Selective peak inference: Unbiased estimation of raw and standardized effect size at local maxima

Samuel J. Davenport*and Thomas E. Nichols†‡§

December 18, 2018

Abstract

The spatial signals in neuroimaging mass univariate analyses can be characterized in a number of ways, but one widely used approach to describe areas of activation is peak inference, the identification of the 3D coordinates of local maxima and the activation magnitude at these locations. These locations and magnitudes provide a useful summary of activation and are routinely reported, however, selection bias is incurred in their magnitudes as these points have both survived a threshold and are local maxima. In this paper we propose the use of bootstrap methods to estimate and correct this bias in order to estimate both the raw units change as well as standardized effect size measured with Cohen's d and partial R^2. We evaluate our method with a massive open dataset, and discuss how the corrected estimates can be used to perform power analyses.

Keywords: fMRI, selective inference, winner's curse, regression to the mean, bias, bootstrap, local maxima, UK biobank, power analyses, massive linear modelling.

1 Introduction

Any time a set of noisy data is scanned for the largest value, this value will be an overestimate of the true, noise-free maximum. This effect is known as regression to the mean or the winner's curse and occurs because, at random, some of the variables get lucky and take on high values. In neuroimaging data, at each voxel we observe a test statistic: giving us a multi-dimensional map the peaks of which can be used to identify areas of the brain where there is activation. We are interested in the underlying true signal at these locations as the observed magnitudes incur a selection bias. This bias is caused by two factors, firstly the observed peaks have been chosen to lie above a threshold and secondly the value at each peak is the largest value in a local region around the peak. In order to determine the true effect sizes we have to account for this bias.

This issue is already well-known in fMRI and is known as a circular inference or double dipping; see Kriegeskorte et al. (2010b). The problem is widespread and a number of articles in the fMRI literature have failed to account for it, as Vul et al. (2011) pointed

*Department of Statistics, University of Oxford, Oxford, OX1 3LB, UK

†Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, UK

‡Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, UK

§Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

out to much controversy. In their meta-analysis of 55 articles, where the test-statistic at each voxel was the correlation between BOLD signal and a personality measure, they found that correlations observed were spuriously high in papers that reported values at peaks, reflecting a bias due to the winner's curse.

The current available solution to this problem in fMRI is data-splitting, where the first half of the data is used to find significant regions and the other half is used to calculate effect sizes; Kriegeskorte et al. (2010a), Kriegeskorte et al. (2010b). This results in unbiased estimates, however as the estimates are calculated using only half of the data they have larger variance. This is especially problematic when the sample sizes are small. For the same reason, the locations of local maxima will be less accurate than if they had been calculated using the whole dataset. Ideally we would like to be able to use all of the data for both purposes, obtaining accurate estimates of the peak locations and unbiased point estimates of the signal magnitude. This type of approach, where you use all of the data, is known as post-model selection or selective inference and has recently generated a lot of interest, see Berk et al. (2013), Lee and Taylor (2014) and in particular Taylor and Tibshirani (2015) for a good overview.

A similar problem arises in genetics, Göring et al. (2001), and there has been much recent work on correcting for selection in this setting: Zhong and Prentice (2008), Zöllner and Pritchard (2007), Ghosh et al. (2008), Siegmund (2002), Xiao and Boehnke (2012). In particular resampling approaches have been applied in a number of papers: Sun and Bull (2005), Wu et al. (2006), Yu et al. (2007), Jeffries (2007). In the imaging literature, Rosenblatt and Benjamini (2014) propose a selective inference approach to obtain obtain unbiased confidence intervals but not point estimates. Under the assumption of constant variance Benjamini and Meir (2014) propose a method to correct all voxels above a threshold, however this doesn't take account of the effect of selecting peaks or the dependence between voxels. Esterman et al. (2010) use a leave one out cross validation approach to provide corrected estimates. This approach has the disadvantage that each resample has a different estimate of the significant locations. We employ a bootstrap resampling method that provides point estimates of local maxima, accounting for both the peak height and the location within the image. Additionally we use all of the data to determine significant locations meaning that these locations are consistent across resamples.

The idea of using an estimate other than the sample mean to provide an estimate for the mean is first due to Stein (1956) and James et al. (1961) who introduced the famous James-Stein esetimator. Recently there has been work to correct for the bias of the largest value observed values of a given distribution. Efron (2011) use an empirical Bayes technique to correct for this bias, an approach that has been applied in the genetics literature: Ferguson et al. (2013). In the case of independent random variables that each come from distributions belonging to a known parametric family, Simon and Simon (2013) introduced a frequentist method to correct bias and Reid et al. (2014) details a post-model selection approach.

Brain imaging data is more complicated than these other settings as it has complex spatial and temporal dependencies. However, we can take advantage of the fact that data from different subjects is independent. This allows us to employ a bootstrap approach to resample the data while preserving the spatial dependence structure. Tan et al. (2014) outline an extension of the Simon and Simon (2013) work to allow for dependence using the non-parametric bootstrap to estimate the bias, and then apply this method to calculate effect sizes in genetics data. We provide a detailed framework for this method and show how it can be applied to neuroimaging data. The novel contribution of our

work is to develop point estimates which account for selective inference bias due to thresholding and the use of local maxima. We develop these methods to obtain accurate estimates of Cohen's d and R^2, two quantities that are essential for power analyses; see Mumford (2012) for an overview and Appendix E for the mathematical details.

We use functional and structural magnetic resonance images (MRI) from 8940 subjects in the UK biobank. The size of this dataset allows us to validate our methods in a way that has never been possible before the availability of data of such scale, allowing us to set aside 4000 subjects to provide an accurate estimate of the truth and divide the remaining subjects into small groups in order to test our methods. The importance of these sorts of real data empirical validations is highlighted by recent work on the validity of cluster size inference Eklund et al. (2016).

The structure of this paper is as follows. Section 2 explains the details behind the bootstrapping method and how it can be applied to one-sample and the more general linear model scenario. In the one-sample case our method provides corrected estimates of the %BOLD mean and Cohen's d at the locations of peaks of the one-sample t-statistic found to be significant after correction for multiple comparisons. In the case of the general linear model it provides corrected estimates of partial R^2 values. Section 2.3 discusses the methods used for big data evaluation. Section 3 illustrates the methods on simulated data and Section 4 applies the techniques to one-sample analysis of functional imaging data and GLM analysis of structural gray matter data. In Section 4.4 we apply out method to a dataset from the Human Connectome Project that involves a contrast for working memory and obtain corrected Cohen's d and % BOLD values at significant peaks.

2 Methods

In order to set up some notation and definitions, let $\mathcal{D} = \{1, \ldots, d_1\} \times \cdots \times \{1, \ldots, d_K\}$ be a d_1 by ... by d_K lattice where $K \in \mathbb{N}$ is the number of dimensions and $d = (d_1, \ldots, d_K) \in \mathbb{N}^K$ is the size of the lattice in each of its K directions. Let $\mathcal{V} \subset \mathcal{D}$ be our set of voxels; for our purposes $\mathcal{V}$ will be the brain or a subset under study. Define an **image** to be a map Z on the set of voxels $\mathcal{V}$ which takes real or vector values.[1] Given an image Z and a connectivity criterion that determines the neighbours of each voxel, define the **local maxima** or **peaks** of Z to be the set of voxels in $\mathcal{V}$ such that the value that Z takes at them is larger than the value Z takes at their neighbours; see Appendix D for a rigorous definition.

2.1 One-Sample

Suppose that we have N subjects and for each $n = 1, \ldots, N$ a corresponding random image Y_n on $\mathcal{V}$ such that

$$Y_n(v) = \mu(v) + \epsilon_n(v) \tag{1}$$

for every voxel $v \in \mathcal{V}$, where for each $n = 1, \ldots, N$, $\epsilon_n \overset{iid}{\sim} F$, where F is an unknown zero-mean multivariate distribution on $\mathcal{V}$. Let $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} Y_n$ be the sample mean image and let $\hat{v}_k$ be the location of the kth largest local maximum of $\hat{\mu}$ above a screening threshold u. We are interested in inferring values of μ at the locations $\hat{v}_k$ since the circular estimate $\hat{\mu}(\hat{v}_k)$ will be a biased estimate of $\mu(\hat{v}_k)$.

[1] $Z : \mathcal{V} \longrightarrow \mathbb{R}^m$ for some $m \in \mathbb{N}$.

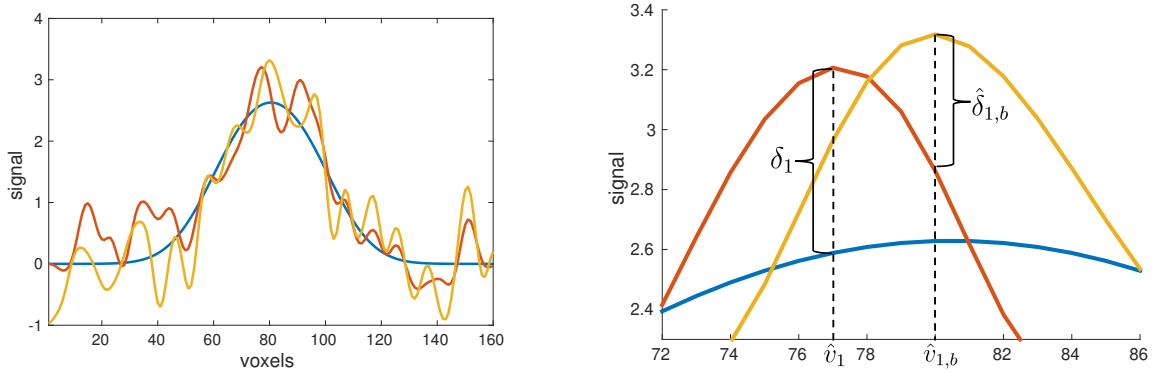
2.1.1 Peak Estimation

F is unknown so in order to estimate the bias of $\mu(\hat{v}_k)$ we bootstrap the data to generate bootstrap samples. The use of the non-parametric bootstrap here means we do not have to make any assumptions on the spatial auto-covariance of the errors. This allows us to obtain an estimate of the bias for each bootstrap iteration as in Tan et al. (2014). For each maxima $\hat{v}_k$ we estimate the intensity as $\tilde{\mu}(\hat{v}_k) = \hat{\mu}(v_k) - \delta_k$, where δ_k are bias correction terms found as described in Algorithm 1 below.

Algorithm 1 Non-Parametric Bootstrap Bias Calculation

1: **Input**: Images $Y_1, \ldots, Y_N$, the number of bootstraps to do: B and a threshold u.
2: Let $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} Y_n$ and let K be the number of peaks of $\hat{\mu}$ above the threshold u and for $k = 1, \ldots, K$, let $\hat{v}_k$ be the location of the kth largest maxima of $\hat{\mu}$.
3: **for** $b = 1, \ldots, B$ **do**
4: Sample $Y^*_{1,b}, \ldots, Y^*_{N,b}$ independently with replacement from $Y_1, \ldots, Y_N$.
5: Let $\hat{\mu}_b = \frac{1}{N}\sum_{n=1}^{m} Y^*_{N,b}$ and for $k = 1, \ldots, K$, let $\hat{v}_{k,b}$ be the location of the kth largest local maxima of $\hat{\mu}_b$.
6: For $k = 1, \ldots, K$, let $\hat{\delta}_{k,b} = \hat{\mu}_b(\hat{v}_{k,b}) - \hat{\mu}(\hat{v}_{k,b})$ be an estimate of the bias at the kth largest local maxima.
7: **end for**
8: For $k = 1, \ldots, K$, let $\hat{\delta}_k = \frac{1}{B}\sum_{b=1}^{B} \hat{\delta}_{k,b}$.
9: **return** $(\hat{\mu}(\hat{v}_1) - \hat{\delta}_1, \ldots, \hat{\mu}(\hat{v}_K) - \hat{\delta}_K)$.

Figure 1 provides a small 1D simulation on a grid of 160 voxels. Here we have just considered $k = 1$: the global maximum. The bias above the noise-free signal (δ_1) is evident, and is estimated by comparing a bootstrap sample to the original, yielding an estimate of $\hat{\delta}_{1,b}$. Here, $N = 20$ and for each $n = 1, \ldots, N$ the error images are created by simulating i.i.d Gaussians at each voxel with variance 4 and then smoothing this with 6 voxel FWHM. δ_1 is the bias of the empirical mean relative to the true mean.



(a) True Mean vs the Empirical Mean

(b) Zooming in on the peak

Figure 1: Illustration of our method on a simple annotated example. Here our set of voxels is $\mathcal{V} = \{1, \ldots, 160\}$, the true mean μ is shown in blue, the empirical mean $\hat{\mu}$ is shown in red and one sample bootstrap realization (iteration b) is shown in yellow. The data is generated as described by model (1) with $N = 20$. The noise was generated by simulating i.i.d Gaussians at each voxel with variance 4 and then smoothing this with 6 voxel FWHM.

2.1.2 Peak Estimation for Effect Size

Brain mapping has traditionally focused on test statistics and not measures of effect size, like %BOLD directly. In this setting we can focus on estimating two different quantities, the effect size (using Cohen's d), or the %BOLD. However before measuring the effect we need to test

$$H_0(v) : \mu(v) = 0 \text{ versus } H_1(v) : \mu(v) \neq 0$$

at each $v \in \mathcal{V}$ in order to determine whether there is an activation at that voxel. We will now assume that the error, ϵ comes from a multivariate Gaussian distribution. Define $\sigma^2 : \mathcal{V} \longrightarrow \mathbb{R}^+$ to be the population variance image such that $\sigma^2(v) = \text{var}(\epsilon(v))$ for each $v \in \mathcal{V}$. The unbiased estimate of the variance at each v is

$$\hat{\sigma}^2(v) = \frac{1}{N-1} \sum_{n=1}^{N} (Y_n(v) - \hat{\mu}(v))^2.$$

Under $H_0(v)$,

$$t(v) = \frac{\hat{\mu}(v)\sqrt{N}}{\hat{\sigma}(v)} \sim t_{N-1},$$

a t-distribution with $N-1$ degrees of freedom.

As before, we require a screening threshold u. While a threshold u on a mean image is ultimately arbitrary, on a statistic image we can choose a value of u to control false positives at a desired level while controlling for multiple testing. For example, we can use results from the theory of random fields to find a u such the familywise error rate, the chance of one or more false positives over the image, is controlled; Worsley et al. (1996), Friston et al. (1994).

However a statistic value T is not interpretable across studies – it depends on sample size and, in particular, grows to infinity with N. Instead, the goal is estimating a standardised effect size such as Cohen's d (typically a scalar multiple of a T image):

$$\hat{d}(v) = \frac{\hat{\mu}(v)}{\hat{\sigma}(v)}$$

which can be used in power analyses see Appendix E. Fortunately this is just a scalar multiple of the t statistic.

See Section 3.1 for the application of Algorithm 2 to simulated data and Section 4.1 for validation of its use on task fMRI data for the estimation of Cohen's d at local maxima.

