Multiple comparison correction

Klaas Enno Stephan

Translational Neuromodeling Unit

Universität Zürich UZH

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Methods & models for fMRI data analysis
22 October 2019

With many thanks for slides & images to: FIL Methods group, Tom Nichols

Overview of SPM

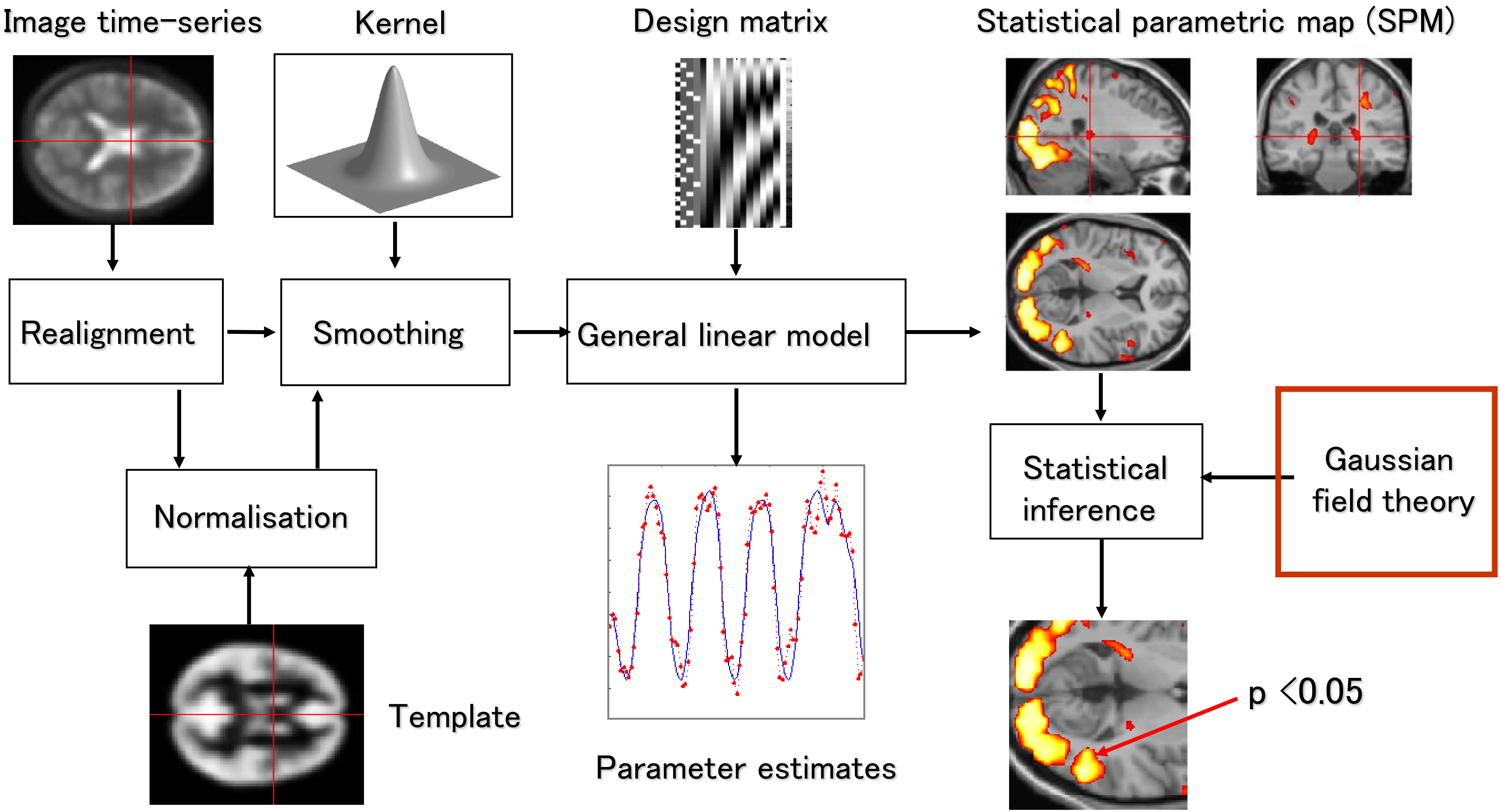


Image time-series

Kernel

Design matrix

Statistical parametric map (SPM)

Realignment

Smoothing

General linear model

Normalisation

Template

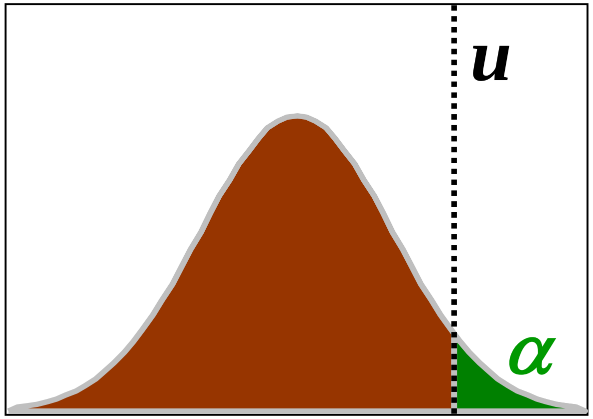
Parameter estimates

Statistical inference

Gaussian field theory

$p < 0.05$

Inference at a single voxel



t distribution

NULL hypothesis
H_0: activation is zero

$\alpha = p(T > u \mid H_0)$

We can choose u to set a voxel-wise significance level of α.

p–value: probability of getting a value of the test statistic t , or a more extreme value, under the null hypothesis.

If the p–value is smaller than u, we reject the null hypothesis.

$$t = \frac{\textbf{\textit{contrast} of estimated parameters}}{\sqrt{\textbf{variance estimate}}}$$

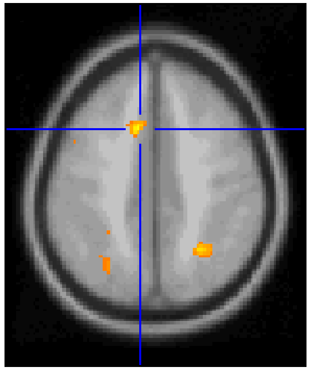
$$\boxed{t = \frac{c^T\hat{\beta}}{s\hat{t}d(c^T\hat{\beta})} = \frac{c^T\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T\left(X^T X\right)^{-1} c}} \sim t_{N-p}}$$

Types of error

		Actual condition	
		H_0 true	H_0 false
Test result	Reject H_0	**False positive (FP)** **Type I error** α	**True positive (TP)**
	Failure to reject H_0	**True negative (TN)**	**False negative (FN)** **Type II error** β
		specificity: 1-α = TN / (TN + FP) = proportion of actual negatives which are correctly identified	**sensitivity (power): 1-β** = TP / (TP + FN) = proportion of actual positives which are correctly identified

Assessing SPMs

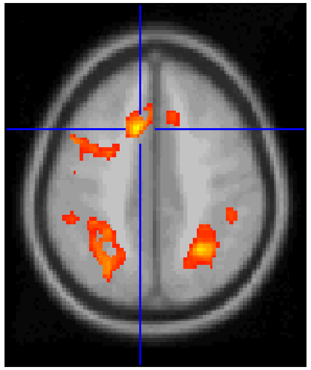
High Threshold



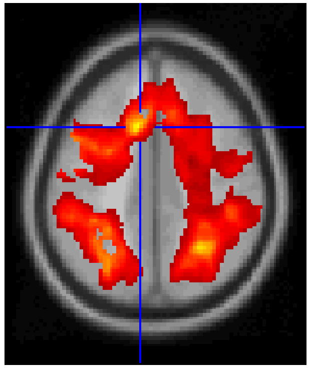
Good Specificity

Poor Power
(risk of false negatives)

