${}_s\mathcal{P}lot$: a statistical tool to unfold data distributions

physics/0402083, to be published in *Nucl. Inst. Meth.*

Muriel Pivk
CERN

30th September 2005
ROOT Workshop, CERN

---

1. Motivation
2. ${}_s\mathcal{P}lot$: the tool, its properties
3. Implementation
4. ${}_s\mathcal{P}lot$ at work
5. Conclusion

1 Motivation (1)

Problem to solve when performing an analysis

Data sample $\equiv$ black box

Few signal events and lots of background

$\Longrightarrow$ How to - distinguish them ?

- extract the physics of the signal ?
- probe the validity of analysis ?

$\rightarrow$ check the distributions of events !

The context of BaBar in 2002

First goal: $\sin 2\beta$, *Phys. Rev. Lett.*89:201802 (2002)

- "Golden mode" decay analysis: $B^0 \rightarrow J/\psi K^0_S$
- Low background

$\Longrightarrow$ No need for a particular tool

1 Motivation (2)

Very rare decay analysis $\sin 2\alpha$ possible thanks to luminosity

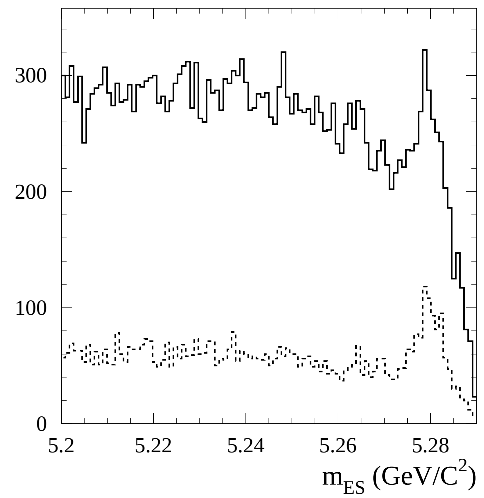
$\Longrightarrow$ Decay channel $B^0 \to h^+h^-$ $(h = \pi, K)$

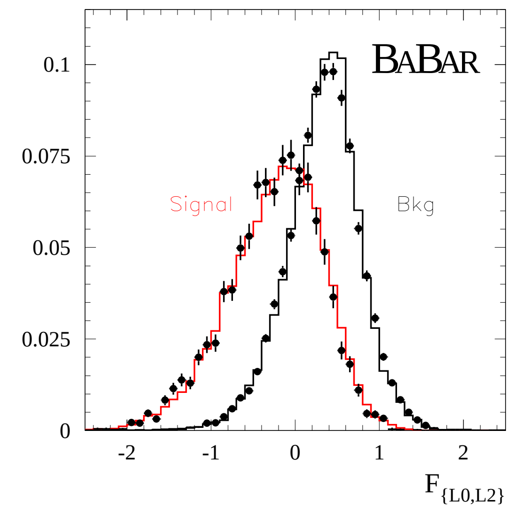
Event selection:

- $m_{\rm ES}$: reconstructed mass of the B candidate
- ΔE : difference of energy between B candidate and $\sqrt{s}/2$

Signal/background discrimination:

- Huge $e^+e^- \to q\bar{q}$ background
- $\mathcal{F}$: Fisher discriminant, uses topology difference of the events





