$_s$Plot

M. Pivk & F. R. L D

Outline

1) The context
2) How to use ${}_s\mathcal{P}lot$
3) General proof
4) (properties)
5) Conclusion

The context

$\mathfrak{E}$: an experiment

that is to say a set of N events, labelled $e = 1, ..., N$

the events are statistically independent

each event can be summarized by a random (multi-dimensional) variable y_e

The statistical distribution (PDF) of y is assumed to be known up to a (multi-dimensional) parameters t

The parameters t are estimated from the experiment $\mathfrak{E}$ by means of the maximum Likelihood method

$$\mathrm{L}_{\mathfrak{E}}(t) = \Pi_{e=1}^{N} \; f_t(y_e)$$

$$\int f_t(y)\mathrm{d}y = 1 \; \boxed{\forall t}$$

If the total yields (N) depends on t, one should rather use the "extended" Likelihood method:

$$\mathrm{L}_{\mathfrak{E}}(t) = \left(\Pi_{e=1}^{N}\ f_t(y_e)\right) \times \frac{\mu_t^N}{N!} e^{-\mu_t}$$

where μ_t is the expected yield, given t.

$$\int f_t(y)\mathrm{d}y = 1\ \forall t$$

Technically, one rather uses the (extended)LogLikelihood

$$\mathcal{L}_{\mathfrak{E}}(t) = \sum_{e=1}^{N} \ln f_t(y_e) \;+\; N \ln \mu_t \;-\; \mu_t \;-\ln N!$$

Since $\ln N!$ does not depend on t, it is an irrelevant constant that can be ignored.

One also notices that $N \ln \mu_t = \sum_{e=1}^{N} \ln \mu_t$, and thus:

$$\boxed{\mathcal{L}_{\mathfrak{E}}(t) = \sum_{e=1}^{N} \ln(\mu_t f_t(y_e)) \;-\; \mu_t}$$

The ${}_s\mathcal{P}lot$ technique applies when f_t results from a sum of N_s "species":

$\mu_t f_t(y) = \sum_{i=1}^{N_s} N_i f_{(i)}(y)$ of N_s (known) contributions $f_{(i)}(y)$

each normalized to unity:

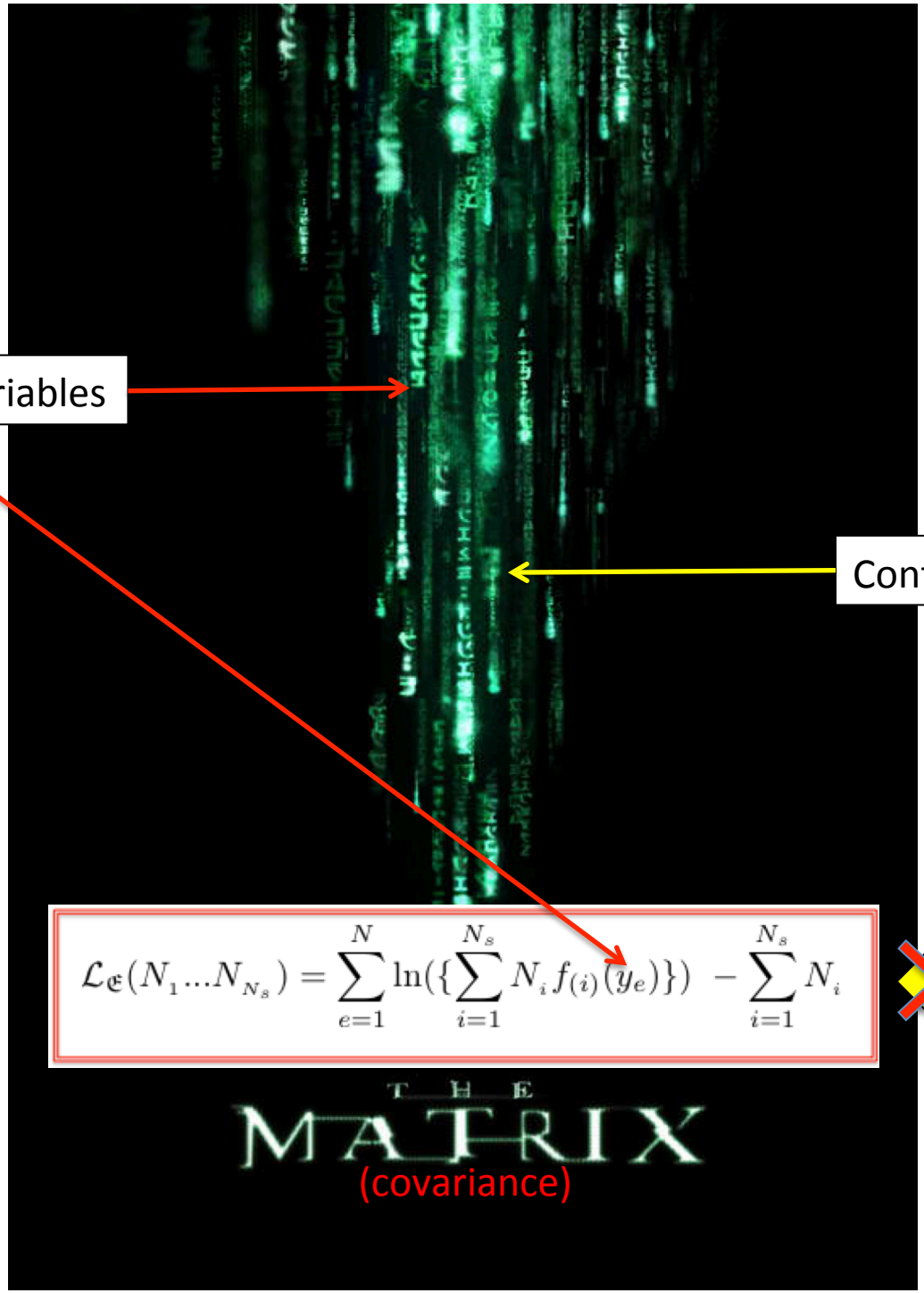
$$\int f_{(i)}(y)\mathrm{d}y = 1 \ \forall i$$

The normalization of $f_t(y)$ implies that $\mu_t = \sum_{i=1}^{N_s} N_i$.

Therefore:

$$\mathcal{L}_{\mathfrak{E}}(t) = \sum_{e=1}^{N} \ln(\mu_t f_t(y_e)) \ - \ \mu_t$$

$$\boxed{\mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s}) = \sum_{e=1}^{N} \ln(\{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\}) \ - \sum_{i=1}^{N_s} N_i}$$



Discriminating variables

y

Control variables

x

$$\mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s}) = \sum_{e=1}^{N} \ln(\{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\}) - \sum_{i=1}^{N_s} N_i$$

Do not enter in any way in $\mathcal{L}_{\mathfrak{E}}$

(covariance)

