

Statistisk analys av fMRI-data

Jonas Hultin, Gustaf Broström, André Svensson, Svante Rosenlind
Handledare: Johan Lindström

Lärosäte: Lunds Universitet

8 januari 2020

Sammanfattning

I denna rapport används fMRI-data för att avgöra vilka delar av hjärnan som är aktiva då en försöksperson avgör om ord rimmar. Detta görs med hjälp av multipel linjär regression och ett antal statistiska tester. Resultatet visar på tre aktiva områden, ett i bakre delen av hjärnan, ett i främre delen av vänster hjärnhalva och ett i främre delen av hjärnan.

Innehåll

1	Inledning	2
1.1	Experiment	2
1.2	Bakgrund	2
1.2.1	Hjärnan	2
1.2.2	fMRI	2
1.3	Syfte	3
1.4	Data	3
1.5	Beräkningsmetoder	3
2	Teori	5
2.1	Multipel linjär regression	5
2.2	Hypotestest	6
2.3	Holm-Bonferronis metod	6
2.4	F-test	7
3	Metod	7
4	Resultat	9
5	Diskussion	11
5.1	Slutsats	11
5.2	Diskussion av metoder	12
5.3	Arbetsfördelning	15

1 Inledning

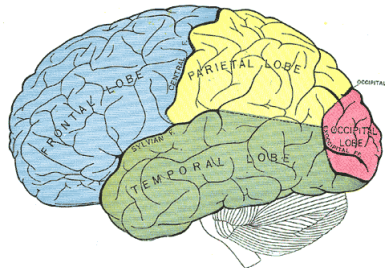
1.1 Experiment

I denna rapport analyseras data ifrån ett experiment där fMRI använts för att studera hur en människas hjärna reagerar på att presenteras två ord och sedan avgöra om dessa rimmar. Detta gjordes i ett intervall av 20 sekunders stimuli och sedan 20 sekunders vila över totalt 320 sekunder. För enkelhetens skull studeras endast ett tvådimensionellt segment av hjärnan strax ovanför personens tinning.

1.2 Bakgrund

1.2.1 Hjärnan

Den mänskliga hjärnan delas in i tre delar, storhjärnan, lillhjärnan och hjärnstammen. Storhjärnan är den största delen av hjärnan och behandlar bland annat syn, hörsel och språkförståelse. Storhjärnan delas själv in i två delar, höger och vänster hjärnhalva. Generellt gäller att höger hjärnhalva kontrollerar kreativitet, konstnärlighet och den rumsliga förståelsen, medan vänster hjärnhalva kontrollerar talförmåga, aritmetik och förmågan att skriva. Varje hjärnhalva kan ytterligare delas in i fyra lobber, pannloben, tinningloben, hjässloben och nackloben, se figur 1. Varje lob styr olika förmågor, exempelvis styr pannloben personlighet och talförmåga, tinningloben styr förståelse för språk, hörsel och minne, hjässloben tolkar språk, känsel och smärta, och nackloben tolkar synintryck (Hines 2018).



Figur 1: Storhjärnan indelad i dess fyra lobber sedd från sidan.

1.2.2 fMRI

Termen fMRI står för funktionell magnetresonanstomografi. Det är en icke invasiv metod för att mäta hjärnaktiviteten. Metoden går ut på att mäta förändringar i blodflöde som uppstår då aktiva delar av hjärnan kräver mer syre. Genom att utsätta testpersonen för olika stimuli kan sedan olika hjärnområden mappas till specifika funktioner (Mayfield 2018).

1.3 Syfte

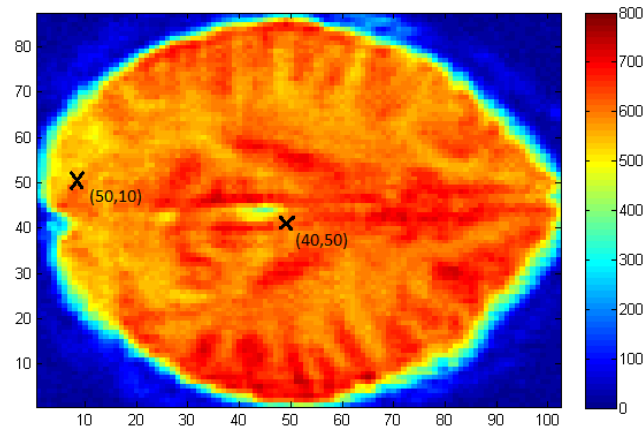
Syftet med arbetet är att använda fMRI-data för att avgöra vilka olika delar av hjärnan som aktiveras då en person avgör om ord rimmar. Under intervallen då testpersonen presenteras med stimuli undersöks hjärnans aktivitet. Hjärnan är alltid aktiv oavsett om den utsätts för stimuli eller ej, så det behövs metoder för att avgöra vilka områden som är mer aktiva under stimuli än under vila. Med ovanstående information om hjärnan kan ökad hjärnaktivitet förväntas finnas i nackloben, samt vänster hjärnhalva då man i denna undersökning tolkar ett ord genom synintryck och sedan med sin ordförståelse analyserar om det rimmar.

