Esercizio su curva ROC

(esempio tratto dall'analisi H->ZZ(*)->4leptoni)

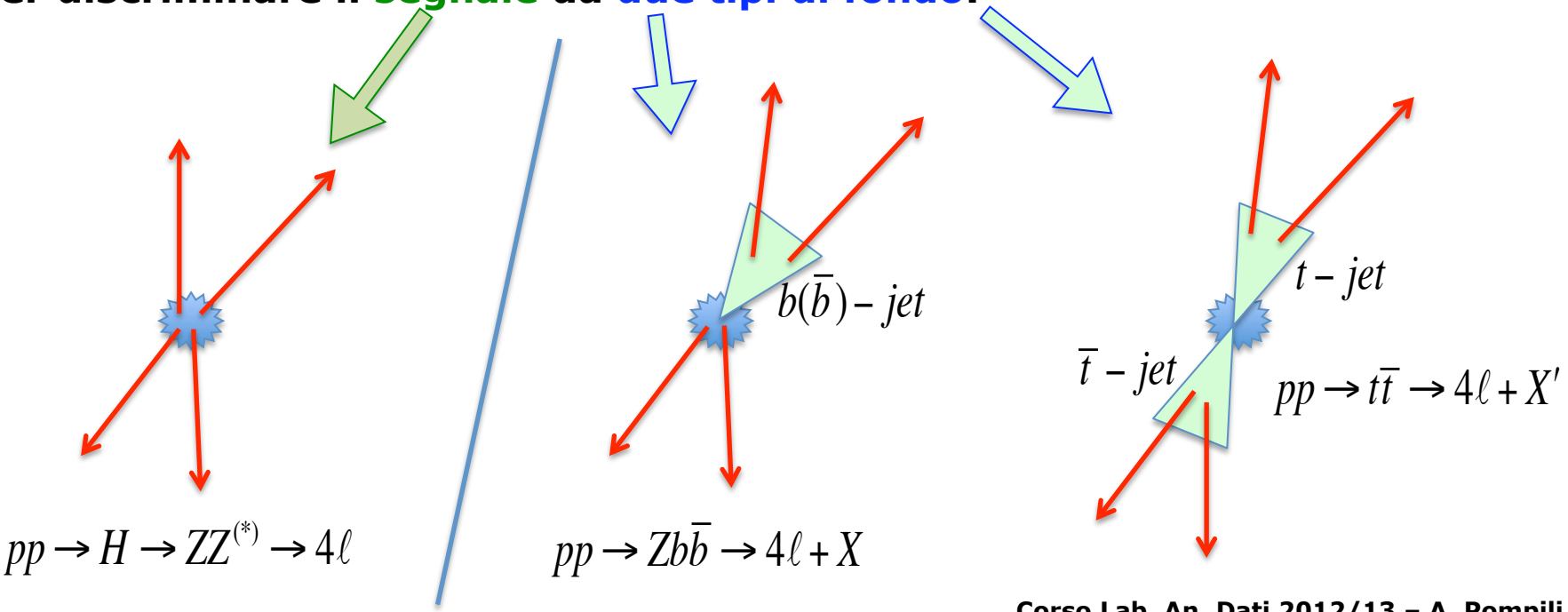
Corso Lab. An. Dati 2012/13 – A. Pompili

Descrizione - I

Nella ottimizzazione dei criteri di selezione del canale $H \to ZZ^{(*)} \to 4\ell$ si sono sviluppati - in CMS - vari algoritmi p.es. di vertexing e isolamento dei leptoni.

Con questo esercizio vediamo un esempio di confronto delle prestazioni di due (leggermenti diversi) algoritmi di "vertexing" aventi lo scopo di rigettare il fondo, effettuato sulla basi di dati simulati di segnale e fondo.