$x\&y$

Are correlated in the sense that they come together for each event

Both variable can be multi-dimensional, with no complication involved

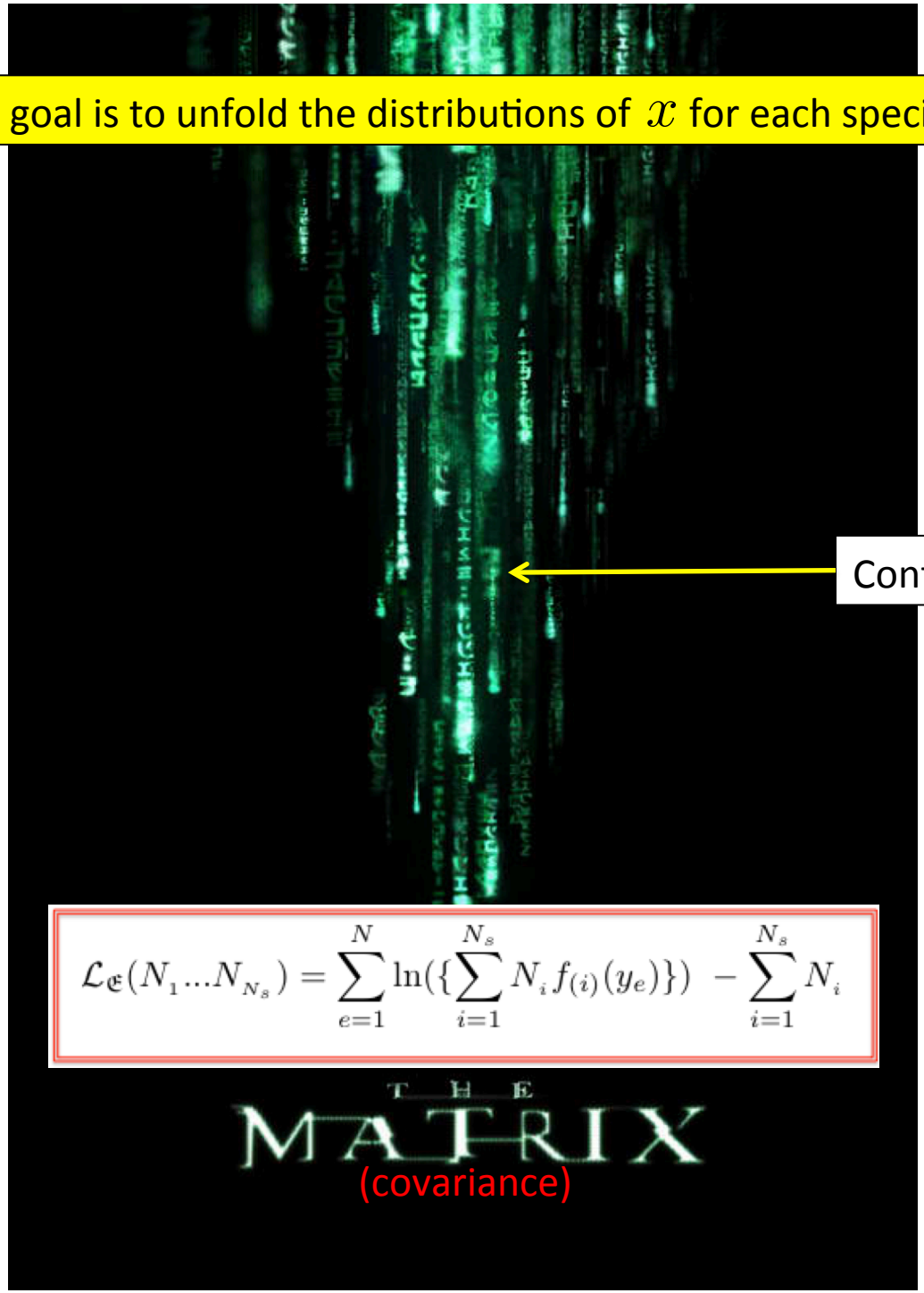
y_1 y_2 x_1 x_2

$$f_{(i)}(y_1, y_2) \times g_{(i)}(x_1, x_2)$$

for a given species

Essential for the method

The goal is to unfold the distributions of x for each species.



Control variables

x

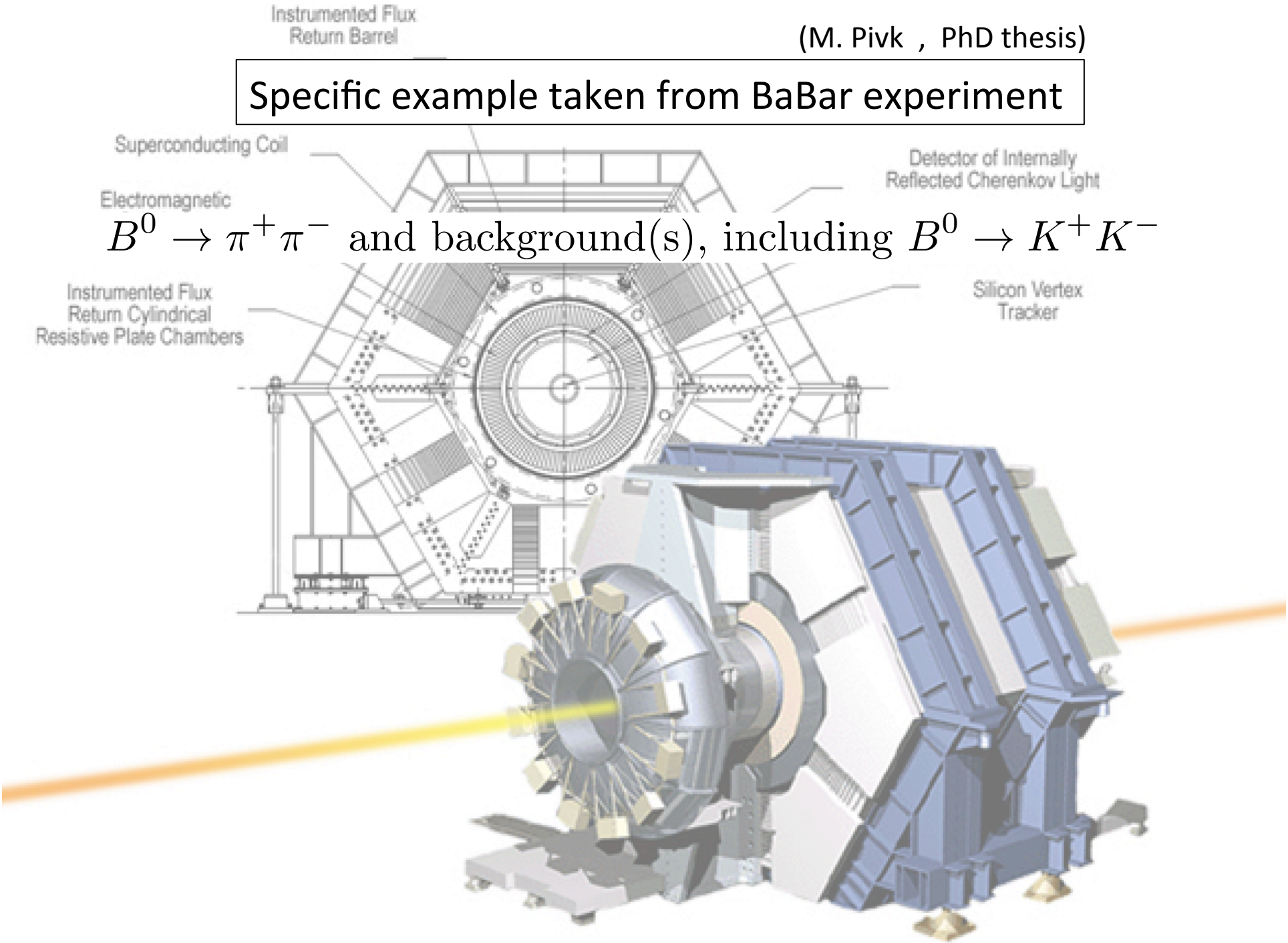
$$\mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s}) = \sum_{e=1}^{N} \ln(\{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\}) - \sum_{i=1}^{N_s} N_i$$

(covariance)