Med. Threshold



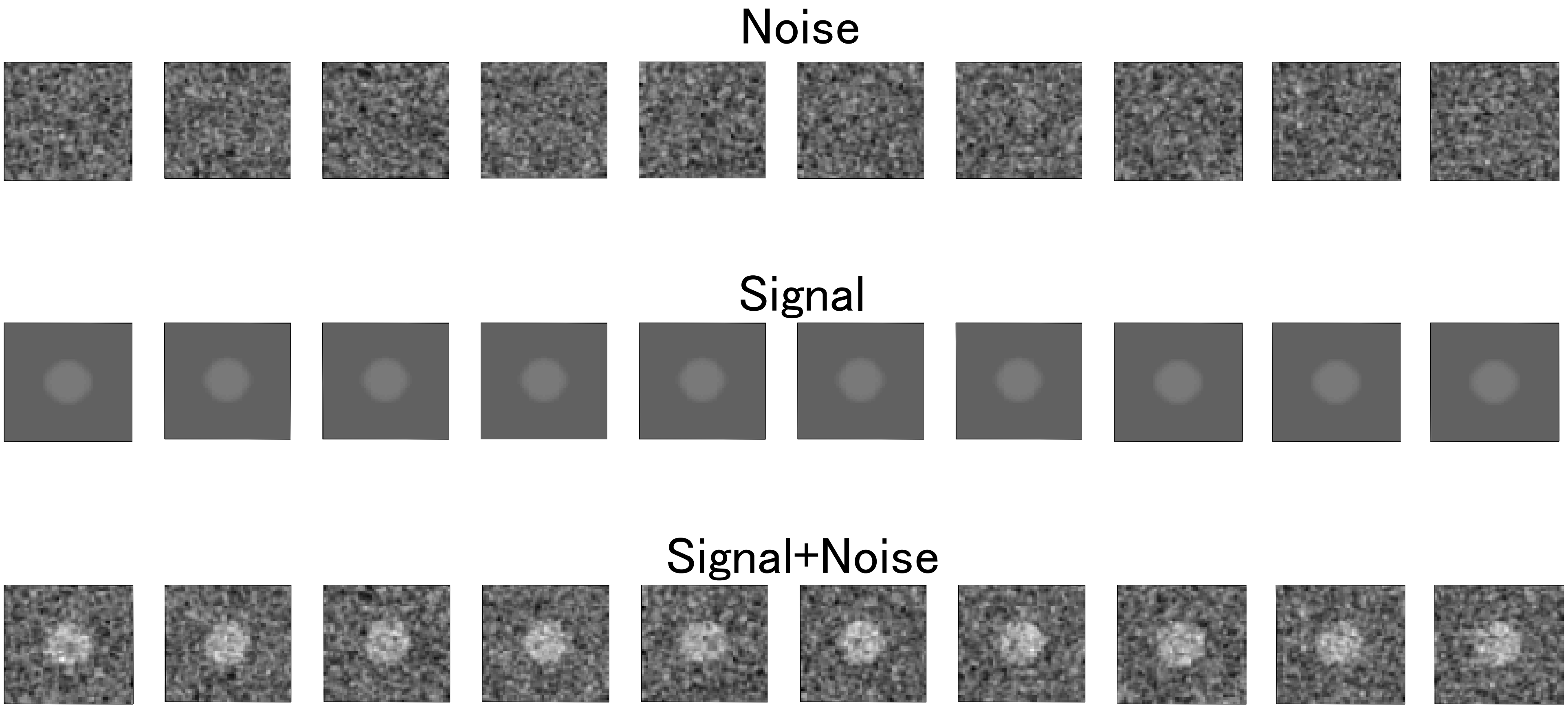
Low Threshold



Poor Specificity
(risk of false positives)

Good Power

Inference on images

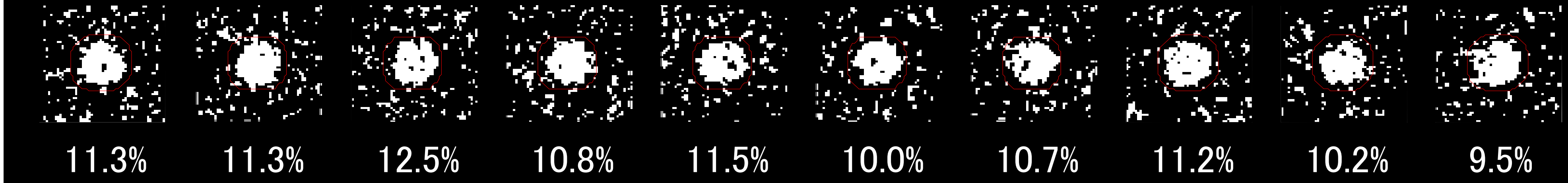


Noise

Signal

Signal+Noise

Use of ‘uncorrected’ p-value, α=0.1



Percentage of Null Pixels that are False Positives

Using an ‘uncorrected’ p-value of 0.1 will lead us to conclude on average that 10% of voxels are active when they are not.

This is clearly undesirable. To correct for this we can define a null hypothesis for images of statistics.

Family-wise null hypothesis

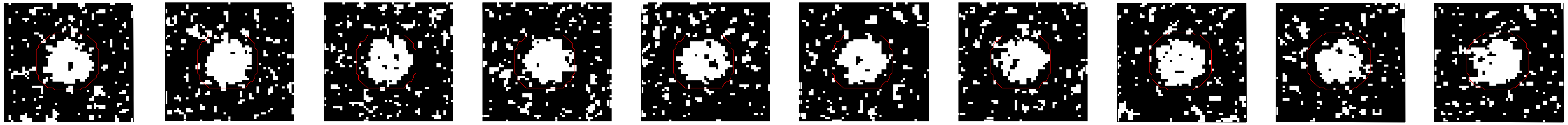
FAMILY-WISE NULL HYPOTHESIS:
Activation is zero **everywhere**.

If we reject a voxel null hypothesis at **any** voxel, we reject the family-wise null hypothesis

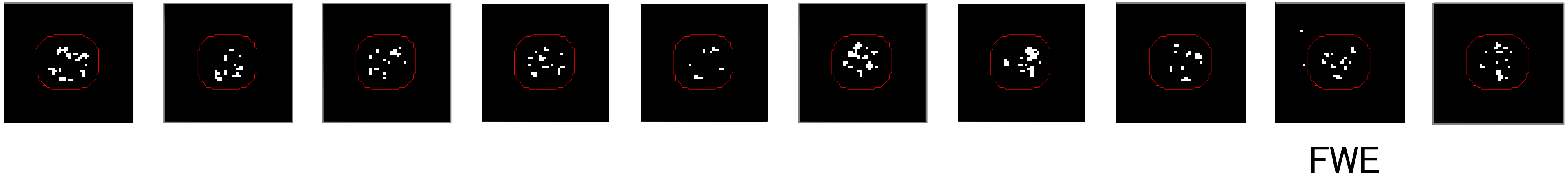
A false-positive **anywhere** in the image gives a Family Wise Error (FWE).

