UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in Scienze Statistiche

Tesi di Laurea

Test multipli per dati di Elettroencefalografia (EEG) evento-correlati

Relatore
Prof. Livio Finos
Dipartimento di Psicologia

Laureanda
Giuliana Lo Presti
Matricola: 1157171

Anno Accademico 2018/2019

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

La rilevazione del segnale elettroencefalografico, oltre a consentire la diagnosi di diversi disturbi dell'attività cerebrale, ha permesso negli ultimi anni di accrescere sempre di più le conoscenze sull'attività elettrica del cervello. Infatti, a partire dagli anni '80, con la nascita della neuroscienza cognitiva e la sempre crescente accessibilità dei computer, molti studiosi provenienti da diversi ambiti di ricerca si sono interessati all'approfondimento di questa tecnica, portando con sé competenze di varia natura ed espandendone l'utilizzo a nuove questioni. Nonostante oggi siano presenti altre tecniche altrettanto valide, come la fMRI o la PET, l'interesse verso l'elettroencefalografia è da attribuire alla sua scarsa invasività, ai suoi costi contenuti e alla sua alta risoluzione temporale.

Il processo che porta all'acquisizione ed elaborazione del segnale è piuttosto articolato e costituito da varie fasi: il corretto svolgimento di ognuna di queste è essenziale per la buona riuscita dello studio e per l'ottenimento di risultati affidabili. Nella prima parte di questo elaborato, si farà cenno ad alcuni particolari concetti che caratterizzano il segnale EEG. In particolare, al Capitolo 1 si farà un'introduzione al significato fisiologico dello stesso e verrà descritto il disegno sperimentale tipicamente utilizzato in questo contesto. Si farà, inoltre, un breve confronto fra le qualità tipiche dell'EEG e quelle di altre tecniche simili e si introdurranno i dati che verranno analizzati nell'elaborato. Il Capitolo 2 è, invece, dedicato alla descrizione teorica di una fase di fondamentale importanza ai fini della buona riuscita dello studio, ovvero quella di filtraggio e di correzione degli artefatti.

La restante parte dell'elaborato è dedicata all'analisi statistica del segnale. Essa è tipicamente finalizzata al confronto di onde rilevate sotto condizioni diverse e si può svolgere seguendo varie strade, talvolta anche molto diverse fra loro. L'obiettivo di questo elaborato è quello di fornire un quadro generale degli approcci di analisi che possono essere adottati, mettendo in luce gli aspetti positivi e quelli negativi che caratterizzano ognuno di essi. L'esigenza di sviluppare tecniche di analisi molto diverse fra loro nasce dal fatto che nessuna di queste è in grado di rispondere a pieno alle esigenze di analisi. Infatti, in alcuni casi sarà necessario fare riferimento soltanto a una sintesi dei dati, perdendo, dunque, una parte del segnale rilevato: è il caso dei cosiddetti "metodi standard", introdotti al Capitolo 3. Questi prevedono che il confronto fra il segnale nelle diverse condizioni si svolga attraverso lo sviluppo di modelli quali

il modello ANOVA per misure ripetute, il modello a effetti misti o il modello ANOVA multivariata. Al Capitolo 4, invece, si procederà allo studio del segnale in ogni singola rilevazione: se da un lato questo libera l'analisi da scelte a priori che ne possono influenzare i risultati, dall'altro fa sì che si raggiunga un livello di precisione dell'analisi più alto di quello richiesto e rende indispensabile l'utilizzo di tecniche di correzione dell'errore. Al Capitolo 5, l'analisi sarà sviluppata ancora sulle osservazioni singole, ma verrà introdotto un approccio di correzione dell'errore non parametrico, cioè quello basato sulle permutazioni. Il Capitolo 6 è, invece, dedicato alla valutazione degli eventuali cambiamenti nei risultati delle tecniche che verranno sviluppate, dopo aver provato ad attenuare il problema della molteplicità, che accomunerà tutte la procedure che verranno sviluppate.

Capitolo 1

Cosa è l'elettroencefalografia?

L'elettroencefalografia (EEG) è un sistema di registrazione dell'attività elettrica del cervello durante in un certo intervallo di tempo. Essa può essere utilizzata in ambito clinico nella diagnostica di disturbi dell'attività neuronale quali, ad esempio, l'epilessia, gli squilibri del sonno, i tumori, gli ictus o il coma, oppure può essere utilizzata in un contesto sperimentale al fine di ampliare le conoscenze sull'attività di determinate aree funzionali del cervello. Nel primo caso, ci si concentra sull'attività che il cervello svolge in modo costante e indipendentemente da qualsiasi azione venga eseguita dal soggetto, mentre nel secondo caso si cerca di individuare la risposta cognitiva a un particolare evento (interno o esterno al soggetto), quindi si parla di onda ERP (Event-Related Potential o Potenziale evento-correlato). Il contesto in cui si sviluppano le tecniche che verranno approfondite in questo elaborato è di tipo sperimentale, quindi l'oggetto di studio sono proprio le onde ERP.

1.1 Il disegno sperimentale

La registrazione del segnale riguarda solitamente un certo gruppo di soggetti opportunamente selezionati, che vengono sottoposti a due o più stimoli di un certo tipo, ad esempio visivi o uditivi. In questa fase, ogni stimolo viene sottoposto al singolo soggetto più di una volta, secondo uno schema prestabilito che, in alcuni casi, porta ad avere uno stimolo più frequentemente di un altro. In questo modo, per ogni soggetto e per ogni stimolo si avranno diversi segnali di risposta a cui fare riferimento: ognuna di queste sessioni di registrazione del segnale viene detta *trial* o esperimento.

In uno studio di questo tipo, l'interesse può essere quello di capire quali zone della corteccia si attivano in una determinata circostanza, quanto tempo passa prima di poter osservare tale attività, la sua intensità o la sua durata complessiva. A seconda della particolare caratteristica che è di interesse il disegno sperimentale assumerà particolari sfumature. Ad esempio, se si sottopone ai soggetti uno stimolo visivo, le immagini possono differenziarsi fra loro non solo in base al contenuto ma anche in base alla

loro luminosità, alle dimensioni dell'oggetto in esse contenuto ecc. Inoltre, può essere utile far svolgere al soggetto un determinato compito come, ad esempio, fargli premere un pulsante subito dopo aver percepito un particolare stimolo. Un esempio di disegno sperimentale è quello che ha portato alla scoperta del cosiddetto *effetto Stroop* [1], dove l'interesse è stato posto sui tempi di risposta al segnale. In esso, sono state mostrate ai soggetti partecipanti diverse immagini contenenti delle parole scritte in diversi colori e il compito che questi dovevano svolgere era quello di nominare il colore dell'inchiostro utilizzato. Dunque, in questo caso si possono identificare due condizioni sperimentali: quella in cui la parola scritta corrisponde al colore dell'inchiostro, e quella in cui vi è una "incompatibilità", cioè la parola scritta identifica un colore diverso da quello dell'inchiostro utilizzato. Sotto quest'ultima condizione, si è osservato un effettivo rallentamento da parte dei soggetti nel nominare il colore dell'inchiostro, ma soltanto grazie all'utilizzo del segnale EEG è stato possibile asserire che tale ritardo non riguarda la fase di percezione dell'immagine, ma quelle successive di elaborazione ed espressione dell'informazione.

1.2 Origini neurali del segnale elettrico

La misurazione del segnale si svolge per mezzo degli *elettrodi*, cioè dischi metallici (o di un altro materiale conduttivo) che, con l'aiuto di un gel conduttivo che connette la pelle all'elettrodo, riescono a rilevare la parte più superficiale dell'attività elettrica presente nella corteccia. Il loro posizionamento sullo scalpo segue particolari schemi, universalmente utilizzati, ognuno dei quali è caratterizzato da un certo numero di elettrodi. Infatti, si può utilizzare un totale di 32, 64, 128 o, in alcuni casi molto particolari, 256 elettrodi. Chiaramente, utilizzare un maggior numero di elettrodi consente di ottenere una specificità spaziale maggiore, molto utile in alcune tipologie di studio, ma allo stesso tempo può causare problemi come la formazione di "ponti elettrici" fra elettrodi vicini attraverso il gel conduttivo, o ridurre la probabilità che il ricercatore riesca ad individuare con successo una cattiva connessione fra i singoli elettrodi e lo scalpo.

1.2.1 Elettrodi attivi e di riferimento

Prima di poter comprendere che tipo di attività riescono a rilevare gli elettrodi e come questi interagiscono fra loro, è necessario approfondire alcuni concetti chiave. Infatti, si definisce *elettricità* un flusso di cariche che scorre attraverso un certo mezzo di conduzione; il numero di cariche che fluiscono in un certo conduttore in un determinato intervallo temporale è detto *corrente*, mentre la pressione elettrica, ovvero il potenziale della corrente che scorre da un punto ad un altro, è detta *voltaggio* o *potenziale*. Quando

si parla di segnale elettrico, si fa riferimento in particolar modo al voltaggio, misurato in Volt o, in questo contesto specifico, in microvolt (μV). Si intuisce, dunque, che il voltaggio non è un concetto che riguarda un singolo punto, cioè un neurone o un aggregato di neuroni, ma riguarda due punti. Per questo motivo, per misurare il voltaggio nell'area dove si sceglie di posizionare l'elettrodo, è necessario scegliere un punto, ad una certa distanza dal primo, dove posizionare un secondo elettrodo che possa essere preso come riferimento. Per ottenere una misura assoluta del voltaggio nell'area di interesse, l'ideale sarebbe scegliere un punto di riferimento elettricamente neutro, tuttavia risulta impossibile trovare un punto della testa o del corpo totalmente neutro dal punto di visto elettrico, quindi ci si accontenta di punti come il naso, i lobi delle orecchie, il mento o il collo che sono fra i più vicini alla condizione di neutralità. Dunque, un voltaggio esprime la differenza di attività fra due siti e per questo motivo, come si vedrà più avanti, è molto importante rimuovere eventuali fonti di rumore, non solo dalla zona di interesse ma anche da quella di riferimento. Per evitare di aggiungere al segnale il rumore rilevato nel punto di riferimento, quando si ha la possibilità di posizionare degli elettrodi in tutta (o quasi) la superficie della testa, si può scegliere di utilizzare come riferimento il voltaggio medio di tutti gli elettrodi, evitando, in questo modo, di ridurre la precisione del segnale registrato.

1.2.2 Potenziale di azione e post-sinaptico

L'attività elettrica cerebrale si può distinguere in due fasi: il potenziale di azione e il potenziale post-sinaptico. Per comprenderne la differenza si riporta in Figura 1.1 la forma tipica della cellula che caratterizza il sistema nervoso: il neurone.

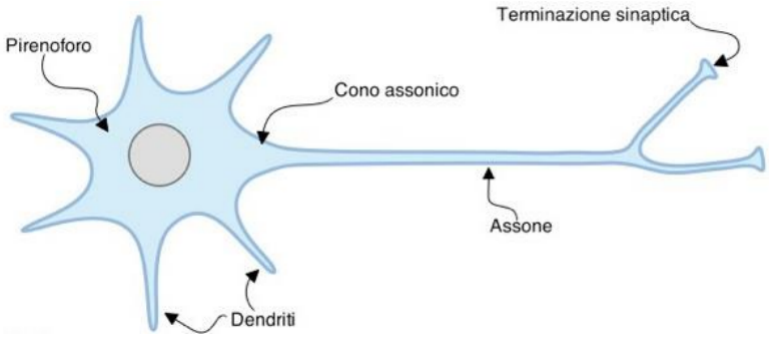

Figura 1.1: Rappresentazione di un neurone

I potenziali d'azione sono picchi di voltaggio discreti che viaggiano lungo l'assone, la durata dell'impulso è molto breve, cioè di circa 1 o 2 ms, ed è tanto più forte quanto più potenziali d'azione vengono rilasciati nell'unità di tempo. L'unico modo per registrare questo tipo di segnale dalla superficie sarebbe quello di misurare l'attività di più neuroni contemporaneamente. Tuttavia, affinché ciò sia possibile è necessario

che i potenziali si manifestino esattamente nello stesso istante e in modo parallelo, altrimenti questi si annullerebbero a vicenda generando un segnale molto più piccolo di quanto sia realmente. Poiché in una situazione reale è molto difficile riscontrare una corrispondenza esatta, il potenziale d'azione risulta misurabile soltanto attraverso elettrodi da posizionare internamente, quindi a diretto contatto con l'area di interesse. Pertanto, questo tipo di segnale può essere rilevato su soggetti vivi soltanto nel caso in cui questi vengano sottoposti ad operazioni chirurgiche.

Il potenziale post-sinaptico, invece, si propaga attraverso i dendriti, subito dopo la manifestazione di un potenziale d'azione. La sua durata è molto più lunga, infatti può andare da decine a centinaia di millisecondi, e può essere registrato con molta più facilità rispetto ai potenziali d'azione. In questa fase, per ogni neurone viene generato un dipolo elettrico, cioè una coppia di cariche, una positiva e una negativa, separati da una certa distanza. Il dipolo generato da un singolo neurone è troppo piccolo per essere misurato ma, in certe condizioni, i dipoli provenienti da diversi neuroni si sommano fra loro, generando un potenziale più forte, che riesce ad essere colto dagli elettrodi posizionati sullo scalpo quasi istantaneamente. Infatti, i potenziali sommati fra loro vengono condotti ad una velocità quasi pari a quella della luce attraverso le meningi, il cranio e la pelle. In particolare, per poter ottenere tale somma, è necessario che il potenziale venga emesso approssimativamente allo stesso momento in migliaia di neuroni e che i dipoli provenienti da ogni neurone siano fra loro allineati perché se questi formano un angolo fino a 90° si ottiene un annullamento parziale, che diventa totale a 180°.

Dati i diversi sistemi di propagazione dei due segnali, la tecnica dell'elettroencefalografia consente di rilevare unicamente il segnale post-sinaptico, pertanto, quando di seguito si parlerà genericamente di "segnale elettrico", si farà un implicito riferimento a quest'ultimo.

1.3 Le componenti di potenziale evento-correlato (ERP)

1.3.1 Attività di fondo e onde ERP

Il voltaggio rilevato nell'EEG, quindi il potenziale post-sinaptico, è a sua volta ottenuto dalla somma di due tipologie di attività neuronale. La prima è quella che si sviluppa in maniera costante e che provoca nel segnale movimenti oscillatori ciclici e continui. Ogni onda di questo tipo può essere rintracciata in una particolare area della corteccia ed è caratterizzata da una particolare frequenza. Ad esempio, l'onda più evidente è quella *alpha*, con una frequenza che va da 8 a 13 Hz, ma si hanno anche le onde *beta* (13-30 Hz), *delta* (<4 Hz), *theta* (4-8 Hz) ecc.

La seconda tipologia di attività neuronale è quella descritta dalle onde ERP. Essa è di particolare interesse in un contesto sperimentale e riguarda la variazione di potenziale che si può osservare in risposta ad un particolare stimolo. Si tratta di un segnale transitorio e relativamente breve, che consente di individuare il funzionamento di una particolare area della corteccia durante un certo processo cognitivo o di evidenziarne delle particolari tempistiche.

La distinzione fra questi due generi di attività esiste, tuttavia, soltanto al livello logico ma non pratico, infatti il segnale rilevato tramite gli elettrodi esprime semplicemente la somma di tutte le singole onde latenti, o *componenti*, che caratterizzano l'attività di una determinata area della corteccia. Questo è uno dei più grandi limiti della tecnica EEG perché, dal momento che l'interesse ricade principalmente sulle onde ERP, la parte di segnale presente indipendentemente da particolari eventi viene considerata come del rumore da cui il segnale deve essere filtrato, ma non si può mai essere certi di ottenere un segnale adeguatamente depurato.

1.3.2 Caratteristiche delle componenti ERP

Il segnale ERP viene rappresentato come un'onda formata da una serie di picchi positivi e negativi di diversa entità, che prendono il nome di *componenti*. Anche se in alcuni casi particolari le componenti presentano forme e proprietà variabili, in genere ogni componente si distingue in base a:

- **Polarità**: positività o negatività del segnale;
- **Latenza**: tempo intercorso dall'inizio dello stimolo alla variazione del potenziale;
- **Ampiezza**: valore più alto di voltaggio raggiunto dalla componente;
- **Distribuzione sullo scalpo**: consente di vedere in quali zone la componente è maggiormente presente.

Ogni componente viene identificata dalla lettera P o N, a seconda che sia caratterizzata da un segnale positivo o negativo, seguito da un numero intero che ne rappresenta l'ordine in cui la componente si presenta rispetto a tutte le altre, oppure i millisecondi che passano prima che questa si manifesti.

In Figura 1.2 si riporta un esempio di come viene rappresentata graficamente un'onda ERP caratterizzata da una componente con polarità negativa, la cui ampiezza è evidenziata in rosso.

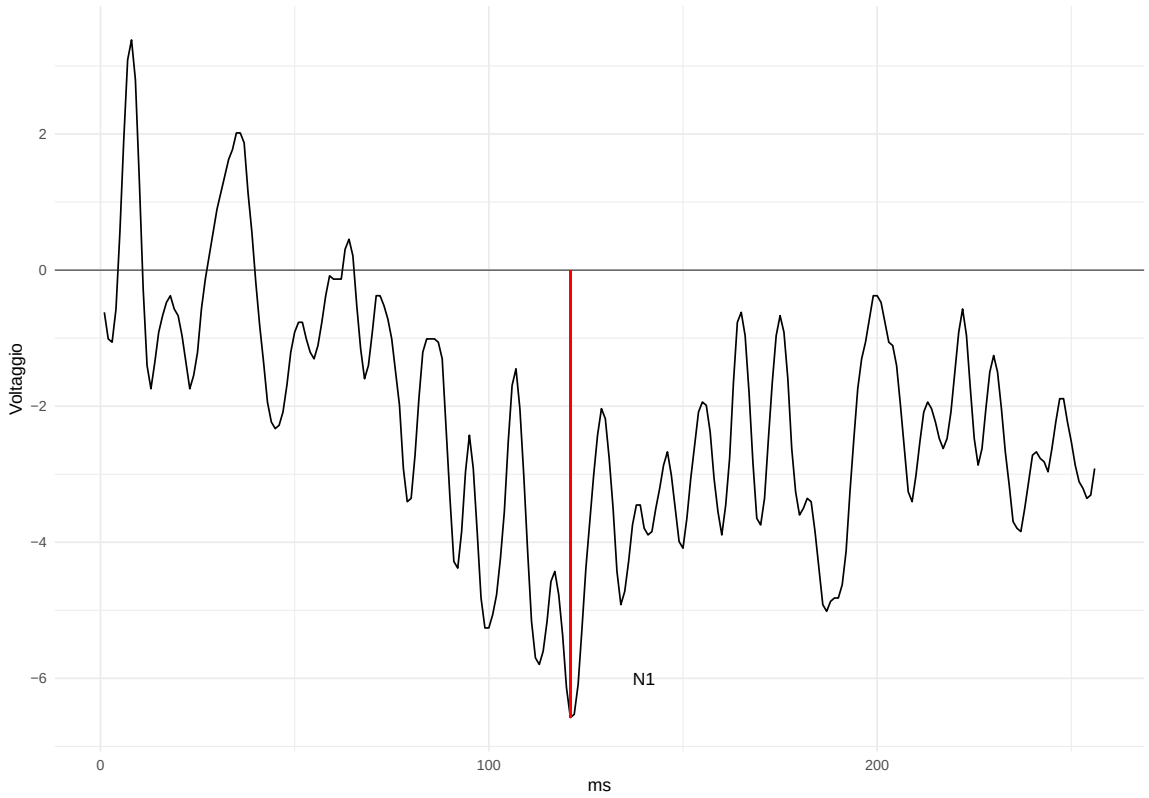

Figura 1.2: Esempio di onda ERP

1.4 Confronto con altre tecniche

Come si è visto finora, quella dell'elettroencefalografia è una misura diretta dell'attività elettrica del cervello. Ad essa possono sostituirsi o affiancarsi altre tecniche dette *indirette* perché si concentrano sull'attività emodinamica della corteccia che segue quella elettrica. Si tratta, ad esempio, della PET (Positron Emission Tomography) o della fMRI (Functional Magnetic Resonance Imaging), che sono in grado di fornire delle immagini della corteccia cerebrale ad alta risoluzione (dell'ordine di millimetri) dove vengono evidenziate le zone cerebralmente attive. Le tecniche di misurazione dirette e indirette vengono spesso affiancate perché risultano complementari in termini di risoluzione spaziale e temporale. Infatti, alla buona risoluzione spaziale delle tecniche indirette non si accompagna una risoluzione altrettanto buona a livello temporale, proprio a causa della lentezza del segnale emodinamico che viene registrato. Infatti, la velocità di registrazione di tale segnale è di diversi secondi. Al contrario, nel caso dell'EEG, grazie alla velocità della risposta elettrica, è possibile registrare il segnale ogni millisecondo, ma non si riesce ad ottenere una rappresentazione spaziale dello stesso altrettanto accurata. Infatti, nonostante si possa arrivare ad utilizzare anche più di 100 elettrodi, questi risulteranno comunque pochi rispetto alle migliaia di neuroni che

si attivano all'interno della corteccia. Dunque, un altro limite della tecnica EEG è la bassa risoluzione spaziale, che rende attualmente molto difficile localizzare in modo attendibile il segnale registrato.

Nonostante questo problema, l'EEG è molto utilizzata grazie ai suoi costi relativamente bassi rispetto alle altre tecniche, alla sua scarsa invasività e al fatto che offre la possibilità di misurare il segnale in diverse situazioni sperimentali.

1.4.1 Presentazione dei dati

I dati a cui si farà riferimento nell'intero elaborato, sono quelli utilizzati da Zhang et al. (1995). Si tratta di un ampio studio in cui l'obiettivo era quello di valutare la correlazione dei segnali EEG con la predisposizione all'alcolismo. Le misurazioni sono state effettuate tramite 64 elettrodi posizionati sui punti standard (Standard Electrode Position Nomenclature, American Electroencephalographic Association 1990), riportati in Figura 1.3.

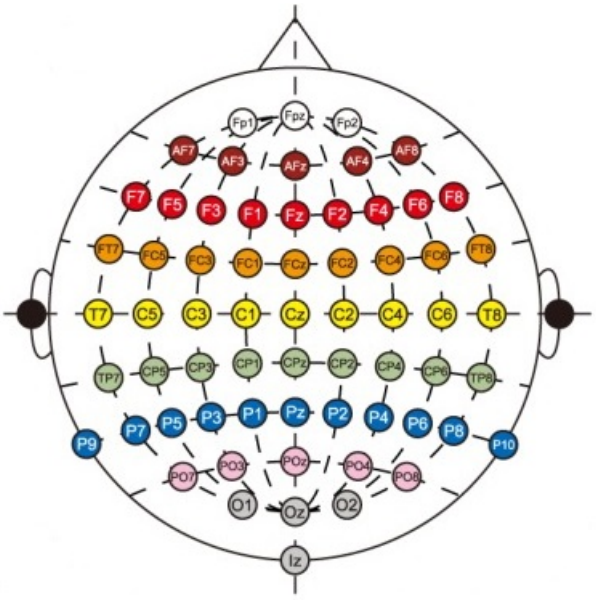

Figura 1.3: Sistema standard a 64 elettrodi

Tipicamente, il segnale EEG viene convertito in voltaggio in una sequenza discreta di istanti temporali chiamati *campioni*. In questo caso, il tasso di campionamento, ovvero il numero di campioni considerati per ogni secondo è di 256 Hz, e il tempo di osservazione complessivo per ogni trial è di 1 secondo. Dunque, per ogni trial si hanno 256 osservazioni del segnale in istanti temporali successivi. Nei dati originari, i gruppi di soggetti sono due, alcolisti e controlli, e gli stimoli sono di tipo visivo e provenienti dal gruppo di immagini di Snodgrass et al. (1980). Alcuni soggetti sono stati sottoposti soltanto ad uno stimolo, S1, mentre altri sono stati sottoposti a due stimoli, S1 e S2.

