UNIVERSITY OF SOUTHAMPTON

Optimising Objective Detection Methods for the Auditory Brainstem Response

by

Michael A. Chesnaye

A thesis submitted in partial fullfillment for the degree of Doctor of Philosophy

in the

Faculty of Engineering and the Environment

Institute of Sound and Vibration Research

Abstract

The transient Auditory Brainstem Response (ABR) is a change in neural activity along the auditory pathway in response to a brief acoustic stimulus. It is typically recorded non-invasively using electroencephalography (EEG), and has become an important diagnostic tool in the clinic, e.g. for diagnosing various neurological disorders and hearing screening in new borns. Detecting the ABR, however, can be a challenging task, and is still strongly dependent on highly trained individuals who are given the task to visually examine the EEG data. Besides incurring additional training costs, visual inspection has limitations in terms of specificity and sensitivity. Consequently, test time for some ABR examinations can be quite extensive, and information is often incomplete. This can have significant clinical implications, and has a large impact on the parents/carers of the infants.

The limitations associated with visual inspection have led to the development of many different objective measures for assisting the examiner during the visual inspection task, and improving the reliability and efficiency of the test. The overall goal for this thesis is to further improve the reliability and efficiency of ABR examinations by improving the specificity, sensitivity, and test time of objective ABR detection methods. To achieve this, the focus is firstly on the objective ABR detection methods themselves, i.e. on exploring, evaluating, optimising, and comparing new and existing detection methods, with the goal to find or develop methods with a high sensitivity, a low test time, and a controlled specificity. Important elements within this analysis include the assumptions underlying the detection methods, along with the adopted test and pre-processing parameters. Results demonstrate that the main concern for specificity is the independence assumption between epochs, which is violated as a function of the stimulus rate and the filter's high-pass cut-off frequency. The best performing method in terms of sensitivity and test time was furthermore a new bootstrapped statistic, consisting of a combination of the Hotelling's T^2 test and a correlation coefficient.

A second route in this thesis for improving the performance of objective ABR detection methods is through the development and optimisation of a new sequential testing framework for ABR detection methods. The approach, called the Convolutional Group Sequential Test (or CGST), controls the specificity of sequentially applied statistical tests, and permits data-driven adaptations (using previously analysed data) to test pa-

rameters following each stage of the sequential analysis. This allows the statistical analysis to be tailored specifically to the subject and recording in question, which offers new opportunities to speed up testing with high statistical power and controlled specificity. Results demonstrate relatively large reductions in test time when compared to a 'single shot' test where the detection method is applied to the data just once.

A final route in this thesis for improving the performance of objective detection methods is through a new adaptive ensemble size re-estimation procedure, integrated within the sequential testing framework. Besides further reductions in test time (relative to non-adaptive sequential test procedures), the adaptive approach can help bring ABR examinations to an unambiguous test outcome in terms of 'ABR present' or 'ABR absent or abnormal'.

Overview

The thesis starts with a brief introduction and background on the auditory brainstem response (ABR), and describes the standard model underlying almost all objective ABR detection methods. A summary of the main findings and contributions from this thesis are is presented in Chapter 1 (section 1.2). Following the introduction, a review of the literature is presented (Chapter 2) where the focus is on some of the more widely used ABR detection methods. A description of the methods and data used throughout this thesis is then presented in Chapters 3 and 4, respectively, after which an in-depth exploratory analysis of the specificity of objective ABR detection methods is presented in Chapter 5. In particular, Chapter 5 explores the main statistical assumptions underlying ABR detection methods, with the goal to identify (and potentially compensate, remove, or modify) test parameters or pre-processing strategies that contribute towards a poor control of specificity. Following the specificity assessment is a sensitivity and test time assessment (Chapter 6). The focus here is on evaluating and comparing new and existing objective detection methods, with the overall goal of finding the most sensitive method, with the shortest test time for some fixed specificity. In the second half of the thesis (Chapters 7, 8, and 9), the focus is on sequential testing for ABR detection. Chapter 7 first presents a brief literature review on sequential test procedures (section 7.1), and introduces a novel method for controlling the FPR of sequentially applied statistical tests, called the Convolutional Group Sequential Test (CGST; section 7.2). The performance of the CGST is then explored and optimised for ABR detection in Chapter 8. Finally, Chapter 9 integrates a new, online sample size re-estimation procedure within the sequential testing framework, and briefly evaluates the performance of the approach using simulations. The thesis then ends with an overview of the main conclusions and some directions for future work in Chapter 10.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

ABR	Auditory brainstem response
AR	Autoregressive
ASSR	Auditory steady state response
b_i	The ith autoregressive parameter
CAEP	Cortical auditory evoked potential
CC	Correlation coefficient
CGST	Convolutional Group Sequential Test
CSM	Component synchrony measure
CLT	Central Limit Theorem
CIs	Confidence intervals
dB	Decibel
DOF	Degrees of freedom
EEG	Electroencephalography
FFT	Fast Fourier Transform
Fmp	F for multiple points
FNR	False-negative rate
FPP	Fitted parametric peaks
Fsp	F for a single point
GG	Greenhouse Geisser correction for sphericity violations

H_0	The null hypothesis
H_1	The alternative hypothesis
HF	Huyn Feldt correction for sphericity violations
Hz	Hertz
i.i.d	Independently and identically distributed
Modified q-sample V2	A modification to the q-sample uniform scores test, applied to the ranks of the phases and amplitudes of the Fourier components of multiple spectral bands
Modified q-sample V4	A modification to the q-sample uniform scores test, applied to the actual values of the phases and amplitudes of the Fourier components of multiple spectral bands
ms	Millisecond
MSC	Magnitude squared coherence
MSSB	Mean sum of squares between
MSSE	Mean sum of squares error
MSSW	Mean sum of squares within
MVN	Multivariate normal
PDF	Probability density function
O	The order of the autoregressive model
PSD	Power spectral density
PSM	Phase synchrony measure
RM ANOVA	Repeated Measures Analysis of Variance
SDR	Standard deviation ratio
SFT	Spectral F-test
SL	Sensation level
SNR	Signal-to-noise ratio
SP	Single point
SPL	Sound pressure level
SPRT	Sequential Probability Ratio Test

T2 Freq	The Hotelling's T^2 test, applied in the frequency domain
T2 Time	The Hotelling's T^2 test, applied in the time domain
T2 Time + CC	A bootstrapped statistic, consisting of the Hotelling's T^2 test (applied in the time domain) and a correlation coefficient
T2 RM	The Hotelling's T^2 test, applied in the time domain as a repeated measures approach
T2C	Circular T^2 test
TPR	True-positive rate
TVMs	Time-voltage means
$y(t)$	The measured voltage in EEG recording y at time point t

Greek

α	The nominal significance level of the test
α_i	The stage i type-I error rate
β_i	The fraction of tests to be rejected for futility at stage i
β_{R_i}	The available β that can be spent at stage i for the CGST
γ	Statistical power
$\hat{\gamma}_i$	The estimated statistical power at stage i
γ_{T_i}	The total estimated statistical power across stages 1 to i
δ	The non-centrality parameter for the non-central F-distribution
$\hat{\delta}$	The estimated non-centrality parameter for the non-central F-distribution
$\Delta\alpha$	The rate at which α is varied from 0 to 1
ΔFx	The rate at which the F-value is varied from 0 to $FMax$
$\tilde{\epsilon}$	The Huyn Feldt correction factor
$\hat{\epsilon}$	The Greenhouse Geiser correction factor
λ	The true effect size

μV	Micro Voltage
$\boldsymbol{\mu_0}$	A vector containing a set of hypothesized values to test against
pi	3.14159265359...
σ	Standard deviation
σ_e^2	Error score variance
σ_t^2	True score variance
$\sigma_{\bar{T}}$	Standard deviation calculated from some template $\bar{T}$
$\sigma_{\bar{X}}$	Standard deviation calculated from the ensemble coherent average
Σ_k	The stage k summary statistic for the CGST
Σ	Either summation, or the true population covariance matrix
τ	Time shift
ϕ_i	The stage i null distribution for transformed p value p_i for the CGST
ϕ_i^0	The null distribution for transformed p value p_i for the CGST
ϕ_i^1	The alternative distribution for transformed p value p_i for the CGST
$\phi^{T[B_i,A_i]}$	Distribution ϕ, truncated to the $[B_i, A_i]$ interval
ϕ_{Σ_i}	The stage i null distribution for stage i summary statistic Σ_i for the CGST
$\phi_{\Sigma_i}^0$	The null distribution for the stage i summary statistic Σ_i for the CGST
$\phi_{\Sigma_i}^1$	The alternative distribution for the stage i summary statistic Σ_i for the CGST
Φ_{Σ_k}	The cumulative distribution function for stage k summary statistic Σ_k for the CGST
$\chi_{v_i}^2$	A chi-squared distribution with v_i degrees of freedom
Ω	Electrical resistance

Roman

$\boldsymbol{A}$	A contrast matrix
A_i	The stage i critical boundary for efficacy for the CGST
AR_k	The total area under the distribution in question after the stage k truncations for the CGST
A_{v_Σ}	The critical boundary for the class of self-designing tests in Hartung & Knapp (REF) when using degrees of freedom v_Σ
B_i	The stage$i$ critical boundary for futility for the CGST
$\boldsymbol{C}$	A scaling matrix
c_1	A constant used for specifying the steepness of a cosine ramp
c_2	A constant used for specifying the steepness of the exponential ramp
$c3$	A scaling factor to account for uncertainty within $\boldsymbol{S_2}$
D	A matrix of data points
$f_{cos}(x)$	A cosine ramp function, defined on the 1.5π to 2π interval (x is restricted to this interval)
$f_{exp}(x)$	An exponential ramp function, defined on the 0 c_2 interval (x is restricted to this interval)
$f_i(p_i)$	The stage i transformation function for p value p_i for the CGST
$FMax$	The upper limit for the x-axis of a simulated F-distribution
F_{v_1,v_2}	An F-distribution with v_1 and v_2 degrees of freedom
$F^{-1}(.)$	The inverse of the central F-distribution
F_{nc}	The cumulative distribution function for the non-central F-distribution
H	The number of columns (of D) to include for the Fmp
N	The ensemble size
N_i	The ensemble size for stage i
P_{Global}	The global p value for the CGST

R	A reliability coefficient
$\boldsymbol{R}$	A rotation matrix
$\boldsymbol{S}$	A feature covariance matrix
$\boldsymbol{S2}$	A feature covariance matrix, estimated from inter-epoch intervals
$\mid \boldsymbol{S_{Max}} \mid$	The largest covariance matrix (as determined by the determinant) from a population of resampled covariance matrices
$\bar{T}$	Some ABR template
V	A matrix of feature values
v_i	The degrees of freedom for the probability density function in question
W	The number of spectral bands to include for some ABR detection method
$W(\Sigma,\, N-Q)$	A Wishart distribution with covariance structure Σ and $N-Q$ degrees of freedom
$\bar{\boldsymbol{x}}$	A vector containing the mean feature values
$\bar{X}$	The ensemble coherent average
$\boldsymbol{x_{min}}$	A vector containing the minimum *a priori* assumed feature means

Acknowledgements

I would first and foremost like to acknowledge my supervisors, David M. Simpson, Steven L. Bell, and James M. Harte. David, thank you for digging me out of all the rabbit holes, and keeping me rolling down the right track. I am grateful for your many suggestions and ideas, which have formed the backbone of this thesis. Needless to say, this work would not have been possible without you. Steve, thank you for your kind supervision, the timely feedback, and the many discussions and meetings, which have often led to new insights and ideas. James, thank you for the many discussions and your insightful feedback on reports and publications. Your presence and connections with the industry have been an inspiration.

I would also like to give thanks to Jordan Cheer, for reviewing my 18-month report, and to give a special thanks to Thomas Blumensath for reviewing both my 9- and my 18-month report. I am also grateful towards Debbie Cane and Sarah M.K. Madsen for collecting the subject recorded ABR data and the recordings of EEG background noise, respectively, and towards Anisa Visram, Jo Brooks, Kevin Munro, Michael Stone, Helen Whiston, and Sara Al-Hanbali for collecting the cortical data. Finally, I would like to acknowledge Jing Lv, whos previous work has provided the necessary foundation for many sections throughout this thesis.

I am of course also grateful to my family and friends; for their love and support.

This work was supported by the Oticon Fonden and the Engineering and Physical Sciences Research Council ([EPSRC, grant No. EP/M026728/1)].

Chapter 1

Introduction

Transient auditory brainstem responses (ABRs) are defined as short changes in neural activity along the auditory pathway in response to a brief acoustic stimulus, such as a click, chirp or tone burst. Typically recorded non-invasively using surface mounted electrodes, they are used primarily for diagnosing abnormalities within the auditory system, such as hearing loss (e.g. Stevens et al, 2013) and various neurological disorders (e.g. Robinson & Rudge, 1980). Usually, the first step for these applications is to determine whether an ABR is present or not, after which additional analysis can be performed on, for example, the morphology of the response.

The main focus for this thesis is on detecting the ABR using objective detection methods. Before turning to objective detection methods, it is worth noting that ABR detection has historically, and continues to be, realised through visual inspection, i.e. by manually inspecting the acquired EEG data. Although potentially quite sensitive (Arnold, 1985), visual inspection has been found to vary substantially between examiners (Vidler & Parker, 2004). As a result, the false-positive and false-negative rates (further defined below) are also dependent on the examiner, which makes quantifying these properties problematic (Don & Elberling, 1996). The process of visual inspection thus introduces a variable and subjective element to what could potentially be a consistent, reliable, and objective measure.

Many researchers have therefore turned to more objective methods for detecting the ABR, i.e. methods with a firm foundation in statistics, capable of producing consistent and highly sensitivity measures for the presence or absence of a response. The primary goal for these methods is still to assist the examiner during the visual inspection task, and, in particular, to improve the reliability of the test, and reduce the required time for response detection. It is can also be envisioned that an objective detection method with a sufficiently high performance would allow examination to be carried out by staff without specialist training, or, in some applications, may allow the human observer to be removed entirely. Either way, this places high requirements on the performance of objective detection methods in terms of specificity, sensitivity, and test time (the

required time for detecting a response), which can be considered as their three most important properties.

With respect to specificity (see also Chapter 5), this is directly related to the false-positive-rate (FPR), also known as the type-I error rate, i.e the rate at which the null hypothesis H_0 of 'no ABR present' is incorrectly rejected. In other words, a false-positive is when it is concluded that a response is present, when there is, in fact, just noise. Specificity is furthermore controlled through the nominal α-level of the test, which is the theoretical or assumed FPR. In practice, deviations from α can occur due to random fluctuations, or due to a violation to the statistical assumptions underlying the detection method. When the observed FPR is larger than α, the test is called liberal, whereas when the observed FPR is smaller than α, it is called conservative.

For many ABR-related applications, it is generally accepted that a conservative test performance is less detrimental than a liberal one. In ABR hearing screening applications, for example, a higher than expected FPR can result in cases of undetected hearing loss (it is incorrectly concluded that the subject heard the acoustic stimulus), which, when left untreated, have been associated with an impaired language development in children (Ramkalawan & Davis, 1992), along with various other more obvious handicaps, such as discrimination, less effective education, a reduced life expectancy and higher unemployment rates (Miziara et al., 2012), to name a few. With respect to a conservative test performance, although this is less detrimental than a liberal one, it is still far from desirable. In particular, a conservative test performance tends to result in a reduced statistical power, and consequently in a prolonged test time (i.e. the reduced statistical power needs to be compensated for by increasing the sample size). In short, the FPR of the test should ideally be controlled as intended, that is, it should match with the nominal α-level of the test.

With respect to sensitivity, this is the detection rate of the test, and is directly related to the false-negative rate (FNR), also known as the type-II error rate, i.e. the rate at which H_0 is incorrectly accepted. In other words, a false-negative is when it is concluded that a response is absent, when a response is, in fact, present. Sensitivity is furthermore closely related to test time, as a more sensitive test will tend to detect the response sooner. Ideally, sensitivity should be as high as possible for some set type-I error rate, and test time as low as possible for some set type-I and type-II error rate. In ABR audiometry, for example, one would expect an increased sensitivity to allow the detection of lower signal-to-noise rate (SNR) responses (evoked by lower amplitude acoustic stimuli), which may lead to greater convergence between behavioural and estimated hearing thresholds. A relatively low type-II error rate is also important when fitting hearing aids, as a type-II error can potentially result in the hearing aid being fitted too loudly. With respect to a reduced test time, this is desirable as some ABR examinations are currently quite long, e.g. it can take well over an hour to measure hearing thresholds in both ears for a range of frequencies. This is not ideal, particularly so for new born hearing screenings as the infant may become restless and introduce movement artefacts to the

EEG measurements. A reduced test time is of course also desirable as available resources are limited, and due to a reduced cost of service delivery.

The main goal for this thesis is to improve the performance of ABR detection methods in terms of specificity, sensitivity, and test time. There are, broadly speaking, two routes through which this might be achieved: (1) increasing the SNR of the ABR, or (2) improving the performance of the objective detection method. Starting with the SNR, a typical ABR has a duration of approximately 15 ms following the onset of an acoustic stimulus, with a peak amplitude of around 0.5 μV (Hall, 2006, p.95). This is in contrast to the noise (also known as the EEG background activity), which consists of a conflux of unwanted potentials with amplitudes typically in the range of at least 10 μV after filtering. The EEG background activity can furthermore originate from a wide range of sources, of which the most common include muscle artefacts (due to e.g. moving, blinking, breathing, and heart beats), along with electro-magnetic interference from power lines, lighting, and a wide range of electronic equipment in general. A first step to increase the SNR is hence to remove the source of the EEG background activity, e.g. by switching off unnecessary electronic equipment. Many noise sources, such as breathing and heart beats, are of course unavoidable (assuming no drastic measures are taken), and suitable artefact rejection and pre-processing strategies (e.g. differential amplification, filtering, and signal averaging, to name a few) are considered common practice. Alternatively, the SNR might be improved by increasing the amplitude of the ABR. The most obvious way to do so is through the amplitude of the acoustic stimulus, which tends to be positively correlated with the amplitude of the evoked response (see e.g. Picton & Hillyard, 1974; Starr & Achor, 1975). Other stimulus parameters that affect the ABR include the stimulus rate (the rate at which the stimulus is presented to the subject; Don et al., 1977; Fowler & Noffsinger, 1983), and the type of stimulus in general (Hood, 1998; Elberling et al., 2010), e.g. the stimulus might be a click or a chirp, or it might be spectrally complex or simple, have a short or long duration, etc. Various additional factors that can affect the ABR include whether the stimulus is presented monaurally or binaurally (e.g. Blegvad, 1975), the age (Fria & Doyle, 1984) and gender (Allison et al., 1983; Jerger & Hall, 1980; Michalewski et al., 1980; Don et al., 1994) of the subject, and potentially even body temperature (Jones et al., 1980) or some forms of medication (Hood, 1998). The (observed) ABR is also affected by the electrode placement (Moore, 1977; Mizrahi et al., 1983). Factors that do not affect the ABR that are worth mentioning include sleep (Jewett & Williston, 1971), coma (Starr et al., 1977), and attention (Picton & Hillyard, 1974). Finally, a second route for improving the performance of ABR detection methods, and coincidentally the main focus for this thesis, is through an improved design and application of the objective detection method.

1.1 The standard model for objective ABR detection methods

The conventional model underlying almost all ABR detection methods is built around the following three basic assumptions: (i) the evoked response is deterministic (it does not change over time) within the recoding session, (ii) the evoked response is independent of the EEG background activity, and (iii) the EEG background activity is a stationary, random, ergodic process. The observed signal following the onset of each stimulus can then be described as (see e.g. Elberling & Don, 1984; Raz et al, 1988):

$$x(t) = ABR(t) + Noise(t) \tag{1.1}$$

Where $x(t)$ is the observed voltage measurement at time t following stimulus onset, $ABR(t)$ is the true value of the evoked response at time point t, and $Noise(t)$ is the observed value of the EEG background activity, similarly at time point t. The time windows following the onsets of the stimuli are typically referred to as 'epochs'. By presenting many stimuli to the subject, an ensemble of epochs are collected, which are pre-processed and analysed using the adopted statistical detection method. Note that each epoch is 'time-locked' to an acoustic stimulus.

Many ABR detection methods also make use (either implicitly or explicitly) of the coherently averaged epoch $\bar{X}(t)$, which (following the above notation) is given by:

$$\bar{X}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) = ABR(t) + \overline{Noise}(t) \tag{1.2}$$

Where N is the ensemble size, $x_i(t)$ is the observed value for epoch i at time point t, and $\overline{Noise}(t)$ is the coherently averaged background noise at time point t. Note that $ABR(t)$ remains constant within the coherent average, as ensemble coherent averaging does not affect deterministic signals. When the samples between epochs are independent and normally distributed, then the power of the background noise within the coherent average (called the residual background noise) is decreased, on average, by $\sqrt{N}$.

It is generally accepted that the aforementioned assumptions are not entirely true for EEG measurements, most notably so for the stationarity assumption, e.g. the variance of the EEG background activity can vary significantly within recordings, which coincidentally brings the normality assumption into question. With respect to independence, the spectral content of EEG measurements introduces correlations between samples, which jeopardises the assumed independence between epochs. Finally, with respect to

the assumption that the ABR is a deterministic signal, this is generally true (Salamy & McKean, 1977), although short term depression of the ABR has been observed for fast stimulus rates (Salamy et al., 1975; Terkildsen et al., 1975; Thornton & Coleman, 1975).

Although the stationarity, normality and independence assumptions are not always satisfied for EEG measurements, it is worth pointing out that the standard model does not necessarily break down when these assumptions are violated, i.e. the main adverse effect is that the residual power in the ensemble coherent average is no longer decreased by $\sqrt{N}$. When using visual inspection for ABR detection, it might therefore take more time to reach an unambiguous decision in terms of ABR present or absent. It might also be pointed out that it is also not clear how violations to the aforementioned assumptions might affect the specificity and sensitivity of ABR examinations when using visual inspection for response detection (as noted earlier, quantifying the TPRs and FPRs for visual inspection is problematic). When using a formal parametric statistical test for ABR detection, on the other hand, then this is a different matter, i.e. it is well known that specificity (and consequently sensitivity and test time) is compromised when the aforementioned assumptions are violated, which then brings the reliability and efficiency of objective detection methods into question. The statistical assumptions underlying objective ABR detection methods play an important role throughout all Chapters in this thesis, and are the main focus for Chapter 5.

Finally, with a few exceptions (Stürzebecher et al., 2005; Cebulla & Stürzebecher, 2013; Stürzebecher & Cebulla, 2015), objective ABR detection methods are designed under the assumption of a 'single shot' application, i.e. it is assumed that they are applied to the data just once. This is, however, not how ABR detection methods are used in practice. Instead, they tend to be used in conjunction with visual inspection, and are thus applied repeatedly to the accumulating data until a decision in terms of response present or response absent has been made. The advantage of doing so is that the higher SNR responses can be detected early (thus reducing test time), whereas the test can still be prolonged in the case of a lower SNR response. The caveat is that the probability of finding patterns in noise is increased with the number of interim 'looks' at the data (e.g. Armitage et al., 1969; Wassmer, 2000). The latter is also known as an 'inflated FPR', and adjusted critical boundaries (for rejecting or accepting H_0) are required in order to preserve the nominal α-level of the test. Controlling the FPR of sequentially applied statistical tests is the main focus for Chapter 7 in this thesis, which describes a new approach, called the Convolutional Group Sequential Test (CGST). When the assumptions underlying the CGST are satisfied, then the CGST also permits data-driven adaptations to test parameters, i.e. previously anaysed data can be used to optimise test parameters for the remaining stages of the sequential analysis. Data-driven adaptations are further explored in Chapter 9, which describes a new approach for choosing the ensemble size online as data becomes available. The main advantage for the adaptive approach is a reduced test time, along with an improved control over

both the TPR and the FNR.

1.2 Research hypotheses

This section presents an overview of the main research hypotheses that developed throughout the project.

The specificity of objective ABR detection methods

1. Pre-processing parameters and test settings (e.g. the filter cut-off frequencies, the stimulus rate, and the artefact rejection method) will partly determine the extent to which the assumptions underlying most objective detection methods are satisfied for ABR detection, which will, in turn, affect the specificity of the test.

2. A set of pre-processing strategies and test parameters can be found, such that the specificity of objective detection methods remains controlled as intended for ABR detection.

The sensitivity and test time of objective ABR detection methods

Exploratory research: *Optimise, evaluate, and compare the specificity, sensitivity and test time of various new and existing objective detection methods, such that a recommended detection method can be provided for ABR examinations.*

Sequential testing

1. The convolution theorem (Grinstead & Snell, 1997) can provide a foundation for the development of a new, flexible, and intuitive approach (the CGST) for controlling the FPR of sequentially applied statistical tests.

2. The CGST will allow the specificity of sequentially applied objective detection methods to be controlled for ABR detection, and will provide reductions in test time relative to a 'single shot' test.

An adaptive ensemble size re-estimation procedure

1. Ensemble size re-estimation using previously collected data will allow for a more informed decision regarding the required ensemble size for detecting the ABR. This will reduce test time, and give an improved control over the TPR.

1.3 Original contributions

Contributions are made towards improving the performance of objective methods for ABR detection in terms of specificity, sensitivity, and test time.

1.3.1 Specificity

This work firstly presents in-depth exploratory assessment of the specificity of single shot ABR detection methods (Chapter 5), with the overall goal of obtaining a more robust evaluation of test significance. In particular, the Chapter uses simulations and recordings of EEG background activity to isolate and explore the main statistical assumptions underlying ABR detection methods, which include the normality, the stationarity, and the independence assumption. Results demonstrate significant violations to the independence assumption (between epochs), as a function of the high-pass cut-off frequency f_c and the stimulus rate, i.e. specific combinations of f_c and the stimulus rate resulted in relatively large deviations from α, ranging from 0.0385 to 0.0985 for $\alpha = 0.05$. Significant violations to the normality and stationarity assumptions were also observed, which resulted in a tendency towards a conservative test performance. For some recordings, the normality and stationarity violations were relatively severe, giving maximum deviations of 0.0161 and 0.0335 (for $\alpha = 0.05$) for normality and stationarity violations, respectively, whereas for other recordings stationarity and normality were more or less satisfied. Finally, various methods and data transformations for removing or compensating for the aforementioned violations were explored, which include (i) bootstrapping in blocks for a more robust evaluation of test significance under independence violations, (ii) normalisation of the epoch variances for removing stationarity violations, and (iii) increasing the ensemble size and/or artefact rejection for compensating and removing normality violations. Further details and results are presented in Chapter 5.

1.3.2 Sensitivity and test time

Various new and existing ABR detection methods were evaluated and compared across a range of feature sets and test conditions (Chapter 6), with the overall goal of finding or designing an ABR detection method with good sensitivity and low test time, for some fixed type-I error rate. With respect to new methods, this work explores (1) the performance of the Hotelling's T^2 test in the time domain, where it is applied as either a standard multivariate approach, or as a multivariate approach for analysing repeated measurements (see section 3.2.4 for details), (2) the Repeated Measures Analysis of Variance test, using the Greenhouse Gesier and Huyn Feldt corrections for sphericity violations, and (3) a new bootstrapped statistic, consisting of a combination of the Hotelling's T^2 test and a correlation coefficient. The latter was designed using a modified bootstrap approach (section 3.6.2). The performance of these methods were evaluated

and compared to various existing methods, which include the Fsp and the Fmp (evaluated using either theoretical F-distributions or with the bootstrap approach), the bootstrapped max-difference and mean power statistics from Lv et al (2007), the q-sample uniform scores test and its modifications from Cebulla et al (2006), Friedman's test, and the Hotelling's T^2 test in the frequency domain. The main results firstly demonstrate a more robust control of specificity for the Fsp and Fmp when evaluating test significance with the bootstrap approach, as opposed to using theoretical F-distributions with assumed DOF. The improved specificity coincidentally resulted in an improved test sensitivity, e.g. for the Fsp, evaluating test significance with the bootstrap (as opposed to using theoretical F-distributions) resulted in a maximum increase in test sensitivity of ∼40% for the simulations, and ∼25% for the subject recorded data. With respect to the remaining methods, an overall sensitive and robust performance (across test conditions) for the Hotelling's T^2 test was observed. When compared to the Fsp (evaluated using theoretical F-distributions), maximum increases in test sensitivity of ∼60% were observed for the simulations, and ∼40% for the subject recorded data. Finally, the best performing method throughout this work was the new bootstrapped statistic, composed of the Hotelling's T^2 test and a correlation coefficient. When compared to the Fsp (evaluated using theoretical F-distributions), a maximum increase in test sensitivity of 70-75% was observed for the simulations, and ∼50% for the subject recorded data.

1.3.3 Sequential testing

A novel method (the CGST) for finding the stage-wise critical decision boundaries and controlling the overall type-I error rate for sequential testing is described in Chapter 7. Various connections with existing methods are also discussed. The main advantage for the CGST over some alternative methods is flexibility, ease of understanding, and low computational load. The specificity, sensitivity and test time of the CGST were assessed across a range of CGST design parameters for ABR detection (Chapter 8). In terms of specificity, results emphasize that care is required to ensure that the assumptions underlying the objective ABR detection method (used for analysing the data) are satisfied, else additional violations originating from the CGST might be introduced. With respect to sensitivity and test time, various trade-offs are demonstrated as a function of CGST design parameters, which include predominantly the number of stages used for the sequential analysis and the choice of the stage-wise critical decision boundaries. When compared to the single shot test (where the objective detection method is applied just once, i.e. not sequentially), results demonstrate reductions in *mean* test time (taken across a large number of tests) for the CGST of up to 40-45%, with no loss in statistical power. The latter came at the cost of an increased *maximum* test time, i.e. for some subjects test time was prolonged. The increased maximum test time also has consequences for the no-stimulus condition, i.e. when a response is absent, the test will proceed to the final stage of the trial in (1-α)100% of the cases. Test time for the no-stimulus condition was therefore prolonged (potentially by a factor of ∼250%). This

emphasizes the importance of futility stopping (early acceptance of the null hypothesis), not only for the CGST, but for sequential test procedures in general. Results indeed demonstrate significant reductions in mean test time for the no-stimulus condition when early stopping in favour of H_0 was permitted. Further details are presented in Chapter 8.

1.3.4 Adaptive ensemble size re-estimation

A new online ensemble size re-estimation procedure, integrated within a sequential testing framework, is proposed for ABR detection (chapter 9). The main advantage of the adaptive approach over a conventional non-adaptive approach (where the ensemble size is fixed in advance) is an improved control over both statistical power and the true-negative rate, along with a reduced test time. In other words, the approach can help bring ABR examinations to an unambiguous test outcome in terms of 'ABR present' or 'ABR absent (or abnormal)', whilst using as little test time as possible. Simulation results demonstrate a reduced test time of ∼10-30% for the stimulus condition and ∼25-45% for the no-stimulus condition.

1.3.5 Publications

Journals

Published

- Chesnaye M.A., Bell S.L., Harte J.M., & Simpson D.M. 2018. Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity and detection time. *International Journal of Audiology*, 57(6), pp. 468-478. DOI: https://doi.org/10.1080/14992027.2018.1447697
- Vanheusden F, Bell S.L., Chesnaye M.A., & Simpson D.M. (2018). Improved detection of vowel envelope frequency-following responses using Hotelling's T2 analysis. *Ear and Hearing.*

Submitted

- Chesnaye M.A., Bell S.L., Harte J.M., & Simpson D.M. (2018). The Convolutional Group Sequential Test: reducing test time for evoked potentials, *IEEE: transactions on bio-medical engineering.*
- Chesnaye M.A., Bell S.L., Harte J.M., & Simpson D.M. (2018). A Group Sequential Test for ABR detection, *International Journal of Audiology.*

- Vanheusden F, Bell S.L., Chesnaye M.A., & Simpson D.M. (2018). Envelope Frequency Following Responses Are Stronger For High-Pass Than Low-Pass Filtered Vowels, *International Journal of Audiology.*

Planned submissions

- Chesnaye M.A., Bell S.L., Harte J.M., & Simpson D.M. An adaptive, online ensemble size re-estimation procedure for ABR detection. *IEEE: transactions on bio-medical engineering.*

- Chesnaye M.A., Bell S.L., Harte J.M., & Simpson D.M. Non-parametric evaluations of test significance for ABR detection. *International Journal of Audiology.*

Conferences

- Chesnaye M.A., Bell S.L., Harte J.M., & Simpson D.M. (2107). A group sequential test strategy for objective auditory brainstem response detection methods. International Evoked Response Audiometry Study Group. Warshaw. *Oral presentation.*

Chapter 2

A review of objective ABR detection methods

This chapter presents a review of some of the more widely used ABR detection methods. The review starts in 1892 with a short story about Hans Berger (section 2.1), the original inventor of the electroencephalogram, after which some of the earliest methods for assisting clinicians and researchers during evoked response detection are presented (section 2.2). From the late 60s onwards, formal parametric statistical tests for ABR detection began to emerge (section 2.3), ultimately resulting in the well known Fsp and Fmp statistics, which are up to the current day still some of the most frequently used ABR detection methods. From 1976 onwards, researchers began to explore single-band frequency domain analysis for ABR detection (section 2.4), which was soon extended to multiple bands in order to cover the broadband spectral content of a typical ABR (section 2.5).

Besides giving an overview of some of the most well known ABR detection methods, the goal for this chapter is to provide justification for choosing or rejecting ABR detection methods for further evaluations and comparisons throughout this thesis. In particular, the review is used to make a selection of some of the best performing ABR detection methods (section 2.6), which are evaluated in terms of specificity, sensitivity, and test time in Chapters 5, 6, and the Appendix. Various new methods that have not yet been explored for ABR detection are also included. Finally, it should be stressed that this review is on some of the more frequently used and well known ABR detection methods. Various less conventional methods such as Neural Networks (Freeman, 1992; Alpsan & Özdamar, 1992a; Alpsan & Özdamar, 1992b; Habraken et al., 1993; Sánchez et al., 1995) and advanced denoising methods (e.g. wavelet de-noising; Popescu et al., 1999; Zhang et al., 2006) are excluded from the review.

2.1 The origin of EEG

The origin of the electroencephalogram dates back to a man falling from his horse one morning in Würzburg. It was during a military training exercise in 1892 when Hans Berger's horse reared, throwing Hans to the ground. Hans landed right in front of the wheel of a horse drawn artillery gun, which was luckily stopped by the driver in the nick of time, thus saving Hans from what would otherwise be certain death. That very evening, and for the first time ever, Hans received a telegram from his father enquiring about his health. He later discovered that it was his sister who, overwhelmed by a sudden sensation that something was wrong, had urged their father to contact him. Hans was convinced that the coincidence could not be accounted for by mere chance alone, and that some form of mental telepathy between him and his sister must have taken place. He hereby developed an interest in psychophysics, and set out on what would be a 40 year journey to find empirical evidence for psychic energies within the brain.

To cut a long story short, Hans was never successful in his quest for detecting psychic energies. However, his research ultimately led him to the discovery that neural potentials could be measured non-invasively from the scalp of a subject, i.e. the first electroencephalographic measurements. After numerous control studies, Berger published his findings in 1929 (Berger, 1929), and coined the observed potentials alpha and beta waves. Due to the German journals being inaccessible to many British and American researchers at the time, and otherwise due to scepticism from colleagues, Berger's findings initially went unnoticed (Millet, 2001), and it wasn't until after they were confirmed by Adrian and Mattews (1934) that his discovery gained recognition from the scientific community.

"Is it possible that I might fulfill the plan I have cherished for over 20 years and even still, to create a kind of brain mirror: the Elektrenkephalogramm!"

A quote from Hans Berger's diary following his discovery. From Millet, 2001

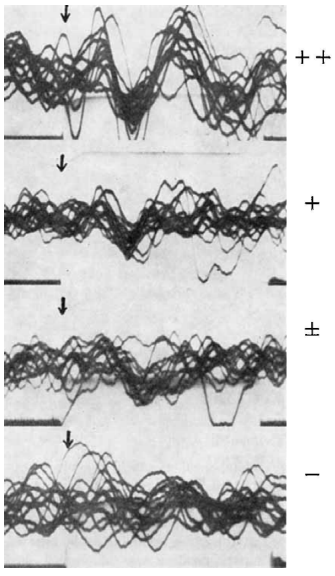
2.2 Early methods and averaging

Early research following Berger's discovery helped to further characterize his alpha and beta waves, and, in particular, to describe how they were affected by various factors such as sleep (Loomis et al., 1937; Davis et al., 1938), drugs (Berger, 1931, 1934, 1937), pathologies (Berger 1931), and different types of stimuli (Adrian and Matthews, 1934; Davis, 1939; Davis et al., 1939; Adrian, 1941). New potentials were also discovered, such as transient potentials induced by sound (Davis, 1939), sleep spindles (Loomis et al., 1935) and K-complexes (Loomis et al., 1937). These potentials were soon exploited by researchers in an attempt to develop the first EEG hearing screening programs (Marcus, Gibbs and Gibbs, 1949; Gidoll 1952; Perl et al., 1953; Derbyshire et al., 1956).

Techniques for assisting clinicians and researchers at the time were still unavailable, and these early experiments were dependent on the researchers ability to visualy detect patterns in the waveforms. Averaging techniques for reducing the EEG background noise and increasing the SNR were also not yet in practice, and the relatively low amplitude potentials were often lost in the fluctuating EEG background noise. This hence restricted applications to potentials with relatively large amplitudes, such as the K-complex.

One of the first averaging techniques, called photographic superposition, was developed by G.D. Dawson in 1947. The technique relied on photographing all responses following stimulus onset, and rephotographing the superimposed records as a single image so that deflections time-locked to the stimulus were easier to detect (Dawson, 1947, 1950). The technique was first applied to somatosensory evoked responses induced by electrically stimulating the peripheral nerve (Dawsen, 1947, 1950), and was later used in the visual and auditory domain for detecting high intensity flashes (Cobb and Morton, 1952) and continuous tones and clicks (Abe, 1954). Suzuki and Asawa (1957) furthermore applied it to tone stimuli of various intensities in an attempt to estimate behavioural hearing thresholds in a group of subjects (Fig. 2.1).

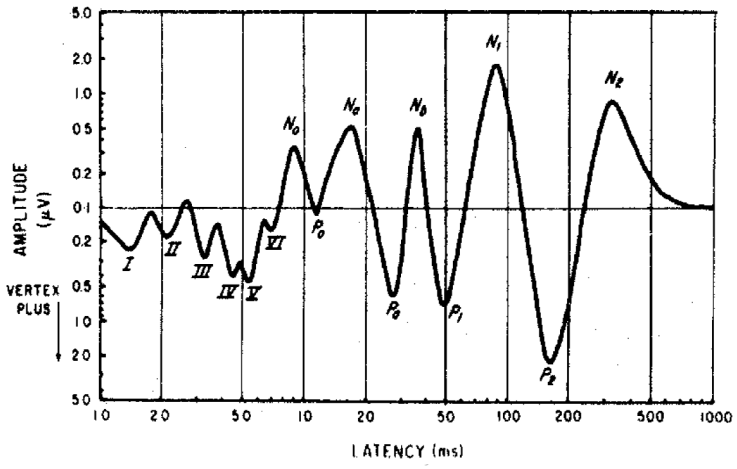
Figure 2.1: Four examples of ten superimposed responses following stimulus onset (indicated by the arrow) measured from the vertex. The first set (++) represents an example of a response that is strongly present, the second (+) a response that is not as strongly represent, the third (±) is inconclusive, and in the fourth (-) the response is considered to be absent. *Reprinted from Publication 'Evoked Potential of Waking Human Brain to Acoustic Stimuli: A Clinical Study on its Application to Objective Audiometry', Vol. 48(5-6), Suzuki T., Asawa I., pp. 508-515 (1957), with permission from Taylor & Francis.*



With the development of computers, Dawsen was able to extend his photographic superposition technique to the digital domain (Dawsen, 1951, 1954), which allowed a much finer grained analysis of the auditory response (Barlow and Brown, 1955; Clark et al., 1961). Small potentials hidden in relatively large fluctuations of EEG background noise (and indetectable to the human eye) could now be extracted, resulting in the identification and characterization of the auditory mid-latency (Geisler et al., 1958; Geisler, 1960;

Goldstein and Rodman, 1967) and auditory brain stem response (Jewett et al., 1970; Jewett & Williston, 1971). Moreover, the findings allowed the development of various evoked response detection methods based on peak amplitudes and latencies (Schimmel et al., 1974; Morley and Liedke, 1977; Aunon, 1978), (for a coherent overview of the identified waveforms and their peaks, see Picton et al., (1974), see also Fig. 2.2).

Figure 2.2: The waveform morphology of the auditory evoked response, plotted on a logarithmic axis of time. *Reprinted from Publication 'Human Auditory Evoked Potentials I: Evaluation of Components', Vol. 36, Picton T.W., Hillyard S.A., Krausz H.I., Galambos R., pp. 179-190 (1974), with permission from Elsevier.*



2.3 Estimating the signal to noise ratio.

A more informed decision on the presence or absence of a response can be made if the EEG background activity is not only reduced, but also quantified, thus giving a measure of the 'quality' of the waveform (Schimmel et al, 1967; Wong and Bickford, 1980; Elbering and Don, 1984). In particular, this would help determine whether peaks and valleys in the ensemble coherent average might be attributed to either the background noise or to the ABR. The latter would be particularly beneficial when considering abnormal responses, or when a response is near threshold and the clinicians do not know exactly what they are looking for.

A first measure for quantifying the residual background noise within the coherent average was provided by Schimmel et al. (1967), and is called the ±-reference. The ±-reference is a special type of averaging that alternates between adding and subtracting each successive epoch. Consequently, deterministic signals within the coherent average will cancel out, and the user is left with the residual background noise. The *pm*-reference was first used in the 'split-sweep' technique developed by Lowy and Weiss (1968), and was later transformed into a formal parametric statistical test by Schimmel et al. (1974). In particular, Schimmel et al. (1974) constructed a power ratio, called the P ratio, between the estimated power of the evoked response (given by the mean square of the coherent

average) and the estimated power of the residual noise (given by the mean square of the *pm*-reference; Schimmel, 1974; Wong and Bickford, 1980). The significance of the P ratio can be assessed using an F-distribution with v_1 and v_2 degrees of freedom, both of which, however, are dependent on the spectral content of the EEG, and are typically unknown (Picton et al., 1983; Elbering and Don, 1984). Taking the square root of the numerator and denominator of the P ratio furthermore gives the standard deviation ratio (SDR), which is loosely related to the correlation coefficient (CC), albeit when the CC is obtained from two replicates of the coherent average (Picton et al., 1983).

The performance of the P ratio in detecting click ABRs was evaluated by Arnold (1985), who compared it to the CC, a 'multiple pre-post Z test' (consisting of multiple univariate tests between the pre- and post-stimulus coherent average), and visual detection by clinicians. Their results show that for low stimulus intensities, the CC was the most sensitive method, but that for high stimulus intensities visual detection by clinicians was the most sensitive. Others have similarly observed either a small advantage for the CC over the P ratio (or, equivalently, the SDR) for click-evoked ABR detection (Picton et al., 1983; Valdes et al., 1987), or more or less equivalent performance (Mijares et al., 2013).

The P ratio was later further modified by Elberling and Don (1984), who replaced the estimated variance of the residual noise with the 'single point' (SP) variance, found by drawing a single sample at some fixed index from all epochs, and calculating the variance of the resulting sample. Under the assumption of independence between the SP values, the DOF for the residual background noise (denoted by v_2) will be equal to the number of epochs $N - 1$. The resulting statistic, called the Fsp, can be evaluated with an F-distribution where v_2 is now known. The DOF for the estimated evoked response (contaminated by EEG background activity) is, however, still unknown, and a conservative approach (to minimise false-positives) is recommended by setting v_1 to 5 (Elberling & Don, 1984).

A final modification to the Fsp consists of replacing the SP variance with the mean of multiple SP variances (Martin et al., 1994; Özdamar & Delgado, 1996; Neely & Pepe, 1997). The reasoning behind this modification, according to Neely & Pepe (1997), is that significant correlations can still be observed for samples separated by up to 3 seconds, in which case v_2 would again depend on the spectral content of the EEG. Replacing the SP variance by the mean of multiple SP variances would presumably reduce this dependency, resulting in a more accurate estimation of the power of the noise (Neely & Pepe, 1997). A second advantage is that the underlying distribution of the mean of multiple SP variances will generally be less disperse than the distribution underlying the SP variance, i.e. the mean of multiple SP variances may provide more consistent estimations of the residual power of the EEG background activity.

With respect to auditory steady-state response (ASSR) detection, the sensitivity of the Fsp has been compared to the SDR, and was found to be more or less equivalent when

detecting the auditory steady-stage middle-latency response (Picton et al., 1987). Both were able to detect responses evoked by 40 dB SL tones, which could be decreased to 20-30 dB SL when a band pass filter of 20-100 Hz was used (as opposed to the initial 1-300 Hz band). They were, however, both outperformed by various frequency domain methods. It is worth noting here that ASSRs tend to have a dominant response at specific modulation frequencies, whereas the ABR is more broadband. Results from studies on objective ASSR detection might therefore not always generalise well to ABR detection. With respect to the ABR, the Fsp has been compared to the Fmp and the Scor statistic (Gentiletti et al., 2003), where the Scor statistic is essentially the Fmp combined with a CC, and where the CC is obtained from the ensemble coherent average and some template (Neely & Pepe, 1997). Results suggest that the Fsp outperformed both the Fmp and the Scor statistic when detecting clicks of various intensity levels.

Further comparisons have been drawn between the Fsp, Friedman's test, Cochran's Q test, and the modified q-sample uniform scores test (Cebulla et al., 2000a) when detecting both simulated data and real click-evoked ABR data. For the simulations, the performance of the modified q-sample test and the Fsp was more or less equivalent, whereas for subject ABR data, the modified q-sample test was found to be more sensitive. The authors speculate that the decrease in the sensitivity of the Fsp for real ABR data relative to their simulations may be due to their subject data following a non-Gaussian distribution. Valderrama et al (2014) furthermore compared the Fsp to the CC (obtained from either two replicates of the coherent average, of from the coherent average and a template), along with a novel method based on peak identification called fitted parametric peaks (FPP). Their FPP method proved to be most sensitive in detecting click-evoked ABRs, followed by the FSP, and lastly by the two CCs.

Finally, the problem of unknown DOF of EEG measurements has been circumvented by Lv et al. (2007) by means of a bootstrap approach. The bootstrap is also a resampling with replacement approach, that Lv et al use to approximate the underlying null distribution of some statistic of interest. Statistical inference is then carried out using the approximated null distribution (see section 3.6). The main advantage for this approach is that the null distribution for the statistic of interest does not have to be assumed *a priori*. Besides circumventing the unknown DOF, the bootstrap also permits a large amount of freedom when choosing which features to use for objective detection. In Lv et al (2007), the bootstrap was used to evaluate and compare the specificity and sensitivity of the FSP, the SDR, the 'Peak-to-Peak Difference' and the 'Mean Power' statistics when detecting ABRs in both simulations and in a small sample of normal hearing adults. For the simulations, the Peak-to-Peak Difference was most sensitive method, whereas for the subject data, the Mean Power was the most sensitive. The performance of all four statistics was however quite similar.

As an alternative to the bootstrap, the permutation test might also be considered (Fisher, 1935; Efron & Tibshirani, p 202, 1993). The permutation test is resampling without replacement approach, and can similarly be used as a non-parametric approach

for evaluating test significance. It has been applied to evoked response detection by Maris & Oostenveld (2007), who used it to evaluate 'clusters of t-statistics', also known as the cluster mass test (Bullmore et al., 1999). The permutation test is further considered in the Appendix (section A.5.4).

2.4 Single-band ABR detection

The majority of the aforementioned ABR detection methods are applied in the time domain, and essentially strive to detect an offset (from zero) in voltage as a function of the stimulus. Detection can also be carried out in the frequency domain, in which case the evoked response is detected through the phases and amplitudes of the Fourier components of the spectral bands. In particular, the presence of an evoked response tends to result in an aggregation of the phase components, which are otherwise uniform on the $[0, 2\pi]$ interval under H_0.

The majority of the research on phase coherence for evoked response detection has been directed towards ASSRs, as the spectral band within which the ASSR can be found is both narrow and known *a priori*. The spectral content of a typical transient ABR, on the other hand, is smeared across multiple bands (Elberling, 1976; Kevanishvili & Aphonchenko, 1979; Elberling, 1979; Suzuki et al., 1982). Besides less ideal filter settings due to an increased overlap between noise and signal, the broadband spectral content of the ABR is problematic for single-band spectral coherence techniques, as these would need to be applied multiple times to cover the bandwidth of a typical ABR, potentially resulting in an inflated FPR.

One of the first studies on frequency domain analysis for evoked response detection was conducted by Sayers et al in 1973, who indeed observed an aggregation of phase values of the Fourier components as a function of stimulus intensity. Based on their findings, they speculate that the evoked response might be a reordering of the phases of the Fourier components of the EEG background noise, as opposed to an evoked additive component, superimposed on the background noise. Although an interesting query, it is presumably irrelevant for ABR detection methods, as phase coherence can be expected under both models (Jervis et al, 1983).

Following their 1973 study, Sayers et al (1979) developed the first frequency domain auditory evoked response detection method, which was built around phase variance (and applied a 'rotating χ^2 test' to deal with the circularity of phase). Their test was later improved by Jervis et al., (1983), and extended to ABRs by Fridman et al (1982). The method now goes by the name component synchrony measure (CSM) (Fridman et al., 1982, 1984) or phase synchrony measure (PSM) (Simpson et al., 2000), and is equivalent to the Rayleigh test (Picton et al., 1987; Champlin, 1992). Other methods that test for phase coherence in a single spectral band include Kuiper's statistic (Bachen,

1986; Cebulla et al., 1996; Stürzebecher & Cebulla, 1997), the Hodges-Ajne test (Jervis et al., 1983; Cebulla et al., 1996; Stürzebecher & Cebulla, 1997), and Watson's U2 test (Cebulla et al., 1996; Stürzebecher & Cebulla, 1997). These tests have furthermore been compared in their ability to detect click-evoked ABRs (Cebulla et al., 1996; Stürzebecher & Cebulla, 1997), from which it was concluded that the Rayleigh and Watson's U2 test were more sensitive in detecting ABRs relative to the Hodges-Ajne and Kuiper's test.

The aforementioned methods are applied exclusively to the phase values of Fourier components, and neglect the amplitudes. From a theoretical point of view, methods that use both phase and amplitude values are more powerful than those that use just phase alone (Dobie & Wilson, 1993). The Rayleigh test has therefore been modified to also take either the ranks of the spectral amplitudes into account (Moore, 1980) or the actual values of the spectral amplitudes (Cebulla et al., 2006). Other tests that depend on both phase and amplitude information include the spectral F test (SFT) for hidden periodicity (Valdes et al., 1997), the magnitude squared coherence (MSC) (Dobie and Wilson, 1989), the Hotelling's T2 test (Hotelling, 1931; Picton et al., 1987), a multivariate version of the MSC called Multiple Coherence (Miranda de Sá et al., 2004), and several modifications of the q-sample uniform scores test (Cebulla et al., 2006).

Simulations have indeed demonstrated an advantage for methods that use both phase and amplitude over those that use just phase (Dobie & Wilson, 1993; 1994a). For evoked response detection, however, the advantage is less obvious, and in some cases, a more or less equivalent performance is observed (Jervis et al., 1983; Dobie & Wilson, 1994a). This suggests that the evoked response can be found primarily in the spectral phases, which is supported by various findings that show a relatively poor performance for methods that use just amplitude information (Jervis et al., 1983; Greenblatt et al., 1985; Champlin, 1992). Nevertheless, there is still somewhat of a consensus that at least a small increase in sensitivity for evoked response detection can be gained by including amplitude information, as opposed to using just the phases (Picton et al., 1987; Champlin, 1992; Dobie & Wilson, 1993; Cebulla et al., 1996; Stürzebecher & Cebulla, 1997; Valdes et al., 1997; Simpson et al., 2000; Picton et al., 2001; Cebulla et al., 2006).

Various comparisons between the aforementioned methods (using both phase and amplitude information) have been drawn, i.e. the modified Rayleigh test (using the ranks of the spectral amplitudes), the MSC, the Hotelling's T2 test and the SFT have been compared in their ability to detect ASSRs within a single-band (Cebulla et al., 2001). Findings show that the modified Rayleigh test was the most sensitive method, followed closely by the MSC, and lastly by the Hotelling's T^2 test and the SFT, although performance between all four statistics was quite similar. The (single-band) Hotelling's T^2 test has furthermore taken a variety of forms for evoked response detection. It was first used for detecting ASSRs, where it was applied to the real and imaginary parts of the Fourier components of a single-band (Rodriguez et al., 1986; Picton et al., 1987). The Hotelling's T^2 test was later modified by Victor and Mast (1991) to exploit the assumption that the real and imaginary parts within a single-band are uncorrelated and have

equal variance. Their modification, called circular T2 (T2C), grants a small increase in power, most noticeably for small ensemble sizes (Victor and Mast, 1991). The T2 test was later modified again by Valdes et al (1997), who estimate the critical boundaries using noise estimated from the spectral bands adjacent to the spectral band of the modulation rate. A potential advantage for the latter is an increased robustness to noise artefacts, i.e. it does not require the mean of the noise to be zero. Their modification was compared to T2C, the CSM, and the SFT, but no consistent differences in test performance were observed.

Finally, when the expected value of the phase of a spectral band is known *a priori*, then a bias can be introduced by projecting the phase values onto an expected phase vector (Dobie & Wilson, 1994b; Lins et al., 1996; Picton et al., 2001). Lins et al (1996) tested a phase-weighted version of the SFT and an amplitude- and phase-weighted version of T2C, where the expected phase and amplitude values were obtained from the grand coherent average (taken across all subject coherent averages). The phase- and amplitude-weighted versions performed better than their original counterparts, with the increase in performance for the SFT being larger than for the T2C. Further tests were performed by Picton et al. (2001), who compared a standard phase coherence test with it's phase-weighted version, along with the SFT, and a phase-weighted t-test for detecting ASSRs. Results again showed a small but significant advantage in sensitivity for the phase-weighted methods.

2.5 Multi-band ABR detection

The broadband spectral content of the ABR has led scientific investigations towards methods for analysing multiple-bands within a single test, i.e. multi-band detection methods. Multi-band detection methods previously explored for ABR detection include the Synchrony Measure (Fridman, 1984), the q-sample uniform scores test (Mardia, 1972) and it's modifications (Stürzebecher et al., 1999; Cebulla et al., 2006), the q-sample Analogue of Watson's U2 Statistic (Maag, 1966; Stúrzebecher et al., 1999), the Hotelling's T^2 test (Hotelling, 1931; Valdes et al., 1987), and Multiple Coherence (Miranda de Sá et al., 2004). An and/or decision rule for combining p values from multiple univariate tests has also been proposed (Stürzebecher & Cebulla, 1997).

With respect to the detection of transient click-evoked ABRs, the sensitivity of the q-sample uniform scores test has been compared to the q-sample Analogue of Watson's U2 Statistic, along with a modified version of the q-sample uniform scores test that, in addition to the ranks of the phases, also takes the ranks of the spectral amplitudes into account (Stürzebecher et al., 1999). The modified version of the q-sample uniform scores test proved to be most sensitive, followed by the original q-sample uniform scores test, and lastly by the q-sample Analogue to Watson's U2 Statistic. The q-samples uniform scores test was later further modified to include combinations of (i) phase values

and amplitude ranks, (ii) amplitude ranks and phase values, or (iii) phase values and amplitude values (Cebulla et al., 2006). These modifications have been compared to the SFT, the Rayleigh test, and the modified Rayleigh test (where the modification uses either the ranked spectral amplitudes or the actual values of the amplitudes) in their ability to detect ASSRs. Findings show that the q-samples uniform scores test (using the actual values of the phases and amplitudes) was the most sensitive method (Cebulla et al., 2006).

With respect to multivariate applications of the Hotelling's T^2 test, it was first applied to the real and imaginary parts of the Fourier components from multiple spectral bands (under the assumption of independence between spectral bands) by Valdes et al (1987) for detecting click-evoked ABRs in infants. Results show a superior test performance for the T^2 test over the SDR and the CC. The Hotelling's T^2 test has also outperformed the Fmp when detecting ABRs extracted from quasi ASSRs (Lachowska et al., 2012), and has recently been applied in the time domain for detecting speech-evoked cortical auditory evoked potentials (CAEPs; Golding et al., 2009; Carter et al., 2010; Chang et al., 2012; Van Dun et al., 2012; Van Dun et al., 2015). These time domain features are defined as the means of segments of epochs (later referred to as 'time-voltage means', or TVMS). Findings from Golding et al. (2009) and Carter et al. (2010) show that, when using these time-domain features, the sensitivity of the Hotelling's T^2 test is at least equivalent to that of a group of experienced examiners when detecting CAEPs. Finaly, Van Dun (2015) applied the Hotelling's T^2 test in combination with a decision tree for sequential testing in an attempt to automate the process for approximating behavioural hearing thresholds in CAEP audiometry.

2.6 Discussion

Based on the review, a selection of methods is now made, which are further evaluated in terms of specificity, sensitivity, and test time in Chapters 5, 6, and the Appendix. Some new methods that have not yet been explored for ABR detection are also included. With respect to the frequency domain methods, the selection is restricted to multi-band methods (due to the broadband spectral content of the ABR) that use both phase and amplitude information.

The first two methods that are of interest include the Fsp and the Fmp. Although the Fsp is not necessarily the most sensitive method (e.g. Cebulla et al., 2000a, 2000b; Lv et al, 2007), it is perhaps the most widely used for ABR detection. Including the Fsp (and the Fmp) may therefore allow findings to be related (to some degree) to previous studies. Moreover, the literature shows some inconsistency in the sensitivity of the Fsp relative to e.g. the Fmp. In particular, Cebulla et al (2000b) observed a small advantage for the Fmp over the Fsp for small ensemble sizes, and more or less equivalent performance for large ensemble sizes, which might be attributed to a

decreased reliability (or an increased variability) of the SP variance (within the Fsp) for small ensemble sizes (see also Methods, section 3.1). Gentiletti et al (2003), on the other hand, observed an advantage in test performance for the Fsp over the Fmp. Both the Fsp and the Fmp are therefore included in the selection, which are evaluated using either theoretical F-distributions with assumed DOF, or with the bootstrap approach. The bootstrap approach will presumably allow a more fair and consistent comparison of test performance, as it does not require the DOF of the data to be assumed in advance.

A method that has evoked much interest throughout the literature is the CC (Arnold, 1985; Picton et al., 1983; Valdes et al., 1987; Neely & Pepe, 1997; Mijares et al., 2013; Valderrama et al, 2014). A recurring complication, however, is again the unknown DOF of the data, which complicates statistical inference. The CC is therefore also included in the selection, and its significance is evaluated using the bootstrap approach. Additional bootstrapped statistics that are included are the 'Peak-to Peak Difference' and the 'Mean Power' statistics, both of which outperformed the Fsp in Lv et al. (2007), but have not yet been compared to alternative methods.

With respect to multi-band ABR detection methods, a first set of methods to include are various modifications to the q-sample uniform scores test (Cebulla et al, 2006), which have shown a good performance for ASSR detection, but have not yet been evaluated for ABR detection. The modifications selected here are (following the notation in Cebulla et al, 2006) the 'Modified q-sample V4' test, which is applied to the actual values of the phases and amplitudes, along with the 'Modified q-sample V2' test, which is applied to the ranks of the phases and amplitudes (see also section 3.5). A competitor to the modified q-sample statistics is the Hotelling's T^2 test. When applied in the frequency domain, the Hotelling's T^2 test essentially uses the same information as the Modified q-sample V4 test, i.e. it is applied to the real and imaginary parts of the Fourier components, whereas the modified q-sample V4 test is applied to the phases and amplitudes. An important difference, however, is that the Hotelling's T^2 test weights the feature means according to the variance and covariance of the features, whereas the modified q-sample V4 test does not. Weighting the feature means by the variance and covariance results in a hyper-ellipsoid as H_0 rejection region for the Hotelling's T^2 test, where the shape of the ellipsoid is determined by the variance and covariance of the features (see section 3.2). The advantage of having an ellipsoid as H_0 rejection region is that the null hypothesis is more easily rejected in some directions relative to others, i.e. it has the potential of providing a more powerful test relative to, for example, tests with a spherical rejection region. Finally, recent studies also demonstrate a good performance for the Hotelling's T^2 test for CAEP detection when applied in the time domain (Golding et al., 2009; Carter et al., 2010; Chang et al., 2012; Van Dun et al., 2012; Van Dun et al., 2015). Time domain features have not yet been explored for the Hotelling's T^2 test for ABR detection, and are also included in the selection.

Some alternative methods that have not yet been explored for ABR detection include Repeated Measures Analysis of Variance (RM ANOVA), along with a Multivariate Anal-

ysis of Variance approach, for which the Hotelling's T^2 test can again be used. Repeated measurements are of interest as they are insensitive to mean voltage offsets due to some types of artefacts or noise, i.e. they may have a more robust control of specificity relative to some alternative methods that just look for non-zero mean voltages in the EEG recording. It is also worth noting here that RM ANOVA is, in theory, more powerful than the multivariate approach for analysing repeated measurements, but requires an additional assumption called sphericity to be satisfied. When sphericity is violated, then the DOF of the test need to be corrected (achieved using the Greenhouse Geiser or Huyn Feldt correction), resulting in a reduced statistical power. The performance of RM ANOVA for ABR detection will hence depend, in part, on the extent to which sphericity is satisfied for EEG measurements (further explored in the Appendix, sections A.3 and A.4). A final method for analysing repeated measurements that is included in the analysis is Friedman's test, which is the non-parametric equivalent to RM ANOVA. Friedman's test requires neither the sphericity nor the normality assumption, and might therefore have an increased robustness to noise and artefacts.

Finally, section 3.6.2 describes a variation to the standard bootstrap approach in Lv et al (2007), which allows multiple statistics to be combined (and tested for significance) efficiently. The approach is used to combine the Hotelling's T^2 test (applied in the time domain) with the CC, henceforth referred to as 'T2 Time + CC'. A useful property for 'T2 Time + CC' is that it can be biased towards detecting specific response morphologies through the CC, without losing the (potentially) robust, non-template specific detection through the Hotelling's T^2 test.

Chapter 3

Objective detection methods

This section provides a description of the ABR detection methods, which were chosen based on the review and discussion in Chapter 2. The methods are further evaluated and compared in terms of specificity, sensitivity, and test time in Chapters 5, 6, and the Appendix.

The data to which the methods are applied consists of ensembles of epochs, structured according to matrix D:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1J} \\ d_{21} & d_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ d_{N1} & \cdots & \cdots & d_{NJ} \end{bmatrix}$$

where N is the ensemble size, J is the number of samples per epoch, and d_{ij} is the jth sample of the ith epoch. The mean epoch $\bar{X}$ (also known as the coherent average), is found by taking the J averages across the columns. The frequency domain representation of D is furthermore obtained by taking the Fast Fourier Transform (FFT) of each row. Features can then be extracted from either the time or frequency domain representations of the data. Extracting Q features from each epoch results in the NxQ-dimensional feature matrix V:

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1Q} \\ v_{21} & v_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ v_{N1} & \cdots & \cdots & v_{NQ} \end{bmatrix}$$

where v_{ij} is the j_{th} feature extracted from the i_{th} epoch.

3.1 The Fsp and the Fmp

The Fsp and the Fmp are defined as the ratio between the variance of the mean epoch $\bar{X}$ (found by taking the J averages across the columns of data matrix D) and the estimated variance of the EEG background noise. For the Fsp, the variance of the EEG background noise is estimated by the 'single point' (SP) variance, which is defined as the variance down a single arbitrarily chosen column of data matrix D. The Fsp is given by Elberling and Don (1984):

$$\text{Fsp} = N\frac{\text{VAR}(\bar{X})}{\text{VAR(SP)}} \tag{3.1}$$

where VAR denotes variance, and SP refers to the values along an arbitrarily chosen column of D. For the Fmp, the variance of the EEG background noise is approximated by taking the average of multiple 'SP variances' (the average of the variances of multiple columns of D). The Fmp is given by (Martin et al., 1994):

$$\text{Fmp} = N\frac{VAR(\bar{X})}{\frac{1}{H}\sum_{i=1}^{H} VAR(SP_i)} \tag{3.2}$$

where $\text{VAR(SP}_i)$ is the variance of the i_{th} included column of D, and H is the number of columns of D to include.

Under the null hypothesis of no response present, it is assumed that the Fsp and the Fmp follow F-distributions with v_1 and v_2. DOF v_2 is equal to $N-1$, under the condition that consecutive epochs are sufficiently distant in time to be uncorrelated, i.e. they are independent. DOF v_1 is more difficult to determine, and depends on the extent to which consecutive samples within $\bar{X}$ are correlated, which, in turn, depends on the spectral content of the data. A conservative recommendation (a FPR smaller than the nominal α-level of the test) is given by Elberling & Don (1984) by setting v_1 to 5. Alternatively, the significance of the Fsp and the Fmp can be evaluated with the bootstrap approach (section 3.6).

3.2 The one-sample Hotelling's T^2 test

The one-sample Hotelling's T^2 test plays a big role throughout this thesis, and is therefore given a slightly more in depth description. The most important equations are

first provided below, after which further insight into Hotelling's T^2 test is established by exploring its relationship with the one-sample t-test, and illustrating its H_0 rejection region. The time and frequency domain features for the Hotelling's T^2 test used throughout this thesis are then also described (sections 3.2.2 and 3.2.3), along with its application to repeated measurements (section 3.2.4).