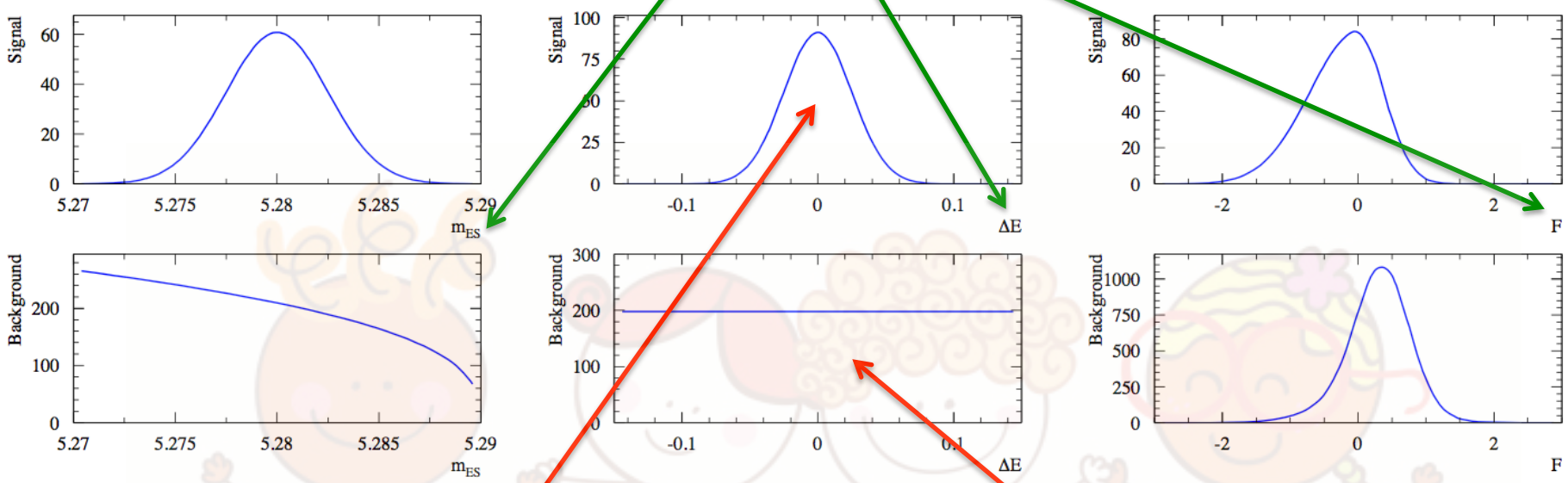
(M. Pivk , PhD thesis)

Specific example taken from BaBar experiment

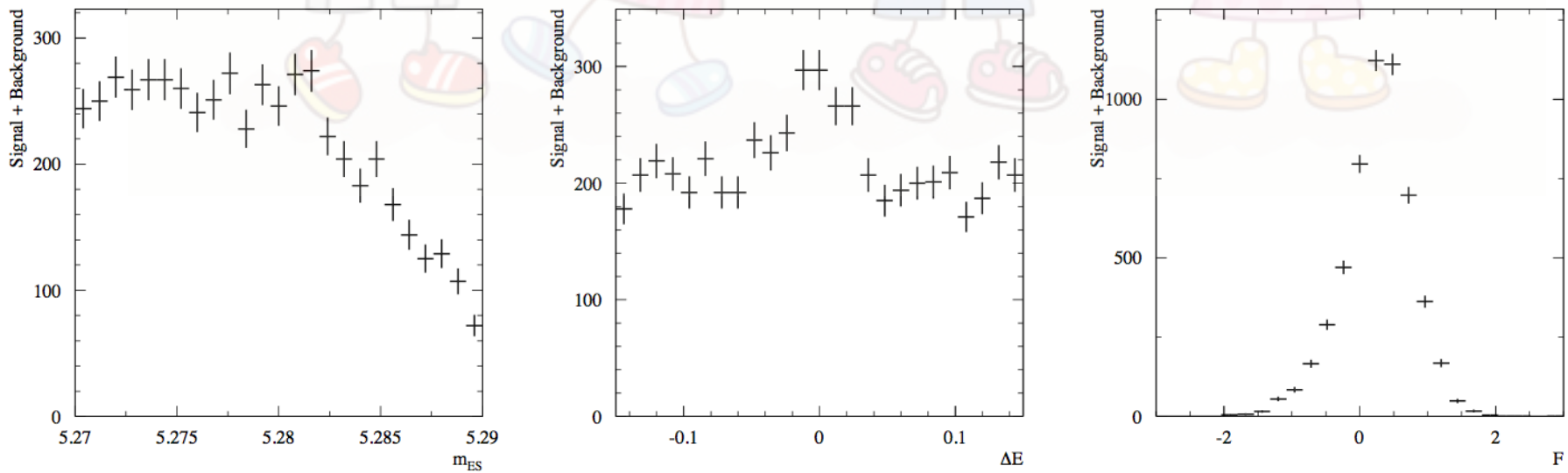
$$B^0 \to \pi^+\pi^- \text{ and background(s), including } B^0 \to K^+K^-$$

