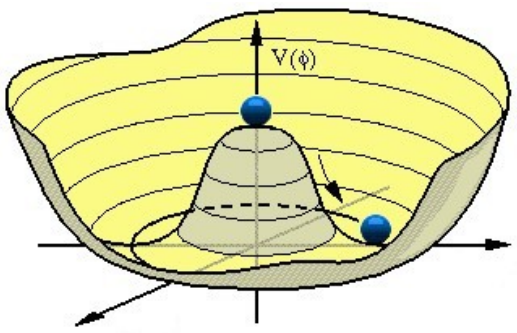
Higgs Hunting

Separation af Simulerede Data i Søgen efter Higgs-bosonen

Jerôme Baltzersen, Morten Hornbech, Mona Kildetoft og Kim Petersen



Førsteårsprojekt i fysik ved Niels Bohr Instituttet i København.

6. februar - 24. marts, 2006

Forord

Denne opgave er skrevet af 1. års studerende på fysik med en matematisk baggrund til andre 1. års stunderende på fysik, og det antages derfor kun, at læseren har kendskab til fysik på det niveau, som er opnået efter tre blokke på fysik. Af og til benyttes statistiske metoder, som fysikstuderende på 1. år ikke stifter bekendtskab med, men disse er forsøgt udførligt beskrevet, og man vil i referencerne kunne finde hjælp til videre forståelse. Vi har valgt at skrive om Higgs-bosonen, da det er en boson, som der stadig arbejdes på at opdage, og dermed giver projektet en realistisk opfattelse af forskning i fysik.

Vi valgte at indsnævre projektet til at finde Higgs-bosonen med størst mulig signifikans, da vi mener, at det er her vores matematiske baggrund kommer mest til sin ret. Desuden ville projektet uden indsnævring simpelthen blive for omfattende.

I arbejdet med en hvilken som helst større opgave er det naturligt, at der opstår småproblemer løbende. Især i projekter, hvori programmering udgør en væsentlig del af arbejdet, vil der ofte opstå fejl i ens kode. Idet vi har fået udleveret Fortran- og PAW-koden, manglede vi af og til den fulde forståelse for deres virkemåde, hvilket langsommeliggjorde fejlretningen af disse. Der er ingen tvivl om, at projektet kræver forståelse for programmering og/eller en vis potion stædighed. I vores arbejde med projektet er vi naturligvis også stødt på forståelsesmæssige problemer. Både disse og fejlene i koden vil dog være helt individuelle og til en vis grad sikkert også tilfældige fra gruppe til gruppe.

Vi ønsker ydermere at takke Troels C. Petersen for tålmodig og kompetent vejledning igennem hele projektet. Med sin store entusiasme har Troels en stor del af æren for, at projektet har været så spændende at arbejde med, og der er ingen tvivl om, at hans begejstring for fysikken har haft en afsmittende effekt på os alle.

Figur 1: Troels C. Petersen; hhv. før og efter endt vejledning.

Indhold

1. Indledning

Vi arbejder i dette projekt med simulerede data bestående af resultater fra en række sammenstød mellem to protonstråler. Simulationen er foretaget under antagelse af, at den såkaldte Higgs-boson eksisterer, og at den udelukkende henfalder til e^+/e^- par. Det er nu vores opgave at nå frem til, hvordan vi udfra vores datamateriale kan observere denne partikel, og hvor statistisk signifikant observationen kan gøres. Hertil benyttes en række forskellige metoder, både meget simple og mere komplicerede, og vi vil i rapporten vurdere resultatet opnået ved hver enkelt metode samt diskutere metodens begrænsninger samt fordele og ulemper. Udover de rå data har vi en program-skabelon i Fortran [2] til rådighed, hvor vi forholdsvis let kan tilføje ny og ændre i den oprindelige kode til brug for vores dataanalyse. Desuden har vi et sæt af makroer til PAW [1] som plotter og fitter vores data på forskellig vis, og som vi selv kan gå ind og ændre i. Dette er vores udgangspunkt for alt arbejde med dataene. De resterende dele af rapporten bygger på vores egne ideer og påfund, naturligvis under løbende inspiration af vores vejleder Troels Petersen [3].

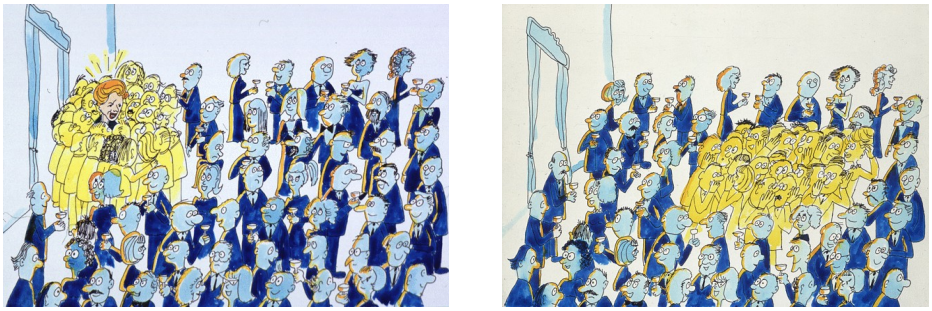
2. Higgs-bosonen

Higgs-bosonen er en partikel som, hvis den eksisterer, vil være ansvarlig for den mekanisme, der giver partikler masse. Standardmodellen giver ingen forklaring på, hvorfor partikler har den masse de har, eller hvordan de har fået den, da denne af symmetrigrunde forudsiger, at de skal være masseløse. Peter Higgs har opstillet en model [4], [5], i hvilken masse opstår som en vekselvirkning. Forestil dig, at hele rummet er dækket af et felt, som vi kalder for Higgsfeltet. Ideen er nu, at det vi fortolker som en partikels masse i virkeligheden er dens tendens til at vekselvirke med Higgsfeltet. Partiklen der formidler denne vekselvirkning er så Higgs-bosonen. Denne teori kan måske virke ubegribelig, men det er ikke desto mindre den, som nutidens fysikere regner for den mest sandsynlige.

Ved at analysere de ligninger, som vekselvirkningen måtte opfylde, fandt Peter Higgs frem til, at det laveste energistadium ved vekselvirkningen ikke lå ved en feltstyrke på nul, og resultatet heraf er, at alle partikler som vekselvirker med Higgs-bosonen kan tillægges en masse, som kan fortolkes som styrken af deres vekselvirkning med Higgs-bosonen. Man kan forestille sig Higgsfeltets potentiale som en krone (se billede på forsiden) og en partikel, der vekselvirker med feltet vil søge mod den laveste energitilstand. Den vil altså „trille ned til en af siderne“. Symmetrien i potentialet er netop en af grundene til, at mange fysikere bifalder Higgsmekanismen, da denne bevarer den såkaldte gaugesymmetri i standardmodellen. Gaugesymmetrien går ud på, at alle de fysiske love skal være invariante under lineære transformationer i rummet – altså drejninger og flytninger. Andre forsøg på at få masse ind i standardmodellen har brudt denne symmetri, men med Higgsmekanismen er der ingen asymmetri i modellen. Symmetribruddet sker først ved selve vekselvirkningen, altså når partiklen „triller ned“. Gaugesymmetrien har tidligere vist sig som en yderst frugtbar tankegang, og ønskes derfor bevaret i standardmodellens forklaring af masse.

Lad os prøve at give en lidt mere intuitiv forklaring. Forestil dig, at du er til en fest, hvor menneskemængden er jævnt fordelt i rummet. Nu træder en kendt person ind i rummet, og de nærmeste personer samler sig om hende (se figur 2, venstre). Som hun bevæger sig gennem lokalet vil hun tiltrække de mennesker, der er tættest på hende, mens dem hun bevæger sig væk fra vil vende tilbage til deres tidligere gøremål og position. Pga. gruppen af mennesker omkring hende, har hun „øget sin masse“. Hun vil altså være sværere at påvirke i sin bevægelse og sværere at sætte i gang efter at hun er stoppet, og dette er jo netop definitionen på træg masse. Generaliseret til tre dimensioner og med relativistiske komplikationer er dette Higgsmekanismen. For at give partikler masse, bliver et Higgsfelt lokalt forstyrret, når en partikel bevæger sig igennem.

Det kan undre, hvorfor der overhoved er behov for en beskrivelse af Higgsfeltet, men uden denne vil det ikke være muligt at forklare, hvorfor Z^0- og $W^\pm$-partikler, der er bærere af den svage kernekraft, er forholdsvis tunge, mens fotoner er masseløse. Men hvad med Higgs-bosonen selv? Hvordan får den masse? I rummet fra før passerer nu et rygte, og dem nærmest døren hører det først og samles, hvorefter de henvender sig til de andre gæster, der er nærmest. Således går rygtet som en bølge af samlinger igennem rummet (figur 2, højre), og da disse samlinger som før er et udtryk for masse, har rygtet altså også en vis masse. Higgs-bosonen er forudsagt til at være et sådant rygte i Higgsfeltet.



Figur 2: Illustration af Higgsmekanismen

3. ATLAS – Detektoren

A ***T***oroidal ***L***HC ***A***pparatu***S*** kort kaldet ATLAS er en af de partikeldetektorer, der skal bruges ved LHC (Large Hadron Collider – den nye accelerator ved CERN). Mere information om ATLAS findes under [6]. Partikelacceleratorer er bygget i cylindriske lag omkring sammenstødspunktet (engelsk: interaction point), og ATLAS har inderst en tracker, dernæst et elektromagnetisk kalorimeter, et muonspektrometer og et magnetsystem, hvilket ses på figur 3. Da vi i vores opgave kun beskæftiger os med elektroner og jets[1], vil vi ikke beskrive muonspektrometeret her. Det er nødvendigt at bygge en større accelerator i form af LHC for at kunne opnå højere energi i partikelsammenstød, og den vil kunne nå op på 14 TeV. ATLAS er bygget, så den ikke fokuserer på én bestemt fysisk proces, men på at indsamle så stor en mængde data som muligt, da dette øger antallet af mulige eksperimenter. Modsat tidligere acceleratorer benytter LHC sig af to protonstråler. Vi ønsker sammenstød af kvarker og anti-kvarker. Umiddelbart virker det mystisk, at en proton skulle indeholde en antikvark, men indeni en proton kan der opstå kvark-antikvark par som en følge af den kvantemekaniske tunneleffekt, der muliggør at partikler kan låne energi kortvarigt. Ved lave energier eksisterer disse kvark-antikvark par ikke længe nok til, at de kan blive ramt ved sammenstødene, men LHC kører med så høj energi, at dette bliver muligt. Det ville naturligvis være bedre at bruge en proton- og en antiprotonstråle, men det er dyrt og teknisk svært at producere en antiprotonstråle. Desuden vil en sådan være sværere at fokusere, og derfor vil man få færre sammenstød - også kaldet events.

