An Implementation of Empirical Bayesian Inference and Non-Null Bootstrapping for Threshold Selection and Power Estimation in Multiple and Single Statistical Testing

Bahman Nasseroleslami*[1]

[1]Academic Unit of Neurology, Trinity College Dublin, the University of Dublin, Dublin, Ireland.

June 8, 2018

Abstract

The majority of conclusions and interpretations in quantitative sciences such as neuroscience are based on statistical tests. However, the statistical inferences commonly rely on the p-values, but not on more expressive measures such as posterior probabilities, false discovery rates (FDR) and statistical power (1 - β). The aim of this report is to make these statistical measures further accessible in single and multiple statistical testing. For multiple testing, the Empirical Bayesian Inference (Efron et al., 2001; Efron, 2007) was implemented using non-parametric test statistics (Area Under the Curve of the Receiving Operator Characteristics Curve or Spearman's rank correlation) and Gaussian Mixture Model estimation of the probability density function of the original and bootstrapped data. For single statistical tests, the same test statistics are used to construct and estimate the null and non-null probability density functions using bootstrapping under null and non-null grouping assumptions. Simulations were used to test the reliability of the results under a wide range of conditions. The results show conformity to the real truth in the simulated conditions, which is held under various conditions imposed on the simulation data. The open-source MATLAB codes are provided and the utility of the approach has been discussed for real-world electroencephalographic signals. This implementation of Empirical Bayesian Inference and informed selection of statistical thresholds are expected to facilitate more realistic scientific deductions in versatile fields, especially in neuroscience, neural signal analysis and neuro-imaging.

1 Introduction

The majority, if not all, of the conclusions and interpretations in quantitative sciences, especially in neuroscience and neuro-imaging are based on statistical tests. While the traditional hypothesis tests based on p-values are still dominant, there has been legitimate remarks on the need for more reliable and thorough statistical procedures and practices (Nuzzo, 2014). It is therefore vital that more meaningful statistical measures be accessible for statistical inference, namely: posterior probabilities, false discovery rates (FDR), and statistical power ($1 - \beta$). These measures are especially useful in statistical inferences involving high-dimensional data in neural signal connectivity or imagining.

Empirical Bayesian Inference (EBI) has shown promise in large-scale between-group comparisons (Efron, 2007b, 2004), especially in genomics (Efron, Tibshirani, Storey, and Tusher, 2001) and to some extent in the applications of neuroelectric signal and connectivity analysis (Singh, Asoh, Takeda, and Phillips, 2015). In EBI, constant prior probabilities are estimated from the data in large-scale multi-variable inferences or hypothesis testing and these priors are subsequently used to find the posterior probabilities using the estimated probability density functions of the pooled test statistics and the null distribution. It is possible to relate the posterior probabilities to frequentist concepts such as False Discovery Rate, FDR (Benjamini and Hochberg, 1995; Benjamini, Krieger, and Yekutieli, 2006), as well as power. While the theoretical framework is adequately established, the existing numerical implementations Efron, 2007b are optimised and suitable only for specialised applications (i.e. statistical genetics, where only a small fraction of the tests are real findings), and some of the essential measures such as FDR and power are not immediately available for informed threshold selection. From a practical viewpoint, the existing software package [locfdr] in R (**r_core_team_r:_2016**) may requires selection of several parameters and not immediately available for neuro-electro-magnetic signal and connectivity analysis in packages such

* Room 5.43, Trinity Biomedical Sciences Institute, 152-160 Pearse Street, Dublin D02 R590, Ireland. E-Mail: nasserob@tcd.ie

as FieldTrip **oostenveld_fieldtrip:_2010** or for neuroimaging analysis in packages such as SPM. Consequently, there is the need for new implementations to facilitate the application of EBI that work in wider range of situations (e.g. small or large proportions of test variables belonging to the affected group), to more explicitly relate the posterior probabilities to FDR and power (allowing informed decision on threshold selection), and to intrinsically account for data with non-normal distributions.

Such informed selection of statistical threshold is also challenging in complex statistical inferences (e.g. with non-normal data distributions) involving single or only a few comparisons or inferences. It would be desirable to similarly select a threshold value for the test statistic that corresponds to a known combination of Type I (α) and Type II (β) errors in a single comparison.

Here, we address these needs using an implementation of EBI using non-parametric test statistics, Gaussian Mixture Models and null bootstrapping. This implementation readily handles one-sample, two-sample (between-group comparison) and correlation problems in multi-dimensional data with arbitrary distributions, which is usable for a wide range of applications. Furthermore, for threshold selection in univariate testing (in the absence of prior probabilities), the non-null distribution is estimated using a non-null bootstrapping. This approach approximates the non-null probability density functions in order to enable the threshold selection for a desired combination of α and β values, regardless of the distribution of data.

2 Methods

2.1 Empirical Bayesian Inference (EBI) for Multiple Inferences

2.1.1 EBI framework

EBI, initially used in genomic applications (Efron et al., 2001) was subsequently expanded theoretically and in terms of computational implementation (Efron, 2004, 2007a). Here, the fundamentals are briefly explained.