The one-sample Hotelling's T^2 test is the multivariate extension to Student's t-test and can be used to test whether the means of Q features are significantly different from Q hypothesised values. The statistic itself is a weighted sum of the Q feature means where the weights are determined by the variances and covariances of the features. These weights have the convenient property of normalising the Q feature means, which allows features with different scales and units to be combined appropriately. The T^2 statistic is given by (Hotelling, 1931; Rencher, 2001, p.118):

$$T^2 = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})\mathbf{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})^H \tag{3.3}$$

where $\bar{\boldsymbol{x}}$ is the Q-dimensional vector of means (found by taking the means down the Q columns of V), $\boldsymbol{\mu_0}$ the Q-dimensional vector of hypothesized values to test against, $\boldsymbol{S^{-1}}$ the inverse of the covariance matrix of the NxQ-dimensional feature matrix V, and ^{H} superscript denotes Hermitian transpose. The T^2 statistic can then be transformed into an F statistic with:

$$\mathrm{F} = \frac{N-Q}{Q(N-1)}T^2, \quad \sim F_{v_1,v_2} \tag{3.4}$$

which follows an F-distribution with v_1 and v_2 DOF under H_0 (denoted by $\sim F_{v_1,v_2}$), where $v_1 = Q$ and $v_2 = N - Q$. Note that in order to calculate $\boldsymbol{S^{-1}}$, the number of epochs N should be larger than the number of features Q. Note also that when the features are highly correlated, that $\boldsymbol{S^{-1}}$ can be close to singular, in which case rounding errors might occur. A solution would then be to use the pseudoinverse (e.g. the Moore-Penrose pseudoinverse; Moore 1920; Penrose 1955) instead of the regular inverse.

3.2.1 Relationship with the t-test and the H_0 rejection region

Further insight in the one-sample Hotelling's T^2 test can be gained by drawing an analogy with its univariate counterpart, Student's one-sample t-test. The t-statistic

(Student, 1908) can be written as:

$$t = \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{N}}} = \sqrt{N}(\bar{x} - \mu_0)s^{-1} \tag{3.5}$$

where N is the sample size, $\bar{x}$ the sample mean, μ_0 the hypothesized value to test against, and s is the sample standard deviation. The t-statistic is hence the distance between $\bar{x}$ and μ_0, standardized in units of the estimated standard error (where one standard error is given by $\frac{s}{\sqrt{(N)}}$). The relationship with the T^2-statistic becomes more apparent by squaring the t-statistic:

$$t^2 = N(\bar{x} - \mu)s^{-2}(\bar{x} - \mu_0)' \tag{3.6}$$

By comparing Eq. 3.3 and 3.5, it is readily seen that the multivariate counterpart to $\bar{x}$ is $\bar{\boldsymbol{x}}$, and the multivariate counterpart to s^{-2} is $\boldsymbol{S^{-1}}$. The multivariate counterpart to s^{-1} (in Eq. 3.5), however, is still somewhat obscure. To find it, spectral decomposition can be used to decompose $\boldsymbol{S^{-1}}$ into its rotation matrix $\boldsymbol{R}$ and scaling matrix $\boldsymbol{C}$, where $\boldsymbol{R}\,\sqrt{\boldsymbol{C}}\,\sqrt{\boldsymbol{C}}\,\boldsymbol{R} = \boldsymbol{S^{-1}}$. It can then be seen that the multivariate counterpart to s^{-1} is $\boldsymbol{R}\,\sqrt{\boldsymbol{C}}$.

$\boldsymbol{S^{-1}}$ in Eq. 3.3 essentially has the same role as s^{-1} in Eq. 3.6, i.e. to normalise the distance between $\bar{\boldsymbol{x}}$ and $\boldsymbol{\mu_0}$. Note that normalisation is an important step if the distance is to be evaluated using a theoretical distribution. In the univariate case, normalising the observed feature values through s^{-1} and re-calculating the standard deviation (from the now normalised feature values) gives a standard deviation of 1. For the multivariate case, the only difference is that the normalisation also takes covariance into account. An example is given for the bivariate case ($Q = 2$) in Fig. 3.1: Data A has covariance matrix

$$\begin{pmatrix} 20 & 10 \\ 10 & 20 \end{pmatrix}$$

Transforming data A using its rotation matrix $\mathbf{R}$ removes the covariance between the features, giving data B, which now has covariance matrix

$$\begin{pmatrix} 30 & 0 \\ 0 & 10 \end{pmatrix}$$

Data B is then further transformed using scaling matrix $\sqrt{\mathbf{C}}$, which normalises the feature variances, giving data C. The covariance matrix for data C is now the 2x2

identity matrix:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

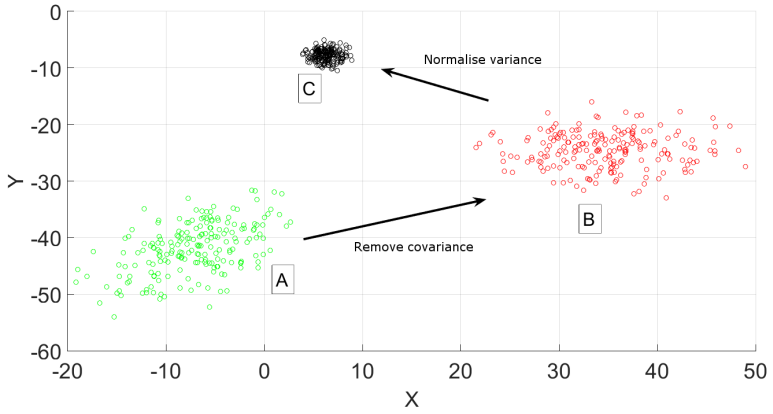


Figure 3.1: An illustration for demonstrating the normalisation process underlying the T^2-statistic for a bivariate data set. In particular, the original feature values (data A) are first rotated, such that the correlation between feature X and feature Y is zero (data B). The resulting rotated feature values are then rescaled, such that the variances of X and Y are both one (data C). The covariance matrix for data C is now the 2x2 identity matrix.

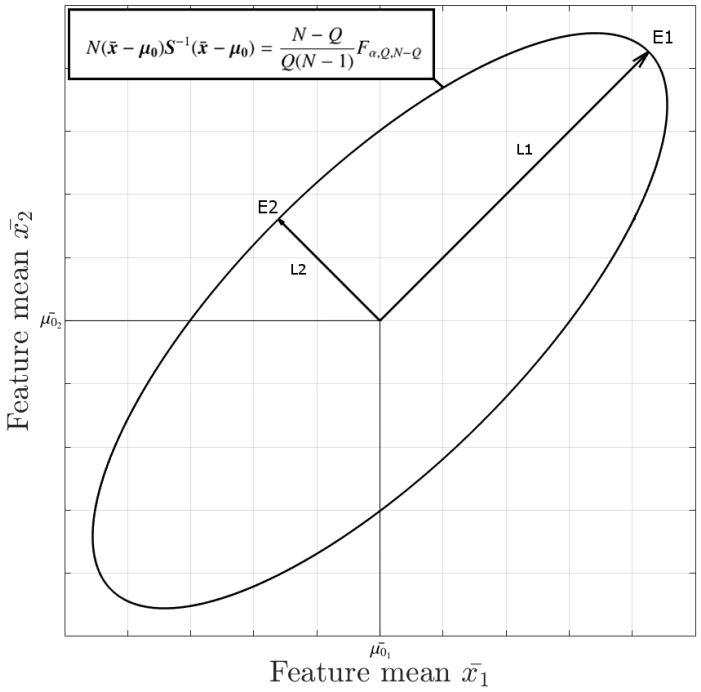


Figure 3.2: The confidence ellipse for a hypothetical bivariate population with positively correlated variables. The semi-axes of the ellipse (E1 and E2) are determined by the eigenvectors of the data's covariance matrix $\boldsymbol{S}$ and have lengths (L1 and L2) proportional to the corresponding eigenvalues. Alternatively, the ellipse can be defined by all possible combinations of sample means ($\bar{x_1}$ and $\bar{x_2}$) that satisfy the given equation, where $F_{\alpha,Q,N-Q}$ is the critical value at level α for an F-distribution with Q and $N-Q$ DOF.

Finally, how covariance might affect test significance is demonstrated for the bivariate case in Fig. 3.2. The confidence ellipse (the H_0 acceptance rejection) is centred around the hypothesized values to test against, denoted by μ_{01} and μ_{02}. The shape of the ellipse is determined by the variance and covariance of the data, i.e. the semi-major (E1) and semi-minor (E2) axes are given by the eigenvectors of $\boldsymbol{S}$, and have lengths (L1 and L2) proportional to the largest and second largest eigenvalues respectively. When the two feature means ($\hat{x_1}$ and $\hat{x_2}$) fall within the confidence ellipse, the null hypothesis is accepted, else H_0 is rejected (with certainty $\geq$ (1-α) x 100%). As can be seen, the result of having an ellipsoidal rejection region is that H_0 is more easily rejected for certain combinations of $\hat{x_1}$ and $\hat{x_2}$, relative to others. Note that when the covariance is zero, that the H_0 rejection region is spherical.

3.2.2 Time domain features

When applied in the time domain, the features consist of 'time-voltage means' (TVMs), which are defined as mean voltages, calculated across short time-intervals within each epoch (see e.g. Golding et al. 2009; Carter et al. 2010; Chang et al. 2012; Van Dun et al., 2012; Van Dun et al., 2015). Note that the direct current component is removed from the EEG recordings with a high-pass filter, meaning the expected values for the TVMs under H_0 will be zero. The hypothesised values to test against (denoted above as $\boldsymbol{\mu_0}$) are therefore given as a Q-dimensional vector of zeros.

To clarify with an example, when using Q TVMs, each epoch is divided into Q segments of approximately equal duration, and the mean is taken across each segment, resulting in the NxQ-dimensional feature matrix V. The length of each TVM segment requires a compromise, such that the segments are neither too long, thus covering both peaks and troughs (resulting in a loss of information) nor too short, thus leading to poor statistical robustness and a reduced test sensitivity. The loss of information due to too few TVMs is further illustrated in Figure 3.3 below.

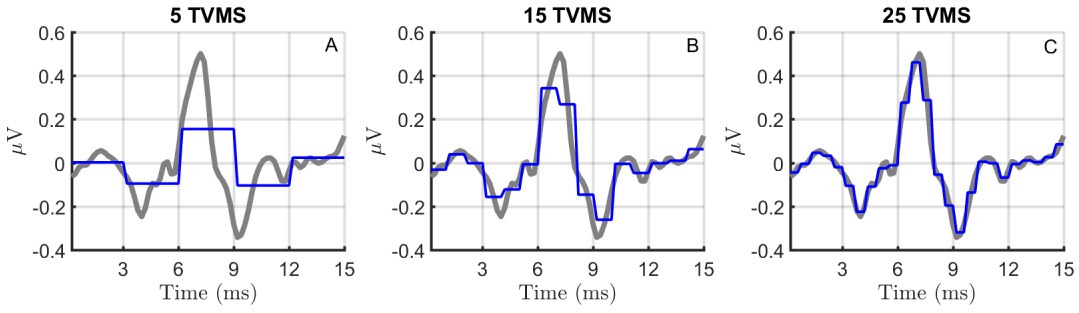


Figure 3.3: Plots A, B, and C show the loss of information when using 5, 15, and 25 TVMs, respectively. The gray plots show an ABR template obtained from the 40 dB SL condition from data set **D2** (see Chapter 4), whereas the blue plots show the values of the TVMs when plotted across the time-segments from which they were obtained. When too few TVMs are used then consecutive peaks and valleys in the waveform start to cancel out (the time-segment across which each TVM is calculated is too long), resulting in a loss of information.

3.2.3 Frequency domain features

When using the frequency domain approach, the features are the real and imaginary parts of the Fourier components of W spectral bands, giving a $Nx2W$-dimensional feature matrix V. Note that when using the frequency domain approach, that $Q = 2W$ in Eq. 3.3 and 3.4. The phases within each spectral band are furthermore assumed to be uniformly distributed between 0 and 2π under H_0. The hypothesised values $\boldsymbol{\mu_0}$ are therefore given as a $2W$-dimensional vector of zeros.

3.2.4 Repeated measurements

Observations are called repeated measurements when multiple measurements are taken from the same sampling unit. For evoked response detection, a sampling unit can be defined is an epoch, in which case the repeated measurements is the voltage over time. Tests for analysing repeated measurements attempt to evaluate the null hypothesis H_0 that, on average, the repeated measurements are constant over time. More formally, for Q repeated measurements on N sampling units, the null hypothesis H_0 is given by: $\bar{x_1} = \bar{x_2} = ... = \bar{x_Q}$, where $\bar{x_i}$ is the mean value for the ith feature.

Repeated measurements are readily analysed in the time domain with the Hotelling's T^2 test. In particular, the columns in V are redefined using $V_i = V_i - V_{i-1}$ for $(i = 2, 3, ..., Q)$, where V_i denotes the ith column of V. The resulting matrix of 'difference features' can then be analysed using the standard Hotelling's T^2 test in Eq. 3.3 (note that column V_1 is removed, thus reducing the dimension of the feature set by 1). Alternatively, the Hotelling's T^2 test can be applied to the original Q-dimensional vector of means $\bar{\boldsymbol{x}}$, and a contrast matrix $\boldsymbol{A}$ can be inserted into the T^2 equation (Eq. 3.3), giving (Rencher, 2001, p.208):

$$T^2 = N(\boldsymbol{A}\bar{\boldsymbol{x}}) \left[\boldsymbol{A}\mathbf{S}^{-1}\boldsymbol{A}^{-1}\right] (\boldsymbol{A}\bar{\boldsymbol{x}})' \tag{3.7}$$

The requirements for $\boldsymbol{A}$ is that it is of rank $Q - 1$ and that it's rows sum to zero, e.g.:

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & -1 \end{pmatrix} \tag{3.8}$$

The T^2-statistic in Eq. 3.7 is then again transformed to an F-statistic, now using $\frac{N-(Q+1)}{(Q-1)(N-1)}T^2$, which follows an F-distribution with $Q - 1$ and $N - Q + 1$ DOF under

H_0. It might be noted that the T^2 test for repeated measurements also goes by the name 'Profile analysis' (Rencher, 2001, p.208).

3.3 Friedman's test

Friedman's test is a non-parametric test for analysing repeated measurements (Friedman, 1937). The test is applied to the ranks of the features, where the ranking is performed within sampling units. To clarify, when Q TVMs are extracted (per epoch), then a separate ranking (across the Q TVMs) is performed for each epoch, thus transforming each epoch into integer values ranging from 1 to Q. Friedman's test is then used to test whether the columns in V (now containing the ranks of the features) share the same underlying distribution. To do so, note first that when H_0 is true (the feature rankings are random) then the theoretical mean rank down an arbitrarily chosen column of V is given by (Friedman, 1937):

$$\mu = \frac{1}{2}(Q+1) \tag{3.9}$$

which has expected variance

$$\sigma^2 = \frac{Q^2-1}{12} \tag{3.10}$$

A measure for evaluating the null hypothesis H_0 can then be constructed using the difference between the theoretical mean feature rank μ and the observed mean feature ranks $\bar{x}_i$ (for $i = 1, 2, ...Q$). In particular, a sum of squares term is constructed using $\sum_{i=1}^{Q}(\bar{x}_i - \mu)^2)$, which follows a (scaled) χ^2 distribution under H_0, that is:

$$\chi^2_{Q-1} = \frac{Q-1}{Q\sigma^2}\sum_{i=1}^{Q}(\bar{x}_i - \mu)^2 \tag{3.11}$$

where $\frac{Q-1}{Q\sigma^2}$ functions as a normalisation factor for the sum of squares term $\sum_{i=1}^{Q}(\bar{x}_i-\mu)^2$. When H_0 is true, then χ^2_{Q-1} follows a χ^2 distribution with $Q-1$ DOF.

3.4 Repeated Measures Analysis of Variance

Repeated measures Analysis of Variance (RM ANOVA) is essentially a modification of the standard between subjects ANOVA, which is therefore described first. For a standard between subjects ANOVA with Q between subject levels and N subjects, there will be NxQ observations (feature matrix V is again NxQ-dimensional), all of which are assumed to be independent. The null hypothesis states that the Q columns in V share the same distribution, i.e. that the between subjects factor (time, in this case) does not affect feature means $\bar{x}_i$ for $i = 1, 2, ..., Q$. The null hypothesis can be evaluated using a variance ratio (see e.g. Cardinal, 2004), where the nominator represents an estimate of the variance of the feature means, denoted by 'mean sum of squares between' (or $MSSB$), and is given by:

$$MSSB = \frac{1}{Q-1} \sum_{i=1}^{Q} N(\bar{x}_i - \bar{x})^2 \tag{3.12}$$

where $\bar{x}$ is the grand mean (the mean taken across feature means $\bar{x}_i$ for $i = 1, 2, ..., Q$). The denominator of the variance ratio is also an estimate of the variance of the feature means, and is denoted by the 'mean sum of squares within' (or $MSSW$):

$$MSSW = \frac{1}{N-Q} \sum_{i=1}^{Q} \sum_{j=1}^{N} (v_{ij} - \bar{x}_i)^2 \tag{3.13}$$

Note that $MSSW$ depends exclusively on the variance *within* each column of V. It is therefore invariant under data translations, which implies that it is a true estimate of sample variance under H_0, regardless of H_0 being true or not (albeit under the assumption of homogeneity of variance amongst the columns in V). The $MSSB$, on the other hand, estimates data variance using the means of the columns (it is the variance *between* the column means), and is not invariant under data translations. $MSSB$ is therefore an unbiased estimate of sample variance under H_0, only when H_0 is indeed true. When H_0 is false, then the $MSSB$ will tend to be larger than $MSSW$, and the variance ratio will tend to be large, giving a higher probability of rejecting H_0. In particular, the null hypothesis can be evaluated using $\frac{MSSB}{MSSW}$, which is F-distribution with $Q-1$ and $N-Q$ DOF under H_0.

The assumptions underlying ANOVA include the normality assumption, the homogeneity of variance assumption (amongst the columns in V), and the independence assumption (between all observations). For repeated measurements, the independence assumption is typically violated, and a modification to the standard ANOVA is required,

achieved by taking the mean of each sampling unit is taken into account. The data should also satisfy an additional assumption called 'sphericity', which is the assumption of 'equal variance of difference scores' (see Appendix, section A.4). When sphericity is violated, then the DOF of the test need to be corrected, achieved using either the Greenhouse Geisser (GG, section A.4.1) and/or the Huyn Feldt correction (HF, section A.4.2). The modification to the standard ANOVA is called the 'mean sum of squares error' (or $MSSE$), and is given by (Cardinal, 2004):

$$MSSE = \frac{1}{(N-1)(Q-1)} \left[\sum_{i=1}^{Q} \sum_{j=1}^{N} (x_{ij} - \bar{x_i})^2 - Q \sum_{i=1}^{N} (\bar{x_j} - \bar{x})^2 \right] \tag{3.14}$$

where $\bar{x_j}$ is the mean taken across the Q TVMs in epoch j. The ratio $\frac{MSSB}{MSSE}$ is then again F-distribution under H_0, now with $Q-1$ and $(N-1)(Q-1)$ DOF.

3.5 The q-sample uniform scores test and its modifications

The original q-sample uniform scores test (Mardia, 1972) is a non-parametric test for evaluating whether the phases of W spectral bands share the same distribution. The modification proposed by Stürzebecher et al (1999) uses the ranks of the amplitudes in addition to the ranks of the phases, and is given by:

$$W^* = c \sum_{j=1}^{W} \left[\left[\sum_{i=1}^{N} r_{ij} cos(\beta_{ij}) \right]^2 + \left[\sum_{i=1}^{N} r_{ij} sin(\beta_{ij}) \right]^2 \right] \tag{3.15}$$

where r_{ij} is the rank of the amplitude of the ith Fourier component (obtained from the ith epoch) of the jth spectral band, c is an additional scaling factor given by:

$$c = \frac{4}{W^2(W+1)^2} \frac{2}{N} \tag{3.16}$$

and β_{ij} is given by:

$$\beta_{ij} = \frac{a_{ij} 2\pi}{NW} \tag{3.17}$$

where a_{ij} is the rank of the phase of the ith Fourier component (similarly obtained from

the ith epoch) of the jth spectral band. This modification will henceforth be referred to as 'Modified q-sample V2' (in accordance with Cebulla et al, 2006).

In addition to the modified q-sample V2 test, the 'Modified q-sample V4' test (Cebulla et al, 2006) is also included in the analysis. The latter uses the actual values of the phases and amplitudes as opposed to their ranks, in which case r_{ij} in Eq. 3.15 refers to the amplitude of the ith Fourier component of the jth spectral band and β_{ij} to the (untransformed) phase value of the ith Fourier component of the jth spectral band. The significance of these statistics can furthermore be evaluated with pre-determined critical values determined using simulations (Stürzebecher et al (1999), Cebulla et al, 2000; Cebulla et al, 2006). Deviating from the literature, the significance of the modified q-sample V2 and V4 statistics in this work are evaluated using the bootstrap. How the critical decision thresholds might differ between these two approaches is further considered in the section 6.1.3 (discussion). Some results on the reliability of pre-determined thresholds and bootstrapped confidence intervals are also presented in the Appendix (sections A.8 and A.5.1, respectively).

3.6 Bootstrapping

The bootstrap (Efron & Tibshirani, 1993) is a resampling with replacement procedure for generating additional (resampled) datasets. Each resampled dataset is constructed by choosing N sampling units (with replacement) from the original sample, where each observation has an equal probability of being selected. Some parameter of interest is then calculated from each resampled dataset, giving a population or histogram of (resampled) parameters, which can then be used to estimate additional parameters of interest (e.g. confidence intervals or a standard error)

In Lv et al (2007), the bootstrap is used to approximate the null distribution for some feature of interest, which can then be used to construct confidence intervals for accepting or rejecting H_0. To achieve this, the resampled data sets should represent EEG measurements under H_0. This is realised by randomly resampling epochs from within the continuous EEG recording without regard to where the stimuli occur, such that the resampled epochs are (on average) no longer time-locked to the stimuli. Note that the resampled epochs may overlap, in accordance with the principles of bootstrapping where samples are picked at random with replacement, i.e. without removing that data from what can be picked later. The feature of interest is then calculated from all re-sampled datasets, giving a population or histogram of values. It is assumed that the distribution of resampled feature values is an accurate approximation of the features true null distribution, under the condition that the number of resampled data sets M and the original ensemble size N is sufficiently large.

An example using the Fsp

Consider evaluating the test significance of an Fsp value of 1.5, calculated from an

ensemble of N = 500 epochs. The bootstrap would then proceed by randomly resampling (with replacement) many additional ensembles (each containing 500 epochs) from the original recording, without regard to where the stimuli occur (for illustration purposes, the number of resampled data sets in this example is set to 10 000). Calculating the Fsp from each bootstrapped ensemble then gives a population of Fsp values, which can be used to approximate the null distribution of the Fsp (see Figure 3.4). The approximated null distribution is then used to construct confidence intervals for rejecting or accepting H_0. Alternatively, a p value can be generated by finding the location of the observed Fsp value (in this case a value of 1.5) along the bootstrapped null distribution. In particular, the p value is given by the percentile under the null distribution to the right of the observed Fsp value, and is for this hypothetical example equal to $p = 0.1261$.

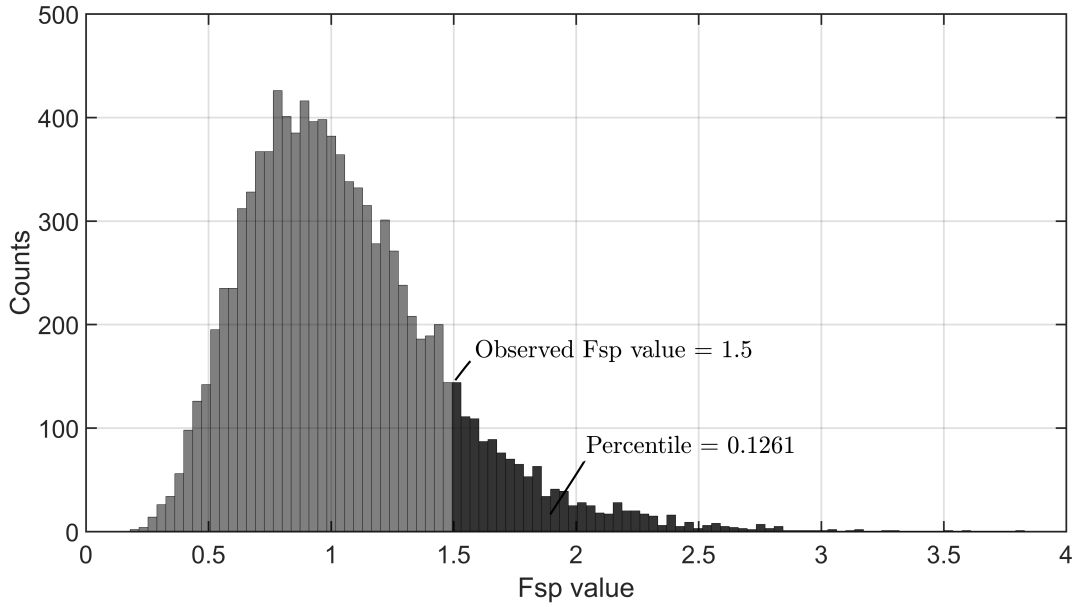


Figure 3.4: An example of how the approximated null distribution is used to evaluate the significance of an observed Fsp value of 1.5, achieved by finding the percentile under the approximated null distribution to the right of the observed Fsp value. For this hypothetical example, the percentile is 0.1261, which is the probability of observing the given Fsp value if the null hypothesis H_0 (no response present) was indeed true, i.e. the p value.

Discussion

The bootstrap approach in Lv et al (2007) makes the assumption that the bootstrapped null distribution approximates the features true null distribution for sufficiently large N and M. This first raises the question as to how large N and M should be before the approximation is sufficiently accurate. Note that when N is too small, the probability of obtaining a sampling error is increased, i.e. the observed sample may not be representative of the true population. Efron & Bradely (1993) suggest that a sample size of at least 30 is usually sufficient for avoiding sampling errors. Although this suggests that N is not an issue for ABR detection (as N is typically much larger than 30), it is not clear how sampling errors might be affected by highly non-stationary, correlated data. With respect to M, this should be sufficiently large to ensure a good consistency

or reliability of the estimated critical thresholds for rejecting H_0. In particular, when M is too small, then the bootstrapped null distribution will be both unreliable and insufficiently smooth, resulting in variable critical thresholds, i.e. repeating the bootstrap procedure may result in different decision boundaries (and hence a different test result). How large M should be before the critical thresholds are sufficiently reliable is further considered using additional simulations in the Appendix (section A.5.1).

A second assumption in Lv et al (2007) is that the evoked response (when present) either cancels out in the resampled data sets, or that its power is negligible. When this is not the case, then parameters generated from the resampled data sets will be biased towards a response, with the magnitude of the bias depending on the SNR of the response within the resampled data sets. As a result, the critical threshold for rejecting H_0 is increased, and test sensitivity is reduced. A possible solution is to approximate the evoked response with the ensemble coherent average, and to subtract it from all epochs prior to resampling (see Appendix, section A.5.2). A second solution (not explored in this work) is to randomly invert half of the epochs within the resampled data sets.

A final complication is the independence assumption between epochs, which takes two forms for the bootstrap. First, note that the random resampling with replacement procedure disrupts the correlation (if present) between the original epochs. The bootstrapped null distribution will then deviate from the true null distribution, with the extent of the deviation depending on the independence violation between the original epochs. Secondly, the resampling with replacement procedure may result in some EEG segments being selected multiple times, which introduces a new violation to the independence assumption, now between the epochs in the resampled data sets. The independence violation between the resampled epochs is further considered in the Appendix (section A.5.3), and a partial solution to independence violations between the original epochs is presented in section 3.6.1 below.

3.6.1 Bootstrapping in blocks

A more robust evaluation of test significance under independence violations between epochs (within the original ensemble) is to resample epochs in blocks, as opposed to resampling on an epoch to epoch basis. In particular, resampling in blocks preserves the correlations between epochs within each block, i.e. the bootstrapped ensembles will retain some degree of the original violation. Note however that the correlations between the resampled blocks is still disrupted. The extent to which the independence violation is preserved is therefore dependent on the number of epochs within each resampled block, which suggests that increasing the number of epochs within each block could be beneficial. This does however come at the price of a reduced variation in the starting positions of the resampled epochs. When this variation is too small, then (i) some epochs may remain partly time-locked to the stimulus, thus reducing test sensitivity, and (ii) insufficient variation amongst the resampled data sets may result in an inaccurate

bootstrapped null distribution. A potential solution to (i) might be to again subtract the ensemble coherent average from the epochs prior to resampling, or to randomly invert half of the resampled epochs within each bootstrapped ensemble. Bootstrapping in blocks for a more robust evaluation of test significance under independence violations is further considered in section 5.1.4.

3.6.2 Bootstrapping multiple features

The bootstrap approach gives the user a great deal of freedom when choosing which features to use for objective detection. Note that the bootstrapped feature can itself be a summary statistic, composed of multiple features or even of multiple statistical tests, e.g. the Fsp might be combined with the T^2 statistic through summation, in which case the bootstrapped statistic would be: $N(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})\mathbf{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})^H + N\frac{VAR(\bar{X})}{VAR(SP)}$.

The main challenge for bootstrapping multiple features is how to combine them, such that test sensitivity is optimised. An obvious disadvantage with summation, for example, is that it weights the summary statistic heavily in favour of features with large scales (the T^2 statistic is generally much larger than the Fsp). The core of the issue is however not that the feature scales are different, but that the underlying distributions of the features are not taken into account appropriately. In particular, the probability of observing the given feature values under H_0 needs to be considered, else the importance of outliers is underestimated, and the sensitivity of the summary statistic reduced.

In what follows, a variation of the standard bootstrap approach described in Lv et al (2007) is described, which allows multiple features to be combined appropriately. The approach is illustrated in Fig. 3.5, and uses the Fsp and the T^2 statistic again as example, i.e. the goal is to construct a single p value, representing the probability of observing the given Fsp and T^2 values under H_0.

Starting with data matrix D (see Fig. 3.5), bootstrapping first proceeds as usual by resampling M bootstrapped ensembles ($M = 1000$ here) from the continuous recording of D, giving bootstrapped ensembles D^*_1, D^*_2, ..., D^*_{1000}. Both the Fsp and the T^2-statistic are then calculated from each bootstrapped ensemble, giving Fsp values Fsp_1, Fsp_2, ..., Fsp_{1000}, and T^2 values T^2_1, T^2_2, ..., T^2_{1000}. These values are used to approximate the underlying null distributions for both the Fsp (plot A) and the T^2 statistic (plot B). Next, the observed Fsp values (Fsp_1, Fsp_2, ..., Fsp_{1000}) and T^2 values (T^2_1, T^2_2, ..., T^2_{1000}) are transformed into p values, achieved by finding their locations (percentiles) along their bootstrapped null distribution, giving p values p_{Fsp_1}, p_{Fsp_2}, ..., $p_{\mathrm{Fsp}_{1000}}$ for the Fsp, and p values $p_{T^2_1}$, $p_{T^2_2}$, ..., $p_{T^2_{1000}}$ for T^2. The resulting p values are then log-transformed, and combined through summation, giving combined values -ln(p_{Fsp_1})-ln($p_{T^2_1}$), -ln(p_{Fsp_2})-ln($p_{T^2_2}$), ..., -ln($p_{\mathrm{Fsp}_{1000}}$)-ln($p_{T^2_{1000}}$), from which the null distribution of the summary statistic is constructed (plot C). The Fsp and the T^2-statistic are then also calculated from the original ensemble D, and the resulting values are processed

using the same procedure: they are first transformed into p values (say p_{Fsp} and p_{T^2}) by finding their location along their bootstrapped null distributions, after which the p values are log-transformed and combined through summation, giving summary statistic $-2ln(p_{Fsp})$ $-2ln(p_{T^2})$. Finally, the significance of $-2ln(p_{Fsp})$ $-2ln(p_{T^2})$ is evaluated by finding its location (percentile) along its bootstrapped null distribution.

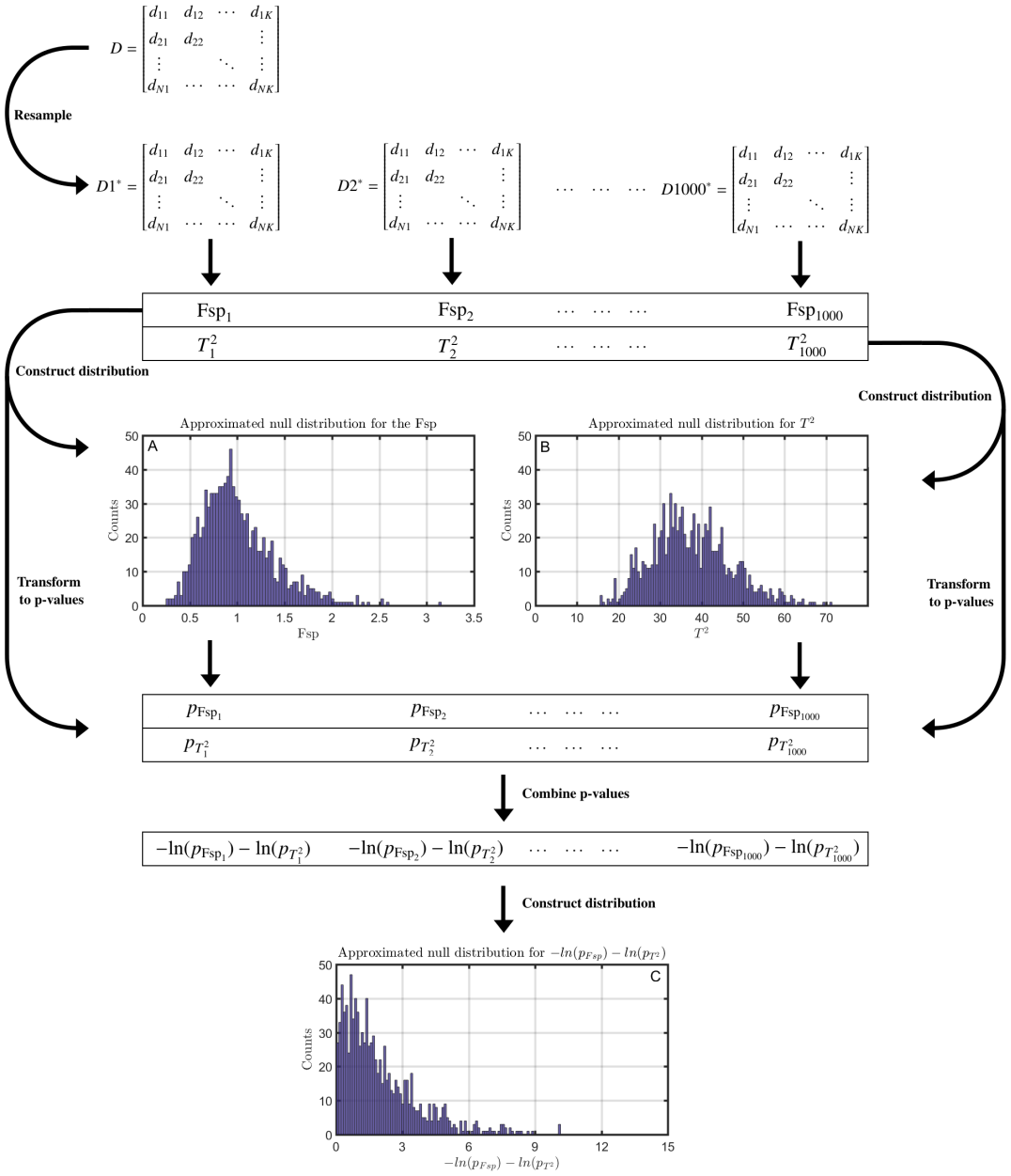


Figure 3.5: A variation of the bootstrap approach for evaluating the significance of multiple features and/or statistical tests simultaneously. In the example presented here, the goal is to evaluate the significance of the Fsp and the T^2 statistic, i.e. to generate a single p value, representing the probability of observing the given Fsp and T^2 values under H_0. Further details are presented in the text.

It is worth noting that if the individual features are independent, that the significance of each feature can be evaluated separately (using the standard bootstrap approach in Lv et

al, 2007), and the resulting p values combined using e.g. Fisher's method (Fisher, 1932). Independence between features and/or statistical tests is nevertheless quite unlikely, particularly so when the features are obtained from the same data set. Taking the correlation between p values into account is hence necessary, which is essentially the goal for the previously described bootstrap approach.

3.6.3 Bootstrapped parameters for objective detection

This section describes various bootstrapped features, including the 'Max Difference' and the 'Mean Power' statistics in Lv et al (2007), along with the CC, and a summary statistic composed of the Hotelling's T^2 test and the CC, called 'T2 Time + CC'. Previously described test statistics that are evaluated with the bootstrap approach include the Modified q-sample uniform scores test (section 3.5), and both the Fsp and the Fmp (section 3.1).

The Peak-to-Peak Difference

The Peak-to-Peak Difference (henceforth 'Max Diff'), is defined as the difference between the maximum and minimum value within the ensemble coherent average, and is given by (Lv et al., 2004; Lv et al., 2007):

$$\text{Max Diff} = max(\bar{X}) - min(\bar{X}) \tag{3.18}$$

Mean Power

The 'Mean Power' is given by the mean square of the ensemble coherent average (Lv et al., 2007):

$$\text{Mean Power} = \frac{1}{K}\sum_{i=1}^{K}\bar{X_i^2} \tag{3.19}$$

where $\bar{X_i}$ is the i^{th} value of the ensemble coherent average.

The correlation coefficient

The correlation coefficient (CC) gives the linear correlation between two variables (Pearson, 1895). It takes values ranging from -1 to 1, with -1 representing perfect negative

correlation (the variables follow identical but opposite trends around their respective means), 1 representing perfect positive correlation (the variables mirror each other perfectly around their respective means), and 0 representing no correlation at all, i.e. perfectly random. In this work, the CC is used to calculate the correlation between the ensemble coherent average $\bar{X}$ and some template $\bar{T}$, in which case the CC is given by:

$$\text{CC} = \frac{Cov(\bar{X}, \bar{T})}{\sigma_{\bar{X}} \sigma_{\bar{T}}} \tag{3.20}$$

where $Cov(\bar{X}, \bar{T})$ is the covariance between $\bar{X}$ and $\bar{T}$, $\sigma_{\bar{X}}$ is the standard deviation of $\bar{X}$, and $\sigma_{\bar{T}}$ is the standard deviation of $\bar{T}$.

T2 Time + CC

As the name suggests, the 'T2 Time + CC' is a summary statistic, composed of the Hotelling's T^2 test (applied in the time domain) and the CC. The statistic is constructed and evaluated using the bootstrap approach described in section 3.6.2, and is given by:

$$\text{T2 Time + CC} = -ln(p_{T2}) - ln(p_{CC}) \tag{3.21}$$

where p_{T2} is the p value obtained from the Hotelling's T^2 test (applied in the time domain) and p_{CC} the p value from the CC, where the CC represents the correlation between $\hat{X}$ and $\hat{T}$ (see Eq. 3.20). The p value p_{T2} can furthermore be generated by evaluating the significance of the T^2 statistic using either theoretical F-distributions or the bootstrap approach. Either way, the underlying null distribution for T2 Time still needs to be approximated (using the bootstrap), else the null distribution for the summary statistic cannot be constructed (see section 3.6.2).

Chapter 4

Data

This chapter provides a description of the data used throughout this work. The two most important data sets are (1) a relatively large database of no-stimulus EEG background noise recordings (data set **D1**), and (2) an ABR threshold series obtained from a small sample of normal hearing adults (data set **D2**). In addition to real data, many sections use simulations to explore, evaluate, and compare the performance of objective ABR detection methods. The data for these simulations consists of realistic simulated coloured noise for representing the EEG background activity, along with ABR templates for representing a response (section 4.4).

4.1 Data set D1: No-stimulus EEG recordings

Recordings of spontaneous EEG background activity (no stimulus was used) were previously collected by Madsen et al. (2017) and Madsen (2010) from 17 subjects (12 males and five females) under four conditions. The conditions were (i) *asleep*, where the subjects were asked to try and fall asleep, though sleep was not confirmed, (ii) *still*, where the subjects were instructed to lie still with their eyes closed, but not to fall asleep, (iii) *blink*, where the subjects were instructed to blink every 1-3 s as a circle appeared on a screen in front of them, and (iv) *move*, where the subjects were asked to move according to a random animation, also shown on a screen in front of them. Measurements were then obtained using a Compumedics Neuroscan II EEG amplifier at a sampling rate of 20 kHz with three silver–silver chloride (Ag/AgCl) electrodes placed on the left mastoid, the right cheek (ground) and the upper forehead (reference). The electrode impedances remained below 1 kΩ throughout the recording for all subjects. A total of 149 continuous EEG recordings were available, with an average of 6800 pre-processed epochs per recording, resulting in a grand total of ∼8 hours of EEG.

It should be noted here that no distinction is made throughout this work between the different noise conditions. This keeps the results concise and is justified as all four

conditions occur in clinical practice, and the methods should ideally perform adequately under each of them. To give an impression of how the different noise conditions may affect the EEG background activity, the variance of the recordings are presented as box-and-whisker diagrams in Fig 4.1 (Figure obtained from Madsen et al., 2017). The 'box' gives the intervals for the first and third quantiles, along with the median (center line), whereas the 'whiskers' show the minimum and maximum values (after outlier removal). As noted in Madsen et al (2017), the given variances were calculated from the full recordings, prior to artefact rejection, and hence represent the average or long term power of the EEG background activity per recording.

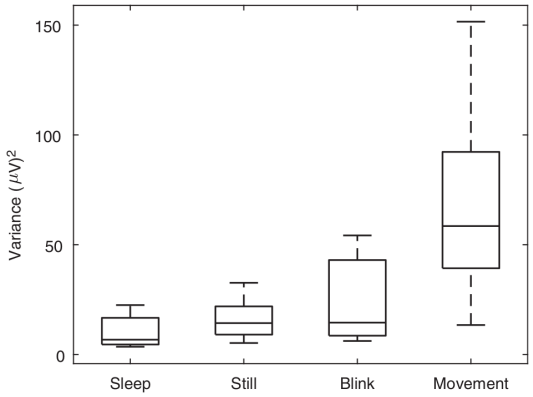


Figure 4.1: Box-and-whisker diagrams, representing the variance of the EEG recordings in data set D1, per noise condition. As noted by Madsen et al, each variance was calculated from the full recording, prior to artefact rejection, and thus represents the average or long term power of the EEG background activity. *Reprinted from Publication 'Accuracy of averaged auditory evoked potential amplitude and latency estimates', Vol. 57(2), Madsen S.M.K., Harte J.M., Elberling C. & Dau T., pp. 1-9 (2017), with permission from Taylor & Francis.*

4.2 Data set D2: Subject recorded ABR threshold series

Subject recorded ABR data, previously described in Lv, Simpson, and Bell (2007), was collected from 12 subjects (six female and six males) ranging from 18 to 30 years of age. The stimulus was a rectangular 100 μs click delivered at a stimulus rate of 33.3 Hz through ER-2 insert phones (Etymotic, Elk Grove Village, IL). The click intensities ranged from 0 to 50 dB SL (sensation level, i.e. relative to individual hearing thresholds) in steps of 10 dB. The behavioural thresholds were estimated using a simple "up-down" approach where the click intensity was reduced in steps of 10 dB for every correct response and increased in steps of 5 dB for every missed response. ABRs were recorded with the active electrode placed at vertex, a reference electrode at the nape of the neck and a ground electrode placed at mid-forehead. Measurements were obtained at a sampling rate of 10 kHz using a Cambridge Electronic Design (CED) micro 1401 data acquisition unit along with a CED 1902 amplifier. Electrode impedances remained below 5 kΩ throughout the recording. Approximately, 3600 clicks were delivered per subject and per stimulus condition.

As described below (section 4.3), the ensemble coherent averages of data set **D2** are used throughout this work to simulate a response. When doing so, it is important that the ensemble coherent averages do, in fact, contain a clear response (else the simulations would just be simulating noise). The criteria for a 'clear response' is further defined for data sets **D3** and **D4** below. The estimated SNRs for the ensemble coherent averages are furthermore shown in Table 4.1 below. These were generated using Eq. 4.1, where $P_{Template}$ is the mean square of the ensemble coherent average from the subject and dB SL condition in question, and P_{Noise} the mean square of the ensemble of epochs when treated as a continuous recording. The mean SNR (taken across subjects) is also presented per dB SL condition. It is worth noting that the residual background noise within the ensemble coherent average is not zero, which implies that the SNRs are likely over-estimated.

Table 4.1: The estimated SNRs for the ensemble coherent averages, for subjects S1 to S12, along with the mean SNR (taken across subjects, per dB SL condition). The SNRs were estimated using Eq. 4.1, where $P_{Template}$ is the mean square of the ensemble coherent average from the subject and dB SL condition in question, and P_{Noise} the mean square of the ensemble of epochs when treated as a continuous recording.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**		**Mean**
0 dB SL	-33.4	-38.4	-41.4	-34.7	-34.5	-34.7	-37.5	-34.5	-39.2	-35.1	-35	-36.1		-36.2
10 dB SL	-33	-34.5	-39.9	-37	-32.5	-26	-38.7	-31.6	-34.8	-32.8	-29.8	-33.1		-33.6
20 dB SL	-24.5	-31.6	-33.9	-30.5	-30.3	-26.6	-33.8	-31.4	-28.7	-32.4	-28.6	-25.6		-29.8
30 dB SL	-26	-31.7	-31.4	-31.1	-23.6	-27.1	-32.5	-27.4	-30	-32.1	-25.9	-24.9		-28.7
40 dB SL	-25.3	-30.8	-28.5	-28.7	-22.8	-26.3	-32.6	-28.3	-28.9	-28.2	-27.4	-23		-27.6
50 dB SL	-28.2	-30.9	-29.1	-23.3	-22.9	-34.6	-30.9	-30.3	-28.6	-25.7	-25.8	-23.3		-27.8

4.3 ABR templates

Various sections throughout this work use simulations to evaluate and compare the test performance of ABR detection methods. When simulating the stimulus condition, a response is represented by rescaling an ABR template, and adding it to noise. The ABR templates are obtained from the ensemble coherent averages from data set **D2**, under the condition that they contain a clear response. The criteria for a clear response is defined as either a significant detection (using $\alpha = 0.05$) with the Hotelling's T^2 test (data set **D3**), or using visual inspection by an experienced audiologist (data set **D4**). The scaling factors for the ABR templates are furthermore chosen such that a specific SNR is obtained, which is calculated using:

$$SNR = 10log_{10}\frac{p_{Template}}{p_{Noise}} \tag{4.1}$$

where $P_{Template}$ is the mean square of the scaled ABR template in question, and P_{Noise} the mean square of the ensemble of epochs (containing just noise). The chosen SNR for the simulated response is typically in the range of -23 to -28 dB, which was based on

both pilot simulations (which show a good coverage of detection rates for these SNRs without having to simulate excessively large ensembles) and on the estimated SNRs presented in Table 4.1 above.

4.3.1 ABR templates: data set D3

For the first set of ABR templates, the criteria for a 'clear response' was a positive detection (using an α-level of 0.05) with the Hotelling's T^2 test (applied in the time domain, using 25 TVMs). The 0 and 10 dB SL conditions were furthermore excluded entirely in an attempt to avoid templates contaminated by significant amounts of noise. Prior to calculating the T^2 statistic, data were band-pass filtered (from either 30-2000 Hz or from 100-2000 Hz) using a 3rd-order Butterworth filter (see Appendix section A.16 for further details on these filters), and artefact rejection was applied by throwing away 10% of the noisiest epochs, as determined by their maximum absolute values. The resulting templates for band-pass filter settings of 100-2000 Hz are shown in Fig. 4.2 for the 20, 30, 40, and 50 dB SL conditions, along with the grand coherent average (the mean of the subject coherent averages).

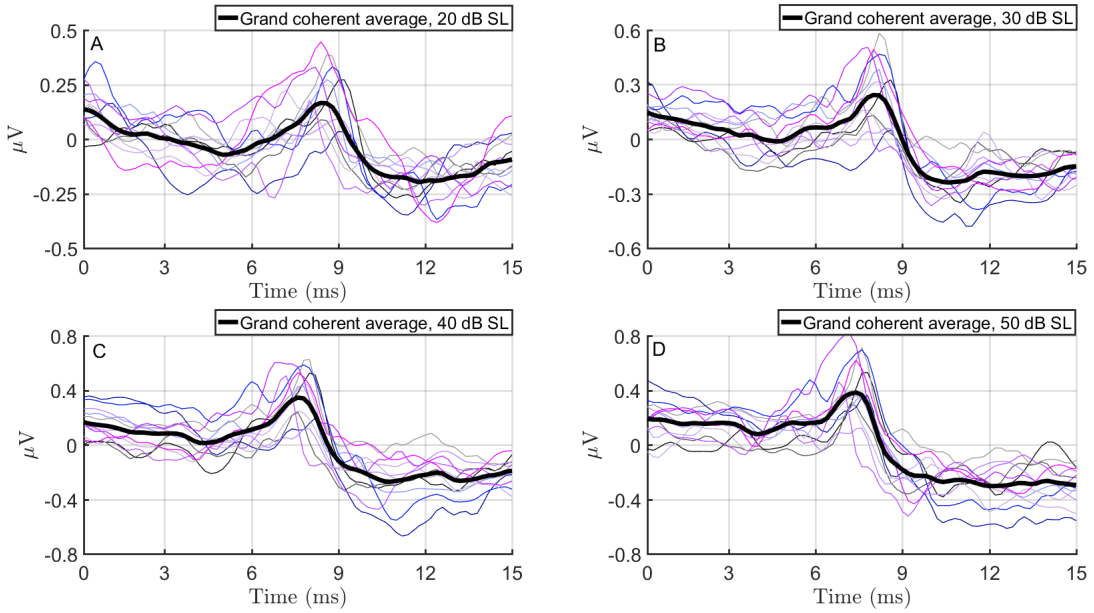


Figure 4.2: The ABR templates from data set **D3**, per dB SL condition. The templates were obtained from the subject ensemble coherent averages, under the condition that these contained a clear response. The criteria for a clear response was a significant ($\mathbf{p<0.05}$) detection by the Hotelling's T^2 test. The grand coherent average (the mean of the subject coherent averages) are also shown per dB SL condition.

4.3.2 ABR templates: data set D4

The second set of ABR templates were similarly obtained from data set **D2**, except that the criteria for a 'clear response' was now determined through visual inspection

by an experienced audiologist. As guidance for determining the presence of a clear response, the audiologist inspected the repeatability of the waveform by comparing two replicates of the coherent average (obtained by averaging across epochs 1-1500, and again across epochs 1501-3000). The audiologist also used the 3-1 signal to noise criterion as additional guidance (see Sutton et al. 2013), but was ultimately left free to decide whether a response was present or not. This process resulted in a total of 34 ABR templates with a clear response: 4, 7, 8, 7 and 8 from the 10, 20, 30, 40 and 50 dB SL conditions, respectively. The templates are presented in Fig. 4.3 (using band-pass filter settings of 100-2000 Hz) per dB SL condition, along with the grand coherent averages (taken across subjects), also per dB SL condition.

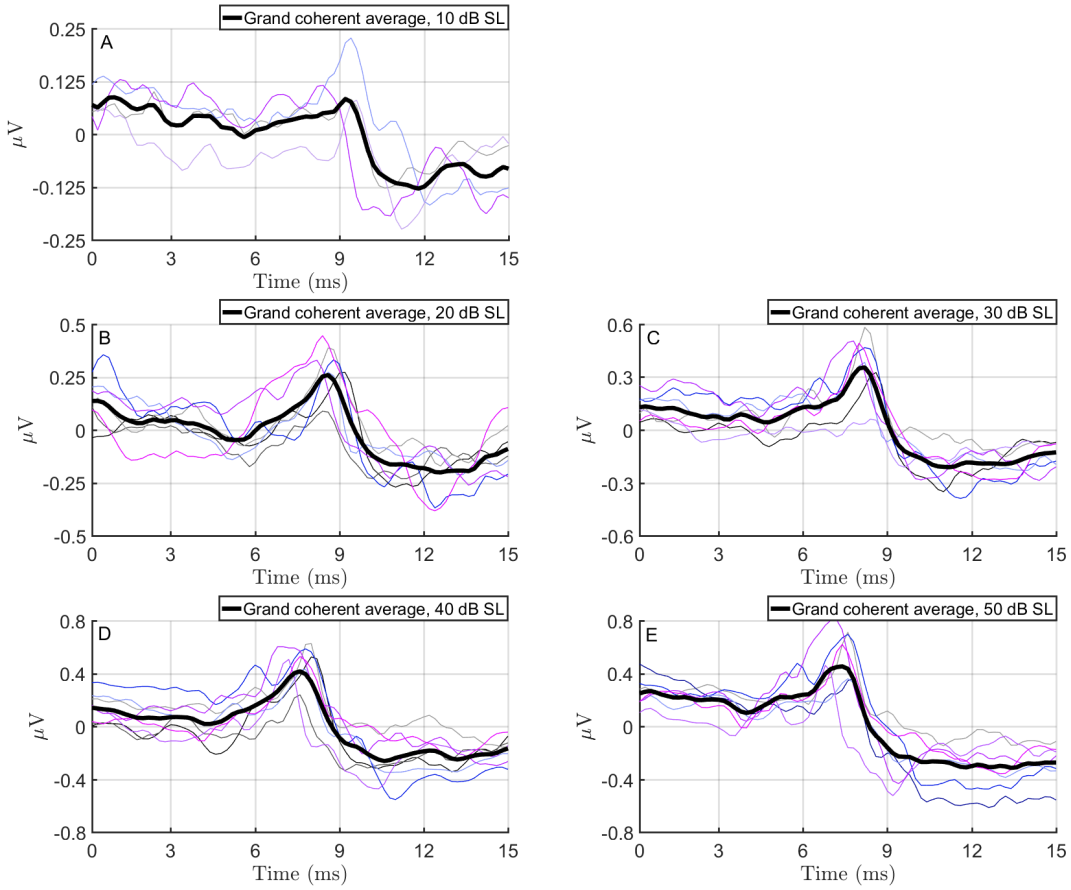


Figure 4.3: The ABR templates from data set **D4**, per dB SL condition. The templates were obtained from the subject ensemble coherent averages, under the condition that these contained a clear response. The latter was determined by an experienced audiologist (further details are presented in the text). The grand coherent average (the mean of the subject coherent averages) is also shown per dB SL condition.

4.4 Simulated coloured noise

Many simulations throughout this work use zero-mean stationary coloured noise to represent the EEG background activity. The coloured noise is generated by filtering Gaus-

sian white noise with an all-pole filter, where the poles of the filter are given by the parameters of an autoregressive (AR) model. The AR models were estimated from the recordings of EEG background activity (data set **D1**) using the Modified Covariance method (Marple, 1987), with a new AR model being fit to each recording. The general form for the resulting AR process is:

$$y(t) = \sum_{i=1}^{O} b_i x(t-i) \tag{4.2}$$

where O is the order of the model, $y(t)$ is the generated signal at time point t, O is the order of the model, and b_k (for $i = 1, 2, ..., O$) are the AR parameters (estimated from the original EEG recording being simulated). Many recordings (typically between 2000 and 50 000) of zero-mean stationary coloured noise are then generated and by filtering Gaussian white noise, using the b_i values as poles in an all pole filter.

The order of the AR models was determined by visually comparing the power spectral densities (PSDs) from the original EEG recordings to the simulated EEG recordings. A relatively high model order of $O = 60$ was then chosen to ensure a close match in terms of the spectral content of the original and the simulated recordings. An example is presented in Fig. 4.4: plot A shows the PSD, estimated using Welch's (1967) FFT method, from one of the original recordings of EEG background activity. Plot B then shows the PSD estimated from a simulated recording, which was simulated using the AR model estimated from the recording in plot A.

Limitations

An obvious shortcoming for this approach is that the simulated noise is stationary, which is not the case for real EEG background activity. Visual inspection of the simulated recordings and their PSDs also suggests that various artefacts, such as movement artefacts and the mains interference, are not simulated adequately. Simulating coloured noise is nevertheless still very useful, as it (i) allows more powerful evaluations and comparisons amongst methods to be drawn (large amounts of data can be generated), and (ii) it provides a more controlled environment for exploring various underlying assumptions of the ABR detection methods.

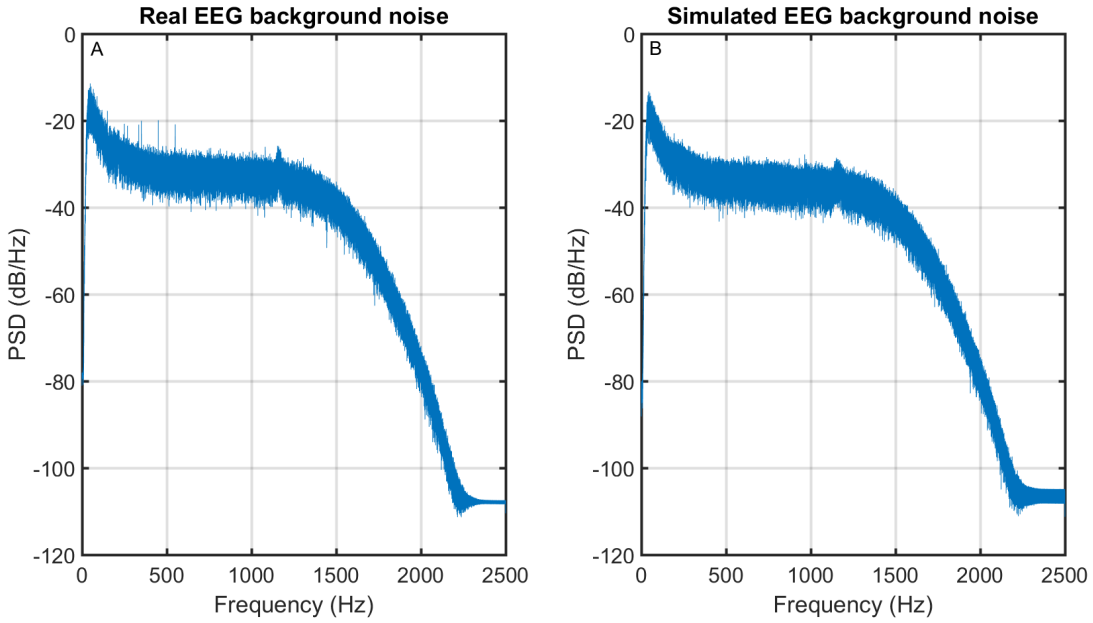


Figure 4.4: An example, illustrating the PSD estimated (estimated using Welch's 1967 FFT method) from one of the original recordings of EEG background activity (plot A) along with the PSD estimated from the corresponding simulated recording (plot B). Further details presented in the text.

Chapter 5

Specificity

The specificity of a test determines how 'specific' the test is when rejecting H_0, i.e. a highly specific test will reject H_0 only under what would be considered highly improbable conditions if H_0 were true. Specificity is therefore directly related to the rate at which H_0 is incorrectly rejected, that is, it is related to the FPR through 1 - specificity.

Specificity is specified *a priori* through the nominal significance level of the test α, which is the theoretical or assumed FPR. In practice, deviations from α can occur due to (i) random variations and (ii) violations to the statistical assumptions underlying the test. A convenient tool for evaluating these deviations is the binomial distribution, which can be used to construct confidence intervals for α. When the observed FPR falls outside the confidence intervals, then a violation of the underlying statistical assumptions is more probable, with the extent of the probability depending on the coverage of the confidence intervals (for more on the binomial distribution and confidence intervals for α, see the Appendix, section A.2).

The aim for this Chapter is to explore the extent to which the main three statistical assumptions underlying most ABR detection methods are satisfied for EEG measurements. These include the independence assumption between epochs (section 5.1), (ii) the normality assumption (section 5.2), and (iii) the stationarity assumption (section 5.3). In the case of significant violations, various solutions are explored for removing of compensating for the violation, with the overall goal of obtaining a more robust control of specificity. The sphericity assumption underlying RM ANOVA is not explored in this Chapter, but is instead addressed in sections A.3 and A.4 of the Appendix. Various assumptions underlying the bootstrapped statistics are also not included, but are considered in section A.5 of the Appendix.

5.1 Independence

Independently and identically distributed (i.i.d.) data implies that the probability of observing any set of data is unaffected by any previously collected data. In terms of conditional probabilities, the i.i.d. assumption is defined as (Jean-Yves Le Boudec, 2015, p39):

$$\Pr(X_i \in A \mid X_1 = x_1, ..., X_{i-1} = x_{i-1}) = \Pr(X_i \in A) \tag{5.1}$$

where A is any set of real data of size N with elements X_i and values x_i (for $i = 1, 2, ..., N$).

Violations to the independence assumption can result in an under- or overestimation of both the sample variance and the DOF of the data (Thiébaux, 1984). This results in a mismatch between the assumed theoretical null distribution and the true null distribution of the test statistic, resulting in a conservative or a liberal test performance.

The independence assumption underlying objective ABR detection methods can take two forms: (i) independence between epochs, and (ii) (some degree of) independence between samples within epochs. Independence between epochs is assumed by more or less all known ABR detection methods. Some statistics (e.g. the Fsp, the Fmp, and the CC) make the additional assumption of some degree of independence between samples within epochs. In particular, these statistics assume the DOF of the samples within epochs when evaluating test significance using some theoretical distribution (see also methods section). As mentioned before, this is problematic, as the DOF of the data can vary both within and between recordings, and is typically not known in advance.

The remainder of this section is structured as follows: a brief literature review on studies that have looked at independence violations for auditory evoked response detection in the past is first presented in section 5.1.1 below, after which some results from a brief exploratory analysis using the Turning Point test (Heyde & Seneta, 1972; Bienaymé, 1874) are described in section 5.1.2. Follow up simulations are then conducted with the goal to quantify potential independence violations in terms of increased or decreased FPRs for the Hotelling's T^2 test (section 5.1.3). Finally, section 5.1.4 presents some early results from the 'bootstrapping in blocks' approach, which is used for a more robust evaluation of test significance under independence violations.

5.1.1 Literature review

Despite the potentially harmful effects from independence violations, few authors have investigated it for auditory evoked response detection. Geisler (1960) looked at the

autocorrelogram of a 6 minute 8-600 Hz band-pass filtered EEG recording, obtained from a subject playing chess, to which 40 dB clicks were presented at a rate of 5 clicks per second. Samples separated by less than 10 ms were quite strongly correlated, but autocorrelations decreased rapidly for larger distances (Fig. 5.1).

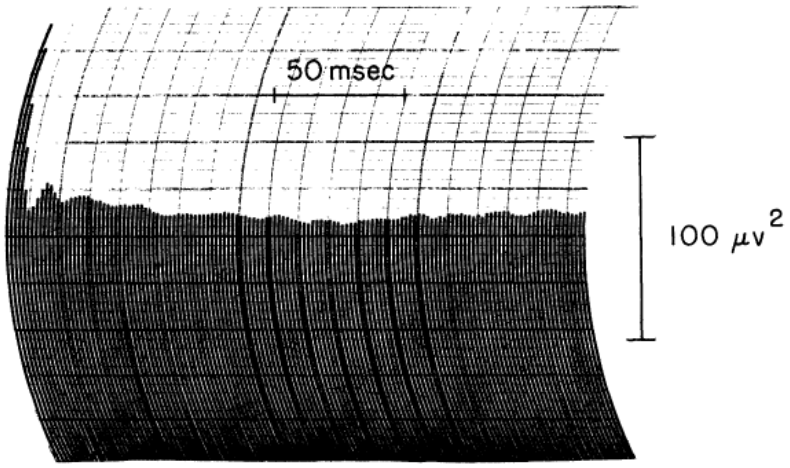


Figure 5.1: Figure from: Geisler C.D. 1960. Average responses to clicks in man recorded by scalp electrodes. *Massachusetts Institute of Technology, Research Laboratory of Electronics.* Technical report 380. The autocorrelogram of a 6 minute 8-600 Hz band-pass filtered EEG recording, obtained from a subject playing chess, to which 40 dB clicks were presented at a rate of 5 clicks per second.

Others have observed significant autocorrelations within EEG recordings for much longer periods of time. Neely & Pepe (1997) looked at 65 dB SPL clicks in 14 babies, and observed significant autocorrelations for samples separated by distances of up to 3 seconds or more. They do, however, also mention a 60 Hz line noise along with low frequency spectral peaks of 10.7 and 11.8 Hz, and do not mention filtering the data prior to the analysis (note that filtering affects the spectral content of the data, which, in turn, determines the autocorrelation function). Victor & Mast (1991) looked at the required length of adjacent EEG segments before the Fourier components of some spectral band could be considered independent. They show that it is safe to assume independence when the power spectra around the frequency in question is approximately flat within the frequency window $\frac{2\pi}{L}$, where L is the length of the (adjacent) EEG segments on which the Fourier analysis was performed (Victor & Mast, 1991; Mast & Victor, 1991). Victor & Mast (1991) calculated L for 16 frequencies in the range of 2.5 to 40 Hz, and found that the required length for independence of adjacent EEG segments was in the range of 3 to 6 seconds.

The time required for independence observed by both Neely & Pepe (1997) and Victor and Mast (1991) is on a different scale as that observed by Geisler (1960). A possible explanation might be found in the spectral content of the data, i.e. data from Geisler was band-passed filtered from 8-600 Hz, whereas Victor & Mast considered spectral bands as low as 2.5 Hz, and Neely & Pepe presumably applied no filtering. The additional time

required for independence observed by Victor & Mast and Neely & Pepe may therefore be due to lower frequencies in their data. Alternatively, the different time scales might be due to different methodologies, i.e. Geisler visually inspected the autocorrelogram, whereas Victor & Mast looked at the power spectra around specific frequencies as a function of the length of adjacent EEG segments. In the following section, the independence assumption is explored in more detail using the Turning Point test, and results are discussed and compared to the aforementioned studies.

5.1.2 Exploring independence violations: the Turning Point test

It can be expected that independence between samples will depend on the distance in time between the samples, along with the dominant frequency within the data (the frequency with the largest amplitude), which is determined primarily by the high-pass cut-off frequency. Independence in this section is therefore explored as a function of the high-pass cut-off frequency and the separating distance between consecutive samples. The latter is achieved using the Turning Point test, which is a simple non-parametric test, applied to the number of local minima and maxima (called 'turning points') in a series of observations. In particular, for a series of N observations $x(i), x(i+1), ..., x(N)$, a local maxima at index j is defined as $x(j-1) < x(j) > x(j+1)$, and a local minima as $x(j-1) > x(j) < x(j+1)$. The expected number of turning points for i.i.d. data is $\frac{1}{3}(2N-1) \approx \frac{2}{3}N$, with expected variance $\frac{16N-29}{90}$ (Heyde & Seneta, 1972; Bienaymé, 1874).