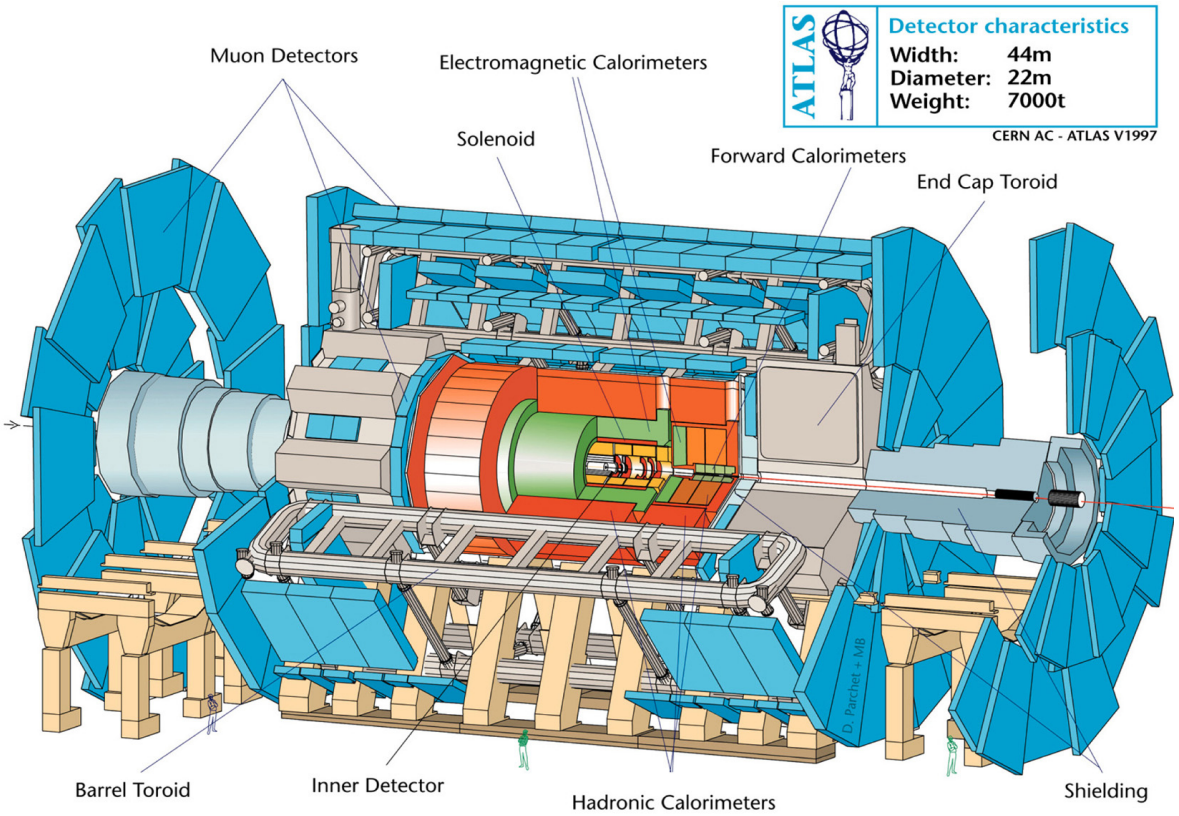
3.1. Den Indre Detektor / Trackeren

Det er trackerens opgave at fastlægge banerne for de partikler, der dannes ved kollisionen. På grund af det magnetfelt, som trackeren er anbragt i vil banerne for ladede partikler krumme, og krumningsgraden og -retningen kan benyttes til at bestemme partiklens impuls og ladning. Banerne kan ydermere benyttes til at se, om der har været skabt partikler, som efterfølgende er henfaldet til andre partikler. Dette må nemlig være tilfældet, såfremt der ikke gælder impulsbevarelse sammenlagt for de observerede partikler. Målingen af partiklernes impuls er særdeles relevant for os, da den muliggør vores beregning af den invariante masse for systemerne bestående af jets/partikel par, som dannes ved sammenstødene.

Trackeren fastlægger den enkelte partikels bane ved at interpolere mellem en række punkter svarende til hits, som partiklen har afsat i trackeren. Det viser sig, at elektroner har en generel tendens til at afsætte flere såkaldte „high threshold“ hits (engelsk: høj tærskel) i trackeren end jets, og målingen af antallet af hits kan derfor bruges til separationen af jets og elektroner. Årsagen hertil er ganske kompliceret, men forsøgt forklaret nedenfor.

Trackeren er opbygget af mange tynde lag af materiale med forskellige brydningsindeks. Når en ladet partikel passerer overgangen mellem to materialer er der en vis sandsynlighed for, at den udsender fotoner i form af „transition radiation“ (dansk: overgangsstråling). Fotonerne giver „high threshold“ hits, og det er dette antal vi måler som hits i trackeren. Det har vist sig, at en ladet partikels sandsynlighed for at udsende en foton ved en given overgang er proportional med den relativistiske gammafaktor, dvs. den er proportional med E/m. For en given energi vil partikler med lille masse derfor generelt udsende

[1] Se Data-afsnit.



Figur 3: ATLAS – detektoren der bruges ved LHC.

flere fotoner, og dermed flere „high threshold“ fotoner. De partikler vi finder i vores jets er stort set kun pioner, der er ladede partikler opbygget af en kvark og en antikvark, og selv den letteste pion er ca. 300 gange så tung som en elektron.

3.2. Det Elektromagnetiske Kalorimeter

Det elektromagnetiske kalorimeter har til formål at registrere mængden af energi, der afsættes af partikler, som udsender elektromagnestisk stråling dvs. elektroner, fotoner og andre ladede partikler. Dataet fra det elektromagnetiske kalorimeter indeholder derudover også positionen for, hvor energien er blevet afsat. Begge disse målinger har en meget lille usikkerhed forbundet med sig. Det er relevant for vores projekt, at det har vist sig, at elektroner generelt afsætter en større andel af deres energi i elektromagnetiske kalorimeteret end jets.

3.3. Magnetsystemet (Solenoid / Barrel Toroid)

Det er kun igennem det todelte magnetsystemet (på figur 3: solenoid og barrel toroid) og den præcise tracker, at det er muligt at bestemme partiklernes impuls. Magnetfeltet bliver nødt til at være meget kraftigt, fordi vores partikler bevæger sig med relativistiske hastigheder, og da det kun kan påvirke partiklerne over en kort strækning. Magnetfeltet har derfor en styrke på 2 Tesla.

4. Data

4.1. Teori og Beregninger

Da masse og energi er ækvivalente størrelser jævnfør relationen $E = \gamma mc^2$, kan der ved partikelsammenstød dannes nye partikler. I en del af tilfældene vil en kvark indeholdt i en proton kollidere med en

tilsvarende antikvark fra en anden proton og danne to modsatrettede klynger af partikler kaldet jets. Et væsentligt fokus for dette projekt vil blive at identificere sammenstødene

$$q + \bar{q} \longrightarrow \text{jet} + \text{jet}$$

og sortere disse fra originaldataene, idet vores primære interesse som tidligere nævnt er at påvise dannelsen af Higgs-bosonen, H^0. Denne forventes imidlertid at henfalde til et elektron-positronpar efter kort tid ved processen

$$q + \bar{q} \longrightarrow H^0 \longrightarrow e^- + e^+.$$

Det skal bemærkes, at Higgs-bosonen i virkelighedens verden typisk vil henfalde til tungere partikler, men vi kigger i vores simulation kun på henfald af ovennævnte type. Situationen kompliceres nu, da der ved en del af kollisionerne vil genereres såkaldte Z^0-partikler, der ligesom Higgs-bosonen henfalder til et elektron-positronpar ved den tilsvarende proces

$$q + \bar{q} \longrightarrow Z^0 \longrightarrow e^- + e^+.$$

4-impulsen, $\mathbf{P} = (p_E, p_x, p_y, p_z)$, for elektronen og 4-impulsen, $\mathbf{Q} = (q_E, q_x, q_y, q_z)$, for den tilsvarende positron kan bestemmes ved at transformere vores måledata som det er beskrevet i appendiks 1. Elektron-positron-systemets invariante masse Bmass kan da udregnes ved hjælp af formlen

$$\text{Bmass} = \frac{\sqrt{(\mathbf{P}+\mathbf{Q})^2}}{c} = \frac{\sqrt{(p_E+q_E)^2 - (p_x+q_x)^2 - (p_y+q_y)^2 - (p_z+q_z)^2}}{c}.$$

Idet det forventes, at Higgs-bosonens masse er større end Z^0-partiklens, kan det dermed afgøres, hvorvidt nogle af elektron-positron parrene stammer fra henfald af Higgs-bosonen. Man bør således ved at plotte antal partikler som funktion af den invariante masse observere en peak for Higgs-bosonen. Det er klart, at ovenstående teori er en forholdsvis grov simplifikation af virkeligheden. Her tænker vi på vores indskrænkning til kun at se på en type henfald, og at vi i vores simulation ikke har deciderede fejlmålinger, men dette er nødvendigt for at kunne beskæftige sig med emnet på dette niveau og inden for den afsatte tidsramme.

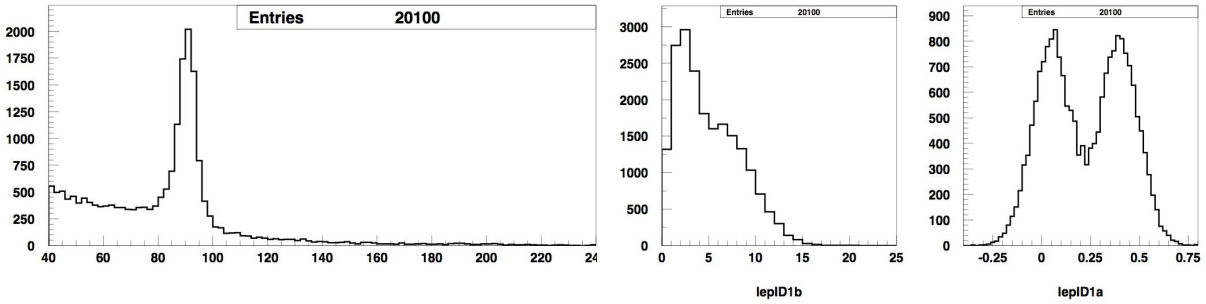
Statistisk bemærkning: I resten af projektet vil der blive brugt en lang række statistiske begreber og metoder. Disse vil ikke blive udførligt beskrevet her, men der findes et utal af glimrende materiale såvel i bogform som på internettet. Vi har valgt at bruge [7], [8] og [9].

