \mu_k^{(r)}, \sigma_k^{2(r)}\right) \mathrm{d}z}{\int_{B_j} f\left(z; \boldsymbol{\theta}^{(r)}\right) \mathrm{d}z}, \tag{1}$$

$$\beta_{jk}^{(r+1)} = \frac{\pi_k^{(r)} \delta_{ik}^{(r+1)}}{\int_{B_j} f\left(z; \boldsymbol{\theta}^{(r)}\right) \mathrm{d}z}, \tag{2}$$

and

$$\gamma_{jk}^{(r+1)} = \frac{\pi_k^{(r)} \kappa_{ik}^{(r+1)}}{\int_{B_j} f\left(z; \boldsymbol{\theta}^{(r)}\right) \mathrm{d}z}. \tag{3}$$

Here

$$\delta_{jk}^{(r+1)} = \mu_k^{(r)} \int_{B_j} \phi\left(z; \mu_k^{(r)}, \sigma_k^{2(r)}\right) \mathrm{d}z - \sigma_k^{2(r)} \upsilon_{jk}^{(r+1)}$$

and

$$\begin{aligned} \kappa_{jk}^{(r+1)} &= \sigma_k^{2(r)} \left[\int_{B_j} \phi\left(z; \mu_k^{(r)}, \sigma_k^{2(r)}\right) \mathrm{d}z + \left(2\mu_k^{(r+1)} - \mu_k^{(r)}\right) \upsilon_{jk}^{(r+1)} - \omega_{jk}^{(r+1)} \right] \\ &\quad + \left[2\mu_k^{(r+1)} - \mu_k^{(r)}\right]^2 \upsilon_{jk}^{(r+1)}, \end{aligned}$$

where

$$\upsilon_{jk}^{(r+1)} = \phi\left(b_j; \mu_k^{(r)}, \sigma_k^{2(r)}\right) - \phi\left(b_{j-1}; \mu_k^{(r)}, \sigma_k^{2(r)}\right),$$

$$\omega_{jk}^{(r+1)} = b_j \phi\left(b_j; \mu_k^{(r)}, \sigma_k^{2(r)}\right) - b_{j-1} \phi\left(b_{j-1}; \mu_k^{(r)}, \sigma_k^{2(r)}\right),$$

and $b_m = \infty$. Then, on the $(r+1)$th M-step (maximization-step), compute

$$\pi_k^{(r+1)} = n^{-1} \sum_{j=1}^{m} n_j \alpha_{jk}^{(r+1)}, \tag{4}$$

$$\mu_k^{(r+1)} = \left[\sum_{j=1}^{m} n_j \alpha_{jk}^{(r+1)}\right]^{-1} \sum_{j=1}^{m} n_j \beta_{jk}^{(r+1)}, \tag{5}$$

and

$$\sigma_k^{2(r+1)} = \left[\sum_{j=1}^{m} n_j \alpha_{jk}^{(r+1)}\right]^{-1} \sum_{j=1}^{m} n_j \gamma_{jk}^{(r+1)}, \tag{6}$$

for each $k \in \{0,1\}$. The E-step and M-steps are repeated until some predetermined stopping criterion is met; see Lange (2013, Sec. 11.5) regarding stopping criteria. Upon stopping, the final iterate of the EM algorithm is declared the MML estimate $\hat{\boldsymbol{\theta}}_n$.

Since the algorithm composing of updates (1)–(6) constitutes an EM algorithm under the strict definition of Dempster, Laird and Rubin, D.B. (1977) (see also McLachlan and Krishnan, 2008, Sec. 1.5), the usual properties of the EM algorithm, as proved by Wu (1983), are conferred upon it. That is, starting from some initial value $\boldsymbol{\theta}^{(0)}$, if we let $\boldsymbol{\theta}^{(\infty)} = \lim_{r\to\infty} \boldsymbol{\theta}^{(r)}$ be a limit point of the EM algorithm, then $\boldsymbol{\theta}^{(\infty)}$ is a stationary point of the log-marginal likelihood

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \log \int_{B_j} f(z;\boldsymbol{\theta})\,\mathrm{d}z, \tag{7}$$

and the sequence $l\left(\boldsymbol{\theta}^{(r)}\right)$ is monotonically increasing in r; see McLachlan and Krishnan (2008, Ch. 3) for details regarding the properties of EM algorithms. We note that the EM algorithm given above is that which is implemented in the mixdist package.

1.2 Consistency of the estimator

As discussed in Bickel and Doksum (2001, Ch. 5), one of the most important properties of any large-sample estimator is that it is consistent (i.e. it converges to something meaningful as more data are obtained). We note that if one observes the data X_i and not P_i or Z_i, for $i \in [n]$, then we can write the individual log-mass for each X_i, given fixed bins B_j, as

$$\log \mathbb{P}(X_i = x;\boldsymbol{\theta}) = \prod_{j=1}^{m} \left[\int_{B_j} \pi_0 \phi\left(z;\mu_0,\sigma_0^2\right) + \pi_1 \phi\left(z;\mu_1,\sigma_1^2\right)\mathrm{d}z\right]^{x_{ij}}. \tag{8}$$

Substitution of (8) into (7) yields the log-marginal likelihood

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \mathbb{P}(X_i = x;\boldsymbol{\theta}).$$

Under mild assumptions regarding the dependence structure of the data $X_1,\ldots,X_n$, we can establish the consistency of the MML estimator $\hat{\boldsymbol{\theta}}_n$ via Theorem 5.14 of van der Vaart (1998).