Algorithm 2 Non-Parametric Bootstrap Bias Calculation

1: **Input**: Images $Y_1, \ldots, Y_N$, the number of bootstraps to do: B and a threshold u.
2: Let $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} Y_n$ and define an image $\hat{\sigma}$ such that $\hat{\sigma}^2(v) = \frac{1}{n-1}\sum_{n=1}^{N}(Y_n(v)-\hat{\mu}(v))^2$ for each $v \in \mathcal{V}$.
3: Let K be the number of peaks of $\hat{\mu}\sqrt{N}/\hat{\sigma}$ above the threshold u and for $k = 1, \ldots, K$, let $\hat{v}_k$ be the location of the kth largest maxima of $\hat{\mu}/\hat{\sigma}$.
4: **for** $b = 1, \ldots, B$ **do**
5: Sample $Y^*_{1,b}, \ldots, Y^*_{N,b}$ independently with replacement from $Y_1, \ldots, Y_N$.
6: Let $\hat{\mu}_b = \frac{1}{N}\sum_{n=1}^{N} Y^*_{n,b}$ and let $\hat{\sigma}^2_b(v) = \frac{1}{N-1}\sum_{n=1}^{N}(Y^*_{n,b}(v) - \hat{\mu}_b(v))^2$ for each $v \in \mathcal{V}$.
7: For $k = 1, \ldots, K$, let $\hat{v}_{k,b}$ be the location of the kth largest local maxima of $\hat{\mu}_b/\hat{\sigma}_b$.
8: Let $\hat{\delta}_{k,b} = \hat{\mu}_b(\hat{v}_{k,b})/\hat{\sigma}_b(\hat{v}_{k,b}) - \hat{\mu}(\hat{v}_{k,b})/\hat{\sigma}(\hat{v}_{k,b})$ be an estimate of the bias.
9: **end for**
10: For $k = 1, \ldots, K$, let $\hat{\delta}_k = \frac{1}{B}\sum_{b=1}^{B}\hat{\delta}_{k,b}$.
11: **return** $(d(\hat{v}_1) - \hat{\delta}_1, \ldots, d(\hat{v}_K) - \hat{\delta}_K)$.

2.1.3 Estimation of the Mean at the Location of Effect Size Peaks

Some authors, Chen et al. (2017) most recently, have argued that the attention given to statistic images is misguided, and more focus should be given to results in the interpretable units, i.e. %BOLD for fMRI. In order to estimate the mean while still controlling for false positives one needs to use the statistic image to identify locations of interest and then measure the mean at those locations. In our framework this is easily accomplished with a small modification to Algorithm 2, computing in Step 8 instead a bias of the form:

$$\hat{\delta}_{k,b} = \hat{\mu}_b(\hat{v}_{k,b}) - \hat{\mu}(\hat{v}_{k,b})$$

and returning $(\hat{\mu}(\hat{v}_1) - \hat{\delta}_1, \ldots, \hat{\mu}(\hat{v}_K) - \hat{\delta}_K)$ instead. See Section 4.2 for validation of this approach on the estimation of % BOLD mean at local maxima of the t-statistics of task fMRI data.

2.1.4 Existing One-Sample Methods

We will be comparing the bootstrap approach to circular inference and data-splitting which are the main two approaches used in the literature. After performing thresholding to find the number of peaks above a threshold as in Algorithm 2, the circular inference uncorrected estimates are simply $d(\hat{v}_1), \ldots, d(\hat{v}_K)$.

In contrast data-splitting is as follows. First we divide the images into two groups: $Y_1, \ldots, Y_{N/2}$ and $Y_{N/2+1}, \ldots, Y_N$. Let d_1 and d_2 be the image estimates of Cohen's d from the first and second half of the subjects respectively. Using a threshold u (adjusted for the fact that we are now working with $N/2$ rather than N subjects) we find the peaks of $d_1\sqrt{N/2}$ that lie above u, at locations $\hat{w}_1, \ldots, \hat{w}_J$ for some number of peaks J. The data-splitting estimates of the peak values are $d_2(\hat{w}_1), \ldots, d_2(\hat{w}_J)$. See Section 2.3.2, Figure 3 for an illustration of the different methods applied to a sample consisting of 50 subjects. Note that in general the number of significant peaks found by data-splitting will be lower than the number found using all of the data as with half the number of subjects there is less power to detect activation.

2.2 General Linear Model

Having introduced the method in the simplified setting of a one-sample model, we now turn to the regression setting. Here, we will often have no practical meaningful units; for example, for a covariate of age, the units of the coefficient are clear (expected change in response per year) but awkward, and more typically users will want to reference the partial coefficient of determination, partial R^2, the proportion of variance explained by one (or more) predictors. Hence we now explain how our method extends to peak estimation of peak partial R^2.

Let Y be an N-dimensional random image such that for each $v \in \mathcal{V}$, we assume the following linear model structure,

$$Y(v) = X\beta(v) + \epsilon(v),$$

for an $N \times p$ design matrix X and a parameter vector $\beta \in \mathbb{R}^p$ where ϵ is the random N-dimensional image of the noise such that $\epsilon(v) = (\epsilon_1(v), \ldots, \epsilon_N(v))^T$ for each $v \in \mathcal{V}$. Then we are interested in testing

$$H_0(v) : C\beta(v) = 0 \text{ versus } H_1(v) : C\beta(v) \neq 0$$

for some contrast matrix $C \in \mathbb{R}^{m \times p}$ for some positive integer m. We can test this at each voxel with the usual F-test, the value of which we denote by $F(v)$ for each $v \in \mathcal{V}$. This image is defined as

$$F = \frac{(C\hat{\beta})^T (C(X^T X)^{-1} C^T)^{-1} (C\hat{\beta})/m}{\hat{\sigma}^2}$$

where $\hat{\beta}$ is the least squares estimate of β and $\hat{\sigma}^2$ is the population estimate of the variance at each voxel. Then under the null hypothesis $H_0(v)$, $F(v)$ has an $F_{m,N-p}$ distribution and can therefore be used for testing $H_0(v)$. We will incorporate this into our bootstrap algorithm in order to establish which peaks are significant.

We are interested in the R^2 values, so we define R^2 be the image with the estimated partial R^2 values for comparing the null model against the alternative at each voxel: we then seek a bias-corrected $\tilde{R}^2(\hat{v}_k) = R^2(\hat{v}_k) - \delta_k$. (See Appendix C for details on how these quantities are defined.) Bootstrapping in the general linear model scenario is based on the residuals; see Davison et al. (2003) Chapter 6. This leads to the algorithm below.

Algorithm 3 Non-Parametric Bootstrap Bias Calculation

Input: Images $Y_1, \ldots, Y_N$, the number of bootstraps to do: B and a threshold u.
Let K be the number of peaks of F above the threshold u and for $k = 1, \ldots, K$, let $\hat{v}_k$ be the location of the kth largest maxima of F.
Let $\hat{\beta} = \hat{\beta}(X, Y)$ and let $\hat{\epsilon} = Y - X\hat{\beta}$ be the residuals on the original data.
For each $n = 1, \ldots, N$, let $r_n = \hat{\epsilon}_n / \sqrt{1 - p_n}$ be the standardized residuals (where $p_n = (X(X^TX)^{-1}X^T)_{nn}$). Let $\overline{r} = \frac{1}{N}\sum_{n=1}^{N} r_i$ be their mean.
for $b = 1, \ldots, B$ **do**
    Sample $\epsilon^*_{1,b}, \ldots, \epsilon^*_{N,b}$ independently with replacement from $r_1 - \overline{r}, \ldots, r_N - \overline{r}$ and let $\epsilon^*_b = (\epsilon^*_{1,b}, \ldots, \epsilon^*_{N,b})^T$ and set $Y^*_b = X\hat{\beta} + \epsilon^*$.
    For $k = 1, \ldots, K$, let $\hat{v}_{k,b}$ be the location of the kth largest local maxima of F^*_b: the bootstraped F-statistic image computed using X, Y^*_b and $\hat{\beta}(X, Y^*_b)$. Let R^2_b be the bootstrapped partial R^2 image and set $\hat{\delta}_b = R^2_b(\hat{v}_{k,b}) - R^2(\hat{v}_{k,b})$ to be the estimate of the bias.
end for
For $k = 1, \ldots, K$, let $\hat{\delta}_k = \frac{1}{B}\sum_{b=1}^{B} \hat{\delta}_{k,b}$.
return $(R^2(\hat{v}_1) - \hat{\delta}_1, \ldots, R^2(\hat{v}_K) - \hat{\delta}_K)$.

In fMRI we are often interested in the case where $C^T = c \in \mathbb{R}^p$ is a contrast vector in which case we can also test using the $t-$statistic

$$t = \frac{c^T\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T(X^TX)^{-1}c}} \sim t_{N-p}.$$

which allows us to perform either one or two sided tests in order to determine significance before bootstrapping.

As in the Section 2.1.4 we can define circular inference and data-splitting estimates. See Section 4.3 for validation of the use of the bootstrap and comparisons between the methods in a GLM scenario where VBM images are regressed against the age of the participants and an intercept.

2.3 Methods - Big Data Validation

In order to test our methods we have been able to take advantage of the huge amount of data from the UK Biobank. This has enabled us to set aside 4000 (randomly selected) subjects in order to compute a very accurate estimate of the mean, Cohen's d or partial R^2 value. In order to avoid losing lots of the brain due to drop out we estimate the true value at each voxel using the available data at that voxel. We will refer to this 4000-subject estimate of the effect as the truth. Implementing linear models on such large datasets requires mathematical tricks in order to avoid excessive computational costs especially when the data is missing. In Appendix A we outline efficient methods for dealing with this and describe how the truth is computed in the different settings. We have divided the remaining 4980 subjects into groups of sizes similar to those used in typical fMRI/VBM studies. For each such group we applied all three methods and compared the values obtained to the truth calculated using the 4000 subjects allowing the performance of the methods across groups to be evaluated.

2.3.1 Image Aquisition

We use data drawn from the UK Biobank, a prospective epidemiological resource combining questionnaires, physical and cognitive measures, and biological samples in a sample of 500,000 subjects in the United Kingdom, aged 40–69 years of age at baseline recruitment. The UK Biobank Imaging Extension provides extensive MRI data of the brain, ultimately on 100,000 subjects. We used the prepared data available from the UK Biobank; full details on imaging acquisition and processing can be found in Miller et al. (2016), Alfaro-Almagro et al. (2018) and from UK Biobank Showcase[2]; a brief description is provided here. The task fMRI data uses the block-design Hariri faces/shapes task Hariri et al. (2002), where the participants are shown triplets of fearful expressions and, in the control condition, triplets of shapes, and for each event perform a matching task. A total of 332 T2*-weighted blood-oxygen level-dependent (BOLD) echo planar images were acquired in each run [TR=0.735s, TE=39ms, FA=52°, 2.4mm^3 isotropic voxels in 88x88x64 matrix, x8 multislice acceleration]. Standard preprocessing and task fMRI modelling was conducted in FEAT (FMRI Expert Analysis Tool; part of the FSL software `http://www.fmrib.ox.ac.uk/fsl`). After head-motion correction and Gaussian kernel of FWHM 5mm, a linear model was fit at each voxel resulting in contrast images for each subject.

Structural T1-weighted images were acquired on each subject [3D MPRAGE, 1mm^3 isotopic voxels in 208x256x256 matrix]. Images were defaced and nonlinearly warped to MNI152 space using FNIRT (FMRIB's Nonlinear Image Registration Tool. Tissue segmentation was performed FSL's FAST (FMRIB's Automated Segmentation Tool), producing images of gray matter that were subsequently warped to MNI152 space, and modulated by the Jacobian of the warp field. In order to save space, images were written with voxel sizes of 2mm.

Additional processing consisted of transforming intrasubject contrast maps to MNI space with 2mm using nonlinear warping determined by the T1 image and an affine registration of the T2* to the T1 image. We also applied a smoothing of 3mm FWHM to the modulated gray matter images.

2.3.2 task fMRI data

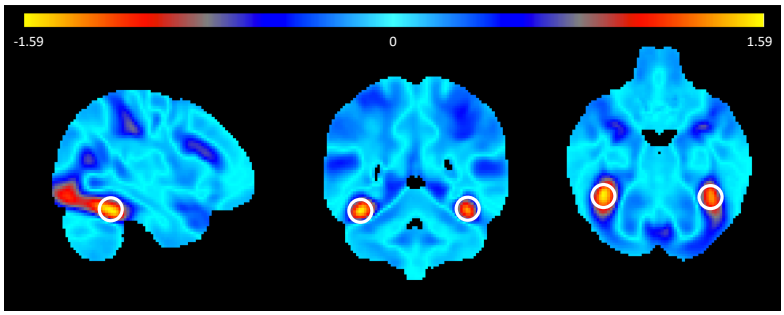
We have faces-shapes contrast images from 8940 subjects and will consider the mean and one sample Cohen's d. The truth is not subject to the circular inference problem due to the sheer number of subjects. Slices through the one-sample Cohen's d truth are shown in Figure 2 below. Each subject has a different subject specific mask. The intersection over all masks give a region whose volume is only around 50% of the entire brain so in order to compute the truth at each voxel we computed the value using the available data at that voxel. See Appendix A.2 for more details.

After calculating the truth there are 4940 subjects left over. For a given sample size N, let $G_N = \lfloor 4940/N \rfloor$ groups[3] be the number of groups of size N that we can divide this remaining data into. This division will enable us to compare the performance of the three available methods, namely circular inference, data-splitting and the bootstrap. See Sections 4.1 and 4.2 for details. Figure 3 illustrates these methods applied to an example sample consisting of 50 subjects.

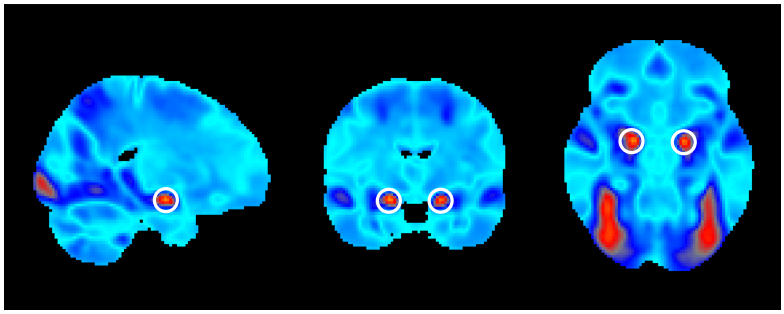
To illustrate the magnitude of the circularity problem we compare maximum peak

[2] `https://biobank.ctsu.ox.ac.uk/crystal/docs/brain_mri.pdf`

[3] For $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer that is less than or equal to x.



(a) Top 2 peaks



(b) 3rd and 4th Highest Peaks

Figure 2: Slices through a radiological view of the maxima of the one-sample Cohen's d truth. The top two local maxima are located at voxels [24, 40, 25] and [64, 39, 26] which correspond to the left and right cerebral cortex and have Cohen's d values of 1.5756 and 1.4326 respectively. The 3rd and 4th local maxima are located at voxels [35, 60, 29] and [54, 60, 29] which are within the left and right amygdalae and have Cohen's d values of 1.3450 and 1.3041 respectively. The locations of these peaks are indicated using white circles.

heights as a function of sample size. We computed the max peak height (of Cohen's d) for different N ranging from 10 to 100, (averaged over the G_N groups), and compare to the true max peak height of Cohen's d, see Figure 4. The bias is substantial for small N but is non-negligible even for moderate N. As N increases the bias decreases to zero as expected and the average peak maximum converges to the true maximum value.

2.3.3 VBM data

We have structural gray matter (VBM) data from the 8940 subjects. As discussed in the methods section, bootstrap methods can be extended to the general linear model scenario. To illustrate this, we regress gray matter images against age, sex and an intercept. In particular let A_n be the age of the nth subject and let S_n be their sex, then consider the model:

$$Y_n(v) = \mu + \beta A_n + \gamma S_n + \epsilon_n(v) \tag{2}$$

for each $v \in \mathcal{V}$, where the ϵ_n are i.i.d random images on $\mathcal{V}$ for $n = 1, \ldots, N$. Using the 4000 subjects, we can calculate an accurate estimate of the partial R^2 for age. The largest maximum is located at voxel [45, 62, 34] and has a partial R^2 of 0.2466. As above we divide the remaining subjects into small subgroups to compare the methods by calculating the partial R^2 for age on each subgroup. See Section 4.3 for the results of this validation.