4.2. Første Simulation

Den første simulation består af et sample på 20.100 events. Der er 10.000 jet/jet events, mens 10.100 er e^+/e^- events, hvoraf 100 kommer fra Higgs-bosoner. Hver event er karakteriseret ved 11 variable: Bmass, $p_{x_1}, p_{x_2}, p_{y_1}, p_{y_2}, p_{z_1}, p_{z_2}$, lepID1a, lepID2a, lepID1b og lepID2b. Den ekstra indicering skyldes, at vi har to jets/partikler for hver event. Bmass er den invariante masse af vores jet/partikel par, mens lepID1a, lepID2a, lepID1b og lepID2b er såkaldte identifikationsvariable. lepID1a og lepID2a er andelen af energi afsat i det elektromagnetiske kalorimeter for hhv. elektronkandidat 1 og 2, og lepID1b og lepID2b er antallet af "High Threshold" hits i trackeren for hhv. elektronkandidat 1 og 2. Det er en fundamental antagelse, at lepID1 og lepID2 er ukorrelerede, når vi betragter et sample af enten elektroner eller jets. Dette er rimeligt at antage, da de to partikler/jets efter sammenstødet ikke kommer i kontakt med hinanden. Fordelingen af lepID1b må følgeligt være den samme som for lepID2b og tilsvarende for lepIDa. Vi kan til gengæld godt have, at lepIDa og lepIDb er korrelerede. Det viser sig dog ikke at være tilfældet i dette sample. Vi vil først gennem simple cuts på vores identifikationsvariable, og derefter ved brug af likelihood-maksimalisering separere jets og elektroner bedst muligt, for dermed at kunne få et så rent elektronsample som muligt. På baggrund af et plot over Bmass variablen vil vi så efterstræbe at gøre vores Higgs-peak så signifikant som muligt.

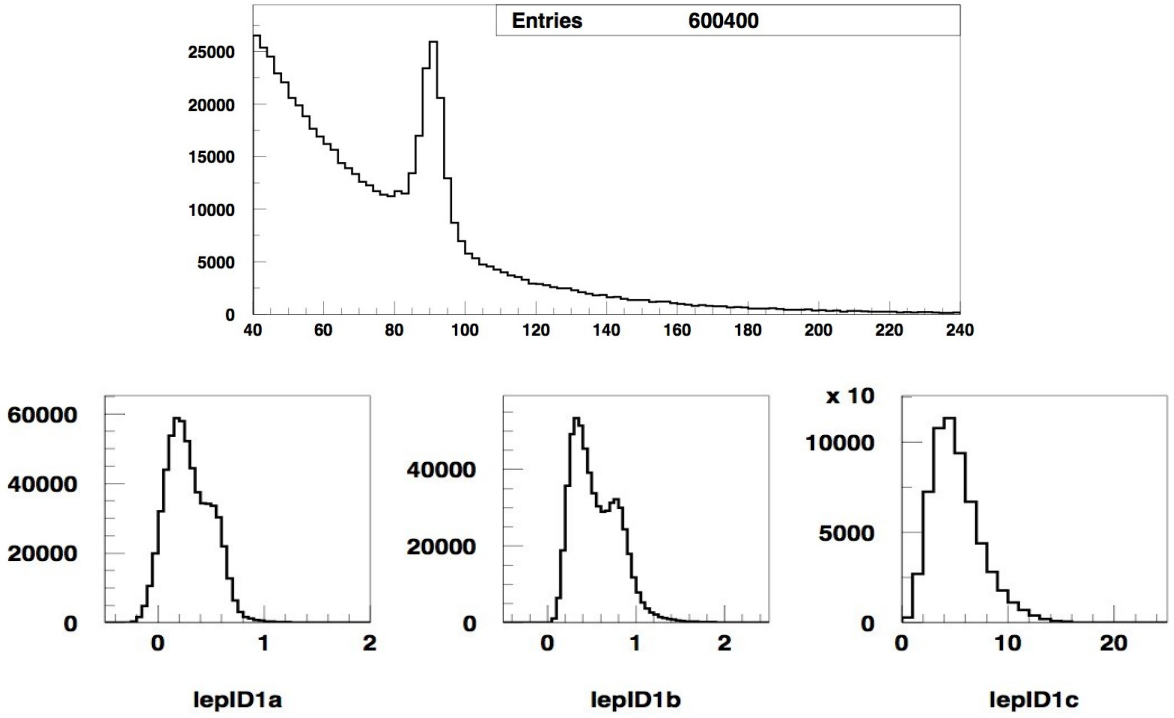
4.3. Anden Simulation

I vores anden simulation er antallet af events øget til 600.400. Der er 500.000 jets og 100.400 elektroner, hvoraf 400 kommer fra Higgs-bosoner, så andelen af baggrund er altså øget betydeligt, hvilket også fremgår af figur 5, øverst. Antallet af ID-variable (figur 5, nederst) er desuden øget til tre, således at



Figur 4: Fordelingen af Bmass, lepID1a og lepID1b for første simulation.

hver event nu er karakteriseret ved i alt 13 variable (hver elektronkandidat har fået én variabel ekstra). Den nye variabel skal forstås som endnu en måling fra kalorimeteret, men det er ikke nærmere specificeret hvorledes den adskiller sig fra vores anden måling. Et naturligt gæt ville være at målingerne kommer fra to forskellige steder i kalorimeteret. I dette sample er lepIDa og lepIDb kalorimetermålingerne, og lepIDc er vores måling fra trackeren. Det viser sig at vores ID variable for det nye sample ikke er ukorrelerrede, faktisk er korrelationen ganske betydelig. Dog har vi naturligvis stadig at lepID1 og lepID2 er ukorrelerede, med samme argument som i sidste afsnit. Korrelationen betyder at vi kun kan cutte på lepID2 og Bmass når vi vil undersøge lepID1, idet vi ellers kan risikere at få misvisende fordelinger. Hvad værre er at det gør vores likelihoodmetode betydeligt svagere, idet udledningen af vores likelihoodfunktion forudsætter uafhængige variable. Man kunne selvfølgelig sagtens forestille sig en likelihoodfunktion for de korrelerede variable, men hvordan denne skulle konstrueres er meget uklart, da vi ikke direkte kan finde den simultane fordeling (fordelingen af alle variablene samtidig) udfra de marginale fordelinger (fordelingen af hver variabel). I vores anden simulation gør vi derfor primært brug af mere avancerede metoder til at separere vores variable. Det drejer sig om neurale netværk og Fischerdiskriminant, som forklares senere.



Figur 5: Fordelingen af Bmass, lepID1a, lepID1b og lepID1c for anden simulation.

5. Første simulation

5.1. Separation ved simple cuts

5.1.1. Fremgangsmåde

Det ønskes at separere elektronerne og positronerne fra jets i datasætte. Strategien er sådan set forholdsvis simpel. Vi finder et passende stort sample med noget, vi mener stort set kun er jets, og ser på hvordan det fordeler sig i vores ID variable. Med denne information kan vi så lave passende cuts på vores ID variable alt efter, hvor rent et elektron-sample vi ønsker.

Vi skal altså skaffe os et rent jet-sample og bruger i den forbindelse vores viden om, at elektronerne i vores sample er dannet ved henfald af enten en Z^0- eller en H^0-boson. Det vides fra forøg ved CERN, at H^0-bosonens masse må være over 115 GeV/c^2, og Z^0-bosonens masse kendes til 91,1 GeV/c^2. Den peak som genereres af Z^0-bosonen kan approksimeres med en Gauß-fordeling med middelværdi $\mu = 91,1$ og spredning $\sigma \approx 3$ (den nøjagtige værdi er ikke væsentlig her). Approksimationen er ikke særlig god i halerne af fordelingen[2], men ser vi på alle events, der opfylder, at Bmass $<$ 60 GeV/c^2 er vi alligevel så tilstrækkeligt mange standardafvigelser væk fra vores Z^0-peak, at andelen af elektroner burde være mindsket drastisk. Vi lægger nu ydermere cuts ind på de to variable lepID2a og lepID2b idet vi ved, at jets overvejende har lavere værdier i begge variable, og plotter lepID1a og lepID1b under disse betingelser (husk at de to par af variable er uafhængige). Hvor strenge cuts man vil bruge afhænger af, hvor rent og hvor stort et sample man ønsker. Strengere cuts giver naturligvis bedre renhed men også et mindre sample. Vi går efter ca. 2.000 events og finder, at lepID2a $< 0,1$ og lepID2b ≤ 2 giver et sample på 1988 events (figur 6, højre). Under processen med at finde de rigtige cuts overbeviser vi os desuden om renheden af vores sample, idet der til sidst ikke er nogen synlig ændring i fordelingen ved yderligere nedjustering af cuts.

Vi føler nu, at vores sample er forholdsvist rent, og mener at de 2.000 events kan antages at repræsentere fordelingen af jets. Vores mål er at nå frem til et elektron-sample med renhed i størrelsesordenen 99,9%. Således ser vi derfor på vores jet-sample, og finder frem til hvilke cuts vi skal lægge på ID1 variablene, for at der er højst 2 events tilbage idet 2/2.000=1-99,9%. Vi når frem til, at hvis vi ydermere kræver, at lepID1a $> 0,3$ og lepID1b ≥ 5, så er der kun en event tilbage og konklusionen er derfor, at højst 1 per 1.988 jets opfylder dette krav. Vi får nu ved at plotte lepID1a og lepID1b under betingelserne lepID2a $> 0,3$ og lepID2b ≥ 5 et elektronsample på 7.083 events (figur 6, venstre), altså en effektivitet på ca. 67%, med den ønskede renhed. Bedre effektivitet kommer på bekostning af renheden, og kan opnås ved at lægge blidere cuts.

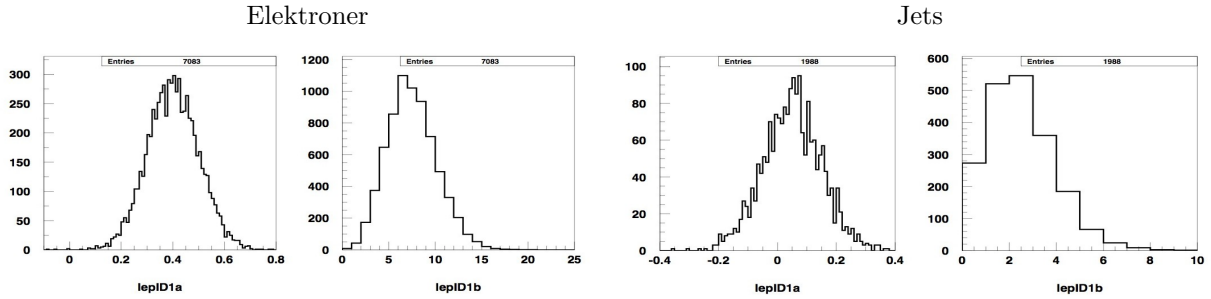
Vi kan nu plotte vores Bmass variabel for dette sample, og nu hvor næsten al baggrunden er væk, kan vi observere endnu en peak ved ca. 200 GeV/c^2. For at finde ud af om peaken er signifikant, og i givet fald hvor signifikant, skal vi lave et fit. Problemet er nu, at vi ikke har nogen garanti for, at bare fordi vi har en meget lav baggrundseffektivitet (urenhed), så fås også den største signifikans. Effektiviteten af vores signal (elektronerne) har utrolig meget at skulle have sagt, og denne er ikke specielt god i vores tilfælde. Derfor prøver vi samtlige kombinationer af cuts for lepID2a $\in \{0,0; 0,2; 0,4\}$ og lepID2b $\in \{2,3,4,5,6,7,8\}$, hvor 0,2 og 6 giver stort set det samme som eksemplet ovenfor. Gruppen med lepID2a $> 0,4$ er primært med for at illustrere betydningen af signaleffektiviteten. Resultaterne er anført i en tabel i næste afsnit, da vores to variable umuliggør et ordentligt 2D plot.