Sostanzialmente l'operazione consiste nell'usare due diverse *statistiche* per discriminare il segnale da due tipi di fondo:



Descrizione - II

$m(H) \equiv 150\,GeV/c^2$

$E_{cms} = \sqrt{s} \equiv 10TeV$

Root-upla di segnale: H150_ZZ_4l_10TeV_GEN_HLT_Presel_glb_2e2mu_2e2mu_merged.root

Root-upla di fondo Zbbar: LLBB_4l_10TeV_Presel_4l_2e2mu_merged.root

Root-upla di fondo ttbar: TT_4l_10TeV_Presel_4l_2e2mu_merged.root

In main.C l'indice i etichetta i 3 casi: i=0(segnale), i=1(ttbar), i=2(Zbbar)

Statistiche: "var1": "LeptIP3D_worst"
"var2": "LeptSTIP_SLIP_worst"

Queste osservabili sono descritte nelle slide seguenti.

Le relative stringhe vengono passate alla macro main.C al momento dell'esecuzione:

```
root> .L main.C
root> main("LeptIP3D_worst","LeptSTIP_SLIP_worst", "30nov","png")
```

...dal momento che in in main.C si ha l'interfaccia:

```
void main(TString var1, TString var2, TString date, TString extens){...}
```

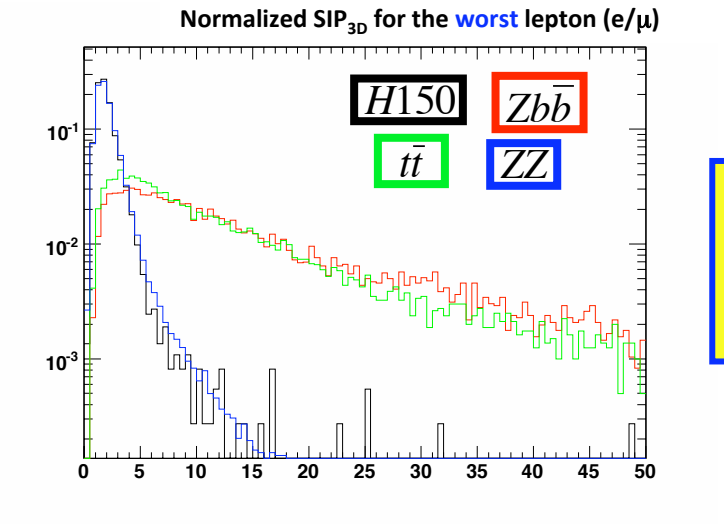
Statistiche (osservabili) - I

Statistica dall'algoritmo1: var1 ("LeptIP3D_worst")

Three-dimensional (IP_{3D}) distance - from the Primary Vertex (PV) - of the point of closest approach to PV for the "back propagated" lepton track is calculated. [Propagators are specific for muons & electrons].

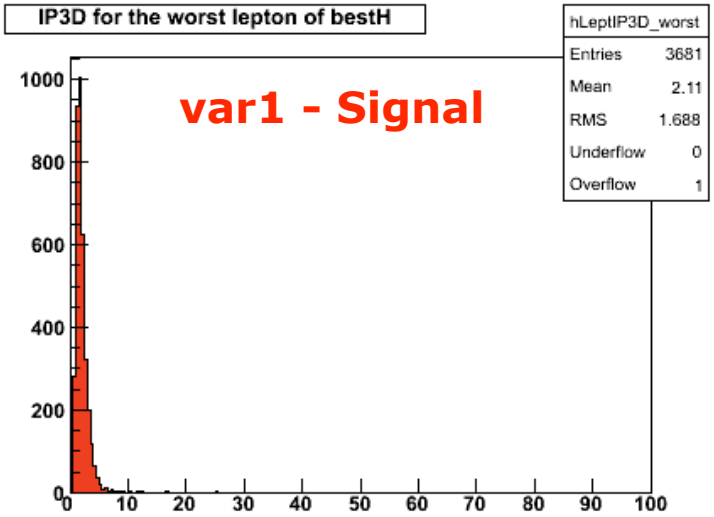
Significance (SIP_{3D}) is obtained by dividing IP_{3D} for the relative uncertainty (by full error computation):