1.4 Data

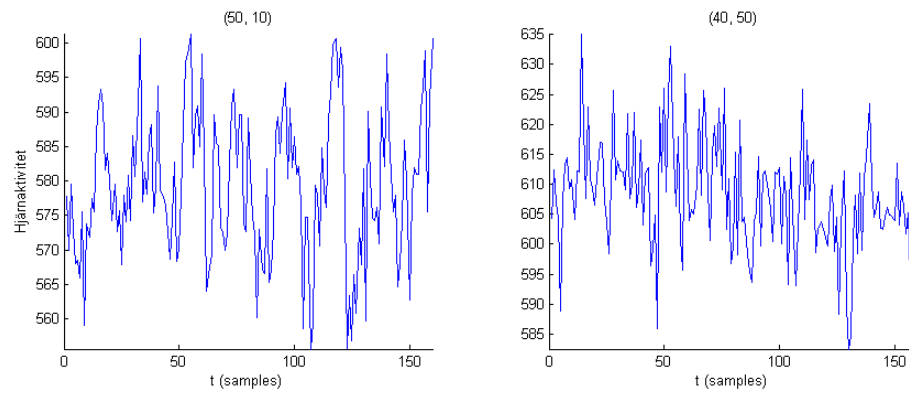
Datan som har använts är 2-dimensionell fMRI-data bestående av 160 bilder tagna med frekvensen 0.5 Hertz. Bilderna är alla tagna några millimeter ovanför tinningen på testpersonen. Datan hade redan förprocesserats, så att bland annat försökspersonens huvudrörelser i så stor mån som möjligt hade kompenserats för. Stimulin som presenteras för testpersonen är text som antingen rimmar eller inte. Stimulin representeras av en binär funktion, en fyrkantsvåg med känd period, presenterad i Figur 4. Starttidpunkten är inte given och behövs därför räknas ut, innan sedan studier kan göras angående vilka pixlar från fMRI:n som är aktiva i fas med fyrkantsvågen. Exempel på fMRI-data presenteras i figur 2 och i figur 3.

1.5 Beräkningsmetoder

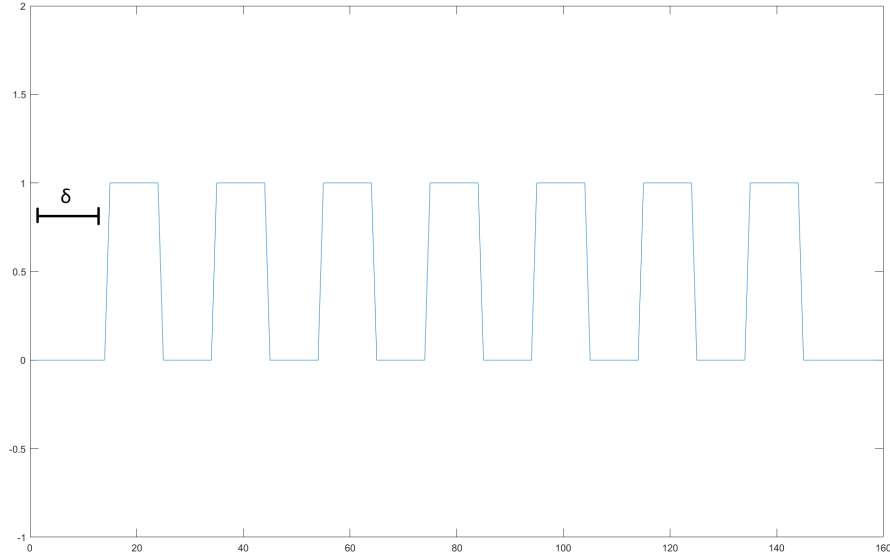
Med hjälp av multipel linjär regression kan man testa det statistiska sambandet att specifika områden i hjärnan är aktiva i samband med att personen utsätts för stimuli. Med denna metod undersöks varje pixel med hypoteser angående aktivitet, och eftersom dessa testas samtidigt riskeras en del fel att uppkomma, vilka behövs motverkas med hjälp av till exempel Holm-Bonferronis metod som förklaras i Teorin. För att sedan säkerställa resultatet används ett F-test som undersöker hur väl regressionen passar datan från fMRI:n.



Figur 2: Ett segment från fMRI av hjärna, där färgskalan visar aktivitet. Markerade punkter undersöks senare. Bilden är tagen uppifrån där bakre delen av hjärnan är till vänster.



Figur 3: Exempel på hjärnaktiviteten under experimentet i två punkter, (50, 10) och (40,50), som också är markerade i figur 2.



