SORT 43 (2) July-December 2019, 237-258 DOI: 10.2436/20.8080.02.87

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

Abstract

False discovery rate (FDR) control is important in multiple testing scenarios that are common in neuroimaging experiments, and p-values from such experiments may often arise from some discretely supported distribution or may be grouped in some way. Two situations that may lead to discretely supported distributions are when the p-values arise from Monte Carlo or permutation tests are used. Grouped p-values may occur when p-values are quantized for storage. In the neuroimaging context, grouped p-values may occur when data are stored in an integer-encoded form. We present a method for FDR control that is applicable in cases where only p-values are available for inference, and when those p-values are discretely supported or grouped. We assess our method via a comprehensive set of simulation scenarios and find that our method can outperform commonly used FDR control schemes in various cases. An implementation to a mouse imaging data set is used as an example to demonstrate the applicability of our approach.

MSC: 62-07, 62F03, 62F35, 62N03, 62P10.

Keywords: Censored data, data quantization, discrete support, empirical-Bayes, false discovery rate control, grouped data, incompletely observed data, mixture model.

1 Introduction

Modern experiments in numerous fields of science now output the results of thousands to millions of hypothesis tests simultaneously. Recent accounts of the theoretical aspects of the phenomenon of simultaneous statistical inference with applications in the life sciences can be found in Dickhaus (2014). Further treatment of the topic can be found in Efron (2010).

We assume that we are operating in a scenario whereupon we (only) observe p-values from $n \in \mathbb{N}$ simultaneous tests of the hypotheses H_i ($i \in [n]$; $[n] = \{1, ..., n\}$), which may

* 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.

Received: December 2018
Accepted: April 2019

be either null or otherwise and may be related in some manner. Suppose that we are conducting well-specified standard significance tests at significance level $\alpha \in (0,1)$. If all of the hypotheses are null, then we can directly compute the expected number of tests declared significant as $n\alpha$. Taking n large (e.g. $n \geq 10^6$) and α at usual levels such as $\alpha \in (0.001, 0.1)$, the number of incorrectly declared hypotheses as not null can be greatly inflated. When there is a potential for large numbers of incorrectly rejected hypotheses, the outcome of using only standard significant tests can lead to spurious conclusions.

In recent years, the leading paradigm for the handling of large-scale simultaneous hypothesis testing scenarios is via the control of the false discovery rate (FDR) of an experiment. The control of FDR was first introduced by Benjamini and Hochberg (1995) and has since been developed upon by numerous other authors. The FDR of an experiment can be defined as FDR=$\mathbb{E}(N_{01}/N_R)\,\mathbb{P}(N_R > 0)$, where N_{01} and N_R denote the number of false positives and the number of rejected hypotheses (hypotheses declared significantly alternative) from the experiment, respectively.

The FDR control method of Benjamini and Hochberg (1995) was first developed to only take an input of n IID (identically and independently distributed) p-values. An extension towards the control of FDR in samples of correlated p-values was derived in Benjamini and Yekutieli (2001). Since these key publications, there have been numerous articles written on the topic of FDR control in various settings and under various conditions; see Benjamini (2010) and the comments therein for an account of the history and development of FDR control.

In most FDR control methods, there is an explicit assumption that the marginal distribution of the p-values of an experiment is uniform over the unit interval, if the hypothesis under consideration is null. This assumption arises via the classical theory of p-values of well-specified tests (cf. Dickhaus, 2014, Sect. 2). However, in practice, there are numerous ways for which the distribution of p-values under the null can deviate from uniformity. In Efron (2010, Sect. 6.4), several causes of deviation from uniformity are suggested. Broadly, these are: failed mathematical assumptions (e.g. incorrect use of distribution for computing p-values), correlation between p-values, and unaccounted covariates or misspecification of null hypotheses. A treatment on the effects of misspecification of the null hypotheses due to unaccounted covariates can be found in Barreto and Howland (2006, Chap. 7 Appendix and Chap. 18).

There are some FDR methods that account for deviation from uniformity in the null distribution. These include the methods of Yekutieli and Benjamini (1999), Korn et al. (2004), Pollard and van der Laan (2004), van der Laan and Hubbard (2006), and Habiger and Pena (2011). Unfortunately, the listed methods all require access to the original data of the experiment in order to compute permutation-based test statistics and thus permutation-based p-values. As mentioned previously, access to the original experimental data lies outside of the scope of this article as we only assume knowledge of the p-values. The empirical-Bayes (EB) paradigm provides a powerful framework under which the deviation of the null away from uniformity can be addressed with only access to the experimental p-values. The EB paradigm for FDR control was first introduced in

Efron et al. (2001). A relatively complete account of the EB paradigm appears in Efron (2010).

We largely follow the work of McLachlan, Bean and Ben-Tovim (2006) and Nguyen et al. (2014). Our novelty and development of the available literature is to present a methodology for addressing the problems that are introduced when p-values are distributed on a discrete support or when the p-values are grouped.

As in Nguyen et al. (2014), we particularly focus on the context of neuroimaging applications. In voxel-based morphometric neuroimaging studies (see, e.g., Ashburner and Friston, 2000), the number of simultaneously tested hypotheses often range in the tens of thousands to the tens of millions. Due to such inflated numbers, the risk of making false discoveries is often unacceptably high. Making inference without FDR control in such situations may lead to an overabundance of absurd conclusions. This is well demonstrated in the infamous results of Bennett et al. (2009), where neuronal activation in the brain of a dead fish was observed in a functional magnetic resonance imaging study, where the FDR was not controlled. Thus, FDR control is an important and ongoing area of research in the neuroimaging literature. A classic treatment regarding FDR control in neuroimaging can be found in Genovese, Lazar and Nichols (2002).

Grouped p-values may arise under incomplete observation; that is, under censoring, grouping, or quantization observation of p-values; see Turnbull (1976) for working definitions of the censored and grouped data and Gersho and Gray (1992) for quantization. We shall elaborate upon these definitions in the sequel.

Neuroimaging data such as MRI and functional MRI volumes are usually stored via one of a number of common storage protocols. Incomplete data may arise when data are compressed using one of these storage algorithms. Some common storage protocols under which neuroimaging data may be compressed include ANALYZE (Robb et al., 1989), DICOM Bidgood et al. (1997), MINC (Vincent et al., 2003), and NIFTI Cox et al., 2004). A good summary of these protocols is presented in Larobina and Murino (2014). In the pursuit of reduced storage sizes, it is not uncommon for neuroimaging data volumes to be stored at the minimum precision specification of any of the aforementioned formats. For example, DICOM volumes can only store data as integers, at a precision level as low as 8-bits (i.e. $2^8 = 256$ unique values). When p-values are stored in such a format, the true values are grouped into bins that are centered on a discrete number of possible values on the unit interval.

Discretely supported p-values may arise from Monte Carlo or permutation tests. In such cases, the p-values for a fixed number of permutations or Monte Carlo replications R, can only take on $R+1$ discrete value. Furthermore, Monte Carlo and permutation tests are both random approximations of exact tests. Such tests can again only output a discrete number of possible p-values that depend on the sample size of the data from which they are computed (cf. Phipson and Smyth, 2010). Monte Carlo and permutation tests are frequently used in neuroimaging studies; see, for example, Winkler et al. (2014).