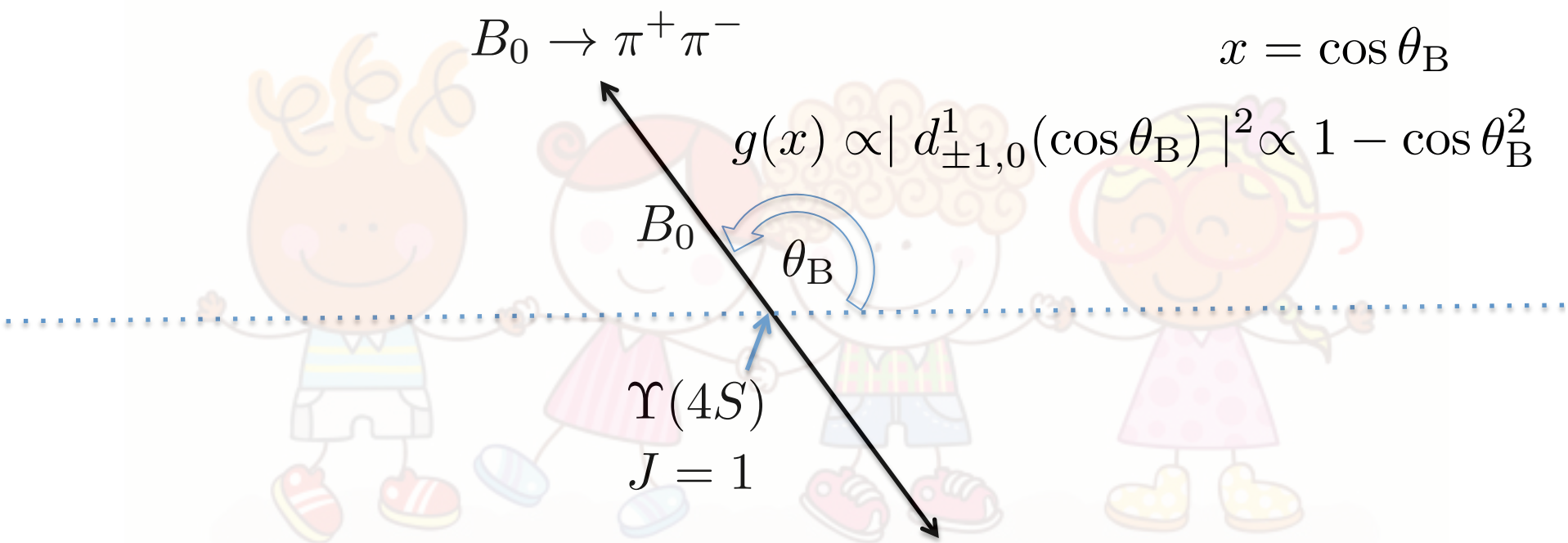


$$y = \{m_{\mathrm{ES}}, \Delta\mathrm{E}, F\}$$



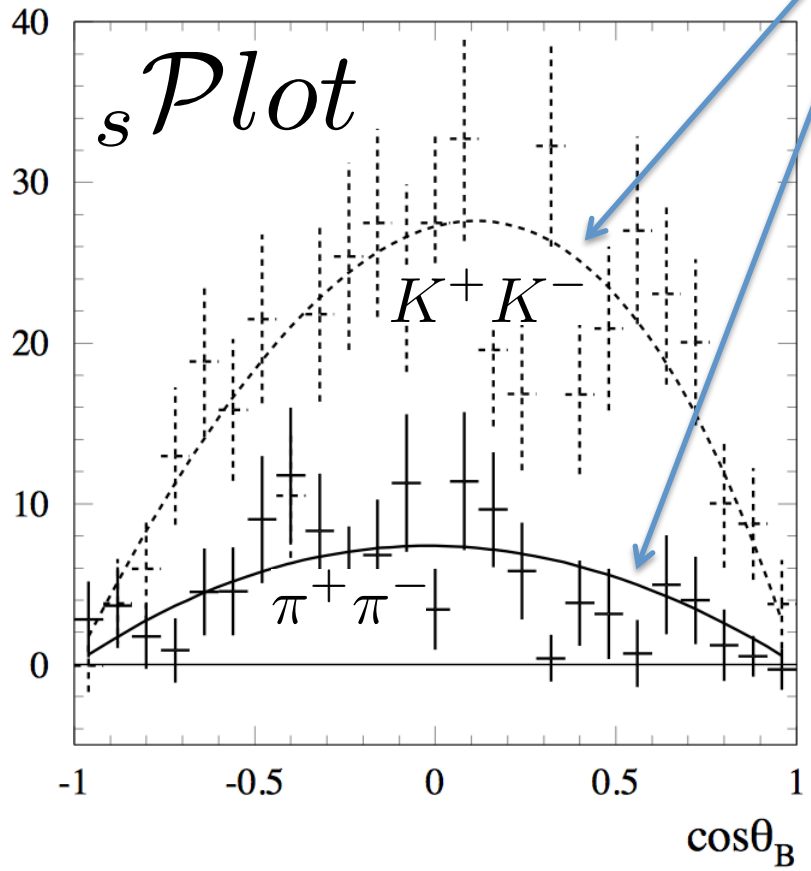
$N_s = 2$ i=1 (for Signal) ; i=2 (for Background)

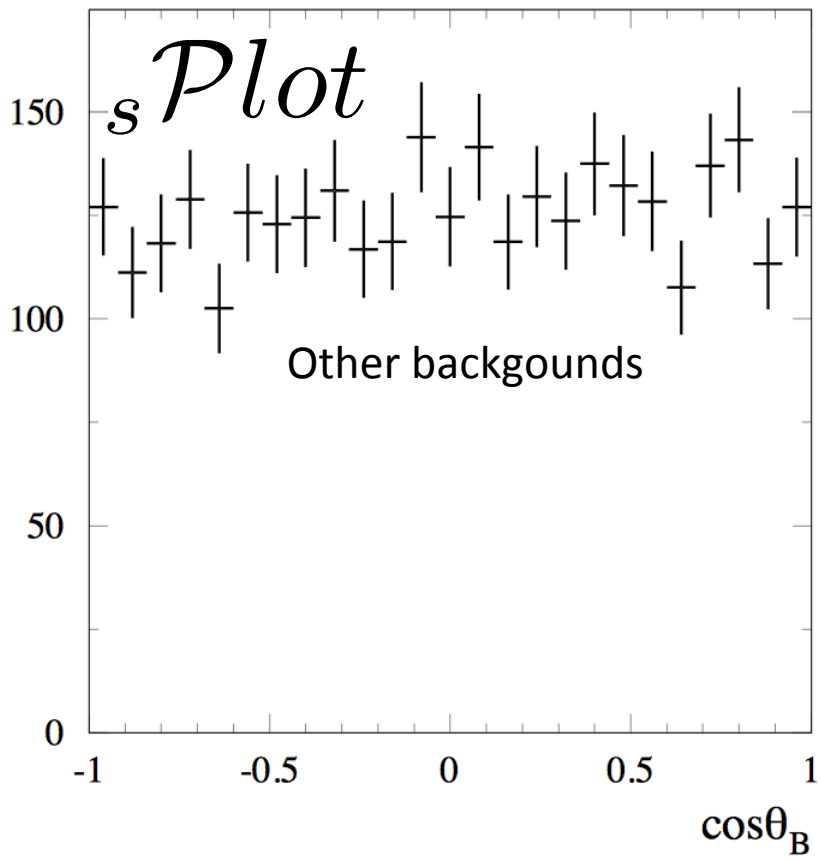



$B_0 \to \pi^+\pi^-$
$x = \cos\theta_{\rm B}$
$g(x) \propto | d^1_{\pm 1,0}(\cos\theta_{\rm B}) |^2 \propto 1 - \cos\theta_{\rm B}^2$
B_0
$\theta_{\rm B}$
$\Upsilon(4S)$
$J = 1$

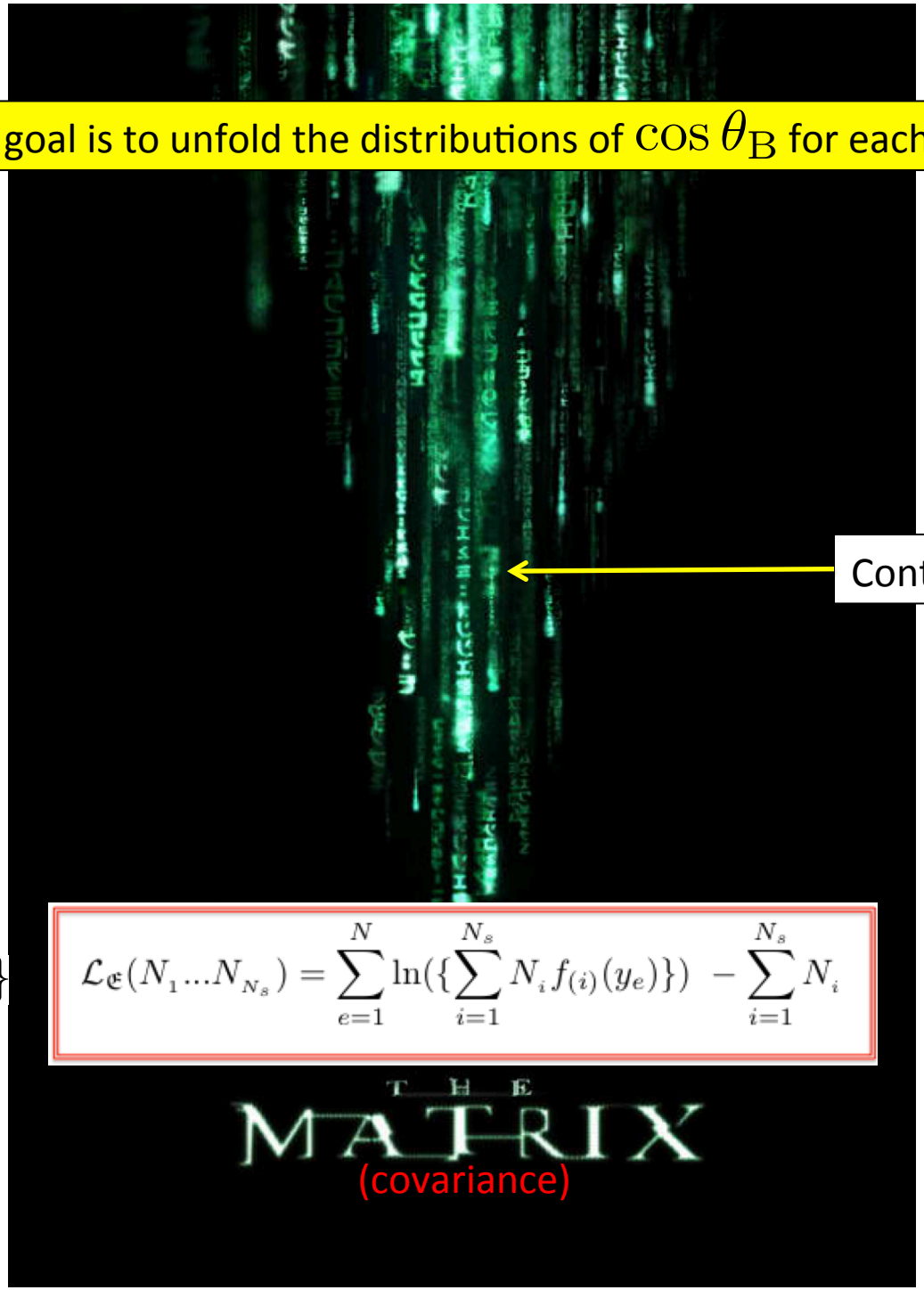
$$g_{\pi^+\pi^-}(\cos\theta_B);\ \ g_{K^+K^-}(\cos\theta_B);\ \ g_{...}(\cos\theta_B)$$

