Edinburgh Imaging
www.ed.ac.uk/edinburgh-imaging

LIMO EEG: advanced designs
Application using continuous variables

D2.A2 & D2.A3 – 3.45 to 5.00

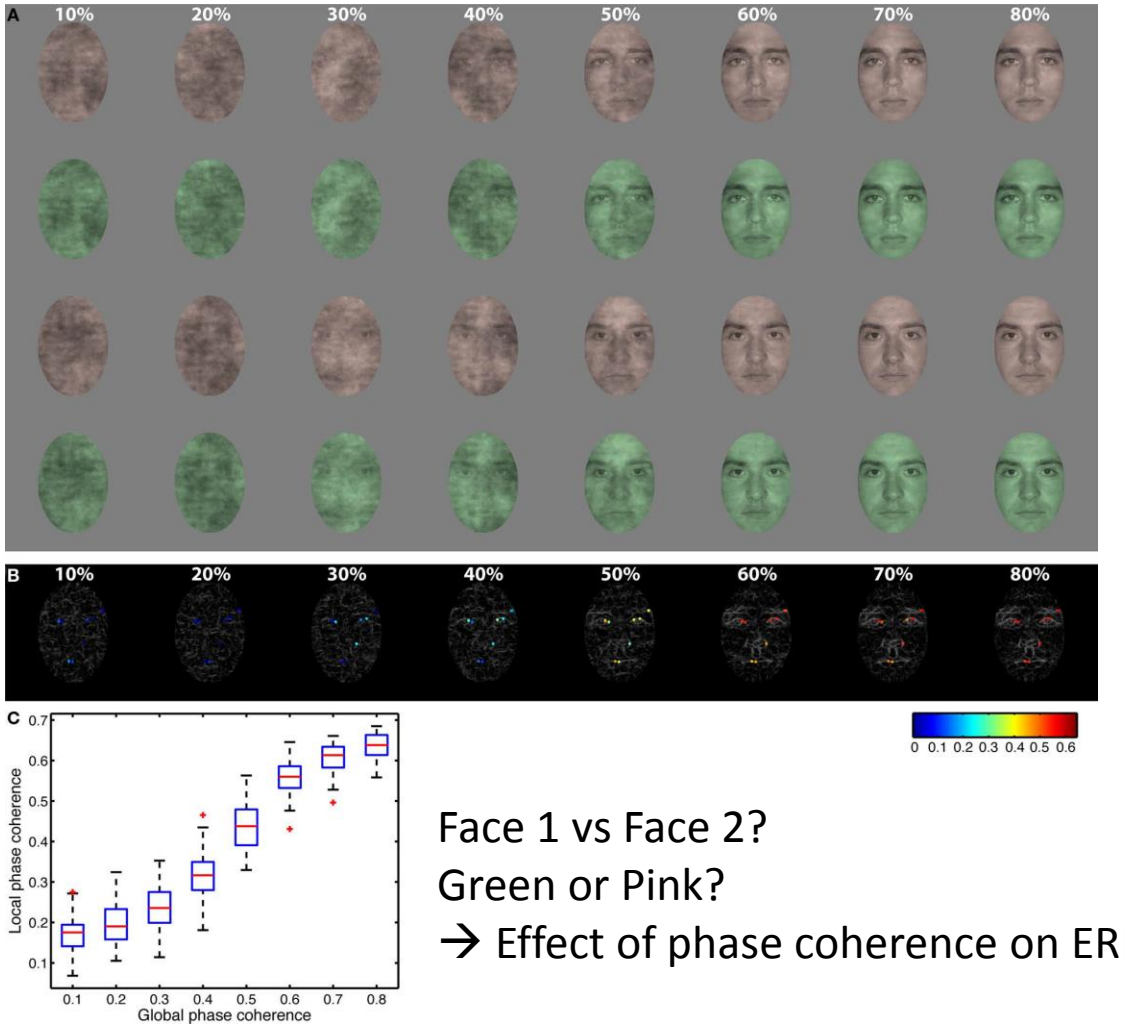
Cyril Pernet, PhD
Centre for Clinical Brain Sciences, The university of Edinburgh, UK

Single subjects or group analysis

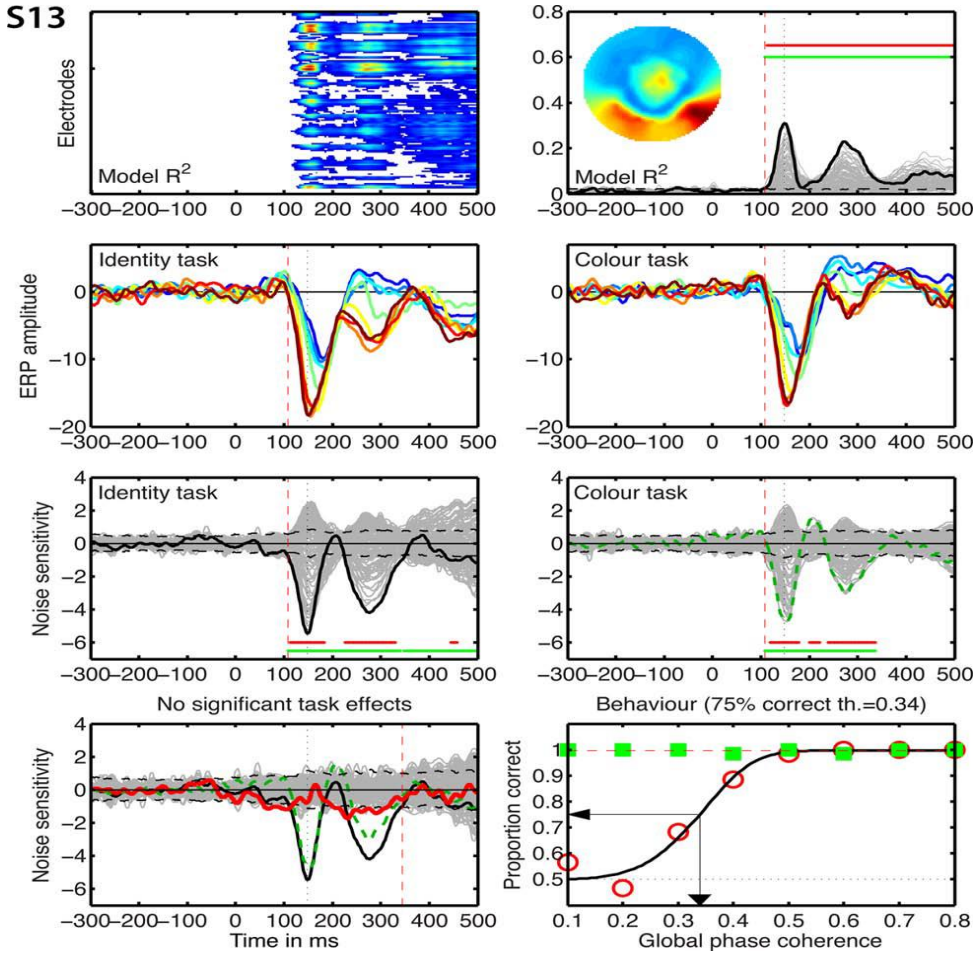
What is the question?

- If the question pertains to dynamic analyses (when things happen) and/or quantitative aspects (how much this variable explains of the data), then single subjects analyses make more sense given the idiosyncratic nature of EEG.
- Yet some group stats are needed for inference – e.g. average cluster onset, average number of subject showing an effect, etc .. + derive group level effect sizes
- If the question is general in nature (is there a measurable difference between these conditions) or pertains to group differences and/or attributes, then group analyses makes sense.

How task constraints modulate the ERP response?

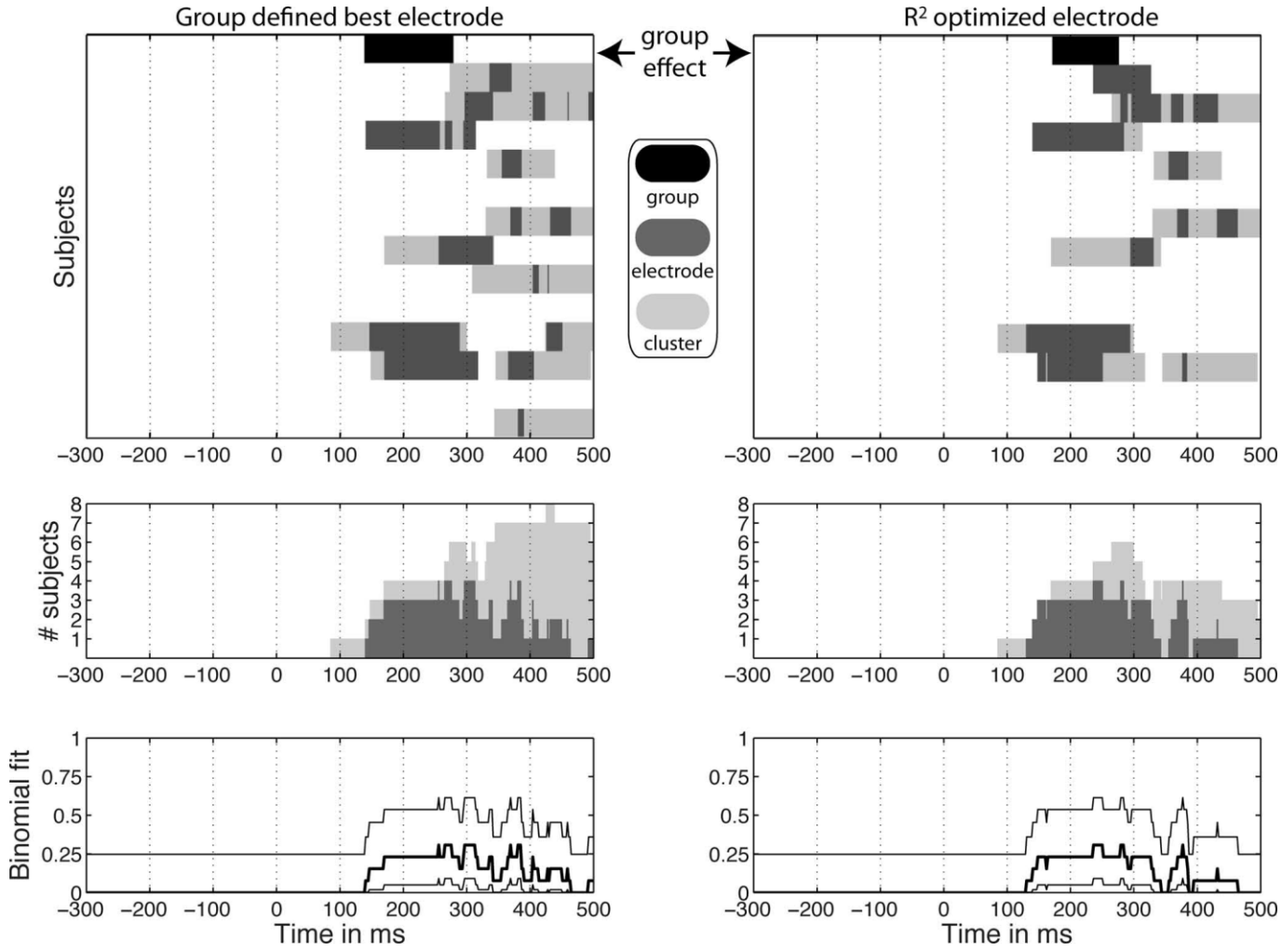


Face 1 vs Face 2?
Green or Pink?
→ Effect of phase coherence on ERP



Rousselet et al. Front Psy 2011

How task constraints modulate the ERP response?

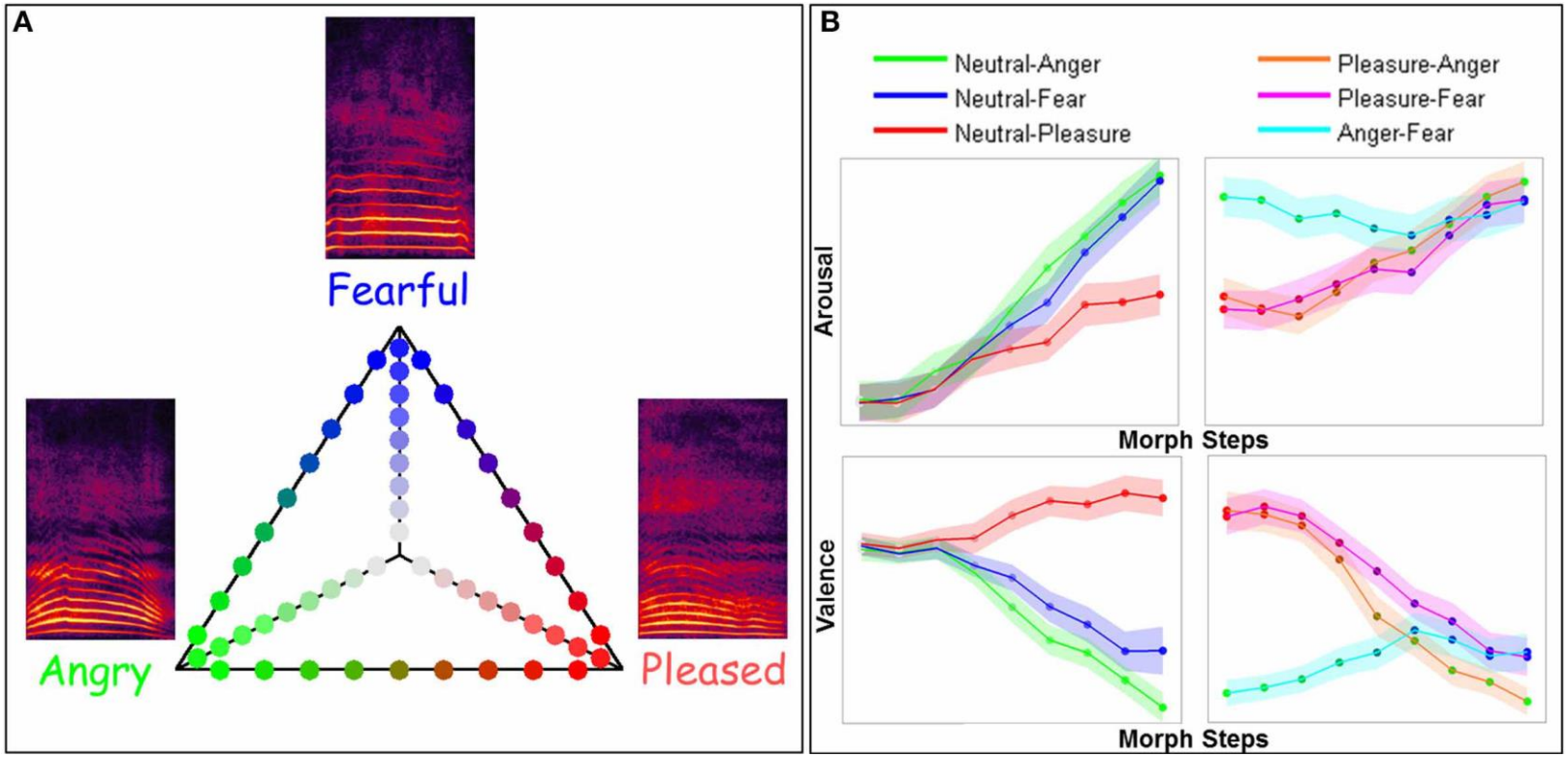


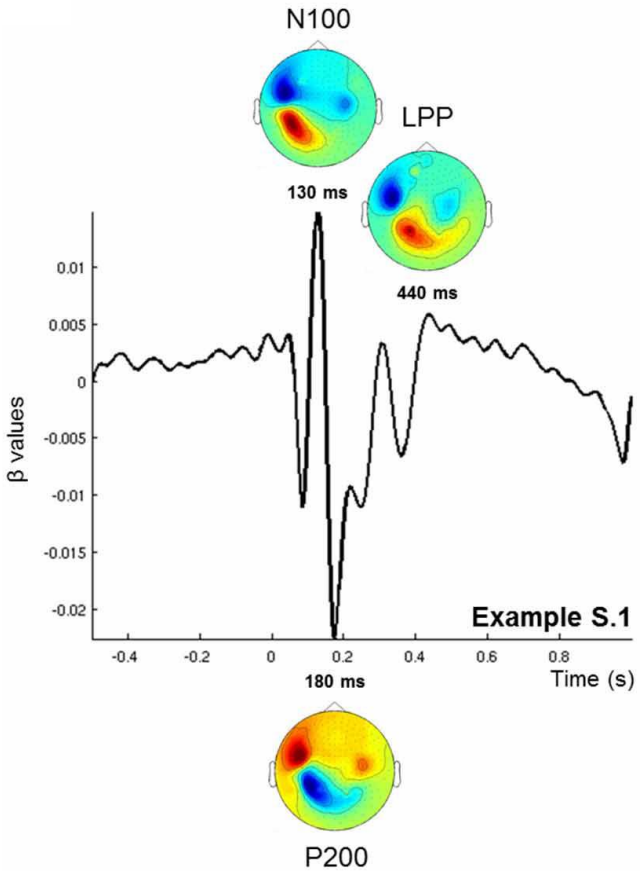
At the group level, ERP sensitivity to phase noise was reduced between about 140 and 300 ms when stimulus phase information was task irrelevant.

we observed a significant task effect in only 60% of subjects, and at any time point only 31% of subjects showed results consistent with group analyses