It is known that grouped observations of real numbers can often lead to inaccuracies in statistical computations. Discussions of some aspects regarding the effects of grouping on statistical computation are discussed in Moschitta, Schoukens and Carbone (2015). The effects of quantization can particularly be ruinous when applying standard EB-based FDR control approaches. The effects of incompleteness in the observation of p-values qualifies as a failure in mathematical assumptions, under the taxonomy of Efron (2010, Sect. 6.4).

In this article, we address the problem of EB-based FDR control using p-values that are discretely supported or grouped, via the use of binned estimation. We demonstrate the effect of grouped p-values on the estimation of the EB model. Making use of the EM (expectation–maximization of Dempster, Laird and Rubin (1977) algorithm from the mix function in the mixdist package (MacDonald and Du, 2012) in the R programming language (R Core Team, 2016), we demonstrate that one can simply and rapidly maximum marginal likelihood (MML) estimation (cf. Varin, 2008) of the EB model. We further prove the consistency of the MML estimator for the EB model. A second numerical study is conducted to demonstrate the performance of our method under incomplete observation of p-values, where a comparison between our method is made against the commonly used methods of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), and Storey (2002). An example application to a mouse brain imaging dataset is then provided to demonstrate the usefulness of our approach in a real data scenario.

The article proceeds as follows. In Section 2, we introduce concepts relating to grouped and discretely observed p-values, and the EB model for p-values. We then demonstrate how the EB model can be used for FDR control. In Section 3, we present a demonstration of the effect of grouped p-values on the naive estimation of the EB model. In Section 4, a numerical study of the performance of our method is presented. In Section 5, the methodology is applied to control the FDR of a mouse imaging data set. Conclusions are drawn in Section 6. Further details regarding our methodology are included in the Supplementary Materials.

2 Binned estimation of the empirical Bayes model for grouped or discretely supported p-values

Let $0 = a_0 < a_1 < \cdots < a_{m-1} < a_m = 1$ be a set of m points along the line segment $[0,1]$. Suppose that we observe n p-values $P_i \in [0,1]$, for $i \in [n]$. Grouping may occur when $P_1, \ldots, P_n$ are subject to rounding (or quantization), such that each p-value to the nearest point a_j, for $j \in [m] \cup \{0\}$, where various measurements of closeness may be used for different applications. Observation of $P_1, \ldots, P_n$ may also be grouped when they are censored. That is, when we only observe the fact that each p-value $P_i \in (a_{j-1}, a_j)$, for some $j \in [n]$, and not its precise value. Under either quantization or censoring, the p-values P_i are each mapped to a discrete set of values, either the $m+1$ quantization

centers a_j or the m intervals (a_{j-1}, a_j), enumerated by the index j. In the case where $P_1, \ldots, P_n$ arise from a Monte Carlo or permutation tests, we may envisage that they are quantized approximations of p-values that arise from an asymptotically large population size and thus can be treated in the same manner as quantized p-values in practice.

2.1 The empirical Bayes model

For $i \in [n]$, let $Z_i = \Phi^{-1}(1 - P_i)$ be the probit transformation of P_i. We refer to Z_i as the z-scores. Here Φ is the cumulative distribution function of the standard normal distribution. Under the EB paradigm, we assume that some proportion $\pi_0 \in [0,1]$ of the n hypotheses are null and thus $\pi_1 = 1 - \pi_0$ are otherwise. Since an alternative (not null) hypothesis generates a p-value that is on average smaller than that of a null hypothesis, we can also assume that the z-scores of null hypotheses arise from some distribution with a mean $\mu_0 \in \mathbb{R}$, where $\mu_0 < \mu_1$ and $\mu_1 \in \mathbb{R}$ is the mean of the alternative z-scores. Under uniformity of the p-values, the z-scores have a standard normal distribution, we can approximate the density of the null z-scores by $f_0(z) = \phi\left(z; \mu_0, \sigma_0^2\right)$, where $\sigma_0^2 > 0$ and $\phi\left(\cdot; \mu, \sigma^2\right)$ is the normal density function with mean μ and variance σ^2. Likewise, we can approximate the density of the alternative z-scores by $f_1(z) = \phi\left(z; \mu_1, \sigma_1^2\right)$, where $\sigma_1^2 > 0$ (cf. Efron, 2004). The marginal density of any z-score, can be approximated by the two-component mixture model

$$f(z; \boldsymbol{\theta}) = \pi_0 f_0(z) + \pi_1 f_1(z), \tag{1}$$

where $\boldsymbol{\theta}^\top = \left(\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2\right)$ is the model parameter vector and $(\cdot)^\top$ is the transpose operator. We say that (1) is the EB model for p-values.

2.2 Statistical model for binned data

Let $-\infty = b_0 < b_1 < b_2 < \ldots < b_{m-1} < \infty$ for some $m \in \mathbb{N} \setminus \{1\}$. We define m bins B_j, for $j \in [m]$, where $B_j = (b_{j-1}, b_j]$ for $j \in [m-1]$ and $B_m = (b_{m-1}, \infty)$. Suppose that we observe n p-values P_i that are converted to z-scores Z_i, which may be infinite in value. Further, define $\mathbb{I}(A)$ as the indicator variable that takes value 1 if proposition A is true and 0 otherwise, and define a new random variable $X_i^\top = (X_{i1}, \ldots, X_{im})$, where $X_{ij} = \mathbb{I}(Z_i \in B_j)$, for each i and $j \in [m]$.

Suppose that the n p-values generate z-scores that are potentially correlated and marginally arise from a mixture model of form (1), with $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, for some valid $\boldsymbol{\theta}^0$. Using the bins and realizations $x_i^\top = (x_{i1}, \ldots, x_{im})$ of each X_i ($i \in [n]$), we can write the marginal likelihood and log-marginal likelihood functions under the mixture model approximation for the z-scores as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\int_{B_j} f(z; \boldsymbol{\theta}) \, \mathrm{d}z \right]^{x_{ij}}$$

and

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \log \int_{B_j} f(z;\boldsymbol{\theta})\,\mathrm{d}z. \tag{2}$$

Write the MML estimator for $\boldsymbol{\theta}^0$ that is obtained from n z-scores as $\hat{\boldsymbol{\theta}}_n$. We can define $\hat{\boldsymbol{\theta}}_n$ as a suitable root of the score equation $\nabla l = \mathbf{0}$, where ∇ is the gradient operator and $\mathbf{0}$ is the zero vector.

The marginal likelihood function is simply an approximation to the likelihood that is constructed under an assumption of independence between the observations X_i (cf. Varin, 2008). In light of not knowing what the true dependence structure between the observations is, the marginal likelihood function can be seen as a quasi-likelihood construction in sense of White (1982). The purpose of a quasi-likelihood construction is to make use of an approximation that is close enough to the true data generative process so that meaningful inference can be drawn. Here its use is to avoid the need to declare an explicit model for potential correlation structures between the observations.

The EM algorithm for MML estimation is the context of this article is provided in Supplementary Materials Section 1. The consistency of the MML estimator is also established in the same section.