Proposition 1 *Assume that $X_1, X_2, \ldots, X_n$ is an identical and strongly-dependent random sequence. Let $-\infty < m < M < \infty$, $0 < s < S < \infty$, and*

$$\Theta = \left\{\boldsymbol{\theta} : \pi_0 > 0, \pi_1 > 0, \pi_0 + \pi_1 = 1, (\mu_0,\mu_1) \in [m,M]^2, \left(\sigma_0^2,\sigma_1^2\right) \in [s,S]^2\right\}.$$

If

$$\Theta_0 = \left\{\boldsymbol{\theta}^0 \in \Theta : \mathbb{E}\log\mathbb{P}\left(X_1 = x;\boldsymbol{\theta}^0\right) = \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\log\mathbb{P}(X_1 = x;\boldsymbol{\theta})\right\},$$

then for every $\epsilon > 0$ and compact set $\mathbb{K} \subset \Theta$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\Theta_0} \left\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right\| \geq \epsilon \text{ and } \hat{\boldsymbol{\theta}} \in \mathbb{K}\right) \to 0.$$

We note that an assumption that implies strong-mixing is M-dependence; see for example Bradley (2005). That is, if for each index i, the datum X_i is dependent on only the data X_j, where $|i-j| \leq M < \infty$. This model is sufficient for many applied settings, such as genome studies and biological imaging.

A caveat to the application of the MML estimator is that one cannot always guarantee that $\hat{\boldsymbol{\theta}}_n$ is in fact the maximal value that is required in Proposition 1. This is because the EM algorithm is only guaranteed to converge to a local maximum of (7) (or a saddle-point that can easily be perturbed to continue onto a local maximum) and not the global maximum required by the theorems. This problem can be largely mitigated by using multiple runs of the EM algorithm from well-selected initial values. The topic of initialization of EM algorithms for mixture models is a complex one and discussions can be found in McLachlan (1988), Biernacki, Celeux and Govaert (2003), Karlis and Xekalaki (2003), and Melnykov and Melnykov (2012).

1.3 Proof of proposition 1

In order to apply van der Vaart (1998, Thm. 5.14), we must check that (i) $\log \mathbb{P}(X_1 = x; \boldsymbol{\theta})$ is continuous for all values of x, and that (ii) the uniform strong law of large numbers holds; that is

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}(X_i = x; \boldsymbol{\theta}) - \mathbb{E} \log \mathbb{P}(X_1 = x; \boldsymbol{\theta}) \right| \xrightarrow{\text{a.s.}} 0.$$

Property (i) is simple to verify since $\mathbb{P}(X_1 = x; \boldsymbol{\theta})$ can be written as an integral of a smooth function for any x. Thus it is continuous and its logarithm is also continuous. To establish property (ii), we utilize Andrews (1992, Thm. 4). This requires that $n^{-1} \sum_{i=1}^{n} \log \mathbb{P}(X_1 = x; \boldsymbol{\theta})$ converges to $\mathbb{E} \log \mathbb{P}(X_1 = x; \boldsymbol{\theta})$, pointwise, almost surely for any $\boldsymbol{\theta} \in \Theta$, and that $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\log \mathbb{P}(X_1 = x; \boldsymbol{\theta})| < \infty$. For any $\boldsymbol{\theta}$, the variance of $\log \mathbb{P}(X_1 = x; \boldsymbol{\theta})$ exists since it is a discrete random variable with only finite outcomes. Thus, we can apply the mixing continuous mapping theorem and the mixing strong law of large numbers (i.e. White, 2001, Thm 3.49 and Cor. 3.48), in order to obtain the pointwise convergence of $n^{-1} \sum_{i=1}^{n} \log \mathbb{P}(X_1 = x; \boldsymbol{\theta})$, almost surely. Next, we again note that $\log \mathbb{P}(X_1 = x; \boldsymbol{\theta})$ is a discrete random variable with finite outcomes for any finite $\boldsymbol{\theta}$. Therefore, the supremum and its expectation are also finite, since Θ contains only finite values. Therefore (ii) is verified and the proposition is proved.

Remark 1 *We note that van der Vaart (1998, Thm. 5.14) only lists the requirement to check assumption (ii) for the proof above. However, the theorem also makes an implicit assumption that the data are independent. Under dependence, we require the additional assumption of the strong law of large numbers (i), as demanded by Andrews (1992, Thm. 4). Here, we utilize the TSE-1D form of the theorem (cf. Andrews, 1992, Eqn. 3.2).*

2 Notes regarding the effect of integer encoding

2.1 The effect of integer encoding on the null distribution

In each case the estimated variance is reduced from the nominal value of $\sigma_0^2 = 1$. This reduction can be explained by the fact that the finite z-scores distribution that is obtained from the probit transformation of the encoded p-values is approximately standard normal distribution that is

truncated to the interval $[-a_\gamma, a_\gamma]$, where $a_\gamma = \Phi^{-1}\left(1 - 1/\left[2^{\gamma+1} - 1\right]\right)$. Using the variance formula for a doubly truncated standard normal distribution (cf. Forbes et al., 2011, Sec. 33.4), we have the variance formula for the z-scores:

$$\text{var}_\gamma = 1 - 2a_\gamma \phi(a_\gamma; 0, 1) / \left[\Phi(a_\gamma) - \Phi(-a_\gamma)\right].$$

Substituting 8 and 9 into γ, we obtain truncated variances of $\text{var}_8 = 9.64\text{E-}1$ and $\text{var}_9 = 9.80\text{E-}1$, respectively. These values are almost identical to those visualized in Figure 1(b).