Vi vil fitte Higgs-peaken med en Gauß-fordeling, men lægger derudover et eksponentialled til i vores fitfunktion. Dette gør vi for at at få den baggrund, der stadig er tilbage, med i fittet, og denne ser umiddelbart ud til at kunne fittes med en eksponentielt aftagende funktion. Vores fitfunktion er derfor givet ved

$$\frac{c}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} + ke^{-\lambda x} \,,$$

hvor vi fitter efter parametrene $p_1 = c$, $p_2 = \mu$, $p_3 = \sigma$, $p_4 = k$ og $p_5 = \lambda$. Resultatet af fittet for førnævnte sample samt for en række andre samples behandles i næste afsnit.

[2] til den interesserede kan det bemærkes at peaken i virkeligheden er en Gauß-fordeling foldet med en Cauchy-fordeling, der har meget tunge haler. Se evt. [7], [8] eller [9].



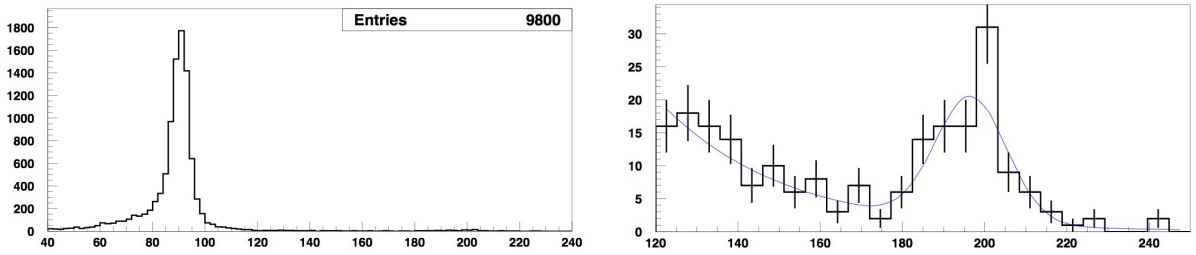
Figur 6: Fordelingen af vores ID-variable for elektron-samplet (venstre) og jet-samplet (højre).

lepID1a $\geq$	lepID1b $\geq$	Events	Signikans (σ)	ε_b (%)	ε_s (%)	χ^2
0,0	2	14.093	6,18	43	$\approx$ 100	1,82
0,0	3	11.974	6,68	23	97	1,38
0,0	4	10.189	6,88	10	92	1,25
0,0	5	8.715	6,78	3,5	84	1,03
0,0	6	7.171	5,32	1,2	72	0,67
0,0	7	5.585	6,26	0,4	56	0.38
0,0	8	4.043	6,01	$<$ 0,1	41	0,74
0,2	2	10.215	6,70	4,2	97	1,04
0,2	**3**	**9.800**	**7,58**	**2,5**	**95**	**1,12**
0,2	4	9.160	7,16	1,0	90	0,96
0,2	5	8.213	6,53	0,4	82	0,73
0,2	6	6.920	5,80	0,1	70	0,52
0,2	7	5.434	6,39	$<$ 0,1	55	0,37
0,2	8	3.936	5,80	$<$ 0,1	40	0,76
0,4	2	5.031	5,12	$<$ 0,1	50	0,92
0,4	3	4.922	5,07	$<$ 0,1	49	0,82
0,4	4	4.660	5,07	$<$ 0,1	45	0,75
0,4	5	4.234	4,74	$<$ 0,1	41	0,64
0,4	6	3.565	4,11	$<$ 0,1	35	0,43
0,4	7	2.810	4,96	$<$ 0,1	28	0,45
0,4	8	2.017	4,15	$<$ 0,1	20	0,58

Tabel 1: Resultatet af vores analyse ved simple cuts. Den fremhævede måling giver den bedste signifikans. ϵ_b og ϵ_s betegner her effektiviteten af baggrund hhv. signal i %.

5.1.2. Resultater

Vores resultater er samlet i tabel 1. Fittet giver os antallet af events i vores Higgs-peak, spredningen på dette antal samt χ^2-værdien for fittet. Signifikansen findes så ved at dividere antallet af events med spredningen, og skal intuitivt forstås som det antal standardafvigelser vores observation afviger fra en fordeling, hvor peaken ikke er der. Hvis vi fx har en signifikans på 2σ betyder det, at der er en sandsynlighed på 5% for, at vores peak bare er et udslag af tilfældighed. Denne sandsynlighed falder drastisk og er ved 4σ allerede nede på 0,01%. I forskningsmæssige sammenhænge kræves mindst 5σ inden man har gjort en opdagelse. Effektiviteten af signal og baggrund er fundet ved at se på rene fordelinger af jets og elektroner, og se hvor stor en andel der er tilbage, når man lægger de respektive cuts på. χ^2-størrelsen bør normalt ligge tæt på 1, men en mindre værdi skal ikke tages for højtideligt. Usikkerhederne er nemlig genereret automatisk af PAW og af tekniske årsager gør PAW ikke dette helt tilfredsstillende. På figur 7 er vist et histogram omkring vores Higgs-peak fittet med den føromtalte funktion. Vi har også inkluderet et Bmass plot, hvor man kan se, at baggrunden er stort set væk. Antallet af søjler er sat til 25, som følge af den anden gruppes resultater. Bemærk i øvrigt, at vi på baggrund af vores fit også kan give et estimat for Higgs-bosonens masse, hvilket dog ikke er inden for dette projekts rammer.



Figur 7: Histogram over Bmass (venstre) og fit af Higgs-peaken (højre)

I tabel 1 ses det tydeligt, at for et fastholdt cut på lepID1a variablen – for eksempel 0.2 – stiger signifikansen umiddelbart som ventet, når effektiviteten af baggrunden falder fra 4.2% med cuttet lepID1b $\geq$ 2 til 2.5% med lepID1b $\geq$ 3. Overraskende er imidlertid, at signifikansen efterfølgende falder, selvom effektiviteten af baggrunden aftager. Dette skyldes dog, at effektiviteten af elektronerne også falder, og det er således en passende vægtning af signal og baggrund, der skal optimeres. Generelt er det bedre at have et meget kraftigt signal end en meget svag baggrund. Dette ses fx i tabellen ved, at vores bedste signifikans opnås ved forholdet 2,5%/95%, hvor vi har kraftigt signal, mens vores mange samples med under 0,1% baggrund ikke giver lige så godt resultat.

5.1.3. Vurdering

Kigger vi blot på den opnåede signikans af vores Higgs-peak er resultatet jo allerede nu ganske godt. Dette er dog ikke så interessant, da vi må huske, at vi arbejder med en forsimplet simulation. Spørgsmålet er altså snarere, hvad og hvor meget vi mister ved denne simple metode. Der er ingen tvivl om, at det sample vi ender med at se på er meget rent, idet det absolutte antal af jets ikke er særlig højt. Det er resultatet af vores sorteringsproces. Imidlertid er vi ikke særligt interesseret i det absolutte antal af jets, men snarere antallet af jets i forhold til antallet af elektroner, for det er størrelsen af denne, der afgør hvor tydelig Higgs-peaken bliver. Idet vi har baggrunden med i vores fit betyder det ikke noget, at der er jets i vores endelige sample, sålænge der blot er mange flere elektroner. De simple cuts kommer her til kort, da de ikke på optimal vis minimerer andelen af jets. Der er en stor risiko for, at vi med vores cuts i bestræbelsen på at få et rent sample skærer alt for mange events væk. Cuts'ene giver os intet grundlag for at vurdere, om vi ser på et optimalt sample. Dette er vores motivation for at anvende mere raffinerede metoder; fx likelihood-maksimalisering.

5.2. Separation ved Likelihood-maksimalisering

5.2.1. Fremgangsmåde

Ideen ved likelihood-maksimalisering er, at vi ved at konstruere en såkaldt likelihood-funktion kan lave en ny variabel der angiver sandsynligheden for, at en given event er en elektron. Dette har den fordel, at vi kun har én variabel at cutte på, og at vi præcist kender betydningen af denne. Sandsynligheds-variablen sammenfatter så at sige al information fra vores fire ID-variable. For at kunne lave vores likelihoodfunktion skal vi kende fordelingen lepIDa og lepIDb for henholdsvis elektroner og jets. Vi skal altså bruge nogle samples af en vis størrelse og renhed, som vi kan fitte med en fordeling. Her er en meget høj renhed at foretrække, da det giver det bedste billede af fordelingen. For elektronerne vælger vi samplet på 7.083 events fra sidste afsnit, og ved at se på dette og finde passende cuts præcis som beskrevet i sidste afsnit, når vi frem til at lepID2a $\leq 0,2$ og lepID2b ≤ 2, giver et jetsample på 6.328 events med højst 0,1% elektroner. Det er disse samples vi vælger som grundlag for vores fits. Vi kan desuden på udfra disse samples beregne korrelationen mellem vores ID-variable for jets og elektroner, ved at bruge formlen fra [8] s. 103. Vi finder for elektroner, at corr(lepIDa,lepIDb) = 0,0025 og for jets, at corr(lepIDa,lepIDb) = -0,0059, så praktisk set er vores variable ukorrelerede som ventet.

Det antages, at lepIDa er Gauß-fordelt, og at lepIDb er Poissonfordelt, hvilket er rimeligt variablernes fremkomst taget i betragtning, da ren måledata typisk vil fordele sig på denne måde i det kontinuerte henholdsvis diskrete tilfælde. Således bliver vores fitfunktioner

$$f_1(x) = \frac{c}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad f_2(x) = \frac{c\lambda^x}{x!} \cdot e^{-\lambda},$$

	μ	σ	λ	χ^2
lepIDa – elektroner	0,4011	0,0992	–	1.04
lepIDa – jets	0,0502	0,1010	–	0,61
lepIDb – elektroner	–	–	7,003	0,67
lepIDb – jets	–	–	2,025	0,19

Tabel 2: Resultatet af fits for ID-variablenes fordelinger i første simulation.

og vi fitter med samtlige parametre. Når vi så har vores fitfunktioner skal de normeres i forhold til, hvor mange jets og elektroner, der var til at begynde med. I dette tilfælde normeres begge fordelinger således til 10.000 for jets og 10.100 for elektroner. Vi betragter nu en vilkårlig event og vil bestemme sandsynligheden for, at denne event er et e^+/e^- par. Sandsynligheden for elektronkandidat 1 beregnes til

$$p_{e_{1a}} = \frac{N_{e_{1a}}}{N_{e_{1a}} + N_{j_{1a}}} ,$$

hvor $p_{e_{1a}}$ naturligvis er en funktion af lepID1a. Beregningen er helt tilsvarende for elektronkandidat 2. N_{e_a} og N_{j_a} er de normerede antalsfordelinger fra vores fits. Sandsynligheden $p_{e_{1b}}$ og $p_{e_{2b}}$ findes naturligvis helt tilsvarende, ligesom at vi også kan finde sandsynligheden for, at en given event er en jet på samme måde. Det viser sig, at vores ID-variable for de rene samples er ukorrelerede, hvilket er vist i næste afsnit[3]. Vi slutter nu, at de også er uafhængige[4]. Vi kan nu finde sandsynligheden for, at kandidat 1 er en elektron som

$$p_{e_1} = \frac{p_{e_{1a}} \cdot p_{e_{1b}}}{p_{e_{1a}} \cdot p_{e_{1b}} + p_{j_{1a}} \cdot p_{j_{1b}}} ,$$

hvor vi har været nødt til at normere sandsynligheden, fordi events med både en jet og en elektron har sandsynlighed nul. På tilsvarende måde kan vi finde p_{e_2}. Den endelige sandsynlighed for, at vores event er et e^+/e^- par, altså at begge kandidater er en elektron (positron) findes så, idet de to kandidaters variable er uafhængige til

$$p_e = \frac{p_{e_1} \cdot p_{e_2}}{p_{e_1} \cdot p_{e_2} + (1 - p_{j_1}) \cdot (1 - p_{j_2})} ,$$

hvor vi lige som før har normeret sandsynligheden, udfra vores viden om, at ingen events indeholder både jets og elektroner. Funktionen p_e er nu vores likelihoodfunktion og ved at lægge cuts ind på den (fx $p_e > 0,5$) kan vi separere jets og elektroner, og finde frem til hvilket cut, der giver den bedste signifikans.