Among 88 million of $B\overline{B}$ pairs

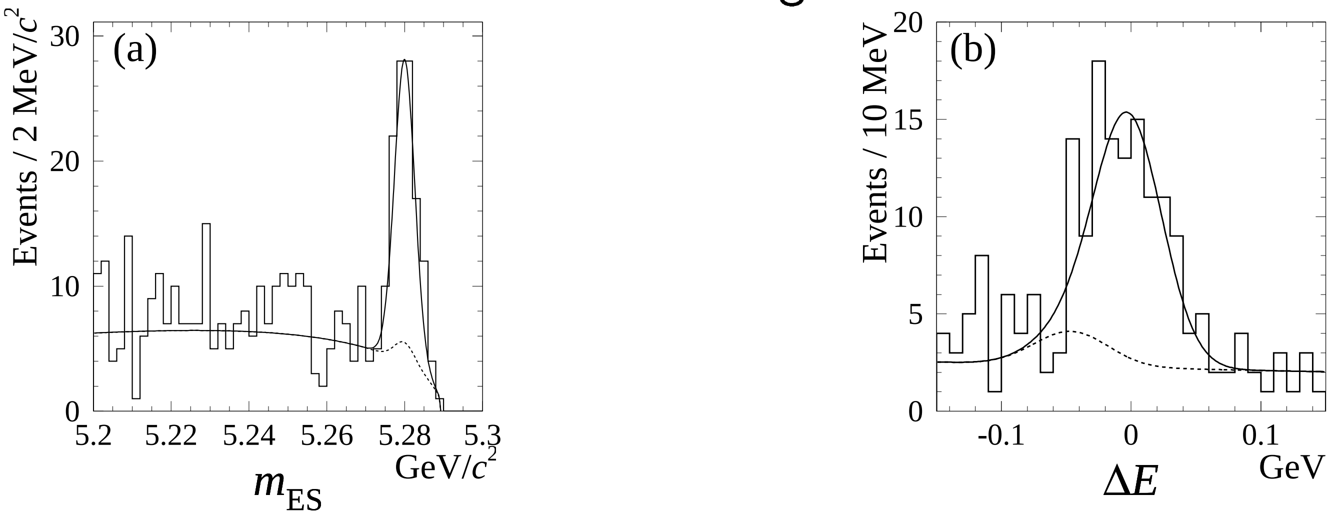
$\Longrightarrow$ 156 $\pi^+\pi^-$ and 588 $K^+\pi^-$ among 26k events

1 Motivation (3)

The question is: how to check the distributions of events ?

Solution ? "Projection plots"

Cut applied on the $\mathcal{L}$ ratio to reduce background



1. subset of sample only
2. signal and background events mixed
3. hard (impossible) if distributions not really different (Fisher ?)

Solution ! ${}_s\mathcal{P}lot$

New tool: firstly meant as projection plots optimization

1. keep all data
2. separate signal and background
3. applicable for ANY variable

2.1 Likelihood analyses

Extended log-likelihood

$$\mathcal{L} = \sum_{e=1}^{N} \ln \left\{ \sum_{i=1}^{\mathrm{N_s}} N_i \mathrm{f}_i(y_e) \right\} - \sum_{i=1}^{\mathrm{N_s}} N_i \tag{1}$$

- N : number of events in the data sample
- e : event number
- $\mathrm{N_s}$: number of species in the data sample
- i : species number (signals, backgrounds)
- y : discriminating variables
- $\mathrm{f}_i(y_e)$: distribution of variables y of species i for event e, normalized to unity

Analysis $B^0 \to h^+h^-$

- $\mathrm{N_s}$: three species
- i : signal $\pi^+\pi^-$ ($N_{\pi\pi}$), signal $K^+\pi^-$ ($N_{K\pi}$), background $q\bar{q}$ ($N_{q\bar{q}}$)
- y : m_{ES}, ΔE, $\mathcal{F}$ $(\ldots)$

2.2 At the beginning where the ${}_{\rm in}\mathcal{P}lot$

Distribution of x for species n, $x \in y$, using the (naive) weight

$$\mathcal{P}_{\rm n}(y_e) = \frac{N_{\rm n}{\rm f}_{\rm n}(y_e)}{\sum_{k=1}^{\rm N_s} N_k{\rm f}_k(y_e)} \quad (2)$$

The reconstructed distribution $\tilde{\rm M}_{\rm n}$ of variable x is defined by:

$$N_{\rm n}\tilde{\rm M}_{\rm n}(x)\delta x \equiv \sum_{e\subset\delta x}^{N} \mathcal{P}_{\rm n}(y_e) \quad (3)$$

Replacing $\sum_{e\subset\delta x}^{N}$ by $\int dy$ (total pdf) $\delta(x(y)-x)\delta x$:

$$\begin{aligned} N_{\rm n}\tilde{\rm M}_{\rm n}(x) &= \int dy \sum_{i=1}^{\rm N_s} N_i{\rm f}_i(y)\delta(x(y)-x)\frac{N_{\rm n}{\rm f}_{\rm n}(y)}{\sum_{k=1}^{\rm N_s} N_k{\rm f}_k(y)} && (4)\\ &= N_{\rm n}\int dy\delta(x(y)-x){\rm f}_{\rm n}(y) && (5)\\ &\equiv N_{\rm n}\mathbf{M}_{\rm n}(x) && (6)\end{aligned}$$

where $\mathbf{M}_{\rm n}(x)$ is the TRUE distribution of variable x for species n

$\Longrightarrow$ Not a clean test:

the Pdf of x is implicitly used to reconstruct itself ... can we avoid it ?