We note that the extra $1/2$ factor in the calculation of a_γ (i.e. $1/2^{\gamma+1}$ rather than $1/2^\gamma$) arises from the fact that half of the p-values in the interval between zero and the next smallest number get rounded towards the zero, and similarly half of the p-values in the interval between one and the next largest number gets rounded towards one. Thus, we lose approximately $1/\left[2^\gamma - 1\right]$ observations from the extreme values of the distribution of the p-values that probit transform to infinite values. Here, the -1 term accounts for a fencepost error.

2.2 The effect on the z-score distribution

We can again provide a reason for the incorrect results that are obtained from the p-type inference. Let $a_\gamma = \Phi^{-1}\left(1 - 1/\left[2^{\gamma+1} - 1\right]\right)$, as in Section 2.1. The ML estimator is estimated using approximately

$$n\left(\pi_0\left[\Phi(a_\gamma) - \Phi(-a_\gamma)\right] + \pi_1\left[\Phi(a_\gamma - 2) - \Phi(-a_\gamma - 2)\right]\right) < n$$

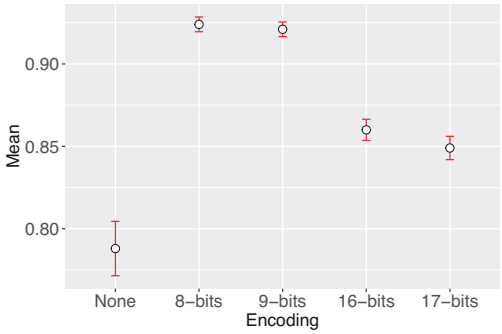
observations from the distribution that is characterized by the density $f(z;\boldsymbol{\theta})$, that is truncated on the interval $[-a_\gamma, a_\gamma]$. Note that no member of the family of densities of form $f(z;\boldsymbol{\theta})$ can perfectly match a truncated version of the density. For example, the two families of densities have different supports. Thus, the ML estimation procedure results in an estimated set of parameter values that yields a member of the untruncated density that best approximates the truncated density, in Kullback-Leibler divergence (cf. White, 1982). This approximation process explains the difference between the estimated parameter values and the generative parameter values. The smaller sample size explains the larger standard errors that are observed, uniformly over the estimates of the parameter elements.

3 Notes regarding the assessment of the binned estimator

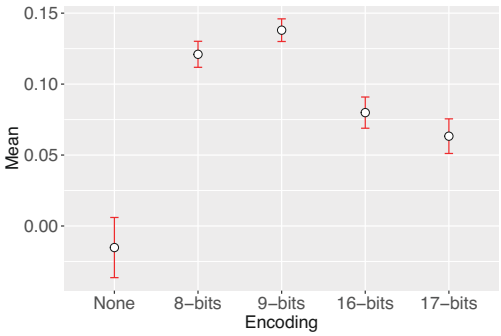
3.1 Accuracy of the z-score distribution

The first set of results from Figure (2) that are labelled N report the MML estimation results when no encodings of testing data are implemented. We observe that the MML estimates appear to be accurate and demonstrate no statistically significant deviation away from the generative parameter elements of the model. The accuracy of the MML estimator appears to be robust to the choice among the three assessed binning schemes. This empirical result supports the theoretical conclusions of Proposition 1.

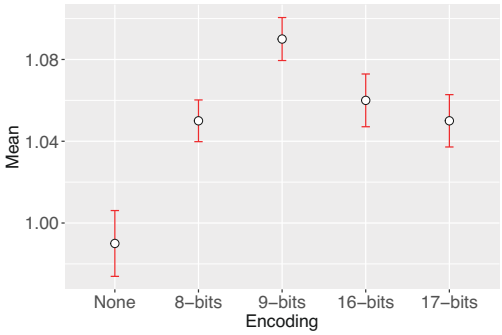
Upon inspection of the results, we found that the reason for the inaccuracy may be due to the fact that the FD and Scott binning methods yielded too many bins, that are of uniform width in the space of the z-scores. The encoding scheme generates uniform width rounding of data in the p-value space, which when converted to z-scores, can sometimes leave FD and Scott-type bins empty. This in turn causes the EM algorithm to fit the idiosyncratic nature of these empty bin



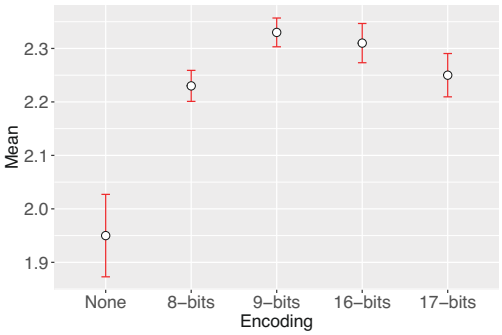
(a) Mean and standard errors from 100 ML estimates $\hat{\pi}_0$ of $\pi_0 = 0.8$.