Suppose $X_{ij}, i = 1...m, j = 1...N$ represents N variables or N-dimensional data sampled from m observations/subjects. The grouping information of data is represented by g_i, which is a binary choice (0 or 1) for a two-sample (2-group) comparison. Statistical testing, performed independently in each variables according to the grouping information (e.g. between-group comparisons), using the test statistic z_j, yields N values. The probability density function of z_j, i.e. the probability of data given the hypotheses, is denoted by:

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \tag{1}$$

where p_0 and p_1 are the prior probabilities of the null and non-null hypotheses ($p_1 = 1 - p_0$) and $f_0(z)$ and $f_1(z)$ are the probability density functions of z under the null and non-null (grouping) assumptions, respectively. The posterior probability, i.e. the probability of the hypotheses given the data, are subsequently given by:

$$P_0(z) = p_0 f_0(z)/f(z) \tag{2}$$
$$P_1(z) = 1 - P_0(z) \tag{3}$$

Comparison of the posterior probability of the non-null hypothesis $P_1(z_j)$ against a threshold P_{crit} provides a Bayesian inference, as well as subsequent frequentist quantities such as local false discovery rate, $fdr_{loc}(z) = P_0(z)$, Type I error α, Type II error β, and the FDR value pertaining to the chosen P_{crit}. The classic EBI includes several stages for estimating the posterior probabilities. First, using a measure of between-group difference (e.g. Student's t-statistic or p-values) and transforming the values to normal (e.g. by inverse normal cumulative distribution function); second, estimation of $f(z)$ from the z_j histogram; third, estimating $f_0(z)$ by theoretical assumptions on the distribution of z_j or bootstrapping; forth, estimation of p_0, usually through the assumption that $f_1(argmax_z f_0(z)) = 0$; and finally, $P_1(z)$ is found by (2) - (3). Here, we will explain the details of each stage for the new implementation of the EBI.

2.1.2 Test Statistic

Instead of using Student's t-statistic or a p-value which reflects the difference of the means of two groups, here a non-parametric measure was used as a test statistic. The Area Under the Receiver Operating Curve (AUROC), A, is closely related to the Mann-Whitney U statistic. AUROC is the probability of data in one group being larger (or smaller) than the other group ($Pr(X_{g=0}) < Pr(X_{g=1})$); hence it is considerably independent of the distribution of the original data, as well as any measure (e.g. mean or median) for comparison. This has been thoroughly discussed elsewhere (Zhou, McClish, and Obuchowski, 2009). AUROC was therefore taken as the test statistic for comparing the m data points in the two groups for each comparison of the N variables.

It is noteworthy that while AUROC is independent of the underlying distribution of the data, the data in all N variables should come from the same null and alternative distributions. The AUROC distributions depend on the number of data in the first and second group, as well as the distribution of the original data. Therefore, the number of data points in each group should be (ideally) the same for all variables, and all of them should come from the same arbitrary distribution (e.g. normal, Beta, Gamma, uniform). This is especially relevant as the curve fitting for null and mixed density

functions, as well as bootstrapping (for construction of null data) rely on pooling data from all variables.

While not essential to transform the AUROC values to normal, it's beneficial to do so computationally. The transformation to normal allows the use of more robust estimation methods such as Gaussian kernel methods that work best in unbounded domain, rather than in bounded ([0 1]) domains. As the AUROC distribution is bounded between 0 and 1, with the expected value of 0.5 under null, a mapping of $2(.) - 1$ combined with Fisher's Z-transform, $z_i = 0.5log_e((.)+1) - 0.5log_e((.)-1)$ (Fisher, 1915; Zhou et al., 2009), can approximately map the data to normal (Zhou et al., 2009; Qin and Hotilovac, 2007):

$$z_j = log_e(\frac{\epsilon + A_j}{\epsilon + 1 - A_j}) \quad (4)$$

$$\sigma^2(z_j) = \frac{\sigma^2(A_j)}{(\epsilon + A_j)^2(\epsilon + 1 - A_j)^2} \quad (5)$$

where $\sigma^2(.)$ indicates the variance. The tuning parameter ϵ which is added to the classic definition here, serves to limit the extreme z values; hence, facilitating numerical integration in later steps. To limit the z values to $[-10, 10]$, $\epsilon = 4.55e - 5$ was adopted here. To avoid sharp distributions where $AUC = 1$ and $z = 10$ ($AUC = -1$ and $z = -10$), the values larger than 9 (smaller than -9) were redistributed to a truncated normal distribution. The redistribution of h extreme values larger than 9, assigned the ith value to $iCDF_{\mathcal{N}(9,0.25,5,13)}((i - 0.5)/h)$, where $iCDF$ is the inverse cumulative distribution function and $\mathcal{N}(9, 0.25, 5, 13)$, is the normal distribution with mean 9, standard deviation 0.25, truncated between $[5, 13]$. The values smaller than -9 were similarly reassigned.

2.1.3 Estimating the $f(z)$ Histogram

Gaussian Mixture Model (GMM) distributions (McLachlan and Peel, 2004) were used to estimate the probability density $f(z)$, using the pool of $z_j, j = 1...N$. Using maximum likelihood estimates of GMM parameters, models with increasing number of Gaussian kernels were set for fitting z_j values. The model with minimum Akaike Information Criterion, AIK, (Akaike, 1974) was eventually considered as the preferred fit. This was concluded when the increasing number of kernels yielded 3 consecutive increases in the AIK. A similar approach has been previously used (Le, Pan, and Lin, 2003) in statistical genetics applications, but not in the context of EBI.