Expected PDFs of $\cos\theta_B$ **NOT** used in the process,
They are just represented to show the agreement.





The goal is to unfold the distributions of $\cos\theta_{\mathrm{B}}$ for each species.



Control variables

$$x = \cos\theta_{\mathrm{B}}$$

$$y = \{m_{\mathrm{ES}}, \Delta\mathrm{E}, F\}$$

$$\mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s}) = \sum_{e=1}^{N} \ln(\{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\}) - \sum_{i=1}^{N_s} N_i$$

(covariance)

How to use ${}_s\mathcal{P}lot$

$$\mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s}) = \sum_{e=1}^{N} \ln(\{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\}) - \sum_{i=1}^{N_s} N_i$$

Maximizing $\mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s})$ provides:

$N_{\pi^+\pi^-}, N_{K^+K^-}, \ldots$

and the covariance matrix V_{ij} between the above N_i estimates

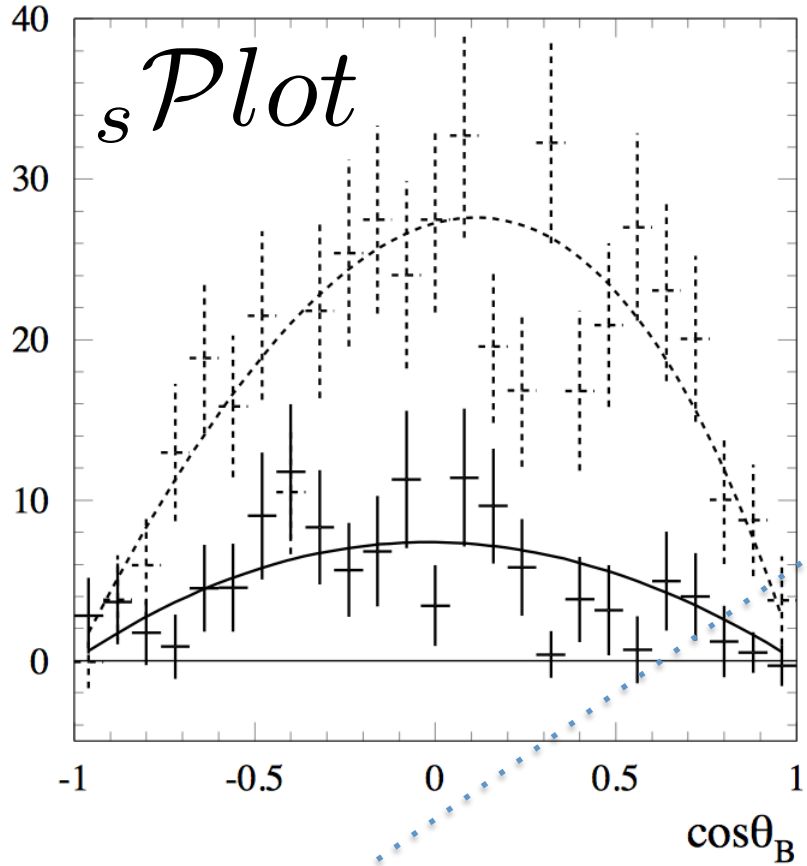
For each event e compute the N_s sWeights ${}_s\mathcal{P}_{(i)}(y_e)$

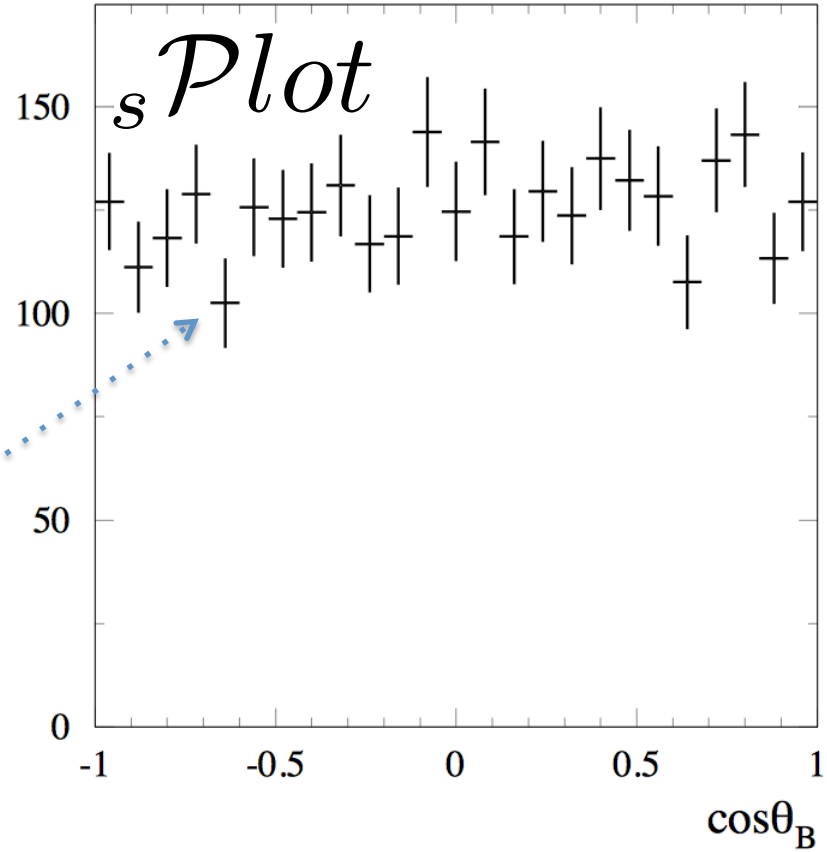
$${}_s\mathcal{P}_{(i)}(y_e) = \frac{\sum_{j=1}^{N_s} V_{ij} f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)}$$

$$ {}_s\mathcal{P}_{(i)}(y_e) = \frac{\sum_{j=1}^{N_s} V_{ij} f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} $$