Figur 4: Regressionsmodellen som användes för stimulit. Här svarar δ mot tidsfördröjningen för den okända starttidpunkten och tiden det tar innan test-personen reagerar på stimuli.

2 Teori

2.1 Multipel linjär regression

Multipel linjär regression är en metod som används för att testa statistiska samband mellan en responsvariabel \hat{y} , och regressionsvariabler, x_1, x_2, \dots, x_n . Modellen ser ut som följande

$$\hat{y} = f(x_1, x_2, \dots, x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

där β_i är parametrar som beskriver den inverkan regressionsvariabeln x_i har på \hat{y} . Feltermen ϵ antas vara normalfördelad och oberoende med någon varians, $\epsilon \in \mathcal{N}(0, \sigma)$ (Blom 2005). Den optimala regressionen anses vara den som minimerar summan av kvadraterna på felen ϵ , vilket är optimalt i både minsta kvadrat- och maximum likelihood-mening. För att hitta de optimala β_i skrivs ekvationen ovan om på matrisform

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{n1} \\ 1 & x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (1)$$

Eftersom man med regressionsanalys vill hitta β_i som passar till många olika värden, kommer y , ϵ och alla x_i i ekvationen ovan att vara kolonnvektorer, med varje rad svarandes mot en datapunkt. Beteckna dem enligt följande:

$$\mathbf{y} = \mathbf{D}\beta + \epsilon \quad (2)$$

En metod för att nu beräkna det optimala värdet på vektorn β är att ta Moore-Penrose-pseudoinversen av matrisen \mathbf{D} och vänstermultiplicera med denna, vilket, eftersom $E(\epsilon) = 0$, ger

$$\beta = \mathbf{D}^+ \mathbf{y} = (\mathbf{D}^T \mathbf{D}^{-1})^{-1} \mathbf{D}^T \mathbf{y} \quad (3)$$

Variansen för detta skattade β ges sedan av

$$V[\beta] = \sigma^2 (\mathbf{D}^T \mathbf{D})^{-1}$$

Här ges σ^2 av följande uttryck, där n är antalet mätpunkter, och p är antalet parametrar i regressionen:

$$\sigma^2 = \frac{Q_0}{n - (p - 1)} = \frac{\mathbf{y}^T \mathbf{y} - (\mathbf{D}\beta)^T \mathbf{y}}{n - (p - 1)}$$

Även om vi bara ämnade beräkna variansen för de enskilda värdena på β_i , får vi nu kovariansmatrisen. Varianserna återfinns dock som diagonalelementen i denna.

2.2 Hypotestest

För att kunna dra en statistisk slutsats från någon data, formulerar man hypoteser, som sedan testas, och antingen förkastas eller inte. Om vi har formulerat någon hypotes H_0 , testar vi den genom att beräkna dess p -värde, vilket är sannolikheten att vi fick den data vi fick, givet att hypotesen stämmer.

Dessa hypotestester utförs på någon given signifikansnivå, vanligtvis 0.05, och man kan då förkasta alla hypoteser som producerar datan som fått eller något mer extremt med sannolikhet lägre än 5%, vilket är detsamma som att dess p -värde är mindre än 0.05.

Om man utför många sådana hypotestester, kommer man oundvikligen att råka ut för typ-1 fel, som syftar på detektion av signal som inte finns genom att förkasta H_0 . För att motverka detta kan man använda Holm-Bonferronis metod.

2.3 Holm-Bonferronis metod

Holm-Bonferronis metod används, som nämnts ovan, för att motverka typ-1 fel som riskerar att uppkomma då flera nollhypoteser testas samtidigt. Metoden

går ut på att lista p -värdena för varje nollhypotes i stigande storleksordning, p_1, p_2, \dots, p_m . För en given signifikansnivå α ska följande olikhet testas

$$p_k > \frac{\alpha}{m + 1 - k}. \quad (4)$$

Först testas olikheten för $k = 1$. Om olikheten inte uppfylls kan nollhypotesen för detta p -värde förkastas. Sedan testas $k = 2$ och samma jämförelse görs. Detta fortsätter för större värden på k tills olikheten uppfylls. Resterande nollhypoteser kan då inte förkastas.

En annan metod som kan användas är Bonferronis metod. Den förkastar nollhypoteser vars p -värde uppfyller $p_k \leq \alpha/m$, där m är antalet nollhypoteser. Bonferronis metod förkastar färre nollhypoteser än vad Holm-Bonferroni gör, det finns alltså kvar fler nollhypoteser som borde förkastas i Bonferronis metod. Därför bör Holm-Bonferroni alltid användas (Holm 1978).

2.4 F-test

För att undersöka huruvida en utförd regression passar datan väl kan ett F -test användas. Ett F -test innebär en jämförelse av summorna av residualerna i två separata fall: med regressionen man använt, och med en enklare version av regression med färre förklarande variabler.