2.1.4 Estimating the Null Distribution $f_0(z)$

For robust estimation of the null distribution, the data labels g_i were set for B_0 times re-sampling with substitution; for each set of the obtained labels, the above-mentioned procedures used for the original data and labels were applied to yield A_j and subsequently the z_j values. The data from all the B_0 bootstraps and all the N tested variables were used for pooling to estimate the null distribution. For computational efficiency, it may be helpful to sparse the null distribution. In this implementation, when the null data points exceeded 20000, the null data were sorted and only the every sk data values were kept for GMM estimation (sk: the integer multiples of 20000 in the number of null data values). Using similar GMM estimation as for $f(z)$, the null distribution $f_0(z)$ was estimated.

2.1.5 Estimating the prior p_0

The approach used by EBI for estimation of the prior p_0, relies on the key assumption that at maximum (peak) value of $f_0(z)$, the value of $f_1(z)$ is zero. Due to the smooth and reliable estimation of $f_0(z)$ and $f(z)$ by AIK-guided GMM fits, we may directly use the values of the estimated probability density functions to find the prior p_0:

$$p_0 = \frac{f(z)}{f_0(z)} \Bigg|_{z=iCDF_{f_0(z)}(0.5))} \quad (6)$$

where $iCDF_{f_0(z)}(0.5)$ is the z value at which the Cumulative Density Function (CDF) of $f_0(z)$ is 0.5, which is the inverse CDF of 0.5, i.e. the median of the null data.

2.1.6 Estimating the Posterior $P_0(z)$

Given the estimates of p_0, $f_0(z)$ and $f(z)$, the calculation of posterior probabilities from (2) and (3) are straightforward. A bound between 0 and 1 was considered to protect against numerical instability at very small probability values.

2.1.7 Estimating FDR and Power ($1 - \beta$)

Following the calculation of p_0, p_1, $f_0(z)$, $f_1(z)$, $f(z)$, $P_0(z)$ and $P_1(z)$, the Type I error α, Type II error β, power (1-β) and FDR (q) can be found by numerical integrations. This is achieved by using a decision threshold value (see Section 2.1.8 for how this threshold is decided on) on either of these measures to infer which variables do or do not show an interesting effect. For a given set decision threshold, a criterion on the Posterior, P_{cr}, we may write:

$$\alpha(P_{cr}) = \int_{P_1(z) \geq P_{cr}} f_0(z)dz \quad (7)$$

$$\beta(P_{cr}) = 1 - \int_{P_1(z) \geq P_{cr}} f_1(z)dz \quad (8)$$

$$q(P_{cr}) = \frac{\int\limits_{P_1(z)\geq P_{cr}} P_0(z)f(z)dz}{\int\limits_{P_1(z)\geq P_{cr}} f(z)dz} \tag{9}$$

as the parameter ϵ in section 2.1.2 limits the values of z, the integration would suffice to take place in the range $[-20, 20]$. Additionally, the global values α_g and β_g show the sperability of the probability density distributions regardless of a chosen P_{cr}:

$$\alpha_g = \frac{\int_{-\infty}^{+\infty} P_1(z)f_0(z)dz}{\int_{-\infty}^{+\infty} f_0(z)dz} \tag{10}$$

$$\beta_g = \frac{\int_{-\infty}^{+\infty} P_0(z)f_1(z)dz}{\int_{-\infty}^{+\infty} f_1(z)dz} \tag{11}$$

2.1.8 Threshold Selection

The threshold selection and subsequent inference is driven by setting a P_{cr} value and comparing the values of $P_1(z)$ against this specific P_{cr} value, as required in the specific context of application. Alternatively, the availability of the computed values of α, β, q, and the detection ratio $\#\{P_1(z_j) \geq P_{cr}\}/N$ as a function of $P_1(z)$, allows setting the P_{cr} values that correspond to specific values of these measures.

2.2 Non-null Bootstrapping for Single Inference

The above-mentioned procedure is applicable for large-scale multiple testing, as this enables the estimation of empirical mixed density $f(z)$, the priors p_0 and p_1, and eventually the Posteriors $P_0(z)$ and $P_1(z)$. For single statistical testing ($N = 1$), similar stages can be followed in order to calculate the test statistic, estimate the histograms or probability density functions, and calculate the null distribution. However, it is not possible to estimate the priors; hence, these variables are not available. Notwithstanding, it is possible to estimate $f_1(z)$ by a different approach, namely the non-null bootstrapping which will make it possible to estimate α and β for a specific threshold, which would be a criterion on z (rather than on $P_1(z)$). This approach is explained below:

2.2.1 Non-Null Bootstrapping & Estimating $f_1(z)$

To estimate the non-null distribution $f_1(z)$ we may rely on bootstrapping when the grouping information of the data g_i are respected (in contrary to the commonly practised null bootstrapping that aims to find the null distribution where the grouping information/labels are not respected). For this purpose, the following re-sampling algorithm was used:

1. If m_{b0} is the number of observations in group $g = 0$, take m_{b0} samples (with substitution) from $\{X_{i.}|g_i = 0\}$ to build Ξ_{b0}.
2. If m_{b1} is the number of observations in group $g = 1$, ($m_{b0} + m_{b1} = m$), take m_{b1} samples (with substitution) from $\{X_{i.}|g_i = 1\}$ to build Ξ_{b1}.
3. find the A (AUC for ROC curve) and consequently the z value according to (4) using the obtained set of the bootstrapped data Ξ_{b0} and Ξ_{b1}.
4. Repeat this procedure (1-3) for B_1 times to obtain the needed samples of $z_b, b = 1...B_1$.

The GMM was then chosen to find the distribution of z_b values, which estimates $f_1(z)$.

