# Statistical Inference in fMRI using Random Field Theory and Resampling Methods

Samuel Davenport

St Peter's College, Oxford

DPhil in Statistics (Oxford Warwick Statistics Program)

# Contents

## 0.1 Acknowledgements

There are a huge number of people that I would like to thank. First and foremost I would like to thank Dad, Mum, Jessie and Gwendy and K/C for being the most lovely family that I could ever wish for. They have always been incredibly supportive of me, putting up with my tendencies to study maths for long days on end. They are my light in the dark and my shoulder to cry on and I will be eternally grateful for their love and support. It is quite amazing to think that I have followed in Mum's fantastical footsteps (Braun, 1994).

I would also like to extend a huge amount of thanks to all of my friends. In particular to Bella who was always there for me when things were difficult. And to Bhavik who was always there to chat with despite being half way across the planet most of the time! And to my many dancing friends from all across the world, especially to Laurel, Alice, Izzie, Harvey, Quintin and the rest of Air Patrol and David, Lucy, Mo, Pete, Vicky, Tayo and Leah who all made my life so full of joy and prevented me from falling into a maths black hole :), despite my best efforts at times! One of the great benefits of being on the OxWaSP program is the amazing cohort and the friendships that have resulted from it. In particular I would like to thank Jeremias, Arne, Giulio, Jack, Petya, Emilia and of course Romek (OxWaSP imposter) for being amazing friends. Also thanks very much to Joanna, Emma and Beverley for their great support and to my amazing office buddies: Rob and Anthony, may we long continue annoying the rest of the world by playing chess at high volume for eternity. I'd also like to thank my fellow Cambridge mathmos, especially Ellie, Jeremy and Nikos (it seems I won't be following you this time Nick haha, unless you're also planning to go to San Diego next year). I would also like to say thanks to Matthew for being a really great friend and to Verena for being an amazing housemate!

I owe a huge debt of thanks to my supervisors Tom Nichols and Armin Schwartzman. Tom has always been incredibly helpful and willing to put in the time and especially to put up with my endless desire to work on Random Field Theory :). Armin has been extremely supportive and has been a huge help, I am very excited to begin my post-doctoral work with him and to continue solving RFT together. I would like to give a special shoutout to Armin's lovely family who I look forward to getting to know even better once I get to San Diego. I also owe a huge amount of thanks to Fabian Telschow, it has been an amazing experience to collaborate so closely together and I very much hope to convince him to remain in academia. I have developed a close and productive working relationship and friendship with these three and have learnt an enormous amount from all of them. This has been incredibly enjoyable and fulfilling and I very much look forward to collaborating closely together in the future. I would of course like to thank all of the members of my (well Tom's) group! In particular, Petya (two mentions!), Alex, Tom, Olivier, Marco, Habib, Soroosh, Brieuc, Simon and Jessie for being a huge support.

Finally I must add that I owe a huge debt of gratitude to GOD AKA Robert Adler for writing the bible AKA the Geometry of Random Fields (Adler, 1981) and for creating the field in which I now work. And of course to Keith Worsley (may he rest in peace), on whose papers I hope to continue building throughout my career, and who I very much hope would be pleased to see me extending his work.

## 0.2 Abstract

This thesis provides a set of tools for analysing random images with a specific focus on applications in functional Magnetic Resonance Imaging (fMRI). To do so we employ Random Field Theory (RFT), a set of theoretical parametric results that can be used to analyse multidimensional random processes (known as random fields), and resampling methods, which draw samples from the data (with or without replacement).

We extend the voxelwise inference framework of Worsley et al. (1992) so that it provides accurate control of the familywise error rate in neuroimaging. We drop the standard RFT assumptions of high smoothness and stationarity and develop a quick parametric framework that provides powerful and valid inference even when the underlying data is non-Gaussian. We validate this using a massive resting state analysis, involving brain imaging data from 7000 subjects from the UK Biobank.

We further use RFT techniques to derive an asymptotic distribution for the extent of a cluster above a threshold $u$ in a non-stationary Gaussian random field as $u \to \infty$. To do so we define the notion of horizontal-window (HW) conditioning and take advantage of recent advances (Cheng and Schwartzman (2015a)) on the HW-distribution of the height of a peak in a non-stationary Gaussian random field. Our results extend those of Nosko (1969) in which the asymptotic cluster size distribution is derived under the assumption of stationarity.

In order to infer upon random fields whose mean is non-zero we derive asymptotic confidence regions for the location of a peak of the true signal given multiple realizations of random fields. These results are valid under non-stationarity and are derived using the theory of extremum estimators. Under the assumption of stationarity we improve upon these asymptotic results using a Monte Carlo approach that provides confidence regions for peaks of the mean which have better coverage in the finite sample.

A second quantity of interest when considering fields whose mean is non-zero is the height of the true signal at the location of a peak in the observed random field. These peaks are typically subject to the winner's curse, which causes inflated effect sizes at peak locations (Vul et al., 2009). We develop a resampling based procedure that obtains low bias estimates of the true signal at the location of the peak. We validate this using task data from over 8000 subjects from the UK Biobank, setting aside 4000 subjects to compute a ground truth, and dividing the remaining subjects into small samples on which to test the results.

# Chapter 1

# Introduction

### 1.0.1 Random Field Theory and its applications

Random Field Theory (hereon RFT) refers to a set of techniques developed by Robert Adler, Keith Worsley, Jonathan Taylor, Jean-Marc Azais and others (Adler (1981), Adler and Taylor (2007), Azaïs and Wschebor (2009)) which are used to analyse properties of random fields such as the expected number of maxima, volumes of level sets above a threshold and many others. Of particular interest are random fields which have a smooth spatial correlation structure because examples of these abound in practice. As such applications of RFT can be found in research areas ranging from astrophysics (Cheng and Schwartzman, 2017) to oceanography ((Longuet-Higgins, 1952),(Longuet-Higgins, 1957)) but have perhaps most significantly been used in neuroimaging (Worsley et al. (1992), Friston et al. (1994), Worsley et al. (1996)) to control false positive rates.

RFT has primarily been used, in fMRI, to perform voxelwise, clustersize and peak based inference. Voxelwise RFT proceeds by controlling the familywise error rate over voxels ($3D$ pixels) using the expected Euler characteristic heuristic (Worsley et al.,

1992). Peak inference uses the distribution for the height of a peak in a stationary random field to control false positive rates over peaks (Chumbley and Friston, 2009). Clustersize inference employs a distribution for the maximum of the size of a cluster in a zero mean stationary random field and uses it to control the familywise error rate over clusters (Friston et al., 1994). All of these methods have historically made assumptions (such as stationarity and a high level of smoothness) that are not guaranteed to hold in practice.

Recently there has been controversy in the neuroimaging community because Eklund et al. (2016) showed that a number of the standard assumptions made, when using RFT in fMRI, do not hold. In particular, they demonstrated that current implementations of RFT fail to control false positive rates when performing clustersize inference. While they raised a number of important concerns, software has lagged behind theory for a long time and many of the assumptions that are currently made can be dropped, as we will discuss. RFT has historically required stationarity, Gaussian fields and very smooth data (Adler (1981), Worsley et al. (1992), Worsley (1994), Worsley et al. (1996)). We extend the voxelwise RFT framework in Chapter 2 and show that these assumptions are no longer essential in order to perform valid voxelwise RFT inference.

The most common alternative approach used to correct for false positives in the spatial setting is based on permutation testing (Nichols and Holmes (2002b), Winkler et al. (2014), Winkler et al. (2016)). This involves resampling or flipping the sign of the observations in order to obtain an empirical distribution of the maximum, or of the size of the largest cluster above a threshold. Resampling methods in general provide a reliable approach for performing inference because they typically rely on fewer assumptions, and we will make good use of them in Chapter 5 for peak height estimation and to perform resting state validations in Chapter 2.

In this introduction we first make some key definitions and then present the model that is most commonly used to analyse fMRI data: essentially how data from scans is combined to form subject level images. We will then introduce and discuss the three different types of inference that are used in fMRI to infer on images namely voxelwise, peakwise and clustersize inference and how these are performed in practice using RFT and permutation. We will end by giving an overview of the contents of this thesis.

### 1.0.2 Random Fields and key definitions

Without further ado, let us start with several key definitions that form the backbone of this thesis.

**Definition 1.0.1.** Given $D \in \mathbb{N}$ (the set of positive integers), $S \subseteq \mathbb{R}^D$ and some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we define a $D$-dimensional **random field** on $S$ to be a measurable function

$$Y : \Omega \to \left\{ f \text{ such that } f : S \to \mathbb{R}^{D'} \right\}$$

for some $D' \in \mathbb{N}$.

Given $k \in \mathbb{N}$ we say that $Y$ is almost surely (a.s.) $k$ times differentiable if almost all sample paths of $Y$ are $k$ times differentiable. Similarly we can define notions of almost sure continuity, smoothness and $C^k$ness (note that we may drop the words almost surely when this is suitably clear). Given a twice a.s. differentiable $D$-dimensional random field $Y : S \to \mathbb{R}$ on a set $S \subset \mathbb{R}^D$, for $s \in \text{int}(S)$, (using int to denote the interior), let

$$\nabla Y(s) = \left( \frac{\partial Y(s)}{\partial s_1}, \ldots, \frac{\partial Y(s)}{\partial s_D} \right) \text{ and } \nabla^2 Y(s) = \left( \frac{\partial^2 Y(s)}{\partial s_i s_j} \right)_{1 \leq i,j \leq D}$$

and use $\nabla^T Y(s)$ to denote $(\nabla Y(s))^T$. This defines random fields: $\nabla^T Y : S \to \mathbb{R}^D$ and $\nabla^2 Y : S \to \mathbb{R}^{D \times D}$. We define the mean and the variance of $Y$ to be the functions that send $s \in S$ to $\mathbb{E}[Y(s)]$ and $\text{var}(Y(s))$ respectively. We say that $Y$ is **stationary** if, for all $s, t \in S$,

$$\text{cov}(Y(s), Y(t)) = R(s - t)$$

for some function $R : \mathbb{R}^D \to \mathbb{R}$ and **isotropic** if

$$\text{cov}(Y(s), Y(t)) = R(\|s - t\|)$$

for some function $R : \mathbb{R} \to \mathbb{R}$, where $\|.\|$ is the $L_2$ norm. We refer to random fields which are not stationary as **non-stationary** and fields which are not isotropic as **non-isotropic**. We say that $Y$ is a Gaussian random field if for all $n \in \mathbb{N}$ and all $t_1, \ldots, t_n \in S$,

$$(Y(t_1), \ldots, Y(t_n))^T$$

is a Gaussian random vector. The most well understood random fields are Gaussian random fields and fields which can be expressed as functions of Gaussian random fields.

### 1.0.3 Voxelwise modeling in fMRI

Functional magnetic resonance imaging (fMRI) is an imaging technique that involves placing a subject in a scanner and taking images of their brain using a powerful magnet that is sensitive to changes in the amounts of oxygenated and deoxygenated blood over time. This is known as the blood oxygenation level dependent or BOLD effect. In general no absolute units of changes are available for this so data is expressed in terms of %BOLD change. Individual subjects are scanned several hundred times over the course of a few minutes while performing a task. Each scan consists of a value at

every voxel of the brain (where the brain is divided into over 200,000 voxels). fMRI data is typically analysed using a hierarchical model which we partially describe in this section; see Mumford and Nichols (2006) and Mumford and Nichols (2009) for further details.

Suppose that we have $n = 1, \ldots, N$ subjects and assume that for the $n$-th subject we have $J_n$ scans and true effect magnitudes $\beta_n(v) \in \mathbb{R}^P$ and for each $j = 1, \ldots, J_n$, an observation $Z_{n,j}(v) \in \mathbb{R}$ at each voxel $v \in \mathcal{V} \subset \mathbb{R}^3$. Here $\mathcal{V}$ is our set of voxels and $P$ is the number of stimulus conditions that the subjects are observed under. We fit a **mass-univariate** model. This consists of fitting the following **first level** linear model to the time series at each voxel for each subject,

$$Z_n(v) = A_n \beta_n(v) + \epsilon_n(v) \tag{1.1}$$

where $Z_n(v) = (Z_{n,1}(v), \ldots, Z_{n,J_n}(v))^T$ and $A_n$ is a $J_n \times P$ design matrix whose $j$-th row is a measure of the expected stimulus at time $j$ over the different conditions. Each column of $A_n$ is constructed as a convolution of the experiment condition indicators and the haemodynamic response function (HRF); the HRF is the idealised response to short stimulus. $\beta_n(v) \in \mathbb{R}^P$ is a vector of coefficients and $\epsilon_n(v) \sim N(0, W(v))$ is the error, usually parametrised as a homoscedastic time series model, here $W(v)$ is the $J_n \times J_n$ temporal covariance matrix. Under the standard linear model pipeline, for each subject, given an estimate for the error covariance, $\hat{W}(v)$, we have the least squares estimate: $\hat{\beta}_n(v) = (A_n^T \hat{W}(v)^{-1} A_n)^{-1} A_n^T \hat{W}(v)^{-1} Z_n(v)$ of $\beta_n(v)$. We refer to the process of estimating $W$ and removing its effect as **whitening**.

Typically we consider a contrast $c \in \mathbb{R}^P$ of the $P$ experimental effects. For instance if $P = 2$, and we are interested in whether there is a difference between two stimulus conditions, then we would take $c = (1, -1)^T$. For each subject $n$ and each $v \in \mathcal{V}$, we

define

$$X_n(v) = c^T \hat{\beta}_n(v).$$

The $X_n$ are known as the **contrast images** or **contrasts of parameter estimates** (**COPES**).

Smoothing is important in fMRI in order to improve the signal to noise ratio and increase the power to detect signal. Some analysis pipelines smooth data before first level analyses, others smooth COPEs produced by the first level analyses (Lohmann et al., 2018). Here we smooth the COPEs, see Chapter 2 for further justification. In fMRI, images are typically spatially smoothed using a Gaussian kernel $K : \mathbb{R}^D \to \mathbb{R}$ such that for $t \in \mathbb{R}^D$,

$$K(t) = \exp(-t^T \Sigma^{-1} t/2)$$

for some positive definite matrix $\Sigma$. If $\Sigma$ is diagonal then in fMRI it is often reported in terms of $\text{FWHM}_i = \sqrt{8\Sigma_{ii} \log 2}$ for $i = 1, \ldots, D$. Applying smoothing yields a random image $Y_n : \mathcal{V} \to \mathbb{R}$ such that, for each $v \in \mathcal{V}$,

$$Y_n(v) = \sum_{v' \in \mathcal{V}} K(v - v') X_n(v'). \tag{1.2}$$

In order to account for variability over subjects we perform the **second level** regression

$$(Y_1(v), \ldots, Y_n(v))^T = A_g \beta_g(v) + \eta(v), \tag{1.3}$$

where we have $G$ groups, $A_g$ is an $N \times G$ group design matrix, $\beta_g(v) \in \mathbb{R}^G$ is a vector of group parameters, and $\eta(v) \in \mathbb{R}^N$ is the group level error. If the first and second level errors are independent and Gaussian then the error in $\beta_g$ is Gaussian. Gaussianity is a standard assumption that is made in fMRI, the validity of which is in practice highly questionable, see Chapter 2.

While the group design matrix $A_g$ can have various forms, we will focus on the one-sample model. Letting $G = 1$ and $A_g = \mathbf{1}_N$ be a vector of ones of length $N$, we can test the null hypothesis that $\beta_1(v) = 0$ with the $t$-statistic:

$$T_L(v) = \frac{\hat{\mu}_N(v)\sqrt{N}}{\hat{\sigma}_N(v)},$$

where $L$ denotes lattice, where for each $v \in \mathcal{V}$,

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

Under the **global null hypothesis**, $\mathbb{E}[Y_n(v)] = 0$ for all $n$ and all $v \in \mathcal{V}$. This can be tested using the $t$-statistic, in a number of ways, as described in following sections.

### 1.0.4 Voxelwise Inference

Once we have pre-processed our data we have a test statistic $T_L(v)$ at each voxel. We have hundreds of thousands of voxels and so if we were to reject each voxel null hypothesis at a rate $\alpha$, then even if the global null hypothesis were true, on average we would reject $100\alpha\%$ of the voxel nulls. We need to take account of this multiple testing problem.

A common solution is to instead control the probability of at least one false rejection (known as the FWER: familywise error rate) to a level $\alpha$. Voxelwise inference takes a threshold $u$ and rejects the global null hypothesis if

$$M = \max_{v \in \mathcal{V}} T_L(v) > u.$$

If $u$ is chosen appropriately (ideally it should be the $100(1-\alpha)\%$ quantile of the maximum) then the probability of a false rejection is $\alpha$ under the global null hypothesis. Unfortunately, the quantiles of the maximum are difficult to obtain. However, by

Markov's inequality:

$$\mathbb{P}(M > u) = \mathbb{P}(M_u(T_L) \geq 1) \leq \mathbb{E}[M_u(T_L)]$$

where $M_u$ is the number of maxima of $T_L$ which are greater than or equal to $u$. At high thresholds $u$ the number of maxima above the threshold is 0 or 1 with high probability and so

$$\mathbb{P}(M_u \geq 1) \approx \mathbb{E}[M_u].$$

The idea behind voxelwise RFT (which was introduced by Worsley et al. (1992) and Worsley et al. (1996)) has been that when the data is smooth enough (the **good lattice assumption**) we can approximate $T_L$ by a random field $T$ so that

$$\mathbb{E}[M_u(T_L)] \approx \mathbb{E}[M_u(T)].$$

There is no closed form for this expectation. However, it is possible to approximate it by the expected Euler characteristic of the excursion set which has a well known closed form. This is the one of the central ideas behind inference using RFT. In this section we will discuss this approximation and some of the current drawbacks of voxelwise inference as well introducing the most commonly used alternative method: permutation testing.

#### 1.0.4.1 The Euler characteristic heuristic

One of the fundamental quantities in RFT is the Euler characteristic. In order to define this properly we will need the notion of a cellular decomposition.

**Definition 1.0.2.** For $1 \leq d \leq D$, a $d$-**cell** is a continuous map $f : \mathbb{D}_d \longrightarrow A$, for some $A \subset \mathbb{R}^D$, where $\mathbb{D}_d = \left\{ x \in \mathbb{R}^{d+1} : \sum_{i=0}^{d} x_i^2 \leq 1 \right\}$ is the $d$-disc.

**Definition 1.0.3.** Given a space $A \subset \mathbb{R}^D$, a **cellular decomposition** is a finite

collection of maps $\{f_1, f_2, \dots\}$ such that each $f_i$ is an $d_i$-cell for some integer $d_i$ and

- For each $i$, $f_i : \text{int}(\mathbb{D}_{d_i}) \longrightarrow A$ is a homeomorphism onto its image which maps the boundary of the disc into the image of lower dimensional cells (except for zero cells)
- $A$ is partitioned by the interiors of the cells.

From this definition we can define the Euler characteristic.

**Definition 1.0.4.** Given a set $A \subset \mathbb{R}^D$ which admits a cellular decomposition: $\{f_1, f_2, \dots\}$, for $d = 1, \dots, D$, let $k_d$ be the number of $d$-cells in the cellular decomposition. Then the **Euler characteristic** of $A$ is defined as

$$\chi(A) = \sum_{d=0}^{D} (-1)^d k_d.$$

See Hatcher (2001) for further details. Adler (1981) shows that for any set $A \subset \mathbb{R}^D$, which admits a cellular decomposition, $\chi(A)$ is uniquely defined. For such a set, enumerating the different numbers of $d$-cells can in general be difficult. Fortunately the Euler characteristic admits relatively easy closed forms. In 1D the $\chi(A)$ is the number of connected components of $A$, in 2D it is the number of connected components minus the number of holes and in 3D it is the number of connected components minus the number of holes plus the number of hollows (Adler (1981), Worsley (1996a)).

**Definition 1.0.5.** Given bounded $S \subset \mathbb{R}^D$ and $f : \mathbb{R}^D \to \mathbb{R}$ and a threshold $u \in \mathbb{R}$, we define the **excursion set**

$$\mathcal{A}_u(f) = \{t \in S : f(t) \geq u\}.$$

It can be shown (see Adler (1981)) that given a function $f$ satisfying suitable regularity conditions, for every $u \in \mathbb{R}$, the excursion set admits a cellular decomposition,

meaning that its Euler characteristic is well-defined. We are now in a position to observe that at high thresholds $u$, the number of components of the excursion set is zero or one with high probability. As a result, at these thresholds,

$$\mathbb{E}[M_u(T)] \approx \mathbb{E}[\chi(\mathcal{A}_u(T))].$$

See Figure 1.1 for an illustration of this in practice. In particular, Taylor (2005) proved the following asymptotic bound on the difference between the probability of an excursion and the expected Euler characteristic of the excursion set of a unit-variance Gaussian random field $Y$ :

$$\liminf_{u\to\infty} -u^{-2}\log|\mathbb{P}(M_u(Y)\geq 1) - \mathbb{E}[\chi(\mathcal{A}_u(Y))]| \geq \frac{1}{2}\left(1+\frac{1}{\sigma_c^2(Y)}\right),$$

where $\sigma_c^2(Y)$ is a fixed positive constant that depends on the covariance structure of $Y$.

For general test-statistics $T$ exact expressions for $\mathbb{E}[M_u(T)]$ exist (by applying the Kac-Rice formula, see Adler and Taylor (2007) Chapter 11) but they involve a very difficult integral. However, $\mathbb{E}[\chi(\mathcal{A}_u(T))]$ has a well studied closed form due to Taylor (2006), Adler and Taylor (2007) and so is much easier to evaluate and can be used to control the voxelwise FWER.

One of the drawbacks of current implementations of standard voxelwise RFT is that it makes the assumption that the data (observed on a discrete lattice) can be approximated by a smooth random field. We show in Chapter 2 that the failure of this assumption causes inference to be conservative, and remove this problem by extending the voxelwise RFT framework to work under arbitrary applied smoothness.

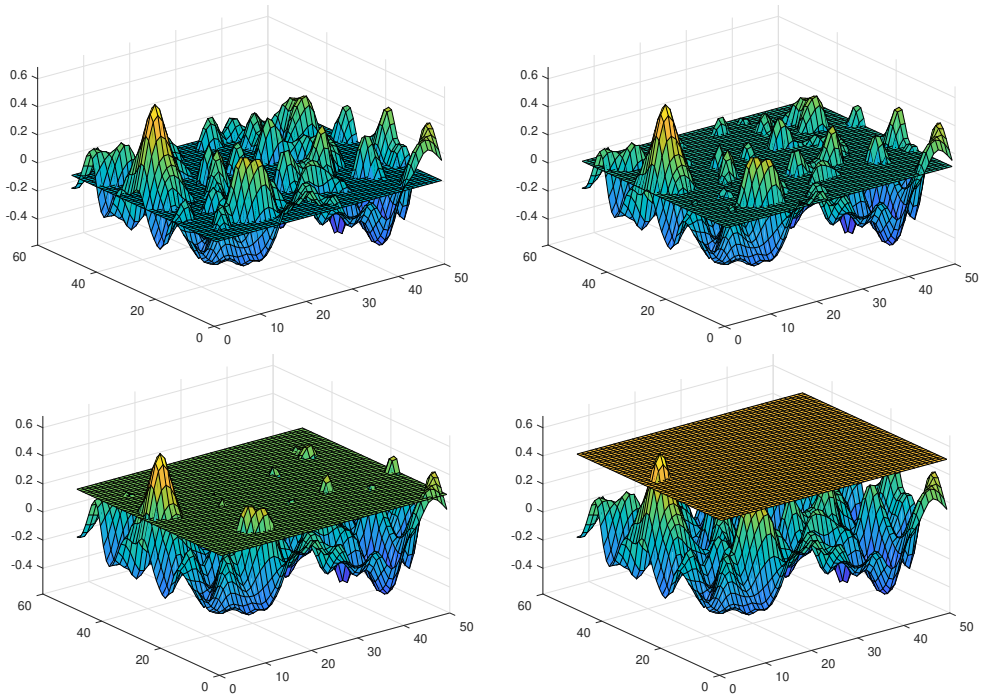


Figure 1.1: Examining the Euler characteristic of the excursion set: $\chi(\mathcal{A}_u)$ of a random field. Here we have generated a realisation of a 2D Gaussian random field and have considered a set of increasing thresholds. At high thresholds $\chi(\mathcal{A}_u)$ equals the number of clusters above the threshold and at very high thresholds this is either 0 or 1 with high probability.

#### 1.0.4.2 Voxelwise Permutation Testing

The good lattice assumption is not a problem for resampling based techniques. The most common of these used in fMRI is known as permutation testing and is described in Algorithm 1 (in the one-sample setting).

The only assumption required is that the distribution of the images is symmetric. This assumption is in fact problematic if the first level design is not randomized, (Eklund et al., 2019). Without randomizing, neither one-sample permutation nor RFT correctly control for false positives as the data is not symmetric. Algorithm 1 describes one-sample permutation inference; it is also possible to perform permutation inference for the linear model, see Winkler et al. (2014) for further details.

The drawback of permutation based methods is that they can be very slow (since

**Algorithm 1** One Sample Permutation - Voxelwise

1: **Input**: Images $Y_1, \ldots, Y_N$ on a set of voxels $\mathcal{V}$, the number of permutations $P$ and desired significance level $\alpha$
2: Let $T_L = T_L(Y_1, \ldots, Y_N)$ be the one-sample $t$-statistic image.
3: **for** $p = 1, \ldots, P$ **do**
4: Generate a vector $B$ of length $N$ such that the entries are independent Bern(0.5) random variables.
5: For $n = 1, \ldots, N$, let $Y_n^* = (-1)^{B(n)} Y_n$
6: Let $T_p = T_L(Y_1^*, \ldots, Y_N^*)$ and let $m_p = \max_{v \in \mathcal{V}} T_p(v)$
7: **end for**
8: Let $k$ be the upper $\alpha$ quantile of the empirical distribution given by $\{m_1, \ldots, m_P\}$.
9: Reject $v$ such that $T_L(v) > k$.

in practice it is common to use $P = 5000$ and the images are typically 3D). In contrast RFT methods are significantly faster. In the era of big data the sample sizes available are very large and the cost imposed by permutation can be prohibitive. The speed up provided by parametric methods (so long as they can be shown to be valid) has the potential to allow for inference to be performed even in these settings.

### 1.0.5 Peak Inference

The second type of inference that is used in spatial signal detection is peak inference. This was introduced into the literature by Chumbley and Friston (2009) and Schwartzman et al. (2011). Chumbley and Friston (2009) used the fact that the height above a threshold, in a stationary Gaussian random field, is asymptotically (as the threshold goes to infinity) exponential to obtain a $p$-value at each peak (Nosko (1969),Adler (1981)). Schwartzman et al. (2011) used the exact distribution for the peak height in a 1D stationary random field. Recent work, Cheng and Schwartzman (2015a), derived the distribution for the height of a peak in a non-stationarity Gaussian random field. In particular they proved the following theorem.

**Theorem 1.0.6.** *Let $Y$ be a $C^3$ Gaussian random field on a bounded set $S \subset \mathbb{R}^D$*

*satisfying mild regularity assumptions. Then, for each $t_0 \in int(S)$, given a CDT $u$ and $w > u$,*

$$\mathbb{P}_{hw}(Y(t_0) > u + w \,|\, t_0 \text{ is a local maximum of } Y \text{ and } Y(t_0) > u) = \frac{\mathbb{E}[|\det \nabla^2 Y(t_0)| 1[Y(t_0) > u + w, \nabla^2 Y(t_0) \prec 0] | Y(t_0) = 0]}{\mathbb{E}[|\det \nabla^2 Y(t_0)| 1[Y(t_0) > u, \nabla^2 Y(t_0) \prec 0] | Y(t_0) = 0]}.$$

Here, given a $D \times D$ matrix $M$, $M \prec 0$ denotes the fact that $M$ is negative definite. ($\mathbb{P}_{hw}$ denotes the horizontal window probability conditional on there being a maximum at $t_0$. We introduce a rigorous definition for this in Chapter 4. Note that we cannot use standard conditional probability distributions here because the event to be conditioned on has probability 0.) This formula is valid for non-stationary and non-mean zero random fields. In practice, however, it can be difficult compute. Cheng and Schwartzman (2015a) use it to prove the following result.

**Corollary 1.0.7.** *Let $Y$ be a mean-zero unit variance $D$-dimensional Gaussian random field on a bounded set $S \subset \mathbb{R}^D$, satisfying mild regularity conditions. Then for each $t_0 \in int(S)$ and each fixed $w > 0$, as $u \to \infty$,*

$$\mathbb{P}_{hw}(Y(t_0) > u + w \,|\, t_0 \text{ is a local maxima of } Y \text{ and } Y(t_0) > u) = \frac{(u+w)^{D-1} e^{-(u+w)^2/2}}{u^{D-1} e^{-u^2/2}} (1 + O(u^{-2})).$$

We show that this result applies even when $w$ is a function of $u$ and use this to prove that the scaled peak height is asymptotically exponentially distributed (see Chapter 4), extending the results of Nosko (1969), Belyaev (1967) and Belyayev et al. (1972) to non-stationarity. This shows that the stationary height distribution used in Chumbley and Friston (2009) is valid for non-stationary random fields so long as the threshold $u$ is taken to be high enough. Given $p$-values for each peak, Chumbley and Friston (2009) control the false discovery rate over peaks using the Benjamini-Hochberg

procedure (Benjamini and Hochberg, 1995). In practice this works reasonably well, however, because the number of peaks that lie above the threshold is random, the false discovery rate is not well defined and Benjamini-Hochberg is not guaranteed to provide meaningful inference. Schwartzman et al. (2011) prove results that provide a more through justification for peak based inference (under stationarity) in 1D and these were extended to arbitrary dimensions in Cheng and Schwartzman (2017). This has been implemented in the context of neuroimaging in Schwartzman and Telschow (2019), though has not yet been rigorously validated using resting state data in an Eklund et al. (2016) type analysis such as the one that we use in Chapter 2.

When conducting peak and/or voxelwise inference in practice, it is typical for papers to report the effect sizes and locations of the largest peaks of the test-statistic. However this incurs a selection bias (Vul et al., 2009) and results in over estimation of the true peak effect size. We present a resampling method to correct for this in Chapter 5. Moreover, simply reporting a point estimate for the peak location is bad practice, instead the estimate should be reported along with a confidence region for the location of the peak activation. In Chapter 3 we develop asymptotically valid confidence regions for the location of peaks in the mean and Cohen's $d$.

### 1.0.6 Clusterwise inference

Clusterwise inference is a widely used technique, for assigning statistical significance in fMRI images, because it generally has greater power to detect activation than other methods (such as voxelwise and peak inference). fMRI activation is often spatially extended and so if one voxel is observed to be active then it is likely that neighbouring voxels are too. As such, techniques that are aimed at detecting regions of activation,

rather than activation at individual voxels, are typically more powerful.

**Definition 1.0.8.** Given a random field $T$ on $S \subset \mathbb{R}^D$ and a threshold $u \in \mathbb{R}$, the clusters of $T$ above $u$ are defined to be the connected components of the excursion set $\mathcal{A}_u(T)$.

In the context of clustersize inference $u$ is referred to as the **cluster defining threshold (CDT)** and $T$ is taken to be the test-statistic map. The distribution of the size of the largest cluster (under the global null hypothesis) is estimated (using RFT or permutation testing) and given a level $\alpha$, clusters that are larger than the $100\alpha\%$ threshold of the clustersize distribution are rejected. This controls the FWER over clusters to $\alpha$.

The clustersize RFT approach combines parametric marginal distributions for the size of each cluster to obtain the distribution for the size of the largest cluster. Nosko (1969), Wilson and Adler (1982) and Wilson (1988) proved the following theorem (see Adler et al. (2010) for a great overview of this theory) regarding the marginal clustersize distribution.

**Theorem 1.0.9.** *Let $Y$ be a mean-zero unit-variance stationary $D$-dimensional Gaussian random field satisfying certain regularity conditions. For $u \in \mathbb{R}$ and $t_0 \in S$, let $c_u$ denote the Lebesgue measure of the component of $\mathcal{A}_u(Y)$ that contains $t_0$. Then for $x \geq 0$,*

$$\lim_{u\to\infty} \mathbb{P}_{hw}\big(u^D 2^{-D/2}(\omega_D)^{-1}\det(\Lambda)^{1/2} c_u \geq x \big| t_0 \text{ is a local maxima of } Y \text{ and } Y(t_0) > u\big) = e^{-x^{2/D}},$$

*where $\Lambda$ is the covariance matrix of the partial derivatives and $\omega_D$ is the volume of the unit ball in $\mathbb{R}^D$.*

We generalize this theorem to non-stationary Gaussian random fields in Chapter

4 (where we also formalise the notion of HW distribution), taking advantage of the results of Cheng and Schwartzman (2015a) discussed in the previous section.

Given this result, and using the fact that the number of clusters above the threshold in a stationary random field has a Poisson distribution at high thresholds, (Adler (1981) Chapter 6.9, Aldous (2013)), Friston et al. (1994) developed a parametric method to control clusterwise false positives. To understand their framework, suppose that $m$ clusters are observed above the CDT $u$ with sizes $n_1, \ldots, n_m$. Then under certain assumptions, that do not exactly hold in practice but are reasonable approximations at high thresholds, we have the following approximate result (Friston et al., 1994).

**Theorem 1.0.10.** *Let $Y$ be a mean-zero variance-one stationary $D$-dimensional Gaussian random field on a bounded set $S \subset \mathbb{R}^D$ and let $u > 0$ be the CDT. Let $m$ be the number of clusters of $Y$ above $u$ with i.i.d[1] sizes $c_1, \ldots, c_m$ that are independent of $m$. Let $c_{\max}$ be the largest cluster size and let $c$ denote a draw from the i.i.d cluster size distribution, then for $k \in \mathbb{N} \cup \{0\}$,*

$$\mathbb{P}(c_{\max} \geq k) \approx 1 - e^{-\mathbb{E}m\mathbb{P}(c \geq k)}.$$

Traditional RFT clustersize inference thus depends on a number of assumptions. Ones of particular note include stationarity, the Poisson distribution of the number of clusters above the threshold and the independence of the size and the number of clusters. Using a resting state validation (the nature of which is discussed in detail in Chapter 2) Eklund et al. (2016) showed that many of these assumptions do not hold in practice meaning that clustersize inference, as currently implemented, leads to inflated false positive rates.

[1]Here i.i.d stands for independent and identically distributed.

#### 1.0.6.1 Permutation Clustersize Inference

If we assume that our images are symmetric, under the null hypothesis, then we can use permutation clustersize inference. In the one-sample case this proceeds as described in Algorithm 2.

**Algorithm 2** One Sample Permutation - Clusterwise

1: **Input**: Images $Y_1, \ldots, Y_N$ on a set of voxels $\mathcal{V}$, the number of permutations $P$ and desired significance level $\alpha$ and CDT $u$
2: Let $T_L = T_L(Y_1, \ldots, Y_N)$ be the test-statistic image and let $c_1, \ldots, c_m$ be the sizes of the clusters of $T_L$ above the threshold $u$.
3: **for** $p = 1, \ldots, P$ **do**
4: Generate a vector $B$ of length $N$ such that the entries are independent Bern(0.5) random variables.
5: For $n = 1, \ldots, N$, let $Y_n^* = (-1)^{B(n)} Y_n$
6: Let $T_p = T_L(Y_1^*, \ldots, Y_N^*)$ and let $c_p^*$ be the size of the largest cluster of $T_p$ above the threshold $u$.
7: **end for**
8: Let $k$ be the upper $\alpha\%$ quantile of the empirical distribution given by $\{c_1^*, \ldots, c_P^*\}$.
9: Reject the clusters $i$ such that $c_i > k$.

This can be extended to more general test-statistics see Winkler et al. (2014). While this framework makes very few assumptions, as with voxelwise permutation, it incurs a high computational cost (one that is slightly larger than for voxelwise inference since the data cannot be vectorized to speed up inference). As such it is highly desirable to develop a parametric framework that works in practice and correctly controls cluster-wise false positives. Unfortunately doing so is beyond the scope of this thesis, however, we are optimistic that combining the results of Chapters 2 and 4 could lead to such a framework at some point in the future.

### 1.0.7 Thesis Overview

This thesis is laid out as follows. In Chapter 2 we introduce our non-stationary voxel-wise RFT inference framework: dropping assumptions of smoothness, stationarity and

(to a large extent) Gaussianity that have historically been made by traditional RFT methods.

In Chapter 3 we develop peak confidence regions for the locations of peaks of the signal (measured in terms of the mean and standardized effect size) in a noisy random field. To do so we use results from the theory of extremum estimation. We additionally show that, by assuming stationarity, we can obtain confidence regions for peaks of the mean that have better finite sample coverage than their asymptotic counterparts.

In Chapter 4 we prove results about the distribution of the size of a cluster above a threshold in a non-stationary Gaussian random field. We extend results, originally due to Nosko (1969) and Wilson (1988) so that they apply under non-stationarity.

In Chapter 5 we introduce a selective framework that uses bootstrap resampling to estimate the true height of the signal at an observed peak in a test-statistic image: countering the inflation that results from the winner's curse (Vul et al., 2009).

Finally, in Chapter 6 we provide some concluding remarks. This thesis is being submitted as an integrated thesis under the new Oxford submission rules, as such each chapter in the main-text is self-contained. As a result there is a very small amount of repetition of definitions across chapters and notation does not always align between chapters. Additionally, all section referencing is entirely within each chapter unless otherwise specified.

In this thesis we hope to show that parametric methods have an important role to play when it comes to analysing and understanding random imaging datasets. We hope to demonstrate that both RFT and resampling methods can be used effectively to perform inference and analysis and that these approaches have a number of interesting applications in neuroimaging.

# Chapter 2

# Accurate voxelwise FWER control in fMRI using Random Field Theory

Samuel Davenport, Fabian Telschow, Armin Schwartzman, Thomas E. Nichols

**Abstract**

In this work we introduce modifications to the standard voxelwise Random Field Theory (RFT) inference framework that dramatically improve its performance in finite samples. Historically applications of RFT in fMRI have relied on assumptions of smoothness, stationarity and Gaussianity. We address these three limitations as follows. Firstly we define convolution fields, enabling RFT inference to work under arbitrary applied smoothness. Secondly, we use the Gaussian Kinematic Formula to estimate the expected Euler characteristic (EEC) under non-stationarity. Thirdly, we show that transforming the data can improve power and allows the EEC to be accurately estimated when the data is non-Gaussian, given a reasonable number of subjects. This allows us to drop the first two assumptions and reduces the impact of the third. These improvements enable us to provide a quick and powerful method that correctly controls the voxelwise false positive rate in fMRI. We employ a big data validation in which we subsample resting state data from 7000 subjects from the UK Biobank to demonstrate that the error rate is correctly controlled.

*Keywords:* Random Field Theory, FWER control, multiple testing, voxelwise inference, spatial statistics, non-stationarity, Gaussianization.

# 1 Introduction

Random Field Theory (RFT) encompasses an advanced set of mathematical techniques for analysing imaging data that has been widely applied in neuroimaging in order to control false positives via cluster, peak and voxel level inference (Worsley et al. (1992), Worsley et al. (1996), Friston et al. (1994), Chumbley and Friston (2009))[1]. RFT has traditionally required that the data is Gaussian, stationary and sufficiently smooth, however as we shall show that none of these assumptions are particularly reasonable in fMRI. Eklund et al. (2016) demonstrated that RFT inference can fail to control false positive rates as a result of the failure of these assumptions. It is thus highly desirable to extend RFT to work without these requirements. There has already been considerable work on extending RFT to work under non-stationarity (see Taylor (2006), Adler and Taylor (2007), Telschow et al. (2019), (Telschow et al., 2020b), Cheng and Schwartzman (2015a)), but it has never been implemented in neuroimaging toolboxes. Here we provide an accurate and fast voxelwise RFT framework that does not require stationarity nor smoothness and no longer relies on the data being Gaussian.

Traditional RFT[2] makes the *good lattice assumption* that observations on the lattice corresponding to the brain have the same properties as a continuous random field (Worsley et al., 1996). This is required in order to provide good estimates of the smoothness of the data (Kiebel et al., 1999) and to correctly infer on the distribution of the maximum. Failure of this assumption leads to voxelwise inference being conser-

[1]In voxelwise inference voxels with test-statistic values lying above a multiple testing threshold are determined to be significant. In both cluster and peak level inference a cluster defining threshold is used to identify clusters/peaks and then thresholding based on cluster extent and peak magnitude is used to determine significant clusters/peaks.

[2]For this article we refer to the body of work of Worsley et al, before the improvements made by Taylor (2006) and Taylor and Worsley (2007b) as **Traditional RFT**. This refers to the RFT inference framework that was built in Worsley et al. (1992), Friston et al. (1994) and Worsley et al. (1996) which assumed stationarity, a high level of smoothness and Gaussianity of the underlying fields.

vative (controlling the false positive rate well below the nominal level, see Eklund et al. (2016)) because the maximum of a continuous random field is always higher than the maximum of the field evaluated on a lattice subset (Worsley (2005)). When analysing fMRI data, in order to improve power, brain images are smoothed before performing inference. In order to adequately satisfy the good lattice assumption, traditional RFT requires a very high level of smoothing, much higher than the levels typically applied in fMRI. Worsley (2005) and Taylor and Worsley (2007a) attempted to solve this problem using discrete local maxima, however their approach is only valid under stationarity and under the assumption that the covariance function is squared exponential. We show that RFT can be modified so that it attains nominal false positive rates even in the low smoothness setting, resulting in more powerful inference on non-null data that is valid under non-stationarity. To do so we observe that in smoothing the data and evaluating on a lattice subset (as is the norm in current software) much information is lost. Smoothing is typically done using a Gaussian kernel which is infinitely smooth and so applying smoothing actually yields a smooth random field (evaluable at every point not just on a lattice subset), known as a convolution field (Telschow et al. (2020b)). Using convolution fields will allow us to drop the smoothness assumption.

RFT has historically required stationarity (Adler (1981), Worsley et al. (1992), Worsley (1994), Worsley et al. (1996)). This allows for estimation of the quantiles of the maximum of a random field, via the expected Euler characteristic (EEC) (Worsley et al., 1992), and is used to control the familywise error rate (FWER). Revolutionary work (Taylor (2006), Taylor and Worsley (2007b)) extended RFT to allow for computation of the EEC under non-stationarity. We have implemented this here in 2D and 3D for use in brain imaging, allowing us to drop the stationarity assumption. Combining this with convolution fields we will show that voxelwise RFT accurately controls the

false positive rate.

RFT requires that the data is Gaussian or that the test-statistic is Gaussian. Given a large enough number of subjects this holds by the Central Limit Theorem (CLT), however the sample sizes typically used in fMRI mean that it can often be an unreasonable assumption. We will show that second level fMRI data has heavy tails meaning that the CLT takes longer to converge and that in practice hundreds of subjects are required before the test-statistic is sufficiently Gaussian. Others have also noted the lack of Gaussianity in fMRI data and have attempted to correct for it. In particular Wager et al. (2005), Woolrich (2008) and Fritsch et al. (2015) considered robust regression models in order to account for non-Gaussianity and Chen et al. (2012), Roche et al. (2007) developed methods for dealing with outliers. So far no work has shown that non-Gaussianity can be compatible with RFT. To address this problem we transform the data to improve levels of Gaussianity under the null hypothesis via a process called Gaussianization which only requires that the distribution of the data is symmetric. We show that, given sufficiently many subjects, this improves the validity of RFT theory and allows us to control the FWER error even when the data is non-Gaussian.

The structure of this paper is as follows. Section 2.1 describes traditional RFT on a lattice and discusses the failure of the good lattice assumption. Section 2.2 introduces the notion of convolution fields and Section 2.3 discusses how to calculate the Euler characteristic under non-stationarity. In Section 2.4 we discuss how to estimate the Lipshitz-Killing curvatures (LKCs) and the smoothness of a convolution random field. We then discuss how we use a big data validation (as in Eklund et al. (2016), Davenport and Nichols (2020)), using resting state data from 7000 subjects from the UK Biobank, to show that the methods perform as desired. Finally we introduce the idea of Gaussianization in Section 2.6.

We discuss our results in Section 3. We demonstrate the gains that can be made by using convolution fields and by Gaussianizing the data. We show that our RFT framework accurately controls the FWER and is more powerful than existing parametric approaches.

Software to implement the methods (including random field generation, estimation of smoothness and LKCs, calculating coverage rates, performing Gaussianization and other RFT analysis) is available in the RFTtoolbox package (`https://github.com/sjdavenport/RFTtoolbox`). Code to reproduce all of the figures is available at `https://github.com/sjdavenport/RFTtoolbox/ConvolutionNeuroPaper`.

# 2 Methods

In this section we will outline the traditional voxelwise RFT framework established by Worsley et al. (1992) and Worsley et al. (1996) and show how it can be extended to work under arbitrary smoothness, non-stationarity and when the underlying data is non-Gaussian.

In what follows we will let $\mathcal{V}$ be a finite subset of $\mathbb{R}^D$ that corresponds to the set of voxels that make up the brain. We will assume that $\mathcal{V}$ is equally spaced in the $d$th direction with spacing $h_d$ for $d = 1, \ldots, D$. We consider group level analyses, in which we have a number of subjects $N$ such that for $n = 1, \ldots, N$ the $n$th subject has a corresponding 3D random image $X_n$, taking values on the lattice $\mathcal{V}$. In neuroimaging $X_n$ is the $c^T\hat{\beta}$ COPE image obtained from running pre-processing and first level analyses for each subject (Mumford and Nichols, 2009) as defined in the introduction to the thesis. These images are typically combined into a test-statistic $T_L$ which must be thresholded in order to identify areas of activation. (Here we have used the subscript $L$

to denote the fact that the test-statistic is only computed on the voxel lattice, in order to distinguish it from the continuous test-statistic field which we subsequently define.)

## 2.1 Traditional voxelwise RFT Inference

Traditional voxelwise RFT takes advantage of the fact that, for $u \in \mathbb{R}$ by Markov's inequality,

$$\mathbb{P}\left(\max_{v \in \mathcal{V}} T_L(v) > u\right) = \mathbb{P}(M_u(T_L) \geq 1) \leq \mathbb{E}[M_u(T_L)]$$

where $M_u(T_L)$ is the number of maxima of $T_L$ which are greater than or equal to $u$. If $u$ is chosen such that the expectation equals $\alpha$ then the FWER will be less than or equal to $\alpha$. At high thresholds $u$ the number of maxima above the threshold is 0 or 1 with high probability and so

$$\mathbb{P}(M_u(T_L) \geq 1) \approx \mathbb{E}[M_u(T_L)].$$

However, even at low thresholds the expected number of maxima still serves as an upper bound. Traditional RFT inference assumes that the data is mean-zero and smooth enough so that we can approximate the lattice based $T_L$ by a mean-zero smooth random field $T$, such that

$$\mathbb{E}[M_u(T_L)] \approx \mathbb{E}[M_u(T)],$$

where $M_u(T)$ is the number of maxima of the continuous field $T$ above $u$. This expectation is not easy to calculate, however it is possible to closely approximate it by the expected Euler characteristic of the excursion set which has a surprisingly simple closed form known as the Gaussian Kinematic Formula (Taylor, 2006), see Section 2.3.

### 2.1.1 Failure of the good lattice assumption

When performing traditional RFT inference, under the global null hypothesis, it is assumed that $T_L$ extends to a mean-zero random field $T$, on a set $S \supset \mathcal{V}$, such that

$$\sup_{v \in \mathcal{V}} T_L(v) \approx \sup_{s \in S} T(s),$$

an assumption known as the **good lattice assumption**. Given any continuous field $T$ defined on $S$ and such that $T(v) = T_L(v)$ for all $v \in \mathcal{V}$,

$$\sup_{v \in \mathcal{V}} T_L(v) \leq \sup_{s \in S} T(s)$$

and equality is usually not achieved. With high enough smoothing this is not a problem because the maximum on the lattice is very close to the maximum of the continuous field. However, at smoothing levels typically used in fMRI, the good lattice assumption does not hold. This causes conservativeness because the threshold needed to control excursions of the continuous process is higher than the threshold needed to control excursions of the process evaluated on the lattice. Application of voxelwise RFT to resting state data confirms this, see Section 3.4.1 and Eklund et al. (2016) Figure 1.

Application of traditional RFT also relies on estimation of the smoothness. Current implementations (e.g. SPM, FSL) are based on Kiebel et al. (1999) and so provide biased estimates, at the typical smoothness levels used in fMRI, because they use discrete derivatives (see Section 3.1). This is plainly evident in Kiebel et al. (1999) Figure 3 but is rarely commented on.

The good lattice assumption is thus a problem for the use of RFT in neuroimaging. However this assumption can be dropped: to do so we introduce the notion of a convolution field which looks at the data continuously rather than restricted to a discrete lattice. This allows us to obtain good estimates of the smoothness as well as correct

false positive rates, without making any assumptions other than that the data (or the test-statistic) is Gaussian and that the kernel that is used for smoothing is three times continuously differentiable. (Note that the assumption that the test-statistic is Gaussian is ensured given enough subjects by the CLT.) We will see that approximating the lattice test-statistic using $T$ is the wrong approach and that in fact one should work in the continuous domain in order to correctly control false positive rates.

## 2.2 Convolution Fields

Typically in fMRI the smoothed images are only evaluated on the lattice, however the original data and the smoothing kernel $K$ together contain unused information. To take advantage of this (as in Telschow et al. (2020b)), for each $n = 1, \dots, N$, given a kernel $K$, a $D$-dimensional lattice $\mathcal{V}$ and a bounded set $S \subset \mathbb{R}^D$ we define the **convolution field** to be $Y_n : S \to \mathbb{R}$ such that

$$Y_n(s) = \sum_{v \in \mathcal{V}} K(s - v) X_n(v) \tag{2.1}$$

for $s \in S$, with pointwise variance $\sigma^2(s) = \mathrm{var}(Y_1(s))$. The convolution field inherits smoothness and Lipschitz properties from the kernel $K$. In fMRI we typically take $K$ to be a Gaussian kernel implying that the resulting convolution fields are infinitely smooth. For each $s \in S$, define the sample mean $\hat{\mu}_N(s) = \frac{1}{N} \sum_{n=1}^{N} Y_n(s)$ and sample variance

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

In order to perform one-sample inference we define $T : S \to \mathbb{R}$ to be the **convolution t-field** sending $s \in S$ to

$$\frac{\sqrt{N} \hat{\mu}_N(s)}{\hat{\sigma}_N(s)}.$$

We take $S$ to be the continuous domain composed of the individual voxels that make up the brain i.e. for $v = (v_1, \ldots, v_D)^T \in \mathcal{V}$, letting

$$S(v) = \left\{ (s_1, \ldots, s_D)^T \in \mathbb{R}^D : \max_{d \in 1,\ldots,D} |s_d - v_d| < \frac{h_d}{2} \right\},$$

$$S = \bigcup_{v \in \mathcal{V}} S(v).$$

We call this canonical choice of $S$ the **voxel domain** and will use this as our domain throughout the remainder of the paper, even though the theory is valid for any bounded set $S \subset \mathbb{R}^D$. The difference between the convolution and lattice approaches is clearly shown in Figure 2.1. The maximum of the convolution $t$-field can be found using optimization algorithms and compared to the threshold obtained using RFT (see Section 2.3) in order to perform inference. This works by finding the local maxima of the convolution $t$-field on the original lattice and using these locations to initialize the optimization. As a result inference is possible at arbitrarily high resolution without causing memory problems.

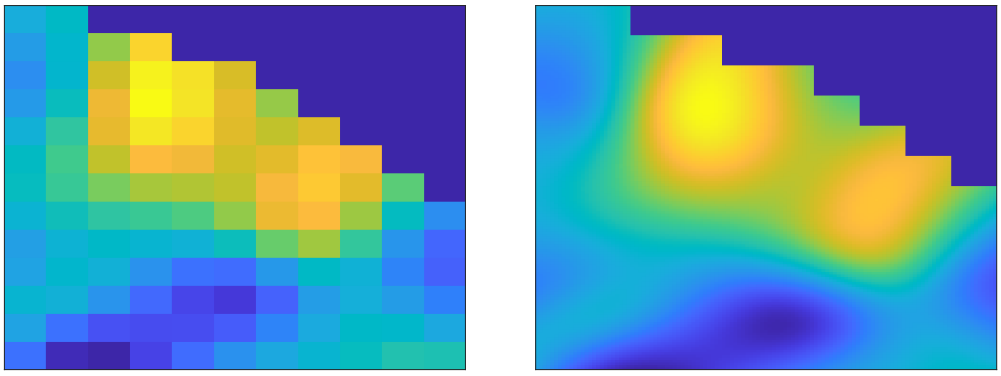


Figure 2.1: Left: A 2D section of the $t$-statistic of 50 resting state fMRI COPE images smoothed with FWHM = 3 voxels on a lattice. These images are pre-processed as discussed in Section 2.5. Right: the corresponding convolution field of the same section of the brain (evaluated on a grid corresponding to resolution $r = 11$, see Section 2.4). The value at the centre of the voxel of the convolution field is the same as the value of that voxel on the original lattice. Points in dark blue in the upper right of the images are points that lie outside of the mask.

In fMRI it is common to smooth the data before performing whitening. With the convolution approach we instead whiten the first level data, compute the first level COPEs, $X_n$, and smooth these. This is not essential: we could instead smooth before whitening and then apply a small amount of additional smoothing in order to obtain a convolution field. If done in this second way the additional smoothing can be relatively small and so will not affect inference. At high smoothness levels the extra smoothing and convolution fields themselves are unnecessary as the good lattice assumption holds. However at the smoothness levels typically applied in fMRI, using convolution fields prevents conservativeness, and can lead to large power increases, see Section 3.4.1.

## 2.3 Obtaining the voxelwise RFT threshold

Computing the EEC is key to controlling the FWER using RFT because for high thresholds $u$ it can be used to closely approximate the expected number of maxima as

$$\mathbb{E}[M_u(T)] \approx \mathbb{E}[\chi(\mathcal{A}_u(T))] \tag{2.2}$$

where $\chi$ is the Euler characteristic of the **excursion set**:

$$\mathcal{A}_u(T) = \{t \in S : T(t) \geq u\}.$$

The approximation (2.2) holds because at high enough thresholds the number of maxima and the Euler characteristic of the excursion set are the same (Worsley et al. (1992), Taylor (2005)). Traditional RFT obtains a closed form for the EEC by assuming stationarity based on the results of Adler (1981) and Worsley (1994). However this is outdated as Taylor (2006) extended these results to non-stationary random fields. His result, known as the **Gaussian Kinematic Formula (GKF)**, is as follows.

**Theorem 2.1.** *(Taylor (2006) Theorem 4.3) Let $Y_1, \ldots, Y_N$ be i.i.d unit-variance*

*Gaussian random fields and let* $F : \mathbb{R}^N \to \mathbb{R}$. *Let* $T$ *be a random field such that* $T(s) = F(Y_1(s), \ldots, Y_N(s))$ *for all* $s \in S$. *Under certain regularity conditions, for all* $u \in \mathbb{R}$,

$$\mathbb{E}[\chi(\mathcal{A}_u(T))] = \sum_{d=0}^{D} \mathcal{L}_d \rho_F(u)$$

*where* $\mathcal{L}_0, \ldots, \mathcal{L}_D$ *are constants and* $\rho_F : \mathbb{R} \to \mathbb{R}$ *is a fixed function that depends on* $F$.

The values $\mathcal{L}_0, \ldots, \mathcal{L}_D$ are known as the **Lipshitz Killing curvatures (LKCs)** and can be estimated using the component random fields: $Y_1, \ldots, Y_N$. The $\rho_F$ are called the **EC densities** and can be computed exactly (Adler et al., 2010). Taking $F : \mathbb{R}^N \to \mathbb{R}$ such that for each $(a_1, \ldots, a_N) \in \mathbb{R}^N$

$$F(a_1, \ldots, a_N) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} a_n \left( \frac{1}{N-1} \sum_{n=1}^{N} \left( a_n - \frac{1}{N} \sum_{n=1}^{N} a_n \right)^2 \right)^{-1/2}$$

and applying the theorem, gives us results about one-sample $t$-fields as

$$T(s) = \frac{\sqrt{N}\hat{\mu}_N(s)}{\hat{\sigma}_N(s)} = F\left( \frac{Y_1(s)}{\sigma(s)}, \ldots, \frac{Y_N(s)}{\sigma(s)} \right).$$

In fMRI the component fields do not have constant variance. However, we have taken advantage of the fact that the $t$-statistic is scale invariant and have divided by $\sigma$ to ensure that the unit-variance condition holds. Taking alternative, appropriate, forms of $F$ gives results about two-sample $t$-fields and $F$-fields (the ones commonly used in neuroimaging) and many others. As such this theorem is strong and widely applicable.

The regularity conditions required by the theorem are relatively mild and can be shown to hold for convolution fields. To do so we need to verify conditions $(G1), (G2)$ and $(G3)$ from Telschow et al. (2019) for the field $Y_1/\sigma$. Since $Y_1/\sigma$ is variance 1, (G1) amounts to almost sure twice differentiability of the field which follows so long as the smoothing kernel is twice differentiable, the lattice is finite and the lattice random

variables $(X_n(v) : v \in \mathcal{V})$ have finite-variance. (G2) is non-degeneracy between the field and its derivatives and holds so long as the lattice random variables are Gaussian and non-degenerate. This holds under relatively mild conditions (see Chapter 4). (G3) can be shown to hold so long as the smoothing kernel is $C^3$ by proving square Lipshitz continuity and using the fact that $|x| \leq |\log(x)|^{-(1+\gamma)}$ for any $\gamma > 0$ and small enough $x$.

Current software packages (e.g. SPM, FSL) still use the stationary version of Theorem 2.1 despite the fact that the noise in fMRI is non-stationary (Eklund et al., 2016). Part of the reason for this is that, under non-stationarity, the LKCs can be difficult to compute. The top two LKCs, however, have well-known closed forms ( Taylor and Worsley (2007a), Adler et al. (2010)):

$$\mathcal{L}_D = \int_S \det(\Lambda(t))^{1/2}\, dt \tag{2.3}$$

and

$$\mathcal{L}_{D-1} = \int_{\partial S} (\det \Lambda_{\partial S}(t))^{1/2} \mathcal{H}_{D-1} \tag{2.4}$$

where $\mathcal{H}_{D-1}$ is the Hausdorff measure on $\partial S$, the boundary of the support, and

$$\Lambda = \mathrm{cov}\left(\nabla \frac{Y_1}{\sigma}\right)$$

is the covariance matrix of the partial derivatives (here we are assuming the global null so that $Y_1$ has mean zero). At each $t \in \partial S$, let $e_1(t), \ldots, e_{D-1}(t)$ be an orthonormal basis to the tangent space to $S$ at $t$, then $\Lambda_{\partial S}$ is defined to be the $(D-1) \times (D-1)$ matrix such for that $1 \leq i, j \leq D-1$,

$$(\Lambda_{\partial S})_{ij}(t) = \mathrm{cov}\left(\frac{\partial(Y_1(t)/\sigma(t))}{\partial e_i(t)}, \frac{\partial(Y_1(t)/\sigma(t))}{\partial e_j(t)}\right)$$

and corresponds to the covariance of the derivatives of the field $Y_1$ with respect to the

tangent vectors. For the voxel domain $S$ the tangent space, at almost all points on the boundary, is the plane parallel to one of the sides of a voxel on the boundary of the mask. At these points, $\Lambda_{\partial S}$ is the $2 \times 2$ subset of $\Lambda$ corresponding to that plane. In one and two dimensions the LKCs are thus relatively easy to calculate. In three dimensions these formulae allow us to calculate $\mathcal{L}_2$ and $\mathcal{L}_3$ (see Section 2.4 for how to estimate them in practice).

An easy to evaluate closed form for $\mathcal{L}_1$ in 3D has not yet been derived in the non-stationary setting however we have made progress on this recently, see Telschow et al. (2020b). At high thresholds (which are the ones that concern us when controlling FWER) $\mathcal{L}_2$ and $\mathcal{L}_3$ dominate because of the form of the EC densities. We can thus roughly approximate $\mathcal{L}_1$ by using the value that it would take under stationarity without affecting the accuracy of the methods.

Our approach follows that of Worsley et al. (1996) but uses the GKF with non-stationary estimates $\hat{\mathcal{L}}_1, \hat{\mathcal{L}}_2, \hat{\mathcal{L}}_3$ of the LKCs ($\hat{\mathcal{L}}_0 = \mathcal{L}_0$ is known as it is the Euler characteristic of $S$). Given an error rate $\alpha$, we choose the voxelwise threshold $u_\alpha$ such that

$$\sum_{d=0}^{D} \hat{\mathcal{L}}_d \rho_F(u_\alpha) = \alpha.$$

which ensures that the FWER is controlled, since at high thresholds,

$$\mathbb{P}\left(\max_{s \in S} T(s) > u_\alpha\right) \leq \mathbb{E}[M_{u_\alpha}(T)] \approx \mathbb{E}[\chi(\mathcal{A}_{u_\alpha}(T))] \approx \alpha \tag{2.5}$$

For very small values of $\alpha$, the number of maxima of $T$ above $u_\alpha$ is typically 0 or 1, meaning that $\mathbb{P}(\max_{s \in S} T(s) > u_\alpha) \approx \alpha$ is a very close approximation. Additionally fMRI images are relatively large so that, for each $u \in \mathbb{R}$, the estimate: $\sum_{d=0}^{D} \hat{\mathcal{L}}_d \rho_F(u)$ for the EEC of $\mathcal{A}_u(T)$ is very accurate even given low numbers of subjects, see Section

3.4.2, as the noise averages out in the integration when estimating the LKCs. For the smoothed signal RFT provides strong control of the FWER under the assumption of subset pivotality, see the discussion for further details.

When $T$ is the $t$-statistic field, the thresholds for a two-tailed test are obtained similarly by observing that

$$\mathbb{P}\left(\max_{s\in S}|T(s)| > u_{\alpha/2}\right) \leq \mathbb{P}\left(\max_{s\in S} T(s) > u_{\alpha/2}\right) + \mathbb{P}\left(\min_{s\in S} T(s) < -u_{\alpha/2}\right) \leq \alpha.$$

For interesting (i.e. low) $\alpha$ levels, $\mathbb{P}\big(\max_{s\in S} T(s) > u_{\alpha/2}, \min_{s\in S} T(s) < -u_{\alpha/2}\big)$ is very small resulting in a good approximation.

## 2.4 LKC estimation

The LKCs are functions of $\Lambda$, the covariance matrix of the partial derivatives of the random fields, so in order to estimate them we need an estimate of $\Lambda$ (see Section 2.4.1 for details on how this is obtained). In practice, in order to approximate the integrals, we subsample the convolution field at set of points within the voxel domain given by

$$\mathcal{V}_r = \left\{ s \in \mathbb{R}^D : s = v + \frac{k \cdot h}{r+1} \text{ for some } k \in \mathbb{Z}^D \cap \left[-\frac{r+1}{2}, \frac{r+1}{2}\right]^D \text{ and } v \in \mathcal{V} \right\}$$

where the **resolution** $r \in \mathbb{N}$ is the number of points between each voxel that is used in the approximation (taking $r = 0$ we obtain the original lattice, i.e. $\mathcal{V}_0 = \mathcal{V}$). Here $h = (h_1, \ldots, h_D)^T$ is the vector of the spacings between voxels and $\cdot$ denotes the dot product. Given an estimate $\hat{\Lambda}$ for $\Lambda$, and a non-negative resolution $r \in \mathbb{N}$, we estimate $\mathcal{L}_D$ using equation (2.3) via

$$\hat{\mathcal{L}}_D = \sum_{s\in\mathcal{V}_r} w_r(s) \Big| \det(\hat{\Lambda}(s)) \Big|^{1/2}.$$

Here $0 < w_r(s) \leq 1$ is the volume of $U_r(s) \cap S$, where $U_r(s)$ is the 3D cuboid centred at $s$ with side length $r$. The weights reflect the contribution of each point to the integral. Since $S$ is the voxel domain, if $r$ is even, $w_r(s) = 1$ for all $s \in \mathcal{V}_r$. When $D = 3$ we can estimate $\mathcal{L}_{D-1}$ using equation 2.4 as

$$\hat{\mathcal{L}}_{D-1} = \sum_{s \in F^r_{xy}} w'_r(s) \left|\det(\hat{\Lambda}_{xy}(s))\right|^{1/2} + \sum_{s \in F^r_{yz}} w'_r(s) \left|\det(\hat{\Lambda}_{yz}(s))\right|^{1/2} + \sum_{s \in F^r_{zx}} w'_r(s) \left|\det(\hat{\Lambda}_{zx}(s))\right|^{1/2}$$

where for distinct $a, b \in \{x, y, z\}$, $F^r_{ab}$ denotes the set of points that lie on the faces of the boundary of $\mathcal{V}_r$ in the $ab$ plane, $\hat{\Lambda}_{ab}$ denotes the $2 \times 2$ matrix $ab$ subset of $\hat{\Lambda}$ and $w'_r(s)$ is the area of the intersection of $U_r(s) \cap \partial S$ with the $ab$ plane. Traditional RFT methods take $r = 0$ (as they do not use convolution fields) and assume stationarity meaning that $\Lambda$ is constant. Under stationarity and taking $r = 0$, the estimates of the top two LKCs reduce to

$$\hat{\mathcal{L}}_D = \left|\det(\hat{\Lambda})\right|^{1/2} |\mathcal{V}|$$

and

$$\hat{\mathcal{L}}_{D-1} = |F_{xy}| \left|\det(\hat{\Lambda}_{xy})\right|^{1/2} + |F_{yz}| \left|\det(\hat{\Lambda}_{yz})\right|^{1/2} + |F_{zx}| \left|\det(\hat{\Lambda}_{zx})\right|^{1/2}.$$

where for distinct $a, b \in \{x, y, z\}$, $|F_{ab}|$ is the volume of the $ab$ faces on the boundary of $\mathcal{V}$. For stationary fields $\mathcal{L}_1$ admits a closed form expression which can be used to estimate it, see Worsley et al. (1996).