Family-Wise Error (FWE) rate = ‘corrected’ p-value

Use of ‘uncorrected’ p-value, α=0.1



Use of ‘corrected’ p-value, α=0.1



The Bonferroni correction

The family-wise error rate (FWE), α, for a family of N **independent** voxels is

$$\alpha = Nv$$

where v is the voxel-wise error rate.

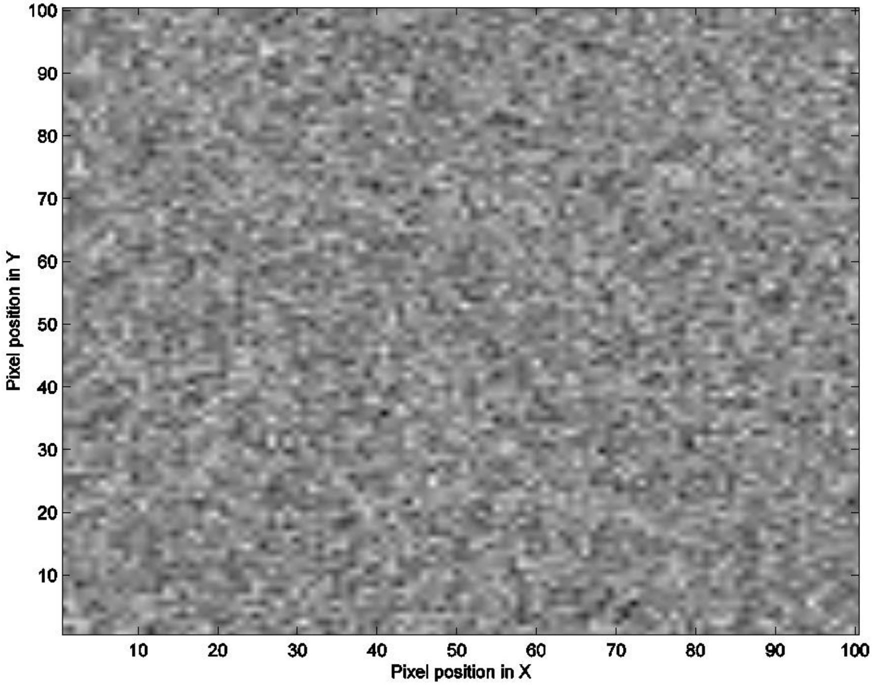
Therefore, to ensure a particular FWE, we can use

$$v = \alpha / N$$

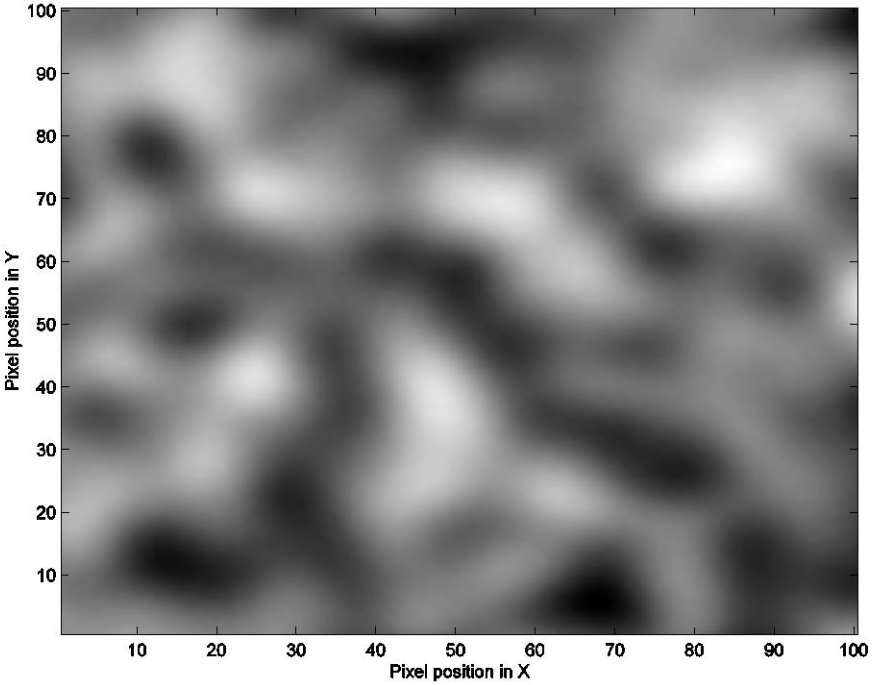
BUT …

The Bonferroni correction

Independent voxels



Spatially correlated voxels



Bonferroni correction assumes independence of voxels

→ this is too conservative for brain images,
which always have a degree of smoothness

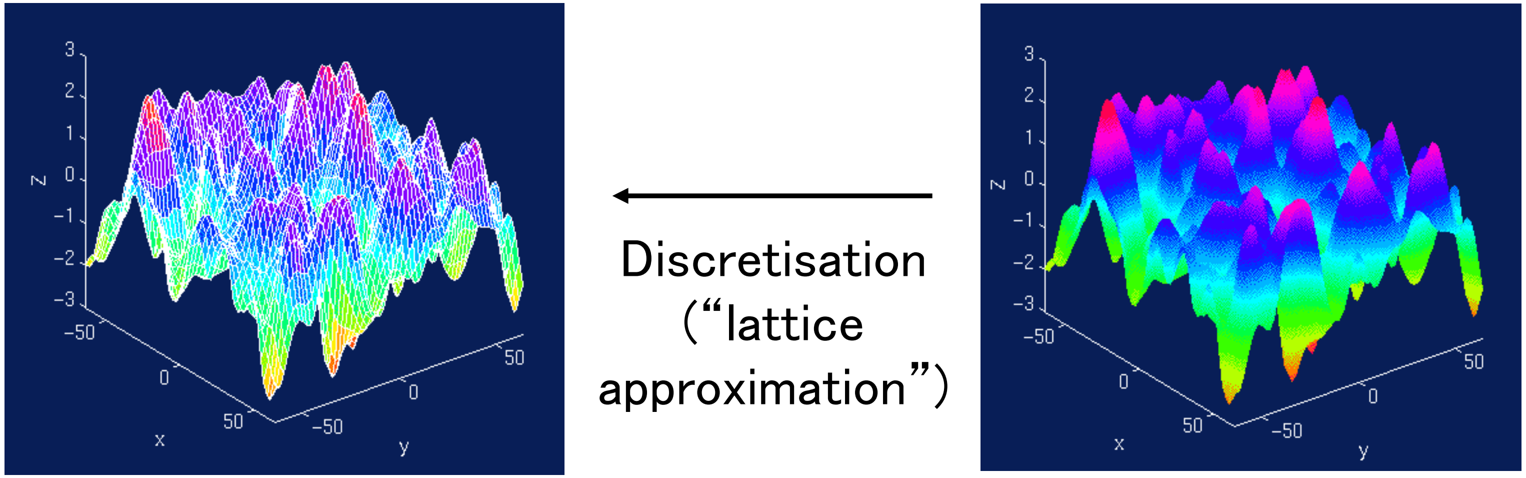
Smoothness (inverse roughness)