In quest'ultimo caso, gli stimoli S1 e S2 possono essere identici (condizione *match*) o diversi (condizione *no match*). I soggetti sono 122 e per ogni soggetto sono stati svolti 120 trial, in ognuno dei quali è stato sottoposto un certo stimolo. Per ragioni di praticità, per questo elaborato si è scelto di utilizzare soltanto i dati relativi a 10 soggetti del gruppo di controlli, sotto la condizione *no match*. Inoltre, per ogni soggetto sono stati considerati soltanto 10 dei 120 trial effettuati.

In Figura 1.4 si riporta, a titolo di esempio, una delle onde su cui si è lavorato: quella ottenuta dalla media dei 10 esperimenti considerati per il soggetto 1, rilevati dall'elettrodo posizionato in FP1 sotto la condizione S1.

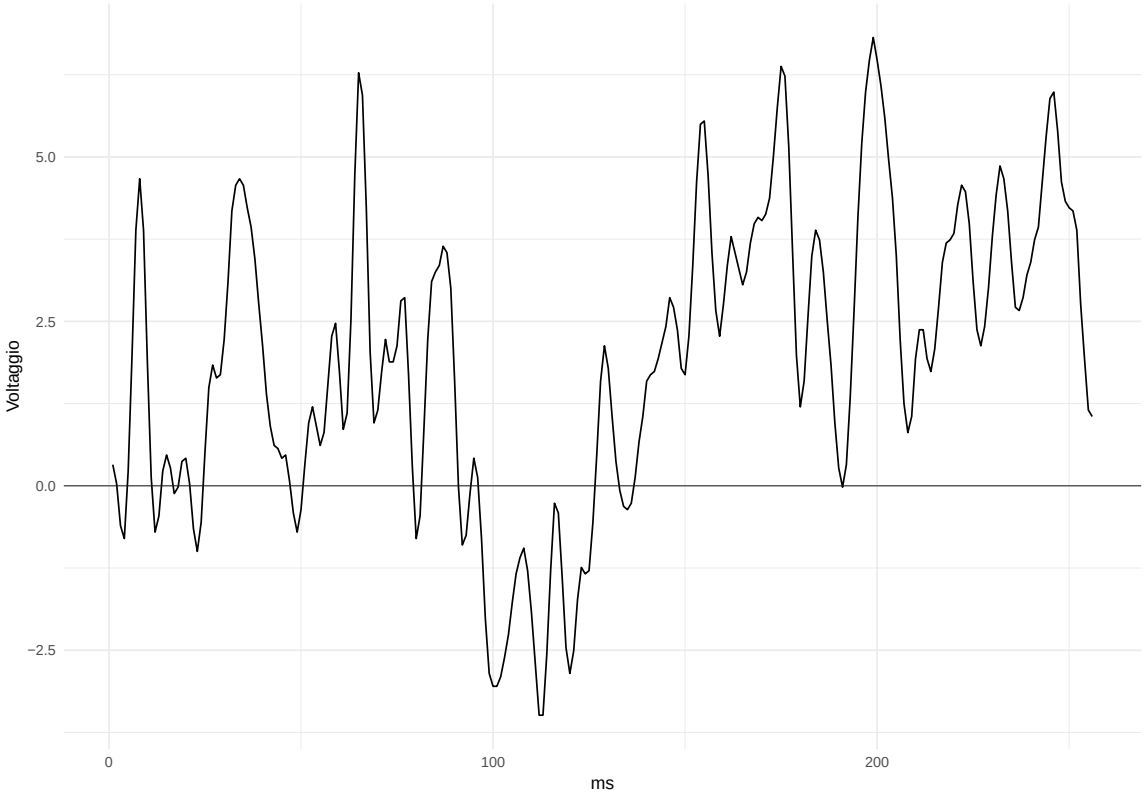

Figura 1.4: Segnale in FP1 sotto la condizione S1 rilevato sul soggetto 1

Di seguito si approfondiranno le tecniche di preprocessamento dei dati, necessarie per ottenere risultati attendibili e facilmente interpretabili. Sui dati che si è scelto di utilizzare, la fase di preprocessamento è, in parte, già svolta, pertanto ci si è limitati a calcolare il segnale medio fra tutti i trial riferiti ad uno stesso soggetto. Gli obiettivi e la logica sottostante a questa scelta verranno approfonditi al Capitolo 2.

Capitolo 2

Prepocessamento e Filtraggio dei dati

Prima di sviluppare qualsiasi analisi del segnale EEG, per poter ottenere risultati attendibili è necessario svolgere il cosiddetto *preprocessamento* dei dati, costituito da una serie di tecniche finalizzate principalmente alla riduzione del *rumore*. Il *rumore* è quella parte di segnale che viene colta dagli elettrodi ma che non è originata direttamente dal processo cognitivo di interesse e, se trascurato, può influenzare i dati a tal punto da portare a risultati del tutto fuorvianti. Il rumore può essere dovuto, ad esempio, a una certa attività elettrica presente nell'ambiente in cui viene svolto l'esperimento, che potrebbe provenire delle apparecchiature utilizzate, o al potenziale della pelle che viene rilevato insieme a quello dalla corteccia, o allo spostamento di alcuni elettrodi che provoca una forte variazione del segnale. In altri casi, invece, il rumore è costituito dall'attività cerebrale stessa, come ad esempio le onde *alpha* o quelle *beta* che vengono registrate insieme al segnale ERP. Il preprocessamento porta spesso ad avere delle forti variazioni nel segnale originario, quindi una certa distorsione dello stesso. Per questo motivo è sempre bene cercare di ottenere dati quanto più puliti è possibile fin dal principio, in modo da limitare l'utilizzo delle tecniche di preprocessamento.

2.1 Il filtraggio

Una delle principali tecniche di preprocessamento è il filtraggio, che può essere definito come la manipolazione di un gran numero di dati al fine di eliminare una parte del rumore che questi contengono. Si tratta di un'operazione tanto necessaria quanto delicata, poiché è molto facile introdurre una certa distorsione nei dati o degli artefatti, pertanto nel filtrare il segnale bisogna fare molta attenzione sia all'adeguatezza del metodo utilizzato, sia a verificare che la distorsione introdotta non sia tanto forte da invalidare totalmente i risultati.

I filtri possono essere distinti in due grandi categorie, in base alla fase di registrazione del segnale in cui vengono applicati: i filtri *analogici* sono utilizzati durante la fase di acquisizione del dato, e i filtri *offline* operano sulle onde ERP già espresse in formato digitale. Infatti, la registrazione del segnale EEG si compone di due fasi preliminari, l'amplificazione del segnale e la sua digitalizzazione.

La prima fase è resa necessaria dal fatto che il segnale EEG originario è talmente piccolo che senza l'amplificazione di un fattore che va da 10000 a 50000, non potrebbe essere colto. Qui si ha un segnale analogico e continuo, e uno dei maggiori rischi in questa fase è quello di amplificare, oltre al segnale EEG stesso, anche segnali elettrici esterni provenienti dall'ambiente, rendendo ancora più forte la distorsione del segnale EEG che ne deriva. Per questo motivo, una parte del filtraggio viene svolta durante il processo di amplificazione.

Una volta amplificato il segnale, si procede con la sua digitalizzazione, cioè le fluttuazioni del voltaggio EEG vengono convertite in rappresentazioni numeriche discrete e in formato digitale, così da poter essere immagazzinate in un computer. Questo processo è svolto per mezzo del cosiddetto *convertitore analogico-digitale* (ADC), caratterizzato da una certa risoluzione che influirà nella qualità finale del processo di digitalizzazione. La risoluzione esprime il numero di valori di voltaggio diversi fra loro che possono essere codificati. Ad esempio, in molti sistemi di digitalizzazione delle EEG, gli ADC hanno una risoluzione di 12 bit: questo vuol dire che il ADC può codificare $2^{12} = 4096$ diversi valori di voltaggio, mentre valori intermedi sono arrotondati al numero più vicino. Supponendo di avere registrato un voltaggio compreso fra $-5V$ e $5V$, una risoluzione a 12 bit porterà a codificare un segnale di $-5V$ come 0 e un segnale di $5V$ come 4096, mentre tutti i voltaggi intermedi (V) verranno codificati come

$$\frac{4096(V+5)}{10}$$

In questo modo, potrebbe accadere che alcuni valori del segnale superino i limiti dell'intervallo imposto dalla risoluzione del ADC (in questo caso 0 e 4096), quindi questi verranno identificati con 0 se il segnale è negativo o con 4096 se il segnale è positivo. La possibilità che ciò accada è tanto più grande quanto più si aumenta il guadagno dell'amplificatore, di conseguenza il ricercatore dovrà scegliere un guadagno dell'amplificatore non troppo grande per ridurre quanto più è possibile il numero di voltaggi che superano l'intervallo del ADC, ma neanche troppo piccolo perché si potrebbe giungere ad una forte perdita di risoluzione.

2.2 Alcune categorie di filtri

In generale, i filtri maggiormente utilizzati sono:

- **low-pass**: attenuano le frequenze alte, lasciando passare quelle più basse;
- **high-pass**: attenuano le frequenze basse, lasciando passare quelle più alte;
- **bandpass**: attenuano sia le frequenze alte che quelle basse, lasciando passare soltanto intervalli di frequenza intermedi;
- **notch**: attenuano specifiche bande di frequenza ristrette, lasciando passare tutte le altre.

Come si può intuire, i filtri vengono identificati in base al tipo di frequenza su cui si concentrano, e questo è in parte dovuto al fatto che spesso i rumori che si vogliono eliminare presentano frequenze molto diverse da quelle del segnale EEG che si vorrebbe rilevare, cioè molto più piccole o molto più grandi. Tuttavia, non sempre è facile risolvere il problema del rumore con questo tipo di filtri, infatti vi sono certe fonti di distorsione, come l'onda *alpha* di cui si è accennato al Capitolo 1, che hanno una frequenza molto simile al segnale EEG e che, quindi, raramente riescono ad essere isolate da esso senza introdurre forti distorsioni.

Spesso i filtri low-pass e high-pass vengono utilizzati in successione, infatti si utilizza il primo per assicurarsi che il tasso di campionamento sia pari ad almeno due volte la frequenza più grande del segnale, altrimenti non sarebbe possibile l'acquisizione stessa del segnale, e il secondo per assicurarsi che potenziali lenti e non neurali non portino il segnale al di fuori dell'intervallo in cui lavora il convertitore. Inoltre, l'impostazione tipica dei filtri dovrebbe essere fra $\frac{1}{3}$ e $\frac{1}{4}$ del tasso di campionamento per il filtro low-pass e pari a 0.01 Hz per quello high-pass. Come qualsiasi tipo di filtro, anche quelli low-pass e high-pass causano una certa distorsione del segnale, più o meno forte in base a come vengono utilizzati. In particolare, entrambi hanno l'effetto di "spalmare" l'onda, distorcendo i tempi di inizio e di fine delle componenti, ma con il filtro high-pass questa sbavatura dei tempi causa anche una serie di oscillazioni verso l'alto e verso il basso aggiungendo, oltre alla distorsione dei tempi, anche delle oscillazioni artificiali. Per questo motivo, come già detto, è sempre meglio cercare di avere dati quanto più puliti è possibile e limitare l'uso dei filtri.

Le figure riportate di seguito [1] forniscono un esempio di come si presenta un'onda grezza contaminata da un rumore di frequenza pari a 60Hz (Figura 2.1) e di come si presenta la stessa onda filtrata da tale rumore (Figura 2.2).

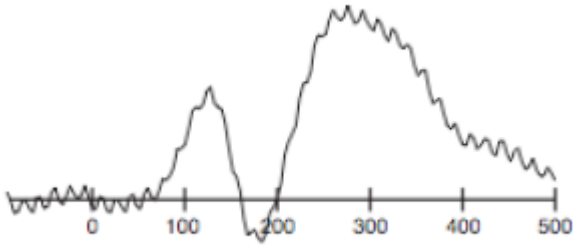

Figura 2.1: Esempio di onda non filtrata

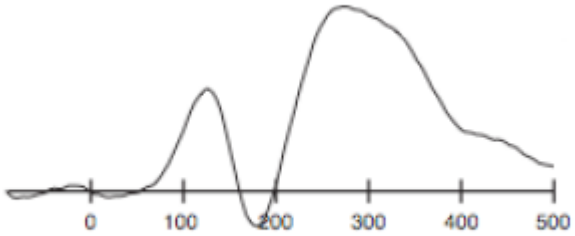

Figura 2.2: Esempio di onda filtrata

2.3 Correzione ed esclusione degli artefatti

Il secondo sistema per ottenere una maggiore pulizia dei dati è quello di correggere o scartare totalmente gli artefatti. Si definiscono artefatti alcuni movimenti dell'onda che potrebbero a tutti gli effetti sembrare delle componenti e che, però, non sono generate da un'effettiva attività neuronale ma, ad esempio, dal potenziale della pelle, dal movimento e il battito degli occhi o dall'attività muscolare. Gli artefatti rappresentano un problema per varie ragioni: spesso sono molto più grandi dei segnali ERP, possono presentarsi in modo sistematico invece che casuale, oppure possono presentarsi maggiormente sotto certe condizioni piuttosto che in altre. Un caso particolarmente problematico, poi, è quello in cui questi artefatti si presentano contemporaneamente all'evento di interesse, perché in tale situazione risulta molto difficile scindere la risposta allo stimolo da ciò che è, appunto, un artefatto. In Figura 2.3 sono rappresentati alcuni esempi di come si presentano le onde EEG in presenza di tre fra le più comuni tipologie di artefatti: un movimento brusco del paziente (A), l'interferenza del battito del cuore (B) e la sudorazione della pelle (C).

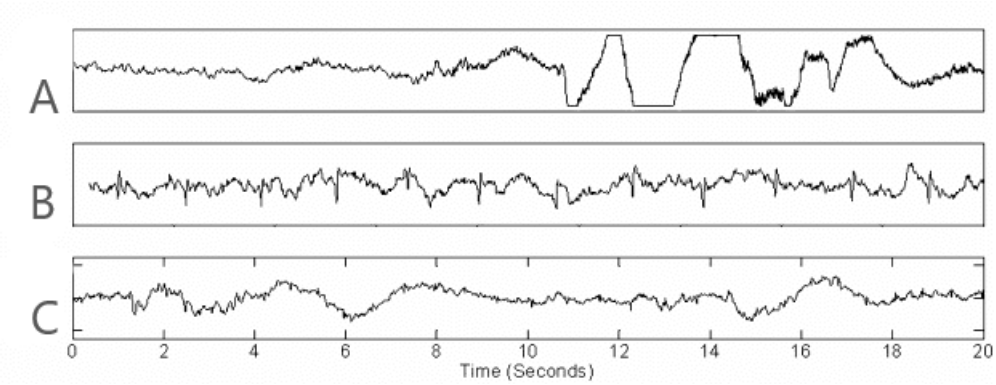

Figura 2.3: Esempi di artefatti

Di seguito verranno illustrati i due approcci più utilizzati per risolvere il problema appena illustrato: l'esclusione e la correzione degli artefatti.

2.3.1 Esclusione degli artefatti

Secondo tale approccio, si cerca di capire se ogni singola onda registrata possa essere caratterizzata da un certo rumore continuo e variabile, e in caso affermativo l'intera onda viene totalmente esclusa dall'analisi. Si procede individuando un certo valore soglia: se il segnale supera tale valore, l'intera onda è considerata influenzata da artefatti, quindi viene esclusa, in caso contrario l'onda viene considerata libera da artefatti rilevanti, quindi viene mantenuta. In questo contesto è possibile individuare quattro situazioni:

- *Scoperta corretta*: l'artefatto è presente e la procedura riesce ad individuarlo;
- *Scoperta errata*: l'artefatto è presente ma la procedura non riesce a coglierlo;
- *Rifiuto corretto*: l'artefatto non è presente e la procedura non coglie nessun artefatto;
- *Rifiuto errato*: l'artefatto non è presente ma secondo la procedura c'è un artefatto.

Il numero di scelte corrette e quello di scelte errate è strettamente legato alla soglia scelta. Infatti, se si sceglie una soglia più bassa il numero di scoperte corrette sarà più grande, ma allo stesso tempo aumenterà anche il numero di rifiuti errati e viceversa. Per aumentare la sensibilità del metodo e renderlo meno dipendente dalla soglia scelta si può provare a migliorare lo strumento con cui si differenzia la presenza o l'assenza degli artefatti. Quindi, si può sviluppare un procedimento a due stadi: nel primo, al posto di fare riferimento direttamente ai dati, si può applicare ad essi una certa funzione, ottenendo un valore specifico che, nel secondo stadio, verrà confrontato con il valore soglia. Ad esempio, si può pensare di misurare la differenza fra il voltaggio

minimo e quello massimo dell'onda e confrontarla con la stessa differenza calcolata in un certo voltaggio utilizzato come soglia. La scelta specifica di quale misura utilizzare è legata al tipo di artefatto che si vuole individuare.

La scelta della soglia rimane, comunque, rilevante ed è generalmente affidata al ricercatore stesso. In particolare, egli, in base anche alla propria esperienza, può scegliere se stabilire una soglia unica per tutti i soggetti oppure se definire una soglia specifica per ogni soggetto. Questa seconda possibilità è dettata dal fatto che, spesso, per ogni soggetto un certo tipo di artefatto produce una deviazione del voltaggio di forma e dimensione diversa rispetto a tutti gli altri, quindi utilizzare una soglia unica non produrrebbe risultati ottimali.

La totale esclusione degli artefatti, tuttavia, rimane sempre un approccio piuttosto grezzo che, in alcuni casi, può portare a lavorare su campioni non rappresentativi. Per questo motivo, quando è possibile, si preferisce correggere gli artefatti individuati.

2.3.2 Correzione degli artefatti

In alcuni casi, l'esperimento viene disegnato in modo tale che la generazione degli artefatti sia del tutto inevitabile. È il caso, ad esempio, di esperimenti che per costruzione prevedono il movimento o lo sbattimento degli occhi: questo genera un'attività molto forte non solo nelle zone vicine all'occhio ma in tutto lo scalpo, dunque risulta impossibile eliminare tutti gli esperimenti in cui è presente questo tipo di deviazione del voltaggio. Per questo motivo, talvolta, è preferibile, se non addirittura necessario, sviluppare delle tecniche più elaborate rispetto alla semplice esclusione degli artefatti. Infatti, le tecniche di correzione degli artefatti si propongono di individuare la parte di voltaggio dovuta ad attività diverse da quella di interesse come, ad esempio, lo sbattimento degli occhi e di eliminare solo questa parte di voltaggio. Dunque, si tratta di un'operazione strettamente legata alla tipologia di artefatto che si intende eliminare.

Questa secondo approccio, seppur meno grezzo della semplice esclusione, ha vari inconvenienti, fra cui il più importante è il fatto che rischia di distorcere in modo significativo le onde ERP rilevate, rendendo molto difficile l'interpretazione dei risultati.

2.4 Il calcolo della media generale

Un ultimo e importante contributo alla riduzione del rumore è dato dal calcolo di una media del segnale rilevato nei diversi trial relativi allo stesso soggetto e condizionatamente all'istante temporale in cui viene svolta la rilevazione. La logica dietro questa operazione si basa sull'ipotesi che l'onda ERP osservata sia data dalla somma dell'onda ERP effettiva più un certo rumore casuale. Ci si basa, inoltre, sulle assunzioni che l'onda ERP reale sia identica in ogni esperimento, che il rumore sia una variabile

casuale con media pari a 0 e che sia incorrelato con il processo cognitivo di interesse. Dunque, calcolare la media di un certo numero di esperimenti implica ottenere un'onda che si avvicina all'onda reale, rendendo l'errore tanto più piccolo quanto più alto è il numero di esperimenti che si utilizzano nel calcolo della media. In particolare, siano R l'ammontare di rumore nel segnale e N il numero di esperimenti considerati, la grandezza del rumore nella media è pari a $(1/\sqrt{N})R$. Le assunzioni fatte sono chiaramente irrealistiche ma, nella maggior parte dei casi, il fatto che queste non vengano rispettate non rappresenta un problema.

In Figura 2.4 [1] sono considerati 8 trial: sulla sinistra si riportano le onde ERP ottenute sui singoli trial, mentre sulla destra sono riportate le onde medie ottenute su un numero crescente di trial, partendo da un solo trial fino ad includerli tutti e 8. L'immagine consente non solo di capire come l'onda media riesca a cogliere i tratti caratteristici più forti delle onde singole, come la presenza di un voltaggio positivo fra i 300 e i 600 ms circa, ma anche come la curva media diventi sempre più "liscia" ogni volta che un trial viene aggiunto al calcolo della media.

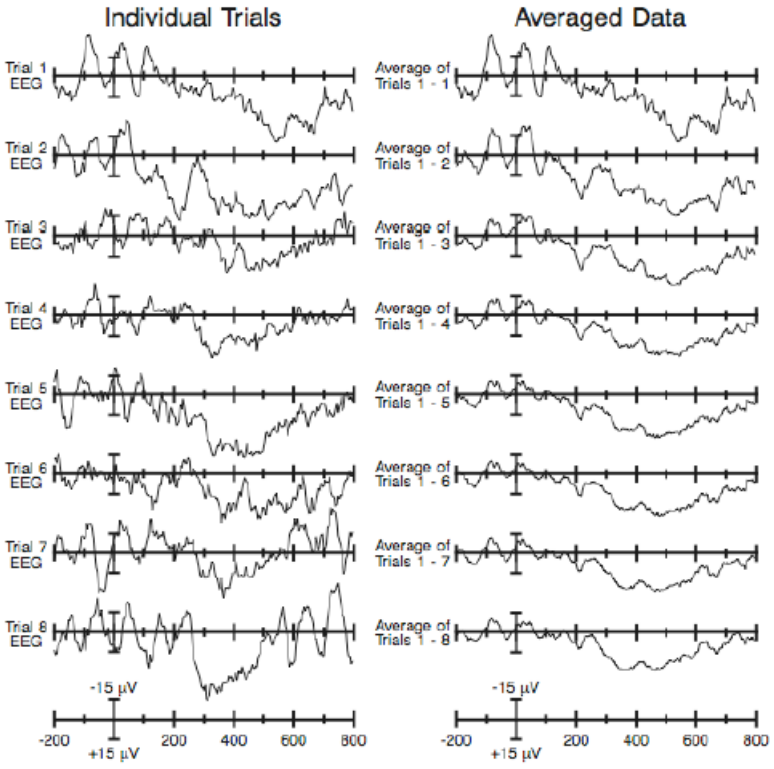

Figura 2.4: Confronto fra i singoli trial e l'onda media

Per particolari forme delle onde, il calcolo di una media generale può portare a risultati distorti. È il caso, ad esempio, di quando le onde osservate su soggetti diversi hanno latenze diverse o di quando le risposte a due stimoli successivi si sovrappongono fra loro. Nel primo caso, l'onda media può cambiare forma rispetto a quelle originarie,

portando ad ipotizzare la presenza di componenti che in realtà non ci sono o hanno caratteristiche diverse da quelle reali. Invece, si può avere una sovrapposizione delle risposte a due stimoli successivi quando questi vengono sottoposti al paziente secondo intervalli temporali non costanti e piuttosto brevi. Questo può facilmente portare ad una errata interpretazione dei risultati, dal momento che l'onda media risulterà più "spalmata", quindi con picchi meno forti rispetto alle onde singole. L'effetto di questa situazione è simile a quello che si ottiene con il filtro low-pass applicato sulla singola onda originaria. La sovrapposizione è tanto più forte quanto più piccolo è il distacco temporale fra i due stimoli ed è particolarmente problematica quando si presenta in misura diversa fra una condizione e l'altra. Dunque, una soluzione per minimizzare gli effetti di tale sovrapposizione può essere quella di utilizzare un intervallo di tempo fra due stimoli consecutivi quanto più ampio è possibile, in modo da rendere la sovrapposizione molto piccola. In alternativa si può pensare di utilizzare un filtro high-pass sulle sovrapposizioni rimanenti, oppure di stimare la sovrapposizione a partire dai dati raccolti e sottrarre questa stima all'onda ERP media.