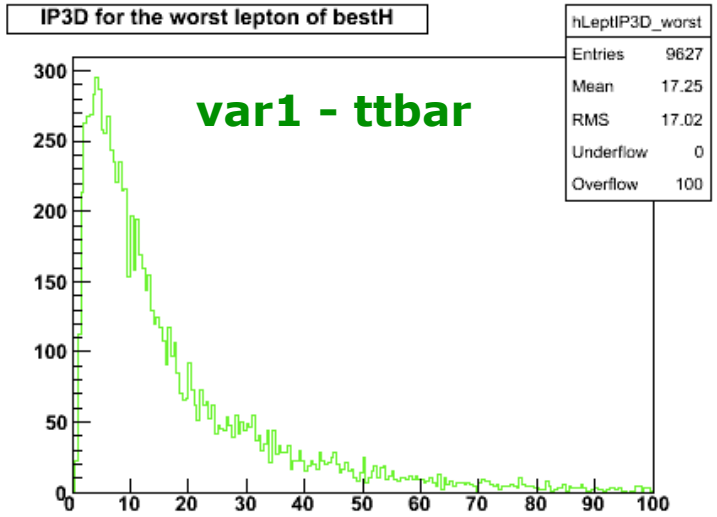
$$SIP_{3D} = \frac{IP_{3D}}{\sigma(IP_{3D})}$$

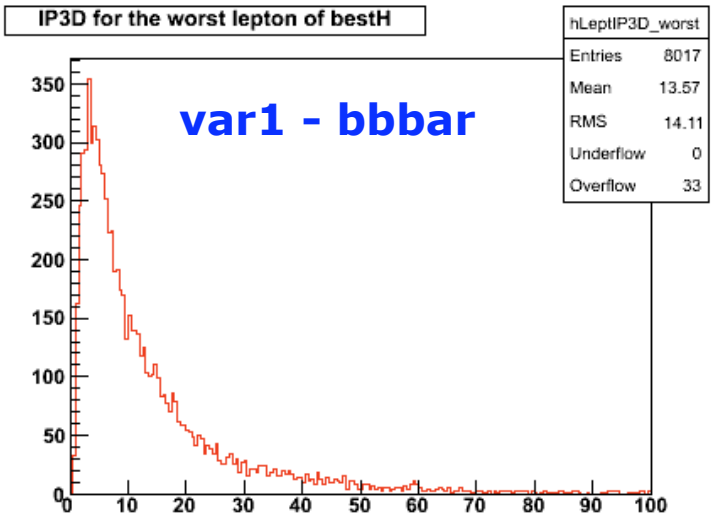


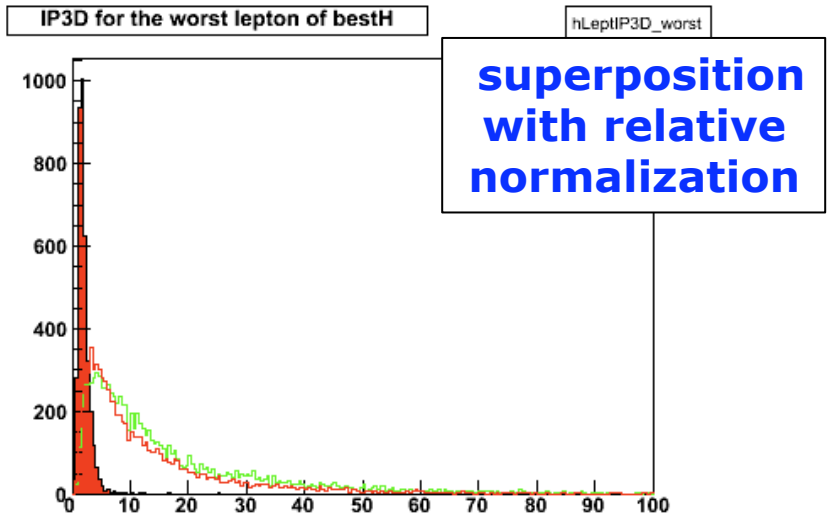
In $Zb\bar{b}$, $t\bar{t}$ and generic multi-jet background events the impact parameters of 2 or 4 leptons (in generic 4-leptons final state) are - in average - naturally larger w.r.t. those of signal.

Statistiche (osservabili) - II









Statistiche (osservabili) - III

Statistica dall'algoritmo2: var2 ("LeptSTIP_SLIP_worst")

Transverse (TIP) and longitudinal (LIP) distances - from PV - of lepton track, "back-propagated" w.r.t. to PV, are calculated.