- roughness = 1/smoothness
- intrinsic smoothness
  - MRI signals are aquired in k-space (Fourier space); after projection on anatomical space, signals have continuous support
  - diffusion of vasodilatory molecules has extended spatial support
- extrinsic smoothness
  - resampling during preprocessing
  - matched filter theorem
    → deliberate additional smoothing to increase SNR
- described in resolution elements: "resels"
- resel = size of image part that corresponds to the FWHM (full width half maximum) of the Gaussian convolution kernel that would have produced the observed image if it had been applied to independent voxel values
- # resels is similar, but not identical to # independent observations
- can be computed from spatial derivatives of the residuals

Random Field Theory

- Consider a statistic image as a discretisation of a continuous underlying random field with a certain smoothness

- Use results from continuous random field theory



Discretisation (“lattice approximation”)

Euler characteristic (EC)

Topological measure

threshold an image at *u*

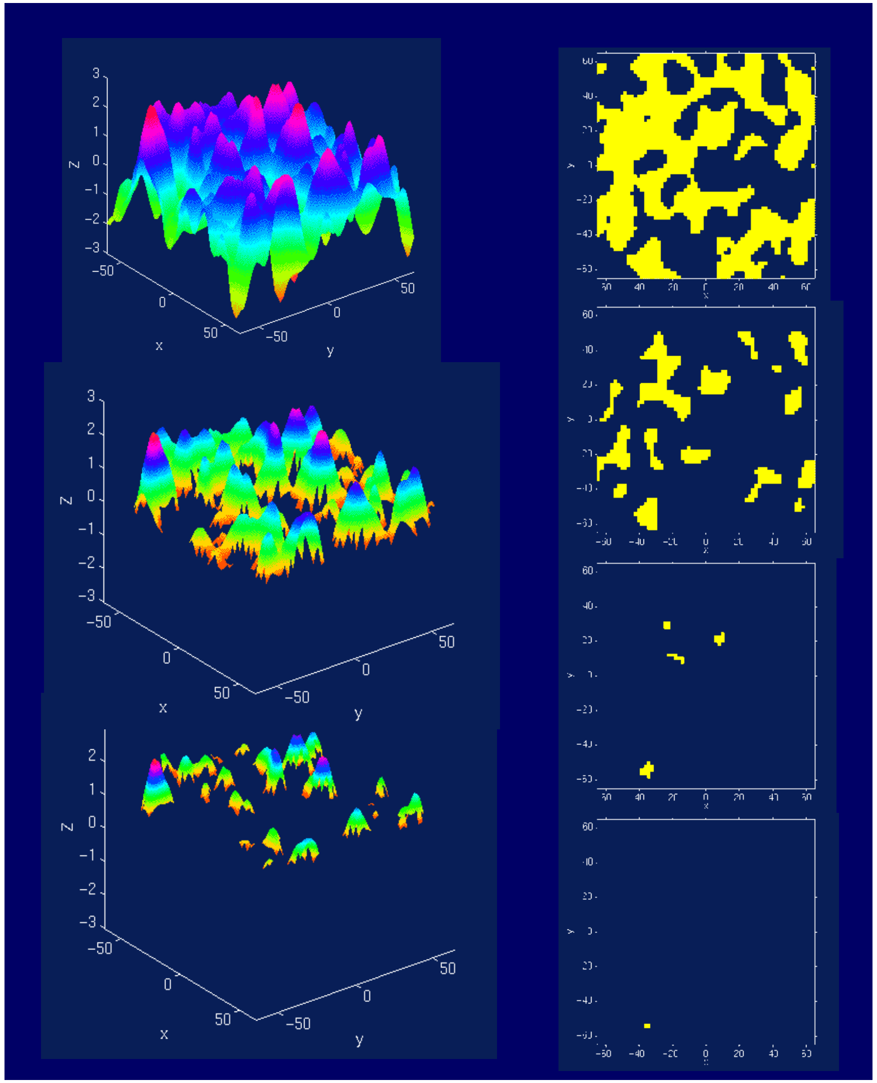
→ EC ∝ # blobs

At high u:

p (blob) = E [EC],

therefore (under H_0):

FWE rate: α = E [EC]



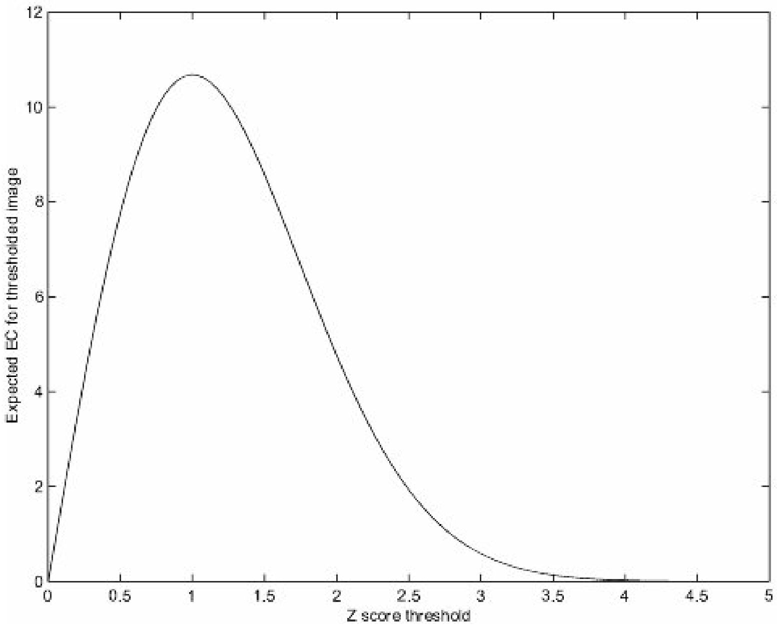
Euler characteristic (EC) for 2D images

$$\mathrm{E}\left[\mathrm{EC}\right] = R(4\log 2)(2\pi)^{-3/2} Z_T \exp(-0.5Z_T^2)$$

R = number of resels
Z_T = Z value threshold

We can determine that Z threshold for which E[EC] = 0.05. At this threshold, every remaining peak represents a significant activation, corrected for multiple comparisons across the search volume.

Example: For 100 resels, E [EC] = 0.049 for a Z threshold of 3.8. That is, the probability of getting one or more blobs where Z is greater than 3.8, is 0.049.



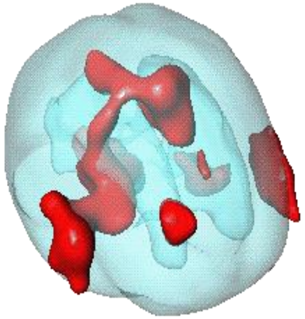
Expected EC values for an image of 100 resels

Euler characteristic (EC) for any image

- Computation of E[EC] can be generalized to volumes of any dimension, shape and size (Worsley et al. 1996).

- When we have an *a priori* hypothesis about where an activation should be, we can (and should) reduce the search volume:
  - mask defined by (probabilistic) anatomical atlases
  - mask defined by separate "functional localisers"
  - mask defined by orthogonal contrasts
  - (spherical) search volume around previously reported coordinates

⇒ small volume correction (SVC)



Worsley et al. 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Human Brain Mapping, 4, 58–83.

