$sPlot$: a statistical tool to unfold data distributions

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

Muriel Pivk, François R. Le Diberder
CERN, LAL Orsay

14th September 2005
PHYSTAT 2005, Oxford

1 Motivation
2 $sPlot$: the tool, its properties
3 Implementation
4 $sPlot$ at work
5 Conclusion
Problem to solve when performing an analysis
Data sample ≡ black box

Few signal events and lots of background

⇒ How - distinguish them ?
- extract the physics of the signal ?
- probe the validity of analysis ?
⇒ check the distributions of events !

“Golden mode” decay analysis

\begin{itemize}
  \item \(B^0 \rightarrow J/\psi K_S^0\)
  \item Low background
\end{itemize}

⇒ No need for a particular tool

\textit{Plot:} a statistical tool to unfold data distributions (page 2)
Motivation (2)

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

\[ B^0 \rightarrow h^+ h^- \ (h = \pi, K) \]

Event selection:

- \( m_{ES} \) : reconstructed mass of the \( B \) candidate
- \( \Delta E \) : difference of energy between \( B \) candidate and \( \sqrt{s}/2 \)

Signal/background discrimination:

- Huge \( e^+ e^- \rightarrow q\bar{q} \) background
- \( F \) : Fisher discriminant, uses topology difference of the events

Among 88 million of \( B \bar{B} \) pairs

\[ \rightarrow 156 \ \pi^+\pi^- \ \text{and} \ 588 \ K^+\pi^- \ \text{among 26k events} \]
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! $s$Plot
New tool: firstly meant as projection plots optimization
1. keep all data
2. separate signal and background
3. possible for ANY variable
2.1 Likelihood analyses

Extended log-likelihood

\[ \mathcal{L} = \sum_{e=1}^{N} \ln \left\{ \sum_{i=1}^{N_s} N_i f_i(y_e) \right\} - \sum_{i=1}^{N_s} N_i \]  

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

Analysis \( B^0 \rightarrow h^+h^- \)

- \( N_s \): three species
- \( i \): signal \( \pi^+\pi^- \) \((N_{\pi\pi})\), signal \( K^+\pi^- \) \((N_{K\pi})\), background \( q\overline{q} \) \((N_{q\overline{q}})\)
- \( y \): \( m_{ES}, \Delta E, \mathcal{F}, \Delta t, \ldots \)
2.2 At the beginning where the plot

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

$$P_n(y_e) = \frac{N_n f_n(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)}$$

(2)

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

$$N_n \tilde{M}_n(x) \delta x \equiv \sum_{e \subset \delta x} P_n(y_e)$$

(3)

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

$$N_n \tilde{M}_n(x) = \int dy \sum_{i=1}^{N_s} N_i f_i(y) \delta(x(y) - x) \frac{N_n f_n(y)}{\sum_{k=1}^{N_s} N_k f_k(y)}$$

(4)

$$= N_n \int dy \delta(x(y) - x) f_n(y)$$

(5)

$$\equiv N_n M_n(x)$$

(6)

where $M_n(x)$ is the TRUE distribution of variable $x$ for species $n$

$\Rightarrow$ Biased if $x \in y$ ... can we avoid it?
2.3 The \textit{sPlot} tool

Distribution of $x, x \notin y$

\[ N_n \tilde{M}_n(x) = \int dy \sum_{i=1}^{N_s} N_i M_i(x) f_i(y) \frac{N_n f_n(y)}{\sum_{k=1}^{N_s} N_k f_k(y)} \]  

\[ = N_n \sum_{i=1}^{N_s} M_i(x) \left( N_i \int dy \frac{f_n(y) f_i(y)}{\sum_{k=1}^{N_s} N_k f_k(y)} \right) \]  

\[ \neq N_n M_n(x) \]

But but but ... !

Variance matrix:

\[ v_{ni}^{-1} = \frac{\partial^2 (-L)}{\partial N_n \partial N_i} = \sum_{e=1}^{N} \frac{f_n(y_e) f_i(y_e)}{(\sum_{k=1}^{N_s} N_k f_k(y_e))^2} \]

\[ = \int dy \frac{f_n(y) f_i(y)}{\sum_{k=1}^{N_s} N_k f_k(y)} \]

Eq. (8) becomes $\tilde{M}_n(x) = \sum_{i=1}^{N_s} M_i(x) N_i v_{ni}^{-1}$

$ \implies $ By inversion:

\[ N_n M_n(x) = \sum_{i=1}^{N_s} v_{ni} \tilde{M}_i(x) \]
New tool $\mathcal{S}P\text{lot}$: weight computed for each event and each species
$N_s$ species in the sample, discriminating variables $y$, $f_i(y)$ their pdfs.