### 2.4.1 Estimating $\Lambda$

In order to estimate the LKCs we need to estimate the covariance matrix of the partial derivatives: $\Lambda$. There are two main methods in the fMRI literature for estimating this: those of Kiebel et al. (1999) and Forman et al. (1995). Both of these assume stationarity and so give suboptimal estimates of $\Lambda$ when applied to fMRI data. We

discuss both of these approaches and show how a better estimator of $\Lambda$ (and one that is valid under non-stationarity) can be obtained using the convolution field framework.

To do so we calculate the residual fields

$$R_n = \frac{Y_n - \hat{\mu}_N}{\hat{\sigma}_N}, \tag{2.6}$$

$n = 1, \ldots, N$, (where operations are performed pointwise) and estimate $\Lambda$ using the sample covariance matrix of their derivatives. In (2.6) we have demeaned the data in addition to scaling. Under the null hypothesis this is strictly unnecessary because the fields have mean zero but under the alternative it becomes important in order to obtain good estimates of the covariance. Given these fields, for each $v \in \mathcal{V}$, for $d = 1, \ldots, D$ their partial derivatives in the $d$th direction can be estimated via

$$Z_{n,d}(v) = (R_n(v + h_d\delta_d) - R_n(v))/h_d$$

where $\delta_d$ is the unit vector in the $d$th direction. Then, under stationarity, the elements of $\Lambda$ can be estimated (Worsley et al. (1992)) by taking (for $i, j = 1, \ldots, D$),

$$\hat{\Lambda}_{ij} = \frac{N-3}{(N-2)(N-1)|\mathcal{V}|} \sum_{n=1}^{N} \sum_{v \in \mathcal{V}} Z_{n,i}(v) Z_{n,j}(v) \tag{2.7}$$

where the inner sum is taken over all $v$ such that $v, v + h_i\delta_i, v + h_j\delta_j \in \mathcal{V}$ and $|\mathcal{V}|$ denotes the number of voxels in $\mathcal{V}$. Note that we have scaled by $N - 1$ instead of $N$ to account for the fact that we have subtracted the mean in (2.6). Here the factor $\frac{N-3}{N-2}$ is needed to obtain an unbiased estimator (accounting for the fact that we divide by $\hat{\sigma}_N$ rather than $\sigma$ in (2.6)), see Kiebel et al. (1999) and Worsley (1996b) for details.

However this estimate is biased because of the use of discrete derivatives see (Kiebel et al., 1999, Figure 3) and our Figure 2.4). A better estimate is obtained by using convolution fields, namely calculating (for each $t$ in some subset $S' \subset S$) the $D \times D$

matrix $\hat{\Lambda}(t)$ such that for $i, j = 1, \ldots, D$,

$$\hat{\Lambda}_{ij}(t) = \frac{N-3}{(N-2)(N-1)} \sum_{n=1}^{N} \frac{\partial R_n(t)}{\partial t_i} \left( \frac{\partial R_n(t)}{\partial t_j} - \frac{1}{N} \sum_{m=1}^{N} \frac{\partial R_m}{\partial t_j} \right) \tag{2.8}$$

where the derivatives, $\frac{\partial R_n(t)}{\partial t_i}$ can be computed exactly. In the stationary case averaging this over all $t \in L$ yields a convolution estimate of $\Lambda$. In the non-stationary case where $\Lambda$ varies over the image, (2.8) provides an unbiased point estimate for $\Lambda$ at each $t \in S'$. This can be seen by arguing as in (Worsley, 1996b).

**Remark 2.2.** *The Kiebel estimate of $\Lambda$ can be modified to work under non-stationarity by taking*

$$\hat{\Lambda}_{ij}(v) = \frac{N-3}{(N-2)(N-1)} \sum_{n=1}^{N} Z_i(v) Z_j(v) \tag{2.9}$$

*for each $v$ such that $v, v + h_i\delta_i, v + h_j\delta_j \in \mathcal{V}$. However it still suffers from the fact that it uses discrete derivatives and thus provides a biased estimate.*

The estimate $\hat{\Lambda}$ can be plugged into the expressions, described above, for the estimated LKCs. Under non-stationarity the estimate for $\hat{\Lambda}$ is noisy at each point. However the noise averages out in the sum, yielding accurate estimates of the LKCs.

### 2.4.2 Estimating the FWHM

Convolving Gaussian white noise with a Gaussian kernel with covariance $\Sigma$ yields a random field that has $\Lambda = \Sigma^{-1}/2$ (Holmes, 1994). In fMRI $\Sigma$ has typically taken to be diagonal i.e.

$$\Lambda = \begin{pmatrix} \Sigma_{11}^{-1}/2 & 0 & 0 \\ 0 & \Sigma_{22}^{-1}/2 & 0 \\ 0 & 0 & \Sigma_{33}^{-1}/2 \end{pmatrix} = \begin{pmatrix} 1/\mathrm{FWHM}_1^2 & 0 & 0 \\ 0 & 1/\mathrm{FWHM}_2^2 & 0 \\ 0 & 0 & 1/\mathrm{FWHM}_3^2 \end{pmatrix} 4\log(2)$$

where $\mathrm{FWHM}_d$ is the smoothness in the $d$th direction. If the kernel is isotropic then the FWHM is the same in each direction and $\Lambda$ is proportional to the identity. This

equivalence has led to an emphasis on FWHM estimation when performing RFT inference in fMRI as the FWHM can be easier to interpret than $\Lambda$. Having smoothed with an isotropic Gaussian kernel, given an estimate $\hat{\Lambda}$, as discussed in Worsley et al. (1992), the FWHM can be estimated via

$$\widehat{\text{FWHM}} = \left( \frac{4\log(2)}{\frac{1}{D}\sum_{d=1}^{D} \hat{\Lambda}_{dd}} \right)^{1/2}.$$

When $\Sigma$ is diagonal but non-isotropic, the FWHM in the $d$th direction can be estimated via $\widehat{\text{FWHM}}_d = \hat{\Lambda}_{dd}^{-1/2}\sqrt{4\log(2)}$. Plugging in the lattice estimate (2.7) for $\hat{\Lambda}$ yields the Kiebel estimate of the FWHM. If we instead plug in equation 2.8 we obtain the convolution estimate of the FWHM under stationarity.

These approaches make use of the relationship between $\Lambda$ and the FWHM that arises from smoothing white noise with a Gaussian kernel. They make the assumption that the underlying process is a continuous random field. This causes problems for the Kiebel estimate because of the bias that occurs due to taking discrete derivatives. An alternative, due to (Forman et al., 1995), under the assumption of stationarity, removes the need for continuous fields and instead derives an estimate that is valid on the lattice. In this case when the kernel is isotropic the estimate of the smoothness is given by

$$\widehat{\text{FWHM}} = \left( \frac{-2\log(2)}{\log\left(1 - \frac{1}{2D}\sum_{d=1}^{D} \hat{\Lambda}_{dd}\right)} \right)^{\frac{1}{2}}.$$

Note that the estimate $\hat{\Lambda}_{dd}$ is the one calculated in equation 2.7 and includes the scaling factor of $\frac{N-3}{N-2}$. In fact the original Forman estimator appears to not have accounted for the scaling factor: we include the factor because it leads to a better estimate (see Figure 2.4). When $\Sigma$ is diagonal but non-isotropic, the smoothness in the $d$th direction can be estimated via $\widehat{\text{FWHM}}_d = \sqrt{2\log(2)}\left(-\log\left(1 - \frac{1}{2}\hat{\Lambda}_{dd}\right)\right)^{-\frac{1}{2}}$. See Jenkinson (2000) for

a detailed derivation of this estimator. We compare the performance of the Forman, Kiebel and convolution estimators for the FWHM in Section 3.1.

Under non-stationarity (or even under stationarity where the smoothing kernel is not a diagonal Gaussian) the FWHM cannot be interpreted so easily, though estimates of it still provide an indicator of the average smoothness levels of the data. We are thus unable to rely on the Forman or Kiebel estimates of the smoothness to give reliable estimates of the LKCs other than in the simple case of white noise convolved with a diagonal Gaussian kernel. This is not a problem in our context because for LKC estimation all we need is $\Lambda$ and we can forget about the FWHM completely.

## 2.5 Resting State Validation Strategy

In the era of Biobank level data, any method that claims to control false positive rates should be rigorously tested to ensure that it does so correctly. In order to validate whether a given method correctly controls the false positive rate in practice, Eklund et al. (2016) and Eklund et al. (2019) introduced the idea of using resting state data. This data is mean zero, as there is no consistent localized activation across subjects, and so provides realistic fMRI data that has no signal. (Note that while resting state data is needed for validation of one-sample analyses, when validating two sample analyses, task data can be used instead.) This approach was applied in Lohmann et al. (2018) to demonstrate false positive control.

Eklund et al. (2016) and Eklund et al. (2019) processed resting state data using a variety of first level designs. They used 3 different datasets: Beijing, Cambridge and Oulu consisting of 198, 198 and 103 subjects respectively. For each dataset they took 6 different first level designs, two block and 4 event related and estimated the FWER

using a resampling approach.

In order to validate FWER control using convolution fields we take a similar approach using resting state data from 7000 subjects from the UK Biobank[3]. For each subject we have a total of 490 T2*-weighted blood-oxygen level-dependent (BOLD) echo planar resting state images [TR = 0.735s, TE = 39 ms, FA = 52°, 2.4mm$^3$ isotropic voxels in $88 \times 88 \times 64$ matrix, $\times 8$ multislice acceleration]. We pre-process the data for each subject using FSL and use a block design at the first level consisting of alternating blocks of 20 time points to obtain first level contrast images. We transform these constants to 2 mm MNI space using nonlinear warping determined by the T1 image and an affine registration of the T2* to the T1 image to obtain our COPEs. The resting state processing used in the Eklund papers was criticized for not being mean-zero (Slotnick, 2017). In order to avoid this criticism the blocks for each subject were randomly phase shifted which ensures that the resulting COPEs are mean zero. We apply our RFT pipeline to this pre-processed data to evaluate whether we obtain the nominal false positive rate.

In order to perform the validation, for $N \in \{10, 20, 50\}$ we choose 5000 subsets of size $N$ from the 7000 subjects, and apply isotropic 3D Gaussian smoothing of $\{2, 3, 4, 5, 6\}$ FWHM per voxel to the COPEs to obtain convolution fields. For each subset and smoothness level we estimate the LKCs under non-stationarity as described in Section 2.4 and use these to obtain one and two tailed $\alpha$-thresholds for the one-sample $t$-statistic for $\alpha \in \{0.05, 0.01\}$. We calculate the maximum and minimum of the $t$-statistic on the original lattice, on the resolution 1 ($r = 1$) lattice and of the convolution field (using optimization algorithms as discussed in Section 2.2). This allows

[3]Full details on imaging acquisition can be found in Miller et al. (2016), Alfaro-Almagro et al. (2018) and from UK Biobank Showcase (`https://Biobank.ctsu.ox.ac.uk/crystal/docs/brain_mri.pdf`), we provide a brief description here. All data were anonymized, and collected with the approval of the respective ethics boards.

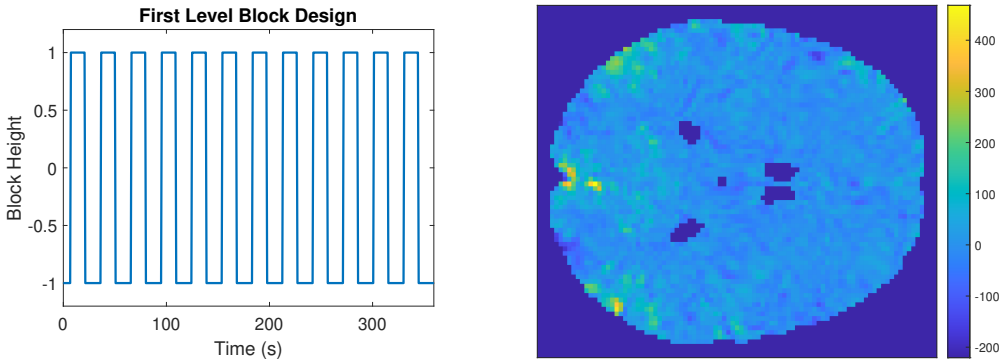



Figure 2.2: The first level block design. The left plot shows one of the random block designs that we used at the first level. The other random blocks used are randomly shifted versions of this design. The right image displays a slice through the brain of the COPE image of one of the subjects that has been processed through FSL using this block design before the data has been smoothed.

us to estimate the FWER for each setting. In order to compare the difference between using stationary and non-stationary LKC estimates we also calculate the EEC under stationarity using the Kiebel and Forman methods.

The intersection of the masks for each subject is different within in each subset of size $N$. In order to compare the EC curves across subsets we run the validation on the intersection mask, taken across all 7000 subjects. (In practice, given a set of subjects, we recommend using the intersection mask which in typical fMRI samples sizes is substantially larger than the mask calculated using 7000 subjects. The latter is smaller due to dropout.)

Eklund et al. (2016) and Eklund et al. (2019) generated the confidence bands for their FWER estimates using stationary Gaussian simulations. However the lack of stationarity Eklund et al. (2016) and Gaussianity (see Section 2.6) in fMRI data means that these bands are likely inaccurate. In our context we have instead opted to treat the draws as independent in order to calculate confidence bands for the coverage rates using the CLT approximation to the binomial distribution. Such an approach would

not have been feasible for Eklund et al. (2016) (as they had to account for dependence between their resamples) as their datasets consist of at most 198 subjects, however it is reasonable in our setting where we are drawing subsets of size at most 50 from a pool of 7000 subjects.

Given $J = 5000$ sets of $N$ subjects, we have test-statistics images $T_1, \ldots, T_J$. For each $j$, we define the Euler characteristic (EC) curve $\chi_j : \mathbb{R} \to \mathbb{R}$ taking $u \in \mathbb{R}$ to

$$\chi_j(u) = \chi(\mathcal{A}_u(T_j)).$$

In order to validate the theory, we can compare the **empirical expected EC curve**:

$$\frac{1}{J}\sum_{j=1}^{J} \chi_j$$

(where the sum is performed pointwise) to the plugin estimate of the expected EC curve calculated using the GKF. Note that the subsets are all calculated using the same mask (namely the intersection mask over all 7000 subjects) so the EC curves are comparable. The estimated LKCs are in fact different for each subset. In order to compare the single average EC curve to the theoretical prediction of the GKF we average the LKC estimates across all 5000 subsets.

We do this for $N \in \{10, 20, 50\}$ and compare the expected EC curves that result from estimating the LKCs using our non-stationary approach to the empirical EC curve calculated using 5000 subsamples. We also calculate the expected EC curve that results from using the Forman and Kiebel approaches to estimating the LKCs. The results are shown in Section 3.4.2.

## 2.6 Gaussianization and robustness to non-Gaussianity

### 2.6.1 Theory

It has been common practice to assume that second-level fMRI data follows a Gaussian distribution (Mumford and Nichols (2006), Worsley et al. (1996)). Analysis of the processed data from the UK Biobank (see Figure 2.3) as well as the Oulu, Beijing and Cambridge datasets used in Eklund et al. (2016) shows that this is far from the case. This non-Gaussianity likely causes problems for a wide range of currently used methods to analyse fMRI data. Given large enough numbers of subjects it is nevertheless reasonable to assume that the test-statistic field follows a normal distribution. This is valid for Biobank level analyses but for the small sample sizes typically used in fMRI studies it turns out be unreasonable.

Theorem 2.1 requires that the underlying fields (i.e. the second level data) used to construct the test-statistic are Gaussian in order to be valid. If we have sufficiently many subjects then the test-statistic is approximately Gaussian because of the CLT and we can apply Theorem 2.1 directly to the resulting test-statistic field. We show (Section 3.4) that in fMRI Gaussianity of the underlying fields is not satisfied and that it seems likely that hundreds of subjects are required before the test-statistic field is sufficiently Gaussian for the GKF to hold. Primarily this occurs because tail outliers have a large effect on the test-statistic. In particular they have a large influence on the distribution of the maximum.

Non-Gaussianity and the lack of convergence in the CLT appear to be big problems for parametric methods which assume Gaussianity; such as RFT. In order to address this we Gaussianize the data via a procedure that we outline in this section. Existing approaches to improving the Gaussianity of the data (such as rank based quantile nor-

malization, (Bartlett (1947), Van der Waerden (1952)) typically work marginally and so applying them in our spatial setting does not take advantage of all of the information that is available. (Rank based quantile normalization itself is additionally not of use in our context as it leaves the test-statistic unchanged.) Our Gaussianization procedure instead works by estimating a null distribution from the data, using information from across all voxels, and transforming the original data (without demeaning it) in order to eliminate heavy tails and improve the level of marginal Gaussianity under the null hypothesis. To obtain the null distribution we standardize and demean the data voxelwise and combine this over the brain. Going back to the original data and standardizing voxelwise without demeaning we can determine the quantile of every voxel relative to this null distribution and use this to convert the data to have approximately Gaussian marginal distributions.

Currently it is standard, when doing one-sample analyses, to compute the $t$-statistic:

$$\frac{\sqrt{N}\hat{\mu}_N(Y_1(v),\ldots,Y_N(v))}{\hat{\sigma}_N(Y_1(v),\ldots,Y_N(v))}.$$

at each voxel $v \in \mathcal{V}$. When the noise is Gaussian, the $t$-statistic is the uniformly most powerful test to detect an effect (Neyman and Pearson, 1936). However when the noise is not Gaussian this is no longer the case and it is possible to obtain test-statistics that are more powerful under the alternative. In particular we consider test-statistics of the form

$$\frac{\sqrt{N}\hat{\mu}_N(f(Y_1(v)),\ldots,f(Y_N(v)))}{\hat{\sigma}_N(f(Y_1(v)),\ldots,f(Y_N(v)))}.$$

We can choose $f$ to make the data more Gaussian under the null hypothesis. If we knew the marginal CDF of the data: $\Psi_v$ under the null at a given voxel $v$ we could then transform our data to $Y'_n(v) = \Phi^{-1}\Psi_v(Y_n(v))$ to ensure that the data was marginally

Gaussian under the null. In fMRI the data has a distribution which is more Laplacian than Gaussian: narrow around the centre and with heavier tails (see Figure 2.3). Of course the marginal CDF is typically unknown in fMRI, however, we can estimate it by making the assumption that it is the same at each voxel (up to scaling by the standard deviation).

More formally, at each voxel $v$ we standardize and demean the fields $X_n$. This yields standardized fields:

$$X_n^{S,D} = \frac{X_n - \hat{\mu}_N(X_1, \ldots, X_N)}{\hat{\sigma}_N(X_1, \ldots, X_N)}. \tag{2.10}$$

In order to calculate the marginal null distribution we combine this data over all voxels and subjects to obtain a null distribution (note that because we subtract the mean in equation (2.10), the estimate of the distribution is not affected by whether the maps are dominated by signal or by noise). This null distribution is illustrated in Figure 2.3 for a subset of 50 subjects from the UK Biobank. As can be seen from the histogram the data is heavy tailed and clearly not drawn from a $t_{50}$ distribution. Going back to the original data we standardize it (without demeaning) to yield:

$$X_n^S = \frac{X_n}{\hat{\sigma}(X_1, \ldots, X_N)}$$

and for each voxel $v$ and subject $n$ we compare $X_n^S(v)$ to the null distribution to obtain a quantile

$$q_n(v) = \frac{1}{N|\mathcal{V}|} \sum_{n=1}^{N} \sum_{v' \in \mathcal{V}} 1[X_n^S(v) \leq X_n^{S,D}(v')].$$

The Gaussianized fields are then given by

$$X_n^G(v) = \Phi^{-1}(q_n(v))$$

for each voxel $v$ and subject $n$, where $\Phi$ is the CDF of the normal distribution. We

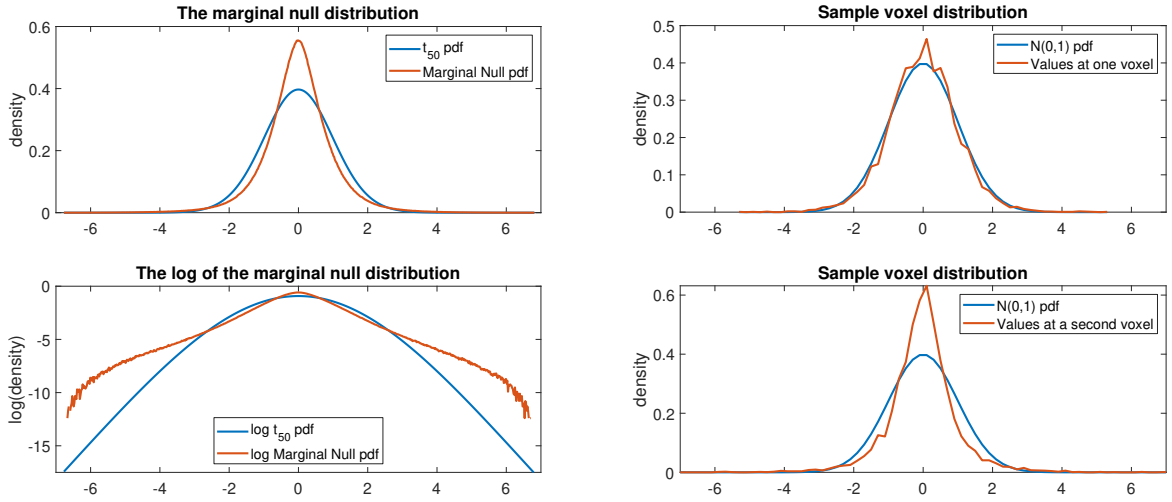



Figure 2.3: The marginal distributions of the resting state data. In the plots on the left we plot the histogram of the marginal null distribution (the density and its log) for a subset of 50 randomly selected subjects (over the values for all subjects and voxels in the 50 subject intersection mask) calculated using (2.10) against the $t_{50}$ pdf. There is substantial discrepancy between the two, illustrating that the data is highly non-Gaussian: instead the distribution has a lot of weight at the centre and has very heavy tails. On the right we plot two histograms of the observed values at two different voxels, calculated over all 7000 subjects (and scaled to have variance 1) against the pdf of the normal distribution. As can be seen from these plots the degree of non-Gaussianity can vary across voxels.

obtain Gaussianized convolution fields

$$Y_n^G(s) = \sum_{v' \in \mathcal{V}} K(s - v') X_n^G(v') \tag{2.11}$$

and corresponding convolution $t$-fields on which we can perform inference.

We perform the standardization before smoothing because otherwise voxels that have a higher variance disproportionately affect voxels with a lower variance. This is actually an issue with existing fMRI analysis pipelines as high and low variance voxels are smoothed together. Standardizing (with the option of Gaussianizing) before smoothing means that voxels contribute equally. Such an approach is only possible if smoothing is performed after whitening.

It is important to note that this transformation does not guarantee that the data is

jointly Gaussian. However joint Gaussianity of the test-statistic is provided by smoothing and the CLT. For our procedure we do not need to make the assumption that the marginal distributions at each voxel are the same (an assumption that is itself weaker than marginal Gaussianity) in order to ensure validity as this is ensured by the CLT given sufficiently many subjects. When the data is heavy tailed the CLT comes into effect much more quickly on the Gaussianized data (compared to the original data) because the heavy tails can lead to slow CLT convergence (see Sections 3.3 and 3.4). The Gaussianization procedure does introduce some dependence between subjects. This will be small given sufficiently many subjects (as it comes about primarily through $\hat{\sigma}$), moreover this is not a problem because the CLT still applies under weak dependence (Bradley Jr, 1981).

### 2.6.2 Simulations

In Section 3.2 we run 2D Gaussian simulations to illustrate the benefits of using convolution fields and the validity of our RFT framework. For these we use a mask that is a 2D slice through the MNI mask and generate random fields by smoothing white noise with an isotropic Gaussian kernel with smoothness ranging from 2 to 6 FWHM per voxel. For each smoothness level we run 5000 simulations: in each applying RFT and calculating the maximum on the lattice, the resolution 1 lattice and of the convolution field in order to calculate the one-tailed FWER in each setting. We compare these to the value of the expected Euler characteristic (which at these thresholds is the expected number of maxima).

In order to illustrate the effect of the Gaussianization transformation we perform these simulations on Gaussianized data. We consider noise distributions which are marginally $t$ with three degrees of freedom, apply the Gaussianization procedure to

the data and measure the false positive rates that result from using RFT. We choose this noise distributions because it results in the data having heavy tails and is thus challenging for RFT. For comparison we also include the effect on false positive rates of applying the Gaussianization to data which is already Gaussian.

Finally we perform the resting state validations in Section 3.4 with and without Gaussianizing the data in order to illustrate the importance of the Gaussianization procedure when applying RFT to fMRI data.

# 3 Results

## 3.1 Smoothness estimation

In this section we compare the different methods for estimating $\Lambda$ by comparing the different estimates of the FWHM. As discussed in Section 2.4.2, when estimating the FWHM of white noise smoothed with a diagonal Gaussian kernel we can estimate the FWHM via the Kiebel, Forman or convolution estimates. In order to compare their performance we generate random fields on a $30 \times 30 \times 30$ lattice by smoothing i.i.d Gaussian white noise with a diagonal isotropic Gaussian kernel. For each applied FWHM in $\{2, 2.5, \ldots, 6\}$ we generate $N = 50$ and 100 random fields to yield an estimate of the smoothness. We do this 1000 times and take the average over the estimates for each $N$ and each FWHM. The results are shown in Figure 2.4. Note that to correct for the edge effect we simulate data on a larger lattice (increased by a size of at least 4 times the standard deviation of the kernel in each direction) and take the central lattice subset.

The results in Figure 2.4 show that the Kiebel estimates are positively biased and

the Forman estimates are negatively biased, though the bias of the Kiebel estimator decreases as the smoothness increases. The convolution estimates have negligible bias except for low FWHM. The reason for the bias at low applied smoothness is that the convolution fields generated are non-stationary. This is not a problem for LKC estimation, however the FWHM is only well defined for stationary random fields and so we wouldn't expect a meaningful estimate in this case. For larger FWHM the convolution fields are still technically non-stationary however in practice they are very close to being stationary so the FWHM can be accurately estimated.

We have also plotted the estimates of the FWHM that result when the estimate for $\Lambda$ is not scaled by $\frac{N-3}{N-2}$. Asymptotically the scaling factor is irrelevant but it has a notable effect at low $N$. For the Kiebel estimate, at these smoothness levels, not scaling leads to an artificial improvement but causes bias when the applied smoothing is higher. The fact that the convolution estimates appear unbiased when the scaling factor is applied provides a clear indication that scaling by the factor is the correct approach: as it provides an unbiased estimate for $\Lambda$ (Worsley, 1996b).

## 3.2 Gaussian Simulations

In order to evaluate the ability of the method to correctly control familywise false positive rates we run the 2D Gaussian simulations described in Section 2.6.2. The results for one-tailed tests at the FWER $\alpha = 0.05$ threshold are shown in Figure 2.5. They show that when using convolution fields the FWER is accurately controlled to the nominal level at all of the $N$ and applied smoothness levels that we considered. The EEC is correctly estimated and provides an upper bound (due to the inequality in equation 2.5). The confidence bands in these and all subsequent figures are calculated

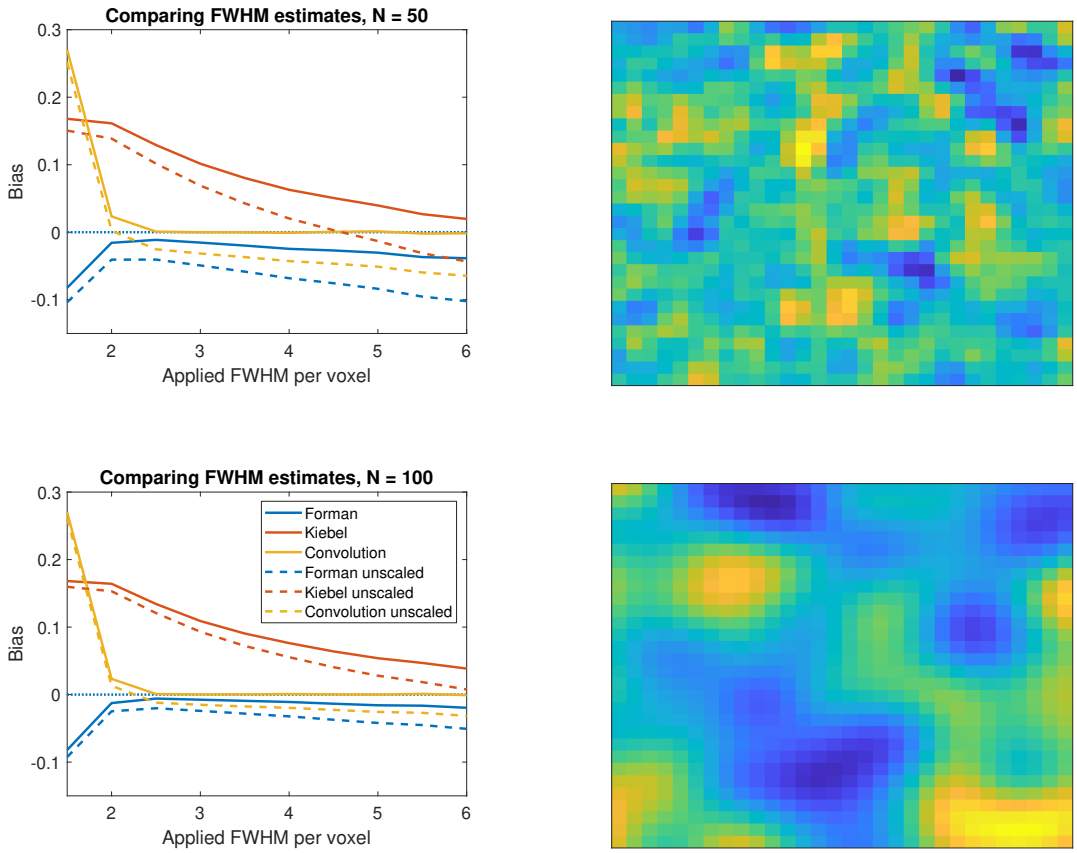



Figure 2.4: Comparing FWHM estimation for Gaussian white noise smoothed with a diagonal isotropic Gaussian kernel. On the left we plot the FWHM estimation bias (estimated-applied) against the applied FWHM for the 3 different methods (obtained using $N = 50$ (top left) and 100 (bottom left) subjects). The convolution estimate is essentially unbiased when the applied FWHM is greater than or equal to 2.5. The Forman and Kiebel estimates are both biased at all FWHM though the bias of the Kiebel estimator decreases as the smoothness increases. These graphs illustrate the importance of the $\frac{N-3}{N-2}$ scaling factor: the unscaled results (dashed lines) omit this factor and result in a downward bias. On the right we plot two 2D slices through 3D white noise smoothed with FWHM = 2 voxels (top right) and 6 voxels (bottom right) on original ($r = 0$) lattice.

using the normal approximation to the binomial distribution.

Figure 2.5 also shows that applying the methods on the original ($r = 0$) lattice without using convolution fields (as in the traditional RFT approach proposed by Worsley et al. (1996)) leads to conservative inference especially at low smoothness levels. Using the resolution 1 ($r = 1$) lattice leads to an improvement as the maximum on the fine lattice is closer to the maximum of the convolution field.

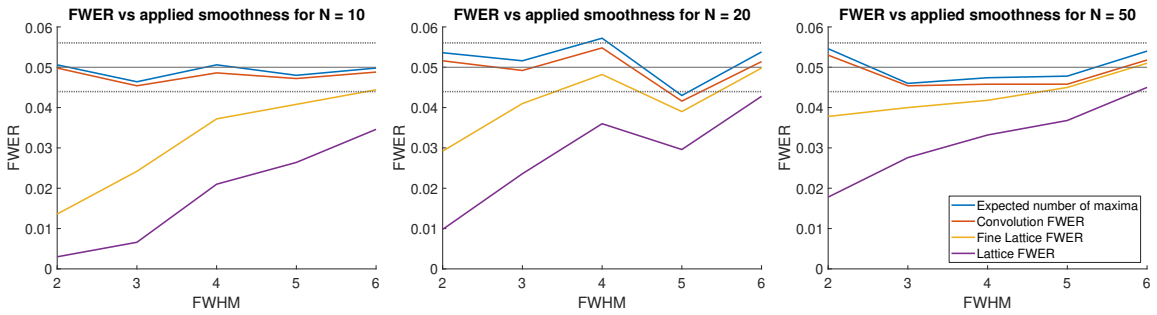



Figure 2.5: Plotting the FWER against the applied smoothness for $N \in \{10, 20, 50\}$ when applying the RFT methods to stationary Gaussian simulations on a slice of the 2mm MNI brain mask. When using convolution fields the expected number of local maxima (shown in blue) is correctly estimated and the FWER (shown in red) is controlled accurately to the nominal level. When the methods are instead applied to the lattice data (shown in purple) the FWER control is conservative, though this improves as the smoothness increases. When the resolution one ($r = 1$) finer lattice (shown in yellow) is used the conservativeness is decreased though not as much as when using convolution fields.

## 3.3 The Gaussianization transform

In this section we consider simulations on non-Gaussian $t_3$ data, as described in Section 2.6.2, in order to evaluate the effects of the Gaussianization transformation.

The results for one tailed tests at the $\alpha = 0.05$ threshold are shown in Figure 2.6 for $N = 20, 50$ and 100. In these figures we plot the FWER that results from applying RFT on the original lattice and to the convolution random field. This is done for the original data and the Gaussianized data. From these plots we see that, while $N = 20$ results are slightly anti-conservative, for $N = 50$ and 100, the convolution RFT framework controls the false positive rate to the nominal level once we Gaussianize even though the original data is very heavy tailed.

In Figure 2.7 we plot the expected EC curve (calculated using the average of the LKCs over the 5000 simulations as described in Section 2.6) and compare it to the empirical EC curve. For the Gaussianized data these curves closely match especially at high thresholds. When the data is not Gaussianized the theory breaks down, (as

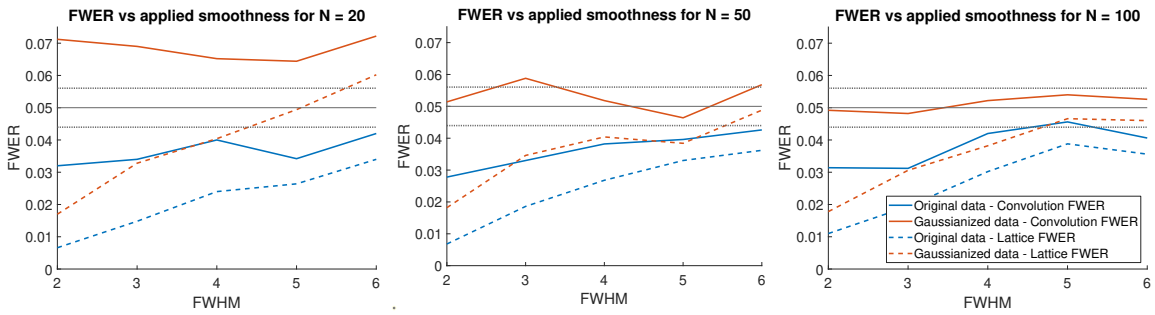



Figure 2.6: Plotting the FWER against the applied smoothness for $N \in \{20, 50, 100\}$ when applying the RFT methods to the original (marginal $t_3$) data and to the Gaussianized data. The methods are conservative when applied to the original data even when the number of subjects is quite large. When the data is Gaussianized and convolution fields are used the FWER (solid red) is accurately controlled to the nominal level once the number of subjects is sufficiently high. If the traditional lattice maximum (dashed red) is used rather than the convolution maximum then a high level of smoothness is required before the methods attain the nominal FWER. For the orignal data (shown in blue), as the smoothness increases the FWER becomes closer to the nominal rate. This is because more averaging is involved and so the CLT comes into affect more quickly.

the data is not Gaussian) even when we use 100 subjects, and the Euler characteristic is overestimated at each threshold leading to conservative inference.

In Figure 2.8 we plot the FWER against the number of subjects, that are used, for applied smoothing levels of 2 and 4 FWHM per voxel. As in Figure 2.6 applying RFT to the original data leads to conservative inference, even when using convolution fields. This illustrates that we cannot rely on the CLT to ensure Gaussianity. After Gaussianization the FWER converges relatively quickly to the nominal level as the number of subjects increases. Using the maximum on the lattice rather than the maximum of the convolution field leads to conservativeness especially at low smoothness levels.

For the original data the EEC is incorrectly estimated (even given $N = 100$) and the FWER control is conservative as a result. This is unsurprising as the data is not Gaussian and so we do not expect Theorem 2.1 to hold. However, applying the RFT

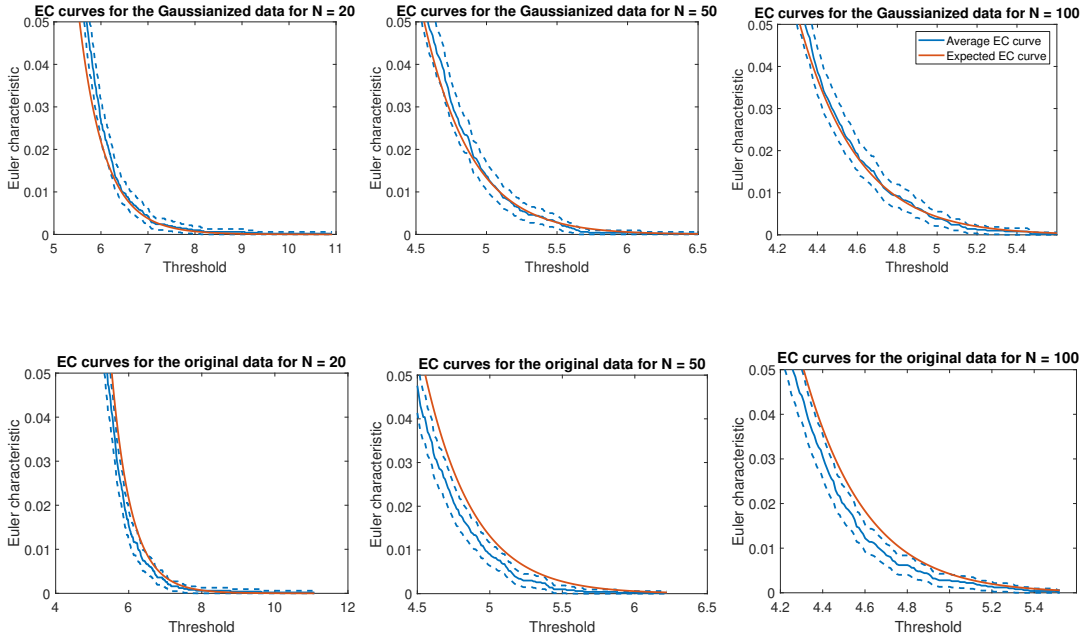



Figure 2.7: Comparing the expected (shown in red) and empirical (shown in blue) tail EC curves for an applied smoothness of 4 FWHM per voxel. The empirical average curve is calculated as the average of the observed EC curves of the one-sample $t$-statistics of 5000 subsets of size $N$. For the Gaussianized data (top row), given sufficiently many subjects, the empirical average EC curve is closely approximated by the EEC (calculated using the average of the LKCs over 5000 simulations) especially at high thresholds. For the original data (bottom row) the expected EC curve is not too far off but requires a much larger number of subjects for it to be an accurate approximation.

framework to the Gaussianized data and using convolution fields we can accurately estimate the EEC and obtain valid and accurate FWER control given sufficient numbers of subjects. For low numbers of subjects the inference (after Gaussianization) is slightly inflated. This occurs for several reasons. Firstly, under the transformation the heavy tailed nature of the data is decreased however the data is not perfectly Gaussian because subtracting the empirical mean (especially for low numbers of subjects is not equivalent to assuming that the data is mean zero); this problem decreases as the number of subjects increases. Secondly while the data (under the null) is transformed to be marginally close to Gaussian it is not jointly Gaussian. Thirdly for low number of subjects the Gaussianization induces dependence between subjects meaning that the

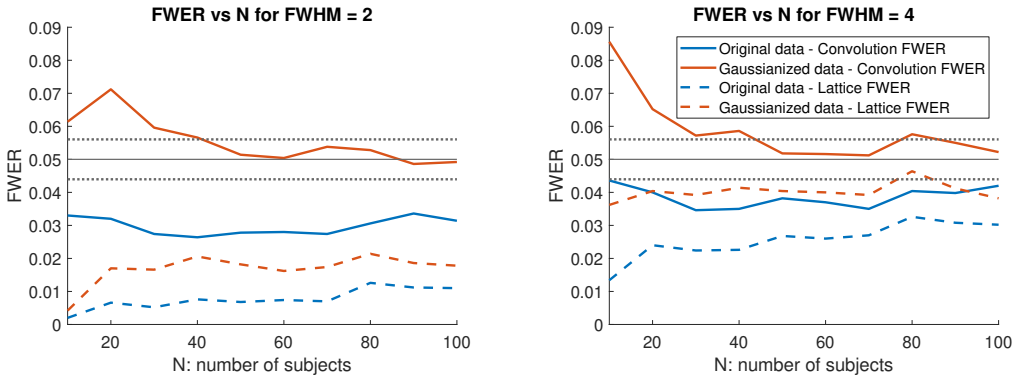



Figure 2.8: Comparing the FWER against the number of subjects when applying the RFT methods to the original (marginal $t_3$) data and to the Gaussianized data for applied smoothing of 2 and 4 FWHM per voxel. The methods are conservative when applied to the original data even when the number of subjects is quite large. When the data is Gaussianized the FWER reaches the nominal level once the number of subjects is around 50 but does reasonably well even for small number of subjects.

results test-statistics are not $t$-fields. Using a reasonable number of subjects means that these issues go away (due improved estimation of the mean, the CLT and reduced dependence) and the FWER is accurately controlled.

For comparison the results of applying the Gaussianization procedure to isotropic Gaussian white noise are shown in Figure 2.11. When RFT is applied to the original data and convolution fields are used, the false positive rate is closely controlled to the nominal level. For the Gaussianized data, the results are similar to when the marginal distribution is $t$, though require a larger number of subjects before the nominal rate is reached.

## 3.4 Resting State Validation

In this section we validate our voxelwise RFT framework using UK Biobank resting state fMRI data, analysed as if it were task data.

### 3.4.1 FWER error rate

We run the error rate evaluations discussed in Section 2.5 (with and without Gaussianizing the data before smoothing) for FWER $\alpha$ levels of 0.05 and 0.01 and show that RFT correctly controls the false positive rates across all settings. The results for $\alpha = 0.05$ are shown in Figures 2.9 and 2.12 and for $\alpha = 0.01$ in Figures 2.13 and 2.14. We also plot the empirical average number of local maxima above the FWER thresholds: from the theory we expect this to have mean $\alpha$ and we expect the FWER to be less than or equal to $\alpha$. (Note that the average number of maxima is the same as the average EC at these thresholds.) From these plots we see that this is indeed the case and that the false positive rate is controlled below the nominal rate in all settings. Using convolution fields, we obtain a huge power increase and obtain coverage that is closer to the nominal rate. This improves as $\alpha$ decreases (compare the results for $\alpha = 0.05, 0.01$) because the approximation in (2.5) becomes more accurate. When the data is not Gaussianized the inference is still valid but is conservative. The average EC is far below the nominal level indicating that the theory fails in this context. This is strong evidence that the test-statistic is not a $t$ random field and that the CLT has not yet come into effect. When the data is Gaussianized the average EC is well predicted by the theory and the FWER that results from using convolution random fields is controlled close to the nominal level (shown in red). This underlines the positive effects of the Gaussianization procedure.

For the Gaussianized data, the Euler characteristic is correctly predicted at all levels of applied smoothness. Traditional RFT methods apply the RFT correction on a lattice ($r = 0$) leading to high levels of conservativeness (the results of doing this are shown in purple) especially at low smoothness levels because of the failure of the good

lattice assumption. By using convolution fields we have eliminated this assumption. To see the effect on the FWER as the resolution increases, we have also included the FWER that results from using the maximum on the resolution 1 ($r = 1$) lattice. This is still conservative relative to the convolution maximum however it is close to the convolution FWER given sufficient smoothness.

We note that there is a discrepancy between the EEC and the FWER of the maximum of the convolution field. This occurs because equation (2.5) provides an upper bound. This upper bound becomes an equality at high thresholds because the number of maxima is either 0 or 1 with high probability but it causes a discrepancy here. This means that the methods are slightly conservative but are always valid. For two tailed testing at the 0.05 level this is much less of a problem (see Figure 2.12) and the issue almost disappears when we seek to control the false positive rate to 0.01: see Figures 2.13 and 2.14. This occurs because the thresholds in the two tail 0.05 case and in the 0.01 cases are higher so that equation 2.5 becomes a better approximation.

This discrepancy gets worse as the number of subjects and/or the smoothness increases. The reason for this may be that the brain is inherently symmetric, meaning that the noise between the two hemispheres of the brain is highly correlated and the event that two or more clusters occur above the threshold occurs with reasonable probability even at high thresholds. This does not affect the validity of the procedure, yet is a property which needs to be studied in greater detail. One advantage of it is that it makes the methods robust to any small sample over-coverage that arises due to Gaussianization. However, we do not see this over-coverage the real-data validations. The methods work (predicting the Euler characteristic and controlling the false positives rates) when $N$ is as low as 10. This may be because the Gaussianization procedure improves, in terms of its ability to transform the data to have Gaussian marginal

distributions, as the number of voxels increases.

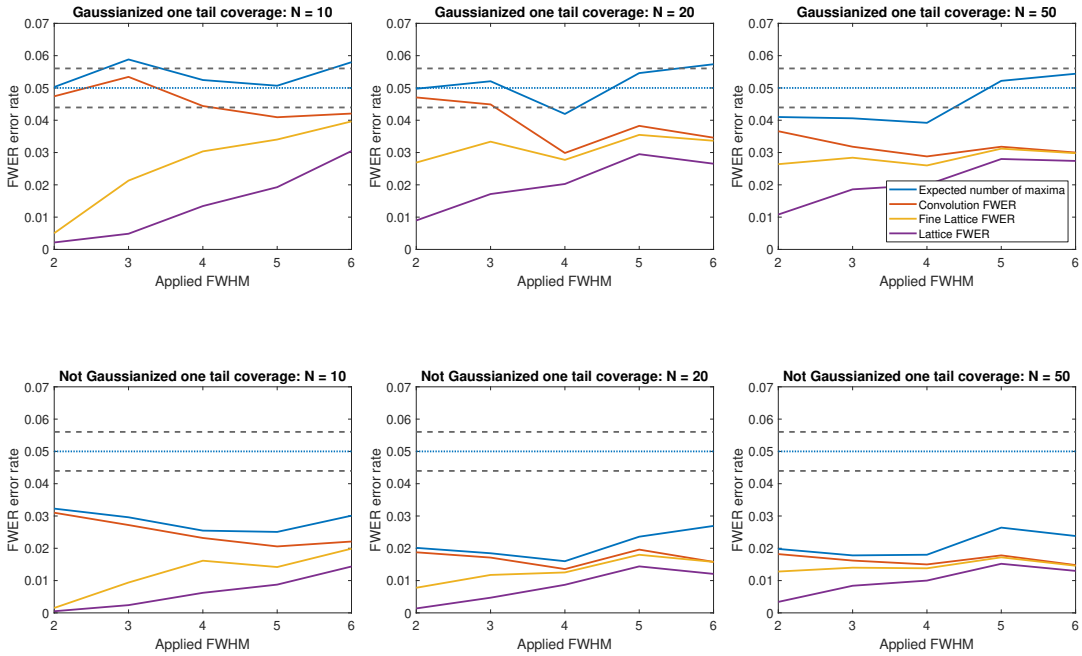



Figure 2.9: Resting state validation of the one tail FWER for $\alpha = 0.05$. For each applied FWHM and $N \in \{10, 20, 50\}$ we plot the average error rate that results from applying the methods to 5000 randomly chosen subsets of our 7000 resting state subjects. We compare FWER control of the data that is Gaussianized before smoothing (top row) to that of the control for the data that is not Gaussianized (bottom row). The FWER using convolution fields (shown in red) is controlled below the nominal rate in all settings though is more accurate for the Gaussianized data. The $(r = 0)$ original lattice FWER (shown in purple) is conservative as is, to a lesser extent, the $(r = 1)$ fine lattice FWER (shown in yellow) though this improves as the smoothness increases. The expected number of maxima (shown in blue) above the $u_\alpha$ threshold is accurately predicted for the Gaussianized data.

### 3.4.2 Empirical vs Expected Euler characteristic

As described in Section 2.5, another way to test how well the theory is doing in practice is to compare the theoretical and empirical EC curves. We have plotted the upper tails of these curves (the sections of the curve that are most relevant for FWER control) for applied smoothnesses of 2 and 5 FWHM per voxel in Figures 2.10 and 2.15. As can be seen from the plots our non-stationary approach gives a very close estimate of

the EEC once the data is Gaussianized. Without Gaussianization the EEC is greatly over estimated. The stationary methods overestimate the EEC in the simulations for $N = 10, 20$ however they perform similarly to the non-stationary estimate for $N = 50$.

Note that it is important not to over interpret these plots. Unlike the FWER plots which are comparable across the resampled subsets, the EC curves rely on the assumption that the subjects are i.i.d. In practice, it seems likely that different subsets may have different covariance structures meaning that it is not necessarily the case that the LKCs are the same across subsets. This may explain any residual discrepancies between the empirical and expected curves that we have plotted because the expected curves are calculated using the average of the LKCs over all subsets. As such while useful as a second way to demonstrate that the theory is working in practice, the most reliable way to see that the methods are working correctly is via the FWER plots.

# 4 Discussion

In this paper we have introduced two innovations that allow for accurate voxelwise RFT inference under low smoothness and non-Gaussianity, making RFT robust to violations of its traditional assumptions. This represents a substantial improvement to the existing voxelwise parametric approaches. Convolution fields bridge the gap between continuous theory and the lattice data that is collected in practice. They have the potential to allow other RFT based approaches, such as peak and clustersize inference, to work without requiring high levels of applied smoothness. The Gaussianization procedure allows the GKF to be valid, when the data is non-Gaussian, given sufficiently many subjects and symmetric marginal distributions. Together these modifications to the standard pipeline allow us to provide a quick and accurate RFT inference frame-

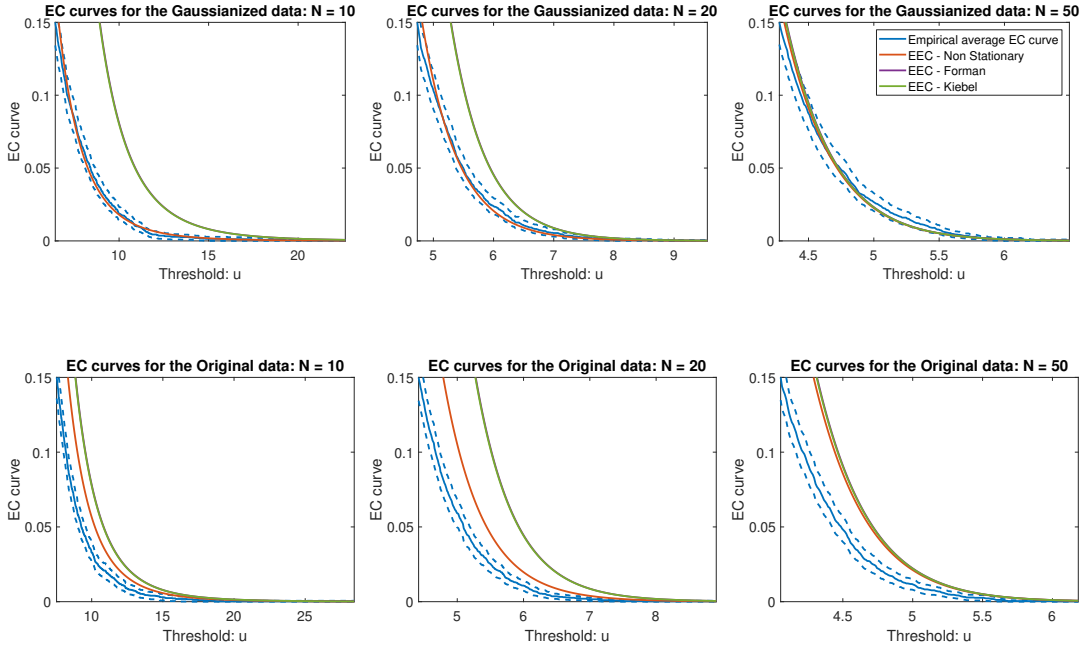



Figure 2.10: Comparing the expected and empirical tail EC curves in different settings using the resting state data for an applied smoothing of 5 FWHM per voxel. The results for the Gaussianized data (top row) and for the original data (bottom row) are noticeably different. For the Gaussianized data, the non-stationary expected EC curve is close to the empirical average curve and always lies within the 95% error bars (calculated using the CLT). For the original data the EEC is overestimated in all cases, which leads to conservative inference. Note that the Kiebel and Forman EC curves are typically so close to each other that they are indistinguishable and provide over estimates other than for $N = 50$.

work, for controlling the voxelwise FWER, that is valid in fMRI. In this work we have primarily considered one-sample test-statistics however the theory is fully general and can be extended to the two-sample and general linear model.

We were able to avoid memory concerns associated with high resolution imaging, by using optimization algorithms to find the peaks of the test-statistic. While this works well to test the global null hypothesis (or within a given region) it is more computational to test the null within a given a voxel. Luckily this is more of a problem for very low sample sizes, where the test-statistic is quite rough, for which it is easier to generate high resolution lattices. For larger sample sizes the solution is to search within the voxels whose initial lattice values lie near to the threshold. There may be

more efficient ways of doing this, such as by using local Slepian models (Adler, 1981) or local derivative approximations, but these are beyond our scope.

Gaussianizing the resting data appears to be effective at improving the underlying Gaussianity. The heavy tails of the original data cause the convergence towards Gaussianity (both marginally and jointly) via the CLT to occur slowly. The marginal distributions of the transformed data are (approximately) Gaussian and the transformation reduces the heavy tails meaning that convergence to Gaussianity occurs faster (both marginally and jointly). Comparison of the EC curves after the transformation shows a close match (between the expected EC curves and the empirical average EC curves), which is strong evidence that it improves the validity of the theory. We have further demonstrated that it allows our parametric methods to accurately control the FWER. The substantial level of non-Gaussianity that we found in the resting state fMRI data raises questions about the validity of parametric methods that assume Gaussianity. This has important implications for a range of other methods that are used to control false positives in fMRI such as clustersize inference Friston et al. (1994) and FDR control Genovese et al. (2002) and means that it is crucial that methods for fMRI inference be robust to non-Gaussianity. We have shown that the power increase that can result from transforming the data is relatively substantial, so we suspect that there are gains to be made by doing so, even when RFT is not used to control the false positive rate.

In our 2D simulations, when combined with RFT, Gaussianization required around 50-70 subjects before nominal coverage was achieved. This is partially due to the dependence that is induced by the procedure in low sample sizes. One possible change to the procedure, which might improve its performance in small samples, would be to average the standard deviation locally in order to reduce the inter-subject dependence.

In the 3D resting state validations 10 subjects were already sufficient to obtain accurate estimation of the EEC and control the false positive rates. In terms of its effect on the data, Gaussianization changes the units of the analysis (because it scales by the standard deviation). This is not an issue in our context, where we want to test the null hypothesis, however it does mean that it is no longer possible to infer directly on the mean %BOLD change. A wider investigation is required to understand the effect of Gaussianization and to compare it to other procedures designed to increase the normality of the data. Further work could look into improving how the transformation works in the tails, in order for the procedure to work well on smaller image sizes. Here we discussed the application of Gaussianization in one-sample models, future work could investigate extending it to work in more general settings.

The discrepancy between the EEC and the convolution FWER, that seems to arise because of symmetry, motivates an alternative procedure that would instead control the false positive rates at the $2\alpha$ level. Given sufficient smoothness and a very large number of subjects it seems that this procedure might still control the FWER to the level $\alpha$ when $\alpha = 0.05$ and would be more powerful. It would be interesting to establish conditions under which this provided correct FWER control and to investigate other methods which take advantage of the symmetry of the covariance structure. Studies that consider voxelwise left-right brain differences do not suffer from the symmetry issue, meaning RFT will could be especially powerful in that setting.

As discussed in Section 2.3, our methods provide strong control of the FWER with respect to the smoothed signal under the assumption of subset pivotality (Westfall and Young, 1993) which is reasonable for fMRI data Nichols and Hayasaka (2003). The Gaussian smoothing kernel has infinite support meaning that the error control is weak relative to the original unsmoothed data (note that this applies to any method that

smoothes the signal - including voxelwise permutation testing). In practice, however, the kernel is truncated (quite substantially). Moreover smoothing could be performed using a kernel which has finite support (such as the quartic kernel) as the RFT framework is valid for any $C^3$ kernel. As such it is possible to make some strong statements about the original signal up to the size of the support of the kernel (or its truncation). Making this precise is non-trivial and it would be interesting to investigate this in future work.

In this work we introduced the idea of smoothing the data after fitting the first level model. This is not essential for the RFT framework as discussed in Section 2.2. However there are a number of advantages that can be gained from doing so. In particular it makes it easier to account for missing data that arises due to dropout, since for each subject the smoothing provides a natural way to extrapolate beyond the support of the subject's mask. In big data studies such as the UK Biobank sometimes up to half of the voxels in the brain are lost, due to dropout, in the intersection mask over all subjects. While it is possible to perform an analysis using the available data at each voxel this is typically slow and does not take advantage of all available information so there is a lot to be gained by postponing the smoothing step. Another advantage is that the data can be masked after the first level inference is performed and before smoothing meaning that there will be less leakage from the CSF into the rest of the mask. One potential issue with this general approach is that the first level autocorrelation estimates may be less accurate when the data is not smoothed beforehand and thus the effect on power needs to be explored; the effect of this is mitigated, however, as the autocorrelation estimates are smoothed before whitening the data (this is the default option in FSL Woolrich et al. (2001)). The departure from the usual fMRI pipeline is in the same spirit as Lohmann et al. (2018) who

considered smoothing the test-statistic instead of applying smoothing before whitening. The optimal point at which to smooth is very much an open area of research.

Future work could look at using these methods to solve problems with RFT clustersize inference: providing a fast and valid parametric approach to clustersize inference using RFT. This is a much harder problem to fix than voxelwise inference because clustersize inference using RFT makes a number of further assumptions which may not be reasonable in practice. Nevertheless it would be exciting to explore this in future research. In this context using the convolution framework means that cluster sizes of the smooth random field can be calculated exactly meaning that the continuous theory should work better in practice. Other interesting extensions of this work include combining it with closed testing Rosenblatt et al. (2018) and applying it to voxelwise GWAS studies: other areas where control of the FWER under dependence is important.

# 5 Acknowledgements

TEN is supported by the Wellcome Trust, 100309/Z/12/Z and SD is funded by the EPSRC. FT, AS and TEN were partially supported by NIH grant R01EB026859. FT thanks the WIAS Berlin for its hospitality where part of the research for this article was performed. The research was carried out under UK Biobank application #34077, with bulk image data shared within Oxford (with UK Biobank permission) from application 8107.

# 6 Further Figures

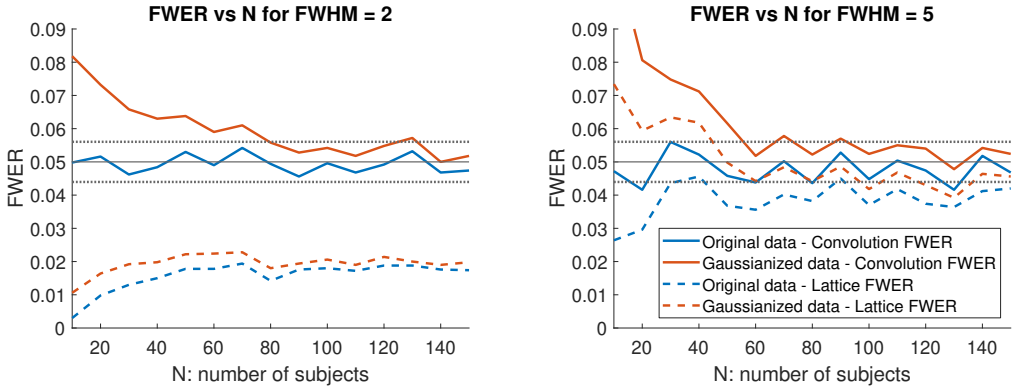



Figure 2.11: Comparing the FWER against the number of subjects when applying the RFT methods to the original (Gaussian) data and to the Gaussianized data for applied smoothing of 2 and 5 FWHM per voxel. When the data is Gaussianized the FWER reaches the nominal level given sufficiently many subjects.

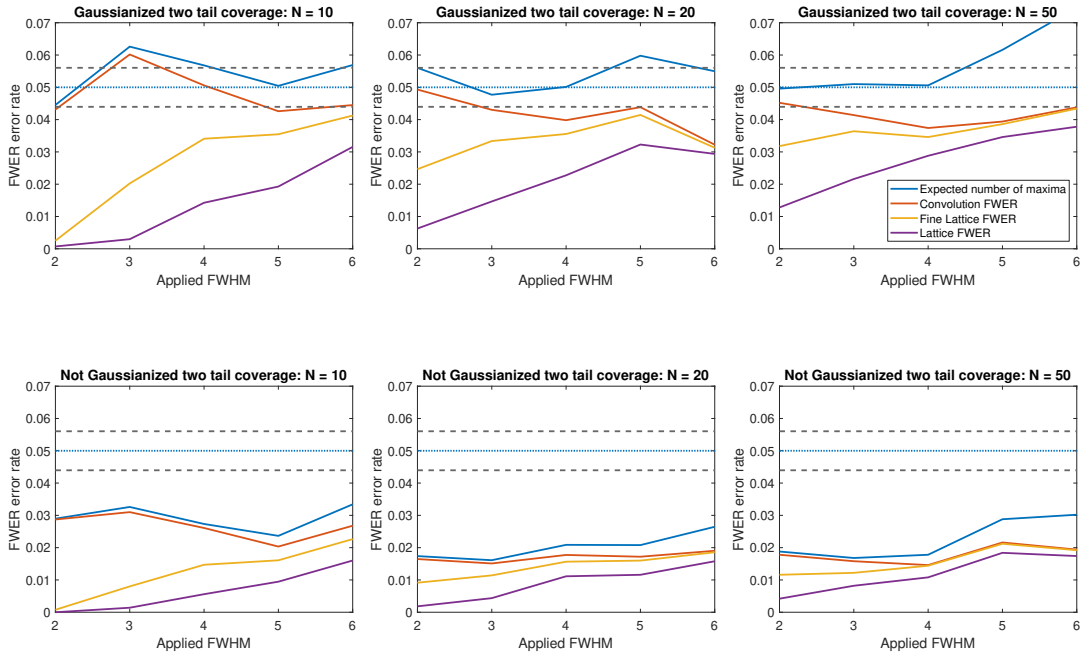



Figure 2.12: Resting state validation of the two tail FWER for $\alpha = 0.05$. Other than being two tailed the settings in each plot are the same as in Figure 2.9. The two tailed thresholds are higher than the one tailed ones meaning that the convolution FWER (shown in red) is closer to the nominal rate.