Rousselet et al. Front Psy 2011

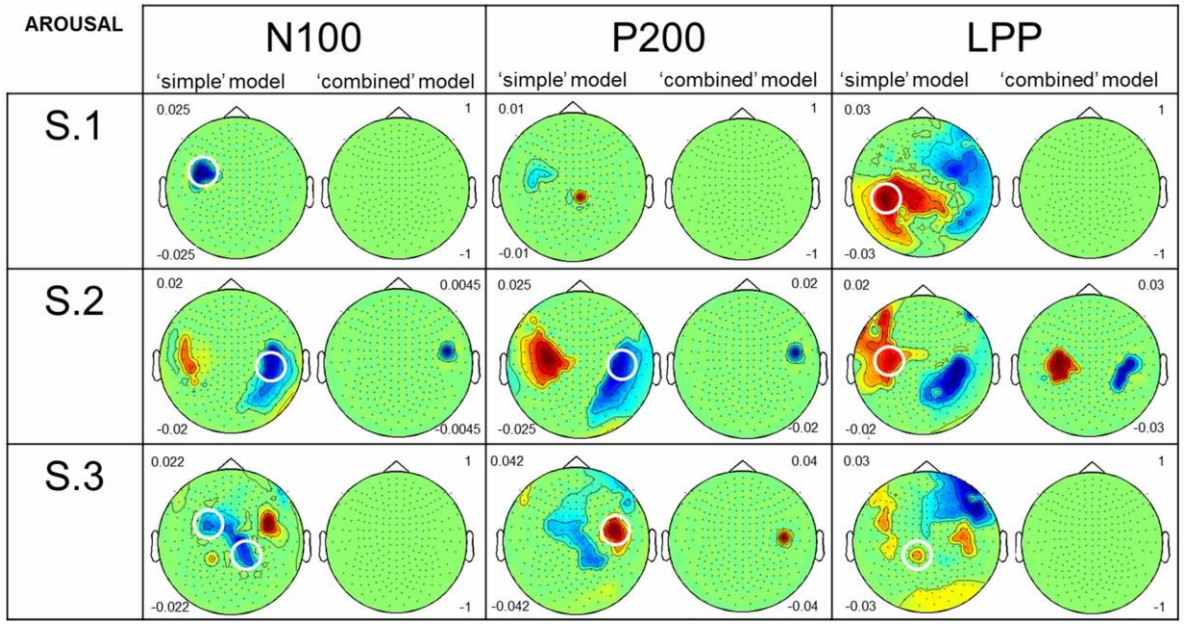
MEG of acoustic properties in affective vocalizations





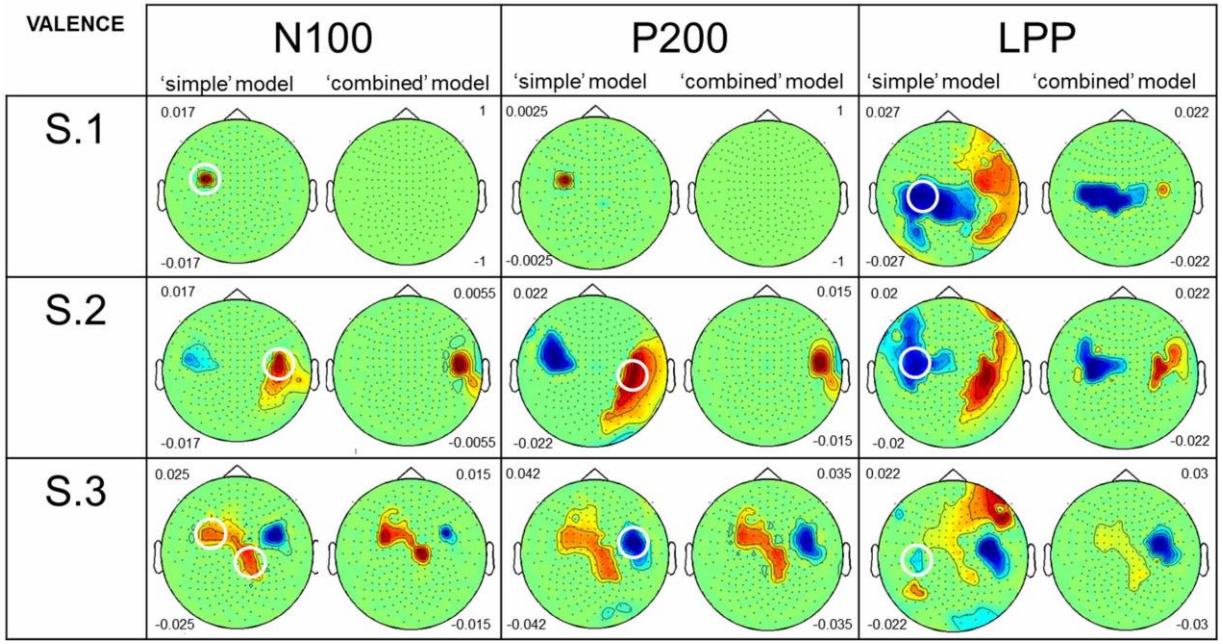
Salvia et al. Front Neurosc 2014

MEG of acoustic properties in affective vocalizations



Simple model: for each sound, input the arousal and valence value - *Combined model*: valence, arousal, and 2 components of a PCA (72% var) from six acoustic parameters: mean/ SD of f0, HNR and percentages of unvoiced frame, jitter and shimmer.

Early effects are largely driven by acoustical variations Once the variance explained by acoustic properties is accounted for, the remaining effects of emotionalv variables (especially valence) are mostly observed at late stages (~400–600 ms).



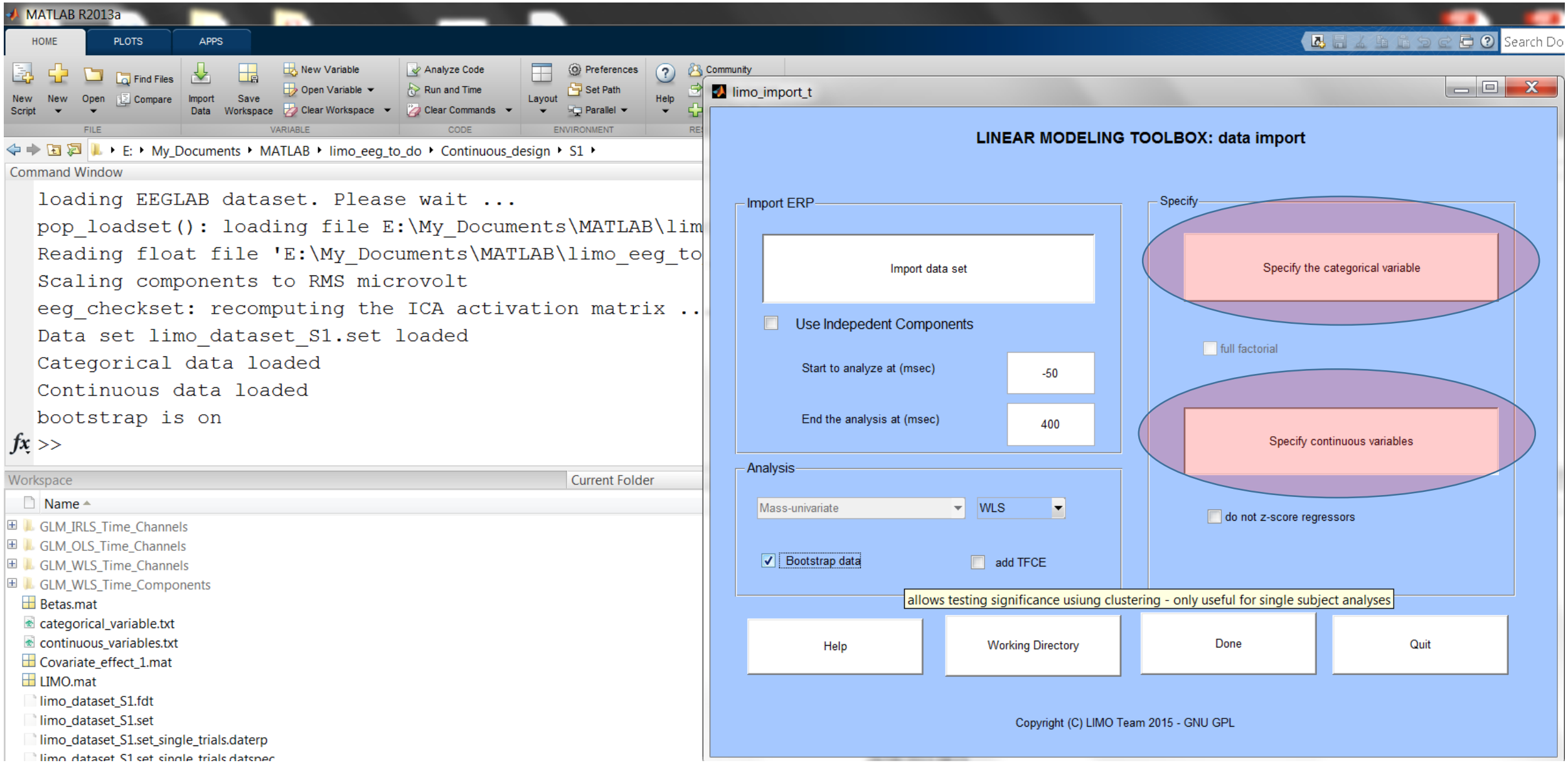
Application to a continuous design

Let’s analyse one subject

- **Design**: 2 faces (cond1/cond2) + a continuous variable related to the phase information in the stimulus space (~noise)

- LIMO EEG – 1st level analysis

= with ERP/means you are limited to ‘categories’, here we have beta1 = face1, beta2 = face2, beta3=local phase coherence, beta4 = constant



The new STUDY allows any type of regressors – sometimes it's difficult to have it all encoded in the .set
Here we use txt file, 1 row per trials encoding faces (1/2) and the actual phase coherence value

Let’s have a look at txt files and edit paths

- edit ‘set_list.txt’, ‘cat_list.txt’ and ‘cont_list.txt’ with the path on your hard drive (find/replace)
- Categorical and Continuous files are nothing but a description of single trials (in the same order as in the .set)

Categorical variable

1
2
2
1
1
2
1
1
2

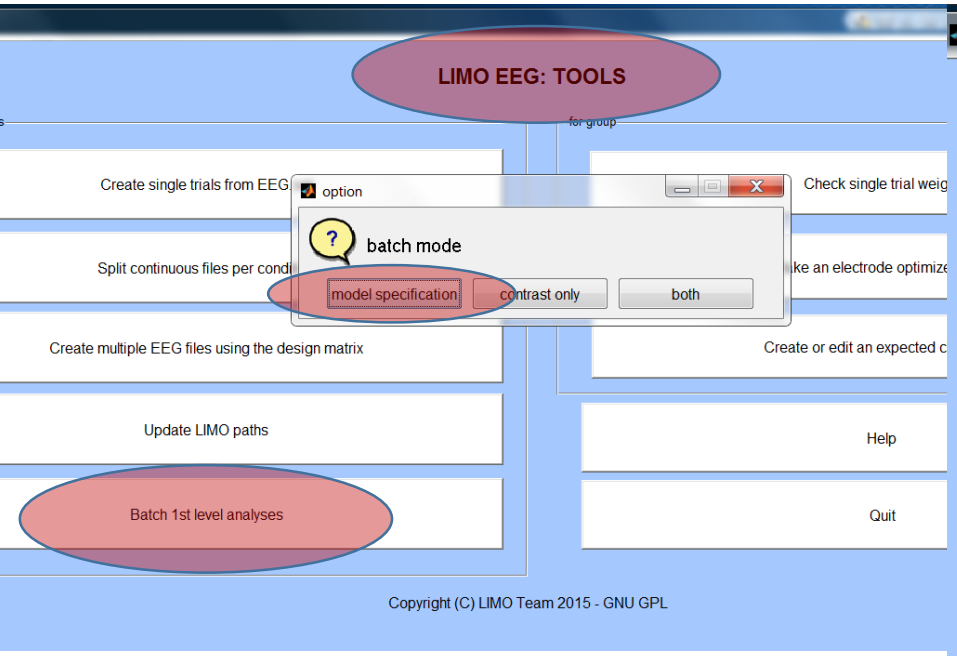
You can use as many conditions as you like – and code them with any number you like

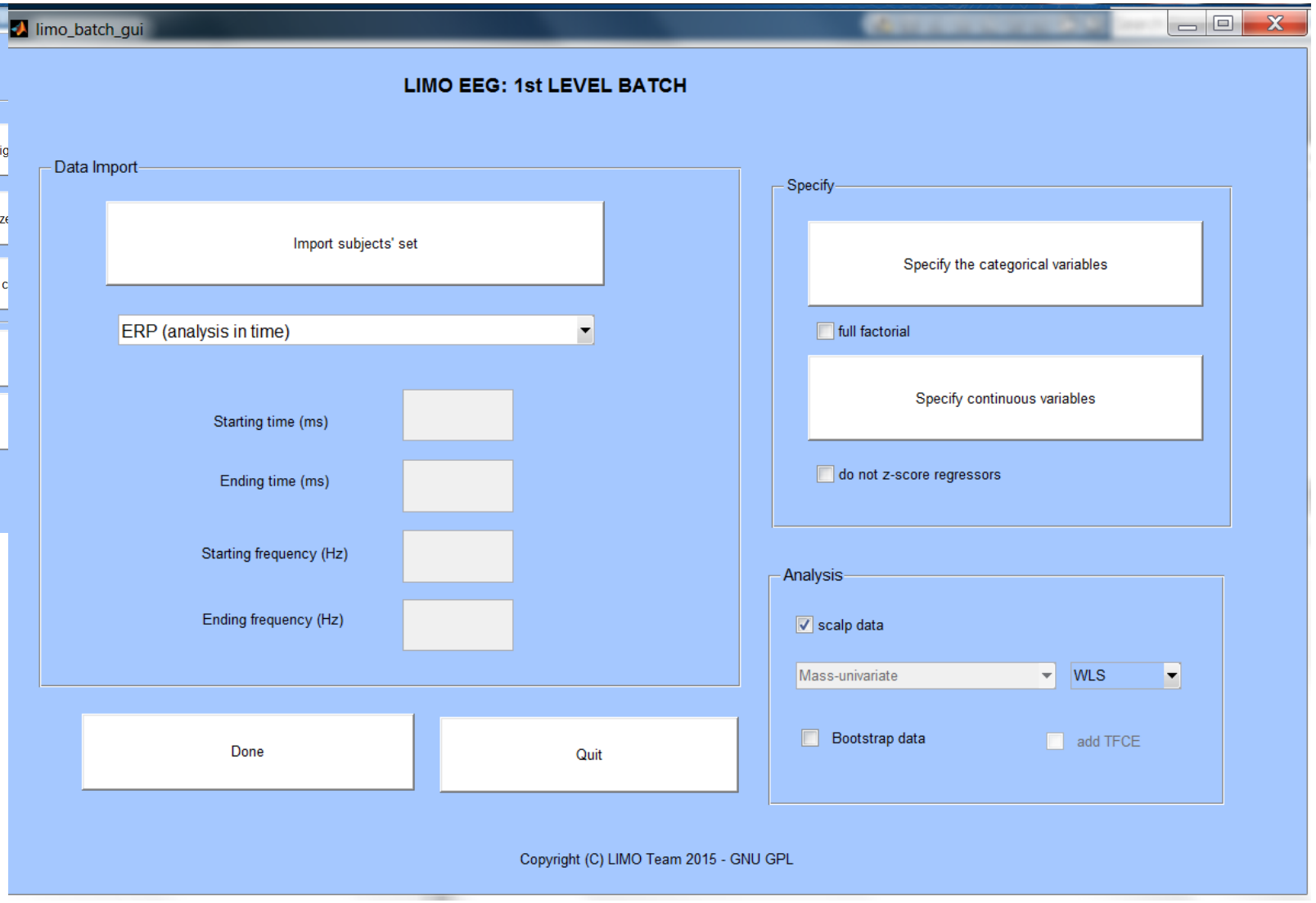
Continuous variable

1.2
2.5
1
0.9
1.2
1.5
1.9
1.6
2.1

You can use as many variable as you like, simply add columns

Use LIMO batch to do all subjects





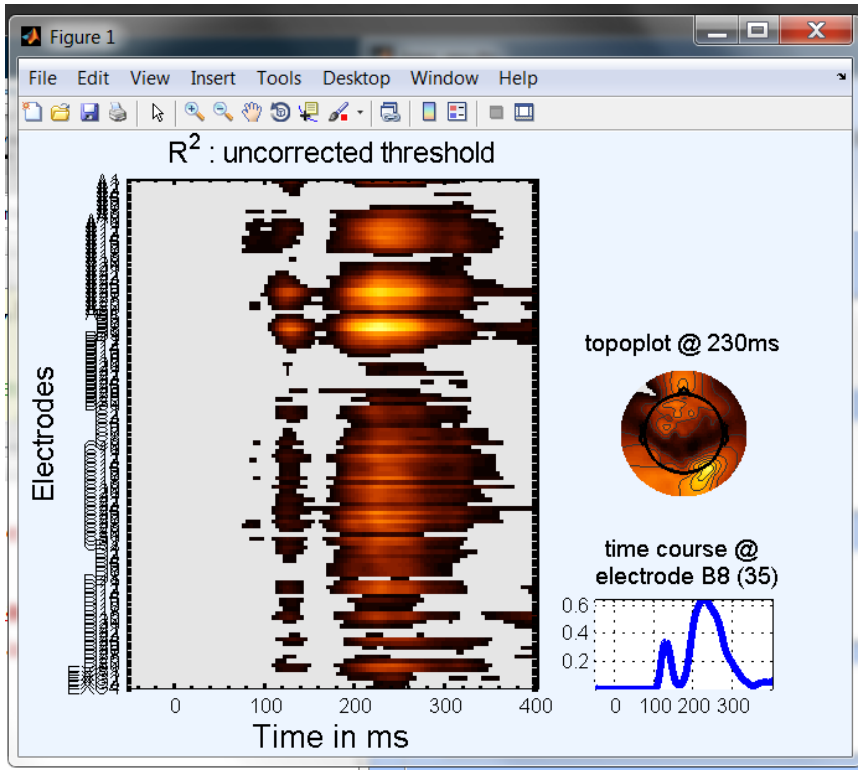
s is the engine behind 'STUDY'