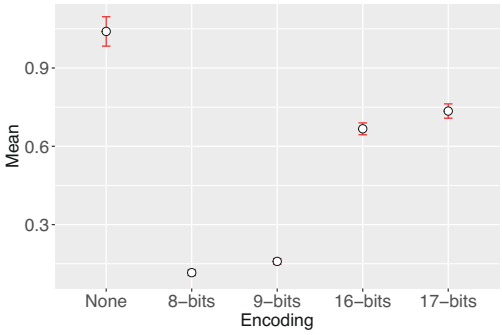
(b) Mean and standard errors from 100 ML estimates $\hat{\mu}_0$ of $\mu_0 = 0$.



(c) Mean and standard errors from 100 ML estimates $\hat{\sigma}_0^2$ of $\sigma_0^2 = 1$.

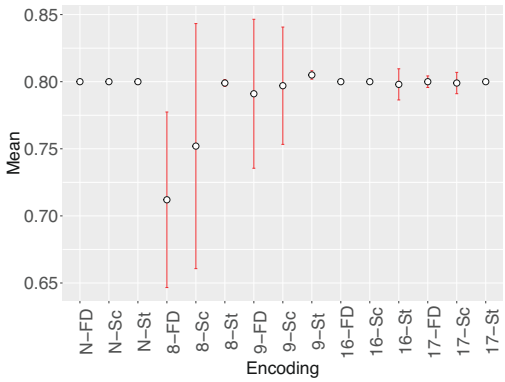


(d) Mean and standard errors from 100 ML estimates $\hat{\mu}_1$ of $\mu_1 = 2$.

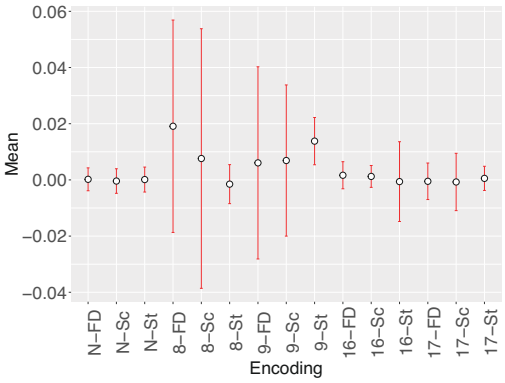


(e) Mean and standard errors from 100 ML estimates $\hat{\sigma}_1^2$ of $\sigma_1^2 = 1$.

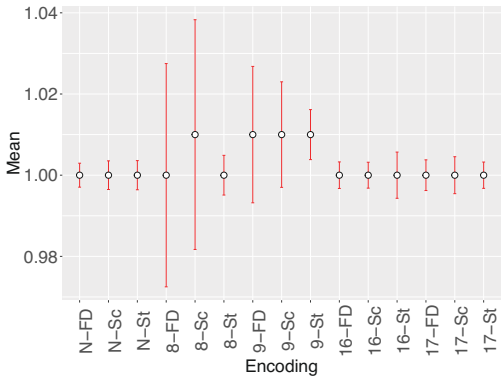
Figure 1: *Monte Carlo study from Section 3.3, regarding the estimation of* $\boldsymbol{\theta}$ *in the presence of integer encodings of p-values. Means are represented by points and standard errors are equal to half the length of the error bars.*



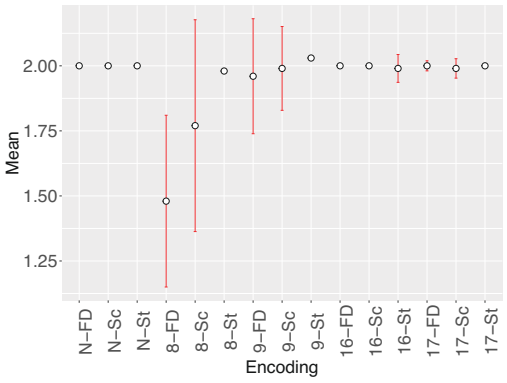
(a) Mean and standard errors from 100 MML estimates $\hat{\pi}_0$ of $\pi_0 = 0.8$.



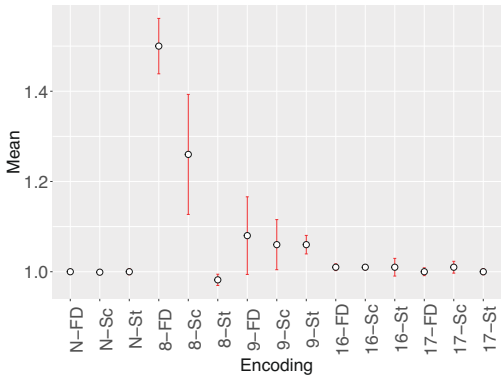
(b) Mean and standard errors from 100 MML estimates $\hat{\mu}_0$ of $\mu_0 = 0$.



(c) Mean and standard errors from 100 MML estimates $\hat{\sigma}_0^2$ of $\sigma_0^2 = 1$.



(d) Mean and standard errors from 100 MML estimates $\hat{\mu}_1$ of $\mu_1 = 2$.



(e) Mean and standard errors from 100 MML estimates $\hat{\sigma}_1^2$ of $\sigma_1^2 = 1$.