Fill N_s histograms of the x variable

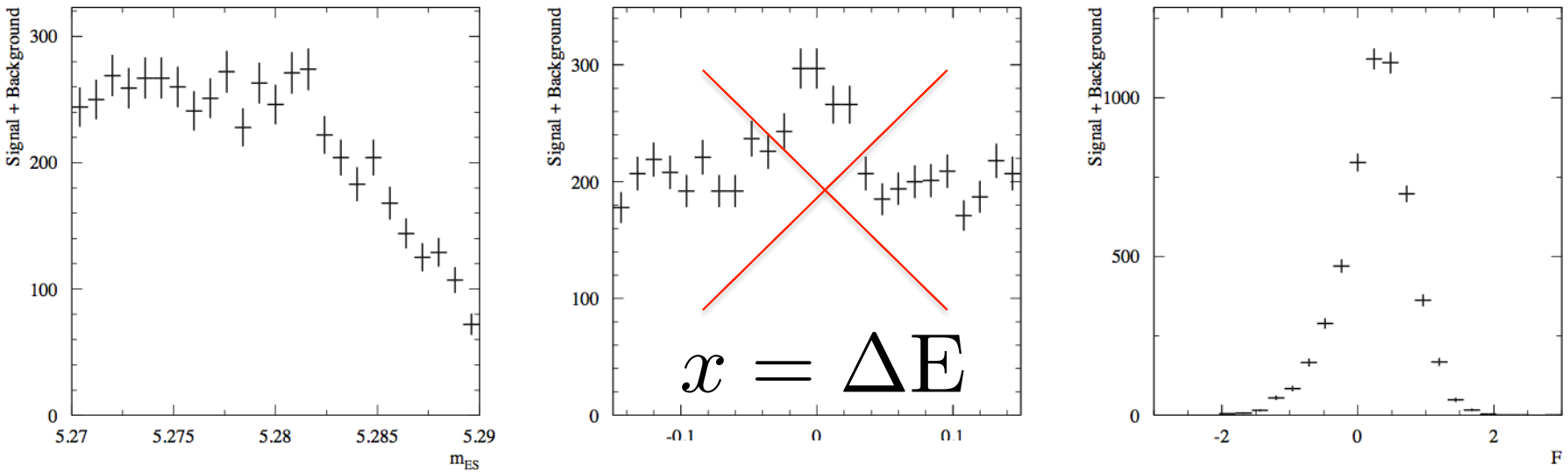
using all events, each weighted by its ith sWeights.

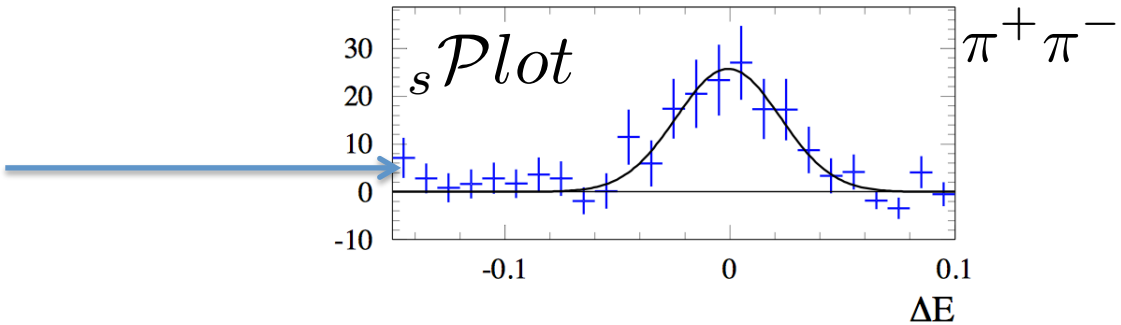


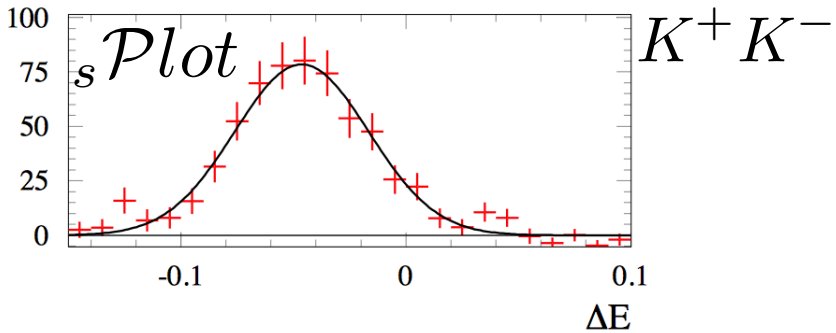


The error bars are given by the quadratic sum of the sWeights.

$$y = \{m_{\mathrm{ES}}, \cancel{\Delta \mathrm{E}}, F\}$$







Proof

Generalities


(in fact, no big deal at all, only patience is needed.)

$$\mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s}) = \sum_{e=1}^{N} \ln(\{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\}) \; - \sum_{i=1}^{N_s} N_i$$

$$0 = \frac{\partial \mathcal{L}_{\mathfrak{E}}}{\partial N_j} = \sum_{e=1}^{N} \frac{f_{(j)}(y_e)}{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)} - 1$$

$$\boxed{\sum_{e=1}^{N} \frac{f_{(j)}(y_e)}{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)} = 1 \quad \forall j}$$

(non-linear set of equations determining the yields)

$$\boxed{\sum_{e=1}^{N} \frac{f_{(j)}(y_e)}{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)} = 1 \quad \forall j}$$

$$\sum_{k=1}^{N_s} N_k = \sum_{k=1}^{N_s} N_k \left(\sum_{e=1}^{N} \frac{f_{(k)}(y_e)}{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)} \right)$$

$$= \sum_{e=1}^{N} \sum_{k=1}^{N_s} N_k \left(\frac{f_{(k)}(y_e)}{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)} \right)$$

$$= \sum_{e=1}^{N} \left(\frac{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)}{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)} \right)_{=1}$$

$$= N$$

$$
\begin{aligned}
\sum_{i=1}^{N_s} N_i V_{ij}^{(-1)} &= \sum_{i=1}^{N_s} N_i \sum_{e=1}^{N} \frac{f_{(i)}(y_e) f_{(j)}(y_e)}{\left(\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)\right)^2} \\
&= \sum_{e=1}^{N} \frac{\left(\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\right) f_{(j)}(y_e)}{\left(\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)\right)^2} \\
&= \sum_{e=1}^{N} \frac{f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} = 1
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{N_s} V_{ij} &= \sum_{i=1}^{N_s} V_{ij} \left(\sum_{k=1}^{N_s} N_k V_{ki}^{(-1)} \right) \\
&= \sum_{k=1}^{N_s} N_k \sum_{i=1}^{N_s} V_{ij} V_{ki}^{(-1)} = N_j
\end{aligned}
$$

THE MATRIX (covariance)

$$\frac{\partial \mathcal{L}_{\mathfrak{E}}}{\partial N_j} = \sum_{e=1}^{N} \frac{f_{(j)}(y_e)}{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)} - 1$$

$$V_{ij}^{(-1)} = \frac{\partial^2(-\mathcal{L}_{\mathfrak{E}})}{\partial N_i \partial N_j} = \sum_{e=1}^{N} \frac{f_{(i)}(y_e) f_{(j)}(y_e)}{(\sum_{k=1}^{N_s} N_k f_{(k)}(y_e))^2}$$

(one does not need Minuit to compute the inverse of the covariance matrix)