2.2.2 Estimating α and Power ($1 - \beta$)

Similar to the calculation of α and β for a given P_{cr} value in (7) and (8), it is possible to use numerical integration to find the relationship between α and β. If $F_0(z)$ is the cumulative density function corresponding to $f_0(z)$, then for a given two-tail decision region $Z(\alpha)$,

$$Z(\alpha) = \{z|(1-F_0(z)) \leq \frac{\alpha}{2}\} \cup \{z|F_0(z) \leq \frac{\alpha}{2}\} \tag{12}$$

it is possible to describe β a as function of α:

$$\beta(\alpha) = 1 - \int\limits_{Z(\alpha)} f_1(z)dz \tag{13}$$

2.2.3 Threshold Selection

The threshold selection in single testing is straightforward, given the known relationship between α and β in (13).

2.3 Numerical Implementation & Simulations

The numerical programming for the proposed EBI implementation was performed in MATLAB (versions 2016b-2018a, Mathworks Inc., Natick, MA, USA). The Empirical Bayesian Inference Toolbox for MATLAB is publicly available at https://github.com/NeuroMotor-org/EBI and is licensed under BSD 3-Clause "New" or "Revised" License. To demonstrate the utility of the proposed implementation and to test its validity, it was applied to simulated data. Simulated data allow comparison of the performance measures to real truth, which is not available in real life applications. All simulations were performed in MATLAB. Different simulations were carried out as detailed below.

2.3.1 A Demonstrating Example

An example similar to applications in neural signal analysis and neuroimaging was considered. The simulation included $N_{var} = 2000$ variables, each with $m_0 = 20$ observations/subjects in first sample (e.g. controls), and $m_1 = 60$ observations/subjects in second sample (e.g. patients), totalling $m_j = 80$ observations/subjects. In control observations, all variables had a normal distribution $\mathcal{N}(0,1)$, while in the other group the first 1600 variables have the same $\mathcal{N}(0,1)$ distribution, the next 300 variables come from $\mathcal{N}(-1,1)$ and the last 100 variables were from $\mathcal{N}(1.5,1)$.

2.3.2 Comparison Against Real Truth

Using the simulated data from previous section, the β and FDR (or q) values were calculated as a function of the posterior threshold P_{cr} and were compared to the true values of β and FDR (q). The real values were found by using the original labels of the variables in the simulations. By comparing them to the detected labels by EBI, the true positive (TP), false positive (FP), false negative and true negative (TN) rates were calculated at each threshold, from which the the real β and FDR were found. Additionally, the same data underwent EBI analysis with a previous implementation of EBI (Efron, 2007b) in R (**r_core_team_r:_2016**), when used with the default value.

2.3.3 Performance Under Different Condition

Several simulations were performed to test the performance of the framework in a broader range of conditions. This controlled variation of the simulation condition can test and inform of the performance in real-life applications, which is not easy with typical experimental data due to the lack of a gold standard or real truth. The parameters for generation of simulated data and the application of the new EBI implementation included: $N_{var} = 200, 2000$, $p_0 = 0.25, 0.75$, $m_0 = 25, 100$, $m_1 = 25, 100$, Normal ($\sigma = 1$) vs. Beta ($a = 2$,$b = 10$) distribution types, difference or effect size (Cohen's $d = 0.2$, $d = 0.9$ for normal distributions or shift $d = 0.05$, $d = 0.09$ for Beta distributions). The estimated p_1, real FDR at expected value of 0.05, and real β at expected value 0.2 were compared across all the 64 simulation conditions. Each simulation condition was repeated 3 times to account for the non-deterministic nature of the implemented bootstrapping and estimation procedures.

2.3.4 Simulation of a Uni-Variate Example

To demonstrate the derivation of the $\alpha - \beta$ curve, a simple simulation with $m_0 = 15$ and $m_1 = 25$, random data with normal distributions for both groups ($\sigma = 1$), and shift value (Cohen's d) of 0.5 was considered.

3 Results

3.1 Example of Multiple Testing with EBI

Figure 1 exemplifies the generated results and a report for a typical simulated case as described in Section 2.3.1.Notice how the two probability density functions $f_0(z)$ and $f_1(z)$, as well as the prior p_1 are the essential components in giving rise to the Posterior distributions $P_0(z)$ and $P_1(z)$. In addition pay attention to how the choice of different threshold levels (color-coded based on the criteria) helps to choose an informed statistical threshold for inference based on the levels of FDR and power they afford.