2.3 Empirical Bayes-based FDR control

Upon estimation of the parameter vector $\boldsymbol{\theta}^0$ via the MML estimator $\hat{\boldsymbol{\theta}}_n$, we can follow the approach of McLachlan et al. (2006) in order to implement EB-based FDR control of the experiment. That is, consider the event $\{H_i \text{ is null } | Z_i = z_i\}$, for each $i \in [n]$. Via Bayes' rule and the MML estimator $\hat{\boldsymbol{\theta}}_n$, we can estimate the probability of the aforementioned event via the expression

$$\hat{\mathbb{P}}(H_i \text{ is null } | Z_i = z_i) = \frac{\hat{\pi}_0 \phi\left(z_i; \hat{\mu}_0, \hat{\sigma}_0^2\right)}{f\left(z_i; \hat{\boldsymbol{\theta}}_n\right)} = \tau\left(z_i; \hat{\boldsymbol{\theta}}\right). \tag{3}$$

Using (3), we can then define the rejection rule

$$r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = \begin{cases} 1, & \text{if } \tau\left(z_i; \hat{\boldsymbol{\theta}}_n\right) \leq c \\ 0, & \text{otherwise,} \end{cases}$$

where $c \in [0,1]$. Here $r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = 1$ if the null hypothesis of H_i is rejected (i.e. H_i is declared significant) and 0 otherwise.

Let the marginal FDR be defined as $m\text{FDR} = \mathbb{E}N_{01}/\mathbb{E}N_R$. We can estimate the mFDR of an experiment via the expression

$$\widehat{m\text{FDR}} = \frac{\sum_{i=1}^{n} \tau\left(z_i; \hat{\boldsymbol{\theta}}_n\right) \mathbb{I}\left(r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = 1\right)}{\sum_{i=1}^{n} \mathbb{I}\left(r\left(z_i; \hat{\boldsymbol{\theta}}_n, c\right) = 1\right)}, \tag{4}$$

which we can prove to converge to the mFDR in probability, under M-dependence (cf. Nguyen et al., 2014, Thm. 1). Subsequently, we can also demonstrate that for large n, the mFDR approaches the FDR (cf. Nguyen et al., 2014, Thm. 2).

Notice that $m\text{FDR} = m\text{FDR}(c)$ is a function of the threshold c. Using the thresholding value, we can approximately control the FDR at any desired level β by setting the threshold c using the rule

$$c_\beta = \arg\max \left\{ c \in [0,1] : \widehat{m\text{FDR}}(c) \leq \beta \right\}. \tag{5}$$

2.4 Choosing the binning scheme

Thus far in discussing the binned estimation of the z-score distribution f, we have assumed that the bin cutoffs $b_1, ..., b_{m-1}$ are predetermined. When the p-values are censored into intervals (a_{j-1}, a_j), for $j \in [m] \cup \{0\}$, as describe at the beginning of the section, we may take the values a_j to inform our bin cutoffs $b_1, \ldots, b_{m-1}$. This can be done by computing the probit transformation of each of the cutoffs. That is, we can set $b_0 = -\infty$, $b_1 = \Phi^{-1}(1 - a_{m-1})$, $b_2 = \Phi^{-1}(1 - a_{m-2}), \ldots, b_{m-1} = \Phi^{-1}(1 - a_1)$. Thus the bin cutoffs are implicitly given by the censoring and thus the problem does not require the user to make a choice regarding the binning scheme, similar to the situation originally encountered in McLachlan and Jones (1988).

When the p-values are quantized or when they are discretely distributed, we must make a non-trivial decision regarding the binning scheme to use. A simple approach to the choice of binning scheme is to use the techniques underlying optimal histogram smoothing on the finite z-scores. In R, there are several optimal histogram smoothing techniques that are deployed in the default hist function. These include the fixed bin width methods of Sturges (1926), Scott (1979), and Freedman and Diaconis (1981).

Under the methods of Sturges (1926), Scott (1979), and Freedman and Diaconis (1981), the number of bins is taken to be $m = \lceil \log_2 n \rceil + 1$,

$$m = \lceil (\text{Range}/h) \rceil \text{ with } h = 2 \times \text{IQR}/n^{1/3},$$

and

$$m = \lceil (\text{Range}/h) \rceil \text{ with } h = 3.5 \times s/n^{1/3},$$

respectively. Here, $\lceil \cdot \rceil$ is the ceiling operator, and Range, IQR, and s are the sample range, interquartile range, and standard deviation, respectively. We compare the effectiveness of each of the binning approaches in the next section.

The binning of data, or the approximation of density functions via histograms, is a nontrivial problem that extends beyond the scope of this article. There is an abundance of methods for data binning that are available within the statistical and machine learning literature. Any of such methods can be used in place of the ones that we have suggested. For example, see the papers of Wand (1997) and Birge and Rozenholc (2006) regarding alternative fixed bin width methods. Examples of variable bin width methods can be found in the works of Kontkanen and Myllymaki (2007) and Denby and Mallows (2009). Further approaches can be found within the references of the cited articles.

3 An integer encoding example

To demonstrate the effects of grouping on p-values, we use the effects of integer encoding of such values as an example. Table 1 of Larobina and Murino (2014) provides a summary of the possible data compression schemes that can be applied when storing data in the ANALYZE, DICOM, MINC, or NIFTI formats. The possible integer storage schemes available for ANALYZE are 8-bits unsigned, or 16 and 32-bits signed. For DICOM, the available schemes are 8, 16, and 32-bits signed or unsigned. For MINC, 8, 16, and 32-bits signed or unsigned, are available. Finally, NIFTI can store data as 8, 16, 32, or 64-bits signed or unsigned.

For reference, 8, 16, 32, and 64 binary bits unsigned can encode 256, 65536, 429496 7296, and 1.84E+19 ($a\text{E}b = a \times 10^b$) unique values, respectively. These numbers are doubled when signed encodings are used. In this article, we only consider integer compression in 8-bits or 16-bits signed and unsigned formats. This is because 32-bits and 64-bits can be used to encode single and double-precision floating points, respectively, which largely mitigate against the reduced precision problems that we discuss in this article.

3.1 Integer encoding of p-values

As noted earlier, we are largely concerned with large scale-hypothesis testing situations that arise from voxel-based experiments (cf. Ashburner and Friston, 2000). In such experiments, a hypothesis test is conducted at each voxel of an imaged volume. For statistical analyses, resulting volumes of p-values are generated. It is these volumes that are then stored, possibly in a reduced precision format, for dissemination or for storage.

Suppose that a γ-bits unsigned integer encoding is used, where $\gamma \in \mathbb{N}$. Note that a γ-bits signed integer encoding is effectively equivalent to a $(\gamma + 1)$-bits unsigned, for all intents and purposes. When the hypothesis testing data are stored as a p-value volume, we suppose that the data are stored such that the smallest integer value encodes the

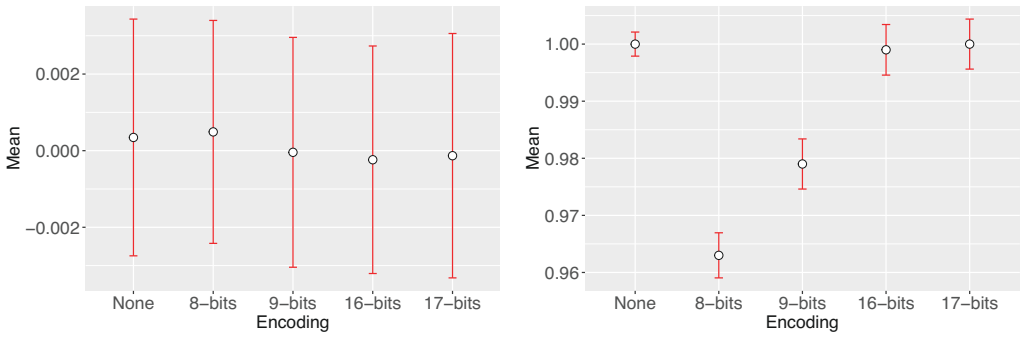
number zero and the largest integer value encodes the number one. The remainder of the integers are used to encode the unit interval at equally-spaced points. The encoding process then rounds the original p-values towards the nearest of these equally-spaced points. We refer to this approach as a γ-bits encoding. Under the storage protocols that we assess, $\gamma \in \{8, 9, 16, 17\}$ generate valid encodings. We note that our considered encoding scheme is only a simplified method of quantization. More complex encoding schemes are possible, such as those considered in Perlmutter et al. (1998).