But one should invert the above to reach the covariance matrix

$$V_{ij} = (V_{ij}^{(-1)})^{(-1)}$$

(and Minuit does it for free)

Asymptotic property of the covariance matrix

$$\langle \sum_{e=1}^{N} [y_e] \rangle = \int \mathrm{d}y \left(\sum_{l=1}^{N_s} \bar{N}_k f_{(k)}(y) \right) [y]$$

$$\begin{aligned}
\bar{V}_{ij}^{(-1)} &= \langle \sum_{e=1}^{N} \frac{f_{(i)}(y_e) f_{(j)}(y_e)}{(\sum_{k=1}^{N_s} N_k f_{(k)}(y_e))^2} \rangle \\
&= \int \mathrm{d}y \left(\sum_{l=1}^{N_s} \bar{N}_k f_{(k)}(y) \right) \frac{f_{(i)}(y) f_{(j)}(y)}{(\sum_{k=1}^{N_s} \bar{N}_k f_{(k)}(y))^2} \\
&= \int \mathrm{d}y \frac{f_{(i)}(y) f_{(j)}(y)}{\sum_{k=1}^{N_s} \bar{N}_k f_{(k)}(y)}
\end{aligned}$$

$$_{s}\mathcal{P}lot$$

Lets consider the sum of the sWeights(i) of events which x values are within a given bin in x:

$$ {}_s\mathcal{P}lot_{(i)}({}_{\mathrm{bin}}x) = \sum_{e=1;x\in_{\mathrm{bin}}x}^{N} {}_s\mathcal{P}_{(i)}(y_e) $$

We want to establish that on the average ${}_s\mathcal{P}lot({}_{\mathrm{bin}}x)$ is identical to the expected number of events from species (i) in the ${}_{\mathrm{bin}}x$.

$$ \langle {}_s\mathcal{P}lot_{(i)}({}_{\mathrm{bin}}x)\rangle = \Delta x \ \bar{N}_i \ g_{(i)}(x) $$

$$
\begin{aligned}
\langle {}_s\mathcal{P}lot_{(i)}({}_{\mathrm{bin}}x)\rangle &= \langle \sum_{e=1;x_e\in{}_{\mathrm{bin}}x}^{N} {}_s\mathcal{P}_{(i)}(y_e)\rangle \\
&= \Delta x \int \mathrm{d}y \left(\sum_{k=1}^{N_s} \bar{N}_k \left[f_{(k)}(y) g_{(k)}(x)\right]\right) {}_s\mathcal{P}_{(i)}(y) \\
&= \Delta x \int \mathrm{d}y \left(\sum_{k=1}^{N_s} \bar{N}_k \left[f_{(k)}(y) g_{(k)}(x)\right]\right) \frac{\sum_{j=1}^{N_s} \bar{V}_{ij} f_{(j)}(y)}{\sum_{l=1}^{N_s} \bar{N}_l f_{(l)}(y)} \\
&= \Delta x \sum_{k=1}^{N_s} \bar{N}_k g_{(k)}(x) \int \mathrm{d}y f_{(k)}(y) \frac{\sum_{j=1}^{N_s} \bar{V}_{ij} f_{(j)}(y)}{\sum_{l=1}^{N_s} \bar{N}_l f_{(l)}(y)} \\
&= \Delta x \sum_{k=1}^{N_s} \bar{N}_k g_{(k)}(x) \sum_{j=1}^{N_s} \bar{V}_{ij} \int \mathrm{d}y \frac{f_{(k)}(y) f_{(j)}(y)}{\sum_{l=1}^{N_s} \bar{N}_l f_{(l)}(y)} \\
&= \Delta x \sum_{k=1}^{N_s} \bar{N}_k g_{(k)}(x) \sum_{j=1}^{N_s} \bar{V}_{ij} \bar{V}_{jk}^{(-1)} = \Delta x \sum_{k=1}^{N_s} \bar{N}_k g_{(k)}(x) \delta_{ik} \\
&= \Delta x \bar{N}_i g_{(i)}(x)
\end{aligned}
$$

(properties)

Property (1)

$$
\begin{aligned}
\sum_{\mathrm{bin}\,x} {}_s\mathcal{P}lot_{(i)}({}_{\mathrm{bin}}x) &= \sum_{e=1}^{N} {}_s\mathcal{P}lot_{(i)}(y_e) \\
&= \sum_{e=1}^{N} \frac{\sum_{j=1}^{N_s} V_{ij} f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} \\
&= \sum_{j=1}^{N_s} V_{ij} \sum_{e=1}^{N} \frac{f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} \\
&= \sum_{j=1}^{N_s} V_{ij} \\
&= N_i
\end{aligned}
$$

Property (2)

$$\sum_{i=1}^{N_s} {}_s\mathcal{P}lot_{(i)}(y_e) = \sum_{i=1}^{N_s} \frac{\sum_{j=1}^{N_s} V_{ij} f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)}$$

$$= \frac{\sum_{j=1}^{N_s} (\sum_{i=1}^{N_s} V_{ij}) f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)}$$

$$= \frac{\sum_{j=1}^{N_s} N_j f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)}$$

$$= 1$$

The sWeights are (kind of) probabilities for an event to stem from the various species.
However:

$${}_s\mathcal{P}lot_{(i)}(y_e) \notin [0,1]$$

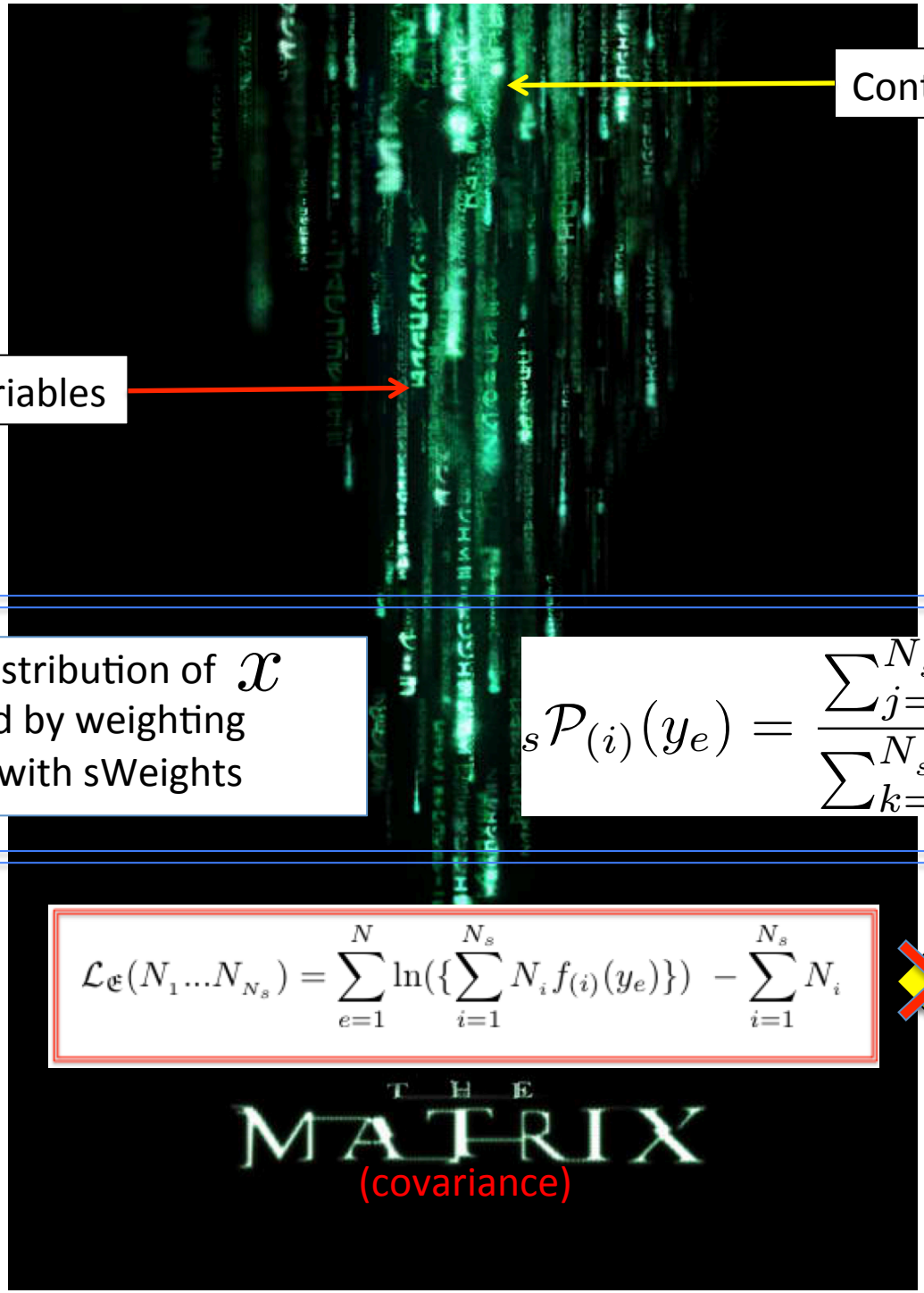
They tend to be negative for “background” and greater than one for “signal”