2.3 The ${}_s\mathcal{P}lot$ tool

Distribution of x, $x \notin y$

$$N_{\rm n}\tilde{\rm M}_{\rm n}(x) = \int dy \sum_{i=1}^{\rm N_s} N_i \mathbf{M}_i(x) {\rm f}_i(y) \frac{N_{\rm n}{\rm f}_{\rm n}(y)}{\sum_{k=1}^{\rm N_s} N_k {\rm f}_k(y)} \tag{7}$$

$$= N_{\rm n} \sum_{i=1}^{\rm N_s} \mathbf{M}_i(x) \left(N_i \int dy \frac{{\rm f}_{\rm n}(y){\rm f}_i(y)}{\sum_{k=1}^{\rm N_s} N_k {\rm f}_k(y)} \right) \tag{8}$$

$$\neq N_{\rm n}\mathbf{M}_{\rm n}(x) \tag{9}$$

But but but ... !

Variance matrix:

$$\mathbf{V}^{-1}_{{\rm n}i} = \frac{\partial^2(-\mathcal{L})}{\partial N_{\rm n} \partial N_i} = \sum_{e=1}^{N} \frac{{\rm f}_{\rm n}(y_e){\rm f}_i(y_e)}{(\sum_{k=1}^{\rm N_s} N_k {\rm f}_k(y_e))^2} \tag{10}$$

$$= \int dy \frac{{\rm f}_{\rm n}(y){\rm f}_i(y)}{\sum_{k=1}^{\rm N_s} N_k {\rm f}_k(y)} \tag{11}$$

Eq. (8) becomes $\tilde{\rm M}_{\rm n}(x) = \sum_{i=1}^{\rm N_s} \mathbf{M}_i(x) N_i \mathbf{V}^{-1}_{{\rm n}i}$

$\Longrightarrow$ By inversion:

$$N_{\rm n}\mathbf{M}_{\rm n}(x) = \sum_{i=1}^{\rm N_s} \mathbf{V}_{{\rm n}i} \tilde{\rm M}_i(x) \tag{12}$$

2.4 ${}_s\mathcal{P}lot$: summary

New tool ${}_s\mathcal{P}lot$: weight computed for each event and each species

$\mathrm{N_s}$ species in the sample, discriminating variables y, $\mathrm{f}_i(y)$ their pdfs.

For species n :

$$ {}_s\mathcal{P}_{\mathrm{n}}(y_e) = \frac{\sum_{i=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}i} \mathrm{f}_i(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \qquad (13) $$

with $\mathbf{V}_{\mathrm{n}i}$ the covariance matrix of the fit (number of events)

The TRUE distribution of x ($x \notin y$) is:

$$ N_{\mathrm{n}} \mathbf{M}_{\mathrm{n}}(x) \equiv \sum_{e \subset \delta x} {}_s\mathcal{P}_{\mathrm{n}}(y_e) \qquad (14) $$

NB

- The most discriminating the variables are, the most powerful ${}_s\mathcal{P}lot$ is.
- The variables must be uncorrelated (already necessary with the $\mathcal{L}$).

2.5 Cute properties

Normalization

1. Each x-distribution is properly normalized:

$$\sum_{e=1}^{N} {}_s\mathcal{P}_{\mathrm{n}}(y_e) = N_{\mathrm{n}} \tag{15}$$

2. The contributions ${}_s\mathcal{P}_{\mathrm{n}}(y_e)$ add up to the number of events actually observed in each x-bin. For any event:

$$\sum_{\mathrm{n}=1}^{\mathrm{N_s}} {}_s\mathcal{P}_{\mathrm{n}}(y_e) = 1 \tag{16}$$