Method

Data for the analysis consists of recordings of EEG background activity from data set **D1**. For each recording in **D1** (149 total), additional recordings were constructed by randomly resampling 20 000 consecutive samples from within the original recording, where each sample was separated from its neighbouring samples by τ ms. The distance τ was then varied from 0.2 to 200 ms, in steps of 0.2 ms (there were 1000 values for τ). The band-pass filter settings (for data set **D1**) were furthermore set to either 30-2000 Hz, or to 100-2000 Hz (filtering was realised using a 3rd-order Butterworth filter, see also the Appendix section A.16 for further details on the filters). The resampled recordings were then tested for independence using the Turning Point test. Under H_0 (data is i.i.d), the PDF for the mean number of turning points for $N = 20000$ samples is approximately normal with mean $\frac{20000 \cdot 2}{3} \approx 13333$ and standard deviation $\sqrt{\frac{16 \cdot 20000 - 29}{90}} \approx 59.63$. The two-sided 95% CIs for the expected number of turning points are therefore $[13216, 13450]$, or in terms of the percentage of tuning points: $[66.08\%, 67.25\%]$ (with mean $\approx 66.67\%$, see Heyde & Seneta, 1972; Bienaymé, 1874): .

Results

The results from the Turning Point test are presented in Fig. 5.2. The two upper plots show the percentage of turning points, for each resampled recording, as a function of the distance in ms τ between samples. The data (in data set **D1**) was band-pass filtered

at either 100-3000 Hz (plot A) or 30-3000 Hz (plot B). The two lower plots show the detection rates (across the 149 recordings) for the Turning Point test, i.e. the fraction of tests where independence was significantly violated (at $\alpha = 0.05$), similarly as a function of the distance between samples, and where data were band-pass filtered at either 100-3000 Hz (plot C) or 30-3000 Hz (plot D).

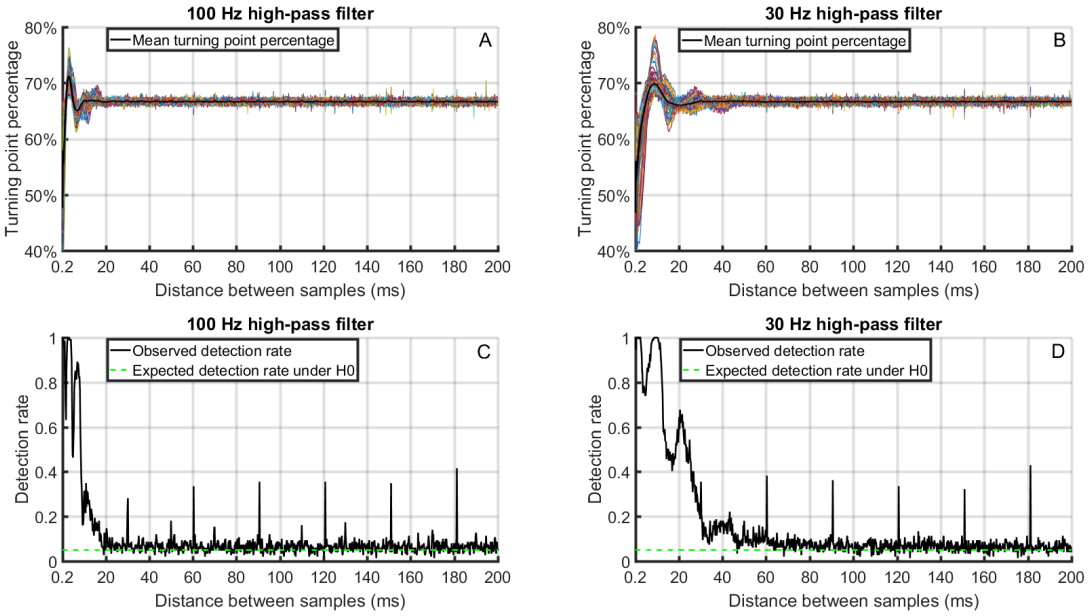


Figure 5.2: Results from the Turning Point test for testing the independence assumption between samples. The two upper plots (plots A and B) show the percentage of turning points, per resampled recording, as a function of the distance in ms between samples, where data were band-pass filtered at either 100-3000 Hz (plot A) or 30-3000 Hz (plot B). The two lower plots (plots C and D) show the fraction of tests (149 in total) where the independence assumption was significantly violated (at $\alpha = 0.05$), similarly as a function of the distance between the samples, and where the data were band-pass filtered at either 100-3000 Hz (plot C) or 30-3000 Hz (plot D).

Discussion

Visual inspection of plots A and B in Fig. 5.2 suggests that independence is satisfied after ∼25 ms when using a 100 Hz high-pass cut-off frequency, and after ∼70 ms when using a 30 Hz high-pass cut-off frequency, corresponding to a ∼40 Hz and a ∼14.29 Hz stimulus rate, respectively. The regular peaks at ∼30, ∼60, ∼90, ∼120, ∼150, ∼180 ms might be attributed to the harmonics of a 50 Hz mains. Overall, results appear to agree with Geisler (1960), who similarly observed autocorrelations between samples separated by ∼50-70 ms (when using a band-pass filter of 8-600 Hz). The different time-scales for independence observed by Victor & Mast (1991) and Neely & Pepe (1997) might therefore be due to different methodologies, e.g. Victor & Mast looked at the required length of *adjacent* EEG measurements before independence was satisfied. Note therefore that the initial correlations between adjacent EEG measurements was not disrupted in Victor & Mast. The independence assumption will hence never be completely satisfied using this approach, i.e. the best one can hope for is to attenuate or 'drown out' the violation by expanding the sample with a sufficiently large amount of i.i.d. data.

5.1.3 Quantifying independence violations

This section explores the extent to which independence violations are relevant for ABR detection methods, achieved by quantifying the violation in terms of increased or decreased FPRs for the Hotelling's T^2 test. The Hotelling's T^2 test is chosen primarily due to (i) fast processing times, and (ii) because realistic EEG data can be simulated, such that all assumptions underlying the Hotelling's T^2 test are satisfied, except the independence assumption between epochs. Independence violations are again explored as a function of the high-pass cut-off frequency and the separating distance between samples (now expressed as a stimulus rate).

Method

Data for the assessment consists of simulated Gaussian, stationary, zero-mean noise with similar spectral content to real EEG background activity, constructed as described in section 4.4. A total of 50 000 recordings were simulated, which were band-pass filtered from f_c to 2000 Hz using a 3^{rd}-order Butterworth filter. The cut-off frequency f_c was varied from 30 to 100 Hz, in steps of 5 Hz. Each filtered recording was then structured into 15 ms epochs, where the distance between epochs was varied from 0 to 40 ms, in steps of 0.4 ms. Note that the analysis window remains constant at 15 ms (it is just the distance between the 15 ms windows that is varied). The latter is related to a (hypothetical) stimulus rate using $\frac{1000}{15+\tau}$, where τ is the distance between the 15 ms windows. The ensemble size was furthermore set to 200 epochs. All resulting ensembles were then analysed using the Hotelling's T^2 (applied in the time domain, using 25 TVMs).

Results

The FPRs (using $\alpha = 0.05$) for the Hotelling's T^2 test are plotted in Fig. 5.3 as a function of the (hypothetical) stimulus rate and the high-pass cut-off frequency. The binomial distribution (section A.2) was used to construct two-sided 95% confidence intervals for the theoretical 0.05 FPR. The large number of tests performed (50 000 total) resulted in relatively narrow confidence intervals, with a lower limit of 0.0481 and an upper limit of 0.0519. FPRs that fell outside the expected boundaries are indicated in Fig. 5.3 by blue (FPR < 0.0481) and red (FPR > 0.0519) cells, whereas the FPRs that fell within the 95% CIs are indicated by green cells.

Discussion

Results (Fig. 5.3) demonstrate potentially large violations to the independence assumption between epochs when using specific combinations of f_c and the stimulus rate, resulting in both conservative and liberal test performances. In particular, when the epochs are positively correlated, the performance of the Hotelling's T^2 test will tend to be liberal, whereas when the epochs are negatively correlated, it will tend to be conservative. These results hence emphasize the importance of choosing suitable values for f_c and the stimulus rate, as an incorrect choice can potentially result in relatively large deviations from the nominal α-level, i.e. using $f_c = 65$ Hz and a 66.67 Hz stimulus rate

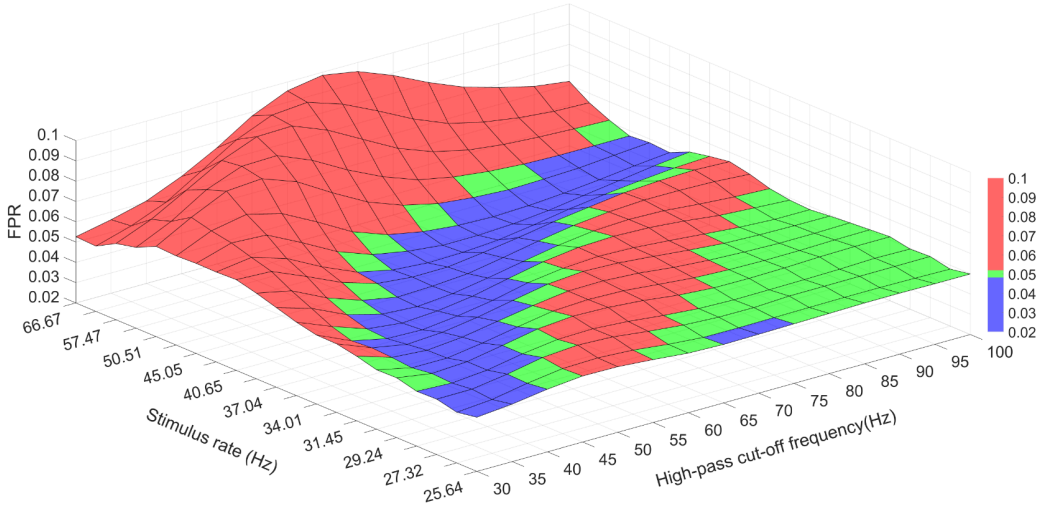


Figure 5.3: The FPRs for the Hotelling's T^2 test ($\alpha = 0.05$), as a function of the (hypothetical) stimulus rate and the high-pass cut-off frequency. Each FPR was generated from 50 000 tests, where the data for the tests consists of simulated Gaussian, stationary, zero-mean noise with similar spectral content to real EEG background activity (details presented in the text). The 95% two-sided confidence intervals for $\alpha = 0.05$ are [0.0481, 0.0519]. FPRs that fall outside the expected boundaries are indicated by blue (FPR < 0.0481) and red (FPR > 0.0519) cells, whereas FPRs that fall within the 95% CIs are indicated by green cells.

resulted in a FPR of 0.0985, as opposed to the theoretical 0.05. It is also worth noting here that a similar simulation was conducted for CAEP detection. Results (presented in the Appendix, section A.15) also demonstrate a relationship between the FPR, the high-pass cut-off frequency f_c, and the stimulus rate.

5.1.4 Compensating for independence violations

The section briefly explores whether 'bootstrapping in blocks' (section 3.6.1) can be used for a more robust assessment of test significance under independence violations.

Method

Data for the assessment was identical to section 5.1.3 above, except that the number of simulated recordings was reduced to 10 000 (due to relatively long processing times). The high-pass cut-off frequency was now also fixed at 65 Hz, and the (hypothetical) stimulus rate took values of 66.67, 58.82, 62.63, 47.62, and 43.48 Hz. These values were chosen based on results from the previous section (Figure 5.3), which showed relatively large independence violations (giving both liberal and conservative test performances) when using these values. The data were again analysed using the Hotelling's T^2 test, which was evaluated using either theoretical F-distributions, or with the bootstrap approach. When using the bootstrap approach, epochs were resampled in either blocks of two or in blocks of four epochs, i.e. each resampled windows had a duration of either 60.06 ms (two 30.03 ms epochs) or 120.12 ms (four 30.03 ms epochs).

Results

The observed FPRs for the Hotelling's T^2 test (evaluated using either theoretical F-distributions of with the bootstrap) under significance independence violations are presented in Fig. 5.4. The nominal level $\alpha = 0.05$ and the two-sided 95% CIs for α (given by [0.0459, 0.0544]) are also shown. Results suggest that the FPRs appear to approach the nominal α-level of the test as the number of epochs per block are increased.

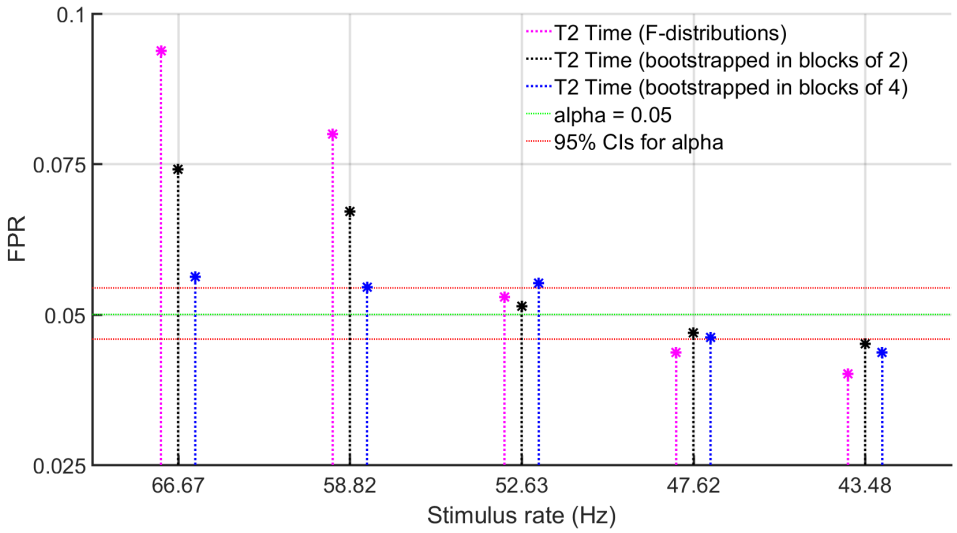


Figure 5.4: The FPRs for the Hotelling's T^2 test for various (hypothetical) stimulus rates when evaluated using either theoretical F-distributions or the bootstrap approach, where random resampling was performed in blocks of epochs or in blocks of four epochs.

Discussion

Results (Fig. 5.4) suggest that 'bootstrapping in blocks' might be a viable option for obtaining a more robust evaluation of test significance under independence violations. As mentioned in section 3.6.1, the advantage of resampling in blocks of epochs is that the correlations between epochs within blocks is preserved. The correlations between blocks, however, is still disrupted. Hence, although deviations from the nominal α-level are significantly reduced, specificity is (at least in theory) still not completely controlled as intended.

Robustness to independence violations might therefore be further improved by increasing the number of resampled epochs per block. However, this comes at the price of a reduced variation in the starting positions of the epochs, which might result in (i) a reduced test sensitivity, as some epochs may remain time-locked to the stimuli, or (ii) insufficient variation in the resampled data sets, which may result in an inaccurate approximation of the null distribution. A limitation for this section is furthermore that the approach was evaluated exclusively under significant independence violations. In future work, the approach should be tested across a more diverse set of test conditions, which should include a stimulus condition to ensure that test sensitivity is not reduced.

5.2 Normality

The goal for this section is to explore the extent to which normality is violated for EEG measurements, and to quantify potential violations in terms of increased or decreased FPRs for the Hotelling's T^2 test. An additional goal is to explore whether violations can be reduced through artefact rejection, or compensated for through Central Limit Theorem (CLT: see Appendix, section A.1) by increasing the ensemble size.

The underlying distribution of a Q-dimensional set of observations $\boldsymbol{x}$ is said to be multivariate normal (MVN) when its density function $f(\boldsymbol{x})$ is described by (Rencher, 2001, p.83):

$$f(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^Q \mid \boldsymbol{\Sigma} \mid^{0.5}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{5.2}$$

where $\boldsymbol{x}$ is the Q-dimensional vector of observed feature means, $\boldsymbol{\mu}$ the Q-dimensional vector containing the true mean values of the Q features, and $\boldsymbol{\Sigma}$ is the true underlying Q-dimensional covariance matrix for the Q features. The normal distribution can also be characterized by its skewness and kurtosis. The kurtosis of a distribution is a measure of how 'peaked' it is, i.e. whether its volume is concentrated around a single peak or whether it is spread out across a larger interval, whereas the skewness of a distribution is a measure of its asymmetry around its mean. The normal distribution has a kurtosis of three and a skewness of zero.

Deviations from normality are a concern for ABR detection firstly due to the assumption (underlying e.g. the Hotelling's T^2 test) that the population feature means and variances are independent, which is only true when the underlying distribution is normal. A second concern is in regards to the theoretical null distribution, which can deviate from the true null distribution when normality is violated, resulting in a conservative or a liberal test performance. That said, deviations from normality are not always an issue, as these can be compensated for through CLT by means of a sufficiently large sample size (Appendix, section A.1). The caveat is the term 'sufficiently large', i.e. it is not clear how large the sample should be before deviations from normality become negligible. The latter is typically left unanswered by authors, or as noted by Mordkoff (2016), authors tend to assume the effects are negligible and then 'look away and whistle'.

5.2.1 Exploring normality violations

This section provides a very brief exploratory analysis, simply by plotting the histograms of the samples of the recordings. The recordings of EEG background activity (data set **D1**) were first band-pass filtered from 30 to 2000 Hz using a 3rd-order Butterworth

filter (see Appendix A.16), and structured into 30.03 ms epochs. Artefact rejection was then applied by throwing away 10% of the noisiest epochs (as determined by their mean square values), and a histogram was constructed from the resulting samples, both before and after artefact rejection.

Results

The histograms of the samples for two subjects are shown in Figure 5.5, both before (plots A and B) and after (plots C and D) artefact rejection. Note that the x-axis in Figure 5.5 was determined by the smallest and largest sample values within the recording in question. Results show that kurtosis is quite extreme prior to artefact rejection (the tails of the histograms are long), but is greatly reduced by artefact rejection. Visual inspection of plots C and D suggests that the histograms are almost perfectly normal after artefact rejection. In the following section, the extent to which these violations are relevant for ABR detection is further explored.

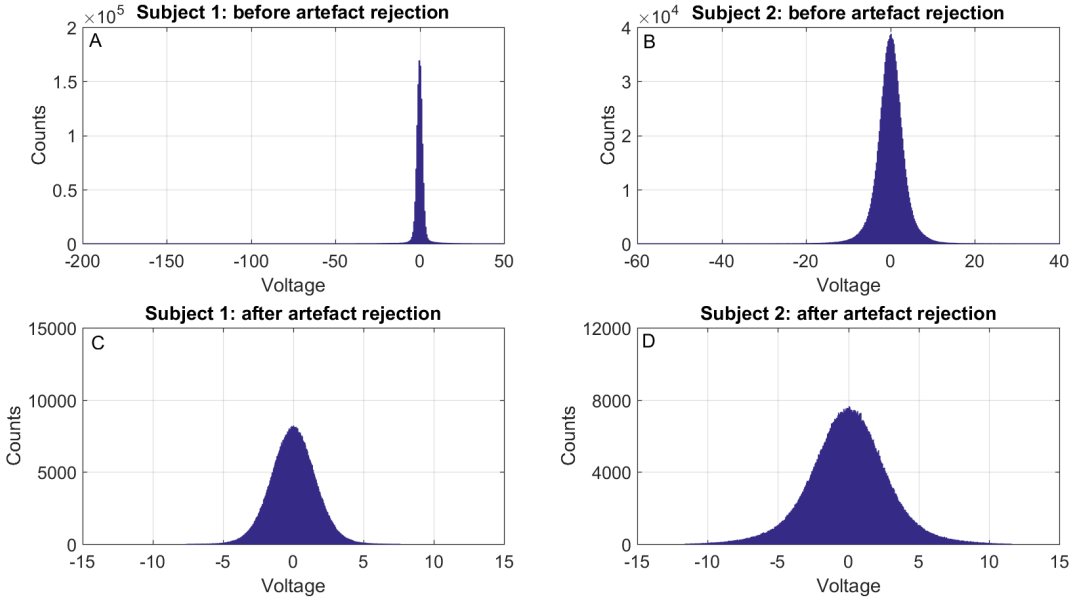


Figure 5.5: Histograms, constructed from two recordings of EEG background activity, both before artefact rejection (plots A and B) and after artefact rejection (plots C and D). The x-axis was determined by the smallest and largest sample values within the recording in question.

5.2.2 Quantifying and compensating for normality violations

This section explores the extent to which normality violations are relevant for ABR detection, achieved by quantifying the violation in terms of increased or decreased FPRs for the Hotelling's T^2 test. Additional goals are to test whether potential violations can be removed for through either artefact rejection or compensated for through CLT (by increasing the ensemble size).

Method

In order to isolate normality violations, stationarity and independence violations need to

be avoided. The latter is achieved by fitting distributions to the samples, and resampling from the fitted distributions. To do so, each recording of EEG background activity from data set **D1** was first compressed into TVMs by taking the mean across each 0.6 ms segment. An Epanechnikov kernel was then fit to the PDF of the resulting TVMs, which was repeated per recording. For each fitted distribution, 10 000 additional 25-dimensional feature sets of size N were simulated, where N took values of either 100 or 500. This procedure was applied both before and after artefact rejection: artefact rejection was applied (prior to compressing the recordings into TVMs) using the same approach described in previous sections, i.e. the recordings were structured into 30.03 ms epochs, and 10% of the noisiest epochs (as determined by their mean square values) were discarded. The resampled feature sets were analysed with the Hotelling's T^2 test.

Results

The FPRs are presented in Fig. 5.6 as a function of the recording index being simulated (there were 149 recordings in total). Plot A first shows the FPRs when using either $N = 100$ or $N = 500$, as a function of the recording index being simulated. Note that artefact rejection was not used for these simulations. The nominal α-level and its approximate two-sided 95% CIs (given by [0.0459, 0.0544]) are also shown. Results demonstrate an overall tendency towards a conservative test performance for both $N = 100$ and $N = 500$. Plot B then shows the FPRs for $N = 100$, where the distributions were now fit after artefact rejection. Results show that the FPRs mostly fall within the two-sided 95% CIs for α, with the exception of a few recordings where the FPR was now exceptionally high.

Post-hoc exploration

A post-hoc analysis was conducted to explore why artefact rejection increased the FPR for some recordings. Results show that the increased FPRs can likely be attributed to the mean of the recording, which was shifted away from zero due to artefact rejection (further clarified below). The latter is illustrated in Fig. 5.7, which shows the mean of the recordings of EEG background activity both before and after artefact rejection.

Summary

When *no artefact rejection* is used, results demonstrate a tendency towards a conservative test performance, which can likely be attributed to excessive kurtosis (resulting in an overestimation of the sample variance). For some recordings, the conservative test performance was quite drastic, i.e. for recording 73, a FPR of 0.0161 was observed (for $N = 100$), as opposed to the theoretical 0.05. The mean FPR (calculated across all recordings) was nevertheless still close to the nominal α-level: for the $N = 100$ condition, the mean FPR was 0.0445, whereas for $N = 500$ it was 0.0456. Note therefore that the effects of normality violations might be difficult to detect when considered across a cohort of recordings. Note also that increasing N from 100 to 500 was insufficient for compensating for the violation through CLT.

When *artefact rejection was used*, the distributions are more or less perfectly normally

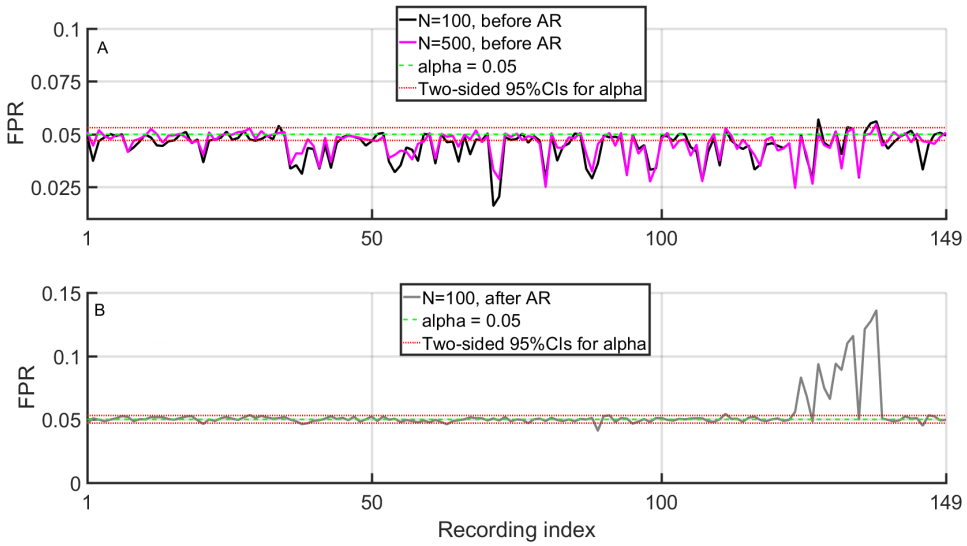


Figure 5.6: The FPRs for the Hotelling's T^2 test, as a function of the recording being simulated. **Plot A**: the ensemble size N was set to either 100 or 500, and no artefact rejection was used. **Plot B**: the ensemble size N was set to 100, and artefact rejection was used. The nominal α-level ($\alpha = 0.05$) and its two-sided 95% CIs are also shown.

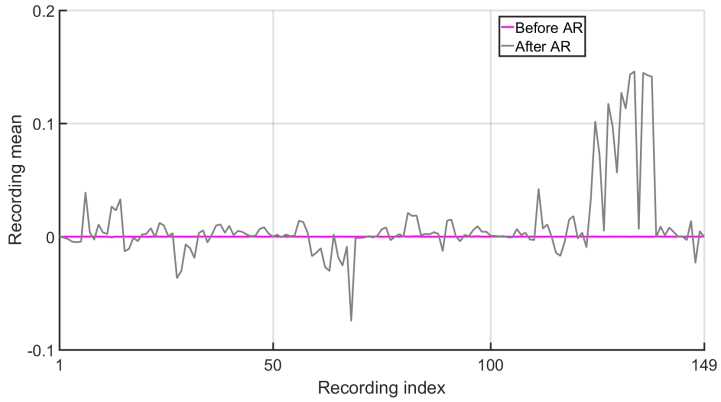


Figure 5.7: The means of the recordings of EEG background activity, both before and after artefact rejection.

distributed (Figure 5.5). The mean of the recordings, however, were now sometimes shifted away from zero (Figure 5.7), resulting in a liberal test performance for some recordings (Figure 5.6, plot B). The non-zero mean for some recordings following artefact rejection can likely be attributed to a combination of skewness and kurtosis, i.e. when the voltage measurements of the outliers are pre-dominantly negative, then artefact rejection results in more negative samples being removed relative to positive samples. The mean of the recording (which is originally zero due to the high-pass filter) is hence increased, resulting in a liberal test performance. For some recordings, the liberal test performance was quite drastic, i.e. for recording 137, the FPR was 0.1359, as opposed to 0.05. The mean FPR (taken across all recordings) was nevertheless again relatively close to the nominal α-level, and was equal to 0.0539.

5.3 Stationarity

Data can be considered stationary when its mean, variance, and autocorrelation are constant over time. More formally, stationarity is satisfied when the joint distribution of $[X(t_1+\tau), X(t_2+\mu), ..., X(t_n+\tau)]$ is independent of time shift τ for any time sequence $t_1 < t_2 < ... < t_n$, where $X(t_i+\tau)$ is the observed value at time $t_i+\tau$ (Le Boudec, 2015, p.223).

It is well known that the EEG background activity can change significantly within and between recordings. Violations to the stationarity assumption are hence commonplace for EEG data analysis. The goal for this section is (i) to briefly demonstrate the extent to which stationarity is violated for EEG background activity, (ii) to quantify the violation in terms of increased or decreased FPRs for the Hotelling's T^2 test, and (iii) to evaluate a data transformation (normalising the epoch variances) for removing the violation.

5.3.1 Exploring stationarity violations

This section briefly demonstrates the extent to which stationarity is violated for EEG background activity. The recordings of EEG background activity (data set **D1**) were downsampled to 5kHz, band-pass filtered from 30-2000 Hz using a 3rd-order Butterworth filter (Appendix A.16), and structured into 30.03 ms epochs. Artefact rejection was then applied by throwing away 10% of the noisiest epochs, as determined by their absolute maximum values. For each recording, the variance was calculated per epoch, both before and after artefact rejection. The resulting epoch variances are plotted as a function of time (seconds) for two subjects in Fig. 5.8, both before and after artefact rejection. As expected, visual inspection suggests relatively large violations to the stationarity assumption, which was reduced by artefact rejection.

5.3.2 Quantifying and compensating for stationarity violations

This section explores the extent to which violations to the stationarity assumption are relevant for evoked response detection, achieved by isolating the stationarity violation, and quantifying it in terms of increased or decreased FPRs for the Hotelling's T^2 test. An additional goal is to test whether specificity can be improved by removing the violation through normalisation of the epoch variances.

Method

Data consists of simulated Gaussian, stationary, zero-mean coloured noise with similar spectral content to real EEG background activity, generated as described in section 4.4. Each recording of EEG background activity (149 total) was used to generate 10 000 additional simulated recordings (using the AR model estimated from the recording in question; see section 4.4). The simulated recordings were then structured into ensembles

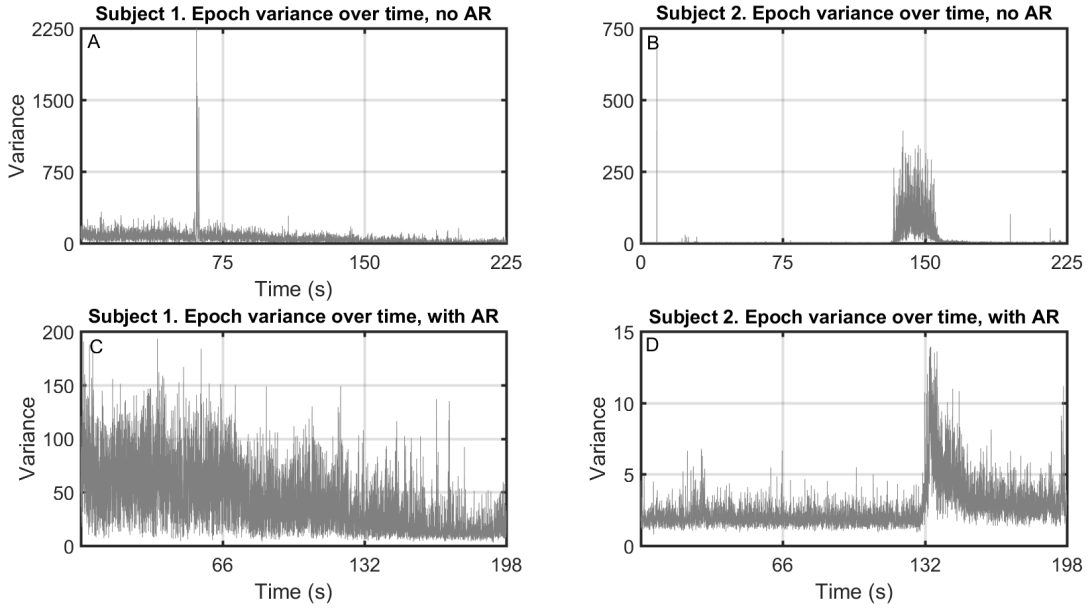


Figure 5.8: The epoch variances over time for two subjects, both before artefact rejection (plots A and B) and after artefact rejection (plots C and D).

of $N = 500$ 30.03 ms epochs. Note that this data is stationary. A violation to the stationarity assumption was therefore introduced by rescaling the epochs. In particular, each epoch was scaled by a specific factor, where the scaling factors were given by the epoch variances calculated from 500 randomly selected (but consecutive) epochs from within the original recording (i.e. the recording from which the AR model in question was obtained). The initial 15 ms of the ensembles were then analysed with the Hotelling's T^2 test using 25 TVMs as features, both before and after introducing the stationarity violation. Finally, the variances of the simulated epochs were also normalised, such that all variances were identical. The initial 15 ms of the normalised ensembles were again analysed using the Hotelling's T^2 test.

To summarise: 10 000 additional recordings were simulated for each original recording of EEG background activity. A violation to the stationarity assumption was then introduced to each simulated recording, after which the violation was removed by normalising the epoch variances. At each stage, the recordings were structured into ensembles of $N = 500$ epochs, which were analysed using the Hotelling's T^2 test.

Results

The FPRs from the Hotelling's T^2 test are presented in Fig. 5.9 as a function of the recording index being simulated. Plot A shows the FPRs before and after introducing the stationarity violation. Note that significant deviations from the nominal α-level can be observed for the *stationary* data, which suggests a violation to the independence assumption between epochs (as all remaining assumptions were satisfied). Note therefore that in order to isolate the stationarity violation, it is necessary to compare the FPRs from the stationary and non-stationary data. To facilitate the comparison, the FPRs

from the stationary data are subtracted from the FPRs from the non-stationary data, and the difference is added to α. Results (Fig. 5.9, Plot B) demonstrate an overall trend towards a conservative test performance, which can now be attributed to the stationarity violation. Finally, Fig. 5.9 Plot C shows the FPRs generated from the normalised epochs. The FPRs generated from the stationary data are also shown for comparison. Results demonstrate a more or less identical performance between the stationary and normalised data, which suggests that normalisation of the epoch variances is a viable option for removing stationarity violations.

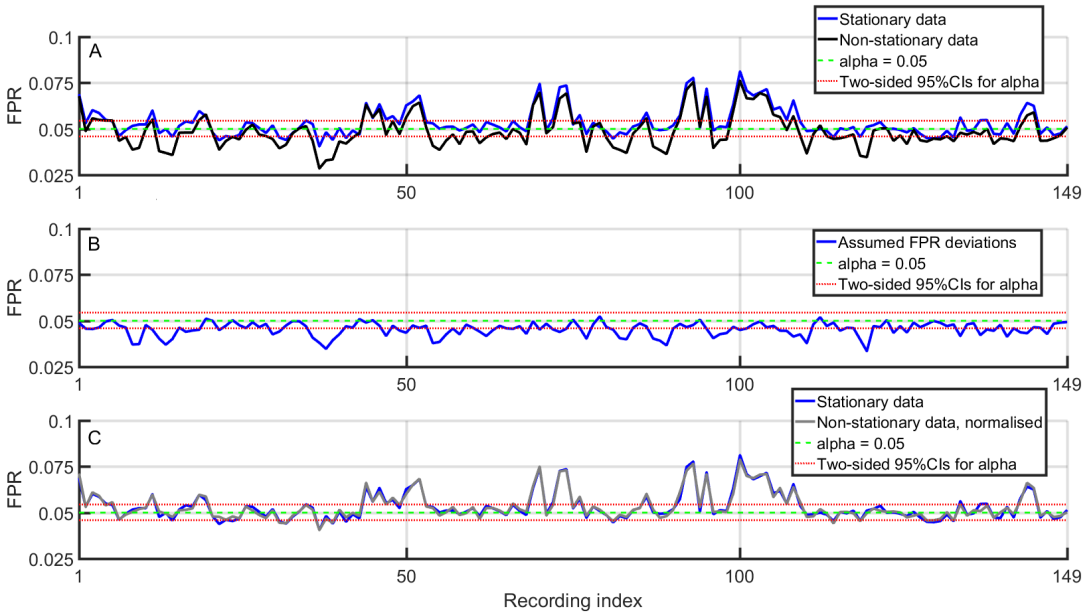


Figure 5.9: The FPRs for the Hotelling's T^2 test as a function of the recording of EEG background activity being simulated. **Plot A**: the FPRs generated from stationary and non-stationary data. **Plot B**: the discrepancy amongst the FPRs presented in Plot A, added to the theoretical $\alpha = 0.05$. **Plot C**: FPRs generated from normalised data, along with the FPRs generated from stationary data, which is repeated here for the sake of comparison. The nominal α-level and its two-sided 95% CIs are also shown.

Summary

This section demonstrated violations to the stationarity assumption (Figure 5.8), which resulted in a tendenecy towards a conservative test performance (Figure 5.9, plot B). The latter can likely be attributed to an overestimated sample variance. The largest deviation from α (in Figure 5.9, plot B) was observed for recording 119, and was equal to 0.0335. The mean FPR across all recordings was furthermore equal to 0.0454, and was hence again relatively close to the expected 0.05. Finally, results also suggest that violations to the stationarity assumption can be removed by normalising the variances of the epochs (Figure 5.9, plot C).

Relationship with Bayesian averaging

Normalising the variances of the epochs is somewhat similar to the Bayesian averaging approach described in Elberling & Walhgreen (1985). For Bayesian averaging, the goal is to increase the SNR within the coherent average for when data is non-stationary.

Following the notation in Elberling & Wahlgreen (1985), the Bayesian-weighted coherent average is given by:

$$\bar{X} = (\frac{S_1}{V_1} + \frac{S_2}{V_2} + ... + \frac{S_n}{V_n}) \cdot \frac{1}{C_n} \tag{5.3}$$

where S_i is the ith block average, obtained by averaging (now using conventional averaging) across a subset of epochs, V_i is the variance of the EEG background noise within the ith block of epochs (estimated using the SP variance), n is the total number of sub-blocks, and $\frac{1}{C_n}$ is a sum of variances, defined as $\sum_{i=1}^{n} V_i$.

Note therefore that Bayesian averaging has a slightly different goal than normalising the epoch variances, i.e. the goal in Bayesian averaging is to maximize the SNR with the coherent average, whereas the goal for normalising the epoch variances is to improve the specificity of the objective detection method by removing stationarity violations. It is nevertheless expected that Bayesian averaging would still reduce non-stationarity violations, but would not remove them completely. Similarly, normalising epoch variances may improve the SNR within the coherent average, but perhaps not as effectively as Bayesian averaging.

5.4 Real EEG background activity

In the previous sections of this Chapter, simulations were used to isolate and evaluate the main assumptions underlying objective ABR detection methods. A shortcoming for these sections is that each assumption was considered in isolation, whereas for real EEG background activity violations to multiple assumptions can occur simultaneously, potentially with interaction effects. A second shortcoming for the simulations is that various real world noise sources were not modelled adequately (e.g. the mains interference and movement artefacts). The goal for this section is therefore to explore the extent to which specificity is controlled for real EEG background activity. The EEG pre-processing parameters included in the assessment are (1) the cut-off frequency for the high-pass filter, (2) the (hypothetical) stimulus rate, (3) artefact rejection, and (4) normalisation of the epoch variances.

Method

Each recording of EEG background activity (data set **D1**) was downsampled to 5 kHz and band-pass filtered from f_c to 2000 Hz (using a 3rd-order Butterworth filter) where f_c took values of 30 to 100 Hz, in steps of 5 Hz. Ensembles of epochs were then constructed by randomly selecting $N = 200$ consecutive windows from within the recording in question. The duration of each window was set to 15 ms, and the distance between windows was varied from 0 ms to 40 ms, in steps of 0.2 ms. As was the case in section

5.1, this can be related to a (hypothetical) stimulus rate of $\frac{1000}{15+\tau}$, where τ is the distance in ms between the 15 ms windows. A total of 50 ensembles were constructed from each recording, both before and after artefact rejection (achieved by throwing away 10% of the noisiest epochs, as determined by the absolute maximum values), resulting in a total of 7450 ensembles, per test condition. The resulting ensembles were analysed with the Hotelling's T^2 test using 25 TVMs as features, extracted from the 15 ms windows, both before and after normalising the variances of the epochs.

To summarize: 7450 ensembles were constructed for the following test conditions: (1) artefact rejection was applied, and the variances of the epochs within each ensemble were normalised, (2) artefact rejection was applied, but the variances of the epochs were not normalised, (3) no artefact rejection was applied, and the variances of the epochs were normalised, and (4) no artefact rejection was applied, and the epoch variances were not normalised. These four conditions were evaluated across all aforementioned high-pass cut-off frequencies and (hypothetical) stimulus rates.

Results

Before presenting the results, it should be mentioned that the FPRs for all four test conditions were quite similar. Hence, to keep this section concise, FPRs from just the second condition (artefact rejection, but no normalisation, which is considered to be the conventional approach) are initially presented in Fig 5.10. After some modifications (further described below), results from all four conditions are presented in Fig. 5.11. The binomial distribution was furthermore used to construct two-sided 95% CIs for $\alpha = 0.05$, giving lower and upper bounds of [0.0452, 0.0552]. Note that these boundaries are approximate, as the random resampling (with replacement) procedure may have resulted in some EEG segments being selected multiple times, resulting in a violation of the independence assumption between tests (underlying the binomial distribution: see Appendix, section A.2). Significant deviations from α are indicated in Fig. 5.10 and 5.11 by red (FPR $>$ 0.0552) and blue (FPR $<$ 0.0452) cells, whereas green cells indicate that the FPR fell within the two-sided 95% CI.

With respect to the big 'spikes' in Fig. 5.10, these can at least partly be attributed to the 50 Hz mains and its harmonics. The 50 Hz mains is indicated in Fig. 5.10 by an 'M', whereas the H1, H2, H3, H4, and H5 (located at 62.5, 41.667, 35.71, 31.25, and 27.78 Hz respectively) correspond to a 250 Hz signal and its harmonics, which might be related to the 50 Hz mains. The source of various additional peaks located at 54.95, 45.05, 34.97, 27.47, and 26.18 Hz (indicated by ?1, ?2, ?3, ?4, and ?5 respectively) are further considered in the discussion below.

Post-hoc analysis

It is first worth emphasizing that the FPRs in Fig. 5.10 show a similar pattern as those from the simulations (Fig 5.3). In order to facilitate this comparison (and to aid visual inspection when comparing the four test conditions), the spikes in Fig. 5.10 were removed. The latter was achieved through interpolation, i.e. by setting the spikes to the

mean value of the neighbouring cells (under the condition that the neighbouring cells did not contain a spike), after which smoothing was applied to the resulting FPRs. The modified results for all four test conditions are presented in Fig. 5.11.

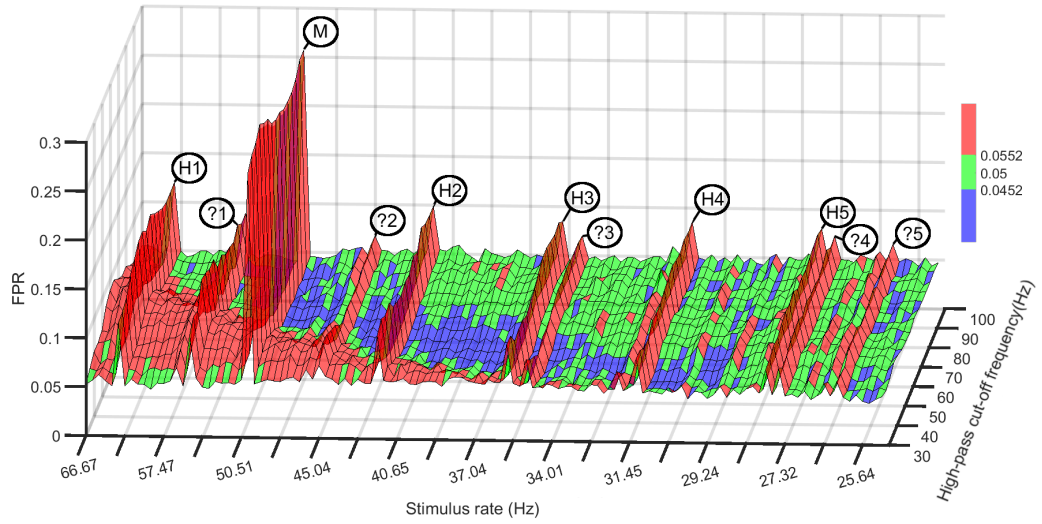


Figure 5.10: The FPRs of the Hotelling's T^2 test, as a function of the high-pass cut-off frequency and (hypothetical) stimulus rate. For this data, artefact rejection was applied, but the epoch variances were not normalised. The source for the 'spikes' in the FPRs (indicated by the M, H1, H2, H3, H4, H5, ?1, ?2, ?3, ?4, and ?5 captions) are further considered in the results and discussion sections.

Discussion

The close correspondence between Fig. 5.3 and 5.11 firstly suggests that the main concern for the specificity of ABR detection methods is the independence assumption between epochs, which is violated primarily as a function of the high-pass cut-off frequency f_c and the stimulus rate. As was the case for the simulations, specific combinations of f_c and the stimulus rate can result in relatively severe deviations from the nominal α-level, e.g. using $f_c = 65$ Hz and a stimulus rate of 66.67 Hz gave a FPR of 0.0896 (Figure 5.10). Certain combinations of f_c and the stimulus rate are nevertheless safe (see Figures 5.10 and 5.11).

With respect to stationarity and normality violations, results from Figure 5.11 demonstrate that the FPRs across all four test conditions were quite similar, which implies that artefact rejection and normalising the epoch variances did not have a huge effect on the FPRs. This suggests that violations to the stationarity and normality assumptions were more or less negligible, albeit when considered across a cohort of subjects. The latter is in agreement with results from the simulations, which show significant deviations from

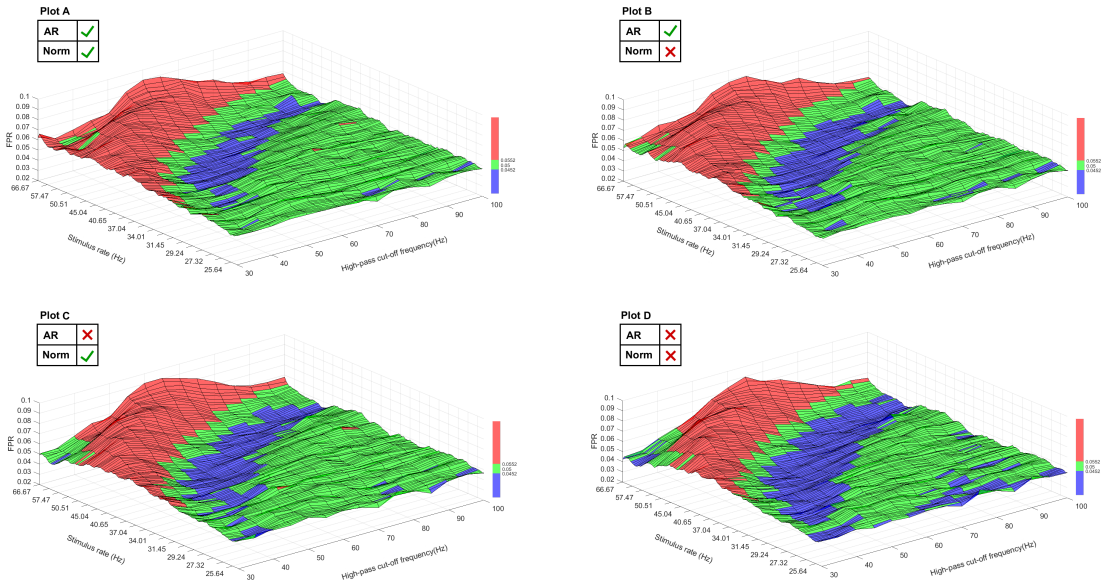


Figure 5.11: The FPRs of the Hotelling's T^2 test after removing the spikes identified in Fig 5.10 and applying a smoothing algorithm, as a function of the high-pass cut-off frequency and the (hypothetical) stimulus rate, for each of the following test conditions: **Plot A**: artefact rejection (denoted here by AR) was used, and the variances of the epochs were normalised (denoted by Norm). **Plot B**: Artefact rejection was used, but the variances of the epochs were not normalised. **Plot C**: artefact rejection was not used, but the variances of the epochs were normalised. **Plot D**: artefact rejection was not used, and the variances of the epochs were not normalised.

α (due to normality and stationarity violations) for some recordings, but a mean FPR (across all recordings) still relatively close to α. It should however be stressed that this does not necessarily imply that normality and stationarity assumptions are negligible, i.e. for many ABR-related applications, a robust and reliable control of specificity is desirable at the level of the individual, not just across a cohort of subjects.

Various spikes in the FPRs in Fig. 5.10 furthermore remain unidentified. The first two spikes are located at 54.9451 and 45.045 Hz (captioned in Fig. 5.10 by ?1 and ?2 respectively). A possible source might be a 500 Hz harmonic, i.e. one full cycle of a 500 Hz signal would have a duration of 2 ms, meaning any epochs separated by distances that are a multiple of 2 ms will be time-locked to multiple cycles of a 500 Hz signal. For a 54.9451 Hz signal, the onsets of the epochs are located at 18.2 ms intervals, whereas for a 45.045 Hz signal they are located at 22.2 ms intervals. Although neither is a multiple of 2 ms, the 4 ms distance between the 18.2 and 22.2 intervals seems suspicious. The remaining three spikes, located at 34.965 Hz, 27.4725 Hz, and 26.178 Hz (captioned by ?3, ?4, and ?5 respectively), however, show no such suspicious relationships. In a nutshell: the source for these spikes remain unknown, and may require a deeper analysis (possibly at the level of the recording) before an answer can be found.

5.5 Summary

This Chapter used simulations and real recordings of EEG background activity to isolate, evaluate, and potentially compensate or remove violations to the main statistical assumptions underlying ABR detection methods, with the overall goal of obtaining a more robust control of specificity. The main statistical assumptions that were explored include the independence assumption (between epochs), the normality assumption, and the stationarity assumption.

The main culprit for increased or decreased FPRs in both the simulations and the real data turned out to be the independence assumption, which was violated as a function of the high-pass cut-off frequency f_c and the stimulus rate. Specific combinations of f_c and the stimulus rate resulted in relatively large deviations from α, ranging from 0.0385 to 0.0985 for $\alpha = 0.05$ (section 5.1.3). Certain combinations of f_c and the stimulus rate are nevertheless safe (see Figures 5.3, 5.10, and 5.11).

With respect to the normality and stationarity assumptions, simulations demonstrate a tendency towards a conservative test performance when these assumptions are violated, with maximum deviations from $\alpha = 0.05$ of 0.0161 for normality violations (using no artefact rejection, and $N = 100$: Figure 5.6, plot A) and 0.0335 for stationarity violations (Figure 5.9, plot B). The mean FPR (across all recordings) was nevertheless still close to $\alpha = 0.05$, and was in the range of 0.045 (for both normality and stationarity violations).

To remove or compensate (through CLT) for normality violations, artefact rejection was used, or the ensemble size was increased. Results show that increasing the ensemble size was insufficient for compensating for the violation. Artefact rejection, on the other hand, resulted in more or less perfectly normally distributed data (Figure 5.5), but sometimes shifted the mean of the recording away from zero, resulting in a liberal test performance (Figures 5.6, plot B, and Figure 5.7). With respect to stationarity violations, these were successfully removed (with no noticeable adverse effects) by normalising the variances of the epochs. Although not explored here, it is also hypothesized that normalising the epoch variances would reduce normality violations.

For the real EEG background activity, violations to the normality and stationarity violations were found to be more or less negligible, albeit when considered across a cohort of subjects. This is in agreement with the simulations, which show that the mean FPR (taken across recordings) is still close to α under significant normality and stationarity violations. The FPRs for individual recordings, however, can still deviate significantly from α, which implies that stationarity and normality violations should ideally not be ignored if a robust control of specificity at the level of the individual is desired.

Finally, early results suggest that 'bootstrapping in blocks' can be used for a more robust evaluation of test significance under independence violations (Figure 5.4). Note that the

bootstrap is also immune to normality and stationarity violations, and might therefore provide a solution to all aforementioned violations.

5.5.1 Limitations

A first limitation for this Chapter is that specificity was explored in isolation, whereas, ideally, it would have gone hand in hand with a sensitivity assessment, i.e. it is not clear how or if some of the methods (e.g. bootstrapping in blocks, and normalising epoch variances) might affect test sensitivity. A second limitation for this Chapter is that results were generated using the Hotelling's T^2 test, and might not generalise well to alternative detection methods. For example, some methods may not distinguish between positively and negatively correlated epochs (as is the case with the Hotelling's T^2 test), potentially resulting in an exclusively liberal test performance for any independence violation. Some methods might also be more susceptible to independence violations relative to others due to the additional assumption of (some degree of) independence between samples within epochs (e.g. the Fsp and the Fmp). The FPRs presented in Figures 5.3, 5.10 and 5.11 might therefore show a similar but more pronounced trend for these methods.

With respect to the simulations for quantifying normality violations (section 5.2.2), a limitation is that a univariate PDF was assumed for the features. As a result, the covariance structure of the features is neglected. In other words, the resampled TVMs can be considered independent, both within and between epochs. Future work might explore a multivariate approach, and strive to model (and resample from) a multivariate distribution. A potential complication with this approach is the sparseness of multivariate datasets, which increases with the dimension of the data. An excessively large data set might therefore be required in order to accurately model the distribution of a 25-dimensional feature set.

Finally, results from the real EEG background activity should ideally have been conducted on a recording to recording basis, as opposed to across a cohort of subjects. The latter was hampered by (i) long computation times, (ii) insufficient data at the level of the recording, and (iii) additional complications in terms of how to present and analyse the results (there would be too many figures).

Chapter 6

Sensitivity and test time

The sensitivity of a test is its detection rate when detecting the effect in question (an ABR). Sensitivity is also referred to as the 'true-positive rate' (TPR), defined as the ratio of significant test outcomes when the alternative hypothesis is true (the ABR is indeed present) over the total number of tests performed. Test time is then defined as the total time spent trying to detect the hypothesized effect size. Test time is furthermore closely related to sensitivity, as a more sensitive test will typically detect the response sooner. An exception is of course when sensitivity is increased by increasing the sample size, which will increase test time.

The goal for this Chapter is to evaluate and compare the sensitivities and test times of various objective ABR detection methods. The sensitivity of a test is influenced by many factors, amongst which is the specificity of the test. In particular, a decrease in the nominal α-level results in a decreased sensitivity, and vice versa for an increase in α. The sensitivity assessment in this Chapter is therefore always accompanied by a specificity assessment, firstly to ensure that specificity is controlled as intended, and secondly to verify that discrepancies in sensitivity are not due to increased or decreased FPRs. Additional factors that may affect test sensitivity include pre-processing parameters, the statistical features selected for the analysis, and (as mentioned above) the ensemble size. A comprehensive comparison in sensitivity should therefore take a range of statistical features, EEG pre-processing parameters and ensemble sizes into account. For this Chapter, the statistical features and pre-processing parameters are selected based on (i) findings or recommendations from the literature, (ii) results from Chapter 5, and (iii) pilot simulations and results from feature optimisations presented in the Appendix (section A.3).

Throughout this Chapter, the performance of 12 different objective detection methods are evaluated and compared. To avoid cluttering the results, the sensitivity assessment is initially split across two simulations (sections 6.1 and 6.2). Based on results from the simulations, a final selection of methods is made for further evaluation in section 6.3, for which the subject recorded ABR threshold series is used. The methods selected for the

assessment throughout this Chapter include: the Hotelling's T^2 test (applied in time or frequency domain), the Fsp and the Fmp (evaluated with theoretical F-distributions or with the bootstrap approach), the Modified q-sample V2 and V4 tests, the bootstrapped correlation coefficient, the bootstrapped Max-Difference and Mean Power statistics, and the bootstrapped 'T2 Time + CC' combination. Additional simulations for comparing the performance of RM ANOVA, Friedman's test, and the Hotelling's T^2 test can also be found in section A.7 of the Appendix.

6.1 Simulations I: comparisons in sensitivity

This section uses simulations to draw comparisons in sensitivity between (i) the Hotelling's T^2 test, applied in both the time and the frequency domain (denoted by 'T2 Time' and 'T2 Freq' respectively), (ii) the Fsp and the Fmp, evaluated using either theoretical F-distributions with assumed DOF (denoted by 'Fsp 5 dof' and 'Fmp 5 dof' respectively), or with the bootstrap approach (denoted by 'Fsp bootstrapped' and 'Fmp bootstrapped', respectively), and (iii) two versions of the modified q-sample uniform scores test, which use either the ranks (Modified q-sample V2) or the actual values (Modified q-sample V4) of the phases and amplitudes of multiple Fourier components.

The primary goal for these simulations is to provide a more powerful comparison of test sensitivity under controlled test conditions. More specific goals include evaluating the hypothesis that 'T2 Freq' will outperform 'Modified q-sample V4', which would be due to the Hotelling's T^2 test taking the correlations between features into account (which are neglected by 'Modified q-sample V4', see also sections 3.2 and 3.5, and discussion section 2.6). It is also hypothesized that the bootstrap approach will improve the specificity and sensitivity of the Fsp and the Fmp, as opposed to evaluating test significance with theoretical F-distributions. In particular, 'Fsp 5 dof' and 'Fmp 5 dof' are expected to give conservative test performances (Elberling & Don, 1984), which would coincidentally result in a reduced test sensitivity, relative to their bootstrapped counterparts.

Method

Data for the simulations consists of recordings of real EEG background activity (data set **D1**) and ABR templates (data set **D4**) for simulating a response. The recordings of EEG background activity were downsampled to 5 kHz and band-pass filtered (using a 3rd-order Butterworth filter, see Appendix A.16) from 30 to 2000 Hz.

Specificity assessment

Each pre-processed recording of EEG background activity was decomposed into ensembles of N epochs, where N took values of 50, 100, 175, 275, 375, 500, 650, 800. Note that the resulting ensembles did not overlap, and can hence be considered (more or less) independent. The duration of each epoch was furthermore set to 30.03 ms, corresponding to

a (hypothetical) stimulus rate of 33.3 Hz. This resulted in a total of 20197, 10060, 5717, 3606, 2640, 1967, 1500, and 1187 ensembles for ensemble sizes of 50, 100, 175, 275, 375, 500, 650, and 800, respectively. The specificity assessment then consists of analysing the initial 15 ms windows of the ensembles using the aforementioned detection methods. The statistical features for the detection methods are described in below.

Sensitivity assessment

For the sensitivity assessment, a random resampling (with replacement) approach was used to resample blocks of N epochs from within the continuous recordings of EEG background activity (data set **D1**). In total, 10 000 ensembles of N epochs were constructed, where N again took values of 50, 100, 175, 275, 375, 500, 650, 800 epochs. A response was simulated by randomly selecting an ABR template (from data set **D4**), rescaling it, and adding it to all epochs within the ensemble in question. The scaling factor was chosen such that the SNR of the response was -23 dB, which was calculated as described in section 4.3. The -23 dB simulated response would correspond to a relatively strong response (from a normal hearing subject) from the 30, 40, or 50 dB SL condition (see Table 4.1 in section 4.2). The initial 15 ms windows of the resulting ensembles were then analysed using the aforementioned detection methods.

Statistical features

The time domain features for T2 Time consist of 25 TVMs (spread equally across the initial 15 ms windows within the epochs). The choice for 25 TVMs was based on pilot simulations, which showed a robust performance for anything between ∼25 and ∼40 TVMs. For the frequency domain methods, all spectral bands within (and including) the 80 and 600 Hz bands were used for the analysis. The latter was based on findings in the literature which show that the majority of the energy within the ABR lies within the 50-250 and 500-600 Hz bands, and (for higher stimulus intensities) also within the 900-1100 Hz band (Elberling, 1976; Kevanishvili & Aphonchenko, 1979; Elberling, 1979; Suzuki et al., 1982). Because of the relatively low dB SL stimulus, it was assumed (for this section) that the energy within the 900-1100 Hz band was negligible (as shown in section 6.2, this is actually not the case). Prior to calculating the FFT, each 15 ms window was first extended to 25 ms through zero-padding, giving a spectral resolution of 40 Hz. For the Modified q-sample V2 and V4 tests, averaging was also used (prior to calculating the FFT) to compress each ensemble into blocks of sub-averages, as recommended by Cebulla et al (2000). For these simulations, averaging was performed across blocks of 25 epochs so that no epochs were excluded from the analysis (each ensemble size is a multiple of 25), which hence compressed each ensemble into $\frac{N}{25}$ sub-averages. The column index (of data matrix D) for calculating the single point variance for the Fsp was furthermore arbitrarily set to 30 (corresponding to the 6th ms following stimulus onset), and the number of columns to include in the Fmp was set to 75 (corresponding to the full analysis window, or 15 ms).

Results

Specificity

The FPRs of the methods (using either $\alpha = 0.01$ or $\alpha = 0.05$) per ensemble size are presented in Table 6.1. Two-sided 95% confidence intervals for nominal levels of either $\alpha = 0.01$ or $\alpha = 0.05$ were found using the binomial distribution (Appendix, section A.2), and are similarly presented in Table 6.1. Significant ($\boldsymbol{p < 0.05}$) deviations from the nominal levels are denoted by red and blue asterisks, indicating a liberal and conservative test performance respectively. Results demonstrate a conservative test performance for 'Fsp 5 dof' and 'Fmp 5 dof'. The remaining methods appear to show a very minor tendency towards a more liberal test performance.

Sensitivity

The percentage of detected responses are presented in Fig 6.1 as a function of the ensemble size N, per method. Results show an overall advantage in sensitivity for the Hotelling's T^2 test (applied in either the time or frequency domain).

Table 6.1: **Simulations I: specificity**. The FPRs of the methods (using either $\alpha = 0.01$ or $\alpha = 0.05$) for the no-stimulus condition, per ensemble size N. The 95% two-sided CIs for α are also shown, per ensemble size. Significantly (**p<0.05**) conservative and liberal test performances are are denoted blue and red asterisks respectively.

Alpha = 0.01								
Ensemble size –>	**50**	**100**	**175**	**275**	**375**	**500**	**650**	**800**
T2 Time	0.0108	0.0125*	0.0098	0.0133	0.0148*	0.0092	0.0113	0.0126
T2 Freq	0.0109	0.0108	0.0114	0.0119	0.0159*	0.0132	0.0173*	0.0126
Fsp 5 dof	0.0054*	0.0053*	0.0051*	0.005*	0.008	0.0056*	0.004*	0.0051*
Fmp 5 dof	0.0023*	0.0036*	0.0037*	0.0044*	0.0061*	0.0056*	0.0033*	0.0042*
Fsp bootstrapped	0.0115*	0.0127*	0.014*	0.0094	0.0144*	0.0117	0.0127	0.0152
Fmp bootstrapped	0.0112	0.0124*	0.0124	0.0097	0.0148*	0.0102	0.012	0.0143
Modified q-sample V2	0.0094	0.0096	0.0117	0.0125	0.0144*	0.0188*	0.0087	0.0152
Modified q-sample V4	0.0086	0.0105	0.0115	0.0089	0.0114	0.0097	0.0113	0.0143
Confidence intervals								
Lower bound	0.0087	0.0083	0.0077	0.0072	0.0068	0.0066	0.0060	0.0059
Upper bound	0.0114	0.0121	0.0128	0.0136	0.0144	0.0153	0.0160	0.0168
Alpha = 0.05								
Ensemble size –>	**50**	**100**	**175**	**275**	**375**	**500**	**650**	**800**
T2 Time	0.0512	0.0542	0.0569*	0.0549	0.0565	0.0504	0.0514	0.0481
T2 Freq	0.052	0.053	0.0565*	0.0541	0.064*	0.057	0.0567	0.0506
Fsp 5 dof	0.0266*	0.0239*	0.025*	0.0214*	0.0254*	0.0254*	0.0233*	0.0278*
Fmp 5 dof	0.0146*	0.0181*	0.0192*	0.0205*	0.0243*	0.0249*	0.022*	0.027*
Fsp bootstrapped	0.0498	0.0509	0.0553	0.0513	0.0553	0.0534	0.0554	0.0489
Fmp bootstrapped	0.048	0.0506	0.0562*	0.0527	0.0534	0.0544	0.0647*	0.0497
Modified q-sample V2	0.0483	0.0493	0.05	0.0491	0.0568	0.0514	0.0554	0.0472
Modified q-sample V4	0.046*	0.046	0.0521	0.0533	0.0549	0.0493	0.0494	0.0481
Confidence intervals								
Lower bound	0.0471	0.0459	0.0446	0.0433	0.0420	0.0412	0.0400	0.0388
Upper bound	0.0531	0.0544	0.0560	0.0574	0.0587	0.0605	0.0620	0.0632

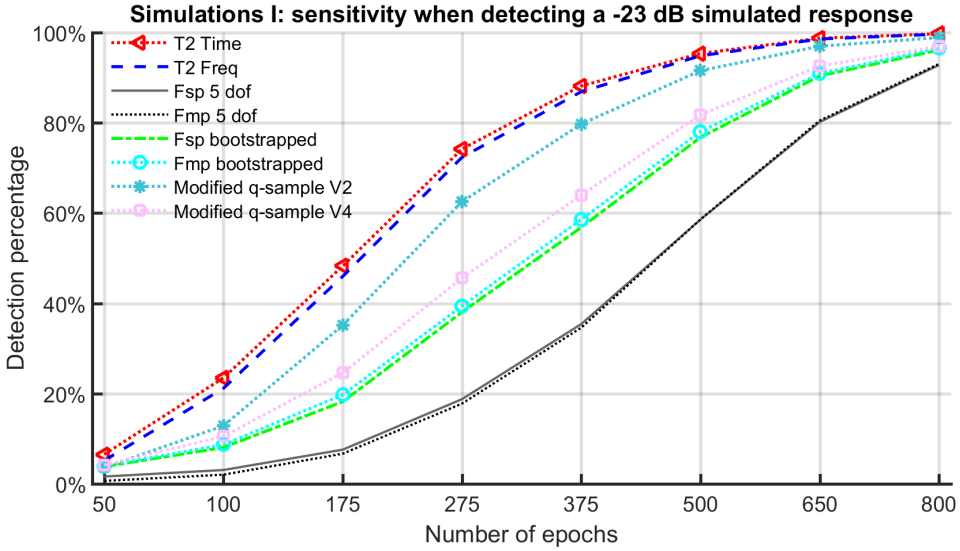


Figure 6.1: The percentage of detected responses when simulating a -23 dB response, as a function of the ensemble size N.

Discussion

In what follows, results from the specificity and sensitivity analysis are further discussed. In particular, various sources that may have contributed to the slight tendency towards a more liberal test performance are first considered, after which the use of pre-determined thresholds for evaluating test significance is discussed. Some pros and cons associated with the use of detection rates for evaluating and comparing sensitivity are then considered. The features for the frequency domain methods are also reviewed, and some comparisons between the Modified q-sample V4 test and the Hotelling's T^2 test are drawn.