Computing EC wrt. search volume and threshold

$$E(\chi_u) \approx \lambda(\Omega)\ \ |\Lambda|^{1/2}\ (u^{\,2} - 1)\ \exp(-u^{\,2}/2)\ /\ (2\pi)^2$$

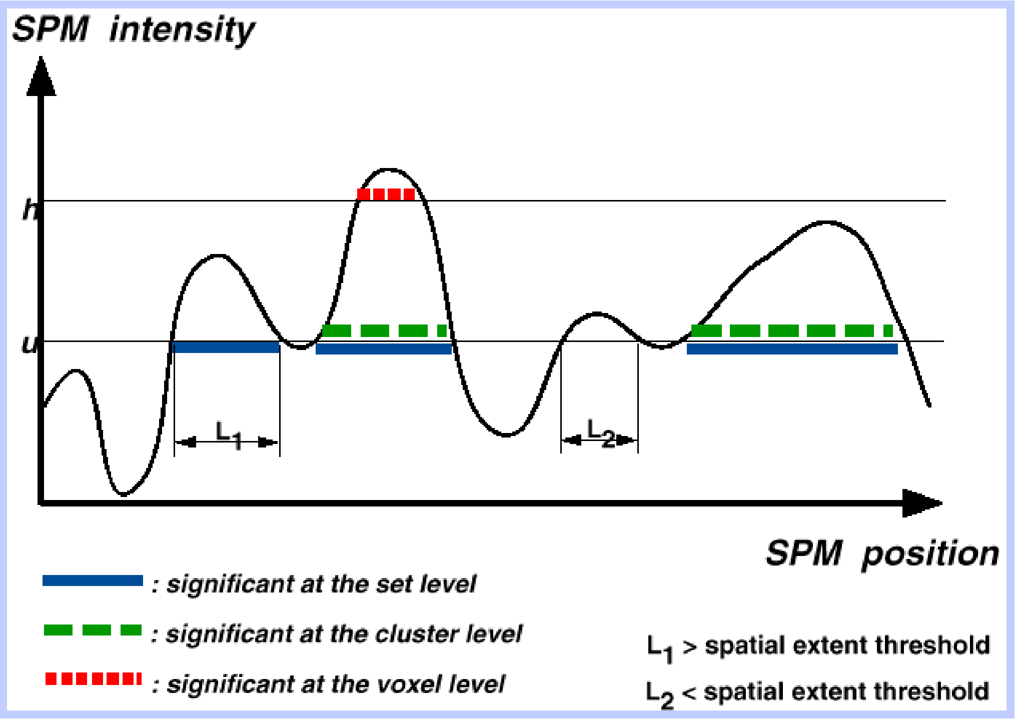
- Ω → Search region $\Omega \subset \mathcal{R}^3$
- $\lambda(\Omega)$ → volume
- $|\Lambda|^{1/2}$ → roughness

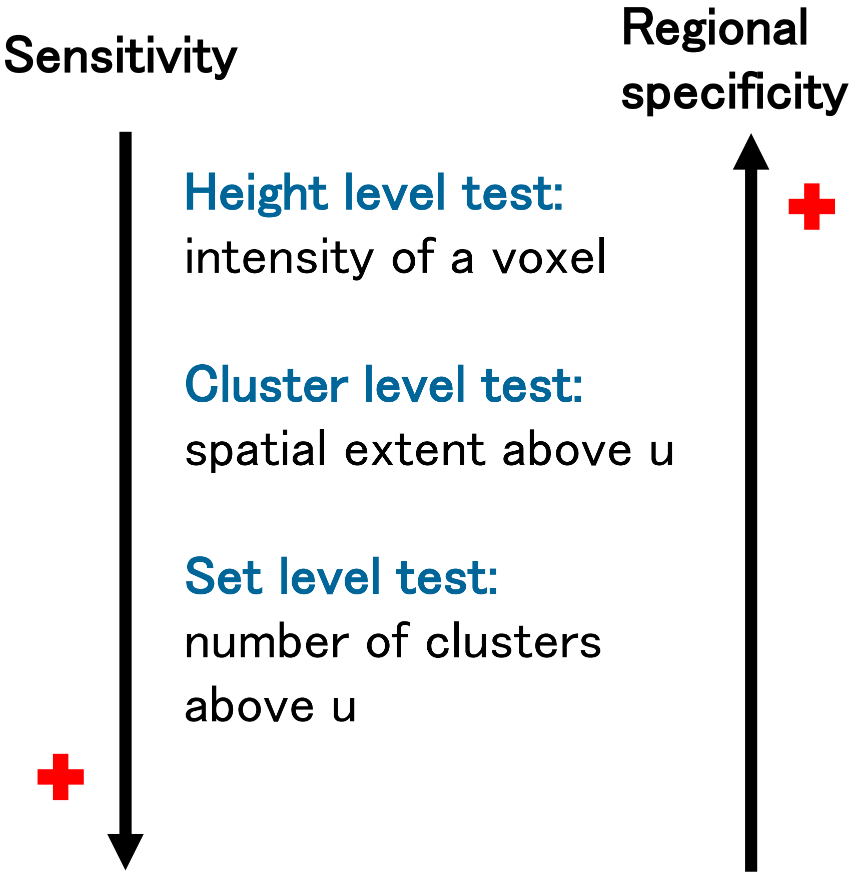
- Assumptions:
  - Multivariate normal
  - Stationary*
  - ACF twice differentiable at 0

* Stationarity

- Results valid w/out stationarity
- More accurate when stationarity holds

Height, cluster and set level tests


SPM intensity
h
u
L1
L2
SPM position
: significant at the set level
: significant at the cluster level
: significant at the voxel level
L1 > spatial extent threshold
L2 < spatial extent threshold