Uncertainties

3. For each species:

$$\sum_{e=1}^{N} ({}_s\mathcal{P}_{\mathrm{n}}(y_e))^2 = \sigma^2(N_{\mathrm{n}}) \tag{17}$$

as given by the fit

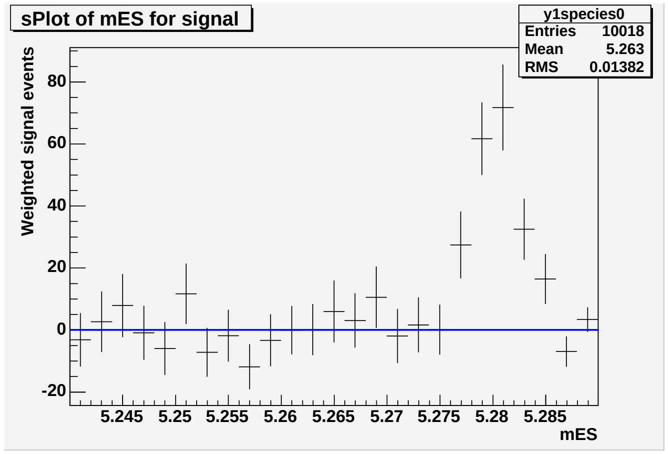
3 Easy implementation

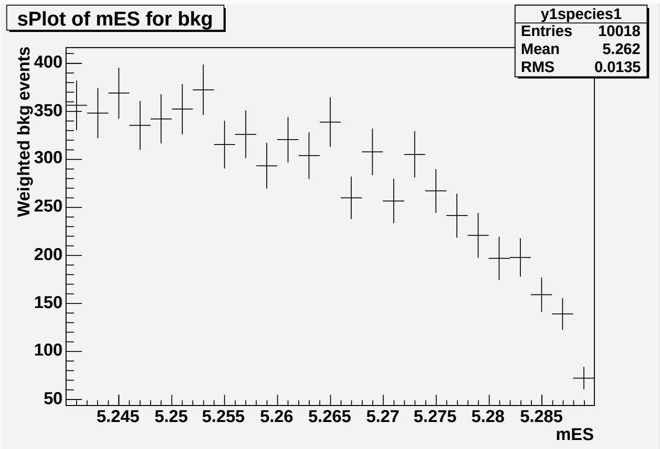
The way to follow

1. Perform the fit to obtain the N_n of each n species present in the data sample without the variable one wants to get the distribution of
2. Compute the sWeights ${}_s\mathcal{P}$ following Eq. 13, using the covariance matrix given by Minuit or computed directly
3. Fill histograms with the value of the variable x weighted with the sWeights ${}_s\mathcal{P}$ for each species present in the data sample