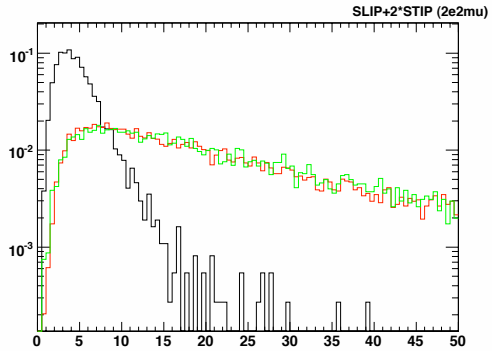
[Propagators are specific for muons & electrons].

Significances (STIP, SLIP) are taken by dividing them for the relative uncertainty (by full error computation - correlation included).

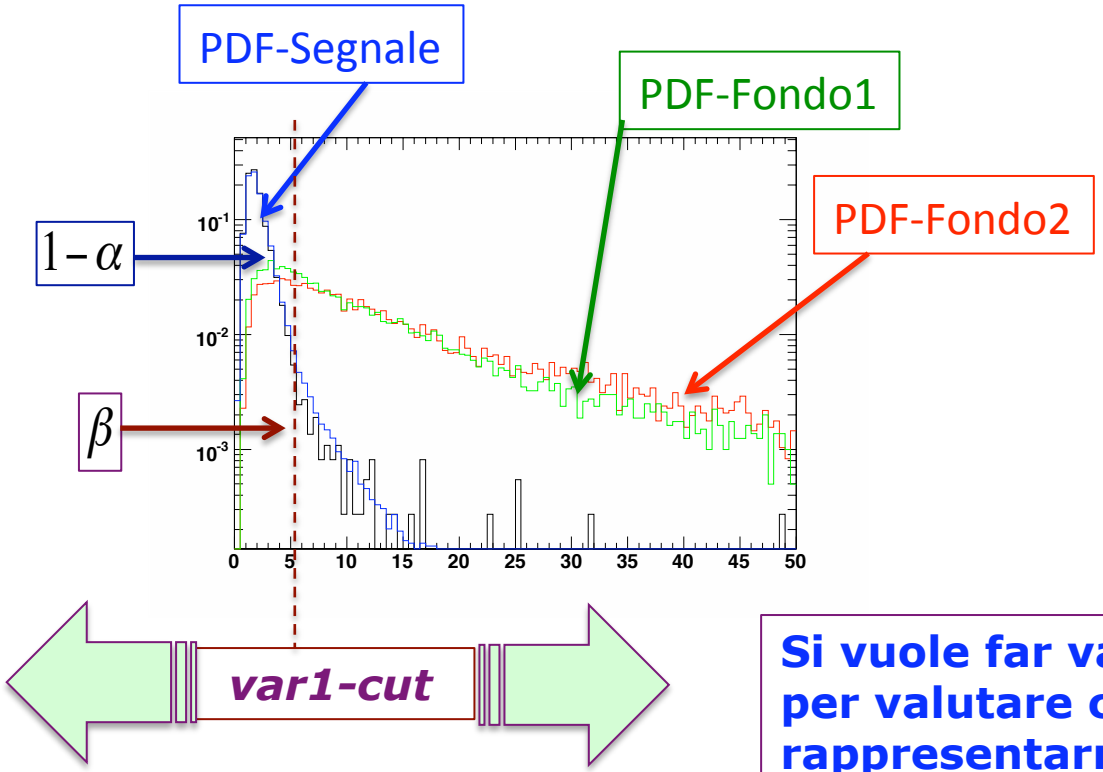
The discriminating observable used here is the following suitable weighted combination of SLIP and STIP:

$$SLIP(4^{th}\ell) + 2 * \left[STIP\left(\ell^{+}\right) + STIP\left(\ell^{-}\right) \right]$$ where ℓ^{+}, ℓ^{-} are from Z^{*}

The idea is that the leptons from b-quarks (having tipically higher STIP value) in both relevant backgrounds events - tend likely to mimic the 2 leptons involved in the Z* reconstruction.