Height level test:
intensity of a voxel

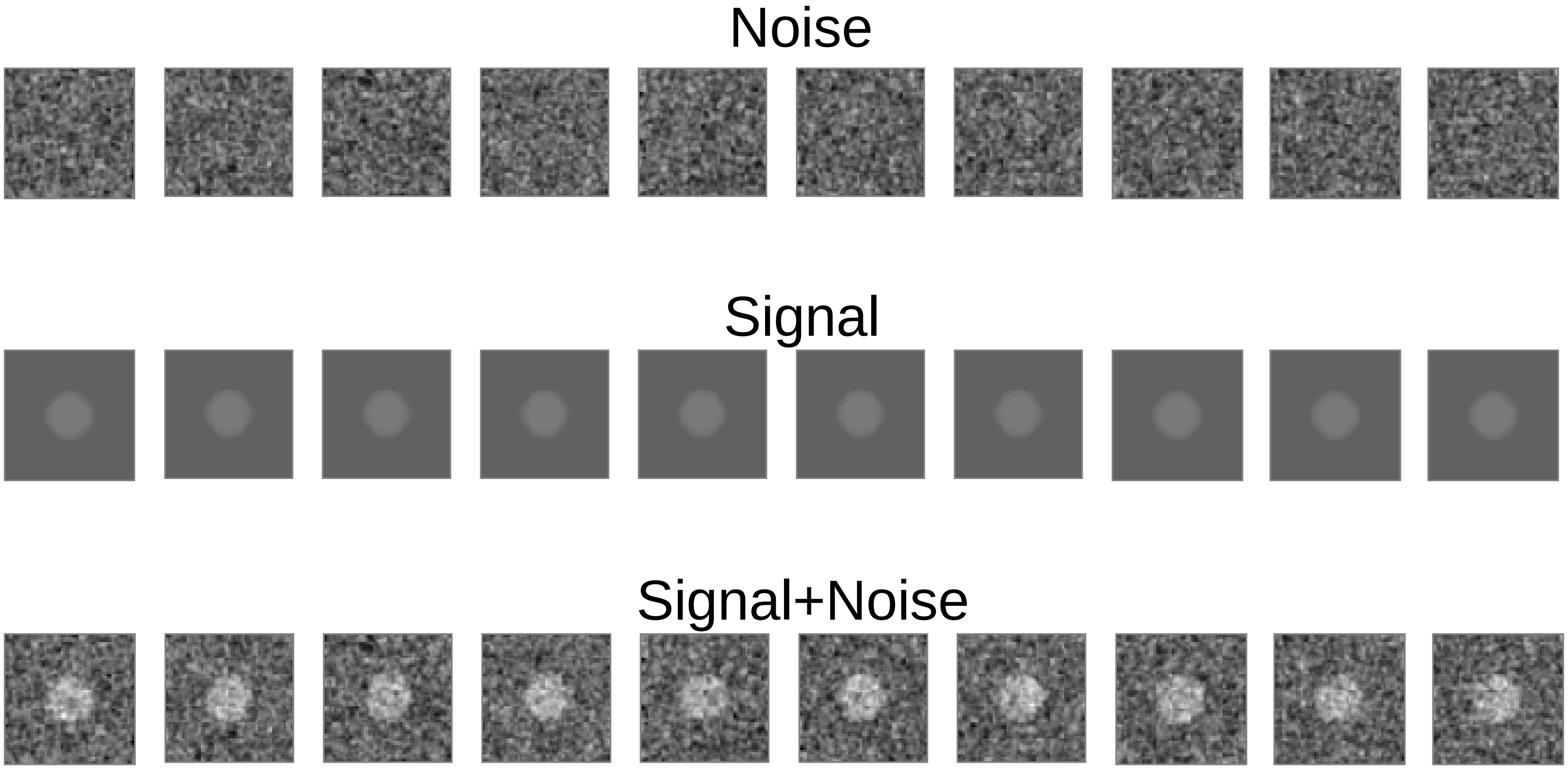
Cluster level test:
spatial extent above u

Set level test:
number of clusters above u

False Discovery Rate (FDR)

- Familywise Error Rate (FWE)
  - probability of one or more false positive voxels in the entire image
- False Discovery Rate (FDR)
  - FDR = E[V/R]
    (R voxels declared active, V falsely so)
  - FDR = proportion of activated voxels that are false positives

False Discovery Rate - Illustration

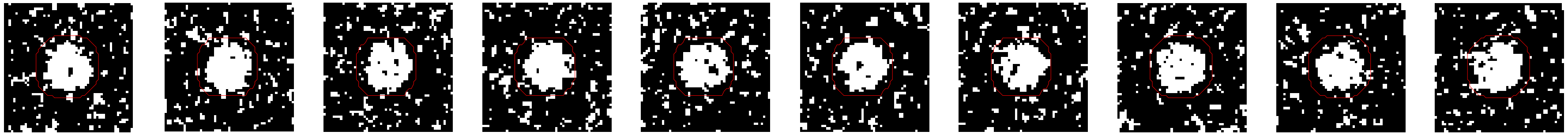


Noise

Signal

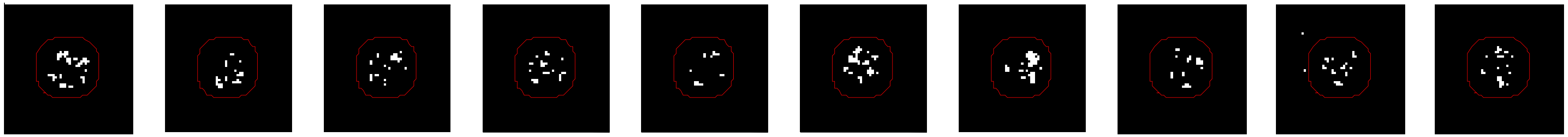
Signal+Noise

Control of Per Comparison Rate at 10%



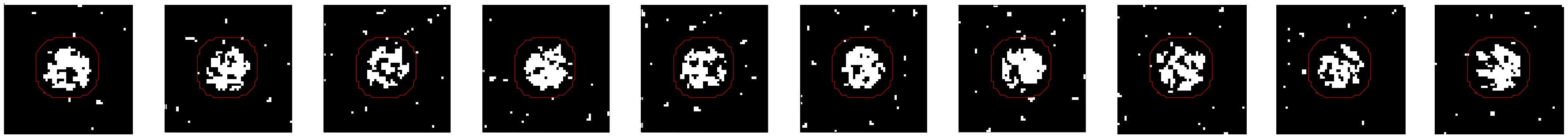
Percentage of False Positives

Control of Familywise Error Rate at 10%



Occurrence of Familywise Error

Control of False Discovery Rate at 10%



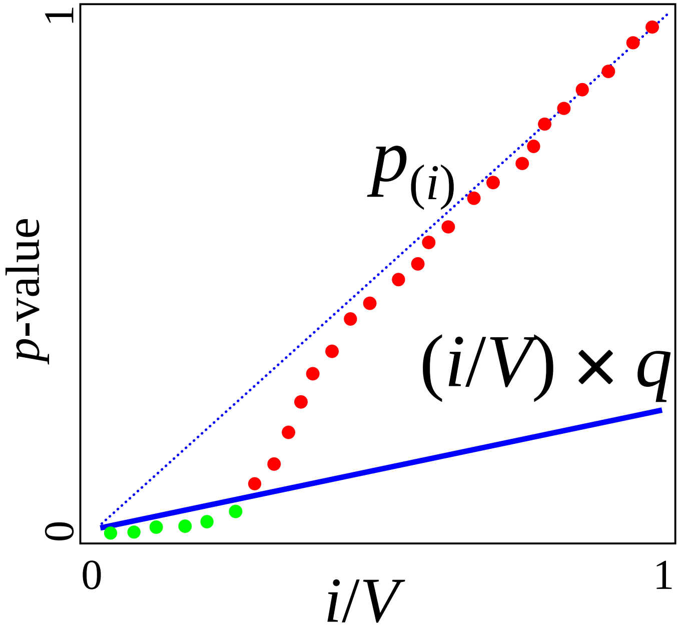
Percentage of Activated Voxels that are False Positives

Benjamini & Hochberg procedure

- Select desired limit q on FDR
- Order p-values, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(V)}$
- Let r be largest i such that

$$p_{(i)} \leq (i/V) \times q$$

- Reject all null hypotheses corresponding to $p_{(1)}, \ldots, p_{(r)}$.