Property (3)

$$
\begin{aligned}
\sum_{\mathrm{bin}\,x} {}_s\mathcal{P}lot^2_{(i)}(\mathrm{bin}\,x) &= \sum_{e=1}^{N} {}_s\mathcal{P}lot^2_{(i)}(y_e) \\
&= \sum_{e=1}^{N} \left(\frac{\sum_{k=1}^{N_s} V_{ik} f_{(k)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} \right) \left(\frac{\sum_{j=1}^{N_s} V_{ij} f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} \right) \\
&= \sum_{k=1}^{N_s} \sum_{j=1}^{N_s} V_{ij} V_{ik} \sum_{e=1}^{N} \left(\frac{f_{(k)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} \right) \left(\frac{f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} \right) \\
&= \sum_{k=1}^{N_s} \sum_{j=1}^{N_s} V_{ij} V_{ik} \sum_{e=1}^{N} \left(\frac{f_{(k)}(y_e) f_{(j)}(y_e)}{(\sum_{k=1}^{N_s} N_k f_{(k)}(y_e))^2} \right) \\
&= \sum_{k=1}^{N_s} \sum_{j=1}^{N_s} V_{ij} V_{ik} V^{(-1)}_{kj} = \sum_{j=1}^{N_s} V_{ij} \delta_{ij} \\
&= V_{ii}
\end{aligned}
$$

Adding in quadrature the ${}_s\mathcal{P}lot$ error bars per bin one recovers the uncertainty on N_i

Conclusion



Control variables

x

Discriminating variables

y

The true distribution of x is obtained by weighting all events with sWeights

$$ {}_s\mathcal{P}_{(i)}(y_e) = \frac{\sum_{j=1}^{N_s} V_{ij} f_{(j)}(y_e)}{\sum_{k=1}^{N_s} N_k f_{(k)}(y_e)} $$

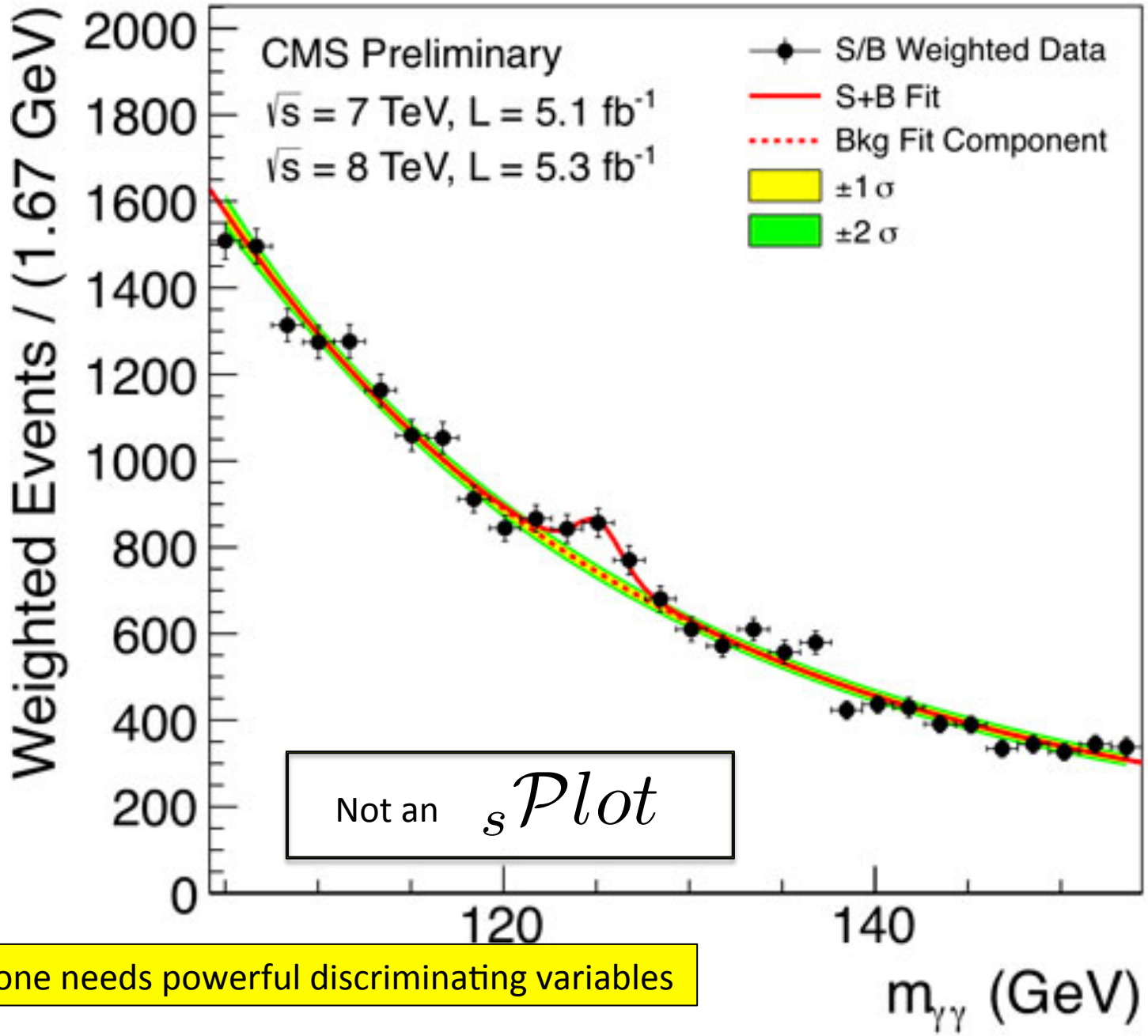
$$ \mathcal{L}_{\mathfrak{E}}(N_1 \ldots N_{N_s}) = \sum_{e=1}^{N} \ln(\{\sum_{i=1}^{N_s} N_i f_{(i)}(y_e)\}) - \sum_{i=1}^{N_s} N_i $$

$\mathcal{L}_{\mathfrak{E}}$

(covariance)

$_s$Plot

A very simple trick, with no handle to tune



… first one needs powerful discriminating variables