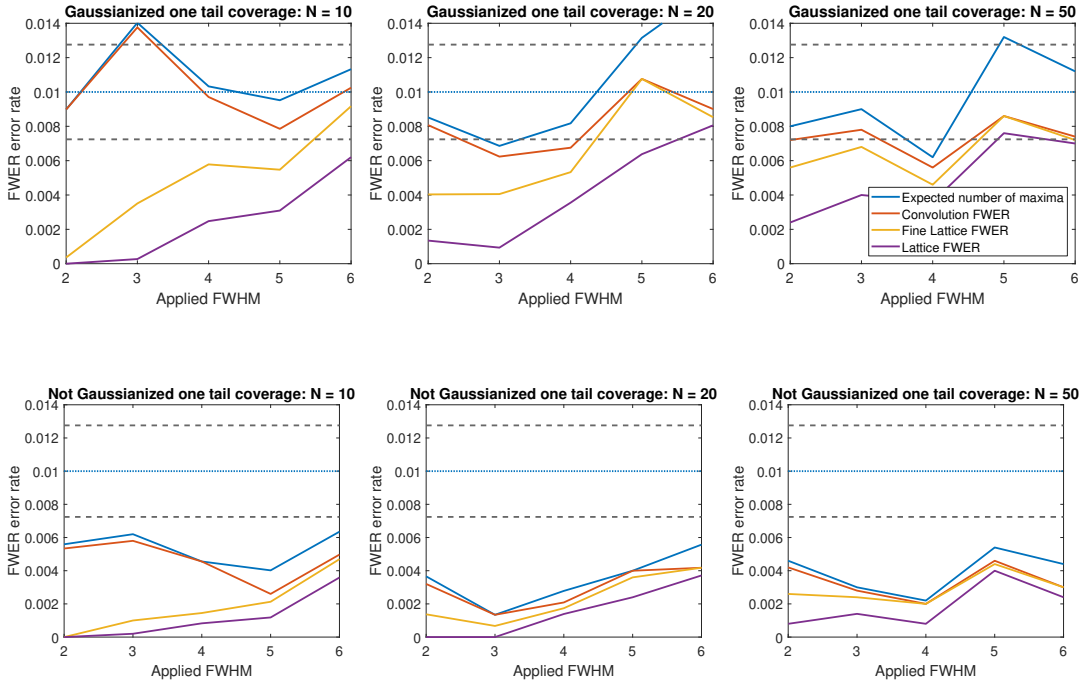



Figure 2.13: Resting state validation of the two tail FWER for $\alpha = 0.01$. The settings in each plot are the same as in Figure 2.9. The FWER is controlled in all settings but is controlled more accurately for the Gaussianized data. Controlling at $\alpha = 0.01$ requires a higher threshold so the convolution FWER (shown in red) is closer to the nominal level than for $\alpha = 0.05$.

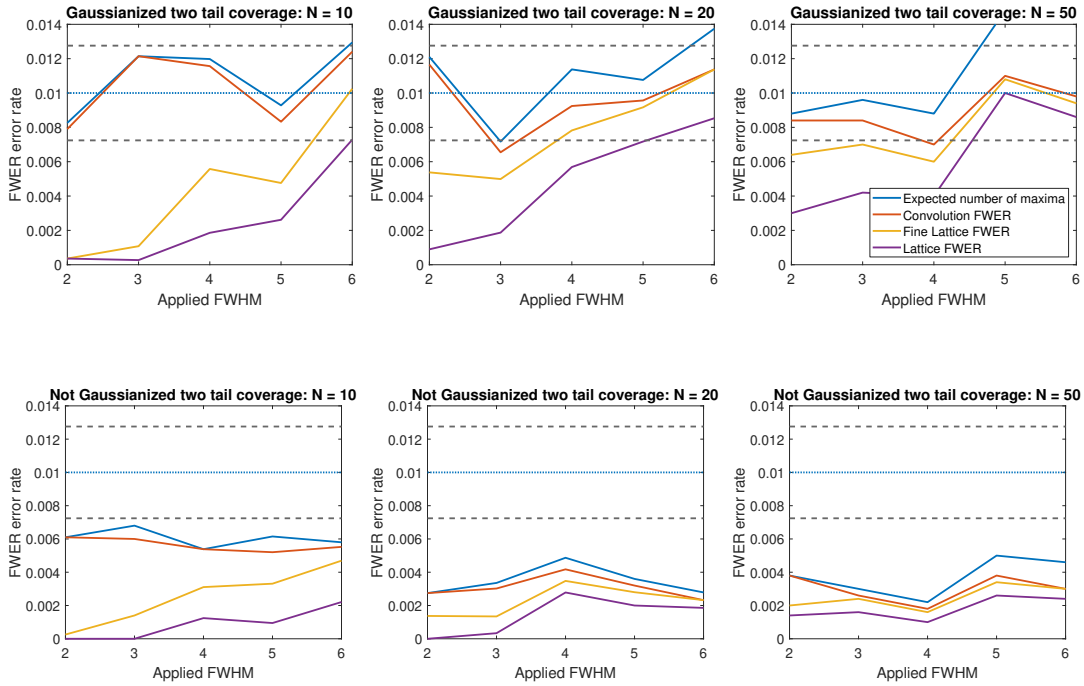



Figure 2.14: Resting state validation of the two tail FWER for $\alpha = 0.01$. Other than being two tailed the settings in each plot are the same as in Figure 2.9. Controlling at $\alpha = 0.01$ and performing a two-requires a higher threshold so the convolution FWER (shown in red) is closer to the nominal level than in the other corresponding Figures.

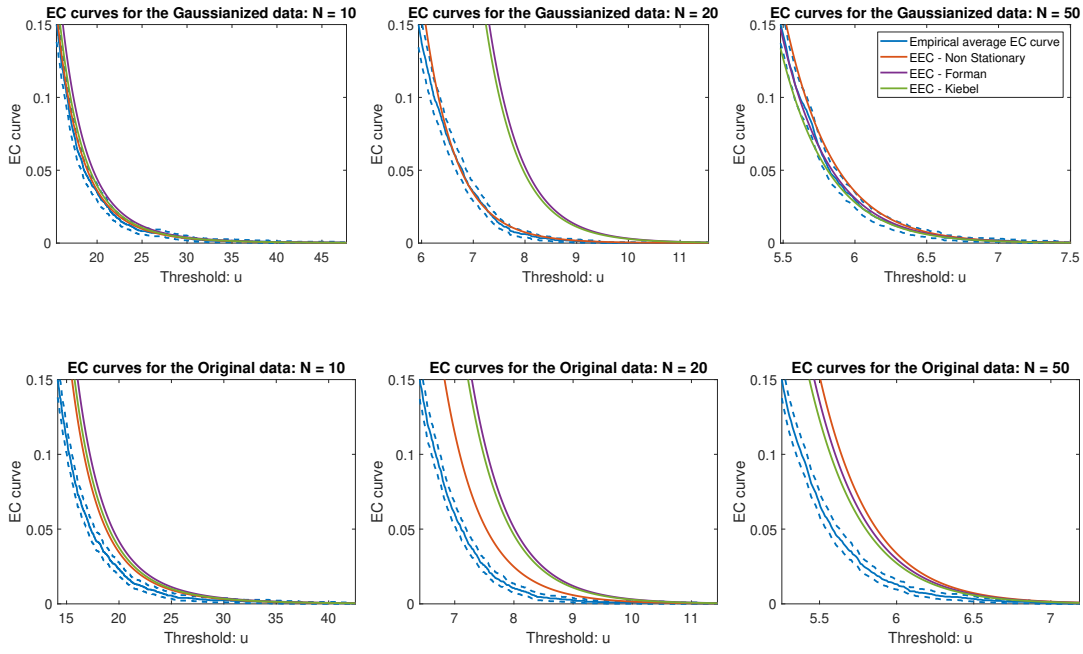



Figure 2.15: Comparing the expected and empirical tail EC curves in different settings using the resting state data for an applied smoothing level of 2 FWHM per voxel. The results are similar to Figure 2.10.

# Chapter 3

# Confidence regions for the location of peaks of a smooth random field

Samuel Davenport, Thomas E. Nichols, Dan Cheng,
Armin Schwartzman

**Abstract**

This article provides confidence regions for the location of peaks of the mean and standardized effect size given multiple realisations of a random process. We prove central limit theorems for the location of the maximum of mean and $t$-statistic random fields and use these to provide asymptotic confidence regions for peak mean and Cohen's $d$. Under the assumption of stationarity we develop Monte Carlo confidence regions for peaks of the mean that have a better finite sample coverage than the ones that are derived based on classical asymptotic normality.

*Keywords:* Confidence regions, peak location, asymptotic normality, maximum likelihood estimation, random fields, spatial statistics.

# 1 Introduction

Detecting the presence of significant peaks in a random field in order to determine areas of activation is an approach used in a number of fields including astrophysics, (Cheng et al., 2017), and neuroimaging (Chumbley and Friston (2009), Chumbley et al. (2010)). Inference is typically performed using results for zero mean random fields where the goal is to determine whether or not a particular peak is high enough to reject the null hypothesis that the mean is zero. These peak detection procedures were formalized in Schwartzman et al. (2011), Cheng and Schwartzman (2017): taking advantage of the peak height distributions derived in Cheng and Xiao (2016), Cheng (2017) and Cheng and Schwartzman (2015a) for stationary zero mean Gaussian random fields.

There has been a huge amount of work investigating properties of zero mean stationary random fields (see Adler (1981) for an overview), however when the mean is non-zero and the assumption of stationarity is dropped, it becomes much more difficult to prove interesting results. Simultaneous confidence bands for the signal have been derived in 1D using resampling approaches and the Gaussian Kinematic Formula (Telschow and Schwartzmann, 2020). Sommerfeld et al. (2018) derived asymptotic sets which provide confidence regions for the points of the signal that lie above a specified threshold (with application to climate data). This framework been applied in neuroimaging to find areas of high activation ((Bowring et al., 2019), (Bowring et al., 2020), Telschow et al. (2020a)). In the context of peaks, Cheng and Schwartzman (2015b) extended results on the distribution for the height of a peak (originally derived under stationarity Nosko (1969), Adler (1981)) to non-stationary and non-zero mean random fields.

We will derive asymptotic confidence regions for peak location, in non-zero mean

random fields, which are valid under non-stationarity. In the astrophysics setting this is important so that peaks can be correctly identified and their effect removed to give an uncorrupted sample of the cosmic microwave background radiation (Cheng et al., 2017). In the neuroimaging setting, as part of mapping the brain, this can be used to determine where the highest peaks of activation are most likely to lie. This is useful to enable studies to be combined correctly using meta-analysis (Eickhoff et al. (2012), Radua et al. (2012)) and so that activation can be compared across different studies. Neuroimaging meta analyses require confidence regions for peak location and unbiased estimates of the effect sizes at peaks (we provided a selective inference resampling based approach to estimating unbiased peak effect sizes in Davenport and Nichols (2020)).

The problem of estimating the location of a peak of a random field is mathematically very similar to that of finding the location of the maximum of the likelihood function for which standard asymptotic theory gives a central limit theorem (CLT). Dugué (1937), Cramer (1946) and Fisher (1925) showed that given i.i.d data, the maximum likelihood estimator is asymptotically normal with variance given by the inverse Fisher information. This result has been developed in many papers, ones of note include Bahadur (1967) who examined rates of convergence and Efron and Hinkley (1978) who investigated whether the observed Fisher information was better than the expected Fisher information as the estimate of the variance. The result has been extended to many other settings over the years. Bradley and Gart (1962), Hoadley (1971), Philippou and Roussas (1975) and Nordberg (1980) extended it to the independent but not identically distributed case and Wald (1948), Heijmans and Magnus (1986a), Sweeting (1980) and others extended it to work in the dependent data setting: see Heijmans and Magnus (1986b) for a full overview of the literature in this area. Amemiya (1985) extended these results to the more general framework of extremum estimators, see also

Hayashi (2000).

We will take advantage of the extremum estimator framework to derive CLTs for the local maxima of mean and $t$-statistic fields. To do so we derive the asymptotic distribution of the derivative of a mean/$t$-statistic field (giving an exact form for this in the case that the underlying fields are Gaussian) and show that the scaled second derivative converges almost surely. Combining these results yields asymptotic confidence regions for the true peak location of the underlying mean and Cohen's $d$. In the finite sample these results inevitably do not perform as well (see Braunstein (1992) for a discussion of the application of these types of asymptotic results in the finite sample). To help solve this problem, under an assumption of stationarity, we will use the joint distribution between the first and second derivatives at the peak to obtain confidence regions for peaks of the mean which have better coverage than the asymptotic confidence regions that are obtained from the CLT.

The structure of this paper is as follows. Section 2 defines mean and $t$-statistic fields, sets out required assumptions on the component fields and proves sufficient conditions for derivative exchangeability. In Section 3 we prove identifiability results: that the number of observed peaks in a sufficiently small ball around each local maximum and minimum converges to 1 and that away from critical points the number of maxima converges to zero. Section 4 states results on the asymptotic normality of the peak location estimators for mean and $t$-statistic fields as well as deriving the asymptotic distribution of the derivative of a $t$-field. There we also describe how to provide improved confidence regions for the location of the mean in a stationary random field. Section 5 demonstrates that the confidence regions achieve the correct coverage in 1D simulations and applies them to obtain confidence intervals for the peak of a 1D MEG power spectrum. Section 8 provides proofs of the theorems.

Simulations and peak inference were conducted using the RFTtoolbox: (`https://github.com/sjdavenport/RFTtoolbox`).

# 2 Model Set-up and Assumptions

## 2.1 Notation

Throughout we will take $(Y_n)_{n\in\mathbb{N}}$ (where $\mathbb{N}$ denotes the set of positive integers) to be i.i.d random fields on some bounded domain $S \subset \mathbb{R}^D$, $D \in \mathbb{N}$ and write $Y_n = \mu + \sigma\epsilon_n$ for some bounded functions $\mu, \sigma : S \longrightarrow \mathbb{R}$, $\sigma > 0$ and zero mean, variance 1 i.i.d random fields $(\epsilon_n)_{n\in\mathbb{N}}$. For $r > 0$ and $s \in S$, we define $B_r(s)$ to be the $D$-dimensional ball of radius $r$ that is centred at $s$ and will drop the subscript $r$ when wanting to describe a ball of arbitrary radius. Throughout we will perform operations on random fields (such as addition, multiplication, division) pointwise. Given $N \in \mathbb{N}$ samples we define the sample mean field as:

$$\hat{\mu}_N = \frac{1}{N}\sum_{n=1}^{N} Y_n \tag{3.1}$$

and the sample variance field to be

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

We will refer to the fields $Y_n$ as the **component fields**. We can then define the $t$-statistic field to be

$$T_N = \frac{\sqrt{N}\hat{\mu}_N}{\hat{\sigma}_N} = \frac{\sqrt{N}\mu + Z_N}{\sqrt{\frac{1}{N-1}V_N}} \tag{3.2}$$

where $Z_N := \frac{\sigma}{\sqrt{N}}\sum_{n=1}^{N}\epsilon_n$ and $V_N := \sum_{n=1}^{N}(Y_n - \hat{\mu}_N)^2$. If the component fields are Gaussian then this field is a non-central $t$-field with $N-1$ degrees of freedom. We

define the Cohen's $d$ field to be $d_N = T_N/\sqrt{N}$. As $N$ tends to infinity the Cohen's $d$ field converges uniformly almost surely to $\mu/\sigma$ (see Lemma 3.6). When they are defined for each $s \in S$, let

$$\Lambda(s) := \text{cov}(\nabla^T Y_1(s)) \text{ and } \Gamma(s) := \mathbb{E}\big[(Y_1(s) - \mu)(\nabla Y_1(s))^T\big].$$

We will use $\xrightarrow{d}$ and $\xrightarrow{\mathbb{P}}$ to denote convergence in distribution/probability respectively. And use $\overset{d}{\Longrightarrow}$ and $\overset{\mathbb{P}}{\Longrightarrow}$ to denote uniform convergence in distribution/probability respectively.

## 2.2 Derivative Exchangeability

In what follows we will need to be able to exchange expectation and differentiation. In the context of random fields this can be done so long as the fields have the following property.

**Definition 2.1.** We say that a random field $f : S \longrightarrow \mathbb{R}^{D'}$, some $D' \in \mathbb{N}$, is $L_1-$**Lipschitz at s** $\in \text{int}(S)$ (where $\text{int}(S)$ denotes the interior of $S$) if there exists an integrable real random variable $L$ and some ball $B(s) \subset S$ centred at $s$ such that

$$\|f(t) - f(s)\| \leq L\|t - s\| \text{ for all } t \in B(s).$$

We define $f$ to be $L_1-$**Lipschitz** on a subset $S' \subset S$ if it is $L_1-$Lipschitz at $s$ for all $s \in \text{int}(S')$. If $S' = S$ then we will not specify the subset.

More generally we will say that a random field $f$ on $S$ satisfies the **DE (derivative exchangeability) condition** at $s \in S$ if $\mathbb{E}[f(t)]$ is differentiable at $t = s$ and $\mathbb{E}[\nabla f(t)] = \nabla\mathbb{E}[f(t)]$, i.e. such that we can exchange the integral and derivative. We say that $f$ satisfies the **DE condition** on $S$ if this holds for all $s \in S$. Arguing as in

the proof of Telschow and Schwartzmann (2020)'s Lemma 4, in the following Lemma we show that it is sufficient for $f$ to be $L_1$-Lipschitz for this to hold.

**Lemma 2.2.** *(Expectation-derivative exchangeability.) Let $f : S \to \mathbb{R}^{D'}$ be an a.s. differentiable random field that is $L_1$-Lipschitz at $s \in S$. Then $f$ satisfies the DE condition at $s$.*

*Proof.* Let $e_i, i = 1, \ldots, D$, be the standard basis vectors in $\mathbb{R}^D$ and let $L, B(s)$ be the Lipschitz constant and the ball around $s$ on which the Lipschitz property holds. Then for $1 \leq i \leq D$ and $1 \leq j \leq D'$,

$$\mathbb{E}[(\nabla f)_{ij}] = \mathbb{E}\left[\frac{\partial f_j(s)}{\partial s_i}\right] = \mathbb{E}\left[\lim_{h\to 0} \frac{f_j(s + he_i) - f_j(s)}{h}\right].$$

We can thus apply the Dominated Convergence Theorem to obtain the result, using $L$ as a dominating function, since for $h$ small enough such that $s + he_i \in B(s)$,

$$\left|\frac{f_j(s + he_i) - f_j(s)}{h}\right| \leq \frac{\|f(s + he_i) - f(s)\|}{|h|} \leq \frac{L\|he_i\|}{|h|} = L.$$

□

Because of the mean value inequality, finiteness of the expected value of the supremum of the local derivative is a sufficient condition for $L_1$-Lipschitzness.

**Lemma 2.3.** *Let $f$ be a random field on $S$ which is a.s. differentiable on some ball $B(s)$, centred at $s \in S$, and suppose that*

$$\mathbb{E} \sup_{t\in B(s)} \|\nabla f(t)\| < \infty.$$

*Then $f$ is $L_1$-Lipschitz at $s$.*

The $L_1$-Lipschitz condition is also satisfied by the broad class of convolution random fields. These fields have been used to control the familywise error rate using the

Gaussian Kinematic Formula (Telschow et al. (2020b), Davenport et al. (2021)) and are important because they bridge the gap between data on a lattice and theory describing continuous random fields. They are defined as follows.

**Definition 2.4.** Given observations $X$ on a lattice $\mathcal{V} \in \mathbb{R}^D$ and some continuous kernel function $K : \mathbb{R}^D \to \mathbb{R}$, we define the convolution field $Y : S \to \mathbb{R}$ sending $s \in S$ to

$$Y(s) = \sum_{l \in \mathcal{V}} K(s-l)X(l).$$

**Proposition 2.5.** *Let $Y$ be a $D$-dimensional convolution field on $S$ generated from observations $X$ on a finite lattice $\mathcal{V}$ and using a kernel $K$ which is Lipschitz with constant $c$. If $\mathbb{E}[|X(l)|] < \infty$ for all $l \in \mathcal{V}$, then $Y$ is $L_1$-Lipschitz. In particular, if $K$ is differentiable then $Y$ satisfies DE conditions on $S$.*

*Proof.* For $s, t \in S$,

$$\begin{aligned} |Y(s) - Y(t)| &\leq \sum_{l \in \mathcal{V}} |K(s-l) - K(t-l)||X(l)| \\ &\leq c \sum_{l \in \mathcal{V}} |X(l)| \|s-t\| = \left( c \sum_{l \in \mathcal{V}} |X(l)| \right) \|s-t\|. \end{aligned}$$

The result follows by taking $c \sum_{l \in \mathcal{V}} |X(l)|$ as the Lipschitz constant and applying Lemma 2.2. □

In what follows we will need to be able to exchange first and second derivatives as well as apply the functional strong law of large numbers (fSLLN) (Ledoux and Talagrand (2013) Corollary 7.10). To ensure that we can do this we will want to impose the following conditions on a random field $f : S \to \mathbb{R}$:

**Assumption 2.6.**

*a. $f$ is a.s. twice continuously differentiable and for all $s \in S$, var$(f(s))$ and cov$(\nabla^T f(s))$ are finite.*

*b.* $\mathbb{E}[\sup_{s\in S}|f(s)|], \mathbb{E}\big[\sup_{j,s\in S}|f_j(s)|\big]$ *and* $\mathbb{E}\big[\sup_{j,k,s\in S}|f_{jk}(s)|\big]$ *are all finite.*

*c.* $(i)\,\mathbb{E}[\sup_{s\in S} f^2(s)]$, $(ii)\,\mathbb{E}\big[\sup_{j,s\in S} f_j^2(s)\big]$ *and* $(iii)\,\mathbb{E}\big[\sup_{j,k,s\in S} f_{jk}^2(s)\big]$ *are finite.*

If $f$ satisfies Assumption 2.6 then the DE conditions hold for $f, f^2$ and their first derivatives on the interior of $S$ by Lemma 2.3 (and applying Cauchy-Swartz). In order for Assumption 2.6 to hold, it is sufficient that the field is a convolution field on a finite lattice with twice continuously differentiable kernel and finite observation expectation.

**Proposition 2.7.** *Let $Y$ be a convolution field defined as in Definition 2.4 and restricted to $S$. Suppose $var(X(l)) < \infty$ for each $l \in \mathcal{V}$ and that $K$ is $C^2$, then $Y$ satisfies Assumption 2.6.*

*Proof.* 2.6a follows easily as the lattice is finite, the variance is finite at each point and $K$ is $C^2$. For 2.6b,c note that, as $K$ is $C^2$, we may assume that the absolute value of $K(s-l)$ and its derivatives are bounded over all $s \in S$ and $l \in \mathcal{V}$ by compactness, as $S$ is bounded and $\mathcal{V}$ is finite. Let $K^*$ be an upper bound on $|K|$, then for $s \in S$,

$$|Y(s)| = \left|\sum_{l\in\mathcal{V}} K(s-l)X(l)\right| \leq K^* \sum_{l\in\mathcal{V}} |X(l)|$$

and

$$\begin{aligned}\big|Y(s)^2\big| &= \left|\sum_{l,l'\in\mathcal{V}} K(s-l)X(l)K(s-l')X(l')\right| \\ &\leq \sum_{l,l'\in\mathcal{V}} |K(s-l)K(s-l')||X(l)X(l')| \leq (K^*)^2 \sum_{l,l'\in\mathcal{V}} |X(l)X(l')|.\end{aligned}$$

so 2.6b,c follow for $Y$. Similar arguments hold for its derivatives. □

**Remark 2.8.** *If a random field $f$ is $L_1$-Lipschitz on $S$ then there exists an integrable random variable $L$ such that (given any $s \in S$)*

$$\left|\sup_{t\in S} f(t) - f(s)\right| \leq L \sup_{t\in S}\|t-s\| = L\mathrm{diam}(S).$$

*In particular, as $S$ is bounded,*

$$\mathbb{E}|\sup_{t\in S} f(t)| \leq \mathbb{E}[L]\text{diam}(S) + \mathbb{E}|f(s)| < \infty.$$

# 3 Local convergence of the number of peaks

Our first set of results relate to the number of peaks that lie within small regions around each critical point. We show that in the signal plus noise model (3.3) the number of peaks of the observed field, in small regions around peaks of the signal, converges in probability to 1. Furthermore, the number of critical points of the field, away from critical points of the signal, tends to 0 in probability. To do so we introduce Assumption 3.2 and require that the first two derivatives of the noise converge uniformly in probability to zero as $N \to \infty$. These results are important because they show that given a large enough sample size we can be sure that any observed peak is a true peak of the underlying signal, providing peak identifiability. In order to make the notion of a peak rigorous we have the following definition.

**Definition 3.1.** Given $f : S \to \mathbb{R}$, we say that $s \in S$ is a **critical point** of $f$ if $\nabla f(s) = 0$. Given a critical point $s$, we define $s$ to be a **local maximum** of $f$ if there is some $r > 0$ such that $f(s) = \sup_{t\in B_r(s)} f(t)$ and call a local maximum $s$ **non-degenerate** if $\nabla^2 f(s) \prec 0$ (we write $A \prec 0$ iff $A$ is a negative definite matrix). Local minima (and their non-degeneracy) are defined similarly.

## 3.1 Identifiability

In order to prove identifiability (Proposition 3.4) we need to make some assumptions. To do so, consider the general signal plus noise model

$$\hat{\gamma}_N = \gamma + \eta_N, \tag{3.3}$$

for $N \in \mathbb{N}$, where $\gamma : S \to \mathbb{R}$ is a fixed function and $\eta_N$ are $D$-dimensional random fields on $S$. Unlike in the standard signal plus noise model, we do not require the $\eta_N$ to have mean zero. Instead, we will impose specific conditions on $\gamma$ and $\eta_N$ and use this more general signal plus noise model to describe both mean and Cohen's $d$ fields. We will want $\gamma$ to be sufficiently nice in terms of smoothness and will want the derivatives of $\eta_N$ to converge to zero in probability as $N \to \infty$. In particular we will impose the following conditions on $\gamma$ to ensure identifiability.

**Assumption 3.2.**

*(a)* $\gamma$ *is* $C^2$.

*(b)* $\gamma$ *has* $J \in \mathbb{N}$ *critical points at locations* $\theta_1, \dots, \theta_J \in S$, *such that for each* $j = 1, \dots, J$ *there exist non-overlapping compact balls* $B_j \subset S$ *with radii* $\delta_j$ *such that* $\theta_j \in int(B_j)$ *and*

$$C := \inf_{t \in S \setminus \cup_j B_j} \|\nabla \gamma(t)\| > 0.$$

*(c) Let* $B = \bigcup_j B_j$ *and let* $P_{\max}$ *and* $P_{\min}$ *be the subsets of* $\{1, \dots, J\}$ *corresponding to the non-degenerate local maxima and minima of* $\gamma$, *respectively. Define*

$$B_{\max} = \bigcup_{j \in P_{\max}} B_j \text{ and } B_{\min} = \bigcup_{j \in P_{\min}} B_j.$$

*Assume that*

$$D_{\max} := -\sup_{t \in B_{\max}} \sup_{\|x\|=1} x^T \nabla^2 \gamma(t) x > 0$$

*and that*

$$D_{\min} := -\sup_{t \in B_{\min}} \sup_{\|x\|=1} x^T \nabla^2 \gamma(t) x < 0.$$

On $S \setminus B$, Assumption 3.2b ensures that $\gamma$ is not flat, as critical points (albeit with low magnitudes) always have a chance of manifesting in regions where the signal is flat no matter how low the variance of the noise. Assumption 3.2c provides bounds on the eigenvalues of the Hessian of $\gamma$, within the specified regions, and ensures that, for $j = 1, \ldots, J$, if $\theta_j$ is a local maximum of $\gamma$, then $\nabla^2 \gamma(s) \prec 0$ for all $s \in B_j$ and so $\theta_j = \text{argmax}_{t \in B_j} \gamma(t)$ (and similar uniqueness holds if $\theta_j$ is a local minimum). These bounds are needed to show that, with high probability, only one peak of $\hat{\gamma}_N$ is found within each region corresponding to a maximum or a minimum. If $\gamma$ is $C^2$ then the conditions on $D_{\max}$ and $D_{\min}$ follow immediately from peak non-degeneracy and by choosing $B_j$ to be sufficiently small.

The sample mean field described in equation (3.1) can be written as

$$\hat{\mu}_N = \mu + \frac{\sigma}{N} \sum_{n=1}^{N} \epsilon_n$$

and so fits the signal plus noise model very naturally. The Cohen's $d$ field can also be written in this form simply by taking $\gamma = \frac{\mu}{\sigma}$ and $\eta_N = \left(d_N - \frac{\mu}{\sigma}\right)$. We will respectively require that $\mu$ and $\frac{\mu}{\sigma}$ satisfy Assumption 3.2. Moreover, we will show in Section 3.3.1 that $\frac{\sigma}{N} \sum_{n=1}^{N} \epsilon_n$ and $\left(d_N - \frac{\mu}{\sigma}\right)$ and their derivatives converge to zero uniformly in probability.

## 3.2 Peak Convergence

We will require the following lemma which provides a bound for the probability that the derivative is non-zero in a signal plus noise model.

**Lemma 3.3.** *Given $S' \subset \mathbb{R}^D$, let $\gamma : S' \to \mathbb{R}$ be differentiable and suppose that $\hat{\gamma}$ is some estimate of $\gamma$ on $S'$ such that $\hat{\gamma} = \gamma + \eta$ for some random field $\eta$ on $S'$. Then*

$$\mathbb{P}(\inf_{t \in S'} \|\nabla \hat{\gamma}(t)\| > 0) \geq 1 - \mathbb{P}\bigg(\sup_{t \in S'} \|\nabla \eta(t)\| > C'\bigg),$$

*where $C' \leq \inf_{t \in S'} \|\nabla \gamma(t)\|$ is a lower bound on the norm of the derivative of $\gamma$ over $S'$.*

*Proof.*

$$\|\nabla \hat{\gamma}(t)\| = \|\nabla \gamma(t) + \nabla \eta(t)\| \geq \|\nabla \gamma(t)\| - \|\nabla \eta(t)\| \geq C' + \inf_{t \in S'}(-\|\nabla \eta(t)\|) \text{ and so}$$

$$\mathbb{P}(\inf_{t \in S'} \|\nabla \hat{\gamma}(t)\| > 0) \geq \mathbb{P}\bigg(\sup_{t \in S'} \|\nabla \eta(t)\| < C'\bigg) = 1 - \mathbb{P}\bigg(\sup_{t \in S'} \|\nabla \eta(t)\| > C'\bigg).$$

□

Using this lemma we can prove the following proposition on identifiability which shows that the number of local maxima/minima of a non-central random field in a region around the true peak converges to 1 in probability and that the number of critical points outside of $B$ converges to zero in probability. The proof adapts that of Cheng et al. (2017)'s Lemma A.1.

**Proposition 3.4.** *Suppose that $(\hat{\gamma}_N)_{N \in \mathbb{N}}$ is a sequence of random fields on $S$ such that $\hat{\gamma}_N = \gamma + \eta_N$ for some sequence of random fields $(\eta_N)_{N \in \mathbb{N}}$ such that $\nabla \eta_N \xrightarrow{\mathbb{P}} 0$ and differentiable $\gamma : S \to \mathbb{R}$ which satisfies Assumption 3.2b . Suppose that for each $N$, $\eta_N$ is a.s. differentiable, then as $N \longrightarrow \infty$,*

$$\mathbb{P}(\#\{t \in S \setminus B : \nabla \hat{\gamma}_N(t) = 0\} = 0) \longrightarrow 1.$$

*Additionally assume that all the conditions of Assumption 3.2 apply to $\gamma$, and that $\eta_N$ is a.s. $C^2$ with $\nabla^2\eta_N \overset{\mathbb{P}}{\Longrightarrow} 0$, and let*

$$M_N = \left\{t \in S : \nabla\hat{\gamma}_N(t) = 0 \text{ and } \nabla^2\hat{\gamma}_N(t) \prec 0\right\}$$

*be the set of non-degenerate local maxima of $\hat{\gamma}_N$. Then, as $N \longrightarrow \infty$, for each $B_j$ containing a non-degenerate local maximum of $\gamma$:*

$$\mathbb{P}(\#\{t \in M_N \cap B_j\} = 0) \longrightarrow 1.$$

*Proof.* See Section 8.1. □

Of course by symmetry an analogous result holds for the non-degenerate local minima.

**Remark 3.5.** *The condition that $\eta_N$ is a.s. $C^2$ holds in a number of situations. In particular it will hold for any convolution field with a $C^2$ kernel derived from a finite lattice. Alternatively, if $\eta_N$ is Gaussian then the conditions for this to hold are well studied: see for instance Adler (1981).*

## 3.3 Verifying convergence

The requirement in Proposition 3.4 that $\nabla\eta_N, \nabla^2\eta_N \overset{\mathbb{P}}{\Longrightarrow} 0$ can be shown to hold in a number of reasonable settings. In this section we will show that it holds for mean and $t$-statistic fields and in the context of the linear model. To demonstrate this we will need to able to exchange integration and differentiation and then apply the fSLLN for which we will require Assumption 2.6.

### 3.3.1 Mean and Cohen's $d$

Assume that the random fields $(Y_n)_{n\in\mathbb{N}}$ satisfy Assumption 2.6a,b. Then we can apply the fSLLN and derivative exchangeability to yield

$$\hat{\mu}_N - \mu \overset{a.s.}{\Longrightarrow} 0, \quad \nabla\hat{\mu}_N - \nabla\mu \overset{a.s.}{\Longrightarrow} 0 \text{ and } \nabla^2\hat{\mu}_N - \nabla^2\mu \overset{a.s.}{\Longrightarrow} 0.$$

where $\overset{a.s.}{\Longrightarrow}$ denotes uniform almost sure convergence over $S$. In the same setting but for Cohen's $d$ we have the following results.

**Lemma 3.6.** *Suppose that $(Y_n)_{n\in\mathbb{N}}$ satisfy Assumption 2.6b(i) and c(i), then $\hat{\sigma}_N^2 \overset{a.s.}{\Longrightarrow} \sigma^2$. If additionally $\inf_{s\in S}\sigma^2(s) > 0$, then as $N \longrightarrow \infty$,*

$$\frac{1}{\hat{\sigma}_N} \overset{a.s.}{\Longrightarrow} \frac{1}{\sigma} \text{ and so } \frac{\hat{\mu}_N}{\hat{\sigma}_N} \overset{a.s.}{\Longrightarrow} \frac{\mu}{\sigma}.$$

*Proof.* Applying Lemma 8.1 pointwise and the fSLLN multiple times (and scaling by $\frac{N}{N-1}$) it follows that $\hat{\sigma}_N^2 \overset{a.s.}{\Longrightarrow} \sigma^2$. If $\inf_{s\in S}\sigma^2(s) > 0$ then the inverse is well-defined and so the final results follow by the continuous mapping theorem and by noting that $\hat{\mu}_N \overset{a.s.}{\Longrightarrow} \mu$. □

**Proposition 3.7.** *Suppose that $(Y_n)_{n\in\mathbb{N}}$ satisfies Assumption 2.6 and that $\inf_{s\in S}\sigma(s) > 0$. Then as $N \longrightarrow \infty$,*

$$\nabla\left(\frac{\hat{\mu}_N}{\hat{\sigma}_N} - \frac{\mu}{\sigma}\right) \overset{a.s.}{\Longrightarrow} 0 \text{ and } \nabla^2\left(\frac{\hat{\mu}_N}{\hat{\sigma}_N} - \frac{\mu}{\sigma}\right) \overset{a.s.}{\Longrightarrow} 0.$$

*Proof.* By Lemma 2.3 we can exchange both first and second derivatives of $\sigma^2\epsilon_1^2$ with the expectation so that

$$\nabla\sigma(s)^2 = \nabla\mathbb{E}\big[(\sigma(s)\epsilon_1(s))^2\big] = \mathbb{E}\big[\nabla(\sigma(s)\epsilon_1(s))^2\big]$$

and

$$\nabla^2\sigma(s)^2 = \nabla^2\mathbb{E}\big[(\sigma(s)\epsilon_1(s))^2\big] = \mathbb{E}\big[\nabla^2(\sigma(s)\epsilon_1(s))^2\big].$$

As such, differentiating the expansion from Lemma 8.1 and applying the fSLLN multiple times it follows that $\nabla\hat{\sigma}_N^2 \overset{a.s.}{\Longrightarrow} \nabla\sigma^2$. As such, applying Lemma 3.6, we have

$$\nabla\hat{\sigma}_N = \nabla\left(\hat{\sigma}_N^2\right)^{1/2} = \frac{\nabla\hat{\sigma}_N^2}{2\hat{\sigma}_N} \overset{a.s.}{\Longrightarrow} \frac{\nabla\sigma^2}{2\sigma} = \nabla\sigma.$$

Similarly, $\nabla^2\hat{\sigma}_N \overset{a.s.}{\Longrightarrow} \nabla^2\sigma$. $\frac{1}{\hat{\sigma}_N} \overset{a.s.}{\Longrightarrow} \frac{1}{\sigma}$ by Lemma 3.6, so it follows that

$$\nabla\left(\frac{\hat{\mu}_N}{\hat{\sigma}_N} - \frac{\mu}{\sigma}\right) = \frac{\nabla\hat{\mu}_N}{\hat{\sigma}_N} - \nabla\left(\frac{1}{\hat{\sigma}_N}\right)\hat{\mu} - \frac{\nabla\mu}{\sigma} + \nabla\left(\frac{1}{\sigma}\right)\mu \overset{a.s.}{\Longrightarrow} 0.$$

The proof for the second derivative is similar. □

These results mean that Proposition 3.4 can be applied to mean and $t$-fields.

**Remark 3.8.** *It is also possible to prove a CLT for Cohen's d, see Telschow et al. (2020a) for further details.*

### 3.3.2 Linear Model

The linear model falls naturally into the signal plus noise framework (e.g. see Sommerfeld et al. (2018), Telschow et al. (2019)) and so the identifiability results of Proposition 3.4 can be shown to apply. To formalize this, let $p \in \mathbb{N}$ be the number of predictors and let $\mathcal{X}$ be a multivariate distribution on $\mathbb{R}^p$ with finite second moments, with density that is bounded above and such that if $x \sim \mathcal{X}$ then $\mathrm{cov}(x)$ is positive definite. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of independent random vectors in $\mathbb{R}^p$ such that $x_n \sim \mathcal{X}$ for all $n$ and for each $N \in \mathbb{N}$ set $X_N = (x_1 \dots x_N)^T \in \mathbb{R}^{N\times p}$. Define a sequence of random fields $(Y_n)_{n\in\mathbb{N}}$ on $S$ such that for $s \in S$,

$$Y_n(s) = x_n^T\beta(s) + \sigma(s)\epsilon_n(s) \tag{3.4}$$

where the $\epsilon_n$ are i.i.d real-valued mean-zero and variance-one random fields and $\beta(s) \in \mathbb{R}^p$. Let $Y^N = [Y_1, \dots, Y_N]^T$ and $\boldsymbol{\epsilon}_N = [\epsilon_1, \dots, \epsilon_N]^T \in \mathbb{R}^N$. Given some contrast vector

$w \in \mathbb{R}^p$ let $\gamma = w^T\beta$ and define

$$\hat{\gamma}_N = w^T\hat{\beta}_N = w^T(X_N^TX_N)^{-1}X_N^TY^N = w^T(X_N^TX_N)^{-1}X_N^T(X_N\beta + \boldsymbol{\epsilon}_N). \tag{3.5}$$

This model thus falls under the framework of (3.3) and we have the following result. (Treating the linear model as a signal plus noise model is relatively common, see e.g. Sommerfeld et al. (2018).)

**Proposition 3.9.** *Suppose that the $\epsilon_n$ are independent of the $x_n$ and satisfy Assumption 2.6, then*

$$\nabla(X_N^TX_N)^{-1}X_N^T\boldsymbol{\epsilon}_N \overset{a.s.}{\Longrightarrow} 0 \text{ and } \nabla^2(X_N^TX_N)^{-1}X_N^T\boldsymbol{\epsilon}_N \overset{a.s.}{\Longrightarrow} 0 \text{ as } N \longrightarrow \infty.$$

*In particular for $w \in \mathbb{R}^p$, as $N \longrightarrow \infty$,*

$$\nabla(\hat{\gamma}_N - \gamma) = w^T\nabla(X_N^TX_N)^{-1}X_N^T\boldsymbol{\epsilon}_N \overset{a.s.}{\Longrightarrow} 0 \text{ and } \nabla^2(\hat{\gamma}_N - \gamma) \overset{a.s.}{\Longrightarrow} 0.$$

*Proof.*

$$\frac{1}{N}X_N^T\nabla\boldsymbol{\epsilon}_N = \frac{1}{N}\Big[x_1, \cdots, x_N\Big]\nabla\boldsymbol{\epsilon}_N$$

and so for $i = 1, \ldots, p$ and $j = 1, \ldots, D$,

$$\left(\frac{1}{N}X_N^T\nabla\boldsymbol{\epsilon}_N\right)_{i,j} = \frac{1}{N}\sum_{n=1}^{N}\big((X_N)_{in}^T(\nabla\boldsymbol{\epsilon}_N)_{nj}\big) = \frac{1}{N}\sum_{n=1}^{N}(x_n)_i\frac{\partial\epsilon_n}{\partial t_j} \overset{a.s.}{\Longrightarrow} \mathbb{E}\left[(x_1)_i\frac{\partial\epsilon_1}{\partial t_j}\right]$$

as $N \longrightarrow \infty$. For each $i$, $(x_n)_i$ for $n = 1, \ldots, N$ are i.i.d as are $\frac{\partial\epsilon_n}{\partial t_j}$ for each $j$, so for all $i, j$, $(x_n)_i\frac{\partial\epsilon_n}{\partial t_j}$ are i.i.d for $n = 1, \ldots, N$. Additionally by independence and since the noise satisfies Assumption 2.6,

$$\mathbb{E}\left[\sup_{s\in S}\left|(x_n)_i\frac{\partial\epsilon_n}{\partial t_j}\right|\right] \leq \mathbb{E}|(x_n)_i|\mathbb{E}\left[\sup_{s\in S}\left|\frac{\partial\epsilon_n}{\partial t_j}\right|\right] < \infty$$

so the convergence above occurs by the fSLLN and the limit equals

$$\mathbb{E}\left[(x_1)_i \frac{\partial \epsilon_1}{\partial t_j}\right] = \mathbb{E}[(x_1)_i]\mathbb{E}\left[\frac{\partial \epsilon_1}{\partial t_j}\right] = 0.$$

Now $N(X_N^T X_N)^{-1} = (\frac{1}{N} X_N^T X_N)^{-1} \xrightarrow{a.s.} \Sigma^{-1}$ as $N \longrightarrow \infty$ (using Lemma 8.2) and so

$$\nabla (X_N^T X_N)^{-1} X_N^T \boldsymbol{\epsilon}_N = \left(\frac{1}{N} X_N^T X_N\right)^{-1} \left(\frac{1}{N} X_N^T \nabla \boldsymbol{\epsilon}_N\right) \xRightarrow{a.s.} 0 \text{ as } N \longrightarrow \infty.$$

The result for the second derivative follows similarly. Since

$$w^T \hat{\beta}_N - w^T \beta = w^T (X_N^T X_N)^{-1} X_N^T (X\beta + \epsilon) - w^T \beta = w^T (X_N^T X_N)^{-1} X_N^T \boldsymbol{\epsilon}_N,$$

the second set of results follow immediately. □

Thus, if we assume that $w^T \beta$ satisfies Assumption 3.2, Proposition 3.4 applies in this linear model setting.

# 4 Confidence Regions

In this section we will prove CLTs for peaks of the mean and Cohen's $d$ of i.i.d random fields. This approach takes advantage of the extremum estimator framework of Amemiya (1985), in particular we will make use of Theorem 4.1 of Shi (2011) which states conditions (which we will refer to as the **CLT conditions**) under which asymptotic normality occurs. Applying these extrema results in the neighbourhood of each peak in the case of sample mean fields, we relatively easily obtain the following theorem.

**Theorem 4.1.** *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d random fields satisfying Assumption 2.6a,b on $S$ that have mean $\mu$ which satisfies Assumption 3.2. For each $j = 1, \ldots, J$ corresponding to a non-degenerate local maximum of $\mu$, let $\hat{\theta}_{j,N} = \text{argmax}_{t\in B_j} \hat{\mu}_N(t)$,*

*then*

$$\sqrt{N}(\hat{\theta}_{j,N} - \theta_j) \xrightarrow{d} N(0, (\nabla^2\mu(\theta_j))^{-1}\Lambda(\theta_j)(\nabla^2\mu(\theta_j))^{-1})$$

*as* $N \longrightarrow \infty$.[1]

*Proof.* Sample mean fields fall under the $M$-estimation framework described in Hayashi (2000) and van der Vaart (1998). The proof proceeds by verifying that the CLT conditions hold. This approach uses a Taylor expansion about $\theta_j$ to give,

$$0 = \nabla\hat{\mu}_N(\hat{\theta}_{j,N}) = \nabla\hat{\mu}_N(\theta_j) + (\hat{\theta}_{j,N} - \theta_j)^T\nabla^2\hat{\mu}_N(\theta_j^*) \tag{3.6}$$

(for some $\theta_j^* \in B_{\|\theta_j - \hat{\theta}_{j,N}\|}(\theta_j)$). So the result follows by writing

$$\sqrt{N}(\hat{\theta}_{j,N} - \theta_j) = -\big(\nabla^2\hat{\mu}_N(\theta_j^*)\big)^{-1}\Big(\sqrt{N}\nabla^T\hat{\mu}_N(\theta_j)\Big),$$

applying the CLT and noting that $\nabla^2\hat{\mu}_N(\theta_j^*)$ converges to $\nabla^2\mu(\theta_j)$. In Section 8.3 we verify that the CLT conditions hold. Note that by symmetry we immediately obtain an analogous result for the location of local minima of $\mu$. □

In this section we will prove a corresponding result for the location of the peaks of $t$-statistic fields by showing that the CLT conditions hold in this context. These results will allow us to build asymptotic confidence regions for peak location. When our random fields are stationary we will show that these can be improved to provide better coverage in the finite sample by taking advantage of the joint distribution between the first and second derivatives.

[1] This result is analogous to the standard asymptotic normality result for the MLE. In that context, our random fields are $Y_n = \log g(X_n, s)$ for some pdf $g$ and random variables $(X_n)_{n\in\mathbb{N}}$ and parameter $s$ varying over some parameter set $S$. The fact that $g$ is a pdf allows the form of the variance to simplify because the Fisher information: $\text{var}(\nabla \log g(X_n, s)) = -\mathbb{E}\big[\nabla^2 \log g(X_n, s)\big]$ for each $s \in \text{int}(S)$. Such a simplification is not valid in general.

## 4.1 Cohen's $d$

Proving an analogous result to Theorem 4.1 for Cohen's $d$ is a little more complicated. Showing that the CLT conditions hold can be broken down into two main steps: proving a pointwise CLT for the distribution of the derivative of a $t$-statistic field and proving convergence of the Hessian in probability.

### 4.1.1 Distribution of the derivative of a $T$-field

If our component random fields are Gaussian then we can obtain finite sample distributions for the derivatives of $Y_n$. To do so we extend Worsley (1994)'s Lemma 5.1a to non-central and non-stationary $t$-fields to derive the distribution the gradient of the $t$-statistic field $T_N$ defined in equation (3.2). We start by assuming that the variance is constant, which simplifies the expressions, and then use this to prove the general result.

**Lemma 4.2.** *Let $(Y_n)_{n\in\mathbb{N}}$ be constant variance a.s. differentiable Gaussian random fields with differentiable mean. Assume that $Y_1^2$ satisfies the DE conditions, and that $\Lambda(s) = cov(\nabla^T Y(s))$ is finite for all $s \in S$. Then for each $s \in S$,*

$$\nabla T_N(s) =_d \left(\frac{N-1}{V_N(s)}\right)^{1/2} N\left(\sqrt{N}\nabla\mu(s), \left(1 + \frac{T_N(s)^2}{N-1}\right)\Lambda(s)\right)$$

*where $T_N(s) \sim t_{N-1}$ and $\frac{1}{N-1}V_N(s) \sim \sigma^2\chi^2_N$ are independent of the normal distribution used above.*

*Proof.* Define $Z_N, V_N$ as in equation (3.2) and note that for ease of notation we will drop the dependence on $s$ in what follows. Differentiating $T_N$, we have,

$$\nabla T_N = \left(\frac{N-1}{V_N}\right)^{1/2}\left(\sqrt{N}\nabla\mu + \nabla Z_N\right) - \frac{\sqrt{N}\mu + Z_N}{2V_N^{3/2}/\sqrt{N-1}}\nabla V_N \text{ and so}$$

$$\nabla T_N | Z_N, V_N \stackrel{d}{=} \left(\frac{N-1}{V_N}\right)^{1/2}\left(\sqrt{N}\nabla\mu + z_X\right) - \frac{\sqrt{N}\mu + Z_N}{2V_N^{3/2}/\sqrt{N-1}} 2V_N^{1/2} z_V$$

where $z_X$ and $z_V \stackrel{iid}{\sim} N(0,\Lambda)$. Note that the equality in distribution follows by Lemma 3.2 of Worsley (1994) which takes advantage of the independence between a constant variance Gaussian random field and its derivative and requires that the square of the fields satisfy the DE conditions (see Appendix 9.1 for a generalization of this to non-stationary $\chi^2$ random fields). $z_X$ and $z_V$ are independent of $Z_N$ and $V_N$ since the former are functions of the derivatives and the later are functions of the component fields. Thus

$$\begin{aligned}
\nabla T_N | Z_N, V_N &\stackrel{d}{=} \left(\frac{N-1}{V_N}\right)^{1/2}\left(\sqrt{N}\nabla\mu + N(0,\Lambda) + \frac{\sqrt{N}\mu + Z_N}{\sqrt{V_N}} N(0,\Lambda)\right) \\
&\stackrel{d}{=} \left(\frac{N-1}{V_N}\right)^{1/2} N\left(\sqrt{N}\nabla\mu, \left(1 + \frac{(\sqrt{N}\mu + Z_N)^2}{V_N}\right)\Lambda\right) \\
&\stackrel{d}{=} \left(\frac{N-1}{V_N}\right)^{1/2} N\left(\sqrt{N}\nabla\mu, \left(1 + \frac{T_N^2}{N-1}\right)\Lambda\right).
\end{aligned}$$

□

So as long as the fields are Gaussian and constant variance it follows that as $N \longrightarrow \infty$,

$$\nabla T_N - \frac{\sqrt{N}}{\sqrt{V_N/(N-1)}}\nabla\mu \stackrel{d}{\longrightarrow} N\left(0, \left(1 + \frac{\mu^2}{\sigma^2}\right)\frac{\Lambda}{\sigma^2}\right)$$

since $\frac{1}{N-1}V_N \stackrel{a.s.}{\longrightarrow} \sigma^2$ and $\frac{T_N^2}{N-1} \longrightarrow \frac{\mu^2}{\sigma^2}$ as $N \to \infty$. For Cohen's $d$ we thus have the following pointwise CLT as $N \longrightarrow \infty$,

$$\sqrt{N}\left(\nabla d_N - \frac{\nabla\mu}{\sqrt{V_N/(N-1)}}\right) \stackrel{d}{\longrightarrow} N\left(0, \left(1 + \frac{\mu^2}{\sigma^2}\right)\frac{\Lambda}{\sigma^2}\right).$$

When the mean is zero this does not depend on the variance, as we would expect, as neither does the $t$-statistic. When the mean is non-zero, the dependence on $\sigma$ is captured via Cohen's $d$ as the variance cancels out in the fraction $\Lambda/\sigma^2$. Let us now

drop the constant variance condition and write

$$T_N = \frac{\frac{1}{\sqrt{N}}\sum_{n=1}^{N} Y_n}{\left(\frac{1}{N-1}\sum_{n=1}^{N}(Y_n - \frac{1}{N}\sum Y_n)^2\right)^{1/2}} = \frac{\frac{1}{\sqrt{N}}\sum_{n=1}^{N} Y_n/\sigma}{\left(\frac{1}{N-1}\sum_{n=1}^{N}(Y_n/\sigma - \frac{1}{N}\sum Y_n/\sigma)^2\right)^{1/2}} \quad (3.7)$$

which is the $t$-statistic derived from component Gaussian random fields $Y_n' = Y_n/\sigma$ which are independent and have constant variance 1. We can thus apply the constant variance result to yield the following corollary.

**Corollary 4.3.** *Assume that the $(Y_n)_{n\in\mathbb{N}}$ are a.s. differentiable Gaussian random fields with differentiable mean and variance, $(V_N)_{N\in\mathbb{N}}$ are defined as in (3.2), and suppose that $\Lambda(s)$ is finite for all $s \in S$ and that $\frac{Y_1^2}{\sigma^2}$ satisfies the DE conditions. Then for each $s \in S$*

$$\nabla T(s) =_d \left(\frac{\sigma^2(s)(N-1)}{V_N(s)}\right)^{1/2} N\left(\sqrt{N}\nabla\frac{\mu(s)}{\sigma(s)}, \left(1 + \frac{T^2}{N-1}\right)\Lambda'(s)\right),$$

*where*

$$\Lambda'(s) := cov\left(\nabla^T \frac{Y_1(s)}{\sigma(s)}\right) = \frac{\Lambda(s)}{\sigma(s)^2} - \frac{\nabla^T\sigma(s)^2\Gamma(s)}{\sigma(s)^4} + \frac{\nabla^T\sigma(s)^2(\nabla\sigma(s)^2)}{4\sigma(s)^4}.$$

*Proof.* Applying Lemma 4.2 to the $t$-statistic from equation (3.7) we obtain the distributional result with $\Lambda'(s) = \mathrm{cov}\left(\nabla^T\frac{Y_1(s)}{\sigma(s)}\right)$. Dropping dependence on $s$ and expanding,

$$\begin{aligned}
\mathrm{cov}\left(\nabla^T\frac{Y_1}{\sigma}\right) &= \mathbb{E}\left[\left(\frac{\nabla^T Y_1}{\sigma} - \frac{Y_1}{2\sigma^3}\nabla^T\sigma^2\right)\left(\frac{\nabla^T Y_1}{\sigma} - \frac{Y_1}{2\sigma^3}\nabla\sigma^2\right)^T\right] \\
&= \frac{\mathbb{E}\left[(\nabla Y_1)^T\nabla Y_1\right]}{\sigma^2} - \frac{\nabla^T\sigma^2\mathbb{E}[Y_1(\nabla Y_1)]}{2\sigma^4} - \frac{\mathbb{E}\left[(\nabla Y_1)^T(Y_1\nabla\sigma^2)\right]}{2\sigma^4} + \frac{\mathbb{E}[Y_1^2](\nabla\sigma^2)^T\nabla\sigma^2}{4\sigma^6} \\
&= \frac{\Lambda}{\sigma^2} - \frac{\nabla^T\sigma^2\Gamma}{\sigma^4} + \frac{(\nabla\sigma^2)^T\nabla\sigma^2}{4\sigma^4}.
\end{aligned}$$

□

As such for all $s \in S$, as $N \to \infty$

$$\sqrt{N}\left(\nabla d_N(s) - \frac{\sigma(s)\nabla(\mu(s)/\sigma(s))}{\sqrt{V_N(s)/(N-1)}}\right) \xrightarrow{d} N\left(0, \left(1 + \frac{\mu^2(s)}{\sigma^2(s)}\right)\Lambda'(s)\right). \quad (3.8)$$

Note that if $\sigma^2$ is constant we recover the constant variance expression.

In practice it is likely that our random fields are not Gaussian. In this case there is no easy closed form for the finite sample distribution of the derivative of the $t$-statistic, however, it is still possible to derive an asymptotic limit for the distribution of the derivative. To do so we first require the following lemma.

**Lemma 4.4.** *Assume that $Y_1$ is $L_1$-Lipschitz and has unit variance and that $\Lambda$ is pointwise finite. Then*

$$\begin{pmatrix} \nabla^T Z_N \\ \nabla^T V_N/\sqrt{N} \end{pmatrix} = \sqrt{N}\begin{pmatrix} \frac{1}{N}\sum_{n=1}^N \nabla^T \epsilon_n \\ \nabla^T \hat{\sigma}_N^2 \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Lambda & 0 \\ 0 & 4\Lambda \end{pmatrix}\right)$$

*pointwise for all $s \in S$, as $N \longrightarrow \infty$.*

*Proof.* See Appendix 8.4.1. □

Using this lemma we can prove the following theorem which generalizes (3.8) to non-stationarity.

**Theorem 4.5.** *Suppose that $Y_1$ is a.s. differentiable with differentiable mean and variance such that $Y_1$ and $Y_1^2$ satisfy the DE conditions and $\Lambda = cov(\nabla^T Y_1)$ is finite over $S$. Then we have the following pointwise CLT on $S$:*

$$\sqrt{N}\left(\nabla d_N - \frac{\sigma\nabla(\mu/\sigma)}{\sqrt{V_N/(N-1)}}\right) \xrightarrow{d} N\left(0, \left(1+\frac{\mu^2}{\sigma^2}\right) cov\left(\nabla^T \frac{Y_1}{\sigma}\right)\right).$$

*Proof.* See Appendix 8.4.2. □

As we would expect the asymptotic variance in the CLT is the same as for when the component fields are Gaussian. Putting the pieces together, as with the mean, gives us the following theorem.

**Theorem 4.6.** *Suppose that $(Y_n)_{n\in\mathbb{N}}$ satisfy Assumption 2.6 with Cohen's d: $\frac{\mu}{\sigma}$ satisfying Assumption 3.2 and such that $\inf_{s\in S}\sigma^2(s) > 0$. For each $j = 1, \ldots, J$ corresponding to a maximum of $\frac{\mu}{\sigma}$, let $\hat{\theta}_{j,N} = \operatorname{argmax}_{t\in B_j} d_N(t)$, then*

$$\sqrt{N}(\hat{\theta}_{j,N} - \theta_j) \xrightarrow{d} N\left(0, \left(1 + \frac{\mu(\theta_j)^2}{\sigma(\theta_j)^2}\right)\left(\nabla^2\frac{\mu(\theta_j)}{\sigma(\theta_j)}\right)\Lambda'(\theta_j)\left(\nabla^2\frac{\mu(\theta_j)}{\sigma(\theta_j)}\right)^T\right).$$

*Proof.* The proof proceeds by using Theorem 4.5 and Proposition 3.7 to show that the CLT conditions hold, see Section 8.5 for details. □

## 4.2 Asymptotic Confidence Regions

Given these results we can obtain confidence regions for peak location which have the correct asymptotic coverage. For the mean, letting

$$\Sigma = (\nabla^2\mu(\theta_j))^{-1}\operatorname{cov}(\nabla^T Y_1(\theta_j))(\nabla^2\mu(\theta_j))^{-1}$$

and applying Theorem 4.1 we have

$$\sqrt{N}\Sigma^{-1/2}(\hat{\theta}_j - \theta_j) \sim N(0, I_D) \implies N(\hat{\theta}_j - \theta_j)\Sigma^{-1}(\hat{\theta}_j - \theta_j) \sim \chi^2_D.$$

Thus for $\alpha \in (0,1)$, letting $\chi^2_{D,1-\alpha}$ be the $1-\alpha$ quantile of the $\chi^2_D$ distribution,

$$\left\{\theta : N(\hat{\theta}_j - \theta)\Sigma^{-1}(\hat{\theta}_j - \theta) < \chi^2_{D,1-\alpha}\right\}$$

is a $(1-\alpha)\%$ asymptotic confidence region for $\theta_j$. In practice $\Sigma$ is unknown however taking $\hat{\Lambda}(\hat{\theta}_j)$ to be the sample covariance of $\nabla Y_1(\theta_j), \ldots, \nabla Y_N(\theta_j)$ and estimating the Hessian of the mean at $\theta_j$ by $\nabla^2\hat{\mu}_N(\theta_j)$ we obtain an asymptotic $(1-\alpha)\%$ confidence region as:

$$\left\{\theta : N(\hat{\theta}_j - \theta)\hat{\Sigma}^{-1}(\hat{\theta}_j - \theta) < \chi^2_{D,1-\alpha}\right\} \tag{3.9}$$

where $\hat{\Sigma} = (\nabla^2\hat{\mu}(\theta_j))^{-1}\hat{\Lambda}(\theta_j)(\nabla^2\hat{\mu}(\theta_j))^{-1}$. This confidence region performs well asymptotically however typically gives undercoverage in the finite sample (see Section 5) because, amongst other factors, it doesn't account for the extra variability that occurs because the second derivative hasn't converged. Assuming that our random fields are stationary it is in fact possible to obtain better finite sample coverage by taking account of the joint distribution between $\nabla\hat{\mu}_N$ and $\nabla^2\hat{\mu}_N$. From the Taylor expansion (3.6), for each $j$ corresponding to a maximum,

$$\hat{\theta}_{j,N} - \theta_j = \left(\nabla^2\hat{\mu}_N(\theta^*)\right)^{-1}\nabla^T\hat{\mu}_N(\theta_j). \tag{3.10}$$

Letting $\mathbb{V}$ denote the **vech** operation sending $D$ dimensional symmetric matrices to $\mathbb{R}^{D(D+1)/2}$,

$$\begin{pmatrix} \nabla^T\hat{\mu}_N(\theta_j) \\ \mathbb{V}(\nabla^2\hat{\mu}_N(\theta_j^*)) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mathbb{V}(\nabla^2\mu_N(\theta_j^*)) \end{pmatrix}, \frac{1}{N}\begin{pmatrix} \Lambda(\theta_j) & \Delta(\theta_j,\theta_j^*) \\ \Delta(\theta_j,\theta_j^*)^T & \Omega(\theta^*) \end{pmatrix}\right)$$

where for $s,t \in S$, $\Delta(s,t) = \text{cov}(\nabla Y_1(s), \mathbb{V}(\nabla^2 Y(t)))$ and $\Omega(s) = \text{cov}(\mathbb{V}(\nabla^2 Y(s)))$. This distribution contains a number of unknown quantities so we cannot simulate directly from it. Instead we can estimate these quantities and simulate random variables from the distribution

$$N\left(\begin{pmatrix} 0 \\ \mathbb{V}(\nabla^2\hat{\mu}_N(\hat{\theta}_{j,N})) \end{pmatrix}, \frac{1}{N}\begin{pmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Omega} \end{pmatrix}\right) \tag{3.11}$$

where $\hat{\Lambda} = \widehat{\text{cov}}(\nabla Y(\hat{\theta}_{j,N}))$, $\hat{\Omega} = \widehat{\text{cov}}(\mathbb{V}(\nabla^2 Y(\hat{\theta}_{j,N})))$ and since we are assuming that our data is stationary, $\Delta = 0$[2]. We can take advantage of the fact that $\Lambda$ and $\Omega$ are constant over $S$ to obtain very good estimates of these quantities based on data from the whole image (rather than from just around the peak). Simulating from (3.11) and plugging into equation (3.10) we can build an approximate parametric distribution for

[2]Note that $\widehat{\text{cov}}$ denotes the estimate of the variance obtaining using $Y_1, \ldots, Y_N$.

$\hat{\theta}_{j,N} - \theta_j$, allowing us to obtain more accurate quantiles, and provide a confidence region that has improved finite sample coverage. Asymptotically, by continuity of the quantities involved, the distributions are equivalent however the Monte Carlo distribution has improved finite sample performance (see Section 5).

For peaks of Cohen's $d$, we can derive analogous confidence regions using Theorem 4.6. In this case, taking

$$\hat{\Sigma} = \left(1 + \frac{\hat{\mu}_N(\hat{\theta}_{j,N})^2}{\hat{\sigma}_N(\hat{\theta}_{j,N})^2}\right)\left(\nabla^2 \frac{\hat{\mu}_N(\hat{\theta}_{j,N})}{\hat{\sigma}_N(\hat{\theta}_{j,N})}\right)\hat{\Lambda}'(\hat{\theta}_{j,N})\left(\nabla^2 \frac{\hat{\mu}_N(\hat{\theta}_{j,N})}{\hat{\sigma}_N(\hat{\theta}_{j,N})}\right)^T$$

we obtain a $(1-\alpha)\%$ asymptotic confidence region via equation (3.9). Note that no simple analogue of the stationary Monte Carlo approach is available for Cohen's $d$.