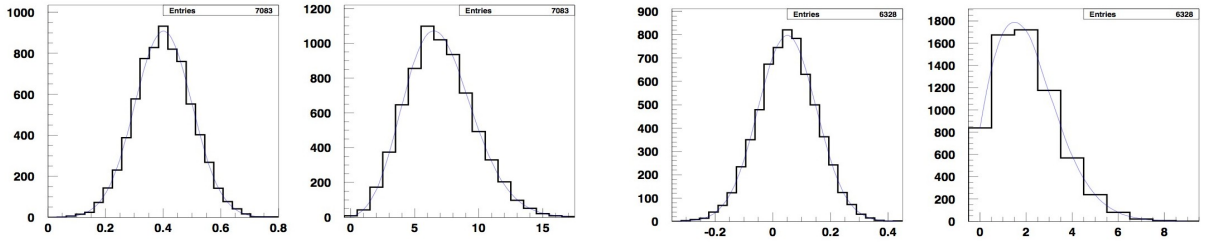
5.2.2. Resultater

På figur 9 ses fordelingerne af vores ID-variable fittet med deres respektive funktioner. Alle fittene giver lave χ^2-værdier (med samme forklaring som tidligere bør man ikke bekymre sig om at værdien evt. ligger under 1), og vi kan dermed betragte vores fitfunktioner med de estimerede parametre indsat, som repræsentative for fordelingerne. Resultatet af vores fits er opsamlet i tabel 2. Med disse funktioner kan vi nu konstruere vores likelihoodfunktion og dermed vores nye variabel. Figur 10 viser fordelingen af vores nye variabel, og man bemærker straks, at den giver en imponerende separation, hvilket overflødiggør en tabel som den vi havde i sidste afsnit. Hvis vi tager udgangspunkt i de samples vi brugte til vores fits, finder vi at allerede ved cuttet $p_e \geq 0,5$ er der kun 3 events tilbage i jetsamplet, mens elektronsamplet kun har mistet 2 events. Idet disse samples kun var vurderet til promille-renhed, så er disse tal ikke signifikante i statistisk forstand. Sagt med andre ord kunne man meget vel forestille sig, at de 3 events i jetsamplet rent faktisk er elektroner, og analogt at de 2 events i elektronsamplet er jets. Praktisk set giver dette cut altså så godt som perfekt separation. Dette bekræftes yderligere af, at hvis vi anvender cuttet på hele vores datasæt forbliver der 10.099 events, hvilket passer godt idet vores simulation indeholder 10.100 e^+/e^--events. Vi prøver nu at lave et fit af vores Higgs-peak med dette cut og får en signifikans på $7,69\sigma$ samt en χ^2-værdi på 1,10. Dette er den højeste signifikans vi har fået, og den ligger som forventet meget tæt på den vi fik ved brug af cuts for $(\varepsilon_b, \varepsilon_s) = (2.5\%, 95\%)$. Vi kan nu også med ret stor sikkerhed fastslå, at dette er den højeste signifikans vi kan få. Den begrænsende faktor er nu halen

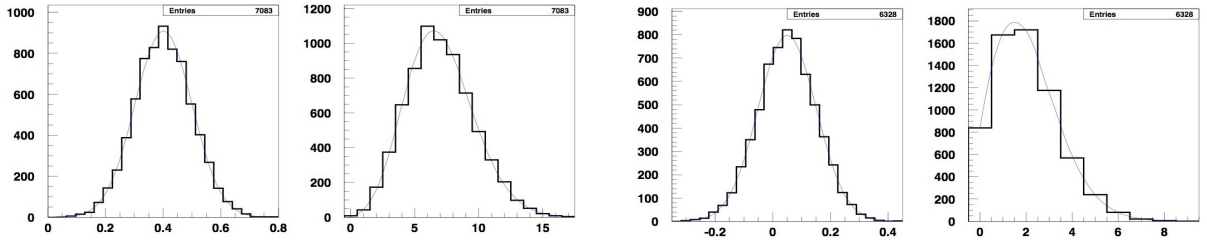
[3] Det er klart, at for hele samplet vil der være en korrelation, da elektroner har overvejende højere værdi for begge variable.

[4] Dette er ikke matematisk helt korrekt, men en god approksimation.

fra Z^0-peaken der går ind vores Higgs-peak, og den kan vi ikke fjerne, da vi i vores simulation ikke har nogen variable, der separerer "Higgs-elektroner" fra "Z^0-elektroner".



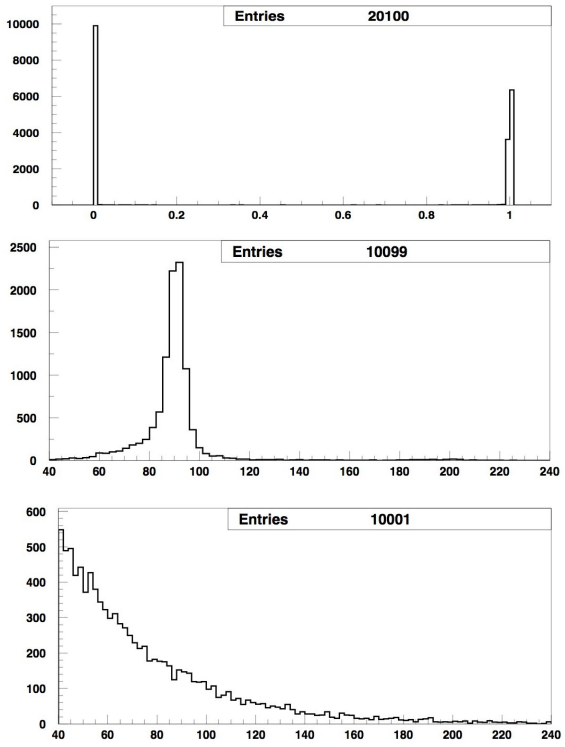
Figur 8: Histogrammer for vores ID-variable med dertilhørende fits. Elektroner t.v. og jets t.h.



Figur 9: Histogrammer for vores ID-variable med dertilhørende fits. Elektroner t.v. og jets t.h.

5.2.3. Vurdering

Vi har nu forsøgt os med en mere raffineret metode til separation af signal og baggrund, og tager endnu engang vores resultater op til kritisk vurdering. Det første vi konkluderer er, at vi har opnået en markant forbedring af separationen i forhold til de simple cuts. Dette skyldes, at likelihood-estimationen på meget mere optimal vis kombinerer den viden vi får om en event ved at bruge alle vores ID-variable, og det virker også ret intuitivt, at sandsynligheden for, at en given event er en elektron, må være noget nær den optimale variabel at separere på. Det viser sig faktisk at give så godt som fuldstændig separation af vores variable. Ved at bruge likelihoodfunktionen får vi desuden sammenfattet vores fire lidt uhåndterlige ID-variable til en enkelt meget letforståelig variabel, som vi let kan justere på. Konklusionen er altså, at likelihood-estimation for denne første simulering, er en meget stærk metode, men den står og falder med et meget vigtigt punkt. Det er helt afgørende, at vores ID-variable er ukorrelerede, idet vi ellers ikke kan finde den simultane fordeling (fordelingen af dem alle på en gang) ved at gange de marginale fordelinger (deres fordelinger hver for sig) sammen, og dermed bliver det meget kompliceret at konstruerer vores likelihoodfunktion.

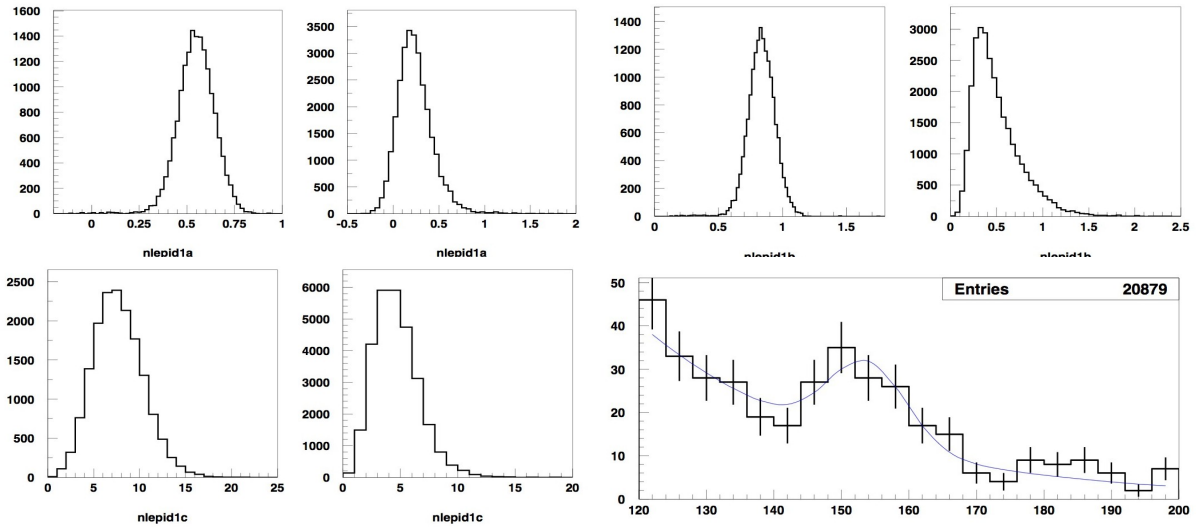


Figur 10: Fordelingen af likelihood, og fordeling af Bmass for jets og elektroner (cut ved $p_e = 0.5$).