Specificity: elevated FPRs

For the specificity assessment, a very minor tendency towards a more liberal test performance was observed for most methods, which might be attributed to various factors, including: (i) violations to the underlying statistical assumptions, (ii) random fluctuations, and (iii) inaccurate CIs for the nominal α-levels. Starting with potential violations to the underlying statistical assumptions, these were evaluated extensively in Chapter 5. The most likely culprit, the independence assumption, was explored using simulations in section 5.1. Results (based on 50 000 tests) show a FPR (using $\alpha = 0.05$ for the Hotelling's T^2 test) of 0.053 when using a (hypothetical) stimulus rate of 33.3 Hz and a high-pass cut-off frequency of 30 Hz (which were the adopted values for this section). The two-sided 95% confidence intervals for $\alpha = 0.05$ were furthermore [0.0481, 0.0519], thus suggesting a very minor violation of the independence assumption for these settings. With respect to stationarity and normality, note that the bootstrapped statistics are (at least in theory) immune to such violations. Because the elevated FPR was also

observed for some of the bootstrapped statistics, it might be concluded that stationarity and normality violations did not play a significant role here (at least not when considered across a cohort of subjects).

Moving on to random fluctuations, although this may seem an easy 'hand waving' explanation, the deviations were very small, which make it quite reasonable. Moreover, the performance of the tests can be considered highly correlated, i.e. if one method has an elevated FPR, then it is more likely that the remaining methods will also have elevated FPRs. The many significant deviations observed in Table 6.1 are then no longer quite as worrying.

Finally, the CIs for the nominal α-levels might be slightly inaccurate. The latter is firstly due to the binomial distribution being a discrete probability distribution, meaning rounding errors occur when approximating CIs. As noted in the Appendix (section A.2), a liberal approach is adopted throughout this work, i.e. the CIs are rounded 'inwards', resulting in a minor bias towards too narrow CIs. A second factor is the independence violation between epochs, which may have resulted in an (even smaller) violation to the independence assumption (now underlying the binomial distribution) between statistical tests, which may also have contributed towards too narrow CIs. Finally, note that for the bootstrapped statistics, the critical thresholds for rejecting H_0 are approximations, i.e. they will have their own distributions (see Appendix, section A.5.1). Note therefore that the probability of a false-positive may vary between tests, in which case the assumption (underlying the binomial distribution) that the probability of a false-positive is fixed at α is violated. This might, again, have resulted in slightly inaccurate CIs for α for the bootstrapped statistics.

To summarise, the minor tendency towards a liberal test performance can most likely be attributed to a minor violation of the independence assumption between epochs, potentially in combination with slightly too narrow CIs for the nominal α-level of the test.

Specificity: pre-determined thresholds

As mentioned earlier, the DOF of EEG measurements are known to vary across recordings. This means that any *a priori* choice for the critical decision boundaries cannot be accurate across all recordings, which is coincidentally why Elberling & Don (1984) have recommended a conservative approach for the Fsp and the Fmp, i.e. by setting DOF v_1 to 5. Note also that the DOF are strongly correlated with the cut-off frequency of the high-pass filter. In particular, increasing the cut-off frequency results in fewer low frequency components. Because the lower frequency components tend to dominate the signal due to their high power, removing these frequencies tends to reduce the correlations amongst the samples within epochs, thus increasing the DOF of the data, and removing it farther from the assumed DOF $v_1 = 5$. It can therefore be expected that

the performance of 'Fsp 5 dof' and 'Fmp 5 dof' would be even more conservative when the high-pass cut-off frequency is increased, which was confirmed when repeating the specificity assessment with an adjusted high-pass cut-off frequency of 100 Hz. In particular, DOF v_1 ranged from 3 to 15 (with mean 8.4 and standard deviation [SD] 2.6) when using a high-pass cut-off frequency of 30 Hz, and from 3 to 20 (with mean 11.3 and SD 4.9) when using a high-pass cut-off frequency of 100 Hz (the latter was achieved by fitting F-distributions to each bootstrapped null distribution and finding the best fitting function). It might be noted here that the conservative estimate of $v_1 = 5$ was originally intended for the higher cut-off frequency of 100 Hz (Elberling & Don, 1984). Note also that the Hotelling's T^2 test and the bootstrapped statistics are immune to independence violations amongst samples within epochs (but not between epochs). In particular, the Hotelling's T2 test accounts for correlated samples within epochs by scaling the feature means by the covariance matrix of the features, whereas the bootstrapped statistics account for correlated samples by resampling on an epoch to epoch basis, which preserves the correlations between samples within epochs. This further allows the bootstrapped confidence intervals to more accurately reflect test dependent factors, such as the EEG background noise, the electrode impedances and ultimately the DOF of the data. The latter is important for many ABR applications where the objective detection methods are expected to perform adequately across EEG recordings with varying DOF.

A similar argument might be made in favour of the bootstrap approach over the use of pre-determined thresholds generated from no-stimulus data (see e.g. Stürzebecher et al., 1999; Cebulla et al., 2000a; Cebulla et al., 2006), i.e. pre-determined thresholds may not generalise well across recordings with varying DOF, whereas the bootstrap approach estimates CIs specifically for the recording in question. The accuracy of pre-determined thresholds calculated from no-stimulus data is further considered in the Appendix (section A.8). Results indeed demonstrate relatively large variation in the critical decision boundaries, even when calculated under more or less identical test conditions.

Sensitivity: detection rates and adjusted α values

Sensitivity was evaluated using detection rates, which have the desirable properties of being intuitive and simple. A potential risk of using detection rates, however, is that methods with higher FPRs receive an unfair advantage over those with lower FPRs. The latter is most notably the case for 'Fsp 5 dof' and 'Fmp 5 dof', which were indeed designed to have lower FPRs. The problem can be resolved by adjusting the nominal α-levels, such that the FPRs are equal across methods. Note however that although this allows for a more fair comparison, it is not necessarily a realistic one as adjustment of the FPR may need to be carried out on an individual basis using prior knowledge that is not always available. The detection rates using the adjusted α-levels are nevertheless presented in the Appendix (section A.6). Results demonstrate an advantage for the Hotelling's T^2 test when using both the adjusted and unadjusted critical α values. This

is also supported by FPRs in Table 6.1: consistent differences in detection rates (Fig. 6.1) cannot be readily explained from relatively inconsistent FPRs.

Sensitivity: the Hotelling's T^2 test and the Modified q-sample V4

With respect to the frequency domain features for the Hotelling's T^2 test, it is worth noting that these are essentially the same as those used by the Modified q-sample V4 test (the Hotelling's T^2 test is applied to the real and imaginary parts of the Fourier components, whereas the Modified q-sample V4 test is applied to the phases and amplitudes), and yet a relatively large discrepancy in performance was still observed. This can likely be attributed to the way in which features are weighted and combined into a single statistic. In particular, the Hotelling's T^2 test weights the features according to their variance and covariance, whereas the Modified q-sample V4 test does not. The latter results in a Q-dimensional hyper-ellipsoid (centred at the features means) as H_0 rejection region for the Hotelling's T^2 statistic, where the shape of the ellipsoid is determined by the variance and covariance of the features (see section 3.2.1). Having an ellipsoid as rejection region means that the null hypothesis is more easily rejected in some directions relative to others, meaning it has the potential of providing a more powerful test relative to, for example, a test with a spherical rejection region.

Based on the preceding paragraph, an identical performance between the Modified q-sample V4 test and the Hotelling's T^2 test might be expected when applied to uncorrelated features with equal variance, which was tested with additional simulations. In particular, simulations described in Stürzebecher et al (1999) and Cebulla et al (2000) were implemented, which used Gaussian zero mean white noise with stationary variance to represent the EEG background noise, along with a sinewave multiplied with a Gaussian window for representing a response. The detection methods included for these simulations were (i) the original q-sample uniform scores test (Mardia, 1972), (ii) both the Modified q-sample V2 and V4 tests (Stürzebecher et al, 1999; Cebulla et al, 2006), and (iii) the Hotelling's T^2 test using the frequency domain approach. As predicted, the Hotelling's T^2 test and the Modified q-sample V4 test both came out on top in terms of sensitivity (with very similar performances), followed by the Modified q-sample V2 test (using ranks rather than measured values), and lastly by the original q-sample uniform scores test (which only uses phase ranks). Further details and results can be found in the Appendix (section A.9).

Finally, it was assumed *a priori* that the power of the ABR within the 900-1100 spectral band would be negligible. A post-hoc feature optimisation (Appendix, section A.3) show that this is not the case, i.e. significant improvements in test sensitivity were gained by including the ∼900-1100 spectral bands. This might also explain why the Hotelling's T^2 test in the time domain slightly outperformed its frequency domain counterpart. It should however be noted that the time-domain features for the Hotelling's T^2 test were also not optimal. Results (Appendix, section A.3) indeed suggest a small increase in

test sensitivity when using 35 TVMs, as opposed to 25 TVMs, albeit when simulating a response using ABR templates from data set **D3**.

Sensitivity: bootstrapping

The bootstrap approach was successful in improving the sensitivity of the Fsp and the Fmp, as opposed to evaluating test significance using theoretical F-distributions. The latter is not surprising when considering the conservative performance of 'Fsp 5 dof' and 'Fmp 5 dof'. When re-plotting the detection rates using the adjusted alpha values (Appendix, section A.6), the sensitivities of 'Fsp 5 dof' and 'Fmp 5 dof' and their bootstrapped counterparts are more similar. Note that it is just the critical boundaries that differ between 'Fsp 5 dof', 'Fmp 5 dof' and their bootstrapped counterparts. Evaluating test significance using the adjusted α-levels would then indeed result in more similar critical boundaries, and hence a more similar test performance.

6.2 Simulations II: comparisons in sensitivity

For the second set of simulations, comparisons in sensitivity are drawn between: (i) the CC, obtained from the ensemble coherent average and some template, (ii) the bootstrapped 'Max Diff' and 'Mean Power' statistics, and (iii) 'T2 Time + CC', i.e. the Hotelling's T^2 test (applied in the time domain) combined with the CC (see section 3.6.3). The Fsp and T2 Time were also included so that comparisons with previous results can be drawn. For T2 Time, test significance was evaluated using either theoretical F-distributions (the standard approach) or with the bootstrap. This allows comparisons to be drawn between the bootstrapped T^2 statistic, and the T^2 statistic evaluated with theoretical F-distributions. The latter is important to verify that the bootstrap approach is not reducing test sensitivity. Finally, the statistical features selected for the analysis were chosen based on feature optimisations (section A.3 of the Appendix), and can be considered more or less optimal.

Method

Data for the simulations is similar to the data used in 'Simulations I' (section 6.1), and consists of recordings of real EEG background activity (data set **D1**), along with click-evoked ABR templates (now using data set **D4**) for simulating a response. The recordings of EEG background activity were downsampled to 5 kHz and band-pass filtered (using a 3rd-order Butterworth filter, see Appendix A.16) from 100 to 2000 Hz. Ensembles of N epochs were then constructed by randomly resampling N consecutive 30.03 ms epochs from within a randomly selected and pre-processed recording of EEG background activity. A response was then simulated at -28 dB using a stimulus rate of

33.3 Hz. A response with a SNR of -28 dB corresponds to the SNR of a typical response to a 30-40 dB SL click (see Table 4.1). The 100 Hz high-pass cut-off frequency was furthermore chosen based on results from Chapter 5, which show that independence is satisfied when using a 100 Hz high-pass cut-off frequency in combination with a 33.3 Hz stimulus rate. The ensemble size N again took values of 50, 100, 175, 275, 375, 500, 650, and 800 epochs, and the initial 15 ms windows of the ensembles were analysed using the aforementioned detection methods, both before and after simulating a -28 dB response.

Statistical features

As mentioned above, the statistical features selected for the methods were chosen based on results from feature optimisations presented in section A.3 of the Appendix. The number of TVMs for both T2 Time and 'T2 Time + CC' was set to 35, and the column index for the Fsp was arbitrarily set to 20 (corresponding to the 4th ms following stimulus onset). With respect to the CC, ABR templates (for correlating with the ensemble coherent average) were constructed per subject, and per dB SL condition. To avoid confusion, the templates for calculating the CC (i.e. the correlation between the ensemble coherent average in question and a template) are henceforth denoted by 'CC templates', whereas the templates for simulating a response (using data set **D4**) are referred to as 'ABR templates'. For each subject and each dB SL condition, a 'CC template' was constructed by taking the grand coherent average across all 'ABR templates' from the dB SL condition in question, after excluding the 'ABR template' that was currently being used to simulate a response. The latter is necessary to avoid an unfair bias towards the simulated response. With respect to the bootstrapped statistics, the ensemble coherent average was subtracted from each epoch prior to random re-sampling, as described in the Appendix (section A.5.2).

Results

Specificity

The observed FPRs (using $\alpha = 0.01$ or $\alpha = 0.05$) for each ensemble size N are presented in Table 6.2. The binomial distribution was used to construct two-sided 95% CIs, giving [0.0076, 0.013] for $\alpha = 0.01$, and [0.0442, 0.0564] for $\alpha = 0.05$. Note however that for these simulations, a random resampling (with replacement) procedure was used to construct the ensembles, which may have resulted in some segments being selected multiple times. The latter may have resulted in an independence violation between tests (underlying the binomial distribution), giving slightly too narrow CIs (see also section A.2 of the Appendix). Significant deviations from the nominal α-levels are nevertheless denoted in Table 6.2 by red and blue asterisks, indicating a liberal and conservative test performance respectively.

Sensitivity

The percentage of detected responses (using $\alpha = 0.01$) are presented in Fig. 6.2, as a function of the ensemble size N. Results show that the bootstrapped CC came out

on top for $N = 50$, but that test sensitivity decreased rapidly (relative to alternative methods) for larger ensemble sizes. For all $N > 100$, the bootstrapped 'T2 Time + CC' statistic came out on top by a relatively large margin. Results also show an identical performance between 'T2 Time (F-distributions)' and 'T2 Time (bootstrapped)', which implies that bootstrapping did not decrease the sensitivity of the Hotelling's T^2 test. Note that for the bootstrap, the ensemble coherent average was subtracted from the epochs prior to random resampling, which is necessary to avoid a small decrease in test sensitivity (see section A.5.2 of the Appendix).

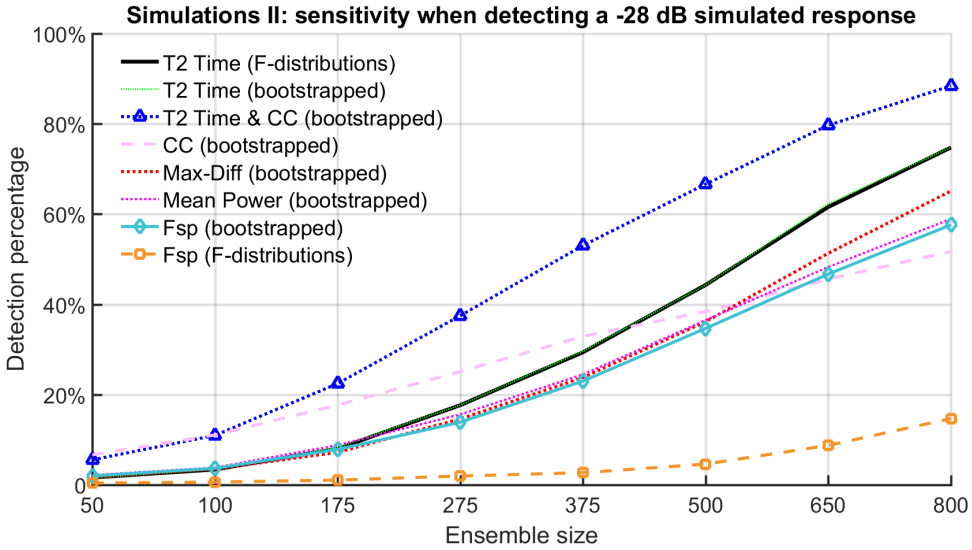


Figure 6.2: The percentage of detected responses when simulating a -28 dB response, as a function of the ensemble size N. Note that the detection rates for T2 Time (bootstrapped) and T2 Time (F-distributions) overlap, and may be difficult to distinguish from each other.

Table 6.2: **Simulations II: specificity**. The FPRs of the methods (using $\alpha = 0.05$ or $\alpha = 0.01$) for the no-stimulus condition, per ensemble size N, using a nominal. Significantly (**p<0.05**) conservative and liberal test performances are denoted blue and red asterisks respectively.

Alpha = 0.01								
Ensemble size –>	**50**	**100**	**175**	**275**	**375**	**500**	**650**	**800**
T2 Time (F-distributions)	0.0114	0.0086	0.0086	0.0076	0.0092	0.0096	0.0096	0.0108
T2 Time (bootstrapped)	0.0102	0.0088	0.01	0.0094	0.0094	0.0108	0.0094	0.0112
T2 Time & CC (bootstrapped)	0.011	0.0102	0.0124	0.0124	0.0128	0.013	0.0126	0.0112
CC (bootstrapped)	0.013	0.0122	0.0116	0.011	0.0128	0.0124	0.011	0.0094
Max Diff (bootstrapped)	0.0080	0.0114	0.0116	0.0110	0.0104	0.0114	0.0110	0.0126
Mean Power (bootstrapped)	0.0086	0.0096	0.0086	0.0112	0.0104	0.0108	0.0098	0.0108
Fsp (bootstrapped)	0.0122	0.0106	0.0124	0.0114	0.0114	0.0102	0.0102	0.012
Fsp 5 dof	0.0032*	0.0012*	0.0012*	0.0018*	0.0002*	0.0002*	0.001*	0.001*
Alpha = 0.05								
Ensemble size –>	**50**	**100**	**175**	**275**	**375**	**500**	**650**	**800**
T2 Time (F-distributions)	0.05	0.0444	0.0444	0.049	0.0492	0.0522	0.0544	0.051
T2 Time (bootstrapped)	0.0458	0.0438*	0.0458	0.05	0.0496	0.0536	0.0544	0.052
T2 Time & CC (bootstrapped)	0.0536	0.0472	0.0522	0.0496	0.0556	0.0562	0.0536	0.0524
CC (bootstrapped)	0.0526	0.0552	0.0522	0.049	0.054	0.0538	0.051	0.0494
Max Diff (bootstrapped)	0.0456	0.0546	0.0540	0.0528	0.0490	0.0548	0.0562	0.0572*
Mean Power (bootstrapped)	0.0466	0.0536	0.0548	0.0512	0.06*	0.0596*	0.0584*	0.0540
Fsp (bootstrapped)	0.0534	0.0588*	0.0554	0.0548	0.0546	0.0634*	0.0588*	0.0544
Fsp 5 dof	0.017*	0.0156*	0.0112*	0.0126*	0.0124*	0.009*	0.0138*	0.0114*

Discussion

With respect to specificity, results (Table 6.2) show that the FPRs were close to the nominal α-levels, although various minor (but significant) deviations from α were still observed. With respect to sensitivity, results demonstrate a relatively good sensitivity for the CC for small N, but performance dropped rapidly (relative to alternative methods) for larger N. The best performing method for this section was the 'T2 Time + CC' combination. These results are now discussed in more detail below.

Sensitivity

Starting with the bootstrapped CC, results show a relatively good test sensitivity for small ensemble sizes, but performance dropped rapidly (relative to alternative methods) as the ensemble size was increased. This can likely be attributed to the choice for the 'CC templates', which may have correlated well with some 'ABR templates' (resulting in an early detection), but poorly with others (resulting in a late or no detection). If this is indeed the case, then the bootstrapped CC may provide an exceptionally sensitive test statistic, under the condition that the exact waveform morphology for the subject in question is known *a priori*. In practice, this information is typically not available, and using the bootstrapped CC as ABR detection method might be a bit of a gamble, i.e. it may perform exceptionally well for some subjects, but poorly for others. A potential solution might be to calculate the CC using multiple templates (and to bootstrap the sum or the mean of the correlations). Alternatively, the CC can be combined with a non-template specific method, such as the Hotelling's T^2 test. Results from this section indeed demonstrate a highly sensitive and robust (across ensemble sizes) performance for the 'T2 Time + CC' combination.

Additional observations worth mentioning include results from T2 Time, which show that sensitivity was not reduced when evaluating test significance with the bootstrap approach, as opposed to using theoretical F-distributions. Note that for the bootstrap approach in this section, the ensemble coherent average was subtracted from the epochs prior to re-sampling. The latter is necessary if a (minor) decrease in sensitivity is to be prevented (Appendix, section A.5.2). With respect to the bootstrapped 'Mean Power', 'Max-Diff', and Fsp, a minor advantage was observed for the 'Mean Power' and 'Max Diff' statistics over the Fsp, which is in agreement with results from Lv et al (2007).

Specificity

With the exception of 'Fsp 5 dof', just a single conservative test performance was observed for the bootstrapped T2 Time statistic (using $\alpha = 0.05$, $N = 100$). For the bootstrapped Fsp, Max Diff, and Mean Power statistics, a slightly liberal test performance was observed for the $N = 100$, $N = 375$, $N = 500$, $N = 650$ and $N = 800$

conditions when using $\alpha = 0.05$. Various factors contributing towards significantly liberal or conservative test performances were previously discussed in section 6.1, and include (i) violations to the statistical assumptions underlying the test, (ii) random fluctuations, and (iii) inaccurate CIs for α. Starting with the CIs, the resampling with repacement procedure used to construct the ensembles in this section may have resulted in some EEG measurements being used multiple times, resulting in an independence violation (underlying the binomial distribution) between statistical tests, which might have contributed towards too narrow CIs (note that this is in addition to the factors discussed in section 6.1). With respect to the independence assumption between epochs, simulation results from section 5.1 (based on 50 000 tests) show a FPR (using $\alpha = 0.05$) for the Hotelling's T^2 test of 0.0484 when using a (hypothetical) stimulus rate of 33.3 Hz and a high-pass cut-off frequency of 100 Hz. The two-sided 95% confidence intervals for $\alpha = 0.05$ were [0.0481, 0.0519], suggesting that independence is satisfied for these settings (albeit when considered across a cohort of recordings). The latter is confirmed in section A.5.1, which shows a FPR (using $\alpha = 0.05$ and 175 000 tests) for the Hotelling's T^2 test of 0.0496, along with two-sided CIs of [0.0490, 0.0510].

6.3 Subject ABR data: comparisons in sensitivity and test time.

Based on simulation results in sections 6.1 and 6.2, a final selection of methods is now made for further comparison using the subject recorded ABR threshold series (data set **D2**). The methods selected for the analysis include (i) the Hotelling's T^2 test, applied in either the time or frequency domain, (ii) the bootstrapped CC, (iii) the bootstrapped T2 Time + CC statistic, and (iv) the Fsp and the Fmp, evaluated using either theoretical F-distributions with assumed DOF, or with the bootstrap approach.

Method

The recordings (data set **D2**) were downsampled to 5 kHz, band-pass filtered from 100 to 1500 Hz using a 3rd-order Butterworth filter (see Appendix, A.16), and structured into ensembles of 30.03 ms epochs. The methods were then applied to the initial 1-16 ms windows of the epochs (the first ms was excluded to avoid potential contaminations from a stimulus artefact), which was repeated per subject, and per dB SL condition. The methods were furthermore applied to the ensembles sequentially, every 50 epochs, from 50 epochs onwards. To clarify - the first test was performed using an ensemble size of 50, then again using an ensemble size of 100, etc., until all 3000 epochs had been analysed (a total of 60 tests, per subject, and per dB SL condition).

Statistical features

The features for the detection methods are the same as those described in section 6.2.

The 'CC templates' were similarly constructed as described in section 6.2.

Results

The detection rates ($\alpha = 0.01$) when using an ensemble size of $N = 3000$ epochs are first presented in Fig. 6.3, per dB SL condition. The required time for detecting a response was then found by finding the number of stimuli (expressed in seconds) required for the p value to drop and remain below the 0.01 threshold for the remainder of the test. The additional requirement that the p value remains below the 0.01 threshold ensures that the FPR is not inflated due to multiple tests being performed. If a test did not drop below the 0.01 significance threshold, then the full ∼90 seconds test time was used (corresponding to 3000 epochs), which may have resulted in an underestimation of test time in the case of a false-negative. The mean of the resulting detection times (taken across subjects) are presented as bar graphs in Fig. 6.4, per method and per dB SL condition.

Visually inspecting the distributions of the detection rates and detection times suggests that both were strongly non-Gaussian, which was confirmed with the Kolmogorov-Smirnov goodness of fit test ($\boldsymbol{p < 0.01}$ for all distributions). Non-parametric statistical analysis was therefore used to test whether the discrepancy amongst the methods in terms of detection rates and detection times was significant. With respect to detection rates, Cochran's Q test was first used to test for equivalence in performance across all 7 methods, per dB SL condition. Results indicate a significant difference in performance for the 20, 30, and 50 dB SL conditions ($\boldsymbol{p < 0.01}$), and for the 10 dB SL condition ($\boldsymbol{p < 0.05}$). As a follow-up, Fishers exact test was used to draw pairwise comparisons amongst the methods for the 10, 20, 30, and 50 dB SL conditions. Results show that the performance of the bootstrapped 'T2 Time + CC' statistic differed significantly ($\boldsymbol{p < 0.05}$) from 'Fsp 5 dof' for the 30 dB SL condition. The remaining comparisons were not significant. Similarly, with respect to detection times, non-parametric statistical analysis was first used to test for equivalence in performance across all 7 methods (now using Friedman's test), per dB SL condition. Results indicate a significant difference in performance for the 20, 30, 40, and 50 dB SL conditions (all $\boldsymbol{p < 0.001}$). The Wilcoxon rank sum test was then used to draw pairwise comparisons between all methods for the 20, 30, 40, and 50 dB SL conditions. Results show that Fsp 5 dof was significantly outperformed by T2 Freq ($\boldsymbol{p < 0.05}$) and T2 Time + CC ($\boldsymbol{p < 0.0001}$) for the 40 dB SL condition, and again by T2 Time + CC for the 20 and 30 dB SL conditions ($\boldsymbol{p < 0.05}$). T2 Time + CC also significantly ($\boldsymbol{p < 0.05}$) outperformed the bootstrapped CC for the 50 dB SL condition, and the bootstrapped Fsp ($\boldsymbol{p < 0.05}$) for the 40 dB SL condition. A borderline significant advantage was observed for the bootstrapped CC over Fsp 5 dof for the 40 dB SL condition ($\boldsymbol{p = 0.0525}$), for T2 Time + CC over the Fsp 5 dof for the 50 dB SL condition ($\boldsymbol{p = 0.0559}$), and for T2 Time over Fsp 5 dof for the 40 dB SL condition ($\boldsymbol{p = 0.0528}$). The remaining comparisons

were not significant.

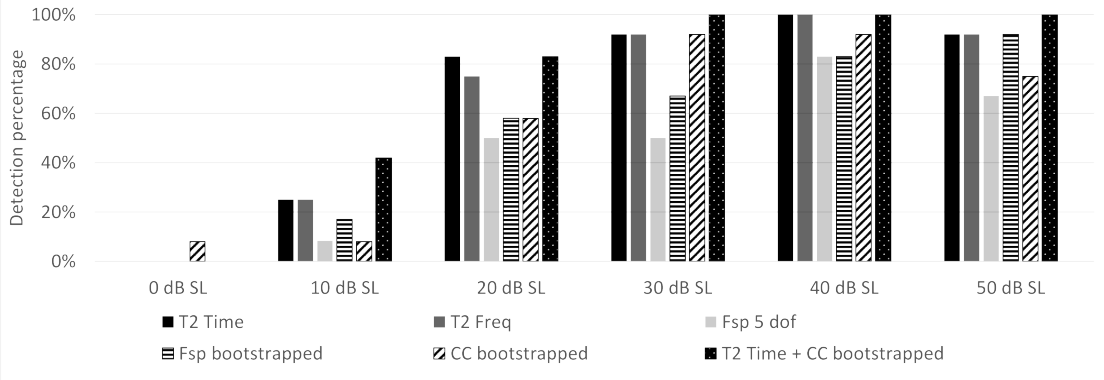


Figure 6.3: The percentage of detected responses ($\boldsymbol{p < 0.01}$) in a small sample of normal hearing adults (data set **D2**), per method and per dB SL condition.

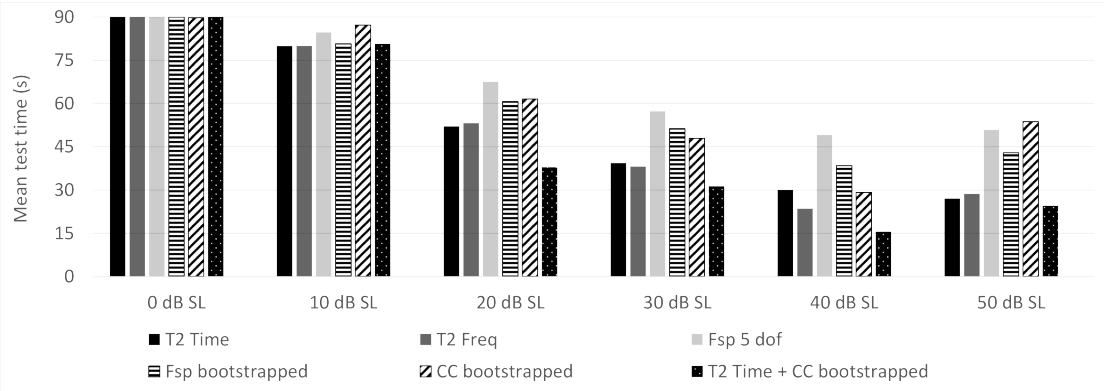


Figure 6.4: The mean of the detection times (calculated across 12 subjects) when detecting ABRs in a small sample of normal hearing adults (data set **D2**), per method and per dB SL condition.

Discussion

Starting with the CC, performance is relatively poor, most notably so for the 50 dB SL condition. This is somewhat in agreement with the simulations, which show a maximum detection rate of around ∼50% when simulating a -28 dB response (Fig. 6.2). As noted in the discussion (section 6.2), a possible explanation is that some responses did not match well with the templates. An additional explanation might be due to the way in which the templates were constructed. In particular, visual inspection by an experienced audiologist was used to first pre-select ensemble coherent averages that contained a 'clear response'. The 'CC templates' were then constructed using these 'clear responses'. This procedure may have resulted in a selection of 'stereotypical' responses by the audiologist. The similarity between the ABR waveforms (and hence the CC templates) might therefore have been larger than what might be typically observed in practice.

Despite the relatively poor performance of the bootstrapped CC in this section, the

'T2 Time + CC' combination was still the best performing ABR detection method. In terms of detection rates, it significantly outperformed Fsp 5 dof, whereas in test time, it significantly outperformed Fsp 5 dof, the bootstrapped CC, and the bootstrapped Fsp. Results also suggest a small advantage for 'T2 Time + CC' over the standard Hotelling's T2 test (applied in either the time or frequency domain), which, although not significant for a small sample size of just 12 subjects, was consistent across test conditions.

With respect to the detection rates, note that just a single comparison between methods was found to be significant. This can partly be attributed to a loss of information when using detection rates. In particular, the p values are transformed into 'all or nothing quantities', i.e. a 0 for $p < \alpha$ or a 1 for $p > \alpha$. The 'all or nothing' values thus no longer discriminate between weakly significant and highly significant p values, e.g. a p value of 0.0501 and a p value of 0.9 would both be rounded up to 1. When also considering the small sample size of just 12 subjects, along with a relatively high sample variance, then it is no longer surprising that the majority of the comparisons between methods were not significant. This is in contrast to the detection times, which did not suffer from the loss of information, and where many comparisons were indeed significant.

6.4 Summary

This Chapter used simulations and subject data to evaluate and compare the performance of various existing and new ABR detection methods. The methods selected for the analysis were chosen primarily based on the literature review presented in Chapter 2, whereas the statistical features were selected based on pilot simulations and results from feature optimisations (Appendix, section A.3), and otherwise on findings or recommendations from the literature. The overall goal for this Chapter was to find or design an ABR detection method with good sensitivity and low test time, for some fixed type-I error rate. The main results in terms of specificity, sensitivity, and test time are summarised below.

Specificity

Throughout this Chapter, the FPRs of the methods mostly fell within the two-sided 95% CIs for α. For a few cases, weakly conservative and liberal test performances were observed, which were attributed to (i) violations to the underlying statistical assumptions, (ii) random fluctuations, and (iii) inaccurate CIs for the nominal α-levels. With respect to the statistical assumptions, the main concern for ABR detection methods was shown in Chapter 5 to be the independence assumption between epochs. When using a high-pass cut-off frequency of 30 Hz and a stimulus rate of 33.3 Hz, results from Chapter 5 suggest a small tendency towards a liberal test performance, which is in agreement with results presented in Table 6.1. When using a high-pass cut-off frequency of 100 Hz

and a stimulus rate of 33.3 Hz, results from Chapter 5 (and the Appendix, section A.5.3) suggest that independence is satisfied. The latter is also in agreement with the FPRs from section 6.2 (where the high-pass cut-off frequency was 100Hz), i.e. no consistently conservative or liberal test performances were observed (with the exception of Fsp 5 dof). With respect to inaccurate CIs for the nominal α-level, it is expected that the CIs throughout this Chapter were slightly too narrow. As discussed in sections 6.1 and 6.2, this may have been the result of a multitude of factors (see discussions in sections 6.1 and 6.2).

With respect to Fsp 5 dof and Fmp 5 dof, results demonstrate an overall conservative test performance, as predicted by Elberling & Don (1984). Comparing results from section 6.1 ($f_c = 30$) to section 6.2 ($f_c = 100$) also confirms that the specificity of 'Fsp 5 dof' and 'Fmp 5 dof' is dependent on the spectral content (and hence the DOF) of the data. This is of course not desirable for ABR detection, i.e. specificity should ideally be controlled as intended across recordings with varying DOF.

With respect to the Hotelling's T^2 test and the bootstrapped statistics, results demonstrate a relatively robust control of specificity throughout this Chapter. For the Hotelling's T^2 test, this can at least partly be attributed to the way in which features are weighted and combined using the feature covariance matrix, which takes the correlations between samples within epochs into account. For the bootstrapped statistics, the random resampling procedure preserves the correlations between samples within epochs, and thus also takes the correlations between samples into account. It is therefore hypothesized that the specificity of the Hotelling's T^2 test and bootstrapped statistics will be more robust across recordings with varying DOF, relative to ABR detection methods where significance is evaluated using (partly) pre-determined significance thresholds, as is the case for e.g. 'Fsp 5 dof' and 'Fmp 5 dof'.

Sensitivity and test time

Starting with the Fsp and the Fmp, results firstly demonstrate an improved sensitivity when evaluating test significance with the bootstrap approach, as opposed to using theoretical F-distributions with assumed DOF. For the simulations, evaluating test significance with the bootstrap approach (as opposed to using theoretical F-distributions) resulted a maximum increase in test sensitivity of ∼40%, whereas for the subject recorded ABR threshold series, the increase in test sensitivity was ∼25%. With respect to the remaining methods, a highly sensitive and robust performance for the Hotelling's T2 test was observed. When compared to the Fsp (evaluated using theoretical F-distributions), a maximum increase in test sensitivity of ∼60% was observed for the simulations, and ∼40% for the subject recorded data. The best performing method throughout this Chapter, however, was the bootstrapped 'T2 Time + CC' combination. For the simulations, a maximum increase in test sensitivity for 'T2 Time + CC' over the Fsp (evaluated using

theoretical F-distributions) of 70-75% was observed, whereas for the subject recorded data this was ∼50%.

With respect to the bootstrapped CC, performance was relatively poor for the subject recored data, and for large N in the simulations. As noted in section 6.2, this can likely be attributed to the adopted 'CC templates', which might have correlated well with some subjects, but poorly with others. This suggests that if the exact ABR waveform morphology for the subject in question is known *a priori*, that the bootstrapped CC may provide an exceptionally sensitive test statistic, else using the CC as ABR detection method might be a bit of a gamble, i.e. it may perform well in some subjects, but poorly in others.

Various additional observations worth mentioning include the performance of T2 Time versus T2 Freq, which was found to be more or less identical, under the condition that the feature sets were optimal (as was the case in section 6.2). Secondly, the discrepancy in sensitivity between T2 Freq and Modified q-sample V4, which demonstrates the importance of taking both the variance and the covariance of the features into account. The latter is achieved by the Hotelling's T^2 test by re-scaling and combining the feature means as a function of the feature covariance matrix. This may similarly have contributed (to some extent) to the advantage for the Hotelling's T^2 test over methods such as the Fsp, the Fmp, and the Mean Power statistics, which similarly neglect feature covariance when evaluating test significance.

Chapter 7

Multi-stage adaptive group sequential tests: the Convolutional Group Sequential Test

When using a conventional approach for statistical hypothesis testing, then the statistical analysis should be specified at the outset. This includes the choice for the statistical test, the features, and the ensemble size. Note that fixing the statistical analysis *a priori* may be inefficient for evoked response detection, as both the signal-to-noise ratio (SNR) and the morphology of the response can vary across subjects, recordings, and test conditions. To deal with uncertainty in the SNR and the response morphology, this Chapter introduces a novel adaptive sequential testing procedure for ABR detection. The main benefit of the procedure over existing test strategies is that the statistical analysis can be optimised online, throughout the sequential analysis, i.e. the test can be tailored specifically to the subject and recording in question. The approach is furthermore built around a new method for controlling the type-I error rate of sequentially applied statistical tests, called the Convolutional Group Sequential Test (CGST). The CGST revolves around the discrete convolution of truncated probability density functions, and allows the null distribution for the test statistic to be constructed at each stage of the sequential analysis. Because the null distribution remains tractable, the procedure for finding the stage-wise critical decision boundaries is greatly simplified.

The remainder of this Chapter is structured as follows: in section 7.1, some background and a very brief literature review on sequential testing (with the emphasis on the adaptive group sequential test) is presented. The CGST and its underlying theoretical framework is then introduced in section 7.2, after which some simulation results are presented in section 7.3. A brief discussion on various design parameters along with some connections to existing methods from the literature is then presented in section

7.4.

7.1 Background

There are many methods available in the literature for constructing the stage-wise critical decision boundaries and controlling the type-I error rate for sequential test procedures. These methods differ primarily in terms of (i) design flexibility, and (ii) how the data is analysed. Starting with the latter, a distinction is typically made between 'fully sequential tests', where data is analysed continuously as it becomes available, and 'group sequential tests', where data is analysed in distinct groups or blocks of observations. A group sequential test is furthermore called adaptive when data-driven adaptations to test parameters are permitted following each stage of the sequential analysis (Wassmer, 2000). The main advantage of an adaptive group sequential test is that the initial assumptions (e.g. the power of the EEG background noise, or the amplitude and morphology of the response) are relaxed, i.e. these can be updated (and the statistical analysis modified accordingly) as new data becomes available, which can help bring the trial to an unambiguous test outcome in terms of 'effect present' or 'effect absent'. With respect to design flexibility, this is related to how free the user is in terms of: choosing the total number of sequential stages for the analysis (e.g. some methods permit just 2 or 3 stages), how the available α-level is 'spread' across the sequential analysis (note that this is also related to statistical power), and the type of adaptations permitted following each stage of the sequential analysis (further discussed in the review below). The main focus for the following section is to provide a very brief review on sequential testing, with the emphasis on the adaptive group sequential test.

7.1.1 Literature review

Early applications for sequential testing with controlled type-I error rates date back to the late 1920s, and were designed for quality control in industrial processes (see e.g. Dodge & Romig, 1929; Shewart, 1931), i.e. for discriminating between defective and non-defective products. Besides improving quality control, sequential testing also allowed causes for potential defects to be identified (and eliminated) at each stage of the sequential analysis, thus resulting in a reduced probability of a defect (Sheward, 1931). Sequential test theory was later further developed by Wald (1947) who introduced the sequential probability ratio test (SPRT); a simple method for constructing decision boundaries (for either accepting or rejecting H_0) for a fully sequential likelihood ratio test, which is still frequently used up to the present day. From the 50s onwards, the (non-adaptive) group sequential test began to emerge (McPherson, 1974; Pocock 1977; O'Brien and Fleming, 1979), which became increasingly popular in clinical trials due to ethical and economical considerations. The design flexibility for these methods, however, was still relatively poor. In particular, the methods were typically restricted

to normally distributed responses, and data-driven adaptations to test parameters were not permitted. The latter was somewhat addressed by Lan & DeMets (1982) who introduced the α-spending function, which allowed decision boundaries to be specified without needing to know the number of stages or the ensemble size per stage. The choice for the ensemble size, however, still had to be made independently of the previously accumulated data, i.e data-driven adaptations were not permitted. This is not the case for the adaptive group sequential test.

The advantage of using an adaptive group sequential test is that the user is allowed to look at previously analysed data (it can be 'unblinded'), and use this data to adjust or optimise test parameters for all remaining stages of the sequential analysis, without compromising the overall type-I error rate (Wassmer, 2000). Examples of the type of adaptations permitted include sample size modifications (see e.g. Proschan & Hunsberger, 1995; Lehmacher & Wassmer, 1999), changes to the number of remaining tests within the trial (e.g. Hartung & Knapp, 2003), and even a change in the choice of statistical test and features selected for the analysis. As noted earlier, adaptive group sequential tests relax the initial assumptions regarding, for example, the effect size and sample variance, and can help bring the trial to an unambiguous outcome in terms of 'effect present' or 'effect absent'. Indeed, as noted by Proschan & Hunsberger (1995), various controversies in the literature may be the result of poorly designed trials, such as an overestimated effect size (and hence an underpowered test), which might have been prevented by use of a suitable adaptive group sequential test.

Many types of adaptive group sequential tests can be found in the literature, the majority of which are built around either (1) conditional error functions (see e.g. Proschan & Hunsberger, 1995; Liu & Chi, 2001; Muller & Shafer, 2001), i.e. the conditional probability of incorrectly rejecting the null hypothesis given the test statistic from the previous stage, or (2) analysing the data in disjoint sub-samples and finding an appropriate critical decision boundary for some combination function of the stage-wise p values or test statistic (Bauer & Köhne, 1994; Lehmacher & Wassmer, 1999; Brannath et al, 2002; Hartung & Knapp, 2003; Chang, 2006; Sheng & Qiu, 2007). The earlier designs in Bauer & Köhne (1994) and Proschan & Hunsberger (1995) allow various data-driven adaptations, but are still limited regarding (i) the number of stages permitted, and (ii) the choice for the stage-wise critical decision boundaries (e.g. how the available α is spent throughout the trial). Methods following these earlier designs strive to either simplify the construction of adaptive group sequential tests (e.g. Sheng & Qiu, 2007), or to provide additional design flexibility in terms of the choice for critical decision boundaries and the type of adaptations permitted (Fisher, 1998; Shen and Fisher, 1999; Lehmacher & Wassmer, 1999; Müller & Shäfer, 2001; Liu & Chi, 2001; Brannath et al, 2002; Hartung & Knapp, 2003; Chang, 2006; Sheng & Qiu, 2007).

Methods for constructing multi-stage adaptive designs that have good flexibility include the 'sum of p values approach' in Chang (2007), the 'new class of completely self-designing tests' described in Hartung & Knapp (2003), 'Recursive combination tests'

described in Brannath et al (2002), and a general approach based on conditional error functions described in Müller & Schäfer (2001). Starting with the method in Chang (2006): as the name suggests, at each stage of the sequential analysis, a p value is generated by the statistical test, which is combined (through summation) with all previously generated p values. At each stage of the analysis, the null hypothesis is evaluated using the summary statistic, and the trial can be stopped for either futility (H_0 is accepted) or efficacy (H_0 is rejected), else the trial proceeds to the next stage. The user is also free to spend the available α across the K stages as desired, and their method permits data-driven adaptations to both the sample size and the statistical test selected for the analysis. A potential disadvantage for their approach is that combining p values through summation can potentially result in a sub-optimal test sensitivity (see also the Appendix, section A.12).

A flexible, potentially powerful, and remarkably simple approach for designing adaptive group sequential tests is given by Hartung & Knapp (2003). The statistic consists of a sum of inverse χ^2-distributed random variables (see also section 7.2.3), and permits adaptations to (i) the sample size, (ii) the statistical test selected for the analysis, and (iii) the number of remaining stages within the trial. A potential disadvantage is that early stopping for futility is not permitted. The stage-wise type-I error rates are also 'hidden' from the user, i.e. it is not transparent in terms of how the available α is 'spread' across the K stages (further details presented in the discussion, section 7.3.2).

Fisher (1998) and Shen and Fisher (1999) have proposed various 'self-designing designs' where the number of stages and sample sizes need not be specified in advance. However, their method only allows the null hypothesis to be rejected after the final stage of the sequential analysis. This is in contrast to the methods described in Müller & Shäfer (2001) and Brannath et al (2002), which provide similar design flexibility as Fisher (1998) and Shen and Fisher (1999), whilst still allowing the test to be stopped early for futility and efficacy. The complexity of these methods, however, is quite high. Moreover, in Brannath et al (2002), the decision to stop early is based exclusively on the p value from the current stage. An overall p value is then computed only after the decision to stop has been made. Although the stage-wise critical decision boundaries can be chosen such that an early stopping will always result in the overall p value being smaller than α, this method might result in a loss of test sensitivity relative to some alternative methods.

In the remainder of this Chapter, a novel method for designing multi-stage adaptive group sequential tests is described, called the Convolutional Group Sequential Test, or CGST. In terms of design flexibility, the CGST is similar to Müller & Shäfer (2001) and Brannath et al (2002), except that data-driven adaptations to the stage-wise critical boundaries are typically not permitted. That said, an exception can be made for evoked response detection under certain conditions (see Chapter 9). It is also worth pointing out that the CGST will likely have an advantage in terms of sensitivity over Brannath et al (2002), as the choice to stop the test early for the CGST is based on all previously

generated p values (not just the p value from the current stage, as is the case in Brannath et al). An additional advantage for the CGST over Brannath et al (2002) and Müller & Schäfer (2001) is that it is considered to be an intuitive and accessible approach.

7.2 The CGST: theoretical framework and graphical illustrations

This section introduces the notation and underlying theoretical framework for the CGST, after which graphical illustrations are used to further clarify the approach. Consider first a sequential test procedure with K stages, i.e. the statistical test is applied to the data K times, with each stage considering a new group of (independent) samples. The choice for the statistical test will depend on the specific problem, but does not affect the CGST itself. The goal is to evaluate the global null hypothesis H_0 at nominal significance level α:

$$H_0 : H_{01} \cap ... \cap H_{0K} \tag{7.1}$$

where H_{0i} (for $i = 1, 2, ..., K$) is the null hypothesis at stage i. It is assumed that all stage-wise null hypotheses H_{0i} pose the same proposition (no ABR present), else the global null hypothesis becomes difficult to interpret. At each stage, a new group of samples is collected, and a p value is generated by analysing the group of samples with a statistical test. As is the case in some methods from the literature (Bauer & Köhn, 1994; Brannath et al., 2002; Chang, 2006; Hartung & Knapp, 2003; Sheng & Qiu, 2007), it is assumed that all stage-wise p values p_i (for $i = 1, 2, ..., K$) are stochastically independent and uniformly distributed on the [0,1] interval under H_0, which implies that the accumulated data cannot be pooled, but must be analysed in disjoint sub-samples. Data analysed in stage i, for example, cannot be re-analysed in the subsequent stages of the trial, neither can it be pooled with previously collected data. However, at each stage of the analysis, all previously generated p values can be combined into a summary statistic, after which the test can be stopped for either futility or efficacy. Futility implies that the summary statistic is sufficiently far from statistical significance, such that additional data collection is deemed futile, and H_0 is accepted, whereas efficacy implies that there is sufficient evidence for rejecting H_0 at level α. The CGST furthermore requires the summary statistic to be a summation of (potentially transformed) p values. The stage k summary statistic is thus defined as:

$$\Sigma_k = \sum_{i=1}^{k} f_i(p_i) \tag{7.2}$$

where $f_i(p_i)$ is the desired transformation at stage i for p_i. A typical transformation that may be used here is that of Fisher (Fisher, 1932), achieved by defining $f_i(p_i) = -2\ln(p_i)$. Note that although transformation is not necessary, combining untransformed p values through summation can potentially result in a small loss of test sensitivity relative to some alternative combination functions (see e.g. Chow &, 2007Chang). Fisher's method in particular has some desirable properties in terms of efficiency (Littel & Folks, 1971), which can be attributed to the $\ln(p_i)$ transform placing more emphasis on small p values (note also that a succession of small p_i is more likely when an evoked response is present). After combining the stage-wise p values, the test can be stopped at stage i for futility when $\Sigma_i < B_i$, or for efficacy when $\Sigma_i > A_i$, where A_i and B_i (for $i = 1, 2, ..., K$) are the stage i critical decision boundaries. Finally, note that it is assumed here that transformation $f_i(p_i)$ gives large values for small p_i, i.e. that $f_i(p_i)$ is monotonic with a negative gradient.

Critical decision boundaries

The method for finding the critical decision boundaries A_i and B_i, such that the nominal α-level of the full test is preserved, is built around the convolution theorem, which states (Grinstead & Snell, 1997):

> *The null distribution for the sum of two independent random variables is given by the convolution of their individual null distributions.*

Hence, if the stage-wise null distributions (the null distributions for $f_i(p_i)$, henceforth denoted by ϕ_i) are known, then these can be iteratively convolved (an additional convolution for each stage of the analysis), to find the null distribution for the combined statistic Σ_i, henceforth denoted by ϕ_{Σ_i}. An important caveat is that ϕ_{Σ_i} changes when proceeding from stage i to stage $i+1$, as it is not possible to enter stage $i+1$ with $\Sigma_i > A_i$ or $\Sigma_i < B_i$, else the trial would already have been stopped. The probability densities for the stage i rejection regions for ϕ_{Σ_i} should therefore be set to zero prior to convolving with ϕ_{i+1}. More formally, the null distribution for the combined statistic at stage two is given by:

$$\phi_{\Sigma_2} = \phi_1^{T[B_1, A_1]} * \phi_2 \tag{7.3}$$

and for all following stages by:

$$\phi_{\Sigma_i} = \phi_{\Sigma_{i-1}}^{T[B_{i-1}, A_{i-1}]} * \phi_i \tag{7.4}$$

where $*$ denotes convolution, and where $\phi^{T[B,A]}$ indicates that distribution ϕ contains non-zero values exclusively for the $[B, A]$ interval, i.e. that the distribution has been truncated to this interval. When using the discrete convolution, the ϕ_i distributions should furthermore be sufficiently smooth to ensure a good accuracy. For Fisher's method (the ϕ_i are χ^2_2-distributed), a good accuracy is obtained by defining ϕ_i on the [0,30] interval with a resolution of $\frac{1}{2000}$.

Once ϕ_{Σ_i} has been generated, then finding A_i and B_i is straightforward. In particular, the stage i critical boundary for efficacy A_i is found by numerically solving:

$$\Phi_{\Sigma_i}[A_i, \infty] = \alpha_i \tag{7.5}$$

where α_i is the desired type-I error rate for stage i, and where $\Phi_{\Sigma_i}[A_i, \infty]$ is the cumulative distribution function for Σ_i, calculated across the interval $[A_i, \infty]$. In practice, ∞ is of course replaced by a sufficiently large value (a value of 30 is sufficiently large when using Fisher's method). The α_i values (for $i = 1, 2, ..., K$) are furthermore chosen freely, under the condition that $\sum_{i=1}^{K} \alpha_i = \alpha$. Similarly, the stage i critical boundary for futility B_i is found by numerically solving:

$$\Phi_{\Sigma_i}[0, B_i] = \beta_i \tag{7.6}$$

where β_i is the desired fraction of tests to be rejected for futility (under H_0) for stage i. The β_i values (for $i = 1, 2, ..., K$) are also chosen freely, under the condition that $\alpha + \sum_{i=1}^{K} \beta_i \leqslant 1$. This procedure is now illustrated graphically using a generic example below.

7.2.1 Illustrations

The goal for this section is to clarify the approach for a three stage sequential design using illustrations presented in Fig. 7.1. First, let the nominal α-level be 0.15 (an unusually high type-I error rate is chosen for illustration purposes only), and be spread equally across 3 stages ($K = 3$), giving stage-wise type-I error rates $\alpha_1 = \alpha_2 = \alpha_3 = 0.05$. The β_i values are furthermore specified as $\beta_1 = 0.2$, $\beta_2 = 0.4$, and $\beta_3 = 0.25$, i.e. for the current example $\alpha + \sum_{i=1} \beta_i = 1$. Further considerations on how to chose the stage-wise α_i and β_i values are given in the discussion. For the p value combination function, the generalized inverse χ^2-method (see e.g. Hartung & Knapp, 2003) is used, which is

defined as:

$$\Sigma_k = \sum_{i=1}^{k} [\chi^2_{v_i}]^{-1}(1 - p_i) \tag{7.7}$$

where $[\chi^2_{v_i}]^{-1}$ is the inverse of a χ^2 distribution with v_i DOF, and where DOF v_i (for $i = 1, 2, ..., K$) are chosen freely by the user. Note that transforming p_i using $[\chi^2_{v_i}]^{-1}(1-p_i)$ results in a $\chi^2_{v_i}$-distributed random variable (under the condition that p_i is uniform on the [0,1] interval under H_0), i.e. for the current example, the ϕ_i distributions are $\chi^2_{v_i}$-distributed. Note also that the v_i values function as weights for the stage-wise p values, with larger values corresponding to a larger weighting, i.e. when DOF v_i are increased, then the $[\chi^2_{v_i}]^{-1}(1 - p_i)$ transform will give larger values, in which case p_i will make a larger contribution towards summary statistic Σ_k. It is also worth mentioning here that when $v_i = 2$ for all i, Fishers method is obtained. For the current example, v_1, v_2, and v_3 are set to 2, 3, and 4 respectively (chosen to illustrate the possibility of using distinct functions at each stage). The choice for statistical test along with the ensemble size for the first stage of the analysis is then also specified, after which data is collected and analysed with the statistical test, thus generating p value p_1. The test can be stopped for efficacy if $p_1 \leqslant \alpha_1$, and for futility if $p_1 \geqslant 1 - \beta_1$, else the trial proceeds to stage two of the analysis. It is worth emphasizing here that ϕ_1 need not be generated for the first stage of the analysis. For completeness, however, the ϕ_1 distribution (given in this example by a χ^2_2 distribution, in accordance with the choice $v_1 = 2$) is shown in Fig. 1 (plot a), along with the stage one critical boundaries A_1 and B_1 for Σ_1. Efficacy boundary A_1 was found by solving Eq. 7.5, i.e. the area under ϕ_1 to the right of A_1 should equal $\alpha_1 = 0.05$, giving $A_1 = 5.992$. Futility boundary B_1 was found by solving Eq. 7.6, i.e. the area under ϕ_1 to the left of B_1 should equal $\beta_1 = 0.2$, solved for $B_1 = 0.446$.

Assuming p_1 fell within the $[B_1, A_1]$ interval, stage 2 is initiated by collecting a second group of samples. Stage two data is then analysed with the statistical test (note again that data analysed in stage one cannot be re-analysed here), giving p value p_2. Results from stages one and two are then combined using Eq. 7.7, giving $\Sigma_2 = [\chi^2_2]^{-1}(1 - p_1) + [\chi^2_3]^{-1}(1 - p_2)$, and the null distribution for Σ_2 is found using Eq. 7.3:

$$\phi_{\Sigma_2} = [\chi^2_2]^{T[B_1,A_1]} * \chi^2_3 \tag{7.8}$$

That is, ϕ_1 (given by a χ^2_2 distribution, in accordance with the choice $v_1 = 2$) is truncated to the $[B_1, A_1]$ interval, and convolved with ϕ_2 (given by a χ^2_3 distribution, in accordance with the choice $v_2 = 3$). This procedure is illustrated in Fig. 1: plot (b) shows $\phi_1^{T[B_1,A_1]}$, i.e. ϕ_1 where the stage one H_0 rejection and acceptance regions have been truncated.

It might be pointed out here that if H_0 were to be true, that 25% of the tests would have already been stopped after stage 1, i.e. $(\alpha_1 \cdot 100)\% = 5\%$ of the tests would have been stopped for efficacy, and $(\beta_1 \cdot 100)\% = 20\%$ for futility. Upon entering stage two, the area under $\phi_1^{T[B_1,A_1]}$ is therefore $1 - \beta_1 - \alpha_1 = 0.75$. The truncated stage one null distribution $\phi_1^{T[B_1,A_1]}$ (plot b) is then convolved with ϕ_2 (plot c), giving ϕ_{Σ_2} (plot d). Stage two critical boundaries A_2 and B_2 can then again be found by solving Eq. 7.5 and 7.6, respectively, i.e. the area under ϕ_{Σ_2} to the right of A_2 should equal $\alpha_2 = 0.05$, giving $A_2 = 9.695$, whereas the area under ϕ_{Σ_2} to the left of B_2 should equal $\beta_2 = 0.4$, giving $B_2 = 4.798$. If $\Sigma_2 \leqslant B_2$ or $\Sigma_2 \geqslant A_2$, the test is stopped for futility and efficacy, respectively, else the trial proceeds to stage three.

Assuming Σ_2 fell within the $[B_2, A_2]$ interval, a third group of samples is collected for the third (and for this example final) stage of the analysis. Stage three data is then analysed, giving p value p_3, which is combined with p_1 and p_2 using Eq. 7.7, now giving $\Sigma_3 = [\chi_2^2]^{-1}(1 - p_1) + [\chi_3^2]^{-1}(1 - p_2) + [\chi_4^2]^{-1}(1 - p_3)$. The null distribution for Σ_3 is then found using Eq. 7.4:

$$\phi_{\Sigma_3} = \phi_{\Sigma_2}^{T[B_2,A_2]} * \chi_4^2 \tag{7.9}$$

That is, ϕ_{Σ_2} is truncated to the $[B_2, A_2]$ interval, after which it is convolved with ϕ_3 (given in this example by a χ_4^2 distribution, in accordance with the choice $v_3 = 4$). The procedure is again illustrated in Fig. 1: plot (e) shows ϕ_{Σ_2} where the stage two rejection regions have been truncated, thus further reducing the area under $\phi_{\Sigma_2}^{T[B_2,A_2]}$ to $1 - \sum_{i=1}^{2} \alpha_i - \sum_{i=1}^{2} \beta_i = 0.3$, and plot (f) shows ϕ_3 (a χ_4^2 distribution). Convolving plots (e) and (f) gives ϕ_{Σ_3}, shown in plot (g). The stage three critical boundaries A_3 and B_3 are then found using the same procedure as in stages one and two: the area under ϕ_{Σ_3} to the left of B_3 should equal to $\beta_3 = 0.25$, giving $B_3 = 13.396$, and the area under ϕ_{Σ_3} to the right of A_3 should equal to $\alpha_3 = 0.05$, giving $A_3 = 13.396$. Note that when $\alpha + \sum_{i=1}^{K} \beta_i = 1$, that the critical boundaries for futility and efficacy at the final stage of the analysis will be the same, i.e. H_0 is either accepted for $\Sigma_3 \leqslant B_3 = A_3$, or rejected for $\Sigma_3 \geqslant B_3 = A_3$.

7.2.2 Simulations

This section describes and presents results from simulations. The goal is to explore sensitivity and test time as a function of the SNR for different choices for K and β_i.

Method

Data for the simulations consists of ensembles of $N = 3000$ epochs of simulated coloured noise (constructed as described in section 4.4) for representing the EEG background activity, along with scaled ABR templates (data set **D4**) for representing a response.

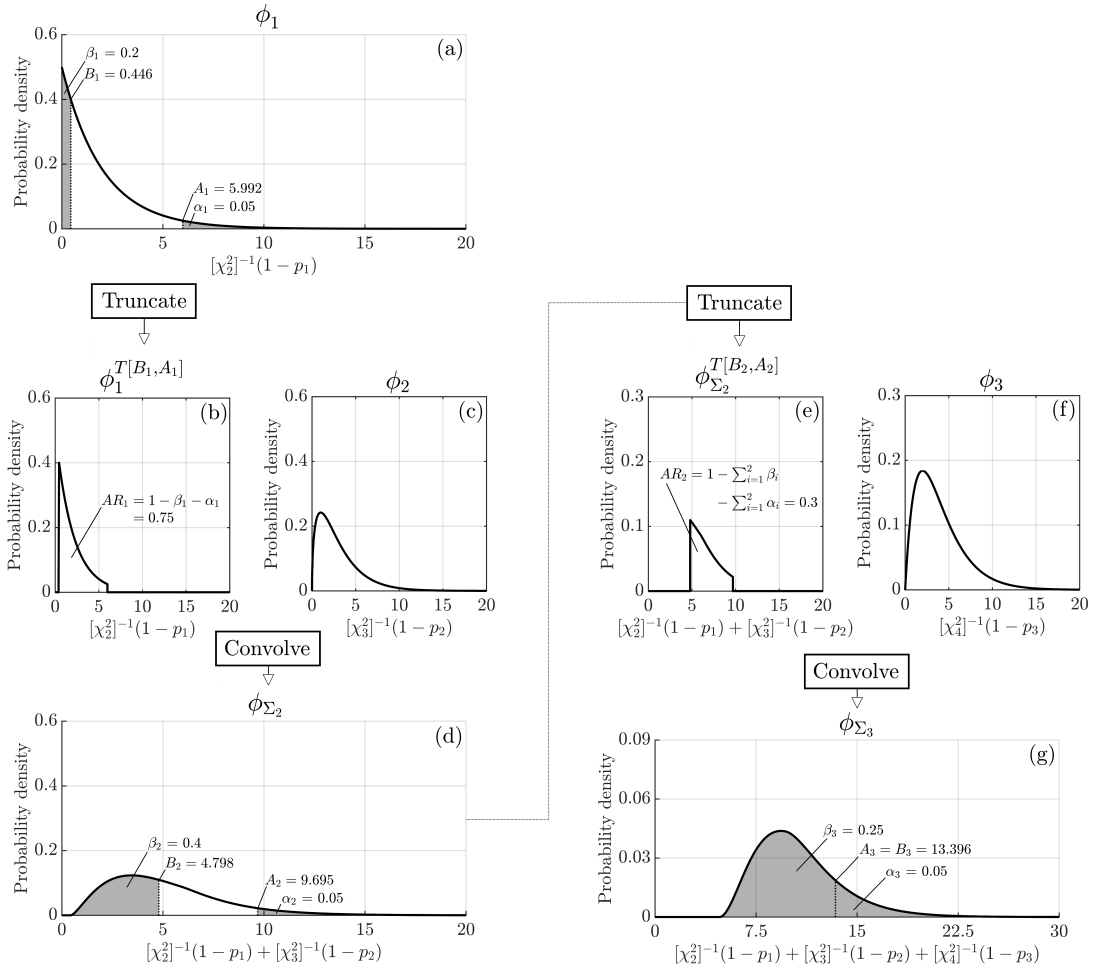


Figure 7.1: An overview of the approach for generating the critical decision boundaries for a three-stage group sequential test design. Details are presented in the text.

The SNR of the simulated response (calculated as described in section 4.3) took values of ranging from -20 dB to -50 dB, in steps of 0.5 dB. The no-stimulus condition was also included, i.e. SNR $= -\infty$. The resulting ensembles were then analysed using the Hotelling's T^2 test (applied in the time domain to 25 TVMs) in K sequential stages, giving stage-wise ensemble sizes of $\frac{3000}{K}$, and where the number of stages K took values of either 1, 2, 4, or 8. The nominal α-level was set to 0.01, which was also split equally across the K stages, giving α_i values of $\frac{\alpha}{K}$ for all i and K. Finally, the analysis was performed both with and without futility stopping. When futility stopping was used, the β_i values were set to $\frac{0.9}{K}$, for all i and K, whereas when no futility stopping was used, the β_i values were all set to zero. It is worth noting here that these values were chosen primarily for exploring the performance of the CGST, and that an optimal selection of design parameters will depend on the specific application in question (see also the discussion in section 7.3).

Results

The true-positive-rates (TPRs) and mean test times (calculated across 100 000 tests) are shown in Figure 7.2 as a function of the SNR, for different K, and for both with

and without futility stopping. Results firstly demonstrate a reduced TPR for increasing K (plots a and c), i.e. analysing N samples using the single shot test will always have a higher statistical power relative to analysing the same N samples using multiple sequentially applied statistical tests (see also Bauer & Köhne, 1994). Test time, however, will tend to be higher for the single shot test (plots b and d), as the test is only applied after the full $N = 3000$ stimuli have been presented. In terms of futility stopping, this had no noticeable effect on the TPR for these simulations (plots a and c). Moreover, when the SNR was relatively large (approximately >-30 dB), futility stopping also had no noticeable effect on the mean test time (plot d). For small SNRs (approximately <-30 dB), however, futility stopping resulted in relatively large reductions in mean test time (plot d). The extent to which futility stopping affects test performance is therefore dependent on the SNR of the response, but also on the choice for the β_i values. In particular, when the evoked response has a high SNR and the β_i values are chosen conservatively, then the Σ_i values will tend to be much larger than the B_i futility boundaries, and the test will typically not be stopped for futility. Vice versa, when the SNR is low (or a response is absent) and the β_i values are chosen more liberally, then the Σ_i values will tend to be closer to the B_i values, and the probability of stopping the test early in favour of H_0 is increased, potentially resulting in an increased false-negative rate. A more liberal choice for the β_i values might therefore result in larger reductions in test time, potentially at the cost of a reduced test sensitivity.