3.2 The effect of integer encoding on the null distribution

Let $n = 10^6$, and for each $i \in [n]$, let H_i be a null hypothesis that is tested using a well-specified test resulting in a p-value P_i arising from a uniform distribution over the unit interval (cf. Dickhaus, 2014, Chap. 2). We simulate and encode the n p-values using γ-bits encodings, for all valid values of γ. The respective z-scores from each encoding scenario are computed, and the parameter elements of $f_0(z) = \phi\left(z; \mu_0, \sigma_0^2\right)$ are then estimated via ML estimation.

Here, we naively omit infinite z-scores. The process is repeated 100 times for each encoding rule. We also estimate the parameter elements of $f_0(z)$ for $n = 10^6$ z-scores that are obtained without encoding in order to provide a benchmark. All computations are conducted in R.

Figure 1 visualizes the results from the numerical study that is set up above. In the figure and elsewhere, we denote the estimate/estimator of any quantity θ as $\hat{\theta}$.



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

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

Figure 1: *Monte Carlo study regarding the estimation of $\mu_0 = 0$ and $\sigma_0^2 = 1$, 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.*

Theoretically, we would anticipate that there is no deviation away from a standard normal distribution when no encoding is introduced. This is exactly what we observe in Figure 1(a), where neither the average of the mean nor variance estimates are outside of a 95% confidence interval (i.e., approximately Mean $\pm 2 \times$ SE, where SE is the

standard error). In fact, only two encoding schemes (8 and 9-bits encodings) resulted in significant differences of any kind, from the anticipated estimated values. Further notes regarding the interpretation of Figure 1 appears in Section 2.1 of the Supplementary Materials.

3.3 The effect on the z-score distribution

Now suppose that the hypotheses H_i are generated from two populations, a null one with probability $\pi_0 = 0.8$, and an alternative one with probability $\pi_1 = 0.2$. Under the null hypothesis, we generate test statistics T_i from a standard normal distribution, and under the alternative, we generate test statistics from a normal distribution with mean $\mu_1 = 2$ and variance $\sigma_1^2 = 1$, instead. The p-values $P_i = 1 - \Phi(T_i)$, for testing the null that the test statistics are standard normal, are also computed. Again, we let $n = 10^6$.

Encoding of the p-values is again conducted under the protocol that are described in Section 3.1. We then compute z-scores and discard any infinite values. The parameter vector $\boldsymbol{\theta}$ is then estimated via ML estimation. The process is again repeated 100 times for each encoding type. ML estimation is conducted via the usual EM algorithm for finite mixtures of normal distributions via the normalmixEM2comp function from the package mixtools (Benaglia et al., 2009). The result of this numerical study is visualized in Figure 1 of the Supplementary Materials.

The estimated parameter elements were uniformly significantly different from the generative values for the model. As γ increases, we observe that the estimated values appear to approach the nominal parameter values. However, this approach appears to be slow and still leads to significantly incorrect estimates, even for the largest considered γ. A quantification of this incorrectness appears in Section 2.2 of the Supplementary Materials.

4 Assessment of the binned estimator

4.1 Accuracy of z-score distribution

We first repeat the experiment from Section 3.3, except instead of ML estimation via the normalmixEM2comp function from the package mixtools, we conduct MML estimation via the mix function from the package mixdist. The results from the experiment, using binning schemes obtained via the histogram binning techniques of Sturges (1926), Scott (1979), and Freedman and Diaconis (1981) are visualized in Figure 2 of the Supplementary Materials. Interpretation of appears in Section 3.1 of the Supplementary Materials.

We note that there is only one set of plots where we do not observe the uniform accuracy of the MML estimator, across the binning schemes that are applied. Under 8-bits encoding, we observe that only the Sturges-binned MML estimator yielded accurate es-

timates of the generative parameter elements. Both the Freedman-Diaconis (FD) and Scott-binned estimators resulted in significantly inaccurate estimates of the null proportion and alternative mean and variance parameters. We note that the Sturges binning leads to faster EM algorithm runtimes due to the fact that fewer numerical integrals are required in the E-step, as described in Section 1.1 of the Supplementary Materials. Since we do not observe any benefits from using FD or Scott-type binning in cases where all three methods yielded accurate estimates, we shall henceforth only consider the use of Sturges bins.

4.2 FDR control experiment

We perform a set of five numerical simulation scenarios, in order to assess the performance of the EB-based FDR control rule that is described in Section 2.3. These studies are denoted S1–S5, and will be described in the sequel.

In each of the scenarios, we generate $n = 10^6$ test statistics $T_1, \ldots, T_n$, with proportion $\pi_0 = 0.8$ that H_i is null ($i \in [n]$). The generative distribution of T_i given H_i is null or alternative differs by the simulation study. However, under each studied scenario, the null hypothesis is assumed to be that T_i is standard normal, and thus p-values are computed as $P_i = 1 - \Phi(T_i)$.

The p-values $P_1, \ldots, P_n$ then undergo the various valid encodings that were previously considered. The EB-based FDR control method is then used to decide which of the hypotheses H_i are significant, at the FDR control level $\beta \in \{0.05, 0.10\}$, based only on the encoded p-values. We compute the false discovery proportion (FDP) and true positive proportion (TPP) from the experiment as measures of performance of FDR control and testing power. The measures FDP and TPP are defined as $\text{FDP} = N_{01}/N_R$ and $\text{TPP} = N_{11}/N_1$, where N_{11} is the number of false positives, N_R is the number of rejected hypotheses (declared significantly alternative), N_{11}, is the number of true positives, and N_1 is the number of alternative hypotheses from the simulated experiment. For each simulation scenario, the experiment is repeated $\text{Reps} = 100$ times and the performance measurements are averaged over the repetitions.

For comparison, we also perform FDR control using the popular methods of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), which we denote as BH and BY, respectively. We also compare our EB-based FDR control to the EB-related FDR control technique of Storey (2002), which is commonly referred to as q-values. We implement the BH and BY methods via the base R `p.adjust` function. The q-values technique is implemented via the `qvalue` package (Storey et al., 2015). Scripts for conducting studies S1–S5 are available at `https://github.com/hiendn/FDR_for_grouped_P_values`.

4.3 Simulation scenarios

In Scenario S1, we independently generate T_i from a standard normal distribution, given that H_i is null, and from a normal distribution with mean 2 and variance 1, otherwise.

This scenario is identical to that which is studied Section 3.3.