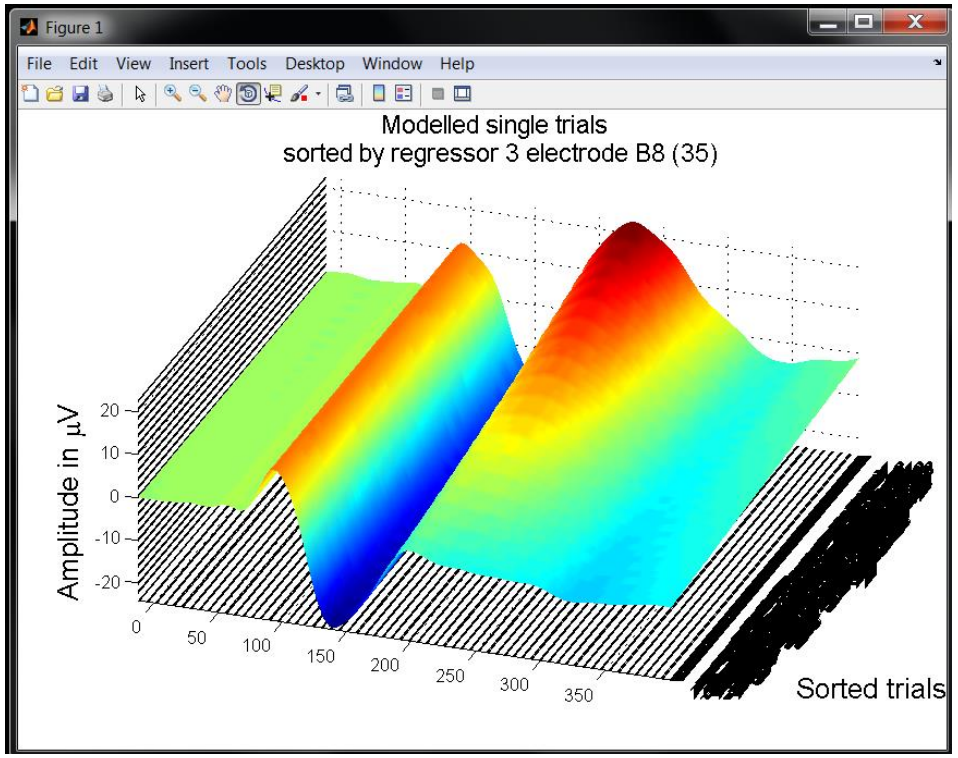
ect the set_list, cat_list, cont_list that u have edited

e limit [-50 450]

What have we done: results

- Image all (R2, condition, covariate)
- Course plots – for continuous variables, make 3D plots !





Group level analysis

- One sample t-test on ‘noise’ regressor

→From the GUI, choose ‘Random Effect’

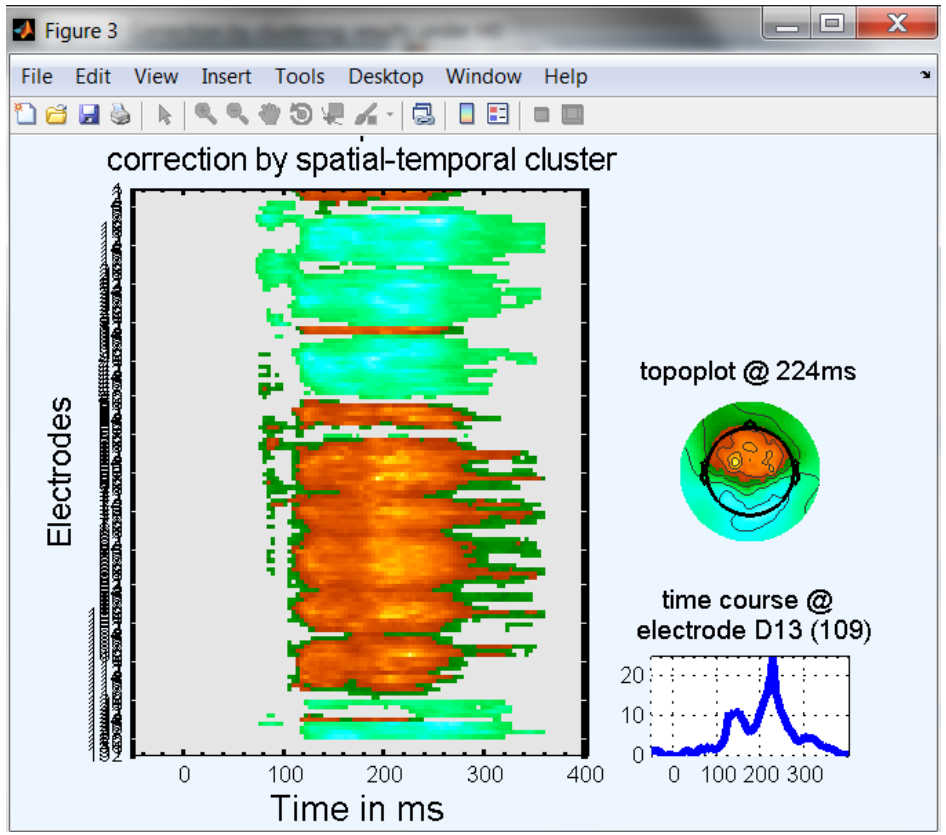
→Load expected_chanlocs

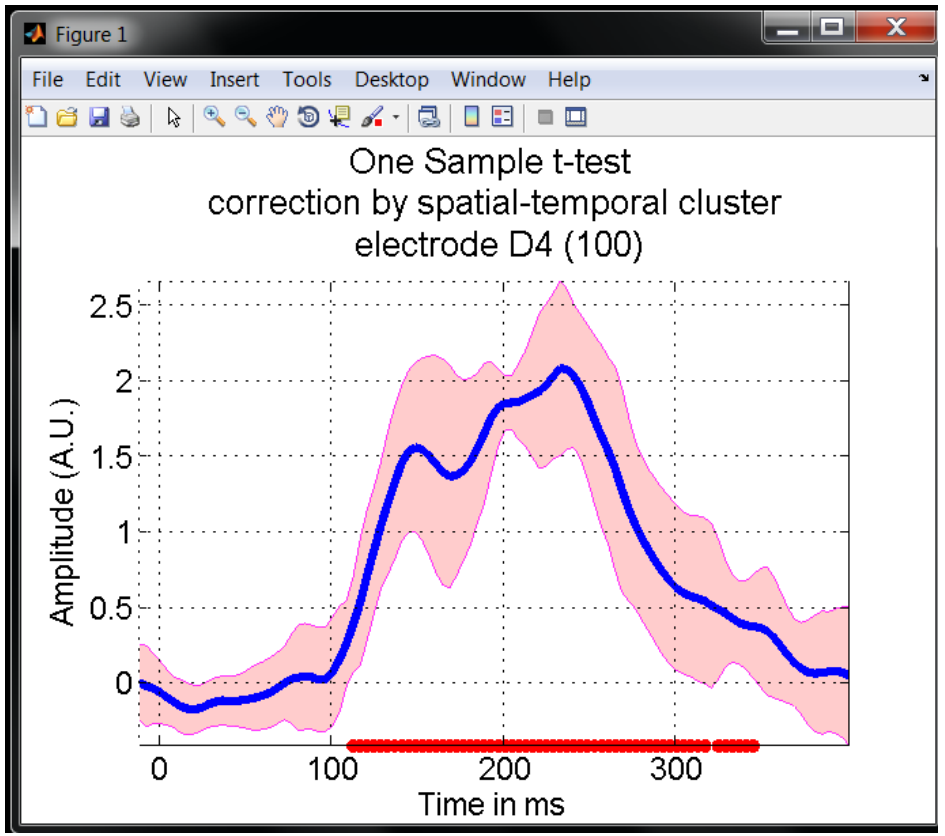
→Run the one sample t-test with bootstrap

→Because we used the batch, we have the list of parameters already there for all subjects (Beta_files_GLM_WLS_Time_Channels.txt) – pick this up or load Beta files one by one !

→Select parameter 3

Review gp level results

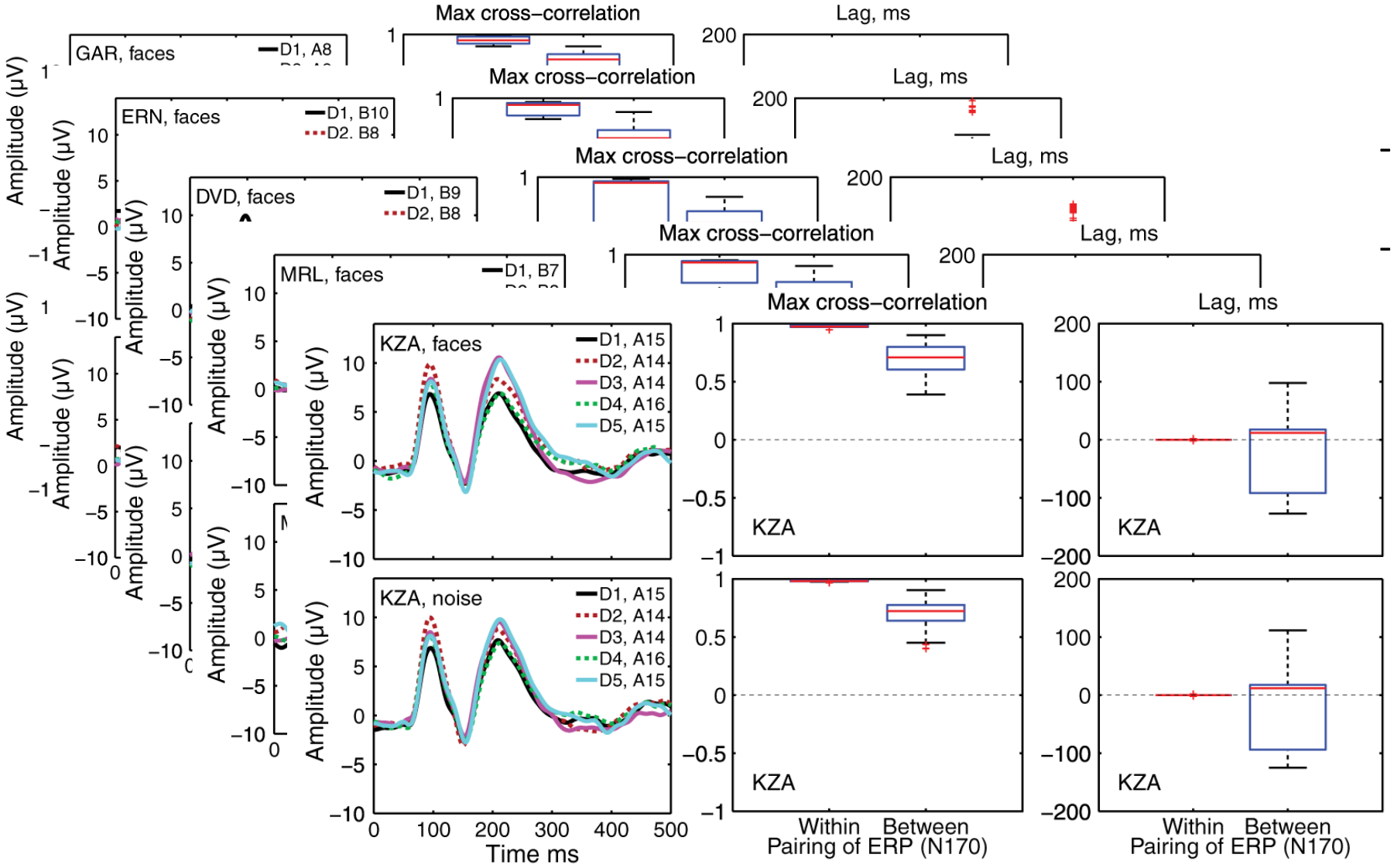




EEG signals are idiosyncratic

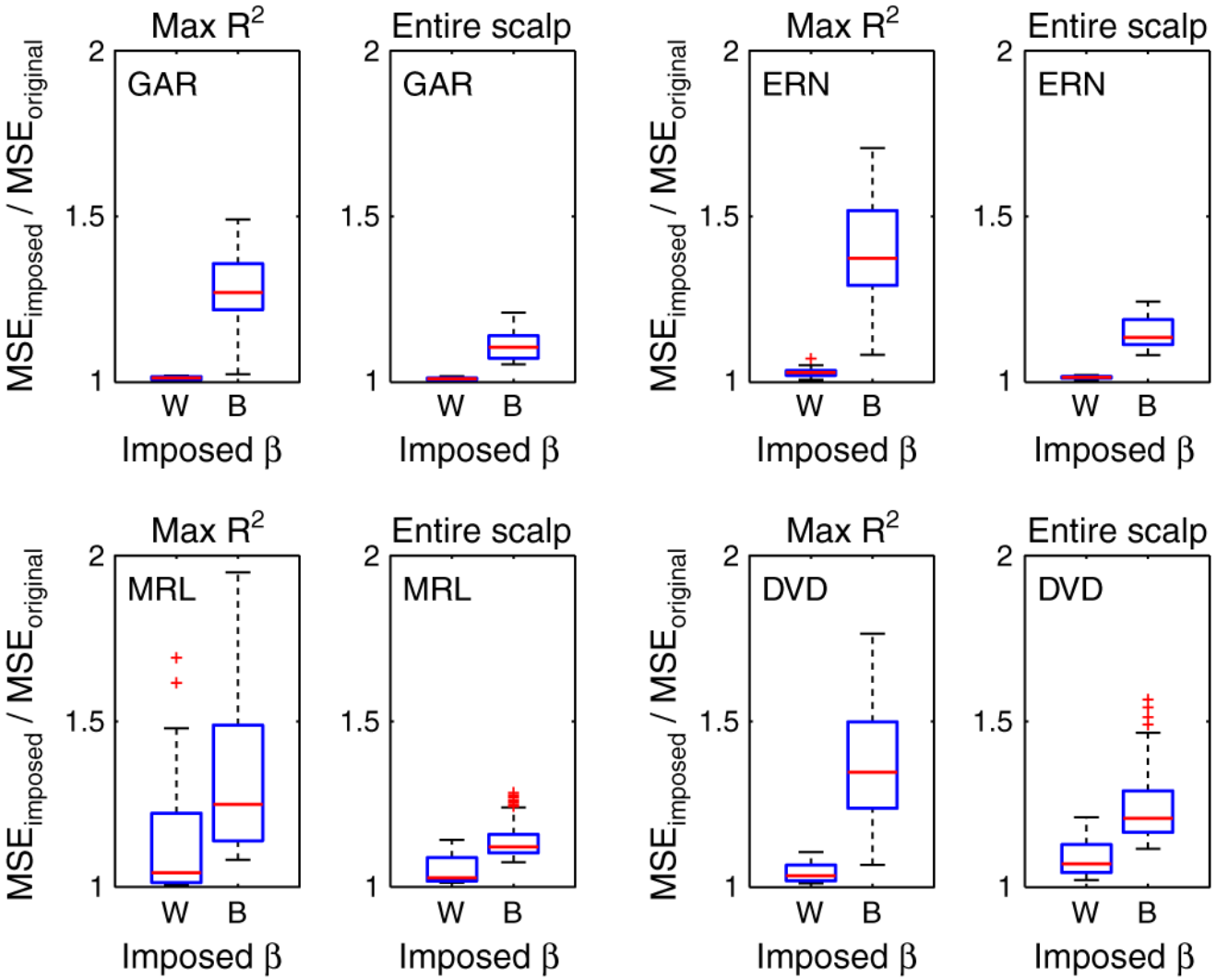
Gaspar et al. 2011 Reliability of ERP and single-trial analyses NeuroImage 58