With respect to the no-stimulus condition: when no futility stopping was used, results show FPRs of 0.00949, 0.00989, and 0.00988 for $K = 2$, $K = 4$, and $K = 8$, respectively, whereas when futility stopping was used, the FPRs were 0.00953, 0.00994, and 0.00992 for $K = 2$, $K = 4$, and $K = 8$, respectively. For the single shot test ($K = 1$), a FPR of 0.0096 was observed. The two-sided 95% confidence intervals for the expected 0.01 FPR are furthermore given by [0.0094, 0.0106]. Hence, no significant deviations from the expected 0.01 FPR were observed for this data. The confidence intervals were furthermore found using a binomial distribution, constructed from 100 000 observations, where the probability of a single 'successful' Bernoulli trial (defined here as a false-positive) was set to 0.01 (the theoretical probability of a false-positive)

7.3 Discussion

This Chapter presented the CGST; a relatively simple and intuitive method for finding the stage-wise critical decision boundaries (for rejecting or accepting H_0) and controlling the type-I error rate for sequentially applied statistical tests. Although originally designed for evoked response detection, it should be stressed that the CGST can potentially be used for a wide range of applications. The only condition for using the CGST is that the following two assumptions are satisfied: (i) the ϕ_i distributions (for $i = 1, 2, ..., K$) are mutually independent under H_0, and (ii) the ϕ_i distributions (for $i = 1, 2, ..., K$) are known *a priori*. With respect to (ii), it was assumed throughout this

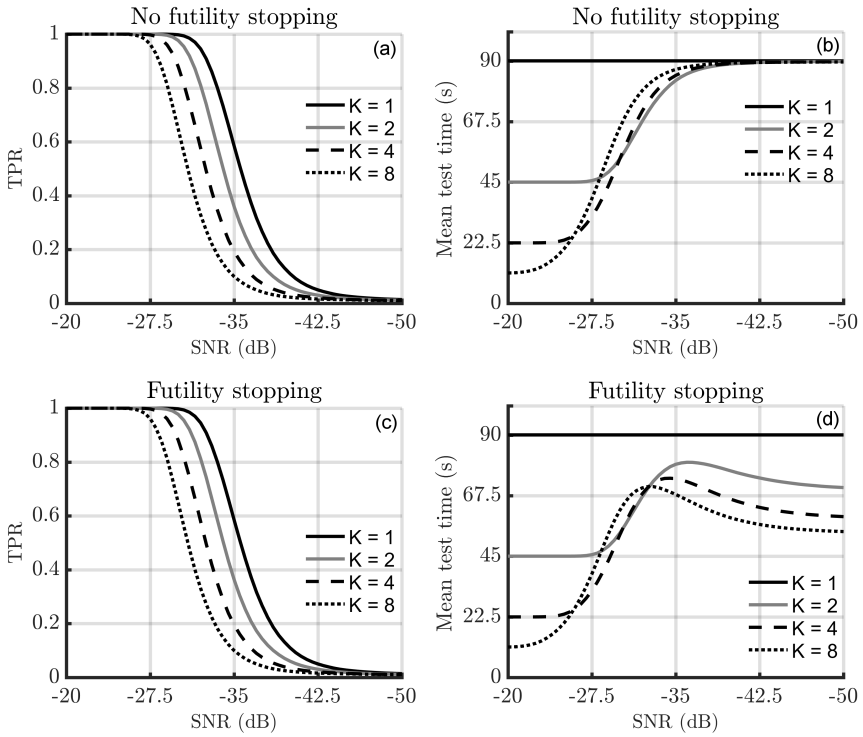


Figure 7.2: Results from the simulations, which include the true-positive-rate (TPR) and mean test time (calculated across 100 000 tests) as a function of the SNR, for various K, both with futility stopping (plots c and d) and without (plots a and b).

work that the stage-wise p values were uniform on the [0,1] interval under H_0. Note that this is only true when the assumptions underlying the statistical test (used for analysing the data) are satisfied. This emphasizes the importance of choosing an appropriate statistical test for the data analysis, i.e. the assumptions underlying the statistical test should be satisfied for the data in question, else additional violations (originating from the CGST) might be introduced, potentially resulting in increased or decreased type-I and type-II error rates.

With respect to test performance, the main advantage of using the CGST over a conventional single shot approach is a reduced mean test time, which comes at the cost of a reduced statistical power. As demonstrated in section three, the trade-off between test time and statistical power is dependent on both the SNR of the response, and the CGST design parameters selected for the analysis. The trade-off can therefore be optimised for the specific application in question by a suitable selection of CGST design parameters, which include; the number of stages K; the ensemble size N; the α_i and β_i values; and the p value transformation functions $f_i()$. Various trade-offs associated with these parameters are further discussed below.

7.3.1 CGST design parameters

Starting with the number of stages K, a trade-off is introduced between a potential reduction in test time versus an increased type-II error rate (a reduced statistical power). As mentioned earlier, analysing N samples using a single test ($K = 1$) will always have a higher statistical power compared to analysing the same N samples using multiple sequentially applied tests (Bauer & Köhne, 1994). Test time, however, will tend to be higher for the single shot test, as the test can only be stopped after the full ensemble of epochs has been collected. For the most part, the optimal number of stages K will depend on the distribution of the SNR of the response (when considered across a cohort of subjects), i.e. when the distribution is disperse, it may be beneficial to use more stages so that the test can be stopped sooner for the higher SNR responses.

With respect to the ensemble size N, this is of course directly related to statistical power and test time, with an increased N going hand in hand with an increased statistical power and an increased test time. How large N should be is therefore dependent on the SNR and the desired TPR, but also on the number of stages K, i.e. if a reduced TPR for increasing K is to be prevented, then N should be increased with K (to compensate for the loss of statistical power). An additional consideration is how to split N across the K stages. Additional simulations (see Appendix A.17) suggest that test time is close to optimal (i.e. as low as possible for some fixed test sensitivity and specificity) when splitting the N epochs equally across the K stages, giving stage-wise ensemble sizes of $\frac{N}{K}$. Alternatively, the user may want to consider choosing N and the stage-wise ensemble sizes adaptively, based on a predictive power analysis using previously analysed data.

With respect to the α_i values, a trade-off is introduced between the type-I and type-II error rate (as is the case with all significance tests), with an increased α_i going hand in hand with an increased type-I error rate and a decreased type-II error rate. The α_i values might therefore be chosen to optimize statistical power throughout the trial. As an example, the user might expect a small effect size for stage ones, and a large effect size for stage two, and might therefore choose to assign more α to the second stage of the analysis. If the effect size is expected to be constant throughout the test, then the safe approach is to split the available α equally across the K stages, giving α_i values of $\frac{\alpha}{K}$.

In terms of the the p value transformation functions f_i, these might similarly be chosen to optimise test sensitivity. The v_i values in the sum of inverse χ^2-distributed random variables in Eq. 7.7, for example, can be used as a weighting for the stage-wise p values (see section 7.2.1). Hence, if the user expects a large effect size in stage two and a small effect size in stage one, then v_2 can be chosen to be larger than v_1, which would give more emphasis to stage two data when evaluating test significance.

Finally, with respect to the β_i values, a trade-off is again introduced between statistical power and test time, i.e. larger β_i values result in an increased probability of stopping

the test in favour of H_0, which decreases test time, potentially at the cost of a type-II error. An additional effect associated with the β_i values is that they reduce the remaining area under the null distribution, which affects the critical decision boundaries (for both efficacy and futility) for the remaining stages. Taking the example presented in section II, and setting $\beta_1 = 0$ (as opposed to $\beta_1 = 0.2$) would give stage two critical boundaries $A_2 = 9.899$ and $B_2 = 3.654$, as opposed to $A_2 = 9.694$ and $B_2 = 4.796$. Note that A_2 is reduced as β_1 is increased. Hence, under the right conditions, increasing the β_i values can sometimes prevent a type-II error. A liberal choice for the β_i values, however, may result in a significantly higher type-II error rate. The choice for the β_i values will therefore again depend on the application in question. In general, a relatively conservative choice would be to either split the available β (equal to 1-α) equally across the K stages (giving β_i values of $\frac{1-\alpha}{K}$), or to choose small β_i for the early stages, and slightly larger β_i for the later stages.

7.3.2 Some connections to existing methods

Various connections between the CGST and methods such as 'self designing tests' described by Hartung and Knapp (2003) can be identified. In Hartung & Knapp, data is analysed in disjoint groups of samples (as is the case with the CGST). At each stage of the analysis, a p value is generated using a statistical test (all p values are again assumed to be independent and uniformly distributed on $[0, 1]$ under H_0), which is combined with all previously generated p values using the generalized inverse χ^2-method. At each stage of the analysis, the test can be stopped for efficacy if the summary statistic exceeds some upper threshold A_{v_Σ}, which is defined as:

$$A_{v_\Sigma} = [\chi^2_{v_\Sigma}]^{-1}(1 - \alpha) \tag{7.10}$$

and where v_Σ are the DOF of a χ^2 distribution. DOF v_Σ should furthermore be specified by the user *a priori*, and functions as a 'currency' that the user is free to 'spend' throughout the trial, i.e. at each stage of the analysis, the user is free to specify how much of the available v_Σ to spend for the next stage. This procedure can be repeated until the test is stopped for efficacy, or until v_Σ has been depleted.

A potential advantage for the approach in Hartung & Knapp (relative to some alternative methods, including the CGST) is that the total number of stages K need not be specified in advance. A potential disadvantage for their approach is that early stopping in favour of H_0 is not permitted. Note also that the stage-wise type-I error rates are 'hidden' from the user. A potential advantage for the CGST over various alternative methods is indeed its clarity in terms of how the type-I error rate is spread across the trial, i.e. the user is given the choice to explicitly specify the desired stage-wise type-I error rates

through the α_i values. Finally, it is worth noting that the critical boundary A_{v_Σ} can also be generated with the CGST, achieved by convolving any number of $\chi^2_{v_i}$ distributions where $\sum_{i=1}^{K} v_i = v_\Sigma$, and by setting all α_i and β_i values to zero, except for α_K which should be set to α. The approach in Hartung & Knapp can therefore be seen as a special case of the CGST.

Connections with additional methods worth mentioning include the 'sum of p values' approach described by Chang (2007), which can be represented by the CGST by setting the combination function to $\Sigma_k = \sum_{i=1}^{k} w_i p_i$, where w_i is the chosen weight for stage i. The ϕ_i distributions are then uniformly distributed on the $[0, w_i]$ interval. The CGST also covers the class of adaptive group sequential tests described by Bauer & Köhne (1994), achieved by using Fisher's method as p value combination function, and by choosing appropriate values for K, α_i, and β_i.

7.4 Conclusion

The CGST is a flexible and intuitive method for finding the stage-wise critical decision boundaries, and controlling the type-I error rate of sequentially applied statistical tests. Although originally designed for ABR detection, the CGST can be used for a wide range of sequential test applications, under the condition that the stage-wise p value null distributions are mutually independent under H_0, and that their null distributions are known *a priori*. The main advantage of using the CGST over a single shot test is furthermore a reduced test time, which comes at the cost of a reduced statistical power. The trade-off between statistical power and test time is dependent on both the SNR or the response, and the selection of CGST design parameters. A suitable selection of CGST design parameters is therefore essential when optimising test performance. The CGST furthermore falls under the class of 'adaptive group sequential tests'; a class of group sequential tests that permit data-driven adaptations to test parameters, without compromising the type-I error rate. For the CGST, adaptations to the sample size and the statistical test selected for the analysis are permitted, which might be exploited when further optimising sequential test procedures. Finally, as shown in the discussion, the CGST is a generalized form of some alternative adaptive group sequential tests found in the literature, and one that facilitates understanding, and allows greater flexibility in the choice of approach.

Chapter 8

The non-adaptive CGST for ABR detection

This chapter explores the specificity, sensitivity, and test time of a sequentially applied Hotelling's T^2 test for ABR detection, where the critical decision boundaries for rejecting or accepting H_0 are found using the CGST. The aim is firstly to verify that the assumptions underlying the CGST remain satisfied (and the FPR controlled as intended) across a range of EEG pre-processing parameters and test conditions. Secondly, the aim is to explore trade-offs betweens statistical power and test time for ABR detection as a function of (i) the number of sequential stages used for the analysis, and (ii) the β_i values. The overall goal is to explore the potential benefit of a sequential test procedure for ABR detection, and to provide general guidelines and/or preliminary recommendations for selecting sequential test design parameters.

The structure for this chapter is as follows: the approach for choosing the β_i values is first described in section 8.1 below, after which a specificity assessment is conducted across a range of CGST design parameters and EEG pre-processing parameters in section 8.2. Section 8.3 then explores test time and sensitivity for the stimulus and no-stimulus conditions as a function of K and β_i. It is worth emphasizing here that data-driven adaptations are not explored in this chapter, but are instead considered in chapter 9.

8.1 The stage-wise true-negative rates

Throughout this chapter, various CGST design parameters are fixed in advance (data-driven adaptations are not explored), which includes the stage-wise true negative rates (TNRs), specified through the β_i values. As mentioned in Chapter 7, the β_i values introduce a trade-off between statistical power and test time, i.e. increasing the β_i values increases the probability of stopping the test early in favour of H_0, thus reducing test time, potentially at the cost of a false-negative.

Ideally, the β_i values would be chosen based on knowledge of both the null and the alternative distribution of the test statistic, as this would allow both the TNR and the FNR to be controlled (further illustrated in chapter 9). In practice, however, the alternative distribution is almost always unknown, and even estimating it from previously analysed data can be problematic. For this chapter, the β_i values are instead chosen as a function of the stage index i. The rationale is that when an ABR is present, statistical power will tend to increase as the ensemble size is increased, i.e. the stage i summary statistic Σ_i will tend to be larger as more data becomes available. Consequently, the B_i values can be chosen more liberally as the trial progresses, without increasing the probability of a false-negative.

The most straightforward approach for choosing β_i is to simply split the available β equally across the K stages, giving β_i values of $\frac{1-\alpha}{K}$ for all i and K. Alternatively, it may be beneficial to assign less β in the early stages, and more in the later stages (or vice versa). In the following section, various functions are described for relating the stage index i to β_i. These functions are henceforth referred to as a 'futility functions'.

8.1.1 Futility functions

The functions for relating stage index i to β_i take the form of either cosine or exponential ramps. A cosine ramp gives values from 0 to 1, and is defined on the [1.5π, 2π] interval. The function is given by:

$$f_{cos}(x) = cos(x)^{c_1} \tag{8.1}$$

where c_1 is a scaling factor that determines how steep the ramp is. Note that x can only take values from 1.5π to 2π. The exponential ramp similarly gives values from 0 to 1, and is defined on the [0, c_2] interval:

$$f_{exp}(x) = 1 - exp(-x) \tag{8.2}$$

where c_2 can again be used to determine the steepness of the slope (see below). Note that x is again restricted to a specific interval, now defined from 0 to c_2.

The β_i values can now be generated as follows; When using the cosine ramp, β_i for stage i is given by:

$$\beta_i = \beta_{R_i}\left[1 - cos\left(1.5\pi + \frac{i(2\pi - 1.5\pi)}{K}\right)^{c_1}\right] \tag{8.3}$$

and when using the exponential ramp, β_i for stage i is given by:

$$\beta_i = \beta_{R_i}\left[1 - exp\left(-i\frac{c_2}{K}\right)\right] \tag{8.4}$$

where β_{R_i} is the maximum available β that can be spent at stage i, given by $\beta_{R_i} = 1 - \alpha - \sum_{j=1}^{i-1}\beta_j$.

Finally, when using a cosine ramp, constant c_1 is set to either 1 (denoted by Cos 1) or to 3 (denoted by Cos 3). When using an exponential ramp, constant c_2 is set to either 5 (denoted by Exp 5) or to 15 (denoted by Exp 15). The actual values for β_i for $K = 2, 3, ..., 9$ and the shape of the futility functions are given in the Appendix (section A.10).

8.2 Specificity

The aim for this section is to test whether the assumptions underlying the CGST are satisfied, and the FPR controlled as intended, across a range of CGST design parameters and EEG pre-processing parameters, and when using the Hotelling's T^2 test (applied in the time-domain) as objective detection method. Throughout this section, the α-level is always set to 0.05, which is spread equally across the K stages, giving stage-wise type-I error rates of $\frac{\alpha}{K}$. For the p value combination function, Fisher's method is used, given by (Fisher, 1932):

$$\Sigma_k = \sum_{i=1}^{k} -2ln(p_i) \tag{8.5}$$

8.2.1 Method

This section first describes simulations for evaluating the independence assumptions between (i) epochs (underlying the Hotelling's T^2 test), and (ii) the stage-wise p values (underlying the CGST). Simulations and recordings of EEG background activity are then used to further test whether specificity remains controlled as intended for ABR detection across a range of β_i values.

Independence assessment

Data consists of 1 000 000 recordings of simulated coloured noise, generated as described in section 4.4. The simulated recordings were all band-pass filtered using a 3rd order Butterworth filter from either 30-1500 Hz, or from 100-1500 Hz (corresponding to typical values used in the literature), after which they were structured into ensembles of $N = 500$

15 ms windows. The distance between the 15 ms windows, denoted by τ, was then varied from 0 to 25 ms, in steps of 0.4 ms, which corresponds to a (hypothetical) stimulus rate of $\frac{1000}{15+\tau}$ (covering stimulus rates of 25.13 Hz up to 66.67 Hz). The 15 ms windows of the ensembles were analysed in K sequential stages using the Hotelling's T^2 test, where K took values ranging from 1 to 9. Finally, the β_i values were set to $\frac{1-\alpha}{K}$, for all i and K. It might be noted here that the large number of simulated recordings were necessary in order to discriminate between relatively small differences in the FPR as a function of K (see also the results section). The IRIDIS High Performance Computing Facility was used for generating and analysing the simulated recordings for this section.

Simulations and EEG background activity

Result from the previous simulations (section 8.2.2 below) suggest that the underlying assumptions are satisfied when using band-pass filter settings of 100-1500 Hz in combination with a stimulus rate to 33.3 Hz. The aim for this section is to verify that specificity is indeed controlled as intended when using these test parameters. The β_i values for this section are furthermore chosen using the previously described futility functions (see section 8.1). The 'no futility stopping' condition was also included, i.e. all β_i were set to zero (early stopping in favour of H_0 was not permitted). For the simulations, data consisted of 10 000 recordings of simulated coloured noise, constructed as described in section 4.4, and structured into ensembles of N epochs, where N was set to either 500 or 3000 epochs. The resulting ensembles were then analysed in K sequential stages with the Hotelling's T^2 test, where K ranging from 1 to 9. For the real EEG background activity, the recordings (data set **D1**) were downsampled to 5 kHz, band-pass filtered from 100-2000 Hz, and structured into ensembles of N epochs, where N was set to either 555 or 3333. Artefact rejection was then applied by throwing away 10% of the noisiest epochs (as determined by their maximum absolute values), resulting in ensemble sizes of $N = 500$ and $N = 3000$ after artefact rejection. Note that the ensembles did not overlap, i.e. data was used at most once. This resulted in a total of 2156 ensembles for $N = 500$, and 324 ensembles for $N = 3000$. The initial 15 ms windows of the ensembles were then analysed in K sequential stages using the Hotelling's T^2 test, where K ranged from 1 to 9.

8.2.2 Results

Independence assessment

The FPRs from the independence assessment are presented in Figure 8.1 as a function of the (hypothetical) stimulus rate, for high-pass cut-off frequencies of either 30 Hz (plot A) or 100 Hz (plot B). The FPRs for $K = 3, 4, ..., 8$ were all quite similar (they fell between the FPRs from $K = 2$ and $K = 9$), and are excluded from the Figure to avoid cluttering. The two-sided 95% confidence intervals for $\alpha = 0.05$ were very narrow, and are given by [0.0496, 0.0504]. For the single shot test ($K = 1$), results demonstrate significant fluctuations around $\alpha = 0.05$ as a function of the stimulus rate and the

high-pass cut-off frequency, which can be attributed to a violation of the independence assumption between epochs. For the sequential test ($K > 1$), the FPRs follow a similar but more pronounced trend, which implies that additional assumptions underlying the CGST were violated. The latter was further explored with a post-hoc analysis.

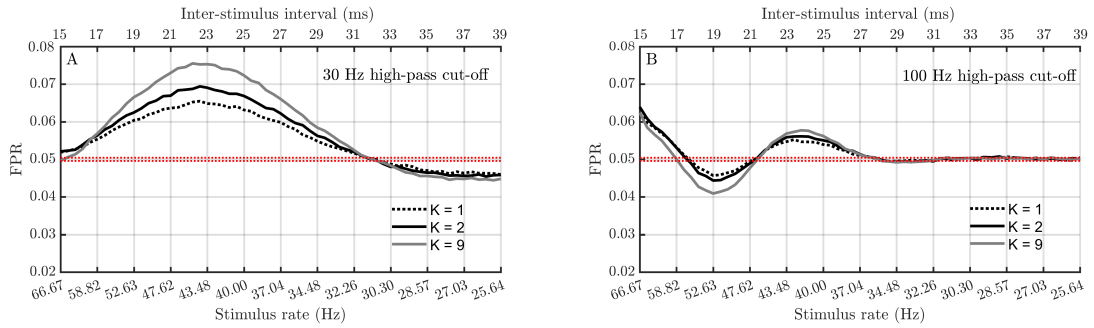


Figure 8.1: FPRs generated by the Hotelling's T^2 test when applied to simulated coloured noise, as a function of the (hypothetical) stimulus rate, when using band-pass filter settings of either 30-1500 Hz (plot A) or 100-1500 Hz (plot B). Results are presented for $K = 1$ (giving a single shot test), $K = 2$, and $K = 9$.

Post-hoc simulations

The goal for the post-hoc simulations is to determine whether the independence assumption (underlying the CGST) between the stage-wise p values was violated or not. This was achieved by inserting a short pause, denoted by τ_2, between each stage of the sequential analysis. The rationale is that as τ_2 is increased, then consecutive blocks of observations will become sufficiently distant in time to be uncorrelated, and independence (between the stage-wise p values) will be satisfied. To test this, The high-pass cut-off frequency f_c was set to 60 Hz, and the stimulus rate to 33.1 Hz. These values were chosen based on Figure 8.1, which show relatively large independence violations (between epochs) when using these parameters. The number of stages K was then set to either 1 or to 9. It is hypothesized that the independence violation between p values will decrease as τ_2 is increased, and hence that the FPR for K = 9 will approach the FPR for $K = 1$ for increasing τ_2.

Results are presented in Figure 8.2: plot A shows the FPR as a function of τ_2. As can be seen, the FPR for $K = 9$ is unaffected by τ_2, which suggests that independence between the stage-wise p values is satisfied, or else that the violation is negligible. The increased FPRs for $K > 1$ observed in Figure 8.1 can therefore be attributed to the stage-wise p value null distributions being no longer uniform on the [0,1] interval (as is assumed by the CGST), which can, in turn, be attributed to the independence violation between epochs. In particular, when independence between epochs is violated, then the null distributions for the p values will be skewed. The latter is demonstrated in Figure 8.2, plot B, which shows a histogram of the p values generated in the first stage of the analysis. When using a high-pass cut-off frequency is increased to 100 Hz and a stimulus rate (SR) of 33.3 Hz, then independence between epochs is satisfied, and the p value null distributions are more or less perfectly uniform on the [0,1] interval (Figure 8.2,

plot C).

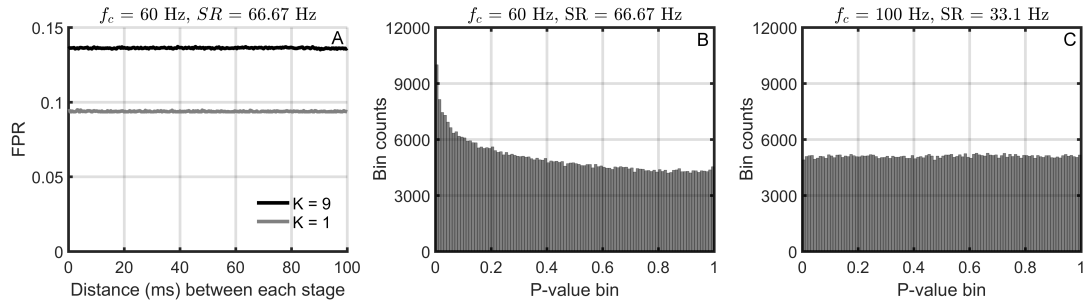


Figure 8.2: Results from post-hoc simulations for exploring the independence assumption (underlying the CGST) between the stage-wise p values. Details are presented in the text.

Simulations and EEG background activity

The FPRs ($\alpha = 0.05$) for the simulations and real EEG background activity are presented in Table 8.1 for different K and N, and for different β_i values. The binomial distribution was used to construct two-sided 95% confidence intervals for the expected FPRs, giving lower and upper boundaries of [0.0459, 0.0544] for the simulations (10 000 tests), and either [0.0413, 0.0598] (for $N = 500$, 2156 tests) or [0.0309, 0.0772] (for $N = 3000$, 324 tests) for the EEG background activity. Significantly (**p<0.05**) conservative and liberal test performances are indicated in Table 8.1 with blue and red asterisks respectively.

8.2.3 Conclusion

Significant violations to the independence assumption between epochs (underlying the Hotelling's T^2 test) were again observed as a function of the high-pass cut-off frequency and stimulus rate (Figure 8.1). The violation resulted in non-uniform p value null distributions, which resulted in an additional violation (now originating from the CGST) to the assumption that the stage-wise p values are uniform on the [0,1] interval under H_0. These results hence emphasize that care is required to ensure that the assumptions underlying the chosen ABR detection method are satisfied, else additional violations (originating from the CGST) might be introduced. The choice for the ABR detection method is therefore important, as some methods have a more robust control of specificity relative to others. The Hotelling's T^2 test, for example, has a good control of specificity relative to 'Fsp 5 dof' and 'Fmp 5 dof'. Bootstrapped statistics are also expected to give a good control of specificity, particularly so as they are robust to normality and stationarity violations.

With respect to the FPRs shown in Table 8.1, a total of 13 (of the 164) test conditions showed a conservative test performance. This can possibly be attributed to random fluctuations in combination with too narrow CIs for α (see also the discussions in sections 6.1-6.2). Alternatively, the stationarity and normality assumptions underlying the

Table 8.1: The FPRs ($\alpha = 0.05$) from the Hotelling's T^2 test when applied to simulated coloured noise and real EEG background activity in K sequential stages. The ensemble size N took values of either 500 or 3000 epochs. The β_i values were furthermore chosen using the futility functions described in section 8.1. The 'no futility stopping' condition was also included, i.e. early stopping in favour of H_0 was not permitted (denoted by **No Fut**). Significant (**p<0.05**) deviations from the nominal α-level are indicated by blue (conservative) and red (liberal) asterisks.

Simulated coloured noise										
	N = 500					**N = 3000**				
	Exp 15	**Exp 5**	**Cos 3**	**Cos 1**	**No Fut**	**Exp 15**	**Exp 5**	**Cos 3**	**Cos 1**	**No Fut**
K=1	-	-	-	-	0.0491	-	-	-	-	0.0465
K=2	0.0490	0.0480	0.0495	0.0490	0.0476	0.0449*	0.0461	0.0473	0.0477	0.0466
K=3	0.0503	0.0489	0.0524	0.0517	0.048	0.0473	0.0460	0.0466	0.0471	0.0468
K=4	0.0470	0.0478	0.0469	0.0474	0.0472	0.0457	0.0433*	0.0459	0.0462	0.0489
K=5	0.0485	0.0496	0.0491	0.0462	0.0478	0.0461	0.0463	0.0483	0.0484	0.0473
K=6	0.0491	0.0492	0.0490	0.0486	0.0469	0.0475	0.0465	0.0454*	0.0489	0.0515
K=7	0.0459	0.0485	0.0474	0.0469	0.0489	0.0480	0.0455*	0.0475	0.0492	0.0455
K=8	0.0487	0.0506	0.0494	0.0486	0.0485	0.0464	0.0454*	0.0464	0.0480	0.0478
K=9	0.0518	0.0491	0.0514	0.0502	0.0493	0.0461	0.0465	0.0476	0.0484	0.0507
EEG background activity										
	N = 500					**N = 3000**				
	Exp 15	**Exp 5**	**Cos 3**	**Cos 1**	**No Fut**	**Exp 15**	**Exp 5**	**Cos 3**	**Cos 1**	**No Fut**
K=1	-	-	-	-	0.0526	-	-	-	-	0.0516
K=2	0.0399*	0.0431	0.0431	0.0427	0.0476	0.0617	0.0586	0.0525	0.0494	0.0516
K=3	0.0427	0.0455	0.0468	0.0445	0.0445	0.0586	0.0494	0.0617	0.0586	0.0645
K=4	0.0533	0.0468	0.0515	0.0496	0.042	0.0556	0.0525	0.0432	0.0556	0.0613
K=5	0.0413	0.0399*	0.0445	0.0436	0.04*	0.0556	0.0679	0.0525	0.0741	0.0613
K=6	0.0519	0.0445	0.0459	0.0468	0.0481	0.0370	0.0309	0.0309	0.0340	0.0645
K=7	0.0510	0.0473	0.0455	0.0492	0.0405*	0.0463	0.0586	0.0556	0.0556	0.0516
K=8	0.0445	0.0422	0.0399*	0.0380*	0.045	0.0340	0.0556	0.0340	0.0340	0.0258*
K=9	0.0436	0.0390*	0.0436	0.0431	0.0425	0.0432	0.0617	0.0494	0.0463	0.0645

Hotelling's T^2 test may have been violated. As was the case with independence violations, violations to stationarity and normality may result in non-uniform p value null distributions, incurring additional violations originating from the CGST. That said, no noticeable relationships can be observed between the FPRs, the number of stages K, and the β_i values (see Table 8.1), which suggests that any additional violations originating from the CGST were negligible for this analysis.

8.3 Sensitivity and test time

The aim for this section is to use simulations and subject ABR data to explore the trade-off between sensitivity and test time, as a function of the number of stages K and the β_i values. The band-pass filter settings in this section were fixed at 100-1500 Hz, and the duration of the epochs to 30.03 ms (corresponding to a stimulus rate of 33.3 Hz).

8.3.1 Method

For the simulations that follow, data consists of simulated coloured noise, constructed as described in section 4.4, along with scaled ABR templates from data set **D4** for simulating a response.

Simulations I: TPR fixed at 0.99
For the first set of simulations, the aim is to obtain a fair comparison of test time for different choices of K. The true-positive rate (TPR) was therefore fixed at 0.99 for all K and all simulated conditions (including the 20, 30, 40, or 50 dB SL condition), achieved by repeatedly generating 10 000 ensembles with increasing or decreasing ensemble sizes N, until a TPR of 0.99±0.005 was obtained, which was repeated per K and per dB SL condition. Needless to say, this approach is not feasible in a clinical setting. The number of sequential stages for the analysis K was varied from 1 to 9. For this analysis, early stopping in favour of H_0 was not permitted (β_i are zero for all i and K).

Simulations II: N fixed at 3000
For the second set of simulations, the ensemble size N was fixed at 3000 epochs, for all K and all simulated conditions (including the 20, 30, 40, or 50 dB SL condition). Contrary to Simulations I, the loss in statistical power for increasing K cannot be compensated for by increasing N, i.e. a reduced TPR can be expected for increasing K. A total of 10 000 recordings were again simulated, and the initial 15 ms windows of the ensembles were analysed in K sequential stages using the Hotelling's T^2 test, where K was varied from 1 to 9. Early stopping in favour of H_0 was again not permitted.

Simulations III: futility stopping
The aim for these simulations is to explore trade-off between statistical power and test time as a function of the β_i values. To do so, the required ensemble sizes for obtaining a 0.99 detection rate are used, i.e. the same N as in Simulations I. The critical decision boundaries were then varied as a function of β_i, which were chosen through the futility functions from section 8.1. The 'no futility stopping' condition was also included. The number of stages K again took values ranging from 1 to 9. The analysis was performed both before and after simulated a response.

Subject ABR data
The subject data were analysed in K sequential stages with the Hotelling's T^2 test (applied to the initial 1-16 ms window), per dB SL condition, and per subject. The β_i values were again chosen using the futility functions from section 8.1, and the number of stages K took values from 1 to 9.

8.3.2 Results

Simulations I: TPR = 0.99
Results from Simulations I (TPR fixed at 0.99) are presented in Figure 8.3 (plots A

and B). Results firstly demonstrate an increased *maximum* test time for increasing K (plot A), i.e. as K is increased, statistical power is decreased, and the ensemble size N needs to be increased in order to maintain the 0.99 TPR. Note that although the ensemble size N was increased with K, the *mean* test time was still decreased (plot B), with reductions in test time of 40-45% when using $K = 6$ (relative to $K = 1$). The decreased mean test time is due to the test being stopped early (and H_0 rejected) for the higher SNR responses, i.e. the final stage of the analysis is typically not reached (and the maximum test time is not used).

Simulations II: N = 3000

Results from Simulations II (N fixed at 3000) are also presented in Figure 8.3 (plots C and D). Note again that for these simulations, the reduced statistical power for increasing K cannot be compensated for by increasing N. Coincidentally, a reduced TPR is observed for increasing K (plot D). The decrease in mean test time for increasing K (plot C) was now also more pronounced; reductions in test time of up to 50-60% are observed for $K = 4$ or $K = 5$, relative to $K = 1$.

Simulations III: futility stopping

Results from Simulations III from the 20 dB SL condition are presented in plots E, F, and G of Figure 8.3. Results first show that when a response is absent (plot G) and early stopping in favour or H_0 is *not* permitted, that the mean test time for the sequential test is increased with K. This is due to the increased *maximum* test time (note again that the maximum test time was increased with K to compensate for the reduced statistical power). For the no-stimulus condition, the trial was allowed to proceed to the final stage of the analysis in $(1 - \alpha)$x100% of the cases for the no-stimulus condition. The *mean* test time is therefore close to the *maximum* test time. When early stopping in favour of H_0 was permitted, on the other hand, then the increased mean test time for the no stimulus condition was greatly reduced (plot G). With respect to the stimulus condition (plots E and F), early stopping in favour of H_0 had no noticeable effect on the TPR or mean test time, under the condition that the β_i values were chosen conservatively, e.g. through the 'Cos 1' or 'Cos 3' futility functions. When the choice for the β_i values was more liberal (e.g. when chosen through the 'Exp 5' and 'Exp 15' futility functions) then a reduced TPR was observed. Finally, it is worth noting that results from the 30, 40, and 50 dB SL conditions demonstrated similar trade-offs between statistical power and test time, and are not presented in order to keep the results concise.

Subject ABR data

Results from the subject ABR data are presented in Fig. 8.4: plots A-F show the detection rates as a function of K for different choices of β_i, whereas plots G-L show the *mean* test time (taken across 12 subjects), similarly as a function of K and for different choices of β_i. The trade-off betweens statistical power and test time as a function of K is similar to that observed for Simulations II where N was also fixed at 3000 epochs. With respect to the β_i values, results again demonstrate reductions in mean test time for increasing β_i, potentially at the cost of a reduced TPR. Test time for the 0 and 10

dB SL conditions was also greatly reduced by increasing the β_i values.

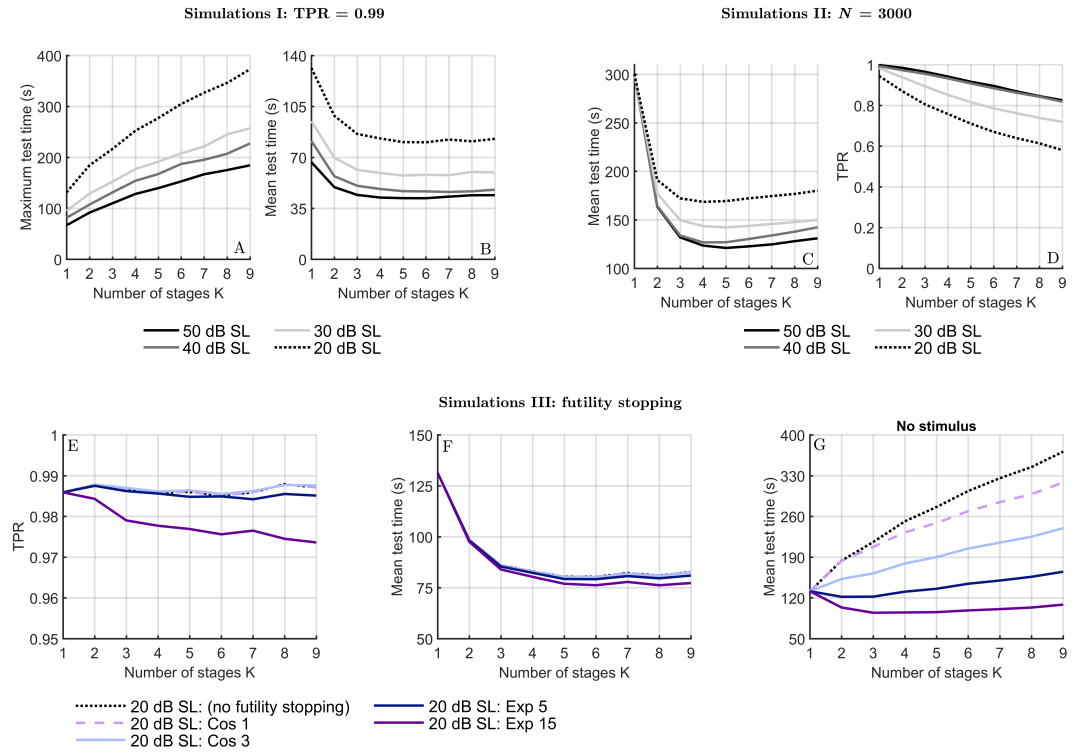


Figure 8.3: Results from Simulations I, II, and III for exploring the trade-off between statistical power and test time per dB SL condition, as a function of K and the β_i values. Details are provided in the text.

8.4 Summary

This chapter explored the specificity, sensitivity, and test time of a sequentially applied Hotelling's T^2 test with critical decision boundaries (for accepting or rejecting H_0) constructed by the CGST. With respect to specificity, results show that the main concern for the CGST for EEG measurements is the assumption that the p value null distributions are uniform on the [0,1] interval under H_0, which is only satisfied when the assumptions underlying the statistical detection method are also satisfied. This emphasizes the importance of using a suitable ABR detection method, i.e. one with a good control of specificity. As shown in chapter 5 (and confirmed in this chapter), the main concern for the specificity of ABR detection methods is the independence assumption between epochs, which is violated as a function of the high-pass cut-off frequency and the stimulus rate. Results from this chapter demonstrate that the FPR of a sequentially applied Hotelling's T^2 test was (more or less) controlled as intended when using a high-pass cut frequency of 100 Hz and a stimulus rate of 33.3 Hz (Table 8.1).

With respect to the trade-offs between statistical power and test time, results firstly demonstrate relatively large reductions for the stimulus condition by increasing K, up

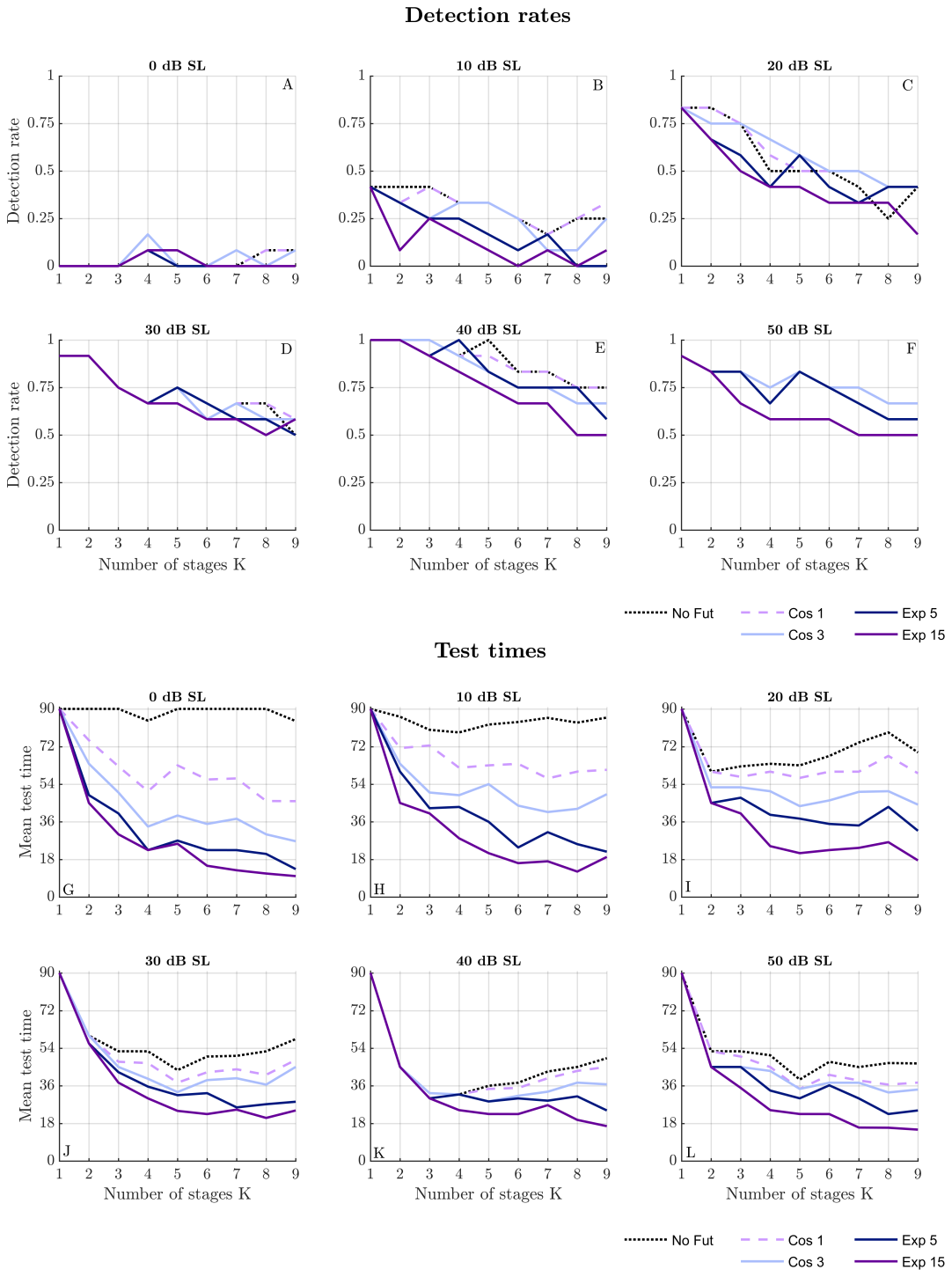


Figure 8.4: Results from the subject ABR data for exploring the trade-off between statistical power and test time per dB SL condition, as a function of K and the β_i values. Details are provided in the text.

to 40-45% for $K = 6$ relative to $K = 1$, with no loss in test sensitivity (Figure 8.3, plot B). In order to achieve these results, the *maximum* test time needs to be increased, else a reduced test sensitivity can be expected. If N cannot be increased with K (due to e.g. an upper limit of, say, 3000 epochs), then a reduced TPR can be expected for

increasing K. The latter was confirmed with both simulations (Figure 8.3) and subject data (Figure 8.4), and was found to be most prominent when the single shot test ($K = 1$) was already under-powered (e.g. for the 10 and 20 dB SL conditions of the subject ABR data). Based on the preceding results, the following rough guidelines might be used for choosing K: if the single shot test ($K = 1$) is expected to be under-powered, and N cannot be increased, then it may be beneficial to keep the number of sequential stages for the analysis low, e.g. 1, 2, or 3 stages might be used. If the single shot test is expected to be over-powered, or if N can be increased with K, then a more efficient approach is to use 4, 5, or 6 stages for the analysis.

Finally, the increased *maximum* test time for increasing K has consequences for the no-stimulus condition, i.e. because the test is allowed to proceed to the final stage of the analysis in (1-α)x100% of the trials, the *mean* test time will be close to the *maximum* test time. This emphasizes the importance of futility stopping for sequential testing. Results indeed demonstrate large reductions in test time for the no-stimulus condition by increasing the β_i values, which can potentially come at the cost of a reduced test sensitivity. The β_i values should therefore be chosen conservatively, if a reduced test sensitivity is to be prevented.

Chapter 9

An adaptive sample-size re-estimation procedure for ABR detection

In all previous chapters, the ensemble size N for the statistical analysis was fixed at the outset. As already mentioned in previous chapters, fixing N in advance may be inefficient for ABR detection as the SNR of the response can vary across recordings, i.e. a fixed N will tend to result in either an over-powered test (and an unnecessarily prolonged test time) for the the higher SNR responses, or an under-powered test (and an increased false-negative rate) for the lower SNR responses. As shown in chapter 8, a partial solution to variability in the SNR is to analyse data sequentially, as this allows the test to be stopped early in the case of a clear response. However, the total ensemble size N (for the full trial) still needs to be fixed in advance. This raises the question as to how N should be chosen, and how detrimental an incorrect choice for N might be towards test performance.

The focus for this chapter is on two approaches for choosing N: (1) a non-adaptive approach where N is chosen at the outset, and (2) an adaptive approach, where N is chosen adaptively, using previously collected data. Note that for the non-adaptive approach, the user should ideally assume an SNR in advance, and then choose N such that some desired TPR is obtained for the assumed SNR. For the adaptive approach, there are various methods available in the literature that can be used for choosing N adaptively (see e.g. Chow & Chang, 2007, Chapter 7; Proschan & Hunsberger, 1995; Lehmacher & Wassmer, 1999; Chow & Chang, 2007; Mehta & Pocock, 2010). These methods essentially use previously analysed data to predict the required N for obtaining the desired statistical power for the trial. The main difference between these methods and the adaptive approach described in this chapter, is that the adaptive approach from this chapter applies the statistical power analysis first. In particular, statistical power is re-estimated continuously (as new data becomes available) for an *a priori* assumed

response. The statistical analysis for response detection is then only applied once some desired statistical power has been obtained. The reader might already be worried that this approach would bias the analysis towards favourable results, thus inflating the FPR. This is indeed a concern, and precautions are necessary to avoid introducing such a bias (further described in the subsequent sections). Note also that the proposed adaptive approach requires both the amplitude and waveform morphology of the response to be assumed *a priori*. This also has some pros and cons, which will be discussed in more detail later on.

The structure of this chapter is as follows: section 9.1 below first gives a brief background on statistical power and the alternative distribution. The general approach for the sample size re-estimation procedure is then described in section 9.2. The formulas and equations for the statistical power and the alternative distribution are then presented specifically for the Hotelling's T^2 test in section 9.2.1. The approach is evaluated and compared to a non-adaptive approach in section 9.3. The results and some directions for future work are further discussed in section 9.4, and the chapter ends with some concluding remarks in section 9.5.

9.1 Background on statistical power and the alternative distribution

Before describing the adaptive approach, a very brief background will be given on statistical power and the alternative distribution, both of which play important roles throughout this chapter. To quote the very first sentence in Chapter one of Cohen's well known book on statistical power analysis (Cohen, 1988, p.1):

The power of a statistical test is the probability that it will yield statistically significant results.

The probability that a test will yield a statistically significant result is given by the area under the true distribution of the test statistic to the right of the critical decision boundary for rejecting H_0. The latter is easily clarified with an example: Figure. 9.1 shows the null distribution for some hypothetical test statistic, along with the critical decision boundary for rejecting H_0 at a 95% confidence ($\alpha = 0.05$), equal to 1.535. If the null hypothesis H_0 is actually false, then the true distribution of the test statistic is not given by the null distribution, but by the 'alternative distribution' (the distribution of the test statistic when anything but H_0 is true). Statistical power is then equal to the area under the alternative distribution, to the right of the critical decision boundary.

In practice, the true distribution of the test statistic is almost always unknown (if this were known, there would be no need for the statistical analysis). For the adaptive approach in this chapter, it is assumed that H_0 is false, and that a response is present. The true distribution for the test statistic for the assumed response is then estimated

using previously collected data. Finally, the estimated alternative distribution can be used to estimate statistical power. This approach is now discussed in more detail in section 9.2 below.

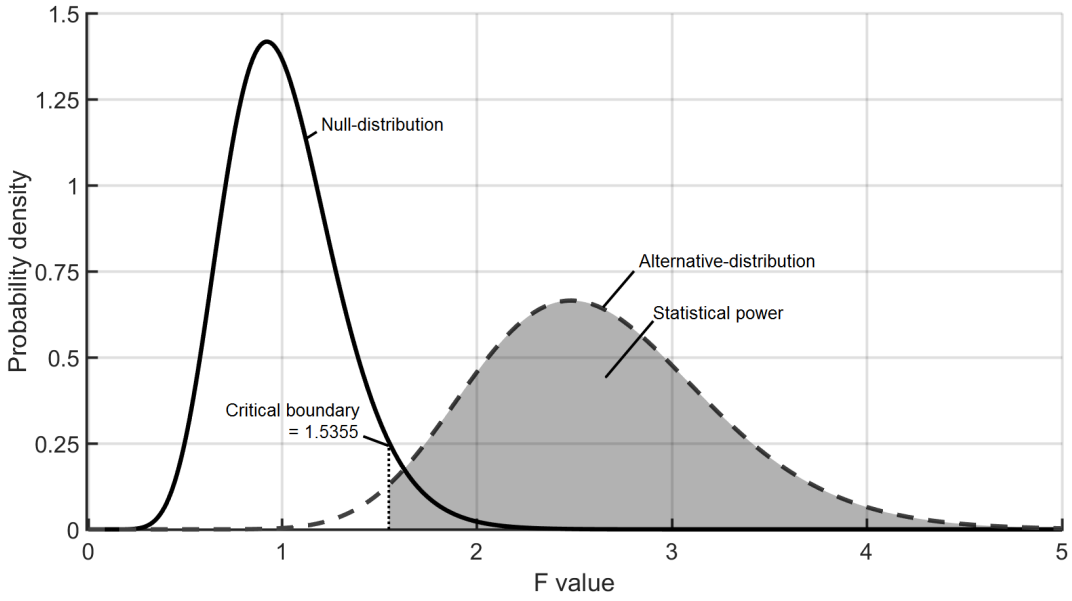


Figure 9.1: An example for illustrating statistical power for some hypothetical test statistic. The null distribution for the test statistic is shown on the left, along with the critical decision boundary for rejecting H_0 at 95% confidence ($\alpha = 0.05$). The distribution to the right is the distribution of the test statistic when H_0 is false, referred to as the alternative distribution. Statistical power is given by the area under the true distribution of the test statistic, to the right of the critical decision boundary.

9.2 Online sample size re-estimation

This section describes a new approach for adaptively choosing N based on a post-hoc power analysis, applied to previously collected data. The core of the approach revolves around estimating the alternative distribution for the test statistic under some *a priori* assumed response. In particular, the alternative distribution is estimated using both the assumed response and the power of the EEG background activity (estimated from previously collected data). The estimated alternative distribution is then used to estimate statistical power, and the statistical analysis is only applied once the desired statistical power has been obtained. The main advantage for this approach over a non-adaptive approach is that the power of the (potentially non-stationary) EEG background activity is taken into account when choosing N. This hence allows a more informed decision with regards to N, potentially resulting in an improved control over the TPR and/or a reduced test time. A second advantage is that a more informed decision can be made with regards to the B_i critical decision boundaries (the critical boundaries for stopping the test early for futility), i.e. these can now be chosen as a function of the estimated alternative distribution, which may give an improved control over the false-

negative rate (FNR). A caveat is that in order to avoid introducing a bias, the EEG background activity should be estimated exclusively from the inter-epoch intervals, i.e the intervals between the windows being analysed by the statistical test (see also Figure 9.2).

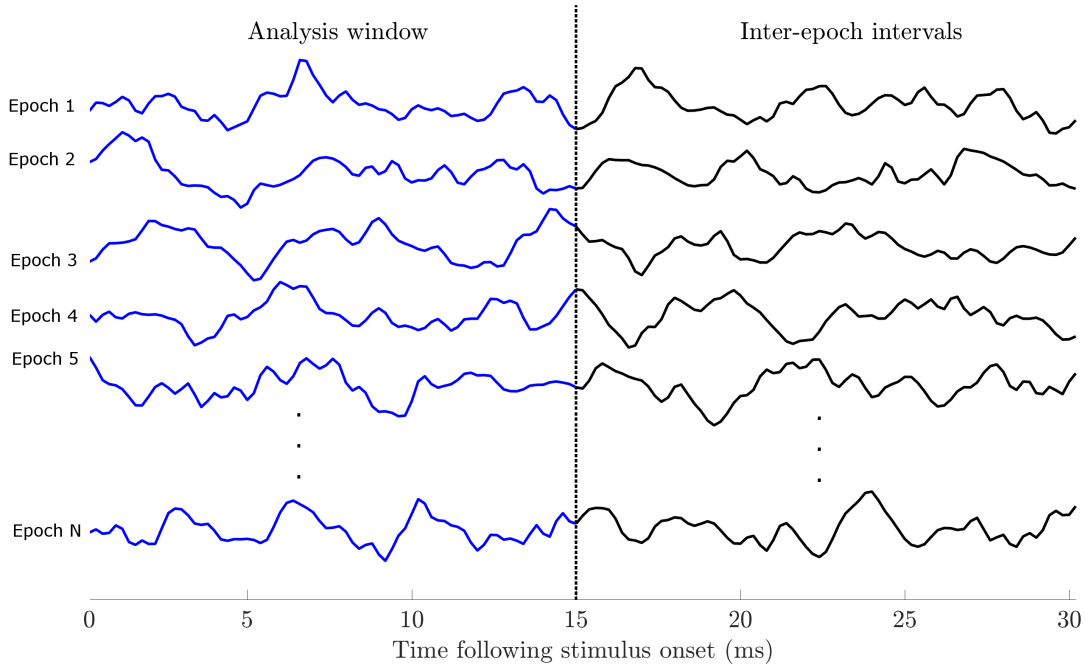


Figure 9.2: The analysis windows (the 0-15 ms window following stimulus onset) and the inter-epoch intervals (the 15-30.03 ms windows following stimulus onset) for N epochs. When using the online sample size re-estimation approach, data within the analysis windows should be kept hidden from the user.

To clarify the approach, the user will be guided through a two-stage sequential test design. The user should first specify the following parameters: the FPR for stage 1 (previously denoted as α_1); the desired TPR for stage 1 (the statistical power for stage 1), say γ_1; and the permitted FNR for stage 1 due to early acceptance of H_0, say Γ_1. The value for β_1 (the stage 1 TNR) is no longer specified, but will instead follow from the previously chosen parameters. For this example, say $\alpha_1 = 0.01$, $\gamma_1 = 0.8$, $\Gamma_1 = 0.02$, and that the alternative distribution is re-estimated after every 50 additional epochs have been collected (using epochs 1-50, 1-100, 1-150, etc.). The resulting null and alternative distributions for this hypothetical example are shown in the upper plots in Figure 9.3 (the null distributions are in black, whereas the alternative distributions are in gray). The critical decision boundary A_1 (for rejecting H_0) is also shown, which was found using the approach described in chapter 7, i.e. the area under the null distribution to the right of A_1 should equal α_1. The estimated statistical power $\hat{\gamma_1}$ is then given by the area under the alternative distribution, to the right of A_1. For this example, the desired statistical power of 0.8 was exceeded once 200 epochs had been collected. Data collection for stage 1 was therefore stopped at $N_1 = 200$. The estimated alternative distribution (using $N_1 = 200$) is then used to find B_1, i.e. B_1 needs to be found such

that the area under the estimated alternative distribution to the left of B_1 is equal to the permitted stage 1 FNR, denoted by Γ_1 (equal to 0.02 for this example).

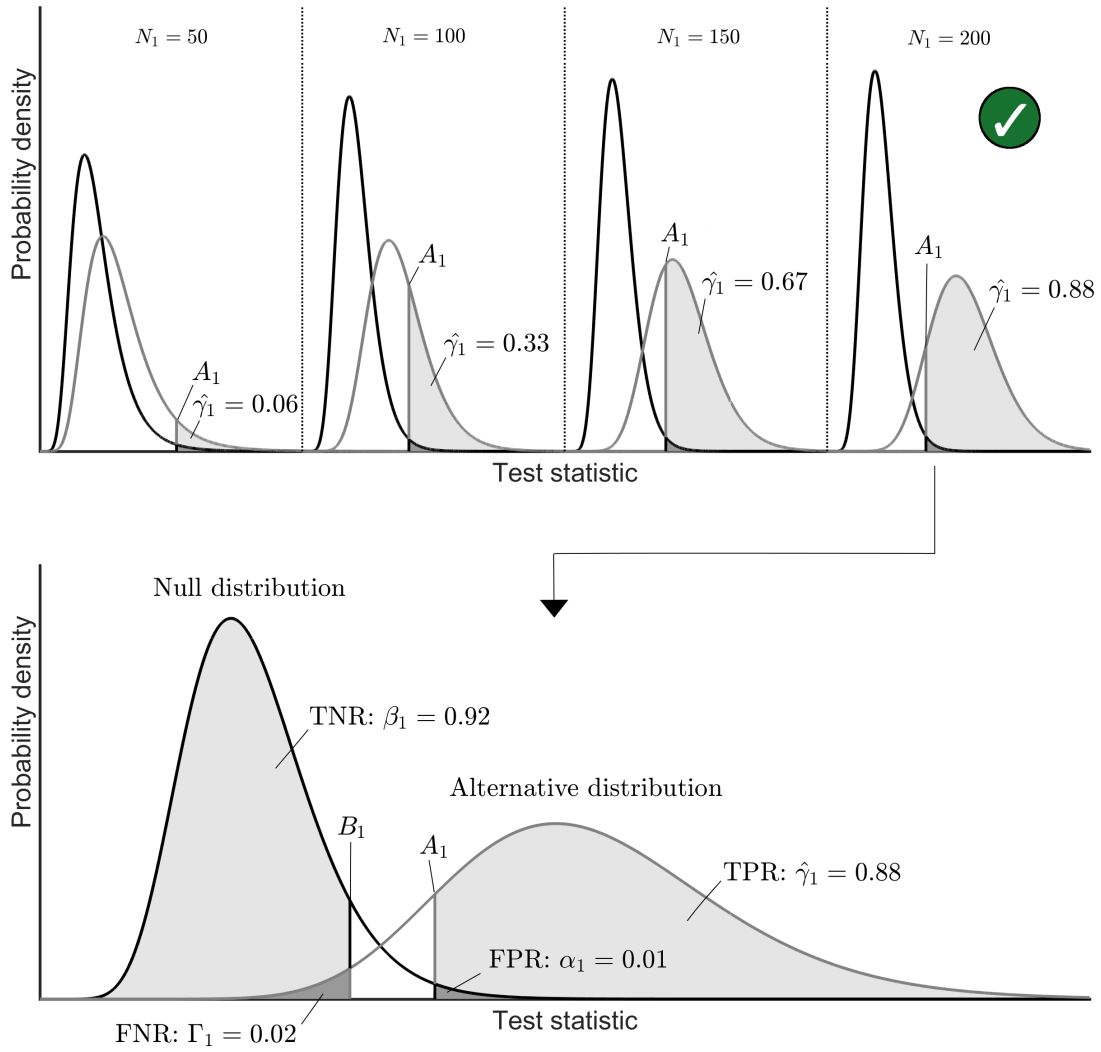


Figure 9.3: An illustration of an online sample size re-estimation procedure for stage 1. The black plots show the null distributions, whereas the gray plots show the alternative distributions. Increasing N_1 shifts the alternative distribution away from the null distribution, thus increasing the estimated statistical power $\hat{\gamma}_1$. For this example, the desired statistical power of 0.8 was exceeded at $N_1 = 200$. Data collection was hence stopped at $N_1 = 200$, after which the estimated null and alternative distributions (using $N_1 = 200$) were used to construct critical decision boundaries A_1 and B_1. Further details are presented in the text.

Assuming the test statistic fell within the $[B_1, A_1]$ interval, the trial proceeds to stage two (the final stage of the analysis for this example). The user should then specify the total desired statistical power (for the full trial), along with the stage two FPR and FNR. Say the total desired statistical power is 0.95, $\alpha_2 = 0.01$, and $\Gamma_2 = 0.02$. Data collection for stage two is then initiated, which follows the same procedure as in stage one, i.e. the alternative distribution is continuously re-estimated (every 50 epochs) until the total estimated statistical power (given by $\hat{\gamma}_1 + \hat{\gamma}_2$) has exceeded the desired statistical power of 0.95 (further illustrated below). Note that in stage two, the procedure is applied to

the null and alternative distributions for the summary statistic (composed of stage 1 and stage 2 transformed p values). These distributions are generated as described in Chapter 7, i.e. by convolving truncated PDFs. Note that the exact same procedure can be applied for the alternative distribution as for the null distribution.

The stage two test procedure is illustrated in Figure 9.4: the upper left plot shows the stage one distributions (for $N_1 = 200$), which have been truncated to the $[B_1,\ A_1]$ interval, whereas the upper right plots show the null and alternative distributions for the stage two test statistic for $N_2 = 50$ and $N_2 = 100$. Convolving the truncated stage 1 distributions with the stage two distributions gives the null and alternative distribution for the stage two summary statistic (middle plots). The distributions for the summary statistic are then used to estimate statistical power: when $N_2 = 50$, the estimated statistical power for stage 2 is $\hat{\gamma}_2 = 0.0452$, giving a total estimated statistical power of $\hat{\gamma}_1 + \hat{\gamma}_2 = 0.9252$; still less than the desired 0.95. Increasing the ensemble size to $N_2 = 100$ gives $\hat{\gamma}_2 = 0.0798$. The total estimated statistical power is now $\hat{\gamma}_1 + \hat{\gamma}_2 = 0.9598$, and has exceeded the desired 0.95. Data collection for stage two can hence be stopped after $N_2 = 100$ epoch have been collected. The null and alternative distributions (for $N_2 = 100$) are then used to find A_2 and B_2, which follows the same procedure as for stage one.

9.2.1 Online sample size re-estimation using the Hotelling's T^2 test

This section describes how statistical power and the alternative distribution are estimated when using the Hotelling's T^2 test as detection method. The alternative distribution, say H_1, for the F-transformed T^2 statistic (Eq. 3.4) is given by a non-central F-distribution with Q and $N - Q$ DOF (Bilodeau & Brenner, 1999, p.100):

$$H_1(F) = F_{nc}\left(F, Q, N - Q, \delta\right) \tag{9.1}$$

where F is the observed F value (the x-axis of the distribution) and δ is a non-centrality parameter. The non-centrality parameter is directly related to the effect size, and is given by (Bilodeau & Brenner, 1999, p.100):

$$\delta = N(\boldsymbol{\mu} - \boldsymbol{\mu_0})\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu_0})^H \tag{9.2}$$

where $\boldsymbol{\mu}$ is a Q-dimensional vector containing the *true* feature means, $\boldsymbol{\mu_0}$ is a Q-dimensional vector containing the hypothesized values to test against, and Σ is the *true* covariance matrix of the features. The alternative distribution H_1 can then be used to determine statistical power, given by the area under the alternative distribution, to the

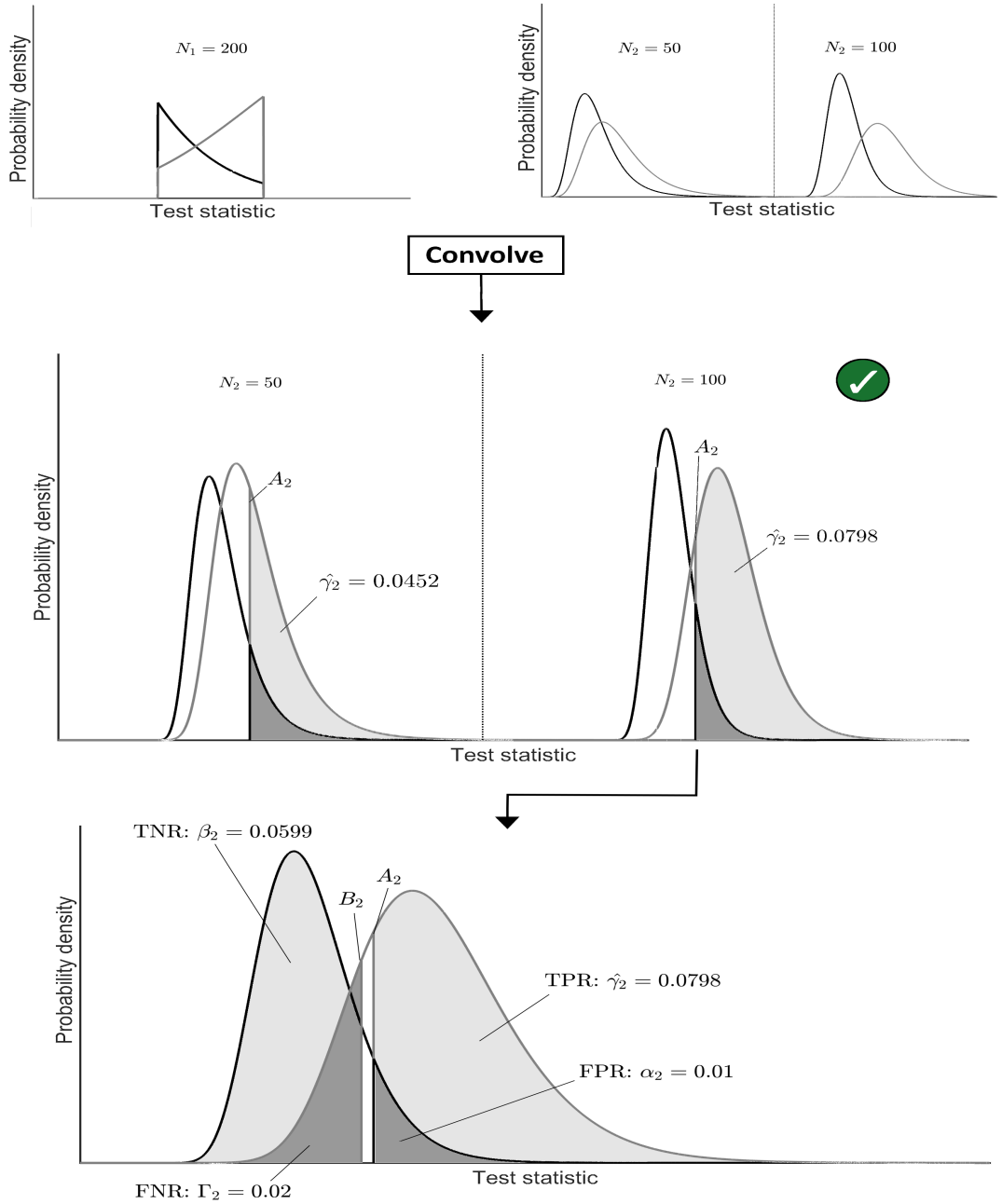


Figure 9.4: An illustration of an online sample size re-estimation procedure for stage two. The black plots show the null distributions, whereas the gray plots show the alternative distribution. **Upper plots:** the upper left plot shows the stage one distributions, which have been truncated to the $[B_1,\ A_1]$ interval. The upper right plots show the null and alternative distributions for the stage two test statistic. **Middle plots**: The null and alternative distribution for the stage two summary statistic, found by convolving the truncated distributions from the stage one with the distributions for the stage two test statistic. **Lower plot**: the final null and alternative distribution for the stage two summary statistic, which are used to construct stage two critical boundaries A_2 and B_2. Further details are presented in the text.

right of the critical decision boundary for rejecting H_0. When using F-distributions,

statistical power γ can be expressed as:

$$\gamma = 1 - F_{nc}\left(F^{-1}(1-\alpha, v_1, v_2), v_1, v_2, \delta\right) \tag{9.3}$$

where $F^{-1}(.)$ is the inverse of the central F-distribution (the inverse of the null distribution) and F_{nc} is the cumulative distribution function (CDF) of the non-central F-distribution. Note that $F^{-1}(1-\alpha, v_1, v_2)$ is the critical boundary for rejecting H_0 and nominal level α. The $F_{nc}\left(F^{-1}(1-\alpha, v_1, v_2), v_1, v_2, \delta\right)$ term is hence the area under the alternative distribution to the *left* of the critical boundary. Subtracting the result from one therefore gives the area to the *right* of the critical boundary, i.e. statistical power.