# 5 Simulations and Data Application

We conduct simulations to evaluate the coverage of the confidence regions in practice. We demonstrate their validity as the sample size increases and investigate their performance relative to the shape of the signal and the smoothness of the noise. We verify that our methods have the correct asymptotic coverage in these settings and illustrate how they can be applied in practice to provide confidence regions for peak location. For our noise distributions we use 1D stationary and non-stationary Gaussian random fields as well as evaluating non-Gaussian settings. To illustrate the theory in practice, we apply it to real 1D data to provide confidence intervals for the locations of peaks in an MEG power spectrum.

The theory we have described above applies to situations where there are multiple peaks. However we assume that we are in a setting where the peaks are identifiable. As such, in our simulations testing the coverage, we consider noise distributed about a

single peak.

## 5.1 Coverage

We simulate data from the signal plus noise model around a single peak in a number of settings (varying the smoothness of the noise, the number of samples $N$ and the shape of the peak). For each setting, given $\alpha > 0$ we run $S \in \mathbb{N}$ simulations. For each simulation $s \in \{1, \ldots, S\}$ we calculate a $(1-\alpha)\%$ confidence region: $R_s^\alpha$ for the true location $\theta$ of a given peak of the mean/Cohen's $d$, as discussed in Section 4.2. Given this we define the **true coverage** to be

$$\mathbb{P}(\theta \in R_1^\alpha).$$

We can approximate this by the **empirical coverage**

$$\frac{1}{S}\sum_{s=1}^{S} 1[\theta \in R_s^\alpha],$$

where $1[\cdot]$ denotes the indicator function. This converges to the true coverage by the SLLN as $S \to \infty$. Since $R_s^\alpha$ is an asymptotic $(1-\alpha)\%$ confidence region the true coverage converges to $1-\alpha$ as $N \to \infty$.

## 5.2 Mean simulations

### 5.2.1 Stationary Gaussian Noise

Our first set of simulations consists of 1D stationary Gaussian noise about two different types of peaks. In this setting we can leverage the stationarity of the fields to obtain good estimates for the variance and $\Lambda$ because they are the same at every point. The peaks we use are sections of the pdfs of the Beta$(1.5, 3)$ and Beta$(1.5, 2)$ distributions.

The first peak is narrow and the second peak is wide relative to the smoothness of the noise. The peaks have been scaled to a domain consisting of 10 voxels and the noise is obtained by smoothing white noise (on the original 10 voxel lattice) with a Gaussian kernel, with a FWHM ranging from 3 to 7 FWHM per voxel, to obtain a convolution field $\epsilon$ (defined in Definition 2.4) for each realization. We add these fields to the mean to obtain simulations with $N \in \{20, 40, 60, 80, 100\}$ and repeat these 1000 times in each setting.

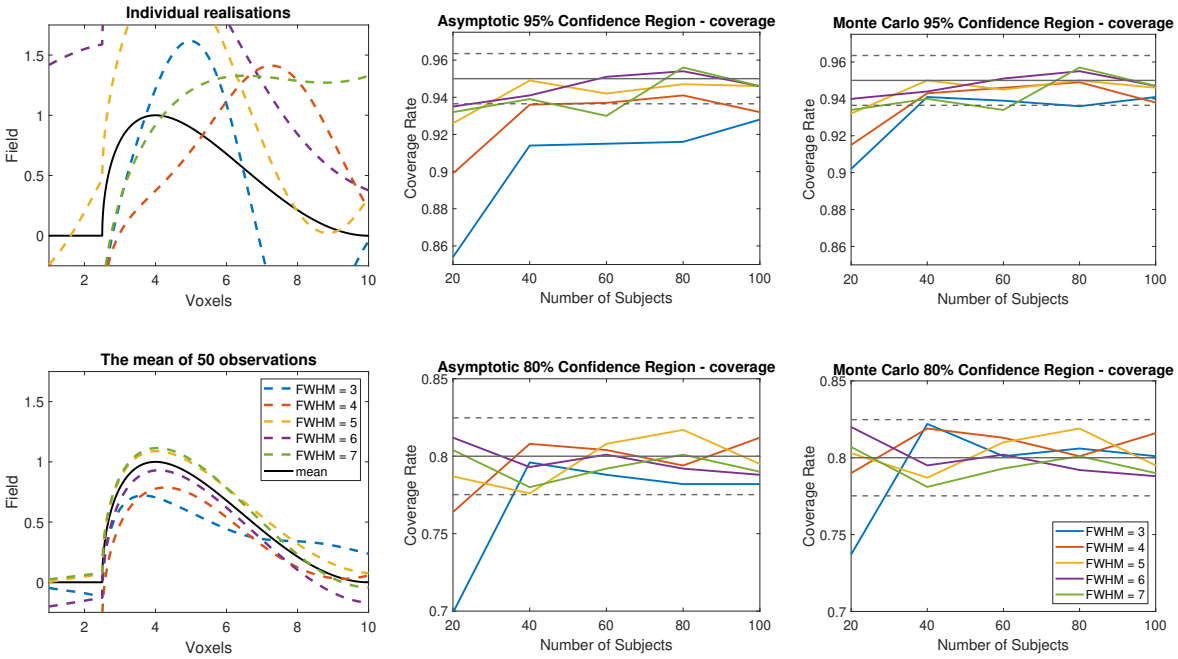



Figure 3.1: Coverage of confidence intervals for the maximum of the mean for stationary, variance-one Gaussian noise added to the narrow beta peak. The upper left panel contains the true mean and single realisations of the processes (which are the true mean plus smooth noise) and the mean of 50 i.i.d realisations is shown in the lower left panel. The upper centre and right panels display the coverage of 95% confidence intervals obtained using the asymptotic confidence regions (3.9) and the Monte Carlo ones described in Section 4.2. The lower centre and right panels display the same coverage results for 80% confidence intervals. Reasonable coverage is generally obtained for $N \geq 40$. From these graphs we can see that the Monte Carlo confidence regions have an improved finite sample coverage. As can be seen from the plots of the data, the peak of the means lie well within the image and so the simulations are not affected by edge effect issues.

The results (shown in Figures 3.1 and 3.2) illustrate that, as the number of subjects

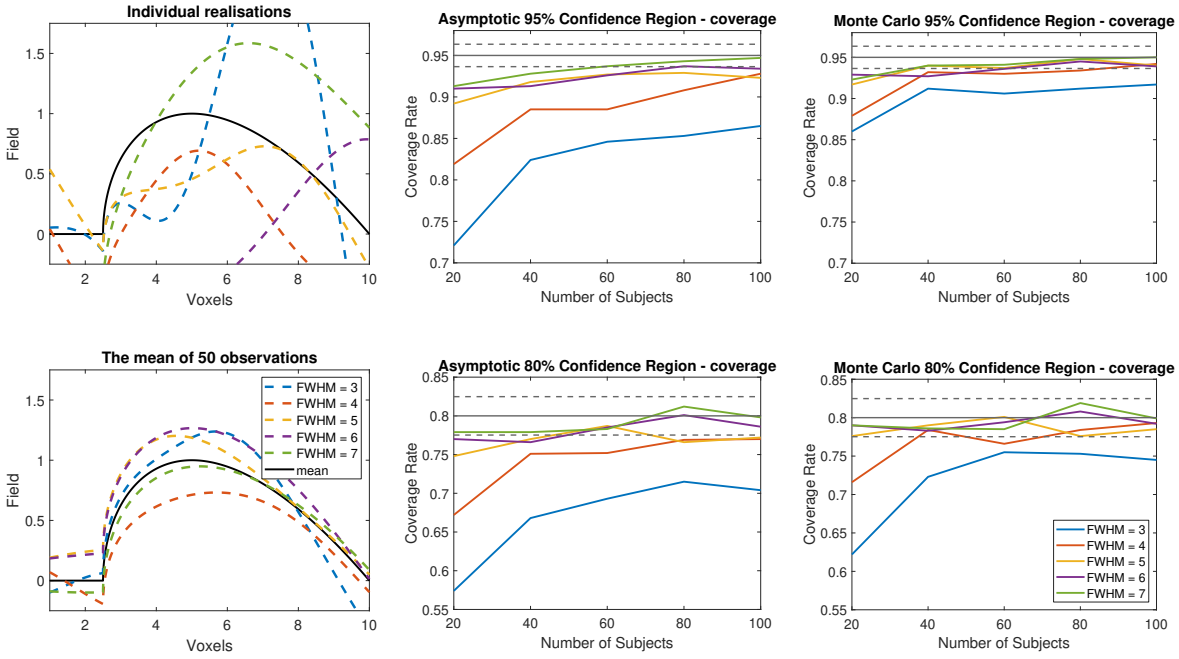



Figure 3.2: Coverage of confidence intervals for the maximum of the mean for stationary, variance 1 Gaussian noise added to the wider Beta peak, scaled such that Cohen's $d$ is 1 at the peak. The layout of the plots is the same as in Figure 3.1. In this setting a larger number of subjects (or higher smoothness) is required before the nominal coverage is obtained.

increases, the coverage converges to the desired level. The dotted lines, in these and all other corresponding Figures, give 95% confidence bands and are obtained using the normal approximation to the binomial distribution. The coverage of the Monte Carlo confidence regions, obtained using the approximation to the joint distribution: (3.11), also converges asymptotically. However they have a better performance in the finite sample, especially for lower smoothness levels.

The lower the FWHM of the noise relative to the shape of the peak, the larger the number of subjects that is needed to obtain the correct coverage. In many settings of interest high smoothness relative to the shape of the peak is a reasonable assumption, allowing us to obtain good coverage given available sample sizes. Here stationarity allows us to obtain better estimates of quantities involved, such as $\Lambda$, as we can average over the whole image. However, asymptotic coverage is achieved regardless of

stationarity for both the asymptotic and Monte Carlo methods.

### 5.2.2 $\chi^2$ Noise

Here we discuss a scenario where the noise is non-Gaussian. In particular if $\epsilon$ is a convolution field smoothed with a given FWHM (we use the same set of fields as in the previous setting) we generate noise fields here as $(\epsilon^2 - 1)/\sqrt{2}$. We add this noise to the same peaks as in the previous section. The results are similar, see Figures 3.6 and 3.7, but the methods take slightly longer to converge to the correct coverage as this is a more challenging setting.

## 5.3 Non-stationary noise

The asymptotic theory we have developed works under non-stationarity. To illustrate its performance in practice, we obtain non-stationary noise fields as follows. First we calculate a random $10 \times 10$ positive definite covariance matrix which we fix. We then generate values on the original (10-voxel) lattice from this covariance matrix and smooth to obtain convolution fields (these are non-stationary as the smoothness varies throughout the image). We add these to the Beta peaks, described above, to obtain our non-mean-zero realizations. The results are shown in Figures 3.8 and 3.9.

In this scenario, the narrow peak now requires a larger number of subjects than the wider peak before coverage is obtained. This occurs because the maxima of the peaks occur are different locations. As such the smoothness of the noise is different (due to the non-stationarity) at each peak: the noise is smoother near the wider peak than at the narrow peak. We illustrate this in Figure 3.3 by plotting the true $\Lambda(t)$, calculated using 20,000 realisations of each field, as $t$ varies across the image. We do this for the

fields generated using applied smoothness of 3 and 7 FWHM. The maximum of the wide peak occurs in a smoother (i.e. lower $\Lambda$) region than the maximum of the narrow peak. Under non-stationarity the estimates of the parameters for the distribution must be estimated locally and as such have higher variance than the corresponding estimates in the stationarity case. The quality of these estimates is thus particularly dependent on the smoothness and higher smoothness improves the coverage rate (as under stationarity). The effect of smoothness thus appears to dominate the effect of the shape of the peak. Note that under non-stationarity a Monte-Carlo distribution is not available.

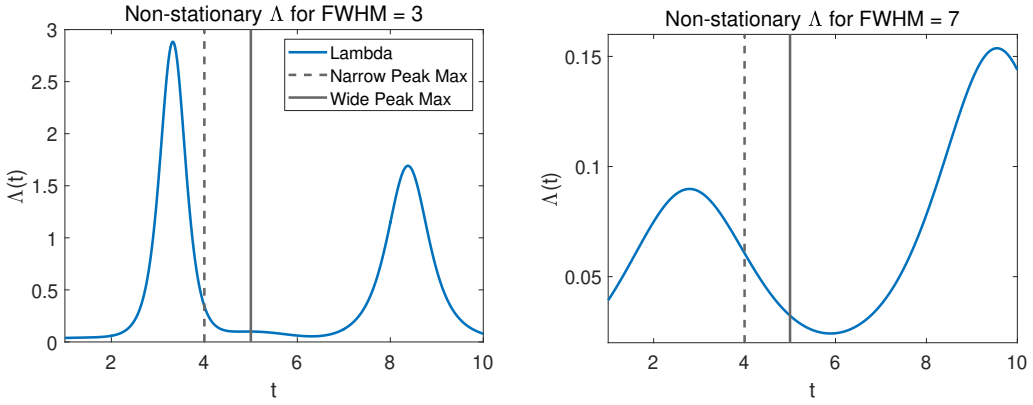



Figure 3.3: Calculating $\Lambda$ for the non-stationary Gaussian random fields as a function of space. The true maximum of the narrow and wide peaks occur at 4 and 5 voxels respectively. As such the smoothness is higher at the wider peak (as $\Lambda$ is lower).

## 5.4 Cohen's $d$ peak simulations

To illustrate the performance of our confidence regions for local maxima of the $t$-statistic we use the Gaussian stationary and non-stationary simulation settings as for the mean but instead infer on Cohen's $d$. For the $t$-statistic, changing the variance across the image is equivalent to changing the mean so, without loss of generality, we take the variance to be constant across the image. As can be seen from Figures 3.10

and 3.11 (the stationary results), the correct coverage appears to be obtained given sufficiently many subjects. Convergence is slower than for the mean and as before improves the higher the level of applied smoothness. The non-stationary results are shown in Figures 3.12 and 3.13. As with the non-stationary Gaussian simulations, the rate of convergence for the peaks is reversed. Note that no stationary Monte Carlo approach is available for Cohen's $d$.

## 5.5 Application: MEG power spectra

We will now apply our methods in a real data setting. We have 1-dimensional MEG data from 79 subjects (from a single MEG node) and for each subject have around 6 minutes of time series data sampled at a rate of 240Hz (see Quinn et al. (2019) for details on the sample and how it was collected). In order to infer on frequencies of interest in the data we turn each time series into a periodogram using Welch's method (Welch (1967), Solomon Jr (1991)). To do so for each subject $n = 1, \ldots, N$, let $X_n(t)$ denote its regularly sampled time series. Given a segment length $a$, we divide $X_n(t)$ sequentially into segments of length $a$ such that each overlaps by $\lfloor \frac{a}{2} \rfloor$ data points (and ignore the final segment if this does not divide evenly). We window each segment using a Gaussian kernel to eliminate cutoff effects and then take the Fourier transform of these windowed segments. Let $M_n$ denote the number of segments, and let $X_{n,m}$ denote the $m$th segment and let $W \in \mathbb{R}^a$ be a window of Gaussian weights. Then

$$\mathcal{D}(W \cdot X_{n,m}) = \mathcal{D}(W) \star \mathcal{D}(X_{n,m})$$

where $\mathcal{D}$ denotes the (periodic) discrete Fourier transform, $\cdot$ denotes pointwise multiplication and $\star$ denotes convolution. In particular for $x \in F = \left\{\frac{240k}{a} : k \in \mathbb{Z}\right\}$,

$$\mathcal{D}(W \cdot X_{n,m})(x) = \sum_{y \in F} K(x-y)\mathcal{D}(X_{n,m})(y) \tag{3.12}$$

where $K$, is the discrete Fourier transform of $W$ and is thus a Gaussian kernel. Since the Gaussian kernel is continuous (3.12) has a natural extension as a convolution field:

$$\mathcal{D}_{c,n,m}(s) = \sum_{y \in F} K(s-y)\mathcal{D}(X_{n,m})(y)$$

defined on $s \in [-120, 120]$ Hz (it is in fact defined on all $s \in \mathbb{R}$ but is periodic so we restrict to this bounded subset). In particular we can define $\mathcal{P}_{n,m}(s) = \|\mathcal{D}_{c,n,m}(s)\|^2$. Using Welch (1967)'s approach we obtain the power spectrum field

$$\mathcal{P}_n = \frac{10}{M_n} \sum_{m=1}^{M_n} \log_{10}(\mathcal{P}_{n,m})$$

(where addition is performed pointwise) defined on $s \in [-120, 120]$ Hz. We wish to infer on peaks in the power spectrum across subjects. Applying Welch's method, taking $a = 240$ (corresponding to a segment length of 1 minute), to the MEG data we obtain the mean shown in Figure 3.4. In this setting the time series have varying length but all consist of around 70,000 time points, meaning that $M_n$ is around 600 for each subject. This means that a large amount of averaging goes into calculating $\mathcal{P}_n$ which are thus effectively Gaussian random fields; by the functional CLT. Figure 3.4 shows that the noise is smooth relative to the mean so we expect to obtain good coverage in this setting. Applying our asymptotic confidence regions approach, we obtain a 95% confidence interval of (2.26, 2.32) Hz for the location of the highest peak of the mean.

In Figure 3.5 we plot the Cohen's $d$ obtained from the log power spectra. Using our approach, we calculate 95% confidence intervals of (10.82, 11.75) Hz and (5.58, 6.49)

Hz, for the locations of the top two peaks. Since the noise is smooth relative to the shape of the peaks we expect these confidence intervals to give good coverage for the true peak locations. Notably the standard error for both confidence intervals is similar, this occurs because the peaks have a similar shape and the smoothness of the noise is similar around each peak.

# 6 Discussion

In this paper we have derived CLTs for the locations of peaks of the mean and Cohen's $d$ of random fields and used these to obtain asymptotic confidence regions. We tested the coverage of the confidence regions in a variety of different settings for two different 1D peak shapes. We showed that, under stationarity, the coverage obtained can be improved by Monte Carlo simulation of the joint distribution between the first and second derivatives. We found that in all noise settings, wider peaks (relative to the noise) and rougher noise require a larger sample size $N$ before the correct coverage is obtained. This is because the wider the peak and the rougher the noise, the more the location of the maximum will be driven by peaks in the noise rather than peaks in the signal. We only considered 1D simulations and examples here, however the theory holds in any number of dimensions.

When the noise is non-stationary the parameters of the Monte Carlo distribution become more difficult to estimate and may not be the same at the location of the empirical peak and at the true peak. Nevertheless it would be interesting to determine scenarios where the Monte Carlo distribution or a variant can give better coverage even when the noise is non-stationary, such as (for instance) under local stationarity. Another way to improve the coverage would be to look at further terms in the Taylor

expansion and take advantage of their joint distribution which would also be Gaussian (at least asymptotically). If the coverage can be improved then this could useful in the context of maximum likelihood estimation in the finite sample.

It may also be of interest to develop non-parametric bootstrap style confidence regions. Consistency results for these have been developed in the context of M-estimation (see Cheng et al. (2010), Wellner and Zhan (1996), Lahiri (1992) and Abrevaya and Huang (2005) for details), and it would be interesting to extend these results to $t$-statistic fields. Future work could also investigate applying these techniques in larger dimensions and in other settings such as for fMRI data. In particular it It should also be possible to develop confidence regions for the locations of peaks of other random fields, such as $R^2$-fields, using similar techniques.

Our methods rely on identifiability of the peaks and we have shown that this occurs, given large enough sample sizes, under reasonable assumptions. In noisy scenarios there might be more than one peak about a true peak of the signal or even none at all. Moreover, if two peaks were found near to each other then it might be difficult to distinguish them. In small sample sizes it may be difficult to know whether identifiability can be assumed to hold. One heuristic that seems reasonable (in 1D) is to assume that identifiability has occurred if the 95% confidence interval (about a given peak) lies within the inflection points of the peak in the observed mean/Cohen's $d$. It would, however, be desirable to make this more precise and prove further results regarding it. One interesting possibility would be to instead consider the joint coverage over multiple peaks rather than the coverage at a single peak. It may also be interesting to explore other settings in which the peak locations are random (rather than fixed). Assuming fixed peak locations seems reasonable in our setting, however in practice it is likely that the true peak location and even the covariance structure of the noise could vary across

samples. In that case, as long as the data satisfies an fCLT, is seems feasible to prove similar results that would allow inference on the peaks of the population mean/Cohen's $d$.

One very interesting application of our results could be to improve coordinate based meta-analysis (see Eickhoff et al. (2009), Salimi-Khorshidi et al. (2009). These types of meta-analyses typically make use of a confidence region around the peaks reported across studies that represents a combination of within study and between study variation in the peak location. The within study variation is typically approximate and not theoretically justified. Our work enables confidence regions to be generated that have asymptotic theoretical coverage guarantees. Moreover we have shown that, for $0 < \alpha < 1$, the volume of the $(1-\alpha)\%$ confidence region shrinks at a rate of $N^{D/2}$ as the sample size $N$ increases. Using our results should thus allow practitioners to better account for the change in the size of the within study uncertainty as the sample size changes as well as enabling them to make more precise confidence statements and perform more exact meta-analyses.

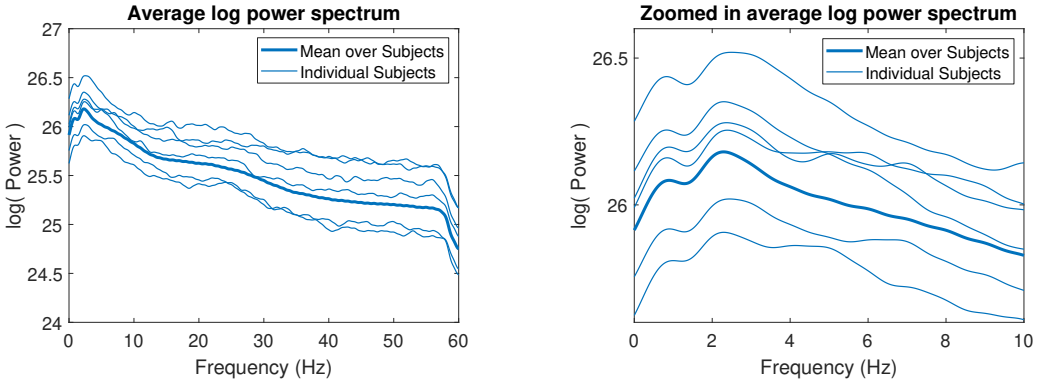



Figure 3.4: The average log MEG power spectrum random fields across subjects and individual subject log power spectra. This is shown from frequencies of 0 to 60 Hz on the left and from 0 to 10 Hz on the right. The individual spectra were calculated using Welch's method with a Gaussian smoothing window. The peak in the mean occurs at 2.29 $\pm 0.025$ Hz where the uncertainty is calculated using the asymptotic 95% confidence region. The noise is very smooth relative to the signal so we expect the confidence interval provide good coverage in this setting.

# 7 Further Figures

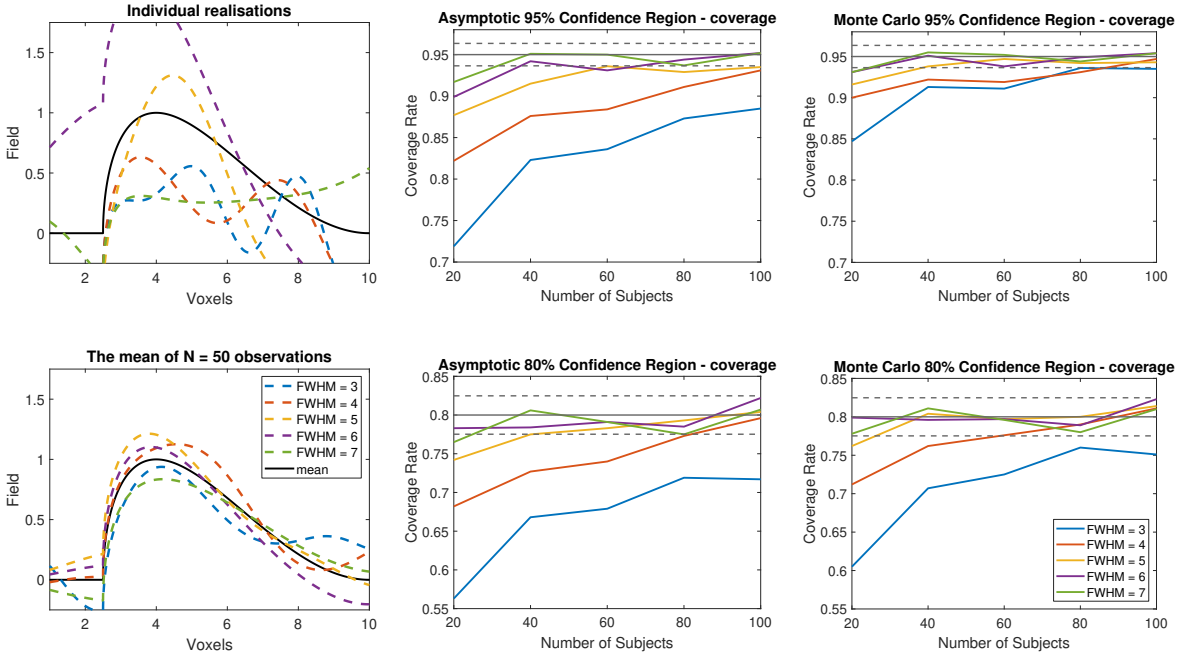



Figure 3.6: Coverage of confidence intervals for the maximum of the mean, obtained for stationary, variance 1, centred $\chi^2$ noise added to the narrow beta peak, scaled such that Cohen's $d$ is 1 at the maximum. The layout of the plots is the same as in Figure 3.1. The correct coverage levels are achieved given sufficiently many subjects.

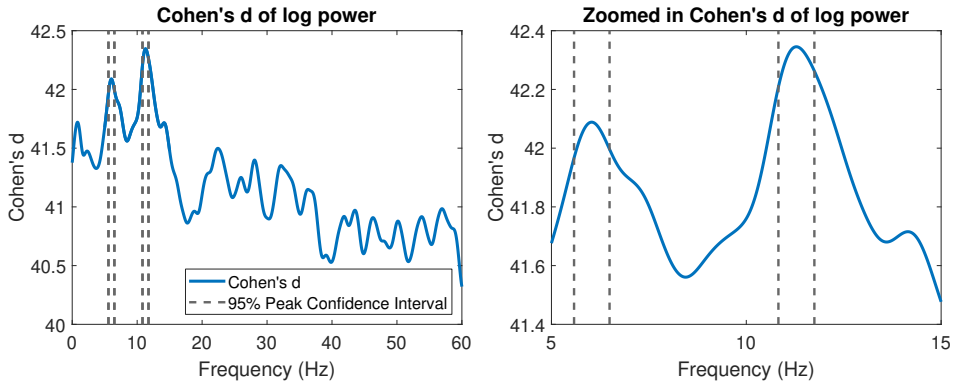



Figure 3.5: 95% confidence interval for the top two peaks of Cohen's $d$ of the log power spectrum. The peaks occur at $11.29 \pm 0.46$ Hz and $6.04 \pm 0.46$ Hz where the uncertainty is calculated using the asymptotic 95% confidence region. The individual spectra were calculated as in Figure 3.4.

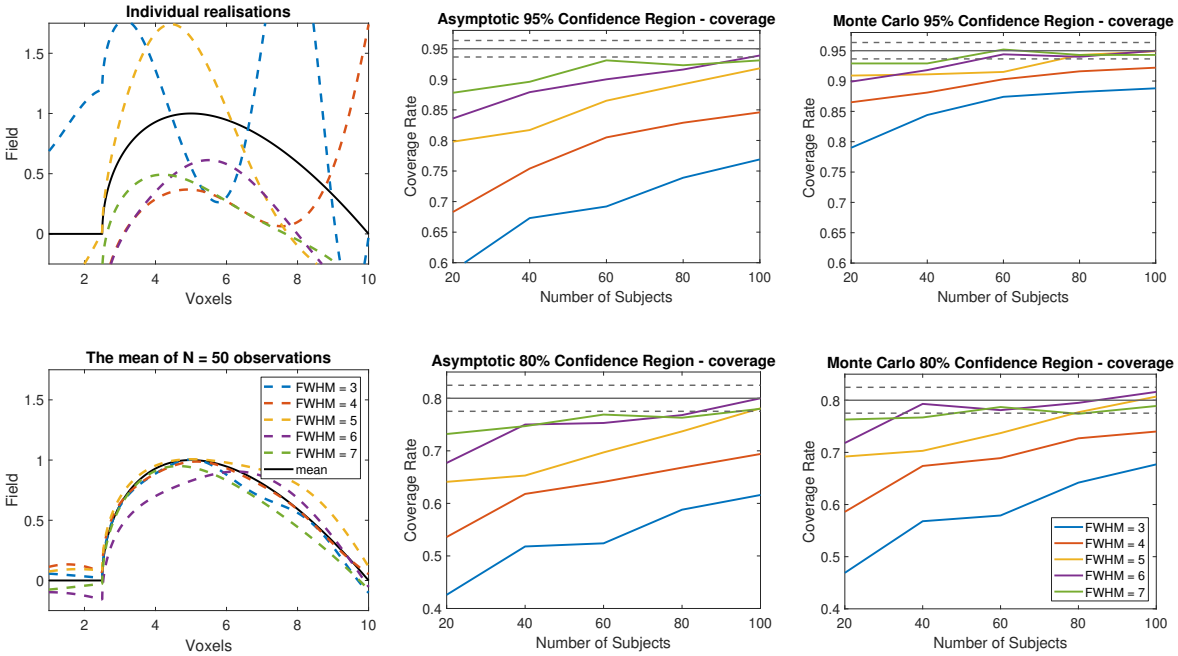



Figure 3.7: Coverage of confidence intervals for the maximum of the mean, obtained for stationary, variance 1, centred $\chi^2$ noise added to the wide beta peak, scaled such that Cohen's $d$ is 1 at the maximum. The layout of the plots is the same as in Figure 3.1. In this setting a larger number of subjects (or higher smoothness) is required before the nominal coverage is obtained.

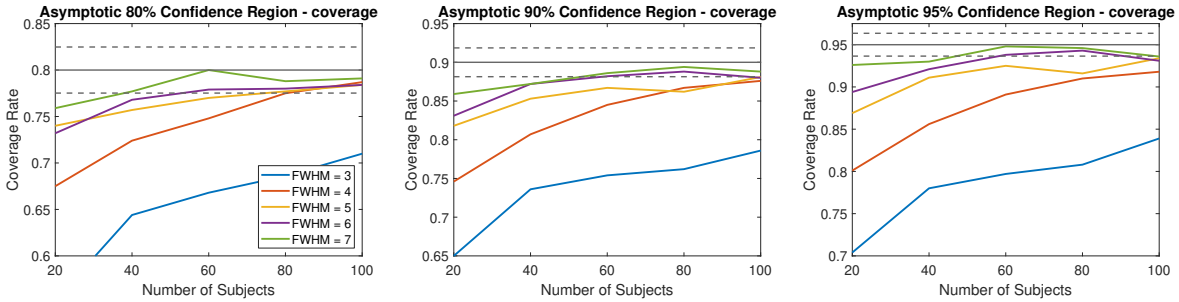



Figure 3.8: Coverage of confidence intervals for the maximum of the mean for variance 1, centred non-stationary Gaussian noise added to the narrow beta peak, scaled such that Cohen's $d$ is 1 at the maximum. Here the FWHM denotes the applied smoothing to the (already correlated) lattice data.

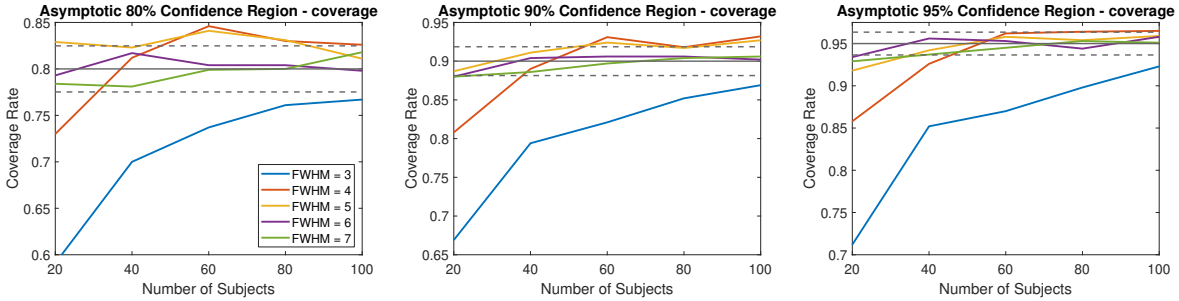



Figure 3.9: Coverage of confidence intervals for the maximum of the mean for non-stationary, variance 1 Gaussian noise added to the wide beta peak, scaled such that Cohen's $d$ is 1 at the peak. Here the FWHM denotes the applied smoothing to the (already correlated) lattice data.

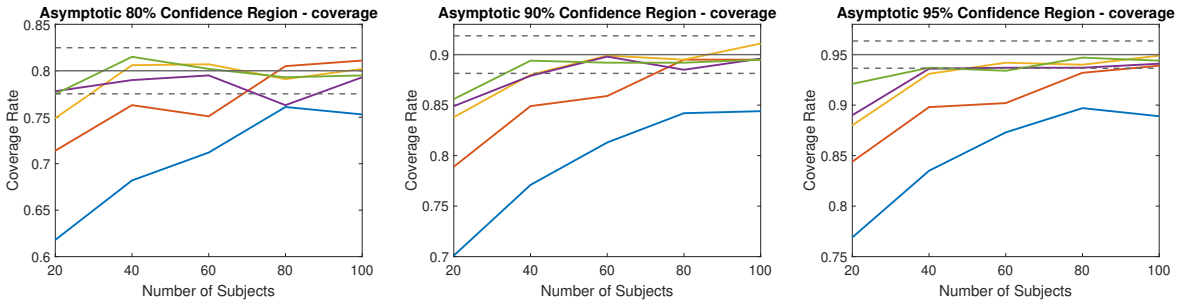



Figure 3.10: Coverage for the maximum of the $t$-statistic obtained for stationary, variance 1, centred Gaussian noise added to the narrow beta peak, scaled such that Cohen's $d$ is 1 at the maximum. In this setting a larger number of subjects (or higher smoothness) is required before the nominal coverage is obtained.

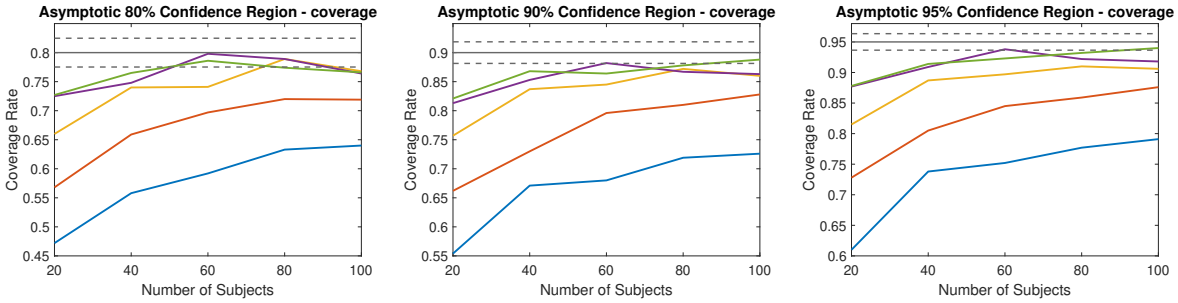



Figure 3.11: Coverage for the maximum of the $t$-statistic obtained for stationary, variance 1, centred Gaussian noise added to the wide beta peak, scaled such that Cohen's $d$ is 1 at the maximum. In this setting a larger number of subjects (or higher smoothness) is required before the nominal coverage is obtained.

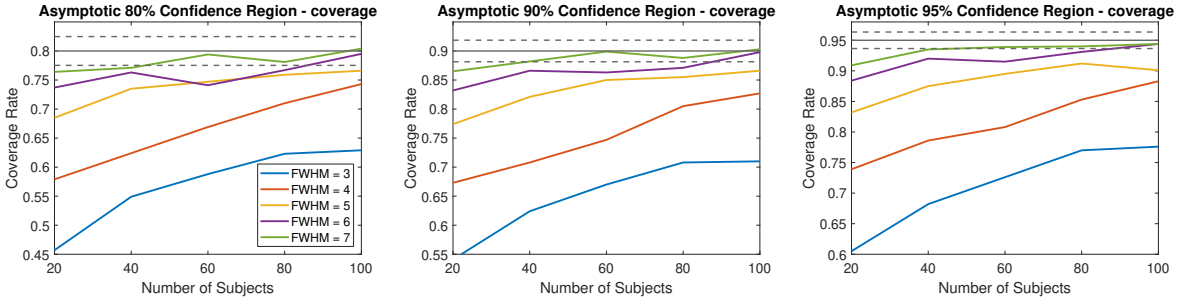



Figure 3.12: Plots of the coverage for the maximum of the $t$-statistic obtained for non-stationary, variance 1, centred Gaussian noise added to the narrow beta peak, scaled such that Cohen's $d$ is 1 at the maximum.

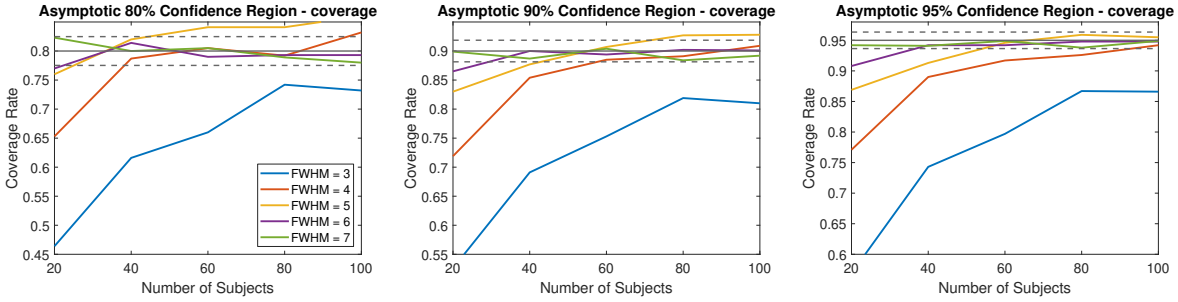



Figure 3.13: Plots of the coverage for the maximum of the $t$-statistic obtained for non-stationary, variance 1, centred Gaussian noise added to the wide beta peak, scaled such that Cohen's $d$ is 1 at the maximum.

# 8 Proofs

## 8.1 Proof of Proposition 3.4

*Proof.* The probability that $\hat{\gamma}_N$ has no critical points in $S \setminus B$ is greater than the probablity that $\inf_{t\in S\setminus B}\|\nabla\hat{\gamma}_N(t)\| > 0$ and by Lemma 3.3,

$$\mathbb{P}(\inf_{t\in S\setminus B}\|\nabla\hat{\gamma}_N(t)\| > 0) \geq 1 - \mathbb{P}\left(\sup_{t\in S\setminus B}\|\nabla\eta_N(t)\| > C\right).$$

By the continuous mapping theorem, $\nabla\eta_N \xrightarrow{\mathbb{P}} 0$ implies that $\|\nabla\eta_N\| \xrightarrow{\mathbb{P}} 0$. We have

$$\sup_{t\in S\setminus B}\|\nabla\eta_N(t)\| \leq \sup_{t\in S}\|\nabla\eta_N(t)\|$$

so in particular,

$$\sup_{t\in S\setminus B}\|\nabla\eta_N(t)\| \xrightarrow{\mathbb{P}} 0$$

from which the first result follows. As in Cheng and Schwartzman (2017), the probability that $\hat{\gamma}_N$ has no maxima is less than or equal to

$$\mathbb{P}\left(\sup_{t\in S\setminus B}\|\nabla\eta_N(t)\| > \delta_j D_{\max}\right)$$

which tends to 0 by a similar argument to above. Note that the argument in Cheng and Schwartzman (2017) requires an application of the Fundamental Theorem of Calculus to the second derivative which is why we require that the second derivatives of $\gamma$ and $\eta_N$ are continuous. The probability of 2 or more local maxima is

$$\mathbb{P}\left(\sup_{t\in S\setminus B}\sup_{\|x\|=1} x^T\nabla^2\eta_N(t)x > D_{\max}\right)$$

which tends to 0, since

$$\sup_{t\in S\setminus B}\sup_{\|x\|=1} x^T\nabla^2\eta_N(t)x \leq \sup_{t\in S\setminus B}\left\|\nabla^2\eta_N(t)\right\| \xrightarrow{a.s.} 0.$$

Combining these last two convergences implies the second result. □

## 8.2 Conditions for convergence

**Lemma 8.1.** *Given $a_1,\ldots,a_N,\mu\in\mathbb{R}$ for some $N\in\mathbb{N}$*

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^N\left(a_n-\frac{1}{N}\sum_{n=1}^N a_n\right)^2 &= \frac{1}{N}\sum_{n=1}^N\left(a_n-\mu+\mu-\frac{1}{N}\sum_{n=1}^N a_n\right)^2\\ &= \frac{1}{N}\sum_{n=1}^N(a_n-\mu)^2+\frac{2}{N}\sum_{n=1}^N\left(\mu-\frac{1}{N}\sum_{n=1}^N a_n\right)(a_n-\mu)+\frac{1}{N}\sum_{n=1}^N\left(\mu-\frac{1}{N}\sum_{n=1}^N a_n\right)^2\\ &= \frac{1}{N}\sum_{n=1}^N(a_n-\mu)^2+\left(\mu-\frac{1}{N}\sum_{n=1}^N a_n\right)\frac{2}{N}\sum_{n=1}^N(a_n-\mu)+\left(\mu-\frac{1}{N}\sum_{n=1}^N a_n\right)^2.\end{aligned}$$

**Lemma 8.2.** *In the setting of Section 3.3.2 let $g$ be the density of $x_1$ and suppose that $g$ is bounded above by $G\in\mathbb{R}$ then $\left(\frac{1}{N}X_N^TX_N\right)^{-1}\xrightarrow{a.s.}\Sigma^{-1}$.*

*Proof.* Define the event $E_N=\left\{X\in\mathbb{R}^{N\times p}:\det(X^TX)=0\right\}$ then, as $x_1,\ldots,x_N$ are independent,

$$\mathbb{P}(X_N\in E_N)=\int_{E_N}\prod_{n=1}^N g(x_n)\,dx_n\leq G^N\int_{E_N}\prod_{n=1}^N dx_n=0.$$

since $E_N$ traces out a lower dimensional subspace. So by countability (as the countable union of measure zero sets has measure zero) we can almost surely assume that $\frac{1}{N}X_N^TX_N$ is invertible for all $N$. The result follows by the continuous mapping theorem. □

## 8.3 Proof of Theorem 4.1

We can prove Theorem 4.1 by verifying the CLT conditions of Shi (2011):

*Proof.* i) In our setting, $B_j$ is compact and $\hat{\mu}_N=\frac{1}{N}\sum_{n=1}^N Y_n$ is continuous, point-

wise measurable and converges uniformly in probability to $\mu$ as shown in Section 3.3.1. Furthermore, $\text{argmax}_{t\in B_j}\,\mu(t) = \theta_j$ is the unique maximum, so it follows that $\hat{\theta}_{j,N}$ is a consistent estimator for $\theta_j$ by Theorem 4.1.1 from Amemiya (1985). $\theta_j$ also lies in the interior of $B_j$.

ii) For $N \in \mathbb{N}$, $\hat{\mu}_N$ is a.s. twice continuously differentiable on $B_j$.

iii) For all $t \in S$,

$$\sqrt{N}(\nabla\hat{\mu}_N(t) - \nabla\mu(t)) = \sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N}\nabla Y_n(t) - \nabla\mu(t)\right) \xrightarrow{d} N(0, \text{cov}(\nabla^T Y_1(t)))$$

since $\text{cov}(\nabla Y_1(t)) < \infty$ and $\mathbb{E}[\nabla Y_n] = \nabla\mu$ by derivative exchangeability. Thus taking $t = \theta_j$, $\nabla\mu(\theta_j) = 0$, so it follows that

$$\sqrt{N}\nabla\hat{\mu}_N(\theta_j) \xrightarrow{d} N(0, \text{cov}(\nabla^T Y_1(\theta_j))).$$

iv) For $\tilde{\theta}_{j,N} \xrightarrow{\mathbb{P}} \theta_j$,

$$\nabla^2\hat{\mu}_N(\tilde{\theta}_{j,N}) = \nabla^2\hat{\mu}_N(\tilde{\theta}_{j,N}) - \nabla^2\mu(\tilde{\theta}_{j,N}) + \nabla^2\mu(\tilde{\theta}_{j,N}) - \nabla^2\mu(\theta_j) + \nabla^2\mu(\theta_j)$$

which converges in probability to $\nabla^2\mu(\theta_j)$, since by the fSLLN,

$$\nabla^2\hat{\mu}_N(\tilde{\theta}_{j,N}) - \nabla^2\mu(\tilde{\theta}_{j,N}) \xrightarrow{a.s.} 0$$

and as $\nabla^2\mu$ is uniformly continuous on $B_j$, $\nabla^2\mu(\tilde{\theta}_{j,N}) - \nabla^2\mu(\theta_j) \xrightarrow{\mathbb{P}} 0$.

v) $\hat{\theta}_{j,N} \xrightarrow{\mathbb{P}} \theta_j$ so in particular the probability of the event that $\hat{\theta}_{j,N}$ lies within the interior of $B_j$ can be taken to be as close to 1 as desired. On this event $\nabla\hat{\mu}_N(\hat{\theta}_{j,N}) = 0$ so in particular $\nabla\hat{\mu}_N(\hat{\theta}_{j,N}) = o_p(N^{-1/2})$.

□

## 8.4 Proofs for Section 4.1

### 8.4.1 Proof of Lemma 4.4

*Proof.* Differentiating $\hat{\sigma}_N^2$ (and evaluating all fields pointwise), we have

$$\begin{aligned}\sqrt{N}\nabla\hat{\sigma}_N^2 &= \frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\left(\epsilon_n - \frac{1}{N}\sum_{j=1}^{N}\epsilon_j\right)\left(\nabla\epsilon_n - \frac{1}{N}\sum_{k=1}^{N}\nabla\epsilon_k\right)\\ &= \frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\epsilon_n\nabla\epsilon_n - \frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\epsilon_n\left(\frac{1}{N}\sum_{k=1}^{N}\nabla\epsilon_k\right)\\ &\quad - \frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\nabla\epsilon_n\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_j\right) + \frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_j\right)\left(\frac{1}{N}\sum_{k=1}^{N}\nabla\epsilon_k\right).\end{aligned}$$

Now $\frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\epsilon_n$ converges in distribution by the CLT (as $\text{var}(\epsilon_n) < \infty$) and $\frac{1}{N}\sum_{k=1}^{N}\nabla\epsilon_k \xrightarrow{a.s.}$ 0 by the SLLN as $Y_1$ satisfies the DE conditions and so

$$\left(\frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\epsilon_n\right)\left(\frac{1}{N}\sum_{k=1}^{N}\nabla\epsilon_k\right) \xrightarrow{\mathbb{P}} 0 \text{ as } N \longrightarrow \infty.$$

Similarly

$$-\frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\nabla\epsilon_n\left(\frac{1}{N}\sum_{n=1}^{N}\epsilon_n\right) \xrightarrow{\mathbb{P}} 0 \text{ and } \frac{2\sqrt{N}}{N-1}\sum_{n=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_j\right)\left(\frac{1}{N}\sum_{k=1}^{N}\nabla\epsilon_k\right) \xrightarrow{\mathbb{P}} 0$$

as $N \longrightarrow \infty$. We can thus write

$$\begin{pmatrix}\nabla^T Z_N \\ \nabla^T V_N/\sqrt{N}\end{pmatrix} = \sqrt{N}\begin{pmatrix}\frac{\sigma}{N}\sum_{n=1}^{N}\nabla^T\epsilon_n \\ \nabla^T\hat{\sigma}_N^2\end{pmatrix} = \sqrt{N}\frac{1}{N}\sum_{n=1}^{N}\begin{pmatrix}\nabla^T\epsilon_n \\ 2\epsilon_n\nabla^T\epsilon_n\end{pmatrix} + \begin{pmatrix}0 \\ B_N\end{pmatrix}$$

where $B_N \xrightarrow{\mathbb{P}} 0$ as $N \longrightarrow \infty$ by Slutsky. Thus by the multivariate CLT and applying Slutsky once more we have

$$\begin{pmatrix}\nabla^T Z_N \\ \nabla^T V_N/\sqrt{N}\end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix}\text{cov}(\nabla^T\epsilon_1) & 0 \\ 0 & 4\text{cov}(\epsilon_1\nabla^T\epsilon_1)\end{pmatrix}\right)$$

as $N \longrightarrow \infty$ since the field and its derivative are independent by the constant variance assumption. The result follows as $\epsilon$ is mean zero and so

$$\text{cov}\big(\epsilon_1\nabla^T\epsilon_1\big) = \text{var}(\epsilon_1)\text{cov}\big(\nabla^T\epsilon_1\big) = \text{cov}\big(\nabla^T\epsilon_1\big) = \Lambda.$$

□

### 8.4.2 Proof of Theorem 4.5

*Proof.* We will take the same approach as we did in the proof of Lemma 4.2 and Corollary 4.3, namely to first prove the result assuming that the variance is constant and then use this to obtain the general result. So assume that $\text{var}(Y_1) = 1$ is constant. Differentiating,

$$\nabla T_N = \sqrt{N-1}\left(\frac{\sqrt{N}\nabla\mu + \nabla Z_N}{V_N^{1/2}}\right) - \sqrt{N-1}\left(\frac{\sqrt{N}\mu + Z_N}{2V_N^{3/2}}\right)\nabla V_N.$$

Thus

$$\begin{aligned}\nabla T_N - \frac{\sqrt{N}\nabla\mu}{\sqrt{V/N-1}} &= \nabla Z_N + \left(\sqrt{\frac{N-1}{V_N}} - 1\right)\nabla Z_N - \frac{N-1}{V_N}\left(\frac{\sqrt{N}\mu + Z_N}{2V_N^{1/2}}\right)\frac{\nabla V_N}{\sqrt{N-1}} \\ &= \nabla Z_N - \frac{\mu\nabla V_N}{2\sqrt{N-1}} + \left(\sqrt{\frac{N-1}{V_N}} - 1\right)\nabla Z_N \\ &\quad + \left(\frac{\mu}{2} - \frac{N-1}{V_N}\left(\frac{\sqrt{N}\mu + Z_N}{2V_N^{1/2}}\right)\right)\frac{\nabla V_N}{\sqrt{N-1}}.\end{aligned}$$

The last two terms converge to zero in distribution (by the usual arguments involving Slutsky by applying the CLT and using the fact that $\frac{\sqrt{N}\mu+Z_N}{\sqrt{V_N}} \xrightarrow{a.s.} \mu$, see Section 4.1.1). Applying Slutsky again and using the joint asymptotic distribution of $(\nabla Z_N, \nabla V_N/\sqrt{N})$ gives the result in the unit-variance case. To see this, applying Lemma 4.4,

$$\begin{pmatrix}\nabla Z_N & \nabla V_N/\sqrt{N}\end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix}\Lambda & 0 \\ 0 & 4\Lambda\end{pmatrix}\right) \text{ as } N \longrightarrow \infty.$$

We have

$$\begin{pmatrix}1, -\mu/2\end{pmatrix}\begin{pmatrix}\Lambda & 0 \\ 0 & 4\Lambda\end{pmatrix}\begin{pmatrix}1 \\ -\mu/2\end{pmatrix} = \Lambda + \mu^2\Lambda,$$

so it follows that, as $N \to \infty$,

$$\nabla T_N - \frac{\sqrt{N}\nabla\mu}{\sqrt{V_N/N - 1}} \xrightarrow{d} N\big(0, \big(1 + \mu^2\big)\Lambda\big).$$

Dropping the assumption of constant variance, the general result follows by considering the fields $Y_n/\sigma$, arguing as in the proof of Corollary 4.3. □

## 8.5 Proof of Theorem 4.6

We can prove Theorem 4.1 by verifying the CLT conditions of Shi (2011).

*Proof.* i) In our setting, $B_j$ is compact and $d_N$ is continuous, pointwise measurable and converges uniformly in probability to $\mu$ as shown in section 3.3.1. Furthermore, $\text{argmax}_{t\in B_j} \frac{\mu(t)}{\sigma(t)} = \theta_0$ is the unique maximum of $\frac{\mu}{\sigma}$, so it follows that $\hat{\theta}_{j,N}$ is a consistent estimator for $\theta_j$ by Theorem 4.1.1 from Amemiya (1985). $\theta_0$ also lies on the interior of $B_j$.

ii) $d_N$ is a.s. twice continuously differentiable as $Y_N$ are and the $Y_n$ are non-degenerate.

iii) $\nabla\frac{\mu(\theta_j)}{\sigma(\theta_j} = 0$ and so by Theorem 4.5,

$$\sqrt{N}\nabla^T d_N(\theta_j) \xrightarrow{d} N\left(0, \left(1 + \frac{\mu(\theta_j)^2}{\sigma(\theta_j)^2}\right)\Lambda'(\theta_j)\right).$$

iv) For $\tilde{\theta}_{j,N} \xrightarrow{\mathbb{P}} \theta_j$,

$$\nabla^2 d_N(\tilde{\theta}_{j,N}) = \nabla^2 d_N(\tilde{\theta}_{j,N}) - \nabla^2\frac{\mu(\tilde{\theta}_{j,N})}{\sigma(\tilde{\theta}_{j,N})} + \nabla^2\frac{\mu(\tilde{\theta}_{j,N})}{\sigma(\tilde{\theta}_{j,N})} - \nabla^2\frac{\mu(\theta_j)}{\sigma(\theta_j)} + \nabla^2\frac{\mu(\theta_j)}{\sigma(\theta_j)}$$

which converges in probability to $\nabla^2\frac{\mu(\theta_j)}{\sigma(\theta_j)}$, since by Proposition 3.7,

$$\nabla^2 d_N(\tilde{\theta}_{j,N}) - \nabla^2\frac{\mu(\tilde{\theta}_{j,N})}{\sigma(\tilde{\theta}_{j,N})} \xrightarrow{a.s.} 0$$

and as $\nabla^2 \frac{\mu}{\sigma}$ is uniformly continuous $\nabla^2 \frac{\mu(\tilde{\theta}_{j,N})}{\sigma(\tilde{\theta}_{j,N})} - \nabla^2 \frac{\mu(\theta_j)}{\sigma(\theta_j)} \xrightarrow{\mathbb{P}} 0$.

v) $\hat{\theta}_{j,N} \xrightarrow{\mathbb{P}} \theta$ so in particular the probability that $\hat{\theta}_{j,N}$ lies within the interior of $B_j$ can be taken to be as close to 1 as required. On this event $\nabla d_N(\hat{\theta}_{j,N}) = 0$ so in particular $\nabla d_N(\hat{\theta}_{j,N}) = o_p(N^{-1/2})$.

□

# 9 Appendix

## 9.1 The derivative of a $\chi^2$ field

**Lemma 9.1.** *Let $Y_1, \ldots, Y_N$ be zero mean i.i.d D-dimensional Gaussian random fields on $S$ (with variance $\sigma^2$ that is not necessarily constant). Let $U = \sum_{n=1}^{N} Y_n^2$ and $Y = (Y_1, \ldots, Y_N)^T$, then*

$$\nabla^T U(s) \big| Y(s) \sim N\left( \frac{2\Gamma(s)U(s)}{\sigma^2}, 4U(s)(\Lambda(s) - \Gamma(s)\Gamma(s)^T/\sigma^2(s)) \right)$$

*and so for all $s \in S$,*

$$\nabla^T U(s) \sim \frac{2\Gamma(s)U(s)}{\sigma(s)^2} + 2U(s)^{1/2} z_U(s)$$

*where $U(s) \sim \chi^2_N$ is independent of $z_U(s) \sim N(0, (\Lambda - \Gamma\Gamma^T/\sigma^2))$.*

*Proof.* Evaluating all quantities pointwise, for each $n = 1, \ldots, D$,

$$(Y_n, \nabla Y_n)^T \sim N\left(0, \begin{pmatrix} \sigma^2 & \Gamma^T \\ \Gamma & \Lambda \end{pmatrix}\right)$$

as such using the formula for the conditional Gaussian distribution,

$$\nabla^T Y_n | Y \sim N\left( \frac{\Gamma}{\sigma^2} Y_n, \Lambda - \Gamma\Gamma^T/\sigma^2 \right).$$

Now differentiating $U$ we find that

$$\nabla^T U|Y = 2\sum_{n=1}^{N} Y_n \nabla^T Y_n \sim N\left(\frac{2\Gamma}{\sigma^2}\sum_{n=1}^{N} Y_n^2, 4\sum_{n=1}^{N} Y_n^2(\Lambda - \Gamma\Gamma^T/\sigma^2)\right)$$
$$\sim N\left(\frac{2\Gamma}{\sigma^2}U, 4U(\Lambda - \Gamma\Gamma^T/\sigma^2)\right)$$

So the final result follows. □

For the final step above we used the fact that given random variables $A, B$,

$$p_{A|B,B}(a,b) = p_{A|B|B}(a|b)p_B(b) = p_{A|B}(a|b)p_B(b) = p_{A,B}(a,b)$$

which means that $A|B$ and $B$ are independent no matter what dependence there is between $A$ and $B$ themselves.

# Chapter 4

# The asymptotic distribution of the size of a cluster in a non-stationary Gaussian random field

Samuel Davenport

**Abstract**

Let $Y$ be a mean-zero $D$-dimensional Gaussian random field on set $S \subset \mathbb{R}^D$ for some $D \in \mathbb{N}$. Given $t_0 \in S$ and a cluster defining threshold $u$, let $c_u$ be the volume of the component of the excursion set of $Y$ above $u$ that contains $t_0$. We show, for $Y$ belonging to a certain class of unit-variance fields, that conditional on $Y$ having a maximum at $t_0$ with height greater than $u$, $c_u^{D/2}$ has an exponential distribution as $u \to \infty$; extending results due to Nosko (1969) to non-stationarity.

*Keywords:* Random Field Theory, clustersize inference, non-stationarity.

# 1 Introduction

The asymptotic distribution of the extent of a cluster of a random field above a threshold has a wide range of applications and in particular has been used to perform clustersize inference in neuroimaging via a framework set up in Friston et al. (1994). Up until now this distribution has only been known for stationary random fields. Friston et al. (1994)'s inference framework has been validated in isotropic Gaussian simulations (Hayasaka and Nichols, 2003), however, recent work (Eklund et al. (2016)) has called into question stationarity assumptions in fMRI data. They tested the performance of clustersize inference using resting state fMRI data and showed that the failure of stationarity (and other) assumptions has led to inflated clusterwise false positive rates. As such it is of great interest to obtain the distribution of the size of a cluster above a threshold under non-stationarity.

Existing approaches to provide a non-stationary clustersize inference framework ((Hayasaka et al., 2004), Worsley et al. (1999)) rely on deforming space to stationarity. While their approach does help to account for local non-stationarity, they do not provide rigorous theoretical results (based on HW conditioning) and instead rely on heuristics to obtain approximate clustersize distributions. The original results for the asymptotic distribution of the size of a cluster above a threshold date back to Nosko (1969). Under stationarity he stated results showing that the powers of the clustersize above a threshold for Gaussian random fields are asymptotically exponential. This was formalized in Wilson and Adler (1982), Wilson (1988) and Nosko (1988) and extended by Aronowich and Adler (1986) to 1D $\chi^2$ random fields and then by Cao (1999) to multidimensional $\chi^2, t$ and $f$-fields. See Adler et al. (2010) Chapter 6 for an overview of this theory. Recently, under minimal assumptions, Cheng and Schwartzman (2015a)

obtained the distribution of the height of a peak of a non-stationary Gaussian field conditional on observing that peak at a given location. Their work extended results, due to Nosko (1969) and Belyaev (1967) (see also Adler (1981) Chapter 6.8), which were originally derived under stationarity and ergodicity assumptions.

In this paper we extend the cluster size results of Nosko (1969) and Wilson (1988) to non-stationary Gaussian fields. Our approach will be to show that the field around a high maximum takes the shape of an elliptic parabaloid and to then apply the results of Cheng and Schwartzman (2015a) to obtain an asymptotic distribution for the size of a cluster as the cluster defining threshold goes to infinity.

This paper is laid out as follows. In Section 2 we set out the assumptions that we will require and prove that they hold for an important class of random fields. We then introduce the notion of HW distributions which are needed for conditioning on events that have zero probability. We show that the height distribution of a peak in a Gaussian random field above a threshold is asymptotically exponential and that HW-conditional on observing a peak above a threshold $u$ we have convergence of the first and second derivative as $u \to \infty$. Section 3 proves an asymptotic functional limit theorem regarding the field around a peak and proves an asymptotic result about the size of a cluster as the threshold goes to infinity. Section 5 discusses the results and Section 6 contains the proofs.

# 2 Assumptions and HW distributions

## 2.1 Assumptions

Throughout we will take $Y : S \to \mathbb{R}$ to be a mean-zero real a.s. thrice continuously differentiable Gaussian random field on a probability space $\mathcal{P} = (\Omega, \mathcal{F}, \mathbb{P})$ with compact domain $S \subset \mathbb{R}^D$ for some dimension $D \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of positive integers. Note that since $S$ is bounded, $\text{diam}(S) := \sup_{s,t\in S}\|s - t\| < \infty$. We will write $\sigma^2 = \text{var}(Y)$, evaluated pointwise, let $\text{int}(S)$ denote the interior of $S$ and for $t \in S$ let

$$Y_i(t) = \nabla_i Y(t) := \frac{\partial Y(t)}{\partial t_i} \text{ and } Y_{ij}(t) = \nabla_{ij} Y(t) := \frac{\partial^2 Y(t)}{\partial t_i t_j} \text{ for } i, j = 1, \ldots, D.$$

For $t \in S$ and $\epsilon > 0$ we will take $B_\epsilon(t)$ to be the open $D$-dimensional ball of radius $\epsilon$ that is centred at $t$ and write $\overline{B}_\epsilon(t)$ to denote its closure. We will also take $\phi$ to be the density of the standard normal random variable. In order to prove our results we will impose the following assumptions on $Y$.

**Assumption 2.1.** *(a) For every $s, t \in S$ with $s \neq t$, the Gaussian random vector*

$$(Y(s), \nabla Y(s), Y_{jk}(s), Y(t), \nabla Y(t), Y_{jk}(t), 1 \leq j, k \leq D)$$

*is non-degenerate.*

*(b) There exists a real integrable random variable $L$ on $\mathcal{P}$ such that*

$$|Y(t) - Y(s)| \leq L\|t - s\|$$

*for all $t, s \in S$.*

*(c) For some $V \in \mathbb{R}$, there exists a real random vector $X \in \mathbb{R}^V$ on $\mathcal{P}$, indexed by a*

*set $\mathcal{V}$ containing $V$ elements, and a constant $c' \in \mathbb{R}$ such that*

$$\|\nabla Y(t) - \nabla Y(s)\| \leq c' \sum_{l \in \mathcal{V}} |X(l)| \|t - s\|$$

*for all $t, s \in S$. Assume further that for each $t \in int(S)$ there exists some $r > 0$ such that $\sup_{l \in \mathcal{V}} \sup_{s \in \bar{B}_r(t)} p_{\nabla Y(t)|X(l),\nabla^2 Y(s),Y(s)} < \infty$. Hereon we will write $L' = c' \sum_{l \in \mathcal{V}} |X(l)|$.*

*(d) For $1 \leq i, j \leq D$, there exist real constants $c_{ij}$ such that for all $t, s \in S$,*

$$|Y_{ij}(t) - Y_{ij}(s)| \leq c_{ij} \sum_{l \in \mathcal{V}} |X(l)| \|t - s\|$$

*Hereon we will write $L_{ij} = c_{ij} \sum_{l \in \mathcal{V}} |X(l)|$. We will further assume that for each $l \in \mathcal{V}, |X(l)|^D$ is integrable, that $\sup_{t \in S} \mathbb{E}\Big[|X(l)|^{2D} | \nabla Y(t) = 0\Big] < \infty$ and that there exists some $\epsilon > 0$ such that*

$$\mathbb{E}\left[\left(\sum_{l \in \mathcal{V}} |X(l)|\right)^{1+\epsilon}\right] < \infty.$$

*(e) $\sigma^2 : S \to \mathbb{R}$ is continuous.*

Let us now define a useful class of random fields that satisfy these assumptions.