Test-retest of ERPs



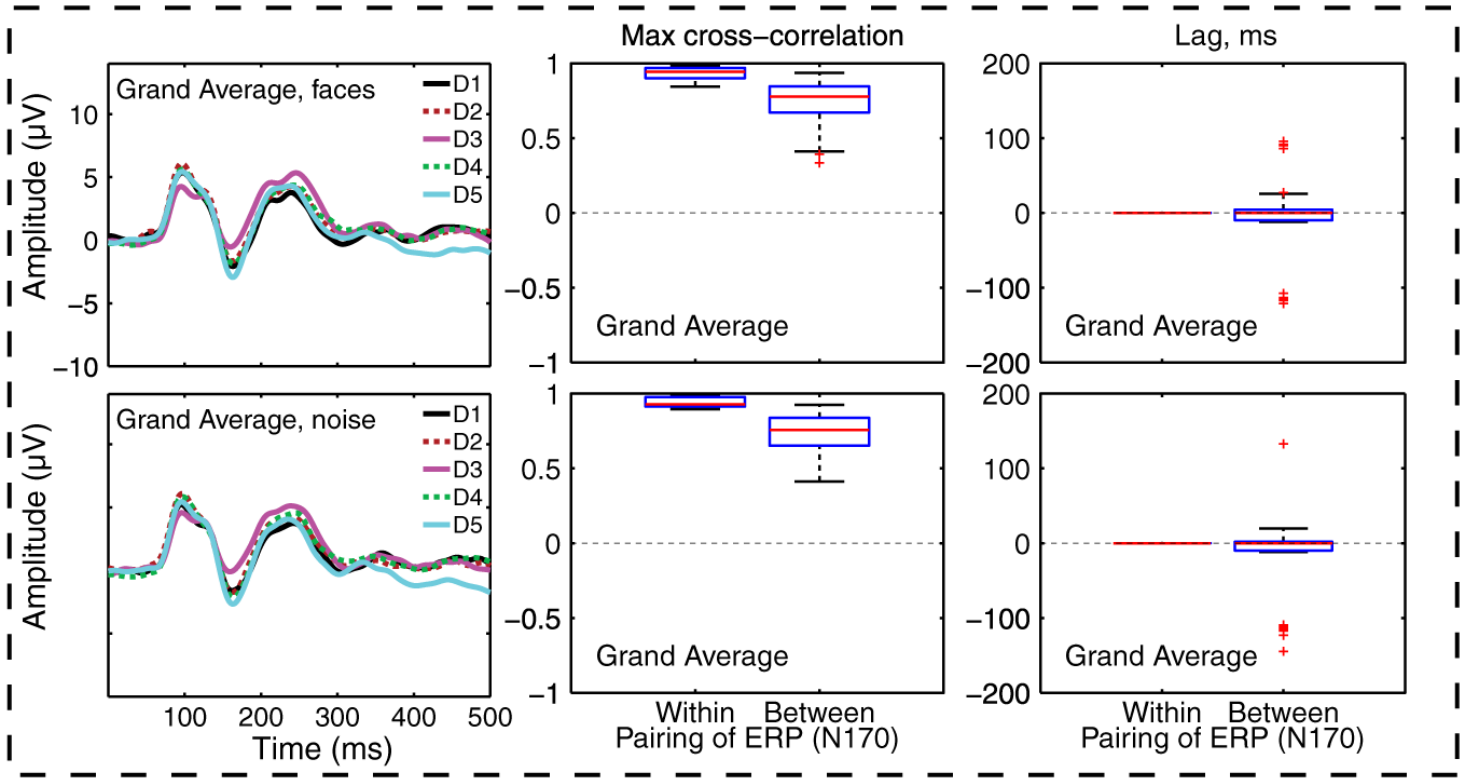
- ERPs are highly reliable within subjects
- xcorr >0.90 with ~4/6 ms lag

Test-retest for parameter estimates



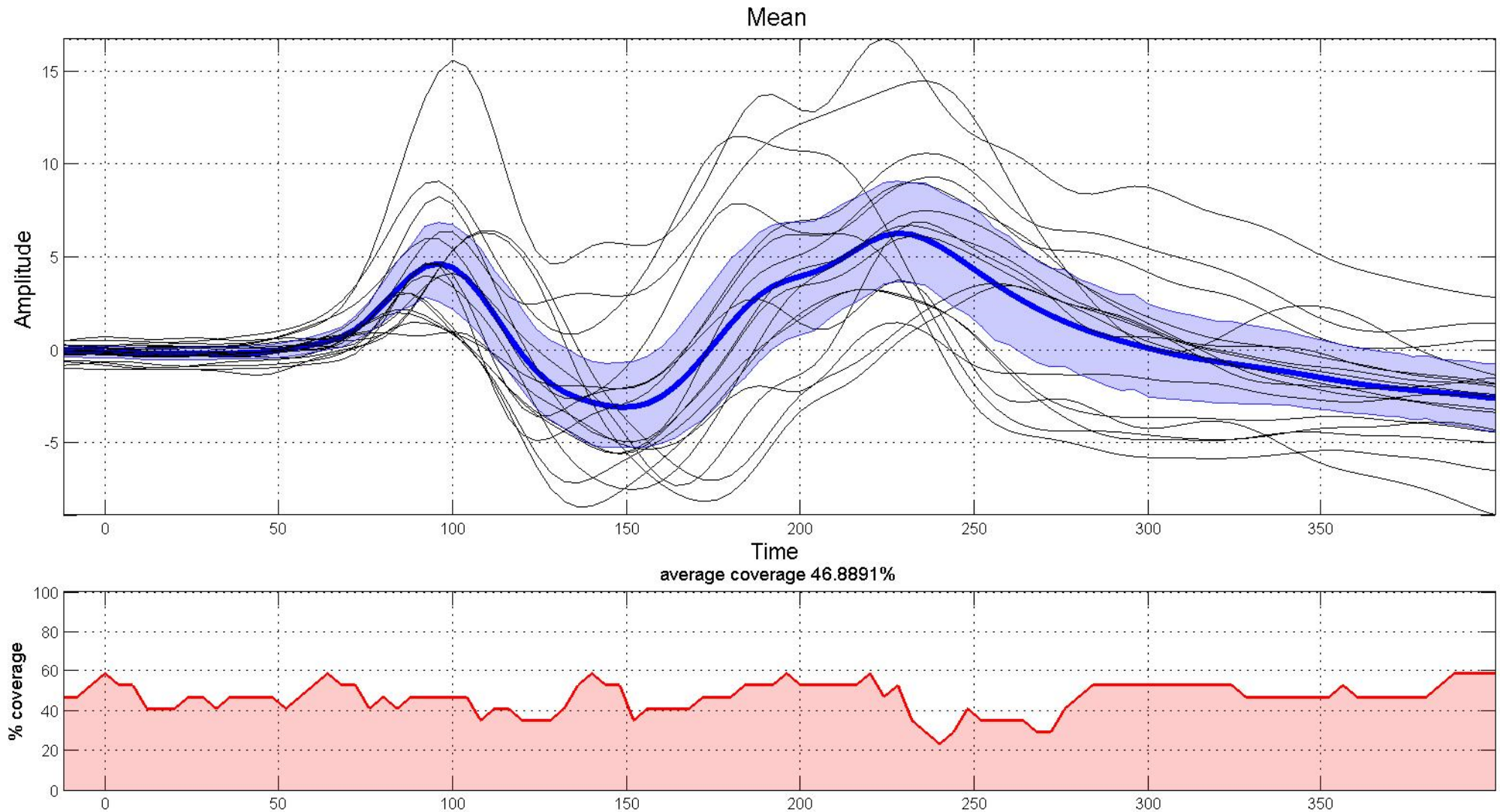
- Beta swapping
- The effect observed one day is the same another day !
- Effects (betas) are idiosyncratic like ERPS

Grand averages do not reflect ERP dynamics

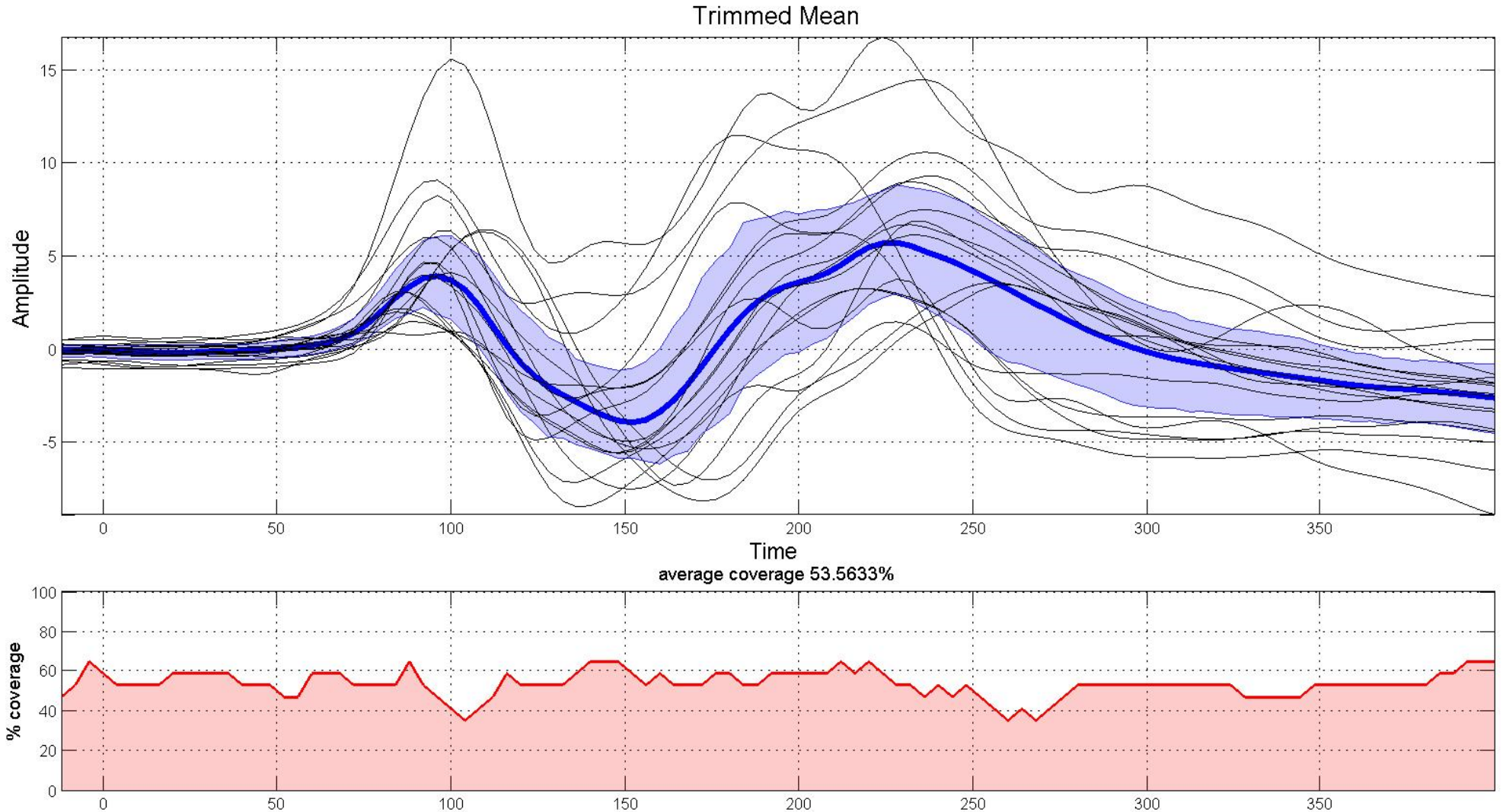


- Because ERPs are highly reliable within subjects, grand averages are also highly reliable.
- However, this 'within-subject' reliability also means that grand averages ERPs are significantly different from individual subjects' ERPs.
- Plots of grand average can be misleading

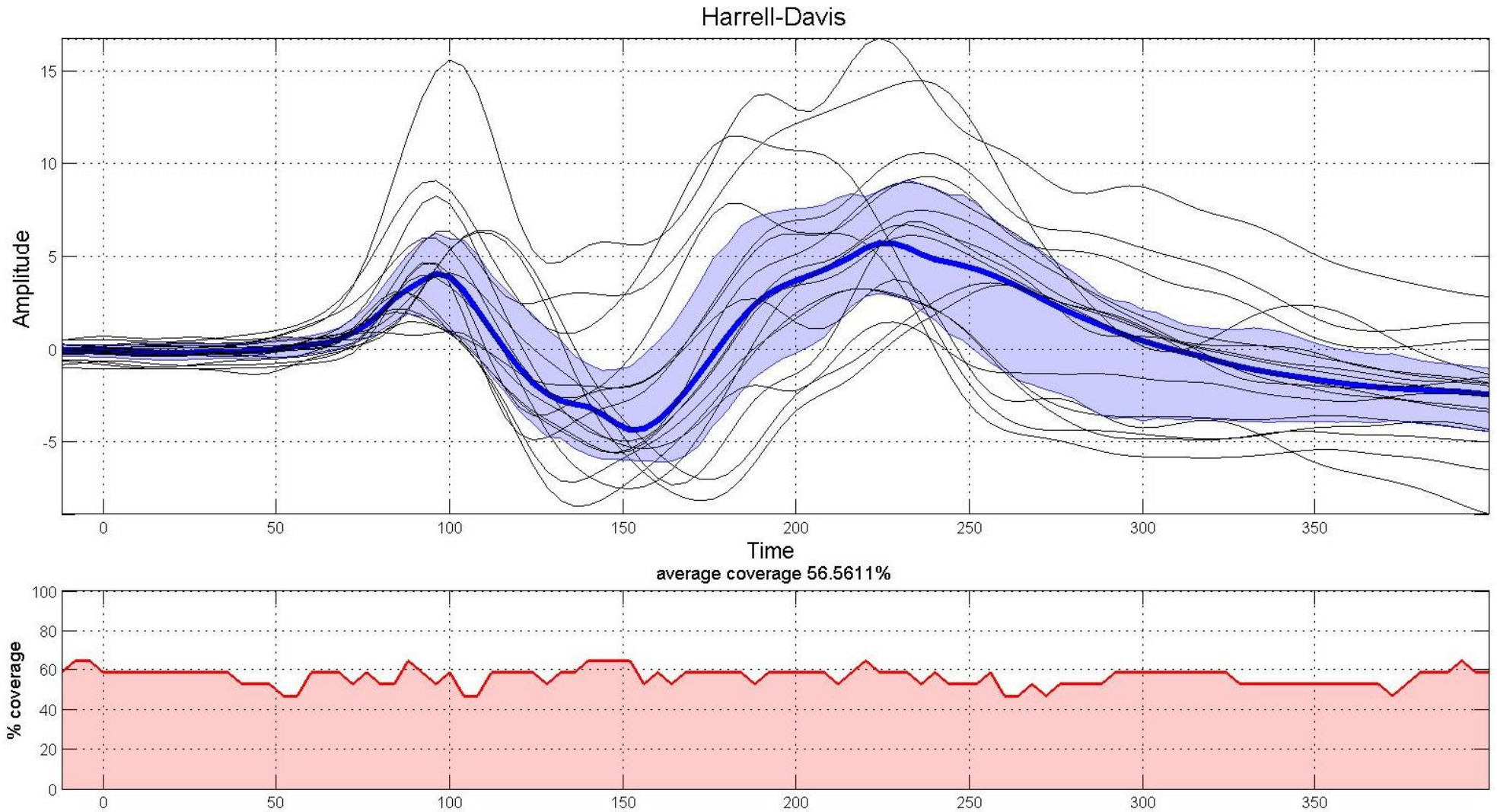
Grand averages do not reflect ERP dynamics



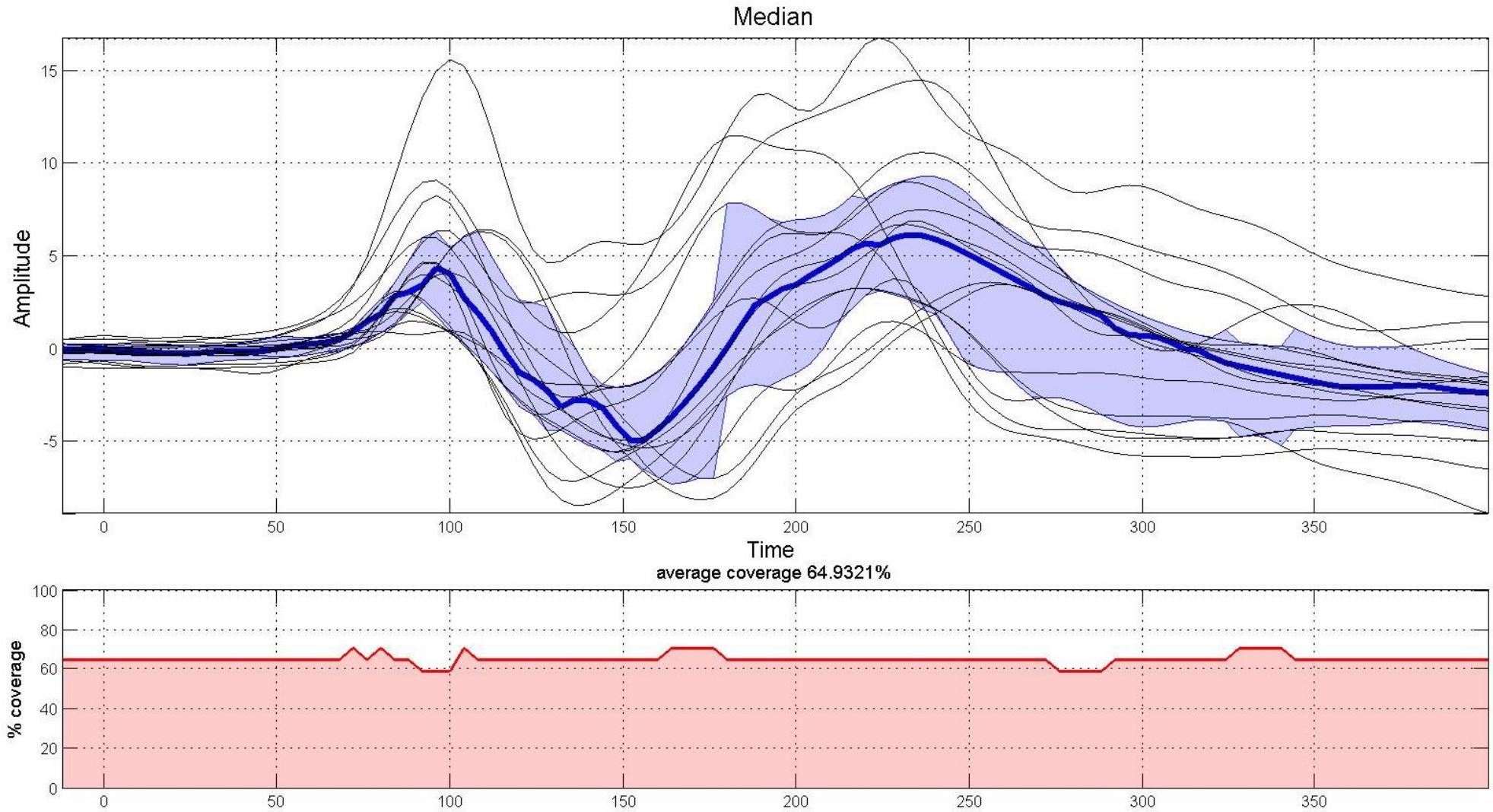
Grand averages do not reflect ERP dynamics