It should be stressed that in practice, both $\boldsymbol{\mu}$ and Σ are almost always unknown, and are instead either assumed *a priori*, or estimated from data using $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$. When estimated from data, the (estimated) non-centrality parameter is given by:

$$\hat{\delta} = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})\boldsymbol{S^{-1}} - (\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})^H \tag{9.4}$$

which is identical to the T^2 statistic in Eq. 3.3.

The caveat with estimating δ is that it can potentially be contaminated by significant amounts of noise. Put differently, $\hat{\delta}$ has its own underlying PDF. As a result, statistical power γ will also be an estimate (denoted by $\hat{\gamma}$), i.e. it will tend to either over-estimate or under-estimate the true statistical power γ, which introduces uncertainty to the approach. As mentioned earlier, the proposed approach assumes a response at the outset. The mean feature vector $\boldsymbol{\mu}$ is therefore assumed to be known, and is given by the feature values extracted from the assumed response (see also section 9.2.2). Uncertainty within $\hat{\delta}$ is therefore solely due to measurement error in $\boldsymbol{S}$. Ideally, uncertainty within $\boldsymbol{S}$ should be taken into account when estimating statistical power. The latter is achieved using a resampling approach, further described below.

Before turning to the resampling approach, it should again be stressed here that in order to avoid introducing a bias, $\boldsymbol{\Sigma}$ should be estimated from just the inter-epoch intervals (the 15-30 ms windows following stimulus onset for this chapter), in which case it will henceforth be referred to as $\boldsymbol{S2}$. In addition, $\boldsymbol{S2}$ should be either independent from the 0-15 ms windows following stimulus onset, or the independence violations should be negligible, else a bias might still be introduced to the analysis. The spectral content of EEG measurements brings this independence assumption into question, and is further considered in section 9.3.

Returning to the resampling approach: the goal is to rescale $\boldsymbol{S_2}$ as a function of uncertainty. In particular, when uncertainty is high, $\boldsymbol{S_2}$ is scaled upwards, giving a liberal

estimate of covariance matrix Σ. The choice to adopt a liberal approach here is to prevent a reduced test sensitivity. In particular, when $\boldsymbol{S_2}$ is under-estimated, $\hat{\gamma}$ will be over-estimated, and data collection will be stopped prematurely (i.e. before the desired statistical power has not been obtained), giving a reduced test sensitivity. The resampling approach itself consists of resampling many covariances matrices from the underlying distribution of $\boldsymbol{S_2}$. The underlying distribution of $\boldsymbol{S_2}$ is given by a scaled Wishart distribution with true covariance structure Σ and $N-1$ degrees of freedom (Rencher, 2001, p.92), denoted by $W(\Sigma, N-1)$. The true covariance structure Σ is of course again unknown, and is instead substituted with $\boldsymbol{S_2}$. The resampling approach then consists of sampling 50 covariance matrices from $W(\boldsymbol{S_2}, N-1)$, ranking them from small to large, as determined by either their determinants or their traces, and selecting the largest resampled covariance matrix. The determinant of a covariance matrix (also known as generalised variance; Wilks, 1932) represent a single value for multi-variate scatter, where larger values correspond to more disperse data. The largest resampled covariance matrix, say $\boldsymbol{S_{Max}}$, is then used to find a re-scaling factor, say c_3, such that:

$$| \boldsymbol{S_{Max}} | \;=\; | \boldsymbol{c_3 S_2} | \tag{9.5}$$

where $| \boldsymbol{S} |$ denotes the generalised variance of $\boldsymbol{S}$. For small N, uncertainty will be relatively large, and the $W(\boldsymbol{S_2}, N-1)$ distribution will be relatively disperse, in which case $| \boldsymbol{S_{Max}} |$ will tend to be larger than $| \boldsymbol{S_2} |$. This will result in $\boldsymbol{S_2}$ being scaled upwards, which reduces $\hat{\delta}$, giving a conservative estimate of statistical power. Consequently, data collection will be prolonged, and an over-powered test can be expected. This is given preference over the alternative, i.e. an under-powered test and a reduced test sensitivity.

Putting it all together, the estimated statistical power in the presence of an assumed response, using the Hotelling's T^2 test (time domain) as detection method, is given by:

$$\hat{\gamma} = 1 - F_{nc}\left(F^{-1}(1-\alpha, Q, N_i - Q), Q, N - Q, \hat{\delta}\right) \tag{9.6}$$

where $\hat{\delta}$ is given by:

$$\hat{\delta} = N(\boldsymbol{x_a} - \boldsymbol{\mu_0})(c_3 \boldsymbol{S_2^{-1}}) - (\boldsymbol{x_a} - \boldsymbol{\mu_0})^H \tag{9.7}$$

and where $\boldsymbol{x_a}$ contains Q TVMs, extracted from the assumed response.

9.2.2 The assumed response

The assumed response is important, as it affects the estimated statistical power, e.g. when the assumed response is smaller than the true response, then the estimated statistical power will tend to be lower than the true statistical power, and data collection will be longer than necessary (giving an over-powered test). For this chapter, the assumed response is a 'minimum response', given by the smallest ABR template from the 40 dB SL condition from data set **D4**. In particular, the ABR templates (the coherent averages from data set **D1**) were all ranked from small to large, as determined by their mean square values, and the ABR template with the smallest mean square value was used as the assumed response. The choice to use a minimum response, as opposed to a mean or maximum response, is to prevent a reduced test sensitivity. To clarify, if the assumed response is *larger* than the true response, then γ will be over-estimated, and data collection will be stopped pre-maturely (before the desired statistical power has been obtained), giving a reduced test sensitivity.

9.2.3 The summary statistic

The summary statistic at each stage of the sequential analysis is given by a sum of inverse F-distributed random variables:

$$\Sigma_k = \sum_{i=1}^{k} F^{-1}(1 - p_i, v_1, v_2) \tag{9.8}$$

where F^{-1} is the inverse of an F-distribution. The choice to use a sum of inverse F-distributed random variables (as opposed to χ^2-distributed random variables in chapter 8) is purely for convenience, i.e. when using the Hotelling's T^2 test, statistical power (and hence the alternative distribution) is expressed directly through F-distributions. It might be noted here that the adaptive approach can still be used for a sum of χ^2-distributed random variables, but the alternative distribution in Eq. 9.1. would then have to be transformed to a non-central χ^2-distribution, achieved by warping the x-axis. It is also worth noting here that the sum of inverse F-distributed random variables appears to give a more or less identical test sensitivity compared to a sum of inverse χ^2-distributed random variables (Appendix, section A.12, Figure A.19).

9.3 Simulations

This section describes simulations for evaluating the specificity, sensitivity, and test time of (1) a non-adaptive approach, where N is optimised in advance, such that some desired

TPR is obtained for an *a priori* assumed SNR, and (2) the previously described adaptive approach.

9.3.1 Method

Data

Data for the simulations consists of ABR templates (from data set **D4**) for simulating a response, along with scaled simulated coloured noise (generated as described in section 4.4) for representing the EEG background activity.

A non-adaptive ensemble size

For the first approach, the ensemble size N is optimised in advance, such that some desired TPR is obtained for an *a priori* assumed SNR. For the current simulations, the assumed SNR is given by the smallest SNR from the 40 dB SL condition from data set **D4**, equal to -32.6, whereas the desired TPR was set to 0.95. The ensemble size N was then optimised, such that a 0.95 TPR was obtained for an SNR of -32.6 dB. The latter was achieved using the same procedure described for 'Simulations I' in chapter 8, i.e. by repeatedly generating 5000 ensembles with increasing or decreasing ensemble sizes, until a TPR of 0.95±0.01 was obtained, which was repeated for difference choices for K (ranging from 1 to 9). Once N was optimised, a second set of simulations were used to evaluate sensitivity and test time (using the optimised N), now across a range of SNRs (shown in Table 4.1, for the 40 dB SL condition). Data were then analysed in K sequential stages using the Hotelling's T^2 test, both before and after simulating a response. Finally, the nominal α-level was set to 0.01, which was spread equally across K stages, giving α_i values of $\frac{\alpha}{K}$ for all i and K, and the β_i values were chosen through the 'Exp 5' futility function (see also section 8.1.1).

An adaptive ensemble size

For the adaptive approach, the desired TPR for the *a priori* assumed minimum response was set to 0.95. The stage-wise TPRs γ_i now need to be chosen, such that $\sum_{i=1}^{K} \gamma_i = 0.95$. As shown in chapter 8, a sensitive and robust performance is obtained by splitting the available N equally across the K stages. The γ_i values are therefore given by the stage-wise TPRs when N is optimised (and split equally across the K stages), such that a 0.95 TPR is obtained. The resulting γ_i values are shown in section A.13 of the Appendix (Table A.12). The nominal α-level was furthermore set to 0.01, which was again spread equally across the K stages, giving α_i values of $\frac{\alpha}{K}$ for all i and K. The total permitted FNR due to early stopping in favour of H_0 was set to 0.04, which was also spread equally across the K stages, giving Γ_i values of $\frac{0.04}{K}$ for all i and K. Finally, data were again analysed with the Hotelling's T^2 test in K sequential stages, both before and after simulating a response.

9.3.2 Results

Results from the simulations are presented in Figure 9.5. Starting with the stimulus condition (plots A and B), results show that when N was optimised (in advance) for an *a priori* assumed minimum SNR, that the test was over-powered, as expected (plot A). For the adaptive approach, results also demonstrate an over-powered test, which can be attributed to (i) the assumption that the response is the 'minimum response' and (ii) the resampling approach, which gives a liberal estimate for the features covariance matrix. Comparing the two approaches shows a similar test performance in terms of TPRs (the largest difference was at $K = 9$, with a TPR of 0.9953 for the non-adaptive approach and a TPR of 0.9878 for the adaptive approach), which suggests that the comparison in test time was relatively fair. The mean test times are shown in plot B, and show a reduced test time for the adaptive approach of ∼10-30% (depending on K) relative to the non-adaptive approach. Results from the no-stimulus condition are also shown in Figure 9.5 (plots C and D). Results first show that the FPRs (for both the adaptive and the non-adaptive approach) fall within the two-sided 99% confidence intervals for the expected 0.01 FPR (plot C), with the exception of a single condition which can likely be attributed to random variation. With respect to test time (plot D), results show large reductions in mean test time for the adaptive approach (relative to the non-adaptive approach) of ∼25-45%, depending on K. The latter can be attributed to a more suitable choice for the B_i futility boundaries.

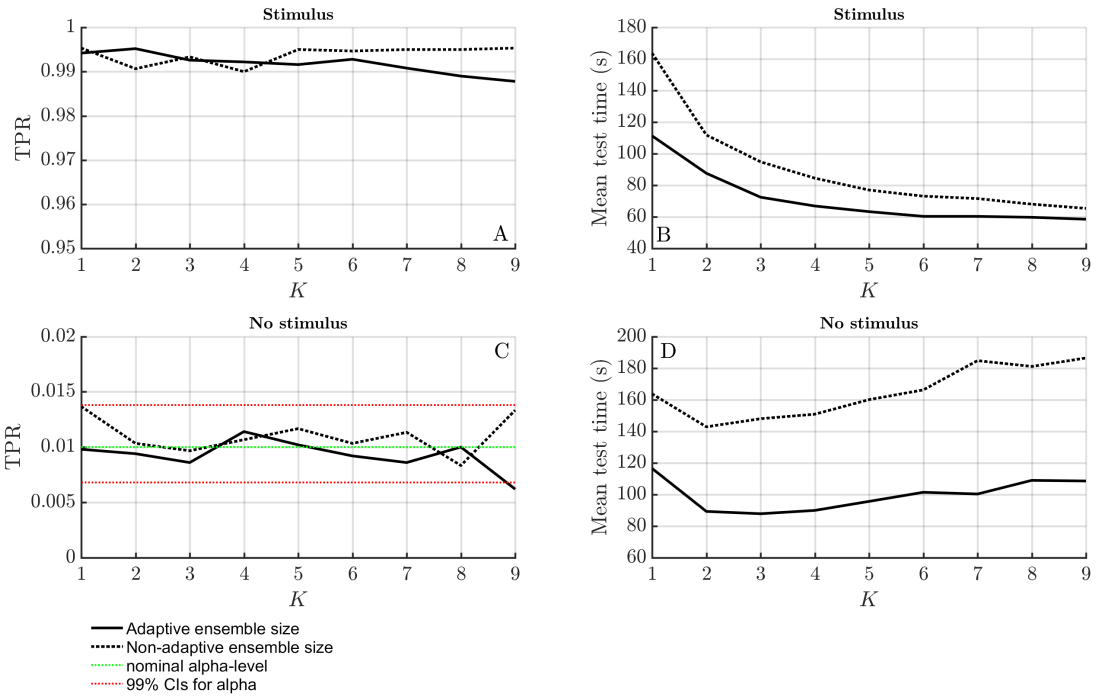


Figure 9.5: Results from the simulations: plots A and B, the TPRs and mean test times (respectively) when a response was present, as a function of K. Plots C and D: the FPRs and mean test times (respectively) when a response was absent, as a function of K.

9.4 Discussion

This chapter described and briefly explored the performance of a new approach for adaptively choosing the ensemble-size N for ABR detection, and compared the results with those from a non-adaptive approach where N was optimised in advance. Results show an advantage for the adaptive approach over the non-adaptive approach, with reductions in mean test time of ∼10-30% for the stimulus condition and ∼25-45% for the no-stimulus condition, with a more or less equal performance in terms of test sensitivity. The advantage for the adaptive approach can be attributed to a reduced uncertainty with respect to the power of the EEG background activity, i.e. the EEG background activity is now estimated for the specific recording in question, which allows for a more informed decision with respect to N. An additional advantage for the adaptive approach is an improved control over the TPR and the FNR, which can now be specified explicitly by the user through the γ_i and Γ_i values. Note that this can help bring ABR examinations to an unambiguous test outcome, to within some confidence limits. In particular, when the test is stopped for futility ($\Sigma_k < B_k$), it can be concluded (with certainty $> 1-\sum_{i=1}^{K}\Gamma_i$) that the response is either absent, or that it is smaller than the minimum response, whereas when the test is stopped for efficacy ($\Sigma_k > B_k$) it can be concluded (with certainty $> 1-\alpha$) that a response is present.

Although results from this chapter look promising, it should be noted that the approach has not yet been tested for real data. Moreover, the extent to which these results are dependent on design parameters and test conditions has not yet been explored. In particular, the following areas require more testing and/or justification: (i) the choice for the assumed response, (ii) the resampling approach, and (iii) specificity in general, i.e. the extent to which the underlying assumptions remain satisfied. Starting with the assumed response, this chapter adopted a conservative approach by assuming a 'minimum response' (the smallest response for normal hearing adults). Besides preventing a reduced test sensitivity, adopting a 'minimum assumed response' also allows more interesting conclusions to be drawn from the analysis, i.e. as mentioned earlier, when the test is stopped for futility, it can now be concluded that the response is either absent or that it is smaller than the minimum response. Note that this is of more interest than 'absent or smaller than the mean normal hearing response', which can be concluded when using a 'mean ABR' as the assumed response. That said, it is not clear what the amplitude and waveform morphology for the 'assumed minimum response' should be. In order to determine this, data may be required from a much larger cohort of normal hearing adults.

With respect to the resampling approach, a potential shortcoming is that the covariance structure Σ of the Wishart distribution $\mathrm{W}(\Sigma, N-1)$ is still unknown, and was instead replaced with $\boldsymbol{S2}$. Consequently, results will again vary as a function of the measurement error within $\boldsymbol{S2}$, which is undesirable. The choice to resample 50 covariance matrices, and not e.g. 100, was chosen primarily to reduce computation time, and might also be

sub-optimal. Future work might further explore the resampling approach, or look into alternative methods for estimating the alternative distribution and/or non-centrality parameters (Meyer, 1967; Spruill, 1986; Li et al, 2009; Neff & Strawderman, 1976; Saxena & Alam, 1982; Perlman & Rasmussen, 1975; Berger et al, 1998; Chow, 1987; Kubokawa et al, 1993; Shao & Strawderman, 1995; Leung & Muirhead, 1987; Kubokawa et al, 2017). A more detailed discussion on estimating the non-centrality parameter and/or the alternative distribution is presented in section A.14 of the Appendix.

With respect to specificity, a potential concern is that the stage-wise critical decision boundaries A_i and B_i were chosen adaptively, whereas the CGST described in chapter 7 states that these should be fixed in advance, which is necessary to avoid introducing a bias. The latter was circumvented in this chapter by assuming the response, and by just using the inter-epoch intervals for estimating the EEG background activity, i.e. the data for the statistical power analysis should be independent of the data being analysed by the statistical test. This implies that the estimated covariance matrix $\boldsymbol{S2}$ should be independent of the data in the 0-15 ms windows. Although this assumption is questionable due to the spectral content of the data (see Figure 5.2), results from this chapter suggests that specificity is still controlled as intended. This is further supported by results from the Appendix (section A.11, and section A.11.3 in particular), which show that the assumptions underlying the CGST remain satisfied when adapting test parameters as a function of sample variance estimated from the inter-epoch intervals. A potential explanation is that sample variance (or the sample covariance matrix) is robust to independence violations as it takes the mean of the data into consideration.

Finally, there are various alternative methods available in the literature for choosing N adaptively (e.g. Chow & Chang, 2007, Chapter 7; Proschan & Hunsberger, 1995; Lehmacher & Wassmer, 1999; Chow & Chang, 2007; Mehta & Pocock, 2010). These methods were not explored here, but might be compared to the adaptive approach from this chapter in future work. There are two important differences between these methods and the current adaptive approach: (i) the assumed response, and (ii) the way in which the statistical power analysis is applied to the data. With respect to the assumed response, this is not required for the methods from the literature, i.e. both the response and the background activity are estimated from previously analysed data. This has both an advantage (there is no need to complicate the approach by assuming a response), and a disadvantage. The disadvantage is that uncertainty within the estimated non-centrality parameter (directly related to the effect size) is increased, which is detrimental towards test performance (see e.g. Levin & Subkoviak, 1977; Bruton et al., 2000; Kanyongo et al., 2007). With respect to (ii), the methods in the literature use a predictive power analysis, whereas the current adaptive approach uses an online, post-hoc power analysis. A potential complication with predictive power analyses is that the EEG background activity is assumed to be a stationary process. When stationariy is violated, the true statistical power will be either smaller or higher than the predicted power, which again introduces additional uncertainty, which is detrimental towards test

performance. For the current adaptive approach, stationarity of the EEG background activity is assumed for just the 0-30 ms windows, i.e. it is assumed that the power of the EEG background activity within the 0-15 window is the same as the power within the 15-30 ms windows.

9.5 Conclusion

A new adaptive sample size re-estimation procedure was proposed and briefly evaluated for ABR detection. Results show that when compared to a non-adaptive approach, reductions in mean test time of ∼10-30% are observed for the stimulus condition, and ∼25-45% for the no-stimulus condition, with a more or less equal test sensitivity. Besides a reduced test time, the adaptive approach gives an improved control over the TPR and the FNR, which can be used to help bring ABR examinations to an unambiguous test outcome in terms of 'ABR present' or 'ABR absent or abnormal'. Future work should further test the approach for real data, and explore test performance across a wider range of test conditions and design parameters.

Chapter 10

Conclusions, limitations, and future work

This work aimed at improving sensitivity, reducing test time, and controlling specificity for objective ABR detection methods. This was achieved by (1) developing, optimising, evaluating, and comparing both new and existing objective detection methods across a range of test conditions and pre-processing settings, (2) by developing, optimising, and evaluating a novel adaptive sequential testing framework for ABR detection, and (3) by developing and evaluating a new adaptive sample-size re-estimation procedure for ABR detection. In what follows, the main conclusions associated with these topics are covered in more detail, after which some directions for future work are layed out.

10.1 Conclusions

10.1.1 Improving the performance of objective ABR detection methods

To improve the performance of objective ABR detection methods, the focus was firstly on an in-depth assessment of specificity (Chapter 5). The emphasis for the specificity assessment was on the main assumptions underlying ABR detection methods, which were evaluated across a range of pre-processing parameters deemed typical for ABR detection. Results show that the main culprit for a poor control of specificity was the independence assumption between epochs, which was violated as a function of the high-pass cut-off frequency and the stimulus rate. Specific combinations of the high-pass cut-off frequency and the stimulus rate resulted in relatively large deviations from the nominal α-level of the test (ranging from 0.0385 to 0.0985 for $\alpha = 0.05$; Figure 5.3), whereas other combinations were found to be safe (e.g. when using a high-pass cut-off frequency of 100 Hz and a 33.3 Hz stimulus rate). Significant violations to the normality and stationarity assumptions were also observed, which resulted in a tendency towards a conservative

test performance with maximum deviations of 0.0161 and 0.0335 (for $\alpha = 0.05$) for normality (Figure 5.6) and stationarity (Figure 5.9) violations, respectively. Violations to the normality assumption were furthermore attributed to excessive kurtosis due to outliers, and were effectively dealt with through artefact rejection, whereas stationarity violations were successfully removed (with no noticeable adverse effects) by normalising the variances of the epochs. Finally, early results suggest that 'bootstrapping in blocks of epochs' can be used for a more robust assessment of test significance under independence violations (Figure 5.4). The bootstrap is also robust to normality and stationarity violations, and might therefore provide a solution to all aforementioned violations, thus giving an improved control of specificity.

In terms of sensitivity and test time, the focus was on developing, optimising, evaluating, and comparing new and existing methods across a range of feature sets and test conditions. Throughout this work, the Hotelling's T^2 test (applied in either the time or frequency domain) gave a sensitive and robust performance across test conditions. The performance of the Hotelling's T^2 test was also optimised for ABR detection in terms of which EEG features to use for the analysis (Appendix, section A.3). An additional method worth mentioning is the bootstrapped correlation coefficient (CC), which has the potential of providing a highly sensitive test statistic (Figure 6.2), under the condition that the true ABR waveform morphology is known *a priori*. This information is, however, typically not available (in practice the true waveform morphology remains unknown). When the match between the template and the true ABR waveform is poor, a low test sensitivity can be expected. A solution is to combine the CC with a non-template specific method, such as the Hotelling's T^2 test, and to evaluate test significance using the bootstrap approach. Results from this work indeed demonstrate a highly sensitive and robust performance for the 'T2 Time + CC' combination. When compared to the Fsp (evaluated using theoretical F-distributions), a maximum increase in test sensitivity of 70-75% was observed for the simulations, and ∼50% for the subject recorded data (Figures 6.2 and 6.3).

10.1.2 Sequential testing

A novel method (the CGST) for finding the stage-wise critical decision boundaries and controlling the FPR for sequential tests was proposed (chapter 7) and evaluated for ABR detection (chapter 8). Results from simulations and real recordings of EEG background activity first confirm that the CGST controls specificity as intended for ABR detection (Figure 8.1, Table 8.1), under the condition that its underlying assumptions remain satisfied. These include (1) that the stage-wise p value null distributions are independent, and (2) that the stage-wise p value null distributions are uniform on the [0,1] interval. As shown in chapter 8, the main concern for ABR detection is assumption (2), which is only satisfied when the assumptions underlying the statistical detection method are also satisfied. This emphasizes the importance of using suitable pre-processing and test

parameters (e.g. the high-pass cut-off frequency and stimulus rate) in combination with detection methods that have a good control of specificity, such as the Hotelling's T^2 test or bootstrapped statistics.

Sensitivity and test time of the CGST were explored using a sequentially applied Hotelling's T^2 test (chapter 8). Results demonstrate various trade-offs between statistical power and test time, primarily as a function of the number of sequential stages K and the choice for the stage-wise critical decision boundaries. Starting with the number of stages K, simulation results demonstrate that the trade-off between statistical power and test time is beneficial for ABR detection, with reductions in *mean* test time for $K = 5$ relative to $K = 1$ of up to 40-45%, with no loss in statistical power (Figure 8.3, plot B). In order to achieve these results, it is necessary to increase the *maximum* test time Figure 8.3, plot A). Hence, when used in practive, test time for some subjects test time will be prolonged, yet the mean test time (across a cohort of subjects) will tend to be decreased (relative to the single shot test). The increased maximum test time (for $K > 1$), however, has consequences for the no-stimulus condition, i.e. when a response is absent, the test will proceed to the final stage of the trial in (1-α)100% of the cases. For the no-stimulus condition, the *mean* test time is therefore close to the *maximum* test time, which emphasizes the importance of futility stopping (early acceptance of H_0) for the sequential test. Results from chapter 8 indeed demonstrate large reductions in mean test time for the no-stimulus condition when early stopping in favour of H_0 was permitted, potentially at the cost of a reduced test sensitivity (Figure 8.3, plots F and G).

10.1.3 Adaptive sample-size re-estimation

A new adaptive sample size re-estimation procedure was proposed and briefly evaluated for ABR detection in chapter 9. The main advantage for the adaptive approach (over a non-adaptive approach) is a reduced uncertainty with respect to the power of the EEG background activity, i.e. the EEG background activity can be estimated for the specific recording in question, which allows for a more informed decision with respect to the ensemble size N. Simulation results show a reduced test time (relative to the non-adaptive approach) of 10-30% for the stimulus condition and 25-45% for the no-stimulus condition, with a more or less equal performance int terms of test sensitivity (Figure 9.5). An additional advantage for the adaptive approach is an improved control over the TPR and the FNR, which can now be specified explicitly by the user through the γ_i values (the stage-wise TPRs) and the Γ_i values (the stage-wise FNRs). This can help bring ABR examinations to an unambiguous test outcome in terms of 'ABR present' or 'ABR absent' (or abnormal). In particular, when the test is stopped for futility ($\Sigma_k < B_k$), it can be concluded with certainty $> (1 - \sum_{i=1}^{K} \Gamma_i)100\%$ that the response is either absent or smaller than the assumed response, whereas when the test is stopped for efficacy ($\Sigma_k > B_k$) it can be concluded with certainty $> (1-\alpha)100\%$ that a response is present.

10.2 Limitations

A first limitation with this work is that the subject ABR threshold data was obtained from just 12 normal hearing adults, and might therefore not be representative of the true population of normal hearing adults. Consequently, many results from this thesis (which were generated either directly or indirectly using this data set) need to be verified using a much larger cohort of normal hearing adults. This is relevant primarily for the feature optimisations in the appendix (section A.3) and the comparisons in sensitivity and test time amongst detection methods (chapter 6). Results from the non-adaptive sequential test procedure and the adaptive ensemble size re-estimation procedure (chapters 8 and 9, respectively) might also be data-dependent, and should similarly be verified using a larger cohort of test subjects.

A second limitation for this work is that the simulations frequently used Gaussian, stationary, coloured noise for representing the EEG background activity, whereas real EEG background activity is not a stationary, Gaussian process. That said, results from chapter 5 suggest that violations to the stationariy and Gausianity assumptions are relatively minor for ABR data (assuming suitable pre-processing and artefact rejection strategies are used). It is also worth pointing out that simulations in this work were typically used to provide additional verification or supporting evidence for results obtained from real data, and that conclusions are seldom drawn from simulations alone. Chapter 9 is an exception to the latter, i.e. the performance of the adaptive ensemble size re-estimation procedure was explored with just simulations. As discussed in chapter 9, this is because the available ABR data were not collected with a view to evaluate a sequential test procedure, i.e. the *maximum* test time for the ABR data was too short. In future work, the adaptive approach should also be tested using suitable real data sets.

More specific limitations for this work were summarised in the discussions and conclusions of their respective chapters, and are not repeated here. Just to mention a few, these limitations are in regards to how the underlying statistical assumptions were evaluated in chapter 5 (see also section 5.5.1), and in regards to the adaptive ensemble size re-estimation procedure (see also the discussion in section 9.4).

10.3 Future work

10.3.1 The bootstrap

Additional work is required to test whether the bootstrap can be used for a more robust evaluation of test significance under independence violations. In particular, the number of epochs in each resampled block needs to be explored. Increasing the number of epochs per block will presumably increase robustness to independence violations, which comes at the cost of a reduced variation in the starting positions of the epochs. When this

variation is too small, then the bootstrapped null distribution may be insufficiently disperse, i.e. it may be an inaccurate representation of the true null distribution, resulting in unreliable and/or biased critical decision boundaries.

10.3.2 The adaptive sequential test

Variability in AER waveform morphologies across subjects and stimuli means that the optimal feature set for response detection (i.e. which EEG features to include in the statistical test) will be both subject- and stimulus-dependent. Future work might explore different approaches for optimising both the selected features and the choice for statistical test throughout the trial, with a view to increase test sensitivity and reduce test time. It is envisioned that the bootstrap approach will play an important role here, as this gives a large amount of freedom when choosing which test statistics to use for AER detection. Moreover, the bootstrap allows multiple EEG features to be combined efficiently into a single test statistic (section 3.6.2). The possibilities for potential feature adaptations and/or optimisations are therefore vast.

10.3.3 Adaptive sample size re-estimation

Future work might further develop and evaluate the adaptive sample size re-estimation procedure from chapter 9 across a wider range of design parameters and pre-processing parameters. This includes different choices for the γ_i and Γ_i values (the stage-wise TPRs and FNRs, respectively), the choice for the assumed response, and the adopted resampling approach. The specificity of the approach should also be evaluated across a wider range of pre-processing parameters (e.g. for different high-pass cut-off frequencies and stimulus rates), and the approach should be tested in subject data. Finally, comparisons in test performance with alternative methods from the literature might also be drawn (e.g. Proschan & Hunsberger, 1995; Lehmacher & Wassmer, 1999; Chow & Chang, 2007; Mehta & Pocock, 2010).

10.3.4 Fast hearing threshold estimation using the CGST

Most test procedures for ABR audiometry aim at estimating the behavioural hearing thresholds using a simple 'up-down' approach, i.e. stimulus intensity is decreased for every correct response, or increased for a missed response. A more efficient approach might be developed by continuously switching between stimulus intensities, such that the amount of information gained (regarding the location of the behavioural hearing threshold) is maximized. Future work might explore how the latter can be realised within an adaptive CGST framework. In particular, at each stage of the trial, a post-hoc power analysis can be conducted to choose the stimulus intensity for the next stage.

10.3.5 New test paradigms for behavioural hearing threshold estimation using the CGST

When using the CGST, the assumption that the response is deterministic is relaxed, i.e. the response is assumed to be deterministic within each block, as opposed to across all acquired data. As a result, the stimulus can potentially be adjusted at each stage of the sequential analysis. This might have some use for CAEP detection, i.e. modifying the type of stimulus at each stage of the analysis might help the subject to focus on the stimuli, thus giving stronger CAEP responses.

Mismatch negativity

Mismatch negativity (MMN) is a response to an 'odd ball' in a sequence of regular events. A MMN response might therefore be expected when changing the stimulus at each stage of the analysis, which might be used as further indication that the subject can hear the acoustic stimulus. In particular, test significance for both the ABR and the MMN could be evaluated simultaneously using the bootstrap approach for multiple features (section 3.6.2). Note that an 'oddball' can potentially be inserted at any moment, not just at the stage transitions. Note also that the bootstrap for multiple features does not require the EEG pre-processing parameters to be identical for each feature, i.e. the MMN could be pre-processing using a different set of parameters relative to the ABR.

Bibliography

[1] Abe M. (1954). Electrical Responses of the Human Brain to Acoustic Stimulus. *The Tohoku Journal of Experimental Medicine*, 60(1), pp. 47-58.

[2] Adrian E.D. (1941). Afferent discharges to the cerebral cortex from peripheral sense organs. *J. Physiol.*, 100, pp. 159-91.

[3] Adrian E.D. & Matthews B.H.C. (1934). The Berger rhythm: Potential changes from the occipital lobes of man. *Brain.* 57, pp. 355-385.

[4] Allison T., Wood C.C. & Goff W.R. (1983). Brain stem auditory, pattern-reversal visual, and short-latency somatosensory evoked potentials: latencies in relation to age, sex, and brain and body size. *Electroencephalography and clinical neurophysiology*, 55(6), pp. 619-636.

[5] Alpsan, D, & Özdamar, Ö. (1992a). Auditory brainstem evoked potential classification for threshold detection by neural networks. I. Network design, similarities between human-expert and network classification, feasibility. *Automedica*, 15(1), pp. 67-82.

[6] Alpsan, D, & Özdamar, Ö. (1992b). Auditory brainstem evoked potential classification for threshold detection by neural networks. II. Effects of input coding, training set size and composition and network size on performance. *Automedica*, 15(1), pp. 83-93.

[7] Armitage P., McPherson C.K. & Rowe B.C. (1960). Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2), pp. 235-244.

[8] Arnold S.A. (1985). Objective versus visual detection of the auditory brainstem response. *Ear and Hearing*, 6(3), pp 144-150.

[9] Aunon J.I. (1978). Computer techniques for the processing of evoked potentials. *Comp. Prog. in Biomed.,* 8(3-3), pp. 243-255.

[10] Bachen N.I. (1986). Detection of stimulus-related (evoked response) activity in the electroencephalogram (EEG). *IEEE. Trans. Biomed. Eng.*, 33(6), pp. 566-571.

[11] Barlow J.S. & Brown R.M. (1955). An analog correlator system for brain potentials. Res. Lab. of Electronics, M.I.T., Cambridge, Mass., Tech. Rept. No. 300.

[12] Bauer P. & K´ohne K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4), pp. 1029-1041.

[13] Berger H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv f´ur Psychiatrie und Nervenkrankheiten*, 87(1), pp. 527-570

[14] Berger H. (1931). Über das Elektrenkephalogramm des Menschen. Dritte Mitteilung. *Archiv f´ur Psychiatrie und Nervenkrankheiten*, 94(1), pp. 16-60

[15] Berger H. (1934). Über das Elektrenkephalogramm des Menschen. Achte Mitteilung. *Archiv f´ur Psychiatrie und Nervenkrankheiten*, 101, 452.

[16] Berger H. (1937). Über das Elektrenkephalogramm des Menschen. Dreizahnte Mitteilung. *Archiv f´ur Psychiatrie und Nervenkrankheiten*, 106, 577.

[17] Berger J.O. Philippe A. & Robert C.P. (1998). Estimation of Quadratic Functions: Noninformative Priors for Noncentrality Parameters. *Statist. Sinica.*, 8(2), pp. 359-375.

[18] Bienaymé I. J. (1874). Sur line question de probabiliés. *Bull. Math. Soc*, Fr. 2, pp. 153-154.

[19] Bilodeau M., & Brenner D. Theory of Multivariate statistics. 1999. Springer, New York.

[20] Blegvad B. (1975). Binaural Summation of Surface-Recorded Electroencephalographic Responses in *Normal-hearing Subjects. Scandanavian Audiology*, 4(4), pp. 233-238.

[21] Bullmore E., Suckling J., Overmeyer S., Rabe-Hesketh S., Taylor E. & Brammer M. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans Med Imag.*, 18(1), pp. 32-42..

[22] Brannath Q.,, Posch M. & Bauer P. (2002). Recursive combination tests. *J. Am. Stat. Assoc.,* 97(457), pp. 236-244,.

[23] Bruton A., Conway J.H. & Holgate S.T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy*, 86(2), pp. 94-99.

[24] Cardinal R.N. Graduate-level statistics for psychology and neuroscience. ANOVA in practice, and complex ANOVA designs. May 2004, Version 2. Available online at https://egret.psychol.cam.ac.uk/psychology/graduate/Guide_to_ANOVA.pdf

[25] Carter L., Golding M., Dillon H. & Seymour J. (2010). The Detection of Infant Cortical Auditory Evoked Potentials (CAEPs) Using Statistical and Visual Detection Techniques. *J. Am. Acad. Audiol.*, 21(5), pp. 347–356.

[26] Cebulla M. & Stürzebecher E. (2015). Automated auditory response detection: Further improvement of the statistical test strategy by using progressive test steps of iteration. *International Journal of Audiology*, 54(8), pp. 568-572.

[27] Cebulla M., Stürzebecher E. & Elberling C. (2006). Objective detection of Auditory Steady State Responses: Comparison of One-Sample and q-Sample Tests. *J. Am. Acad. Audiol.*, 17(2), pp. 93-103.

[28] Cebulla M., Stürzebecher E. & Wernecke K.D. (1996). Objective detection of auditory evoked potentials - comparison of several statistical tests in the frequency domain by means of Monte Carlo simulations. *Scand. Audiol.*, 25(3), pp. 201-206.

[29] Cebulla M., Stürzebecher E. & Wernecke K.D. (2000a). Objective detection of auditory brainstem potentials - Comparison of statistical tests in the time and frequency domain. *Scand. Audiol.*, 29(1), pp. 44-51.

[30] Cebulla M., Stürzebecher E. & Wernecke K.D. (2000b). Comparison of several SNR estimates for objective response detection in noise. *Z. Audiol.*, 39(1), pp. 14-22.

[31] Cebulla M., Stürzebecher E. & Wernecke K.D. (2001). Objective detection of the amplitude modulation following response (AMFR). *Audiology*, 40(5), pp. 245–252.

[32] Champlin C.A. (1992). Methods for detecting auditory steady-state potentials recorded from humans. *Hearing Research*, 58(1), pp. 63-69.

[33] Chang M. (2006). Adaptive design method based on sum of p-values. *Statist.Med.,* 26(14), pp. 2772-2784.

[34] Chang H.W., Dillon H., Carter L., Van Dun B. & Young S.T. (2012). The relationship between cortical auditory evoked potential (CAEP) detection and estimated audibility in infants with sensorineural hearing loss. *International Journal of Audiology*, 51(9), pp. 663-70.

[35] Chow M.S. (1987). A Complete Class Theorem for Estimating a Noncentrality Parameter. *Ann. Statist.*, 15(2), pp. 800-804.

[36] Clark W.A., Goldstein M.H., Brown Jr., R. M., O'Brien D. F. & Zieman H. E. (1961). The average response computer (ARC): A digital device for computing averages and amplitude and time histograms of electrophysiological responses. *Trans. IRE*, 8(1), pp. 46-51.

[37] Cobb W. & Morton H.B. (1952). The human retinogram in response to high-intensity flashes. *Electroencephalography and Clinical Neurophysiology*, 4(4), pp. 547-556.

[38] Cohen J. Statistical Power Analysis for the Behavioral Sciences. 1988. 2nd edition. Lawrence Erlbaum Associates.

[39] Davis P.A. (1939). Effects of acoustic stimuli on the waking human brain. *Journal of Neurophysiology*, 2(6), pp. 494-499.

[40] Davis H., Davis P.A., Loomis A.L., Harvey N. & Hobart G. (1938). Human brain potentials during the onset of sleep. *Journal of Neurophysiology*, 1(1), pp. 24-38.

[41] Davis H., Davis P.A., Loomis A.L., Harvey E.N. & Hobart G. (1939). Electrical reactions of the human brain to auditory stimulation during sleep. *J. Neurophysiol.*, pp. 500-514.

[42] Dawsen G.D. (1947). Cerebral responses to electrical stimulation of peripheral nerve in man. *J Neurol Neurosurg Psychiatry.*, 10(3), 134-140.

[43] Dawsen G.D. (1950). Cerebral responses to nerve stimulation in man. *Brit. Med. Bull.*., 6(4), 326-9.

[44] Dawsen G.D. (1951). A summation technique for detecting small signals in a large irregular background. *J Physiol.* 115(1):2p-3p.

[45] Dawsen G.D. (1954). A summation technique for detection of small evoked potentials. *Electroencephalography & Clinical Neurophysiology,* 6(1), pp. 65-84.

[46] Derbyshire A.J., Fraser A.A., McDermott M. & Bridge A. (1956). Audiometric measurements by electroencephalography. *Electroencephalography and Clinical Neurophysiology*, 8(3), pp. 467-478.

[47] Dobie R.A. & Wilson M.J. (1989). Analysis of Auditory Evoked Potentials by Magnitude-Squared Coherence. *Ear and Hearing*, 10(1), pp. 2-13.

[48] Dobie R.A. & Wilson M.J. (1993). Objective response detection in the frequency domain. *Electroenceph. clin. Neurophysiol.*, 88(6), pp. 516-524.

[49] Dobie R.A. & Wilson M.J. (1994). Objective detection of 40 Hz auditory evoked potentials: phase coherence vs. magnitude-squared coherence. *Electroenceph. clin. Neurophysiol.*, 92(5), pp. 405-413.

[50] Dodge H.F. & Romig H.G. (1929). A method for sampling inspection. *Bell Syst. Tech. J.*, 8(4), pp. 613-631.

[51] Don M., Allen A. R. & Starr A. (1977). Effects of click rate on the latency of auditory brain stem responses in humans. *The Annals of otology rhinology, and laryngology*, 86(2), pp. 186-195.

[52] Don M. & Elberling C. (1994). Evaluating residual background noise in human auditory brainstem responses. *The Journal of the Acoustical Society of America*, 96(5), pp. 2746-2757.

[53] Don M. Ponton C.W. Eggermont J.J. & Masuada A. (1994). Auditory brainstem response (ABR) peak amplitude variability reflects individual differences in cochlear response times. *The Journal of the Acoustical Society of America*, 96(6), pp. 3476-3491.

[54] Efron B. & Tibshirani R.J. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[55] Elberling C. (1976). Action Potentials Recorded from the Promontory and the Surface, Compared with Recordings from the Ear Canal in Man. *Scand. Audiol.*, 5(2), pp. 69-78.

[56] Elberling C. (1979). Auditory electrophysiology: spectral analysis of cochlear and brain stem evoked potentials. *Scand. Audiol.*, 8(1), pp. 57-64.

[57] Elberling C., Callø J. & Don M. (2010). Evaluating Auditory Brainstem Responses to Different Chirp Stimuli at Three Levels of Stimulation. *The Journal of the Acoustical Society of America*, 128(1), pp. 215–223.

[58] Elberling C. & Don M. (1984). Quality estimation of averaged auditory brainstem responses. *Scandinavian audiology*, 13(3), pp. 187-197.

[59] Fisher L.D. (1998). Self-Designing Clinical Trials. *Statistics in Medicine*, 17(14), pp. 1551-1562.

[60] Fisher R.A. *Statistical methods for research workers*, 11th ed. Oliver and Boyd, Edinburgh, 1932.

[61] Fisher R.A. (1935). The Design of Experiments, New York: Hafner

[62] Fowler C.G. & Noffsinger D. (1983). Effects of stimulus repetition rate and frequency on the auditory brainstem response in normal cochlear-impaired, and VIII nerve/brainstem-impaired subjects. *Journal of speech and hearing research*, 26(4), pp. 560-567.

[63] Freeman D.T. (1992). Computer Applications in Otolaryngology: Computer Recognition of Brain Stem Auditory Evoked Potential Wave V by a Neural Network. *Annals of Otology, Rhinology & Laryngology*, 101(9), pp. 82-790.

[64] Fria T.J. & Doyle W.J. (1984). Maturation of the auditory brain stem response (abs): Additional perspectives. *Ear and Hearing*, 5(6), pp. 361-365.

[65] Fridman J., John E.R., Bergelson M., Kaiser J.B. & Baird H.W. (1982). Application of digital filtering and automatic peak detection to brain stem auditory evoked potentials. *Electroenceph. clin. Neurophysiol.*, 53(4), pp. 405-416.

[66] Fridman J., Zapulla R., Bergelson M., Greenblatt E., Malis L., MorRell F. & Hoeppner T. (1984). Application of Phase Spectral Analysis for Brainstem Auditory Evoked

Potential Detection in Normal Subjects and Patients with Posterior Foassa Tumors. *Audiology*, 23(1), pp. 99-113.

[67] Friedman M. 1937. The use of ranks to void the assumption of normality implicit in the analysis of variance. *J. Am. Statist. Assoc.*;32:675-701.

[68] Geisler C.D. 1960. Average responses to clicks in man recorded by scalp electrodes. *Massachusetts Institute of Technology, Research Laboratory of Electronics.* Technical report 380.

[69] Geisler C.D., Frishkopf L.S., Rosenblith W.A. 1958. Extracranial responses to acoustic clicks in man. *Science* 128, 1210-1211.

[70] Gentiletti-Faenze G.G., Yañ ez-Suarez O. & Cornejo-Cruz J.M. (2003). Evaluation of Automatic Identification Algorithms for Auditory Brainstem Response used in Universal Hearing Loss Screening. *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE,* 3, pp. 2857-2860.

[71] Gidoll S. H. (1952). Quantitative determination of hearing of audiometric frequencies in the electroencephalogram. *Arch. Otolaryng.*, 55(5), pp. 597-601.

[72] Girden E.R. (1992). ANOVA: Repeated Measures, *Sage university paper series on qualitative applications in the social sciences, 84, Newbury Park, CA: Sage.*

[73] Golding M., Dilon H., Seymour J. & Carter L. (2009). The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection. *International Journal of Audiology*, 48(12), pp. 833-842.

[74] Goldstein R., Rodman L.B. 1967. Early Components of Averaged Evoked Responses to Rapidly Repeated Auditory Stimuli. *Journal of Speech, Language, and Hearing Research*, Vol. 10, 697-705

[75] Greenblatt E., Zapulla R.A., Kaye S. & Fridman J. (1985). Response threshold determination of the brain stem auditory evoked response: a comparison of the phase versus magnitude derived from the Fast Fourier Transform. *Audiology*, 24(4), pp. 288-296.

[76] Greenhouse S.W. & Geisser S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24(2), pp. 95-112.

[77] Grinstead C.M. & Snell J.L. *Introduction to Probability*, 2nd ed., American Mathematical Society, the United States of America, 1997.

[78] Habraken J.B.A., Van Gils M.J. & Cluitmans P.J.M. (1993). Identification of peak V in brainstem auditory evoked potentials with neural networks. *Computers in Biology and Medicine*, 23(5), pp. 369-380.

[79] Hall J. W. (2006). *New Handbook of Auditory Evoked Responses*. 1st ed. London: Pearson.

[80] Hartung J. & Knapp G. 2003. A new class of completely self-designing clinical trials. *Biometrical J.*, 45(1), pp. 319.

[81] Heyde C.C. & Seneta E. (1972). Studies in the History of Probability and Statistics. XXXI.The simple branching process, a turning point test and a fundamental Inequality: A historical note on I. J. Bienaymé. *Biometrika*, 59(3), pp. 680–683

[82] Hood L.J. (1998). *Clinical Applications of the Auditory Brainstem Response*. Singular Publishing Group, Inc, Aan Diego London.

[83] Hotelling H. 1931. The Generalization of Student's Ratio.*Ann. Math. Statist.*, 2(3), pp. 360-378.

[84] Huynh H. & Feldt L.S. (1970). Conditions under which mean square ratios in repeated measures designs have exact F-distributions. *Journal of the American Statistical Association*, 65(332), pp. 1582 - 1589.

[85] Jerger J. & Hall J. (1980). Effects of age and sec on auditory brainstem response. *Archives of otolaryngology*, 106(7), pp. 387-391.

[86] Jervis B.W., Nichols M.J., Johnson T.E., Allen E. & Hudson N.R. (1983). A fundamental investigation of auditory evoked potentials. *IEEE. Trans. Biomed. Eng.*, 30(1), pp. 43-50.

[87] Jewett D.L., Romano M.N., Williston J.S. 1970. Human Auditory Evoked Potentials possible brain stem components detected on the scalp. *SCIENCE* 167(3924):1517-8

[88] Jewett, D. & Williston, J. 1971. Auditory-evoked far-fields averaged from the scalp of humans. *Brain*, 94, 681–696.

[89] Jones T.A., Stockard J.J. & Weidner W.J. (1980). The effects of temperature and acute alcohol intoxication on brain stem auditory evoked potentials in the cat. *Electroencephalography and clinical neurophysiology*, 49(1-2), pp. 23-30.

[90] Kevanishvili Z. & Aphonchenko V. (1979). Frequency composition of the brain-stem auditory evoked potential. *Scand. Audiol.*, 8(1), pp. 51-55.

[91] Kanyongo G.Y., Brook G.P., Kyei-Blankson L. & Gocmen G. (2007). Reliability and Statistical Power: How Measurement Fallibility Affects Power and Required Sample Sizes for Several Parametric and Nonparametric Statistics. *Journal of Modern Applied Statistical Methods*, 6(1), pp. 81-90.

[92] Kubokawa T., Marchand E. & Strawderman W.E. (2017). A Unified Approach to Estimation of Noncentrality Parameters, the Multiple Correlation Coefficient, and Mixture Models. *Mathematical Methods of Statistics*, 26(2), pp. 134-148.

[93] Kubokawa T., Robert C.P. & Saleh A. K. Md. E. (1993). Estimation of Noncentrality Parameters. *Canad. J. Statist.*, 21(1), pp. 45-57.

[94] Lachowska M., Bohórquez J. & ´Ozdamar ´O. (2012). Simultaneous Acquisition of 80 Hz ASSRs and ABRs from Quasi ASSRs for Threshold Estimation. *Ear and Hearing*, 33(5), pp. 660–671.

[95] Lan K.K.G. & DeMets D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3), pp. 659-663.

[96] Le Boudec J-Y. *Performance and Evaluation of Computer and Communication Systems*. Version 2.3, May 19, 2015. Available at http://perfeval.epfl.ch

[97] Lehmacher W. & Wassmer G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4), pp. 1286-1290.

[98] Leung P.L. & Muirhead R.J. (1987). Estimation of Parameter Matrices and Eigenvalues in MANOVA and Canonical Correlation Analysis. *Ann. Statist.*, 15(4), pp. 1651-1666.

[99] Levin J.R. & Subkoviak M.J. (1977). Planning an Experiment in the Company of Measurement Error. *Applied Psychological Measurement*, 1(3), pp. 331-338.

[100] Li Q., Zhang J. & Dai S. (2009). On Estimating the Noncentrality Parameter of a Chi-Squared Distribution. *Statist. Probab. Lett.*, 79(1), pp. 98-104.

[101] Lins O.G., Picton T.W., Boucher B., Durieux-Smith A., Champagne S.C., Moran L.M., Perez-Abalo M.C., Martin V. & Savio G. (1996). Frequency-specific audiometry using steady-state responses. *Ear Hear.*, 17(2), pp. 81-96.

[102] Liu Q. & Chi G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics*, 57(1), pp. 172-177.

[103] Loomis A. L., Harvey E. N. & Hobart G.A. (1935). Potential rhythms of the cerebral cortex during sleep. *Science*, 81(2111), pp. 597-8.

[104] Loomis A. L., Harvey E. N. & Hobart G. A. (1937). Cerebral states during sleep, as studied by human brain potentials. *Journal of Experimental Psychology*, 21(2), pp. 127-144.

[105] Lowy K. & Weiss B. (1968). Assessing the significance of averaged evoked potentials with an on-line computer. The split-sweep method. *Electroenceph. clin. Neurophysiol.,* 25(2), pp. 177-180.

[106] Lv J., Bell S.L. & Simpson D.M. A statistical test for the detection of auditory evoked potentials. At *IPEM Annual Scientific Meeting IPEM Annual Scientific Meeting*, 06 - 08, Sep., 2004.

[107] Lv J., Simpson D.M. & Bell S.L. (2007). Objective detection of evoked potentials using a bootstrap technique. *Medical Engineering & Physics*, 29(2), pp. 191–198.

[108] Maag U.R. (1966). A k-sample analogue of Watson's U2 statistic. *Biometrika*, 53(3-4), pp. 579-583.

[109] Madsen S.M.K. (2010). Accuracy of averaged auditory evoked potential amplitude and latency estimates. Msc thesis, Dept. Elec. Eng., Tech. Uni of Denmark.

[110] Madsen S.M.K., Harte J.M., Elberling C. & Dau T. (2017). Accuracy of averaged auditory evoked potential amplitude and latency estimates. *International Journal of Audiology*, 57(2), pp. 1-9.

[111] Marcus R.E., Gibbs E.L. & Gibbs F. A. 1949. Electroencephalography in the diagnosis of hearing loss in the very young child. *Dis. Nerv. System*, 10, 170.

[112] Mardia K.V. Statistics of directional data. *London and New York* New York: Academic Press, 1972.

[113] Maris E. & Oostenveld R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), pp. 177-190.

[114] Marple S.L.Jr. *Digital Spectral Analysis with Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1987.

[115] Martin W. H., Schwegler J. W., Gleeson A. L. & Shi Y-B. (1994). New techniques of hearing assessment. *Otolaryngol. Clin. North Am.*, 27(3), pp. 487-510.

[116] Mast J. & Victor J.D. (1991). Fluctuations of steady-state VEPs: interaction of driven evoked potentials and the EEG. *Electroenceph. clin. Neurophysiol.*, 78(5), pp. 389-401.

[117] Mathai A.M. (1972). The Exact Distributions of Three Multivariate Statistics Associated with Wilks' Concept of Generalized Variance. *Sankhyā: The Indian Journal of Statistics, Series A*, 34(2), pp. 161-170.

[118] McPherson K. (1974). Statistics: the problem of examining accumulating data more than once. *New Engl. J. Med.*, 290(9), pp. 501-502.

[119] Meyer P.L. (1967). The maximum likelihood estimate of the non-centrality parameter of a non-central χ^2 variate. *J. Amer. Statist. Assoc.*, 62(320), pp. 1258-1264.

[120] Michalewski H.J., Thompson L.W., Patterson J.V., Bowman T.E. & Litzel-man D. (1980). Sex differences in the amplitude and latencies of the human auditory brain stem potential. *Electroencephalography and clinical neurophysiology*, 48(3), pp. 351-356.

[121] Mijares E., Pérez Abalo M.C., Herrera D., Lage A. & Mayrim Vega. (2013). Comparing statistics for objective detection of transient and steady-state evoked responses in newborns. *International Journal of Audiology*, 52(1), pp. 44-49.

[122] Millet D. (2001). Hans Berger: From Psychic Energy to the EEG. *Perspectives in Biology and Medicine*, 44(4), pp. 522-542.

[123] Miranda de Sá A.M.F.L., Felix L.B. & Infantosi A.F.C. (2004). A matrix-based algorithm for estimating multiple coherence of periodic signal and its application to the multichannel EEG during sensory stimulation. *IEEE Trans. Biomed. Eng.*, 51(7), pp. 1140-1146.

[124] Miziara I.D., Miziara C.S.M.G., Tsuji R.K. & Bento R.F. (2012). Bioethics and medial/legal considerations on cochlear implants in children. *Brazilian Journal of Otorhinolaryngology*, 78(3), pp70-79.

[125] Mizrahi E.M., Maulsby R.L. & Frost J.D. (1983). Improved wave v resolution by dual-channel brain stem auditory evoked potential recording. *Electroencephalography and clinical neurophysiology*, 55(1), pp. 105-107.

[126] Moore B.R. (1980). A modification to the rayleigh test for vector data. *Biometrika*, 67(1), pp. 175-180.

[127] Mordkoff T.J. (2016). The Assumption(s) of Normality. available online at: http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/I.07%20normal.pdf

[128] Morley, G. K., and K. E. Liedke. 1977. Automated evoked potential analysis using peak and latency discrimination. *Proc. San Diego Biomed. Symp.* 16, 291-298.

[129] Ḿuller H.H. & Sh́afer H. (2001). Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches. *Biometrics*, 57(3), pp. 886-891.

[130] Neely S.T. & Pepe M.S. (1997). United States Patent Number: 5,637,379. Method and Apparatus for Objective and Automated Analysis of Auditory Brainstem Response to Determine Hearing Capacity.

[131] Neff N. & Strawderman W.E. (1976). Further Remarks on Estimating the Parameter of a Noncentral Chi-Squared Distribution. *Communications in Statistics - Theory and Methods*, 5(1), pp. 65-76.

[132] O'Brien P.C. & Fleming T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35(3), pp. 549-556.

[133] Özdamar Ö. & Delgado R.E. (1996). Measurement of Signal and Noise Characteristics in Ongoing Auditory Brainstem Response Averaging. *Annals of Biomedial Engineering,* 24(6), pp. 702-715.

[134] Perlman M.D. & Rasmussen V.A. (1975). Some Remarks on Estimating a Noncentrality Parameter. *Comm. Statist. A.*, 4(5), pp. 455-468.

[135] Picton T.W., Dimitrijevic A., Sasha J.M. & Roon Van P. (2001). The use of phase in the detection of auditory steady-state responses. *Clinical Neurophysiology*, 112(9), pp. 1698-1711.

[136] Picton T.W. & Hillyard S.A. (1974). Human auditory evoked potentials. II. Effects of attention. *Electroencephalography and Clinical Neurophysiology*, 36(2), pp. 191-199.

[137] Picton T.W., Hillyard S.A., Krausz H.I., Galambos R. (1974). Human Auditory Evoked Potentials I: Evaluation of Components. *Electroencephalography and Clinical Neurophysiology.* Vol 36: 179-190.

[138] Picton T.W., Linden R.D., Hamel G. & Maru J.T. (1983). Aspects of averaging. *Sem. Hear.,* 4(4), pp. 327:341 .

[139] Picton T.W., Vajsra F., Rodriguez R. & Campbell K.B. (1987). Reliability estimates from steady-state evoked potentials. *Electroenceph. clin. Neurophysiol.*, 68(2), pp. 119-131.

[140] Pearson K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, pp. 240–242.

[141] Penrose R. (1955). A Generalized Inverse for Matrices. *Proceedings of the Cambridge Philosophical Society*, 51(3), pp. 406–413.

[142] Perl E.P., Galambos R. & Glorig A. (1953). The estimation of hearing threshold by electroencephalography. *Electroencephalography and Clinical Neurophysiology*, 5(4), pp. 501-512.

[143] Pocock S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2), pp. 191-199.

[144] Polya G. 1920. tÜber den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Moment problem. *Math. Z.*, 8, pp. 171-178 .

[145] Popescu M., Papadimitriou S., Karamitsos D. & Bezerianos A. (1999). Adaptive Denoising and Multiscale Detection of the V Wave in Brainstem Auditory Evoked Potentials. *Audiology & neuro-otology*, 4(1), pp. 38-50.

[146] Proschan M.A. & Hunsberger S.A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51(4), pp. 1315-1324.

[147] Ramkalawan T.W, & Davis A.C. (1991). The effects of hearing loss and age of intervention on some language metrics in young hearing-impaired children. *British journal of audiology*, 26(2), pp. 97-107.

[148] Raz J., Turetsky B. & Fein G. (1988). Confidence Intervals for the Signal-to-Noise Ratio When a Signal Embedded in Noise is Observed Over Repeated Trials. *IEEE Transactions on Biomedical Engineering*, 35(8), pp. 646-649.

[149] Rencher A.C. Methods of Multivariate Analysis. Second Edition. John Wiley & Sons, Inc., 2001.

[150] Rodriguez R., Picton T, Linden D., Hamel G. & Laframboise G. (1986). Human auditory steady-state responses: effects of intensity and frequency. *Ear Hear.*, 7(5), pp. 300-313.

[151] Rukhin A.L. (1993). Estimation of the Noncentrality Parameter of an F-Distribution. *J. Statist. Plann. Inf.*, 35(2), pp. 201-311.

[152] Salamy A., & McKean C.M. (1977). Habituation and Dishabituation of Cortical and Brainstem Evoked Potentials. *International Journal of Neuroscience*, 7(3), pp. 175-182.

[153] Salamy A., McKean C.M. & Buda F. (1975). Maturational changes in auditory transmission as reflected in human brain stem potentials. *Brain Research*, 96(2), pp. 361- 365.

[154] Sánchez R., Riquenes A. & Pérez-Abalo M. (1995). Automatic detection of auditory brainstem responses using feature vectors. *International journal of bio-medical computing*, vol 39(3), pp. 287-97.

[155] Saxena K.M. L. & Alam K. (1982). Estimation of the Noncentrality Parameter of the Chi-Squared Distribution. *Ann. Statist.*, 10(3), pp. 1012-1016.

[156] Sayers B. McA., Beagley H.A. & Henshall W.R. (1974). The mechanism of auditory evoked EEG responses. *Nature*, 247(5441), pp. 481-483.

[157] Sayers B. McA., Beagley H.A. & Riha J. (1979). Pattern Analysis of Auditory-Evoked EEG Potentials. *Audiology*, 18, pp. 1-16.

[158] Schimmel H. (1967). The (±) reference: accuracy of estimated mean components in average response studies. *Science*, 157(3784), pp. 92-94.

[159] Schimmel H., Rapin I., Cohen M.M. 1974. Improving evoked response audiometry with special reference to the use of machine scoring. *Audiology*, 13: 33-65.

[160] Simpson D.M., Tierra-Criollo C.J., Leite R.T., Zayen E.J.B. & Infantosi A.F.C. (2000). Objective response detection in an electroencephalogram during somatosensory stimulation. *Annuals of Biomedical Engineering.* 28(6), pp. 691-698.

[161] Shao P.Y.S. & Strawderman W.E. (1995). Improving on the Positive Part of the UMVUE of a Noncentrality Parameter of a Noncentral Chi-Square Distribution. *J. Multivariate Anal.*, 53(1), pp. 52-66.

[162] Shen Y. & Fisher L. (1999). Statistical Inference for Self-Designing Clinical Trials With a One-Sided Hypothesis. *Biometrics* 55(1), pp.190-197.

[163] Sheng J. & Qiu L. (2007). p-Value calculation for multi-stage additive tests. *J. Stat. Comput. Sim.,* 77(12), pp. 1057-1064.

[164] Sheward W.A. (1931). *Economic control of Manufactured Product.* Van Nostrand, New York.

[165] Spruill M.C. (1986). Computation of the Maximum Likelihood Estimate of a Noncentrality Parameter. *Journal of Multivariate Analysis*, 18, pp. 216-226.

[166] Starr A. & Achor J. (1975). Auditory brain stem response in neurological disease. *Archives of Neurology*, 32(11), pp.761-768.

[167] Starr A., Amlie R.N., Martin W.H. & Sanders S. (1977). Development of auditory function in newborn infants revealed by auditory brainstem potentials. *Pediatrics*, 60(6), pp. 831-839.

[168] Student. (1908). The Probable Error of a Mean. *Biometrika*, 6(1), pp. 1-25.

[169] Stürzebecher E. & Cebulla M. (1997). Objective detection of auditory evoked potentials. Comparison of several statistical tests in the frequency domain on the basis of near-threshold ABR data. *Scandinavian Audiology*, 26(1), pp. 7-14.

[170] Stürzebecherr E. & Cebulla M. (2013). Automated auditory response detection: Improvement of the statistical test strategy. *International Journal of Audiology*, 52(12), pp. 861-864.

[171] Stürzebecher E., Cebulla M. & Elberling C. (2005). Automated auditory response detection: Statistical problems with repeated testing. *International Journal of Audiology*, 44(2), pp. 110-117.

[172] Stürzebecher E., Cebulla M. & Wernecke K.D. (1999). Objective response detection in the frequency domain: comparison of several q-sample tests. *Audiology & neuro-otology*, 4(1), pp. 2-11.

[173] Sutton G., Lightfoot G., Stevens J., Booth R., Brennan S., Feirn R. & Meredith R. (2013). Guidance for Auditory Brainstem Response Testing in Babies. Version 2.1. Reading, UK: British Society of Audiology.

[174] Suzuki T., Asawa I. (1957). Evoked potential of waking human brain to acoustic stimuli. *Acta. Oto-Laryn.*, 48(5-6), pp. 508-515.

[175] Suzuki T., Sakabe N. & Miyashita Y. (1982). Power Spectral Analysis of Auditory Brain Stem Responses to Pure Tone Stimuli. *Scandinavian Audiology*, 11(1), pp. 25-30.

[176] Terkildsen K., Osterhammel P. & Huis in't Veld F. Far-Field Electrocochleography, Adaptation. *Scandinavian Audiology*, 4(4), pp. 215-220.

[177] Thiébaux H.J. (1984). The Interpretation and Estimation of Effective Sample Size. *Journal of Climate and Applied Meteorology*, 23(5), pp. 800-811.

[178] Thornton A.R. & Coleman M.J. (1975). The adaptation of cochlear and brainstem auditory evoked potentials in humans. *Electroencephalography and Clinical Neurophysiology*, 39(4), pp. 399-406.

[179] Valderrama J.T., de la Torre A., Alvarez I., Segura J.C., Thornton A.R.D. & Sainz M. (2014). Automatic quality assessment and peak identification of auditory brainstem responses with fitted parametric peaks. *Computer Methods and Programs in Biomedicine.*, 114(3), pp. 262-275.

[180] Valdes-Sosa M.J., Bobes M.A., Perez-Abalo M.C., Perera M., Carballo J.A. & Valdes-Sosa P. (1987). Comparison of Auditory-Evoked Potential Detection Methods Using Signal Detection Theory. *Audiology,* 26(3), pp. 166-178.

[181] Valdes J.L., Perez-Abalo M.C., Martin V., Savio G., Sierra C., Rodriguez E. & Lins O. (1997). Comparison of Statistical Indicators for the Automatic Detection of 80 Hz Auditory Steady State Responses. *Ear & Hearing*, 18(5), pp. 420-429.

[182] Van Dun B., Carter L. & Dillon H. (2012). Sensitivity of cortical auditory evoked potential (CAEP) detection for hearing-impaired infants in response to short speech sounds. *Audiology Research*, 2(1), e13.