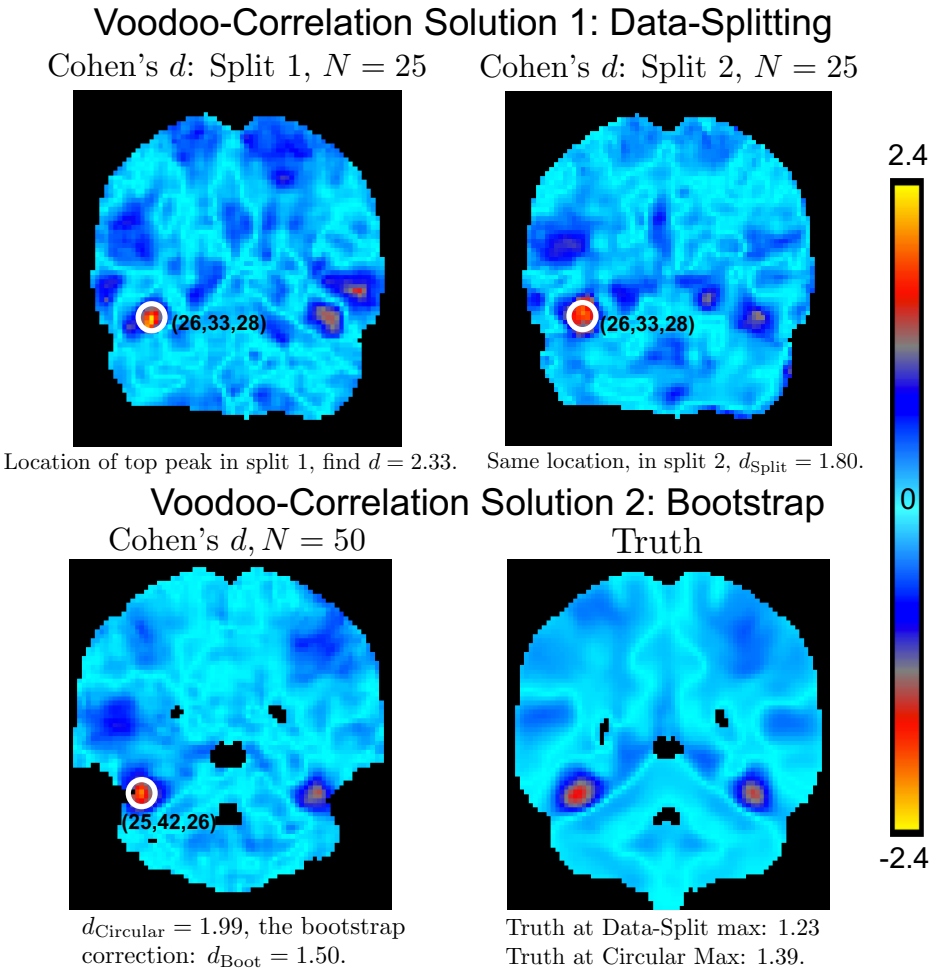


Figure 3: Comparing the different methods. We have taken a sample of size 50 and implemented all the methods in order to yield an estimate of the truth at the location of the maximum. Data-splitting requires us to split the data in half, using the first half of the data to determine the locations. Circular inference and the bootstrap both use all of the data to calculate the locations.

2.4 Computing the Thresholds

Researchers in the field typically either use random field theory (RFT) (Worsley et al. (1996)) or permutation testing (Nichols and Holmes (2001)). Voxelwise RFT controls the false positive rates but is slightly conservative, see Eklund et al. (2016), primarily because the lattice assumption is not valid for small sample sizes. On the other hand voxelwise permutation can also have inflated false positive rates, see Eklund et al. (2018), and has a high computational cost. In our case this cost is prohibitive as we need to perform a big data validation which requires many analyses as described in Section 2.3. We have thus elected to use voxelwise random field theory for our big data analyses. In practise when running a typical fmri/vbm analysis our methods work independently of the method used to choose the threshold.

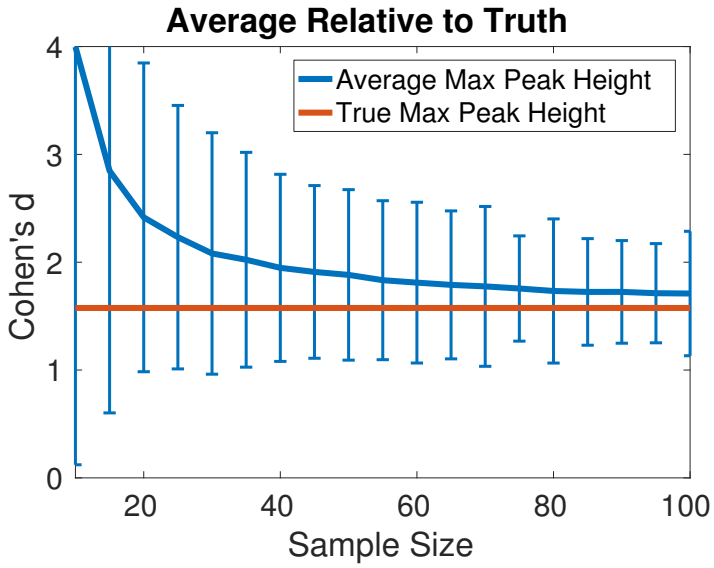


Figure 4: The average selection bias in the one-sample Cohen's d: an illustration of the winner's curse. We have plotted the average peak height of the maximum against N. For each N we computed Cohen's d for each of the G_N groups of size N found the value of the maximum and took the average over the G_N groups. The 95% error bars are based on the 2.5% and 97.5% quantiles for each sample size. The bias is substantial for small N but is non-negligible even for moderate N.

3 Results - Simulated Examples

In this section we illustrate the performance of Algorithm 2. In order to test our methods we have generated 3D simulations on a 91 by 109 by 91 size grid which makes up our set of voxels $\mathcal{V}$. This grid size is that which results from using MNI space and 2mm voxels.[4]

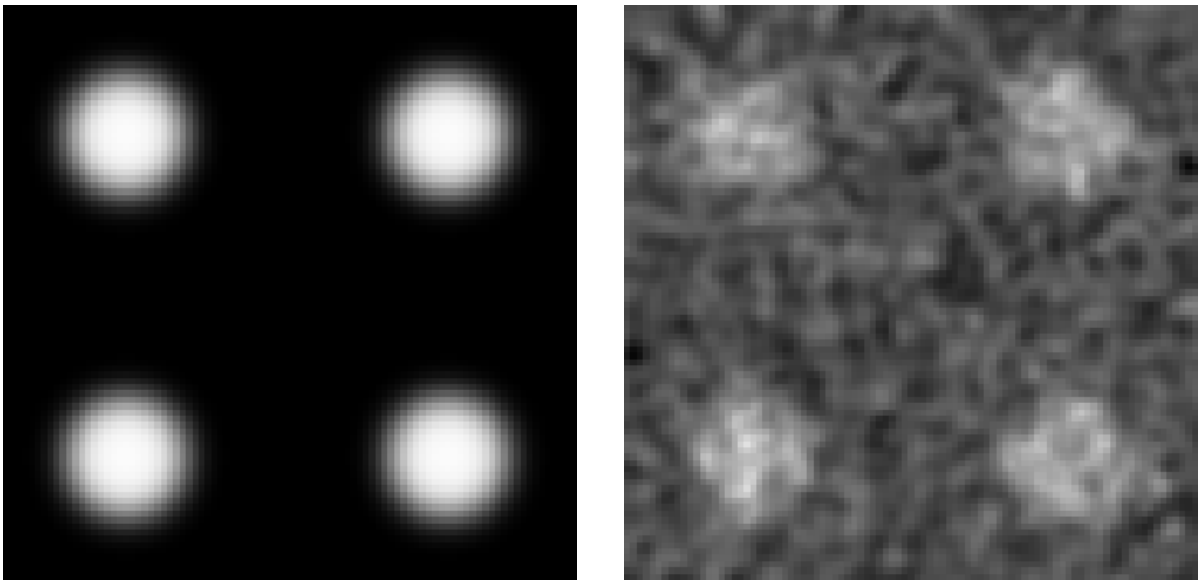
3.1 One Sample Cohen's d

In order to validate Algorithm 2 we generated data according to model (1) with underlying mean consisting of 9 different peaks each with magnitude 1/2, with one located near corner and one at the centre of the image. See Figure 5 for a slice through this signal and an example realization. For the ϵ_n we used centred gaussian noise smoothed with given FWHM, scaled to have variance 1.

We applied Algorithm 2 in order to provide estimates of Cohen's d. To do so we took FWHMs ranging from 0 to 12mm and for each FWHM generated 1000 realizations. For each realization we generated data consisting of 20 (and then 50) random images centred at the underlying mean and with the described error variance. In order to determine the correct thresholds, for each FWHM we generated 5000 null gaussian random fields with the same covariance structure and took the 95% quantile of the distribution of the maximum. This yields a voxelwise threshold that when applied rejects the null hypothesis 0.05 of the time on null data.

The one-sample Cohen's d is a biased estimator for the population Cohen's d and as such requires a correction factor in order to obtain unbiased estimates for both data-

[4]See the supplementary material for simulations that test Algorithm 1 and show that it works well and decreases the MSE. Algorithm 1 is less relevant in the neuroimaging context as in neuroimaging we first have to threshold our data in order to determine the location of significant effect sizes.



(a) True Signal (b) Sample Cohen's d

Figure 5: The true signal and a sample simulation. Panel (a) illustrates a slice through the true signal corresponding to the plane $x = 70$, $z = 20$, panel (b) illustrates a slice through the one sample Cohen's d for 50 subjects where each the data for each subject is computed by adding Gaussian noise with an FWHM of 6mm (scaled to have variance 1) to the signal.

splitting and the bootstrap. This is explained in more detail in Appendix E. Estimates of the bias, variance and MSE resulting from applying each of the three methods have been plotted in Figure 6. In our simulations the circular and bootstrap methods found considerably more peaks than data-splitting. This is to be expected as they use double the data (relative to data-splitting) to locate the peaks and are thus more powerful. Indeed in our simulations for $N = 20$, data-splitting often found no peaks to be significant at all. The bootstrap estimates consistently have the lowest MSE across a range of FWHM and sample sizes.

3.2 Estimating the Mean

As discussed in Section 2.1.3, Algorithm 2 can be applied to estimate the mean and we do this in the same simulation setting as for the Cohen's d estimates. Estimates of the bias, variance and MSE of each of the three methods have been plotted in Figure 7 where as with Cohen's d the bootstrap estimates perform very well.

4 Results - Brain Imaging Data

In order to validate our approach and compare it to existing methods using real data we have considered sample sizes of $N = 20, 50$ and 100, typical of the sample sizes used in fMRI studies. As described in the methods section for each N we divided 4940 subjects into G_N groups of size N which corresponds to 247 groups of size 20, 98 groups of size 50 and 47 groups of size 100. For each N and each group $g = 1, \ldots, G_N$ we applied the circular, bootstrapping and data-splitting methods to produce estimates. Note that because the empirical mean is only an estimate of the true mean the bootstrap estimates

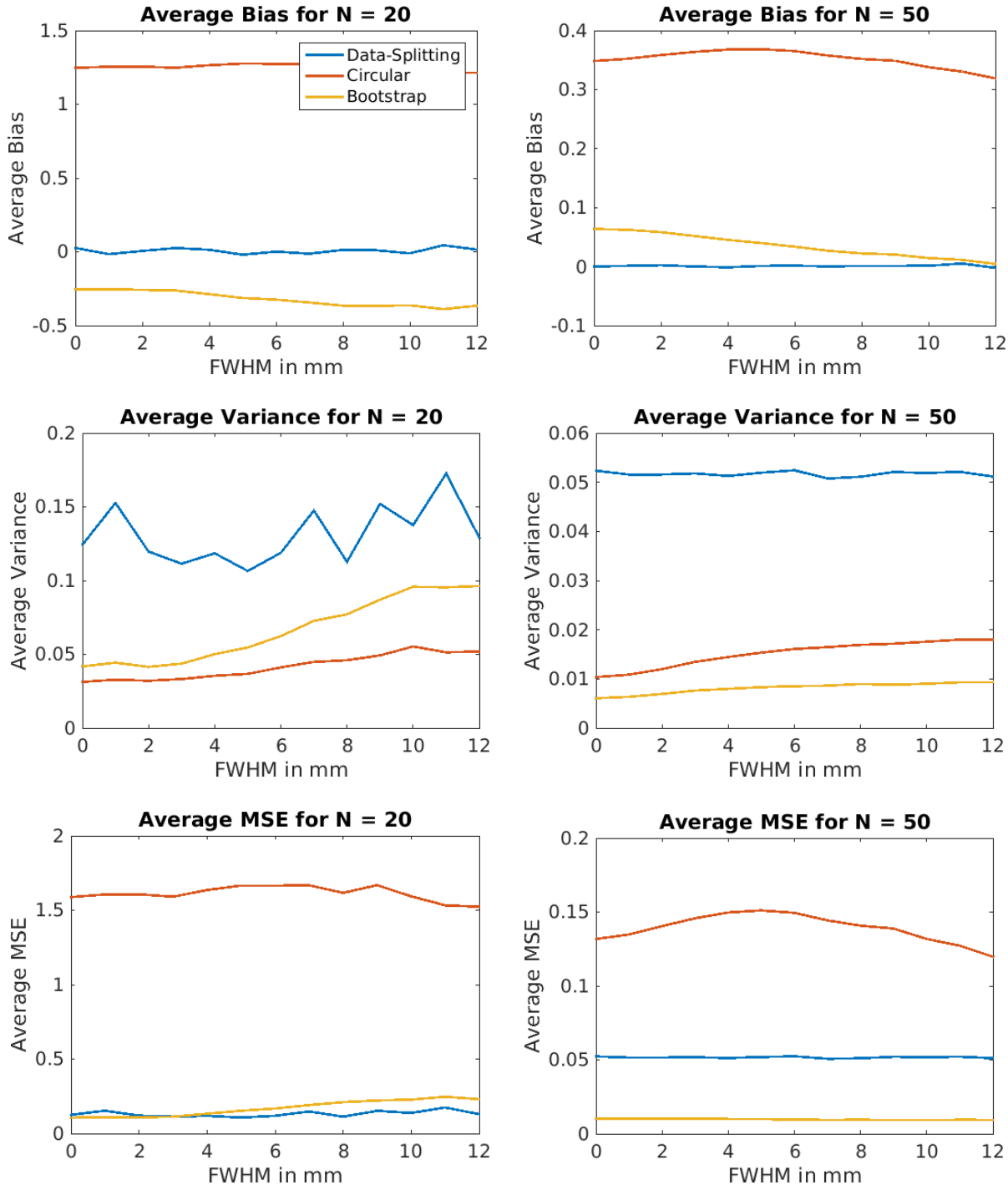


Figure 6: Implementing Algorithm 2 on simulated data to estimate Cohen's d. We have generated data from model (1) on a $91 \times 109 \times 91$ size lattice with signal which has 9 peaks and smooth gaussian noise. For the noise we took FWHMs of: 0, 0.5, 1, ..., 5.5, 6 per voxel, which for voxels of size 2mm correspond to FWHMs of 0, 1, ..., 12. For each FWHM we generated 1000 realizations (for each realization generating either $N = 20$ or $N = 50$ images), and calculated the estimates of the local maxima of Cohen's d for each realization using the three methods. We then calculated the MSE, bias and variance (see Appendix B for precise definitions of these quantities). The bootstrap estimates have significantly lower bias than the circular estimates and have a lower variance than the data-splitting estimates. This leads to a decrease in the MSE of the estimates. This comparison illustrates the improvement of the bootstrap as the sample size increases.