F -testet går då ut på att man bildar följande kvot

$$F = \frac{RSS_1 - RSS_2}{p_2 - p_1} \bigg/ \frac{RSS_2}{n - p_2}$$

där n är antalet mätpunkter, p_1 och p_2 är antalet parametrar i de två fallen, och RSS_1 och RSS_2 summan av kvadratfelen i de två fallen. Eftersom den enklare regressionen är ett specialfall av den fullständiga, kommer ökningen av antalet parametrar alltid innebära en minskning av kvadratfelen, och F är därmed positiv.

Om vi då formulerar nollhypotesen att regressionerna är lika bra, kommer sannolikheten att vi fick det F -värde som vi fick eller högre ges av F -fördelningen, med $p_2 - p_1$ respektive $n - p_2$ som parametrar. Med en given signifikansnivå kan vi med därmed avgöra om den större modellen passar datan signifikant bättre, genom att jämföra F -värdet med signifikansnivån. Eftersom F -fördelningen inte är symmetrisk och speciellt bara är definierad för positiva värden, måste testet göras ensidigt (Rawlings, Pantula och Dickey 1998).

3 Metod

Till att börja med identifieras tidsfördröjningen δ . Detta görs först med hjälp av att ta ett medelvärde av hjärnans aktivitet då vi tror att stimulit ska vara

presenterat och sedan förflytta stimulit. Tanken är att då medelvärdet är som störst, är tidsfördröjningen innan första stimuli ges. Detta går också att göra genom att utföra multipel linjär regression med olika δ och skatta δ till det värde som minimerar felet ϵ .

För varje pixel i bilden utförs multipel linjär regression där 10 olika regressionsfunktioner används. Den första regressionsfunktionen bestod en konstant term, den andra svarade mot ett linjärt tidsberoende. De övriga 8 motsvarade stimulen och bestod av rektangelpulser med längd av 10 sampels. De placerades vid de tidpunkter som orden visades för försökspersonen, men var förskjutna med δ . Regressionen såg alltså ut på följande sätt:

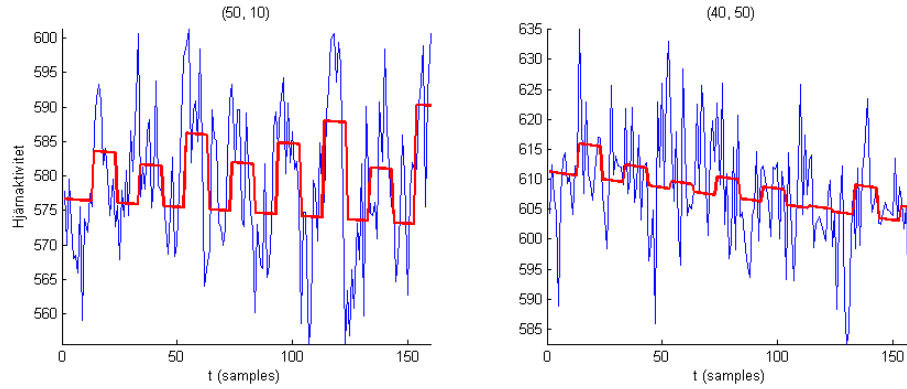
$$\hat{y} = \beta_0 + \beta_1 t + \sum_{n=1}^8 (\gamma_n f_n(t - \delta)) + \epsilon \quad (5)$$

där β_0, β_1 och γ_i var regressionsparametrarna som skulle bestämmas. Här står funktionen $f_n(t)$ för en rektangelpuls med längd 10 svarandes mot det n :te stimulit. Mer tekniskt utfördes regressionen på följande sätt:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix} + \beta_1 \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ \vdots \end{pmatrix} + \gamma_1 \underbrace{\begin{pmatrix} \vdots \\ 0 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ 0 \\ \vdots \end{pmatrix}}_{10}^{\delta} + \gamma_2 \underbrace{\begin{pmatrix} \vdots \\ 0 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ 0 \\ \vdots \end{pmatrix}}_{10}^{\delta + 20} + \cdots + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \end{pmatrix} \quad (6)$$

Här antogs ϵ_i , felen, vara normalfördelade kring 0. Genom att sedan samla regressionsfunktionerna i en matris, och parametrarna i en kolonnvektor kunde den optimala lösningen hittas genom att använda Moore-Penrose-pseudoinversen enligt följande:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & \cdots \\ 1 & 2 & 0 & \cdots \\ 1 & 3 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}}_{=\mathbf{D}} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \gamma_1 \\ \gamma_2 \\ \vdots \end{pmatrix} \Rightarrow \beta = \mathbf{D}^+ \mathbf{y}$$



Figur 5: Två exempel på den utförda regressionen, i punkterna (50, 10) och (40, 50). Punkten är markerad i figur 2, och datan syns också utan regression i figur 3.