Capitolo 3

L'approccio statistico standard

Una volta ottenuta un'onda ERP quanto più è possibile libera da rumore e artefatti, si può procedere con l'analisi statistica dei dati, finalizzata a capire se gli eventi a cui il soggetto è stato sottoposto hanno generato una variazione significativa nel voltaggio prodotto e, eventualmente, in che forma e misura. A tal fine, le analisi statistiche convenzionalmente utilizzate si concentrano su una particolare caratteristica dell'onda ERP, in base alle finalità dello studio. Infatti, ci si concentra sull'ampiezza dell'onda se l'interesse è quello di confrontare due o più condizioni in termini di forza del segnale generato, o sulla latenza se l'interesse è quello di valutare i tempi di reazione allo stimolo. Prima di approfondire il funzionamento delle tecniche statistiche convenzionali, verranno descritti i principali metodi di stima di queste due grandezze, valutandone i vantaggi e le problematiche. Come si vedrà, uno dei principali difetti di tali procedure è il fatto che si concentrano solo su un particolare sotto-intervallo temporale, la cui scelta può talvolta compromettere i risultati.

3.1 Misura delle ampiezze ERP

Data una generica onda è possibile misurarne l'ampiezza tramite tecniche come la PCA (Principal Component Analysis) o la ICA (Independent Component Alalisys) che, però, rispondono a esigenze più esplorative che inferenziali come, invece, richiederebbe lo studio delle onde ERP. Inoltre, la ICA genera solitamente soluzioni piuttosto instabili e difficilmente interpretabili, oltre a necessitare di assunzioni molto difficili da verificare [1]. Per questo motivo, spesso si preferisce utilizzare delle misure più semplici ed intuitive: l'ampiezza del picco o l'ampiezza media (o *area*).

3.1.1 Ampiezza del picco

La prima misura è ottenuta semplicemente considerando, all'interno della finestra temporale scelta, il punto in cui si raggiunge il valore di voltaggio più alto. I problemi

che si possono riscontrare sono diversi: si può incorrere in una situazione in cui il punto massimo si trova esattamente all'estremità della finestra temporale, ad esempio a causa di altre componenti che si sovrappongono a quella di interesse, oppure si può avere una misura fortemente influenzata dal rumore, dal momento che secondo tale tecnica l'informazione contenuta in una componente, lunga centinaia di millisecondi, viene sintetizzata da una sola osservazione.

Per questi motivi si preferisce spesso utilizzare una misura di ampiezza locale, costituita dai punti che presentano un voltaggio più alto rispetto ai 3-5 punti che lo circondano su entrambi i lati. In questo modo, si riduce il rischio che la misura venga falsata dal rumore, anche se questo non potrà essere annullato del tutto. Infatti, le misure di ampiezza del picco tenderanno sempre ad essere più grandi se il rumore è molto forte e se la finestra di misurazione è ampia: questo fa sì che non si possano confrontare ampiezze ottenute su intervalli temporali diversi. Un inconveniente di questa misura è la sua non linearità, cioè la media delle ampiezze calcolate a partire dalle singole onde non corrisponde all'ampiezza calcolata sull'onda media, e questo può provocare delle discrepanze fra le analisi statistiche svolte e una eventuale rappresentazione grafica.

3.1.2 Ampiezza media

Nel caso in cui si scelga di calcolare l'ampiezza media, si considerano tutte le osservazioni che ricadono all'interno della finestra temporale individuata e se ne calcola una media. Questa misura corrisponde all'area sottesa all'onda nell'intervallo scelto e rappresenta una buona soluzione al problema accennato al Capitolo precedente che riguarda il caso in cui le latenze non sono le stesse fra i soggetti. Infatti, l'area sotto la curva media corrisponde esattamente alla media delle aree sotto le curve dei singoli esperimenti, quindi la misura dell'area risulterà totalmente indipendente dalla variabilità della latenza.

Inoltre, in questo caso non è importante che l'istante temporale in cui si raggiunge l'ampiezza massima ricada all'interno dell'intervallo prescelto, quindi si può pensare di utilizzare un intervallo più stretto. In questo modo, infatti, si riduce l'influenza da parte delle componenti sovrapposte, anche se non bisogna mai individuare finestre troppo strette (<40ms), perché altrimenti aumenterebbe l'influenza subita da parte della componente di rumore. Per definizione, calcolare l'ampiezza media corrisponde esattamente ad applicare un filtro low-pass. Infatti, misurare l'ampiezza media su dati filtrati per una certa finestra temporale è equivalente a misurare l'ampiezza media in un'onda non filtrata usando una finestra di misurazione più ampia.

La misura dell'ampiezza media ha numerosi vantaggi rispetto a quella dell'ampiezza del picco. Infatti, è una misura che non diventa distorta quando si ha un livello di rumore alto o quando si ha una finestra temporale ampia ed è una misura lineare.

In alternativa, al posto di far riferimento direttamente all'area sotto la curva, ci si può riferire all'istante temporale che divide questa area in due parti esattamente uguali. Questo tipo di misura può non essere adeguato per componenti multifasiche, cioè composte sia da porzioni positive che negative perché queste possono eliminarsi a vicenda in caso di variabilità delle latenze.

3.2 Misure delle latenze ERP

Un semplice metodo per misurare le latenze di una componente ERP è quello di considerare la latenza a cui la componente raggiunge il picco di ampiezza massima, in un intervallo di tempo prestabilito. Si tratta, dunque, di un approccio molto simile a quello utilizzato per l'ampiezza media, quindi si avranno anche gli stessi problemi, come la forte dipendenza dal rumore e dalle componenti sovrapposte, la non linearità ecc. Per questo motivo, nell'effettuare tale misurazione è bene prendere alcune precauzioni, come filtrare il rumore ad alta frequenza, usare una misura di picco locale invece che di picco assoluto e assicurarsi che le onde che vengono confrontate abbiano livelli di rumore simili.

Sotto alcune condizioni, è possibile evitare alcuni problemi che caratterizzano la latenza del picco, utilizzando la cosiddetta *latenza dell'area frazionaria*, un concetto analogo alla misura di ampiezza media. Per ottenere questa misura, ci si condiziona ancora ad un intervallo temporale prefissato e si calcola l'area al di sotto dell'onda ERP: la misura di latenza desiderata è quella che divide tale area in una frazione pre-specificata. La misura più utilizzata è quella della latenza al 50% dell'area, che divide l'area in due parti esattamente uguali fra loro. Soprattutto se l'onda osservata è ottenuta dalla sovrapposizione di diverse componenti, questa misura non è adatta a stimare la latenza assoluta di una componente, a meno che l'intervallo considerato non includa tutta la componente. Fortunatamente, però, nella maggior parte degli studi l'interesse principale non è una misura assoluta della latenza, ma un confronto delle latenze sotto due diverse condizioni, quindi è possibile utilizzare la misura di latenza del 50% dell'area anche se non si considera l'intera finestra temporale e vi sono diverse componenti sovrapposte. Tale misura è particolarmente utile nel caso in cui una certa componente non abbia un picco distinto o abbia picchi multipli, riuscendo a coglierne correttamente le tempistiche, inoltre è indipendente dal livello di rumore presente nei dati ed è una misura lineare. Bisogna, però, fare attenzione che l'intervallo considerato comprenda la maggior parte della componente di interesse e che non includa ampi contributi da diverse componenti, altrimenti si rischierebbe di ottenere risultati distorti, anche se questo preclude l'utilizzo di moltissimi esperimenti.

La Figura 3.1 riporta un esempio delle misure viste finora nell'intervallo compreso fra il 140° e 190° ms (contrassegnato dalle linee rosse tratteggiate). In particolare, si riporta in verde l'ampiezza del picco, pari a 6.36, in viola l'ampiezza media, pari a 3.16, mentre le linee tratteggiate verticali indicano la misura di latenza (blu) e quella di latenza dell'area frazionaria, che divide la regione in due parti esattamente uguali (azzurro).

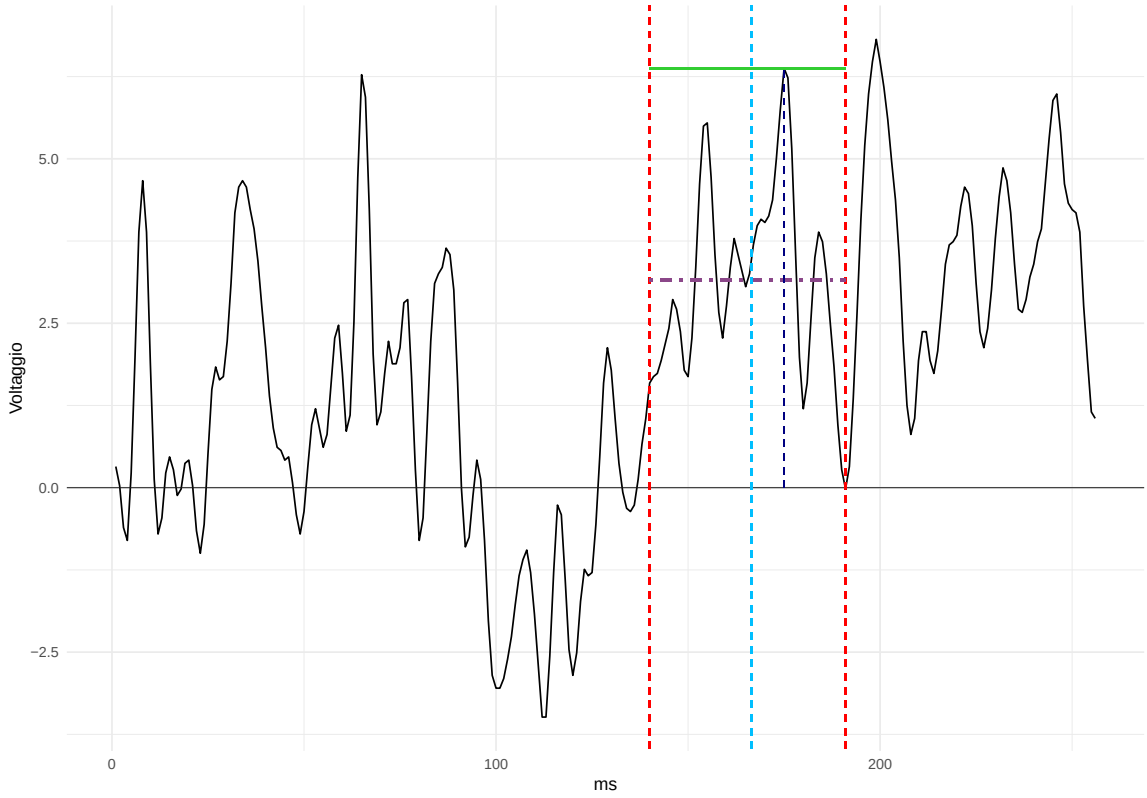

Figura 3.1: Esempio delle misure di ampiezza del picco (verde), ampiezza media (viola), latenza (blu) e latenza dell'area frazionaria (azzurro)

3.3 ANOVA per misure ripetute

Una volta ottenute le misure di ampiezza o di latenza della componente ERP, il principale interesse è di capire se queste siano significativamente diverse nelle varie condizioni a confronto e quindi se gli eventi abbiano generato risposte distinte. Il primo modo per far ciò è sviluppare un modello ANOVA [1], dove la variabile risposta considerata è proprio l'ampiezza o la latenza del segnale. I fattori di tale modello saranno la condizione a cui viene sottoposto il soggetto e una o più variabili che indicano la posizione dell'elettrodo, quindi il *canale* di rilevazione. Questa informazione può essere riportata tramite un solo fattore, contenente tanti livelli quanti sono gli elettrodi utilizzati, oppure tramite due fattori: uno che indica se l'elettrodo è posizionato sull'emisfero

destro, sul sinistro o al centro, e uno che indica se l'elettrodo è posizionato sulla zona anteriore o posteriore dello scalpo, in questo caso si possono avere fino a 3 livelli: frontale, centrale, parietale.

L'utilizzo di un modello ANOVA per misure ripetute è giustificato dalla particolare configurazione del disegno sperimentale già introdotto al Capitolo 1. In un contesto di analisi ERP, infatti, si è soliti raccogliere per ogni soggetto diverse misurazioni del segnale e questo fa sì che nell'insieme dei dati si possano distinguere vari "gruppi" di osservazioni, dove ogni gruppo è riferito ad un soggetto e può presentare particolari caratteristiche che vanno tenute in considerazione in fase di analisi. Dunque, la variabilità della risposta è spiegata in parte dai fattori inseriti nel modello e in parte è da ricondurre al soggetto stesso, pertanto si può pensare di avere la seguente suddivisione:

$$SS_{tot} = SS_{fatt} + SS_{sogg} + SS_{err}$$

Dove SS_{fatt}, SS_{sogg}, SS_{err} sono rispettivamente le componenti di variabilità dovuta ai fattori, quella da ricondurre al soggetto e quella residua.

I dati illustrati al Capitolo 1 presentano soltanto due condizioni e 64 canali, dunque il modello assumerà la seguente forma.

$$Y_{ij} = c + \mu_i + \gamma X_{ij} + \beta_1 Z_{1ij} + ... + \beta_{63} Z_{63ij} + \beta_{64j}[X_j Z_{1ij}] + ... + \beta_{127}[X Z_{63ij}] + \epsilon_{ij}$$

dove:

- $i = 1, ..., 10$ indica il soggetto;
- $j = 1, ..., 128$ indica l'osservazione relativa all'i-esimo soggetto;
- c è l'intercetta comune a tutti i soggetti;
- Y è la variabile risposta, cioè la misura scelta per sintetizzare il segnale registrato nell'intervallo scelto;
- X è la variabile dummy che indica la condizione a cui il soggetto viene sottoposto. Essa vale 0 se ci si riferisce alla condizione S1 e 1 se ci si riferisce alla condizione S2;
- Z_1,...,Z_{63} sono variabili dummy, ognuna delle quali fa riferimento a uno dei 63 elettrodi diversi da quello di riferimento;
- μ_i é il parametro che esprime la variabilità del soggetto i-esimo.

Questo modello si basa sull'assunzione che gli errori ϵ_{ij} siano indipendenti e identicamente distribuiti secondo una Normale di media 0 e varianza σ^2. Negli sviluppi

pratici l'assunzione di normalità è spesso violata ma, grazie al teorema del limite centrale, è comunque possibile sviluppare il modello, anche se i p-value calcolati saranno solo un'approssimazione della reale probabilità di commettere errori di tipo I.

Come si può notare dalla formula, il modello generale contiene delle componenti di interazione fra i fattori riferiti ai singoli elettrodi e la condizione a cui il soggetto è stato sottoposto. In questo caso, se alcune di queste componenti di interazione risultano significative, la loro interpretazione non è del tutto immediata [1]. Infatti, può accadere che un effetto di interazione di questo tipo diventi significativo quando solo una piccola area della corteccia genera un segnale diverso nelle diverse condizioni, mentre se tale diversità riguarda contemporaneamente diversi siti generatori di potenziale, si potrebbe ottenere semplicemente un effetto additivo. Questo comporta che, anche quando le interazioni sono significative, non è possibile trarre conclusioni forti sulle differenze con cui la varie zone della corteccia si attivano sotto diverse condizioni.

Infine, un'importante assunzione che in alcuni casi andrebbe fatta in un generico modello ANOVA, è quella di *sfericità* [1]. Essa riguarda i fattori che contengono almeno 3 livelli, infatti, stabilisce che tutte le possibili coppie dei livelli del fattore abbiano lo stesso livello di correlazione. Tuttavia, questa assunzione è spesso violata nel contesto delle EEG perché se, ad esempio, il fattore riguarda il posizionamento degli elettrodi, è noto che due elettrodi vicini tenderanno ad avere una correlazione maggiore rispetto a due elettrodi più lontani. Il modello non è molto robusto rispetto alla violazione di tale assunzione e quindi se questa non dovesse verificarsi, si otterrebbero p-value artificialmente bassi, rischiando di considerare un effetto significativo anche se, in realtà, non lo è. Per risolvere questo problema si può sfruttare la correzione di Greenhouse-Geisser che contrasta l'aumento dell'errore di tipo I riducendo i gradi di libertà, quindi generando p-value più alti. L'impostazione del modello vista sopra non implica la presenza di fattori con più di due livelli, quindi non necessita di tale assunzione.

3.4 Modello a effetti misti

I modelli a effetti misti consentono un utilizzo più flessibile dei dati [5]. In esso i fattori vengono distinti in *fissi*, cioè stabiliti a priori in base al disegno sperimentale, e *casuali*, se nel modello è inserito solo un campione casuale dei livelli che questi potrebbero assumere. In questo contesto, i parametri fissi sono quelli relativi alla condizione e al posizionamento degli elettrodi, mentre la parte casuale è quella relativa ai soggetti. Dunque, contrariamente al modello ANOVA, dove la componente riferita al soggetto i-esimo è un semplice parametro da stimare insieme a tutti gli altri, qui tale componente, cioè μ_i, è una variabile casuale con una propria distribuzione. Il modello assume la seguente forma:

$$Y_{ij} = c + \gamma X_{ij} + \beta_1 Z_{1ij} + ... + \beta_{63} Z_{63ij} + \beta_{64}[X_j Z_{1ij}] + ... + \beta_{127}[X_{Z63ij}] + u_{ij}$$

$$\text{con } u_{ij} = \mu_i + \epsilon_{ij}$$

Le assunzioni sono quelle già citate per il modello ANOVA a misure ripetute, a cui va aggiunto:

$$\epsilon_{ij} \underset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$$

$$\mu_i \sim N(0, \sigma_\mu^2)$$

$$\epsilon_{ij} \perp \mu_i$$

Il termine di errore u_i presenta i seguenti momenti:

- $var(u_{ij}) = \sigma_\mu^2 + \sigma_\epsilon^2$;
- $cov(u_{ij}, u_{is}) = \sigma_\mu^2$, correlazione fra osservazioni riferite allo stesso soggetto;
- $cov(u_{ij}, u_{rj}) = 0$, le osservazioni di due soggetti diversi riferite alla stessa combinazione di canale e istante di rilevazione sono incorrelate;
- $cov(u_{ij}, u_{rs}) = 0$, le osservazioni riferite a soggetti diversi e a due combinazioni di canale e istante di rilevazione diversi sono fra loro incorrelate.

Dunque, tutte le osservazioni riferite ad uno stesso soggetto hanno una certa variabilità, definita dal parametro σ_μ^2 che è ignoto ma assunto costante per tutti i soggetti (omoschedasticità). Proprio la presenza di questo livello di variabilità aggiuntivo rende necessarie delle tecniche di stima più articolate rispetto a quelle solitamente utilizzate per un semplice modello lineare, come ad esempio la stima di tipo REML (Restricted Maximum Likelihood). Essa si basa sulla possibilità di trovare una combinazione lineare dei dati (Ay) la cui distribuzione non dipenda dai parametri fissi, cioè γ e il vettore dei β. A partire da tale combinazione si effettua una stima della componente di varianza che viene, a sua volta, utilizzata per la stima dei parametri fissi. In questo modo è possibile ottenere delle stime non distorte per i parametri fissi.

3.5 ANOVA Multivariata (MANOVA)

Un ulteriore approccio all'analisi degli effetti nel segnale ERP, è quello del modello MANOVA, cioè una generalizzazione al caso multivariato del modello ANOVA precedentemente illustrato. In questo contesto, il posizionamento degli elettrodi non viene

tenuto in considerazione sotto forma di fattore, ma si sviluppa un modello ANOVA per ogni singolo elettrodo. Diversamente da un semplice modello ANOVA, nel valutare la significatività degli effetti non si dovrà tenere in considerazione la semplice varianza della variabile risposta Y, ma l'intera matrice di varianza e covarianza di tutte le variabili dipendenti. Facendo ancora riferimento ai dati presentati al Capitolo 1, dovranno essere sviluppati contemporaneamente i seguenti modelli:

$$
\begin{aligned}
Y_{1ij} &= c + \mu_{1i} + \gamma_1 X_{ij} + \epsilon_{1ij} \\
Y_{2ij} &= c + \mu_{2i} + \gamma_2 X_{ij} + \epsilon_{2ij} \\
&\dots \\
Y_{64ij} &= c + \mu_{64i} + \gamma_{64} X_{ij} + \epsilon_{64ij}
\end{aligned}
$$

dove $j = 1, 2$ dato che per ogni soggetto e in ogni canale si hanno soltanto le rilevazioni sotto la condizione S1 e quelle sotto la condizione S2.

In genere, si procede sviluppando in un primo momento un test complessivo, cioè che verifichi l'ipotesi di uguaglianza delle risposte fra le due condizioni contemporaneamente su tutti i canali. Assumendo, anche in questo caso, la normalità multivariata delle variabili, la verifica di questa ipotesi può essere svolta tramite la statistica test t^2 di Hotelling, che rappresenta una generalizzazione della statistica test F, utilizzata in un semplice ANOVA. La relazione fra le due statistiche test è la seguente:

$$t^2 \sim T_{p,n-1} = \frac{p(n-1)}{n-p} F_{p,n-p}$$

Nel caso in cui tale ipotesi venga rigettata, per il ricercatore potrebbe essere interessante approfondire l'analisi, indagando su quali particolari canali presentano un segnale diverso dagli altri e quindi portano a tale rigetto. A tal fine, è utile svolgere un test t su ogni canale anche se, dato l'alto numero di test da svolgere (in questo caso 64), risulterebbe necessario l'utilizzo delle cosiddette tecniche di correzione dei p-value per l'aggiustamento dell'errore di tipo I, di cui si parlerà in modo più approfondito al Capitolo 4.

3.6 Analisi secondo l'approccio standard

Per analizzare i dati presentati al Capitolo 1 secondo gli approcci standard presentati finora, si è scelto di sviluppare soltanto gli ultimi due modelli descritti, data la loro maggiore flessibilità. La sintesi delle varie osservazioni effettuate nel corso della rilevazione è stata fatta calcolando l'ampiezza media nell'intervallo di osservazione fra il 70°

e il 110° millisecondo. Dunque, come prima cosa, si è proceduto al calcolo di:

$$Y_{ije} = \sum_{t=70}^{110} \frac{s_{ijet}}{41}$$

con s_{ijet} il livello di segnale misurato sul soggetto $i-$esimo ($i = 1, ..., 10$), nel canale e al t-esimo millisecondo e sotto la condizione j (con $j = 1, 2$) e Y_{ije} l'ampiezza media del segnale nell'intervallo 70-110ms ottenuta per il soggetto i relativamente al canale e e alla condizione j: questa verrà utilizzata come variabile risposta nei modelli che verranno sviluppati di seguito. Per analogia ai metodi che verranno sviluppati nei Capitoli successivi, si è scelto di non raggruppare i canali di rilevazione in base alla zona della corteccia occupata, quindi $e = 1, ..., 64$.

Nel modello a effetti misti il livello di riferimento è quello del segnale misurato sotto la condizione S1 dall'elettrodo TP8, dunque di seguito si riporta, a titolo di esempio, l'effetto che si dovrà valutare per capire se l'ampiezza media nell'elettrodo AF1 è significativamente diversa fra la condizione S1 e quella S2 nell'intervallo scelto.

$$AF1 : (c + \mu_i + \gamma + \beta_1 + \beta_{64}) - (c + \mu_i + \beta_1) = \gamma + \beta_{64}$$

Un effetto simile dovrà essere calcolato e valutato per tutti i 63 elettrodi diversi da quello di riferimento. In questo modello, senza svolgere nessuna correzione sui p-value, si ottengono 42 effetti significativi fra tutti i 64 canali. In Figura 3.2 si fornisce una rappresentazione grafica di tale risultato, dopo aver trasformato i p-value di interesse come $-log_{10}(p)$. I valori al di sopra della soglia pari a $-log_{10}(p) \approx 1.30$ sono quelli da considerare significativi.