**Definition 2.2.** Given a finite lattice $\mathcal{V} \subset \mathbb{R}^D$ and real random variables $\{X(l) : l \in \mathcal{V}\}$ and a continuous kernel $K : \mathbb{R}^D \to \mathbb{R}$, define the **convolution field** $Y : \mathbb{R}^D \to \mathbb{R}$ such that for each $s \in \mathbb{R}^D$,

$$Y(s) = \sum_{l \in \mathcal{V}} K(s - l) X(l).$$

We will use the notation $(Y, X, \mathcal{V}, K)$ to denote the convolution field and will write $\mathbb{V}(X)$ to denote the $|\mathcal{V}|$ length vector each entry of which is a distinct element of $\{X(l) : l \in \mathcal{V}\}$. We say that $(Y, X, \mathcal{V}, K)$ is a **Gaussian convolution field** if $\mathbb{V}(X)$ is a Gaussian random vector.

Convolution fields were introduced in Telschow et al. (2020b). They appear in a number of diverse settings; in particular in brain imaging where it may be most interesting to apply our theoretical results in practice. In Davenport et al. (2021) (Chapter 2) we showed that they can be used to accurately control the familywise error rate in fMRI when conducting voxelwise inference.

**Definition 2.3.** It is often convenient to work with constant variance fields as they have the property that the field is independent of its derivative. Given a convolution field $(Y, X, \mathcal{V}, K)$ we define the **scaled convolution field** to be the variance 1 random field $Z : \mathbb{R}^D \to \mathbb{R}$ such that

$$Z = \frac{Y}{\sigma}.$$

where division is calculated pointwise.

As with $Y$, we will denote the partial derivatives of $K$ and $Z$ using subset indices. Importantly convolution fields typically have non-degenerate marginal distributions. To show this we need to introduce a relevant notion of linear independence of functions.

**Definition 2.4.** Given a $D$-dimensional lattice $\mathcal{V}$, we say that functions $f_1, \dots, f_n : \mathbb{R}^D \to \mathbb{R}$ are $\mathcal{V}$-linearly independent if given constants $a_1, \dots, a_n \in \mathbb{R}$ (some $n \in \mathbb{N}$) and any $s \in \mathbb{R}^D$, the relation

$$\sum_{i=1}^{n} a_i f_i(s - l) = 0$$

holding for all $l \in \mathcal{V}$ implies that $a_i = 0$ for all $i = 1, \dots, n$. We say that they are doubly $\mathcal{V}$-linearly independent if given constants $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{R}$ and any $s \neq t \in \mathbb{R}^D$, the relation

$$\sum_{i=1}^{n} a_i f_i(s - l) + \sum_{i=1}^{n} b_i f_i(t - l) = 0$$

holding for all $l \in \mathcal{V}$ implies that $a_i = b_i = 0$ for all $i = 1, \dots, n$.

Let $\mathbb{V}$ denote the **vech** operation sending $D$ dimensional symmetric matrices to $\mathbb{R}^{D(D+1)/2}$, see Section 6.1 for a formal definition. Given these definitions we have the following Lemma.

**Lemma 2.5.** *Let $(Y, X, \mathcal{V}, K)$ be a convolution field with twice differentiable $D$-dimensional kernel $K$ such that $\mathbb{V}(X)$ is a non-degenerate Gaussian vector and $K, K_i, K_{jk}$ for $1 \leq i, j, k \leq D$ are $\mathcal{V}$-linearly independent. Let $Z$ be the corresponding scaled field. Then for all $s \in \mathbb{R}^D$,*

$$(Y(s), \nabla Y(s), (\mathbb{V}(\nabla^2 Y(s)))^T)^T \text{ and } (Z(s), \nabla Z(s), (\mathbb{V}(\nabla^2 Z(s)))^T)^T$$

*are non-degenerate Gaussian random vectors. If $K, K_i, K_{jk}$ are doubly $\mathcal{V}$-linearly independent then Assumption 2.1a holds for $Y$ and $Z$.*

*Proof.* Given $s \in \mathbb{R}^D$, suppose that there exist sets of real constants $a, a_i, a_{jk}, c$ ($1 \leq i \leq D, 1 \leq j \leq k \leq D$) such that

$$aY(s) + \sum_{i=1}^{D} a_i Y_i(s) + \sum_{1 \leq j \leq k \leq D} a_{jk} Y_{jk}(s) = c.$$

Non-degeneracy of $\mathbb{V}(X)$ then implies that for all $l \in \mathcal{V}$

$$aK(s-l) + \sum_{i=1}^{D} a_i K_i(s-l) + \sum_{1 \leq j \leq k \leq D} a_{jk} K_{jk}(s-l) = 0,$$

which by the linear independence constraint implies that the constants are all zero. This proves non-degeneracy of $(Y(s), \nabla Y(s), \mathbb{V}(\nabla^2 Y(s))^T)^T$. For the scaled field, we note that

$$\nabla \frac{Y}{\sigma} = \frac{\nabla Y}{\sigma} - \frac{Y \nabla \sigma}{\sigma^2} = \frac{\nabla Y}{\sigma} - \frac{\nabla \sigma}{\sigma}\left(\frac{Y}{\sigma}\right)$$

and

$$\nabla^2 \frac{Y}{\sigma} = \frac{\nabla^2 Y}{\sigma} - \frac{2(\nabla Y)^T (\nabla \sigma)}{\sigma^2} - \frac{Y \nabla^2 \sigma}{\sigma^2} + \frac{2(\nabla \sigma)^T (\nabla \sigma) Y}{\sigma^3}.$$

So a linear combination of $\frac{Y}{\sigma}, \nabla_i \frac{Y}{\sigma}, \nabla^2_{ij} \frac{Y}{\sigma}$ yields a linear combination of $Y, Y_i, Y_{jk}$. This means that non-degeneracy of $(Z(s), \nabla Z(s), \mathbb{V}(\nabla^2 Z(s))^T)^T$ is equivalent to non-degeneracy of $(Y(s), \nabla Y(s), \mathbb{V}(\nabla^2 Y(s))^T)^T$. The proof of the second claim follows similarly. □

Non-degeneracy is one of the conditions for the Gaussian Kinematic formula to hold for convolution fields. In our context it is needed in order to ensure that $Y$ satisfies the KR (Kac-Rice) conditions (Adler and Taylor, 2007) and we will use it to prove that the bounds on the conditional pdfs, required in Assumption 2.1, hold. Taking $(X(l))_{l\in\mathcal{V}}$ to be jointly Gaussian and restricting $Y$ to a compact $S \subset \mathbb{R}$ ensures that $Y$ falls under our desired framework. In particular we have the following result.

**Proposition 2.6.** *Let $(Y, X, \mathcal{V}, K)$ be a $D$-dimensional convolution field such that $\mathbb{V}(X)$ is a non-degenerate Gaussian random vector and $K$ is a $C^3$ kernel such that $K, K_i, K_{jk}$ for $1 \leq i, j, k \leq D$ are doubly $(\mathcal{V} \setminus l)$-linearly independent for all $l \in \mathcal{V}$. Then $Y$ restricted to $S$ satisfies Assumption 2.1.*

*Proof.* Assumption 2.1a holds by the previous Lemma. Now let

$$S' = \{s - l : s \in S, l \in \mathcal{V}\},$$

then $S'$ is bounded as $S$ is bounded and $\mathcal{V}$ is finite. Given $t, s \in S$,

$$|Y(t) - Y(s)| \leq \sum_{l\in\mathcal{V}} |X(l)||K(s-l) - K(t-l)| \leq \sup_{s'\in S'} \|\nabla K(s')\| \sum_{l\in\mathcal{V}} |X(l)|\|s - t\|$$

so that Assumption 2.1b is satisfied as $\nabla K$ is continuous. Since $K$ is $C^3$ we can apply the same argument to the first two derivatives of $Y$ to show that the Lipschitz requirements of Assumption 2.1c,d are satisfied with Lipschitz constants of the form $c\sum_{l\in\mathcal{V}}|X(l)|$ for some constant $c$.

Given $t \in \text{int}(S)$, choose $r > 0$ such that $\overline{B}_r(t) \subset S$. Then for all $s \in \overline{B}_r(t)$, for each $l \in \mathcal{V}$, $\left(\nabla Y(t), X(l), Y(s), \mathbb{V}(\nabla^2 Y(s))^T\right)^T$ is a non-degenerate Gaussian random vector (arguing as in Lemma 2.5) with continuous covariance structure so, by compactness of the ball, there exists a bound on the supremum of the required conditional pdf for Assumption 2.1c. The integrability conditions follow as the Gaussian distribution has finite moments and as for each $l \in \mathcal{V}$, $(X(l), \nabla Y(t))$ has a continuous covariance structure. □

In our context convolution fields provide a realistic example that is easy to work with and satisfies all of our assumptions. Importantly, scaled convolution fields also satisfy these assumptions.

**Proposition 2.7.** *Let $(Y, X, \mathcal{V}, K)$ be a $D$-dimensional convolution field satisying the conditions of Proposition 2.6. Let $K$ have support $A$ and suppose that*

$$S \subset \left\{x \in \mathbb{R}^D : x = l + a, \text{ some } l \in \mathcal{V} \text{ and } a \in A\right\}.$$

*Then $\frac{Y}{\sigma}$ restricted to $S$ satisfies Assumption 2.1.*

*Proof.* $S$ is bounded and so contained within a compact set meaning that the variance is bounded above and below. The bound below is greater than 0 since for it to be equal to zero would mean there was a point in $S$ at which the field had zero variance which would imply that $\mathbb{V}(X)$ were non-degenerate. By the mean value inequality, for all $s, t \in S$,

$$\left|\frac{Y(t)}{\sigma(t)} - \frac{Y(s)}{\sigma(s)}\right| \leq \sup_{s' \in S}\left\|\nabla \frac{Y(s')}{\sigma(s')}\right\| \|t - s\|,$$

and we have

$$\sup_{s' \in S}\left\|\nabla \frac{Y(s')}{\sigma(s')}\right\| = \sup_{s' \in S}\left\|\frac{\nabla Y(s')}{\sigma(s')} - \frac{\nabla \sigma(s') Y(s')}{\sigma^2(s')}\right\|$$

$$\leq \frac{L}{\inf_{s' \in S} \sigma(s)} + \sup_{s' \in S} |Y(s')| \frac{\sup_{s' \in S} \|\nabla \sigma(s')\|}{\inf_{s' \in S} \sigma^2(s')}.$$

The kernel is bounded above (as discussed in the proof of Proposition 2.6) by some constant $K^*$ and so

$$\sup_{s' \in S} |Y(s')| \leq K^* \sum_{l \in \mathcal{V}} |X(l)|.$$

This means that an integrable Lipschitz bound exists so that Assumption 2.1b is satisfied. Arguing similarly for the first and second derivatives for each of them we obtain Lipschitz constants of the form

$$c \sum_{l \in \mathcal{V}} |X(l)|.$$

Arguing as in the proof of Proposition 2.6 it follows that the remaining constraints of Assumption 2.1 are satisfied. □

The linear independence condition is satisfied for a wide-range of kernels and lattices. In particular we have the following.

**Proposition 2.8.** *Let $K$ be the $D$-dimensional Gaussian kernel and let $\mathcal{V}$ be a $D$-dimensional lattice such that for each $d = 1, \ldots, D$, there exist at least 3 points of $\mathcal{V}$ with distinct $d$th entries. Then $K, K_i, K_{jk}$ for $1 \leq i, j, k \leq D$ are $\mathcal{V}$-linearly independent.*

*Proof.* For $1 \leq i, j, k \leq D$, and $x \in \mathbb{R}^D$, $K_i(x) = x_i K(x)$ and $K_{jk}(x) = (x_j x_k + \delta_{jk})K(x)$, where here $\delta$ denotes the dirac delta function. So given $s \in \mathbb{R}^D$, if there exist constants $a, a_i, a_{jk}$ such that

$$aK(s-l) + \sum_{i=1}^{D} a_i K_i(s-l) + \sum_{1 \leq j \leq k \leq D} a_{jk} K_{jk}(s-l) = 0,$$

for all $l \in \mathcal{V}$, then by the linear independence condition it follows that

$$a + \sum_{i=1}^{D} a_i (s_i - l_i) + \sum_{1 \leq j \leq k \leq D} a_{jk}((s_j - l_j)(s_k - l_k) + \delta_{jk}) = 0. \tag{4.1}$$

Given $i = 1, \ldots, D$, if we fix $(l_1, \ldots, l_{i-1}, l_{i+1}, \ldots, l_D)$ then (4.1) is a quadratic in $l_i$ so the only way that it can have more than 2 distinct solutions is if $a_{ii} = 0$ and

$$a_i + \sum_{j \neq i} a_{ij}(s_j - l_j) = 0.$$

If $a_{ij} \neq 0$ for some $j \neq i$ then will only be 1 distinct solution for $l_j$ which is a contradiction. As such it follows that $a_i, a_{ij} = 0$ for all $i, j$ and that $a = 0$. □

In order to ensure that the Gaussian kernel and its derivatives are $\mathcal{V} \setminus l$ independent for all $l \in \mathcal{V}$ it is thus sufficient that the lattice has at least 4 points in each direction. In most applications (e.g. in fMRI) this assumption is easily satisfied. Double linear independence holds for the Gaussian kernel but is more difficult to show. However, arguing as in the proof of the above proposition, it can easily be shown to hold for any polynomial kernel given a large enough lattice.

## 2.2 HW distributia

The event that $Y$ has a maximum at a point $t_0 \in \text{int}(S)$ can be written as

$$\mathcal{M}(t_0) := \left\{ \nabla Y(t_0) = 0 \text{ and } \nabla^2 Y(t_0) \prec 0 \right\}$$

This event has probability zero and so in order to condition on it we will invoke HW (horizontal window) distributions as in Adler (1981), Cao (1999) and Cheng and Schwartzman (2015a). (See Kac and Slepian (1959) for a discussion of the motivation behind using horizontal as opposed to vertical windowing.) In particular given $u \in \mathbb{R}$ we will want to condition on observing a maximum above a threshold $u$ at $t_0$: conditioning on the event

$$\mathcal{M}_u(t_0) := \mathcal{M}(t_0) \cap \{Y(t_0) > u\}.$$

In order to do so, for $T \subset S$ let

$$\mathcal{M}(T) := \bigcup_{t \in T} \mathcal{M}(t).$$

Then, given $A \subseteq \mathcal{F}$, we define the probability of $A$ HW-conditional on $\mathcal{M}_u(t_0)$ to be

$$\mathbb{P}(A||\mathcal{M}_u(t_0)) := \lim_{r \to 0} \mathbb{P}(A|\mathcal{M}(U_r(t_0)), Y(t_0) > u) = \lim_{r \to 0} \frac{\mathbb{P}(A \cap \mathcal{M}(U_r(t_0)), Y(t_0) > u)}{\mathbb{P}(\mathcal{M}(U_r(t_0)), Y(t_0) > u)}$$

where for $r > 0$, $U_r(t_0)$ is the $D$-dimensional open cube of side length $r$ centred at $t_0$. We can similarly define the probability of $A$ HW-conditional on $\mathcal{M}(t_0)$ to be

$$\mathbb{P}(A||\mathcal{M}(t_0)) = \lim_{r \to 0} \frac{\mathbb{P}(A \cap \mathcal{M}(U_r(t_0)))}{\mathbb{P}(\mathcal{M}(U_r(t_0)))}.$$

Let $(V_n)_{n \in \mathbb{N}}, V$ be a sequence of real random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then HW-conditional on $\mathcal{M}_u(t_0)$ we say that $V_n$ converges HW in probability to $V$ and write

$$V_n \xrightarrow[hw]{\mathbb{P}} V$$

(where $t_0$ is taken to be implicit) if for all $\delta > 0$,

$$\mathbb{P}(\|V_n - V\| > \delta||\mathcal{M}_u(t_0)) \longrightarrow 0 \text{ as } n \to \infty.$$

We observe that

$$\mathcal{M}(U_r(t_0)) = \{\mu(U_r(t_0)) \geq 1\}.$$

where for $T \subset S$, $\mu(T)$ is the number of maxima of $Y$ that lie within $T$. In order to evaluate HW probabilities we will take advantage of the fact that on a small ball the number of local maxima that occur is zero or one with high probability. As a result the probability of at least one maximum occurring is close to the expected number of maxima. The following Lemma, which is a generalization of part of the proof of Cheng and Schwartzman (2015a)'s Theorem 2.1, formalizes this. Note that the conditions of

Cheng and Schwartzman (2015a) are satisfied given Assumption 2.1a and the fact that $Y$ is $C^3$ and $S$ is compact.

**Lemma 2.9.** *Suppose that $Y$ satisfies Assumption 2.1a, then given $A \subseteq \mathcal{F}$ and $t_0 \in int(S)$, as $r \to 0$,*

$$\mathbb{E}[\mu(U_r(t_0))1[A]] - \mathbb{P}(\mu(U_r(t_0)) \geq 1, A) = o(r^D).$$

*Proof.* Using the fact that $1 = \sum_{i=0}^{\infty} 1[\mu(U_r(t_0)) = i]$ and, for $j \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, letting $p_j = \mathbb{P}(A, \mu(U_r(t_0)) = j)$, we have:

$$\begin{aligned}
&\mathbb{E}[\mu(U_r(t_0))\, 1[A]] - \mathbb{P}(\mu(U_r(t_0)) \geq 1, A) = \\
&\qquad = \mathbb{E}\left[\sum_{j=1}^{\infty} j1[A, \mu(U_r(t_0)) = j)] - \sum_{j=1}^{\infty} 1[A, \mu(U_r(t_0)) = j)]\right] \\
&\qquad = \sum_{j=2}^{\infty}(j-1)p_j \leq \sum_{j=2}^{\infty} \frac{j(j-1)}{2} p_j = \frac{1}{2}\mathbb{E}[\mu(U_r(t_0))(\mu(U_r(t_0)) - 1)1[A]] \\
&\qquad \leq \frac{1}{2}\mathbb{E}[\mu(U_r(t_0))(\mu(U_r(t_0)) - 1)] = o(r^D)
\end{aligned}$$

where the last inequality holds as $\mu(U_r(t_0)) \in \mathbb{N}_0$ so $\mu(U_r(t_0))(\mu(U_r(t_0)) - 1) \geq 0$ and

$$\sum_{j=2}^{\infty} \frac{j(j-1)}{2} p_j \leq \sum_{i=2}^{\infty} \frac{j(j-1)}{2} \mathbb{P}(\mu(U_r(t_0)) = j) = \frac{1}{2}\mathbb{E}[\mu(U_r(t_0))(\mu(U_r(t_0)) - 1)] = o(r^D).$$

□

This result will prove to be very useful in our context as the expected number of maxima is a well studied quantity and we can take advantage of results of Adler and Taylor (2007), such as the Kac-Rice formula, in order to evaluate and bound it.

## 2.3 HW convergence of the field and its derivatives

At high thresholds the (scaled) distribution of the excess height of an excursion above the threshold is asymptotically exponential. This result is stated in Cheng and Schwartz-

man (2015a) without proof as a consequence of their Corollary 2.4. In fact, in order to show this we need to obtain the following small modification of their result (the difference being that our statement allows $w$ to depend on $u$).

**Proposition 2.10.** *Suppose that $Y$ satisfies Assumption 2.1a and is mean zero and has unit-variance. Then given $t_0 \in int(S)$, as $u \to \infty$, given $w = w(u)$,*

$$\mathbb{P}(Y(t_0) > u + w || \mathcal{M}_u(t_0)) = \frac{(u+w)^{D-1} e^{-(u+w)^2/2}(1 + O((u+w)^{-2}))}{u^{D-1} e^{-u^2/2}(1 + O(u^{-2}))}.$$

*Proof.* The proof is essentially the same as that of Cheng and Schwartzman (2015a)'s Corollary 2.4, see Section 6.2 for the details. □

Using this proposition we can prove the following theorem.

**Theorem 2.11.** *Suppose that $Y$ satisfies Assumption 2.1a and is mean zero and has unit-variance. Then given $t_0 \in int(S)$ and $x \geq 0$, as $u \to \infty$,*

$$\mathbb{P}(u(Y(t_0) - u) > x || \mathcal{M}_u(t_0)) \longrightarrow e^{-x}.$$

*Furthermore,*

$$\frac{Y(t_0)}{u} \xrightarrow[hw]{\mathbb{P}} 1.$$

*Proof.* By Proposition 2.10, as $u \to \infty$,

$$\begin{aligned}\mathbb{P}(u(Y(t_0) - u) > x || \mathcal{M}_u(t_0)) &= \mathbb{P}(Y(t_0) - u > x/u || \mathcal{M}_u(t_0)) \\ &= \frac{\left(\frac{x}{u} + u\right)^{D-1} e^{-(u+x/u)^2/2}(1 + O((u + x/u)^{-2}))}{u^{D-1} e^{-u^2/2}(1 + O(u^{-2}))} \longrightarrow e^{-x}.\end{aligned}$$

Since the scaled height above the threshold is asymptotically exponential the second result follows immediately. More formally we see that, for each $\delta > 0$ as $u \to \infty$,

$$\mathbb{P}\left(\frac{Y(t_0)}{u} > 1 + \delta || \mathcal{M}_u(t_0)\right) = \frac{(\delta u + u)^{D-1} e^{-(\delta u + u)^2/2}(1 + O((u + \delta u)^{-2}))}{u^{D-1} e^{-u^2/2}(1 + O(u^{-2}))}$$

$$\sim (1+\delta)^{D-1} e^{-\delta^2 u^2 - 2\delta u^2} \longrightarrow 0,$$

from which the second result follows as $Y(t_0) > u$ on the event $\mathcal{M}_u(t_0)$. □

As observed in Cheng and Schwartzman (2015a), these results generalize those of Nosko (1969) in that stationarity is no longer required for them to be valid. We will make use of this exponential distribution in Section 3 when we bound components of the excursion set using inner and outer ellipsoids whose volumes are proportional to the scaled height.

First however we need to investigate what the first and second derivatives are doing in the neighbourhood of the peak. To do so, let $\Lambda = \mathrm{cov}(\nabla^T Y)$, $\Delta = \mathrm{cov}(\mathbb{V}(\nabla^2 Y), \nabla^T Y)$ and $\Omega = \mathrm{cov}(\mathbb{V}(\nabla^2 Y))$, all pointwise. Suppose that the variance of $Y$ is constant, then for all $t \in S$,

$$\begin{pmatrix} Y(t) \\ \nabla Y(t) \\ \mathbb{V}(\nabla^2 Y(t)) \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma^2 & 0 & -\mathbb{V}(\Lambda(t))^T \\ 0 & \Lambda(t) & \Delta(t)^T \\ -\mathbb{V}(\Lambda(t)) & \Delta(t) & \Omega(t) \end{pmatrix}\right)$$

and as such

$$\begin{aligned} &\mathbb{V}(\nabla^2 Y(t)) | Y(t) = y, \nabla Y(t) = 0 \\ &\qquad \sim N\big(-y\mathbb{V}(\Lambda(t))/\sigma^2, \Omega(t) - \mathbb{V}(\Lambda(t))\mathbb{V}(\Lambda(t))^T/\sigma^2 - \Delta(t)\Lambda(t)^{-1}\Delta(t)^T\big). \end{aligned}$$

Here we have used the formula for the conditional normal since

$$(-\mathbb{V}(\Lambda(t)), \Delta(t)) \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \Lambda(t)^{-1} \end{pmatrix} \begin{pmatrix} y \\ 0_D \end{pmatrix} = -y\mathbb{V}(\Lambda(t))/\sigma^2$$

and

$$\begin{aligned} \Omega(t) - (-\mathbb{V}(\Lambda(t)), \Delta(t)) &\begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \Lambda(t)^{-1} \end{pmatrix} \begin{pmatrix} -\mathbb{V}(\Lambda(t))^T \\ \Delta(t) \end{pmatrix} \\ &= \Omega(t) - \mathbb{V}(\Lambda(t))\mathbb{V}(\Lambda(t))^T/\sigma^2 - \Delta(t)\Lambda(t)^{-1}\Delta(t)^T. \end{aligned}$$

If $t_0$ is a maximum of $Y$, then $\nabla Y(t_0) = 0$ so as the peak height, $Y(t_0)$, goes to infinity we would expect

$$\mathbb{V}(\nabla^2 Y(t_0))/Y(t_0) \longrightarrow -\mathbb{V}(\Lambda)/\sigma^2.$$

We need to show this holds $HW$-conditional on $t_0$ being a maximum. To this end we have the following more formal result.

**Proposition 2.12.** *Suppose that $\frac{Y}{\sigma}$ is a unit-variance field which satisfies Assumption 2.1, then HW-conditional on $\mathcal{M}_u(t_0)$, as $u \to \infty$,*

$$\nabla^2 Y(t_0)/Y(t_0) \xrightarrow[hw]{\mathbb{P}} -cov\big(\nabla^T Y(t_0)\big)$$

*and in particular,*

$$\nabla^2 Y(t_0)/u \xrightarrow[hw]{\mathbb{P}} -cov(\nabla Y(t_0)).$$

*Proof.* As above, let $\Lambda = \text{cov}(\nabla^T Y)$ and for any $\eta > 0$ and $u \in \mathbb{R}$, define the event

$$A_u(t_0) = \left\{ \left\| \frac{\nabla^2 Y(t_0)}{Y(t_0)} + \Lambda(t_0) \right\| > \eta, Y(t_0) > u \right\}.$$

Then, invoking HW conditioning,

$$\begin{aligned}
&\mathbb{P}\left( \left\| \frac{\nabla^2 Y(t_0)}{Y(t_0)} + \Lambda(t_0) \right\| > \eta || \mathcal{M}_u(t_0) \right) \\
&\qquad = \lim_{r \to 0} \frac{\mathbb{P}\left( \left\| \frac{\nabla^2 Y(t_0)}{Y(t_0)} + \Lambda(t_0) \right\| > \eta, \mathcal{M}(U_r(t_0)), Y(t_0) > u \right)}{\mathbb{P}(\mathcal{M}(U_r(t_0)), Y(t_0) > u)} \\
&\qquad = \lim_{r \to 0} \frac{\mathbb{E}[\mu(U_r(t_0)) 1[A_u(t_0)]] + o(r^D)}{\mathbb{E}[\mu(U_r(t_0)) 1[Y(t_0) > u]] + o(r^D)},
\end{aligned}$$

where we have applied Lemma 2.9 twice. The proof proceeds by bounding the numerator using Adler and Taylor (2007)'s Theorem 11.2.3. See Section 6.4 for the full details. Note that the second result follows immediately from the first result by applying Theorem 2.11. □

Because the derivative is zero at the maximum, we have the following result (noting that here we no longer require unit-variance).

**Proposition 2.13.** *Suppose that $Y$ satisfies Assumption 2.1a,b. Then HW-conditional on $\mathcal{M}_u(t_0)$, $\nabla Y(t_0) \overset{\mathbb{P}}{\underset{hw}{=}} 0$.*

*Proof.* See Section 6.5. □

Propositions 2.12 and 2.13 are important because they allow us to understand how the field behaves HW conditional on there being a peak at a given point of a given height.

# 3 Asymptotic distribution of the size of a cluster above a threshold

In this section we will derive the asymptotic HW distribution of the size of a cluster above a threshold in a mean-zero unit-variance Gaussian convolution field. To do so we require the following functional limit theorem which is proved using the results and ideas of the previous section.

**Theorem 3.1.** *Let $(Y, X, \mathcal{V}, K)$ be a mean-zero unit-variance Gaussian convolution field on $S \subset \mathbb{R}^D$ with $C^3$ kernel $K$ that satisfies Assumption 2.1, and assume (without loss of generality, shifting $S$ if necessary) that $0 \in int(S)$. Let $E \subset \mathbb{R}^D$ be a bounded set, then for any $\eta > 0$,*

$$\lim_{u\to\infty} \mathbb{P}\left(\sup_{t\in E}\left|u(Y(t/u) - Y(0)) + \frac{1}{2}t^T\Lambda(0)t\right| > \eta \,\middle|\middle|\, \mathcal{M}_u(0)\right) = 0.$$

*Proof.* Take $u$ large enough so that $B_{u^{-1}\mathrm{diam}(E)}(0) \subset S$. 0 lies within the interior of $S$ and $E$ is bounded so this is always possible and in particular it follows that the

probabilities in the limit above are well-defined for large enough $u$. Then,

$$\mathbb{P}\left(\sup_{t\in E}\left|u(Y(t/u)-Y(0))+\frac{1}{2}t^T\Lambda(0)t\right|>\eta\,\middle|\middle|\,\mathcal{M}_u(0)\right) \tag{4.2}$$

$$\leq \mathbb{P}\left(\sup_{t\in E}\left|u(Y(t/u)-Y(0))-t^T\nabla Y(0)-\frac{t^T\nabla^2 Y(0)t}{2u}\right|>\frac{\eta}{3}\,\middle|\middle|\,\mathcal{M}_u(0)\right) \tag{4.3}$$

$$+\mathbb{P}\left(\sup_{t\in E}|\nabla Y(0)t|>\frac{\eta}{3}\middle|\middle|\,\mathcal{M}_u(0)\right)+\mathbb{P}\left(\sup_{t\in E}\left|\frac{1}{2}t^T\Lambda(0)t+\frac{1}{2}\frac{t^T\nabla^2 Y(0)t}{u}\right|>\frac{\eta}{3}\middle|\middle|\,\mathcal{M}_u(0)\right). \tag{4.4}$$

Our proof will show that each of these terms tends to zero as $u\to\infty$. In each of the last two probabilities we can get rid of the supremum over $E$ as it is less than or equal to the value attained at some $t\in\overline{B}_{\mathrm{diam}(E)}(0)$. For instance, taking $h=\mathrm{diam}(E)$,

$$1\left[\sup_{t\in E}\nabla Y(0)t>\frac{\eta}{3}\right]\leq 1\left[\|\nabla Y(0)\|h>\frac{\eta}{3}\right]. \tag{4.5}$$

As such the second term is HW equal to 0 by Proposition 2.13. It similarly follows that the final term converges to zero by Proposition 2.12.

Showing that the first term converges to zero is more difficult but can be shown using similar arguments. In this case, unlike for the second two terms, it is not possible to remove the supremum from the integral as easily. Instead we Taylor expand the kernel, allowing us to taking the supremum of the kernel. See Section 6.6 for details. □

We are now is a position to prove our main results. To do so, let us first define the excursion set

$$C_u=\{t\in S: Y(t)>u\},$$

and let $c_u$ be the connected component of $C_u$ that contains the origin. We will show that $\lambda(c_u)$ is asymptotically exponential in nature, where $\lambda$ is the Lebesgue measure on $\mathbb{R}^D$. To do so we will use the results of the previous section to bound $c_u$ with an inner and an outer elliptic paraboloid whose volumes are a function of the scaled height

above the threshold. This idea was used in Wilson (1988) and Cao (1999) to obtain clustersize distributions under stationarity. Scaling space, we obtain the following.

**Theorem 3.2.** *Let $(Y, X, \mathcal{V}, K)$ be a unit-variance Gaussian convolution field on $S \subset \mathbb{R}^D$ with $C^3$ kernel $K$ that satisfies Assumption 2.1, and assume (without loss of generality, shifting $S$ if necessary) that $0 \in int(S)$. Given $\epsilon > 0$, and bounded $E \subset \mathbb{R}^D$ which contains the origin. For $u \in \mathbb{R}$, define the sets*

$$F_u^{\pm}(\epsilon, E) = \left\{ t \in E : u(Y(0) - u) \pm \epsilon > \frac{1}{2} t^T \Lambda(0) t \right\}$$

*and*

$$C_u^*(E) = \left\{ t \in E : tu^{-1} \in S \text{ and } Y(t/u) > u \right\}.$$

*L et $c_u^*(E)$ be the connected component of $C_u^*(E)$ that contains the origin. Then*

$$\lim_{u \to \infty} \mathbb{P}\big( F_u^-(\epsilon, E) \subseteq C_u^*(E) \subseteq F_u^+(\epsilon, E) || \, \mathcal{M}_u(0) \big) = 1$$

*and*

$$\lim_{u \to \infty} \mathbb{P}\big( F_u^-(\epsilon, E) \subseteq c_u^*(E) \subseteq C_u^*(E) \subseteq F_u^+(\epsilon, E) || \, \mathcal{M}_u(0) \big) \geq e^{-\epsilon}.$$

*Proof.* Given $\epsilon, \delta > 0$, applying Theorem 3.1, choose $U \in \mathbb{R}$ such that $tU^{-1} \in S$ for all $t \in E$ and such that for all $u \geq U$,

$$\mathbb{P}\left( \sup_{t \in E} \left| u(Y(t/u) - Y(0)) + \frac{1}{2} t^T \Lambda(0) t \right| > \epsilon \, || \mathcal{M}_u(0) \right) < \delta.$$

Then $t \in C_u^*(E) \implies$

$$u(Y(t/u) - Y(0)) + \frac{1}{2} t^T \Lambda(0) t > u(u - Y(0)) + \frac{1}{2} t^T \Lambda(0) t$$

so with HW probability greater than $1 - \delta$,

$$\epsilon > u(u - Y(0)) + \frac{1}{2} t^T \Lambda(0) t$$

and in particular $t \in F_u^+(\epsilon, E)$. Similarly, given $t \in F_u^-(\epsilon, E)$ we have

$$-\epsilon > u(u - Y(0)) + \frac{1}{2}t^T\Lambda(0)t$$

so with HW probability greater than $1 - \delta$,

$$u(Y(t/u) - Y(0)) + \frac{1}{2}t^T\Lambda(0)t > u(u - Y(0)) + \frac{1}{2}t^T\Lambda(0)t$$

so $Y(t/u) \geq u$ i.e. $t \in C_u^*(E)$. In particular taking $u$ large enough we can ensure that

$$\mathbb{P}\big(F_u^-(\epsilon, E) \subseteq C_u^*(E) \subseteq F_u^+(\epsilon, E) || \, \mathcal{M}_u(0)\big)$$

is as close to 1 as we like. Since $F_u^-(\epsilon, E)$ is connected, on the inner event, $F_u^-(\epsilon, E) \subseteq c_u^*(E)$ if $F_u^-(\epsilon, E)$ contains the origin which happens if and only if $u(Y(0) - u) > \epsilon$. So the second result follows by taking the limit as $u \to \infty$ and applying Theorem 2.11. $\square$

Given $h > 0$, define $E(h) = \big\{t \in \mathbb{R}^D : h \geq t^T\Lambda(0)t/2\big\}$ to be a series of compact ellipsoids centred at the origin. Given $u \in \mathbb{R}$ and $\epsilon > 0$, applying Theorem 2.11 it follows that, as $u \to \infty$,

$$\mathbb{P}(u(Y(0) - u) + \epsilon \geq h/2 || \mathcal{M}_u(0)) = \mathbb{P}(u(Y(0) - u) \geq h/2 - \epsilon || \mathcal{M}_u(0)) \to e^{-h/2+\epsilon}.$$

So taking large enough $U$ and $h$, we can ensure that this conditional probability is as small as we like for all $u > U$. As such for all $\epsilon > 0$, applying Theorem 3.2 (with $E = E(h)$) and taking large enough $U$ we can ensure that,

$$\mathbb{P}\big(F_u^-(\epsilon, E(h)) \subseteq c_u^*(E(h)) \subseteq F_u^+(\epsilon, E(h)) \subset E(h/2) || \mathcal{M}_u(0)\big) > 1 - \delta_h(\epsilon) \qquad (4.6)$$

for all $u > U$, where $\delta_h(\epsilon) = 2 - (e^{-\epsilon} + e^{-h/2+\epsilon} - \epsilon)$. Here we have used the fact that $u(Y(0) - u) + \epsilon < h/2$ implies that $F_u^+(\epsilon, E(h)) \subset E(h/2)$. On the inner event in the

probability in (4.6), $c_u^*(E(h)) \subset E(h/2)$ and so the connected component of

$$\big\{t \in \mathbb{R}^D : tu^{-1} \in S, Y(t/u) > u\big\},$$

that contains the origin, lies within $E(h/2)$ and so completely within $E(h)$; meaning that $c_u^*(E(h))$ is in fact equal to this connected component. In particular on this inner event,

$$\lambda(c_u) = \lambda(c_u^*(E(h)))/u^D.$$

As such,

$$\lim_{u\to\infty} \mathbb{P}\big(\lambda(F_u^-(\epsilon, E(h))) \subset \lambda(c_u)u^D \subset \lambda(F_u^+(\epsilon, E(h)))||\mathcal{M}_u(0)\big) \geq 1 - \delta_h(\epsilon). \tag{4.7}$$

The volumes of $F_u^{\pm}(\epsilon, E(h))$ can be obtained as a function of $u(Y(0) - u)$ which has an asymptotic exponential HW distribution. This leads to the following theorem.

**Theorem 3.3.** *Let $(Y, X, \mathcal{V}, K)$ be a mean-zero unit-variance Gaussian convolution field on $S \subset \mathbb{R}^D$ with $C^3$ kernel $K$. Then for all $t_0 \in int(S)$ and $x \geq 0$,*

$$\lim_{u\to\infty} \mathbb{P}\big(u^D 2^{-D/2}(\omega_D)^{-1} \det(\Lambda(t_0))^{1/2}\lambda(c_u) \geq x \,||\, \mathcal{M}_u(t_0)\big) = \exp(-x^{2/D}),$$

*where $\omega_D$ is the volume of the unit ball in $\mathbb{R}^D$.*

*Proof.* Without loss of generality suppose that $t_0 = 0$. For $a > 0$,

$$\lambda\left(\left\{t : a \geq \frac{1}{2}t^T\Lambda(0)t\right\}\right) = (2a)^{D/2}\omega_D \det(\Lambda(0))^{-1/2}.$$

So from (4.7) it follows that for all $\epsilon > 0$,

$$\lim_{u\to\infty} \mathbb{P}\big(\omega_D 2^{D/2}(u(Y(0) - u) - \epsilon)^{D/2} \det(\Lambda(0))^{-1/2} < u^D\lambda(c_u) \tag{4.8}$$

$$< \omega_D 2^{D/2}(u(Y(0) - u) + \epsilon)^{D/2} \det(\Lambda(0))^{-1/2}||\mathcal{M}_u(0)\big) \geq 1 - \delta_h(\epsilon) \tag{4.9}$$

As such the result follows by setting $h = h(\epsilon) = \epsilon^{-1}$ (or any function of $\epsilon$ that converges to $\infty$ as $\epsilon \to 0$) and taking the limit as $\epsilon \to 0$. More formally, let

$$A_\epsilon = \Big\{ \Big| \big(u^D \omega_D^{-1} 2^{-D/2} \det(\Lambda(0))^{1/2} \lambda(c_u)\big)^{2/D} - u(Y(0) - u) \Big| < \epsilon \Big\}.$$

Then (4.8) implies that $\lim_{u\to\infty} \mathbb{P}(A_\epsilon || \mathcal{M}_u(0)) \geq 1 - \delta_{\epsilon^{-1}}(\epsilon)$ and it follows that

$$\begin{aligned}
&\mathbb{P}\big(u^D 2^{-D/2} (\omega_D)^{-1} \det(\Lambda(t_0))^{1/2} \lambda(c_u) \geq x \,||\, \mathcal{M}_u(t_0)\big) \\
&\quad \leq \mathbb{P}\big(u^D 2^{-D/2} (\omega_D)^{-1} \det(\Lambda(t_0))^{1/2} \lambda(c_u) \geq x, A_\epsilon \,||\, \mathcal{M}_u(t_0)\big) + \mathbb{P}\big(A_\epsilon^C || \mathcal{M}_u(0)\big) \\
&\quad \leq \mathbb{P}\Big(\big(u^D 2^{-D/2} (\omega_D)^{-1} \det(\Lambda(t_0))^{1/2} \lambda(c_u)\big)^{2/D} \geq x^{2/D}, A_\epsilon \,||\, \mathcal{M}_u(t_0)\Big) + \delta_{\epsilon^{-1}}(\epsilon) \\
&\quad \leq \mathbb{P}\big(u(Y(0) - u) \geq x^{2/D} - \epsilon \,||\, \mathcal{M}_u(t_0)\big) + \delta_{\epsilon^{-1}}(\epsilon) \longrightarrow e^{\epsilon - x^{2/D}} + \delta_{\epsilon^{-1}}(\epsilon)
\end{aligned}$$

as $u \to \infty$ by Theorem 2.11. Arguing similarly to obtain a lower bound and taking the limit as $\epsilon$ tends to zero yields the result.

□

This result generalizes Theorem 6.5.1 of Adler et al. (2010) to non-stationary Gaussian random fields. Taking $\Lambda$ to be constant throughout $S$ we recover the stationary result.

# 4 Simulations

In order to validate the theory and show that the results work in practice we run simulations consisting of non-stationary, variance-one Gaussian convolution random fields. To do so we generate a random $100 \times 100$ positive definite covariance structure which we fix: this will give us the basis for our non-stationary spatial covariance structure. We use an initial lattice composed of 100 equally spaced voxels and simulate Gaussian random data on this lattice according to this covariance structure. We smooth this

data using a 1D Gaussian kernel (with a specified FWHM) to generate a convolution field which we evaluate on a fine grid (corresponding to the $r = 9$ grid discussed in Chapter 2). In order to ensure that the field is variance 1 we do this 20000 times (for each FWHM) and use these 20000 realisations to obtain the variance of at each fine grid point. This allows us to separately generate variance one simulations (by dividing by the standard deviation at each grid point). Scaling by the standard deviation and taking the derivative (via differences on the fine grid) allows us the calculate the smoothness: $\Lambda$ at each point. To illustrate our results we apply Gaussian kernels with FWHM 2 and 4 voxels respectively. We have plotted the resulting spatial smoothness estimates in Figure 4.1. These plots show that the fields are highly non-stationary - with smoothness varying throughout the image. (Note that the variation in the smoothness is not due to noise, as the smoothness is estimated using 20000 fields.) In this figure we have also plotted (variance one) sample realisations of the fields.

In order to validate the theory we have generated 10000 fields for applied FWHMs of 2 and 4 voxels. For each simulation we calculate the size of the clusters (if any) that do not intersect the edge of the domain) above a cluster defining threshold of $u = 3$. And for each such cluster we calculate the height of the maximum that the field reaches within the cluster. We have shown theoretically that we expect to be able to bound the cluster within an elliptical paraboloid. To illustrate that this holds (approximately) in practice in Figure 4.2 we plot the cluster extent against the scaled peak excursion for each observed cluster above the threshold. Mathematically this corresponds to the following. For each cluster $c$ above the threshold we calculate the location of its maximum: $s^* \in (1, 100)$, and plot its size $\lambda(c)$ against

$$\omega_1 2^{1/2} (u(Y(s^*) - u))^{1/2} \det(\Lambda(s^*))^{-1/2}. \tag{4.10}$$

The results at these thresholds show that there is an approximately linear relationship, which is in line with our results. There is still some noise in this, since the theory only holds asymptotically.

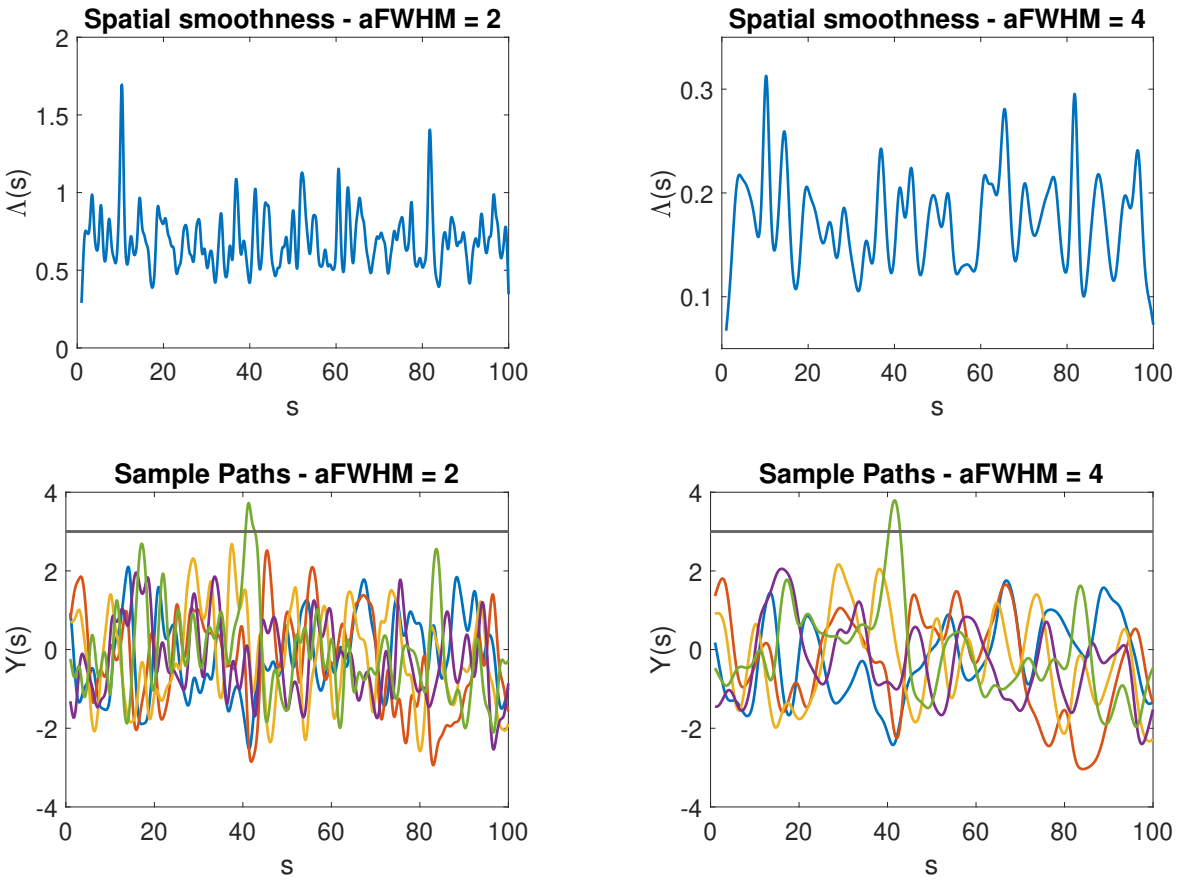



Figure 4.1: The top panels show how the smoothness varies spatially for the two scenarios. The bottom panels each show 5 example realisations of the processes and illustrate excursions above the CDT $u = 3$. Note that these have been chosen so that there was an excursion above the cluster forming threshold (for illustration). In practice in this scenario such an event occurs infrequently.

Ideally we would use a much larger cluster forming threshold to validate the theory. Unfortunately as the threshold increases the number of clusters above the threshold decreases quickly. The rate of decrease is $O(e^{-u^2/2})$; this can be seen by applying the formula for the expected Euler characteristic discussed in Chapter 2 (since at high thresholds the number of maxima equals the Euler characteristic). This means that at high thresholds a huge number of simulations are required in order to validate the

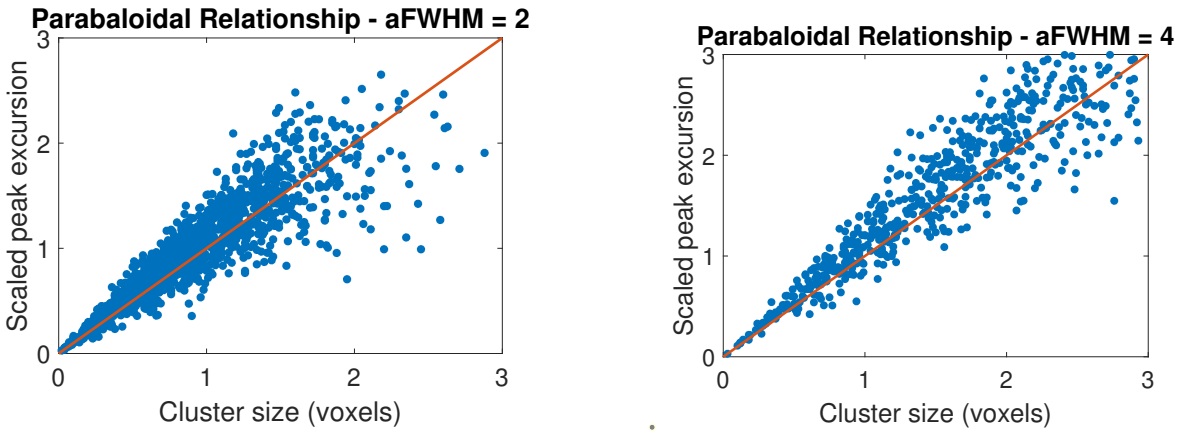



Figure 4.2: Validating theory by plotting the size of clusters above the threshold against the scaled peak excursion height defined in (4.10). The relationship is approximately linear (with noise) which supports the theory. The noise occurs because the theory is only valid asymptotically.

theory. Even in 1D this quickly becomes infeasible. This gets worse as the number of dimensions increases because evaluating the convolution field becomes expensive. In 1D the size of the clusters could be found using Newton-Raphson to determine the points at which the field intersects the level $u$ and this could help reduce the computation time. This may help, however at high thresholds the clusters are very small so in practice a fine grid is still required in order to provide an initialisation. Moreover, in higher dimensions the intersection with the threshold consists of an infinite set and is thus more difficult to easily obtain without finely evaluating the convolution field which is memory intensive. These difficulties are surmountable (to some extent) however require substantial additional computation and analysis to overcome. We plan to investigate this further in future work.

# 5 Discussion

In this work we have derived the asymptotic HW distribution for the size of a cluster above a high threshold in a unit-variance Gaussian convolution field. We hope that this

will form the basis for future work on cluster size inference in Gaussian random fields and its application in neuroimaging, however, this requires considerable further work. Clustersize inference using RFT is much faster than the main alternative: permutation testing and because it makes parametric assumptions it has the potential to be more powerful when the assumptions that it makes holds. In light of the large sample sizes which are increasingly being used in practice (resulting in very large computation time when performing permutation testing to control false positive rates), with the rise of Biobank level data (Alfaro-Almagro et al., 2018), developing a clustersize inference framework that solves the problems of Eklund et al. (2016) is of great interest. Our recent work, Davenport et al. (2021), extended the voxelwise RFT inference framework of Worsley et al. (1992), Worsley et al. (1996) so that it accurately controlled the FWER in fMRI. If it is possible (there are a number of challenges involved as clustersize inference makes a number of additional assumptions relative to voxelwise inference) the next step is to combine these results in order to provide valid clustersize inference.

Future work could extend these results by removing the unit-variance assumption and considering more general forms of Gaussian and other types of random fields (such as $\chi^2, T$ and $F$) that are useful in performing inference. Many of the clustersize results that we have proved can likely be extended to these settings, following arguments in Cao (1999). The primary difficulty lies in extending the peak height distribution (derived in Cheng and Schwartzman (2015a) for non-stationary Gaussian random fields) to these other types of fields.

# 6 Proofs

## 6.1 Vech/$\mathbb{V}$ notation

It is helpful to introduce some further notation in order to discuss the distribution of second derivatives of Gaussian field. These derivatives are $D \times D$ symmetric matrices and so when analysing their distribution we need only consider the elements in the upper triangular part of the matrix. Furthermore we will want to vectorize this so that we consider these elements but in the form of a vector: this will make our notation more compact and allow us to more easily write out joint distributions. To that end, letting $\mathrm{Sym}_D$ be the set of $D \times D$ symmetric matrices, we define the **vech** operation to be

$$\mathbb{V} : \mathrm{Sym}_D \to \mathbb{R}^{D(D+1)/2}$$

such that for $Q \in \mathrm{Sym}_D$, for $i \leq j \leq D$, $\mathbb{V}(Q)_{j(j-1)/2+i} = Q_{i,j}$.

## 6.2 Proof of Theorem 2.10

As in the proof of Cheng and Schwartzman (2015a)'s Theorem 2.3, as $x \to \infty$,

$$\mathbb{E}\big[|\det \nabla Y(t_0)| 1[\nabla^2 Y(t_0) \prec 0] \big| Y(t_0) = x, \nabla Y(t_0) = 0\big] =$$
$$= (-1)^D \mathbb{E}\big[\det \nabla^2 Y(t_0) \big| Y(t_0) = x, \nabla Y(t_0) = 0\big] + o(e^{-\alpha x^2})$$

for some constant $\alpha > 0$. As such, applying their Lemma 4.2, as $x \to \infty$,

$$\mathbb{E}\big[|\det \nabla Y(t_0)| 1[\nabla^2 Y(t_0) \prec 0] \big| Y(t_0) = x, \nabla Y(t_0) = 0\big] = \det(\Lambda(t_0)) x^D (1 + O(x^{-2}))$$

since $O(x^{-2})$ dominates $o(e^{-\alpha x^2})$. Applying their Theorem 2.2, and taking $\phi$ to be the standard normal density, it follows that

$$\begin{aligned}\mathbb{P}(Y(t_0) > u+w || \mathcal{M}_u(t_0)) &= \frac{\int_{u+w}^{\infty} \phi(x)\mathbb{E}[\det \nabla^2 Y(t_0) | Y(t_0)=x, \nabla Y(t_0)=0]\, dx}{\int_u^{\infty} \phi(x)\mathbb{E}[\det \nabla^2 Y(t_0) | Y(t_0)=x, \nabla Y(t_0)=0]\, dx} \\ &= \frac{\int_{u+w}^{\infty} x^D(1+O(x^{-2}))\phi(x)\, dx}{\int_u^{\infty} x^D(1+O(x^{-2}))\phi(x)\, dx} \\ &= \frac{(u+w)^{D-1}e^{-(u+w)^2/2}(1+O((u+w)^{-2}))}{u^{D-1}e^{-u^2/2}(1+O(u^{-2}))}.\end{aligned}$$

## 6.3 Supporting Lemmas

For the proof of Proposition 2.12, we will need the following Lemma.

**Lemma 6.1.** *For $i,j = 1,\dots,D$ let $L_{ij}$ be the Lipschitz constant of $Y_{ij}(t)$. Let $P_D$ be the set of permutations of $\{1,\dots,D\}$, then for all $t_0, t \in S$,*

$$\left|\det(\nabla^2 Y(t)) - \det(\nabla^2 Y(t_0))\right| \le \|t-t_0\| \sum_{p\in P_D}\sum_{j=1}^{D} L_{jp(j)} \prod_{i=1}^{j-1} \left|\nabla^2 Y(t)_{ip(i)}\right| \prod_{h=j+1}^{D} \left|\nabla^2 Y(t_0)_{hp(h)}\right|.$$

*Proof.* Given $p \in P_D$, let $s(p)$ denote its sign, then

$$\begin{aligned}&\left|\det(\nabla^2 Y(t)) - \det(\nabla^2 Y(t_0))\right| = \left|\sum_{p\in P_D} s(p)\left(\prod_{i=1}^{D}(\nabla^2 Y(t))_{ip(i)} - \prod_{k=1}^{D}(\nabla^2 Y(t_0))_{kp(k)}\right)\right| \\ &= \Bigg|\sum_{p\in P_D} s(p)\sum_{j=1}^{D}\big(\nabla^2 Y(t)_{1p(1)}\dots\nabla^2 Y(t)_{jp(j)}\nabla^2 Y(t_0)_{j+1p(j+1)}\dots\nabla^2 Y(t_0)_{Dp(D)} \\ &\qquad\qquad -\nabla^2 Y(t)_{1p(1)}\dots\nabla^2 Y(t)_{j-1p(j-1)}\nabla^2 Y(t_0)_{jp(j)}\dots\nabla^2 Y(t_0)_{Dp(D)}\big)\Bigg| \\ &= \left|\sum_{p\in P_D} s(p)\left(\sum_{j=1}^{D}\prod_{i=1}^{j-1}\nabla^2 Y(t)_{ip(i)}\prod_{k=j+1}^{D}\nabla^2 Y(t_0)_{kp(k)}(\nabla^2 Y(t)_{jp(j)} - \nabla^2 Y(t_0)_{jp(j)})\right)\right| \\ &\le \sum_{p\in P_D}\sum_{j=1}^{D}\prod_{i=1}^{j-1}\left|\nabla^2 Y(t)_{ip(i)}\right|\prod_{k=j+1}^{D}\left|\nabla^2 Y(t_0)_{kp(k)}\right|\left|\nabla^2 Y(t)_{jp(j)} - \nabla^2 Y(t_0)_{jp(j)})\right| \\ &\le \|t-t_0\|\sum_{p\in P_D}\sum_{j=1}^{D} L_{jp(j)}\prod_{i=1}^{j-1}\left|\nabla^2 Y(t)_{ip(i)}\right|\prod_{k=j+1}^{D}\left|\nabla^2 Y(t_0)_{kp(k)}\right|.\end{aligned}$$

□

**Lemma 6.2.** *Let $Y$ be a Gaussian random field on a compact subset $S$ of $\mathbb{R}^D$ with continuous mean and variance. Then* $\sup_{t\in S}\mathbb{E}\left[|Y(t)|^D\right] < \infty$.

*Proof.* Because of Gaussianity, $\mathbb{E}\left[|Y(t)|^D\right]$ is a continuous function of $\mathbb{E}[Y(t)]$ and $\mathrm{var}(Y(t))$ and so is a continuous function of $t$ on $S$ and is therefore bounded as $S$ is compact. □

Note that this Lemma applies under Assumption 2.1e for mean-zero fields.

To demonstrate convergence of the number of critical points we will need to define some notation and to prove the Lemma below which formalizes part of the proof of Adler and Taylor (2007)'s Theorem 11.2.6. For $\epsilon > 0$, let $\delta_\epsilon : \mathbb{R}^D \to \mathbb{R}$ be constant on $B_\epsilon(0)$ and zero elsewhere, normalized so that

$$\int_{B_\epsilon(0)} \delta_\epsilon(t)\, dt = 1.$$

Then we have the following Lemma.

**Lemma 6.3.** *Let $V$ and $W$ be real random vectors (with a well defined joint density) taking values in sets $\mathcal{X} \subseteq \mathbb{R}^D$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ respectively (some $m \in \mathbb{N}$) such that $B_k(0) \subseteq \mathcal{X}$ for some $k > 0$. Suppose that the conditional density $p_{V|W}(\cdot|w)$ is continuous at 0 for all $w \in \mathcal{Y}$ and that $\sup_{v\in B_k(0), w\in\mathcal{Y}} p_{V|W}(v|w) < \infty$. Then, given some $g : \mathcal{Y} \to \mathbb{R}$ such that $\mathbb{E}[|g(W)|] < \infty$, for any $\epsilon < k$ we have*

$$\mathbb{E}[\delta_\epsilon(V)|g(W)|] \leq \sup_{v\in B_k(0), w\in\mathcal{Y}} p_{V|W}(v|w)\mathbb{E}[|g(W)|] < \infty \text{ and}$$

$$\lim_{\epsilon\to 0}\mathbb{E}[\delta_\epsilon(V)g(W)] = \mathbb{E}[g(W)|V = 0]p_V(0).$$

*Proof.* As $\epsilon \to 0$

$$\mathbb{E}[\delta_\epsilon(V)g(W)] = \int_{\mathcal{X}\times\mathcal{Y}} \delta_\epsilon(v)g(w)p_{V,W}(v,w)\, dv\, dw$$

$$= \int_{\mathcal{Y}} g(w) p_W(w) \int_{\mathcal{X}} \delta_\epsilon(v) p_{V|W}(v|w)\, dv\, dw$$
$$\longrightarrow \int_{\mathcal{Y}} g(w) p_{V,W}(0,w)\, dw = \int_{\mathcal{Y}} g(w) p_{W|V}(w|0)\, dw\, p_V(0).$$

Since, as $\epsilon \to 0$,

$$\int_{\mathcal{X}} \delta_\epsilon(v) p_{V|W}(v|w)\, dv \longrightarrow p_{V|W}(0|w)$$

by Lebesgue's Continuity Theorem. Here we have applied the Dominated Convergence Theorem: using

$$|g(w)| p_W(w) \times \sup_{v \in B_0(k), w \in \mathcal{Y}} p_{V|W}(v|w)$$

as the dominating function. This function is measurable as the product of measurable functions and integrable by assumption. The bound on the integral follows as for any $\epsilon < k$,

$$\mathbb{E}[\delta_\epsilon(V)|g(W)|] = \int_{\mathcal{X}\times\mathcal{Y}} \delta_\epsilon(v)|g(w)| p_{V,W}(v,w)\, dv\, dw$$
$$\leq \int_{\mathcal{Y}} |g(w)| p_W(w) \int_{\mathcal{X}} \delta_\epsilon(v) p_{V|W}(v|w)\, dv\, dw \leq \sup_{v \in B_0(k), w \in \mathcal{Y}} p_{V|W}(v|w) \mathbb{E}[|g(W)|].$$

□

Note that if $V$ and $W$ are non-degenerate Gaussian vectors then $p_{V|W}(v|w)$ is continuous and bounded and so the result holds.