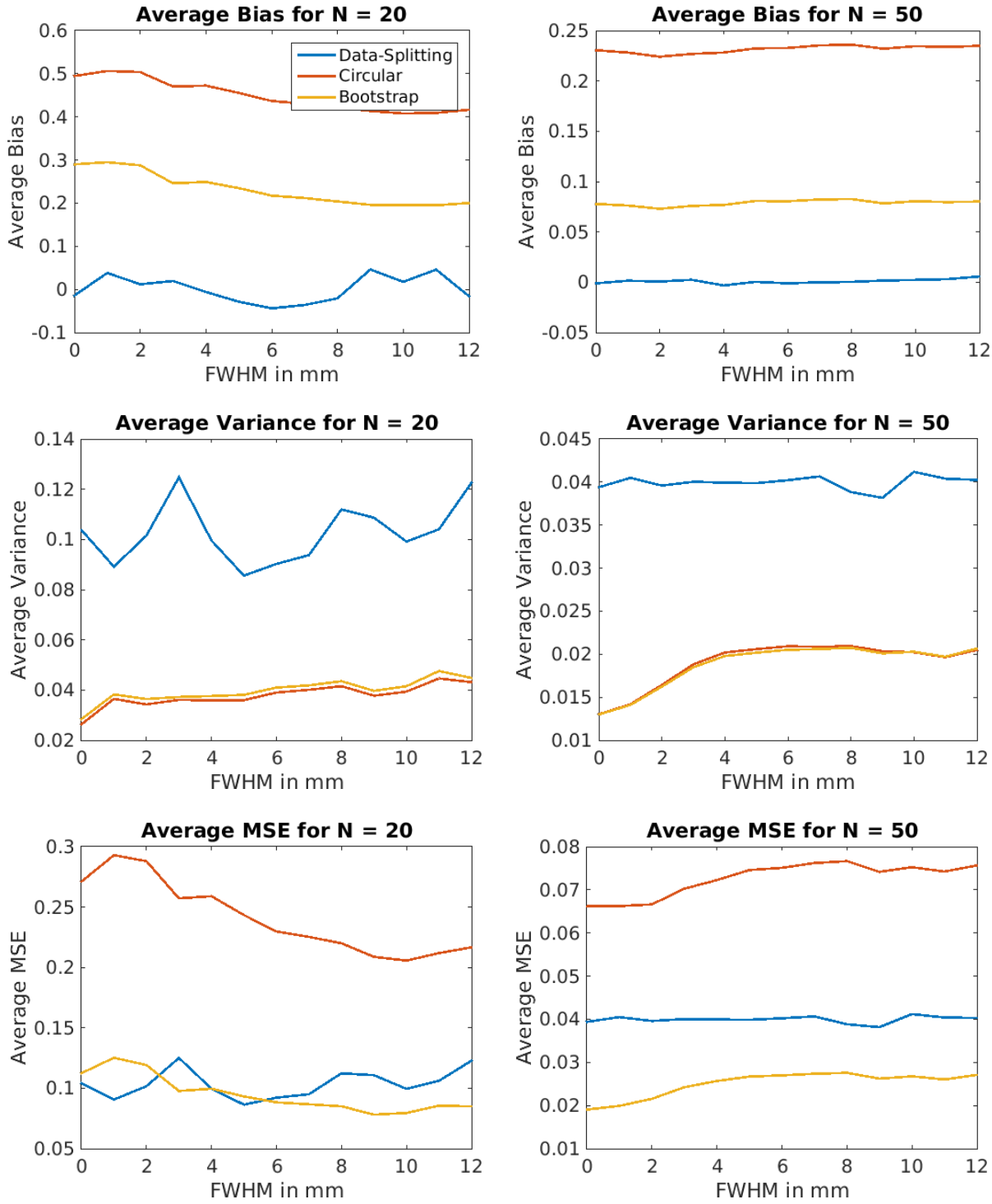


Figure 7: Implementing Algorithm 2 on simulated data to estimate the mean. We have generated data from model (1) on a $91 \times 109 \times 91$ size lattice with signal which has 9 peaks and smooth gaussian noise. For the noise we took FWHMs of: $0, 0.5, 1, \ldots, 5.5, 6$ per voxel, which for voxels of size 2mm correspond to FWHMs of $0, 1, \ldots, 12$. For each FWHM we generated 100 realizations (for each realization generating either $N = 20$ or $N = 50$ images), and calculated the estimates of the mean at the local maxima of Cohen's d for each realization using the three methods. We then calculated the MSE, bias and variance (see Appendix B for precise definitions of these quantities). For $N = 50$, the bootstrap estimates have significantly lower bias than the circular estimates and have a lower variance than the data-splitting estimates. This leads to a decrease in the MSE of the estimates. This comparison illustrates the improvement of the bootstrap as the sample size increases. However even in the $N = 20$ scenario the MSEs are comparable.

are still slightly biased. There is a trade-off to be made between bias and variance. Relative to data-splitting the bootstrap methods have lower variance and this leads to a lower MSE. They improve with increasing sample size relative to independent splitting due to the convergence of the empirical mean to the true mean. In neuroimaging it is very important to have an accurate estimate of the location of the effect. Our bootstrap estimates use all of the data to estimate the locations and provide good estimates of the effect sizes.

For each small group of N subjects $n = 1, \ldots, N$, we take Y_n to be the contrast image from the n-th subject in that group where the contrast is between faces and shapes. Here, $\mathcal{V}$ is the subset of the 3-dimensional 91 by 109 by 91 lattice of voxels corresponding to the brain; we take it to be the intersection of the subject masks within each group in order to perform our analysis.

Due to the need to correct for false positives and the variance of the % BOLD and gray matter signals, we do not recommend using Algorithm 1 on imaging data and recommend that the data is first thresholded and then used to estimated the effect size. Instead we first threshold using the t or F-statistic as in Algorithms 2 and 3. Algorithm 1 can be applied and gives accurate estimates of the underlying signal see Figures 2 and 3 of the supplementary material for its application to the fMRI data.

We look at two different data-sets and applications. In the first we look at task fMRI data from the UK Biobank with the faces-shapes contrast as described in Section 2.3.1 and 2.3.2. We implement our methods to obtain corrected estimates of Cohen's d and the % BOLD change at significant maxima and we compare them to the circular and data-splitting estimates. In the second data-set we consider the gray matter images as described in Section 2.3.3 and implement our methods to obtain corrected estimates at significant local maxima of the partial R^2 for age that arises from fitting a linear model with 2 covariates: age and an intercept. We compare these estimates to the ones that result from circular inference and data-splitting.

4.1 Estimating Cohen's d for the task fMRI images

In this section we apply Algorithm 2 to estimate Cohen's d at the locations of significant maxima, using a significance threshold determined by voxelwise RFT. In order to validate our approach for each $N = 20, 50, 100$ we applied all of the methods to each of the G_N groups. We then compared the resulting estimates to the true Cohen's d map calculated using 4000 subjects.

The bootstrap and circular estimates use all of the data to find the peaks and so are much more powerful. For $N = 20$, data-splitting is only able to use 10 subjects to calculate the locations and finds 7 significant peaks over all 247 groups meaning that it only finds $7/247 \approx 0.0283$ peaks per group of subjects. Smaller sample sizes have less power to identify significant peaks. Additionally voxelwise RFT inference is slightly conservative (see Eklund et al. (2016)) because of the break down of the good lattice assumption. We would thus discourage the use of data-splitting for such small sample sizes as identification of some significant locations is more important. The bootstrap and circular methods on the other hand use all 20 subjects to locate the peaks and find 1565 significant peaks which corresponds to around 6.3 peaks per group.

In order to illustrate how the methods compare for $N = 20, 50, 100$ we have plotted box plots of the bias of the estimators over all the significant peaks over all of the groups, see Figure 8. From these we see that the circular estimates are highly biased whereas the bootstrap estimates are asymptotically unbiased with the bias decreasing as the sample

size increases. The high bias of the circular estimates is particularly bad for the small $N = 20$ case. As expected the data-splitting estimates are unbiased. In Figure 8 we have plotted barplots of the MSE and variance. From these we see that due to the large bias the circular estimates have very high MSE. The bootstrap has the lowest MSE for $N = 50$ and 100 and data-splitting has the largest variance in all cases. For $N = 20$, we only have 7 data points for data-splitting corresponding to the 7 peaks that were above the threshold over all the 247 groups and so it is difficult to compare to the other methods. See Appendix B for precise definitions of the MSE, variance and bias in this context.

To further understand how the estimates compare we have plotted their values against the truth in Figure 9. For each $N = 20, 50, 100$ and each of the methods we have plotted a graph of the estimates that arise under that method against the truth at their locations. Each data point in the graph represents the estimate under that method of a significant peak in one of the G_N groups. (As data-splitting uses half of the subjects to find significant peaks it finds fewer peaks.) For reference we have also plotted the identity line on the plots. If our estimates were perfect then we would expect them all to lie on this line. Given that the sample size is so small, the $N = 20$ case is the most challenging for estimation. The circular estimates are very biased while bootstrap estimates are numerous and give reasonable estimates and the data-splitting estimates are particularly variable. As N increases, the estimates all perform better. The circular estimates are biased, and the data-splitting estimates are variable whereas the bootstrap estimates have low bias and variance.

Note that the shape of the plot of the bootstrap estimates is very dependent on the shape of the plot of the circular estimates. This is because the bootstrap struggles to distinguish between points that have the same observed values but different true values. This is a predictable difficulty of this approach. However the decrease in the variance of the bootstrap approach relative to data-splitting causes data-splitting to have a higher MSE than the bootstrap approach. In all cases data-splitting has less power and therefore finds less significant peaks because it uses half of the data. In practise we observe a random selection of peaks with varying true values meaning that we do not condition on the true values when we observe a peak. As such the bias of the bootstrap estimates is very low (as illustrated in Figure 8) and decreases to 0 asymptotically.

4.2 Estimating the mean for the task fMRI images

Reporting Cohen's d or simply the t-statistic is common practise in fMRI. However it may also be of interest to be able to estimate the underlying mean μ itself; Chen et al. (2017). The bootstrap method can easily be used for this, all that is required is the small modification of Algorithm 2 detailed in Section 2.1.3. The thresholds (and so the number of significant peaks) will be the same as in the previous section as we first threshold the t-statistic to find voxels that are significant and then estimate the effect.

We have produced similar graphs to those of the previous section. For the truth we used the 4000 estimate of the mean rather than Cohen's d: this will be very accurate as an estimate for the true mean by the strong law of large numbers. The box plot in Figure 10 illustrates that the circular estimates are biased whereas the bootstrap and data-splitting estimates have very low bias. The barplots show that the bootstrap estimates have lower MSE than the data-splitting estimates. (Note that for $N = 20$ there are still only 7 data points and so there is not enough data to get a reliable understanding of how it performs.) What is striking however is that for $N = 50$ and 100 the circular estimates

have a lower MSE than the data-splitting estimates indicating that the selection bias is much less severe when estimating the mean such that the variance of the data-splitting estimates dominates. The severity of the selection bias is much less in this scenario. The reason for this is that circularity occurs when you use the same statistic to determine the peak locations and the values observed there. Here the statistics are correlated, but are not the same, meaning that the selection bias is considerably less. The bootstrap estimates are asymptotically unbiased and have the lowest MSE.

We have plotted the graphs comparing the estimates and the truth in Figure 11. As before in the $N = 20$ case, data-splitting finds very few effects whereas the other methods find many significant data-points. The bootstrap is able to correct for the bias very well resulting in estimates that lie along the identity line.

4.3 Results - The GLM on Structural Gray Matter Data

It is of particular interest to be able to obtain unbiased estimates of partial R^2 values. These are widely used though the literature, usually without correction for selection bias; Vul et al. (2009) specifically looked at the bias in correlation values. In this section, we fit model (2) to our gray matter data as discussed in Section 2.3.3. For each group, we generated F-statistic maps in order to test for the presence of age in the model. We took sample sizes $N = 50, 100$ and 200 and as above divided our subjects into G_N groups of size N and applied the three methods using voxelwise RFT to determine significant voxels.[5]

We have produced similar graphs to those of the previous sections. For the truth we used the 4000 subjects to estimate the partial R^2 for age after fitting the linear model. The box plot in Figure 12 illustrates that the circular estimates are highly biased whereas the data-splitting estimates are unbiased and the bootstrap estimates have low bias which tends asymptotically to zero as N increases. The barplots show that the bootstrap estimates have lower MSE than the data-splitting estimates. MSE for the circular estimates in the $N = 50$ case is 0.1161 and so is cut off by the graph. The bootstrap estimates have low bias and have the lowest MSE.

We have plotted the graphs comparing the estimates and the truth in Figure 13. These plots illustrate the convergence of the bootstrap as the sample size increases. In this scenario the estimates for all 3 methods are more variable about the identity line than in the previous examples. The bootstrap overcorrects points with large true values and undercorrects those with smaller true values. This is to be expected given the spread of the circular estimates, with points with a large variety of true partial R^2 values having similar partial R^2 circular estimates. This effect decreases as the sample size increases (as the circular estimates become more parallel to the identity line), however the bootstrap estimates hug the identity more closely than the data-splitting estimates which leads to the drop in variance and MSE shown in Figure 12.

4.4 Application to a Working Memory dataset

In order to illustrate the bootstrap method in action we have applied it to a sample of 80 subjects from the human connectome project and look at one of the working memory contrast. Subjects performed a continuous performance working memory task, an N-back task using alternating blocks of 0-back and 2-back conditions with faces, non-living

[5]We have considered larger sample sizes as estimation (for all the methods) in this setting is more challenging than in the one sample setting.

man-made objects, animals, body parts, house and words. We examined only the average (2-back − 0-back) contrast, identifying brain regions supporting working memory.

We use a group level model comprised of a one-sample t-statistic at each voxel in order to test for activation. Using voxelwise random field theory at the 0.05 level resulting results in a threshold of 5.73 for the t-statistic. The largest peak above the threshold has a t-statistic of 10.38 and lies within the Medial Frontal Gyrus an area commonly associated with working memory. With 80 subjects, 10.38 corresponds to a circular Cohen's d of 1.52 which when corrected using the bootstrap becomes 1.16. In total 192 peaks lay above the threshold with 25 within the Medial Frontal Gyrus. We have displayed the circular and bootstraped Cohen's d as well as the bootstrap estimate of the mean of the top 10 of these 25 peaks in Table 1. Slices through the one-sample t-statistic at the voxel corresponding to the largest peak are shown in Figure 4.4.

To see the effect that these corrections have on power we have plotted a graph of sample size against power for a whole brain analysis using a p-value threshold of 2×10^{-7} in Figure 15. The power is calculated as described in Appendix E.2.