Därefter kunde variansen för dessa skattade parametrar beräknas enligt metoden som finns beskriven i teoriavsnittet, 2.1. Från detta formuleras sedan följande hypoteser för varje pixel:

$$H_0 : \text{Alla } \gamma = 0, \quad H_1 : \text{Åtminstone ett } \gamma \neq 0$$

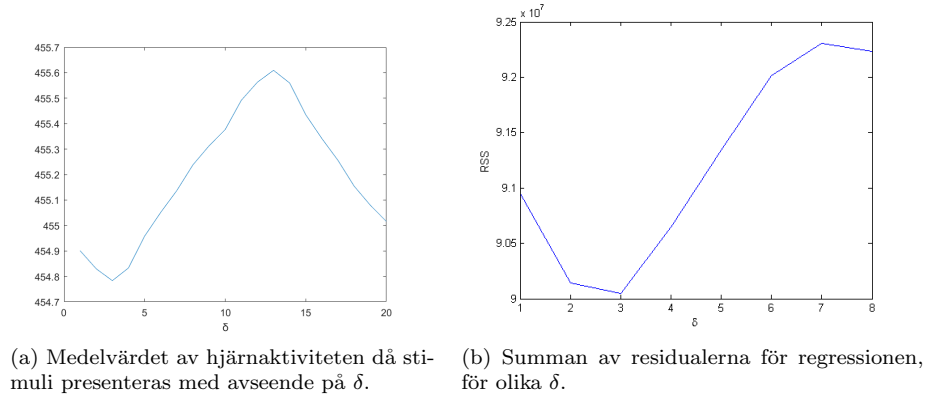
För att avgöra hur väl vår modell passar datan utför vi ett F -test, och jämför då vår fullständiga regression mot en som bara är affin, vilket skulle svara mot $\gamma_i = 0$ för alla i , det vill säga modellen under H_0 . Vi beräknar regressionskoefficienterna för det enklare fallet precis som ovan. Sedan summerar vi kvadraterna av alla fel, för båda modellerna, och skapar F -värdet enligt följande:

$$F = \frac{RSS_1 - RSS_2}{p_2 - p_1} \bigg/ \frac{RSS_2}{n - p_2}$$

Mer specifikt har vi $p_1 = 2$, $p_2 = 10$, och $n = 160$. Detta test utförs sedan individuellt för varje pixel.

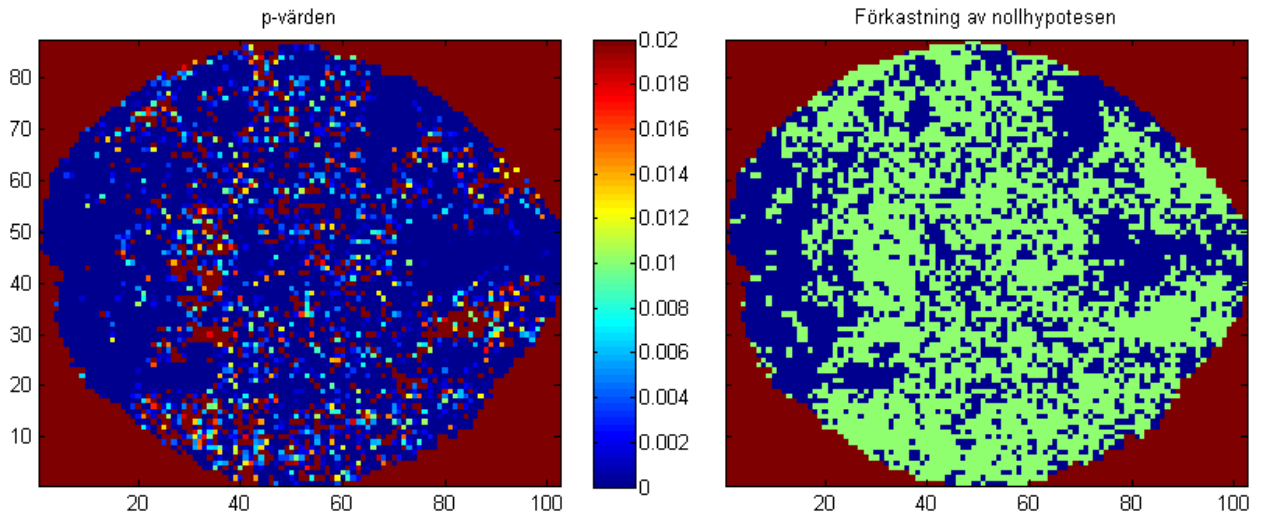
4 Resultat

Då vi skulle skatta δ är detta resultatet vi fick.



Figur 6: Skattning av δ utförd med två olika metoder.

Här ser vi att tidsfördröjningen $\delta = 13$ för både (a) och (b) då stimulen för regressionen börjar med 10 sampels av vila.