## 6.4 Proof of Proposition 2.12

By Theorem 11.2.3 from Adler and Taylor (2007) (the conditions of which hold by Adler's Lemmas 11.2.10, Lemmas 11.2.11 and 11.2.12 under Assumption 2.1) we have $\mu(U_r(t_0)) = \lim_{\epsilon \to 0} N_\epsilon$, where

$$N_\epsilon = \int_{U_r(t_0)} \delta_\epsilon(\nabla Y(t)) 1[\nabla^2 Y(t) \prec 0] \left|\det \nabla^2 Y(t)\right| dt.$$

It follows that

$$\mathbb{E}[\mu(U_r(t_0))1[A_u(t_0)]] = \mathbb{E}\Big[\lim_{\epsilon\to 0} N_\epsilon 1[A_u(t_0)]]\Big] \leq \lim_{\epsilon\to 0}\mathbb{E}[N_\epsilon 1[A_u(t_0)]]$$

where the inequality holds by Fatou's Lemma. Now,

$$N_\epsilon \leq \int_{U_r(t_0)} \delta_\epsilon(\nabla Y(t))\big|\det \nabla^2 Y(t)\big|\, dt.$$

and so

$$\mathbb{E}[N_\epsilon 1[A_u(t_0)]] \leq \mathbb{E}\Big[\int_{U_r(t_0)} \delta_\epsilon(\nabla Y(t))\big|\det(\nabla^2 Y(t))\big|\, dt \times 1[A_u(t_0)]\Big] \tag{4.11}$$

$$= \int_{U_r(t_0)} \mathbb{E}\big[\delta_\epsilon(\nabla Y(t))\big|\det(\nabla^2 Y(t))\big|1[A_u(t_0)]\big]\, dt \tag{4.12}$$

$$\leq \int_{U_r(t_0)} \mathbb{E}\big[\delta_\epsilon(\nabla Y(t))\big|\det(\nabla^2 Y(t_0))\big|1[A_u(t_0)]\big] \tag{4.13}$$

$$+ \mathbb{E}\big[\delta_\epsilon(\nabla(Y(t)))\big|\det(\nabla^2 Y(t)) - \det(\nabla^2 Y(t_0))\big|\big]\, dt. \tag{4.14}$$

The equality above holds by Fubini's theorem as the inner terms are non-negative. The terms in the first integral in the final expression satisfy the conditions of Lemma 6.3. These conditions follow as for each $t \in U_r(t_0)$, $\big(\nabla Y(t), Y(t_0), \mathbb{V}(\nabla^2 Y(t_0))^T\big)$ is a non-degenerate Gaussian vector (so the requisite densities are well-defined and continuous) and by integrating the bound on the conditional pdf in Assumption 2.1c to obtain the bound required for the Lemma. As such, applying the Dominated Convergence Theorem (using the bound from Lemma 6.3 and the fact that all powers of combinations of elements of $\nabla^2 Y(t_0)$ are integrable (and their expectations bounded over $S$ by Lemma 6.2) and $U_r(t_0)$ is bounded), the first integral (4.13) converges to

$$\int_{U_r(t_0)} \mathbb{E}\big[\big|\det(\nabla^2 Y(t_0))\big|1[A_u(t_0)]\big|\nabla Y(t) = 0\big] p_{\nabla Y(t)}(0)\, dt$$

as $\epsilon \to 0$. For the second integral 4.14,

$$\begin{aligned}
&\int_{U_r(t_0)} \mathbb{E}\big[\delta_\epsilon(\nabla(Y(t)))\big|\det(\nabla^2 Y(t)) - \det(\nabla^2 Y(t_0))\big|\big]\, dt \\
&\quad \leq r \sum_{p\in P_D} \sum_{j=1}^{D} \int_{U_r(t_0)} \mathbb{E}\left[\delta_\epsilon(\nabla Y(t)) L_{jp(j)} \prod_{i=1}^{j-1} |\nabla^2 Y(t)_{ip(i)}| \prod_{h=j+1}^{D} |\nabla^2 Y(t_0)_{hp(h)}|\right] dt \\
&\quad \leq r \sum_{p\in P_D} \sum_{j=1}^{D} c_{jp(j)} \sum_{l\in\mathcal{V}} \int_{U_r(t_0)} \mathbb{E}\left[\delta_\epsilon(\nabla Y(t))|X(l)| \prod_{i=1}^{j-1} |\nabla^2 Y(t)_{ip(i)}| \prod_{h=j+1}^{D} |\nabla^2 Y(t_0)_{hp(h)}|\right] dt
\end{aligned}$$

using Lemma 6.1. The requisite densities are well-defined, continuous and bounded by Assumption 2.1 and the expectation of

$$|X(l)| \prod_{i=1}^{j-1} |\nabla^2 Y(t)_{ip(i)}| \prod_{h=j+1}^{D} |\nabla^2 Y(t_0)_{hp(h)}|$$

is bounded (by Lemma 6.2, Assumption 2.1d and Adler and Taylor (2007)'s Lemma 11.2.5). As such, as $\epsilon$ tends to zero, by Lemma 6.3 and the Dominated Convergence Theorem, the upper bound converges to

$$r \sum_{p\in P_D} \sum_{j=1}^{D} c_{jp(j)} \sum_{l\in\mathcal{V}} \int_{U_r(t_0)} \mathbb{E}\left[|X(l)| \prod_{i=1}^{j-1} |\nabla^2 Y(t)_{ip(i)}| \prod_{h=j+1}^{D} |\nabla^2 Y(t_0)_{hp(h)}| \,\middle|\, \nabla Y(t) = 0\right] p_{\nabla Y(t)}(0)\, dt.$$

For each $j$ and $p$ we can bound the inner expectation, arguing as above (but with the expectation conditional on $\nabla Y(t) = 0$). As such, there is some constant $C$ such that we can bound the overall sum by

$$rD \times D! C \lambda(U_r(t_0)) \max_t p_{\nabla Y(t)}(0).$$

As such, there is some constant $C'$ such that the second integral (4.14) is bounded by $r^{D+1}C'$ and so, applying Lebesgue's continuity theorem (which we can as the inner expectation inherits continuity from the densities) to the first integral (4.13), it follows that

$$\lim_{r\to 0} \frac{1}{r^D} \mathbb{E}[\mu(U_r(t_0)) 1[A_u(t_0)]] \leq \mathbb{E}\big[\big|\det(\nabla^2 Y(t_0))\big| 1[A_u(t_0)] \big| \nabla Y(t_0) = 0\big] p_{\nabla Y(t_0)}(0).$$

Following the logic of Cheng and Schwartzman (2015a),

$$\frac{1}{r^D}\mathbb{E}[\mu(U_r(t_0))1[Y(t_0) > u]] \to \mathbb{E}\big[\big|\det \nabla^2 Y(t_0)\big|1[\nabla^2 Y(t_0) \prec 0]\big|\nabla Y(t_0) = 0\big]p_{\nabla Y(t_0)}(0) \tag{4.15}$$

as $r \to 0$. And so

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{\nabla^2 Y(t_0)}{Y(t_0)} + \Lambda(t_0)\right\| > \eta||\mathcal{M}_u(t_0)\right) \\
&\qquad \leq \frac{\mathbb{E}\Big[|\det(\nabla^2 Y(t_0))|1\Big[\Big\|\frac{\nabla^2 Y(t_0)}{Y(t_0)} + \Lambda(t_0)\Big\| > \eta, Y(t_0) > u\Big]\Big|\nabla Y(t_0) = 0\Big]}{\mathbb{E}[|\det(\nabla^2 Y(t_0))|1[Y(t_0) > u, \nabla^2 Y(t_0) \prec 0]|\nabla Y(t_0) = 0]}.
\end{aligned}$$

Let $V = \nabla^2 Y(t_0) + \Lambda(t_0)Y(t_0)$, then the distribution of $V$ conditional on $\nabla Y(t_0) = 0$ is Gaussian and doesn't depend on $Y(t_0)$ (with pdf $p'_V := p_{V|\nabla Y(t_0)=0}$). We can thus write (and expand as in Adler (1981) Theorem 6.3.1) the numerator as

$$\left|\int_u^\infty \phi(y) \int_{\mathbb{R}^{D(D+1)/2}} \det(v - y\Lambda(t_0))1[\|v\| > y\eta]p'_V(v)\,dv\,dy\right|$$

$$\begin{aligned}
&= \left|\int_u^\infty \phi(y) \int_{\mathbb{R}^{D(D+1)/2}} 1[\|v\| > y\eta] \sum_{k=0}^{D} b_k(v)y^k p'_V(v)\,dv\,dy\right| \\
&\leq \mathbb{P}(\|V\| > u\eta|\nabla Y(t_0) = 0)|b_D| \int_u^\infty y^D \phi(y)\,dy \\
&\qquad\qquad + \sum_{k=0}^{D-1} \int_{\mathbb{R}^{D(D+1)/2}} |b_k(v)|p'_V(v)\,dv \int_u^\infty y^k \phi(y)\,dy
\end{aligned}$$

where we have expanded the determinant in terms of some polynomials $b_k, k = 0, 1 \ldots, D$. The last inequality holds because $y \geq u$ implies that $1[\|v\| > y\eta] \leq 1[\|v\| > u\eta]$ and uses the fact that $b_D = (-1)^D \det(\Lambda(t_0))$ is a constant and doesn't depend on $v$ as

$$\det(v - y\Lambda(t_0)) = (-1)^D \det(\Lambda(t_0)) \det\big(y - \Lambda(t_0)^{-1}v\big).$$

Arguing as in the proof of Adler (1981) Theorem 6.3.1, we can expand the bound as

$$\det(\Lambda(t_0))\mathbb{P}(\|V\| > u\eta|\nabla Y(t_0) = 0)u^{D-1}e^{-u^2/2} + O\big(u^{-1}\big)u^{D-1}e^{-u^2/2}. \tag{4.16}$$

Once more arguing as in the proof of Adler (1981) Theorem 6.3.1, we can write

$$\mathbb{E}\big[\big|\det \nabla^2 Y(t_0)\big| 1[Y(t_0) > u, \nabla^2 Y(t_0) \prec 0]\big|\nabla Y(t_0) = 0\big]$$

as

$$a_u u^{D-1} e^{-u^2/2} + O\big(u^{-1}\big) u^{D-1} e^{-u^2/2} \tag{4.17}$$

where $a_u$ is a sequence such that

$$\det(\Lambda(t_0))\big(1 - O(u^{-1})\big) \leq a_u \leq \det(\Lambda(t_0)).$$

Taking the ratio of (4.16) and (4.17), the numerator and the denominator are dominated by the leading terms as $u \to \infty$. In particular, $\mathbb{P}(\|V\| > u\eta|\nabla Y(t_0) = 0) \longrightarrow 0$ and $a_u$ converges to $\det(\Lambda(t_0))$ as $u \to \infty$ so the ratio converges to 0 as $u \to \infty$, as required.

## 6.5 Proof of Proposition 2.13

The proof proceeds in a similar fashion to that of Proposition 2.12. For, $u > 0$ let $A'_u(t_0) = \{\|\nabla Y(t_0)\| > \eta, Y(t_0) > u\}$,

$$\mathbb{P}(\|\nabla Y(t_0)\| > \eta||\mathcal{M}_u(t_0)) = \lim_{r\to 0} \frac{\mathbb{E}[\mu(U_r(t_0))1[A'_u(t_0)]] + o(r^D)}{\mathbb{E}[\mu(U_r(t_0))1[Y(t_0) > u]] + o(r^D)}. \tag{4.18}$$

Arguing as before, we need to evaluate

$$\lim_{\epsilon\to 0} \int_{U_r(t_0)} \mathbb{E}\big[\delta_\epsilon(\nabla Y(t))\big|\det(\nabla^2 Y(t))\big| 1[A'_u(t_0)]\big]\, dt. \tag{4.19}$$

in order to obtain an upper bound. Now,

$$\begin{aligned} 1[A'_u(t_0)] = 1\Big[\|\nabla Y(t_0)\| > \eta, \|\nabla Y(t_0) - \nabla Y(t)\| > \frac{\eta}{2}, Y(t_0) > u\Big] \\ + 1\Big[\|\nabla Y(t_0)\| > \eta, \|\nabla Y(t_0) - \nabla Y(t)\| \leq \frac{\eta}{2}, Y(t_0) > u\Big] \end{aligned}$$

$$\leq 1\Big[L'\|t_0 - t\| > \frac{\eta}{2}, Y(t_0) > u\Big] + 1\Big[\|\nabla Y(t)\| > \frac{\eta}{2}, Y(t_0) > u\Big]$$
$$\leq 1\Bigg[|X(l)|\|t_0 - t\| > \frac{\eta}{2|\mathcal{V}|c'} \text{ for some } l \in \mathcal{V}, Y(t_0) > u\Bigg] + 1\Big[\|\nabla Y(t)\| > \frac{\eta}{2}, Y(t_0) > u\Big]$$
$$\leq \sum_{l\in\mathcal{V}} 1\Bigg[|X(l)|\|t_0 - t\| > \frac{\eta}{2|\mathcal{V}|c'}, Y(t_0) > u\Bigg] + 1\Big[\|\nabla Y(t)\| > \frac{\eta}{2}, Y(t_0) > u\Big].$$

As such we can bound the integral in (4.19) by

$$\sum_{l\in\mathcal{V}} \int_{U_r(t_0)} \mathbb{E}\Bigg[\delta_\epsilon(\nabla Y(t))\big|\det(\nabla^2 Y(t))\big|1\Bigg[|X(l)|\|t_0 - t\| > \frac{\eta}{2|\mathcal{V}|c'}, Y(t_0) > u\Bigg]\Bigg]\, dt$$
$$+ \int_{U_r(t_0)} \mathbb{E}\Bigg[\delta_\epsilon(\nabla Y(t))\big|\det(\nabla^2 Y(t))\big|1\Bigg[\|\nabla Y(t)\| > \frac{\eta}{2|\mathcal{V}|c'}, Y(t_0) > u\Bigg]\Bigg]\, dt.$$

For $\epsilon < \frac{\eta}{2|\mathcal{V}|c'}$ the second term is 0. Thus, using Assumption 2.1 and applying Lemmas 6.2 and 6.3 and the Dominated Convergence Theorem, the limit, (4.19), is bounded above by

$$\sum_{l\in\mathcal{V}} \int_{U_r(t_0)} \mathbb{E}\Bigg[\big|\det(\nabla^2 Y(t))\big|1\Bigg[|X(l)|\|t_0 - t\| > \frac{\eta}{2|\mathcal{V}|c'}, Y(t_0) > u\Bigg]\Bigg|\nabla Y(t) = 0\Bigg] p_{\nabla Y(t)}(0)\, dt$$
$$\leq \sup_{s\in S} \mathbb{E}\big[\det(\nabla^2 Y(s))^2\big|\nabla Y(s) = 0\big]^{\frac{1}{2}} \sum_{l\in\mathcal{V}} \int_{U_r(t_0)} \mathbb{P}\Bigg(|X(l)| > \frac{\eta}{2r|\mathcal{V}|c'}\Bigg|\nabla Y(t) = 0\Bigg)^{\frac{1}{2}} p_{\nabla Y(t)}(0)\, dt$$
$$\leq \sup_{s\in S} \mathbb{E}\big[\det(\nabla^2 Y(s))^2\big|\nabla Y(s) = 0\big] \frac{2r^{D+\frac{1}{2}}|\mathcal{V}|c'}{\eta^{\frac{1}{2}}} \sum_{l\in\mathcal{V}} \sup_{t\in U_r(t_0)} \mathbb{E}[|X(l)||\nabla Y(t) = 0]^{\frac{1}{2}} p_{\nabla Y(t)}(0).$$

The middle inequality follows by Adler and Taylor (2007)'s Lemma 11.2.5 and the final inequality follows by Markov's Theorem. The suprema are bounded above so this term is $o(r^D)$ as $r \to 0$. As such

$$r^{-D}\mathbb{E}[\mu(U_r(t_0))1[A'_u(t_0)]] \longrightarrow 0$$

and so it follows that $\mathbb{P}(A'_u(t_0)||\mathcal{M}_u(t_0))$ equals 0. Here we have used the fact that the limit as $r \to 0$, of $r^{-D}$ times the denominator of (4.18), equals

$$\mathbb{E}\big[\big|\det \nabla^2 Y(t_0)\big|1[Y(t_0) > u, \nabla^2 Y(t_0) \prec 0]\big|\nabla Y(t_0) = 0\big]$$

(as in (4.15)) which is positive and therefore non-zero.

## 6.6 Proof of Theorem 3.1

*Proof.* We can expand the first term (4.3) as

$$\lim_{r\to 0}\frac{\mathbb{P}\Big(\sup_{t\in E}\Big|u(Y(t/u)-Y(0))-t^T\nabla Y(0)-\frac{t^T\nabla^2 Y(0)t}{2u}\Big|>\frac{\eta}{3},\mathcal{M}_u(U_r(0)),Y(0)>u\Big)}{\mathbb{P}(\mathcal{M}(U_r(0)),Y(0)>u)}. \tag{4.20}$$

$Y$ is a convolution field (by assumption) and so Taylor expanding its kernel $K$ about $s_0\in S$, for each $s\in S$,

$$\begin{aligned}
Y(s)&=\sum_{l\in\mathcal{V}}K(s-l)X(l)\\
&=\sum_{l\in\mathcal{V}}K(s_0-l)X(l)+(s-s_0)^T\sum_{l\in\mathcal{V}}\nabla K(s_0-l)X(l)\\
&\qquad+\frac{1}{2}(s-s_0)^T\sum_{l\in\mathcal{V}}\nabla^2K(s_0-l)X(l)(s-s_0)\\
&\qquad+\frac{1}{6}\sum_{l\in\mathcal{V}}\sum_{i,j,k=1}^{D}(s^*-s_0)_i(s^*-s_0)_j(s^*-s_0)_k\nabla^3K(s^*-l)_{ijk}X(l)\\
&=Y(s_0)+(s-s_0)^T\nabla Y(s_0)+\frac{1}{2}(s-s_0)^T\nabla^2Y(s_0)(s-s_0)\\
&\qquad+\frac{1}{6}\sum_{l\in\mathcal{V}}\sum_{i,j,k=1}^{D}(s^*-s_0)_i(s^*-s_0)_j(s^*-s_0)_k\nabla^3K(s^*-l)_{ijk}X(l)
\end{aligned}$$

for some $s^*\in B_{s_0}(\|s-s_0\|)$. Taking $s=tu^{-1}$ and expanding about $s_0=0$, we see that

$$\sup_{t\in E}\Big|u(Y(t/u)-Y(0))-t^T\nabla Y(0)-\frac{t^T\nabla^2Y(0)t}{2u}\Big| \tag{4.21}$$

$$=\frac{u}{6}\sup_{t\in E}\Big|\sum_{l\in\mathcal{V}}\sum_{i,j,k=1}^{D}s_i^*(t)s_j^*(t)s_k^*(t)\nabla^3K(s^*(t)-l)_{ijk}X(l)\Big| \tag{4.22}$$

where for each $t\in E$, $s^*(t)\in B_0(\|tu^{-1}\|)\subseteq B_0(hu^{-1})$, recalling that $h=\mathrm{diam}(E)$. We

can bound (4.22) by

$$\frac{1}{6u^2}\sum_{l\in\mathcal{V}}|X(l)|\sup_{s\in\bar{B}_0(hu^{-1}),ijk}\left|\nabla^3 K(s-l)_{ijk}\right|\sup_{t\in\bar{B}_0(h)}p(|t|)\leq Mu^{-2}\sum_{l\in\mathcal{V}}|X(l)|$$

where $p$ is a $D$-dimensional degree 3 multinomial. Here $M$ is a constant independent of $u$ which exists because of continuity of $p$ and $\nabla^3 K$ and compactness. As such we can bound the numerator within the limit in (4.20) by

$$\mathbb{P}\left(Mu^{-2}\sum_{l\in\mathcal{V}}|X(l)|>\frac{\eta}{3},\mathcal{M}(U_r(0)),Y(0)>u\right).$$

Now, as in the proof of Proposition 2.12, applying Lemma 2.9,

$$\lim_{r\to 0}\frac{\mathbb{P}(A'_u,\mathcal{M}_u(U_r(0)))}{\mathbb{P}(\mathcal{M}_u(U_r(0)),Y(0)>u)}=\lim_{r\to 0}\frac{\mathbb{E}[\mu(U_r(0))1[A'_u]]+o(r^D)}{\mathbb{E}[\mu(U_r(0))1[Y(0)>u]]+o(r^D)} \tag{4.23}$$

where $A'_u=\left\{Mu^{-2}\sum_{l\in\mathcal{V}}|X(l)|>\frac{\eta}{3},Y(0)>u\right\}$ and in particular,

$$\mathbb{E}[\mu(U_r(0))1[A'_u]]\leq\lim_{\epsilon\to 0}\mathbb{E}[N_\epsilon 1[A'_u]].$$

where $N_\epsilon$ is defined as in the proof of Proposition 2.12. For $\epsilon>0$,

$$\mathbb{E}[N_\epsilon 1[A'_u]]\leq\int_{U_r(0)}\mathbb{E}\left[\delta_\epsilon(\nabla Y(t))\left|\det(\nabla^2 Y(t))\right|1[A'_u]\right]dt.$$

Arguing as in the proof of Proposition 2.13,

$$1\left[Mu^{-2}\sum_{l\in\mathcal{V}}|X(l)|>\frac{\eta}{3},Y(0)>u\right]\leq\sum_{l\in\mathcal{V}}1\left[Mu^{-2}|X(l)|>\frac{\eta}{3|\mathcal{V}|},Y(0)>u\right]$$

and so the integral is bounded by

$$\sum_{l\in\mathcal{V}}\int_{U_r(0)}\mathbb{E}\left[\delta_\epsilon(\nabla Y(t))\left|\det(\nabla^2 Y(t))\right|1\left[Mu^{-2}|X(l)|>\frac{\eta}{3|\mathcal{V}|},Y(0)>u\right]\right]dt.$$

The terms inside the integrals satisfy the conditions of Lemma 6.3 and so this converges

to

$$\sum_{l\in\mathcal{V}}\int_{U_r(0)}\mathbb{E}\bigg[\big|\det(\nabla^2 Y(t))\big|1\bigg[Mu^{-2}|X(l)|>\frac{\eta}{3|\mathcal{V}|},Y(0)>u\bigg]\bigg|\nabla Y(t)=0\bigg]p_{\nabla Y(t)}(0)\,dt.$$

By Lebesgue's Continuity Theorem (which we can apply as the inner expectation inherits continuity from the densities) it follows that

$$\lim_{r\to 0} r^{-D}\mathbb{E}[\mu(U_r(0))1[A'_u]]$$

is bounded above by

$$\sum_{l\in\mathcal{V}}\mathbb{E}\bigg[\big|\det(\nabla^2 Y(0))\big|1\bigg[Mu^{-2}|X(l)|>\frac{\eta}{3|\mathcal{V}|},Y(0)>u\bigg]\bigg|\nabla Y(0)=0\bigg]p_{\nabla Y(0)}(0)\,dt.$$

For each $l\in\mathcal{V}$, let $V=\nabla^2 Y(t_0)+\Lambda(t_0)Y(t_0)$ and $W=X(l)-a_l Y(0)$ where $a_l=\operatorname{cov}(Y(0),X(l))$. $Y(0)$ is independent of $V$ and $W$ and so, arguing as in the proof of Proposition 2.12, we can bound the leading term of the $l$th expectation by

$$\int_u^\infty y^D\phi(y)\int_{\mathbb{R}^{D(D+1)/2}}\bigg(1\bigg[M|w|>\frac{\eta u^2}{6|\mathcal{V}|}\bigg]+1\bigg[M|a_l y|>\frac{\eta u^2}{6|\mathcal{V}|}\bigg]\bigg)|b_D|p'_{V,W}(v,w)\,dw\,dv\,dy$$

where $p'_{V,W}$ is the joint pdf of $V$ and $W$ conditional on $\nabla Y(0)=0$. For sufficiently large $u$ (so that $\frac{\eta u^2}{6M|\mathcal{V}||a_l|}>u$), this can be written as

$$\det(\Lambda(t_0))\mathbb{P}\bigg(M|W|>\frac{\eta u^2}{6|\mathcal{V}|}\bigg|\nabla Y(t_0)=0\bigg)u^{D-1}e^{-u^2}$$
$$+\det(\Lambda(t_0))\bigg(\frac{\eta u^2}{6|\mathcal{V}|Ma_l}\bigg)^{D-1}e^{\frac{-\eta^2u^4}{72|\mathcal{V}|^2M^2a_l^2}}.$$

Expressing the denominator of (4.23) as in the proof of Proposition 2.12 and taking the limit as $u\to 0$ it follows that the first term, i.e. (4.20), converges to 0. Combining this with the results for the second and the third term yields the overall result. □

# 7 Acknowledgments

I am immensely grateful to Robert J. Adler, Dan Cheng, Fabian Telschow, Thomas E. Nichols, and Armin Schwartzman for helpful discussions on Random Field Theory.

# Chapter 5

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

Samuel Davenport, Thomas E. Nichols

Now published in Neuroimage (2020)

**Abstract**

The spatial signals in neuroimaging mass univariate analyses can be characterized in a number of ways, but one widely used approach is peak inference: the identification of peaks in the image. Peak locations and magnitudes provide a useful summary of activation and are routinely reported, however, the magnitudes reflect selection bias as these points have both survived a threshold and are local maxima. In this paper we propose the use of resampling 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, an analysis produces a test statistic at each voxel, and there are then a number of inference methods available to assess the evidence for a true activation. Voxel, peak and cluster level inference are the most common.[1] When papers report the effect size at a peak it is biased due to the winner's curse. This bias is typically caused by two factors, firstly the observed peaks have been chosen such that they lie above a threshold and secondly the value at each peak is the largest value in a local region around the peak (though any type of threshold or selection of the peaks based on their magnitude or a another correlated quantity has the potential to cause them to be biased). In order to determine the true effect sizes we have to account for this bias.

This issue is well-known in neuroimaging and is called circular inference or double dipping (Kriegeskorte et al., 2009). Vul et al. (2009) conducted a review and found the problem to be widespread in the fMRI literature, 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 main existing solution to this problem in neuroimaging is data-splitting, where

[1] In voxelwise inference voxels with test statistic values lying above a multiple testing threshold are determined to be significant. In both peak and cluster level inference a primary threshold is used to identify peaks/clusters and then thresholding based on peak magnitude and cluster extent is used to determine significant peaks/clusters.

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. (2010), Kriegeskorte et al. (2009). While this produces unbiased values, the estimates have larger variance as they are calculated using only half of the data. For the same reason, the locations of local maxima will be less accurate than if they had been calculated using the whole dataset. These problems are especially serious when the sample sizes are small. A widely used alternative is to select a voxel or ROI a priori based on previous studies and to only calculate the effect at that location. While this approach is unbiased it has the disadvantage that only the pre-specified voxel or ROI can be considered, and not the peaks found in the observed data. Instead if it were possible to use all of the data to estimate locations and effect sizes whilst still obtaining unbiased point estimates of the signal magnitude we would obtain much more accurate estimates of the peak locations. 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 et al. (2016) and in particular Taylor and Tibshirani (2015) for a good overview.

A similar problem arises in genetics, see Göring et al. (2001), and there has been much recent work on correcting for selection in this setting. Zhong and Prentice (2008), Ghosh et al. (2008) and Xiao and Boehnke (2011) consider pointwise correction by calculating the distribution of the effect statistic conditional on it being significant while Zhou and Wright (2015), Sun and Bull (2005), Wu et al. (2006), Yu et al. (2007) and Jeffries (2006) consider resampling based approaches. In the imaging literature, Rosenblatt and Benjamini (2014) propose a selective inference approach to 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, analogous to the genetics pointwise correction discussed above. 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 however this approach has the disadvantage that each instance of resampled data has a different estimate of the significant locations meaning that these are not identifiable. 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. We use all of the data to determine significant locations, meaning that these locations are consistent across resamples and relate to the original statistic image used for inference.

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 and Stein (1961) who introduced the famous James-Stein estimator. More recently there has been work to correct for the bias in estimating the means of the largest observed values of a given distribution. Efron (2011) uses 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 which involves calculating the distribution of Gaussian random variables conditional on being selected. Using the bootstrap to correct for bias is an idea original due to Efron and Tibshirani (1986), see Efron and Tibshirani (1994) for more details.

Brain imaging data is more complicated than these other settings as it has complex spatial and temporal dependencies. However, for group analyses 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. Our approach is based on an extension of Simon and Simon (2013) to account for dependence proposed by Tan et al. (2014), where a non-parametric bootstrap is used to estimate bias in effect sizes, motivated by a genetics application. We provide a detailed framework for this method and show how it can be applied in the context of neuroimaging. 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 the mean, Cohen's $d$ and $R^2$, quantities that are essential for power analyses and inference. See Mumford (2012) for an overview and Appendix 7.3 for the mathematical details of how to use power analyses in neuroimaging. Unbiased peak estimates are also very important when performing certain types of meta-analysis Radua et al. (2012).

We use functional and structural magnetic resonance images (MRI) from 8,940 subjects from 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 4,000 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 sort 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 raw effect (e.g. %BOLD mean where BOLD stands for the Blood-oxygen-level-dependent signal) 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 2.4

discusses the methods used for big data evaluation. Section 3.1 illustrates the methods on simulated data and Section 3.2 applies the techniques to one-sample analysis of functional imaging data and GLM analysis of structural gray matter data obtained using voxel based morphometry (VBM). In Section 3.3 we apply our 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.

Software to implement the methods and generate figures is available at `https://sjdavenport.github.io/software/`. Simulations and thresholding were conducted using the RFTtoolbox (`https://github.com/sjdavenport/RFTtoolbox`). See Appendix 5 for details.

# 2 Methods

Let $\mathcal{V}$ be the set of voxel locations corresponding to the brain or some subset under study and define an **image** to be a map which takes voxel locations to intensities. Given an image $Z$ and a connectivity criterion that determines the neighbours of each voxel (we use a connectivity criterion of 18 in our 3D analyses), define a **local maxima** or **peak** of $Z$ to be a voxel such that the value that $Z$ takes at that location is larger than the value $Z$ takes at neighbouring voxels; see Section 8.7 of the supplementary material for a rigorous definition of this.

## 2.1 One-Sample

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

$$Y_n(v) = \mu(v) + \epsilon_n(v) \tag{5.1}$$

where $\mu(v)$ is the common mean intensity, and the noise terms $\epsilon_1, \ldots, \epsilon_n$ are i.i.d mean zero random images from some unknown 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 $k$th largest local maximum of $\hat{\mu}$ above a screening threshold $u$. For each $k$, we are interested in inferring on $\mu(\hat{v}_k)$, the value of $\mu$ at the location $\hat{v}_k$, improving on the biased circular estimate $\hat{\mu}(\hat{v}_k)$.

### 2.1.1 Peak Estimation

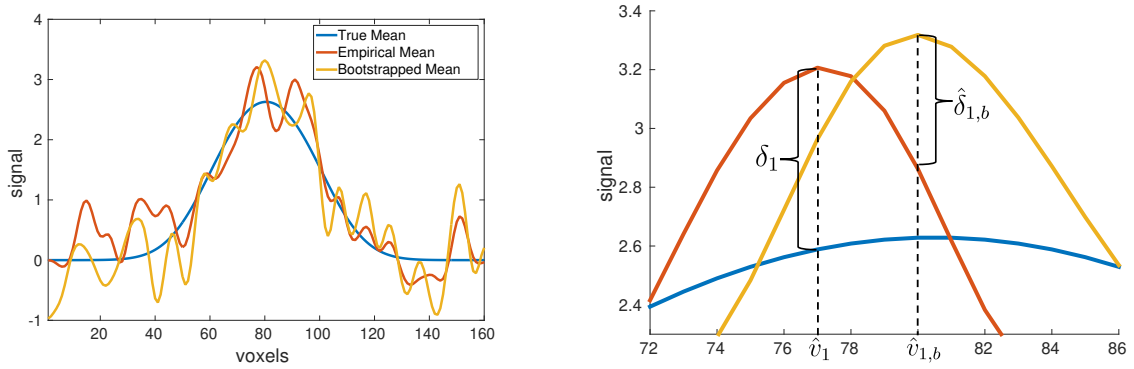
The noise distribution in model (5.1) is unknown so in order to estimate the bias of $\hat{\mu}(\hat{v}_k)$ we use the data to generate bootstrap samples without making any distributional assumptions. This allows us to obtain an estimate of the bias for each bootstrap iteration as in Tan et al. (2014). For each maxima $\hat{v}_k$ we estimate the bias-corrected value as $\tilde{\mu}(\hat{v}_k) = \hat{\mu}(\hat{v}_k) - \delta_k$, where $\delta_k$ are bias correction terms found as described in Algorithm 3 below. See Table 5.1 for a variable key.

**Algorithm 3** Non-Parametric Bootstrap Bias Calculation

1: **Input**: Images $Y_1, \ldots, Y_N$, the number of bootstrap samples $B$ and screening 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 $u$, and for $k = 1, \ldots, K$, let $\hat{v}_k$ be the location of the $k$th 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 $\hat{\mu}_b = \frac{1}{N}\sum_{n=1}^{N} Y^*_{N,b}$ and for $k = 1, \ldots, K$, let $\hat{v}_{k,b}$ be the location of the $k$th 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 $k$th 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 5.1 provides an illustrative 1D simulation on a grid of 160 voxels where we consider the case $k = 1$: the global maximum. 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 a 6 voxel FWHM Gaussian kernel. The bias above the noise-free signal ($\delta_1$) is evident, and is estimated by comparing a bootstrap sample to the original, yielding an estimate of $\hat{\delta}_{1,b}$. $\delta_1$ is the bias of the empirical mean relative to the true mean.




(a) True vs. sample and bootstrap

(b) Zooming in on the peak

Figure 5.1: Illustration of our bootstrap peak bias correction method on a simple annotated example. Here our set of voxels is $\mathcal{V} = \{1, \ldots, 160\}$, the true mean $\mu$ 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 figure on the right is a zoomed in version of the figure on the left from voxel 72 to 86. The top peak is biased above the true value by $\delta_1$. Using this realization the height of the bootstrapped peak is compared to the height of the empirical mean at the location of the peak ($\hat{v}_{1,b}$), resulting in an estimate $\hat{\delta}_{1,b}$ of the bias.

### 2.1.2 Peak Estimation for Effect Size

While the above method is based on the sample mean, neuroimaging studies typically base their inferences on statistic images. In the simplest setting, a one-sample analysis of fMRI contrast data, we are testing

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

at each $v \in \mathcal{V}$, using the statistic image, in order to determine whether there is an activation at that voxel. Given the voxels that have been determined to be active we

are interested in estimating two different quantities, the effect size (measured in terms of Cohen's $d$) and the raw unit, i.e. %BOLD change for fMRI.

We first need to define the test-statistic image. Define $\sigma^2$ to be the population variance image, estimated in an unbiased manner using

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

In order to perform one-sample hypothesis testing, the $t$-statistic

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

is typically used. If the data are Gaussian at each voxel this follows a $t$-distribution with $N-1$ degrees of freedom.

For a voxel $v$, $H_0(v)$ is rejected if $t(v)$ lies above 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 that 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).

While ubiquitously reported, $t$-statistic values are not interpretable across studies, as they depend on the sample size and grow to infinity with $N$. Good practice, and in particular to facilitate power analyses (see Appendix 7.3), requires computation of a standardized effect size such as Cohen's $d$, which at each voxel $v$ is defined as

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

As this is just the one-sample $t$-statistic divided by $\sqrt{N}$, the peaks in the $t$-statistic image will be at the same locations as those of the one-sample Cohen's $d$. Algorithm

4 describes how to compute bias-corrected estimates of Cohen's $d$ peaks, which we evaluate with simulated data (Section 3.1.1) and real task fMRI data (Section 3.2.1). The one-sample Cohen's $d$ is a biased estimator for the population Cohen's $d$:

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

with $\mathbb{E}[\hat{d}] = C_N d$ where $C_N$ is a correction factor that depends on the degrees of freedom. $C_N \longrightarrow 1$ as $n \longrightarrow \infty$ but for finite samples we need to account for it; see Appendix 7.3 for details.

**Algorithm 4** Non-Parametric Bootstrap Bias Calculation

1: **Input**: Images $Y_1, \ldots, Y_N$, the number of bootstrap samples $B$ and threshold $u$.
2: Compute mean and standard deviation images, $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} Y_n$ and $\hat{\sigma}$ with $\hat{\sigma}^2(v) = \frac{1}{N-1}\sum_{n=1}^{N}(Y_n(v) - \hat{\mu}(v))^2$ for each $v \in \mathcal{V}$.
3: Let $K$ be the number of peaks of $t$ above $u$ and for $k = 1, \ldots, K$, let $\hat{v}_k$ be the location of the $k$th largest maxima of $\hat{d} = \hat{\mu}/\hat{\sigma}$.
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 $k$th largest local maxima of $\hat{d}_b = \hat{\mu}_b/\hat{\sigma}_b$.
8: Let $\hat{\delta}_{k,b} = (\hat{d}_b(\hat{v}_{k,b}) - \hat{d}(\hat{v}_{k,b}))/C_N$ 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** $(\hat{d}(\hat{v}_1)/C_N - \hat{\delta}_1, \ldots, \hat{d}(\hat{v}_K)/C_N - \hat{\delta}_K)$.

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

In fMRI the underlying mean $\mu$ corresponds to the true %BOLD signal which is the expected value of the contrast image for each subject, see Mumford and Nichols (2009) for more details. We assume that first level models have been fit to give contrast images for each subject. At the first level a number of pre-processing steps are implemented including registration, motion correction, and normalization which affect the %BOLD signal. 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 with interpretable units, i.e. %BOLD. In order to estimate the mean while still controlling for false positives one needs to use the $t$-statistic image to identify significant peaks and then estimate the raw unit (e.g. %BOLD) change at these locations. This is easily accomplished with a small modification to Algorithm 4, computing in Step 8 instead a bias in raw effect units of:

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

and returning $(\hat{\mu}(\hat{v}_1) - \hat{\delta}_1, \ldots, \hat{\mu}(\hat{v}_K) - \hat{\delta}_K)$ instead. See Section 3.1.2 for simulated evaluations of this approach and Section 3.2.2 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 compare the bootstrap approach to circular inference (no correction) and data-splitting, the main two approaches used in the literature. After finding the number of peaks above the threshold as in Algorithm 4, the circular inference uncorrected estimates are simply $\hat{d}(\hat{v}_1)/C_N, \ldots, \hat{d}(\hat{v}_K)/C_N$.

Data-splitting proceeds 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 $\hat{d}_1$ and $\hat{d}_2$ be the image estimates of Cohen's $d$ from the first and second half of the subjects respectively. Using a threshold $u$ we find the peaks of the one-sample $t$-statistic $\hat{d}_1\sqrt{N/2}$ that lie above $u$, at locations $\hat{w}_1, \ldots, \hat{w}_J$ for some number of peaks $J$ (note that $u$ must be adjusted to account for the fact we are using half the data). The data-splitting estimates of the peak values are $\hat{d}_2(\hat{w}_1)/C_{N/2}, \ldots, \hat{d}_2(\hat{w}_J)/C_{N/2}$. See Figure 5.4 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 not already explained by other terms in the model. Hence we now generalize our method to obtain corrected estimates of the peak partial $R^2$.

Let $Y : \mathcal{V} \to \mathbb{R}^N$ be a random image such that for each $v \in \mathcal{V}$, we assume the following linear model,

$$Y(v) = X\beta(v) + \epsilon(v), \tag{5.2}$$

for an $N \times p$ design matrix $X$ and a parameter vector $\beta(v) \in \mathbb{R}^p$ where $\epsilon$ is the random image of the noise such that $\epsilon(v) = (\epsilon_1(v), \ldots, \epsilon_N(v))^T$ for each $v \in \mathcal{V}$ (where the $\epsilon_i$ are i.i.d zero mean zero random images). 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 $C \in \mathbb{R}^{m \times p}$ where $m$ is the number of contrasts that we simultaneously test. We can test this at each voxel with the usual $F$-test, which at

each voxel $v$ is

$$F(v) = \frac{(C\hat{\beta}(v))^T (C(X^T X)^{-1} C^T)^{-1} (C\hat{\beta}(v))/m}{\hat{\sigma}(v)^2} \tag{5.3}$$

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

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 estimate of the partial $R^2$ at local maxima. See Appendix 7.1 for details on how partial $R^2$ is formally defined. Bootstrapping in the general linear model scenario is based on the residuals; see Davison et al. (2003) Chapter 6. This leads to Algorithm 5.

**Algorithm 5** Non-Parametric Bootstrap Bias Calculation

1: **Input**: Images $Y_1, \ldots, Y_N$, the number of bootstrap samples $B$ and threshold $u$.
2: 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 $k$th largest maxima of $F$.
3: Let $\hat{\beta} = \hat{\beta}(X, Y) = (X^T X)^{-1} X^T Y$ and let $\hat{\epsilon} = Y - X\hat{\beta}$ be the residuals.
4: For each $n = 1, \ldots, N$, let $r_n = \hat{\epsilon}_n / \sqrt{1 - p_n}$ be the modified residuals, where $p_n = (X(X^T X)^{-1} X^T)_{nn}$. Let $\overline{r} = \frac{1}{N} \sum_{n=1}^{N} r_i$ be their mean.
5: **for** $b = 1, \ldots, B$ **do**
6: 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 $Y^*_b = X\hat{\beta} + \epsilon^*$.
7: Let $F^*_b$ be the bootstrapped $F$-statistic image computed using $Y^*_b$ and $\hat{\beta}(X, Y^*_b)$ (in both numerator and denominator of equation 5.3). For $k = 1, \ldots, K$, let $\hat{v}_{k,b}$ be the location of the $k$th largest local maxima of $F^*_b$. Let $R^2_b$ be the bootstrapped partial $R^2$ image and set $\hat{\delta}_{k,b} = R^2_b(\hat{v}_{k,b}) - R^2(\hat{v}_{k,b})$ to be the estimate of the bias.
8: **end for**
9: For $k = 1, \ldots, K$, let $\hat{\delta}_k = \frac{1}{B} \sum_{b=1}^{B} \hat{\delta}_{k,b}$.
10: **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

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

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 3.2.3 for validation of the use of the bootstrap and comparisons between the methods in a GLM scenario where gray matter images are regressed against the age of the participants and an intercept. Note that there is no known analogous correction factor $C_N$ for $R^2$ and so even the data-splitting estimates will not be completely unbiased as estimates for the population $R^2$. However as can be seen from implementation of the algorithms in simulations (see Supplementary Material Figures 5.19, 5.20) and on real data this bias is comparatively small.

| **Variable** | **Definition** |
|---|---|
| $\hat{\mu}, \hat{\mu}_b$ | $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Y_n$ and $\hat{\mu}_b$ is the $b$th bootstrapped version of $\hat{\mu}$. |
| $\hat{\sigma}^2, \hat{\sigma}_b^2$ | $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(Y_n - \hat{\mu})^2$ and $\hat{\sigma}_b^2$ is the $b$th bootstrapped version of $\hat{\mu}$. |
| $\hat{d}, \hat{d}_b$ | $\hat{d} = \hat{\mu}/\hat{\sigma}$ and $\hat{d}_b = \hat{\mu}_b/\hat{\sigma}_b$ |
| $R^2, R_b^2$ | $R^2$ is the partial $R^2$ image and $R_b^2$ is the $b$th bootstrapped version of $R^2$. |
| $\hat{\delta}_{k,b}$ | $\hat{\delta}_{k,b}$ is the $b$th bootstrap estimate of the bias at the $k$th largest peak. |
| $\hat{\delta}_k$ | $\hat{\delta}_k = \frac{1}{B}\sum_{b=1}^{B} \hat{\delta}_{k,b}$ is the estimate of the bias at the $k$th largest peak. |
| $\hat{v}_k$ | $\hat{v}_k$ is the location of the $k$th largest peak in the observed effectsize image i.e. $\hat{\mu}, \hat{d}, R^2$ respectively. |
| $\hat{v}_{k,b}$ | $\hat{v}_{k,b}$ is the location of the $k$th largest peak in the $b$th bootstrap image i.e. $\hat{\mu}_b, \hat{d}_b, R_b^2$ respectively. |

Table 5.1: Variable Key: Note that for clarity we have dropped the index $v$ above, all operations are done pointwise on the images at each of their voxels.

## 2.3 Simulations

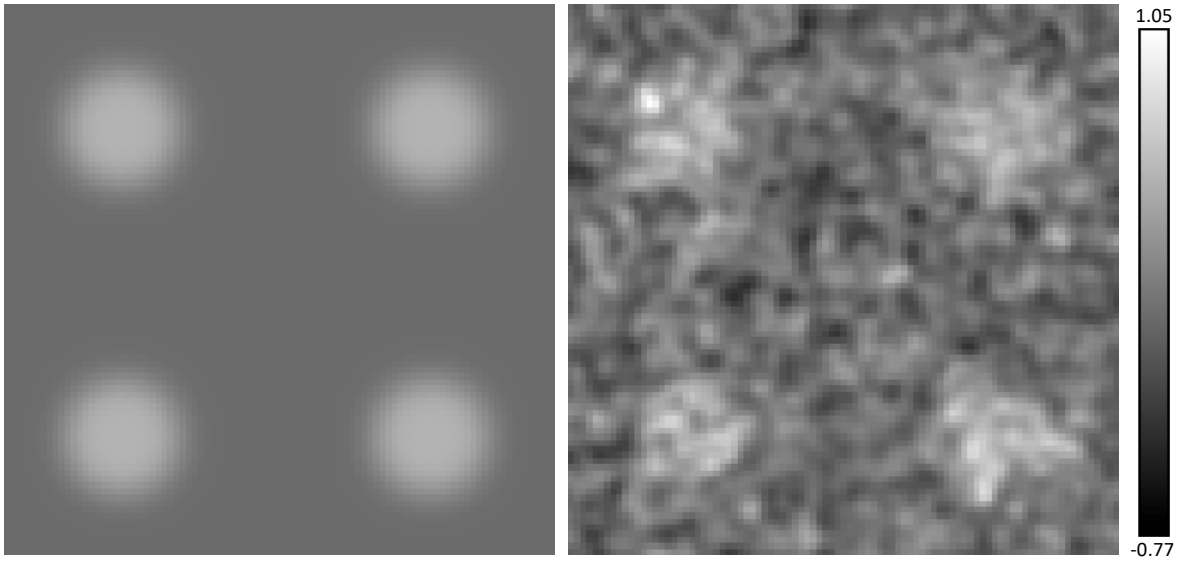
### 2.3.1 One Sample Mean

In order to test Algorithm 3 we generate 3D simulations on a $91 \times 109 \times 91$ 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. We generate data according to model (1) with underlying mean consisting of 3 different peaks with magnitudes of 2, 4, 4 placed at different points of the image. For the $\epsilon_n$ we use mean zero Gaussian noise smoothed with an FWHM of 3 voxels, scaled to have variance 1. In order to evaluate the methods we consider sample sizes of $N = 20, 30, \ldots, 100$ and for each sample size generate 1000 realizations. We use a threshold $u = 2$, this value has been chosen arbitrarily, in practice it could be chosen based on domain knowledge about the underlying signal.

### 2.3.2 One Sample

In order to test the performance of Algorithm 4 for estimation of Cohen's $d$ and the mean we generate 3D simulations (on the same set of voxels $\mathcal{V}$ as above) according to model (5.1) 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 5.2 for a slice through this signal and one realization. For the $\epsilon_n$ we use mean zero Gaussian noise smoothed with an FWHM of 6mm, scaled to have variance 1.

In order to evaluate how the methods compare as the sample size increases we consider $N$ random images, $N = \{20, 30, \ldots, 100\}$, generating 1,000 realizations (of the simulations described above) for each $N$. We use an additional simulation to find the voxelwise threshold that controls the familywise error rate at 5%; for each $N$ we generate 5,000 null $t_{N-1}$ random fields (computed by taking the one-sample $t$-statistic




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

Figure 5.2: Simulation true signal and one realization. Panel (a) illustrates a slice through the true signal corresponding to the plane $y = 20$, with maximum intensity $d = 1/2$. Panel (b) illustrates the same slice through the one sample Cohen's $d$ for 50 subjects. Data for each subject is computed by adding Gaussian noise (with 3 voxel FWHM) to the signal.

of $N$ zero mean Gaussian random fields with 3 voxel FWHM) and take the 95% quantile of the distribution of the maximum.

In order to evaluate how the methods compare as the variance changes we generate 1000 realizations (for each realization we generate 50 subjects) and adjust the variance (which is constant over the image) such that the peak Cohen's $d$ takes the values: $\{0.1, 0.2, \ldots, 0.7\}$ rather than just 0.5. We controlled the voxelwise familywise error rate as described above.

### 2.3.3 General Linear Model

In order to test the performance of Algorithm 5 for estimation of partial $R^2$ we generate 3D simulations (on the same set of voxels $\mathcal{V}$ as above) according to the model

$$Y_n(v) = 1 + \mu(v)x_n + \epsilon_n(v) \tag{5.4}$$

where $x_n \overset{iid}{\sim} N(0,1)$, $n = 1, \ldots, N$ and the $\epsilon_n$ are i.i.d random images which are mean zero Gaussian, with 3 voxel FWHM and scaled to have variance 1. $\mu$ consists of 9 different peaks each with magnitude 0.5822, with one located near corner and one at the centre of the image. The value 0.5822 has been chosen so that the power matches that of the one-sample simulations, see Supplementary Material Section 8.2 for details. As for the one-sample simulations, for $N \in \{20, \ldots, 100\}$ we generate 1,000 realizations of the above model and calculate a voxelwise threshold using additional simulations.

In order to evaluate how the methods compare as the variance changes we generate 1000 realizations and change the variance (which is constant over the image) such that the peak $R^2$ takes the values: $\{0.1, 0.2, \ldots, 0.6\}$. We consider $N = 50$ and $N = 100$ in order to illustrate what happens when you have a sufficiently large number of subjects. We controlled the voxelwise familywise error rate as described above.

## 2.4 Big Data Validation

In order to test our methods we take advantage of the large sample sizes in the UK Biobank. This enables us to set aside 4,000 (randomly selected) subjects in order to compute a very accurate estimate of the mean, Cohen's $d$ or partial $R^2$ value. We will refer to this 4,000-subject estimate of the effect as the ground truth. See Appendix 7.2 for details on how the ground truth is computed in the different settings. Implementing large linear models with missing data can be computationally burdensome so we outline efficient methods for dealing with this in Appendices 7.2.3 and 7.2.4.

We divide the remaining 4,940 subjects into groups similar in size to those used in typical fMRI/VBM studies. For each such group we apply all three methods and compare the values obtained to the ground truth calculated using the 4000 subjects,

allowing the performance of the methods across groups to be evaluated. In each small sample, we consider only complete-data voxels, as is typical in neuroimaging analyses.

### 2.4.1 Image Acquisition

The UK Biobank is 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 use 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. All data were anonymized, and collected with the approval of the respective ethics boards.

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 $88 \times 88 \times 64$ matrix, $\times 8$ multislice acceleration]. Standard preprocessing and task fMRI modeling 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$^3$

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

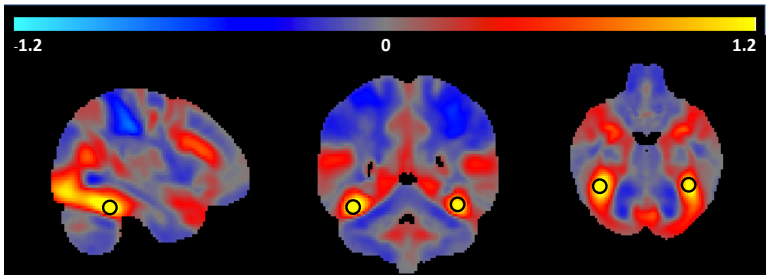
isotopic voxels in $208 \times 256 \times 256$ matrix]. Images were defaced and nonlinearly warped to MNI152 space using FNIRT (fMRIB's Nonlinear Image Registration Tool). For VBM, tissue segmentation was performed with FSL's FAST (fMRIB's Automated Segmentation Tool), producing images of gray matter that were subsequently warped to MNI152 space, and modulated by the Jacobian of the warp field. Warped modulated 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 additionally apply a smoothing of 3mm FWHM to the modulated gray matter images.

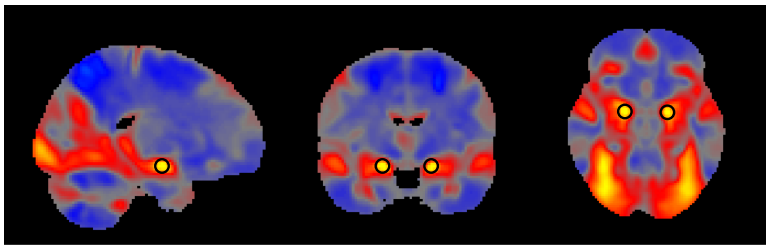
### 2.4.2 Task fMRI analysis

We have faces-shapes contrast images from 8,940 subjects and consider the mean and one sample Cohen's $d$. We compute the 4,000-subject Cohen's $d$ ground truth image for voxels with data for at least 100 subjects (Figure 5.3). For a given sample size $N$, let $G_N = \lfloor 4940/N \rfloor$ be the number of groups of size $N$ into which we can divide the 4,940 remaining subjects[3]. This division enables a comparison of the performance of the three available methods, circular inference, data-splitting and the bootstrap. As in the simulations, we measure the performance in terms of bias, standard deviation and root mean squared error (RMSE) as defined in Section 2.5. We use sample sizes of 20, 50 and 100 to illustrate the performance of the methods, these sizes have been chosen since they are representative of those typically used in fMRI. Sections 3.2.1 and 3.2.2 present the results of applying the methods and Figure 5.4 illustrates these methods applied to an exemplar sample consisting of 50 subjects.

[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 5.3: Slices through the top four maxima of the one-sample Cohen's $d$ ground truth. The top two local maxima are located at voxels (42, -46, -22) and (-38, -48, -20) (mm MNI space) which correspond to the left and right temporal occipital fusiform cortices and have Cohen's $d$ values of 1.5756 and 1.4326 respectively. The 3rd and 4th largest local maxima are located at voxels (20,-6,-14) and (-18, -6, -14) 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 black circles.

### 2.4.3 VBM data

We have structural gray matter (VBM) data from the 8,940 subjects to illustrate how our bootstrap method performs in the context of the general linear model. We regress these gray matter images against age, sex and an intercept. Using the 4,000 subjects, we calculate a ground truth estimate of the partial $R^2$ for age at every voxel for which the mean of the VBM values at that voxel over all 4,000 subjects are greater than 0.1. The maximum of this ground truth is located at voxel (45, 62, 34) and has a partial $R^2$ of 0.2466. As above we divide the remaining subjects into $G_N = \lfloor 4940/N \rfloor$ groups, for $N = 50, 100$ and 150 to compare the methods by calculating the partial $R^2$ for age on each subgroup. Here we use larger sample sizes as this setting is more challenging for inference as the effect size is considerably smaller than in the task fMRI data. See

Section 3.2.3 for the results of this validation.

### 2.4.4 Threshold Computation

For real data analyses, researchers typically either use random field theory (RFT) (Worsley et al., 1996) or permutation testing (Nichols and Holmes, 2002a) to compute screening thresholds. Voxelwise RFT controls the false positive rates but is slightly conservative, (Eklund et al., 2016), primarily because the lattice assumption is not valid for low smoothness levels. (NOTE: this Chapter, despite being last sequentially, was written before the other chapters and is the text of our paper that was accepted to NeuroImage. As such we do not make use of the improved voxelwise RFT framework that we developed in Chapter 2. All references to voxelwise RFT in this chapter refer to the traditional Worsley et al. (1992) RFT that we defined in Chapter 2 and, as discussed there, is conservative when applied to fMRI data.) On the other hand one-sample permutation can have slightly inflated false positive rates, (Eklund et al., 2019), and has a high computational cost. In our case this cost is very large as we need to perform a big data validation which requires many analyses as discussed in Section 2.4. We have thus elected to use voxelwise random field theory for our big data analyses. In practice when running a typical fMRI/VBM analysis, our methods work independently of the method used to choose the threshold.

## 2.5 Method Comparison

In order to compare the bootstrap, data-splitting and circular inference (in simulations and in the big data validation), we consider their bias, standard deviation and root mean squared error (RMSE) calculated over the 1,000 realizations for each $N$. Here we are computing the bias, standard deviation and RMSE in a non-standard context, in

that the true parameter values vary in each instance. Traditionally in inference, one estimates a common $\theta$ with estimators $\hat{\theta}_1, ..., \hat{\theta}_{K^*}$, giving us the usual MSE decomposition for a sample of size $K^*$,

$$\begin{aligned}\text{MSE} &= \frac{1}{K}\sum_{k=1}^{K^*}(\hat{\theta}_k - \theta)^2 \\ &= \frac{1}{K}\sum_{k=1}^{K^*}\left(\hat{\theta}_k - \frac{1}{K^*}\sum_{k=1}^{K^*}\hat{\theta}_k\right)^2 + \left(\frac{1}{K^*}\sum_{k=1}^{K^*}\hat{\theta}_k - \theta\right)^2\end{aligned}$$

into variance and squared bias. However in our context we have estimators $\hat{\theta}_1, \ldots, \hat{\theta}_{K^*}$ of parameters $\theta_1, \ldots, \theta_{K^*}$. In our setting, $K^*$ is the number of significant peaks that are found over all realizations. For $k = 1, \ldots, K^*$, $\hat{\theta}_k$ is the value of one of these significant peaks, and $\theta_k$ is the true value at the voxel corresponding to the peak. The $\theta_k$ are different because the locations of the peaks are random. As such we instead define

$$\tilde{\theta}_k = \hat{\theta}_k - \theta_k$$

and use the fact that the noise-free value of $\tilde{\theta}_k$ is 0 for each $k$. This allows us to define the MSE as

$$\begin{aligned}\text{MSE} &= \frac{1}{K^*}\sum_{k=1}^{K^*}(\tilde{\theta}_k - 0)^2 \\ &= \frac{1}{K^*}\sum_{k=1}^{K^*}(\tilde{\theta}_k - \frac{1}{K^*}\sum_{k=1}^{K^*}\tilde{\theta}_k)^2 + \left(\frac{1}{K^*}\sum_{k=1}^{K^*}\tilde{\theta}_k\right)^2\end{aligned}$$

where the second equality follows by bias-variance decomposition. This leads us to define the variance

$$\frac{1}{K^*}\sum_{k=1}^{K^*}(\tilde{\theta}_k - \frac{1}{K^*}\sum_{k=1}^{K^*}\tilde{\theta}_k)^2 = \frac{1}{K^*}\sum_{k=1}^{K^*}\left(\hat{\theta}_k - \theta_k - \frac{1}{K^*}\sum_{k=1}^{K^*}(\hat{\theta}_k - \theta_k)\right)^2,$$

and the bias

$$\frac{1}{K^*}\sum_{k=1}^{K^*}\tilde{\theta}_k = \frac{1}{K^*}\sum_{k=1}^{K^*}(\hat{\theta}_k - \theta_k)$$

in this context. These expressions represent the bias/variance/MSE at a randomly selected peak in the sense that if a peak is selected at random we expect it to be biased above the true effect size by this amount on average. The root mean squared error or RMSE is defined to be the square root of the MSE and the standard deviation to be the square root of the variance.

In the context of the big data analysis described above, for a given sample size $N$, $K^*$ is the number of significant peaks over all the $G_N$ subsets. In Figures 5.10, 5.12, 5.14 for each set of estimates we have made boxplots for the bias $\hat{\theta}_k - \theta_k$ over all $K^*$ significant peaks. For the bar plots in these figures we have plotted the RMSE and standard deviation as defined above.

# 3 Results

## 3.1 Results - Simulations

We found that Algorithm 3, bias correction for peak height in a sample mean image alone, had similar performance as for bias correction of statistic peaks (Algorithm 4), detailed below. However, as direct assessment of the mean image provides no way to control false positives, we regard it as a less useful method and have relegated its evaluation (on simulations and real data) to the Supplementary Material (Section 8.1). In this section we illustrate the performance of Algorithm 4. The performance of the methods in the GLM setting (Algorithm 5) is very similar and so this has also been left to the Supplementary Material (Section 8.2).

### 3.1.1 One Sample Cohen's $d$

Figure 5.5 left column plots the estimates of the bias, standard deviation and root mean squared error from each of the three methods as the sample size increases where the peak effect size is 0.5 as discussed in Section 2.3.2. As expected the circular method has the worst bias, but has low standard deviation, while data splitting is unbiased, but has the highest standard deviation. Bootstrapping has a bias which decreases to 0 as the sample size increases. Summarising mean and standard deviation with RMSE, we find that the bootstrap method has the lowest RMSE for all sample sizes except $N = 20$. The $N = 20$ exception likely occurs because resampling methods perform better for larger sample sizes.

Figure 5.5 right column plots the estimates of the bias, standard deviation and root mean squared error from each of the three methods for $N = 50$ for a range of peak effect sizes. At all except the smallest effect size the bootstrap outperforms the others in terms of RMSE. The small effect size deviation occurs because the bootstrap correction is based on the rank order of the peaks, and when SNR is low the sample rank orders is a poor approximation of the noise-free rank order. As such for lower effect sizes a larger number of subjects is required for the bootstrap to outperform the other methods. For all methods the bias decreases as the peak effect size increases, this occurs because the peaks in the signal are more prominent and therefore are less subject to the winner's curse.

In our simulations the circular and bootstrap methods find 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 many of our simulations for small sample sizes, data-splitting often found no peaks to be significant

at all. In order to compare the power of the methods we computed the average number of significant voxels found across all realizations for each sample size (Figure 5.6). Circular inference and the bootstrap both use the one-sample $t$-statistic to determine significance so they both find the same number of voxels above the threshold, which is substantially more than the number found by data-splitting.

### 3.1.2 Estimating the Mean

As discussed in Section 2.1.3, a bias correction for the mean at locations of peaks in the statistic image can be obtained from a variant of Algorithm 4. Estimates of the bias, standard deviation and RMSE of each of the three methods (see Figure 5.7), show a very similar performance as for the previous setting, with the bootstrap method having the lowest RMSE across all sample sizes. When the SNR is low a larger number of subjects is required before the bootstrap outperforms data-splitting in terms of RMSE (see Figure 5.7, bottom right plot). To illustrate that this occurs we have included plots that illustrate the relative performance of the algorithms (for a larger number of subjects) in Figure 5.21 of the Supplementary Material.

## 3.2 Results - Real Data

In this section we apply the methods to task fMRI and VBM data as described in Section 2.4. However, before discussing these these results, we illustrate the magnitude of the circularity problem we compare maximum peak heights as a function of sample size. To do so we compute the maximum 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 5.8. 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.

### 3.2.1 Evaluation: Task fMRI Cohen's $d$ peak height estimation

Figure 5.10 presents the results of applying Algorithm 4 to the one-sample task fMRI data, and is analogous to the simulated data results in Figure 5.5. (Note that as bias can be measured at each peak, it can be presented via boxplots; whereas only a single standard deviation and RMSE can be computed per setting, see Section 2.5 for details) As in the simulations we find that the circular estimates are highly biased whereas the bootstrap estimates have low bias with the bias decreasing as the sample size increases. The bootstrap has the lowest RMSE for each sample size.

In order to compare the power of the methods, as with the simulations, we have computed the average number of voxels above the threshold over all $G_N$ groups (for $N \in \{10, 20, \ldots, 100\}$) see Figure 5.9. From this we see that, for a given sample size, circular inference and the bootstrap find many more peaks than data-splitting which illustrates the considerable difference in power. We note that for data splitting with $N = 20$ we observe only a total of 7 peaks that were above the threshold over all the $G_{20} = 247$ groups, so the results in Figure 5.10 could be unstable in this case.

To further understand how the estimates compare we plot their values against the ground truth in Figure 5.11. For each $N$, each data point in the corresponding graph shows the estimated peak intensity of a significant peak from one of the $G_N$ groups (ordinate), and the ground truth intensity at the location of the peak (abscissa). The $N = 20$ case is the most challenging for estimation. Here the circular estimates are very biased while bootstrap estimates give reasonable estimates and the data-splitting estimates are particularly variable and are fewer in number. As $N$ increases, all of the methods perform better: the circular estimates are biased, and the data-splitting

estimates are variable whereas the bootstrap estimates have low bias and variance. The effect of the threshold is particularly evident in the plots for the circular method and, to a lesser extent, for the bootstrap method.

Note that in Figure 5.11, for large $N$, the circular method suffers a bias that is relatively constant with respect to the true Cohen's $d$. The bootstrap method corrects this bias, with the point cloud having roughly the same shape as the circular method's, only shifted downward. In contrast, for the lowest $N$, there is a greater mismatch in the circular and bootstrap plots. This is because the bootstrap correction is based on the rank order of the peaks, and when SNR is low the sample rank orders is a poor approximation of the noise-free rank order.

### 3.2.2 Evaluation: Task fMRI mean estimation at Cohen's $d$ peak location

Figure 5.12 illustrates the results of applying Algorithm 4 to the one-sample task fMRI data to estimate the mean at Cohen's $d$ (or $t$-statistic) peak locations. Here, all the methods perform well and have relatively low little bias, though circular inference is the most biased. Data-splitting has the worst standard deviation and RMSE for $N = 50$ and 100; it has best RMSE for $N = 20$, but we again note that this is based on only 7 peaks.

The selection bias is much less severe in this scenario. This is due to the selection being based on Cohen's $d$, which is correlated with but not the same as the mean. Notably for $N = 50$ and 100 the circular estimates have a lower RMSE than the data-splitting estimates, but the bootstrap estimates have the lowest RMSE.

Scatter plots comparing the estimates and the ground truth (Figure 5.13) reflect the observation of a much reduced problem of circularity bias; however, the bias in the circular estimates is still evident.

### 3.2.3 Evaluation: Gray matter VBM $R^2$ peak height estimation

Figure 5.14 illustrates the results for estimating the partial $R^2$ of age with the VBM data. The performance here resembles that of Cohen's $d$ peak estimation: there is little bias for data-splitting and the bootstrap, the circular and bootstrap methods have lower standard deviation, and the bootstrap consistently has the lowest RMSE.

While the boxplot and bar plot summaries (Figure 5.14) are consistent, the analogous Cohen's $d$ scatter plots (Figure 5.15) have a very different character. As the circular results (left column) make clear, most of the $R^2$ estimates are close to the threshold, indicating a severe selection effect. As discussed above, when SNR is low the observed rank order can differ substantially from the noise-free rank order, reducing the accuracy of the bootstrap method. However, as the sample size increases the bootstrap estimates fall closer to the identity line more closely and in terms of RMSE, the bootstrap still outperforms circular inference and data-splitting.

## 3.3 Demonstration on HCP Task fMRI dataset

In order to illustrate the bootstrap method in action we apply it to a sample of 80 unrelated subjects from the Human Connectome project and look at one of the working memory contrasts. Subjects performed an $N$-back task using alternating blocks of 0-back and 2-back conditions with faces, non-living man-made objects, animals, body parts, house and words. We examine the average (2-back − 0-back) contrast, identifying brain regions supporting working memory in general.

We use a group level model and compute a one-sample $t$-statistic at each voxel in order to test for activation. Voxelwise permutation testing is used to control the familywise error rate to 5% resulting in a threshold of 5.10 for the $t$-statistic. The largest

peak above this threshold has a $t$-statistic value of 13.58 and lies within the Medial Frontal Gyrus an area commonly associated with working memory. At the largest peak the circular Cohen's $d$ is $13.58/\sqrt{80} = 1.52$; the bootstrap corrected Cohen's $d$ estimate is 1.161. In total 234 peaks lie above the threshold, with 25 peaks falling within the Medial Frontal Gyrus region (Harvard-Oxford Atlas). Table 5.2 reports the circular and bootstrapped Cohen's $d$ as well as the bootstrap estimate of the mean for the top 10 of these 25 (Cohen's $d$/$t$-statistic) peaks. Slices through the one-sample $t$-statistic at the voxel corresponding to the largest peak are shown in Figure 5.16.