Grand averages do not reflect ERP dynamics



Grand averages do not reflect ERP dynamics



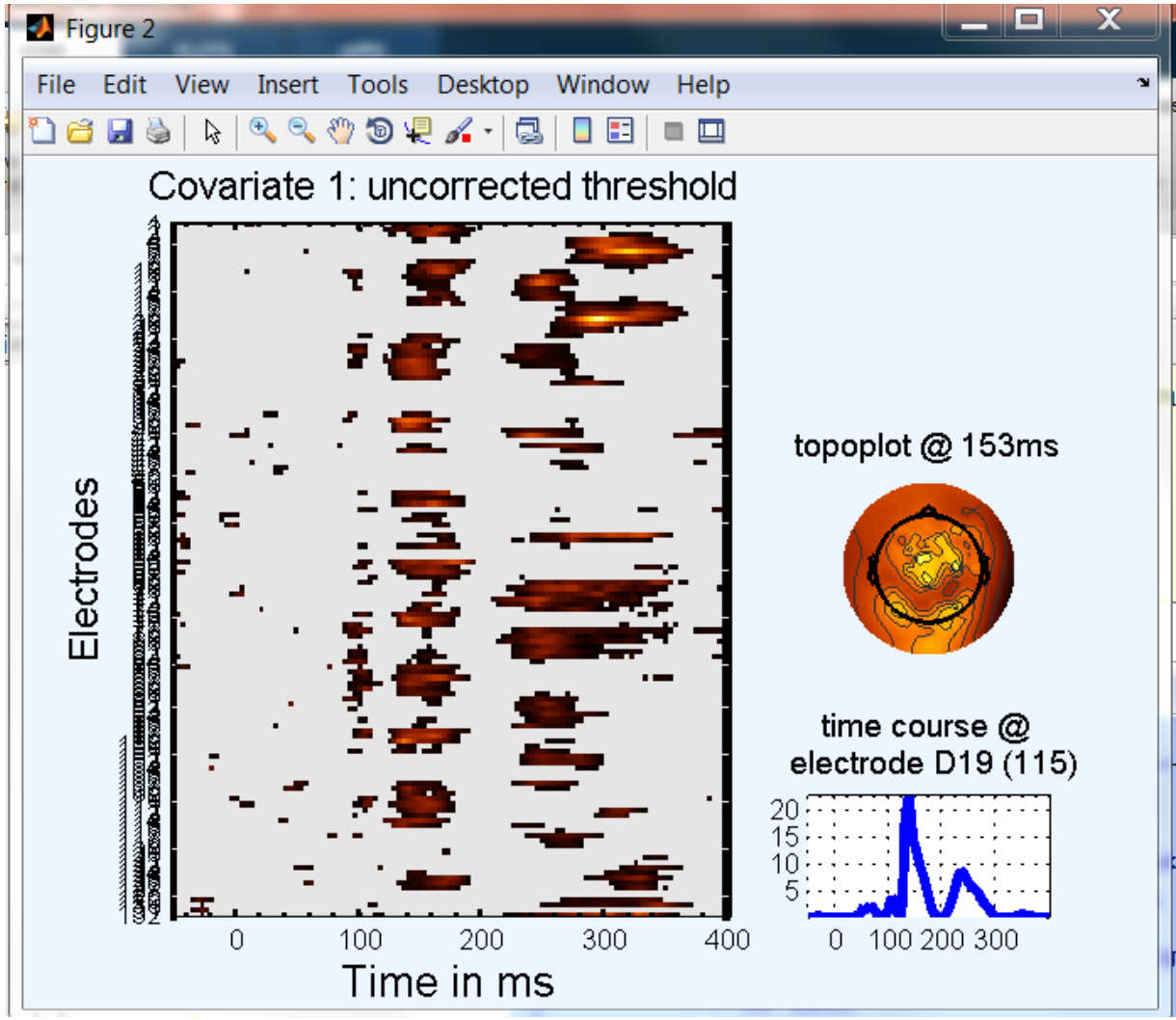
Always good to check distributions !

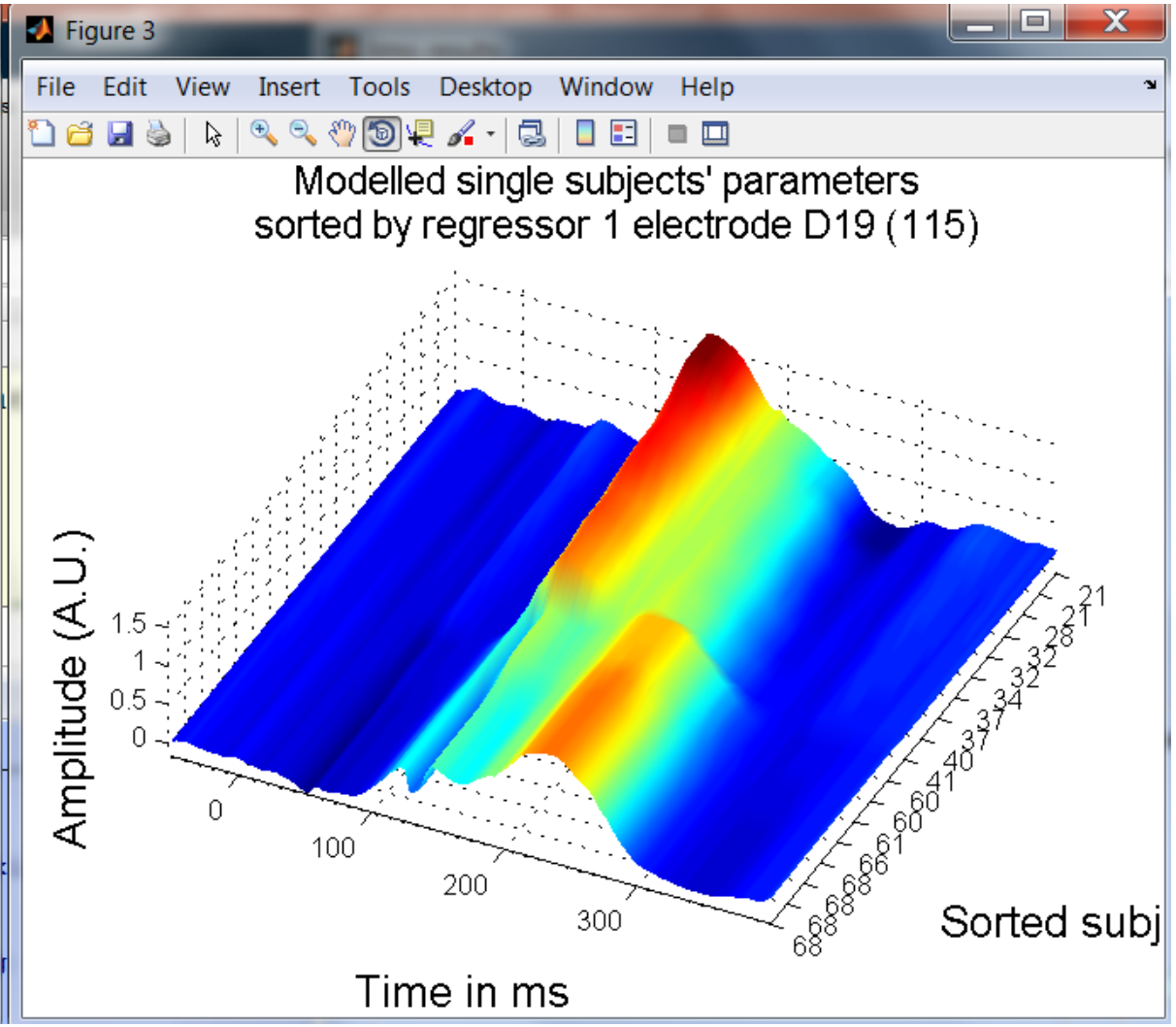
- LIMO RANDOM EFFECTS
- Central tendency and CI also gives you data for all subjects
- Parameter Plots → a set of tool to visually explore how ‘things’ are distributed

Going further with parametric analyses

- Group level regression on the phase regression parameter !
- So far we looked at how much phase coherence explains of the ERP (R2 values and semi-partial coefficients per subjects)
- At the group level we looked at when phase coherence influence the ERP
- Now we can test if the phase coherence influence on the ERP is a function of age

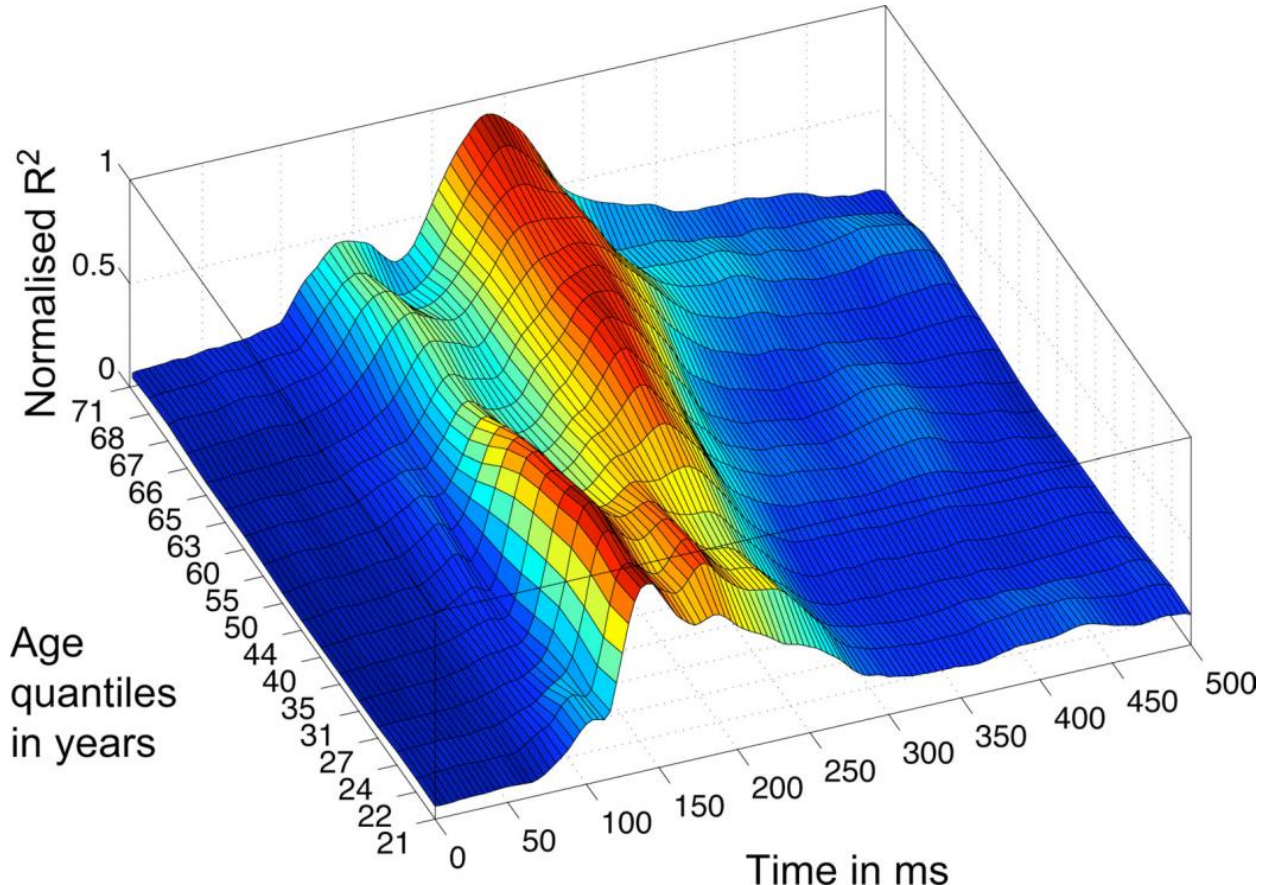
Group level regression





Going further with parametric analyses

- Alternative analyses to quantify effects are possible. For instance, use the R2, cumulate, normalize across subjects and regress age



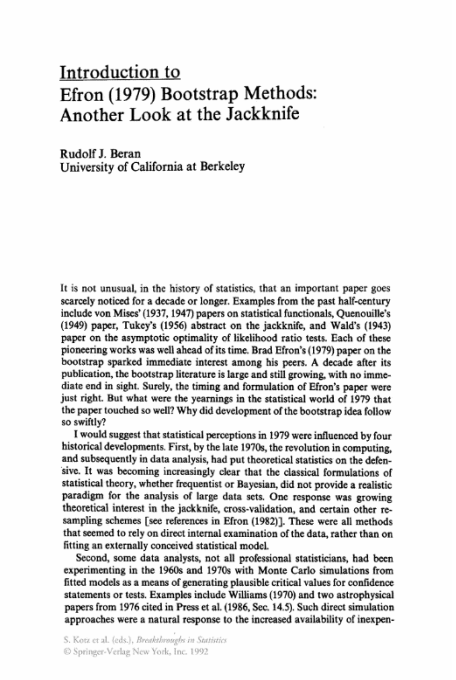
→ tells you how fast one accumulates face information in noise and this changes as we get older (with a big shift around 45 y.o.)

Rousselet, et al. (2010), Front Psy

Almost the end

The maths behind the GUI

- If you want to go now, it’s fine
- Review bootstrap and application to CI
- Details further application to multiple comparisons correction



Introduction to
Efron (1979) Bootstrap Methods:
Another Look at the Jackknife

Rudolf J. Beran
University of California at Berkeley

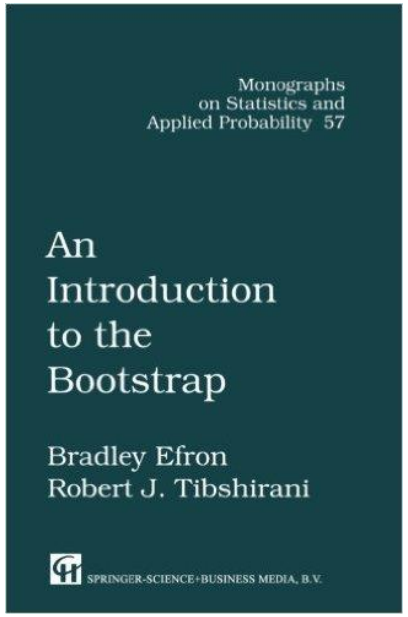
It is not unusual, in the history of statistics, that an important paper goes scarcely noticed for a decade or longer. Examples from the past half-century include von Mises' (1937, 1947) papers on statistical functionals, Quenouille's (1949) paper, Tukey's (1956) abstract on the jackknife, and Wald's (1943) paper on the asymptotic optimality of likelihood ratio tests. Each of these pioneering works was well ahead of its time. Brad Efron's (1979) paper on the bootstrap sparked immediate interest among his peers. A decade after its publication, the bootstrap literature is large and still growing, with no immediate end in sight. Surely, the timing and formulation of Efron's paper were just right. But what were the yearnings in the statistical world of 1979 that the paper touched so well? Why did development of the bootstrap idea follow so swiftly?

I would suggest that statistical perceptions in 1979 were influenced by four historical developments. First, by the late 1970s, the revolution in computing, and subsequently in data analysis, had put theoretical statistics on the defensive. It was becoming increasingly clear that the classical formulations of statistical theory, whether frequentist or Bayesian, did not provide a realistic paradigm for the analysis of large data sets. One response was growing theoretical interest in the jackknife, cross-validation, and certain other resampling schemes [see references in Efron (1982)]. These were all methods that seemed to rely on direct internal examination of the data, rather than on fitting an externally conceived statistical model.