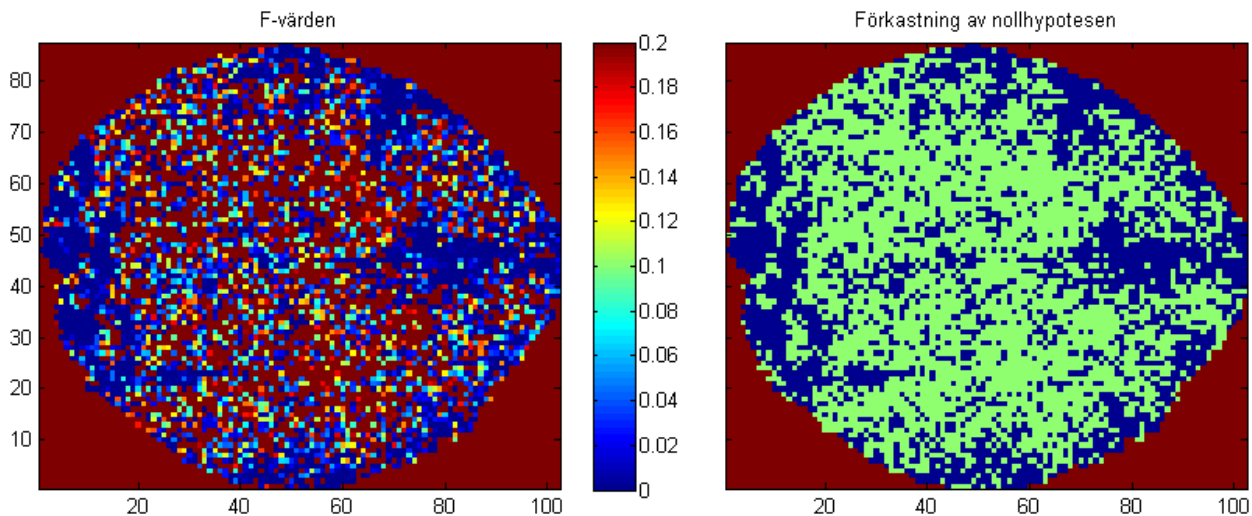


Figur 7: I den vänstra bilden visas p -värdena för nollhypoteserna i varje pixel, med mörkblå svarandes mot 0, och mörkröd svarandes mot p -värden över 0.02. I den högra visas de pixlar där nollhypotesen kunde förkastas i mörkblått, och övriga i grönt. Holm-Bonferronis metod användes för att lösa problemet med multipla hypoteser.

I figur 7 syns resultatet från den statistiska undersökningen för alla pixlar i hjärnan. Det syns tydligt att det finns tre större distinkta områden, där

nollhypotesen kunde förkastas för nästan alla pixlar, och där p -värdena också är förhållandevis låga.

I figur 8 syns i stället utfallet av F -testet. Man ser att modellen reducerar felen i stor utsträckning i de områden som innan hade låga p -värden, och att F -värdena blir större i övriga delar av hjärnan. Detta är att vänta sig, eftersom felen i ett område som inte svarar på stimuli huvudsakligen beror på mätbrus. Att låta regressionen inbegripa aktivitet relaterat till stimuli kommer då inte att minska felen avsevärt. I ett område som svarar på stimuli, kommer den linjära regressionen inte kunna representera denna mer komplicerade karaktär, och tillökningen av rektangelpulserna kommer då att minska felen avsevärt.



Figur 8: F -värdena för varje pixel visas i den vänstra bilden, och i den högra visas de som är under signifikansnivån 0.05 i blått, och de där F -värdet är över 0.05 i grönt. Med nollhypotesen menas här hypotesen att den större modellen inte ger någon förbättring.

5 Diskussion

5.1 Slutsats

I figur 7 så ser vi tydligt tre större områden där både p -värdena är relativt låga samtidigt som nollhypotesen kunde förkastas. Detta innebär, säkerställt genom F -testet, att dessa är de områden i hjärnan som aktiveras när personen ser och analyserar orden som denne har presenterats. Områdena befinner sig som vi antog både i bakhuvudet och i vänster hjärnhalva. Dessutom syns en klar

aktivitet även i den främre centrala delen av hjärnan, vilket vi inte förväntade oss från den information vi samlat om hjärnans funktioner.