Table 1: *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

We consider hypothesis tests that generate dependent test statistics in Scenarios S2 and S3. In S2 two first-order autoregressive sequences of n observations are generated. The null sequence is generated with mean coefficient 0, autoregressive coefficient 0.5, and normal errors with variances scaled so that the overall variance of the sequence is 1. The second chain is the same, except that the mean coefficient is 2 instead of zero. If H_i is null, then T_i is drawn from the first chain; otherwise T_i is drawn from the second chain. See Amemiya (1985, Sect. 5.2) regarding autoregressive models. Scenario S3 is exactly the same as Scenario S2, except that the autoregressive coefficient is set to -0.5 instead of 0.5.

In Scenario S4, we independently generate T_i from a normal distribution with mean 0.5 and variance 1, given that H_i is null, and from a normal distribution with mean 2.5 and variance 1, otherwise. This scenario is misspecified in the sense that the p-values P_i are not computed under the correct null hypothesis. Thus, the distribution of the P_i will not be uniform and thus the well-specified testing assumption of BH, BY, and q-values is not met.

Lastly, in Scenario S5, we independently generate T_i from a Student t-distribution with mean 0.5 and variance 1 and degrees of freedom 25, given that H_i, and from a Student t-distribution with mean 2.5 and variance 1 and degrees of freedom 25, otherwise. Justifications regarding the choices for the five scenarios appear in Section 3.2 of the Supplementary Materials.

4.4 Results

The results for Scenarios S1–S5 are reported in Tables 1–5 of the Supplementary Materials, respectively. We provide the results for S5 in the main text, as it can be viewed as the scenario that is most difficult and is thus most interest.

From Table 1, we observe that q-values is anti-conservative uniformly over all encoding types and FDR control levels in Scenario S5. Furthermore, BH was also uniformly anti-conservative when used to control the FDR at $\beta = 0.05$. The BH method also yielded anti-conservative control of the FDR at $\beta = 0.10$, when the data were encoded using p-type encodings. Both EB and BY were equally anti-conservative for control of FDR at $\beta = 0.05$, when the data were encoded using 8-bits or 9-bits encodings. However, the control at the $\beta = 0.10$ level from both methods for the two aforementioned encoding schemes were both equal and approximately at the correct rate. For all other encoding types, both EB and BY correctly controlled the FDR, for both levels of β. BY appeared more powerful than EB although by only a small amount.

The results above demonstrate that EB along with BY were somewhat more robust to misspecification and data compression via integer encoding than the two other tested methods. Thus, as we had anticipated, there was an observable practical effect to FDR mitigation via conventional methods when p-value data were observed on a discrete support. However, our EB method, and to an extent, the BY method, were able to mitigate against the negative effects of discretization induced by censoring, grouping, and truncation, and thus should be preferred over the other assessed methods in such settings.

For a discussion of results regarding Scenarios S1–S4, we direct the reader to Section 4.4 of the Supplementary Materials. From the results of Scenarios S1–S5, we can conclude that the EB method can correctly control the FDR when the tests were well-specified, and are also somewhat robust to misspecification, otherwise.

5 Example application

5.1 Description of data

Correlations between the structural properties of brain regions, as measured over a sample of subjects, are being increasingly studied as a means of understanding neurological

development (Li et al., 2013) and diseases (Seeley et al., 2009, Wheeler and Voineskos, 2014, Sharda et al., 2016). These correlation patterns, which are often referred to as structural covariance in the neuroimaging literature, are widely studied in humans (Alexander-Bloch, Giedd and Bullmore, 2013, Evans, 2013), as well as in animal models such as mice (Pagani, Bifone and Gozzi, 2016).

For our example application, we study neurological magnetic resonance imaging (MRI) data from a sample of 241 mice. The MRI sample of both female and male adult mice were obtained by taking the control data from a phenotyping study (Ellegood et al., 2015) in order to create a representative wildtype population with variability. All mice were scanned ex-vivo after perfusion with a gadolinium-based contrast agent, and all images were obtained at the same location (i.e. the Mouse Imaging Centre). Scanning was performed on a Varian 7T small animal MR scanner that was adapted for multiple mouse imaging.

The preparation and image acquisition followed a standard pipeline that is similar to the one described in Lerch, Sled and Henkelman (2010). Specifically, a T2-weighted fast-spin echo sequence was used to produce whole-brain images that have an isotropic resolution of 56 micrometers. After images were acquired, the data were corrected for distortions and then registered together by deformation towards a common nonlinear average. The registration pipeline included corrections for nonuniformities that were induced by radio frequency inhomogeneities or gradient-related eddy currents (Sled, Zijdenbos and Evans, 1998). The registered images had a volume of $x \times y \times z = 225 \times 320 \times 152$ voxels, of which $n = 2818191$ voxels corresponded to neurological matter. The exported data were stored in the MINC format.

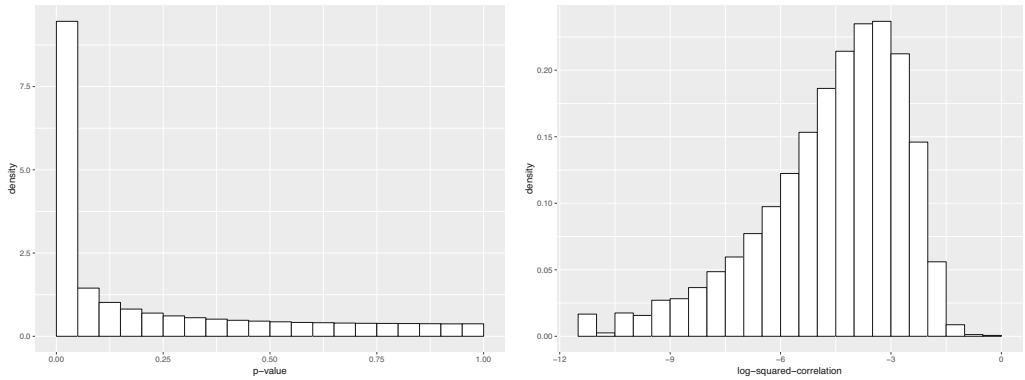
As an output, the registration process produces a set of Jacobian determinants that provide a measure of the extent in which a voxel from the average brain must expand or contract in order to match each of the individual brains of the sample. The Jacobian determinants field of each sample individual is thus a measure of local volume change. For further processing, the Jacobian determinants are log-transformed in order to reduce skewness.

5.2 Hypothesis testing

Upon attainment of the sample of 241 Jacobian determinant fields from the registered mice brain MRIs, we can assess whether or not the local volume change at any particular voxel is correlated with some region of interest. To do so, we select a "seed" voxel within the region of interest and compute the voxelwise sample (Pearson) correlation between the log-transformed Jacobian determinant of the seed voxel and those at every other voxel in the sample of MRIs. This correlation measure can then be used as a measure of structural covariance of the region of interest and the rest of the brain. In the past, structural covariance methods have been used to draw inference regarding a broad array of phenomena such as cortical thickness (Lerch et al., 2006), and cortical maturation and development (Raznahan et al., 2011).

Thus at each of the $n = 2818191$ voxels we computed a correlation coefficient. Using the correlation coefficients, we conducted voxelwise tests of the null hypothesis that the true correlation between the log-transformed Jacobian determinants of the seed voxel and voxel $i \in [n]$ is zero versus the two-sided alternative. The p-values of each test were computed using the Fisher z-transformation and normal approximation (Fisher, 1921).