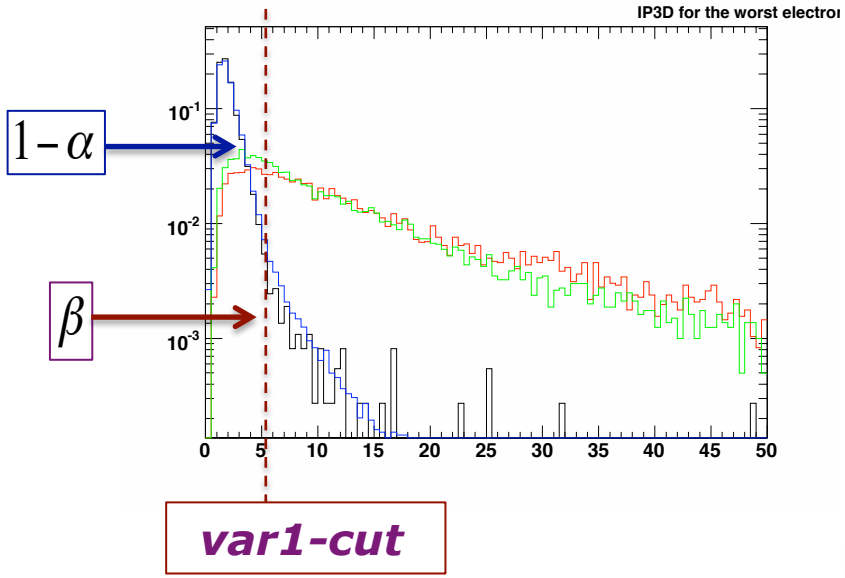


Discriminazione Segnale-vs-Fondo - I



Si vuole far variare *var1-cut* con continuita' per valutare come variano i 2 integrali e rappresentarne la variazione una contro l'altra (curva ROC)!

Discriminazione Segnale-vs-Fondo - II



Siccome le "PDF" sono "binnate" gli integrali parziali 1-α e β corrispondenti a valori progressivi di *var1-cut* saranno calcolati dal primo bin al j-esimo bin parziale fatto crescere progressivamente in un for-loop.

Siccome queste "PDF" sono PDF effettive solo se normalizzate, devo dividere gli integrali parziali per quelli totali !

```
for (int i=0; i<3; i++)
  {
    TotIntegral1[i] = hVar1[i]->Integral(1,201);
    nBins1 = hVar1[i]->GetNbinsX();

    for(int j=1; j<nBins1+1; j++)
        {
          PartIntegral1[i] = hVar1[i]->Integral(1,j);
          Eff1[i] = PartIntegral1[i]/TotIntegral1[i];
          IntegralCounts1[i][j-1]=Eff1[i]; hVarRatio[i]->Fill(Eff1[i]);
        }
  }
```

Discriminazione Segnale-vs-Fondo - III

Nel doppio loop della slide precedente …
viene riempita la matrice 3*200 IntegralCounts1(i,j) per la var1, con i 200 valori degli integrali:
- (1-α) per i=0
- β per i=1
- β per i=2

Analogamente nella macro si procede con la var2.

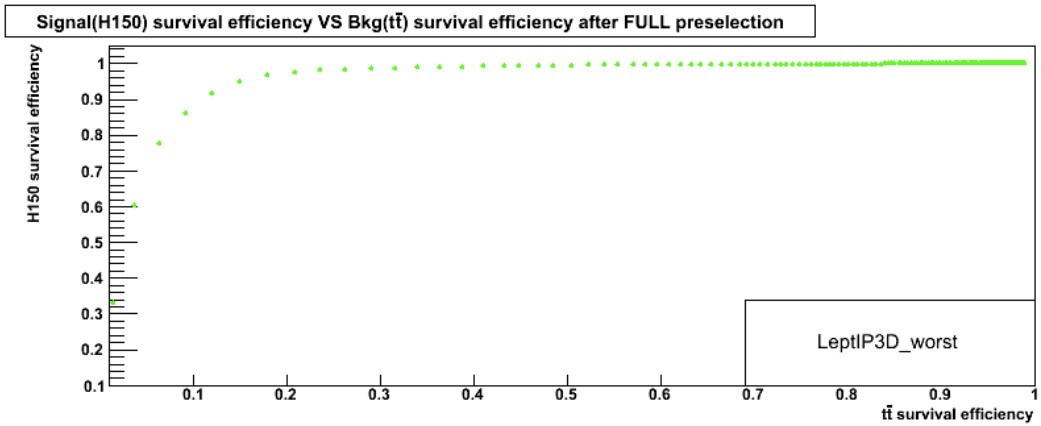
Infine si fanno i cosidetti scatter plot per le varie combinazioni.
Per esempio, per la var1:

```
gr_h150_TT_vtx = new TGraph(nBins1, IntegralCounts1[1], IntegralCounts1[0]);
```