Figure 2: *Monte Carlo study regarding the binned estimation of $\boldsymbol{\theta}$, in the presence of integer encodings of p-values. Means are represented by points and standard errors are equal to half the length of the error bars. The result of each experiment and its binning is reported along the x axis. Here, N indicates no encoding, where as 8, 9, 16, or 17 indicates the level of γ. The abbreviations FD, Sc and St indicate that Freedman Diaconis, Scott or Sturges binning was used, respectively.*

patterns, that leads to overfitting and biased estimation. This problem diminishes as γ increases, since there is more overlap between the encoded p-values and the FD and Scott-type bins, which leads to fewer numbers of empty bins, and thus less overfitting. The Sturges binning mitigates against this empty bins problem by having much larger bin sizes than the other two assessed methods.

One may remedy the empty bins and small bins problems of the FD and Scott-type methods by using some kind of heuristic for joining together adjacent bins to produce bins that contain larger numbers of observations. Such methods include the strategy of combining frequency classes that are discussed by Lewis and Burke (1949). Due to the ad-hoc nature of such methods, and the number of different approaches, the pursuit of their application falls outside the scope of this article.

3.2 FDR control experiment and simulation scenarios

Scenario S1 is ideal, in the sense that it fulfills the situation whereupon the hypotheses are generate test statistics that are IID and well-specified in the sense that the p-values P_i are uniformly distributed under the null. All methods should adequately control the FDR in this case.

Scenarios S2 and S3 are designed to test the performance of the methods when there are dependencies between the hypotheses. Since S1 only induces a positive correlation structure on the test statistics, all of the methods should be able to correctly control the FDR level in this case. In S2, negative correlations are induced between consecutive test statistics. Thus, there are no theoretical guarantees of the performance of BH in this case. The robustness of BH to positive correlation is proved in Benjamini and Yekutieli (2001) (see also Yekutieli, 2008). Robustness of BY to all forms of correlation is proved in Benjamini and Yekutieli (2001) and the performance of q-values under weak dependence is discussed in Storey and Tibshirani (2003).

Scenario S4 The is somewhat ideal for our EB-based method, since the z-scores distribution under the null is a normal distribution. However, it violates the uniformity assumption that the other methods depend upon.

Scenario S5 is misspecified in the sense that the the p-values P_i are not computed under the correct null hypothesis. It is also not ideal for our EB-based method, since the distribution of the z-scores under the null is not normal. Thus, there are no performance guarantees for any of the assessed methods in this case.

3.3 Results

The results for Scenarios S1–S5 are reported in Tables 1–5, respectively.

We begin by making some general observations. Firstly, in terms of power (i.e. TPP), the FDR control methods follow the order: BY, EB, BH, and q-values, from least to most powerful. Similarly, with respect to conservatism of their FDR control (i.e. how much smaller FPP is to the nominal value β), we observe the same order: BY, EB, BH, and q-values, from most conservative to least. Across the three well-specified testing scenarios (S1–S3), we observe that EB, BH, and BY were all conservative. These initial observations were uniform across the different encoding methods.

We observe that q-values can often result in anti-conservative control of the FDR (i.e. FDP consistently exceeding the nominal value β) in many scenarios and encoding types. For example in S1–S3, we observe that q-values is anti-conservative for both values of β when we use 8-bits encoding. We recommend that q-values should be avoided when data are compressed using 8-bits integers encoding.

Table 1: *Average FDP and TPP results (Reps = 100) for Scenario S1. The best outcome under each encoding for each value of β is highlighted in boldface. Here, the best FDP proportion is one that is closest to the nominal value without exceeding it and the best TPP value is highest value given that the FDP does not exceed the nominal value. FDP values that exceed the nominal value are emphasized in italics.*

		FDP		TPP	
Encoding	Method	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.05$	$\beta = 0.10$
None	EB	2.46E-02	4.66E-02	1.22E-01	2.14E-01
	BH	**3.99E-02**	7.96E-02	**1.89E-01**	3.25E-01
	BY	3.19E-03	6.63E-03	1.12E-02	2.84E-02
	q-values	*5.03E-02*	**1.00E-01**	2.28E-01	**3.81E-01**
8-bits	EB	**4.01E-02**	4.11E-02	**1.90E-01**	1.93E-01
	BH	3.98E-02	**9.64E-02**	1.88E-01	**3.70E-01**
	BY	3.98E-02	3.98E-02	1.88E-01	1.88E-01
	q-values	*7.23E-02*	*1.13E-01*	3.01E-01	4.12E-01
9-bits	EB	2.88E-02	4.33E-02	1.42E-01	2.00E-01
	BH	4.92E-02	**8.29E-02**	2.25E-01	**3.34E-01**
	BY	2.75E-02	2.75E-02	1.36E-01	1.36E-01
	q-values	**4.94E-02**	*1.02E-01*	**2.26E-01**	3.85E-01
16-bits	EB	2.47E-02	4.63E-02	1.22E-01	2.13E-01
	BH	**3.99E-02**	**7.98E-02**	**1.88E-01**	**3.25E-01**
	BY	3.32E-03	6.72E-03	1.67E-02	3.08E-02
	q-values	*4.96E-02*	*9.98E-02*	2.26E-01	3.80E-01
17-bits	EB	2.50E-02	4.64E-02	1.21E-01	2.13E-01
	BH	**4.02E-02**	7.95E-02	**1.88E-01**	3.25E-01
	BY	2.73E-03	7.07E-03	1.31E-02	3.01E-02
	q-values	*5.06E-02*	**1.00E-01**	2.28E-01	**3.80E-01**