Using the seed voxel at spatial location $(x,y,z) = (125,124,64)$ – within the bed nucleus of the stria terminalis – we conducted the hypothesis tests, as described above. Histograms of the p-values and log-squared correlation coefficients can be found in Figure 2. We note that the histogram of the log-squared correlation coefficients omits 35856 voxels that had zero correlation with the seed voxel. Further note that a correlation of one yields a log-squared coefficient of ≈ -0.69.



(a) Histogram of p-values. (b) Histogram of log-squared correlation coefficients.

Figure 2: *Histograms of p-values and log-squared correlation coefficients for the structural covariance experiment with seed voxel* $(x,y,z) = (125,124,64)$ *are presented in subplots (a) and (b), respectively.*

An inspection of Figure 2 reveals that the p-value distribution from the experiment deviates significantly from a uniform distribution. The magnitude of the deviation indicates that there may be a potentially large number of voxels that are strongly correlated with the seed voxel, and thus with the region of interest that the seed voxel represents. Using FDR control, we can attempt to identify these correlated voxels in a manner that limits the potential number of false discoveries that are made.

Using the unique function in R, we observed that there were only 66249 discrete and unique numerical values that made up the sample of p-values. These discrete values include zero and one, making up 311575 and 6 voxels of the p-value sample, respectively. Our observations indicate that the data were censored and grouped, at some stage in processing pipeline. It is difficult to tell how such incompleteness were induced, since there may have been multiple encodings of the data along the pipeline that has resulted in the final reported outputs. As such, from our earlier discussions, it would be prudent to apply our EB-based FDR control methodology, since it explicitly accounts for the encoded nature of the data. Furthermore, due to the mathematical approximation via the

use of the Fisher z-transformation as well as the omission of other variables that may contribute to the analysis such as covariates describing the mice (e.g. gender and model strain), the null hypothesis that the population correlation is equal to zero is likely to be misspecified. From Section 4.3, we have observed that the EB-based method is effective in such a setting.

5.3 FDR control

We firstly transform the p-values p_i to the z-scores $p_i = \Phi^{-1}(1-p_i)$, for each $i \in [n]$. A histogram of the z-scores that is obtained is presented in Figure 3. We note that the z-scores that are obtained from the 311581 with p-values equal to zero or one are omitted in this plot. There is a clear truncation of the histogram at the z-score value of 4.169 which corresponds to the smallest non-zero p-value of 1.53E-05.

Using the methods from Section 2, we fit the EB mixture model and obtain the parameter vector

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\top} &= \left(\hat{\pi}_0, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_1, \hat{\sigma}_1^2\right) \\ &= \left(0.5035, 0.5141, 1.200^2, 2.9568, 1.785^2\right), \end{aligned} \tag{6}$$

which corresponds to the mixture model,

$$f\left(z;\hat{\boldsymbol{\theta}}\right) = 0.5035 \times \phi\left(z; 0.5141, 1.200^2\right) + 0.4965 \times \phi\left(z; 2.9568, 1.785^2\right). \tag{7}$$

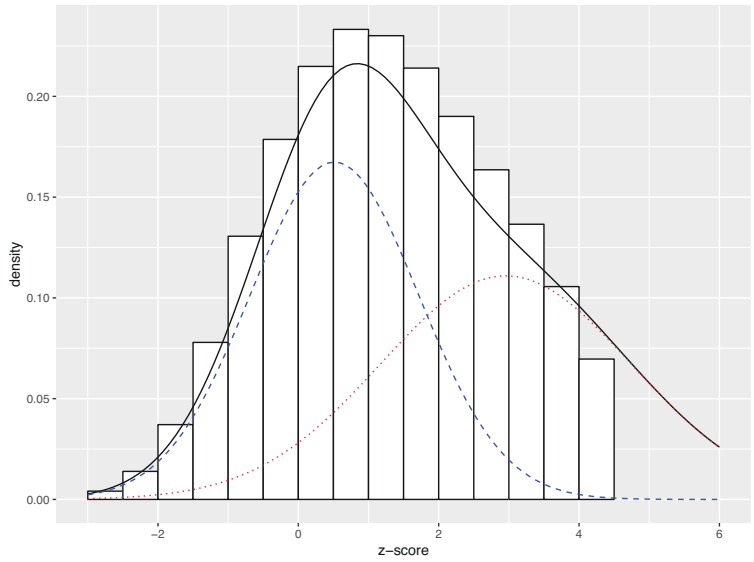


Figure 3: *The functions $f\left(\cdot;\hat{\boldsymbol{\theta}}\right)$, $\hat{\pi}_0\hat{f}_0$, and $\hat{\pi}_1\hat{f}_1$ are plotted with solid, dashed, and dotted lines, respectively.*

As in Sect. 4, we use the Sturges binning scheme that was previously described in Section 2.4. Let $\hat{f}_0(z) = \phi\left(z; 0.5141, 1.200^2\right)$ and $\hat{f}_1(z) = \phi\left(z; 2.9568, 1.785^2\right)$ be the estimates of f_0 and f_1, respectively. We visualize $f\left(\cdot; \hat{\boldsymbol{\theta}}\right)$, $\hat{\pi}_0\hat{f}_0$, and $\hat{\pi}_1\hat{f}_1$ together in Figure 3. A discussion regarding the goodness-of-fit of (7) is provided in Section 4.1 of the Supplementary Materials.

Upon inspection of Figure 3, we observe that mixture model (7) provides a good fit to the suggested curvature of the histogram. The estimated parameter vector from (6) indicates that the null distribution is significantly shifted to the right. This may be due to a combination of the effects of encoding and the effects of mathematical misspecification of the test and omission of covariates. We further observe that there is a large proportion (almost 50%) of potentially alternative hypotheses. Given such a high number, there is potentially for numerous false positives if we were to reject the null using the p-value (or z-score) alone. Thus, we require FDR control in order to make more careful inference.

Using Eqs (4) and (5), we controlled the estimated *m*FDR at the $\beta = 0.1$ level by setting the threshold $c_{0.1} = 0.09986$. This resulted in 608685 of the voxels being declared significantly correlated with the seed, under FDR control, which equates to 21.60%.

For comparison, using BH, BY, and q-values to control the FDR at the same $\beta = 0.1$ level, we obtain 1314429, 727102, and 1718143 significant voxels, respectively. Correspondingly, these numbers respectively translate to 46.64%, 25.80%, and 60.97% of the total number of hypotheses tested. Given the similarity of this testing scenario to simulation study S4, we can expect that the BH and q-values methods are grossly anti-conservative in their control and are would therefore would yield a greater FDR level than that which is desired. We observe, as in our simulations, that our method and BY tend to result in similar numbers of rejections. Whether one method or the other is overly conservative or anti-conservative in this case cannot be deduced without further assessment of the true significance of the rejected hypotheses.

Figure 4 displays visualizations of the significant voxels using our EB method at the perpendicular cross-sections intersecting the seed point $(x, y, z) = (125, 124, 64)$. Upon inspection of Figure 4 we observe that significant correlation with the seed vector appears to be exhibited across the brain. The displays A2 and A3 in Figure 4 further show that the correlation appears to be symmetric between the two hemispheres. Furthermore, the correlation patterns appear in contiguous and smooth regions.

The observations of whole-brain correlation with the bed nucleus of the stria terminalis are well supported in the literature. For example, similar connectivity observations were made by Dong et al. (2001) and Dong and Swanson (2006) in mouse studies, and by McMenamin and Pessoa (2015) and Torrisi et al. (2015) in human studies.