Circular Cohen's d	Corrected Cohen's d	Circular Mean	Corrected Mean	Peak Location
1.519	1.161	45.039	43.315	(31, 67, 64)
1.137	0.922	34.660	32.106	(69, 66, 57)
1.096	0.889	56.135	53.313	(62, 63, 68)
1.091	0.888	27.870	25.720	(31, 70, 60)
1.079	0.883	46.134	43.431	(23, 80, 52)
1.078	0.883	35.137	32.768	(61, 64, 67)
1.078	0.882	37.759	35.628	(25, 80, 54)
1.067	0.876	37.807	35.586	(69, 67, 55)
0.994	0.817	33.937	31.836	(67, 76, 54)
0.979	0.807	28.038	26.042	(65, 66, 64)

Table 1: The circular and corrected estimates of Cohen's d and the mean at the top ten significant peaks in the Medial Frontal Gyrus.

5 Discussion

When conducting a neuroimaging study the first priority is being able to identify significant effects. Given the small size of many studies it is highly undesirable to have to divide the data in half in order to perform accurate inference as this leads to a large decrease in power. In this paper, we have introduced a bootstrapping based method which avoids this problem. We have compared it to data-splitting and circular inference with simulations and real data and have shown that it is asymptotically unbiased and leads to a decrease in the MSE. Relative to data-splitting it results in a large decrease in the variance of the estimates and is thus able to yield a balance in the trade off between bias and variance. We have provided an in-depth analysis of the effect of sample size on selection bias which has enabled us to determine the impact that sample size has on the accuracy of the estimates of peak values. The bootstrap method has the advantage that

it is able to use all of the data to compute the peak locations meaning that it finds more peaks and that its estimates of their locations are more accurate.

Large data repositories of neuroimaging data enabled us to validate our methods in a way that has never been possible before. In this paper we have outlined an easy way of doing so using the UK Biobank, by using a large number of subjects to compute an accurate version of the truth and dividing the remaining subjects into small groups on which to test the methods. We recommend that all emerging statistical methods be tested in this manner. In the interests of reproducibility is it just as important to validate existing methods where such validation has not yet taken place. Additionally it is important that, whenever possible, researchers make their data available so that their results are reproducible and can be improved upon as methods improve. We suggest that researchers store their data on Open fMRI or other such databases.

The circular inference problem is one that is ever present in neuroimaging. In order to provide unbiased estimates of the effect size we recommend using the bootstrap method with 5000 bootstraps, though of course it comes to the number of bootstraps the more that can be performed the better. There is much ongoing research in the field of selective inference and there is lots of potential for other methods to be modified for use in the fMRI setting. There are a number of potential options for further work on the bootstrap. Currently our method corrects locally at the location of the empirical sample maximum. This is appropriate because when someone comes to replicate your results you want them to test the effect at a given location. However using the bootstrap it would also be possible to compare the maximum observed peak value to that of the maximum of the empirical mean, thereby allow an estimate of the true maximum of the process across the whole brain. This problem is something that cannot be solved using data-splitting approaches. One of the difficulties here is that it is hard to precisely match bootstrap peaks to the peaks in the original mean. One approach would be to estimate the bias using only data from some small radius around the peak location. This wouldn't account for the effect of the maxima over the whole image but could still allow for an improved estimate of the bias. It would particularly desirable to derive estimates corrected using random field theory. However it is theoretically very difficult to estimate the peak height distribution of a non mean zero random process, Cheng and Schwartzman (2015).

In the fMRI setting the bootstrap works best in the one-sample scenario due to the high signal to noise ratio. In order to apply it in more general settings where the noise has a larger variance a larger number of subjects is required. In the general multiple regression setting a reasonable number of subjects is needed in order for the estimates to perform well. The bootstrap estimates lower the MSE while allowing for a more accurate estimate of the location. However as the VBM data shows there is still room for improvement and there is lots of scope for future research.

6 Acknowledgements

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Appendices

A Masking and Calculating the Truth

With access to the UK Biobank we have an unprecedented amount of neuroimaging data. This enables us to set aside a large number of subjects in order to get a very accurate estimate of the true effect size, be it the mean, Cohen's d or a coefficient or partial R^2 in a linear model. However dealing with so many subjects requires us to deal with a number of problems. In particular on a normal computer it is not possible to load all of the subjects into memory so we need to take a different approach. In this Appendix we describe how we have dealt with masking and detail how we computed the truth given that we have a very large (in our case 4000) number of subjects.

A.1 Masking

Given a subject: j and a corresponding image Y_j, define its **mask** to be the image: $M_j : \mathcal{D} \longrightarrow \mathbb{R}$ such that

$$M_j(v) = \begin{cases} 1 & \text{if } Z_j(v) \neq \text{na ie if subject } j \text{ has data at voxel } v \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathcal{D}$ is the underlying lattice that we are working on. Given this definition, define the **subject mask** of a subset S of subjects to be the image M such that

$$M(v) = \begin{cases} 1 & M_j(v) = 1 \text{ for all } j \in S \\ 0 & \text{otherwise.} \end{cases}$$

A.2 Full Mean and Full Cohen's d

We chose a random subset S of $1, \ldots, 8940$ of size 4000 and estimated the true mean and Cohen's d using the the available data at each voxel. Define the **full mean** image by

$$\mu_F(v) = \frac{\sum_j Z_j(v) M_j(v)}{\sum_j M_j(v)} \times \mathbb{1}(M_j(v) = 1 \text{ for some } j)$$

and **full population variance** estimate to be:

$$\sigma_F^2(v) = \frac{\sum_j (Z_j - \mu_F(v))^2 M_j(v)}{\sum_j M_j(v) - 1} \times \mathbb{1}(M_j(v) = 1 \text{ for at least 2 } j)$$

and the **full Cohen's** d estimate as

$$d_F(v) = \frac{\mu_F(v)}{\sigma_F(v)}.$$

Since we are interested in the brain itself these images are each multiplied by a mask of the 2mm MNI brain. We follow brain imaging conventions and given a small sample S use the subject mask corresponding to S (multiplied by the MNI mask) in order to perform inference on the S, and use the full mean or Cohen's d as our estimates of the ground truth.

A.3 Full Linear Model

Our images have 902,629 = 91*109*91 voxels and for 4000 subjects this data occupies 27GB RAM at double precision, presenting serious computational challenges. Here we outline a method for computing linear models when the data cannot be loaded into RAM all at once. Fitting a separate linear model at each voxel is very computationally intensive as it requires one to load all of the images again at each of the 902,629 voxels in order to extract the data at that voxel. Loading all of the images is time consuming and so this is not a practical approach. One method to speed up the procedure is to divide the brain image into blocks that can fit in memory and be dealt with in reasonable amounts of time. This works but still takes a substantial amount of time. Instead it is possible to take advantage of the form of the linear model in order to quickly calculate the estimates in a linear model for arbitrarily large datasets. Let J be the number of subjects and let $Y_1, \ldots, Y_J$ and $M_1, \ldots, M_J$ be the corresponding images and masks respectively. Suppose that we have a design matrix $X = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_J \end{bmatrix}$ corresponding to an intercept and a single regressor, and that there is no missing data.

Then in order to estimate the true coefficients for the linear model we need to compute:

$$(X^T X)^{-1} X^T Y(v)$$

at each voxel v in the MNI brain, where $Y = [Y_1, \ldots, Y_J]^T$. To do so we can use the fact that

$$X^T Y(v) = \begin{pmatrix} \sum_j Y_j(v) \\ \sum_j X_j Y_j(v) \end{pmatrix}$$

loading one image at a time and summing this allows us to vectorize and quickly calculate $X^T Y$. All that is then required is a quick run through the images, calculating the sums as you go along, and pre-multiplying $X^T Y(v)$ by the inverse of

$$X^T X = \begin{pmatrix} J & \sum_j X_j \\ \sum_j X_j & \sum_j X_j^2 \end{pmatrix}$$

which only has to be calculated once. Running though the images sequentially σ^2, t and partial R^2 values can easily be calculated. This method can easily be extended to multiple regressors.

A.4 Full Linear Model with Masking

So far we have been assuming that all of our images have data in the same locations, however in reality due to subject specific factors such as distortion this is not the case. Thus we need to be able to take account of the individual subject masks. There are missing data approaches to this, however here we take the complete data approach which is unbiased under the assumption that the data is missing at random. At each voxel we estimate the relevant statistics using the data that is available at that voxel. For each voxel v, let $C(v) := \{j : M_j(v) = 1\}$ then we need to compute

$$(X_{C(v)}^T X_{C(v)})^{-1} X_{C(v)}^T Y_{C(v)}.$$

Now,

$$X_{C(v)}^T Y_{C(v)} = \begin{pmatrix} \sum_j M_j(v) Y_j(v) \\ \sum_j M_j(v) X_j(v) Y_j(v) \end{pmatrix} \text{ and}$$

$$(X_{C(v)}^T X_{C(v)})^{-1} = \begin{pmatrix} \sum_j M_j(v) & \sum_j M_j(v)X_j(v) \\ \sum_j M_j(v)X_j(v) & \sum_j M_j(v)X_j(v)^2 \end{pmatrix}^{-1}$$

so we can perform the same trick as in the full linear model without masking: running through the images and summing as you go along. Similar tricks can be used to compute the standard deviation, t-statistic and partial R^2 values. This works efficiently when we have p regressors and p is small, however the matrix $(X_{C(v)}^T X_{C(v)})^{-1}$ is different for each voxel and so has to be stored at each voxel and updated 4000 times. Storing a matrix of size p^2 at each voxel becomes highly memory intensive as p gets much larger than 15. In our examples the linear models only have a few regressors so this is not a problem.

B Bias, MSE and Variance Computations

We compute bias, variance and MSE in a non-standard context, in that the true parameter values vary in each instance. Traditionally, one estimates a single θ based with estimators $\hat{\theta}_1, ..., \hat{\theta}_n$, giving us the usual MSE decomposition for a sample of size n:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i)^2 + \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \theta\right)^2$$

into variance and squared bias. However in our context we have estimators $\hat{\theta}_1, \ldots, \hat{\theta}_n$ of parameters $\theta_1, \ldots, \theta_n$. In our setting, for a sample size N, n is the number of significant peaks that are found over all realizations. For $i = 1, \ldots, n$, $\hat{\theta}_i$ is the value of one of these significant peaks. Suppose that it occurs at voxel $\hat{v}_i$, then θ_i is the value of the truth at that voxel. The θ_i are different because the location of the empirical peak is different for each group and for each significant peak in that group. To determine how far off the estimators are on average we can consider the mean squared error (MSE):

$$\text{MSE} \ = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2.$$

Let $\text{eVar} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i - \hat{\mu})^2$ and let $\hat{\mu} := \text{eBias} := \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)$. Then

$$\begin{aligned}
\text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\theta}_i - \theta_i\Big)^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\theta}_i - \theta_i - \hat{\mu} + \hat{\mu}\Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i - \hat{\mu})^2 + \frac{2}{n}\hat{\mu}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i - \hat{\mu}) + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i - \hat{\mu})^2 + \frac{2}{n}\hat{\mu}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i) - \frac{2}{n}n\hat{\mu}^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i - \hat{\mu})^2 + 2\hat{\mu}^2 - 2\hat{\mu}^2 + \hat{\mu}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i - \hat{\mu})^2 + \hat{\mu}^2 = \text{eVar} + \text{eBias}^2.
\end{aligned}$$

Under the assumption of a linear offset such that $\hat{\theta}_i = \mu + \theta_i + \epsilon_i$ for some mean μ and some error terms ϵ_i which have variance σ^2, we have that $\hat{\mu}$ is an unbiased estimator for

μ and eVar is an (asymptotically) unbiased estimator for σ^2. Which means that eBias is the empirical average bias of the estimators and the eVar is interpretable as the empirical variance of each bias $\hat{\theta}_i - \theta_i$ about this average.

In the context of our estimates in Section 4, for a given sample size N, n is the number of significant peaks over all the G_N subsets. This allows us to define the MSE, variance and bias in this context. In the box plots in Figures 8, 10, 12 for each set of estimates we have made a box plot for the bias $\hat{\theta}_i - \theta_i$ over all n significant peaks. For the bar plots in these figures we have plotted the MSE and variance as defined above.

C partial R^2 in terms of F

Let RSS_Ω and RSS_ω respectively be the residual sum of squares for the overall model Ω and some sub-model $\omega \subset \Omega$ with p_0 degrees of freedom. Then we can write the F statistic for comparing ω and Ω as

$$F = \frac{(\text{RSS}_\omega - \text{RSS}_\Omega)/m}{\text{RSS}_\Omega/N - p}$$

where $m = p - p_0$ and the partial coefficient of determination is:

$$R^2 = 1 - \frac{\text{RSS}_\Omega}{\text{RSS}_\omega}.$$

So,

$$F = \frac{n-p}{m}\left(\frac{\text{RSS}_\omega}{\text{RSS}_\Omega} - 1\right) = \frac{n-p}{m}\left(\frac{1}{1-R^2} - 1\right) = \frac{n-p}{m}\left(\frac{R^2}{1-R^2}\right)$$

and rearranging this we have that

$$\frac{1}{1-R^2} = \frac{m}{n-p}F + 1 \implies$$

$$R^2 = 1 - \left(\frac{m}{n-p}F + 1\right)^{-1} = 1 - \frac{n-p}{mF+n-p} = \frac{mF}{mF+n-p}.$$

The F-statistic above has a different form to the F-statistic defined in Section 2.2. For every contrast matrix C taking the sub-model $\omega_C = \{\beta : C\beta = 0\}$ and applying the General Linear Hypothesis establishes their equivalence.

D Neighbourhoods and Local Maxima

Suppose that the vertices in $\mathcal{V}$ are connected by a set of edges. Let the collection of these edges be denoted by E. Then we define two vertices u and v to be **neighbours** in the graph $\mathcal{G} = (\mathcal{V}, E)$ if the edge connecting u and v which we denote by uv is contained in the set of edges E. Given $v \in \mathcal{V}$, define the **neighbourhood** of v to be the set of voxels that are neighbours to v and denote this by $\mathbf{ne(v)}$.

Now given an image $Z : \mathcal{V} \longrightarrow \mathbb{R}$, we define a voxel v to be a **local maxima** if $Z(v) \geq Z(v')$ for all $v' \in ne(v)$. Strictly speaking v is a local maximum with respect to the image Z but this is almost always clear from context. Given a vector valued image we can use a similar definition to define local maxima but need to ensure that $\geq$ is defined differently so that it can be used to compared two vectors.

Typically in 3D brain images we take the edge set to be defined by a connectivity criterion of either 6, 18 or 26, which if our voxels are represented by cubes correspond to those surrounding voxels which share surfaces, edges and corners respectively. As a result the neighbourhood of each voxel and so the definition of the local maxima are dependent on the connectivity criterion.

E Non-Central Distributions and Power Analyses

E.1 Non-Central Distributions

E.1.1 One-Sample t-statistic

Following the model from Section 2.1.2 we have that

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} Y_n \sim N(\mu, \sigma^2/N) \text{ independently of } \hat{\sigma}^2 \sim \frac{\sigma^2}{N-1}\chi^2_{N-1}$$

at each voxel. As such $\hat{\mu}\sqrt{N} =_d \mu\sqrt{N} + N(0, \sigma^2)$ independent of $\hat{\sigma}$ and so the t-statistic $\hat{\mu}\sqrt{N}/\tilde{\sigma}$ has a non-central t-distribution with non-centrality parameter $\mu\sqrt{N}/\sigma$ and $N-1$ degrees of freedom. As such

$$\mathbb{E}\left[\frac{\hat{\mu}\sqrt{N}}{\tilde{\sigma}}\right] = \frac{\mu}{\sigma}\sqrt{\frac{N-1}{2}}\frac{\Gamma((N-2)/2)}{\Gamma((N-1)/2)} = \frac{C_N\mu}{\sigma}$$

for $N > 2$, where Γ is the gamma function, where C_N is the correction factor. This uses the formula for the mean of the non-central t distribution. In particular it follows that

$$\frac{\hat{\mu}\sqrt{N}}{\tilde{\sigma}C_N}$$

is an unbiased of the population Cohen's $d : \frac{\mu}{\sigma}$.