3.2 Comparison of Typical Behaviour Against Truth

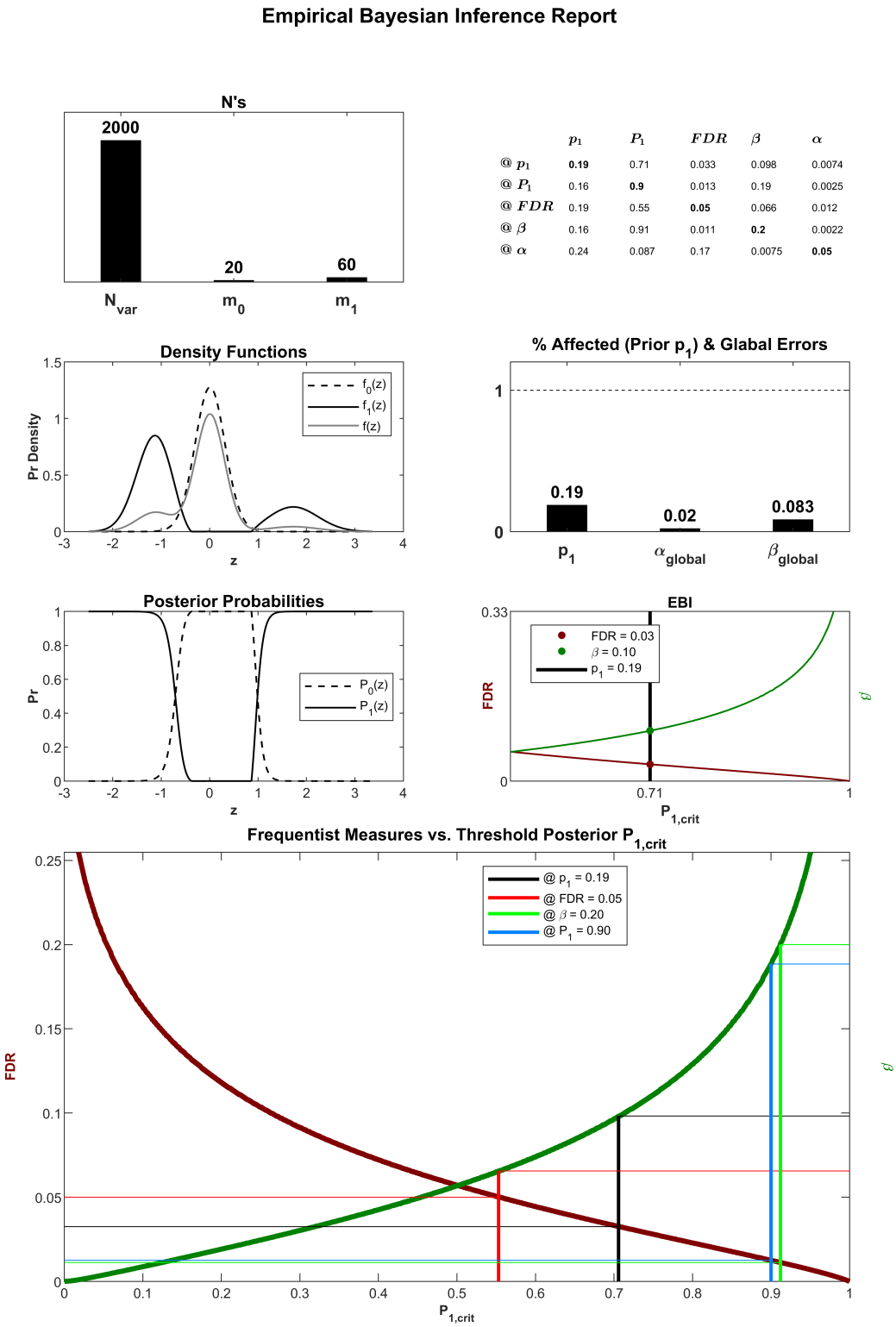
Figure 2 compares the estimated FDR and β as a function of the threshold values (P_1), against the real truth, using simulation labels as described in Section 2.3.2. Notice the similarity of real truth curves and estimates by the new implementation. In addition, as the simulated condition had a low prior p_1, the results from the previous *locfdr* implementation in R showed a good conformity to the real truth and the new implementation.

3.3 Performance of the EBI under Various Simulation Conditions

Figure 3 compares the estimated prior p_1 against the real values, as well as the real FDR and β values when estimated at nominal values at 0.05 and 0.2 (Described in Section 2.3.3). In the majority of conditions, the estimated measures were very close to the real values and there was negligible difference between the 5 different iterations of the simulation. The exception is at low effect sizes, combined with low number of observation/subjects and extreme prior values, where the estimation errors increase (possibly due to dissociations between the affected and non-affected variables and density functions). In the majority of the simulation cases the *locfdr* R package did not converge; therefore, the results were not included.

3.4 Example Single Testing

Figure 4 shows the correspondence of different α values to β values, for a an exemplary simulated data (Section 2.3.4).



	p_1	P_1	FDR	β	α
@ p_1	**0.19**	0.71	0.033	0.098	0.0074
@ P_1	0.16	**0.9**	0.013	0.19	0.0025
@ FDR	0.19	0.55	**0.05**	0.066	0.012
@ β	0.16	0.91	0.011	**0.2**	0.0022
@ α	0.24	0.087	0.17	0.0075	**0.05**

Figure 1. Report of an Empirical Bayesian Inference, applied on typical simulated data. The probability density functions of the mixed ($f(z)$) and null ($f_0(z)$) data, estimated from the original and permuted, as well as of the non-null ($f_1(z)$) data, estimated from $f(z)$ and $f_0(z)$ by Bayesian inference are plotted as a function of the z (transformed to z from original test statistics). Following the estimation of the fixed prior probability $p_1 = 0.19$ (the ratio of affected to non-affected variables), the Posterior probabilities for null ($P_0(z)$) and non-null ($P_1(z)$) are estimated. The $\alpha_{global} = 0.02$ and $\beta_{global} = 0.083$ show the global level of Type I and II errors regardless of a specific threshold, calculated from the probability density functions and Posterior Probabilities. The threshold selection is facilitated by the plots of FDR and β as a function of $P_{1,crit}$. The table indicates common criteria as a function of p_1, P_1, FDR, β and α (thresholds as bold diagonal values), and other corresponding values afforded by the selected criterion. For example, at $p_1 = 0.19$, the estimated FDR and β values will be 0.03 and 0.10 respectively. In the large plot, for each chosen criterion labelled by colour-coding in the legend, the projections to left and right indicate the afforded FRR and β by each criterion, respectively.