Table 2: *Average FDP and TPP results (Reps = 100) for Scenario S2. The best outcome under each encoding for each value of β is highlighted in boldface. Here, the best FDP proportion is one that is closest to the nominal value without exceeding it and the best TPP value is highest value given that the FDP does not exceed the nominal value. FDP values that exceed the nominal value are emphasized in italics.*

		FDP		TPP	
Encoding	Method	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.05$	$\beta = 0.10$
None	EB	2.44E-02	4.62E-02	1.23E-01	2.16E-01
	BH	3.96E-02	7.96E-02	1.90E-01	3.26E-01
	BY	2.99E-03	6.16E-03	1.16E-02	2.87E-02
	q-values	**4.96E-02**	**1.00E-01**	**2.28E-01**	**3.81E-01**
8-bits	EB	**3.99E-02**	3.99E-02	**1.88E-01**	1.88E-01
	BH	**3.99E-02**	**9.65E-02**	**1.88E-01**	**3.69E-01**
	BY	**3.99E-02**	3.99E-02	**1.88E-01**	1.88E-01
	q-values	*7.26E-02*	*1.12E-01*	3.02E-01	4.09E-01
9-bits	EB	2.81E-02	4.23E-02	1.37E-01	1.97E-01
	BH	4.92E-02	**8.23E-02**	2.25E-01	**3.33E-01**
	BY	2.81E-02	2.81E-02	1.37E-01	1.37E-01
	q-values	**4.95E-02**	*1.02E-01*	**2.26E-01**	3.85E-01
16-bits	EB	2.45E-02	4.67E-02	1.23E-01	2.14E-01
	BH	**4.00E-02**	8.05E-02	**1.89E-01**	3.26E-01
	BY	3.96E-03	6.90E-03	1.74E-02	3.12E-02
	q-values	*5.04E-02*	**1.00E-01**	2.27E-01	**3.80E-01**
17-bits	EB	2.42E-02	4.62E-02	1.21E-01	2.13E-01
	BH	4.00E-02	8.01E-02	1.88E-01	3.25E-01
	BY	4.63E-03	7.27E-03	1.31E-02	3.00E-02
	q-values	**5.00E-02**	**1.00E-01**	**2.27E-01**	**3.80E-01**

Table 3: *Average FDP and TPP results (Reps = 100) for Scenario S3. The best outcome under each encoding for each value of β is highlighted in boldface. Here, the best FDP proportion is one that is closest to the nominal value without exceeding it and the best TPP value is highest value given that the FDP does not exceed the nominal value. FDP values that exceed the nominal value are emphasized in italics.*

		FDP		TPP	
Encoding	Method	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.05$	$\beta = 0.10$
None	EB	2.52E-02	4.71E-02	1.22E-01	2.14E-01
	BH	**4.04E-02**	7.98E-02	**1.89E-01**	3.25E-01
	BY	3.11E-03	7.10E-03	1.16E-02	2.88E-02
	q-values	*5.06E-02*	**9.97E-02**	2.28E-01	**3.81E-01**
8-bits	EB	**4.02E-02**	4.05E-02	**1.88E-01**	1.89E-01
	BH	**4.02E-02**	**9.60E-02**	**1.88E-01**	**3.69E-01**
	BY	**4.02E-02**	4.02E-02	**1.88E-01**	1.88E-01
	q-values	*7.21E-02*	*1.12E-01*	3.02E-01	4.10E-01
9-bits	EB	2.90E-02	4.51E-02	1.41E-01	2.05E-01
	BH	4.97E-02	**8.20E-02**	2.25E-01	**3.31E-01**
	BY	2.78E-02	2.78E-02	1.36E-01	1.36E-01
	q-values	**4.98E-02**	*1.01E-01*	**2.26E-01**	3.84E-01
16-bits	EB	2.48E-02	4.60E-02	1.22E-01	2.14E-01
	BH	3.99E-02	7.97E-02	1.89E-01	3.25E-01
	BY	4.20E-03	6.93E-03	1.71E-02	3.09E-02
	q-values	**4.98E-02**	**9.98E-02**	**2.28E-01**	**3.80E-01**
17-bits	EB	2.50E-02	4.66E-02	1.22E-01	2.14E-01
	BH	4.00E-02	7.98E-02	1.88E-01	3.24E-01
	BY	3.42E-03	7.05E-03	1.28E-02	2.97E-02
	q-values	**4.99E-02**	**1.00E-01**	**2.27E-01**	**3.80E-01**

Table 4: *Average FDP and TPP results (Reps = 100) for Scenario S4. The best outcome under each encoding for each value of β is highlighted in boldface. Here, the best FDP proportion is one that is closest to the nominal value without exceeding it and the best TPP value is highest value given that the FDP does not exceed the nominal value. FDP values that exceed the nominal value are emphasized in italics.*