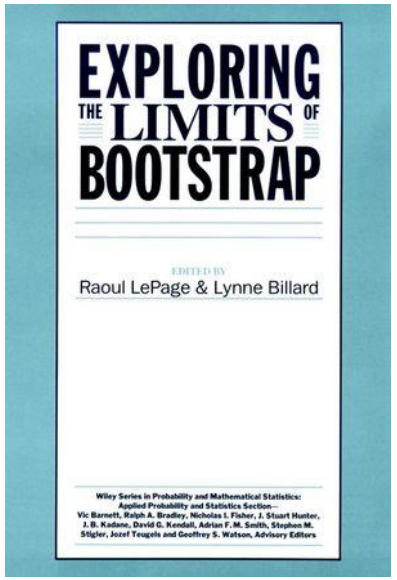
Second, some data analysts, not all professional statisticians, had been experimenting in the 1960s and 1970s with Monte Carlo simulations from fitted models as a means of generating plausible critical values for confidence statements or tests. Examples include Williams (1970) and two astrophysical papers from 1976 cited in Press et al. (1986, Sec. 14.5). Such direct simulation approaches were a natural response to the increased availability of inexpen-

S. Kotz et al. (eds.), *Breakthroughs in Statistics*
© Springer-Verlag New York, Inc. 1992

Efron , B. (1979). Bootstrap methods; another look at the jackknife . *Ann. Statist* . **7** , 1 – 26



Efron , B. , and Tibshirani , R. (1993). *An Introduction to the Bootstrap* . Chapman & Hall , New York



LePage, R & Billard L (Ed)
Exploring the Limits of Bootstrap, 1992

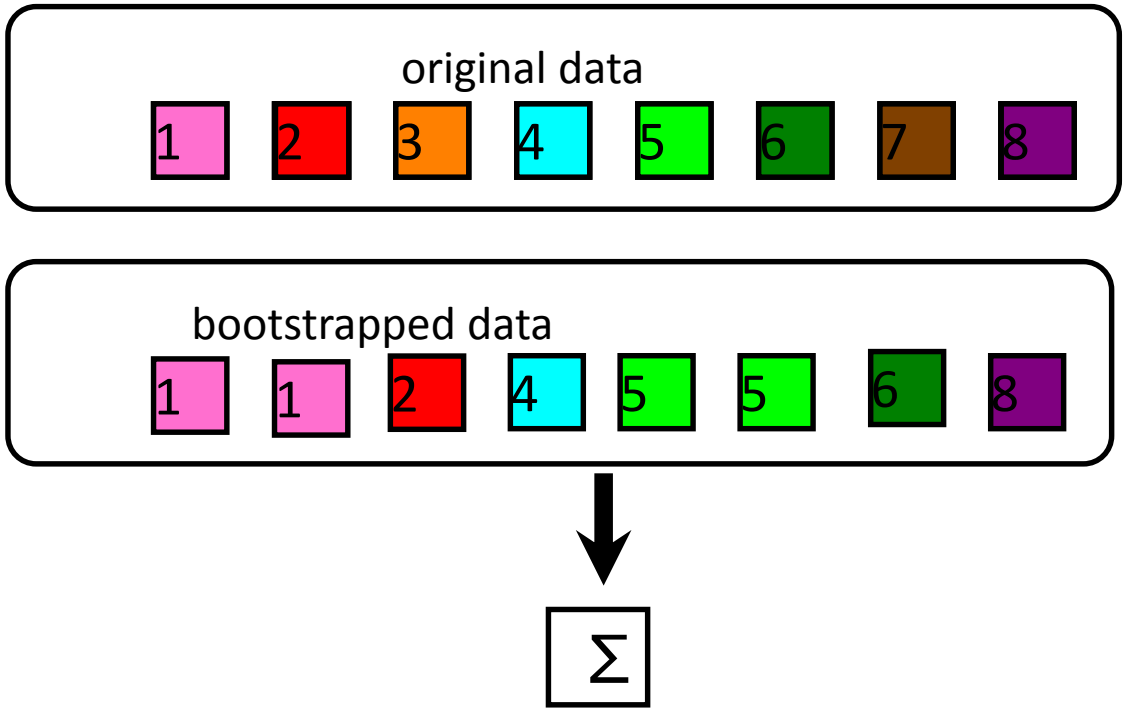
Bootstrap: central idea

- Statistics rely on estimators (e.g. the mean) and measures of accuracy for those estimators (standard error and confidence intervals)
- “The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates.” Efron & Tibshirani, 1993
- The bootstrap is a type of resampling procedure along with jack-knife and permutations.

- Bootstrap is particularly effective at estimating accuracy (bias, SE, CI) but it can also be applied to many other problems – in particular to estimate distributions.

General recipe

(1) sample WITH replacement n observations (under H1 for CI of an estimate, under H0 for the null distribution)



(2) compute estimate
e.g. sum, trimmed mean

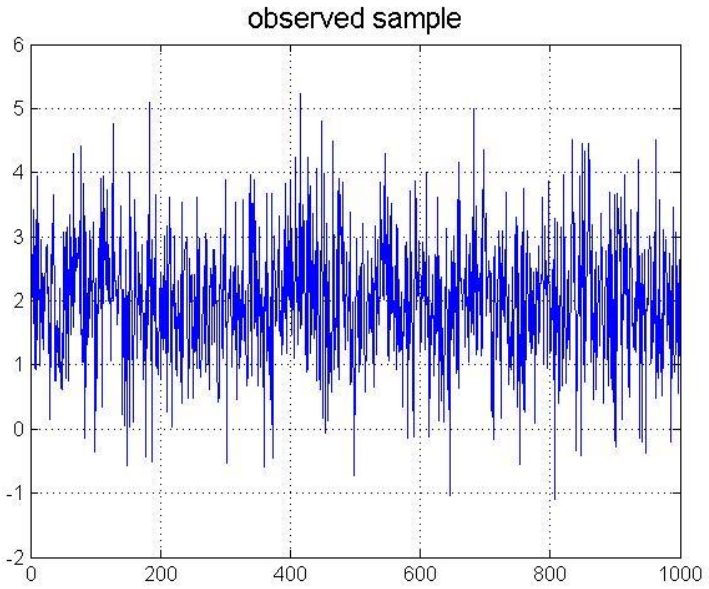
(3) repeat (1) & (2) b times

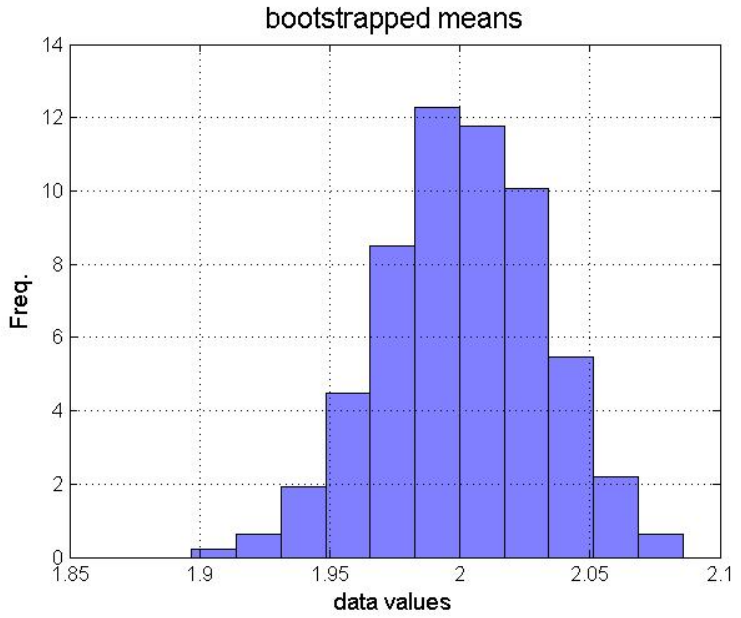
Σ_1 Σ_2 Σ_3 Σ_4 Σ_5 Σ_6 . . . Σ_b

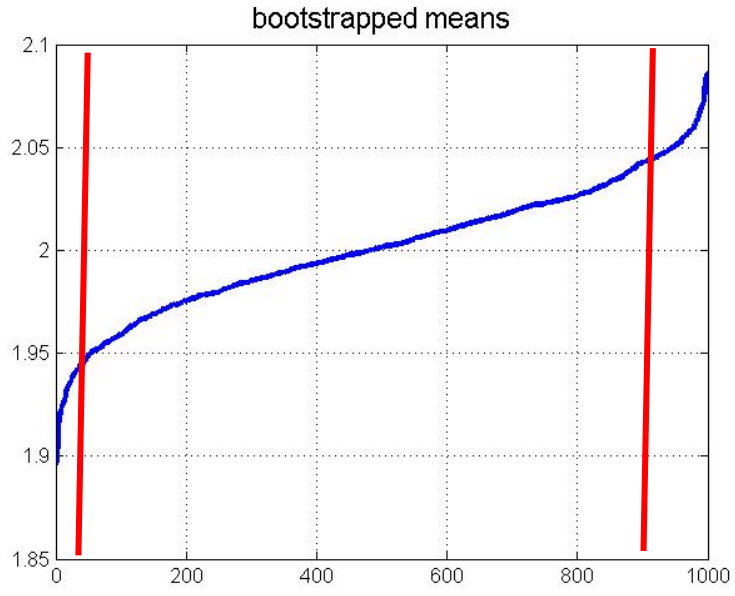
(4) get bias, std, confidence interval, p-value

Percentile boot Confidence Interval

- Let ϑ *be an estimator, and we* want the 1-alpha CI(ϑ)
- *Bootstrap the data computing* ϑ^* *to obtain a distribution of this parameter and take the 1-alpha/2 upper and lower percentile*







upper and lower percentiles

The Annals of Statistics
1981, Vol. 9, No. 1, 130–134

THE BAYESIAN BOOTSTRAP

BY DONALD B. RUBIN

Educational Testing Service

The Bayesian bootstrap is the Bayesian analogue of the bootstrap. Instead of simulating the sampling distribution of a statistic estimating a parameter, the Bayesian bootstrap simulates the posterior distribution of the parameter; operationally and inferentially the methods are quite similar. Because both methods of drawing inferences are based on somewhat peculiar model assumptions and the resulting inferences are generally sensitive to these assumptions, neither method should be applied without some consideration of the reasonableness of these model assumptions. In this sense, neither method is a true bootstrap procedure yielding inferences unaided by external assumptions.

Bayesian bootstrap

- In the bootstrap, we sample each *x i* with replacement, with a probability 1/ *n – the assumption is that only the observed value are possible values in the parent population*
- In the Bayesian bootstrap, we use a posterior probability distribution for the *X i '* s.
- Rubin's algorithm: (1) draw u=1:n-1 from uniform
  (2) sort u u(0) =0 and u(n) = 1
  (3) gap = u(i)-u(i-1)

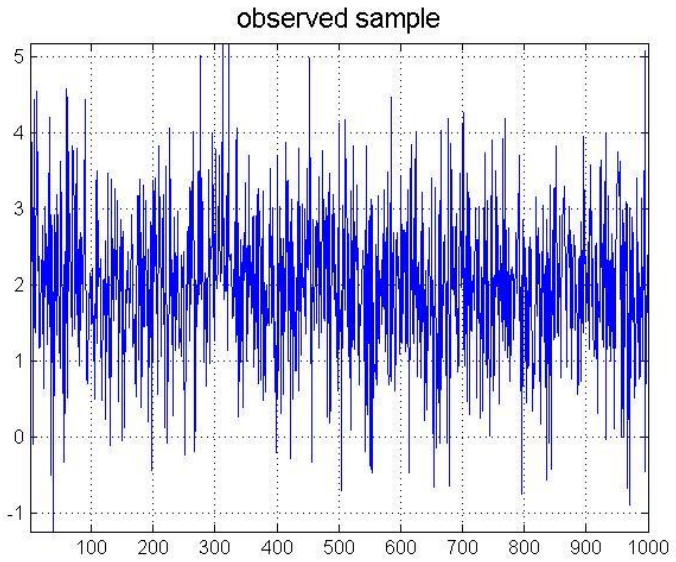
  } Substitute by a Dirichlet

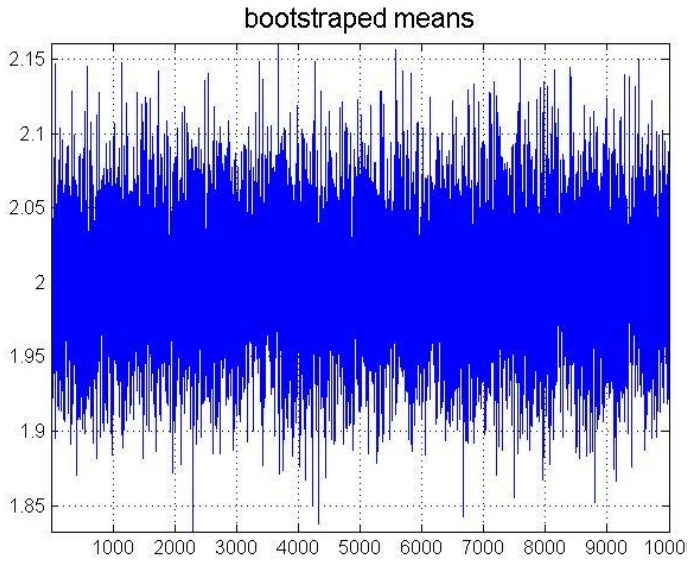
  (4) resample X using prob of xi = gap(i)
  → repeat B times

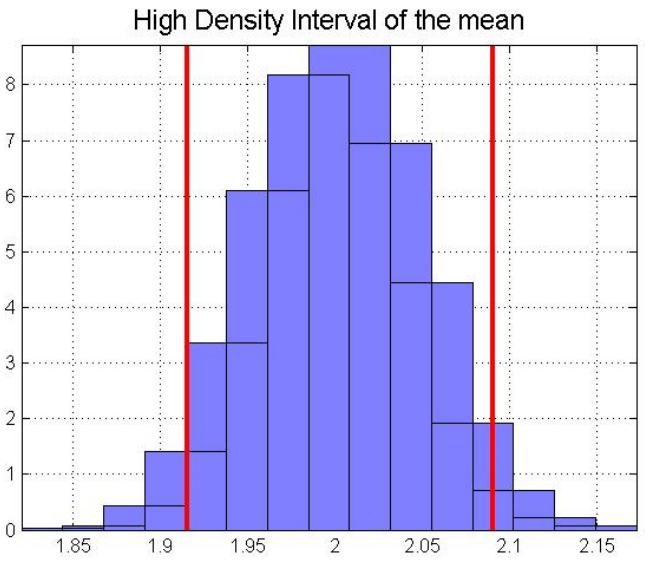
High Density Intervals

- Having the posterior density of means – we can compute the most dense intervals = credible intervals

→ compute the centile distances between bootstrap estimates and take the smallest (i.e. densest)

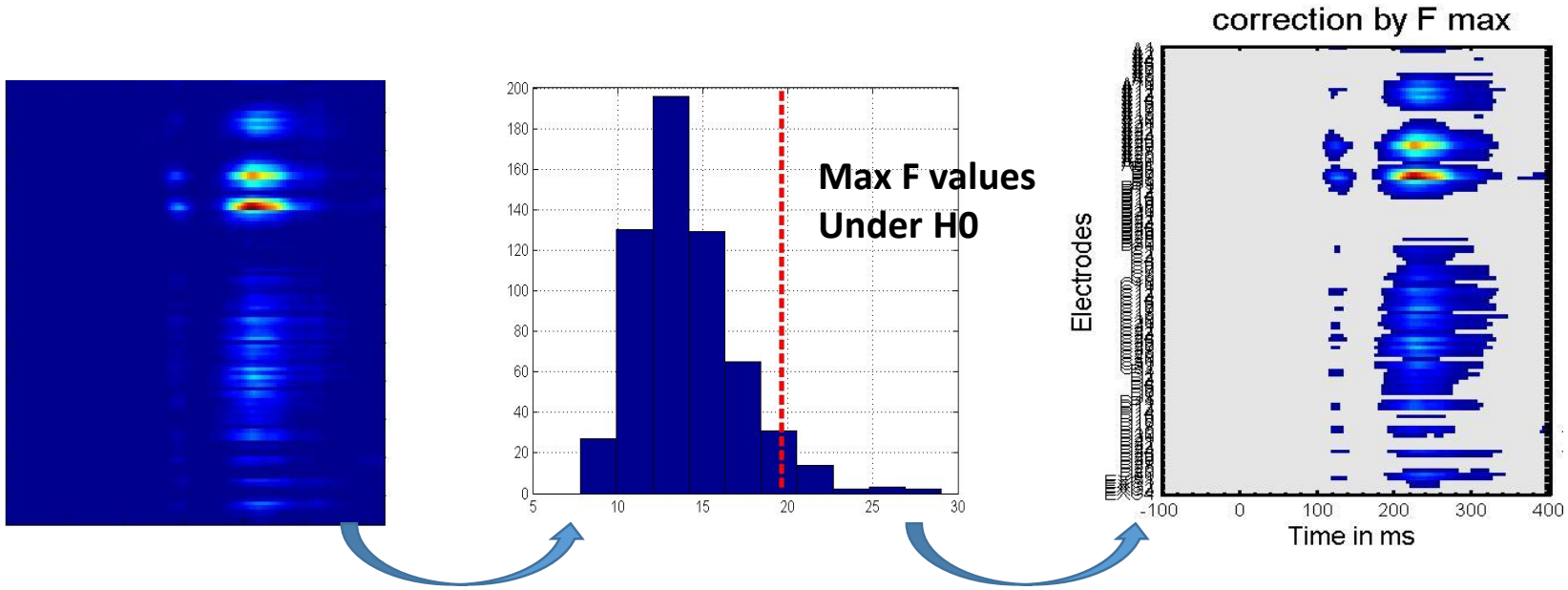






Correction for multiple testing using Maximum Statistics

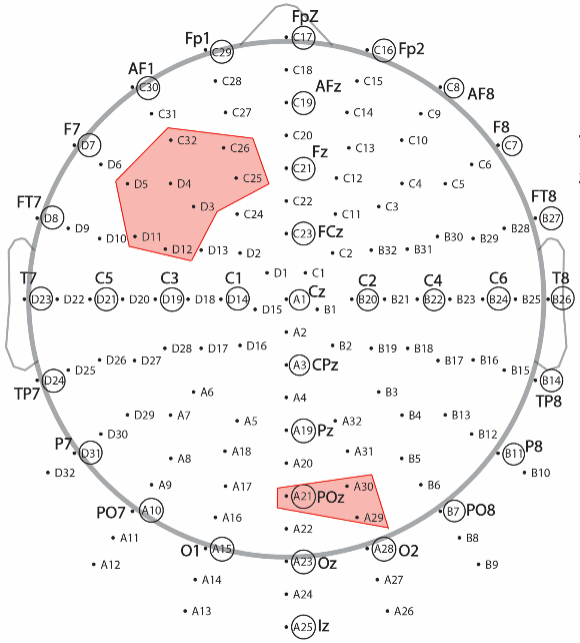
- Since the FWER is the prob that any stats > u, then the FWER is also the prob. that the max stats > u
- Estimate the distribution of max under H0 (bootstrap) and simply threshold the observed results a threshold u -- Still assumes all tests are independent