E.1.2 General Linear Model

We have that $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$ independently of $\hat{\sigma}^2 \sim \frac{\sigma^2}{N-p}\chi^2_{N-p}$ and so

$$(C(X^TX)^{-1}C^T)^{-1/2}C\hat{\beta} \sim N\big((C(X^TX)^{-1}C^T)^{-1/2}C\beta, \sigma^2 I_m\big)$$

and so $(C\hat{\beta})^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta})$ has a non-central chi-squared distribution with m degrees of freedom and non-centrality parameter $(C\beta)^T(C(X^TX)^{-1}C^T)^{-1}(C\beta)$ and as such

$$F = \frac{(C\hat{\beta})^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta})/m}{\hat{\sigma}^2}$$

has a non-central F distribution with non-centrality parameter $(C\beta)^T(C(X^TX)^{-1}C^T)^{-1}(C\beta)/\sigma^2$ and degrees of freedom m and $N-p$. In particular

$$\mathbb{E}[F] = \frac{(N-p)(m + (C\beta)^T(C(X^TX)^{-1}C^T)^{-1}(C\beta)/\sigma^2)}{m(N-p-2)}$$

In the case where $C = c^T$ is just a single contrast vector and we want to perform inference using the t-statistic instead of the F-statistic, the t-statistic

$$\frac{c^T\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T(X^TX)^{-1}c}}$$

has a non-central t-distribution with $N-p$ degrees of freedom and non-centrality parameter $c^T\beta/\sqrt{\sigma^2 c^T(X^TX)^{-1}c}$.

E.2 Power Analyses

E.2.1 One Sample

In the one sample scenario, for a sample size N' and an estimate of the non-centrality parameter: λ, the power is:

$$\mathbb{P}(T_{N'-1,\lambda} > t_{1-\alpha,N'-1})$$

where $t_{1-\alpha,N'-1}$ is chosen such that $\mathbb{P}(T_{N'-1,0} > t_{1-\alpha,N'-1}) = \alpha$ and $T_{N'-1,\lambda}$ has a non-central T distribution with $N'-1$ degrees of freedom and non-centrality parameter λ.

E.2.2 Multiple Regression - Cohen's f^2

Cohen's f^2 is defined to be $f^2 := \dfrac{R^2}{1-R^2} = \dfrac{m}{N-p}F$ (as shown in Appendix C) where R^2 is the partial R^2. Now,

$$\frac{m}{N-p}F = \frac{(C\hat{\beta})^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta})/(N-p)}{\hat{\sigma}^2} = \frac{(C\hat{\beta})^T(C(\frac{1}{N-p}X^TX)^{-1}C^T)^{-1}(C\hat{\beta})}{\hat{\sigma}^2}.$$

Suppose that $X = \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}^T$ where $\{x_n\}_{n\in\mathbb{N}}$ is a sequence of iid random vectors from some multivariate distribution D. Then $(\frac{1}{N-p}X^TX)_{i,j} \to \mathbb{E}[x_{1i}x_{1j}]$ as $N \to \infty$ by the strong law of large numbers. Also, $\hat{\beta} \to \beta$ and $\hat{\sigma}^2 \to \sigma^2$ as $N \to \infty$. As such f^2 converges to a population value of

$$f_p^2 = \frac{(C\beta)^T(C(\mathbb{E}X^TX)^{-1}C^T)^{-1}(C\beta)}{\sigma^2}$$

and this also implies convergence of R^2. Given a new sample of N' subjects with design matrix X' an $N' \times p$ matrix such that the rows are iid with distribution D, so long as N' is sufficiently large, we can obtain reasonable estimates of the power. To do so note that the (new) F-statistic has a non-central F distribution with non-centrality parameter:

$$\frac{(C\beta)^T(C(X'^TX')^{-1}C^T)^{-1}(C\beta)}{\sigma^2} = N'\frac{(C\beta)^T(C(\frac{1}{N'}X'^TX')^{-1}C^T)^{-1}(C\beta)}{\sigma^2} \approx N'f_p^2 \approx N'f^2.$$

Let $\lambda = N'f^2$ be the estimate of the non-centrality parameter. Then the power is:

$$\mathbb{P}(F_{m,N'-p,\lambda} > f_{1-\alpha,m,N'-p})$$

where $f_{1-\alpha,m,N'-p}$ is chosen such that $\mathbb{P}(F_{m,N'-p,0} > f_{1-\alpha,N'-1}) = \alpha$ and where $F_{m,N'-p,\lambda}$ has a non-central T distribution with $N'-1$ degrees of freedom and non-centrality parameter λ.

E.2.3 Multiple Regression - Cohen's f

In the case that $C = c^T$ is a contrast vector, we often use the t-statistic as this allows us to perform one-sided tests. In which case we can use Cohen's f which is defined as

$$f = \frac{c^T \hat{\beta}/(N-p)}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}$$

and use $\sqrt{N'} f$ as our estimate of the non-centrality parameter using this to calculate an estimate of the power analogously to above.

References

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper L R Andersson, Stamatios N Sotiropoulos, Saad Jbabdi, Ludovica Griffanti, Moises Hernandez-fernandez, Emmanuel Vallee, Diego Vidaurre, Matthew Webster, Paul Mccarthy, Christopher Rorden, Alessandro Daducci, Daniel C Alexander, Hui Zhang, Iulius Dragonu, Paul M Matthews, Karla L Miller, and Stephen M Smith. NeuroImage Image processing and Quality Control for the fi rst 10 , 000 brain imaging datasets from UK Biobank. 166 (April 2017):400–424, 2018. doi: 10.1016/j.neuroimage.2017.10.034.

Yoav Benjamini and Amit Meir. Selective Correlations - the conditional estimators. page 18, 2014. URL `https://arxiv.org/abs/1412.3242`.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013. ISSN 00905364. doi: 10.1214/12-AOS1077.

Gang Chen, Paul A. Taylor, and Robert W. Cox. Is the statistic value all we should care about in neuroimaging? *NeuroImage*, 147(October 2016):952–959, 2017. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.09.066.

Dan Cheng and Armin Schwartzman. Multiple Testing of Local Extrema for Detection of Change Points. (2008):34, 2015. ISSN 0090-5364. doi: 10.1214/11-AOS943.

A C Davison, D V Hinkley, and G A Young. Recent Developments in Bootstrap Methodology. *Statistical Science*, 18(2):141–157, 2003. ISSN 0883-4237. doi: 10.1214/ss/1063994969.

Bradley Efron. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm11181.

Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1602413113.

Anders Eklund, Hans Knutsson, and Thomas E Nichols. Cluster Failure Revisited: Impact of First Level Design and Data Quality on Cluster False Positive Rates. 2018. doi: 10.1101/296798.

Michael Esterman, Benjamin Tamber-Rosenau, Yu-Chin Chiu, and Steven Yantis. Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage*, 50(2):572–576, 2010. doi: 10.1016/j.neuroimage.2009.10.092.Avoiding.

John P. Ferguson, Judy H. Cho, Can Yang, and Zhao Hongyu. Empirical Bayes Correction for the Winner's Curse in Genetic Association Studies. *Genetic Epidemiology*, 37 (1):60–68, 2013. doi: 10.1002/gepi.21683.Empirical.

K J Friston, K. J. Worsley, R S J Frackowiak, J C Mazziotta, and A C Evans. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214–220, 1994.

Arpita Ghosh, Fei Zou, and Fred A Wright. Estimating Odds Ratios in Genome Scans: An Approximate Conditional Likelihood Approach. *The American Journal of Human Genetics*, (May):1064–1074, 2008. doi: 10.1016/j.ajhg.2008.03.002.

H. Göring, Joseph D. Terwilliger, and John Blangero. Large Upward Bias in Estimation of Locus-Specific Effects from Genomewide Scans. *The American Journal of Human Genetics*, (69):1357–1369, 2001.

Ahmad R. Hariri, Alessandro Tessitore, Venkata S. Mattay, Francesco Fera, and Daniel R. Weinberger. The amygdala response to emotional stimuli: A comparison of faces and scenes. *NeuroImage*, 17(1):317–323, 2002. ISSN 10538119. doi: 10.1006/nimg.2002.1179.

W. James, W. James, C. Stein, and C. Stein. Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, pages 361– 379, 1961. ISSN 0097-0433.

Neal O Jeffries. Multiple comparisons distortions of parameter estimates. *Biostatistics*, pages 500–504, 2007. doi: 10.1093/biostatistics/kxl025.

Nikolaus Kriegeskorte, Martin A Lindquist, Thomas E. Nichols, Russell A Poldrack, and Edward Vul. Everything You Never Wanted to Know about Circular Analysis, but Were Afraid to Ask. *Journal of Cerebral Blood Flow & Metabolism*, 30(9):1551–1557, 2010a. ISSN 0271-678X. doi: 10.1038/jcbfm.2010.86.

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience – the dangers of double dipping. 12(5): 535–540, 2010b. doi: 10.1038/nn.2303.Circular.

Jason D Lee and Jonathan Taylor. Exact Post Model Selection Inference for Marginal Screening. *In Advances in Neural Information Processing Systems*, 1(2):1–9, 2014.

Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper L R Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M Matthews, and Stephen M Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016. ISSN 1546-1726. doi: 10.1038/nn.4393. URL `http://www.nature.com/doifinder/10.1038/`

nn.4393{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27643430{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5086094.

Jeanette A. Mumford. A power calculation guide for FMRI studies. *Social Cognitive and Affective Neuroscience*, 7(6):738–742, 2012. ISSN 17495016. doi: 10.1093/scan/nss059.

Thomas E. Nichols and Andrew P Holmes. Nonparametric Permutation Tests for {PET} functional Neuroimaging Experiments: A Primer with examples. *Human Brain Mapping*, 15(1):1–25, 2001. ISSN 1065-9471. doi: 10.1002/hbm.1058.

Stephen Reid, Jonathan Taylor, and Robert J Tibshirani. Post selection point and interval estimation of signal sizes in Gaussian samples. *arXiv preprint arXiv:1405.3340*, (1):1–22, 2014. ISSN 1708945X. doi: 10.1002/cjs.11320.

J.D. Rosenblatt and Y. Benjamini. Selective correlations; not voodoo. *NeuroImage*, 103:401–410, dec 2014. ISSN 10538119. doi: 10.1016/j.neuroimage.2014.08.023. URL http://linkinghub.elsevier.com/retrieve/pii/S1053811914006910.

D Siegmund. Upward Bias in Estimation of Genetic Effects. *The American Society of Human Genetics*, pages 1183–1188, 2002.

Noah Simon and Richard Simon. On Estimating Many Means, Selection Bias, and the Bootstrap. 2013.

Charles Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206, 1956.

Lei Sun and Shelley B. Bull. Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology*, 28(4):352–367, 2005. ISSN 07410395. doi: 10.1002/gepi.20068.

Kean Ming Tan, Noah Simon, and Daniela Witten. Selection Bias Correction and Effect Size Estimation under Dependence. 2014.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences of the United States of America*, 112 (25):7629–34, 2015. ISSN 1091-6490. doi: 10.1073/pnas.1507583112.

E Vul, C Harris, P Winkielman, and H Pashler. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3):274–290, 2009.

Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashlet. Reply to Comments on "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition". 4(3):274–290, 2011. ISSN 17456916. doi: 10.1111/j.1745-6924.2009.01132.x.

K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K.J. Friston, and A. C. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1):58–73, 1996. ISSN 10659471. doi: 10.1002/(SICI)1097-0193(1996)4:1⟨58::AID-HBM4⟩3.0.CO;2-O.

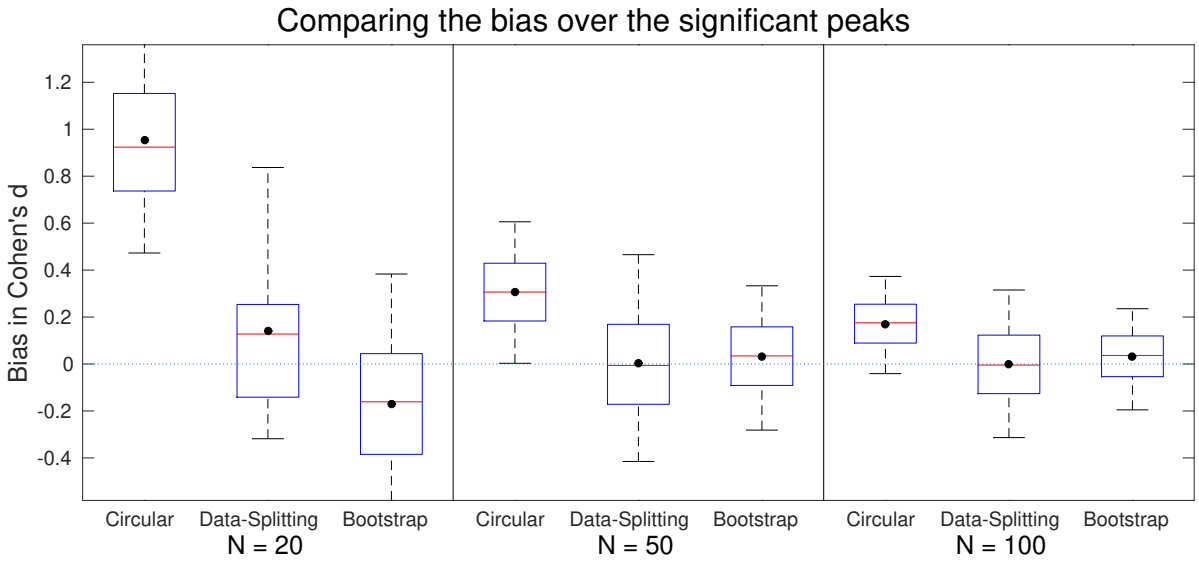
Long Wu, Lei Sun, and Shelley B. Bull. Locus-Specific Heritability Estimation via the Bootstrap in Linkage Scans for. *Human Heredity*, 9:84–96, 2006. doi: 10.1159/000096096.