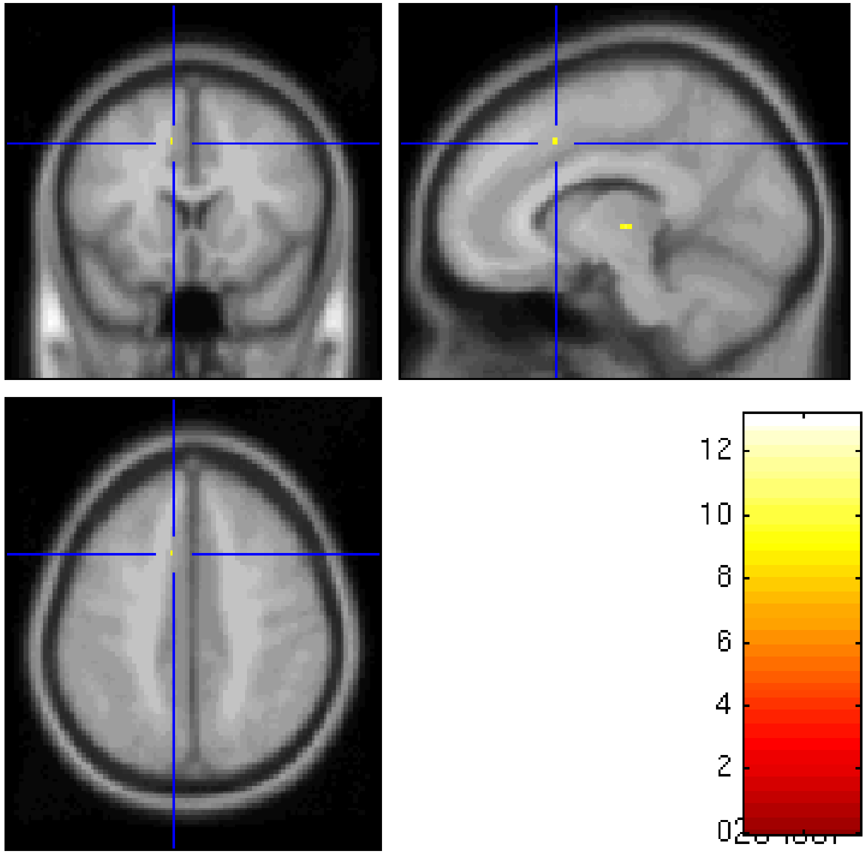


i/V = proportion of all selected voxels

Benjamini & Hochberg, *JRSS-B* (1995) 57:289-300

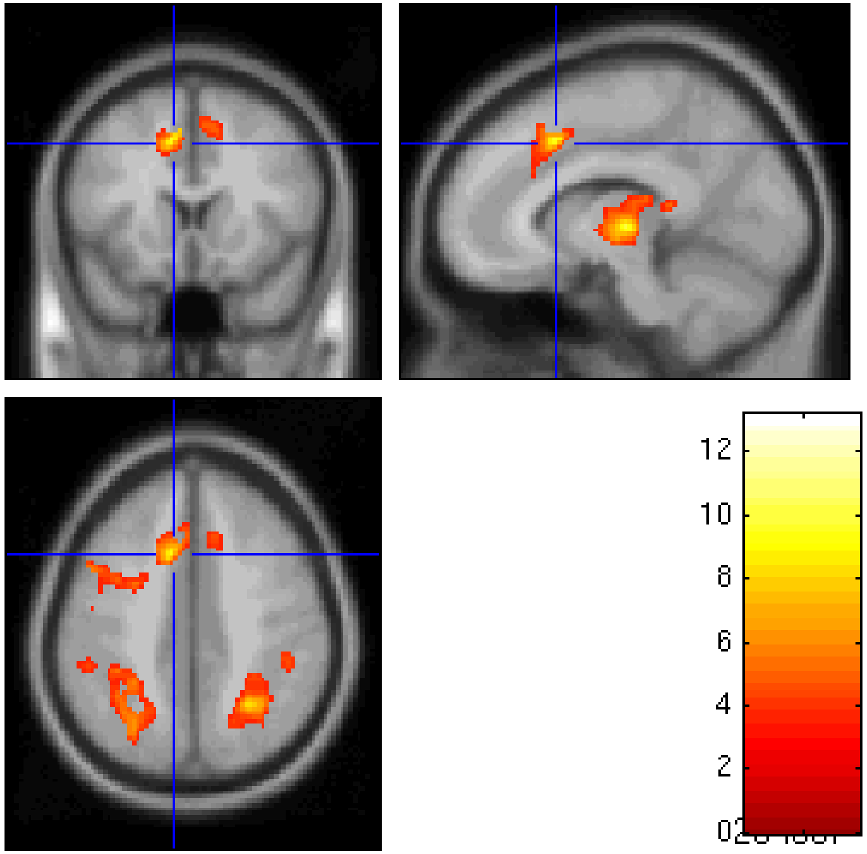
Real Data: FWE correction with RFT

- Threshold
  - S = 110,776
  - 2 × 2 × 2 voxels
5.1 × 5.8 × 6.9 mm
FWHM
  - u = 9.870
- Result
  - 5 voxels above the threshold



Real Data: FWE correction with FDR

- Threshold
  - $u = 3.83$
- Result
  - 3,073 voxels above threshold



Caveats concerning FDR

- questionable whether voxel-wise FDR implementations are suitable for neuroimaging data

- Chumbley & Friston 2009 argue that:
  - the fMRI signal is spatially extended, it does not have compact support
  - inference should therefore not be about single voxels, but about topological features of the signal (e.g. peaks or clusters)

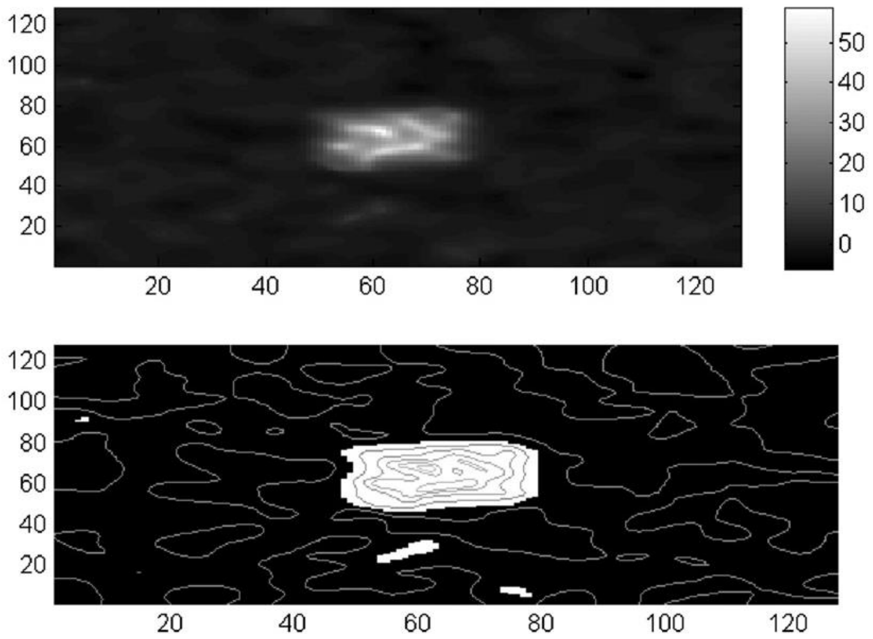
Chumbley & Friston 2009: example of FDR failure