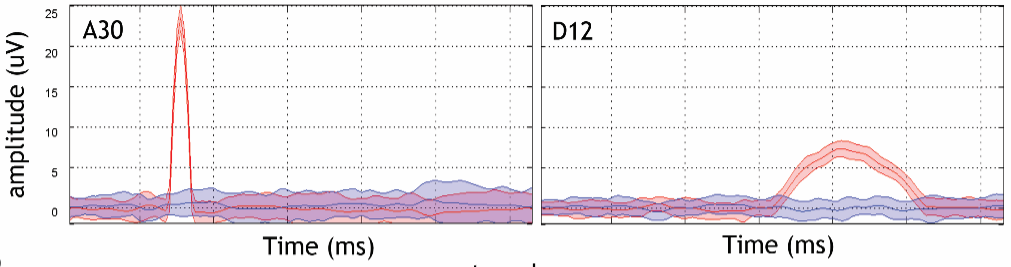


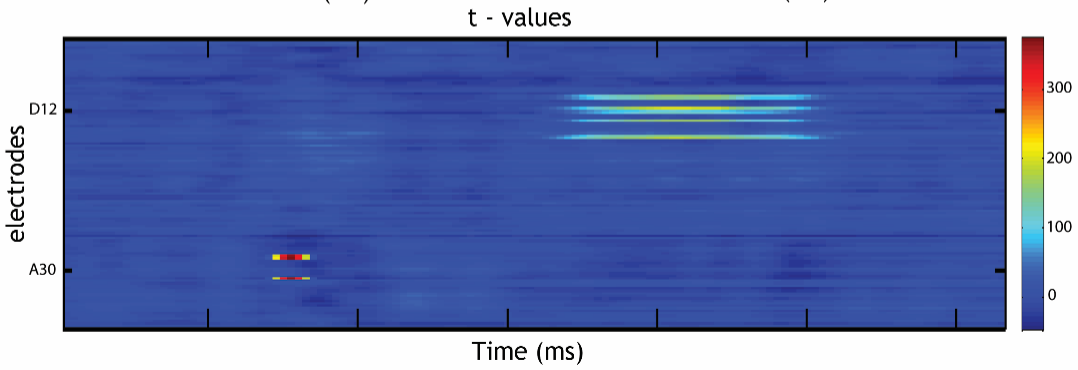
The clustering solution

- Clustering is a good option because it accounts for topological features in the data. Techniques like Bonferroni, FDR, max(stats) control the FWER but independently of the correlation between tests.

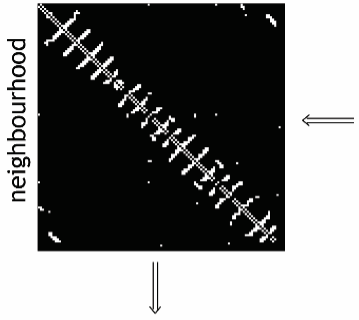
- To use clustering we need to consider cluster statistics rather than individual statistics
- Cluster statistics depend on (i) the cluster size, which depends on the data at hand (how correlated data are in space and in time/frequency), and (ii) the strength of the signal (how strong are the t, F values in a cluster) or (iii) a combination of both.







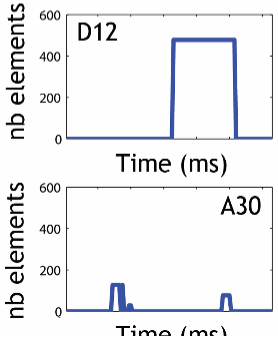
Spatial - Temporal clustering



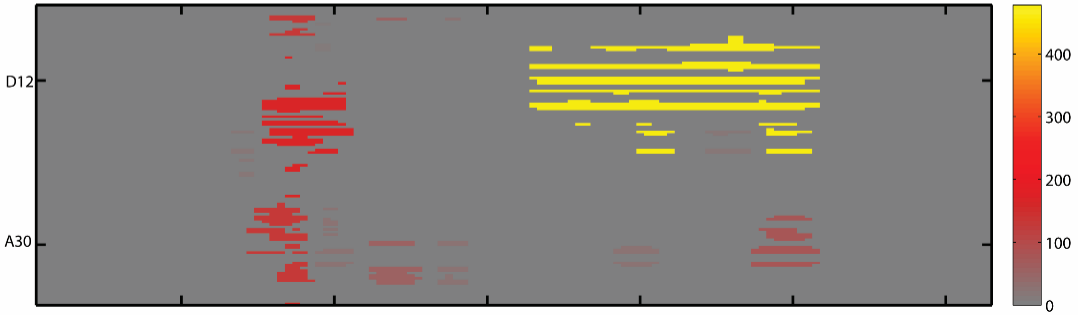
p values < 0.05

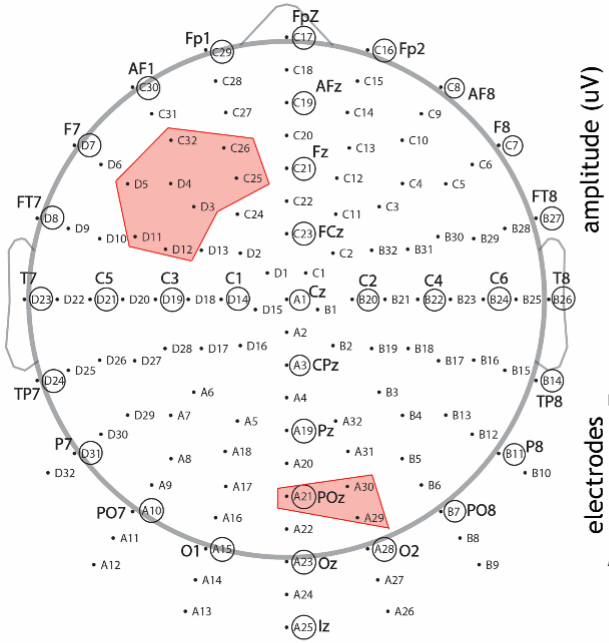
maximum extent
= number of
electrodes and
time points

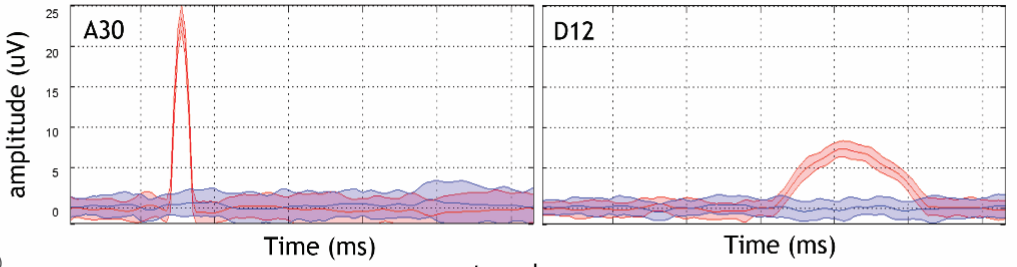
cluster 1 = 478
cluster 2 = 127

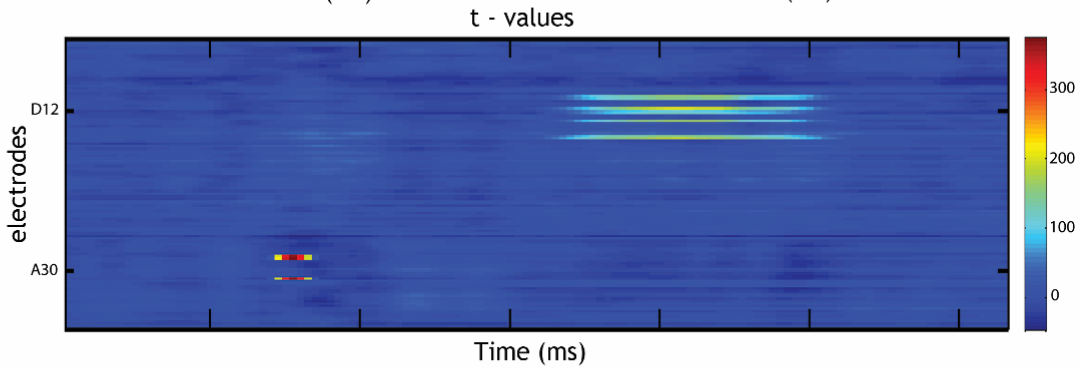


Cluster extent

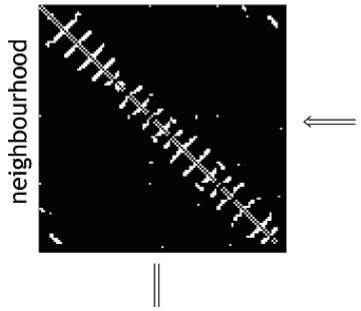






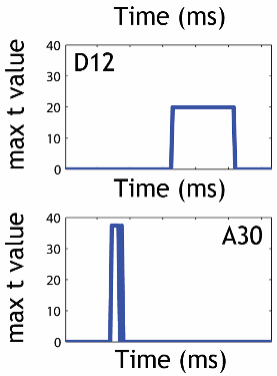


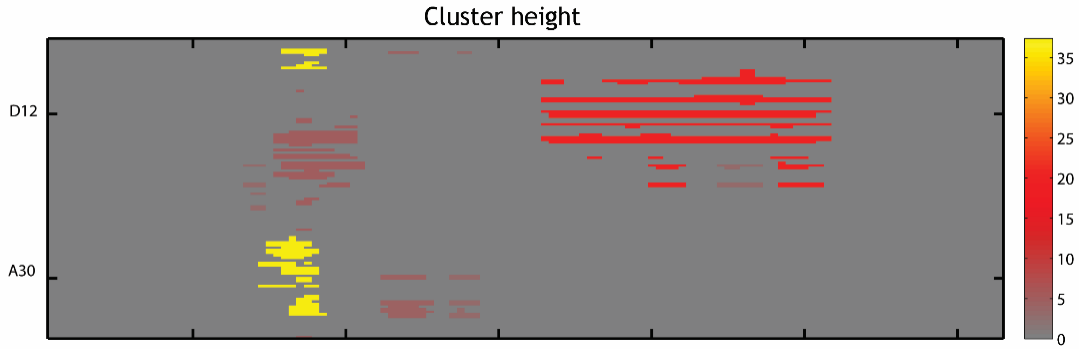
Spatial - Temporal clustering

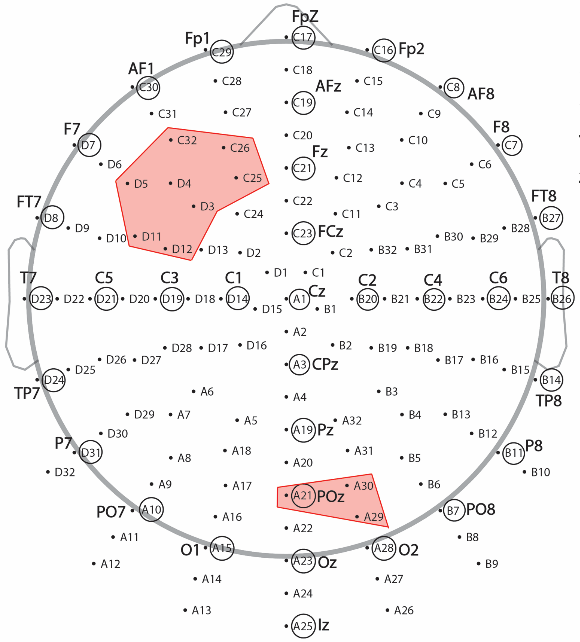


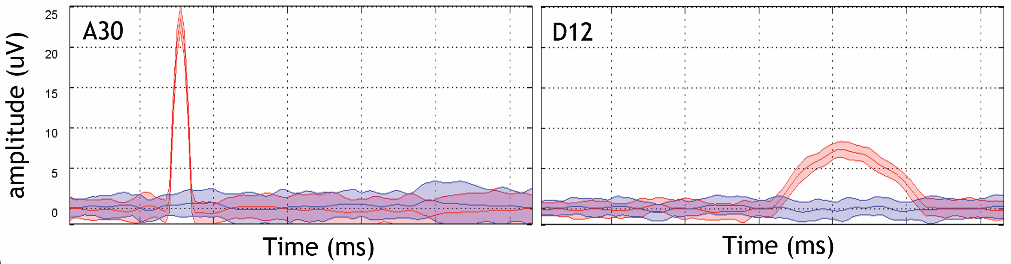
maximum height
within a cluster
of electrodes
and time points