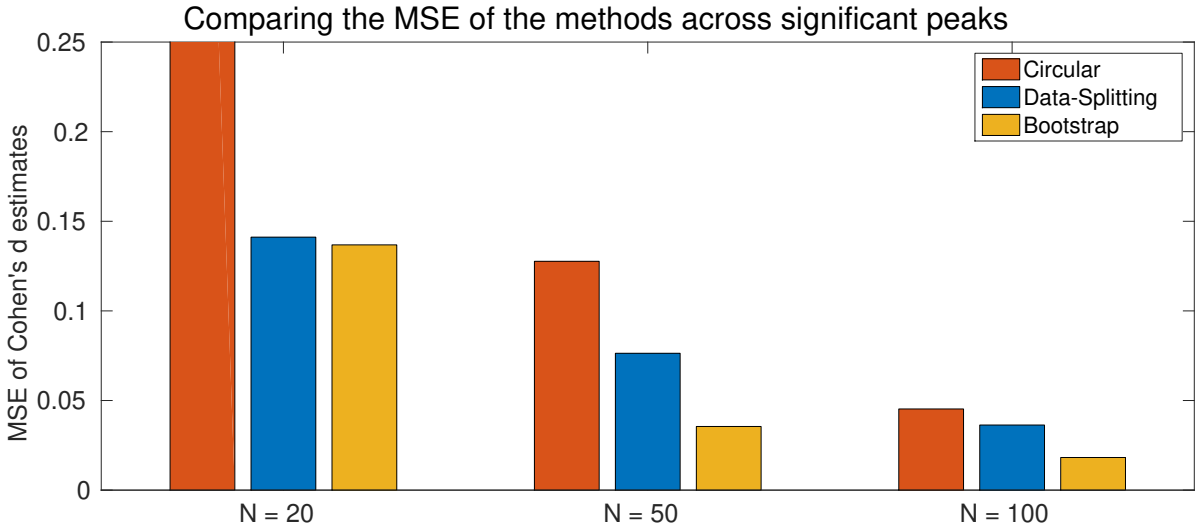
Rui Xiao and Michael Boehnke. Quantifying and correcting for the winner's curse in quantitative- trait association studies Rui. *Genetic Epidemiology*, 35(3):133–138, 2012. doi: 10.1002/gepi.20551.Quantifying.

Kai Yu, Nilanjan Chatterjee, William Wheeler, Qizhai Li, Sophia Wang, Nathaniel Rothman, and Sholom Wacholder. Flexible Design for Following Up Positive Findings. *The American Journal of Human Genetics*, 81(September):540–551, 2007. doi: 10.1086/520678.

Hua Zhong and Ross L Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics (Oxford, England)*, 9(4): 621–34, oct 2008. ISSN 1468-4357. doi: 10.1093/biostatistics/kxn001.

Sebastian Zöllner and Jonathan K. Pritchard. Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data. *The American Journal of Human Genetics*, 80(4):605–615, 2007. ISSN 00029297. doi: 10.1086/512821.





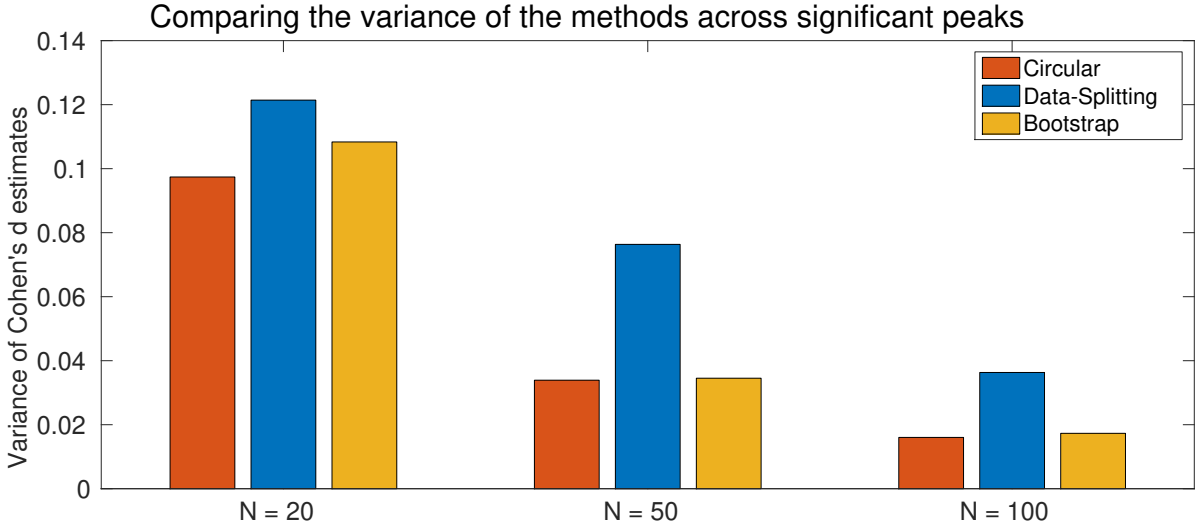


Figure 8: Comparing estimates of the one sample Cohen's d for task fMRI. For $N = 20, 50$ and 100 we implemented the methods on each of the G_N groups. We plotted estimates of the average variance and MSE for each N. For the bias we plotted box plots of the bias where the average bias is indicated by a black dot on each box plot. The bootstrap estimates are asymptotically unbiased and have low MSE. Note that the $N = 20$ data-splitting MSE and variance is computed using only 7 data points so may not be representative. See Appendix B for definitions of the MSE, variance and bias.

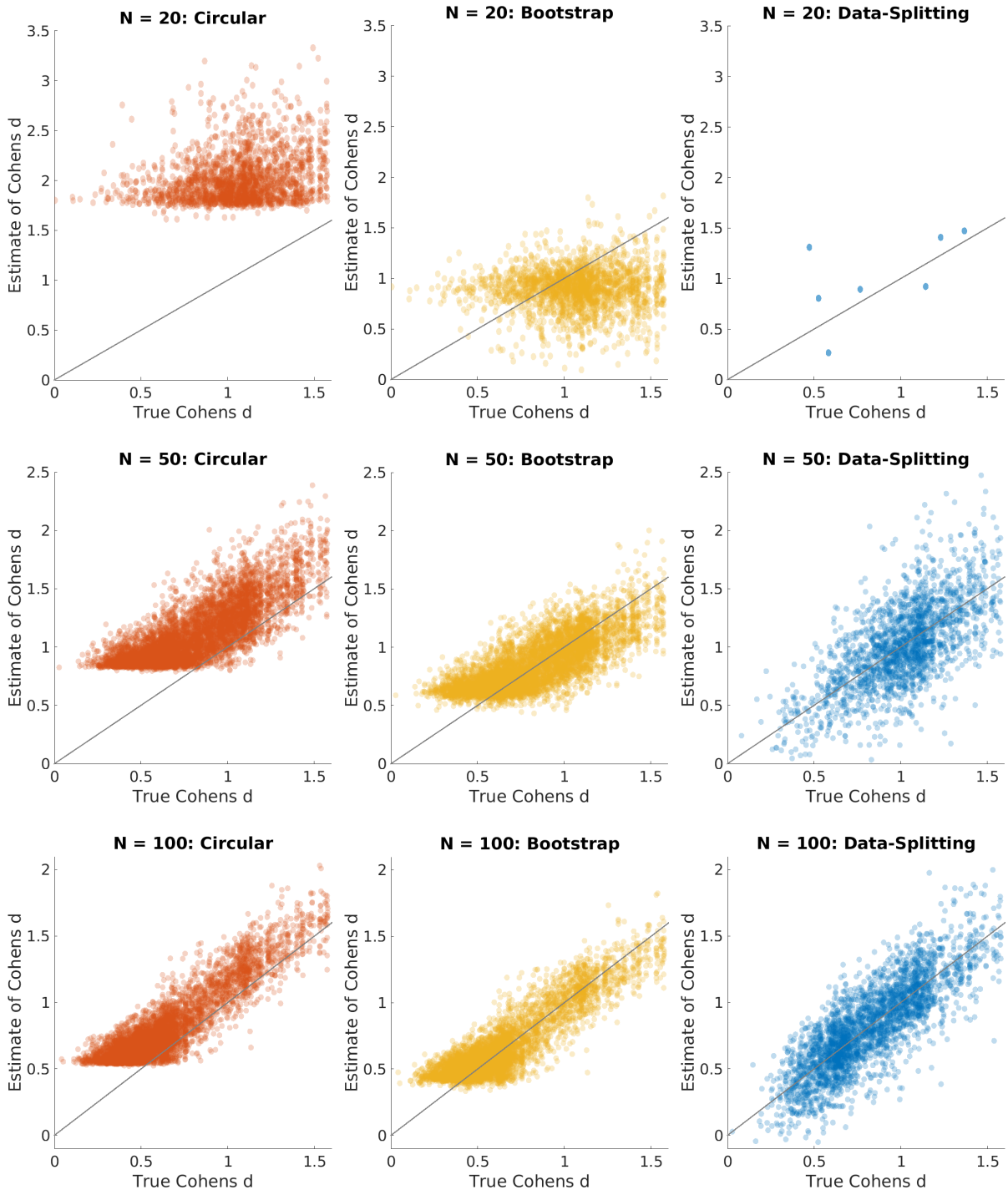


Figure 9: Plotting estimates of the one-sample Cohen's d for task fMRI images against the true values at the estimated locations. For each $N = 20, 50, 100$ we implemented our methods on the G_N samples and for each sample found a number of peaks deemed to be significant; each dot in the above plots represents one of these peaks. For each peak we have plotted the estimate of Cohen's d at its location against the truth at that location. Note that the number of peaks and their locations are the same for circular inference and the bootstrap but is different for data-splitting because it uses the first half of the subjects in order to determine significant peaks. From these plots we can see that the bootstrap estimates have low bias and variance and improve as the sample size increases. The data-splitting estimates are unbiased but are more variable and less numerous.

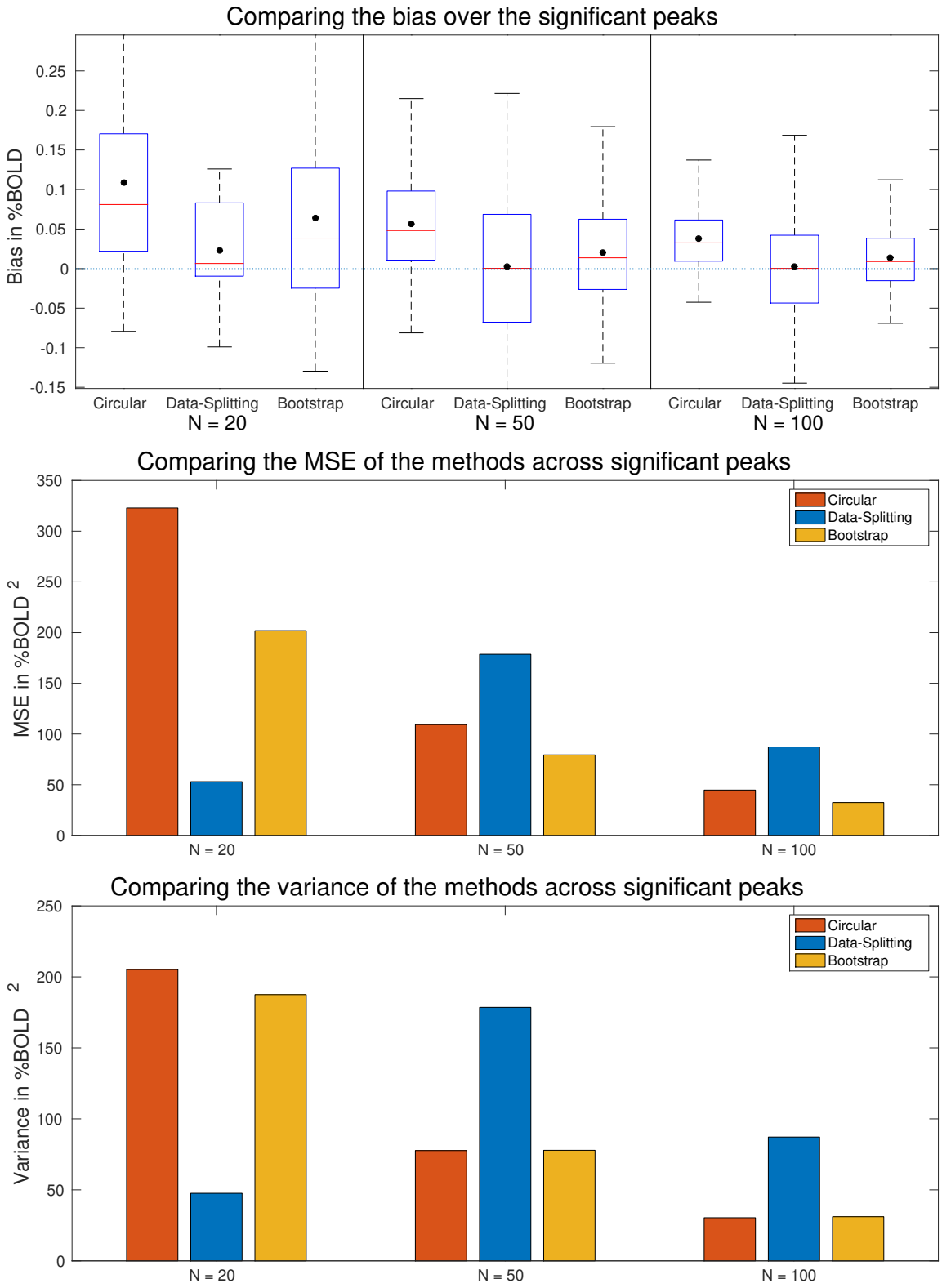


Figure 10: Comparing estimates of the significant local maxima of the one sample mean for task fMRI. For $N = 20, 50$ and 100 we have implemented the methods on each of the G_N groups. We have plotted estimates of the average variance, MSE and bias for each N as in Figure 8. Note that the $N = 20$ data-splitting MSE and variance is computed using only 7 data points so may not be representative. See Appendix B for definitions of the MSE, variance and bias. Particularly of note here is that the circular estimates already have lower MSE than data-splitting.

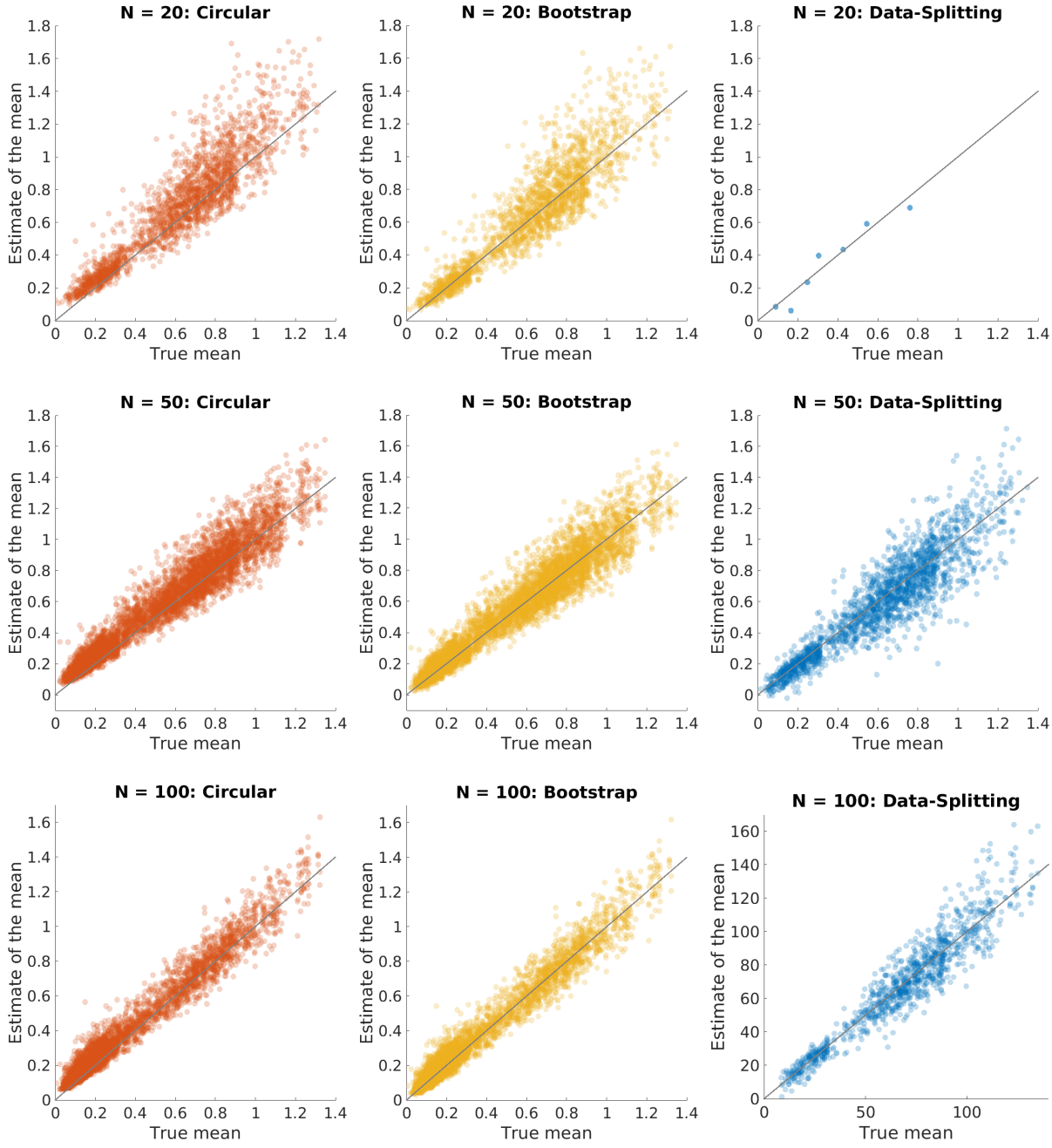


Figure 11: Plotting estimates of the one-sample mean against the true values at the estimated locations using task fMRI. For each $N = 20, 50, 100$ we implemented our methods on the G_N samples and for each sample find a number of a peaks to be significant: each dot in the above plots represents one of these peaks. For each peak we have plotted the estimate of the mean (in % BOLD) at its location against the truth at that location. Note that the number of peaks and their locations are the same for circular inference and the bootstrap but is different for data-splitting because it uses the first half of the subjects in order to determine significant peaks. From these plots we see that the circularity bias is much less than for Cohen's d and that the bootstrap estimates perform very well. The data-splitting estimates are unbiased but are more variable and less numerous.