Figure 5.17 shows the effect that these corrections have on power, where we have plotted a graph of sample size against power for a whole brain analysis using a $p$-value threshold, corresponding to taking $T = 5.10$, of $1.39 \times 10^{-6}$. Using the raw value would suggest that only 24 subjects are need to attain 80% power, when in fact the corrected estimate shows that 34 subjects are needed to provide this level of power. See Appendix 7.3.2 for details on how the power is calculated.

| Circular Cohen's $d$ | Corrected Cohen's $d$ | Circular Mean (%BOLD) | Corrected Mean (%BOLD) | Peak Location |
|---|---|---|---|---|
| 1.519 | 1.161 | 0.450 | 0.433 | (28, 8, 56) |
| 1.137 | 0.922 | 0.347 | 0.321 | (-48, 6, 42) |
| 1.096 | 0.889 | 0.561 | 0.533 | (-34, 0, 64) |
| 1.091 | 0.888 | 0.279 | 0.257 | (28, 14, 48) |
| 1.079 | 0.883 | 0.461 | 0.434 | (44, 34, 32) |
| 1.078 | 0.883 | 0.351 | 0.328 | (-32, 2, 62) |
| 1.078 | 0.882 | 0.378 | 0.356 | (40, 34, 36) |
| 1.067 | 0.876 | 0.378 | 0.356 | (-48, 8, 38) |
| 0.994 | 0.817 | 0.339 | 0.318 | (-44, 26, 36) |
| 0.979 | 0.807 | 0.280 | 0.260 | (-40, 6, 56) |

Table 5.2: The circular and corrected estimates of Cohen's $d$ and the mean at the top ten significant peaks of Cohen's $d$ in the Medial Frontal Gyrus. There is appreciable bias for the largest Cohen's $d$ peaks, while the %BOLD values at the Cohen's $d$ peaks have relatively little bias.

# 4 Discussion

Unbiased estimation of effect size is essential yet absent from most neuroimaging studies. We have evaluated three methods for assessing the signal magnitude at peaks in neuroimaging analyses. The bootstrap method that we have introduced provides circularity-corrected estimates from an analysis using all of the data. Compared to uncorrected, circular inference our method has dramatically less bias and lower RMSE. While data-splitting is unbiased by construction, our method has lower standard deviation and RMSE in most settings. Given the small size of many studies, using data-splitting, and thereby having to divide the data in half, may produce unacceptable reductions in power.

Even for small sample sizes the bootstrap has similar or better RMSE relative to data-splitting. However, we note that in neuroimaging it is very important to have an accurate estimate of the location of the effect. For this reason we assert, that even in this scenario, the bootstrap is to be preferred over data-splitting since it uses all data to compute the peak locations. It thus identifies a greater number of significant peaks and its estimates of their locations are more accurate.

The dramatic difference in the plots comparing the estimates with the ground truth for Cohen's $d$ and $R^2$ (Figures 5.11 & 5.15, respectively) should be viewed in terms of the dramatic difference in power between these two settings. Consider that, in these evaluations, we have a typical Cohen's $d$ of 1.0 and $R^2$ of 0.1. If we consider a power analysis with a whole brain $\alpha = 1.39 \times 10^{-6}$ (here we use the $\alpha$ level of Section 3.3 as representative of typical threshold) and a target power of $1 - \beta = 80\%$, a one-sample $t$-test with this Cohen's $d$ would require 42 subjects; in contrast, a simple linear regression with this $R^2$ would require 306 subjects. As shown in our simulations (cf.

Figure 5.5 (left column) and Figure 5.19) when the effect sizes are comparable the bootstrap requires a larger number of subjects before it outperforms data-splitting in terms of RMSE. In this light, we find the $R^2$ results even more impressive, providing adequate performance even with negligible power.

Large scale repositories of neuroimaging data have enabled us to validate our methods in a way that has (to our knowledge) not been done before in the neuroimaging setting. This involves setting aside 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 performance of methods relative to the ground truth. This approach enables methods to be rigorously tested and ensures that they work on real data rather than just on simplified simulations. We recommend this sort of evaluation for all new statistical imaging methods, as well as for existing methods that have not been rigorously tested on real data.

At present our method provides a bias correction for the intensity at the location of the observed peaks $\hat{v}_k$. This is appropriate because when a researcher comes to replicate the results they should be able to test the effect at a given location. However another direction would be to obtain estimates for the signal intensity at true peak locations. Let $v_k$ be the location of the $k$th largest peak in the noise-free image. In the setting of Algorithm 3, at present we infer on $\mu(\hat{v}_k)$ by estimating the bias $\hat{\mu}(\hat{v}_k) - \mu(\hat{v}_k)$, but we could instead infer on $\mu(v_k)$ by estimating the bias $\hat{\mu}(\hat{v}_k) - \mu(v_k)$. It may be possible to estimate this bias using a bootstrap approach: comparing bootstrapped peaks to peaks of the empirical mean. Peaks could be matched according to their order statistics or to the nearest large empirical peak within a certain radius in order to obtain an estimate of the bias. The challenge of this approach would be to obtain a good criteria for peak matching. Algorithms 4 and 5 could be extended similarly.

There is much ongoing research in the field of selective inference and there is much potential for other methods to be modified for use in the fMRI setting. It would particularly desirable to derive theoretical corrections using random field theory, however it is at the moment difficult to estimate the peak height distribution of a non mean zero random process (Cheng and Schwartzman, 2015a), so this is an important area for future research. The bootstrap approach provides approximately unbiased estimates of the effect sizes (note that these estimates are not completely unbiased as the bootstrap is centred at the empirical mean rather than the true mean). In practice it greatly reduces bias in the estimates (as shown in the simulations and our validations). However, it would be of interest to prove results that determined which settings the MSE of the bootstrap approach is lower than that of data-splitting. We have shown that in practice this holds so long as you have sufficiently many subjects (in these settings 20-50 subjects is typically sufficient).

Clusterwise inference is commonly used in fMRI and it is also of interest to develop selective inference approaches that allow for power analyses in this context. One approach would be to obtain an unbiased estimate of clusterwise mean (which typically suffers from selection bias) and to report the mask of where the activity lies. Our method cannot be directly used to estimate of the cluster mean because of a lack of pivotality however it may be possible to modify it in such a way that this it not a problem. Methods such as that developed in Hayasaka et al. (2007) could then be used to perform power analyses. This would provide an approximate estimate of the power though two potential issues with this approach are that in reality the mean is not constant over each cluster and that not every voxel within a significantly cluster is active.

We have motivated peak-level inference for its use in power analysis, however it

also forms an essential part of how results are presented in SPM. Ever since a revision of SPM5 that introduced FDR inference for peaks, the "voxel-level" column label has been replaced with with "peak-level" in the inference table. (Confusingly, FWE $p$-values are identical for voxels and peaks, while FDR $p$-values differ substantially, with peak $p$-values notably depending on a screening threshold). Chumbley et al. (2010) have stridently argued against voxel-level inference, asserting that only peaks (and clusters) should be objects of inference in neuroimaging, as these are topological characteristics that can be unambiguously identified in a continuous process analogue of the statistic image. In general we see the value of voxel, peak and clusterwise inference, however the ubiquity of reporting peaks in statistic images in SPM and other packages is an important motivation for this work.

One important finding of our work was that the circular estimates of the mean at peaks of the test-statistic were relatively unbiased and already had lower MSE than the data-splitting estimates. This is quite an exciting finding as it shows that (in neuroimaging) practitioners do not have to worry so much about the bias in the mean when reporting the raw effect size. Our bootstrap estimates still provide a better estimate - with lower bias and MSE - and so we still recommend using the bootstrap estimates in this setting.

In sum the bootstrap approach provides a method to remove the bias while using all of the data to obtain accurate estimates of the locations. Relative to data-splitting and circular inference this results in estimates which have similar or generally better RMSE.

# 5 Software Availability and Reproducibility

The analysis in this paper was performed using MATLAB 2015a. Scripts to implement the bootstrap, circular inference and data-splitting methods are available at `https://github.com/sjdavenport/SIbootstrap`. Code to perform power analyses and large-scale linear modeling has also been included. For reproducibility scripts to reproduce the figures in the results section are also available in the Results_Figures folder.

Simulations and thresholding were performed using code from the RFTtoolbox available at `https://github.com/sjdavenport/RFTtoolbox`. Brain imaging figures where created using FSLeyes (McCarthy, 2019).

# 6 Acknowledgments

We would like to thank the 3 anonymous reviewers for their comments which have helped to improve the quality of this manuscript.

TEN is supported by the Wellcome Trust, 100309/Z/12/Z and SJD is funded by the EPSRC. 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.

# 7 Appendix

## 7.1 Computing partial $R^2$ from an $F$-statistic

For a general linear model, let $\Omega$ denote the overall model and let $\omega \subset \Omega$ denote some sub-model with $p_0$ degrees of freedom. Define $\text{RSS}_\Omega$ and $\text{RSS}_\omega$ to be the residual sum of squares for each of the models. Then we can write the $F$ statistic for comparing $\omega$ and $\Omega$ 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:

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

Thus, with some algebra $F$ can be expressed in terms of $R^2$ as

$$F = \frac{N-p}{m}\left(\frac{R^2}{1-R^2}\right)$$

and conversely, $R^2$ can be expressed in terms of $F$ as,

$$R^2 = \frac{mF}{mF + N - p}.$$

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

## 7.2 Masking and Calculating the Ground Truth

The UK Biobank 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 regression coefficient or partial $R^2$ in a linear model. This appendix describes the details of how

we use such a large sample to create ground truth and how we deal with practical challenges, including masking and the inability to load all images into memory at once.

### 7.2.1 Masking

Any neuroimaging analysis requires a mask to define voxels that are to be included in the modeling process. In practice, the mask for each subject is unique. Let $\mathcal{D}$ be the set of all possible voxels in the image, then given a subject: $n$, define its **mask** to be the image: $M_n : \mathcal{D} \longrightarrow \mathbb{R}$ such that $M_n(v)$ is 1 if subject $n$ has data at voxel $v$, 0 otherwise. Given this definition, define the **intersection mask** of a subset $\mathcal{S}$ of subjects to be the image $M_{\mathcal{S}}$ such that

$$M_{\mathcal{S}}(v) = \begin{cases} 1 & M_n(v) = 1 \text{ for all } n \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

The mask used for a small sample analysis on the subset $\mathcal{S}$ is the product of the intersection mask $M_{\mathcal{S}}$ with the 2mm MNI brain mask, the image MNI152_T1_2mm_brain_mask in FSL. We refer to this as the **analysis mask**.

### 7.2.2 One-Sample Ground Truth Mean and Cohen's $d$

We choose a random subset $\mathcal{S}$ of $\{1, \ldots, 8940\}$ of size 4,000 to estimate a ground truth mean and Cohen's $d$ using the available data at each voxel. Note that 99.99% of voxels had data from from at least 100 subjects, while 98% had data from at least 3,000 subjects. In order that each voxel be a reliable estimate we require that at least 100 subjects have data at that voxel in order that it be included. Given subject images $Y_n, n \in \mathcal{S}$, define the **ground truth mean** to be

$$\mu(v) = \frac{\sum_{n \in \mathcal{S}} Y_n(v) M_n(v)}{\sum_{n \in \mathcal{S}} M_n(v)} \times \mathbb{1}(M_n(v) = 1 \text{ for at least 100 } n \in \mathcal{S}),$$

where $\mathbb{1}(\cdot)$ is the indicator function. Define the **ground truth variance** to be:

$$\sigma^2(v) = \frac{\sum_{n\in\mathcal{S}}(Y_n - \mu(v))^2 M_n(v)}{\sum_{n\in\mathcal{S}} M_n(v) - 1} \times \mathbb{1}(M_n(v) = 1 \text{ for at least 100 } n \in \mathcal{S}),$$

and the **ground truth Cohen's** $d$ estimate as

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

Finally each of these are additionally masked with the 2mm MNI brain mask.

### 7.2.3 Linear Model for Big Data - No Missingness

The full unmasked images comprise $902,629 = 91 \times 109 \times 91$ voxels and for 4,000 subjects this data would occupy 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 separate linear models at each voxel sequentially is slow as it requires access to all of the images for each of the voxels (in the subset of the brain that is of interest) in turn. Loading all of the images at once is generally not feasible due to memory constraints. An improvement would be to divide the brain image into blocks that can fit in memory, however this still requires each image be accessed multiple times. Instead it is possible to write the estimates of the linear model in terms of individual contributions of each subject, allowing arbitrarily large datasets by only reading one subject's data at a time. Suppose that we have $N_{\text{all}}$ subjects (when computing the ground truth $N_{\text{all}} = 4{,}000$), that we have an $N_{\text{all}} \times p$ design matrix $X$ and that there is no missing data. Let $Y$ be the $N_{\text{all}} \times V$ matrix of all the subject images where $V$ is the number of voxels in each subject image $Y_n$. For the mass univariate linear model $Y = X\beta + \epsilon$, we want to compute

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

at each voxel. Instead of computing this directly we observe that for each $v \in \mathcal{V}$,

$$X^T Y(v) = \begin{pmatrix} x_1, \ldots, x_{N_{\text{all}}} \end{pmatrix} \begin{pmatrix} Y_1(v) \\ \vdots \\ Y_{N_{\text{all}}}(v) \end{pmatrix} = \sum_{n=1}^{N_{all}} Y_n(v) x_n,$$

where $x_n^T$ is the $n$th row of $X$, and so $X^T Y$ can computed by loading one image at a time. This $p \times V$ matrix can then be pre-multiplied by $(X^T X)^{-1}$, which only has to be calculated once, in order to calculate $\hat{\beta}$. The sample variance image can then be computed by a second pass through the data as

$$\hat{\sigma}^2 = (N_{\text{all}} - p)^{-1} \sum_{n=1}^{N_{all}} (Y_n - x_n^T \hat{\beta})^2.$$

The $F$-statistic can then be computed as usual and this allows calculation of the ground truth partial $R^2$ using the transformation from Appendix 7.1.

### 7.2.4 Linear Model for Big Data - Accounting for Missingness

The previous section assumed identical masks for all subjects, which is not realistic due to susceptibility drop-out in fMRI, variation in field of view in structural MRI, and simply random variation in the exact brain boundary in each subject. For each subject $n = 1, \ldots, N_{\text{all}}$, suppose that we have a binary mask image $M_n$ which denotes the missingness in the response (we will assume that there is no missingness in the predictors). Since there is no missingness in the covariates and they are fixed, we can obtain an unbiased estimator of the regression coefficient at each voxel using the complete data at that voxel, so long as we assume that the missingness mechanism is independent of the image data (White and Carlin, 2010). Assuming this is reasonable, given that the missingness is typically due to acquisition and technical artifacts.

For each voxel $v$, let $C(v) := \{n : M_n(v) = 1\}$, and let $C(v)$ as a subscript indicate subsetting the corresponding rows of a matrix. Then for each voxel $v$ we need to compute

$$\hat{\beta}(v) = (X_{C(v)}^T X_{C(v)})^{-1} X_{C(v)}^T Y_{C(v)}.$$

The first and second parts of this expression can be computed as

$$(X_{C(v)}^T X_{C(v)})^{-1} = \left( \sum_{n=1}^{N_{\text{all}}} M_n(v) x_n x_n^T \right)^{-1}$$

and

$$X_{C(v)}^T Y_{C(v)} = \sum_{n=1}^{N_{\text{all}}} M_n(v) Y_n(v) x_n$$

so this can also be computed by loading one image at a time. This requires storage of a $p \times p$ matrix at each voxel which, for moderate $p$, is not problematic. The same masking computation can be used when computing residual variance, which then allows computation of the $F$-statistic and this allows calculation of the ground truth partial $R^2$ using the transformation from Appendix 7.1.

## 7.3 Non-Central Distributions and Power Analyses

### 7.3.1 Non-Central Distributions

**One-Sample $t$-statistic** Following the model from Section 2.1, under the assumption of Gaussian noise,

$$\hat{\mu}\sqrt{N} \sim N(\mu\sqrt{N}, \sigma^2)$$

is independent of $\hat{\sigma}$ and so the $t$-statistic $\hat{\mu}\sqrt{N}/\hat{\sigma}$ has a non-central $t$-distribution with non-centrality parameter $\mu\sqrt{N}/\sigma$ and $N - 1$ degrees of freedom. The mean of the

non-central $t$ is not the non-centrality parameter, instead

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

for $N > 2$, where $\Gamma$ is the gamma function and $C_N$ is a bias correction factor (Hogben et al., 1961). Thus we use

$$\frac{\hat{\mu}}{\hat{\sigma}C_N}$$

as an unbiased of the population Cohen's $d$. Note that here and henceforth whenever we have two images $A$ and $B$ we write $\frac{A}{B}$ to be the image which takes the values $\frac{A(v)}{B(v)}$ at each voxel $v$.

**Non-Central $F$ and $t$ distributions in the General Linear Model** For the general linear model at a given voxel (note we suppress the voxel $v$ index here), we have $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$ independently of $\hat{\sigma}^2 \sim \frac{\sigma^2}{N-p}\chi^2_{N-p}$ which implies that

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

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

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

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 $N-p$ and so

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

as derived in Patnaik (1949). 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 $c^T\beta/\sqrt{\sigma^2 c^T(X^TX)^{-1}c}$.

### 7.3.2 Power Analyses

**One Sample** In the one sample scenario, for a potential future sample size $N'$ and an estimate of the non-centrality parameter: $\lambda$, 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 $\lambda$.

**Multiple Regression - Cohen's $f^2$** Calculation of power in the general linear model scenario is slightly more complicated as it requires distribution assumptions and approximations. To do so define Cohen's $f^2$ to be

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

where $R^2$ is the partial coefficient of determination and we have used the fact that $\frac{R^2}{1-R^2} = \frac{m}{N-p}F$ as derived in Appendix 7.1. In the framework of the general linear model (5.2), for $N \in \mathbb{N}$ suppose we observe an $N$-dimensional image $Y_N$ such that

$$Y_N = X_N\beta + \epsilon^N$$

for some $p$-dimensional parameter image $\beta$ and $N$-dimensional noise image $\epsilon^N = (\epsilon_1, \ldots, \epsilon_N)^T$ where $\{\epsilon_n\}_{n \in N}$ is an i.i.d sequence of noise images which have finite variance. Let $X_N = \left[ x_1, \ldots, x_N \right]^T$ be the design matrix, where $\{x_n\}_{n \in \mathbb{N}} \in \mathbb{R}^p$ is a sequence of finite variance, i.i.d random vectors (independent of the noise process) each with multivariate distribution $D$. For each $N$, let $\hat{\beta}_N$ be the $p$-dimensional image linear least squares estimator and let $\hat{\sigma}^2_N$ be the image estimate of variance.

Then $\frac{1}{N} X_N^T X_N \xrightarrow{a.s.} \mathbb{E}\big[x_1 x_1^T\big]$, $\hat{\beta}_N \xrightarrow{a.s.} \beta$ and $\hat{\sigma}^2_N \xrightarrow{a.s.} \sigma$ as $N \longrightarrow \infty$ (where $\xrightarrow{a.s.}$ denotes pointwise almost sure convergence) see the supplementary material Section 8.6 for proofs. Let $f^2_N$ be Cohen's $f^2$ for the $N$th model. Then combining the above results,

$$f^2_N \xrightarrow{a.s.} f^2_p := \frac{(C\beta)^T (C(\mathbb{E}\big[x_1 x_1^T\big])^{-1} C^T)^{-1} (C\beta)}{\sigma^2}$$

as $N \longrightarrow \infty$. This also implies almost sure convergence of $R^2$.

Given a new sample of $N'$ subjects from model 7.3.2 with corresponding design matrix $X'$ (an $N' \times p$ matrix whose rows are i.i.d with distribution $D$), then as 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:

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

where $f^2$ is the estimate of $f^2_p$. 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 $F$ distribution with $m$ and $N' - p$ degrees of freedom and non-centrality parameter $\lambda$.

**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

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

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

# 8 Supplementary Material

## 8.1 Application of Algorithm 3 to Simulated Data

The 3D simulations to test Algorithm 3 are described in Section 2.3.1 of the main text. The results (shown in Figure 5.18) are similar to those of the simulations in the main text.

In order to evaluate how the methods compare as the variance changes we generate 1000 realizations (for each realization we generate 50 subjects) and change the variance (which is constant over the image) such that $\frac{1}{\sigma}$ takes values in $\{0.2, 0.4, \ldots, 1.4\}$. The results are plotted in right column of Figure 5.18.

## 8.2 GLM simulations

The 3D simulations to test Algorithm 5 are described in Section 2.3.3 of the main

text. In order to approximately match the power of the one-sample simulations, in model (4) we take $\mu$ to have a peak value of 0.5822. The power is only ever approximately the same as it changes (for the one-sample $t$-statistic versus the $F$-statistic) over sample size and thresholding levels (thus it depends on the FWHM of the noise process). In order to derive the power for model (4) we need

$$f_p^2 = \left( \frac{c^T \beta}{c^T (\mathbb{E}[X^T X])^{-1} c} \right)^2 = \beta^2$$

where $X$ is the model design matrix. The second equality holds as $\mathbb{E}(X^T X)$ is the identity matrix here as the random variables $x$ are mean 0 variance 1 and are independent of the intercept term (as that's just a constant). This yields a corresponding population $R^2 = \frac{\beta^2}{1+\beta^2}$ and allows us to determine the $\beta$ value required to attain a certain level of power, see Section 7.3.2.

The results (shown in Figures 5.19 and 5.20) are similar to those of the simulations in the main text. In Figure 5.19 the bootstrap requires a slightly larger number of subjects (relative to the one sample simulations) before the RMSE drops below that of data-splitting. This is likely because the $R^2$ is a rather complicated function of the subject images and so the bootstrap needs a larger sample size in order to be as effective.

## 8.3 Additional Simulations for Estimating the Mean at Cohen's $d$ peaks

Figure 5.21 plots graphs in the same setting as the graphs in Figure 5.7 (right column) but take $N = 100$ instead of $N = 50$ in order to illustrate that the bootstrap improves (relative to data-splitting) in terms of RMSE for a larger number of subjects.

## 8.4 Application of Algorithm 3 to fMRI Data

We implemented Algorithm 3, using a threshold of 1.2% BOLD on the UK Biobank fMRI data to obtain barplots, boxplots and graphs which can be interpreted in the same manner as the ones in the main text. See Figures 5.22 and 5.23.

## 8.5 Comparing the Bootstrap and Circular Inference at top peaks

The peak locations found by the bootstrap approach and circular inference are the same. As such we can directly compare the bias/RMSE at the location of the $n$th largest maxima. We have done this in the graph below for the top 15 maxima for the fMRI and VBM datasets. In order to compute these graphs for each $n$ we found the $n$th largest peak in the effect size image ($t$-statistic or partial $R^2$). For instance taking $n = 1$ gives us the maximum, $n = 2$ the second largest peak etc. Given a number of subjects $N$ and a peak rank $n$, we get $G_N = \lfloor 4940/N \rfloor$ peaks of rank $n$ and obtain estimates $\hat{\theta}^n_1, \dots \hat{\theta}^n_{G_N}$ for the values of the underlying effects $(\theta^n_1, \dots \theta^n_{G_N})$ at the locations of these peaks (using circular inference and the bootstrap methods). As in the main text (Section 2.5) the underlying effects take different values since the locations are different. As such for $k = 1, \dots, G_N$ we compare the differences $\hat{\theta}^n_k - \theta^n_k$ to 0 (where the $\theta^n_k$ are computed using the 4000 subject held out ground truth) and compute

$$\text{Bias}_n = \frac{1}{G_N} \sum_{k=1}^{G_N} \left( \hat{\theta}^n_k - \theta^n_k \right) \quad \text{and} \quad \text{RMSE}_n = \left( \frac{1}{G_N} \sum_{k=1}^{G_N} (\hat{\theta}^n_k - \theta^n_k)^2 \right)^{1/2}.$$

We have plotted these quantities against $n$ in Figures 5.24 and 5.25. We note that the graphs are somewhat variable because we have used the 4940 subjects to calculate

them meaning that for each $N = 50, 100, 150$ we only have $98, 49, 32$ (respectively) peaks for each $n$. From the graphs we see the bootstrap significantly outperforms circular inference (at all peak ranks). The bootstrap estimates are relatively unbiased and have low RMSE whereas the circular estimates are substantially positively biased and have a larger RMSE.

## 8.6 Derivations

### 8.6.1 Proofs for Section 7.3.2

Under the framework of Section 7.3.2, we have:

**Proposition 8.1.**

$$\frac{1}{N}X_N^T X_N \xrightarrow{a.s.} \mathbb{E}\big[x_1 x_1^T\big],$$

$$\hat{\beta}_N \xrightarrow{a.s.} \beta$$

*and*

$$\hat{\sigma}_N^2 \xrightarrow{a.s.} \sigma$$

*as $N \longrightarrow \infty$, where $\xrightarrow{a.s.}$ denotes pointwise almost sure convergence.*

*Proof.*

$$\frac{1}{N}X_N^T X_N = \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T \xrightarrow{a.s.} \mathbb{E}\big[x_1 x_1^T\big]$$

as $N \longrightarrow \infty$ by the strong law of large numbers (SLLN) as the variance of the $x_i$ is finite. As such

$$\hat{\beta}_N = (X_N^T X_N)^{-1} X_N^T Y_N = \beta + \left(\frac{1}{N}X_N^T X_N\right)^{-1} \frac{1}{N} X_N^T \epsilon_N \xrightarrow{a.s.} \beta$$

by applying Slutsky since the SLLN implies that $X_N^T \epsilon_N / N = \sum_{k=1}^{N} \epsilon_k x_k / N \xrightarrow{a.s.} 0$ as $N \longrightarrow \infty$ since by independence the expectation is 0. Note Cauchy-Schwartz and the finite variance conditions are used here in order to show that the expected absolute first moment is finite and thereby justify the convergence.

It follows that for every $\eta > 0$ there is some large enough $N$ such that $\|\hat{\beta}_N - \beta\|^2 < \eta$ and in particular for large enough $N$,

$$\frac{1}{N-p}\left\|X_N\beta - X_N\hat{\beta}_N\right\|^2 = \frac{1}{N-p}\sum_{i=1}^{N}\left(x_i^T\hat{\beta}_N - x_i^T\beta\right)^2 \leq \frac{\eta}{N-p}\sum_{i=1}^{N}\left\|x_i^T\right\|^2 \longrightarrow \eta\mathbb{E}\left\|x_i^T\right\|^2$$

which tends to zero as $N \longrightarrow \infty$ since $\eta$ can be made arbitrarily small. Also, for any

$\eta$ and large enough $N$, by Cauchy-Schwartz,

$$\frac{1}{N-p}\Big|(Y_N - X_n\beta)^T(X_N\beta - X_N\hat{\beta})\Big| = \frac{1}{N-p}\sum_i \epsilon_i |x_i^T\beta - x_i^T\hat{\beta}_N| \leq \frac{\eta^{1/2}}{N-p}\sum_i \epsilon_i \big\|x_i^T\big\|$$

so this converges to zero as $N \longrightarrow \infty$ by the SLLN.

$$\begin{aligned}\hat{\sigma}_N^2 &= \frac{1}{N-p}\|Y_N - X_N\hat{\beta}_N\|^2 \\ &= \frac{1}{N-p}\|Y_N - X_N\beta\|^2 + \frac{1}{N-p}\|X_N\beta - X_N\hat{\beta}\|^2 + \frac{2}{N-p}(Y_N - X_N\beta)^T(X_N\beta - X_N\hat{\beta}_N).\end{aligned}$$

So $\hat{\sigma}_N^2 \xrightarrow{a.s.} \sigma^2$ as $N \longrightarrow \infty$ since the last two terms tend to 0 and the first equals $\frac{1}{N-p}\sum_i^N \epsilon_i^2$ and converges by the SLLN. □

## 8.7 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 $v$ 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 $ne(v)$.

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 $v' \in ne(v)$. In 3D brain images we have a rectilinear grid of voxels and 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 such which voxels are defined to be local maxima is dependent on the connectivity criterion. In this paper we have used a neighbourhood criterion of 18 which is the default in SPM.

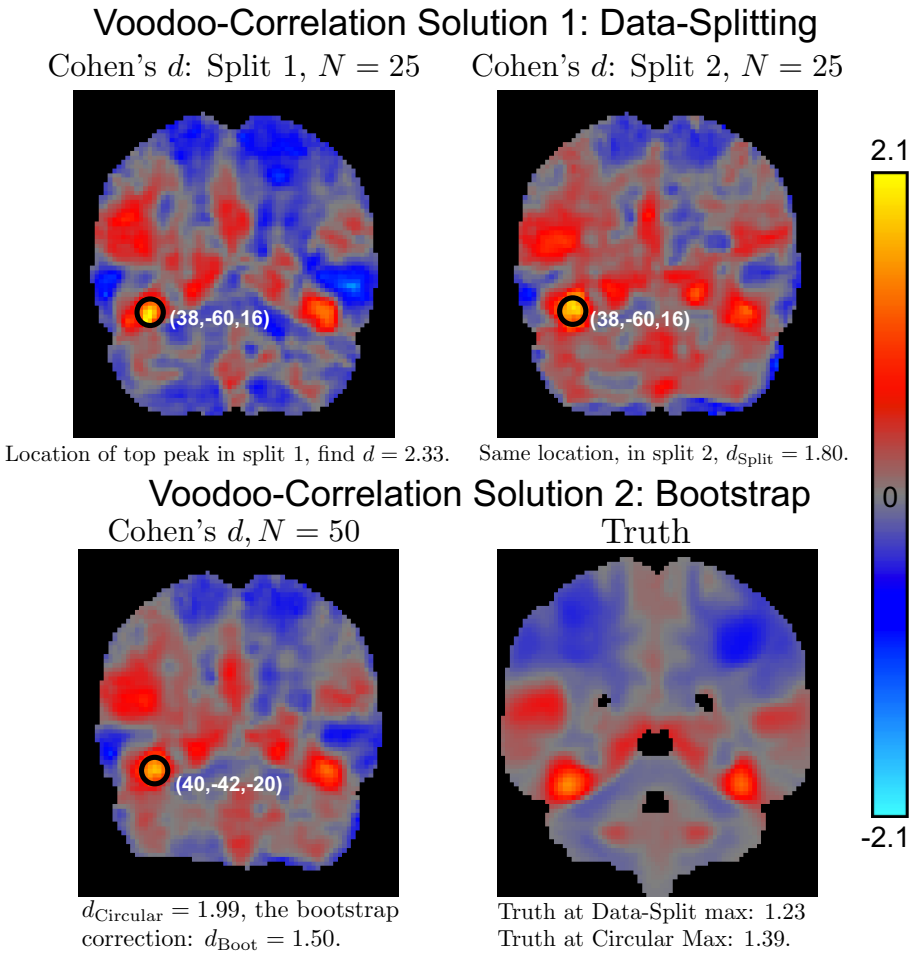



Figure 5.4: Illustration of a 50 subjects analysis using circular inference, the proposed bootstrap method and data-splitting. Data-splitting requires splitting the data in half, using the first half of the data to determine the locations, but provides a non-circular estimate. Circular inference and the bootstrap both use all of the data to calculate the locations. This figure illustrates the extra variability in data-splitting, with a noisier map and greater variability (estimated $d$ of 1.80 vs truth of 1.23); the bootstrap estimate had smaller bias (estimated $d$ of 1.50 vs truth of 1.39). This illustration is indicative of the extensive evaluations reported below.

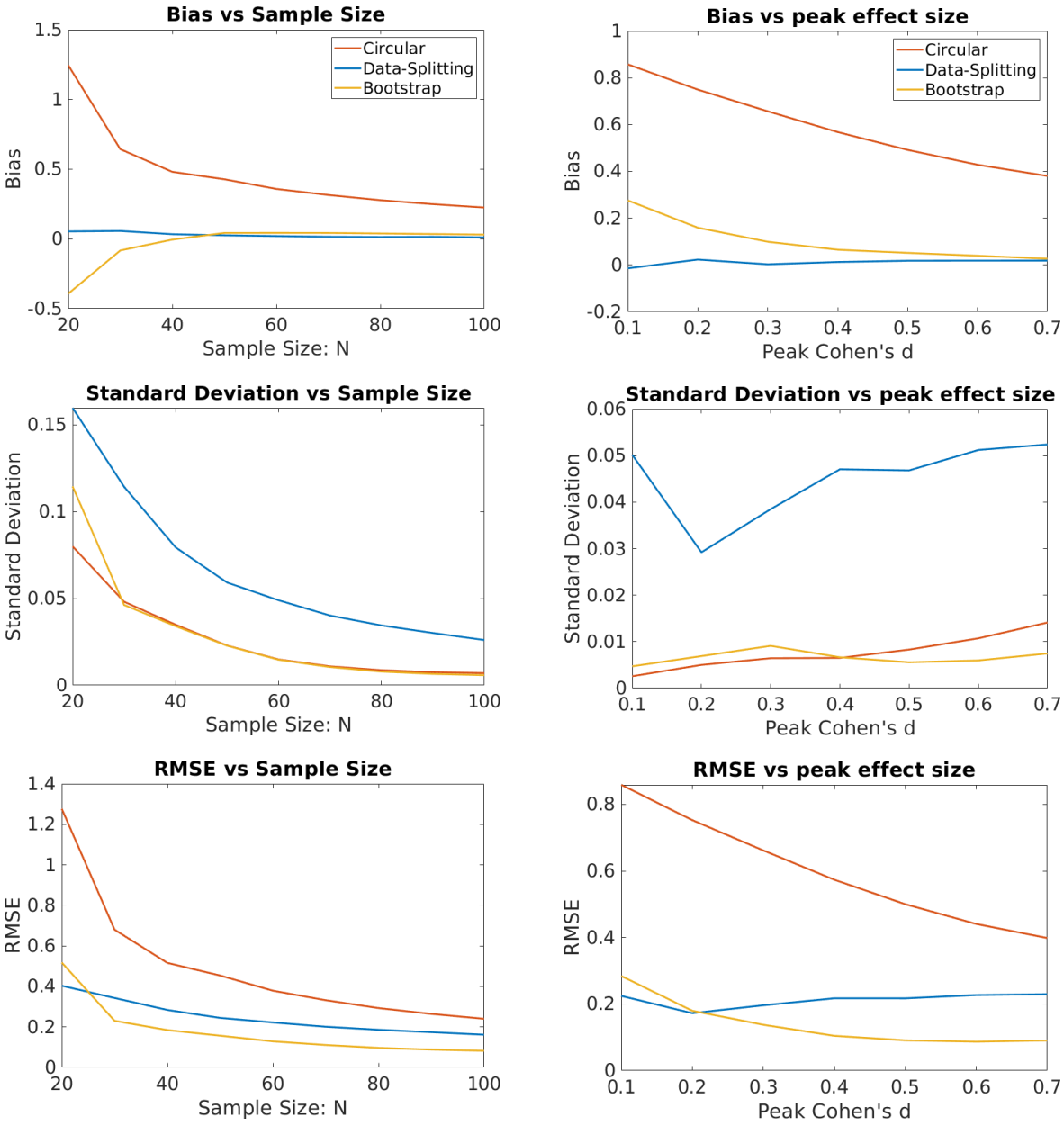



Figure 5.5: Evaluation as sample size and peak effect size changes of bias correction for Cohen's $d$ peaks (Algorithm 4) on simulated data generated as described in Section 2.3. Left column takes the effect size to be fixed and looks at how the measures change with sample size. Right column takes the number of subjects to be 50 and looks at how the measures change when the effect size is scaled. Each plot shows the bias (top), standard deviation (middle) and RMSE (bottom) calculated over 1,000 realisations. By the overall measure of RMSE, the bootstrap method performs the best except for the smallest effect sizes and sample sizes.

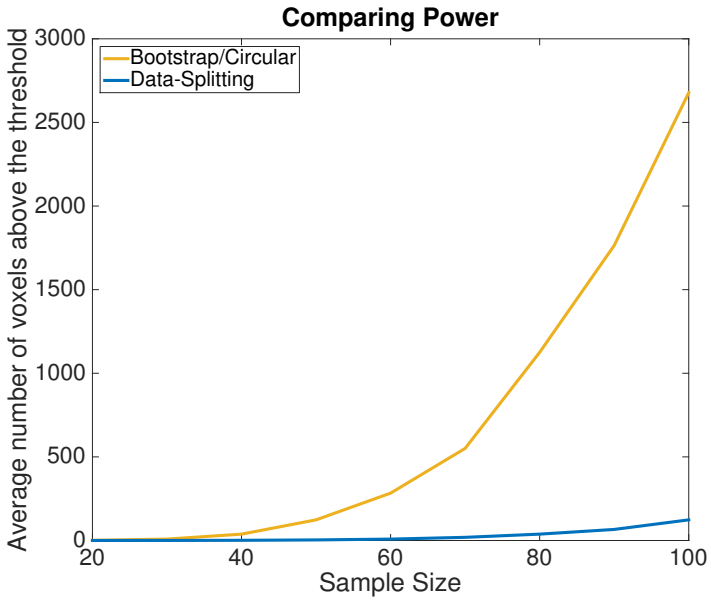



Figure 5.6: Comparing the power of the methods applied to the one-sample Cohen's $d$ simulations. We have computed the average number of significant voxels found above the threshold per realization for $N \in \{20, 30, \ldots, 100\}$ over all 1,000 realizations. Circular inference and the bootstrap find considerably more voxels to be significant than data-splitting.

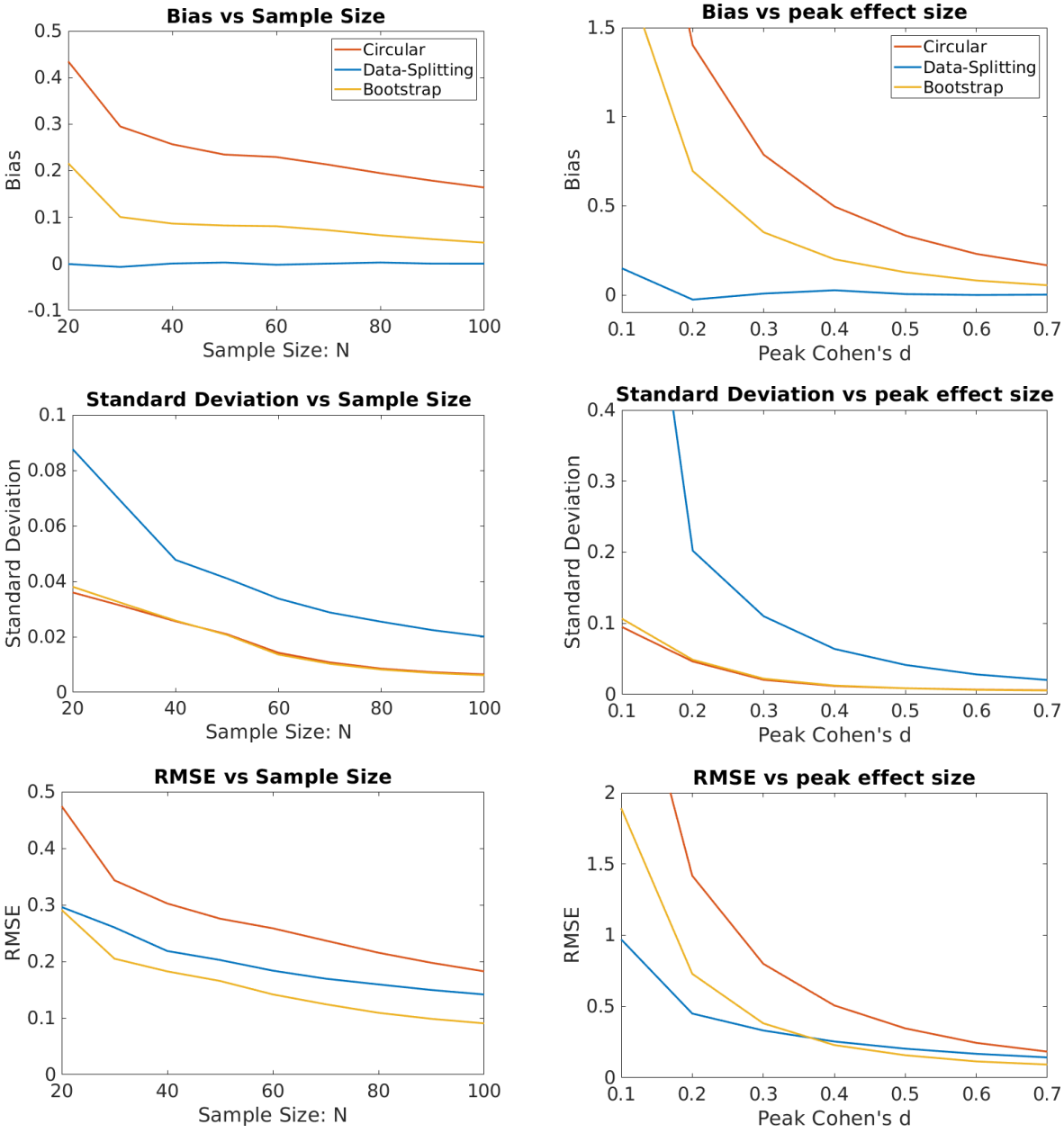



Figure 5.7: Evaluation as sample size and peak effect size changes of bias correction for the %BOLD mean at locations of Cohen's $d$ peaks (Algorithm 4) on simulated data generated as described in Section 2.3. Left column takes the effect size to be fixed and looks at how the measures change with sample size. Right column takes the number of subjects to be 50 and looks at how the measures change when the effect size is scaled. Each plot shows the bias (top), standard deviation (middle) and RMSE (bottom) calculated over 1,000 realisations. By the overall measure of RMSE, the bootstrap method performs the best except for the smallest effect sizes and sample sizes. The small effect sizes in the bottom right graph require a larger number of subjects before the bootstrap estimates attain a lower RMSE than data-splitting.

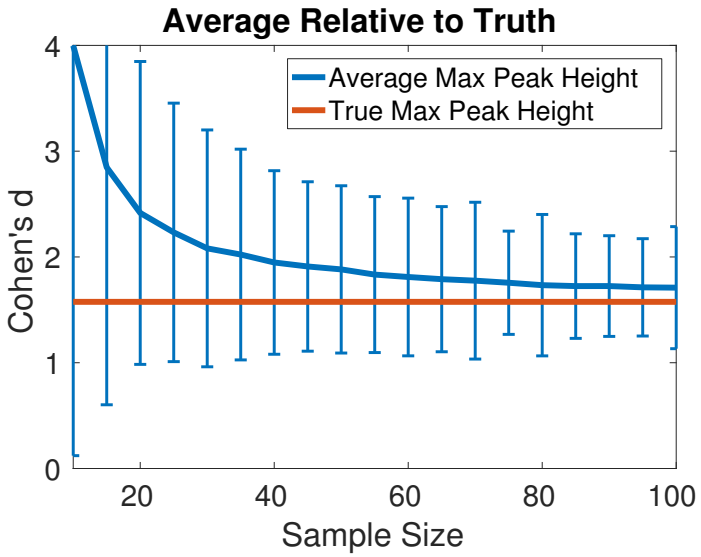



Figure 5.8: Illustration of the winner's curse: average selection bias in the one-sample Cohen's $d$: the average peak height of the maximum is plotted against $N$. For each $N$ we compute Cohen's $d$ for each of the $G_N$ groups of size $N$ find the value of the maximum and take the average over the $G_N$ 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 $N$.

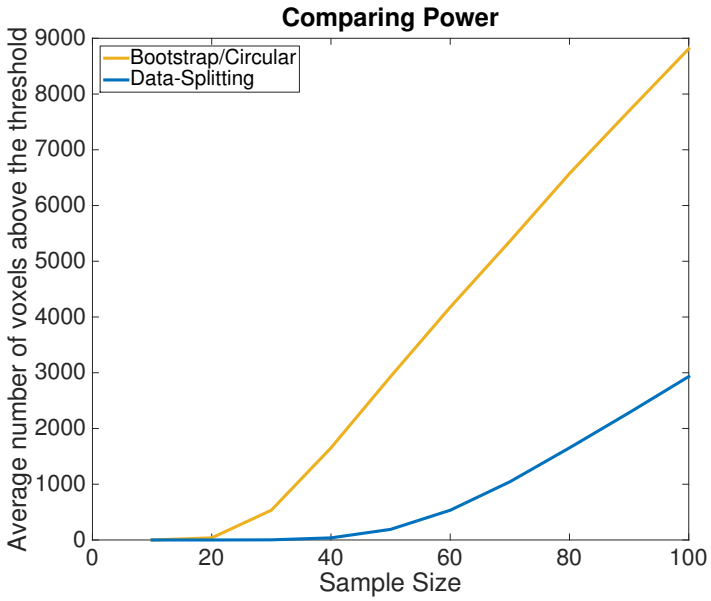



Figure 5.9: Comparing the power of the three methods. For each $N \in \{10, 20, \ldots, 100\}$ we consider the average (over the $G_N$ independent groups) number of voxels above the screening threshold. Circular inference and bootstrap find considerably more peaks than data-splitting.

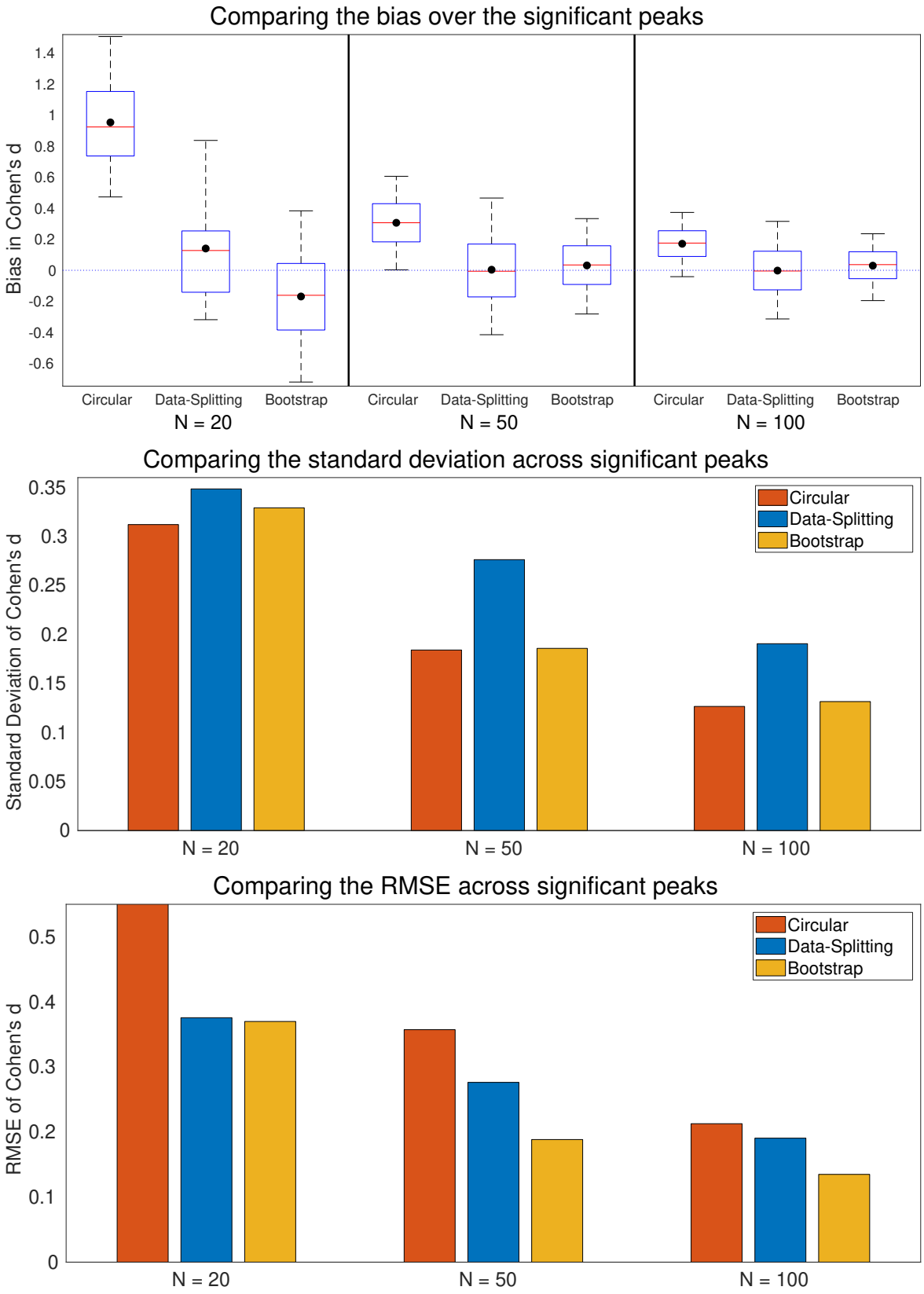



Figure 5.10: Comparison of one sample Cohen's $d$ estimation for task fMRI. Bias (top), standard deviation (middle), and RMSE (bottom) are shown for $N = 20, 50$ and 100 sample sizes, based on $G_N$ samples. While both data-splitting and bootstrap are generally unbiased, the bootstrap has the smallest RMSE. Note that the $N = 20$ data-splitting values are computed using only 7 data points so may not be representative. Note also that the RMSE for circular inference for $N = 20$ is cutoff and actually has a value of 1.0029.

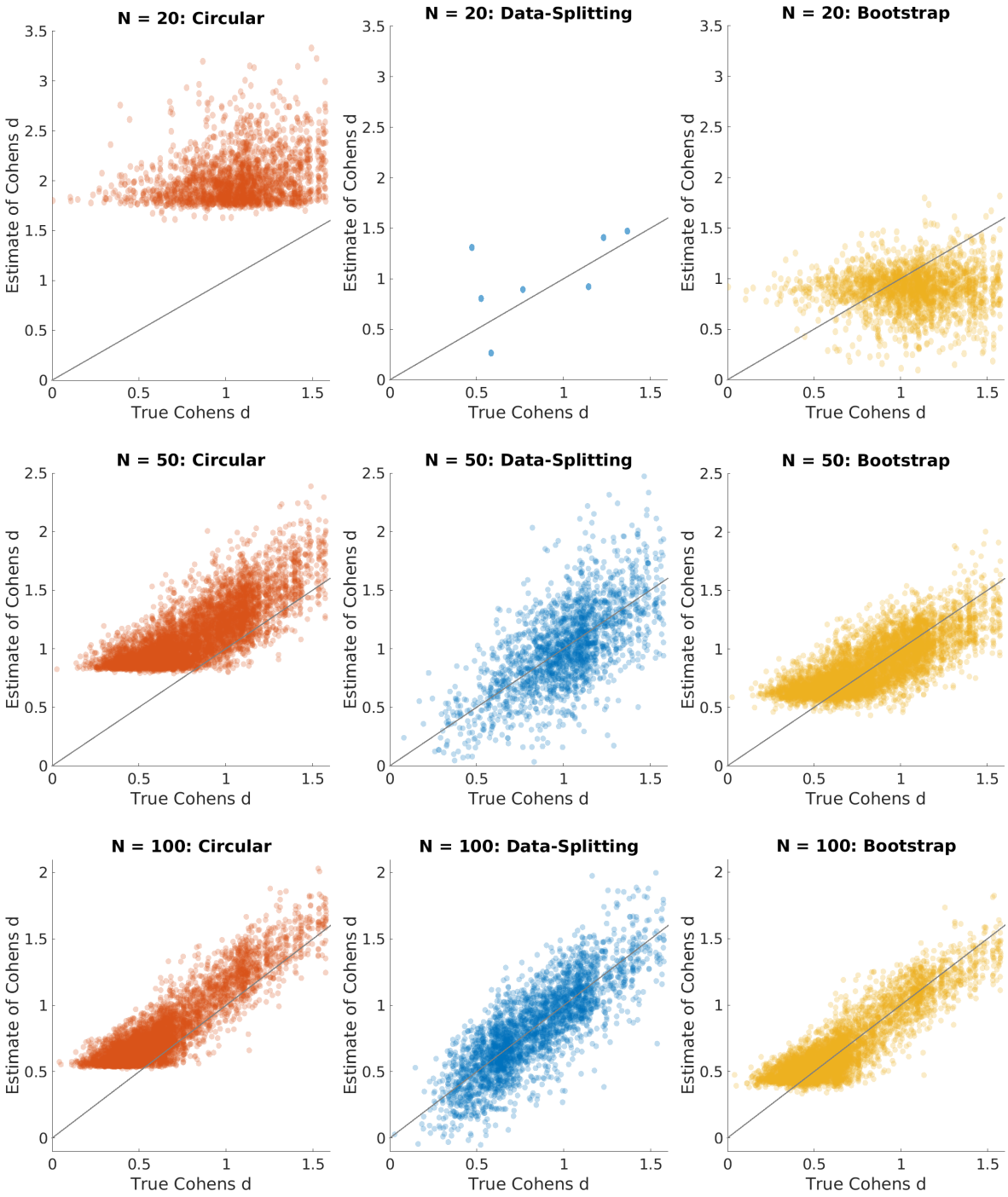



Figure 5.11: Plots of estimated versus true value of one-sample Cohen's $d$ for task fMRI images, for circular (left), data-splitting (middle), and bootstrap (right). Plots show all peaks found over the $G_N$ samples for each sample size, $N = 20, 50, 100$ (top to bottom). For each peak the true Cohen's $d$ is obtained at that location from the held-out $4,000$ subject Cohen's $d$ image. Note that the number of peaks and their locations are the same for circular inference and the bootstrap but are different for data-splitting as 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 standard deviation and improve as the sample size increases. The data-splitting estimates are unbiased but are more variable and reflect fewer detected peaks.

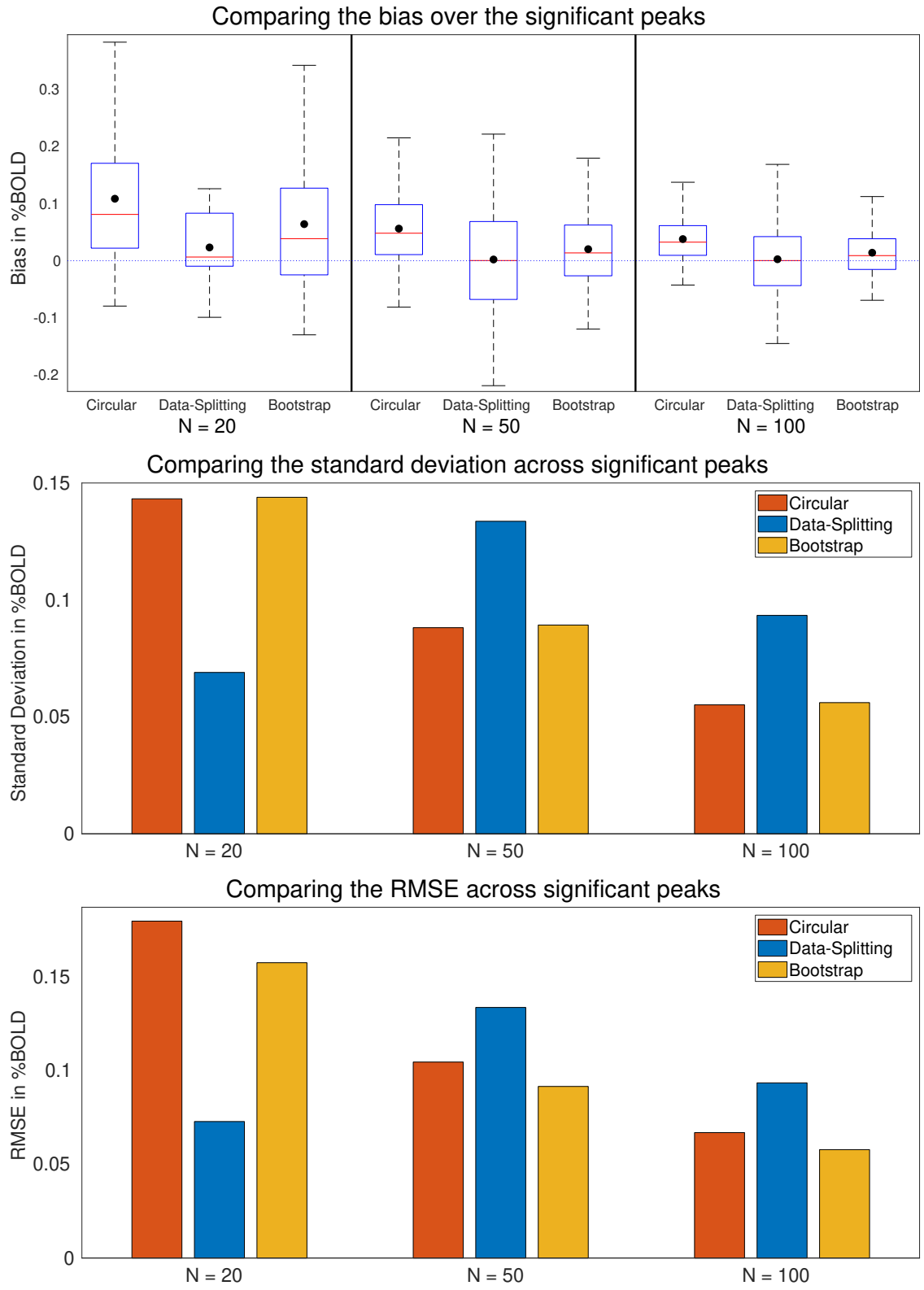



Figure 5.12: Comparison of one sample mean estimation for task fMRI. Bias (top), standard deviation (middle), and RMSE (bottom) are shown for $N = 20, 50$ and 100 sample sizes, based on $G_N$ samples. While both data-splitting and bootstrap are generally unbiased, the bootstrap has the smallest RMSE for larger sample sizes. Note that the $N = 20$ data-splitting RMSE and standard deviation is computed using only 7 data points so may not be representative. What is particular of note here is that the circular estimates have lower RMSE than data-splitting for $N = 50$ and 100.

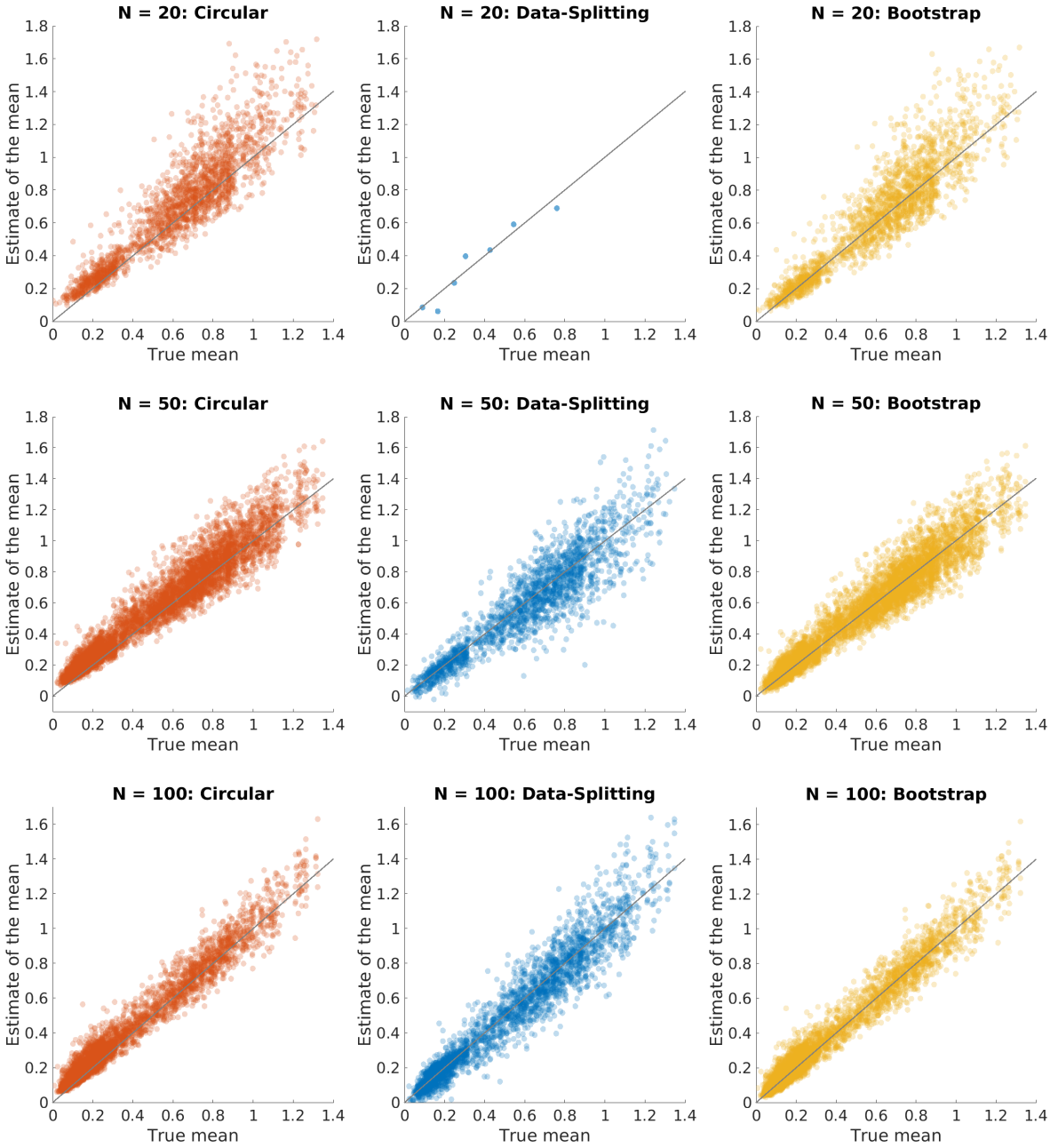



Figure 5.13: Plots of estimated versus true value of the one-sample mean (in %BOLD) for task fMRI images, for circular (left), data-splitting (middle), and bootstrap (right). Plots show all peaks found over the $G_N$ samples for each sample size, $N = 20, 50, 100$ (top to bottom). For each peak the true sample mean is obtained at that location from the held-out $4,000$ subject sample mean image. Note that the number of peaks and their locations are the same for circular inference and the bootstrap but are different for data-splitting as 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 reflect fewer detected peaks.

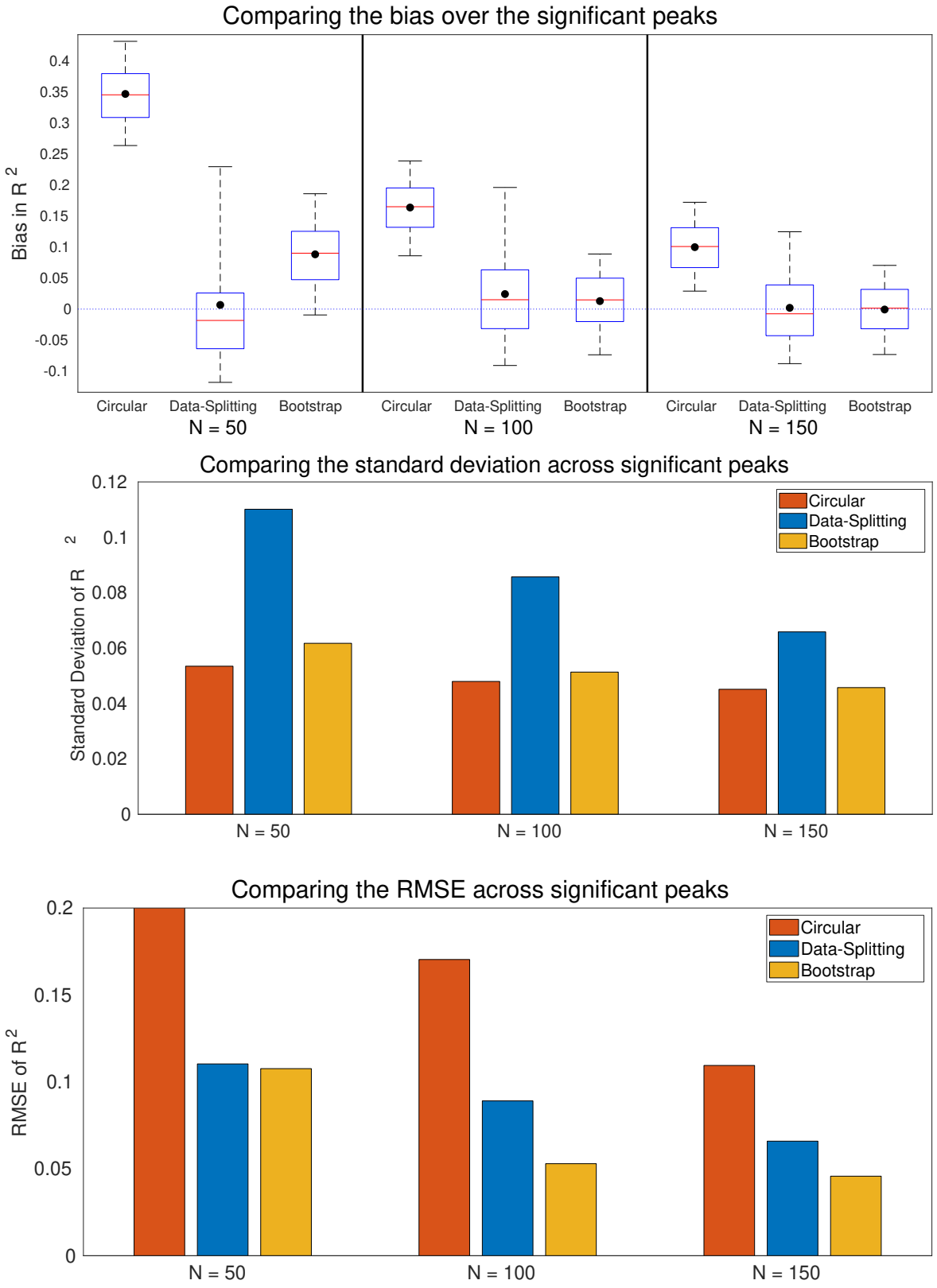



Figure 5.14: Comparison of estimates for the partial $R^2$ for age in the presence of sex and an intercept on VBM data. Bias (top), standard deviation (middle), and RMSE (bottom) are shown for $N = 50, 100$ and 150 sample sizes, based on $G_N$ samples. While both data-splitting and bootstrap are generally unbiased, the bootstrap has the smallest RMSE for all sample sizes. Note that RMSE for the circular estimates in the $N = 50$ case is 0.3407 and so is cut off by the graph.