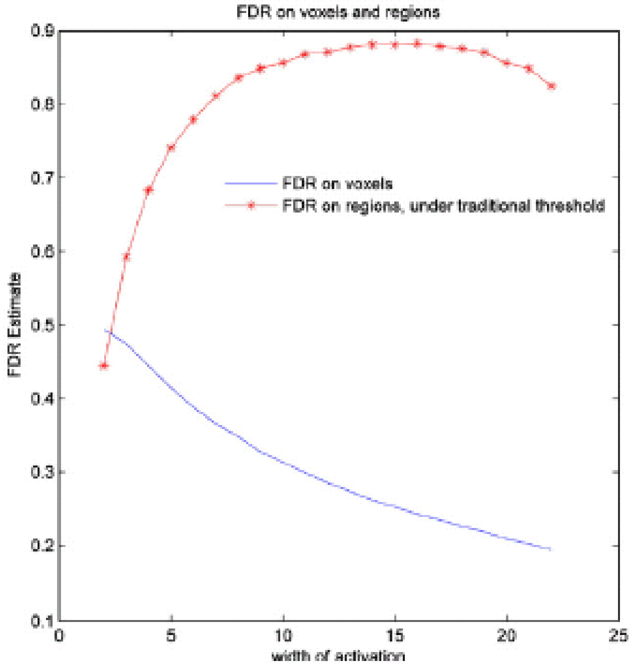
- “Imagine that we declare 100 voxels significant using an FDR criterion. 95 of these voxels constitute a single region that is truly active. The remaining five voxels are false discoveries and are dispersed randomly over the search space.

  In this example, the false discovery rate of voxels conforms to its expectation of 5%. However, the false discovery rate in terms of regional activations is over 80%. This is because we have discovered six activations but only one is a true activation.”

  (Chumbley & Friston 2009, NeuroImage)

Chumbley & Friston 2009: example of FDR failure





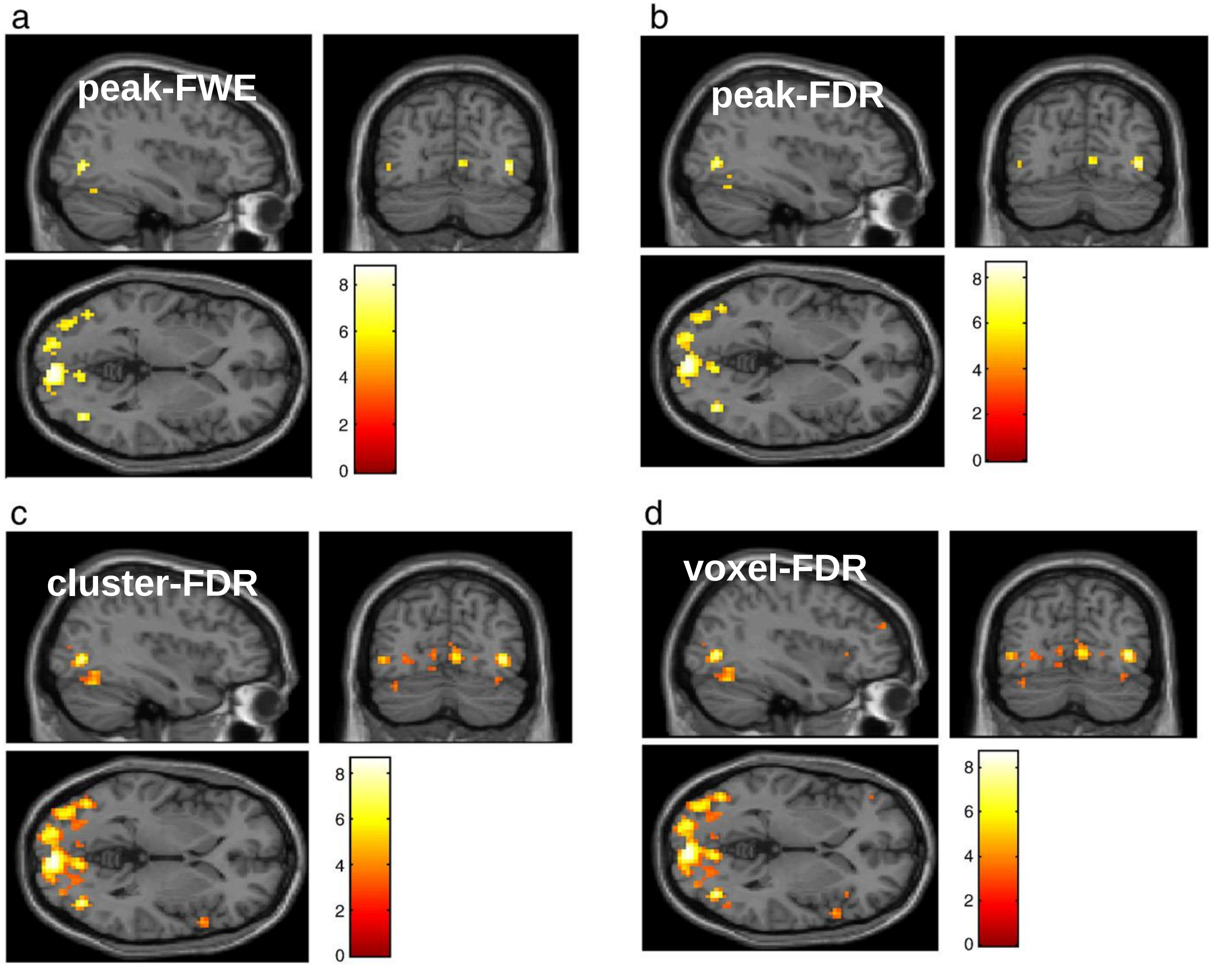
- simulated data with intrinsic smoothness: 8 images with true signal in centre and background noise
- one-sample t-test, FDR-threshold at voxel-level (q=0.05)
- result: both voxel- and cluster-wise FDR bigger than expected (due to smoothness)

Chumbley & Friston 2010: Topological FDR

- instead of p-values of individual voxels, apply FDR to p-values of topological features of the signal (peaks or clusters)

- simulations: peak-FDR is more sensitive than peak-FWE

- empirical analysis: number of sign. peaks increases monotonically: peak-FWE, peak-FDR, cluster-FDR, voxel-FDR



Conclusions

- Corrections for multiple testing are necessary to control the false positive risk.
- FWE
  - Very specific, not so sensitive
  - Random Field Theory
    - Inference about topological features (peaks, clusters)
    - Excellent for large sample sizes (e.g. single-subject analyses or large group analyses)
    - Afford littles power for group studies with small sample size → consider non-parametric methods (not discussed in this talk)
- FDR
  - Less specific, more sensitive
  - Interpret with care!
    - represents false positive risk over whole set of selected voxels
    - voxel-wise FDR may be problematic (ongoing discussion)
    - topological FDR now available in SPM

Further reading

- Chumbley JR, Friston KJ. False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage. 2009;44(1):62-70.
- Chumbley J, Worsley K, Flandin G, Friston K (2010) Topological FDR for neuroimaging. Neuroimage 49:3057-3064.
- Friston KJ, Frith CD, Liddle PF, Frackowiak RS. Comparing functional (PET) images: the assessment of significant change. J Cereb Blood Flow Metab. 1991 Jul;11(4):690-9.
- Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage. 2002 Apr;15(4):870-8.
- Worsley KJ Marrett S Neelin P Vandal AC Friston KJ Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. Human Brain Mapping 1996;4:58-73.

Thank you