[183] Van Dun B., Dillon H. & Seeto M. (2015). Estimating Hearing Thresholds in Hearing-Impaired Adults through Objective Detection of Cortical Auditory Evoked Potentials. *Journal of the American Academy of Audiology*, 26(4), pp. 370-83.

[184] Victor J.D. & Mast J. (1991). A new statistic for steady-state evoked potentials. *Electroencephalography and clinical neurophysiology*, 78(5), pp. 378-388.

[185] Vidler M. & Parker D. (2004). Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. *International Journal of Audiology*, 43, pp. 417-429.

[186] Wald A. (1947). *Sequential Analysis*. New York, Wiley.

[187] Wassmer G. (2000). Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical papers*, 41(3), pp. 253-279.

[188] Welch, P. (1967). 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, 70-73. DOI: http://dx.doi.org/10.1109/TAU.1967.1161901

[189] Wilks S. S. (1932). Certain generalisations in the analysis of variance. *Biometrika*, 24, pp. 471-494.

[190] Wong P.K.H & Bickford R.G. (1980). Brain stem auditory evoked potentials: The use of noise estimate. *Electroenceph. clin. Neurophysiol.,* 50(1-2), pp. 25-34.

[191] Zhang R., McAllister G., Scotney B., McClean S. & Houston G. (2006). Combining wavelet analysis and Bayesian networks for the classification of auditory brainstem response. *IEEE. Trans. Inf. Technol. Biomed.*, 10(3), pp. 458-67.

Appendix

A.1 Central Limit Theorem

The Central Limit Theorem (coined by Polya in 1920, and built around the work of Laplace in 1810) states that the sample mean is normally distributed, irrespective of the true underlying distribution from which is was obtained. The only assumptions underlying the Central Limit Theorem are that the samples are independent, and that the sample size is sufficiently large. More formally, the Central Limit Theorem is defined as:

$$\sqrt{N}(\bar{X} - \mu) \xrightarrow{\mathrm{L}} \mathbf{N}(0, \sigma^2) \tag{1}$$

where $\bar{X}$ is the sample mean, μ is the true mean of the underlying population, σ^2 is the true variance of the underlying population, N is the sample size, $\mathbf{N}(0, \sigma^2)$ denotes a normal distribution with mean 0 and variance σ^2, and L denotes Limit (as N goes to ∞).

A.2 The binomial distribution

A Bernoulli trial is a random experiment with exactly two possible outcomes, typically interpreted as 'success' and 'failure'. When X Bernoulli trials are performed and the probability of a successful trial is P, then the binomial distribution gives the probability densities of observing x successful trials. The distribution is given by:

$$B(x|X, P) = \frac{X!}{x!(X-x)!} P^x (1-P)^{X-x} \tag{2}$$

The binomial distribution is used extensively throughout this work when constructing CIs for the nominal α-level of the test. A 'successful' Bernoulli trial is hence defined as

a false-positive, where the probability of a successful trial P is equal to α. The total number of Bernoulli trials X is furthermore given by the total number of independent tests performed. The distribution is then used to generate the probability densities of observing x false-positives.

As an example, say 10 000 independent tests are performed under H_0 at nominal level $\alpha = 0.05$. The expected number of false-positives observed under H_0 is then 500 (5% of 10 000). In practice, deviations from the theoretical 500 false-positive observations occur due to random fluctuations (it is assumed for now that the underlying statistical assumptions of the test are satisfied). The binomial distribution describes the spread of these random fluctuations. An example of a binomial distribution using $X = 10000$ and $P = 0.05$ is presented in Fig. A.1. The two-sided 95% CIs are given by [459, 544], and are also presented. Note that the binomial distribution is a discrete distribution, which means that rounding errors occur when approximating the CIs. Throughout this work, a slightly liberal approach is adopted, i.e. the boundaries are rounded 'inwards', giving slightly too narrow CIs.

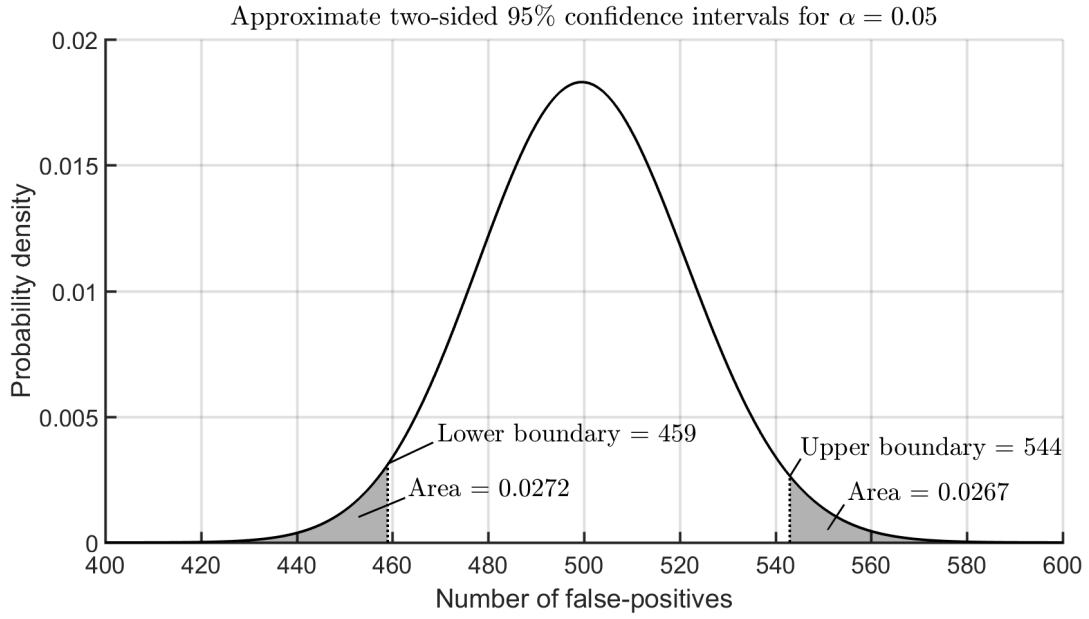


Figure A.2.1: The binomial distribution, representing the expected distribution for the number of false-positives when 10 000 independent tests are performed at nominal significance level $\alpha = 0.05$. The two-sided 95% CIs (for the expected 500 false-positives) are also shown, and are given by [459, 544]. Note that the binomial distribution is a discrete distribution, meaning rounding errors occur when approximating the CIs.

Independent observations

An important underlying assumption of the binomial distribution is that the X observations are independent. In this work, the observations are essentially the p-values generated by the statistical tests (giving a 1 for $p < \alpha$ and a 0 otherwise). The independence assumption hence requires these p-values to be independent. When independence

is violated, then the 'effective number of observations' is smaller than the assumed number of observations. As a result, the estimated CIs will be too narrow.

Many sections throughout this thesis furthermore evaluate specificity using real EEG background activity (data set **D1**). When doing so, one of the following two approaches is typically adopted. In the first approach, each recording is split into ensembles of epochs, using each EEG measurement at most once. The ensembles (and the resulting p-values) can hence be considered more or less independent. In the second approach, blocks of epochs are resampled repeatedly from within the continuous recording. Contrary to the first approach, data can now be used multiple times, potentially resulting in a violation of the independence assumption between ensembles.

A.3 Feature optimisations

This section uses simulations to optimise the sensitivity of various time and frequency domain ABR detection methods. The data for this section consists of simulated coloured noise along with ABR templates (data set **D3**) for simulating a response.

A.3.1 Time domain

Starting with the time domain methods, the goal for this section is to optimise sensitivity in terms of the number of TVMs. The number of TVMs introduces a trade-off between statistical robustness versus a potential loss of information. In particular, when the number of TVMs are too low, then consecutive peaks and valleys within the ABR waveform might cancel each other out. When the number of TVMs is too high, on the other hand, then the TVMs will be highly correlated, resulting in redundant features, and potentially a reduced statistical power. The time domain ABR detection methods included in the optimisation are: (i) T2 Time (the Hotelling's T^2 test applied in the time domain), (ii) T2 RM (the Hotelling's T^2 test when applied in the time domain as a repeated measures approach), (iii) RM ANOVA, and (iv) Friedman's test. When using RM ANOVA, sphericity violations are accounted for by adjusting the DOF of the F-distribution using a correction factor, as described in section A.4

Method

Simulated coloured noise was generated as described in section 4.4, using a band-pass filter of 100-2000 Hz (a 3rd-order Butterworth filter, see also section A.16). A total of 5000 recordings were simulated, which were structured into ensembles of N 30.03 ms, where the ensemble size N was set to 200. A -24 dB response was then simulated as described in section 4.3, using the ABR templates from data set **D3**. The initial 15 ms windows of the epochs were compressed into Q TVMs, where Q ranged from 2 to 75, and the resulting TVMs were analysed (both before and after simulating a response) using the aforementioned ABR detection methods.

Results

The TPRs and FPRs are presented as a function of the number of TVMs in Fig. A.3. For the no-stimulus condition (plot B), the nominal α-level and its two-sided 95% CIs is also plotted. The mean correction factor (calculated across 5000 tests) for RM ANOVA is furthermore presented in Fig. A.4, also as a function of the number of TVMs. Note that a large correction factor implies a small violation to the sphericity assumption, and vice versa for a small correction factor.

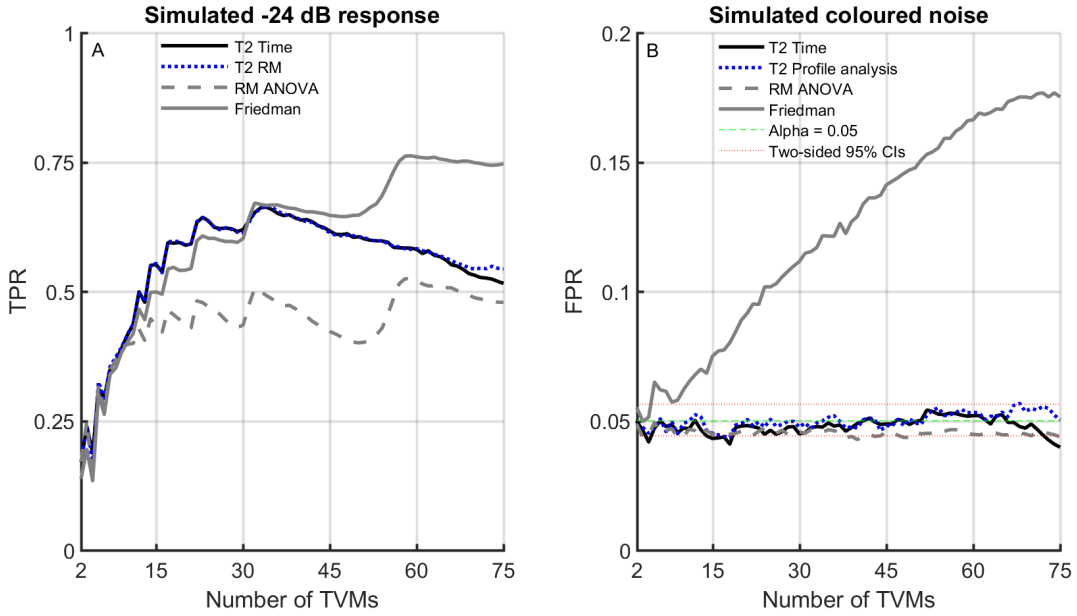


Figure A.3.1: The TPRs (plot A) and FPRs (plot B) as a function of the number of TVMs.

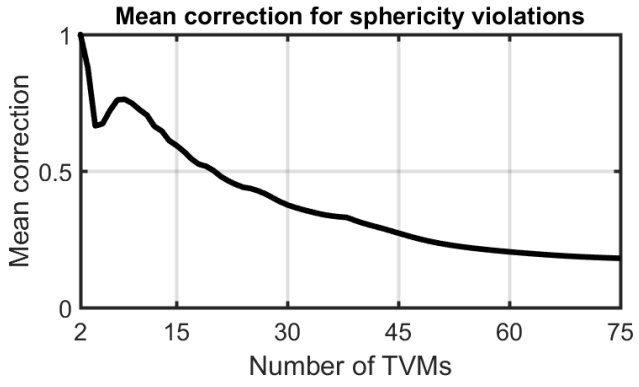


Figure A.3.2: The mean correction factor (calculated across 5000 tests) for RM ANOVA as a function of the number of TVMs. Note that when the correction factor is relatively large (close to one), that the sphericity violation was small.

Discussion

For both T2 Time and T2 RM, the optimal number of TVMs (for this data) was 35. Further increasing the number of TVMs resulted in a loss of test sensitvitiy. For RM ANOVA, the TPR peaked at ∼51% when using 32 TVMs, and again at ∼51-52% when using anything between 57-63 TVMs. For Friedman's test, the specificity was exceptionally poor when using anything more than three TVMs. It worth noting that the

implementation for Friedman's test was verified using the example presented in Friedman's original paper (Friedman, 1937).

A.3.2 Frequency domain

This section explores how many and which spectral bands to include for the frequency domain methods. In particular, it explores which spectral bands contain significant contributions from the ABR, and how large the contribution should be before the spectral band should be included in the analysis. Ideally, the spectral bands selected for the analysis should include only those bands that contain significant contributions from the ABR, whilst excluding others that contain pre-dominantly noise. The optimisation in this section is furthermore restricted to just the Hotelling's T^2 test.

Approach

Simulated coloured noise was generated as described in section 4.4, using a band-pass filter of 100-2000 Hz. A total of 5000 recordings were simulated, which were structured into ensembles of N 30.03 ms, where the ensemble size N was set to 200. A -24 dB response was then simulated as described in section 4.3, using the ABR templates from data set **D3**. The initial 15 ms windows of the epochs were then extended to 25 ms using zero-padding, after which the Hotelling's T^2 test was applied separately to each spectral band (using the real and imaginary parts as features). The detection rate was then calculated per spectral band, after which the spectral bands were ranked from high to low (Table A.1). Finally, the Hotelling's T^2 Test was applied to the top W ranked spectral bands, where W was varied from 2 to 30, i.e. it was first applied to just the top ranking spectral band, then the top 2 ranking spectral bands, etc., until all 30 top ranking spectral bands had been analysed.

Table A.3.1: The top 30 ranked spectral bands, where the ranking was performed as a function of the percentage of detected responses using the Hotelling's T^2 test as detection method

Rank	Frequency	Detection	Rank	Frequency	Detection	Rank	Frequency	Detection
1	120	27.3%	**11**	320	12.5%	**21**	1000	9.6%
2	440	21.8%	**12**	560	12.2%	**22**	840	9.6%
3	480	20.8%	**13**	1040	12.2%	**23**	720	9%
4	520	17.4%	**14**	640	12%	**24**	800	9%
5	200	16.4%	**15**	960	11.22%	**25**	1120	9%
6	240	15.3%	**16**	880	10.9%	**26**	760	8.8%
7	920	15.1%	**17**	600	10.6%	**27**	1320	8%
8	400	14.5%	**18**	1080	10.2%	**28**	1280	7.6%
9	280	13.1%	**19**	360	10%	**29**	1240	7%
10	160	13%	**20**	680	10%	**30**	1360	7%

Results

The detection rate is first plotted as a function of the spectral band being analysed in Fig. A.4, both after (plot A) and before (plot B) simulating a response. For the no-stimulus condition (plot B), the nominal α-level is also plotted along with its two-sided

95% CIs. The detection rates are then also plotted as a function of the number of top ranked spectral bands included in the Hotelling's T^2 test, similarly before (plot D) and after (plot C) simulating a response. Results show that detection peaks (for this data) when using the top ∼24 ranked spectral bands in Table A.1.

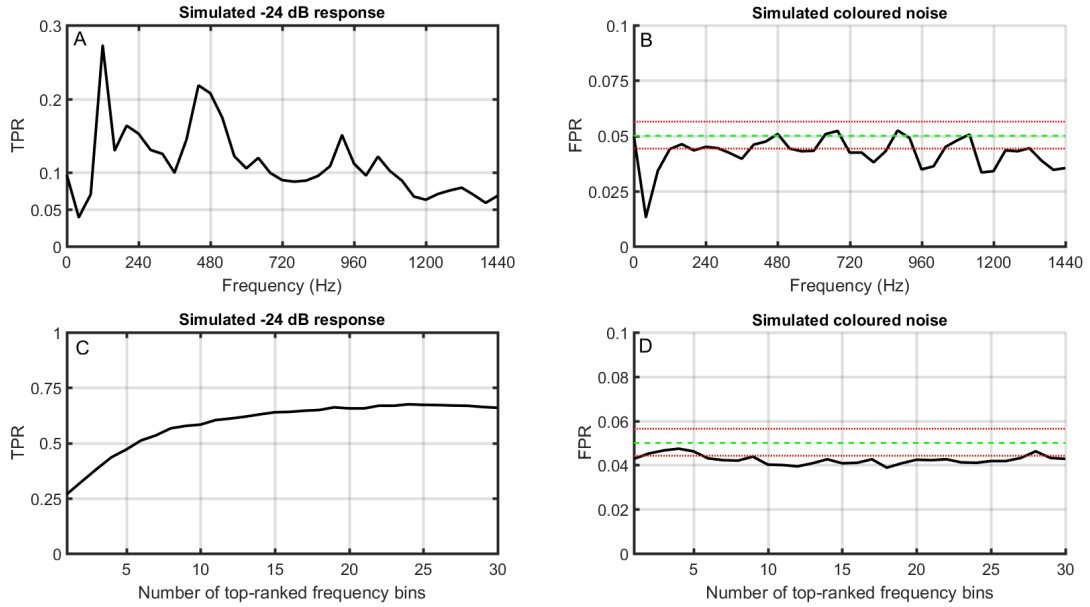


Figure A.3.3: Results from the frequency domain optimisation for the Hotelling's T^2 test. Plots A and B: the detection rates as a function of the spectral band being analysed, both before (plot B) and after (plot A) simulating a response. Plots C and D: the detection rate, now as a function of the number of top-ranked spectral bands (see Table A.1), similarly before (plot D) and after (plot C) simulating a response.

A.4 Corrections for sphericity violations

This section provides a brief description of the Greenhouse Geisser (GG) and Huyn Feldt (HF) corrections for sphericity violations, denoted by $\hat{\epsilon}$ and $\tilde{\epsilon}$, respectively. In particular, sphericity violations are accounted for by multiplying the DOF of the F-distribution with either $\hat{\epsilon}$ or $\tilde{\epsilon}$. The GG correction has previously been found to be conservative for $\hat{\epsilon} < 0.75$ (Huyn & Feldt, 1976), whereas the HF correction $\tilde{\epsilon}$ is slightly liberal. Girden (1992) therefore recommends using the GG method for $\hat{\epsilon} > 0.75$, and the HF method otherwise, which is the adopted approach throughout this work.

A.4.1 The Greenhouse Geiser correction

Following the original notation in Greenhouse & Geiser (1959), the GG correction $\hat{\epsilon}$ is given by:

$$\hat{\epsilon} = \frac{p^2(\bar{\sigma_{tt}} - \sigma_{..})^2}{(p-1)\sum\sum\sigma_{ts}^2 - 2p\sum\bar{\sigma_t}^2 + p^2\bar{\sigma_{..}^2}} \tag{3}$$

where p is the number of within subjects levels (the number of TVMs), σ_{ts} are the elements of the feature covariance matrix $\boldsymbol{S}$, $\hat{\sigma_{tt}}$ is the mean of the diagonal of $\boldsymbol{S}$, $\sigma_{..}$ is the overall mean of $\boldsymbol{S}$, and $\hat{\sigma_t}$ is the mean of row (or column) t of $\boldsymbol{S}$.

A.4.2 The Huyn Feldt correction

Following the notation of Huyn & Feldt (1976), the HF correction $\tilde{\epsilon}$ is found by modifying the Greenhouse Geiser correction $\hat{\epsilon}$, and is given by:

$$\tilde{\epsilon} = \frac{Nr\hat{\epsilon} - 2}{r(N - g - (r\hat{\epsilon}))} \tag{4}$$

where N is the total number of sampling units, r is the number of within subjects factors (the number of TVMs), and g is the number of levels of the between subjects factor (equal to 1 for evoked response detection).

A.5 Assumptions underlying the bootstrap and the permutation test

The goal for this section is to provide a more in depth assessment of the performance of the bootstrap approach for ABR detection. In particular, a brief assessment of the reliability of bootstrapped CIs is first provided in section A.5.1 below, after which the assumption that the ABR cancels out in the resampled data sets (i.e. that its SNR is zero) is explored in section A.5.2. A minor variation to the standard approach (where the ensemble coherent average is subtracted from the epochs prior to resampling) is also explored. Finally, a very brief assessment of the independence assumption between resampled epochs is provided in section A.5.3.

A.5.1 The reliability of bootstrapped confidence intervals

This section briefly explores the reliability (the consistency or repeatability) of bootstrapped critical decision boundaries, as a function of the number of resampled ensembles M. Note that the expected type-I error rate for unreliable CIs will still be α. It is instead the robustness (in terms of test performance) of the ABR detection method that is affected. In particular, unreliable critical boundaries can contribute towards either a conservative or a liberal evaluation of test performance for some subjects. The reliability of bootstrapped CIs should hence ideally be as high as possible. Reliability is furthermore directly related to the number of resampled datasets M. Increasing M will increase the reliability of the CIs, which comes at the cost of an increased processing time. The goal for this section is to explore how the reliability of bootstrapped CIs is affected by M. Finally, the reader might have noticed that 'CIs' and 'critical decision boundaries' are being used interchangeably, which is justified for this section as the CIs are one-sided.

Method

The data for the assessment, say D, consists of a single ensemble containing $N = 200$ 30.03 ms epochs. The epochs are composed of simulated coloured noise, generated as described in section 4.4 (using a 3rd-order Butterworth band-pass filter of 100-2000 Hz). The first step for the assessment is to approximate the true *bootstrapped* null distribution (for the T^2 statistic). The latter is achieved by resampling $M = 50\,000$ ensembles from D, and calculating the T^2 statistic from each resampled ensemble. The histogram of the resulting T^2 values (which is assumed to be the true bootstrapped null distribution due to the large number of resampled data sets) is shown in Fig. A.5 (upper plot), along with its 95% percentile (equal to 44.6419). Next, 200 critical values ($\alpha = 0.05$) for the T^2 statistic were calculated using M resampled data sets, where M took values of 500, 1000, 2500, 5000, 75000, or 10 000. To clarify, when using e.g. $M = 500$, a total of 500 ensembles of $N = 200$ epochs were resampled from D, which were analysed using the T^2 test. A histogram was then constructed from the 500 T^2 values, which was used to find the 95% percentile for the T^2 statistic. This procedure was repeated 200 times, resulting in a distribution of critical values (for the $M = 500$ condition), which was similarly repeated for each M.

Results

The resulting histograms (each constructed from 200 critical values) are shown in Fig. A.5 for different M.

Discussion

The bootstrap controls the type-I error rate when evaluated across a large number of tests, but can still result in unreliable CIs when the number of resampled data sets M is too low. In particular, when using $M = 500$, the smallest estimated 95% CI was given by 41.35 (note that the assumed true CI was 44.6419), corresponding to the 0.9148 percentile of the (assumed) true bootstrapped null distribution. The largest estimated

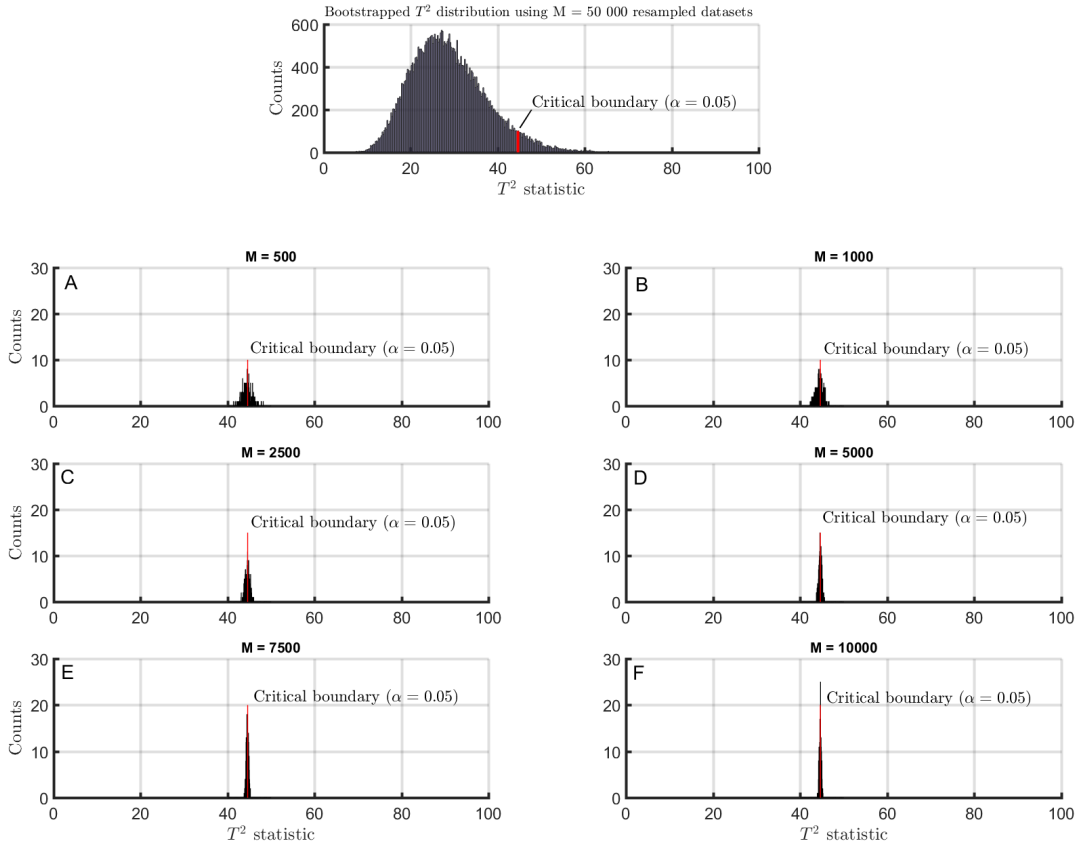


Figure A.5.1: The histograms (each constructed from 200 critical values) for different M. The assumed true *bootstrapped* null distribution for the T^2 statistic is also shown (upper plot). The latter was generated using $M = 50\ 000$ resampled data sets.

95% CI was furthermore equal to 48.1772, corresponding to the 0.9741 percentile. Hence, in the very unlikely scenario that all bootstrapped CIs are underestimated (<41.35), then the expected type-I error rate will be <0.0259, whereas when all bootstrapped CIs are overestimated (>48.1772) it will be >0.0852. The upper and lower boundaries for the expected spread of the type-I error rate for a worst case scenario (using $M = 500$) is hence given by [0.0259, 0.0852]. Increasing M reduces the spread, i.e. using $M = 1000$ gives [0.0345, 0.0736], $M = 2500$ gives [0.0387, 0.0645], $M = 5000$ gives [0.0421, 0.0581], $M = 7500$ gives [0.0449, 0.0571], and $M = 10000$ gives [0.0449, 0.0559]. An exceptionally reliable bootstrapped CI would hence require a relatively large number of resampled datasets. When using $\alpha = 0.01$, then M would ideally be even larger due to the sparseness of the outer tails of the bootstrapped null distributions.

A.5.2 Subtracting the coherent average prior to resampling

An important assumption underlying Lv et al (2007) is that the ABR cancels out in the resampled data sets, i.e. that its SNR is zero. When this is not the case, then the bootstrapped ensembles will still contain a small response, meaning the resulting bootstrapped null distribution will be slightly biased towards the alternative distribution.

As a result, the critical boundary for rejecting H_0 will be increased, and test sensitivity reduced. This section explores the simple solution of subtracting the ensemble coherent average from the epochs prior to resampling. Data for this section consists of the recordings of EEG background activity (data set **D1**), along with ABR templates for simulating a response (data set **D3**).

Method

The recordings of EEG background activity (data set **D1**) were structured into ensembles of $N = 200$ 30.03 ms epochs, resulting in a total of 5448 ensembles (note that these ensembles can be considered more or less independent). A -27 dB response was then simulated as described in section 4.3 using ABR templates from data set **D3**. The initial 15 ms windows of the epochs were analysed with the Hotelling's T^2 test (applied in the time domain) both before and after simulating a response. The significance of the T^2 statistic was then evaluated using either (i) theoretical F-distributions (the standard approach), (ii) the bootstrap approach in Lv et al (2007, or (iii) the bootstrap approach in Lv et al where the ensemble CA was subtracted from the epochs prior to resampling.

Results

The FPRs and TPRs (using $\alpha = 0.01$) are first presented in Table A.2. The two-sided 99% confidence interval for $\alpha = 0.01$ is furthermore given by [0.007, 0.0138] (5448 tests were performed). The TPRs are then also plotted as a function of the theoretical α-level in Fig 11.6, which is essentially a (modified) Receiver Operating Characteristic (ROC) curve. Note that this ROC curve deviate from a standard ROC curve as it shows the TPR as a function of the theoretical α-level as opposed to the observed type-I error rate.

Table A.5.1: The FPRs and TPRs (using $\alpha = 0.01$) when evaluating the test significance of the T^2 statistic using either theoretical F-distributions, with the standard bootstrap approach in Lv et al, or with the standard bootstrap approach when subtracting the ensemble CA from the epochs prior to resampling. The two-sided 99% confidence interval for the theoretical 0.01 FPR is furthermore given by [0.007, 0.0138] (5448 tests were performed).

	FPR ($\alpha = 0.01$)	**TPR** ($\alpha = 0.01$)
T2 Time (F-distributions)	0.009	0.2150
T2 Time (bootstrapped)	0.0099	0.2034
T2 Time (bootstrapped, CA subtracted)	0.0101	0.2144

Discussion

Results from Fig. A.6 suggest a small increase in sensitivity by subtracting the CA from the epochs prior to resampling, which suggests that the power of the evoked response in the resampled data sets is not zero. Although results suggest that the benefit is relatively small, additional simulations (details not presented) demonstrate that when the inter-epoch intervals are decreased (i.e. the stimulus rate is increased) or when the SNR of the response is increased, that the benefit of subtracting the CA prior to resampling is larger, e.g. an increase in TPR from ∼0.3 to ∼0.36 was observed when using a stimulus rate of ∼60 Hz.

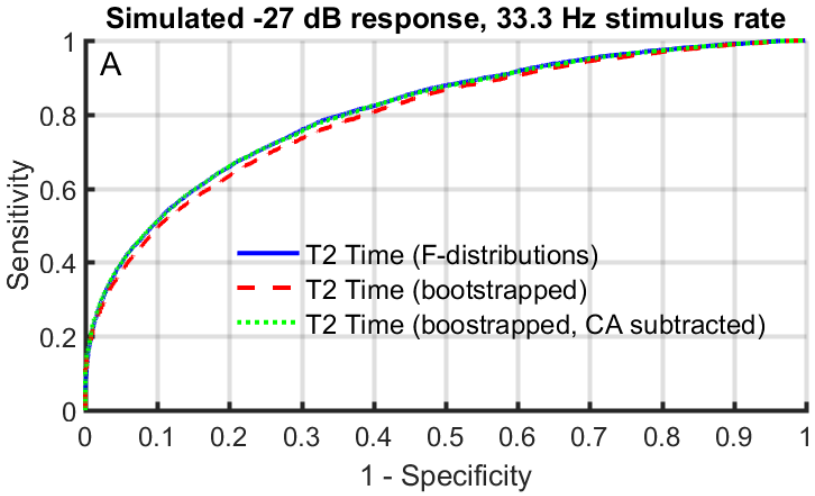


Figure A.5.2: The TPRs plotted as a function of the theoretical α-level of the test when evaluating the significance of the T^2 statistic using either (1) theoretical F-distributions, (2) the bootstrap approach in Lv et al (2007), or (3) the bootstrap in Lv et al (2007) when subtracting the ensemble CA from the epochs prior to resampling.

A.5.3 Independence violations

The goal for this section is to provide a more powerful assessment of the specificity of the Hotelling's T^2 test when evaluating test significance using (i) theoretical F-distributions, (ii) the standard bootstrap in Lv et al (2007), and (iii) the standard bootstrap where the ensemble CA is subtracted from the epochs prior to resampling. The assessment is conducted using 175 000 simulated tests, using a high-pass cut-off frequency of 100 Hz and a (hypothetical) stimulus rate of 33.3 Hz.

The sub-goals for this section are furthermore three-fold. First, various sections throughout this work suggest a very minor tendency towards a conservative test performance when using a high-pass filter of 100 Hz and hypothetical stimulus rate of 33.3 Hz. The large number of simulated tests used in this section is used to either verify or rule out the possibility that this is due to a violation of the independence assumption between epochs. Secondly, the resampling with replacement procedure used by the bootstrap means that some EEG segments might be used multiple times, potentially resulting in an additional violation of the independence assumption between (resampled) epochs, and ultimately in an inaccurate approximation of the null distribution for the bootstrapped test statistic. Because the independence violation is expected to be relatively minor, a large number of tests may be required to expose it. Finally, subtracting the ensemble coherent average from the epochs prior to resampling may result in a reduced random variation for the resampled data sets (relative to the original data set), again potentially resulting in an inaccurate estimation of the null distribution.

Method

Ensembles of coloured noise were simulated, as described in section 4.4 (using a band-

pass filter of 100-2000 Hz). A total of 175 000 recordings were simulated, which were structured into ensembles of $N = 200$ 30.03 ms epochs. The initial 15 ms of the ensembles were then analysed using the Hotelling's T^2 test (applied to 25 TVMs). The test significance of the T^2 statistic was evaluated using one of the three aforementioned methods, i.e. theoretical F-distributions, the standard bootstrap, or the bootstrap after subtracting the ensemble CA from the epochs prior to resampling.

Results

The observed FPRs using either $\alpha = 0.01$ or $\alpha = 0.05$ are presented in Table A.3. The 95% CIs for $\alpha = 0.01$ were given by [0.0095, 0.0105], and for $\alpha = 0.05$ by [0.0490, 0.0510]. No significant deviations were observed.

Table A.5.2: The observed FPRs (calculated from 175 000 simulated tests) using either $\alpha = 0.01$ or $\alpha = 0.05$. The 95% CIs for $\alpha = 0.01$ are given by [0.0095, 0.0105], and for $\alpha = 0.05$ by [0.0490, 0.0510].

	$\alpha = 0.01$	$\alpha = 0.05$
T2 Time (F-distributions)	0.0098	0.0496
T2 Time (bootstrapped)	0.0103	0.0497
T2 Time (bootstrapped, CA subtracted)	0.0102	0.0493

Discussion

Results for 'T2 Time (F-distributions)' firstly suggests that independence between epochs was indeed satisfied when using a high-pass cut-off frequency of 100 Hz and a hypothetical stimulus rate of 33.3 Hz. Results from 'T2 Time (bootstrapped)' suggest that the independence violation between resampled epochs is also negligible for these settings. Note however that the latter might not generalise to alternative stimulus rates. In particular, higher stimulus rates will reduce the inter-epoch intervals, resulting in an increased probability of an overlap in data, and thus in increased independence violation within the resampled data sets. Finally, results from 'T2 Time (bootstrapped, CA subtracted)' suggest that random variation is not significantly decreased (and that it otherwise has no noticeable effect on the approximated null distribution) by subtracting the ensemble CA from the epochs prior to resampling.

A.5.4 The permutation test

A potential competitor to the bootstrap is the permutation test, which dates back to R.A. Fisher (1935), who introduced it as a theoretical argument in support for Students *t*-test (Efron & Tibshirani, p 202, 1993). It is similar to the bootstrap in that it attempts to construct a reference distribution for the parameter of interest, which can then be used for statistical inference. It has also been used for evoked response detection by Maris & Oostenveld (2007), who used it to evaluate clusters of t-statistics, also known as the cluster mass test (Bullmore et al., 1999).

The permutation test differs from the bootstrap in that (i) it applies a resampling

without replacement approach, and (ii) it requires two independent samples, say $X1$ and $X2$, as opposed to just a single sample. The permutation test may also require a minor modification to the test statistic. Ideally, a two sample test should be used to compare $X1$ and $X2$. Alternatively, a one-sample test can be applied to both $X1$ and $X2$ separately, and the difference used as test statistic. The goal for the permutation test is then to evaluate the null hypothesis H_0, that $X1$ and $X2$ share the same underlying distribution.

In practice, the permutation test can be implemented using a few simple steps (Maris & Oostenveld, 2007): (1) $X1$ and $X2$ are pooled to construct a single pooled sample space, (2) the pooled sampled space is randomly split amongst two new samples, which is repeated many times, and (3) the statistic of interest is calculated from the resampled data sets. Like this bootstrap, this allows a reference distribution to be constructed, which can then be used to construct e.g. confidence intervals.

The main assumption underlying the permutation test is the equal probability of observing any subset from the pooled sample space of $X1$ and $X2$ under H_0. This assumption is somewhat questionable for evoked response detection. In particular, when one of the samples (say $X1$) contains a response and the other (say $X2$) does not, then randomly dividing $X1$ and $X2$ into two new samples will inevitably result in an unequal distribution of $X1$ and $X2$ values for some of the resampled data sets. As a result, some resampled data sets will contain a relatively large response, whereas others will contain a relatively small response. Note that the resampled data sets are then representative (to some degree) of the alternative hypothesis, which states that the resampled data sets are obtained under different conditions. Note also that the core of the issue is that the permutation test does not disrupt the time-locking between the epochs and the stimuli. A potential solution might therefore be to re-sample from within the continuous recordings. It is, however, not clear how this might be achieved using a resampling *without* replacement approach. Note also that if the permutation test is modified so that it uses a resampling *with* with replacement approach (and that it resamples from within the continuous recording), that it is essentially identical to the bootstrap in Lv et al (2007), except that it still requires two independent samples as opposed to one. Finally, note that obtaining two independent samples under similar test conditions is problematic for evoked response detection due to non-stationary data and varying DOF between recordings.

A.6 Detection rates using adjusted α-levels

This section presents the detection rates (from simulations in sections 6.1-6.3) when using adjusted critical α-levels, i.e. the critical α-level (for a significant detection) was adjusted, such that the FPRs were equal across methods, which thus allows a more fair comparison in test sensitivity.

Simulations I

For the first set of simulations (Simulations I, section 6.1), the critical α-levels were adjusted, per method, such that the FPR (across all ensemble sizes) was 0.01. The adjusted α-levels were 0.0087 (T2 Time), 0.0088 (T2 Freq), 0.021 (Fsp 5 dof), 0.0321 (Fmp 5 dof), 0.0071 (Fsp bootstrapped), 0.0074 (Fmp bootstrapped), 0.0088 (modified q-sample V2) and 0.009 (Modified q-sample V4). The detection rates using the adjusted α-levels are presented in Fig A.7.

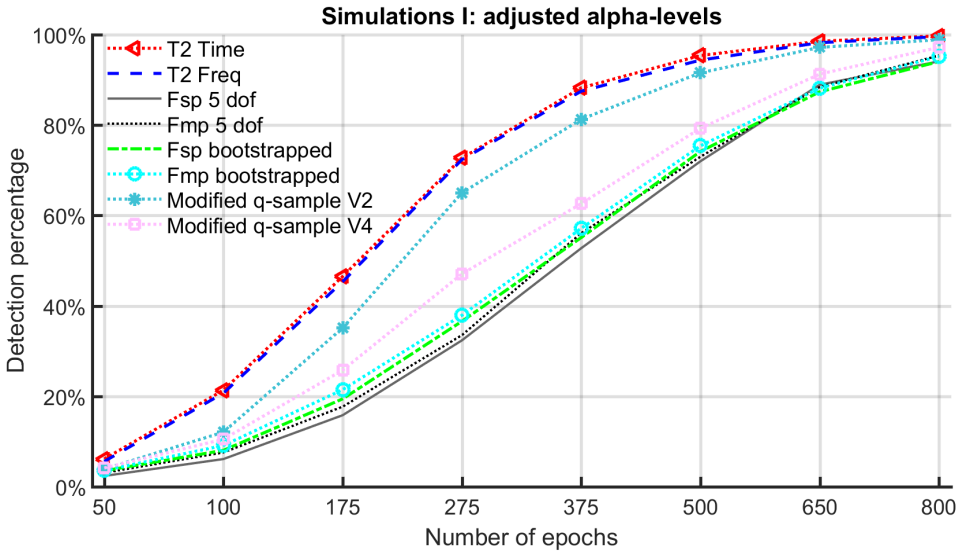


Figure A.6.1: The detection rates for the methods from Simulations I (section 6.1) when using the adjusted α-levels.

Simulations II

For the second set of simulations (section 6.2), the critical α-levels were also adjusted, now per ensemble size. The adjusted α-levels (for obtaining FPRs of 0.01) are presented in Table A.4, and the detection rates (using the adjusted α-levels) in Fig. A.8.

Table A.6.1: The required α-levels, per ensemble size N, for obtaining a FPR of 0.01 in simulations presented in section 6.2.

Ensemble size –>	**50**	**100**	**175**	**275**	**375**	**500**	**650**	**800**
T2 Time (F-distributions)	0.0087	0.0112	0.0110	0.0125	0.0117	0.0104	0.0101	0.0089
T2 Time (bootstrapped)	0.0091	0.0111	0.0091	0.0101	0.0101	0.0091	0.0111	0.0080
CC (bootstrapped)	0.0071	0.0071	0.0071	0.0080	0.0061	0.0071	0.0080	0.0101
T2 Time & CC (bootstrapped)	0.0091	0.0091	0.0071	0.0080	0.0071	0.0071	0.0071	0.0080
Fsp (bootstrapped)	0.0080	0.0080	0.0080	0.0091	0.0091	0.0091	0.0091	0.0080
MP (bootstrapped)	0.0111	0.0101	0.0111	0.0071	0.0091	0.0091	0.0101	0.0080
Fsp (F-distributions)	0.0315	0.0367	0.0462	0.0429	0.0447	0.0524	0.0406	0.0447

A.7 Comparisons in sensitivity: additional simulations

This section uses simulations to compare the sensitivity of (i) the Hotelling's T^2 test, applied in both the time and frequency domain, (ii) RM ANOVA, using the GG and HF corrections as compensation for sphericity violations (see section A.4), and (iii)

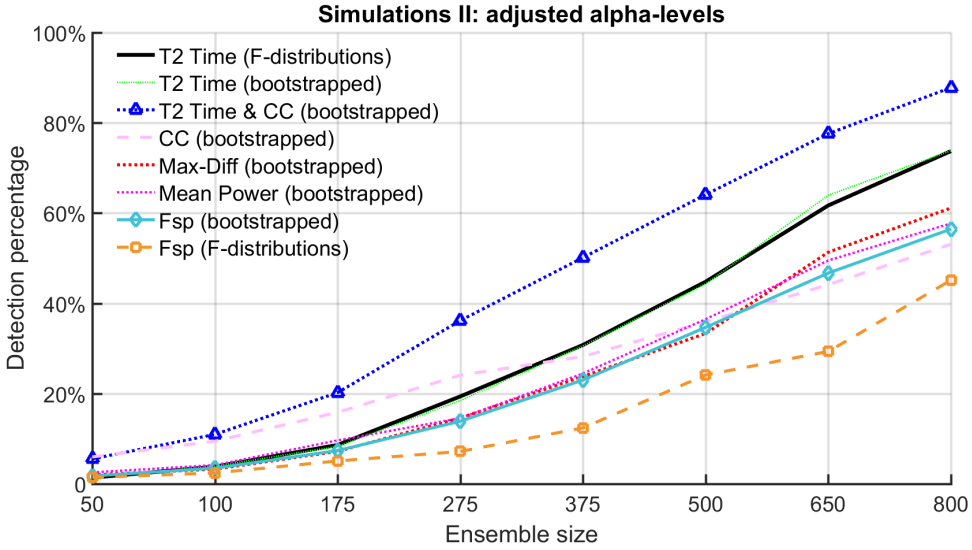


Figure A.6.2: The detection rates for the methods from Simulations II (section 6.2) when using the adjusted α-levels.

Friedman's test. When used in the time domain, the Hotelling's T^2 test is applied as either a repeated measures approach (denoted by T2 RM; see also section 3.2.4), or as the standard time domain approach (T2 Time). The statistical features used for this section were chosen based on feature optimisations presented in the section A.3, and can be considered close to optimal in this section. As was the case for section 6.1, the main goal for these simulations is to provide a powerful comparison between the sensitivities of the methods.

Method

Data for the simulations is similar to the data used in 'Simulations I' (section 6.1), and consists of recordings of real EEG background activity (data set **D1**), along with click-evoked ABR templates (data set **D3**) for simulating a response. The recordings of EEG background activity were downsampled to 5 kHz and band-pass filtered (using a 3rd-order Butterworth filter) from 100 to 2000 Hz.

Specificity assessment

Ensembles of N epochs were constructed by randomly resampling N consecutive 30.03 ms epochs from within a randomly selected and pre-processed recording of EEG background activity. The ensemble size N took values of 50, 100, 175, 275, 375, 500, 650, 800 epochs. A total of 5000 ensembles containing just EEG background noise were constructed, per ensemble size. The initial 15 ms windows of the ensembles were analysed using the aforementioned ABR detection methods.

Sensitivity assessment

A response was simulated by randomly selecting an ABR template (from data set **D3**), rescaling it, and adding it to all epochs within the ensemble in question. The scaling

factor was chosen such that the SNR of the response was -24 dB, calculated as described in section 4.3. The initial 15 ms windows of the resulting ensembles were analysed using the aforementioned detection methods.

Statistical features

The statistical features selected for the analysis were chosen based on results presented in section A.3. The time domain features consist of 35 TVMs for both T2 Time and T2 RM, and 32 TVMs for RM ANOVA. For Friedman's test, just 3 TVMs were used, which was based on results from the specificity assessment (section A.3), which show a significantly liberal test performance for Friedman's test when using anything more than three TVMs. Frequency domain features for T2 Freq consist of the real and imaginary parts of the Fourier components of the top 18 ranked spectral bands presented in Table A.1.

Results

Specificity

The observed FPRs (using $\alpha = 0.01$ or $\alpha = 0.05$) for each ensemble size N are presented in Table A.5. The binomial distribution was used to construct two-sided 95% CIs, giving [0.0076, 0.013] for $\alpha = 0.01$, and [0.0442, 0.0564] for $\alpha = 0.05$. Note however that the re-sampling with replacement procedure may have resulted in some segments being selected multiple times, resulting in an independence violation between ensembles (underlying the binomial distribution), giving too narrow CIs (see also section A.2). Significant deviations from the nominal α-levels are nevertheless denoted in Table A.5 by red and blue asterisks, indicating a liberal and conservative test performance respectively.

Sensitivity

The percentage of detected responses are presented in Fig. A.9 as a function of the ensemble size N. Note that the performances of T2 Time, T2 Freq, and T2 RM is more or less identical, and cannot easily be distinguished from each other through visual inspection.

Adjusted critical α-levels

The critical α-levels were adjusted, per method, such that the FPR was 0.01, per ensemble size. The adjusted critical α-levels are presented in Table A.6, and the detection rates (using the adjusted α-levels) are presented in Fig. A.10.

Discussion

A brief discussion on the results from the specificity and sensitivity assessment follows.

Table A.7.1: The FPRs of the methods (using either $\alpha = 0.01$ or $\alpha = 0.05$) for the no-stimulus condition, per ensemble size N. Significantly (**p<0.05**) conservative and liberal test performances are indicated by blue and red asterisks respectively.

Alpha = 0.01								
Ensemble size –>	**50**	**100**	**175**	**275**	**375**	**500**	**650**	**800**
T2 Time	0.0074*	0.009	0.0086	0.0096	0.0072*	0.0072*	0.0074*	0.0112
T2 Freq	0.0068*	0.0102	0.0096	0.0074*	0.0084	0.0086	0.007*	0.0072*
T2 RM	0.0074*	0.009	0.0094	0.0094	0.0064*	0.0076	0.0078	0.0106
RM ANOVA	0.0068*	0.0088	0.0104	0.0086	0.0104	0.012	0.0118	0.0132*
Friedman	0.0082	0.0084	0.0102	0.0108	0.0094	0.0108	0.0102	0.0132*
Alpha = 0.05								
Ensemble size –>	**50**	**100**	**175**	**275**	**500**	**650**	**800**	
T2 Time	0.0442	0.0468	0.0436*	0.0508	0.0444	0.0528	0.0554	0.0512
T2 Freq	0.0488	0.0476	0.0468	0.0472	0.0468	0.0492	0.053	0.054
T2 RM	0.0406*	0.0458	0.0442	0.0512	0.0454	0.0524	0.0522	0.0516
RM ANOVA	0.0356*	0.0422*	0.0408*	0.045	0.0514	0.0532	0.0562	0.0604*
Friedman	0.0474	0.0582*	0.058	0.0516	0.0512	0.052	0.0634*	0.0614*

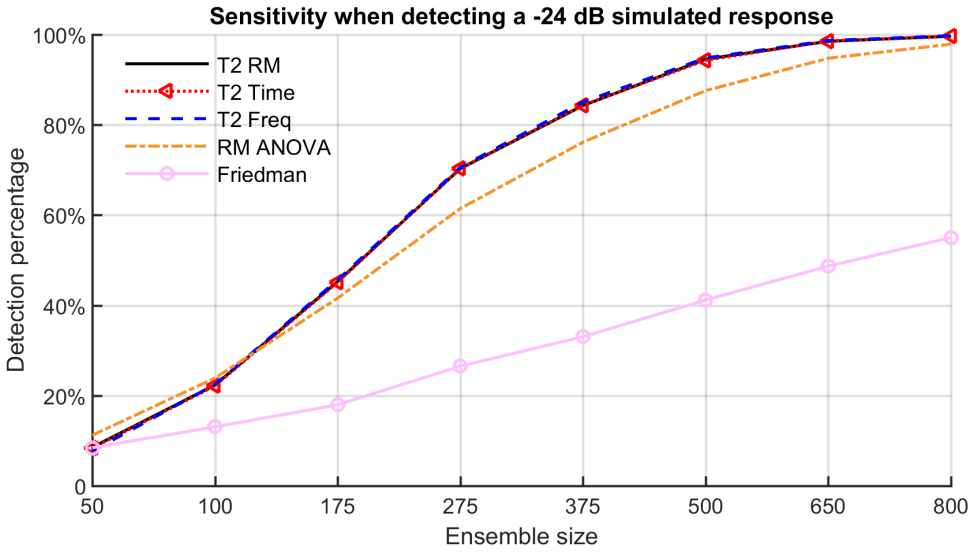


Figure A.7.1: The percentage of detected responses as a function of the ensemble size N when simulating a -24 dB response. Note that the performances of 'T2 Time', 'T2 RM', and 'T2 Freq' are all very similar, and may be difficult to distinguish from each other.

Table A.7.2: The adjusted α-levels for obtaining FPRs of exactly 0.01.

Ensemble size –>	**50**	**100**	**175**	**275**	**375**	**500**	**650**	**800**
T2 Time	0.0131	0.0122	0.0113	0.0115	0.0133	0.0134	0.0130	0.0085
T2 Freq	0.0146	0.0093	0.0102	0.0132	0.0110	0.0113	0.0130	0.0122
T2 RM	0.0126	0.0122	0.0093	0.0090	0.0106	0.0093	0.0099	0.0075
RM ANOVA	0.0141	0.0128	0.0097	0.0123	0.0092	0.0082	0.0081	0.0079
Friedman	0.0148	0.0110	0.0112	0.0106	0.0129	0.0121	0.0115	0.0092

Sensitivity

In terms of test sensitivity, RM ANOVA came out on top for the small ensemble sizes ($N = 50$ and $N = 100$), but was outperformed by the Hotelling's T^2 test for larger

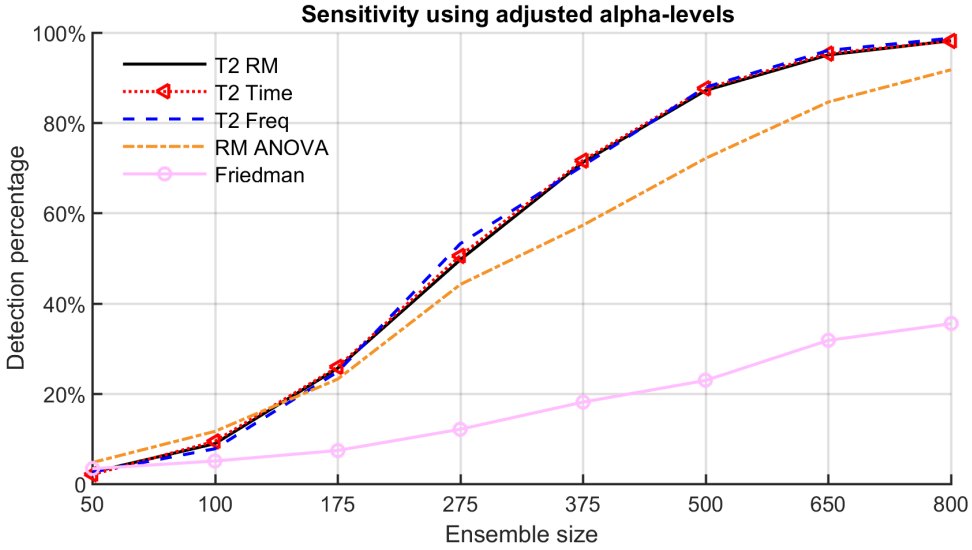


Figure A.7.2: Sensitivity when detecting a -24 dB simulated response using adjusted α-levels.

ensemble sizes ($N > 100$). This might be attributed to the sphericity assumption, i.e. sphericity violations may have become more robust as the ensemble size was increased, resulting in larger corrections to the DOF, and a loss of statistical power.

With respect to the Hotelling's T^2 test, the performance of all three test statistics (T2 Time, T2 Freq, and T2 RM) was very similar, suggesting that time and frequency domain analysis are more or less identical when using optimal feature sets. The very small advantage for T2 Time over T2 Freq observed in section 6.1 (Fig. 6.1) can hence be attributed to a sub-optimal feature set for T2 Freq. With respect to T2 Time and T2 RM, note that, at least in theory, T2 Time should be the more sensitive approach, as it can detect any voltage offset from 0, whereas T2 RM can only detect changes in voltage over time. A potential advantage for T2 RM, on the other hand, is that it may have superior specificity when the mean EEG background activity is not zero. Results presented in this section nevertheless suggest that their performance is more or less identical.

Finally, with respect to Friedman's test, sensitivity was relatively poor, which was to be expected when using just 3 TVMs as features (consecutive peaks and valleys in the ABR waveform would have cancelled out, to some extent). The choice for three TVMs was nevertheless necessary, as anything more than three resulted in a significantly liberal test performance (see section A.3).

Specificity

Results from the specificity assessment now suggest a minor tendency towards a conservative test performance (as opposed to the liberal performance observed in section 6.1). Various factors contributing towards significantly liberal or conservative test per-

formances were previously discussed in sections 6.1 and 6.2, and also apply here. With respect to RM ANOVA, results (Table A.5) suggest that the FPR is somewhat dependent on the ensemble size N, i.e. the FPR tends to increase with N. The latter might be due to random variation. Alternatively, violations to the sphericity assumption might have became more pronounced as the ensemble size was increased, meaning the correction for the sphericity violation (using the GG or HF methods) would be larger. Note also that the GG and HF corrections are approximate, and might vary as a function of the test condition (e.g. the ensemble size). A more robust test performance for RM ANOVA might therefore be obtained by evaluating test significance with the bootstrap, as opposed to using theoretical distributions. Finally, with respect to Friedman's test, results similarly suggest a (weak) correlation between the FPR and the ensemble size N (Table A.5). As was the case with RM ANOVA, specificity might be improved using the bootstrap approach, as opposed to using theoretical distributions.

A.8 Pre-determined thresholds from no-stimulus data

In the literature, statistical inference has been performed using thresholds calculated from recordings of EEG background activity (Stürzebecher et al., 1996; Stúrzebecher et al., 1999; Cebulla et al (2000); Cebulla et al., 2006). The caveat associated with this approach is that the critical boundaries may not generalise well across recordings with varying DOF. Moreover, the spread of the true critical decision boundaries (per recording) might be large relative to the true *mean* critical decision boundary (obtained across recordings), which is detrimental towards the consistency or robustness of the performance of the ABR detection method. This section briefly evaluates the reliability of pre-determined thresholds calculated from no-stimulus data.

Method

The recordings of EEG background activity (data set **D1**) were downsampled to 5 kHz, band-pass filtered from 100-2000 Hz, and decomposed into ensembles of $N = 500$ 30 ms epochs. Each ensemble was then split in two: the first half consists of the initial 0-15 ms windows of the epochs, whereas the second half consists of the 15-30 ms windows. Various statistical tests were applied to the resulting ensembles, giving a population or histogram of values per method. The resulting histograms were then used to construct 95% or 99% critical thresholds for rejecting H_0.

Results

The 95% or 99% critical thresholds are presented in Table A.6 for either the 0-15 ms segments (Set 1) or the 15-30 ms segments (Set 2).

Discussion

Results suggest that the reliability of the thresholds for the rank-based methods (Modified q-sample V2 and the original q-sample test) was not too bad, i.e. thresholds differed

Table A.8.1: The 95% or 99% coverage intervals calculated from two sets of no-stimulus data for various detection methods.

	99% interval		**95% interval**	
Methods	**Set 1**	**Set 2**	**Set 1**	**Set 2**
Fsp	2.37	2.53	1.81	1.78
Original q-sample	51.24	49.59	41.1	41.57
Modified q-sample V2	79.27	85.355	29.59	27.77
Modified q-sample V4	79.99	95.53	30.55	27.89

by ∼1-3%. The reliability for the Fsp and the modified q-sample V4 test was however relatively poor, i.e. thresholds differed by ∼6-16%. Using a larger α level of 0.05 reduced the variability (relative to $\alpha = 0.01$), which is due to the sparseness of the tails of the histograms. More consistent thresholds might therefore be obtained if a larger database were to be used. It should however also be noted that the reliability of pre-determined thresholds was most likely overestimated here. In particular, data set 1 was obtained under identical test conditions (with more or less identical DOF) as data set two (set one was obtained from the 0-15 ms windows whereas set two was obtained from the adjacent 15-30 ms windows). In a more realistic scenario, the thresholds might vary more. Finally, note that this analysis was based on just two critical values. A more robust evaluation would require a larger number of observations before an accurate estimate of the spread of pre-determined thresholds can be established.

A.9 Replicated simulations from Stúrzebecher et al (1999) & Cebulla et al (2000a)

This section describes additional simulations, which were included to explore the sensitivities of various frequency domain methods when detecting simulated ABRs in Gaussian White Noise. The methods included in the analysis are the Hotelling's T^2 test, the original q-sample uniform scores test, and both the Modified q-sample V2 and V4 tests. The simulations follow the same design as described in Stúrzebecher et al (1999) and Cebulla et al (2000a).

Method

As described in Stúrzebecher et al (1999) & Cebulla et al (2000a), data for the simulations consists of zero mean Gaussian random variables with a variance of one, where each pair of random variables represents the real and imaginary parts of some spectral band. Ensembles of 50 'epochs', represented by the real and imaginary parts of W spectral bands, were thus constructed. The signal to detect was furthermore a sine wave,

multiplied by a Gaus curve, which is defined (in the time domain) as:

$$S(x) = Asin\left(\frac{2\pi.x}{L}\right).e^{-a^2\left(x-\frac{L}{2}\right)^2} \tag{5}$$

where A was set to 0.25, a to 0.1, and L is the length of the signal (128 samples). The signal S was then transformed to the frequency domain with the FFT. The real and imaginary parts of 14 spectral bands containing the largest portion of the response were then arbitrarily added to half of the epochs within each ensemble. A total of 5000 ensembles of 50 epochs were thus constructed, which were analysed using the aforementioned detection methods.

Results

The ROC curve of each method is plotted in Fig. A.10. Note that the x-axis shows the theoretical FPR, as opposed to the empirical FPR. The latter is justified, as all statistical assumptions underlying H_0 are satisfied for Gaussian zero-mean white noise. Results show an almost identical performance between T2 Freq and the Modified q-sample V4 test, with a small advantage for the Modified q-sample V4 when α was smaller than 0.08, and vice versa when α was larger than 0.08. In second place came the Modified q-sample V2 test, followed by the original q-sample test, as predicted by Stúrzebecher et al (1999).

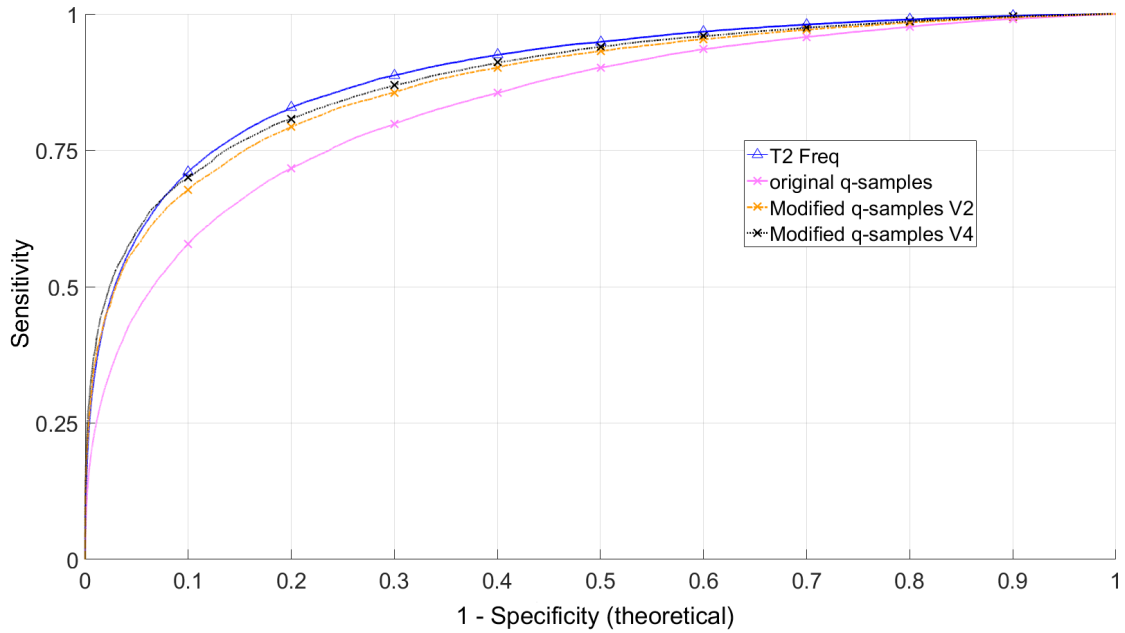


Figure A.9.1: The ROC curves for various frequency domain detection methods when detecting a sine wave multiplied by a Gaus curve in Gaussian White noise. Note that the x-axis shows the theoretical *alpha*-level, as opposed to the empirical type-I error rate, which is justified as all underlying assumptions are satisfied by definition for Gaussian White Noise.

A.10 β_i values for the non-adaptive CGST

This section presents the β_i values for the non-adaptive CGST. As described in Chapter 8 (section 8.1), the β_i values are chosen as a function of the stage index i using various 'futility functions'. The futility functions (described in section 8.1) include two cosine ramps and two exponential ramps. The resulting β_i values are shown in Table A.7 for different K, per futility function.

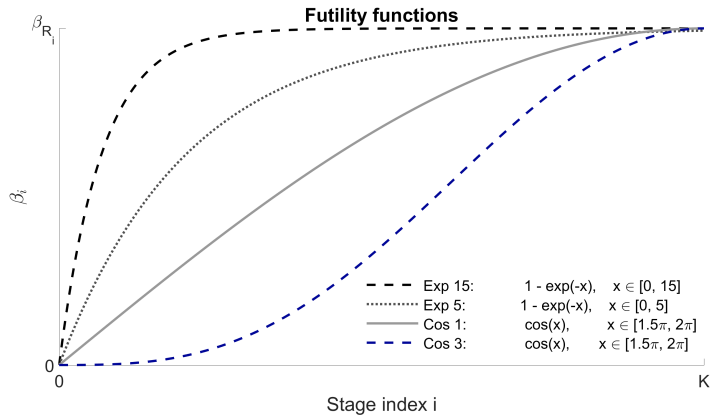


Figure A.10.1: The 'futility functions' for relating stage index i to β_i. Further details are presented in Chapter 8.

A.11 Adaptation criteria for the CGST

In this section, various adaptation criteria for modifying test parameters in the CGST are explored, which include (i) the number of stages K, and (ii) the stage-wise futility boundaries, which are modified through the β_i values. The aim for this section is firstly to establish some intuition in regards to how certain adaptation criteria violate the underlying assumptions of the CGST (in particular, the assumption that the null distribution of the p-values is uniform on [0, 1]), and to explore the extent to which these violations might be relevant for ABR detection. The latter is achieved by quantifying the violation in terms of deviations from (1) the expected FPRs and (2) the expected number of tests rejected for futility (given by β_i). A second goal for this section is to find a useful adaptation criteria where the underlying assumptions of the CGST remain satisfied.

The first adaptation criteria explored in this section uses the stage-one p-value p_1 to modify either (1) the number of stages K, or (2) the β_i values. The adaptation criteria and test protocol are kept as simple as possible. In particular, the following adaptation is used for K:

Table A.10.1: The β_i values used for the non-adaptive CGST in Chapter 8. The values are chosen as a function of the stage index i. The relationship between stage index i and the β_i values is given by the adopted 'futility function' (see section 8.1). The futility functions adopted for the analysis include two cosine ramps and two exponential ramps.