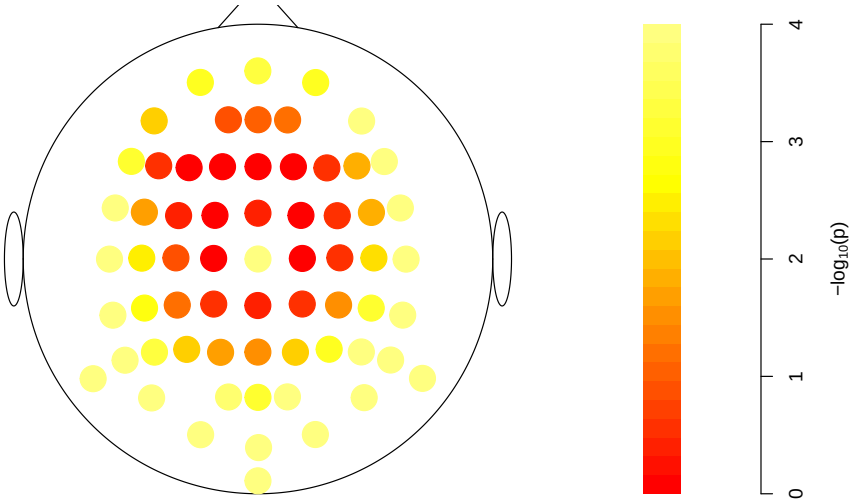

Figura 3.2: P-value grezzi del modello a effetti misti

Per quanto riguarda il modello MANOVA, invece, si è scelto di sviluppare direttamente i 64 modelli ANOVA per il confronto diretto del segnale fra le due condizioni. Nel far ciò, non sarà necessario scegliere un livello di riferimento e ogni elettrodo avrà una propria stima dell'effetto di interesse, γ. Fra questi, 39 sono risultati significativi come mostrato in Figura 3.3.

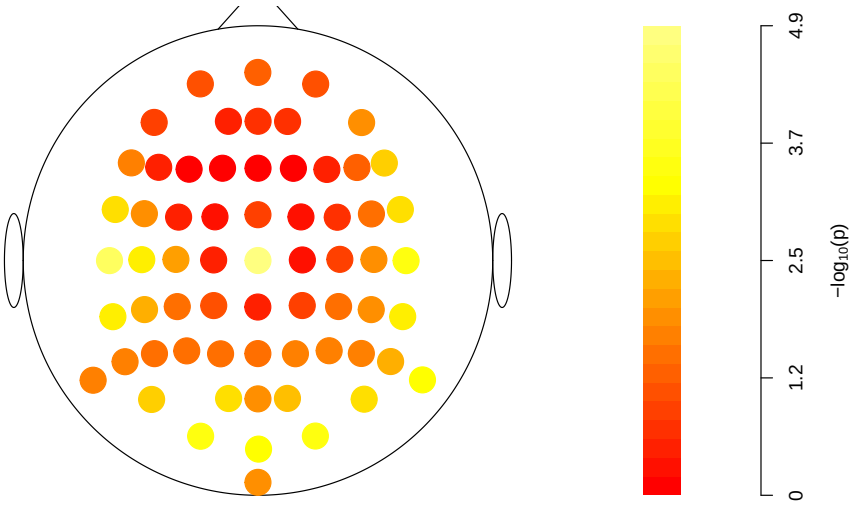

Figura 3.3: P-value grezzi del modello MANOVA

Nella valutazione dei risultati ottenuti con i metodi standard è molto importante tener conto del fatto che essi sono ottenuti soltanto su un sottoinsieme di osservazioni scelto a priori e in modo arbitrario. Infatti, tale scelta influisce sull'esito dell'analisi in un modo che è difficile da prevedere, specialmente per particolari forme delle onde studiate o se la quantità di rumore presente nei dati è molto alta. Insieme a quella dell'intervallo temporale, può risultare di fondamentale importanza per la determinazione dei risultati finali anche la scelta della misura di sintesi su cui sviluppare l'analisi.

Capitolo 4

Test multipli e controllo dell'errore di I tipo

I metodi visti finora rappresentano solo una delle prospettive che è possibile assumere nell'analisi di un'onda ERP, ovvero quella in cui si sceglie di concentrare l'attenzione semplicemente su una sintesi dei dati. Un approccio alternativo, invece, è quello che porta a svolgere l'analisi direttamente sul segnale osservato, consentendo, in questo modo, di ottenere risultati molto più precisi al livello temporale. In particolare, dato un certo soggetto e un certo canale di rilevazione, il confronto del segnale nelle diverse condizioni si svolge su ogni singolo istante di osservazione attraverso lo sviluppo di un test. In questo modo il numero totale di test da sviluppare sarà molto elevato e questo rende indispensabile, anche in questo caso, l'utilizzo delle tecniche di controllo della probabilità di commettere errori. Alcune fra le tecniche più diffuse in questo ambito saranno illustrate più avanti in questo Capitolo, dopo aver illustrato nel dettaglio il tipo di ipotesi da verificare e le varie tipologie di errore in cui si può incorrere in un'analisi multivariata.

4.1 Il test t

Il primo passo previsto da tale approccio è proprio lo svolgimento dei test. In particolare, dati $\mu_{et,S1}$ e $\mu_{et,S2}$, che indicano le medie fra le osservazioni sui 10 soggetti all'istante t-esimo (con t=1,..., 256), con $e = 1, ..., 64$ indice del canale di rilevazione, rispettivamente sotto le condizioni S1 e S2, si ha $\delta_t = \mu_{et,S1} - \mu_{et,S2}$. Quindi, il sistema di ipotesi da verificare è:

$$\begin{cases} H_0 : \delta_{et} = 0 \\ H_1 : \delta_{et} \neq 0 \end{cases}$$

Se H_0 è accettata, il segnale rilevato al canale e all'istante t è considerato statisticamente uguale fra le due condizioni, altrimenti, se vale H_1 il segnale rilevato sotto la

condizione S1 nel canale e all'istante t è diverso da quello rilevato sotto la condizione S2. La verifica di tale ipotesi avviene tramite un semplice test t per dati appaiati, dato che la rilevazione sotto le due diverse condizioni viene svolta sugli stessi soggetti. Dunque, la statistica test di riferimento è:

$$t_{et} = \frac{\delta_{et}}{SE(\delta_{et})} \sim t_{n-1,\frac{\alpha}{2}}$$

Con t_{n-1} distribuzione t di student, con n-1 gradi di libertà, dove n è il numero di differenze calcolate, in questo caso pari a 10.

Nel verificare un certo sistema di ipotesi è sempre possibile commettere due tipologie di errore:

- **Errore di tipo I** (o Falsi positivi) quando si rigetta un'ipotesi che, invece, è vera. La probabilità di commettere questo tipo di errore è:

$$\alpha = \mathbb{P}(p \leq 0.05|H_0)$$

- **Errore di tipo II** (o Falsi negativi) quando non si rigetta un'ipotesi che, invece, è falsa. La probabilità di commettere questo tipo di errore è:

$$\beta = \mathbb{P}(p > 0.05|H_1)$$

  A partire da β viene definita la **potenza** di un test, cioè la probabilità di rifiutare l'ipotesi quando questa è falsa, quindi di fare una scelta corretta:

$$1 - \beta = \mathbb{P}(p \leq 0.05|H_1) = 1 - \mathbb{P}(p > 0.05|H_1)$$

Generalmente è considerato più grave l'errore di tipo I, quindi si è soliti imporre a priori un limite massimo alla probabilità di commettere questo tipo di errore: di seguito tale limite sarà $\alpha = 0.05$.

4.2 Test multipli e misure dell'errore

Nel caso particolare dei dati presentati Capitolo 1, l'analisi su ogni istante temporale richiede lo svolgimento simultaneo di 256×64=16384 test. Inoltre, come tipicamente accade nelle osservazioni di ERP, i livelli di segnale registrati risultano fortemente correlati fra loro sia al livello spaziale, dato che gli elettrodi sono posizionati in punti talvolta molto vicini fra loro, sia a livello temporale, dato che il segnale rilevato in un certo istante sarà dipendente da quello rilevato all'istante immediatamente precedente. Di conseguenza, anche i test avranno fra loro un certo livello di correlazione.

Lo svolgimento simultaneo di un numero così alto di test porta a commettere degli errori con una probabilità molto più alta di quella che ci si aspetterebbe, pertanto risultano necessarie delle tecniche di correzione *post hoc* dell'errore che consentano di correggere o, quanto meno, quantificare la probabilità di commettere un errore. Le principali tecniche utilizzate si distinguono in base alla tipologia di errore che sono in grado di controllare, pertanto, prima di analizzarne il funzionamento, si approfondiranno le varie tipologie di errore che si possono commettere nel caso di test multipli.

Seguendo la notazione di Goeman et al. (2014), si considerino genericamente m ipotesi nulle da testare: fra queste m_0 sono vere, mentre $m_1 = m - m_0$ sono false. Chiaramente, m_0 e m_1 sono quantità non note, ma se ne può ottenere una stima. Sia R l'insieme di ipotesi considerate false in base alle tecniche per i test multipli: questo insieme dovrà essere quanto più simile è possibile a m_1, ma è ammessa la possibilità che esso contenga sia ipotesi correttamente rifiutate sia altre ipotesi che vengono rigettate erroneamente, perché in realtà sono vere. Quest'ultimo sottoinsieme viene identificato con V. Pertanto, si distinguono:

- **FDP** (False Discovery Proportion): fra tutte le ipotesi rigettate, esprime la proporzione di ipotesi rigettate erroneamente, perché vere.

$$\text{FDP}= \begin{cases} Q = \frac{V}{R}, & \text{se } R > 0 \\ 0 \text{ altrimenti} \end{cases}$$

- **FWER** (Familywise Error Rate): esprime la probabilità che il gruppo di ipotesi rigettate contenga qualche errore.

$$\text{FWER}=\mathbb{P}(V > 0) = \mathbb{P}(Q > 0)$$

- **FDR** (False Discovery Rate): esprime la proporzione attesa di errori fra le ipotesi rifiutate.

$$\text{FDR}=E(Q)$$

Esse rappresentano una generalizzazione alle ipotesi multiple dell'errore di tipo I quindi, così come con una sola ipotesi si controlla l'errore di tipo I, in un contesto di ipotesi multiple si può scegliere se controllare FDP, FWER o FDR. In particolare, nel caso del controllo di FWER e FDR si fa in modo che il gruppo di ipotesi rigettate R sia scelto in modo tale che sia $\mathbb{P}(Q > 0)$ che $E(Q)$ siano al massimo pari ad α.

Le similitudini fra i metodi di correzione di FWER e di FDR sono molteplici. Ad esempio, generalmente, entrambe le procedure di controllo si sviluppano calcolando i cosiddetti *p-value aggiustati*, cioè il più piccolo livello α a cui la procedura porterebbe al rifiuto dell'ipotesi. Esse, inoltre, risultano in relazione fra loro, infatti, dato che

$Q \in [0,1]$ si ha:

$$E(Q) \leq \mathbb{P}(Q > 0)$$

Dunque, il controllo del FWER risulta equivalente a quello del FDR anche se, in alcuni casi, risulta preferibile il primo perché gode della proprietà di *subsetting*. Infatti, se secondo la procedura di controllo del FWER un certo gruppo di ipotesi viene rifiutato, allora il controllo è garantito anche per ogni sottoinsieme di tale gruppo, quindi si può essere certi di controllare FWER anche per le singole ipotesi che possono essere considerate come sottoinsiemi del gruppo originario. Pertanto, se FWER di un certo gruppo di ipotesi è minore di α, allora per ogni ipotesi di questo gruppo la probabilità di fare un errore di tipo I è minore di α. La stessa proprietà non vale per il controllo del FDR, che è valido soltanto per l'intero gruppo di ipotesi considerate dal principio. Infatti, controllare FDR vuol dire semplicemente controllare la media dell'errore di tipo I rispetto a tutte le ipotesi considerate.

A loro volta, le procedure di controllo di FWER e FDR si distinguono dal controllo della FDP perché nelle prime il ricercatore sceglie il tasso di errore che deve essere controllato e la procedura si occupa di trovare un certo insieme di ipotesi R da rigettare in base a tale criterio, tramite il calcolo dei *p-value aggiustati*. Nel secondo tipo di procedure, invece, l'insieme R è stabilito dal ricercatore, e in base a questo la proporzione di false scoperte, Q o FDP (False Discovery Proportion), viene stimata costruendo un intervallo di confidenza. Queste ultime, infatti, non contemplano l'utilizzo di p-value aggiustati.

4.3 Metodi per il controllo del FWER

4.3.1 Correzione di Bonferroni

Il più semplice metodo di correzione del FWER è quello di Bonferroni. Esso è valido sotto ogni struttura di dipendenza dei p-value e prevede il rifiuto della j-esima ipotesi se [7]:

$$p_j \leq \frac{\alpha}{m}$$

dove m è ancora il numero totale di ipotesi da testare, α esprime la probabilità massima di commettere errori di tipo I che si è disposti ad accettare per l'intera procedura, solitamente posta pari a 0.05, p_j è il p-value *grezzo* ottenuto per la j-esima ipotesi, cioè quello ottenuto senza alcuna correzione. Questa disuguaglianza può essere riformulata in termini di *p-value aggiustato*, $\tilde{p_j}$, che sarà confrontato direttamente con la probabilità di errore di tipo I originariamente scelta, α, infatti:

$$\tilde{p_j} = p_j m \leq \alpha$$

Indicando con q_i il p-value della i-esima ipotesi nulla vera, il FWER può essere espresso come:

$$\text{FWER} = \sum_{i=1}^{m_0} P(q_i \leq \frac{\alpha}{m})$$

Dunque, ipotizzando che i p-value q_i siano distribuiti uniformemente fra 0 e 1, si ha:

$$\text{FWER} \leq m_0 \frac{\alpha}{m} \leq \alpha$$

La procedura di Bonferroni è molto conservativa, cioè tende a generare p-value aggiustati molto grandi portando, quindi, al rigetto di poche ipotesi. In effetti, come si può osservare dall'ultima disuguaglianza, il controllo del FWER, di fatto, non è svolto rispetto ad α, ma rispetto alla quantità più piccola $m_0 \frac{\alpha}{m}$. È comunque da tenere in considerazione il fatto che il livello di conservatività è strettamente legato alla struttura di dipendenza dei test. Infatti, se questi sono correlati fra loro negativamente la conservatività è quasi nulla, mentre risulta molto bassa se i test sono indipendenti. Al contrario, se la correlazione è positiva, la procedura risulterà molto conservativa. La procedura di correzione di Bonferroni può essere utilizzata per qualsiasi valore del rapporto $\frac{m_0}{m}$, che definisce la proporzione di ipotesi vere sul totale di quelle analizzate, ma se questo rapporto è basso, cioè molte ipotesi non sono vere, la conservatività della procedura rappresenta un vero e proprio difetto perché porta a ridurre artificialmente il numero di *scoperte* (cioè di p-value significativi) che si ottengono. La procedura di Bonferroni consente di individuare con precisione quali fra le ipotesi testate siano accettate e quali no, e per questo motivo si parla di controllo *forte* del FWER.

4.3.2 Correzione di Holm

Il metodo di correzione di Holm è una variante meno conservativa di quello di Bonferroni e anch'esso può essere sviluppato direttamente sui p-value grezzi o attraverso la costruzione dei p-value aggiustati.

Nel primo caso, si può intendere la procedura come uso reiterato del metodo di Bonferroni dove ad ogni passo cambia il valore critico considerato. Infatti, al primo passo si rifiutano tutte le ipotesi che presentano un p-value grezzo (p_j) tale che:

$$p_j \leq \frac{\alpha}{m_0}$$

dove m_0 è il numero totale di ipotesi da testare. Nei passi successivi tale valore è sostituito rispettivamente da m_1, m_2,... cioè dal numero di ipotesi che fino al passo precedente non sono state rigettate. La procedura continua finché tutte le ipotesi sono state rigettate o finché nessuna di queste può più esserlo.

La costruzione dei p-value aggiustati, invece, si sviluppa attraverso i seguenti passi [7]:

1) Si riordinano in senso crescente i p-value grezzi, ottenendo il vettore $p_{(1)}, p_{(2)}, ..., p_{(m)}$;

2) Si calcolano le quantità $m_j p_{(j)}$, dove m_j è la posizione occupata dal j-esimo p-value nel vettore ordinato;

3) Se queste quantità mantengono l'ordinamento originario si pone

$$\tilde{p_{(j)}} = m_j p_{(j)}$$

altrimenti il p-value aggiustato sarà pari a

$$\tilde{p_{(j)}} = \max_{j=1,...,i} m_j p_{(j)}$$

4) Poiché in alcuni casi il p-value ottenuto potrebbe essere maggiore di 1, si pone

$$\tilde{p_{(j)}} = \min(\tilde{p_{(j)}}, 1)$$

Una volta ottenuti i p-value aggiustati, questi vengono confrontati con il livello di significatività scelto a priori, solitamente pari ad $\alpha = 0.05$.

Così come il metodo di Bonferroni, anche quello di Holm è utilizzabile con qualsiasi schema di correlazione fra i test ma, come già anticipato, risulta meno conservativo rispetto ad esso, infatti porterà al rifiuto di almeno tante ipotesi quante ne vengono rigettate col metodo di Bonferroni.

Una prima applicazione del metodo di correzione di Holm può essere quella rivolta ai p-value calcolati al Capitolo 3 sui modelli a effetti misti e MANOVA. Si tratta infatti, in entrambi i casi del calcolo di 64 p-value non indipendenti che, quindi, se lasciati "grezzi" possono portare a conclusioni fuorvianti. Nella Tabella 4.1, infatti, si fa un confronto tra i risultati ottenuti con i due modelli finora sviluppati, prima e dopo aver corretto i p-value col metodo di Holm. In particolare, per semplicità vengono riportati soltanto i canali che hanno dato i risultati più rilevanti, cioè quelli dove almeno uno dei p-value calcolati è risultato significativo. Nel passaggio dai p-value grezzi a quelli corretti si può osservare per entrambi i modelli una forte riduzione del numero totale di scoperte che, infatti, si riduce da 42 a 19 nel modello a effetti misti e da 39 a 8 nel modello MANOVA. Dunque, nel primo caso la correzione di Holm esclude circa il 55% dei canali in cui inizialmente si era individuato un effetto significativo della condizione, mentre nel secondo caso si arriva ad escluderne quasi l'80%.

Inoltre, volendo confrontare fra loro i risultati ottenuti con i due modelli, si arriverebbe a conclusioni molto diverse a seconda che si faccia riferimento ai p-value grezzi o a quelli corretti. Infatti, prima di applicare la correzione i due metodi sembrano dare quasi lo stesso numero totale di scoperte (42 e 39), mentre a seguito della correzione i risultati sembrano molto diversi fra loro, dato che col modello a effetti fissi si ottiene più del doppio del numero totale di scoperte ottenute col modello MANOVA (rispettivamente 19 e 8). Una tale differenza nei risultati può essere spiegata dal fatto che il modello a effetti misti è basato su ipotesi che in un contesto reale di studio delle onde ERP difficilmente si verificano come, ad esempio, l'assunzione che la componente riferita al soggetto μ_i abbia la stessa variabilità per tutti i soggetti.

Un elemento di contatto fra i due modelli, invece, è dato dal fatto che con entrambi le scoperte ottenute sono disposte principalmente nella zona più esterna della corteccia posteriore, a cui si aggiunge un unico canale (CZ) posizionato esattamente al centro. La disposizione spaziale delle scoperte ottenute a seguito della correzione di Holm viene mostrata nelle Figure 4.1 e 4.2 dove si riporta la trasformazione $-log_{10}(p)$ dei p-value. Anche in questo caso, il valore critico è $-log_{10}(p) \approx 1.30$, dunque al di sopra di tale soglia l'effetto è da considerarsi significativo.

Tabella 4.1: P-value per i canali più rilevanti ottenuti dai modelli a effetti fissi e MANOVA prima e dopo la correzione di Holm

	Effetti misti grezzi	**Effetti misti Holm**	**MANOVA grezzi**	**MANOVA Holm**
AF7	0.008	1.000	0.155	1.000
AF8	<0.00	**0.038**	0.015	0.624
C3	0.148	1.000	0.009	0.413
C5	0.003	1.000	0.001	0.057
C6	0.005	1.000	0.016	0.624
CP3	0.069	1.000	0.029	0.892
CP4	0.034	1.000	0.037	1.000
CP5	0.001	1.000	0.004	0.207
CP6	0.001	1.000	0.014	0.595
CZ	<0.00	**<0.00**	<0.00	**0.001**
F6	0.013	1.000	0.065	1.000
F7	0.001	1.000	0.024	0.776
F8	<0.00	**0.006**	0.002	0.107
FC5	0.018	1.000	0.011	0.480
FC6	0.014	1.000	0.038	1.000
FP1	0.001	1.000	0.074	1.000
FP2	0.001	1.000	0.080	1.000
FPZ	<0.00	1.000	0.064	1.000
FT7	<0.00	0.100	0.001	0.071
FT8	<0.00	**0.008**	0.001	0.069
IZ	<0.00	**0.017**	0.015	0.624
O1	<0.00	**<0.00**	<0.00	**0.022**
O2	<0.00	**<0.00**	<0.00	**0.017**
OZ	<0.00	**<0.00**	<0.00	**0.029**
P1	0.019	1.000	0.028	0.892
P10	<0.00	**<0.00**	0.001	**0.029**
P2	0.007	1.000	0.023	0.776
P3	0.006	1.000	0.031	0.921
P4	0.001	1.000	0.022	0.773
P5	<0.00	0.966	0.033	0.960
P6	<0.00	**0.022**	0.019	0.708
P7	<0.00	**0.016**	0.021	0.758
P8	<0.00	**<0.00**	0.005	0.245
P9	<0.00	**<0.00**	0.019	0.708
PO1	<0.00	0.242	0.001	0.071
PO2	<0.00	0.077	0.003	0.165
PO7	<0.00	**<0.00**	0.002	0.087
PO8	<0.00	**<0.00**	0.001	0.076
POZ	0.001	1.000	0.012	0.510
PZ	0.026	1.000	0.035	0.968
T7	<0.00	**0.017**	<0.00	**0.002**
T8	<0.00	**0.001**	<0.00	**0.022**
TP7	<0.00	**0.020**	0.001	**0.044**
TP8	<0.00	**<0.00**	0.001	0.055

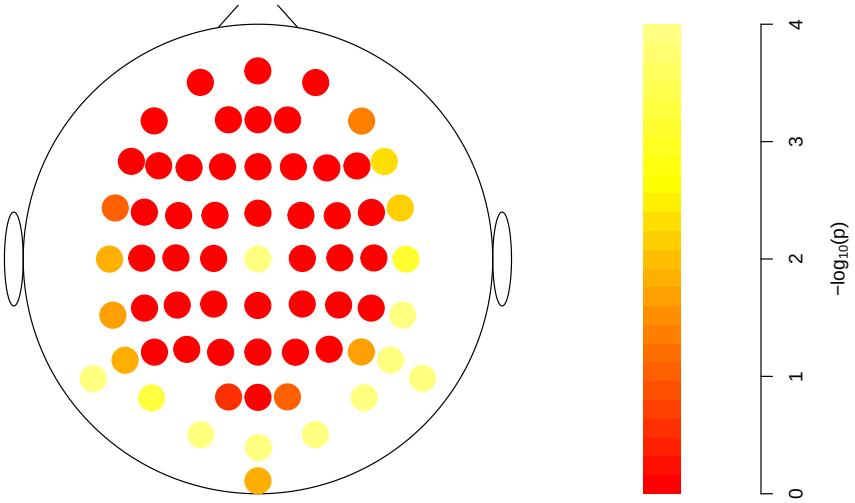

Figura 4.1: P-value del modello a effetti misti corretti col metodo di Holm

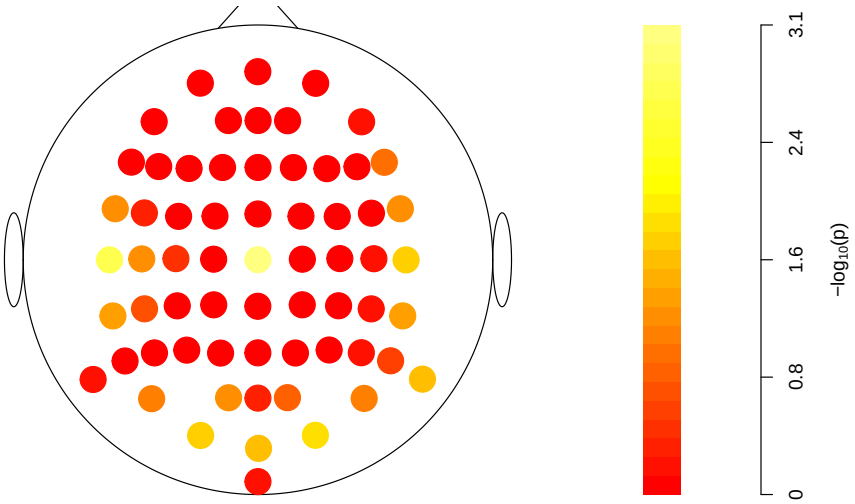

Figura 4.2: P-value del modello MANOVA corretti col metodo di Holm

4.4 Controllo del FDR: la procedura di Benjamini&Hochberg

Una delle procedure più utilizzate per il controllo del FDR (False Discovery Rate) è quella elaborata da Benjamini&Hochberg. Questa è valida sotto l'assunzione di indipendenza dei p-value o sotto l'assunzione di dipendenza positiva rispetto all'ordinamento stocastico (PRDS) secondo cui, dato $q_i > \frac{\alpha}{m_0}$ per $i = 1, ..., m_0$, la quantità $E[f(p_1, p_2, ..., p_m)|q_i = u]$ è non decrescente in u e per ogni funzione f non decrescente [7].