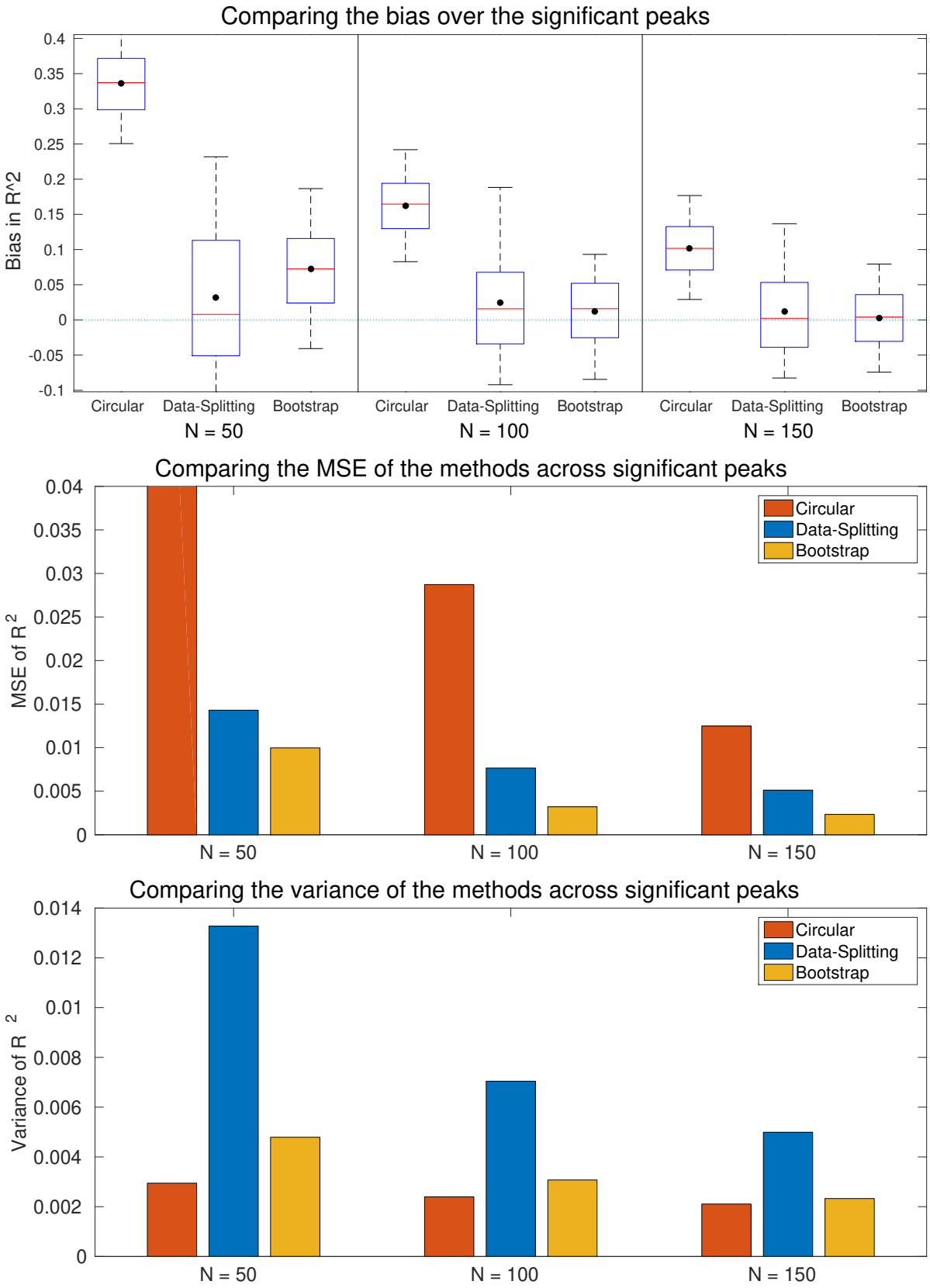


Figure 12: Comparing estimates of the significant local maxima of the partial R^2 for age on VBM data. For $N = 50, 100$ and 150 we have implemented the methods on each of the G_N groups. We have plotted estimates of the average variance and MSE for each N. For the bias we have plotted boxplots of the bias over each significant peak. The average bias is indicated by a black dot on each boxplot. From these graphs we see that the bootstrap estimates are asymptotically unbiased and have the lowest MSE in all cases while the variance of the data-splitting estimates is comparatively high.

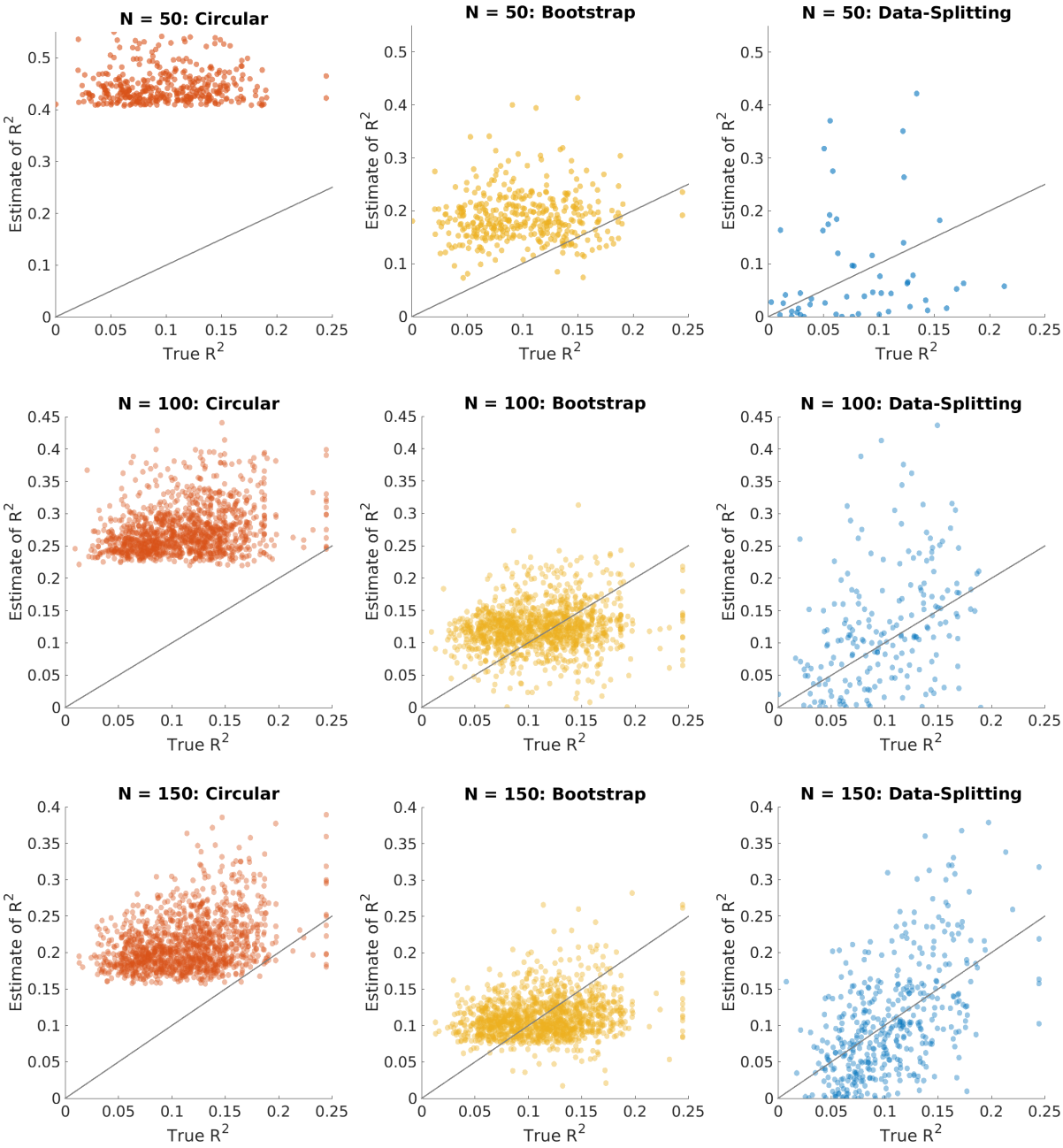


Figure 13: Plotting estimates of the estimated partial R^2 for age (obtained using a regression on VBM data) against the true values at the estimated locations. For each $N = 50, 100, 150$ we implemented our methods on the G_N samples and for each sample calculated the F and partial R^2-statistic maps in order to find significant peaks: each dot in the above plots represents one of these peaks. For each peak we have plotted the estimate of the mean at its location against the truth at that location. Note that the number of peaks and their locations are the same for circular inference and the bootstrap but is different for data-splitting because it uses the first half of the subjects in order to determine significant peaks. From these plots we see that the naive estimates are biased high while the bootstrap and data-splitting estimates are unbiased on average. Data-splitting is the most variable, though the bootstrap overcorrects values with a large true partial R^2 and undercorrects those with a low partial R^2 for $N = 100, 150$. The bootstrap estimates hug the identity line more closely than the data-splitting estimates resulting in the decrease in MSE, see Figure 12.

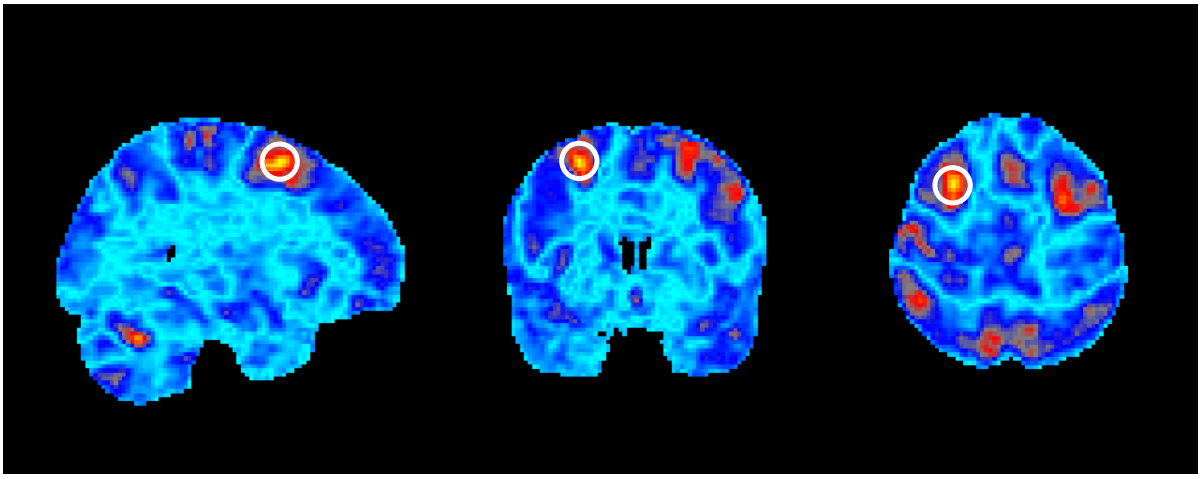


Figure 14: Slices through the one-sample t-statistic for the working memory contrast (2-back − 0-back) for subjects from the Human Connectome Project. White circles indicate the location of the largest peak of activation which lies at the voxel (31, 67, 69) at the edge of the Medial Frontal Gyrus.

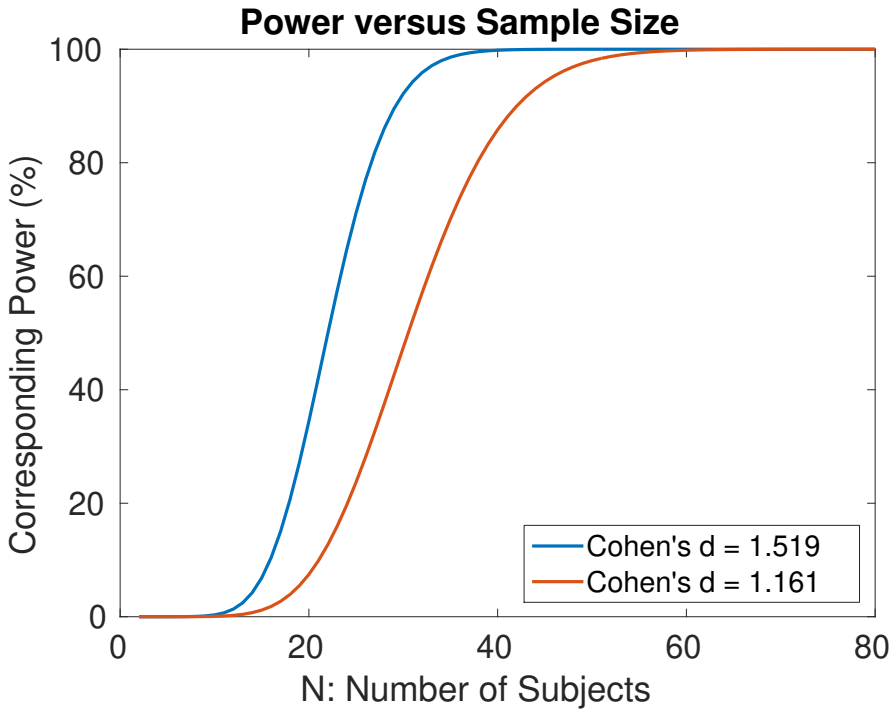


Figure 15: The power corresponding to a given sample size for the one-sample t-statistic. The blue curve is the power curve corresponding to the circular estimate of the Cohen's d for the HCP working memory dataset and the read curve is the power curve corresponding to the corrected Cohen's d.