Tool ${}_s\mathcal{P}lot$ in ROOT

Class TSPlot: implemented by Anna Kreshuk, to be released soon

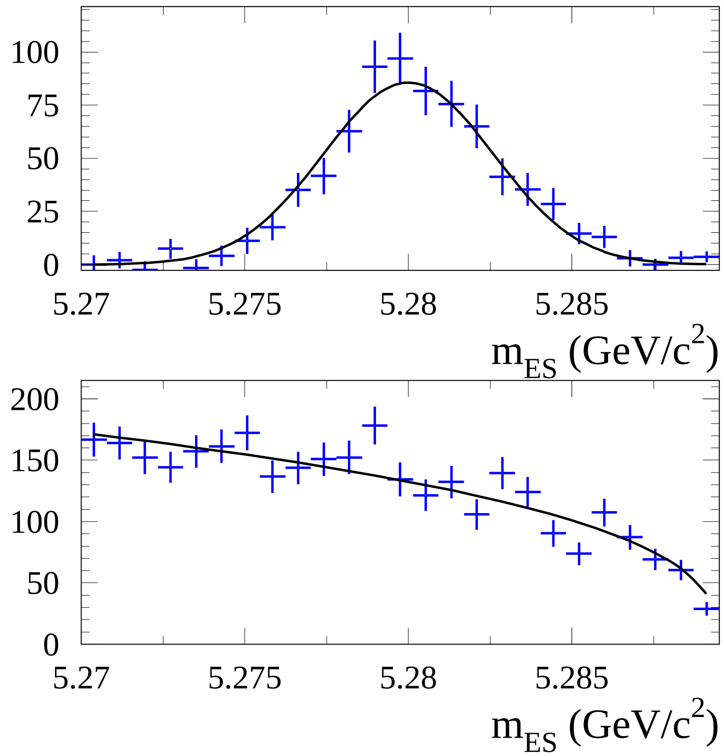


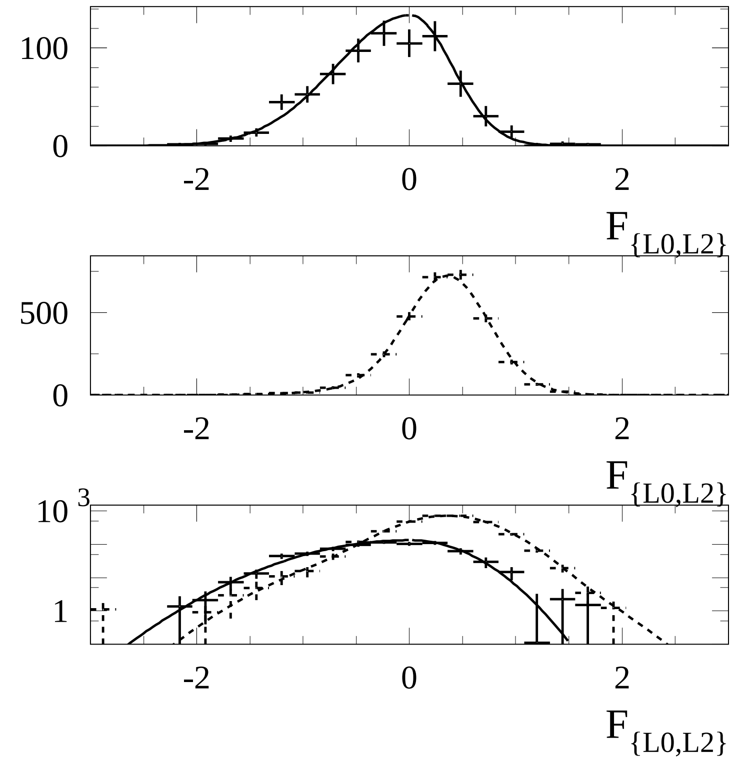


4.1 ${}_s\mathcal{P}lot$ at work: $B^0 \to \pi^+\pi^-$ (1)

BABAR data: ${}_s\mathcal{P}lots$ of $m_{\rm ES}$ and $\mathcal{F}$

Distributions used in the fit are superimposed





- ΔE and $\mathcal{F}$ only
- $m_{\rm ES}$ not in the fit

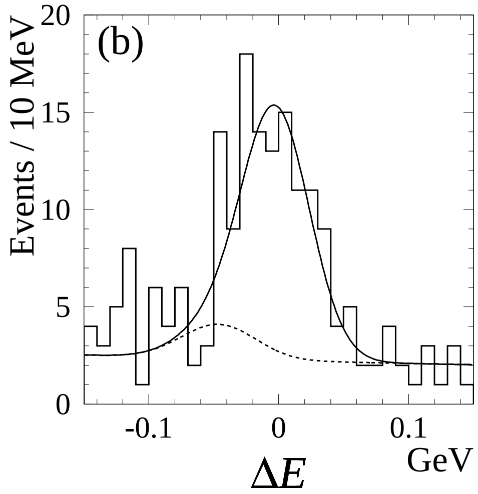
- $m_{\rm ES}$ and ΔE only
- $\mathcal{F}$ not in the fit

$\Longrightarrow$ Very good agreement

$\Longrightarrow$ Optimal tool to validate an analysis ! Still for Fisher !

4.2 $_s\mathcal{P}lot$ at work: $B^0 \to \pi^+\pi^-$ (2)

Comparison with "projection plots"



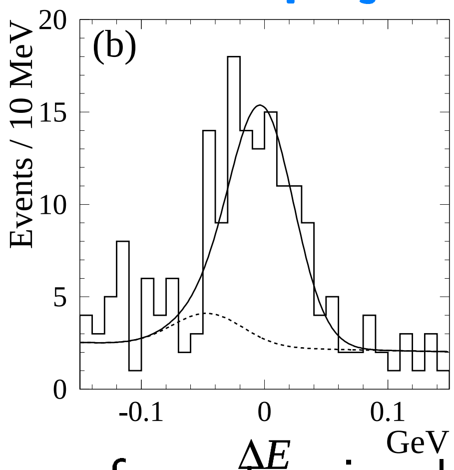
Projection plot :

- Cut on the $\mathcal{L}$ ratio: signal loss and remaining background
- Uncertainties related to signal + background

$\Longrightarrow$ Excess of events: signal ? background ?