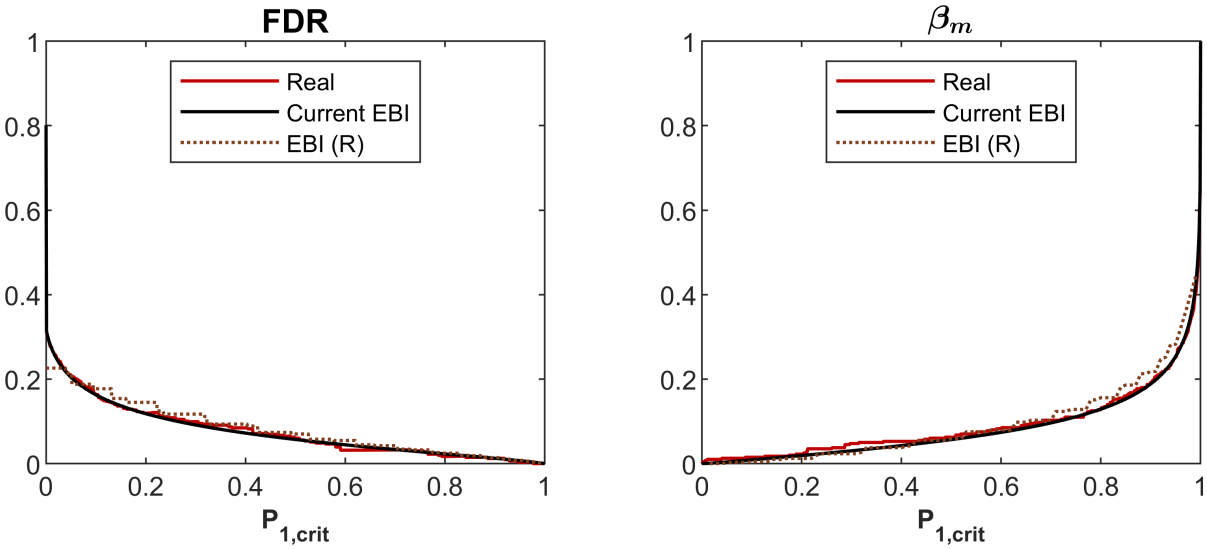


Figure 2. Comparison of FDR and β found by the EBI implemntation against the real truth in simulations. The data and conditions correspond to Figure 1. The dotted lines show the estimations from the *locfdr* package in R.

4 Discussion

While EBI (Efron et al., 2001; Efron, 2007b) provided a comprehensive theoretical framework for multivariate high-dimensional inference, the previous numerical implementation of EBI provided valid results only in limited conditions, namely, low prior p_1 values and for rather high threshold values. Importantly, the previous implementation required numerous adjustments and parameter selection. The new implementation eliminates the need for parameter tuning (especially by using AIK for GMD fitting) and allows the method to be used in broader range of conditions. Importantly, the statistical power is explicitly estimated and made available for inference.

4.1 Applications

The new approach suits applications involving neural signal analysis, such as electromyography (EMG), Electroencephalography (EEG), as well as neuroimaging, e.g. Magnetic Resonance Imaging (MRI). More specifically, spectral, time-frequency, as well as functional and effective connectivity analyses can benefit most from the new statistical implementation. In applications such as fMRI, the need for improved statistical inference has been explicitly emphasised by highlighting the limitations with existing techniques that lead to high false discovery rates (Eklund, Nichols, & Knutsson, 2016). The existing attempts to improve the statistical inference in EEG connectivity analysis (Singh, Asoh, & Phillips, 2011; Singh et al., 2015), have yield only partial success to date. Here, we used simulations to compare the EBI reports against the real truth. Importantly, in 2 recent studies, we showed that EBI is reasonably cross-validated against traditional frequentist methods. EBI was cross-validated against the correction of significance level α according to the number of principal components in data (Iyer et al., 2017) when testing the significance of EEG time series. Moreover, the inference of EBI was cross-validated against adaptive False Discovery Rates (aFDR) (Benjamini & Hochberg, 1995; Benjamini et al., 2006) in comparing the average EEG connectivity patterns between healthy individuals and patient groups (Nasseroleslami et al., 2017).

A unique advantage of EBI is its ability to implicitly account for potential positive and negative correlations that may be present in the data. It is therefore a suitable candidate for situations where positive or negative correlations exist in multi-dimensional data (e.g. EEG/MEG network connectivity analysis or structural or functional MR imaging). This is afforded by the way the individual z-values pertaining to each variable are aggregated (i.e. the independent calculation of the test scores) and by the chosen approach for the calculation of a null distribution that similarly corresponds to the same data with or without correlation structures (Efron, 2007a). The flexible estimation of null distribution from permuted data by GMM affords such flexibility for inference in rather broad conditions.

In applications where only simple statistical testing is required, the calculated FDR is an accurate estimation equivalent to the procedures for pooled multivariate permutation tests, which can be used without reference to Bayesian inference.

4.2 Limitations

The practical range for the number of variables is between 100-10000. The performance beyond this range degrades. This limitation originates from the EBI framework, rather than a specific numerical implementation.