La procedura prevede, innanzitutto, l'ordinamento dei p-value grezzi, da cui si ottiene il vettore $p_{(1)}, p_{(2)}, ..., p_{(m)}$. Ogni elemento di questo vettore viene confrontato col valore critico:

$$c_j = \frac{j\alpha}{m}$$

Successivamente, si trova il più grande j per cui $p_{(j)}$ è minore del suo corrispondente valore critico, e si rifiutano le j ipotesi con i p-value più piccoli.

In generale, le procedure di controllo del FDR sono molto più potenti rispetto a quelle che controllano il FWER, soprattutto quando molte delle ipotesi da testare sono false. Tuttavia, il controllo che esse svolgono non riguarda le singole probabilità di errore ma il loro valore atteso, e questo può portare in alcuni casi a comprometterne il risultato. Potrebbe accadere, infatti, che i p-value dei singoli test (quindi le probabilità di commettere un errore) si compensino fra loro, facendo in modo che il livello di errore complessivo sia proprio quello desiderato, α, anche se la probabilità di errore di alcuni test singoli supera quella imposta. Questo problema risulta ancora più accentuato quando i test sono dipendenti fra loro, proprio come nel caso del segnale ERP.

Un secondo difetto delle procedure di correzione del FDR è il fatto che esse non consentono di adattare a posteriori la correzione ottenuta a sottoinsiemi di ipotesi. Si consideri, ad esempio, l'insieme di ipotesi $H = \{H_1, ..., H_n\}$ su cui si vuole effettuare un controllo del FDR tramite il test $\varphi = (\varphi_1, ..., \varphi_n)$ e il suo sottoinsieme $H' = \{H_1, ..., H_r\}$ [11]: non è detto che il corrispondente sottoinsieme di test $\varphi' = (\varphi_1, ..., \varphi_r)$ sia in grado di svolgere un controllo del FDR al livello α, come stabilito originariamente. Per questo motivo, nell'utilizzo delle procedure di correzione del FDR bisogna sempre fare attenzione a riferire i risultati specificatamente al gruppo di ipotesi su cui queste sono state sviluppate e non a dei sottoinsiemi.

In un contesto di analisi delle ERP, le procedure di correzione del FDR risultano comunque utili quando l'interesse non è quello di cogliere la significatività dell'effetto in ogni singola rilevazione, ma semplicemente di svolgere un confronto generale fra due condizioni. Per questo motivo, anche se è ritenuta poco attendibile per i problemi appena visti, la correzione del FDR è stata ugualmente svolta attraverso il metodo di

Benjamini &Hochberg. I risultati ottenuti sono riportati e discussi in Appendice A.

4.5 Controllo dell'errore nei dati in esame

Come già anticipato, l'analisi dei dati introdotti al Capitolo 1 richiede lo svolgimento di 16384 test non indipendenti, pertanto l'utilizzo delle tecniche di controllo dell'errore per test multipli risulta necessario per non ottenere risultati del tutto fuorvianti.

La Tabella 4.2 riporta il numero dei rifiuti ottenuti in base ai p-value grezzi e con le tecniche di correzione finora introdotte, sia in termini assoluti che in termini percentuali.

Tabella 4.2: Numero di scoperte ottenute dai p-value grezzi, dalla correzione di Bonferroni e da quella di Holm

p-value grezzi	Correzione di Bonferroni	Correzione di Holm
3298	1	1
20.00%	0.006%	0.006%

Si può notare, innanzitutto, la forte influenza delle tecniche di correzione che, infatti, consentono di ridurre fortemente il tasso di scoperte, cioè da un totale del 20% di ipotesi rigettate secondo i p-value grezzi, si passa al rifiuto soltanto dello 0.006% delle osservazioni sia col metodo di Bonferroni che con quello di Holm, i cui risultati coincidono esattamente. Questo indica che la maggior parte dei p-value grezzi significativi sono, in realtà, delle *false scoperte*. Utilizzando le correzioni di Bonferroni e Holm, invece, soltanto un p-value, riferito al canale PO7, è risultato significativo: la probabilità che si tratti di una *falsa scoperta* è molto bassa, cioè pari a $\alpha = 0.05$.

Le Figure 4.3 e 4.4, rappresentano la differenza del segnale medio fra la condizione S1 e quella S2, nei 256 istanti temporali osservati e nei 64 canali di rilevazione e in esse è evidenziata l'unica differenza risultata significativa.

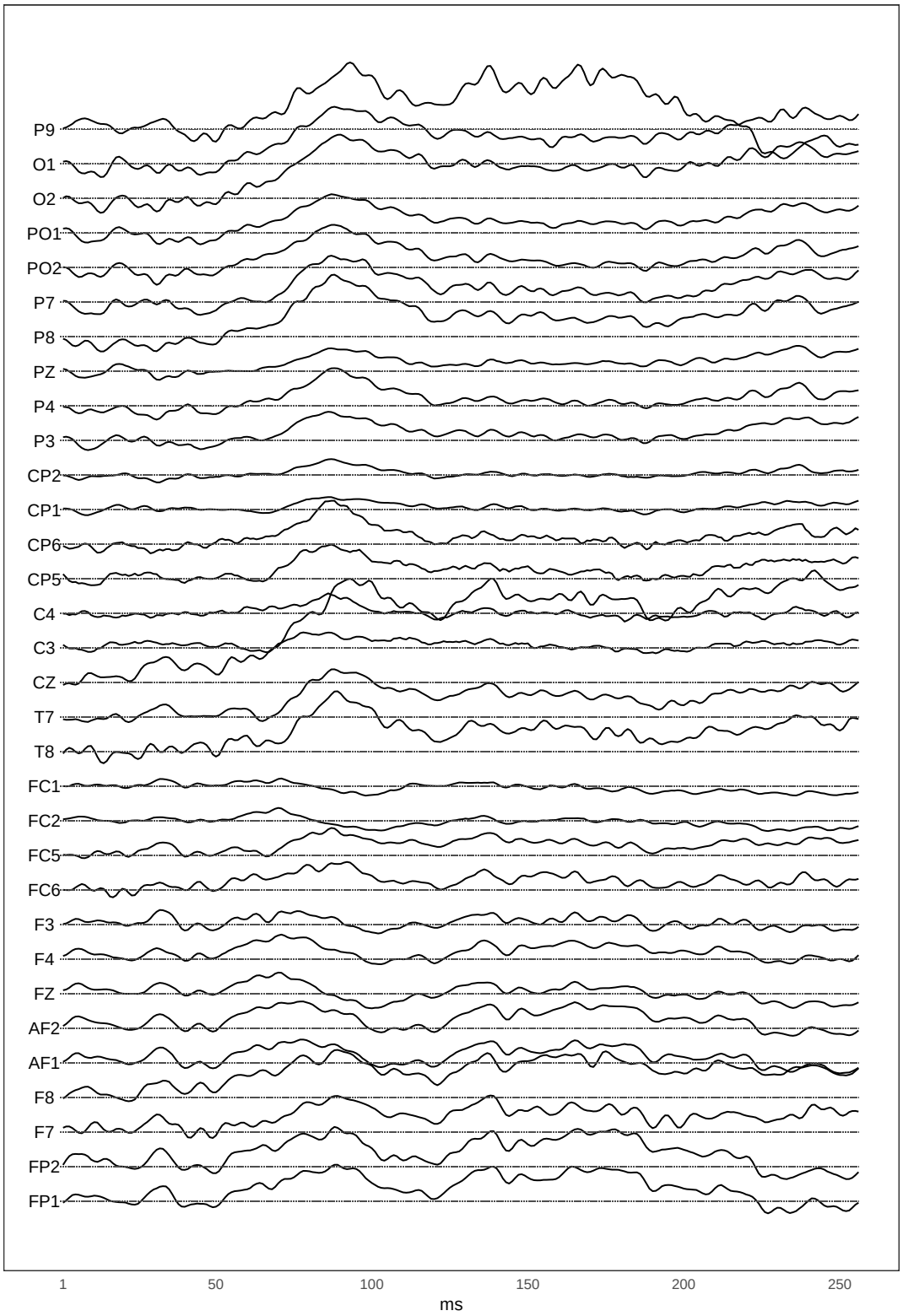

Figura 4.3: Scoperte ottenute tramite i metodi di Bonferroni e Holm sui primi 32 canali

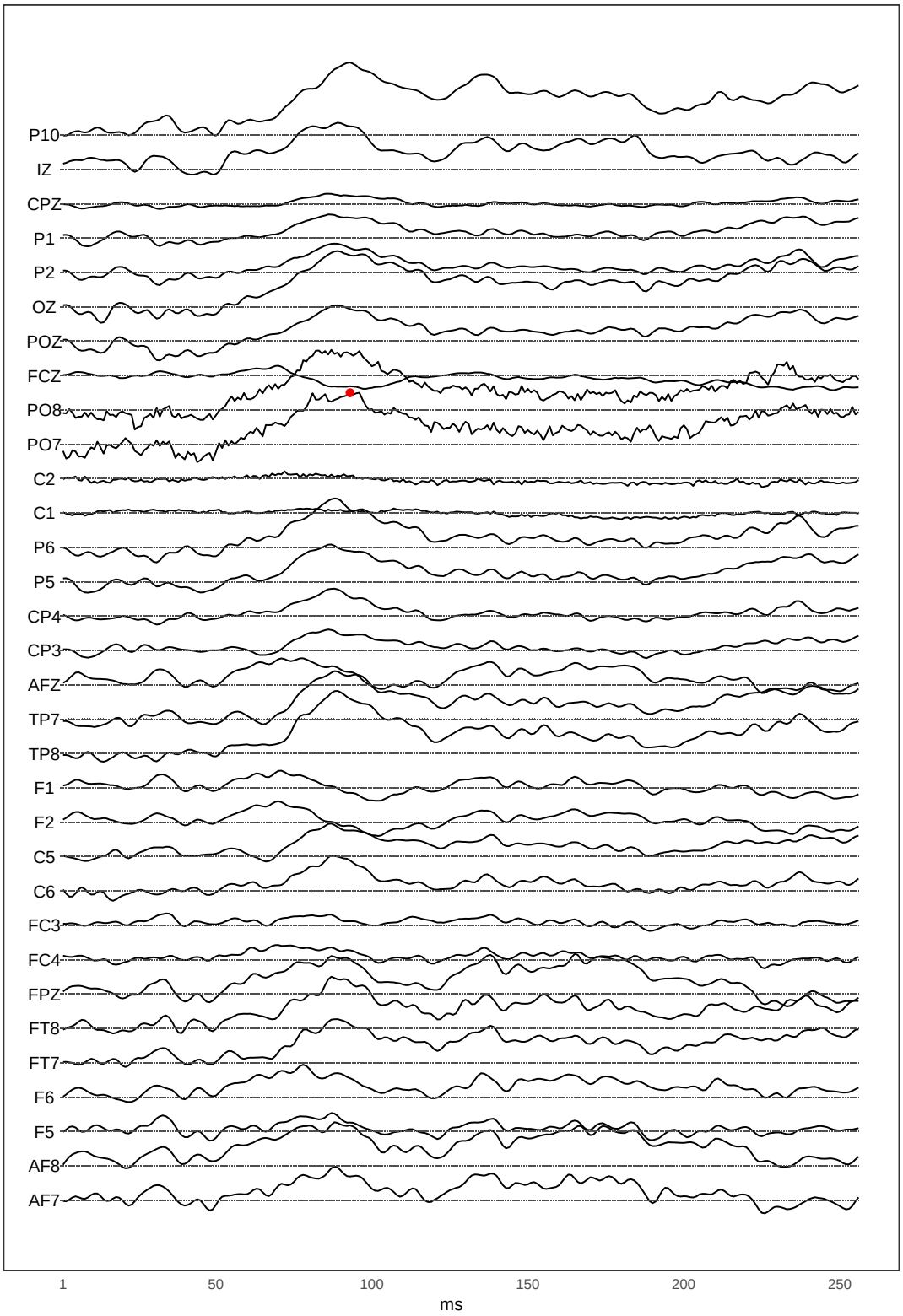

Figura 4.4: Scoperte ottenute tramite i metodi di Bonferroni e Holm sugli ultimi 32 canali

4.6 Controllo del FDP: la All-Resolutions Inference

Una procedura di controllo dell'errore di I tipo molto più flessibile rispetto quanto visto finora, è quella proposta da Goeman e Solari (2011) e ribattezzata come ARI (All-Resolutions Inference) da Rosenblatt et al. (2018) nella sua prima applicazione allo studio delle immagini di fMRI. Attraverso la procedura ARI si è potuto, in parte, superare uno dei principali problemi delle varie tecniche utilizzate nell'analisi dell'fMRI, cioè quello del *paradosso della specificità spaziale*, di cui si parlerà nel seguito. Un beneficio simile può essere ottenuto anche in un contesto di analisi delle onde ERP, infatti la procedura ARI non solo consente di ottenere una stima del TDP (True Discovery Proportion) in una particolare area di interesse, ma dà anche la possibilità di modificare a posteriori il livello di approfondimento dell'analisi, in modo da individuare con una precisione sempre maggiore gli istanti temporali in cui, in un determinato canale, l'effetto della condizione a cui il soggetto viene sottoposto in fase di rilevazione del segnale risulta significativo.

Il controllo dell'errore operato attraverso il metodo ARI si sviluppa in un'ottica opposta rispetto a quella che ha caratterizzato i metodi visti finora. Infatti, se all'origine dei metodi di Bonferroni, Holm e Benjamini&Hochberg vi è la definizione del livello massimo di errore che si è disposti ad accettare, sulla base del quale viene definito l'insieme di ipotesi significative, nella procedura ARI è possibile definire un certo sottoinsieme di ipotesi che potrebbero essere considerate delle scoperte (e quindi false) e soltanto dopo si effettua una stima dell'errore che si otterrebbe da quel sottoinsieme. In particolare, per il sottoinsieme scelto si stima il False Discovery Proportion (FDP) o, equivalentemente, il True Discovery Proportion (TDP), cioè rispettivamente la proporzione di ipotesi rigettate erroneamente e quella delle ipotesi rigettate correttamente sul totale delle ipotesi considerate.

Per capire come poter arrivare a tale stima, si consideri l'insieme m contenente un certo numero di differenze di segnale osservate e si ipotizzi di voler individuare quante di queste differenze sono realmente significative nel sottoinsieme di m, S. Fra i $2^{|m|}$ sottoinsiemi che è possibile costruire con gli elementi contenuti in m, il sottoinsieme S è scelto in modo arbitrario in base a considerazioni di natura teorica o legate agli obiettivi dello studio, o in base all'osservazione dei dati stessi. Si identifichino, inoltre, con A e $a(S)$ i sottoinsiemi, rispettivamente, di m e di S contenenti solo le osservazioni in cui vi è un effetto reale. Il sistema di ipotesi attorno al quale si sviluppa la procedura è:

$$H_S : a(S) = 0$$

Chiaramente, sia A che $a(S)$ sono ignoti, pertanto sarà necessario svolgere una verifica di ipotesi per ogni possibile costruzione dell'insieme $a(S)$: esso può essere

costituito da una sola osservazione, e in questo caso H_S è detta *ipotesi elementare*, oppure può comprendere diverse osservazioni, e in questo caso l'ipotesi H_S è costruita sull'intersezione delle varie ipotesi elementari. Ad esempio, se $S = (\delta_1, \delta_2, \delta_3)$, dove δ indica una differenza di segnale fra due condizioni, le ipotesi *elementari* sono:

$$H_1 : \delta_1 = 0$$

$$H_2 : \delta_2 = 0$$

$$H_3 : \delta_3 = 0$$

mentre quella che comprende tutte e tre le osservazioni è data da:

$$H_{123} : H_1 \cap H_2 \cap H_3$$

Altre ipotesi possono essere costruite considerando gruppi di due differenze.

In ogni caso, se H_S viene rigettata, allora il sottoinsieme $a(S)$ contiene almeno una osservazione in cui è presente un effetto, al contrario, se $a(S)$ viene accettata si assume che tutte le osservazioni che ne fanno parte siano prive di effetto.

La verifica di queste ipotesi viene svolta tramite i cosiddetti *test locali* che possono assumere diverse forme a seconda del numero totale di osservazioni che vengono tenute in considerazione nell'intera procedura. Infatti, se l'insieme S contiene fino a 20 osservazioni è possibile utilizzare qualsiasi tipo di test sia ritenuto più adatto ai dati in esame, ma se il numero di osservazioni è più alto si avrebbero in totale troppi test da sviluppare, rendendo i tempi computazionali troppo lunghi. Pertanto, in questo secondo caso, è preferibile sfruttare algoritmi in grado di velocizzare la procedura, come quello delle combinazioni di Fisher o quello dei test basati sulla disuguaglianza di Simes. Il primo può essere utilizzato solo con ipotesi indipendenti fra loro, mentre il secondo può essere utilizzato sia in caso di indipendenza sia nel caso di correlazioni non negative fra le ipotesi (PRDS). Data la presenza di correlazione nei dati di ERP, di seguito si approfondirà solo la seconda tipologia di test.

Si consideri la seguente disuguaglianza, detta *disuguaglianza di Simes*, che riguarda l'insieme $m \setminus A$, cioè solo le osservazioni in cui non vi è un effetto:

$$P(p_{m \setminus A} \leq \alpha) \leq \alpha$$

Se si assume vera l'ipotesi di correlazioni non negative fra i test (PRDS), la disuguaglianza di Simes è valida, quindi può essere utilizzato il test di Simes che porta al rifiuto di H_S se:

$$p_S \leq \alpha$$

$$\text{dove } p_S = \min_{1 \leq i \leq |S|} \frac{|S|}{i} p_{(i)}$$

Lo sviluppo della procedura ARI, tuttavia, prevede che in un primo momento gli insiemi m ed S coincidano, quindi le ipotesi H_S saranno costituite da tutti i $2^{|m|}$ sottoinsiemi ottenibili a partire da m, mentre la scelta del sottoinsieme di interesse S viene fatta soltanto in un secondo momento. In questo modo il numero totale di ipotesi da verificare può diventare molto elevato, quindi molto spesso è utile affiancare ai test di Simes una procedura di controllo del FWER, detta *closed testing*. Essa prevede che l'ipotesi H_S possa essere rifiutata se e solo se H_I è rifiutata per tutte le $I \supseteq a(S)$. Ad esempio, si consideri ancora $S = (\delta_1, \delta_2, \delta_3)$ e una delle ipotesi elementari che dovrà essere testata

$$H_1 : a(S) = \delta_1 = 0$$

Essa può essere rigettata solo se lo sono anche H_{123}, H_{12}, H_{13}, infatti I identifica gli insiemi che contengono δ_1, cioè $(\delta_1, \delta_2, \delta_3)$, (δ_1, δ_2), (δ_1, δ_2).

Inoltre, secondo il closed testing, il rifiuto di H_S avviene se e solo se:

$$\min_{1 \leq i \leq |S|} \left\{ \tfrac{h}{i} p_{(i:S)} \right\} \leq \alpha$$

$$\text{dove } h = max \left\{ i \in \{0, ..., m\} : i p_{(m-i+j:m)} j\alpha, \quad \text{per} \quad j = 1, ..., i \right\}$$

La quantità h può essere interpretata come la dimensione del più grande insieme di osservazioni non rigettate dal test di Simes. Anche sotto questa correzione, la procedura ARI è valida solo se vale la disuguaglianza di Simes.

Attraverso la verifica delle ipotesi appena viste, dato l'insieme S, si può procedere alla stima di:

$$\widehat{TDP}(S) = \frac{\hat{a}(S)}{|S|}$$

Generalmente, tale stima viene espressa sotto forma di intervallo di confidenza unilaterale di livello $1 - \alpha$, dove il limite inferiore è dato dalla stima $\widehat{TDP}(S)$ ottenuta e quello superiore è fisso e pari a 1. Analogamente, può essere costruito un intervallo per il FDP dove il limite inferiore è fisso e pari a 0 e quello superiore è pari a 1-TDP.

La procedura appena descritta è molto potente e allo stesso tempo flessibile perché non impone nessun rifiuto ma dà soltanto un'indicazione su quale sottoinsieme di osservazioni S riuscirebbe a dare un valore alto dell'indice TDP. Come già accennato, dato un insieme S, se la stima di $a(S)$ (o equivalentemente del TDP) è diversa da 0, anche se di poco, allora è possibile affermare che almeno una osservazione contenuta in S è caratterizzata da un effetto significativo. Tuttavia, se l'interesse è quello di individuare con precisione quali osservazioni hanno portato ad una stima di $a(S)$ (o

del TDP) diversa da 0, considerare un insieme S molto grande potrebbe rappresentare un problema (*paradosso della specificità spaziale* [8]) che è tanto più grave quanto più piccola è la stima del TDP. In questi casi, infatti, l'utilizzo della procedura ARI risulta particolarmente utile dal momento che, consentendo una scelta a posteriori dell'insieme S, l'analisi non è limitata ad una sola scelta ma può essere estesa anche a eventuali restrizioni di S scelte in base al livello di precisione che si desidera ottenere nel risultato, senza che questo comprometta la qualità dell'inferenza ottenuta.

4.6.1 Applicazione ai dati

Come già visto nei paragrafi precedenti, i dati presi in esame in questo elaborato richiedono il confronto fra le condizioni S1 e S2 di ben 16384 osservazioni, pertanto nello sviluppo della procedura ARI l'insieme m è costituito proprio da 16384 test. Fra tutti i sottoinsiemi S che possono essere costruiti con tali osservazioni, ne sono stati selezionati solo alcuni ritenuti di particolare interesse.

Per analogia con i metodi visti nei paragrafi precedenti, si è scelto, innanzitutto, di studiare i risultati senza alcuna riduzione dell'insieme di osservazioni, cioè considerando $S \equiv m$. Ne risulta che $TDP \in (0.0054, 1)$, cioè che se si dovessero considerare tutte le 16384 ipotesi come delle scoperte (cioè come osservazioni con un effetto significativo) almeno lo 0.54% di esse sarebbe una vera scoperta (cioè solo 88). Questo risultato è stato ottenuto tramite la procedura di correzione del closed testing, vista al paragrafo precedente, dove fra le ipotesi elementari soltanto una è stata caratterizzata da un p-value significativo, cioè quella relativa al 93° ms sul canale PO7.

In base a quanto appena ottenuto, cioè mantenendo $m = 16384$, si è poi proceduto a varie restrizioni dell'insieme S. Il primo criterio seguito è quello secondo cui S è costituito dalle osservazioni relative a un particolare canale e questo ha, quindi, portato alla costruzione di 64 insiemi S diversi. Come ci si poteva aspettare dai risultati precedentemente ottenuti, in quasi tutti i canali la stima del limite inferiore del TDP è risultata esattamente pari a 0. Gli unici due canali per cui tale stima è risultata diversa da 0 sono CZ e PO7, i cui risultati sono riassunti nella Tabella 4.3.