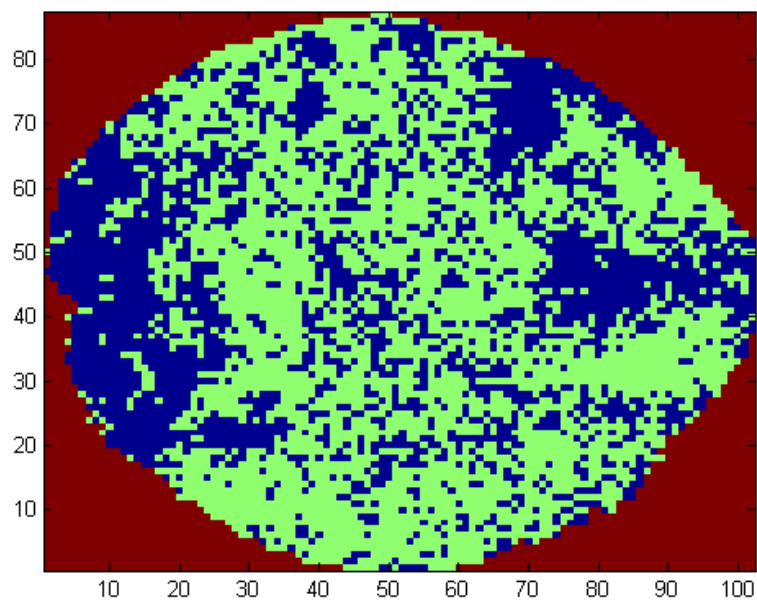
5.2 Diskussion av metoder

Vi använde Holm-Bonferronis metod för att kompensera för att vi samtidigt försökte testa många hypoteser. För att förklara dess behov tänker vi oss först ett enkelt experiment, där vi vill förkasta någon nollhypotes, som vi vet faktiskt stämmer. Om vi bara upprepar experimentet tillräckligt många gånger kommer vi förr eller senare lyckas förkasta den på någon signifikansnivå, trots att den stämmer. Holm-Bonferronis metod kompenserar då för denna effekt, då många hypoteser testas.

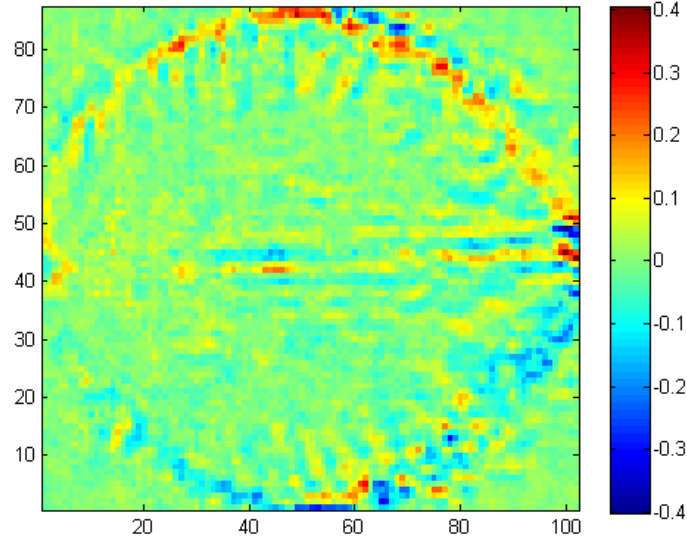
En enklare metod för att lösa problemet med multipel hypotestestning är att använda Bonferronis metod. Den går ut på att man delar signifikansnivån med antalet hypoteser man testat, och jämför p -värdena mot denna nya signifikansnivå. I vårt fall blir alla pixlar med p -värde under $0.05/8874 = 5.6 \cdot 10^{-6}$. Detta kan ses i figur 9.

Holm-Bonferronis metod är strikt bättre än Bonferronis om fler än en hypotes testas, vilket lätt kan ses om man jämför ekvation 4 med metoden i stycket ovan. I Holm-Bonferronis metod kommer det lägsta p -värdet att testas mot samma nivå som i Bonferronis metod, men alla efterföljande kommer möta lättare krav för signifikans.

I figur 10 syns hur termen β_1 , som svarar mot en linjär förändring i tiden, varierar i hjärnan. Den är statistiskt signifikant huvudsakligen kring den övre kanten av hjärnan och också vid punkten precis i mitten. I figur 2 syns att den uppmätta aktiviteten i denna punkt, och även utanför hjärnan, är lägre. Därför skulle de negativa värdena på β_1 i underkanten kunna förklaras med att hjärnan till en början ockuperar dessa pixlar, men sedan långsamt flyttar sig uppåt under experimentet. På samma sätt ser vi att i mitten av hjärnan har punkten med låg aktivitet flyttat sig uppåt, vilket resulterar i en positiv linjär trend i pixlarna den lämnade, och en negativ i de som den nådde.



Figur 9: De pixlar där nollhypotesen kunde förkastas, med Bonferronis metod.