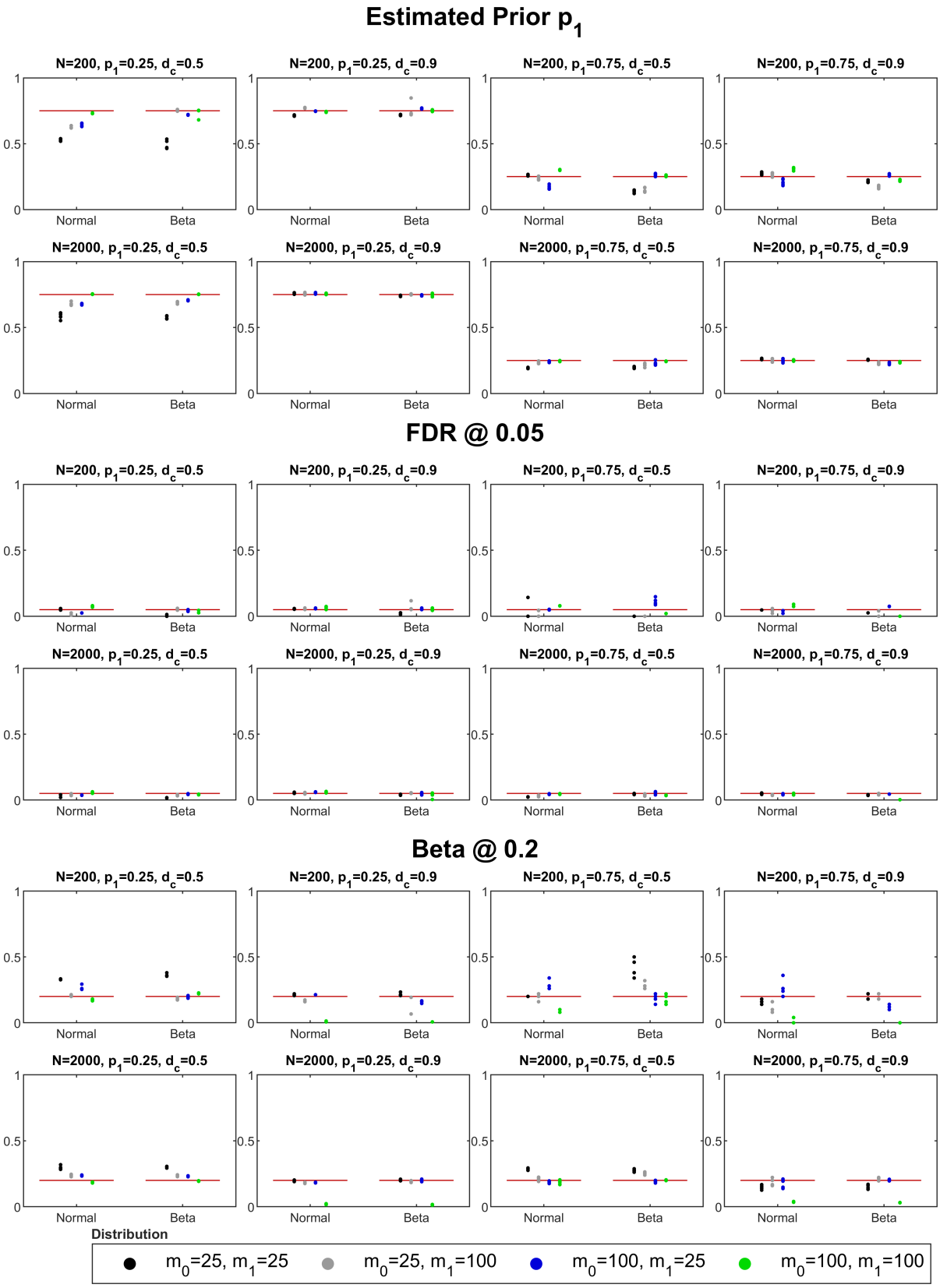


Figure 3. Comparison of estimated prior p_1 against real truth, and real FDR and β values estimated at nominal values 0.05 and 0.2 in simulations. See Section 2.3.3 for details of methods..

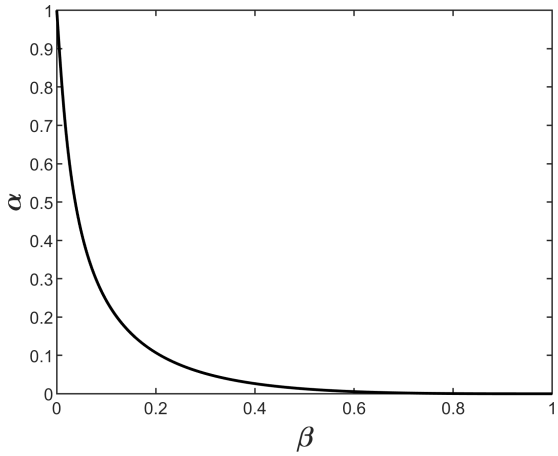


Figure 4. Correspondence of different α values to β values. The plot can be used for informed selection of threshold, similar to multivariate EBI.

Too few variables lead to inaccurate probability density estimations where few isolated data points are not adequately represented by continuous distributions. In this situation, the extreme values of prior probabilities would correspond to fewer data points with real effect; hence, the probability densities fitted to these values will not be very representative and accurate.

On the other hand, too many variables lead to unwanted spread of the null distribution to the extent that inference at low FDR values does not yield significant results. This situation, however, can be partly remedied for by applying the EBI as several independent batches of analyses on the mutually exclusive chunks of data, each containing different variables. This is permissible as the quantities such as FDR and Posterior probability (and to a reasonable extent the power) are not affected by multiple testing (as is the case for p-values).