Tabella 4.3: Stime di TDP e FDP secondo ARI per singoli canali

Canali	*Vere Scoperte*	TDP	FDP
CZ	7	0.0273	0.973
PO7	1	0.004	0.996

Dunque, è possibile dire che ipotizzando di rigettare l'intero gruppo di ipotesi sulle osservazioni relative rispettivamente ai canali CZ e PO7, si avrebbe che il 27% di queste

sarebbe una reale scoperta nel canale CZ, mentre nel canale PO7 lo sarebbe lo 0.004% del totale delle osservazioni avvenute su quel canale.

In Figura 4.5 sono riportati i 64 canali con un'intensità di colore tanto più bassa quanto più alto è il valore del TDP stimato, infatti la maggior parte dei canali, che ha presentato una stima del TDP pari a 0, è rappresentata con il colore più scuro (arancione), mentre si riescono a distinguere i canali CZ (al centro della corteccia) e quello PO7.

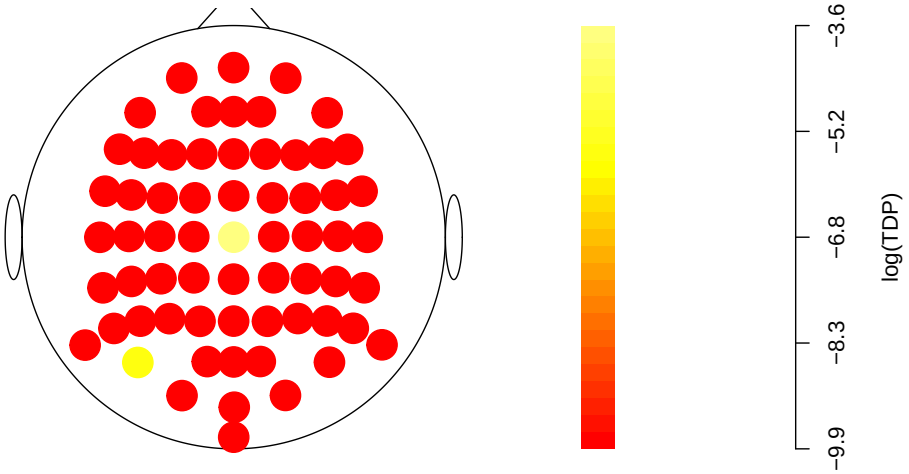

Figura 4.5: Rappresentazione grafica dell'intensità del TDP stimato

Una seconda scelta dell'insieme S è quella che include tutte le osservazioni avvenute nell'intervallo temporale compreso fra il 70° e 110°ms senza alcuna distinzione fra i canali, analogamente a quanto fatto al Capitolo 3 per lo sviluppo dei modelli standard. Successivamente, si è provato a restringere sempre di più tale intervallo, fino a renderlo molto piccolo, in modo da cercare di capire se le eventuali osservazioni con un effetto fossero tutte concentrare in un intervallo ancora più ristretto o meno.

Dai risultati ottenuti, mostrati in Tabella 4.4, sembra che le osservazioni con effetto significativo siano abbastanza distribuite in tutto l'intervallo, infatti la riduzione dell'intervallo temporale è accompagnata, quasi sempre, da una riduzione anche del numero di *vere scoperte*, pertanto non si ha modo di osservare un aumento nella stima del TDP. L'unica eccezione è nel passaggio dall'intervallo 85-95ms a quello 87-93ms, dove si osserva un aumento del TDP da 0.0043 a 0.0067. Dunque, è possibile affermare che, almeno in un canale, tutto l'intervallo fra il 70° e 110°ms è caratterizzato da un'attività cerebrale statisticamente diversa fra le condizioni S1 e S2.

Tabella 4.4: Stime di TDP e FDP secondo ARI su sotto-intervalli temporali

ms	Vere Scoperte	TDP	FDP
70-110	19	0.0072	0.9928
75-105	12	0.0060	0.9940
80-100	9	0.0067	0.9933
85-95	3	0.0043	0.9957
87-93	3	0.0067	0.9933
89-91	1	0.0052	0.9948

4.7 Risultati a confronto

La Tabella 4.5 riporta una sintesi dei risultati più rilevanti ottenuti con i metodi sviluppati finora. In particolare, si riportano i risultati ottenuti con i modelli a effetti misti e MANOVA a seguito della correzione effettuata col metodo di Holm, indicando, per semplicità interpretativa, la significatività del p-value con un asterisco, mentre se il p-value non è risultato significativo, il canale è contrassegnato da un punto. Inoltre, sono stati omessi i risultati ottenuti col metodo di Bonferroni perché esattamente corrispondenti a quelli ottenuti col metodo di Holm. Infine, sono stati riportati anche in questo caso solo 32 canali, che sono quelli per cui sono stati ottenuti risultati di rilievo con almeno uno dei metodi adottati, mentre i restanti 32 canali non hanno manifestato alcuna significatività.

Dalla Tabella 4.5 sembra che l'effetto della condizione sul segnale generato non sia, in nessuno dei canali di rilevazione, tanto forte da risultare significativo secondo tutte le procedure utilizzate. Tuttavia, i canali dove tale effetto sembra esistere sono CZ e PO7, dove almeno una osservazione è risultata significativa attraverso i metodi di Holm e/o ARI. Queste ultime due procedure sono risultate particolarmente conservative, infatti hanno evidenziato un numero di scoperte molto basso rispetto al totale delle osservazioni, al contrario di quanto fatto dai metodi standard che, invece, nonostante i p-value generati siano anch'essi soggetti alla correzione *post hoc*, evidenziano un effetto significativo della condizione in diversi canali.

Tabella 4.5: Sintesi dei risultati ottenuti

Canali	**Modello a effetti misti**	**Modello MANOVA**	**Holm**	**ARI**
AF8	*	.	0	0
C3	.	.	0	0
C5	.	.	0	0
C6	.	.	0	0
CP3	.	.	0	0
CP5	.	.	0	0
CP6	.	.	0	0
CZ	*	*	0	7
F4	.	.	0	0
F6	.	.	0	0
F8	*	.	0	0
FT7	.	.	0	0
FT8	*	.	0	0
IZ	*	.	0	0
O1	*	*	0	0
O2	*	*	0	0
OZ	*	*	0	0
P10	*	*	0	0
P3	.	.	0	0
P5	.	.	0	0
P6	*	.	0	0
P7	*	.	0	0
P8	*	.	0	0
P9	*	.	0	0
PO1	.	.	0	0
PO2	.	.	0	0
PO7	*	.	1	1
PO8	*	.	0	0
T7	*	*	0	0
T8	*	*	0	0
TP7	*	*	0	0
TP8	*	.	0	0

Capitolo 5

Metodo basato sulle permutazioni

Quelli visti finora fanno parte di un insieme di metodi parametrici che, per essere considerati affidabili, richiedono la validità di determinate assunzioni. In molti casi, tuttavia, queste possono risultare piuttosto stringenti e poco realistiche e se non vengono rispettate dai dati raccolti, il rischio di commettere degli errori può diventare alto. Si fa riferimento, ad esempio, all'ipotesi di normalità delle variabili, alla loro indipendenza, all'omoschedasticità e alla necessità di avere un campione casuale. Per questo motivo, di seguito verrà introdotto un approccio di tipo *non parametrico* che consente di rilassare alcune fra queste ipotesi, rendendo la procedura di analisi sfruttabile in contesti con caratteristiche molto varie.

Come introdotto al Capitolo 1, il tipico disegno sperimentale in uno studio sulle componenti ERP prevede la costruzione e il confronto di due campioni non indipendenti, perché costituiti dagli stessi soggetti e sottoposti a due condizioni sperimentali diverse, che nello studio presentato al Capitolo 1 sono chiamate S1 e S2. La procedura che verrà illustrata di seguito riguarda tutte le possibili combinazioni $W = e \times t$ con $W = 1, ..., 16384$, $e = 1, ..., 64$ e $t = 1, ..., 256$, dove e e t indicano rispettivamente i canali di rilevazione e gli istanti temporali: per semplificare la notazione questi indici verranno omessi.

Il sistema di ipotesi di interesse è:

$$\begin{cases} H_0 : \bar{\delta} = 0 \\ H_1 : \bar{\delta} \neq 0 \end{cases}$$

con $\bar{\delta} = \frac{\sum_{i=1}^{10} \delta_i}{10}$ che è la media del vettore $\delta = (\delta_1, \delta_2, ..., \delta_{10})$ il cui generico elemento è $\delta_i = y_{i,S1} - y_{i,S2}$, cioè la differenza fra le osservazioni sotto le condizioni S1 e S2, condizionatamente al soggetto i. Come già visto nei capitoli precedenti, la verifica di tale ipotesi si svolge tramite il calcolo di un p-value. In un contesto parametrico, per ottenere un'approssimazione del p-value e quindi poter sviluppare l'analisi, risulta necessario assumere che i dati siano distribuiti normalmente, infatti solo sotto questa

assunzione è possibile ipotizzare che la distribuzione nulla sia una t di Student, da cui è facile ricavare il p-value.

Tuttavia, può accadere in alcuni contesti che non si abbia alcuna idea di quale sia il processo generatore dei dati o che esso sia molto distante da una distribuzione normale: nel primo caso sarebbe impossibile svolgere l'analisi secondo un approccio parametrico, mentre nel secondo caso si potrebbero ottenere soltanto risultati fortemente approssimati. Al contrario, un approccio non parametrico dà la possibilità di costruire una distribuzione nulla esatta a partire dai dati osservati e, per questo motivo, risulta spesso preferibile quando non si hanno informazioni a sufficienza riguardo il processo generatore dei dati. Un altro importante vantaggio dell'utilizzare un approccio non parametrico è il fatto di consentire un perfetto adattamento della distribuzione nulla alla struttura di dipendenza empiricamente osservabile nei dati. In questo modo, si può tener conto di tale struttura nello svolgimento dell'analisi senza dover avanzare ipotesi talvolta troppo stringenti, come quella dell'indipendenza fra le osservazioni.

Per capire come si può arrivare ad una stima esatta della distribuzione nulla secondo l'approccio di permutazione è necessario, innanzitutto, definire alcuni concetti. Dato il contesto di dati appaiati, si definiscono il fattore z che identifica i 10 soggetti sottoposti allo studio e che sarà considerato come fattore di stratificazione e il fattore x che identifica la condizione a cui viene sottoposto il soggetto, che in questo caso assume i due livelli S1 e S2.

Sotto l'ipotesi nulla, qualsiasi differenza δ_i osservata è dovuta unicamente al caso, quindi risulta ugualmente probabile osservare una differenza positiva o una negativa. Di conseguenza, per ogni combinazione di e e t si ha un vettore $g = (g_1, g_2, ..., g_{10}) = \{-1, 1\}^{10}$ che determina il segno da attribuire alle differenze calcolate e che può assumere 2^{10} conformazioni diverse [13]. Dunque, se si sceglie di seguire un approccio permutativo con dati appaiati l'oggetto di primario interesse è il vettore delle differenze permutato, cioè $\delta^* = (g_1\delta_1, g_2\delta_2, ..., g_{10}\delta_{10})$. L'insieme di tutte le possibili conformazioni che può assumere il vettore δ^* è detto *orbita* e ognuno degli elementi che ne fanno parte ha la stessa probabilità di presentarsi. L'*orbita* è un sottoinsieme di tutti i possibili vettori di differenze che si potrebbero osservare.

In questo contesto non è necessario assumere che i dati provengano dalla stessa distribuzione ma, affinché la procedura sia valida, basta assumere che ci sia una certa simmetria attorno alla media sia nel vettore di osservazioni originario δ, sia in quello permutato δ^*. Dunque, poiché sotto H_0 la media delle differenze δ è nulla, si ha:

$$f(y_i|x_i = S1, z_i) - f(y_i|x_i = S2, z_i) = f(y_i|x_i = S2, z_i) - f(y_i|x_i = S1, z_i) \quad \forall z_i$$

Questa assunzione implica la scambiabilità dei dati condizionatamente ai soggetti e non necessita l'assunzione poco realistica di omoschedasticità fra gli stessi.

Dunque, è possibile riscrivere il sistema di ipotesi come

$$\begin{cases} H_0 : f(y_i|x_i = S1, z_i) = f(y_i|x_i = S2, z_i) & \forall z_i \\ \iff f(\delta_i) = f(\text{-}\delta_i) & \forall z_i \\ H_1 : f(\delta_i) \neq f(\text{-}\delta_i) & \text{per almeno un i} \end{cases}$$

5.1 La procedura di Westfall&Young

In questo elaborato le tecniche permutative verranno sfruttate per la correzione dell'errore nello svolgimento dei test multipli. In particolare, verrà approfondito il metodo del maxT proposto da Westfall&Young [7]. Esso può essere sviluppato soltanto sotto l'assunzione di scambiabilità e si basa sul calcolo della statistica t già introdotta al Capitolo 4, per ogni combinazione di istante temporale e canale di rilevazione. L'idea di base su cui si fonda la procedura è quella per cui se i due gruppi di segnale registrato S1 e S2 sono statisticamente uguali fra loro, il segno attribuito alla differenza di segnale calcolata fra le due condizioni è essenzialmente dovuto al caso, quindi non dovrebbe provocare alcuna differenza nella stima della statistica t ottenuta.

Il primo passo per sviluppare la procedura di Westfall&Young è quello di stabilire il numero M di permutazioni da svolgere. Successivamente, per la generica permutazione m, si percorrono i seguenti passi:

1) Costruzione del vettore di differenze permutato $\delta^*_W = (g_1\delta_1, g_2\delta_2, ..., g_{10}\delta 10)$ per ogni W con $W = 1, ..., 16384$;

2) Calcolo della statistica test t_W;

3) Fra le W statistiche calcolate, si sceglie quella con valore più alto in modo da ottenere il vettore t_{max} di dimensione M, costituito da tutti i valori t più alti ottenuti nelle M permutazioni.

In questo modo, si è ottenuta una stima della distribuzione nulla di t_{max}, cioè della distribuzione del più alto valore di t fra i W calcolati ipotizzando che H_0 sia vera. A partire da tale distribuzione si individua il valore critico α_0, cioè quello al di sopra del quale vi è solo il 5% dei valori t_{max} calcolati: le statistiche t calcolate verranno confrontate con questo valore critico, in modo da rigettare tutte le realizzazioni che in valore assoluto risultano più grandi di esso, cioè che ricadono nella regione di rifiuto. Successivamente, la procedura riprende direttamente dal punto 3), cioè dall'individuazione degli M test t più grandi che verranno sfruttati per una nuova stima della distribuzione nulla e quindi per l'individuazione di un nuovo valore critico, α_1, che porterà all'esclusione di altri test. La procedura continua finché nessun test ricade più nella zona di

rifiuto o finché sono stati già rigettati tutti i test a disposizione. Nello svolgimento di tale procedura è da considerare la possibilità che la rimozione di alcune ipotesi riduca il valore di alcuni t_{max} individuati, per cui potrebbe accadere che $\alpha_1 < \alpha_0$.

Un'alternativa al calcolo dei valori critici α è il calcolo diretto dei p-value per ognuno dei test ottenuti. Essi possono essere ottenuti, a partire dalla distribuzione nulla stimata, come

$$P_{t_m} = \frac{1}{M} \sum_{m=1}^{M} I_{t_m \geq t_0}$$

Si otterranno, dunque, dei p-value discreti e multipli di $\frac{1}{M}$, infatti il valore minimo che possono assumere non è 0 ma proprio $\frac{1}{M}$.

Nella procedura appena illustrata si può notare come il numero M di permutazioni svolga un ruolo di fondamentale importanza per definire l'accuratezza con cui la distribuzione nulla viene stimata, infatti la situazione ideale sarebbe quella di svolgere tutte le permutazioni possibili. Nella maggior parte dei casi, però, questo numero è talmente elevato da rendere la procedura troppo complessa dal punto di vista computazionale, quindi ci si accontenta di sfruttare soltanto un certo sottoinsieme di permutazioni di dimensione M, selezionato in modo casuale. Una volta scelto il livello di significatività teorico α, si tende a porre come numero minimo di permutazioni da considerare $\frac{1}{\alpha}$. Secondo Goeman et al. (2011), se α=0.05 un buon numero di permutazioni è 1000.

Un'alternativa al maxT è il cosiddetto metodo del minP, anch'esso formulato da Westfall&Young. Esso prevede, innanzitutto, la costruzione di una matrice P di dimensione W, che contenga i p-value grezzi ottenuti per tutte le ipotesi ad ogni permutazione. Successivamente, si costruisce una seconda matrice delle stesse dimensioni, $\tilde{P}$, che contenga i p-value grezzi permutati rispetto alle ipotesi. Il generico elemento di $\tilde{P}$ è:

$$p_{kj} = \#\{l : p_{kl} \leq p_{kj}\}$$

Analogamente alla procedura del maxT, su questa matrice verranno ricavati i valori critici ed individuate, nei diversi passi della procedura, le ipotesi da rigettare. La procedura del minP risulta meno pratica rispetto a quella del maxT sia perché, per ottenere risultati affidabili, richiede un numero di permutazioni molto più alto, cioè almeno pari a $\frac{W}{\alpha}$, sia perché i p-value che vengono generati assumono tipicamente un numero di valori molto limitato, quindi molti di questi valori saranno ripetuti fra le varie ipotesi.

5.2 Quando conviene?

Nonostante talvolta possa risultare piuttosto complesso dal punto di vista computazionale, l'utilizzo di un approccio permutativo risulta spesso preferibile agli approcci parametrici per diverse ragioni. Si tratta, ad esempio, del caso in cui il numero di osservazioni che si hanno a disposizione è molto basso, infatti in questo caso la loro distribuzione empirica risulta particolarmente lontana dalle distribuzioni asintotiche che vengono tipicamente prese come riferimento nei metodi parametrici, introducendo nei risultati un livello di approssimazione talvolta molto alto.

Un'altra importante caratteristica dell'approccio non parametrico è il fatto che in esso non è necessario effettuare alcuna assunzione sulla struttura di dipendenza fra i dati. Infatti, mentre procedure come quella di Benjamini&Hochberg e ARI valgono soltanto sotto l'assunzione di totale indipendenza o di correlazione non negativa fra i test (PRDS), in un contesto permutativo la distribuzione nulla stimata a cui si fa riferimento si adatta alla struttura di dipendenza empiricamente osservata. Questo risulta particolarmente utile quando non si è in grado di stabilire a priori come possano esere messi in relazione fra loro i dati, come nel caso del segnale ERP [7].

Infine, alcuni studi di simulazione [14] hanno evidenziato in generale una maggiore potenza delle procedure non parametriche rispetto a quelle parametriche. Ad esempio, il metodo di Westafall&Young, tramite il calcolo di p-value esatti, garantisce una maggiore potenza nel controllo del FWER rispetto a quanto possa garantire la procedura di Bonferroni. Quest'ultima, infatti, pone come limite massimo al valore soglia la quantità $\frac{\alpha}{W}$, mentre la procedura del maxT, anche grazie al fatto che il valore soglia α è stimato a partire da una distribuzione esatta, genera spesso una stima più alta di $\frac{\alpha}{W}$ e quindi meno conservativa [7].

5.3 Sviluppo e valutazione della procedura maxT

Nell'applicazione del metodo del maxT ai dati in esame, si è scelto ancora una volta di stabilire come livello di significatività massimo $\alpha = 0.05$. Quindi, secondo le indicazioni di Goeman et al. (2011), per avere dei risultati sufficientemente attendibili basterebbero M=1000 permutazioni. Tuttavia, dal momento che i dati in esame riguardano soltanto 10 soggetti, si è scelto di calcolare e sfruttare tutte le permutazioni che si possono costruire con questo campione, cioè $2^{10} = 1024$, in modo tale da evitare l'eventuale compromissione dei risultati dovuta alla casualità con cui si sarebbe dovuto scegliere il sottoinsieme di permutazioni da utilizzare.

Con questo approccio si è ottenuto il rigetto di 8 test relativi al canale CZ nell'intervallo di osservazione fra 88 e 100 ms, e di un solo test per il canale PO7, al 93° millisecondo di osservazione. Questo risultato è riportato, insieme agli altri nella

Tabella 5.1: esso è in linea con quanto ottenuto da Winkler et al. (2014) secondo cui il metodo maxT risulta più potente rispetto ad Holm e ad ARI, anche se solo leggermente.

Nelle Figure 5.1 e 5.2 si riporta lo stesso grafico già visto al Capitolo 4, con l'aggiunta delle scoperte ottenute tramite il metodo del maxT di Wetfall&Young.

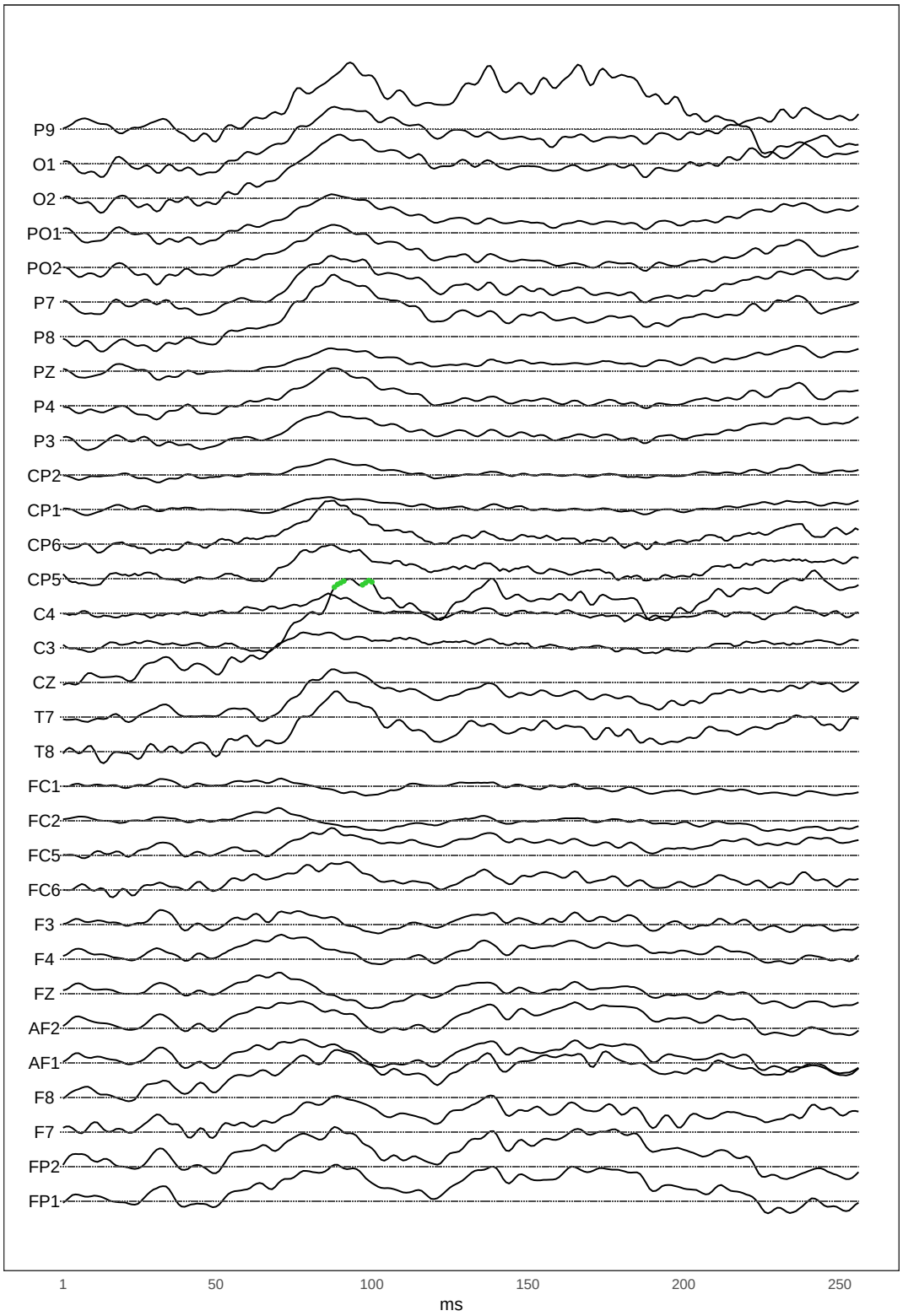

Figura 5.1: Scoperte ottenute tramite i metodi di Holm e maxT sui primi 32 canali