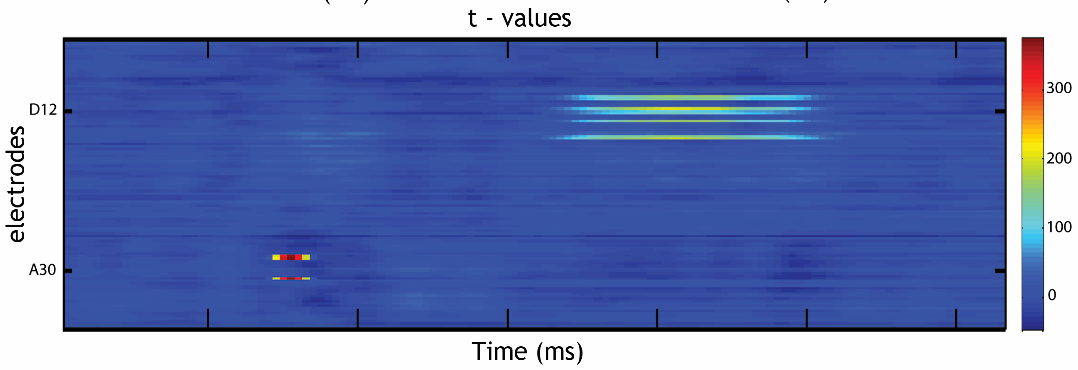
cluster 1 = 19.7
cluster 2 = 37.4



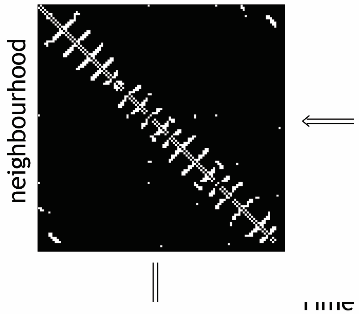






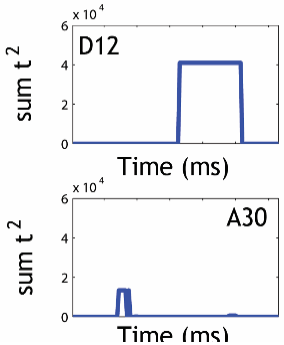


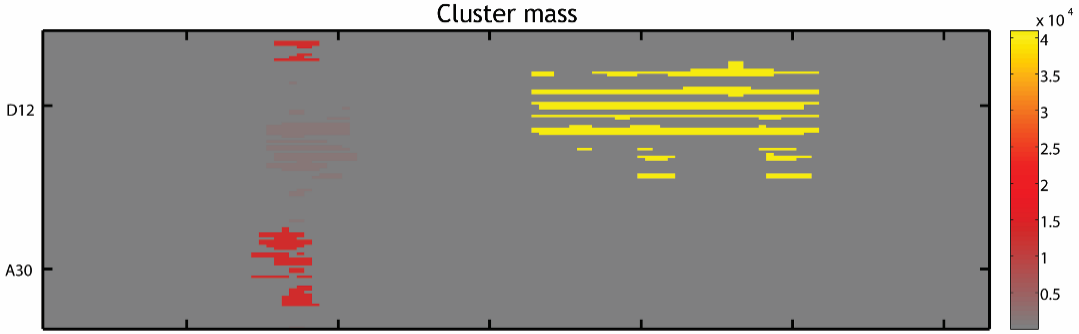
Spatial - Temporal clustering



mass (sum t^2) of values within a cluster of electrodes and time points

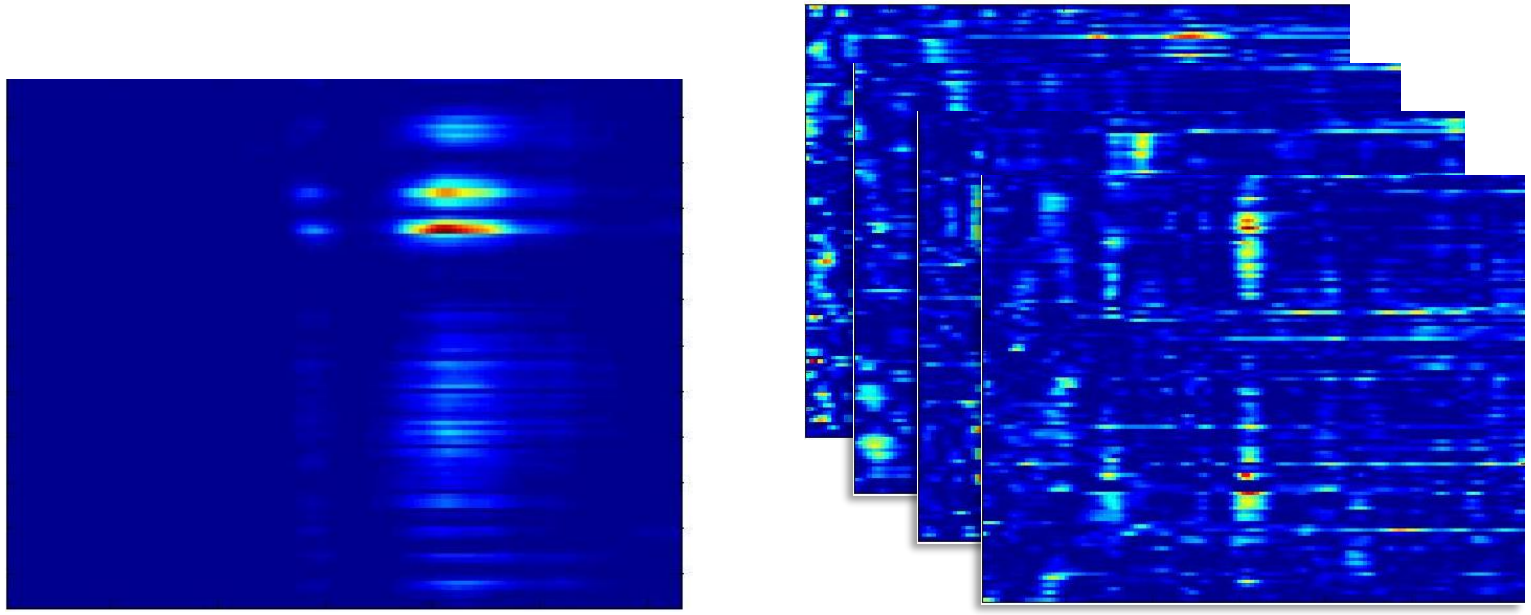
cluster 1 = 40984
cluster 2 = 13386





The clustering solution

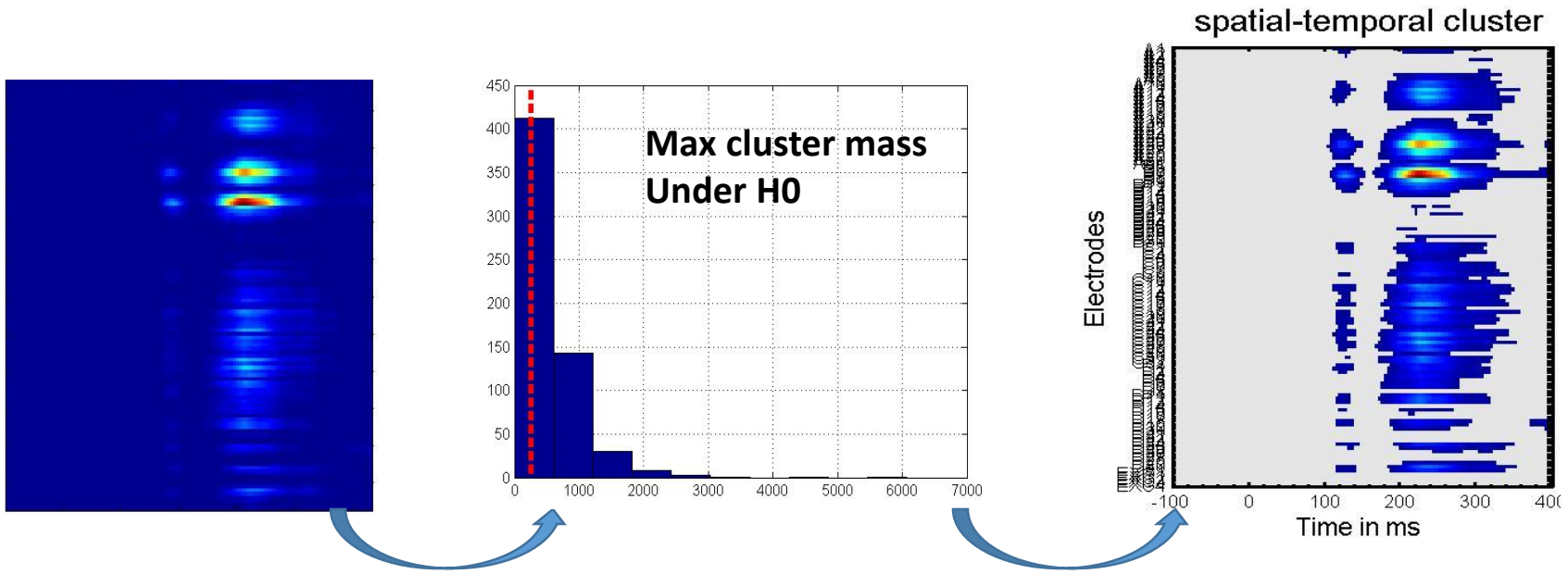
- In LIMO EEG, we bootstrap the data under H0: center the data or break the link between the design matrix and the data and then resample and test. This way we can find u for a single bin, the the whole space, or for clusters.



Observed F values → F values under H0

The clustering solution

- Spatial-Temporal clustering: for each bootstrap, threshold at alpha and record the max(cluster mass), i.e. sum of F values within a cluster. Then threshold the observed clusters based on there mass using this distribution → accounts for correlations in space and time.



Loss of resolution: inference is about the cluster, not max in time or a specific electrode !

Threshold Free Cluster Enhancement

- Threshold Free Cluster Enhancement (TFCE): Integrate the cluster mass at multiple thresholds. A TFCE score is thus obtain per cell but the value is a weighted function of the statistics by it's belonging to a cluster.

Smith & Nichols 2009 NeuroImage 44

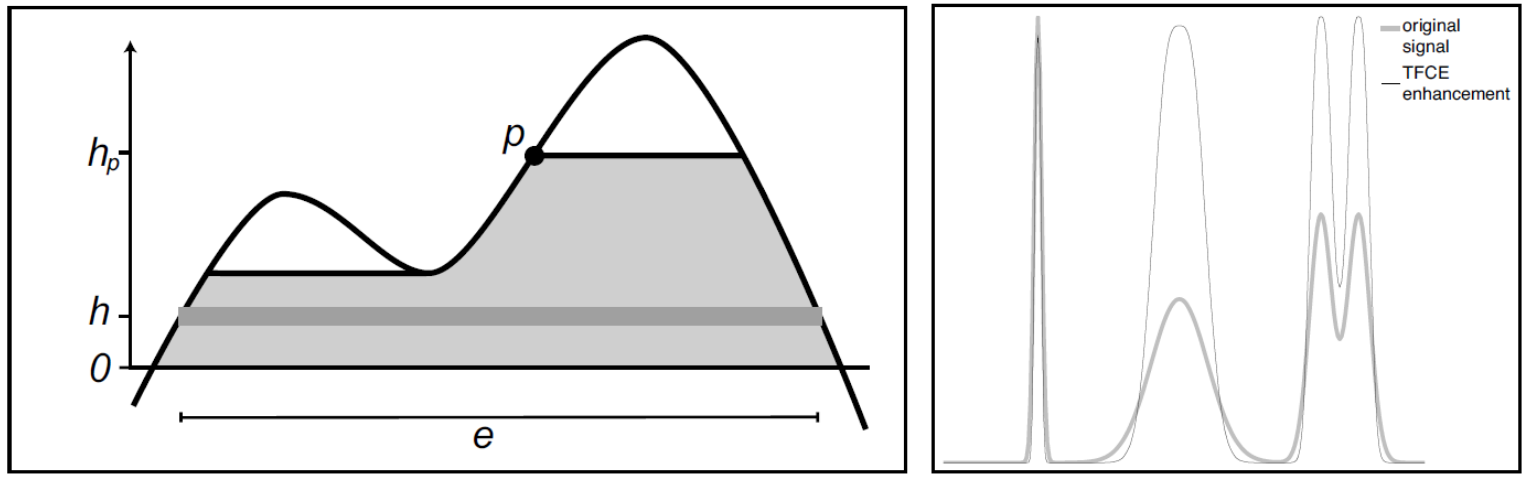
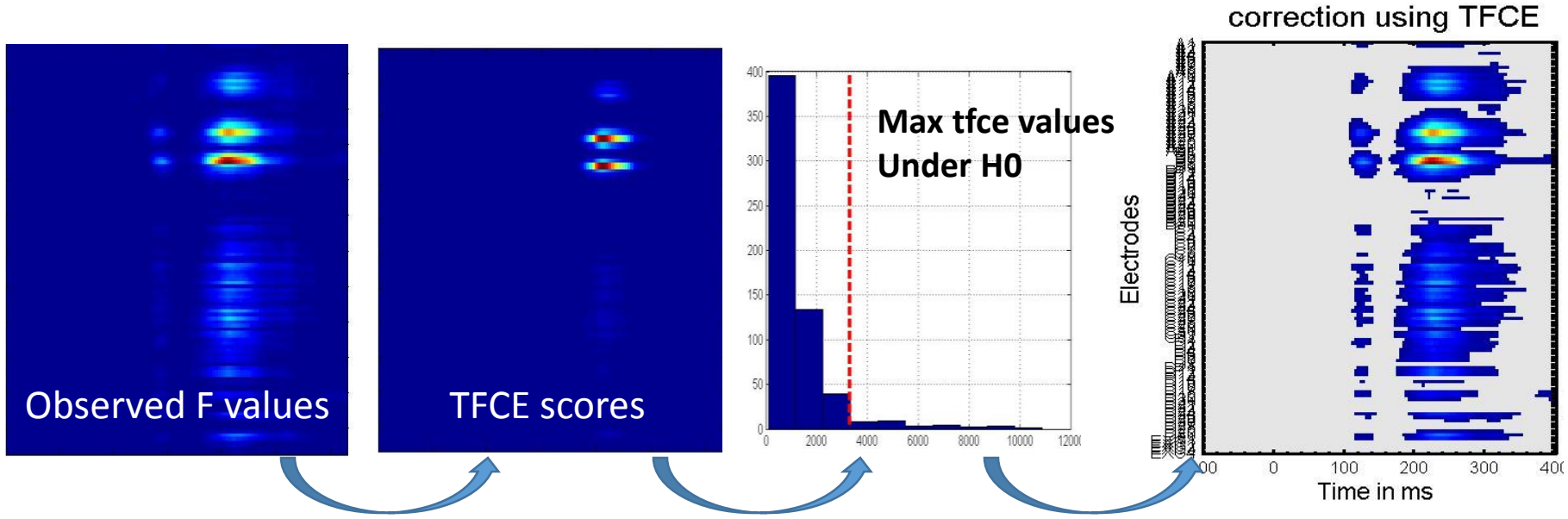


Figure 1: Illustration of the TFCE approach. Left: The TFCE score at voxel p is given by the sum of the scores of all incremental supporting sections (one such is shown as the dark grey band) within the area of "support" of p (light grey). The score for each section is a simple function of its height h and extent e. Right: Example input image and TFCE-enhanced output. The input contains a focal, high signal, a much more spatially extended, lower, signal and a pair of overlapping signals of intermediate extent and height. The TFCE output has the same maximal values for all three cases, and preserves the distinct local maxima in the third case.

Threshold Free Cluster Enhancement

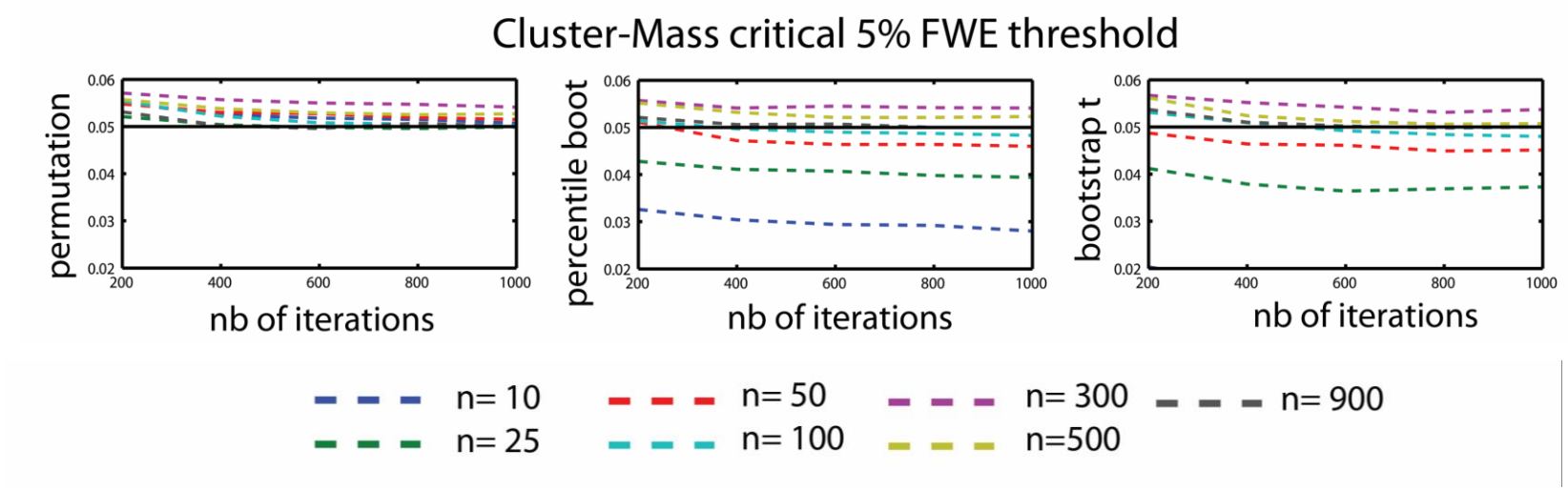
- Threshold Free Cluster Enhancement (TFCE): Integrate the cluster mass at multiple thresholds. A TFCE score is thus obtain per cell but the value is a weighted function of the statistics by it's belonging to a cluster. As before, bootstrap under H0 and get max(tfce).



Excellent resolution: inference is about cells, but we accounted for space/time dependence

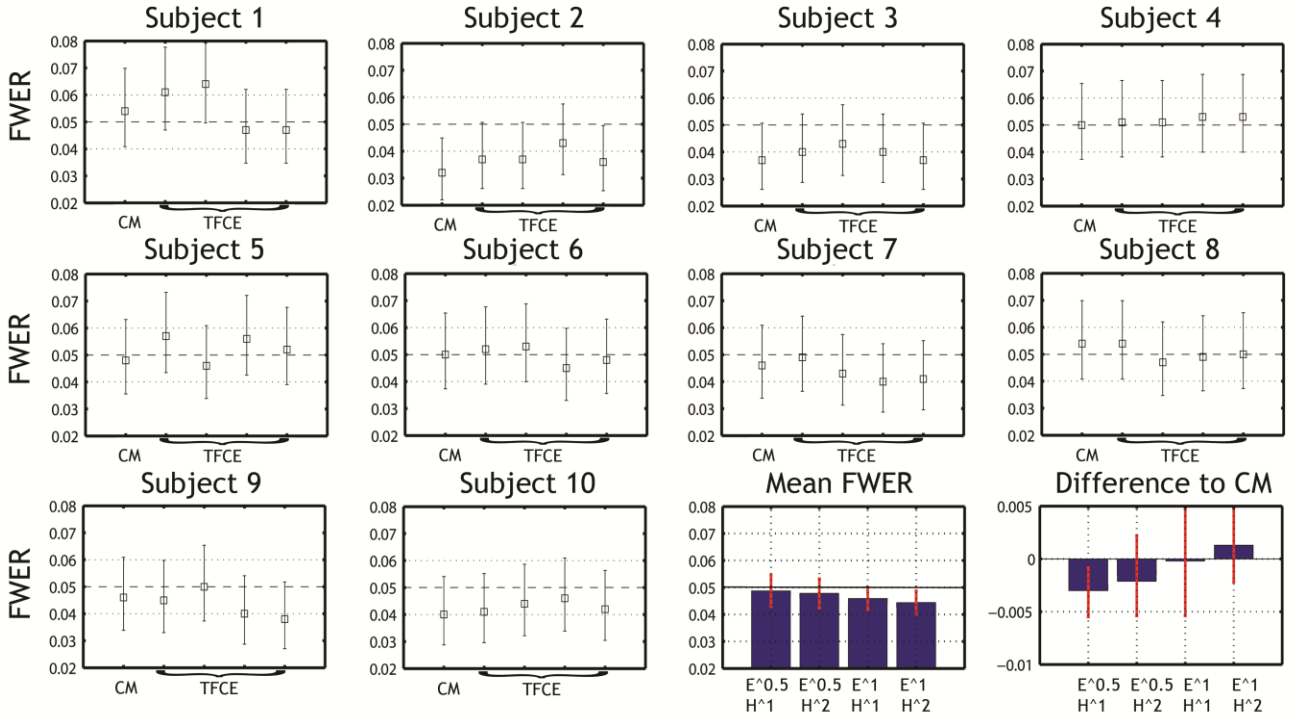
Review of techniques

- All techniques (including permutation not shown here) control well the FWER under H0 with some limitations for small sample sizes



Review of techniques

- All techniques (including permutation not shown here) control well the FWER under H0 with some limitations for small sample sizes



MCC summary

- Simulation work show that overall permutation / bootstrap / cluster-mass / TFCE control well the type 1 FWER.
- a minimum of 800 iterations are necessary to obtain stable results
- for low critical family-wise error rates (e.g. p = 1%), permutations can be too liberal;
- For within subject bootstrap, a min of 50 trials per condition is requested at the risk to be too conservative