As the complete procedure for EBI relies on permutations for building the null distribution, the procedure would depend on random number generations and some variability in each run. Additionally, the numerical procedures for estimating the GMM fits to the distributions are subject to minor variability in each run. These 2 factors make the inference a non-deterministic procedure, which is subject to some variability. While important to take this into consideration, the results in Figure 3 idicate that this variability does not change the nature of the results.

Future studies are expected to focus on the factors that lead to inaccurate numerical estimations, further extending the range of operating conditions, as well as theoretical developments for robust estimation of prior when extreme data and conditions are processed.

5 Conclusion

The implementations of statistical inferences such as EBI that can inform of the posterior probabilities and statistical power need to be converted to common practice. This implementation of threshold selection for EBI and single testing has potential to add value to the neural signal analysis and neuroimaging studies by enabling realistic inference on high-dimensional multivariate data.

Acknowledgement

The author would like to thank the students and staff in the Academic Unit of Neurology, at Trinity College Dublin, the University of Dublin for facilitating and supporting this work. The study was supported by Irish Research Council (Government of Ireland Postdoctoral Research Fellowship GOIPD/2015/213 to the author) and by Science Foundation Ireland (SFI/16/ERCD/3854).

Appendix A: Extension of the Tests from Comparison to Correlation Coefficients and Beyond

The original EBI has been primarily used for two-sample one-dimensional location problems (between-group comparison) such as gene discovery by comparing a control group to a treatment or affected group, or similarly by comparison of the healthy individuals against patients as intended in neuro-electro-magnetic signal analysis. Notwithstanding, the framework can be similarly used for virtually any statistical test, such as one-sample location problems (where comparison of data against zero or paired comparison of data are intended), as well as correlation analysis. These options have been implemented in the EBI Toolbox for MATLAB.

5.1 One Sample Inference

For a one-sample 1-dimensional location problem, including n data points x_i, the Wilcoxon's Signed Rank test statistic is defined as $W = \Sigma_{i=1}^{n} i.x_{R(i)}.1_{\{x_{R(i)}>0\}}$, where $R(i)$ is the rank of the $\{|x_i||x_i \neq 0\}$. The normalised test static may be defined as $W_n = W/(\Sigma R(i))$ which is bounded between 0 and 1. Therefore, W_n can be transformed to the z space using (4), as in the case of AUC and the subsequent procedures will be similar to the two-sample problem. The bootstrapping procedure for building the null distribution, is carried out by performing random sign flips (multiplying data by random -1 or 1 values) and recalculation of the test statistics for the number of bootstrapping cycles.

5.2 Correlation Coefficient

To use the same framework for analysis of correlation coefficients, the Spearman's correlation coefficient ρ can be mapped to the [0 1] range (as for AUROC, A) by the transformation $(\rho + 1)/2$. In this case, the grouping information will have equal number of paired zeros and ones. The null permutation of the data shall consist of using separate re-sampling (with substitution) from the first and second groups of observations for the same data points as the original data, which disregards their pairing information.

References

Akaike, H. (1974, December). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705

Benjamini, Y. & Hochberg, Y. (1995, January 1). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006, January 9). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. doi:10.1093/biomet/93.3.491

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, *99*(465), 96–104.

Efron, B. (2007a). Correlation and Large-Scale Simultaneous Significance Testing. *Journal of the American Statistical Association*, *102*(477), 93–103.

Efron, B. (2007b, August). Size, power and false discovery rates. *The Annals of Statistics*, *35*(4), 1351–1377. doi:10.1214/009053606000001460

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association*, *96*(456), 1151–1160. doi:10.1198/016214501753382129

Eklund, A., Nichols, T. E., & Knutsson, H. (2016, December 7). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905. doi:10.1073/pnas.1602413113. pmid: 27357684

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, *10*(4), 507–521. doi:10.2307/2331838. JSTOR: 2331838

Iyer, P. M., Mohr, K., Broderick, M., Gavin, B., Burke, T., Bede, P., … Vajda, A. (2017). Mismatch negativity as an indicator of cognitive sub-Domain Dysfunction in amyotrophic lateral sclerosis. *Frontiers in Neurology*, *8*, 395.

Le, C. T., Pan, W., & Lin, J. (2003, July 1). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics*, *3*(3), 117–124. doi:10.1007/s10142-003-0085-7

McLachlan, G. & Peel, D. (2004, April 5). *Finite Mixture Models*. John Wiley & Sons.

Nasseroleslami, B., Dukic, S., Broderick, M., Mohr, K., Schuster, C., Gavin, B., … Hardiman, O. (2017, November). Characteristic Increases in EEG Connectivity Correlate With Changes of Structural MRI in Amyotrophic Lateral Sclerosis. *Cerebral Cortex*, 1–15. doi:10.1093/cercor/bhx301

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*(7487), 150–152.

Qin, G. & Hotilovac, L. (2007, August 14). Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research*, *17*(2), 207–221. doi:10.1177/0962280207087173

Singh, A. K., Asoh, H., & Phillips, S. (2011, September 1). Optimal detection of functional connectivity from high-dimensional EEG synchrony data. *NeuroImage*, *58*(1), 148–156. doi:10.1016/j.neuroimage.2011.05.082

Singh, A. K., Asoh, H., Takeda, Y., & Phillips, S. (2015, March 30). Statistical Detection of EEG Synchrony Using Empirical Bayesian Inference. *PLoS ONE*, *10*(3), e0121795. doi:10.1371/journal.pone.0121795

Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2009, September 25). *Statistical Methods in Diagnostic Medicine*. Hoboken, NJ, USA: John Wiley & Sons.