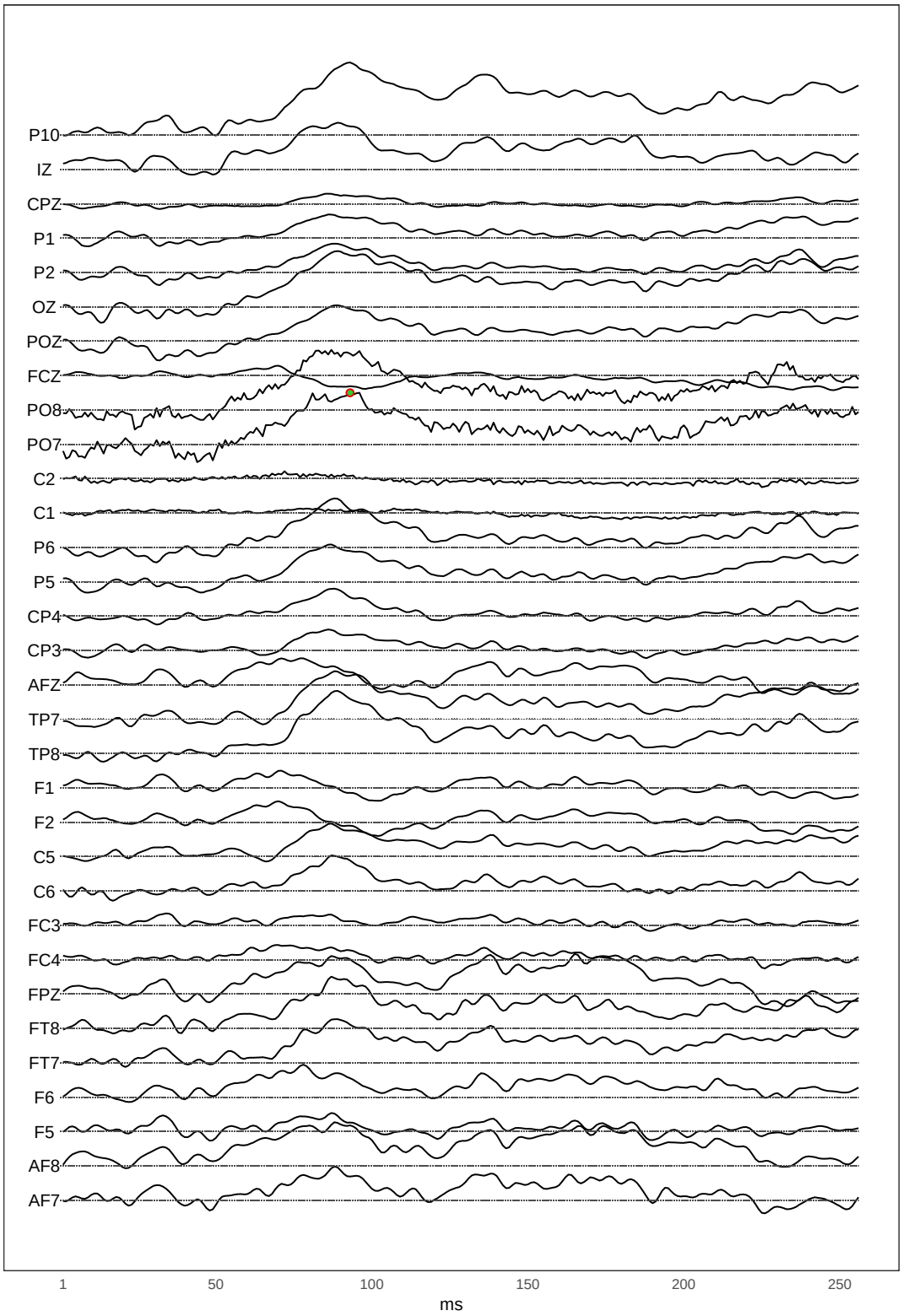

Figura 5.2: Scoperte ottenute tramite i metodi di Holm e maxT sugli ultimi 32 canali

5.3.1 Analogie col modello MANOVA

Fra tutti i metodi di analisi illustrati finora, si possono notare alcune analogie fra il modello MANOVA e il metodo basato sulle permutazioni. Si consideri, innanzitutto, la formula generale del modello MANOVA che si utilizza una volta calcolate le ampiezze Y, per ogni canale e:

$$Y_{eij} = c + \mu_{ei} + \gamma_e X_{ij} + \epsilon_{eij}$$

dove i e j sono rispettivamente gli indici dei soggetti e del numero di osservazione relativo al soggetto, con i=1,...,10 e j=1,2, e μ_{ei} rappresenta quella parte di segnale dovuta direttamente alle caratteristiche intrinseche del soggetto. Nel confrontare il segnale registrato sotto le due condizioni su un certo soggetto, quest'ultima componente si annulla. Infatti, per il generico canale e e considerando la condizione S1 come baseline, il modello MANOVA assume le seguenti forme rispettivamente sotto le condizioni S1 e S2 per il soggetto i-esimo:

$$Y_{ei1} = c + \mu_{ei} + \epsilon_{ei1}$$

$$Y_{ei2} = c + \mu_{ei} + \gamma_e X_{i2} + \epsilon_{ei2}$$

da cui:

$$\begin{aligned}\delta_{ei} = Y_{ei2} - Y_{ei1} &= c + \mu_{ei} + \gamma_e X_{i2} + \epsilon_{ei2} - (c + \mu_{ei} + \epsilon_{ei1}) \\ &= \gamma_e X_{i2} + \epsilon_{ei2} - \epsilon_{ei1}\end{aligned}$$

Il metodo basato sulle permutazioni, invece, si sviluppa a partire dal vettore delle differenze di segnale nelle due condizioni $\delta = (\delta_1, \delta_2, ..., \delta_{10})$, calcolate sui singoli soggetti. Dunque, in entrambi gli approcci le caratteristiche intrinseche del soggetto sono considerate una quantità ignota ma fissa che, quindi, può essere trascurata. Oltre al fatto di consentire un confronto del segnale diretto, i due metodi sono accomunati dalla possibilità di riuscire ad ottenere delle stime senza dover necessariamente assumere una distribuzione normale dei dati.

Dunque, si è scelto di svolgere una seconda volta la procedura del maxT, non più sulle singole osservazioni ma sulle ampiezze già calcolate al Capitolo 3. In questo modo, si è ottenuto per ogni canale un singolo p-value aggiustato che indica la presenza o meno di un certo effetto, come mostrato in Tabella 5.1. Questo metodo ha rilevato una differenza significativa del segnale fra le due condizioni in 17 canali, contrariamente a quanto ottenuto al Capitolo 4 con la correzione dei p-value ottenuti dal modello MANOVA attraverso il metodo di Holm, che aveva portato a rigettare l'ipotesi di uguaglianza dei segnali soltanto in 8 canali. Questa differenza potrebbe essere attribuita

alla maggiore capacità dei metodi permutativi di gestire le dipendenze fra i test.

Inoltre, si può osservare come il segnale sembri essere diverso fra le condizioni in un numero di canali maggiore rispetto a quanto si ottiene dall'applicazione dello stesso metodo (maxT) sulle singole osservazioni.

Tabella 5.1: Sintesi dei risultati ottenuti

Canali	**Effetti Misti Holm**	**MANOVA Holm**	**maxT ampiezze**	**Holm**	**ARI**	**maxT**
AF8	*	.	.	0	0	0
C3	.	.	.	0	0	0
C5	.	.	*	0	0	0
C6	.	.	.	0	0	0
CP3	.	.	.	0	0	0
CP5	.	.	*	0	0	0
CP6	.	.	.	0	0	0
CZ	*	*	*	0	7	8
F4	.	.	.	0	0	0
F6	.	.	.	0	0	0
F8	*	.	*	0	0	0
FT7	.	.	.	0	0	0
FT8	*	.	*	0	0	0
IZ	*	.	.	0	0	0
O1	*	*	*	0	0	0
O2	*	*	*	0	0	0
OZ	*	*	*	0	0	0
P10	*	*	*	0	0	0
P3	.	.	.	0	0	0
P5	.	.	.	0	0	0
P6	*	.	.	0	0	0
P7	*	.	*	0	0	0
P8	*	.	*	0	0	0
P9	*	.	.	0	0	0
PO1	.	.	.	0	0	0
PO2	.	.	.	0	0	0
PO7	*	.	*	1	1	1
PO8	*	.	*	0	0	0
T7	*	*	*	0	0	0
T8	*	*	*	0	0	0
TP7	*	*	*	0	0	0
TP8	*	.	*	0	0	0

5.4 Il metodo *cluster mass*

L'ultimo approccio che si è scelto di sviluppare per l'analisi del segnale ERP è il cosiddetto *cluster mass*, proposto da Maris et al.(2007) e che, in qualche modo, riesce a coniugare i vari punti di forza dei metodi visti finora. Si tratta, infatti, di un approccio che parte dall'analisi della singola osservazione per poi svilupparsi attraverso la formazione e la valutazione di gruppi di osservazioni accomunati da certe caratteristiche, detti *cluster* [17]. L'idea di base è analoga a quella dei metodi standard sviluppati al Capitolo 3, cioè si assume che se in una singola osservazione è presente un effetto reale, allora è molto probabile che questo effetto sia presente anche nelle osservazioni temporalmente adiacenti ad essa. Tuttavia, mentre nei metodi standard l'analisi si concentra sulle osservazioni che ricadono in un particolare intervallo temporale ritenuto di interesse, ad esempio per ragioni teoriche, nel caso del *cluster mass* l'individuazione dei cluster avviene esclusivamente in base ai livelli di segnale effettivamente registrati e alla scelta di un certo *valore soglia*, come si vedrà più avanti. Infatti, in questo secondo caso non si è in grado di stabilire a priori il numero di cluster che verranno individuati né la loro dimensione. Inoltre, mentre per i primi l'analisi si concentra esclusivamente sul gruppo di osservazioni scelto, svolgendo, quindi, una forte sintesi delle informazioni a disposizione, nel secondo questo non avviene, dal momento che tutte le osservazioni saranno sfruttate durante l'intero svolgimento della procedura.

Il metodo *cluster mass* si sviluppa a partire dalla definizione di un modello ANOVA per misure ripetute, che può essere caratterizzato da componenti fisse o casuali, e da variabili *di disturbo*, oltre a quelle di diretto interesse. Dato il disegno di studio con cui è stato raccolto il segnale studiato in questo elaborato, si farà riferimento a un semplice modello a effetti casuali con una sola variabile di interesse, cioè quella riferita alla condizione. Il modello ha, quindi, forma:

$$Y_{ij} = c + \gamma X_{ij} + u_{ij}$$
$$\text{con } u_{ij} = \mu_i + \epsilon_{ij}$$

con X_{ij} che è una variabile a due livelli riferita alla condizione e μ_i che è l'effetto casuale riferito al soggetto i-esimo. Le assunzioni su cui si basa il modello sono quelle già espresse al Paragrafo 3.4 per un modello a effetti misti. Il parametro di interesse è γ, che viene stimato attraverso la statistica F e su cui si vuole verificare il sistema di ipotesi:

$$\begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma \neq 0 \end{cases}$$

Come già anticipato, il calcolo della statistica F riguarda ogni singolo istante in cui è stato rilevato il segnale e proprio su tali statistiche si basa l'individuazione dei cluster. Infatti, un insieme di punti temporalmente adiacenti costituisce un cluster se tutti questi superano il valore soglia stabilito a priori, τ. Se non si hanno informazioni a favore della scelta di un particolare valore soglia, si può scegliere il 95° percentile della distribuzione nulla della statistica F. Una volta definiti i cluster, si può procedere con il calcolo di una misura sintetica che ne esprima la grandezza, cioè che tenga conto non solo del numero di osservazioni che ne fanno parte, ma anche della grandezza delle statistiche test F ottenute per ognuna di tali osservazioni. In altre parole, si calcola $m_i = f(C_i)$, dove $f(\cdot)$ è una funzione che può assumere qualsiasi forma e C_i è l'i-esimo cluster, con $i = 1, ..., n_c$, dove n_c è il numero totale di cluster ottenuti sulle n statistiche calcolate. Quando si lavora con le statistiche F, una scelta molto comune per la funzione $f(\cdot)$ è la semplice somma di tali statistiche per ogni punto incluso nel cluster.

Il primo passo, dunque, è quello di ricavare la distribuzione nulla delle statistiche F attraverso lo sviluppo di n_P permutazioni del campione originario. Questo consentirà l'individuazione del valore soglia τ attraverso cui, nel campione originario e per ogni permutazione P, verranno individuati i cluster. Dunque, una volta calcolate le misure di sintesi $m_i = f(C_i)$ per ognuno di questi cluster, sarà possibile costruire la distribuzione nulla M di tale misura, considerando per ogni permutazione il valore m_i più alto. Con riferimento alla distribuzione nulla ottenuta, si possono calcolare i seguenti p-value per ogni cluster individuato nei dati originari:

$$p_i = \frac{1}{n_P} \sum_{m* \in M} I(m* \geq m_i)$$

dove n_P è il numero di permutazioni svolte e m^* è la misura di grandezza del cluster ottenuta nel campione permutato.

Dunque, a tutti i punti che fanno parte del cluster C_i sarà attribuito lo stesso p-value, quindi se questo risulta significativo, vuol dire che in tutti i punti interni al cluster l'effetto è significativo.

Le principali proprietà del metodo *cluster mass* sono il fatto che la procedura di individuazione dei cluster e la loro distinzione in significativi e non è interamente basata sui dati empiricamente osservati, e il fatto che nella definizione di cluster vengano tenuti in considerazione contemporaneamente diversi aspetti dell'onda studiata, come la grandezza del segnale e la sua persistenza nel tempo: in questo modo è possibile evitare

che l'eventuale presenza di rumore o di artefatti che si estendono per pochi millisecondi influiscano in modo determinante sui risultati. Inoltre, la procedura garantisce un controllo del FWER che però, diversamente dai metodi sviluppati in precedenza, non riguarda le singole osservazioni ma direttamente i cluster individuati. Nonostante i risultati finali ne siano in qualche modo influenzati, il controllo è garantito per qualsiasi valore scelto per la soglia τ.

Infine, per alcuni aspetti il metodo *cluster mass* può essere considerato analogo alla procedura ARI, trattata al Capitolo 4. In entrambi i casi, infatti, si giunge all'individuazione di uno o più insiemi di dati in cui si può supporre che sia presente un qualche effetto. Tuttavia, mentre con ARI una volta stabilita la presenza di un effetto (TDP>0) è possibile arrivare ad una sua quantificazione attraverso la stima del TDP ed eventualmente restringere il cluster fino ad ottenere la stima del TDP desiderata, il *cluster mass* consente semplicemente di ammettere la significatività o meno dell'effetto sull'intero cluster, senza poter avanzare alcuna ipotesi riguardo la sua estensione o la sua localizzazione. Questo problema risulta tanto più grave quanto più estesi sono i cluster individuati, infatti anche in questo caso si può parlare di *paradosso della specificità spaziale*.

5.4.1 Applicazione ai dati

L'applicazione del metodo *cluster mass* ai dati introdotti al Capitolo 1 si è sviluppata, come già anticipato, attraverso l'impostazione di un modello ANOVA per misure ripetute con un solo effetto casuale e una variabile di interesse, riferita alla condizione a cui il soggetto viene sottoposto durante la registrazione del segnale, identificata con S1 o S2. Il valore soglia scelto è quello corrispondente al 95° percentile della distribuzione delle statistiche F, cioè $\tau = 5.12$ e la funzione di sintesi utilizzata è la semplice somma della statistiche F, cioè:

$$m_i = \sum_{j=1}^{n_i} F_j$$

dove F_j è la statistica F ottenuta all'osservazione j-esima del cluster e n_i è il numero totale di osservazioni contenute nel cluster i-esimo. Infine, per il calcolo del p-value sono state sviluppate 5000 permutazioni.

Sulle n=16384 osservazioni totali, sono stati individuati 252 cluster di varia grandezza, di cui soltanto due sono risultati significativi, con p-value pari a 0.0286 e 0.0002. Essi si riferiscono ai canali O2 e CZ, rispettivamente negli intervalli temporali compresi fra il 68° e il 168° e fra il 68° e il 256° millisecondo di rilevazione. Questo risultato è rappresentato nelle Figure 5.3 e 5.4, dove si riportano ancora una volta le differenze

medie di segnale fra le due condizioni per ogni istante di osservazione e per ogni canale e dove si evidenziano in blu i cluster risultati significativi col metodo del *cluster mass*.

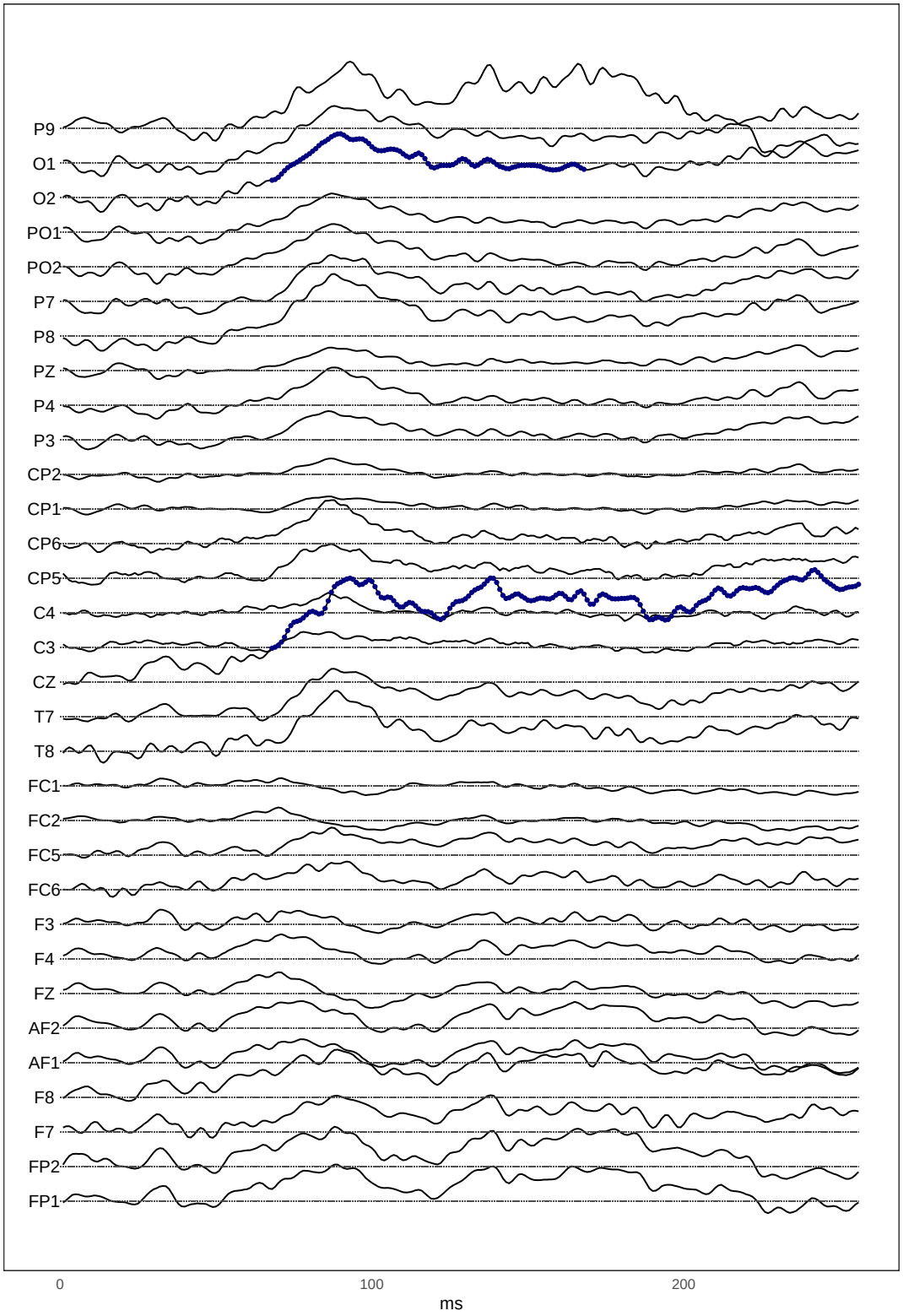

Figura 5.3: Cluster significativi col metodo *cluster mass*

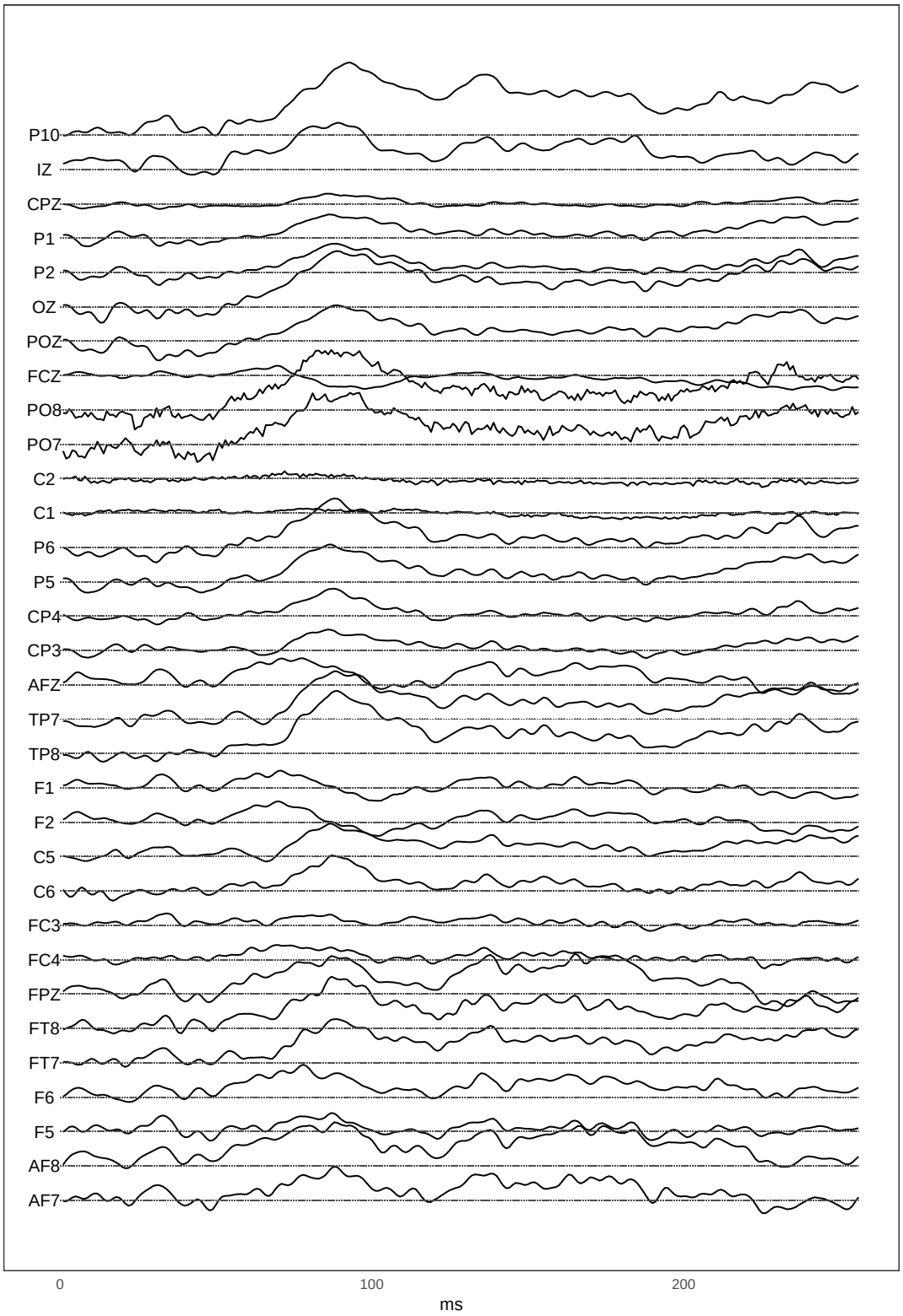

Figura 5.4: Cluster significativi col metodo *cluster mass*

Capitolo 6

Riduzione della molteplicità

Le procedure illustrate nei capitoli precedenti si concentrano sul controllo degli errori FWER, FDR e FDP che, come già detto, rappresentano una generalizzazione dell'errore di tipo I al caso di test multipli. Di seguito, invece, si approfondirà un altro concetto di fondamentale importanza in un contesto di verifica di ipotesi, ovvero la *potenza*. Si tratta della capacità della procedura di rigettare un'ipotesi che è effettivamente falsa. Essendo un'azione corretta, una qualsiasi procedura per lo svolgimento di test è tanto migliore quanto più alta è la potenza che la caratterizza.