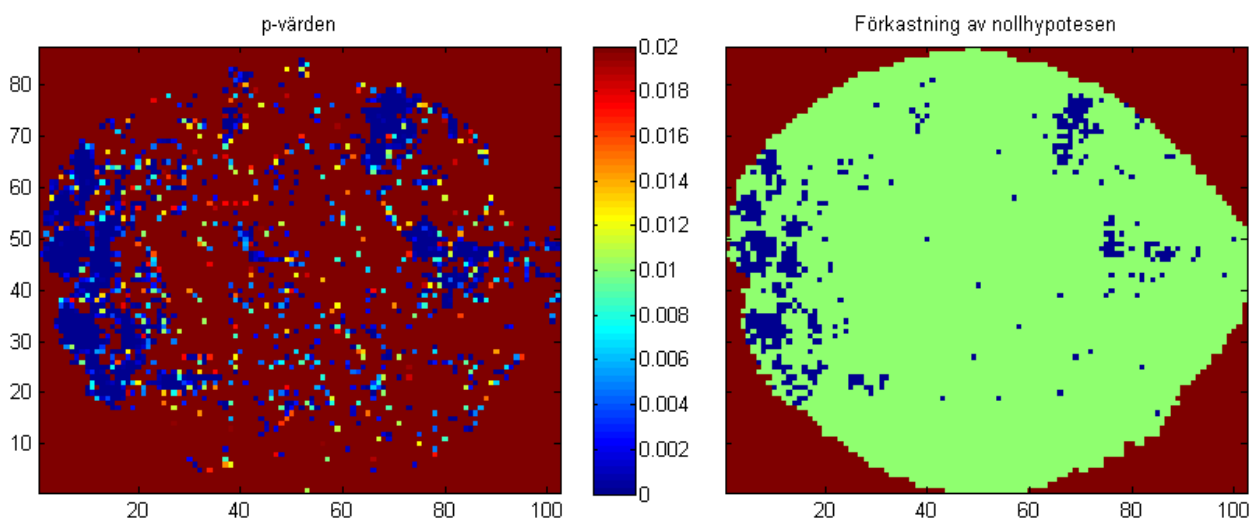


Figur 10: Parametern β_1 , som svarar mot en linjär förändring av hjärnaktivitet.

I vår modell skapade vi en separat parameter för varje rektangelpuls, men en enklare modell hade också kunnat vara att gå till väga på samma sätt, fast med alla pulser knutna till en parameter. Om vi jämför med ekvation 5 får vi i stället följande ekvation

$$y = \beta_0 + \beta_1 t + \gamma \sum_{n=1}^8 (f_n(t - \delta)) + \epsilon \quad (7)$$

Vi kan bilda grafer motsvarande de i figur 7 för den nya regressionen också, med nollhypotesen $H_0 : \gamma = 0$. De syns i figur 11. Om vi kontrasterar de två bilderna, och fokuserar på förkastningen av nollhypoteserna, kan vi konstatera att med den enklare modellen var det färre pixlar där vi kunde förkasta nollhypotesen. Detta stämde generellt, både för de områden som svarade på stimuli och de som inte gjorde det, så ingen modell var klart bättre än någon annan.



Figur 11: I den vänstra bilden visas p -värdena för nollhypoteserna i varje pixel, med mörkblå svarandes mot 0, och mörkröd svarandes mot p -värden över 0.02 alternativt pixlar utanför hjärnan. I den högra visas de pixlar där nollhypotesen kunde förkastas i mörkblått, och övriga i grönt.

5.3 Arbetsfördelning

Under projektarbetets gång har vi alla suttit tillsammans i en datorsal och arbetat via Overleaf, under ett tiotal eftermiddagar. Under arbetets gång har vi tillsammans kodat i Matlab samtidigt som vi undersökt matematiska modeller så som linjär regression och statistiska tester. Vissa tester var svårförstådda och svåra att tillämpade i kod, så vi delade upp arbetet genom att studera olika tester och försökte applicera dessa på koden. Svante hade bäst koll på MATLAB, vilket innebar att han stod för majoriteten av appliceringsarbetet, eftersom han oftast var först med att hitta lösningen vi sökte. När sådana stunder nåddes påbörjade vi andra tre att skriva rapporten, samla information om de olika metoderna och framtida testerna samt hitta information om hjärnans funktioner.

Vår handledares roll i projektet var att presentera metoder och tester som vi skulle kunna använda för att få fram ett tydligt och säkert resultat. Om vi stötte på problem med att förstå något av dessa kunde han även förklara dem tydligare när vi bad om det. När vi hade problem med att få struktur och tydlighet i rapporten gav han oss även tips angående detta.

Referenser

- Blom, Gunnar (2005). *Sannolikhets teori och statistik teori med tillämpningar*. ISBN: 9789144024424.
- Hines, Tonya (2018). *Anatomy of the Brain*. URL: <https://mayfieldclinic.com/pe-anatbrain.htm> (hämtad 2019-02-06).
- Holm, Sture (1978). *A Simple Sequentially Rejective Multiple Test Procedure*. URL: https://www.jstor.org/stable/4615733?seq=1#metadata_info_tab_contents (hämtad 2019-02-13).
- Mayfield (2018). *Brain mapping: functional MRI and DTI*. URL: https://mayfieldclinic.com/PE-fmri_DTI.htm (hämtad 2019-02-08).
- Rawlings, John O., Sastry G. Pantula och David A. Dickey (1998). *Applied Regression Analysis - A Research Tool*. 2. utg. Springer, s. 120–122.