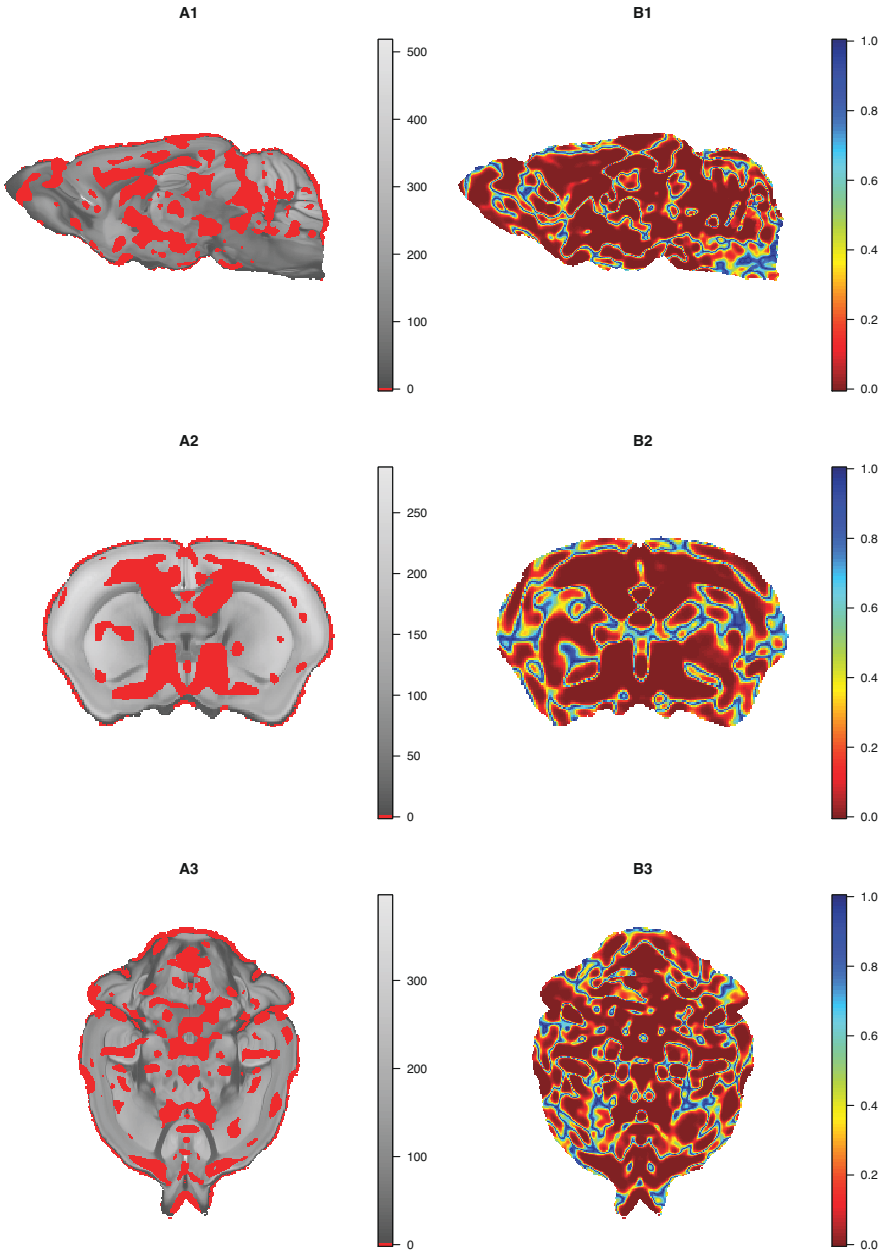


Figure 4: *A1 and B1 display the anatomic background MRI intensities and p-values for the $x = 125$ slice, respectively. Similarly A2 and B2 display the respective quantities for the $y = 124$ slice, and A3 and B3 display the respective quantities for the $z = 64$ slice. In A1–A3, red voxels indicate those that are significant when controlled at the $\beta = 0.1$ FDR level.*

6 Conclusions

We have presented an EB-based FDR control method for the mitigation of false positive results in multiple simultaneous hypothesis testing scenarios where only p-values are available from the hypothesis tests, and when these p-values are distributed on a discrete support. Due to the nature of the construction of our method, it is robust to situations where the hypothesis tests are also misspecified or when there may be omitted covariates that have not been included in the testing procedures for regression models.

In order to handle the discretization induced by censoring, grouping, or quantization of p-value data, we utilized a finite mixture model that can be estimated from binned data. We proved that the parameter vector of the mixture model can also be estimated consistently, even when the testing data may be correlated. A simulation study was used to demonstrate that our methodology was competitive with some popular methods in well-specified testing scenarios, and outperformed these methods when the testing data arise from misspecified tests.

Finally a brain imaging study of mice was conducted to demonstrate our methodology in practice. The study constituted a whole-brain voxel-based study of connectivity to the bed nucleus of the stria terminalis, consisting of $n = 2818191$ tests. The p-values for the study were obtained from a complex pipeline that resulted in a set of quantized values, which included zeros and ones. Furthermore, the p-values were correlated (due to the spatial nature of imaging and subsequent processing) and the hypothesis tests were conducted under mathematical assumptions that may have lead to misspecification. As such, the use of our methodology was most suitable for the study. As a result of the study, we found whole-brain correlation patterns that were consistent with those found in the literature.

Conducting FDR control when p-values are distributed on a discrete support, such as when the values are incompletely observed or when tests are conducted via Monte Carlo or permutation schemes, is an interesting inferential problem and requires careful attention. Our developed methodology provides a simple and robust solution when performing inference with such p-value data.

Acknowledgements

HDN is funded by Australian Research Council project DE170101134. GJM was funded partially by the Australian Government through the Australian Research Council (project numbers IC170100035, DP170100907). The authors are also thankful for the many enlightening and useful comments from the Editorial Board and from the Referees that have greatly improved the expositional quality of the paper.

Supplementary Materials

The Supplementary Materials for the article can be found online at `https://github.com/hiendn/FDR_for_grouped_P_values`.

References

Alexander-Bloch, A., Giedd, J. N. and Bullmore, E. (2013). Imaging structural co-variance between human brain regions. *Nature Reviews Neuroscience*, 14, 322–336.

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge: Harvard University Press.

Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry—the methods. *NeuroImage*, 11, 805–821.

Barreto, H. and Howland, F. M. (2006). *Introductory Econometrics Using Monte Carlo Simulation with Microsoft Excel*. Cambridge: Cambridge University Press.

Benaglia, T., Chauveau, D., Hunter, D. R. and Young, D. S. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32, 1–29.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society B*, 72, 405–416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57, 289–300.

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

Bennett, C. M., Baird, A. A., Miller, M. B. and Wolford, G. L. (2009). Neuro correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. In *15th Annual meeting of the Organization for Human Brain Mapping*.

Bidgood, W. D., Horri, S. C., Prior, F. W. and Van Syckle, D. E. (1997). Understanding and using DICOM, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association*, 4, 199–212.

Birge, L. and Rozenholc, Y. (2006). How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, 10, 24–45.

Cox, R. W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C. J., Lancaster, J. L., Rex, D. E., Smith, S. M., Woodward, J. B. and Strother, S. C. (2004). A (sort of) new image data format standard: Nifti-1. *Neuroimage*, 22, e1440.