Una caratteristica delle tecniche di correzione per test multipli è il fatto che più è basso il numero di test da svolgere, più è alta la potenza che le caratterizza. Per questo motivo, in questo Capitolo verranno nuovamente sviluppate le procedure viste finora su un numero totale di test minore, ottenuto riducendo il numero di livelli delle variabili. In questo modo, sarà possibile valutare se e come cambia il numero di rifiuti ottenuti. In particolare, ci si è concentrati sulla variabile che definisce il tempo di osservazione, infatti, considerando l'intero periodo di rilevazione, si è ridotto il numero totale di istanti temporali calcolando una semplice media aritmetica su osservazioni contigue. Questa operazione è stata svolta 3 volte, ogni volta svolgendo la media su un numero di osservazioni sempre maggiore in modo da ridurre sempre di più la molteplicità. In Tabella 6.1, per ogni riduzione, sono riportati il numero di osservazioni incluse nel calcolo della media, il numero di osservazioni che si è ottenuto per ogni canale, e il numero totale di osservazioni a disposizione, tenendo conto che il numero di canali di rilevazione è stato mantenuto sempre fisso a 64.

Il tentativo di guadagno in potenza riguarda i p-value grezzi e tutti i metodi di correzione per test multipli precedentemente sviluppati, cioè quelli di Bonferroni e Holm, la procedura ARI e quella del maxT di Westfall&Young, ma non riguarda i metodi standard perché, come si è visto al Capitolo 3, essi si concentrano su un particolare sotto-intervallo in cui si calcola l'ampiezza complessiva del segnale: questa è stata

Tabella 6.1: Medie calcolate e osservazioni totali

	n osservazioni per media	n osservazioni per canale	n osservazioni totali
m	1	256	16384
m2	2	128	8192
m4	4	64	4096
m8	8	32	2048

ottenuta come media aritmetica del dato originario nell'intervallo scelto, quindi ridurre gli istanti temporali tramite una media delle osservazioni non porterebbe a nessun cambiamento nel calcolo dell'ampiezza.

Dato che non si è in grado di stabilire quali siano gli effetti reali, e quindi di valutare in modo diretto la potenza, ci si limiterà ad osservare come cambia la proporzione degli "effetti" (o *scoperte*) ottenuti al variare della molteplicità. In particolare, si calcolano i rapporti fra numerosità che sono l'una il doppio dell'altra, quindi si confrontano rispettivamente: m con m2, m2 con m4 e m4 con m8. Un rapporto circa pari a 2 indica che il tasso di effetti ottenuti rimane lo stesso anche dopo la riduzione perché, in questo caso, la riduzione del numero di scoperte è esattamente proporzionale a quella del numero di test svolti.

Tabella 6.2: Rapporti fra il numero di scoperte complessive fra i diversi livelli di molteplicità

	Grezzi	Bonferroni	Holm	ARI	maxT
m/m2	2.00	0.33	0.33	1.83	1.50
m2/m4	1.99	1.50	1.50	1.71	1.50
m4/m8	1.92	1.00	1.00	1.47	0.8

Dalla Tabella 6.2 si osserva che sui p-value grezzi il numero di rifiuti si riduce in modo quasi esattamente proporzionale alla riduzione della molteplicità, per tutte le riduzioni effettuate. Ad eccezione della procedura ARI, i risultati non sono molto chiari perché per alcune riduzioni (da m a m2 per Bonferroni e Holm e da m4 a m8 per maxT), si ha un aumento del numero totale di rifiuti, mentre in altri casi si ha un effettivo aumento della proporzione del numero di rifiuti rispetto al totale di test svolti.

Per quanto riguarda la procedura ARI, invece, si può osservare una riduzione del numero di scoperte meno che proporzionale su tutte le riduzioni, quindi il tentativo di miglioramento del metodo sembra funzionare.

Analogamente, questi rapporti possono essere calcolati sui risultati ottenuti attraverso la procedura ARI sui singoli canali e su un sotto-intervallo temporale definito a priori.

Nel primo caso, come si è visto già al Capitolo 4, gli unici due canali con dei risultati rilevanti sono CZ e PO7, anche se in quest'ultimo, già dalla prima riduzione, il numero di *vere scoperte* è diventato nullo. Per questo motivo, in Tabella 6.3 si riportano solo i rapporti relativi al canale CZ. In esso, si osserva fin da subito un aumento del tesso di rifiuti che si mantiene pressoché costante, dato che i rapporti sono tutti <2 e simili fra loro.

Tabella 6.3: Rapporti fra le vere scoperte nel canale CZ

	CZ
m/m2	1.40
m2/m4	1.25
m4/m8	1.33

Per quanto riguarda, invece, l'analisi su un sotto-intervallo temporale, si è fatto riferimento all'area di osservazione compresa fra 70 e 110ms nei dati originari, $I_{(70-110)}$, quindi per m2, m4, e m8 sono stati considerati rispettivamente i seguenti intervalli: 35-55ms, 18-28ms, 9-14ms. I risultati sono riportati in Tabella 6.4.

Tabella 6.4: Rapporti fra le vere scoperte in un sottointervallo temporale

	$I_{(70-110)}$
m/m2	1.90
m2/m4	1.25
m4/m8	0.80

Anche in questo caso si osserva un aumento del tasso di rifiuti che, però, è sempre più forte al ridursi degli istanti temporali considerati.

Conclusione

Nel corso dell'elaborato sono stati trattati alcuni fra i metodi più comunemente utilizzati per l'analisi del segnale ERP. Come si è visto, la principale distinzione che può essere fatta fra questi riguarda il livello di precisione temporale che si riesce a raggiungere. In tal senso, infatti, si hanno da un lato i cosiddetti metodi "standard", che sono in grado di stabilire la presenza di un effetto soltanto in riferimento a un insieme di osservazioni temporalmente contigue, mentre dall'altro vi è l'approccio "punto per punto" che si riferisce al confronto fra le condizioni su ogni singola registrazione di segnale effettuata, quindi punta ad individuare in modo preciso i particolari istanti temporali in cui si è potuto osservare un effetto.

Le due tipologie di approccio si presentano fortemente diverse sotto tanti punti di vista come la base teorica su cui si sviluppano, le assunzioni richieste, la forma e la precisione dei risultati, gli strumenti di analisi attraverso cui si sviluppano. Al contrario, uno dei pochi elementi che li accomuna è il fatto che entrambi implicano lo svolgimento di un numero di test molto alto, da cui nasce l'esigenza di controllare la probabilità di commettere un errore di I tipo che, altrimenti, diventerebbe troppo alto. Questo problema riguarda principalmente l'approccio "punto per punto", per cui nel corso dell'elaborato si è provato ad adottare diverse procedure di controllo del FWER, come quella di Bonferroni, di Holm e il metodo maxT di Westfall&Young, mentre per i metodi standard ci si è limitati soltanto alla correzione di Holm e all'applicazione del metodo di maxT sulle ampiezze calcolate nell'intervallo selezionato.

Tuttavia, nessuno dei due approcci adottati risulta chiaramente migliore dell'altro, dato che entrambi presentano diverse qualità ma anche vari inconvenienti. Per quanto riguarda i metodi standard, infatti, essi fanno spesso riferimento ad assunzioni difficilmente verificabili empiricamente come, ad esempio, quella di normalità o quella di omoschedasticità, ma il suo problema principale risiede nel fatto che l'inferenza che si ottiene è fortemente influenzata da due scelte a priori che non possono essere modificate nel corso della procedura: quella dell'intervallo temporale su cui concentrare l'analisi e quella della misura di sintesi da adottare per tale intervallo. Queste criticità riguardano anche il metodo *cluster mass*, presentato al Capitolo 5, che nonostante i diversi vantaggi di cui si è parlato, non consente alcuna revisione a posteriori dei cluster individuati ed è fortemente condizionato dal valore soglia scelto.

Per quanto riguarda l'approccio "punto per punto", invece, il principale difetto riguarda la precisione eccessiva dei risultati a cui esso conduce. Infatti, in un contesto di analisi ERP risulta poco rilevante se la condizione genera risposte diverse soltanto in singoli istanti temporali, perché si è interessati ad individuare effetti leggermente più duraturi. Infatti, bisogna sempre tener conto del fatto che un effetto potrebbe risultare significativo in un singolo istante non perché sia reale, ma solo a causa della presenza di rumore nel segnale o di qualche artefatto. Questa eccessiva precisione, oltre a essere poco utile, contribuisce ad aumentare notevolmente il numero di test da costruire e quindi a ridurre la potenza delle procedure sviluppate e la qualità dell'inferenza ottenuta.

Il metodo che, fra gli altri, si adatta maggiormente alle esigenze di analisi del segnale ERP, è la All-Resolutions Inference (ARI). Essa, infatti, fin dall'inizio si sviluppa sull'intero gruppo di dati rilevati e soltanto dopo consente l'individuazione di uno o più sottogruppi di interesse su cui concentrare l'analisi. In questo modo, nessuna informazione viene trascurata e si evita il rischio che l'intera procedura venga influenzata da scelte a priori in un modo che è difficile da prevedere. Allo stesso tempo, però, si possono ottenere risultati con una precisione temporale più bassa di quanto sia possibile fare con i metodi che lavorano sulle singole osservazioni. Inoltre, la procedura ARI non punta a dare indicazioni precise su quali singole osservazioni hanno generato un segnale diverso fra le due condizioni, ma soltanto a dare un'idea di quanto sia esteso l'effetto della condizione in un particolare sottogruppo di dati, attraverso una stima del TDP. Come si è visto, tale stima è ottenuta per mezzo dei cosiddetti *test locali*, il cui controllo del FWER è garantito dalla procedura del *closed testing*. A differenza delle altre, questa non fa riferimento soltanto al p-value relativo al singolo test e al totale dei test sviluppati, ma valuta anche la significatività degli effetti congiunti, attraverso l'intersezione delle ipotesi. Inoltre, il metodo ARI è caratterizzato da una forte flessibilità poiché, senza compromettere il controllo del FWER, consente di rivedere a posteriori il sottoinsieme scelto e quindi, eventualmente, di aumentare o diminuire il livello di precisione dei risultati fino ad ottenere quello ritenuto ottimale.

Tuttavia, se da un lato ARI sembra raggiungere il giusto compromesso fra le varie esigenze di analisi del segnale ERP, dall'altro basa i propri risultati sul test di Simes che, come si è visto, risulta valido solo in caso di indipendenza o di correlazione non negativa fra le ipotesi. La condizione ideale, invece, sarebbe quella di poter adattare i test alla struttura di dipendenza empirica dei dati, come si riesce a fare attraverso un approccio di permutazione. Dunque, un possibile sviluppo futuro è quello introdurre nella procedura ARI l'utilizzo di test basati su permutazioni dei dati.

Appendice A

Le procedure per la correzione del FDR sono solitamente caratterizzate da una potenza molto più forte rispetto a quelle utilizzate per il controllo del FWER. In effetti, dai risultati in esame si evince una fortissima differenza nel numero totale di scoperte ottenute che, come si è visto, è pari a 1 per i metodi di Bonferroni e Holm, mentre per il metodo di Benjamini&Hoschberg risulta pari a 521, ovvero il 3.18% del totale delle differenze analizzate. In Tabella 6.5, sono riportati i risultati già visti al Capitolo 4 insieme a quelli ottenuti tramite la correzione del FDR. Si può osservare che, oltre ad essere in numero fortemente più alto, le scoperte ottenute col metodo di Benjamini&Hochberg sono anche molto più sparse fra i canali, infatti in 31 dei 64 canali di rilevazione si è potuto osservare almeno una differenza significativa fra le due condizioni.

Poiché questo risultato è stato ottenuto fissando il FDR al 5%, è possibile affermare che fra le 521 scoperte ottenute, in media il 5% di esse è una *falsa scoperta*. Inoltre, per i problemi già visti al Capitolo 4, il risultato ottenuto è valido esclusivamente per l'intero gruppo delle 16384 osservazioni, mentre se si volessero analizzare dei sottogruppi come, ad esempio, i singoli canali o un intervallo temporale specifico, sarebbe necessario sviluppare nuovamente l'intera procedura solo sulle osservazioni di interesse. Le Figure 6.1 e 6.2 riportano i grafici costruiti sulle differenze medie fra le condizioni S1 e S2, dove sono, però, poste in evidenza anche le scoperte ottenute col metodo di Benjamini&Hochberg.

Tabella 6.5: Sintesi dei risultati ottenuti con Benjamini&Hochberg

Canali	**Modello a effetti misti**	**Modello MANOVA**	**Holm**	**ARI**	**BH**
AF8	*	.	0	0	2
C3	.	.	0	0	4
C5	.	.	0	0	18
C6	.	.	0	0	5
CP3	.	.	0	0	6
CP5	.	.	0	0	21
CP6	.	.	0	0	4
CZ	*	*	0	7	140
F4	.	.	0	0	3
F6	.	.	0	0	4
F8	*	.	0	0	7
FT7	.	.	0	0	6
FT8	*	.	0	0	14
IZ	*	.	0	0	4
O1	*	*	0	0	27
O2	*	*	0	0	35
OZ	*	*	0	0	23
P10	*	*	0	0	24
P3	.	.	0	0	2
P5	.	.	0	0	6
P6	*	.	0	0	6
P7	*	.	0	0	5
P8	*	.	0	0	19
P9	*	.	0	0	0
PO1	.	.	0	0	2
PO2	.	.	0	0	2
PO7	*	.	1	1	11
PO8	*	.	0	0	20
T7	*	*	0	0	22
T8	*	*	0	0	25
TP7	*	*	0	0	22
TP8	*	.	0	0	32

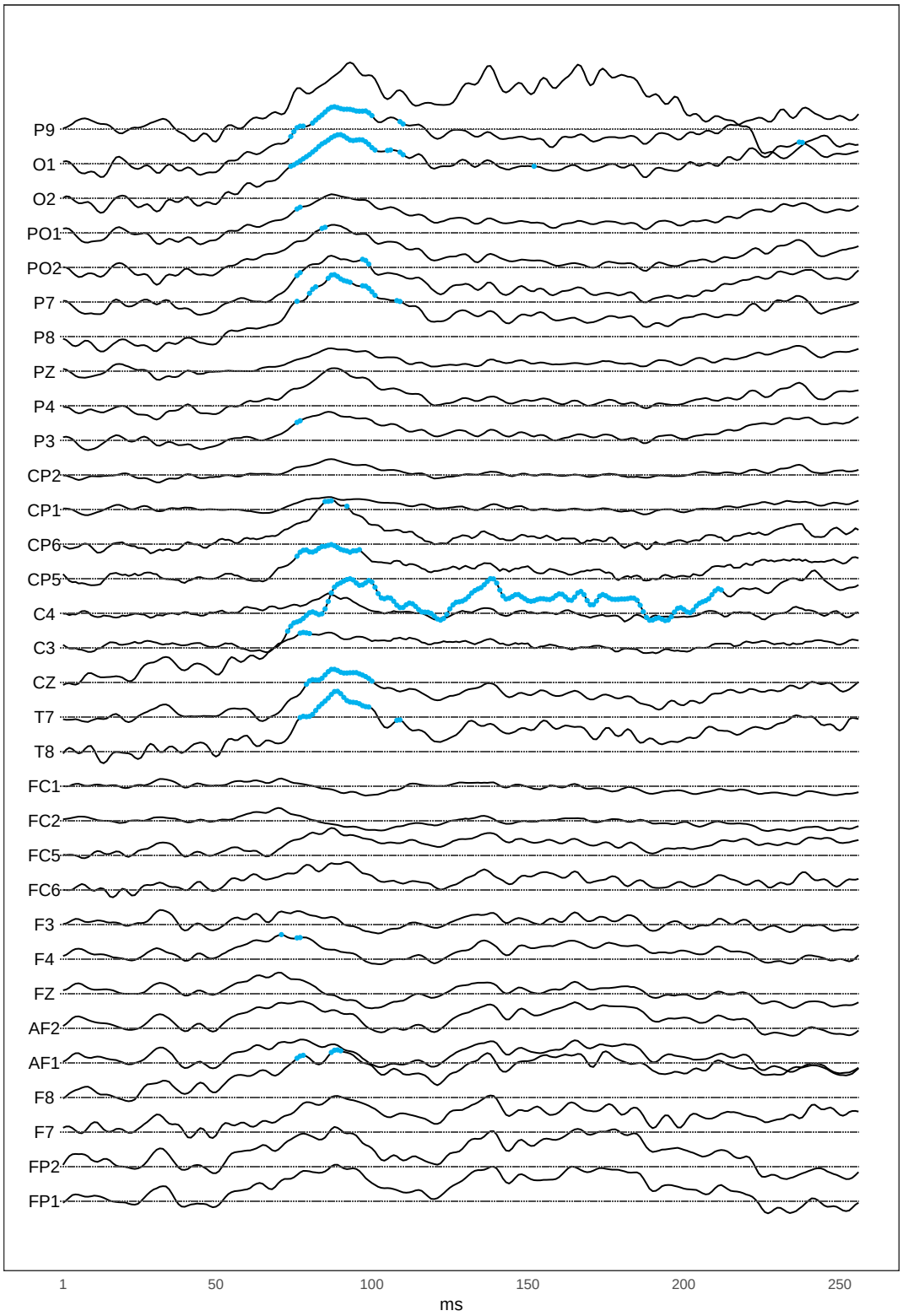

Figura 6.1: Scoperte ottenute tramite i metodi di Holm e Benjamini&Hochberg sui primi 32 canali

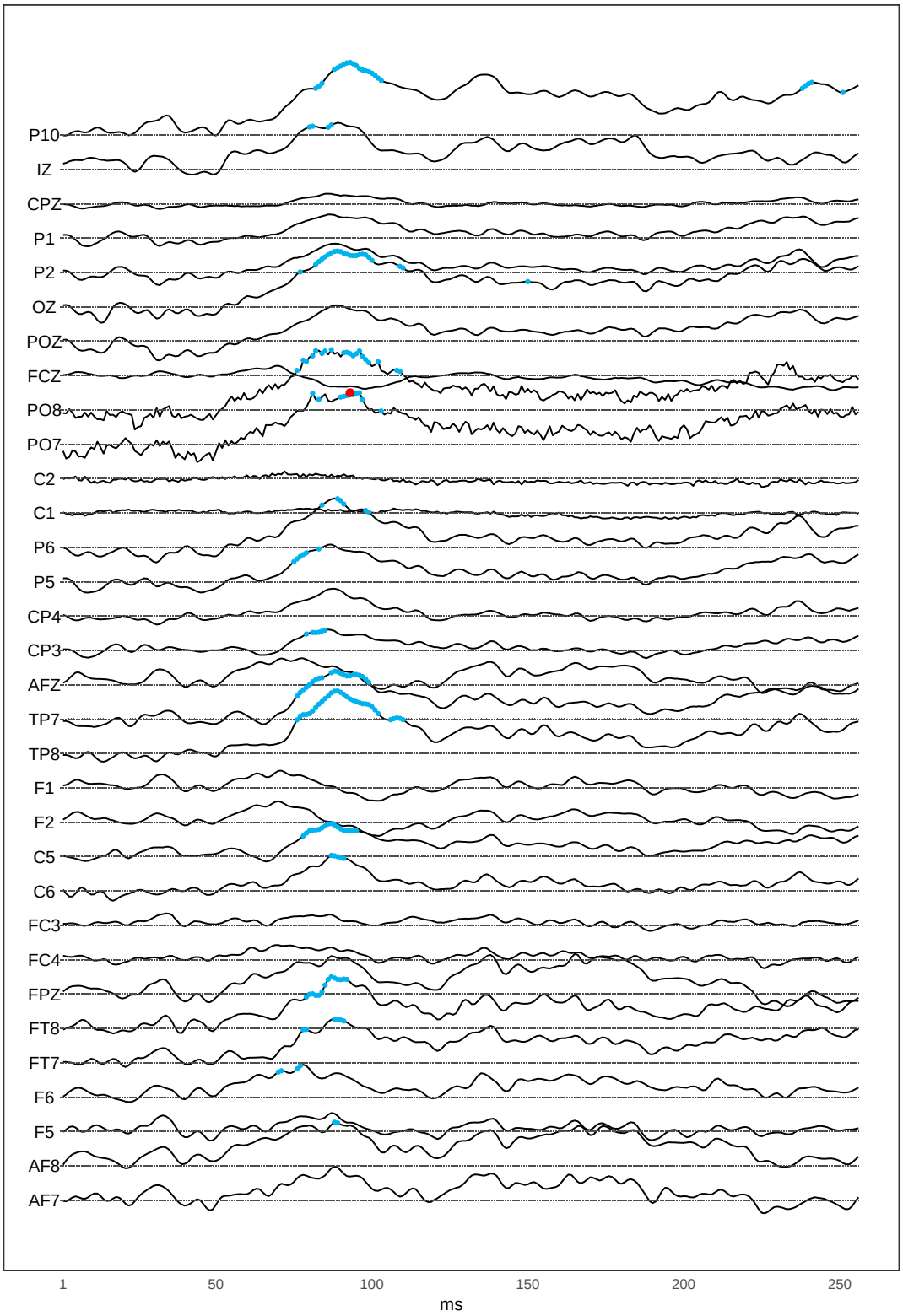

Figura 6.2: Scoperte ottenute tramite i metodi di Holm e Benjamini&Hochberg sugli ultimi 32 canali

Bibliografia

[1] Luck S. An Introduction to the Event-Related Potential Technique, 2014.

[2] Zhang X, Begleiter H, Porjesz B, Wang W, Litke A. Event related potentials during object recognition tasks *Brain Research Bulletin*, 1995; 38(6):531-538. **URL:** http://archive.ics.uci.edu/ml/datasets/EEG+Database

[3] Snodgrass J, Vanderwart M. A standardized set of 260 pictures: norms for the naming agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 1980; 6:174-215.

[4] Bates M. lme4: Mixed-effects modeling with R; 2010.

[5] Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, 2014; 67

[6] Hothorn T,Bretz F, Westfall P. Simultaneous Inference in General Parametric Models, 2019.

[7] Goeman J, Solari A. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 2014; 33: 1801-1980.

[8] Rosenblatt J, Finos L, Weeda W, Solari A, Goeman, J. All-resolutions Inference for brain imaging. *NeuroImage*, 2018; 181: 786-796.

[9] Goeman J, Solari A. Multiple Testing for Exploratory Research. *Statistical Science*, 2011; 26(4): 584-597.

[10] Goeman J, Mejier R, Krebs T, Solari A. Simultaneous Control of All False Discovery Proportions in Large-scale Multiple Hypothesis Testing, 2017. arXiv:11611.06739.

[11] Finner H, Roters M. On the False Discovery Rate and Expected Type I Errors. *Biometrical Journal*, 2001; 43(8): 985–1005.

[12] Goeman J, Meijer R, Krebs T. Methods for Closed Testing with Simes Inequality, in Particular Hommel's Method, 2018.

[13] Pesarin F, Salmaso L. Permutation tests for complex data, 2010.

[14] Winkler A, Ridgway G, Webster M, Smith S, Nichols T. Permutation inference for the general linear model. *NeuroImage*, 2014; 92(100): 381-397.

[15] Comerlati D. Conditional hypothesis testing with prior information, Università degli studi di Padova, Tesi di laurea, 2018/2019.

[16] Pollard K, Gilbert H, Ge Y, Taylor S, Dudoit S. Resampling-based multiple hypothesis testing, 2019.
URL: https://git.bioconductor.org/packages/multtest

[17] Renaud O, Frossard J. Permutation tests for regression, ANOVA and comparison of signals: the permuco package, 2019.

[18] Maris E, Oostenveld R. Nonparametric Statistical Testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 2007; 164(1): 177–190.