		FDP		TPP	
Encoding	Method	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.05$	$\beta = 0.10$
None	EB	**2.48E-02**	**4.63E-02**	**1.22E-01**	**2.14E-01**
	BH	*1.48E-01*	*2.47E-01*	4.89E-01	6.50E-01
	BY	2.17E-02	3.69E-02	1.07E-01	1.77E-01
	q-values	*3.74E-01*	*5.69E-01*	7.93E-01	9.33E-01
8-bits	EB	*8.95E-02*	**8.95E-02**	3.50E-01	**3.50E-01**
	BH	*1.50E-01*	*2.49E-01*	4.92E-01	6.53E-01
	BY	*8.95E-02*	**8.95E-02**	3.50E-01	**3.50E-01**
	q-values	*3.80E-01*	*5.72E-01*	7.99E-01	9.34E-01
9-bits	EB	*6.46E-02*	**6.46E-02**	2.74E-01	**2.74E-01**
	BH	*1.61E-01*	*2.53E-01*	5.13E-01	6.58E-01
	BY	*6.46E-02*	**6.46E-02**	2.74E-01	**2.74E-01**
	q-values	*3.76E-01*	*5.69E-01*	7.95E-01	9.33E-01
16-bits	EB	**2.51E-02**	**4.74E-02**	**1.25E-01**	**2.17E-01**
	BH	*1.48E-01*	*2.47E-01*	4.89E-01	6.50E-01
	BY	2.17E-02	3.70E-02	1.10E-01	1.77E-01
	q-values	*3.74E-01*	*5.69E-01*	7.93E-01	9.33E-01
17-bits	EB	**2.45E-02**	**4.65E-02**	**1.22E-01**	**2.13E-01**
	BH	*1.49E-01*	*2.47E-01*	4.90E-01	6.51E-01
	BY	2.19E-02	3.73E-02	1.09E-01	1.78E-01
	q-values	*3.74E-01*	*5.68E-01*	7.92E-01	9.32E-01

Table 5: *Average FDP and TPP results (Reps = 100) for Scenario S5. The best outcome under each encoding for each value of β is highlighted in boldface. Here, the best FDP proportion is one that is closest to the nominal value without exceeding it and the best TPP value is highest value given that the FDP does not exceed the nominal value. FDP values that exceed the nominal value are emphasized in italics.*

		FDP		TPP	
Encoding	Method	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.05$	$\beta = 0.10$
None	EB	**4.35E-02**	6.15E-02	**1.02E-01**	1.85E-01
	BH	*6.10E-02*	**9.62E-02**	1.82E-01	**3.20E-01**
	BY	2.48E-02	2.80E-02	1.51E-02	3.09E-02
	q-values	*7.14E-02*	1.16E-01	2.26E-01	3.85E-01
8-bits	EB	*1.03E-01*	*1.03E-01*	3.46E-01	3.46E-01
	BH	*1.56E-01*	*2.45E-01*	4.93E-01	6.58E-01
	BY	*1.03E-01*	*1.03E-01*	3.46E-01	3.46E-01
	q-values	*3.65E-01*	*5.62E-01*	8.00E-01	9.33E-01
9-bits	EB	*8.23E-02*	**8.23E-02**	2.70E-01	**2.70E-01**
	BH	*1.66E-01*	2.50E-01	5.15E-01	6.66E-01
	BY	*8.23E-02*	**8.23E-02**	2.70E-01	**2.70E-01**
	q-values	*3.63E-01*	5.59E-01	7.97E-01	9.32E-01
16-bits	EB	3.97E-02	5.67E-02	8.55E-02	1.62E-01
	BH	*1.55E-01*	*2.43E-01*	4.92E-01	6.56E-01
	BY	**4.45E-02**	**5.86E-02**	**1.06E-01**	**1.72E-01**
	q-values	*3.61E-01*	*5.60E-01*	7.97E-01	9.33E-01
17-bits	EB	4.12E-02	5.73E-02	8.53E-02	1.63E-01
	BH	*1.54E-01*	*2.42E-01*	4.90E-01	6.55E-01
	BY	**4.48E-02**	**5.88E-02**	**1.03E-01**	**1.70E-01**
	q-values	*3.60E-01*	*5.58E-01*	7.95E-01	9.32E-01

In Scenario S4, we observe that BH and q-values are highly anti-conservative. Here applications of the two methods resulted in FDP values that greatly exceeded the nominal value of β, uniformly over the encoding methods. Both EB and BY were also anti-conservative when the data were compressed by either 8-bits or 9-bits encodings, for control of the FDR at rate $\beta = 0.05$, although the amounts exceeded were much less than those of the BH and q-values results. At the $\beta = 0.10$ level, both methods were conservative for the two previously mentioned encoding rates. At all other encoding rates, EB and BY were conservative. EB was less conservative and more powerful than BY in each of the cases where they both correctly controlled the FDR level, and thus should be preferred.

4 Notes regarding the example application

4.1 Goodness-of-fit of the empirical Bayes model

In order to assess the goodness-of-fit of (7) to the z-score data, we can compute the corresponding log-ML, which equals $l\left(\hat{\boldsymbol{\theta}}\right) = -3252.598$. We can then compare this value to the log-ML of the EB model under the so-called theoretical null model, under the assumption that each p-value P_i is uniformly distributed in the unit interval when H_i is null, for each $i \in [n]$. This corresponds to fixing $f_0(z) = \phi(z;0,1)$.