6. Anden Simulation

6.1. Separation ved Cuts

Den fremgangsmåde, vi benytter i vores anden simulation, for at separere elektroner og jets ved brug af simple cuts på vores ID-variable er stort set identisk med den beskrevet under første simulation, og kun ændringer vil blive fremhævet. Vi når på samme måde som tidligere frem til, at lepID2a $\geq 0,5$, lepID2b $\geq 0,8$ og lepID2c ≥ 8 giver et elektron-sample på 23.022 events med højst 0.6% jets, mens lepID2a $\leq 0,3$, lepID2b $\leq 0,5$, lepID2c ≤ 2 og Bmass $\leq$ 60GeV giver et jet-sample på 28.756 events med højst 0.7% elektroner. Dette kan vi dog gøre bedre. Det viser sig nemlig, at fordelingen af jets for lepIDa og lepIDb har en meget tung hale i højre side (se figur 11), og det kan derfor betale sig også at lægge et øvre cut på elektronsamplet. Således viser det sig, at $0.5 \leq$ lepID2a $\leq 0,9$, $0,8 \leq$ lepID2b $\leq 1,2$ og lepID2c ≥ 8 giver et elektron-sample på 20.879 events med højst 0,2% baggrund, og den korrigerede mængde for jet-samplet bliver omkring 0,1%. Prisen for denne renhed har imidlertid været høj, idet vores effektivitet for elektron-samplet er nede på ca. 20%. På baggrund af vores elektron-sample finder vi en peak i vores Bmass plot ved en masse på ca. 150 GeV/c^2, hvilket naturligvis må være Higgs-peaken. Den ligger altså betydeligt tættere på Z^0-peaken i denne simulation, hvilket gør situationen besværligere, da vi så har mere "baggrund" fra denne, som jo ikke kan fjernes. Laver vi et fit (figur 11, nederst til højre) med de nævnte cuts, får vi en signifikans på 4,60σ og en χ^2-værdi på 1,15.



Figur 11: Fordeling ID-variable for elektroner (venstre) og jets (højre), samt fit af H^0-peaken.

Det er en meget omstændelig proces at optimere vores cuts, som er helt identisk med den tidligere udførte, og vi vil derfor ikke beskrive den her. En ting, vi dog bruger cuts-metoden til, er at få lavet nogle forholdsvis små, men meget rene samples. Disse kan vi nemlig bruge til at træne vores neurale netværk, der er en af de metoder vi senere vil anvende. Til dette formål vil vi, for at øge renheden af elektronsamplet, tilføje cuttet $|\text{Bmass} - 91.1\text{GeV}| \leq 10\text{GeV}$, så vi kun ser på events lige omkring Z^0-peaken, hvor andelen af elektroner er størst. Ellers er det eneste, vi yderligere vil tage med fra cuts-metoden under denne simulation, et $(\varepsilon_s, \varepsilon_b)$-plot. Dette er anført til sidst i rapporten, hvor det sammenlignes med tilsvarende plots for de andre metoder.

6.2. Separation ved Likelihood-maksimalisering

Her kan vi som udgangspunkt ikke regne med, at vores ID-variable er ukorrelerede. På samme måde som tidligere finder vi derfor korrelationerne for variablene parvist, og de er opsummeret i tabel 3. Som det ses, er det praktisk taget kun lepID1a og lepID1b, der er korrelerede, og dette åbner nogle muligheder, for selvom vi på grund af korrelationen ikke kan lave en likelihoodfunktion baseret på alle tre variable, kunne vi prøve at lave en baseret udelukkende på fx lepID1b og lepID1c. Konstruktionen vil dermed være fuldstændig analog med den, vi tidligere foretog, bortset fra at normeringerne af vores fordelinger.

	lepID1a / lepID1b	lepID1a / lepID1c	lepID1b / lepID1c
elektroner	0,3707	0,0019	-0,0097
jets	0,4567	0,0092	-0,0014

Tabel 3: Korrelationer mellem vores ID-variable i anden simulation.

	μ	σ	λ	α	β	χ^2
lepIDb – elektroner	0,8384	0,1001	–	–	–	1,33
lepIDb – jets	-0,8177	0,5038	–	–	–	3,83
lepIDc – elektroner	–	–	7,181	–	–	0,77
lepIDc – jets	–	–	–	4,449	0,9382	32,1

Tabel 4: Resultatet af fits for ID-variablenes fordelinger i anden simulation.

Som følge af, at vi ikke inddrager den sidste variabel, kan vi i denne situation ikke på samme måde fortolke likelihoodvariablen som sandsynligheden for, at en given event er en elektron. Vi tager derfor ikke den nøjagtige fordeling så højtideligt i dette tilfælde, men interesserer os blot for variablens evne til at separere jets fra elektroner.

Vi betragter nu figur 11 for at få en ide om, hvilke funktioner vi bør fitte med. Det ser umiddelbart ud til, at vi for elektronernes vedkommende kan bruge en Gauß-fordeling for lepID1b og en Poisson-fordeling for lepID1c, ligesom tidligere. Helt så enkelt ser det ikke ud til at være for jets. Vi har prøvet nogle forskellige funktioner og fik en rimelig god tilpasning ved at fitte lepID1b med en logaritmisk normalfordeling. lepID1c kunne også for jets se ud til at være Poisson-fordelt, men dette giver en utrolig ringe tilpasning. Den bedste tilpasning fik vi ved at benytte en gammafordeling, uden at det dog blev særlig godt. Af nye fitfunktioner har vi således

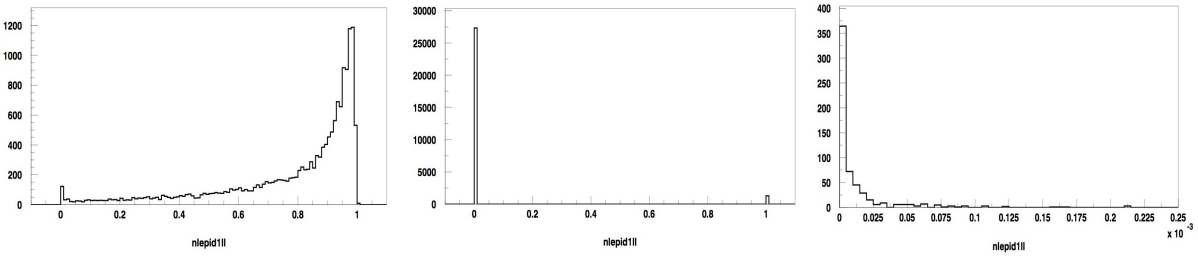
$$f_3(x) = \frac{c}{x\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}\,, \quad f_4(x) = cx^{\alpha-1}e^{-x/\beta}\,.$$

Resultatet af vores fits er opsummeret i tabel 4 og figur 13. Figur 12 viser fordelingen af vores likelihoodvariabel (nlepID1ll) for henholdvis jets og elektroner. For elektroner er fordelingen rimelig som forventet, men ved nærmere undersøgelse af jetfordelingen, opdager vi noget mystisk. På figuren kan man se, at jetfordelingen udover peaken ved nul også har en peak ved 1. Dette behøvede sådan set ikke at være mærkeligt, da det kunne være en urenhed i samplet, men i så fald burde den ligne vores elektronfordeling, når man zoomede ind på den, hvilket ikke er tilfældet. Det viser sig, at jetfordelingen faktisk er udartet i 1, og dermed er peaken et resultat af en række events, der har værdien præcis 1 (eller i hvert fald så tæt på at programmet ikke kan regne med det). Hvad forklaringen på dette er, har vi ikke været i stand til at nå frem til, men det svækker naturligvis vores likelihoodfunktion en smule. På figur 12 (højre) ses det, at dette problem ikke forekommer i 0. Dette til trods viser vores likelihoodfunktion sig faktisk at give gode resultater. Det ses umiddelbart af de førnævnte fordelinger, at et cut i intervallet $(0.0, 0.1)$ nok vil give den bedste separation. Vi ser ingen grund til at lave en større tabel med vores undersøgelser, men nævner blot at cuttet nlepID1ll ≥ 0.03 umiddelbart giver de bedste resultat. For dette cut giver et fit (figur 13) signifikansen 7.55σ med en χ^2-værdi på 1,05 og effektiviteter $\varepsilon_b = 5,1\%$ og $\varepsilon_s = 95\%$. Her skal ε_b givetvis tages med et gran salt grundet det tidligere nævnte problem med jetfordelingen. Konkluderende set får vi nogle overraskende gode resultater med vores likelihoodmetode, men det skyldes primært, at vi var i stand til at se på et par bestående af ukorrelerede variable. Havde dette ikke været muligt, var resultatet helt sikkert ikke blevet så godt. Man kan selvfølgelig altid lave en likelihoodfunktion baseret på en variabel alene, men så er vi ikke nået meget længere end med almindelige cuts.

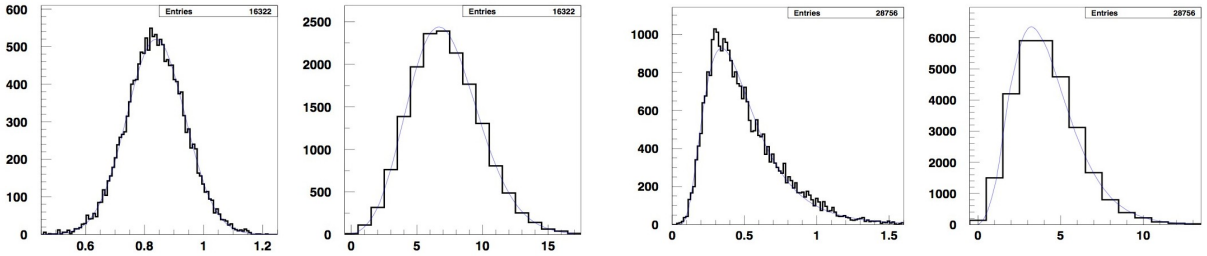
6.3. Separation ved Fischers Lineære Diskriminant

6.3.1. Fremgangsmåde

Under afsnittet om separation ved simple cuts begrænsede vi os til at cutte på hver variabel for sig, således at et cut på én ID-variabel ikke påvirkede cuttet på en anden, men dette er ikke nøvendigvis den bedste metode. Forestil dig et 3-dimensionalt koordinatsystem med lepID1a, lepID1b og lepID1c



Figur 12: Fordelingen af likelihood for elektroner (venstre) og jets (midt og højre).