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.

Denby, L. and Mallows, C. (2009). Variations on the histogram. *Journal of Computational and Graphical Statistics*, 18, 21–31.

Dickhaus, T. (2014). *Simultaneous Statistical Inference: With Applications in the Life Sciences*. New York: Springer.

Dong, H.-W., Pretrovich, G. D., Watts, A. G. and Swanson, L. W. (2001). Basic organization of projections from the oval and fusiform nuclei of the bed nuclei of the stria terminalis in adult rat brain. *Journal of Comparative Neurology*, 436, 430–455.

Dong, H.-W. and Swanson, L. R. (2006). Projections from bed nuclei of the stria terminalis, anteromedial area: cerebral hemisphere integration of neuroendocrine, autonomic, and behavioral aspects of energy balance. *Journal of Comparative Neurology*, 494, 142–178.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99, 96–104.

Efron, B. (2010). *Large-scale Inference*. Cambridge: Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151–1160.

Ellegood, J., Anagnostou, E., Babineau, B., Crawley, J., Lin, L., Genestine, M., Dicicco-Bloom, E., Lai, J., Foster, J., Penagarikano, O., Geshwind, H., Pacey, L. K., Hampson, D. R., Laliberte, C. L., Mills, A. A., Tam, E., Osborne, L. R., Kouser, M., Espinosa-Becerra, F., Xuan, Z., Powell, M., Raznahan, A., Robins, D. M., Nakai, N., Nakatani, J., Takumi, T., van Eede, M. C., Kerr, T. M., Muller, C., Blakely, R. D., Veenstra-VanderWeele, J., Henkelman, R. M. and Lerch, J. P. (2015). Clustering autism: using neuroanatomic difference in 26 mouse models to gain insight into the heterogeneity. *Molecular Psychiatry*, 20, 118–125.

Evans, A. C. (2013). Networks of anatomical covariance. *Neuroimage*, 80, 489–504.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3–32.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 453–476.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15, 870–878.

Gersho, A. and Gray, R. M. (1992). *Vector Quantization and Signal Compression*. New York: Springer.

Habiger, J. D. and Pena, E. A. (2011). Randomised P-values and nonparametric procedures in multiple testing. *Journal of Nonparametric Statistics*, 23, 583–604.

Kontkanen, P. and Myllymaki, P. (2007). MDL histogram density estimation. In *Artificial Intelligence and Statistics* (pp. 219–226).

Korn, E. L., Troendle, J. F., McShane, L. M. and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Statistical Planning and Inference*, 124, 379–398.

Larobina, M. and Murino, L. (2014). Medical image file formats. *Journal of Digital Imaging*, 27, 200–206.

Lerch, J. P., Sled, J. G. and Henkelman, R. M. (2010). *Magnetic Resonance Neuroimaging*, chapter MRI phenotyping of genetically altered mice, (pp. 349–361). Springer: New York.

Lerch, J. P., Worsley, K., Shaw, W. P., Greenstein, D. K., Lenroot, K. L., Giedd, J. and Evans, A. C. (2006). Mapping anatomic correlations across cerebral cortex (MACACC) using cortical thickness from MRI. *NeuroImage*, 31, 993–1003.

Li, X., Pu, F., Fan, Y., Niu, H., Li, S. and Li, D. (2013). Age-related changes in brain structural covariance networks. *Frontiers in Human Neuroscience*, 7, 98.

MacDonald, P. D. M. and Du, J. (2012). *mixdist: Finite Mixture Distribution Models*. Comprehensive R Archive Network.

McLachlan, G. J., Bean, R. W. and Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22, 1608–1615.

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

McMenamin, B. W. and Pessoa, L. (2015). Discovering networks altered by potential threat ("anxiety") using quadratic discriminant analysis. *Neuroimage*, 116, 1–9.

Moschitta, A., Schoukens, J. and Carbone, P. (2015). Information and statistical efficiency when quantizing noisy DC values. *IEEE Transactions on Instrumentation and Measurement*, 64, 308–317.

Nguyen, H. D., McLachlan, G. J., Cherbuin, N. and Janke, A. L. (2014). False discovery rate control in magnetic resonance imaging studies via Markov random fields. *IEEE Transactions on Medical Imaging*, 33, 1735–1748.

Pagani, M., Bifone, A. and Gozzi, A. (2016). Structural covariance networks in the mouse brain. *Neuroimage*, 129, 55–63.

Perlmutter, S. M., Cosman, P. C., Tseng, C.-W., Olshen, R. A., Grey, R. M., Li, K. C. P. and Bergin, C. J. (1998). Medical image compression and vector quantization. *Statistical Science*, 13, 30–53.

Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values whenn permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9, 1–12.

Pollard, C. S. and van der Laan, M. J. (2004). Choice of a null distribution in resampling-based multiple testing. *Statistical Planning and Inference*, 125, 85–100.

R Core Team (2016). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

Raznahan, A., Lerch, J. P., Lee, N., Greenstein, D., Wallace, G. L., Stockman, M., Clasen, L., Shaw, P. W. and Giedd, J. N. (2011). Patterns of coordinated anatomical change in human cortical development: a longitudinal neuroimaging study of maturational coupling. *Neuron*, 72, 873–884.

Robb, R. A., Hanson, D. P., Karwoski, R. A., Larson, A. G., Workman, E. L. and Stacy, M. C. (1989). Analyze: a comprehensive, operator-interactive software package for multidimensional medical image display and analysis. *Computerized Medical Imaging and Graphics*, 13, 433–454.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66, 605–610.

Seeley, W. W., Zhou, R. K. C. J., Miller, B. L. and Greicius, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron*, 62, 42–52.

Sharda, M., Khundrakpam, B. S., Evans, A. C. and Singh, N. C. (2016). Disruption of structural covariance networks for language in autism is modulated by verbal ability. *Brain Structure and FUnction*, 221, 1017–1032.

Sled, J. G., Zijdenbos, A. P. and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17, 87–97.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, 64, 479–498.

Storey, J. D., Bass, A. J., Dabney, A. and Robinson, D. (2015). *qvalue: Q-value estimation for false discovery rate control*.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21, 65–66.

Torrisi, S., O'Connell, K., Davis, A., Reynolds, R., Balderston, N., Fudge, J. L., Grillon, C. and Ernst, M. (2015). Resting state connectivity of the bed nucleus of the stria terminalis at ultra-high field. *Human Brain Mapping*, 36, 4076–4088.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society B*, 38, 290–295.

van der Laan, M. J. and Hubbard, A. E. (2006). Quantile-function based null distribution in resampling based multiple testing. *Statistical Applications in Genetics and Molecular Biology*, 5, 14.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92, 1–28.

Vincent, R. D., Janke, A., Sled, J. G., Baghdadi, L., Neelin, P. and Evans, A. C. (2003). MINC 2.0: a modality independent format for multidimensional medical images. In *10th Annual Meeting of the Organization for Human Brain Mapping*.

Wand, M. P. (1997). Data-Based Choice of Histogram Bin Width. *The American Statistician*, 51, 59–64.

Wheeler, A. L. and Voineskos, A. N. (2014). A review of structural neuroimaging in schizophrenia: from connectivity to connectomics. *Frontiers in Human Neuroscience*, 8, 653.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. and Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92, 381–397.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Statistical Planning and Inference*, 82, 171–196.