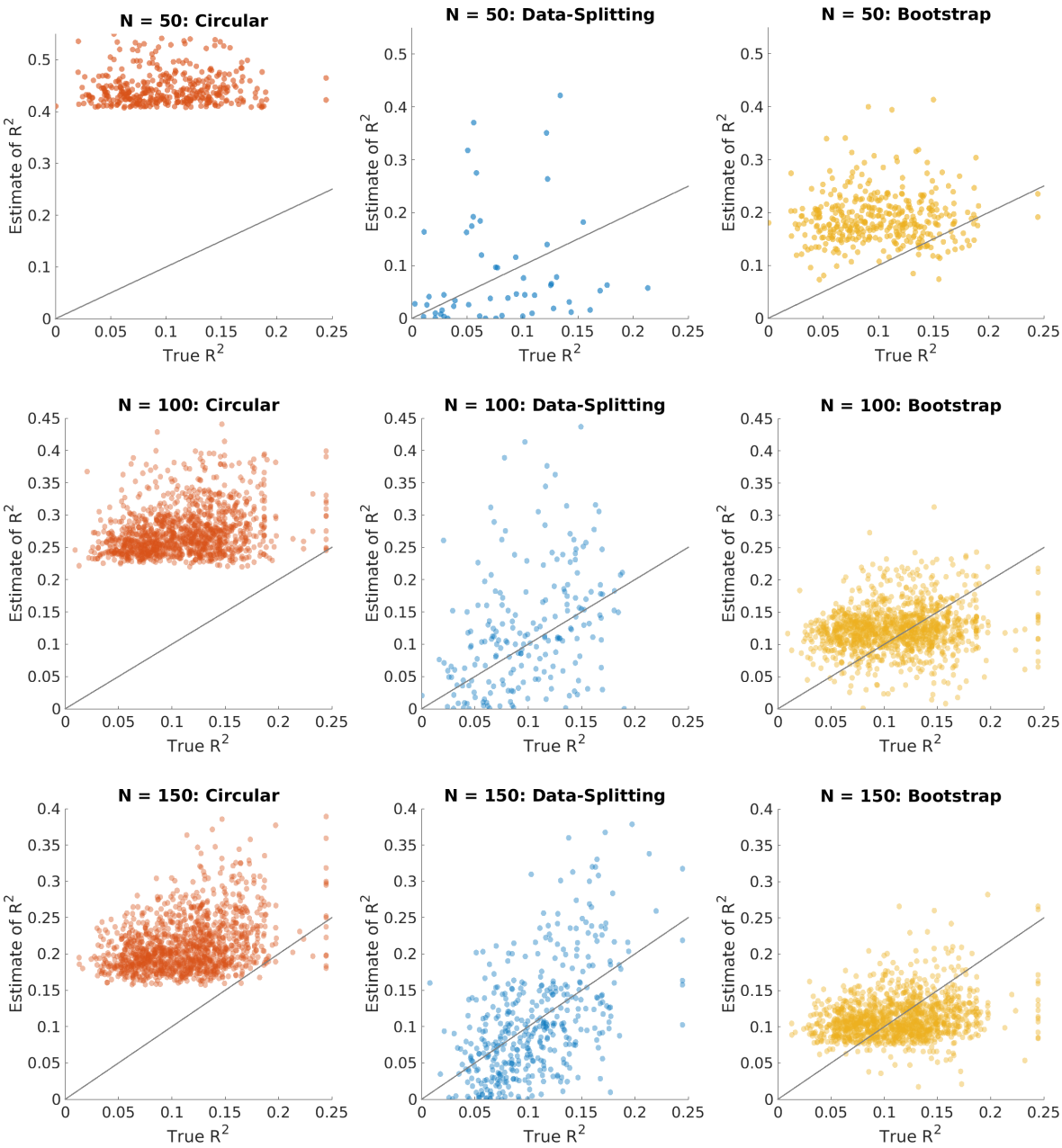



Figure 5.15: Plots of estimated versus true value of the partial $R^2$ for age, obtained using a GLM regression on VBM data, for circular (left), data-splitting (middle), and bootstrap (right). Plots show all peaks found over the $G_N$ samples for each sample size, $N = 50, 100, 150$ (top to bottom). For each peak the true partial $R^2$ for age is obtained at that location from the held-out $4,000$ subject partial $R^2$ image. Note that the number of peaks and their locations are the same for circular inference and the bootstrap but are different for data-splitting as 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 while the bootstrap and data-splitting estimates are unbiased on average. Data-splitting is the most variable, though the bootstrap over corrects values with a large true partial $R^2$ and under corrects those with a low partial $R^2$ for $N = 100, 150$. On average the bootstrap estimates lie closer to the identity line than the data-splitting estimates resulting in the decrease in RMSE, see Figure 5.14.

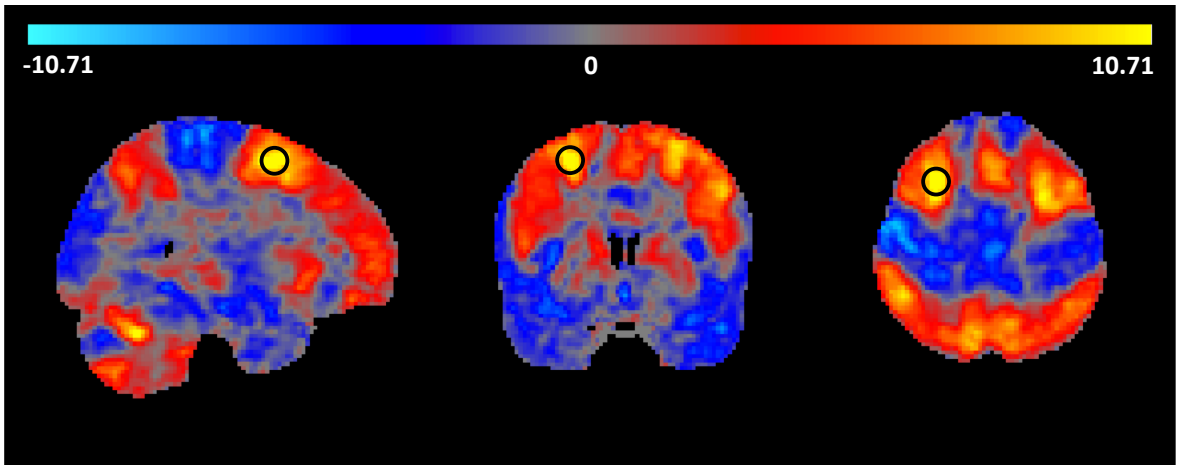



Figure 5.16: Slices through the one-sample $t$-statistic for the working memory contrast (2-back − 0-back) for subjects from the Human Connectome Project. Black circles indicate the location of the largest peak of activation which lies at the voxel (28, 8, 56) at the edge of the Medial Frontal Gyrus. At this location the observed (circular) Cohen's $d$ is 1.519, while the bootstrap-corrected value is 1.161; the observed %BOLD change at this voxel is %0.450 and corrected estimate is %0.433.

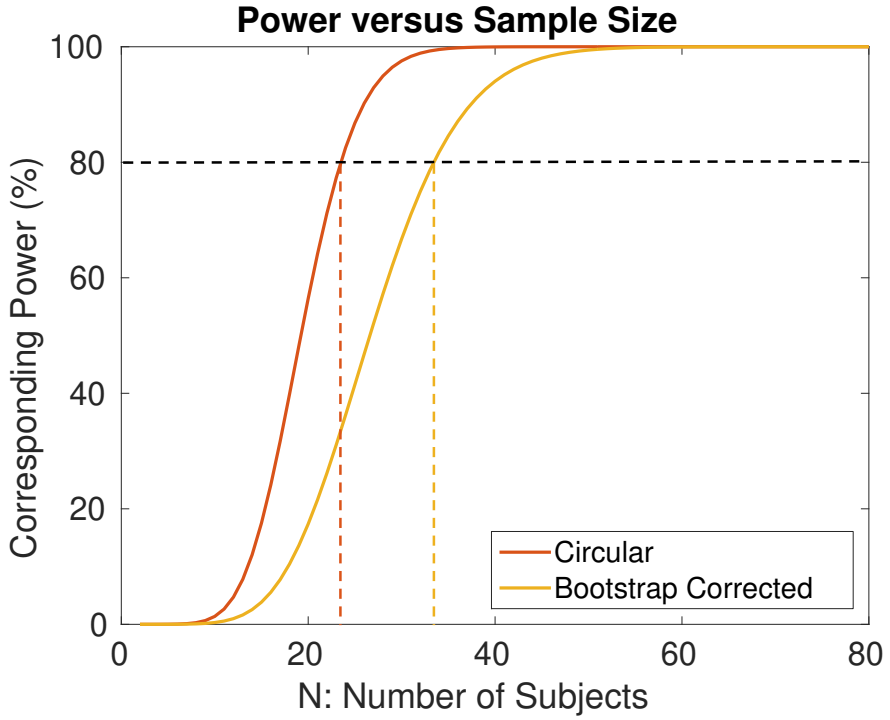



Figure 5.17: The power corresponding to a given sample size for the one-sample $t$-statistic when the $T$-statistic threshold is 5.10. The blue curve is the power curve corresponding to the circular estimate of the Cohen's $d$ for the HCP working memory dataset and the red curve is the power curve corresponding to the corrected Cohen's $d$. Using the raw value would suggest that only 24 subjects are need to attain 80% power, when in fact the corrected estimate shows that 34 subjects are needed to provide this level of power.

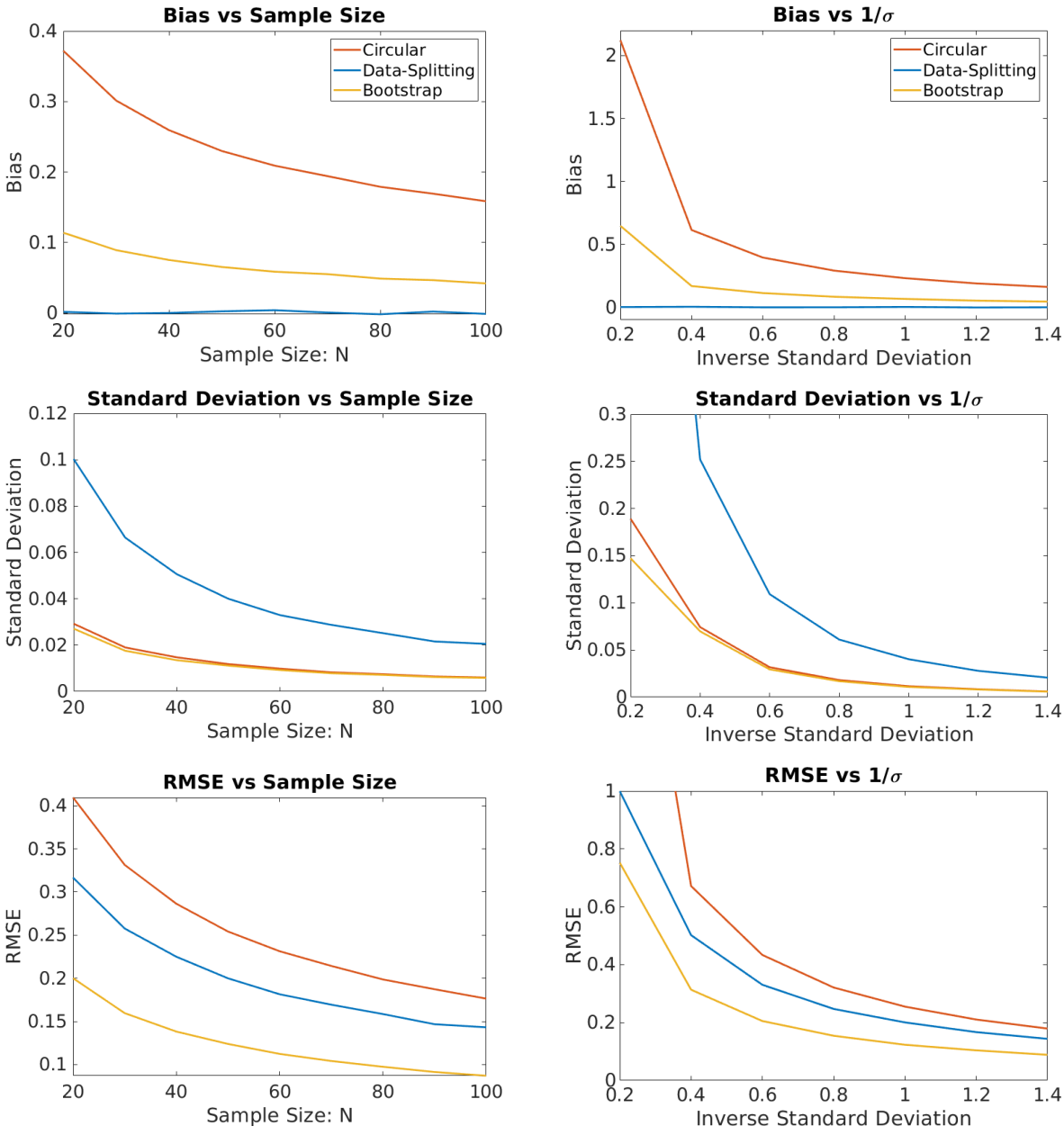



Figure 5.18: Evaluation as sample size and variance change of bias correction for peaks of the mean (Algorithm 3) on simulated data generated as described in Section 2.3.1. Left column looks at how the measures change with sample size (where the underlying signal and variance are fixed). Right column takes the number of subjects to be 50 and looks at how the measures change with the variance. Each plot shows the bias (top), standard deviation (middle) and RMSE (bottom) calculated over 1,000 realisations. By the overall measure of RMSE, the bootstrap method performs the best.

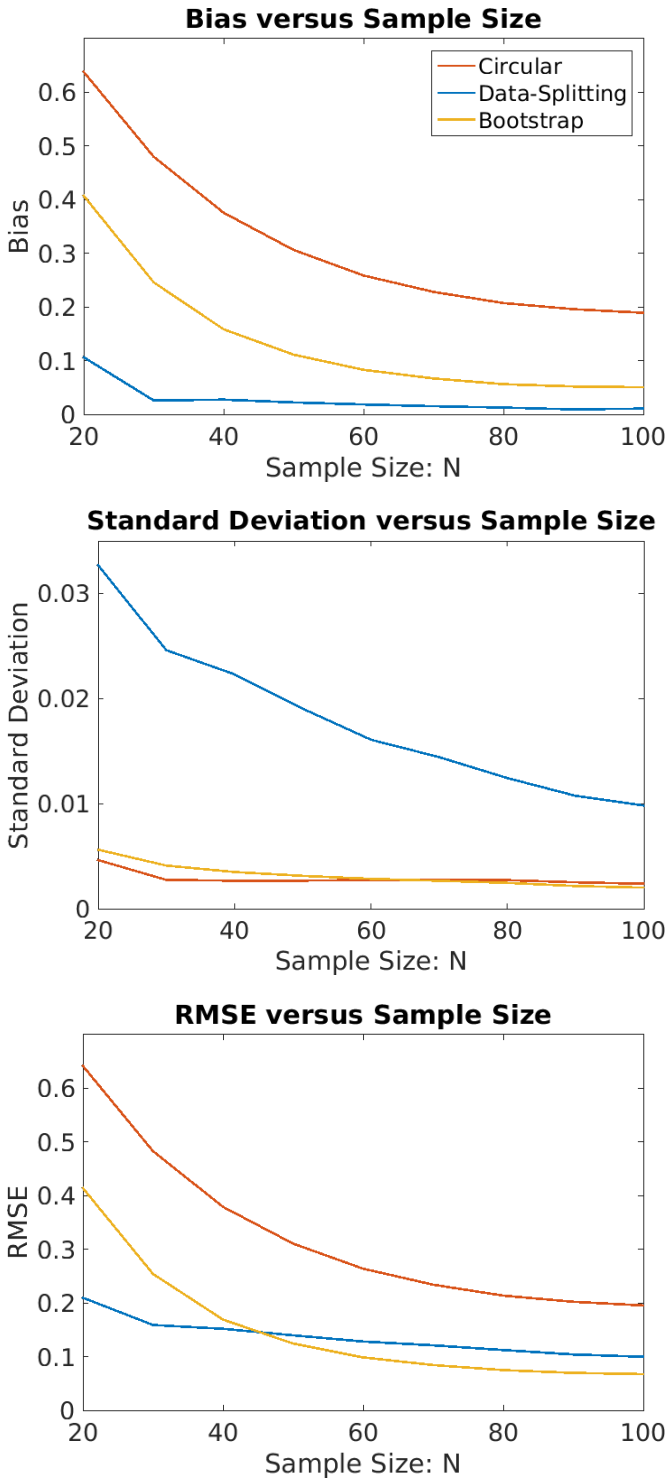



Figure 5.19: Evaluation of bias correction for $R^2$ peaks (Algorithm 5) as the number of subjects increases on simulated data generated as described in Section 8.2 (for a peak effect size of $\mu = 0.5822$). Each plot shows the bias (top), standard deviation (middle) and RMSE (bottom) averaged over 1000 realisations, for samples of size $N = 20, 30, \ldots, 100$. By the overall measure of RMSE, the bootstrap method performs the best so long as the sample size is sufficiently large.

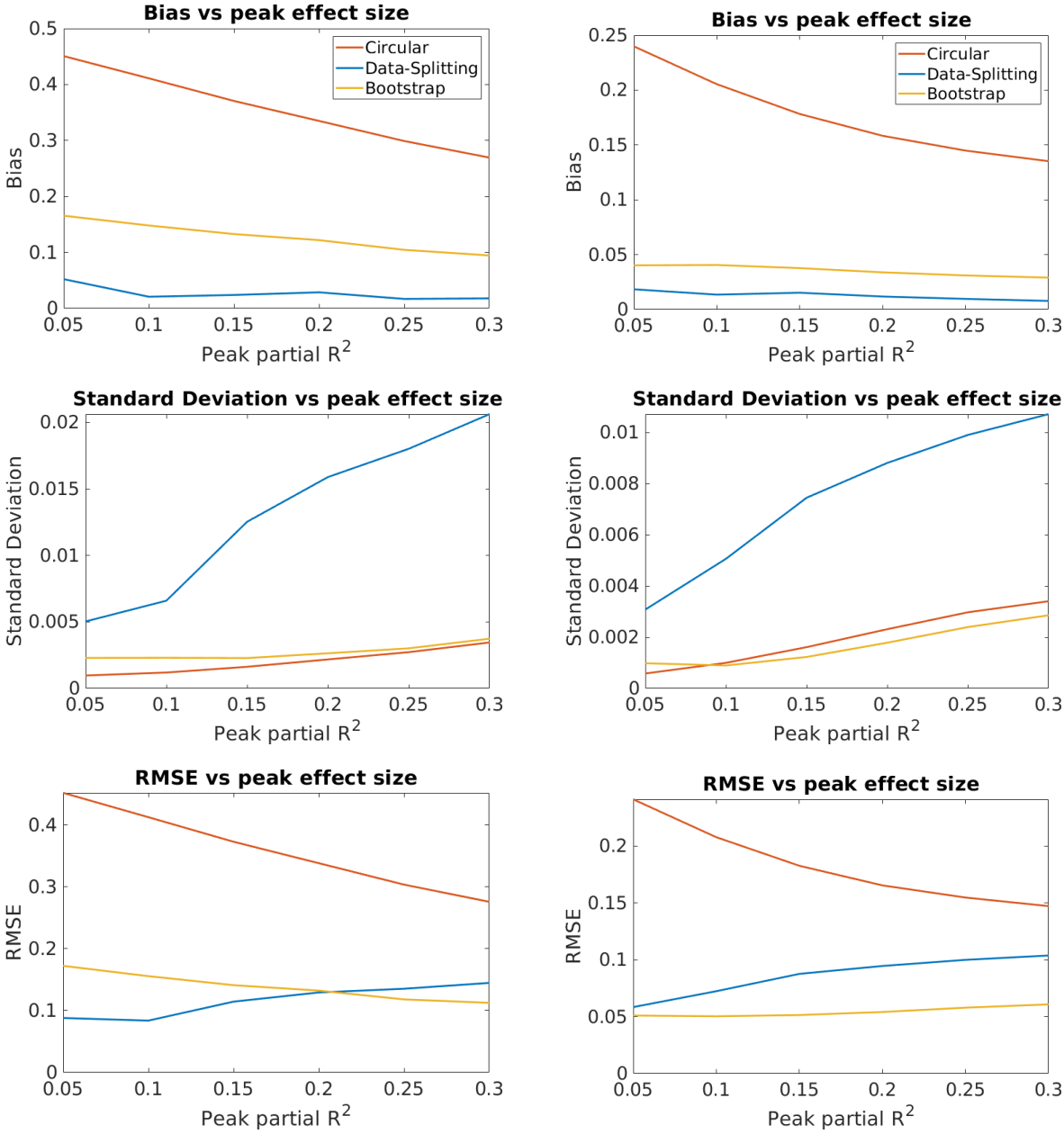



Figure 5.20: Evaluation as the variance (measured via the peak effect size) changes of bias correction for peaks of $R^2$ (Algorithm 5) on simulated data generated as described in Section 2.3.1. Left column takes $N = 50$ and the right column takes $N = 100$. Each plot shows the bias (top), standard deviation (middle) and RMSE (bottom) calculated over 1,000 realisations. Smaller effect sizes require a larger number of subjects before the bootstrap outperforms data-splitting in terms of RMSE.

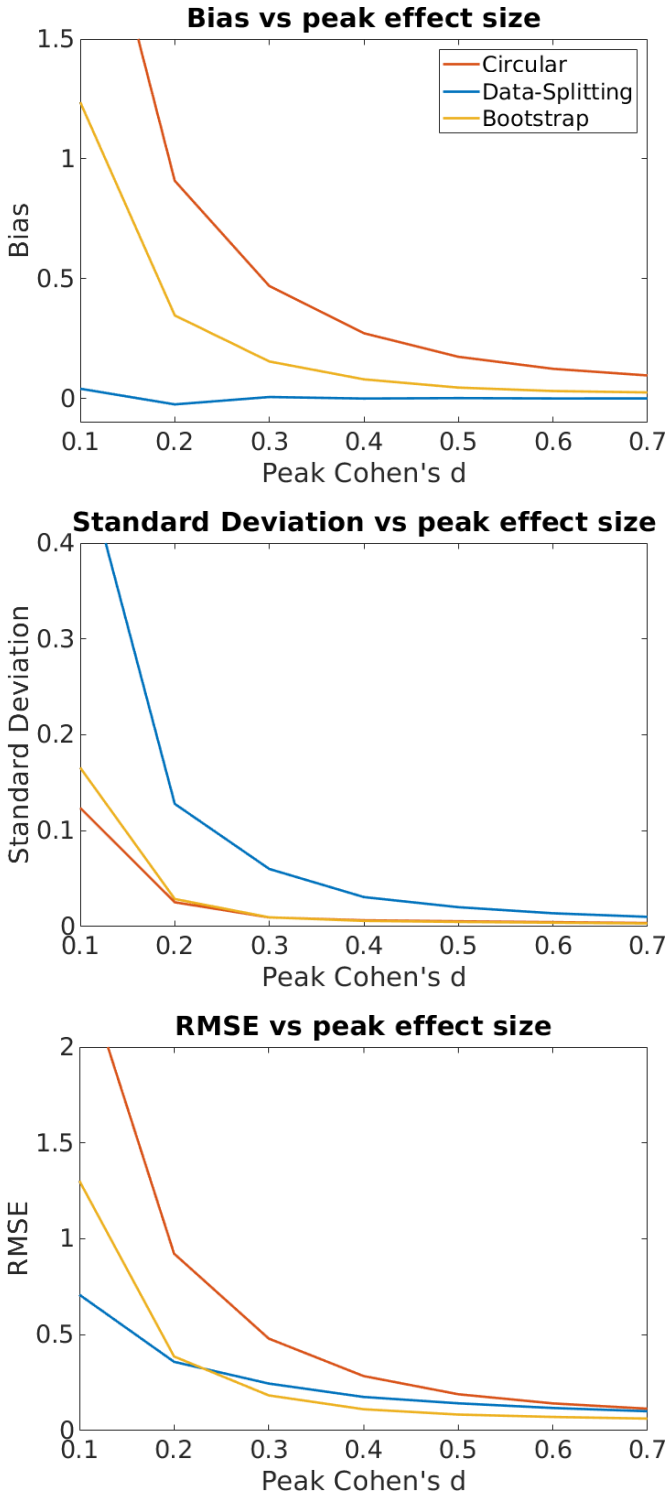



Figure 5.21: Evaluation as the variance (measured via the peak effect size) changes of bias correction for the %BOLD mean at locations of Cohen's $d$ peaks (Algorithm 4) on simulated data. These graphs are in the same setting as the graphs in Figure 5.7 (right column) but take $N = 100$ instead of $N = 50$ in order to illustrate that the bootstrap improves (relative to data-splitting) in terms of RMSE for a larger number of subjects. This supports the trend shown in Figure 5.7 (left column).

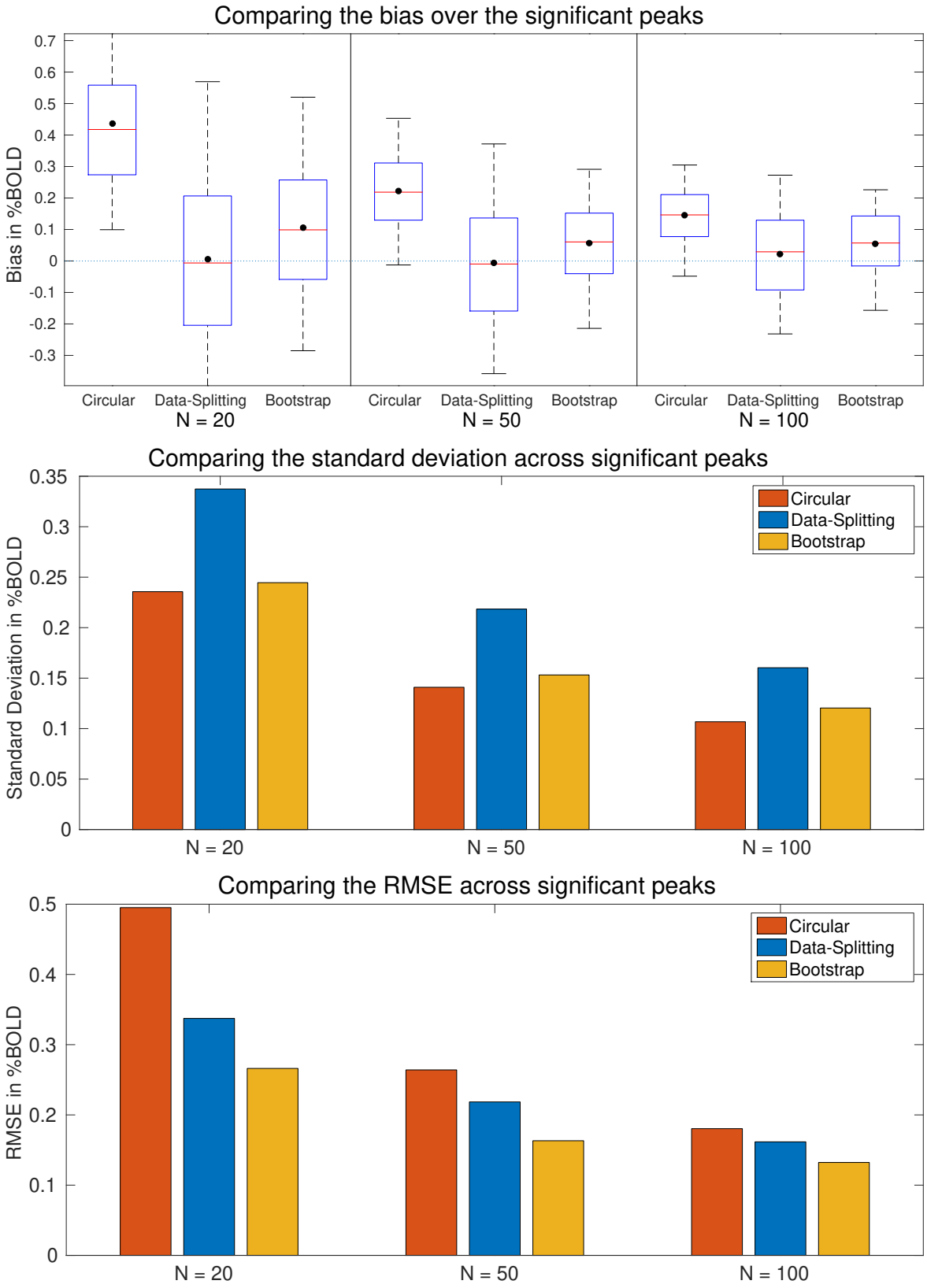



Figure 5.22: Comparison of estimates for the sample mean via Algorithm 3. Bias (top), variance (middle), and MSE (bottom) are shown for $N = 20, 50$ and 100 sample sizes, based on $G_N$ samples. The bootstrap estimates display some bias but have lower MSE than the data-splitting estimates.

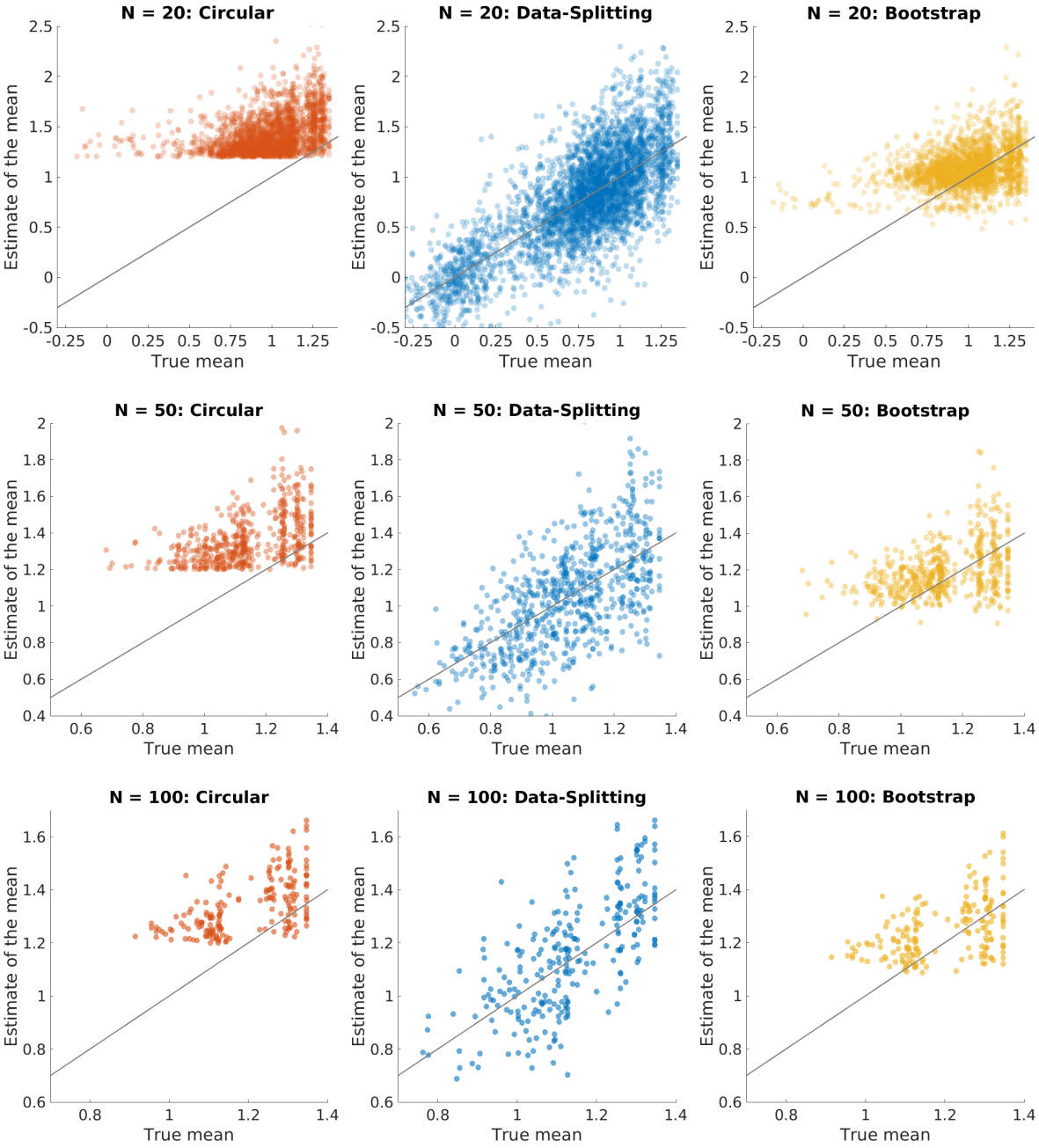



Figure 5.23: Plots of estimated versus true value of the sample mean, for circular (left), data-splitting (middle), and bootstrap (right). Plots show all peaks found over the $G_N$ samples for each sample size, $N = 20, 50, 100$ (top to bottom). For each peak the true sample mean is obtained at that location from the held-out 4000 subject mean image. Note that the number of peaks and their locations are the same for circular inference and the bootstrap but are different for data-splitting as it uses the first half of the subjects in order to determine significant peaks. From these plots we see that the data-splitting estimates are unbiased but are more variable than the bootstrap and circular estimates.

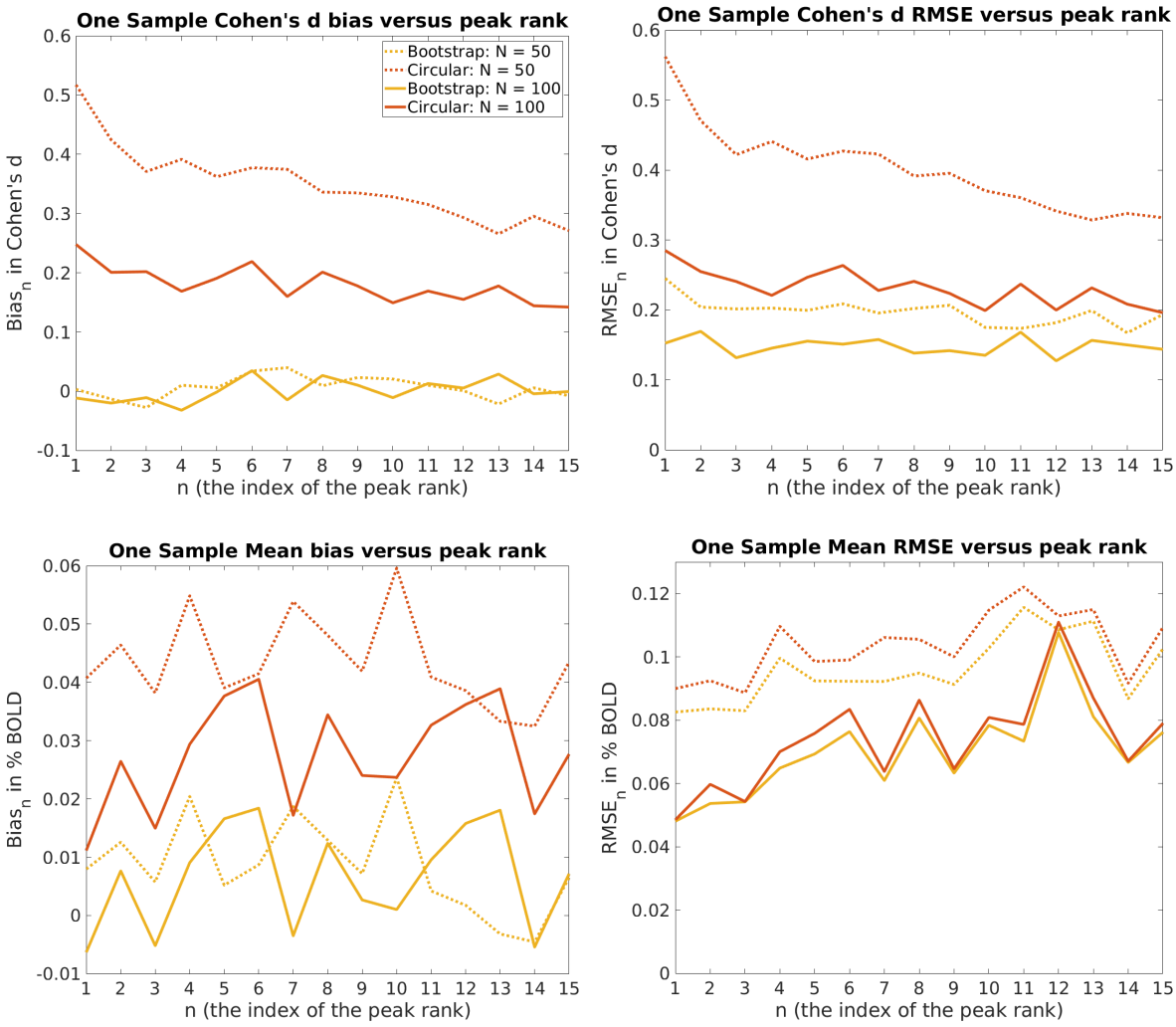



Figure 5.24: Comparison of estimates, at the top peaks of the t-statistic of the one-sample mean and Cohen's $d$, for task fMRI images. Here the bootstrap estimates are calculated using Algorithm 4. Plots of the bias are shown on the left and plots of the RMSE are shown on the right. The bootstrap curves (shown in yellow) always lie below those of the circular curves (shown in red). In particular they show that at each peak rank the bootstrap has low bias and low RMSE relative to circular inference.

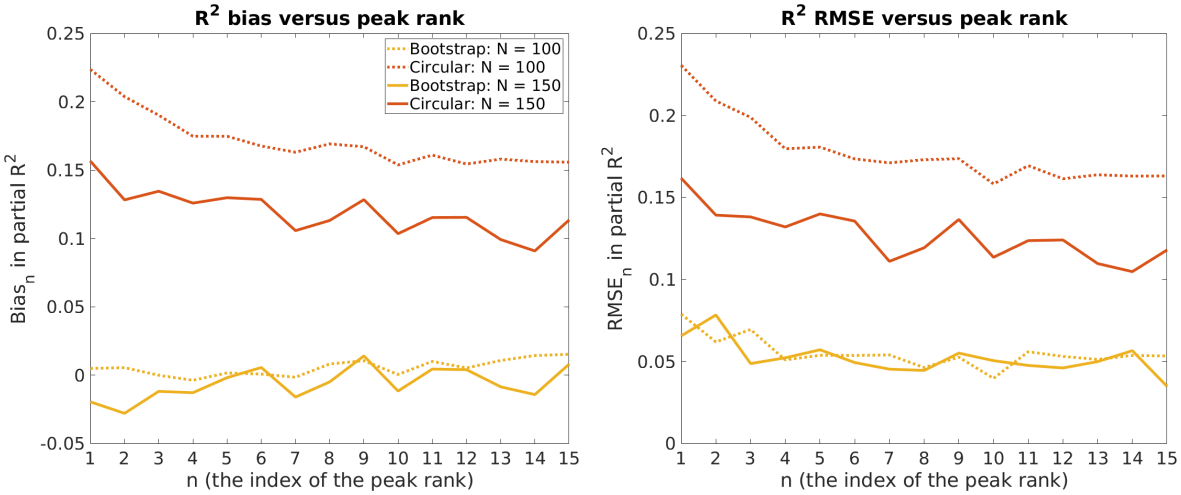



Figure 5.25: Comparison of estimates at the top peak ranks for the partial $R^2$ for age obtained using a GLM regression on VBM data. Plots of the bias are shown on the left and plots of the RMSE are shown on the right. The bootstrap curves (shown in yellow) always lie below those of the circular curves (shown in red). In particular they show that at each peak rank the bootstrap has low bias and low RMSE relative to circular inference.

# Chapter 6

# Conclusion

In this thesis we have provided a range of methods to infer on random fields and have demonstrated how they can be applied in practice, with a particular focus on applications in neuroimaging.

Our voxelwise RFT framework (Chapter 2) provides parametric inferences that work under weaker assumptions than are traditionally made, allowing it to work well in practice. There we developed the notion of convolution random fields, which we first introduced in Telschow et al. (2020b). These can be used to bridge the gap between the lattice observations and continuous theory. Their introduction into the literature represents an important step forward because the theory behind continuous random fields is well developed (Adler (1981), Adler and Taylor (2007)), while much less is known about realizations of random fields on a discrete lattice.

We also introduced a Gaussianization procedure that allows RFT to work in practice. RFT methods have been used in fMRI for many years under the assumption that the data was sufficiently Gaussian for them to work well. We have shown here that this is not the case: fMRI data is highly non-Gaussian and this causes problems for RFT inference when it is not accounted for. When the original data is used the

Euler characteristic is miscalculated because the heavy tails affect the distribution of the maximum of the test-statistic field. When the data is Gaussianized we are able to correctly predict the EEC and accurately control the FWER given sufficiently many subjects. The extent to which the failure of the Gaussianity assumption affects other models currently used in fMRI analysis remains to be investigated and this is an important question for future work. The Gaussianization procedure itself deserves a lot more attention and we hope to investigate it in detail in future papers.

The availability of large amounts of data from the UK Biobank have enabled us to perform extensive validation of our methods. In Chapter 2 we did so in order to show that we correctly controlled the false positive rates and did so in Chapter 5 to show that our methods correctly estimated the effect size at peaks. These validations were only possible because of the large amount of data available in the UK Biobank. We recommend that any and every analysis method (new and old) is tested using these types of validations in order to ensure that they perform as expected (and control the false positive rate/provide correct estimation of the effect size). As far as we know these validations are the largest of their kind that have been used to test the performance of methods in fMRI. The others that we are aware of, that have been used to estimate the false positive rate, are those used in Eklund et al. (2016), Eklund et al. (2019) and Lohmann et al. (2018) (which resampled subsets from datasets consisting of at most only 198 subjects). It is our sincere hope that this type of rigorous large scale testing will become standard practice.

In Chapters 3 and 5 we provided methods to infer on the location and the height of a peak of the signal in a random field. This is especially important in the Biobank era, where the large number of subjects means that it becomes more interesting to understand features of the underlying signal. In fMRI this is useful because the data

has a mean that is different from zero everywhere. As such, when the effect size is small, a null hypothesis testing framework is appropriate but when it is larger, depending on the effect size, this may no longer be the right approach. One way to get around this difficulty when doing hypothesis testing is to change the null to be about activation above a given level and there has been some work on this (Sommerfeld et al. (2018), Bowring et al. (2019), Bowring et al. (2020)) however there is much more theory to be developed in this area. Moreover, their approach must be tested using a big data validation on real fMRI data, before we can recommend its use in practice. Further development of these methods and others, to infer on the signal, represents a much needed area of research.

In Chapter 4 we derived distributions for the size of clusters above high thresholds which are valid under non-stationarity. We also proved results on the non-degeneracy of convolution fields that are necessary to show that they satisfy the Gaussian Kinematic Formula and justify the use of our voxelwise inference framework. We hope that this work (in combination with the results from Chapter 2) will provide a basis with which it will be possible to develop a valid non-stationary parametric clustersize inference framework. One way of doing so would be to extend the work of Chumbley and Friston (2009), Schwartzman et al. (2011) and Cheng and Schwartzman (2017), using the marginal clustersize distribution to obtain the cluster based $p$-values and to control the FDR over clusters. Clustersize inference is typically more powerful than other methods for detecting activation in spatial data and so these are exciting avenues for future research.

One of the difficulties here is that the distribution for the number of clusters in a non-stationary Gaussian random field is unknown: modelling this as Poisson seems reasonable though this has only been shown to hold under stationarity (Aldous (2013),

Adler (1981)) and in 1D (Azaïs and Mercadier (2003)). Whether this is reasonable in fMRI data remains to be seen. A second issue is that the covariance structure of the brain is symmetric, between hemispheres, resulting in substantial long range dependence. This caused some conservativeness in Chapter 2 at higher levels of smoothness. It may be possible to account for this in the model and improve the level of voxelwise FWER control provided by RFT. However, clustersize inference has traditionally required the distribution of the size of the clusters be independent. This is probably not reasonable under long range dependence and so could cause problems in the analysis pipeline.

It is our hope that the methods we have developed provide useful tools for neuroimagers and others with which to discover and infer on signal in random fields. We anticipate extending these methods to work in a variety of other settings.

# Bibliography


Jason Abrevaya and Jian Huang. On the bootstrap of the maximum score estimator. *Econometrica*, 73(4):1175–1204, 2005.

Robert J. Adler. *The Geometry of Random Fields*. 1981.

Robert J. Adler and Jonathan Taylor. *Random Fields and Geometry*. 2007. ISBN 9780387481128.

Robert J Adler, Jonathan E Taylor, and Keith J. Worsley. *Applications of random fields and geometry: Foundations and case studies*. 2010.

David Aldous. *Probability approximations via the Poisson clumping heuristic*, volume 77. Springer Science & Business Media, 2013.

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166: 400–424, 2018.

Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

M Aronowich and Robert J Adler. Extrema and Level Crossings of Chi2 Processes. 18 (4):901–920, 1986.

Jean-Marc Azaïs and Cécile Mercadier. Asymptotic poisson character of extremes in non-stationary gaussian models. *Extremes*, 6(4):301–318, 2003.

Jean-Marc Azaïs and Mario Wschebor. *Level sets and extrema of random processes and fields*, volume 6. John Wiley & Sons, 2009.

Raghu Raj Bahadur. Rates of convergence of estimates and test statistics. *The Annals of Mathematical Statistics*, 38(2):303–324, 1967.

Maurice S Bartlett. The use of transformations. *Biometrics*, 3(1):39–52, 1947.

Yuri Konstantinovich Belyaev. Bursts and shines of random fields. In *Doklady Akademii Nauk*, volume 176, pages 495–497. Russian Academy of Sciences, 1967.

Yu K Belyayev et al. Point processes and first passage problems. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Probability Theory*. The Regents of the University of California, 1972.

Yoav Benjamini and Yosef Hochberg. 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, 1995.

Yoav Benjamini and Amit Meir. Selective correlations-the conditional estimators. *arXiv preprint arXiv:1412.3242*, 2014.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Spatial confidence sets for raw effect size images. *NeuroImage*, 203:116187, 2019.

Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Confidence sets for cohen's d effect size images. *bioRxiv*, 2020.

Ralph A Bradley and John J Gart. The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214, 1962.

Richard C Bradley Jr. Central limit theorems under weak dependence. *Journal of Multivariate Analysis*, 11(1):1–16, 1981.

Lavinia Braun. South Africa's Stepchild: Reading the work of Sarah Gertrude Millin with emphasis upon its Biblical and Judaic aspects (1909-1951). *PhD Thesis*, 1994.

Samuel L Braunstein. How large a sample is needed for the maximum likelihood estimator to be approximately gaussian? *Journal of Physics A: Mathematical and General*, 25(13):3813, 1992.

J Cao. The size of the connected components of excursion sets of $\chi$ 2, t and f fields. *Advances in Applied Probability*, 31(3):579–595, 1999.

Gang Chen, Ziad S Saad, Audrey R Nath, Michael S Beauchamp, and Robert W Cox. Fmri group analysis combining effect estimates and their variances. *Neuroimage*, 60 (1):747–765, 2012.

Gang Chen, Paul A Taylor, and Robert W Cox. Is the statistic value all we should care about in neuroimaging? *Neuroimage*, 147:952–959, 2017.

Dan Cheng. Excursion probabilities of isotropic and locally isotropic gaussian random fields on manifolds. *Extremes*, 20(2):475–487, 2017.

Dan Cheng and Armin Schwartzman. Distribution of the height of local maxima of Gaussian random fields. *Extremes*, 18(2):213–240, 2015a. ISSN 1572915X. doi: 10.1007/s10687-014-0211-z.

Dan Cheng and Armin Schwartzman. Distribution of the height of local maxima of gaussian random fields. *Extremes*, 18(2):213–240, 2015b.

Dan Cheng and Armin Schwartzman. Multiple testing of local maxima for detection of peaks in random fields. *Annals of Statistics*, 45(2):529–556, 2017.

Dan Cheng and Yimin Xiao. The mean Euler characteristic and excursion probability of Gaussian random fields with stationary increments. *Annals of Applied Probability*, 26(2):722–759, 2016. ISSN 10505164. doi: 10.1214/15-AAP1101.

Dan Cheng, Armin Schwartzman, et al. Multiple testing of local maxima for detection of peaks in random fields. *The Annals of Statistics*, 45(2):529–556, 2017.

Guang Cheng, Jianhua Z Huang, et al. Bootstrap consistency for general semiparametric m-estimation. *The Annals of Statistics*, 38(5):2884–2915, 2010.

Justin Chumbley, Keith Worsley, Guillaume Flandin, and Karl Friston. Topological fdr for neuroimaging. *Neuroimage*, 49(4):3057–3064, 2010.

Justin R. Chumbley and Karl J. Friston. False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(1):62–70, 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2008.05.021. URL `http://dx.doi.org/10.1016/j.neuroimage.2008.05.021`.

Harold Cramer. *Mathematical Methods of Statistics*. Asia Publishing House, 1946.

Samuel Davenport and Thomas E. Nichols. Selective peak inference: Unbiased estimation of raw and standardized effect size at local maxima. *NeuroImage*, 2020.

Samuel Davenport, Fabian Telschow, Armin Schwarzman, and Thomas E. Nichols. Accurate voxelwise fwer control in fmri using random field theory. 2021.

Anthony C Davison, David V Hinkley, and G Alastair Young. Recent developments in bootstrap methodology. *Statistical Science*, pages 141–157, 2003.

Daniel Dugué. Application des propriétés de la limite au sens du calcul des probabilités à l'étude de diverses questions d'estimation. 1937.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–483, 1978.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

S. B. Eickhoff, D. Bzdok, A. R. Laird, F. Kurth, and P. T Fox. Activation likelihood estimation revisited. *NeuroImage*, 59(3):2349–2361, 2012. doi: 10.1016/j.neuroimage.2011.09.017.Activation.

Simon B Eickhoff, Angela R Laird, Christian Grefkes, and Ling E Wang. Coordinate-based ALE meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. 30(9):2907–2926, 2009. doi: 10.1002/hbm.20718.Coordinate-based.

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.

Anders Eklund, Hans Knutsson, and Thomas E Nichols. Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates. *Human brain mapping*, 40(7):2017–2032, 2019.

Michael Esterman, Benjamin J 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.

John P Ferguson, Judy H Cho, Can Yang, and Hongyu Zhao. Empirical bayes correction for the winner's curse in genetic association studies. *Genetic epidemiology*, 37 (1):60–68, 2013.

Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press, 1925.

Steven D. Forman, Jonathan D. Cohen, Mark Fitzgerald, William F. Eddy, Mark A. Mintun, and Douglas C. Noll. Improved Assessment of Significant Activation in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster-Size Threshold. *Magnetic Resonance in Medicine*, 33(5):636–647, 1995. ISSN 15222594. doi: 10.1002/mrm.1910330508.

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.

Virgile Fritsch, Benoit Da Mota, Eva Loth, Gaël Varoquaux, Tobias Banaschewski, Gareth J Barker, Arun LW Bokde, Rüdiger Brühl, Brigitte Butzek, Patricia Conrod, et al. Robust regression for large-scale neuroimaging studies. *Neuroimage*, 111:431–441, 2015.

Christopher R. Genovese, Nicole A. Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4):870–878, 2002. ISSN 10538119. doi: 10.1006/nimg.2001.1037.

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*, 82(5):1064–1074, 2008.

Harald HH 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(6):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.

Allen Hatcher. *Algebraic topology*. Online Copy, 2001.

Satoru Hayasaka and Thomas E. Nichols. Validating cluster size inference: Random field and permutation methods. *NeuroImage*, 20(4):2343–2356, 2003. ISSN 10538119. doi: 10.1016/j.neuroimage.2003.08.003.

Satoru Hayasaka, K. Luan Phan, Israel Liberzon, K. J. Worsley, and Thomas E. Nichols. Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*, 22(2):676–687, 2004. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.01.041.

Satoru Hayasaka, Ann M Peiffer, Christina E Hugenschmidt, and Paul J Laurienti. Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage*, 37(3):721–730, 2007.

Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

Risto DH Heijmans and Jan R Magnus. Asymptotic normality of maximum likelihood estimators obtained from normally distributed but dependent observations. *Econometric Theory*, pages 374–412, 1986a.

Risto Donald Henri Heijmans and Jan Rudolf Magnus. On the first–order efficiency and asymptotic normality of maximum likelihood estimators obtained from dependent observations. *Statistica Neerlandica*, 40(3):169–188, 1986b.

Bruce Hoadley. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of mathematical statistics*, pages 1977–1991, 1971.

David Hogben, RS Pinkham, and MB Wilk. The moments of the non-central t-distribution. *Biometrika*, 48(3/4):465–468, 1961.

Andrew P Holmes. Statistical Issues in Function Brain Imaging. 1994.

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

Neal O Jeffries. Multiple comparisons distortions of parameter estimates. *Biostatistics*, 8(2):500–504, 2006.

Mark Jenkinson. Estimation of Smoothness from the Residual Field. *FMRIB Technical Report TR00MJ3*, 2000.

Mark Kac and David Slepian. Large excursions of gaussian processes. *The Annals of Mathematical Statistics*, 30(4):1215–1228, 1959.

S J Kiebel, J B Poline, K J Friston, Andrew P Holmes, and K. J. Worsley. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage*, 10(6):756–766, 1999. ISSN 1053-8119. doi: 10.1006/nimg.1999.0508.

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535, 2009.

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, 2010.

SN Lahiri. On bootstrapping m-estimators. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 157–170, 1992.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Gabriele Lohmann, Johannes Stelzer, Eric Lacosse, Vinod J Kumar, Karsten Mueller, Esther Kuehn, Wolfgang Grodd, and Klaus Scheffler. Lisa improves statistical analysis for fmri. *Nature communications*, 9(1):1–9, 2018.

Michael S Longuet-Higgins. On the statisticaldistribution of the height of sea waves. *JMR*, 11:245–266, 1952.

Michael Selwyn Longuet-Higgins. The statistical analysis of a random, moving surface. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 249(966):321–387, 1957.

Paul McCarthy. Fsleyes. Apr 2019. doi: 10.5281/zenodo.2630502.

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 LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523, 2016.

Jeanette A Mumford. A power calculation guide for fmri studies. *Social cognitive and affective neuroscience*, 7(6):738–742, 2012.

Jeanette A Mumford and Thomas Nichols. Simple group fmri modeling and inference. *Neuroimage*, 47(4):1469–1475, 2009.

Jeanette A. Mumford and Thomas E. Nichols. Modeling and inference of multisubject fMRI data. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):42–51, 2006. ISSN 07395175. doi: 10.1109/MEMB.2006.1607668.

Jerzy Neyman and Egon Sharpe Pearson. Contributions to the theory of testing statistical hypotheses. *Statistical Research Memoirs*, 1936.

Thomas E. Nichols and Satoru Hayasaka. Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, 12(5):419–446, 2003. ISSN 0962-2802. doi: 10.1191/0962280203sm341ra.

Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002a.

Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002b.

Lennart Nordberg. Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models. *Scandinavian Journal of Statistics*, pages 27–32, 1980.

V. P. Nosko. Local Structure of Gaussian Random Fields in the Vicinity of High-Level Light Sources, 1969.

VP Nosko. Asymptotic distributions of characteristics of high-level overshoots of a homogeneous gaussian random field. *Theory of Probability & Its Applications*, 32(4): 659–669, 1988.

PB Patnaik. The non-central $\chi^2$ and F-distribution and their applications. *Biometrika*, 36(1/2):202–232, 1949.

Andreas N Philippou and George G Roussas. Asymptotic normality of the maximum likelihood estimate in the independent not identically distributed case. *Annals of the Institute of Statistical Mathematics*, 27(1):45–55, 1975.

Andrew J Quinn, Freek van Ede, Matthew J Brookes, Simone G Heideman, Magdalena Nowak, Zelekha A Seedat, Diego Vidaurre, Catharina Zich, Anna C Nobre, and Mark W Woolrich. Unpacking transient event dynamics in electrophysiological power Spectra. *Brain topography*, pages 1–15, 2019.

J Radua, D Mataix-Cols, Mary L Phillips, W El-Hage, DM Kronhaus, N Cardoner, and S Surguladze. A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European psychiatry*, 27 (8):605–611, 2012.

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.

Alexis Roche, Sébastien Mériaux, Merlin Keller, and Bertrand Thirion. Mixed-effect statistics for group analysis in fmri: a nonparametric maximum likelihood approach. *Neuroimage*, 38(3):501–510, 2007.

Jonathan D Rosenblatt and Yoav Benjamini. Selective correlations; not voodoo. *Neuroimage*, 103:401–410, 2014.

Jonathan D Rosenblatt, Livio Finos, Wouter D Weeda, Aldo Solari, and Jelle J Goeman. All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796, 2018.

Gholamreza Salimi-Khorshidi, Stephen M. Smith, John R. Keltner, Tor D. Wager, and Thomas E. Nichols. Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, 45(3):810–823, 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2008.12.039. URL `http://dx.doi.org/10.1016/j.neuroimage.2008.12.039`.

Armin Schwartzman and Fabian Telschow. Peak p-values and false discovery rate inference in neuroimaging. *NeuroImage*, 197:402–413, 2019.

Armin Schwartzman, Yulia Gavrilov, and Robert J Adler. Multiple testing of local maxima for detection of peaks in 1D. *Annals of statistics*, 39(6):3290, 2011.

Xiaoxia Shi. Lecture notes: Asymptotic normality of extremum estimators. 2011.

Noah Simon and Richard Simon. On estimating many means, selection bias, and the bootstrap. *arXiv preprint arXiv:1311.3709*, 2013.

Scott D. Slotnick. Resting-state fMRI data reflects default network activity rather than null data: A defense of commonly employed methods to correct for multiple comparisons. *Cognitive Neuroscience*, 8(3):141–143, 2017. ISSN 17588936. doi: 10.1080/17588928.2016.1273892.

Otis M Solomon Jr. PSD computations using welch's method.[power spectral density (psd)]. Technical report, Sandia National Labs., Albuquerque, NM (United States), 1991.

Max Sommerfeld, Stephan Sain, and Armin Schwartzman. Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. *Journal of the American Statistical Association*, 1459:0–0, 2018.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford United States, 1956.

Lei Sun and Shelley B Bull. Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 28(4):352–367, 2005.

Trevor J Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, pages 1375–1381, 1980.

Kean Ming Tan, Noah Simon, and Daniela Witten. Selection bias correction and effect size estimation under dependence. *arXiv preprint arXiv:1405.4251*, 2014.

Jonathan Taylor. 2 Random Fields. *Mechanics of Materials*, 16:55–64, 2005. doi: 10.1016/0167-6636(93)90027-O.

Jonathan Taylor. A Gaussian kinematic formula. *Annals of Probability*, 34(1):122–158, 2006. ISSN 00911798. doi: 10.1214/009117905000000594.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

Jonathan E Taylor and K. J. Worsley. Detecting Sparse Signals in Random Fields, With an Application to Brain Mapping. *Journal of the American Statistical Association*, 102(479):913–928, 2007a. ISSN 0162-1459. doi: 10.1198/016214507000000815.

Jonathan E Taylor and Keith J Worsley. Detecting sparse signals in random fields, with an application to brain mapping. *Journal of the American Statistical Association*, 102(479):913–928, 2007b.

Fabian Telschow and Armin Schwartzmann. On Simultaneous Confidence Statements and Inference for Functional Data Using the Gaussian Kinematic Formula. (April): 1–27, 2020.

Fabian Telschow, Armin Schwartzman, Dan Cheng, and Pratyush Pranav. Estimation of Expected Euler Characteristic Curves of Nonstationary Smooth Gaussian Random Fields. *arXiv e-prints*, art. arXiv:1908.02493, August 2019.

Fabian Telschow, Samuel Davenport, and Armin Schwartzman. Functional delta residuals. 2020a.

Fabian Telschow, Samuel Davenport, and Armin Schwartzman. From discrete to continuous land: Using the continuous gaussian kinematic formula on a discrete lattice. *Preprint*, 2020b.

A.W. van der Vaart. *Asymptotic Statistics*. 1998.

BL Van der Waerden. Order tests for the two-sample problem and their power. In *Indagationes Mathematicae (Proceedings)*, volume 55, pages 453–458. Elsevier, 1952.

Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on psychological science*, 4(3):274–290, 2009.

Tor D Wager, Matthew C Keller, Steven C Lacey, and John Jonides. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage*, 26(1):99–113, 2005.

Abraham Wald. Asymptotic properties of the maximum likelihood estimate of an unknown parameter of a discrete stochastic process. *The Annals of Mathematical Statistics*, pages 40–46, 1948.

Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

Jon A Wellner and Yihui Zhan. Bootstrapping z-estimators. 1996.

Peter H Westfall and S Stanley Young. Resampling-based multiple testing. Examples and methods for p-value adjustment., 1993. URL `papers2://publication/uuid/21ABA9C5-67BB-4C6D-BE4C-9EA8CCA16E9D`.

Ian R White and John B Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29 (28):2920–2931, 2010.

Richard J. Wilson. Model fields in crossing theory: A weak convergence perspective. *Advances in Applied Probability*, 20:756–774, 1988.

Richard J. Wilson and Robert J. Adler. The structure of gaussian fields near a level crossing. *Advances in Applied Probability*, 14:543–565, 1982.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.01.060.

Anderson M. Winkler, Gerard R. Ridgway, Gwenaëlle Douaud, Thomas E. Nichols, and Stephen M. Smith. Faster permutation inference in brain imaging. *NeuroImage*, 141:502–516, 2016. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.05.068.

Mark Woolrich. Robust group analysis using outlier inference. *Neuroimage*, 41(2): 286–301, 2008.

Mark W. Woolrich, Brian D. Ripley, Michael Brady, and Stephen M. Smith. Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. *NeuroImage*, 14(6): 1370–1386, 2001. ISSN 10538119. doi: 10.1006/nimg.2001.0931.

K. J. Worsley. Local Maxima and the Expected Euler Characteristic of Excursion Sets of $\chi$ 2, F and t Fields. *Advances in Applied Probability*, 26(1):13–42, 1994.

K. J. Worsley. An improved theoretical P value for SPMs based on discrete local maxima. *NeuroImage*, 28(4):1056–1062, 2005. ISSN 10538119. doi: 10.1016/j.neuroimage.2005.06.053.

K. J. Worsley, A C Evans, S Marrett, and P Neelin. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of cerebral blood flow and metabolism.*, 12(6):900–18, 1992. ISSN 0271-678X. doi: 10.1038/jcbfm.1992.127.

K. J. Worsley, M. Andermann, T. Koulis, D. MacDonald, and A. C. Evans. Detecting changes in nonisotropic images. *Human Brain Mapping*, 8(2-3):98–101, 1999. ISSN 10659471. doi: 10.1002/(SICI)1097-0193(1999)8:2/3⟨98::AID-HBM5⟩3.0.CO;2-F.

Keith J. Worsley. The geometry of random images. *Chance*, 9(1):27–40, 1996a.

Keith J. Worsley. An unbiased estimator for the roughness of a multivariate gaussian random field. 1996b.

Keith J. Worsley, Sean Marrett, Peter Neelin, Alain C Vandal, Karl J Friston, and Alan C Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, 4(1):58–73, 1996.

Long Yang Wu, Lei Sun, and Shelley B Bull. Locus-specific heritability estimation via the bootstrap in linkage scans for quantitative trait loci. *Human Heredity*, 62(2): 84–96, 2006.

Rui Xiao and Michael Boehnke. Quantifying and correcting for the winner's curse in quantitative-trait association studies. *Genetic epidemiology*, 35(3):133–138, 2011.

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(3):540–551, 2007.

Hua Zhong and Ross L Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9(4):621–634, 2008.

Yi-Hui Zhou and Fred A Wright. The projack: a resampling approach to correct for ranking bias in high-throughput studies. *Biostatistics*, 17(1):54–64, 2015.