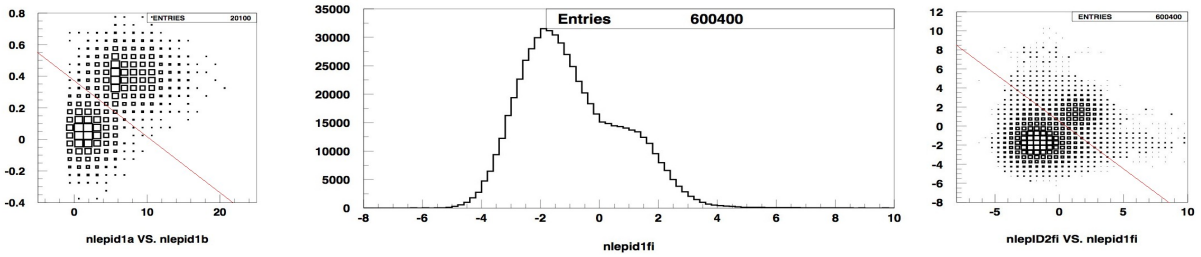
Figur 13: Fordeling ID-variable for elektroner (venstre) og jets (højre), med fits.

ud af hver sin akse, og forestil dig alle vores events plottet i dette koordinatsystem (vi har med vilje ikke inkluderet et 3D plot, da dette viste sig at være helt ubrugeligt). De cuts, vi hidtil har lavet, har svaret til at skære en kasse ud af dette koordinatsystem, men det kunne jo godt være, at man fik en meget bedre separation ved at skære en anden figur ud. En af mulighederne kunne f.eks. være at lægge en plan ind i koordinatsystemet, således at vores data blev delt i to. Dette er svært at illustrere i det 3-dimensionale tilfælde, men lad os som eksempel betragte vores første simulation, hvor vi jo kun havde to variable, således at vi ikke skal skære med et plan, men med en linje. Figur 14 viser de to ID-variable plottet mod hinanden, og her ser vi tydeligt to grupper svarende til jets og elektroner. Hvis vi vil skille de to grupper ad bedst muligt, er det ret klart, at et rektangel ikke er det bedste udsnit at vælge. Derimod ville den indtegnede røde linje give en glimrende separation, og på samme måde kunne man i tre dimensioner indtegne en plan. Spørgsmålet er nu hvilken plan, der giver den bedste separation af datamaterialet, og det er her Fischer-diskriminanten kommer ind i billedet.

Vi ønsker altså at konstruerere en ny variabel $k_1 \cdot \text{lepID1a} + k_2 \cdot \text{lepID1b} + k_2 \cdot \text{lepID1c}$ (og tilsvarende for lepID2), således at et cut på denne svarer til at skære med en plan i vores koordinatsystem. Opgaven er at bestemme konstanterne k_1, k_2 og k_3, således at planen separerer bedst muligt. Til at klare denne numerisk tunge opgave har vi fået et program. Ideen er, at vi konstruerer nogle forholdsvis små (15.000-30.000 events) og meget rene samples af jets og elektroner, som vi giver til programmet, således at det kan genkende dem fra hinanden. Programmet udregner så på baggrund af disse informationer Fischers lineære diskriminant, som er en numerisk størrelse, der kan bruges til at optimere parametrene for vores plan, således at separationen bliver bedst mulig. Metoden virker i øvrigt for et vilkårligt antal variable. Programmet finder et sæt passende parametre og konstruerer to nye variable nlepID1fi og nlepID2fi, en for hver elektronkandidat, og lægger dem ind i et nyt datasæt, som vi herefter kan arbejde med. Det vil føre for vidt at gå yderligere i detaljer om beregningerne, som programmet udfører.

6.3.2. Vurdering og Resultater

Da vi nu har to nye variable, der giver en bedre separation af datamaterialet, kunne vi forsøge at lægge simple cuts på disse for at adskille jets og elektroner. Figur 14 (højre) antyder imidlertid, at en ret linje måske ville være mere passende at cutte med. Dette giver os bare et problem, nemlig at finde effektiviteten af signal og baggrund ved cuttet. Vi kan nemlig ikke cutte på nlepID1fi samtidig med, at vi cutter på nlepID1a, nlepID1b og nlepID1c (i vores nye datasæt er lep erstattet med nlep for alle variable), og noget tilsvarende gælder selvfølgelig for nlepID2fi. Dette skyldes, at nlepID1fi er en linearkombination af de tre gamle ID-variable og er dermed stærkt korreleret med dem. Ser vi derfor



Figur 14: Eksempel på cut med linje (venstre) og fordelingen af Fischer-variablen (midt og højre).

a	σ	ε_b	ε_s	χ^2	a	σ	ε_b	ε_s	χ^2
-1,5	4,21	24,0%	99%	1,67	-1,0	6,03	18,0%	99%	2,23
-0,5	7,32	13,0%	99%	2,28	0,0	7,38	9,5%	98%	2,28
0,5	8,79	7,0%	98%	1,67	1,5	7,32	3,9%	93%	0,86
2,0	7,13	2,8%	85%	0,90	2,5	6,65	2,0%	73%	0,74

Tabel 5: Resultatet af vores analyse med Fischervariablen.

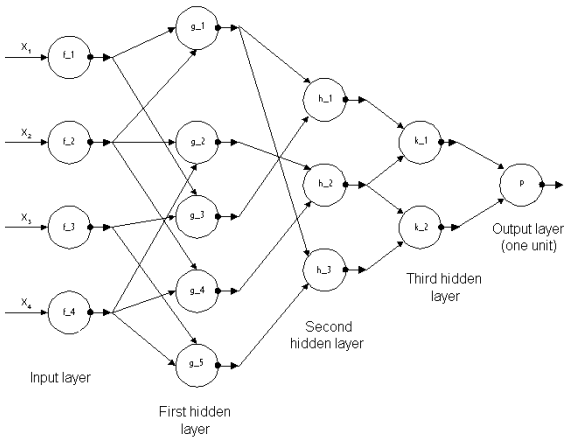
på et rent elektronsample, hvor vi har cuttet på de gamle ID1-variable, og spørger hvor stor en andel, der ligger inden for et givet cut på nlepID1fi, får vi ikke det rigtige svar. Det gør vi derimod, hvis vi i stedet cutter på nlepID2fi, da denne kun er korreleret med de gamle ID2-variable, men da et cut med en linje bruger begge Fischer-variable er dette ikke til nogen nytte. For jets er dette ikke noget problem, da vi blot kan se på et sample, hvor vi kun har cuttet på Bmass, som vi vil kunne få tilstrækkeligt rent til formålet. For elektroner må vi se på et sample, der er mindre korreleret med vores Fischer-variabel, dvs. hvor vi har cuttet på færre variable. Det viser sig, at hvis vi udelader cuttet på nlepID2c, men medtager cuttet $|\text{nBmass} - 91,1| \leq 10$, får vi et sample med omtrent de ønskede egenskaber.

Den røde linje, som er indtegnet på figur 14 (højre), kan udtrykkes som $\text{lepID1fi} + \text{lepID2fi} = 0,5$, og vores cut bliver dermed $\text{lepID1fi} + \text{lepID2fi} \geq 0,5$. Som tidligere er det dog ikke sikkert, at det er denne linje, der giver den bedste signifikans, så vi prøver med cuts $\text{lepID1fi} + \text{lepID2fi} \geq a$ for forskellige værdier af a. Resultaterne er opsummeret i tabel 5. Vi ser, at den bedste signifikans får vi faktisk for $a = 0,5$, som var vores oprindelige gæt. Ellers bekræfter tabellen de tidligere observerede tendenser, nemlig at højt signal og svag baggrund giver bedst signifikans, og at det er vigtigst med stærkt signal. Konklusionen på vores analyse ved brug af Fischer-variablen er, at metoden i høj grad formår at separere jets og elektroner, og den er derfor et godt alternativ til likelihoodmetoden, hvis variablene viser sig at være meget korrelerede. Som med likelihoodmetoden arbejder vi med nogle rimelig intuitive variable, og dermed bevarer vi forståelsen for, hvad der foregår. Dette kan godt blive et problem med den metode, som vi beskriver i næste afsnit.

6.4. Separation ved brug af Artificial Neural Networks (ANN)

6.4.1. Hvad er et ANN?

For at separare data optimalt kan man bruge Artificial Neural Networks (ANN), der er inspireret af biologiske hjerners virkemåde. For mere information se [10] og [11]. På figur 15 ses et eksempel på et ANN. Analogt til biologiske hjerner kræver et ANN ligeledes indlæring og træning inden den egentlige opgave kan udføres. De enkelte funktioner i de forskellige lag af det neurale netværk kaldes vægte, og vi træner vores ANN ved at lade det analysere et trænings-sample i stil med det vi senere ønsker at analysere. I dette trænings-sample ved vi på forhånd, hvad der er baggrund, og hvad der er signal. Vi beder nu vores ANN om at justere vægtene, indtil baggrund og signal separares på optimal måde i vores træningssample. Det er dette endelige resultat af vægtenes justering, vi benytter os af til at analysere vores egentlige måledata.

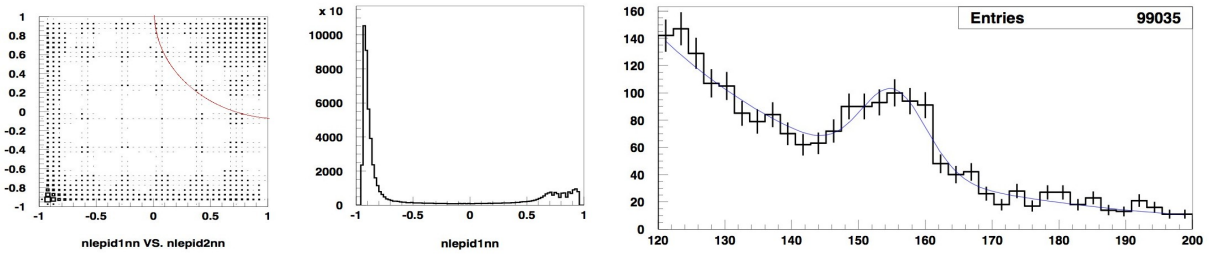


Lad $\mathcal{F}$ betegne mængden af mulige vægte. Da ønskes $C : \mathcal{F} \to \mathbb{R}$ så $C(\mathbf{v}^*) \leq C(\mathbf{v})$ for alle $\mathbf{v} \in \mathcal{F}$. C er altså en på funktion vi definerer og som bruges til at evaluere ANN's effektivitet.

Figur 15: En illustration af, hvordan et ANN kan være bygget op. Begyndelsesværdien $x = (x_1, x_2, x_3, x_4)$, som altså er input, sendes gennem en række layers. Hvert af disse layers indeholder en række vægte, der hver især bearbejder dataen fra foregående layer. Pilene indikerer, hvilke input en given vægt har; fx har h_3 både g_1 og g_5 som input. Neurale netværk kan fx bruges til at genkende billeder af bogstaver, som man efterhånden ofte møder på internettet. [12]

6.4.2. Fremgangsmåde og Resultater

Ligesom ved Fischer-metoden har vi fået udleveret et program, som vi kan give nogle rene samples af jets og elektroner, således at det kan genkende disse fra hinanden. Programmet vil så anvende et neuralt netværk på hele vores datasæt og give os to nye variable, nlepID1nn og nlepID2nn (figur 16), der begge ligger i intervallet [-1,1]. Der er ingen intuition forbundet med disse variable. Vi vil på ingen måde beskrive yderligere, hvordan det neurale netværk bearbejder vores data, men i stedet vil vi nøjes med at forholde os til resultaterne af processen. På figur 16 ses fordelingen af variablene nlepIDnn, og vi bemærker naturligvis straks den store separation. Tilsyneladende kan det bedst betale sig at skære med en kugle omkring punktet (1,1), altså et cut på formen $(\text{nlepID1nn} - 1)^2 + (\text{nlepID1nn-1})^2 \leq r^2$. Ved at prøve os lidt frem finder vi, at $r^2 = 1,05$ umiddelbart giver det bedste resultat, nemlig en signifikans på 9,65σ og en χ^2-værdi på 1,08. Desuden fås effektiviteterne $\varepsilon_b = 2,6\%$ og $\varepsilon_s = 94\%$. Bemærk at det problem, vi havde med at bestemme effektiviteterne ved vores Fischer-metode, også gør sig gældende her, så effektiviteterne er behæftet med en vis usikkerhed.