Vettore = riga della Matrice per i=1

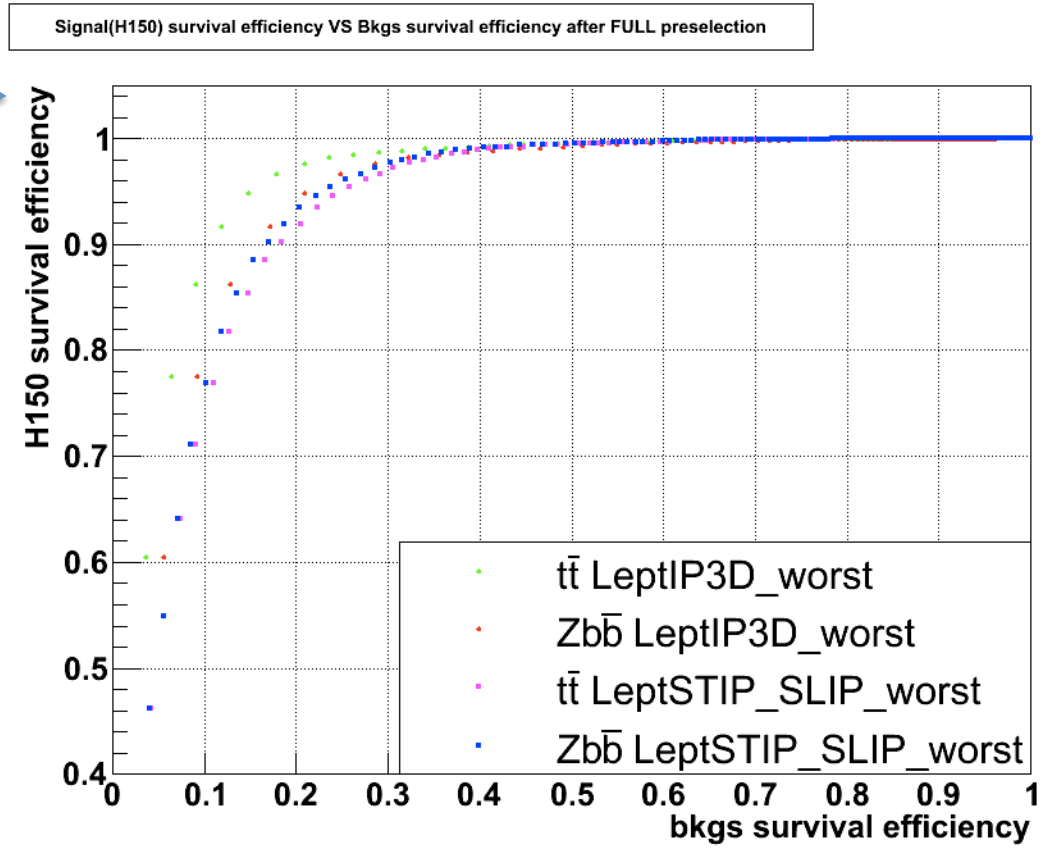
Vettore = riga della Matrice per i=0

… e si ottiene:



Discriminazione Segnale-vs-Fondo - IV

Infine i 4 scatter plot vengono sovrapposti :



Per confrontare i due algoritmi bisogna confrontare:

verde vs rosa (reiezione ttbar)
rosso vs blu (reiezione Zbbar)

Breve conclusione:

1) la statistica var1 e' migliore della var2 ;

2) il vantaggio di var1 e' apprezzabile nella reiezione del fondo ttbar (meno in quella del fondo Zbbar)