For species $n$:

$$\mathcal{S}P_n(y_e) = \frac{\sum_{i=1}^{N_s} V_{ni} f_i(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)}$$ (13)

with $V_{ni}$ the covariance matrix of the fit (number of events)

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

$$N_n M_n(x) \equiv \sum_{e \in \delta x} \mathcal{S}P_n(y_e)$$ (14)

NB
The most discriminating the variables are, the most powerful $\mathcal{S}P\text{lot}$ is.
2.5 Cute properties

Normalization
1. Each $x$-distribution is properly normalized:

$$\sum_{e=1}^{N} sP_n(y_e) = N_n \quad (15)$$

2. The contributions $sP_n(y_e)$ add up to the number of events actually observed in each $x$-bin. For any event:

$$\sum_{n=1}^{N_s} sP_n(y_e) = 1 \quad (16)$$

Uncertainties
3. In each bin, for each species:

$$\sum_{e=1}^{N} (sP_n(y_e))^2 = \sigma^2(N_n) \quad (17)$$

as given by the fit
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 $sP$ 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 $sP$ for each species present in the data sample

Tool $sPlot$ in ROOT

Class TSplot: implemented by Anna Kreshuk, to be released soon
### 4.1 Illustration with simulated $B^0 \rightarrow \pi^+\pi^-$

Distributions of variables $m_{ES}$, $\Delta E$ et $\mathcal{F}$

Two species: signal $\pi^+\pi^-$, background $q\bar{q}$

---

**Plots of** $m_{ES}$, $\Delta E$ et $\mathcal{F}$

- $\Delta E$ and $\mathcal{F}$ only
- $m_{ES}$ not in the fit
- $m_{ES}$ and $\mathcal{F}$ only
- $\Delta E$ not in the fit
- $m_{ES}$ and $\Delta E$ only
- $\mathcal{F}$ not in the fit
4.2 $sPlot$ at work: real data (1)

**Data: $sPlots$ of $m_{ES}$ and $F$**

Distributions used in the fit are superimposed

- $\Delta E$ and $F$ only
- $m_{ES}$ not in the fit

$\Rightarrow$ Very good agreement

$\Rightarrow$ Optimal tool to validate an analysis ! Still for Fisher !
Comparison with “projection plots”

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

$\Rightarrow$ Excess of events: signal ? background ?
**4.3 sPlot at work: real data (2)**

**Comparison with “projection plots”**

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

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

$s$Plot: 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

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

$\Rightarrow B(B^0 \rightarrow h^+ h^-)$ under-estimated by about 10% (!!!)

Confirmed later for different charmless $\textit{BaBar}$ analyses
4.4 Publications

Only BaBar so far ...

1. Branching fractions and CP asymmetries in $B^0 \rightarrow K^+K^-K^0_S$ and $B^+ \rightarrow K^+K^0_SK^0_S$, Phys. Rev. Lett.93:181805, 2004
2. Measurement of neutral B decay branching fractions to $K^0_S\pi^+\pi^-$ final states, Phys. Rev. D70:091103, 2004
3. BF and CP asymmetries in $B^0 \rightarrow \pi^0\pi^0$, $B^+ \rightarrow \pi^+\pi^0$ and $B^+ \rightarrow K^+\pi^0$ decays and isospin analysis of the $B \rightarrow \pi\pi$ system, Phys. Rev. Lett.94:181802, 2005
4. Measurement of CP asymmetries in $B^0 \rightarrow \phi K^0_S$ and $B^0 \rightarrow K^+K^-K^0_S$ decays, Phys. Rev. D71:091102, 2005
5. ...

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


- $N_{K^+\pi^-} + N_{K^-\pi^+} = 1606 \pm 51$
  - $N_{K^+\pi^-} = 910$
  - $N_{K^-\pi^+} = 696$

- $A_{K\pi} = -0.133 \pm 0.030 \pm 0.009$
New tool $sPlot$: optimal for information!

1. Only data involved
2. No bias ($sPlotted$ variable not in the fit)
3. Shows signal and background separately
4. Statistical uncertainties
5. Easy to use!

$\implies$ Excellent tool to validate an analysis
Reveal subtle effects: $B^0 \rightarrow h^+ h^- (\gamma)$

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

More in the reference

• Case where species fixed in the fit

Shall be useful beyond $B$ physics
Higgs searches, SUperSYmetry, …