Figur 16: Fordelingen af nlepIDnn (venstre) og fittet svarende til $r^2 = 1,05$.

Det neurale netværk gav altså det bedste resultat af alle vores metoder, hvilket var ventet, da det er klart den mest avancerede. Prisen har til gengæld været en stor del af intuitionen. Algoritmen er så kompliceret, at den nærmest er umulig at overskue, og derfor kalder mange fysikere et ANN for en "black box", hvor man kommer noget ind og forhåbentlig får noget rigtigt ud.

6.5. Sammenligning af metoderne.

Vi har nu benyttet en række forskellige metoder til at separere vores data og ønsker at sammenligne dem. Dette kan vi gøre ved at lave et plot over $(\varepsilon_s, \varepsilon_b)$ for hver metode. Jo stejlere kurven er jo bedre er metoden. Resultaterne kan ses i tabel 6. Dette fungerer også bedre da alle fire plot oven i hinanden vil blive meget gnidret.

ε_s	ε_b for Cuts	ε_b for Likelihood	ε_b for Fischer	ε_b for ANN
20%	0,46%	6,0%	0,57%	0,23%
40%	2,2%	6,1%	1,0%	0,49%
60%	2,9%	6,1%	1,6%	0,83%
80%	8,1%	6,1%	2,4%	1,4%
99%	73%	22%	18%	12%

Tabel 6: Sammenligner signal- og baggrundseffektiviteten for de fire metoder anvendt i projektet.

Ikke overraskende viser ovenstående tabel, at cuts, der jo er den simpleste metode, også giver de dårligste resultater, mens ANN giver os den bedste separation.

7. Konklusion og Perspektivering

Dette projekt er svært at placere mellem et teoretisk og eksperimentelt projekt. Vi beskæftiger os i projektet udelukkende med en stærkt simplificeret model af, hvad man tror kunne være virkelighed, idet vi forudsætter, at Higgs-bosonen rent faktisk eksisterer. Den første meget grove simplificering gøres, idet vi antager, at Z^0- og Higgs-bosonen udelukkende henfalder til elektroner og positroner, og som det er beskrevet i projektet er vores data-samples forholdsvis simple. Den teoretiske del af projektet er ligeledes marginal, idet der kun skal bruges meget lidt teoretisk viden i forbindelse med separation af vores data.

Dette førsteårsprojekt er derfor langt mere en øvelse i statistisk bearbejdning af fysiske måledata end noget andet. Der er igennem hele projektet lagt vægt på, hvilke overvejelser vi har gjort os, og hvordan vi er nået frem til den endelige metode. Progressionen i projektet bør bemærkes, idet vi i takt med stigende kompleksitet, og dermed mere virkelighedstro samples, har måttet anvende stadigt stærkere metoder. I den første simulation giver likelihood-maksimalisering stort set optimal separation. I den anden simulation benytter vi os også af Fischer-diskriminant og ANN's, som begge giver en bedre separation. Disse metoder kommer dog ikke til deres fulde ret, da vores anden simulation stadig er relativt simpel, idet vi kun ser på tre variable, der er lineært korrelerede. Realistiske fysiske data vil ofte indholde langt flere variable, og vi kan ikke på forhånd udtale os om arten af korrelationerne mellem disse. I en sådan situation vil ANN klart vise sin styrke frem for de øvrige metoder. Dette betyder dog ikke, at de andre metoder er ligegyldige, ubrugelige eller forældede, da de jo netop skal bruges til at fremskaffe et sample, hvormed vi kan træne vores ANN.

Havde vi haft mere tid end disse 7 uger, ville det være oplagt at se på mere komplicerede samples med langt flere variable. Vores partnergruppe har bestemt Higgs-bosonens masse samt bredde og har undersøgt, hvor stor signifikans det er muligt at opnå afhængigt af størrelsen af Higgs-bosonens masse [13]. Dette ville også være en naturlig udvidelse af vores projekt, da en del af denne information bruges i vores opgave. Omvendt benyttede vores partnergruppe vores resultater i deres projekt, og projekterne supplerer derfor hinanden på en naturlig og interessant måde.

I vores projekt har vi været nødsaget til at fitte vores Z^0-peak med en Gauß-fordeling, hvor den sande fordeling faktisk er en Gauß- foldet med Cauchy-fordeling. At fitte med en sådan foldning er dog alt for numerisk krævende, hvorfor vi valgte ikke at gøre det. Normalt er ID-variablene også afhængige af $p_\perp$ og η, hvilket ikke var tilfældet i vores simulation.

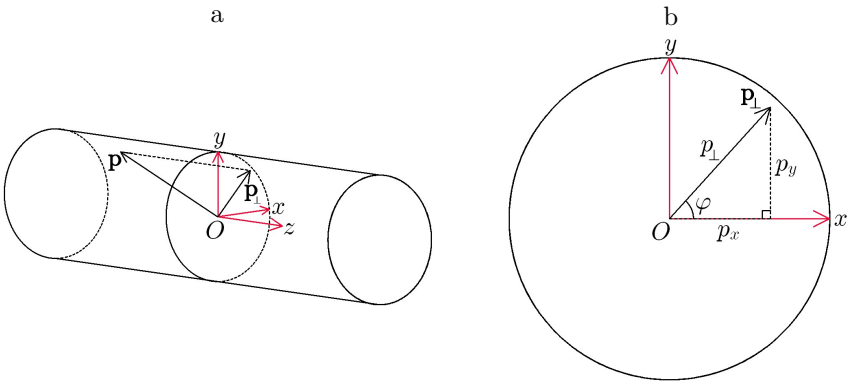
Projektet er altså et eksempel på, hvor langt en gruppe 1. års fysikstuderende med 1 års matematisk baggrund kan komme i løbet af 7 uger. Vi har opnået en viden og erfaring, som vi tror og håber på vil kunne gavne os i fremtiden, når vi beskæftiger os med fysik: Vi har lært, hvordan man analyserer måledata på professionel vis ved brug af mere eller mindre avancerede computerprogrammer.

Appendiks

A. Transformation af Måledata

Igennem hele projektet har vi henvist til p_x, p_y og p_z. Det er dog ikke disse tal ATLAS-detektoren giver os, men det er simpelt at transformere vores måledata, hvilket vi viser i dette appendiks. ATLAS-detektoren har af tekniske grunde begrænsninger på sine målinger, og det er derfor smart at outputte måledata som ATLAS gør det. Omvendt fjernes den del af intuitionen vi har brug for i dette projekt, hvorfor vi ønsker at transformere.

Betragt en elektron, som ved et henfald er opstået i punktet O med impulsen $\mathbf{p}$ (se figur 17a).



Figur 17: a: Elektronen i O har impulsen $\mathbf{p}$, hvis projektion på xy-planen er $\mathbf{p}_\perp$. b: xy-planen.

Med henblik på at afgøre hvorvidt elektronen er et henfaldsprodukt fra en Z^0-boson eller en H^0-boson, måles der tre størrelser, som er naturlige variable for detektoren at måle. Der er her tale om længden $p_\perp$ af $\mathbf{p}$'s projektion på xy-planen, vinklen φ, som denne projektion $\mathbf{p}_\perp$ danner med x-aksen, samt størrelsen

$$\eta = -\frac{1}{2}\log\left(\frac{p_E + p_z}{p_E - p_z}\right),$$

hvor p_E betegner elektronens energi delt med lyshastigheden, og p_z betegner z-komponenten for impulsvektoren. Vi begrunder nu i det følgende, at koordinaterne for elektronens 4-impulsvektor $\mathbf{P} = (p_E, p_x, p_y, p_z)$ i systemet specificeret på figur 17a er givet ved formlerne

$$p_E = \sqrt{p_x^2 + p_y^2 + p_z^2}, \quad p_x = p_\perp \cos\varphi, \quad p_y = p_\perp \sin\varphi \quad \text{og} \quad p_z = -p_\perp \sinh\eta.$$

Idet elektronens masse m_e er meget tæt på 0, kan det konkluderes, at

$$0 = m_e^2 c^2 = \mathbf{P}^2 = p_E^2 - p_x^2 - p_y^2 - p_z^2,$$

hvoraf den første formel oplagt følger. De to efterfølgende sammenhænge fremgår klart af figur 17b, og endelig giver definitionen af η, at

$$p_z = p_E \frac{e^{-2\eta} - 1}{e^{-2\eta} + 1} = p_E \frac{e^{-\eta} - e^{\eta}}{e^{-\eta} + e^{\eta}} = -p_E \tanh\eta = -\sqrt{p_x^2 + p_y^2 + p_z^2}\tanh\eta,$$

hvilket ved kvadrering giver, at

$$p_z^2 = \frac{p_\perp^2(\cos^2\varphi + \sin^2\varphi)\tanh^2\eta}{1 - \tanh^2\eta} = \frac{p_\perp^2 \sinh^2\eta}{\cosh^2\eta - \sinh^2\eta} = p_\perp^2 \sinh^2\eta.$$

Den sidste af de postulerede formler fås da umiddelbart ved udnyttelse af, at $\sinh\eta \leq 0$, netop hvis $0 \leq p_z < p_E$. Analogt med ovenstående kan 4-impulsen $\mathbf{Q} = (q_E, q_x, q_y, q_z)$ for den tilsvarende positron bestemmes.

Litteratur

[1] PAW, tutorial, dokumentation og FAQ: http://paw.web.cern.ch/paw/.

[2] Standard reference for Fortran: http://www.fortran.com/F77_std/rjcnf.html

[3] Troels C. Petersens hjemmeside: http://www.nbi.dk/∼petersen/

[4] Higgs-bosonen, 5 et-siders pædagogiske forklaringer om Higgs-bosonen: http://www.phy.uct.ac.za/courses/phy400w/particle/higgs.htm.

[5] Higgs-bosonen, mere om: http://physicsweb.org/articles/world/17/7/6.

[6] ATLAS. Wikipedia, English: http://en.wikipedia.org/wiki/ATLAS_experiment.

[7] Glen Cowan, *Statistical Data Analysis*, Oxford Science Publications, 1998

[8] Michael Sørensen, *En Introduktion til Sandsynlighedsregning*, 5. udgave, Afdeling for anvendt matematik og statistik, 2004.

[9] Ernst Hansen, *Sandsynlighedsregning på Målteoretisk grundlag*, 4. udgave, Afdeling for anvendt matematik og statistik, 2004.

[10] Neurale Netværk, Wikipedia, English: http://en.wikipedia.org/wiki/Artificial_neural_network.

[11] Neurale Netværk: http://www.doc.ic.ac.uk/ nd/surprise_96/journal/vol4/cs11/report.html#The%20Learning%20Process.

[12] Bogstav-billede. Taget fra GMail.com efter gentagen fejlindlogning.

[13] Higgs Hunting, Hans Georg Hegelund, Martin Barming Nielsen, Lea Hildebrandt Rossander, Joakim Sandroos, 27. marts 2006, NBI.