Exp 15									
	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**	**Stage 6**	**Stage 7**	**Stage 8**	**Stage 9**
K=2	0.9495	0.0005	-	-	-	-	-	-	-
K=3	0.9436	0.0064	0	-	-	-	-	-	-
K=4	0.9275	0.0220	0.0005	0	-	-	-	-	-
K=5	0.9026	0.0450	0.0022	0.0001	0	-	-	-	-
K=6	0.8714	0.0722	0.0059	0.0005	0	0	-	-	-
K=7	0.8381	0.0987	0.0116	0.0014	0.0002	0	0	-	-
K=8	0.8038	0.1237	0.0191	0.0029	0.0004	0.0001	0	0	-
K=9	0.7696	0.1465	0.0275	0.0052	0.0010	0.0002	0	0	0
Exp 5									
	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**	**Stage 6**	**Stage 7**	**Stage 8**	**Stage 9**
K=2	0.8719	0.0717	-	-	-	-	-	-	-
K=3	0.7702	0.1459	0.0275	-	-	-	-	-	-
K=4	0.6772	0.1947	0.0557	0.0159	-	-	-	-	-
K=5	0.6003	0.2210	0.0813	0.0299	0.0110	-	-	-	-
K=6	0.5360	0.2342	0.1017	0.0442	0.0192	0.0083	-	-	-
K=7	0.4843	0.2374	0.1167	0.0569	0.0280	0.0136	0.0067	-	-
K=8	0.4409	0.2363	0.1271	0.0676	0.0362	0.0195	0.0104	0.0056	-
K=9	0.4040	0.2332	0.1330	0.0768	0.0438	0.0253	0.0144	0.0083	0.0048
Cos 1									
	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**	**Stage 6**	**Stage 7**	**Stage 8**	**Stage 9**
K=2	0.6715	0.2785	-	-	-	-	-	-	-
K=3	0.4745	0.3481	0.1274		-	-	-	-	-
K=4	0.3629	0.3085	0.2062	0.0723		-	-	-	-
K=5	0.2934	0.2647	0.2101	0.1350	0.0467	-	-	-	-
K=6	0.2451	0.2294	0.1970	0.1511	0.0950	0.0324	-	-	-
K=7	0.2110	0.2005	0.1807	0.1502	0.1135	0.0702	0.0239	-	-
K=8	0.1850	0.1779	0.1648	0.1437	0.1180	0.0881	0.0540	0.0183	-
K=9	0.1645	0.1604	0.1496	0.1360	0.1168	0.0953	0.0698	0.0431	0.0145
Cos 3									
	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**	**Stage 6**	**Stage 7**	**Stage 8**	**Stage 9**
K=2	0.3355	0.6145	-	-	-	-	-	-	-
K=3	0.1184	0.4985	0.3332	-	-	-	-	-	-
K=4	0.0530	0.2825	0.4136	0.2009		-	-	-	-
K=5	0.0280	0.1647	0.3098	0.3142	0.1333	-	-	-	-
K=6	0.0163	0.1021	0.2171	0.2814	0.2393	0.0938	-	-	-
K=7	0.0104	0.0668	0.1529	0.2233	0.2414	0.1853	0.0699	-	-
K=8	0.0070	0.0460	0.1099	0.1726	0.2098	0.2038	0.1470	0.0540	-
K=9	0.0049	0.0331	0.0804	0.1338	0.1742	0.1905	0.1708	0.1196	0.0428

$$K = \begin{cases} 2 & \text{if } p_1 < p_T \\ 3 & \text{if } p_1 \geqslant p_T \end{cases}$$

where p_T is a freely chosen threshold. With respect to the β_i values, the following adaptation is used:

$$\beta_2 = \begin{cases} 0.8 & \text{if } p_1 < p_T \\ 0.15 & \text{if } p_1 \geqslant p_T \end{cases}$$

Hence, for each adaptation there are just two routes, denoted by Route A (for $p_1 < p_T$) and Route B (for $p_1 \geqslant p_T$). When adaptations to K are permitted, then the design for the CGST is given by Fig. A.11 (note that adaptations to β_i are not permitted here): starting at stage one, data $D1$ is obtained, which is analysed using a statistical test, giving p-value p_1. When $p_1 < \alpha_1$, the test is stopped, else the trial proceeds to the 'adaptation phase'. When $p_1 < p_T$, the trial proceeds to stage two through Route A (in which case $K = 2$), else the trial takes Route B ($K = 3$). In stage two, data $D2$ is obtained, which is analysed using a statistical test, giving p-value p_2. Fisher's method is then used to combine p-values, giving $\sum_2 = -2ln(p_1) - 2ln(p_2)$. When $\sum_2$ exceeds the critical decision boundary, the trial is stopped for efficacy, else the trial may proceed to the third stage (under the condition that the trial took Route B). When adaptations to the β_i values are permitted, then the underlying CGST design is illustrated in Fig. A.12. The design is more or less identical to Fig. A.11, except that K is now always set to three. The only difference is that β_2 is set to 0.8 for Route A, or to 0.15 for Route B.

The aforementioned designs are now used to evaluate different *criteria* for choosing between Routes A and B. The first criteria has already been described, and is built around the stage one p-values p_1 (section A.11.1). In section A.11.2, the criteria is built are the stage one sample variance σ_1^2, and in section A.11.3, the criteria is built around the feature covariance matrix estimated from inter-epoch intervals.

A.11.1 P-values as adaptation criteria

This section uses the stage one p-value p_1 as criteria for adapting either K or β_2. The CGST designs are illustrated in Fig. A.11 and A.12 for adaptations to K and β_2 respectively.

Method

The p-values from 100 000 trials were simulated by sampling from a uniform distribution on the [0,1] interval. The simulated p-values were then evaluated using the CGST designs presented in Fig. A.11 and A.12. The p-value threshold p_T for choosing between Routes A and B was set to either $P_T = 0.5$ or to $p_T = 0.1$. The total α level of the trial was furthermore set to 0.03. When taking route A ($K = 2$), then $\alpha_1 = 0.01$ and $\alpha_2 = 0.02$. When taking route B ($K = 3$), then $\alpha_1 = \alpha_2 = \alpha_3 = 0.01$. Note that when adapting K, no futility stopping was used. Also, when adapting β_2, no adaptations to K were permitted.

Results

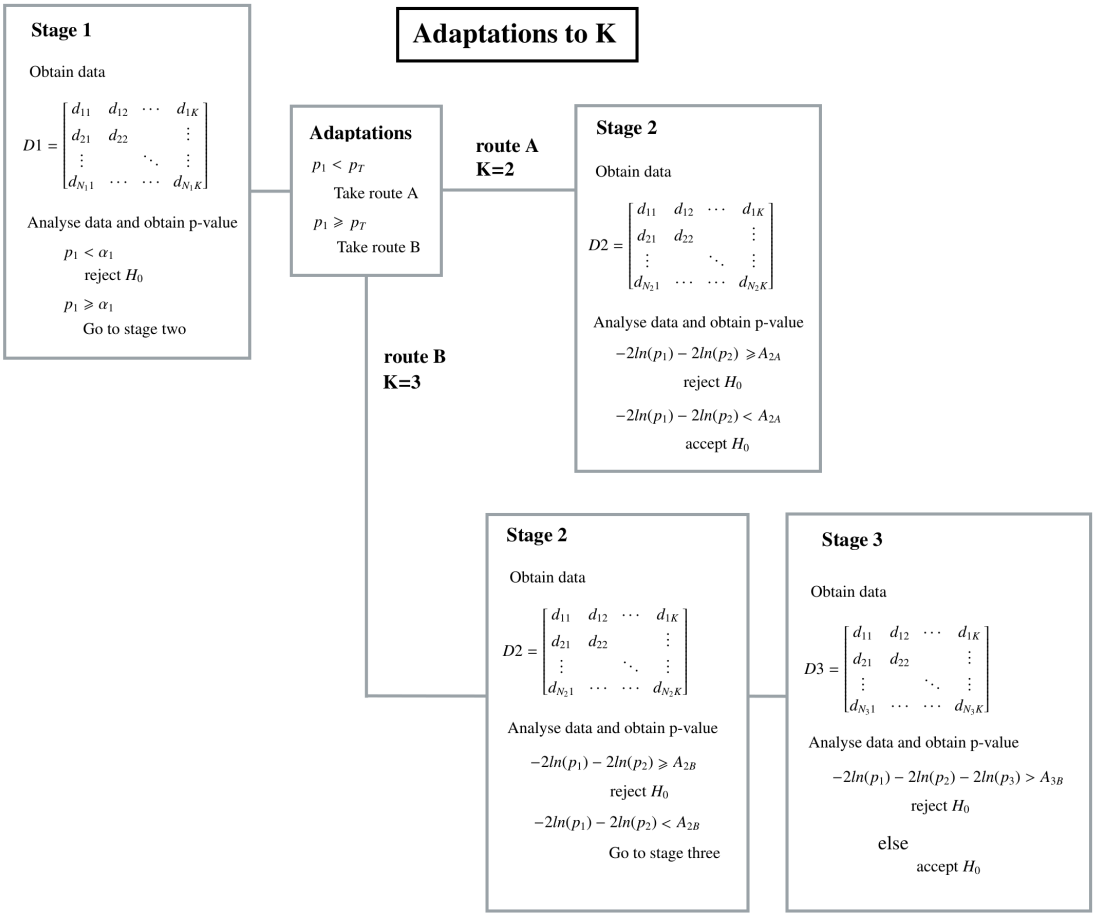


Figure A.11.1: The test procedure used throughout section A.11.1 when adapting the total number of stages K. The criteria is based on the stage one p-value p_1, i.e. when $p_1 < P_T$, the trial takes route A ($K = 2$), else the trial takes route B ($K = 3$). Sections A.11.2 and A.11.3 explore alternative criteria for choosing between routes A and B. Further details are presented in the text.

The approximated underlying null distributions of the p-values for routes A and B are displayed as histograms in Fig. A.13 when using either $p_T = 0.5$ (two upper plots) or $p_T = 0.1$ (two lower two plots) as threshold. As expected, the null distribution for Route A is uniform on the $[\alpha_1, P_T]$ interval, whereas the null distribution for Route B is uniform on the $[P_T, 1]$ interval. The stage-wise FPRs (given by the ratio of the number of times the summary statistic exceeded the efficacy threshold, over the total number of times the trial entered the route in question) and the fraction of tests rejected for futility (calculated using a similar approach) are presented in Table A.8.

Discussion

Results (Table A.8) show that the FPRs for Route A tend to be liberal, whereas the FPRs for Route B are conservative. Although the net result is still close to the nominal α-level of the test, liberal and conservative stage-wise type-I error rates are still undesirable, as this will tend to decrease the robustness or reliability of the performance of the ABR detection method. Ideally, both the stage-wise type-I error rates and the type-I error rate of the full trial should be controlled as intended.

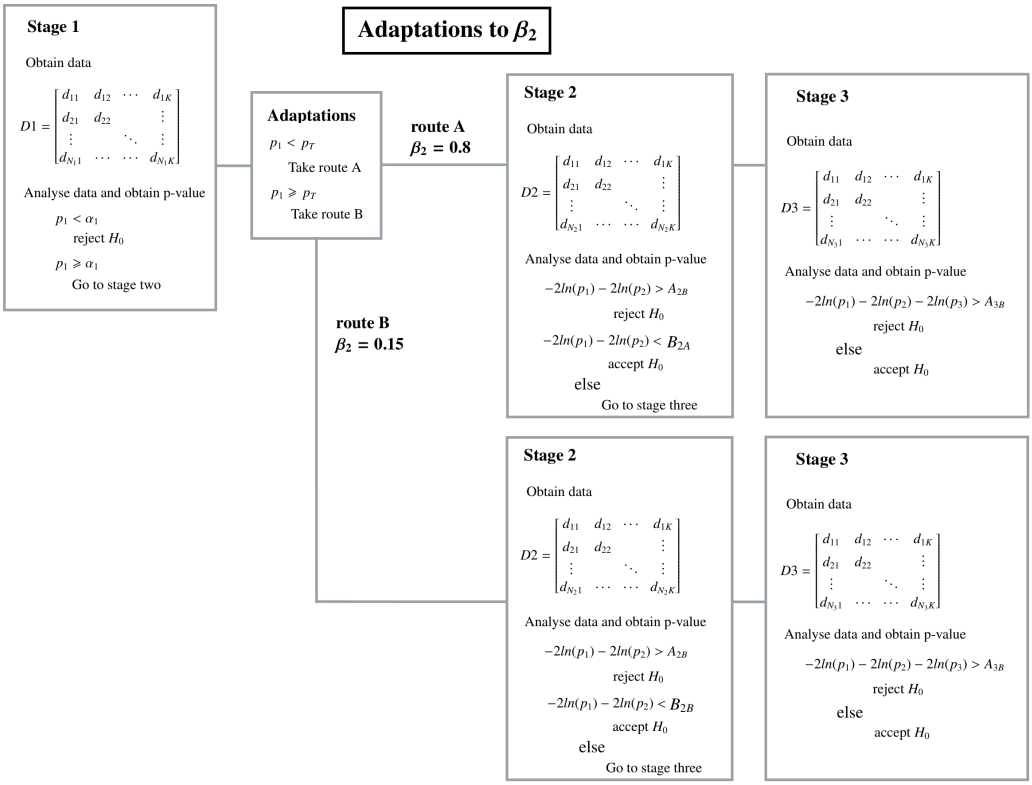


Figure A.11.2: The test procedure used throughout section A.11.1 when adapting the stage two futility boundary through β_2. The criteria is based on the stage one p-value p_1, i.e. when $p_1 < P_T$, the trial takes route A ($\beta_2 = 0.8$), else the trial takes route B ($\beta_2 = 0.15$). Sections A.11.2 and A.11.3 explore alternative adaptation criteria for choosing between routes A and B. Further details are presented in the text.

A.11.2 Feature variance as adaptation criteria

For this section, the criteria for choosing between Routes A and B is given by the stage one sample variance, i.e. Route A is chosen for $\sigma_1^2 < \sigma_T^2$, else route B is chosen, where σ_T^2 is some chosen threshold.

Method

For the first stage of the trial, p-value p_1 was generated using a t-test. Data consists of 50 samples of simulated Gaussian White Noise (with a true mean of zero and a true variance of one). The p-values for the remaining stages of the trial were simulated as described in section A.11.1 above, i.e. by sampling from a uniform distribution on the [0,1] interval. Upon entering stage two, the stage one sample variance σ_1^2 was calculated, which was used to choose between Routes A ($\sigma_1^2 < \sigma_T^2$) and B ($\sigma_1^2 \geqslant \sigma_T^2$). The threshold σ_T^2 was set either 1 or to 0.75.

Results

The approximated underlying null distributions of the p-values for Routes A and B are displayed as histograms in Fig. A.14 when using either $\sigma_T^2 = 1$ two upper plots) or $\sigma_T^2 = 0.75$ (two lower two plots). Results show that the null distributions still deviate

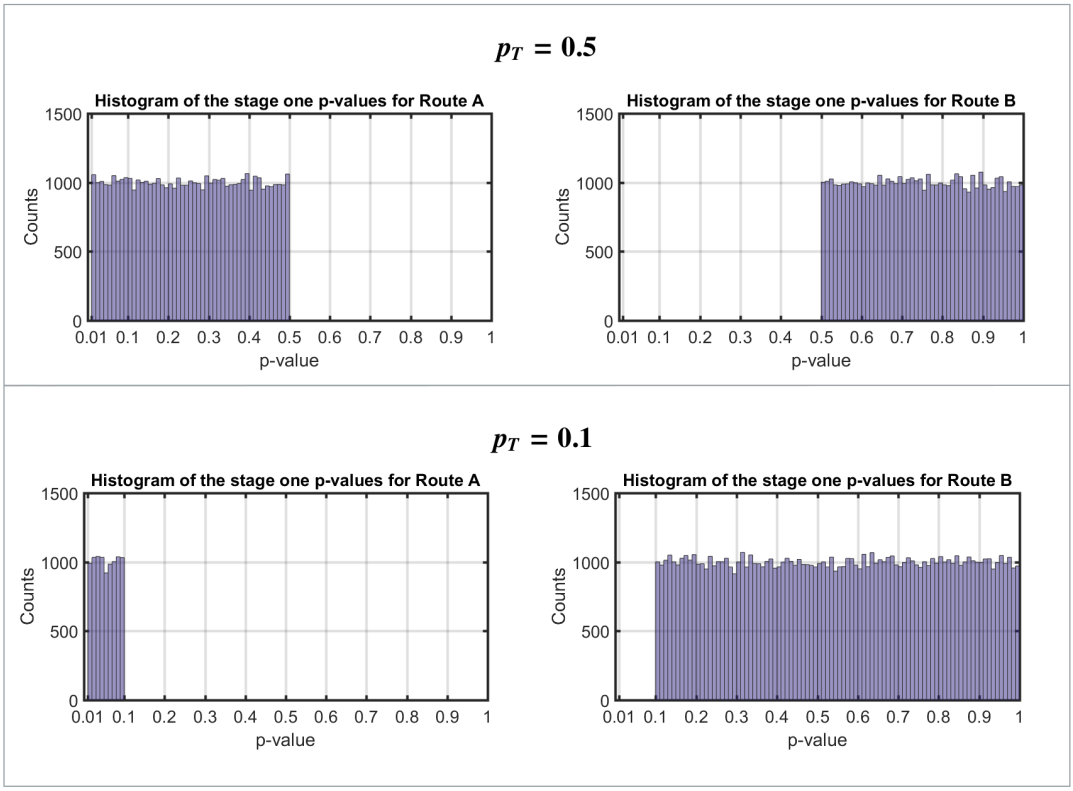


Figure A.11.3: The underlying null distributions of the stage one p-values when using p_1 as criteria for choosing between Routes A and B. The top upper plots show the approximated null distributions when using $p_T = 0.5$ as p_1 threshold for choosing between Routes A and B. The two lower plots show the approximated null distributions for when using $p_T = 0.1$ as threshold. Further details are presented in the text.

from the assumed [0.01, 1] uniform distributions, although the deviations are are much less severe (relative to Fig. A.13). The stage-wise FPRs and the fraction of tests rejected for futility are presented in Table A.9.

Discussion

Although the underlying null distributions for the stage-one p-values are still not uniformly distributed (Fig. A.4), the observed stage-wise FPRs are now relatively close to the expected stage-wise FPRs. The total FPR for the full trial is now also closer to the nominal α-level. Sample variance might therefore be a viable option for data-driven adaptations to the critical decision boundaries. Additional analysis is however required in order to (i) verify the results using real data, and (ii) to verify that these results generalize to alternative CGST designs (e.g. designs that use more than three stages).

A.11.3 Feature variance estimated from inter-epoch intervals as adaptation criteria

The criteria for choosing between Routes A and B explored in this section is built around the feature covariance matrix, estimated from the inter-epoch intervals (denoted by $\boldsymbol{S_2}$).

Table A.11.1: An overview of the expected and observed stage-wise FPRs, along with the expected and observed fraction of tests rejected for futility (at stage two), when using stage one p-value p_1 as criteria for choosing between Routes A and B. The threshold for choosing between Routes A and B was set to either $P_T = 0.5$ or $P_T = 0.1$.

Adaptations to β_2

P_T = 0.5

	Route A (49 089 tests)				**Route B** (49 927 tests)				100 000 tests
	Stage 1	**Stage 2**	**Stage 3**	**Total route A FPR**	**Stage 1**	**Stage 2**	**Stage 3**	**Total route B FPR**	**Total FPR**
Expected FPR	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0098	0.0179	0.0173	0.0450	0.0098	0.0033	0.0041	0.0172	0.0308
Expected futility	-	0.8	-		-	0.15	-		
Observed futility	-	0.6829	-		-	0.2997	-		

P_T = 0.1

	Route A (8992 tests)				**Route B** (89 951 tests)				100 000 tests
	Stage 1	**Stage 2**	**Stage 3**	**Total route A FPR**	**Stage 1**	**Stage 2**	**Stage 3**	**Total route B FPR**	**Total FPR**
Expected FPR	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0106	0.0532	0.0499	0.1137	0.0106	0.0051	0.0069	0.0226	0.0309
Expected futility	-	0.8	-		-	0.15	-		
Observed futility	-	0.1655	-		-	0.1673	-		

Adaptations to the number of stages K

P_T = 0.5

	Route A (49 050 tests)				**Route B** (49 931 tests)				100 000 tests
	Stage 1	**Stage 2**	**Route A FPR**		**Stage 1**	**Stage 2**	**Stage 3**	**Route B FPR**	**Total FPR**
Expected FPR	0.01	0.02	0.03		0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0103	0.0342	0.0445		0.0103	0.0026	0.0046	0.0174	0.0307

P_T = 0.1

	Route A (9076 tests)				**Route B** (89 929 tests)				100 000 tests
	Stage 1	**Stage 2**	**Route A FPR**		**Stage 1**	**Stage 2**	**Stage 3**	**Route B FPR**	**Total FPR**
Expected FPR	0.01	0.02	0.03		0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0101	0.1092	0.1191		0.0101	0.0053	0.0073	0.0226	0.0313

Table A.11.2: An overview of the expected and observed stage-wise FPRs, along with the expected and observed fraction of tests rejected for futility (at stage two), when using stage one sample varianceσ_1^2 as criteria for choosing between Routes A and B. The threshold for choosing between Routes A and B was set to either $\sigma_T^2 = 1$ or $\sigma_T^2 = 0.75$. Further details presented in the text.

Adaptations to the futility boundaries

$\sigma_T^2 = 1$

	Route A (52 008 tests)				**Route B** (47 023 tests)				100 000 tests
	Stage 1	**Stage 2**	**Stage 3**	**Total route A FPR**	**Stage 1**	**Stage 2**	**Stage 3**	**Total route B FPR**	**Total FPR**
Expected FPR	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0097	0.012	0.0123	0.034	0.0097	0.0093	0.009	0.028	0.0309
Expected futility	-	0.8	-		-	0.15	-		
Observed futility	-	0.7858	-		-	0.1647	-		

$\sigma_T^2 = 0.75$

	Route A (9715 tests)				**Route B** (89 280 tests)				100 000 tests
	Stage 1	**Stage 2**	**Stage 3**	**Total route A FPR**	**Stage 1**	**Stage 2**	**Stage 3**	**Total route B FPR**	**Total FPR**
Expected FPR	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0101	0.0149	0.0141	0.0391	0.0101	0.0095	0.01	0.0296	0.0303
Expected futility	-	0.8	-		-	0.15	-		
Observed futility	-	0.757	-		-	0.1559	-		

Adaptations to the number of stages K

$\sigma_T^2 = 1$

	Route A (52 054 tests)				**Route B** (46 901 tests)				100 000 tests
	Stage 1	**Stage 2**	**Route A FPR**		**Stage 1**	**Stage 2**	**Stage 3**	**Route B FPR**	**Total FPR**
Expected FPR	0.01	0.02	0.03		0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0104	0.0240	0.0345		0.0104	0.0083	0.009	0.0277	0.031

$\sigma_T^2 = 0.75$

	Route A (9705 tests)				**Route B** (89 313 tests)				100 000 tests
	Stage 1	**Stage 2**	**Route A FPR**		**Stage 1**	**Stage 2**	**Stage 3**	**Route B FPR**	**Total FPR**
Expected FPR	0.01	0.02	0.03		0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0098	0.0285	0.0384		0.0098	0.0096	0.0095	0.0289	0.0296

Data in this section consists of simulated coloured noise, generated as described in section 4.4, using band-pass filter settings of 100-2000 Hz. Data was now furthermore analysed using the Hotelling's T^2 test (applied to 25 TVMs). To keep the adaptation criteria as

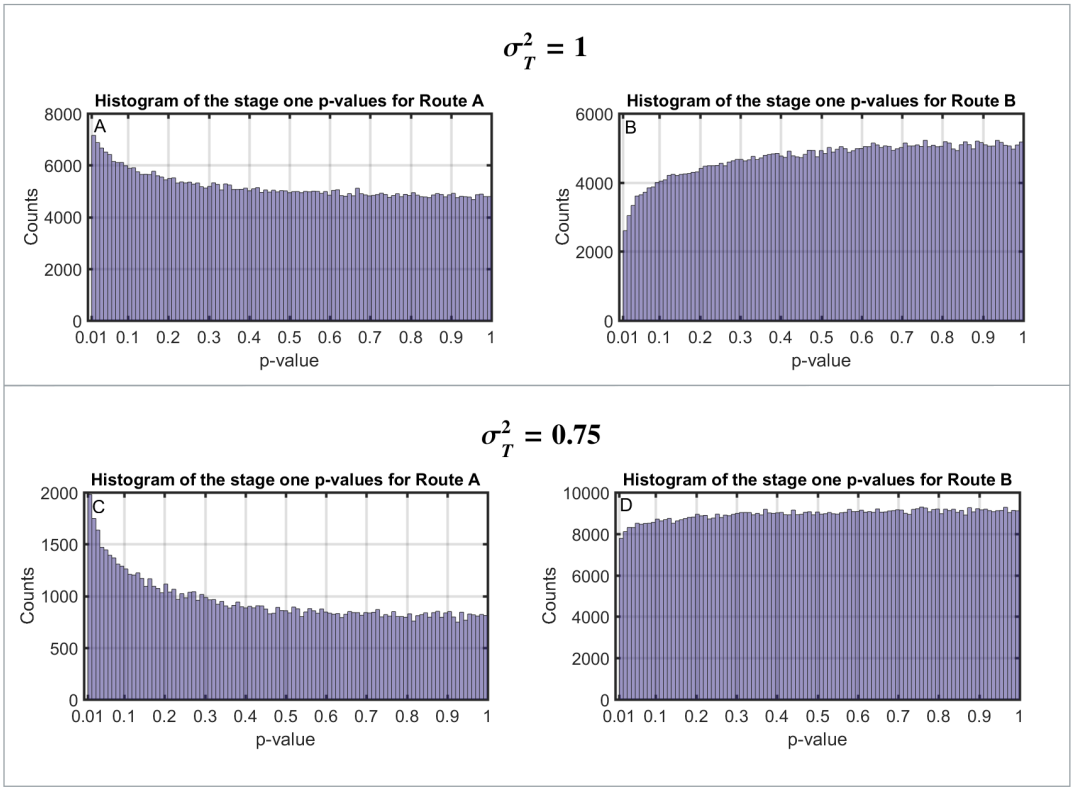


Figure A.11.4: The underlying null distributions of the stage one p-values when using stage one sample variance σ_1^2 as criteria for choosing between Routes A and B. The top upper plots show the approximated null distributions when using $\sigma_T^2 = 1$ as threshold (for σ_1^2) when choosing between Routes A and B. The two lower plots show the approximated null distributions when using $\sigma_T^2 = 0.75$ as threshold. Further details are presented in the text.

simple as possible, it would also be useful to have a single value for representing the feature covariance matrix. One option is to use the trace of the feature covariance matrix, i.e. the sum of the diagonal components of the feature covariance matrix, henceforth $Tr(\boldsymbol{S2})$. Alternatively, the determinant of the feature covariance matrix can be used (henceforth $\mid \boldsymbol{S2} \mid$), which gives a single value for multivariate scatter.

The first goal for this section is to explore the extent to which $\boldsymbol{S_2}$ is independent of the feature covariance matrix estimated from the initial 15 ms analysis window (denoted by $\boldsymbol{S_1}$). Note that independence between $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ implies that $\boldsymbol{S_2}$ can be used as adaptation criteria without introducing a violation to the underlying CGST assumptions. The second goal is to evaluate the underlying null distributions per route, and to quantify potential violations in terms of increased or decreased stage-wise type-I error rates and fraction of tests rejected for futility.

Independence assessment

Data consists of simulated coloured noise, generated as described in section 4.4 (using a band-pass filter of 100-2000 Hz). A total of 10 000 recordings were simulated for each AR

model. Independence between $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ was hence explored separately, per AR model (there were 149 AR models, corresponding to the 149 recordings in data set **D1**)). The simulated recordings were then structured into ensembles of $N = 500$ 30.03 ms epochs, after which feature covariance matrices $\boldsymbol{S_1}$ (extracted from the initial 0-15 ms windows of the epochs) and $\boldsymbol{S_2}$ (extracted from the 15-30 ms windows) were calculated. Both the trace and the determinant were then calculated from all feature covariance matrices, and the resulting values were used to estimate CCs. A CC was hence calculated per AR model, using either the traces or the determinants of the feature covariance matrices.

Results

The resulting CCs are presented in Fig. A.15, as a function of the index of the AR model being simulated. Results suggest that independence is satisfied for all AR models, with the exception of a single AR model, which gave CCs of 0.1117 and 0.1899 for the trace and determinants respectively.

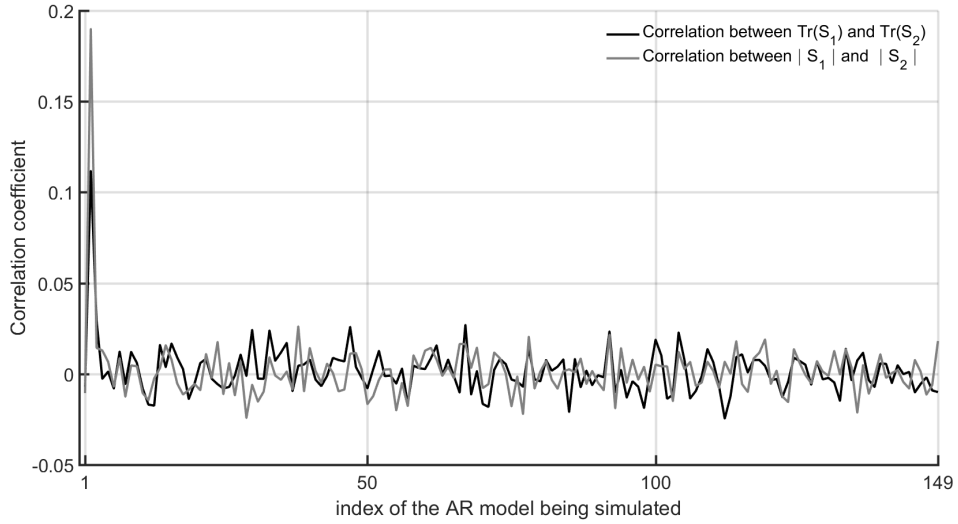


Figure A.11.5: The CCs for the trace and the determinant of feature covarience matrices $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ as a function of the AR model being simulated. Further details are presented in the text.

Performance assessment

This section evaluates the underlying null distributions and the specificity of the CGST when using $Tr(\boldsymbol{S_2})$ as criteria for choosing between Routes A and B. Data in this section consists of simulated coloured noise, generated using AR coefficients from just a single AR model. In particular, the AR model selected for the assessment was the model that resulted in a relatively high CC in Fig. A.15 (AR model located at index two). Results from this section can hence be considered as a 'worst case scenario' for data set **D1**.

Method

A total of 100 000 recordings of coloured noise were simulated as described in section 4.4 (now using just a single AR model, as mentioned above), which were band-pass filtered

from 100-2000 Hz. The simulated recordings were then structured into ensembles of $N = 1500$ 30.03 ms epochs, and the initial 0-15 ms windows of the ensembles were analysed in K sequential stages using the Hotelling's T^2 test. Upon entering stage two, $Tr(\boldsymbol{S_2})$ was calculated from the inter-epoch intervals, which was then used as criteria for choosing between Routes A and B. In particular, when $Tr(\boldsymbol{S_2}) < Tr_T$, the trial took Route A, else the trial took Route B. The threshold Tr_T was set to either 46.5 (resulting in an approximate 1 to 1 entry ratio for Routes A and B respectively), or to 45.5 (resulting in an approximate 1 to 9 entry ratio of Routes A and B respectively).

Results

The approximated underlying null distributions of the p-values for Routes A and B are displayed as histograms in Fig. A.6 when using either $Tr_T = 46.5$ (the two upper plots) or $Tr_T = 46.5$ (the two lower two plots). No noticeable deviations from the assumed uniform distributions are observed. The stage-wise FPRs and the fraction of tests rejected for futility are presented in Table A.10. Results suggest that the stage-wise type-I error rates and fraction of tests rejected for futility are now controlled as intended.

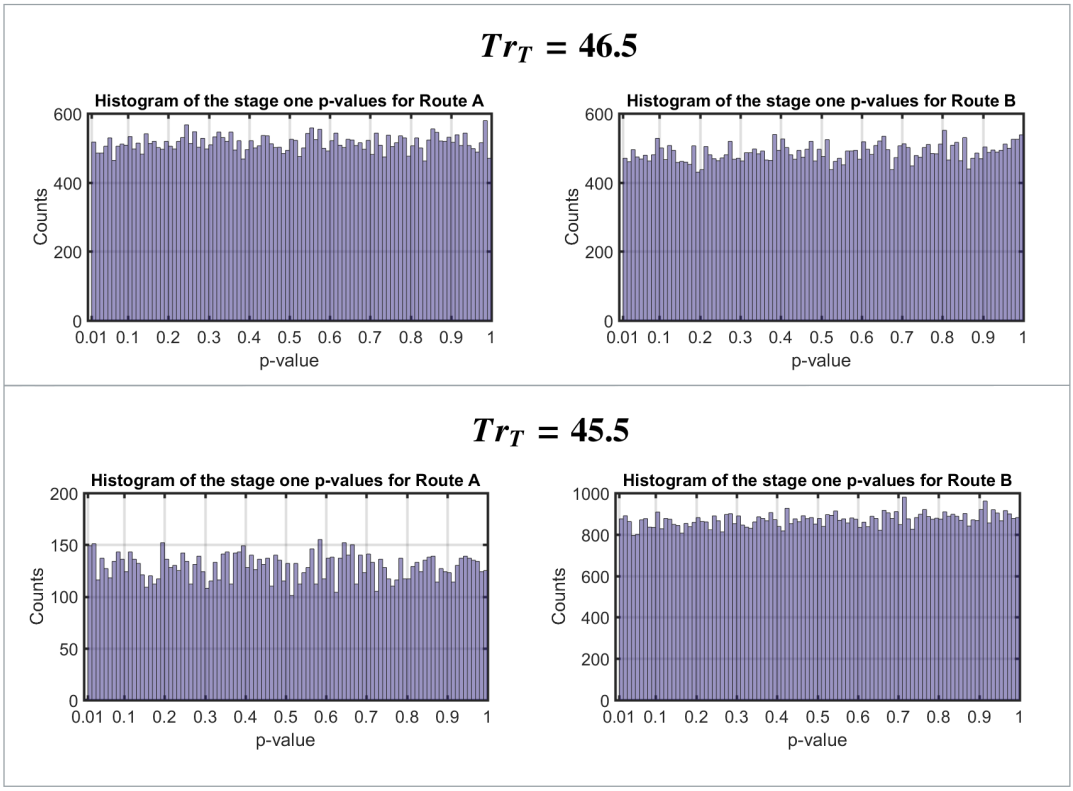


Figure A.11.6: The approximated underlying null distributions of the stage one p-values when using $Tr(\boldsymbol{S_2})$ as criteria for choosing between route A or route B. When $Tr(\boldsymbol{S_2}) < Tr_T$, the trial takes route A, else the trial takes route B. The threshold Tr_T for choosing between routes A and B was set to either 46.5 or to 45.5.

Discussion

Even though results from this section can be considered as a 'worst case scenario',

Table A.11.3: An overview of the expected and observed stage-wise FPRs, along with the expected and observed fraction of tests rejected for futility (at stage two), when using $Tr(\boldsymbol{S_2})$ as criteria for choosing between Routes A and B. When $Tr(\boldsymbol{S_2}) < Tr_T$, the trial takes Route A, else the trial takes Route B, where Tr_T takes value of either 46.5 or 45.5. Further details are presented in the text.

Adaptations to the futility boundaries									
$Tr_T < 46.5$									
	Route A (50 212 tests)				**Route B** (48 780 tests)				100 000 tests
	Stage 1	**Stage 2**	**Stage 3**	**Total route A FPR**	**Stage 1**	**Stage 2**	**Stage 3**	**Total route B FPR**	**Total FPR**
Expected FPR	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0101	0.01	0.0097	0.0298	0.0101	0.009	0.01	0.029	0.0292
Expected futility	-	0.8	-		-	0.15	-		
Observed futility	-	0.8105	-		-	0.1575	-		
$Tr_T < 45.5$									
	Route A (13 122 tests)				**Route B** (85 858 tests)				100 000 tests
	Stage 1	**Stage 2**	**Stage 3**	**Total route A FPR**	**Stage 1**	**Stage 2**	**Stage 3**	**Total route B FPR**	**Total FPR**
Expected FPR	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0102	0.0098	0.0089	0.0289	0.0102	0.0097	0.0094	0.0293	0.0291
Expected futility	-	0.8	-		-	0.15	-		
Observed futility	-	0.812	-		-	0.1547	-		
Adaptations to the number of stages K									
$Tr_T < 46.5$									
	Route A (50 882 tests)				**Route B** (48 172 tests)				100 000 tests
	Stage 1	**Stage 2**	**Route A FPR**		**Stage 1**	**Stage 2**	**Stage 3**	**Route B FPR**	**Total FPR**
Expected FPR	0.01	0.02	0.03		0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0095	0.0206	0.0301		0.0095	0.0097	0.0101	0.0292	0.0295
$Tr_T < 45.5$									
	Route A (12 786 tests)				**Route B** (86 278 tests)				100 000 tests
	Stage 1	**Stage 2**	**Route A FPR**		**Stage 1**	**Stage 2**	**Stage 3**	**Route B FPR**	**Total FPR**
Expected FPR	0.01	0.02	0.03		0.01	0.01	0.01	0.03	0.03
Observed FPR	0.0094	0.0205	0.0299		0.0094	0.0096	0.0095	0.0284	0.0284

the stage-wise type-I error rates and fraction of tests rejected for futility appear to be controlled as intended. This suggests that data-driven adaptations to the stage-wise critical decision boundaries are permitted, under the condition that they are built around the feature covariance matrix estimated from inter-epoch intervals ($\boldsymbol{S2}$), albeit when using a high-pass cut-off frequency of 100 Hz and a stimulus rate of 33.3 Hz.

A.12 P-value combination functions for the CGST

This section explores the test sensitivity of a sequentially applied t-test for various p-value combination functions. In particular, the summary statistic Σ_2 is defined as either a sum of p-values, given by:

$$\Sigma_2 = p_1 + p_2 \tag{6}$$

or as a sum of inverse χ^2-distributed random variables, all with two DOF:

$$\Sigma_2 = [\chi^2_2]^{-1}(1 - p_1) + [\chi^2_2]^{-1}(1 - p_2) \tag{7}$$

of by a sum of inverse F-distributed random variables, all with Q and $N - Q$ DOF:

$$\Sigma_2 = F^{-1}_{v_1,v_2}(1 - p_1) + F^{-1}_{v_1,v_2}(1 - p_2) \tag{8}$$

The goal for this section is to compare test sensitivity for the aforementioned p-value combination functions.

Method

Data for the assessment consists of simulated Gaussian White Noise (with a true mean of 0 and a variance of 1). A total of 100 000 trials were simulated, where the sample size per trial was set to 100. A response was then simulated by adding a constant amplitude signal to the samples, where the amplitude to the signal ranged from 0 to 0.5, in steps of 0.05 (these values were chosen as they gave a good coverage of TPRs). Each sample was then analysed in two sequential stages (using $N_1 = N_2 = 50$) using a t-test. The resulting p-values were then evaluated using critical decision boundaries estimated using the CGST (using $\alpha_1 = \alpha_2 = 0.025$ and $\beta_1 = \beta_2 = 0$).

Results

The TPRs are plotted as a function of the signal amplitude per p-value combination function in Fig. A.17. Results demonstrate a minor advantage for the sum of inverse χ^2 and the sum of inverse F-distributed random variables over a simple summation of p-values.

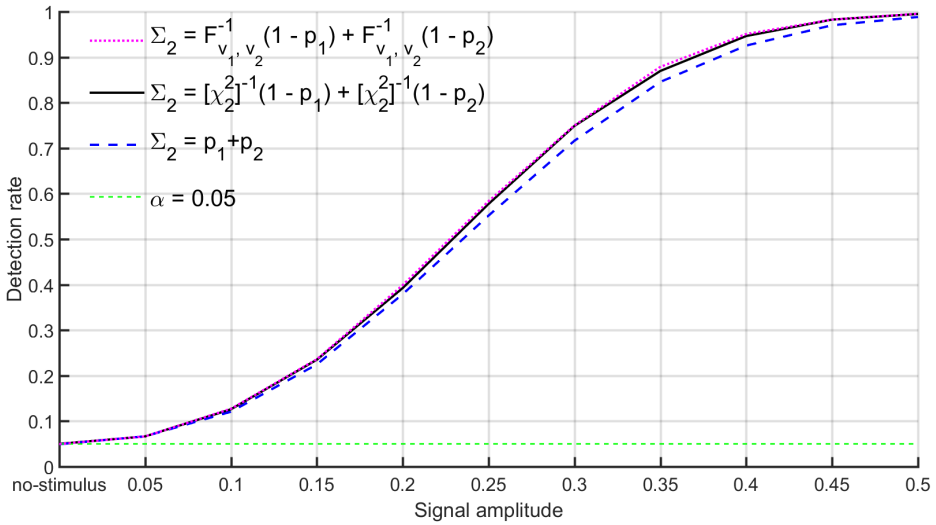


Figure A.12.1: The TPR as a function of the amplitude of the simulated signal for different p-value combination functions.

Discussion

Results from this section suggest an advantage for the sum of inverse χ^2 and the sum of inverse F-distributed random variables over a summation of untransformed p-values.

However, it should be stressed that these results cannot be generalised to all test conditions, as there is no single optimal method when combining p-values across all test conditions, i.e. the optimal method will depend on the underlying distribution of the p-values. Additional simulations (results not shown) in fact demonstrate an advantage for a sum of untransformed p-values over e.g. the sum of inverse χ^2-distributed random variables, under the condition that the underlying p-value null distribution is uniform on e.g. the $[0, P_T]$ interval for $0 > P_T < 1$. For real world applications, the distribution of the p-values will almost never be uniform under the alternative hypothesis, but will instead be skewed towards small values. The latter was similarly the case for this section, in which case using a sum of untransformed p-values results in a minor loss of test sensitivity.

A.13 The stage-wise statistical powers γ_i

This section presents the γ_i values (for the adaptive approach in chapter 9) associated with equal ensemble sizes N_i for all K stages, and where the TPR for the full sequential analysis is equal to 0.95. In particular, 5000 ensembles with increasing or decreasing N (split equally across the K stages) were simulated until a TPR of 0.95 was obtained. The resulting stage-wise TPRs (the γ_i values) are shown in Table A.12 below.

Table A.13.1: The resulting TPRs (the γ_i values for chapter 9) when splitting the available N equally across K stages, and where the TPR for the full sequential analysis was equal to 0.95.

	K=1	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**	**K=8**	**K=9**
K=1	0.95								
K=2	0.7745	0.1754							
K=3	0.5571	0.3129	0.0795						
K=4	0.4140	0.3329	0.1552	0.0481					
K=5	0.3160	0.3125	0.1933	0.0943	0.0343				
K=6	0.2485	0.2795	0.2054	0.1246	0.0650	0.0269			
K=7	0.2014	0.2468	0.2038	0.1413	0.0871	0.0481	0.0218		
K=8	0.1646	0.2154	0.1945	0.1490	0.1030	0.0657	0.0382	0.0190	
K=9	0.1402	0.1916	0.1833	0.1500	0.1116	0.0774	0.0504	0.0303	0.0158

A.14 Estimating the non-centrality parameter δ

The equation for calculating statistical power (Eq. 9.1) assumes that the non-centrality parameter δ is the true non-centrality parameter, which is typically unknown for real world applications. Instead, δ is usually estimated from data, and can therefore be contaminated by significant amounts of noise. As a result, the power calculation can potentially be inaccurate. Uncertainty within $\hat{\delta}$ should hence be taken into account when adapting test parameters. This section present a brief literature review on various

methods in the literature for estimating δ when using the Hotellign's T^2 test as detection method.

There are various methods are available in the literature for approximating δ and/or its underlying probability distribution. Some of these methods have been developed for the non-central χ^2 distribution (Meyer, 1967; Spruill, 1986; Li et al, 2009; Neff & Strawderman, 1976; Saxena & Alam, 1982; Shao & Strawderman, 1995), which might be applicable to non-central F-distributions, under the condition that the F-distributed random variable can be decomposed into two χ^2-distributed random variables (a non-central χ^2 and a central χ^2). Rukhin (1993) has also proposed a method for estimating the non-centrality parameter directly for a non-central F-distribution, but similarly assumes that the F-statistic can be decomposable into two χ^2-distributed random variables. Note that it is not clear how or if the (F-transformed) T^2 statistic can be decomposed into two χ^2-distributed random variables. As shown in Wilk's (1932), the T^2 statistic can actually be decomposed (using maximum likelihood) into a ratio of two generalised variances (a ratio of the determinants of two covariance matrices), of which the underlying distribution is quite complex (and not χ^2, see e.g. Mathai, 1972).

Hence, unless the T^2 statistic can be related to a ratio of a central and non-central χ^2-distributed random variable, the aforementioned methods appear to be inapplicable when using the Hotelling's T^2 test as detection method. Various alternative methods for estimating δ and its accuracy are however available in the literature, and were not explored in this work (Perlman & Rasmussen, 1975; Berger et al, 1998; Chow, 1987; Kubokawa et al, 1993; Leung & Muirhead, 1987; Kubokawa et al, 2017). Future work might explore these methods in an attempt to find a more accurate approach for quantifying the uncertainty within $\hat{\delta}$.

An alternative route for approximating uncertainty within $\hat{\delta}$ is through reliability. In the literature, reliability is typically described from a 'classical test theory' point of view (see e.g. Levin & Subkoviak, 1977; Bruton et al., 2000; Kanyongo et al., 2007). In particular, it is assumed that each observation (obtained from a single sampling unit) is composed of a true score and an error score. The observed variance of a group of observations is similarly assumed to be composed of a true variance σ_t^2 and an error variance σ_e^2. The 'true variance' is hence the variance *between* sampling units, whereas the 'error variance' is the variance *within* sampling units (how this relates to evoked response detection is further addressed below). A reliability coefficient R can then defined as (Kanyongo et al., 2007):

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} = \frac{\sigma_t^2}{\sigma_o} \tag{9}$$

where σ_o is the observed score variance. When measurements are made without error, then the observations will consist of perfectly reliable scores, and R will equal one.

Alternatively, R will approach zero as the measurement error grows to ∞, in which case the observations can be considered completely random.

The concept of reliability can be applied to evoked response detection as follows: each observed value (each measurement within the EEG recording) consists of a true score (given either by zero when no response is present, and otherwise by the amplitude of the evoked response) and an error score (given by the EEG background activity). The true score variance σ_t^2 is then given by the variance within the true ABR waveform, whereas the error score variance σ_e^2 is given by variance due to the EEG background activity. This is closely related to the SNR estimator given by Raz et al (1988):

$$\hat{SNR} = \frac{\hat{\sigma_o}^2 - \hat{\sigma_n}^2}{\sigma_n^2} \tag{10}$$

where $\hat{\sigma_o}^2$ is the total estimated power, given by:

$$\hat{\sigma_o}^2 = \frac{1}{JN} \sum_{j}^{J} \sum_{i}^{N} d_{ij}^2 \tag{11}$$

where d_{ij} is the jth value of the ith epoch, and where $\hat{\sigma_n}^2$ is the estimated power of the noise, given by:

$$\sigma_n^2 = \frac{1}{N(J-1)} \sum_{i=1}^{N} \sum_{j=1}^{J} (d_{ij} - \bar{X}_j)^2 \tag{12}$$

where $\bar{X}_j$ is the jth value of the ensemble coherent average.

A convenient property associated with the method in Raz et al (1988) is that it allows confidence intervals to be constructed for the estimated SNR, which were shown by Raz et al to be quite accurate. In future work, a method for incorporating uncertainty within $\hat{\delta}$ in the presence of an *a priori* assumed minimum response might be designed around the SNR estimator in Raz et al (1988).

A.15 Independence violations for Cortical Auditory Evoked Potential Detection

This section explores the independence violation for CAEP detection, as a function of the cut-off frequency for the high-pass filter and the (hypothetical) stimulus rate. The

data used for this section consists of simulated coloured noise, generated by filtering Gaussian White Noise with an all pole filter where the poles are the parameters of an AR model, estimated from recordings of EEG background activity (see also section 4.4). The recordings of EEG background activity used in this section are further described below.

A.15.1 Data

Recordings of EEG background activity (no stimulus was used) were obtained from 19 subjects. All subjects had normal hearing levels (PTA thresholds <20 dB HL for the 500, 1000, 2000, 4000, and 6000 Hz frequencies) and normal tympanic membrane responses. The subjects sat in an upright position in a comfortable chair, and were watching a DVD on a monitor placed at eye level. EEG measurements were then obtained at a sampling rate of 30 kHz with electrodes places at the high forehead, the right mastoid, and the left mastoid which served as ground. The electrode impedances remained below 1 kΩ throughout the recording. A total of 130 continuous EEG recordings were available collected (approximately 6-7 recordings were collected per subject), with an average duration of ∼2.5 minutes per recording, resulting in approximately 5.5 hours of EEG. Each recording was then downsampled to 1 kHz and band-pass filtered from 0.1-100 Hz. A 60th-order AR model was then fit to each pre-processed recording, as described in section 4.4.

A.15.2 Method

Simulated coloured noise was generated as described in section 4.4, now using the aforementioned AR models. Each simulated recording was then band-pass filtered using 3rd-order Butterworth filters from fc to 100 Hz, and structured into ensembles of $N = 50$ 500 ms epochs. The high-pass cut-off frequency fc was also varied from 0.1 to 10 Hz, in steps of 0.3 Hz. The distance between the 500 ms epochs (denoted by τ) was also varied from 0 to 500 ms, in steps of 50 ms. The latter corresponds to a (hypothetical) stimulus rate of $\frac{500+\tau}{1000}$. A total of 25 000 ensembles of $N = 50$ epochs were simulated for each choice of τ and fc, all of which were analysed in the time domain using the Hotelling's T^2 test (applied to 10 TVMs).

A.15.3 Results

The FPRs (each calculated from 25 000 simulated tests using $\alpha = 0.05$) are presented in Fig. A.18 as a function of the high-pass cut-off frequency fc and the (hypothetical) stimulus rate. The two-sided 95% CIs for α are given by [0.0474, 0.0528]. Significant deviations from α are indicated in Fig. A.18 by blue (< 0.0474) and red (> 0.0528) cells respectively, whereas green cells indicate that the observed FPR fell within the 95% CIs.

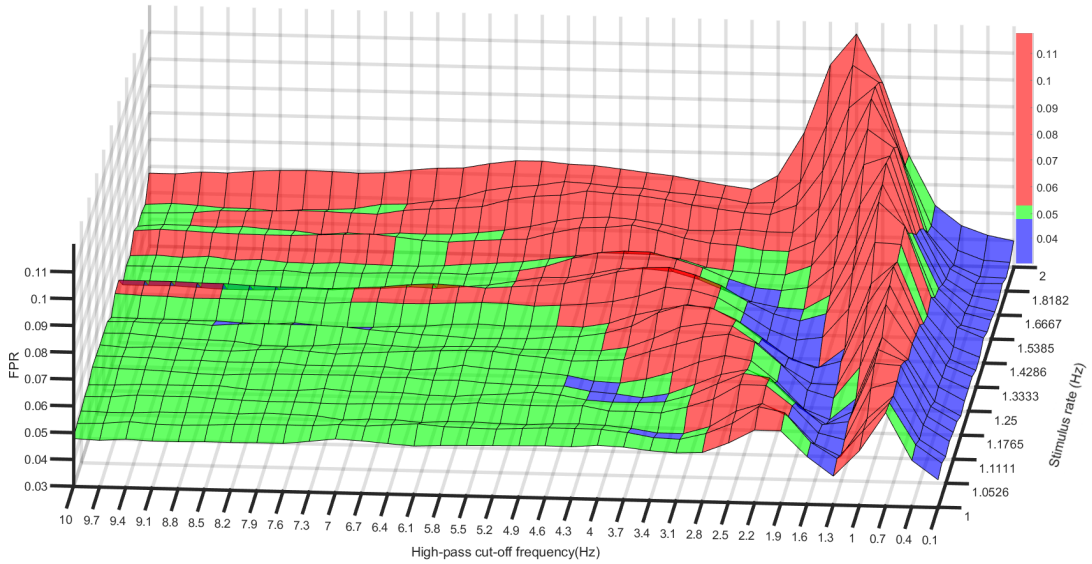


Figure A.15.1: The FPRs (each calculated from 25 000 simulated tests using $\alpha = 0.05$) as a function of the high-pass cut-off frequency fc and the (hypothetical) stimulus rate. Significant deviations from nominal level $\alpha = 0.05$ are indicated by blue (< 0.0474) and red (> 0.0528) cells, whereas green cells indicate that the observed FPR fell within the 95% CIs.

A.15.4 Discussion

Results demonstrate significant violations to the independence assumption as a function of the high-pass cut-off frequency fc and the (hypothetical) stimulus rate. For a more in-depth discussion on the independence assumption, the reader is referred to Chapter 5 (section 5.1).

A.16 Magnitude response of the filter

Throughout this thesis, digital filtering is achieved using 3rd order Butterworth filters. The low-pass cut-off frequency is typically set to 1500 Hz, whereas the high-pass cut-off frequency to set to either 30 or 100 Hz. In some sections, alternative high-pass cut-off frequencies between 30-100 Hz are used. Digital filtering is furthermore always realised using matlab's 'filtfilt' function, which is a forward and reverse filtering technique that introduces zero phase-shift to the filtered signals, i.e. the phase response of the Butterworth filters is always zero. The magnitude response for a 30-1500 and 100-1500 Hz 3rd order Butterworth band-pass filter is shown in figure A.21 below.

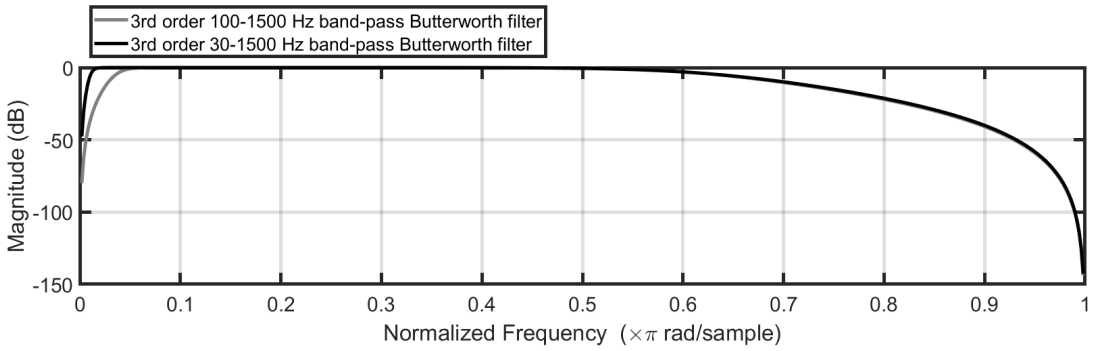


Figure A.16.1: The magnitude response of a 3rd order 30-1500 Hz Butterworth band-pass filter, and a 3rd order 100-1500 Hz Butterworth band-pass filter.

A.17 Optimal stage-wise ensemble sizes for two and three stage sequential tests

This section describes an approach for determining the optimal stage-wise ensemble sizes for a sequentially applied Hotelling's T^2 test, where optimal is defined as the smallest possible test time for a fixed test sensitivity and specificity. For this section, test sensitivity is fixed at 0.95, which is achieved by increasing or decreasing the total ensemble size N until a TPR of 0.95 has been obtained. This is repeated using different choices for the stage-wise ensemble sizes N_i. In particular, a specific percentage of the total N is spent at each stage i. As an example, say $N = 200$, of which 25% is spent at stages 1 (giving $N_1 = 50$) and 75% at stage two (giving $N_2 = 150$). If the TPR (for some effect size, to be defined) is smaller than 0.95, then N is increased, after which it is again split across N_1 and N_2 (using the same 1/4 ratio). This is repeated until a TPR of 0.95 has been obtained, which is then also repeated for different N_i ratios.

Method

The total α-level is set to 0.05, which is spread equally across the K stages, giving stage-wise α_i values of $\frac{\alpha}{K}$ for all i. The summary statistic for evaluating the null hypothesis of 'no effect present' is given by a sum of F-transformed p values, as defined in Eq. 9.8, where $v_1 = 2$ and $v_2 = N_i - 2$, and where N_i is the ensemble size for stage i. To keep the approach as simple as possible, the following assumptions are also made: (1) the noise is a white, Gaussian, zero mean process with a true of variance of 1, (2) the feature set for the Hotelling's T^2 test is two-dimensional, and (3) the true amplitude of the effect to detect is 0.1.

Given the aforementioned assumptions and test parameters, statistical power can now easily be calculated for different choices of N and N_i. Starting at stage one, statistical power γ_1 is given by (Bilodeau & Brenner, 1999):

$$\gamma_1 = 1 - F_{nc}\left(F^{-1}(1-\alpha_1, 2, N_1 - 1), 2, N_1 - 2, \delta_1\right) \tag{13}$$

were δ_1 is the non-centrality parameter, which (for the Hotelling's T^2 test) is given by (Bilodeau & Brenner, 1999):

$$\delta_1 = N_1(\boldsymbol{\mu} - \boldsymbol{\mu_0})\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu_0})^H \tag{14}$$

where $\boldsymbol{\mu}$ is the true mean feature vector, $\boldsymbol{\mu_0}$ is the vector of hypothesizes values to test against (given by zeros), and Σ is the true feature covariance matrix. Note that $\boldsymbol{\mu}$ and Σ are now both known, and are given by $[0.1, 0.1]$ and the 2-dimensional identity matrix $\boldsymbol{I}$, respectively. The stage i non-centrality parameter can therefore be simplified to $\delta_i = N_i 0.02$, i.e. the effect size (given by $\frac{\delta_i}{N_i}$) is equal to 0.02.

For stage two, statistical power can be calculated with help from the CGST. The stage two alternative distribution of the summary statistic (denoted by $\phi^1_{\Sigma_2}$, note that this is now the true distribution of the summary statistic) is given by convolution $\phi_1^{1,T[B_1,A_1]} * \phi_2^1$, where $\phi_1^{1,T[B_1,A_1]}$ denotes the stage one alternative distribution, truncated to the $[B_1, A_1]$ interval, and where ϕ_2^1 denotes the stage two alternative distribution (for the F-transformed stage two p value). Stage two statistical power is then given by the area under $\phi^1_{\Sigma_2}$ to the right of A_2. The total statistical power after stage two, say γ_{Σ_2}, is now equal to $\gamma_1 + \gamma_2$. Statistical power for stage three (and indeed for all subsequent stages) follows the exact same procedure. Once N and N_i have been found, such that the total statistical power is equal to 0.95, then the mean number of samples tested (i.e. the mean test time) is readily given by $\sum_{i=1}^{K} \gamma_i N_i$.

Results

The mean number of samples used (for obtaining a TPR of 0.95) for a two stage design is first plotted in Figure A.17.1 as a function of the percentage of N that was spent in stage one. Results show that the minimum test time is obtained when 47% of N is spent in stage one, and the remaining 53% in stage two. This corresponded to a stage one statistical power of 68.19% and a stage two statistical power of 26.82% (giving a total power of 95.01%). Results for a three stage design are then plotted in Figure A.17.2. The mean number of samples used (for obtaining a TPR of 0.95) is now shown as a function of the percentage of N spent in both stages one and stage two. Results show that the minimum test time is obtained when spending 30% of N in stage one, 30% in stage two, and the remaining 40% in stage three. This corresponded to a stage one statistical power of 45.92%, a stage two statistical power of 32.36%, and a stage three statistical power of 16.72%.

Discussion

Results suggest that for two and three stage sequential tests, test time is minimised when statistical power in the early stages is relatively high, i.e. ∼68% in the first stage for a two stage design, and ∼45% in stage one for a three stage design. This can be attributed to the relationship between statistical power and the ensemble size, which

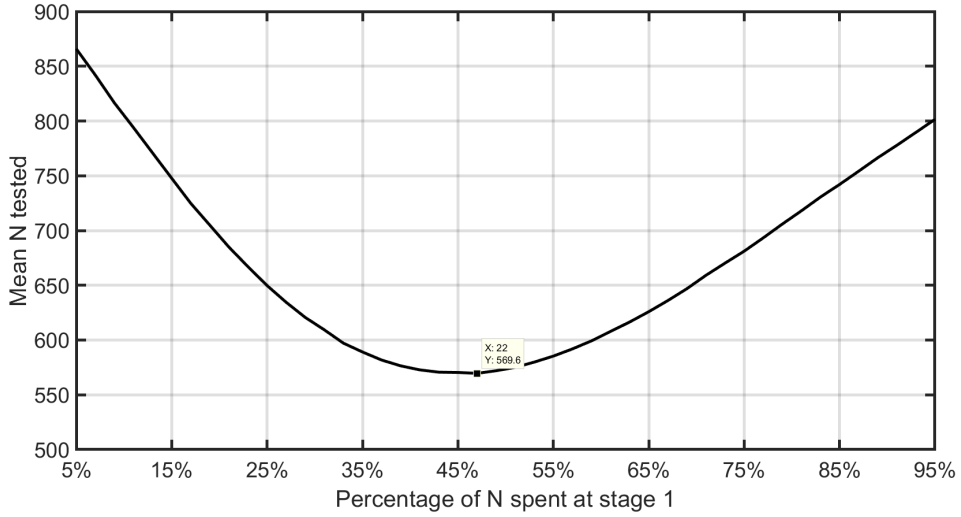


Figure A.17.1: The mean number of samples used for a two stage design (such that the TPR was 0.95), as a function of the percentage of N spent in stage one.

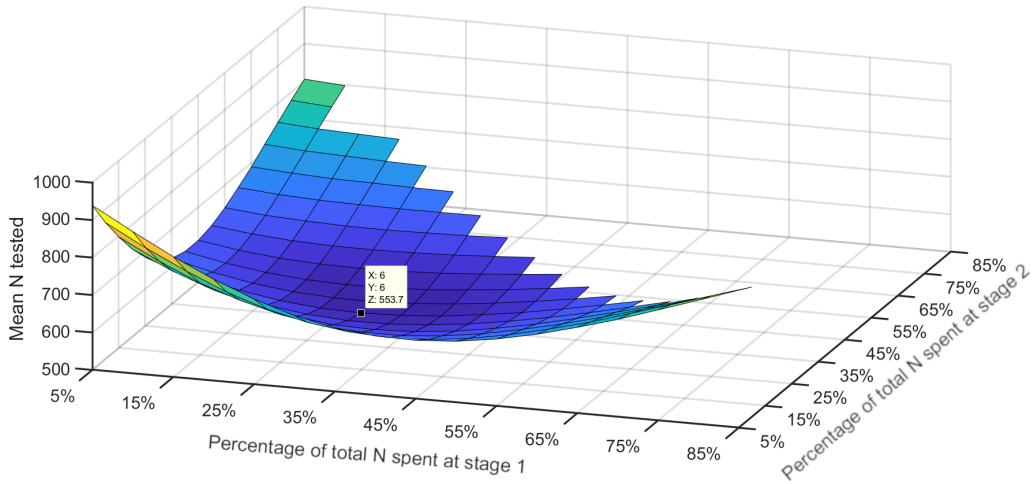


Figure A.17.2: The mean number of samples used for a three stage design (such that the TPR was 0.95), as a function of the percentage of N spent in both stages one and stage two.

is shown in Figure A.17.3 for the single shot test. Note that a much larger increase in N is required in order to increase statistical power from 0.95 to 0.99 (the ensemble size then needs to be increased from ∼800 to ∼1050, or 250 additional samples need to be collected), as opposed to increasing statistical power from 0.5 to 0.55 (in which case the ensemble size should be increased from ∼250 to ∼280, or just 30 additional samples need to be collected). Consequently, increasing the ensemble size for the earlier stages is beneficial only up to a certain point, after which relatively large increases in test time are required for relatively small increases in statistical power. Results from this section also demonstrate that the relationship is close to optimal when the available N is split equally across the K stages. Because the latter greatly simplifies the choice for N_i and is still close to optimal, it was the adopted methodology throughout this thesis. Small

increases in test performance might however still be gained by splitting the available N differently across the K stages.

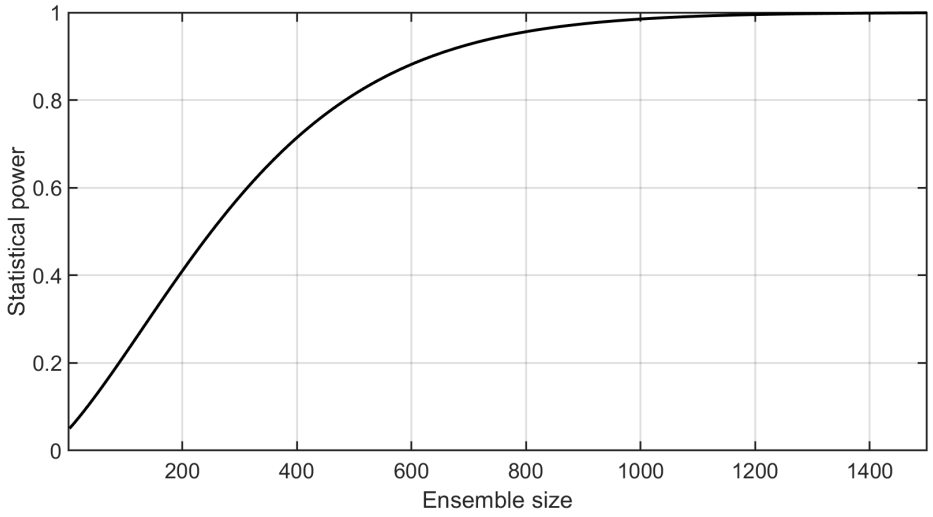


Figure A.17.3: Statistical power for the single shot test (using test parameters defined in this section), as a function of the ensemble size.

Limitations and future work

There are many shortcoming with the approach described in this section. Perhaps most importantly, three assumptions are made regarding the type of noise, the effect size, and the feature set for the Hotelling's T^2 test. Starting with the latter, this section assumed a 2-dimensional feature set, whereas most sections throughout this work use a 25-dimensional feature set. In future work, the test time should be explored using more realistic feature dimensions. With respect to the effect size, an important limitation is that the current approach assumed just a single effect size, whereas in practice, a distribution of effect sizes can be expected (due to varying SNRs across subjects and recordings). It can be envisioned that the approach could be extended to a distribution of effect sizes, e.g. by taking the percentiles of each effect size into account during the optimisation. Finally, the amplitude of the response was assumed to be 0.1, and the noise was assumed to be a Gaussian white process, both of which are unrealistic in real world ABR examinations. Future work might also explore the approach using more realistic background activity, along with real ABR waveforms for representing the response.