4.2 $_s\mathcal{P}lot$ at work: $B^0 \to \pi^+\pi^-$ (2)

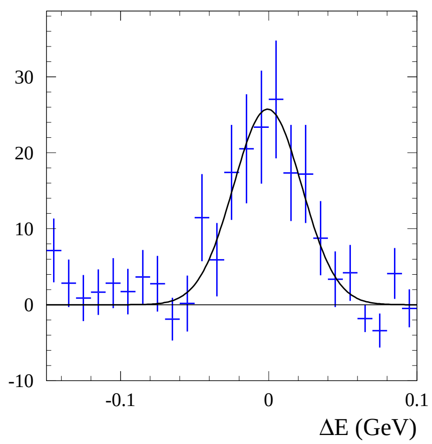
Comparison with "projection plots"



Projection plot :

- Cut on the $\mathcal{L}$ ratio: signal loss and remaining background
- Uncertainties related to signal + background

$\Longrightarrow$ Excess of events: signal ? background ?



$_s\mathcal{P}lot$: Can reveal subtle effects

- No cut applied: keep all the signal events and get rid of all the background ones (statistically)
- Uncertainties related to the signal only

$\Longrightarrow$ Signal ! radiative events ($B^0 \to \pi^+\pi^-\gamma$) ignored in the analysis

$\Longrightarrow$ $\mathcal{B}(B^0 \to h^+h^-)$ under-estimated by about 10% (!!)

Confirmed later for different charmless BABAR analyses (hep-ex/0508046)

4.3 Publications

Only BABAR so far ...

1. *Branching fractions and CP asymmetries in* $B^0 \to K^+K^-K^0_S$ *and* $B^+ \to K^+K^0_SK^0_S$, *Phys. Rev. Lett.*93:181805, 2004
2. *Measurement of neutral* B *decay branching fractions to* $K^0_S\pi^+\pi^-$ *final states*, *Phys. Rev. D*70:091103, 2004
3. *BF and CP asymmetries in* $B^0 \to \pi^0\pi^0$, $B^+ \to \pi^+\pi^0$ *and* $B^+ \to K^+\pi^0$ *decays and isospin analysis of the* $B \to \pi\pi$ *system*, *Phys. Rev. Lett.*94:181802, 2005
4. *Measurement of* CP *asymmetries in* $B^0 \to \phi K^0_S$ *and* $B^0 \to K^+K^-K^0_S$ *decays*, *Phys. Rev. D*71:091102, 2005
5. ...

Observation of direct CP violation in $B^0 \to K^+\pi^-$

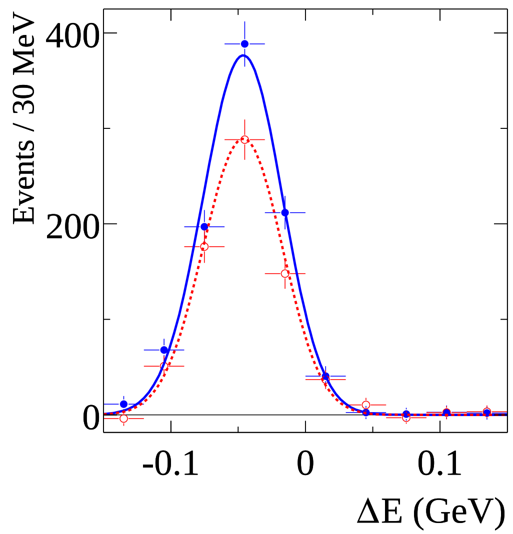
Phys. Rev. Lett..93:131801 (2004)

- $N_{K^+\pi^-} + N_{K^-\pi^+} = 1606 \pm 51$

  $N_{K^+\pi^-} = 910$

  $N_{K^-\pi^+} = 696$

- $A_{K\pi} = -0.133 \pm 0.030 \pm 0.009$



5 Summary and conclusion

New tool $_s\mathcal{P}lot$: optimal for information !

1. Only data involved
2. No bias ($_s\mathcal{P}lot$ted variable not in the fit)
3. Shows signal and background separately
4. Statistical uncertainties
5. Easy to use ! Moreover class TSPlot in ROOT very soon

$\Longrightarrow$ Excellent tool to validate an analysis

Reveal subtle effects : $B^0 \to h^+h^-(\gamma)$

$\Longrightarrow$ Excellent tool to perform an analysis in Dalitz

More in the reference

- Detailed explanations
- Case where species fixed in the fit

Shall be useful beyond B physics

Higgs searches, SUperSYmetry, . . .