Let

$$f(z;\boldsymbol{\vartheta}) = \pi_0\phi(z;0,1) + \pi_1\phi\left(z;\mu_1,\sigma_1^2\right)$$

be the theoretical null EB model, where $\boldsymbol{\vartheta}^\top = \left(\pi_0,\mu_1,\sigma_1^2\right)$ is the restricted parameter vector. Using the mix function in the mixdist package, we estimate the parameter vector $\boldsymbol{\vartheta}$ for the data from Section 5 via MML estimation in order to obtain the fitted model

$$f\left(z;\hat{\boldsymbol{\vartheta}}\right) = 0.2884 \times \phi(0,1) + 0.7116 \times \phi\left(2.376, 1.796^2\right). \tag{9}$$

A visualization of (9) along with its weighted null and alternative components $\hat{\pi}_0 f_0(z) = 0.2884 \times \phi(0,1)$ and $\hat{\pi}_1 \hat{f}_1(z) = 0.7116 \times \phi\left(2.376, 1.796^2\right)$ is provided in Figure 3. We observe that the fit is dramatically poorer in the centre of the plot, when compared to that which was obtained in Figure 3. This fact can be compared by via the log-ML value of the theoretical null EB model $l\left(\hat{\boldsymbol{\vartheta}}\right) = -14399.83$.

We observe that the log-ML values suggest that (7) provides a better fit than (3). However, due to the differences in the number of parameters (i.e., $\left|\hat{\boldsymbol{\theta}}\right| = 5$ and $\left|\hat{\boldsymbol{\vartheta}}\right| = 3$), we cannot directly compare the values $l\left(\hat{\boldsymbol{\theta}}\right)$ and $l\left(\hat{\boldsymbol{\vartheta}}\right)$. Since the marginal likelihood is a kind of pseudolikelihood, we can use the pseudolikelihood information criterion (PLIC) of Stanford and Raftery (2002). For (7) and (3), the PLIC values are

$$-2l\left(\hat{\boldsymbol{\theta}}\right) + \left|\hat{\boldsymbol{\theta}}\right| \log n = 6579.454$$

and

$$-2l\left(\hat{\boldsymbol{\vartheta}}\right) + \left|\hat{\boldsymbol{\vartheta}}\right| \log n = 28844.21,$$

respectively. We again favor (7) over (3), since smaller PLIC values are preferred. Thus, we can be confident that the EB model provides a discernibly better fit to the theoretical null model and is thus is a feasibly better fit to the experimental data.

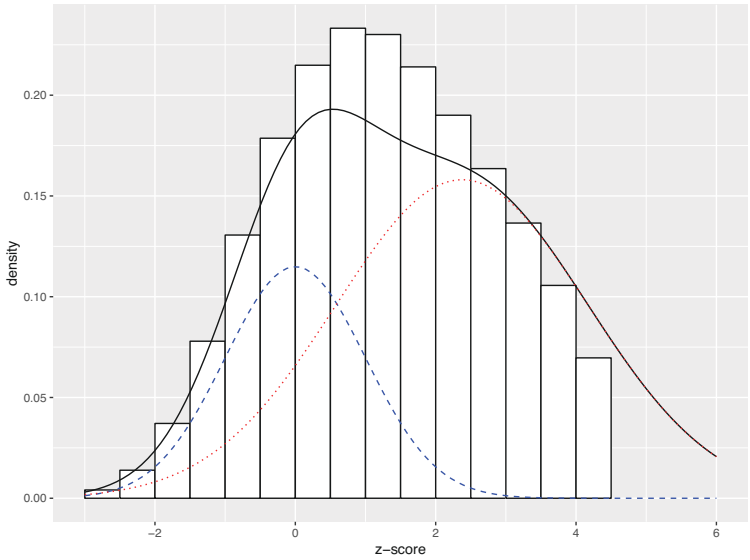


Figure 3: *The theoretical null-based functions* $f\left(\cdot;\hat{\boldsymbol{\vartheta}}\right)$, $\hat{\pi}_0 f_0(z)$, *and* $\hat{\pi}_1 \hat{f}_1$ *are plotted with solid, dashed, and dotted lines, respectively.*

References

Andrews, D. W. K. (1992). Generic uniform convergence. *Econometric Theory*, 8, 241–257.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165–1188.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics*. Upper Sanddle River: Prentice Hall.

Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, 41, 561–575.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2, 107–144.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Forbes, C., Evans, M., Hastings, N. and Peacock, B. (2011). *Statistical Distributions*. New York: Wiley.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, 41, 577–590.

Lange, K. (2013). *Optimization*. New York: Springer.

Lewis, D. and Burke, C. J. (1949). The use and misuse of the chi-squared test. *Psychological Bulletin*, 46, 433–489.

McLachlan, G. J. (1988). On the choice of starting values for the EM algorithm in fitting mixture models. *The Statistician*, 37, 417–425.

McLachlan, G. J. and Jones, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, 44, 571–578.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm And Extensions*. New York: Wiley, 2 edition.

Melnykov, V. and Melnykov, I. (2012). Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Computational Statistics and Data Analysis*, 56, 1381–1395.

Stanford, C. D. and Raftery, A. E. (2002). Approximate Bayes factor for image segmentation: the pseudo-likelihood information criterion (PLIC). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 1517–1520.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100, 9440–9445.
van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.
White, H. (2001). *Asymptotic Theory For Econometricians*. San Diego: Academic Press.
Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11, 95–103.
Yekutieli, D. (2008). False discovery rate control for non-positively regression dependent test statistics. *Journal of Statistical Planning and Inference*, 138, 405–415.