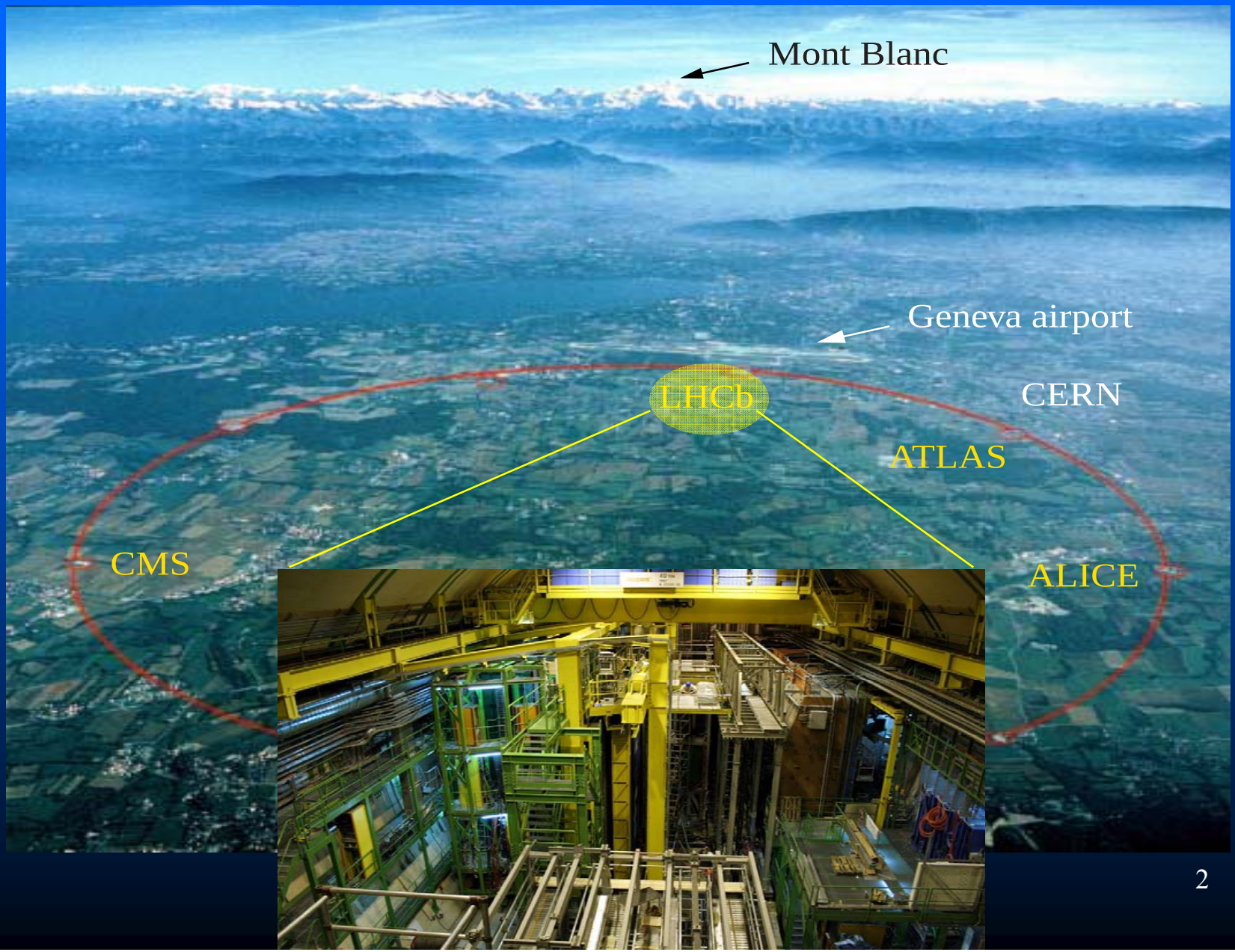
Statistical Issues at LHCb

Yuehong Xie
University of Edinburgh

(on behalf of the LHCb Collaboration)

PHYSTAT-LHC Workshop
CERN, Geneva June 27-29, 2007

LHCb: a dedicated B physics experiment at LHC



Physicists know where to look for new physics in flavour sector

- LHCb will look for effects of new physics in CP violation and rare phenomena in B meson decays
  - CMS and ATLAS will search for new particles directly produced
- In the Standard Model quark flavour mixing and CP violation are fully determined by the CKM matrix with four parameters
  - Over-constraining the CKM matrix is a stringent test of the SM
  - Any inconsistency will mean new source of flavour mixing and CP violation
- FCNC (flavour changing neutral currents) are forbidden at tree level in the SM but new physics can have significant effects in FCNC processes
  - Comparing asymmetry and/or rate measurements with their SM predictions in FCNC processes is a sensitive test of the SM
  - Any discrepancy will indicate presence of new physics particles in FCNC processes

Statisticians know how to quantify new physics effects

- In the language of statistics, LHCb will perform hypothesis testing
- The null hypothesis: the SM is valid at the energy scale relevant to B meson decays
- No alternative hypothesis is given explicitly, but rejecting the SM means new physics is needed to describe B meson decays
- What LHCb need do is
  - Identify a test-statistic which has high power to separate the SM and potential NP models
  - Measure the test-statistic from data
  - Evaluate the extreme probability in the null hypothesis, called *p*-value
  - If the p-value is too small, reject the null hypothesis
  - Otherwise go for another test-statistic and repeat the test

Where statistics/statisticians can help B physics/physicists?

- B flavour tagging on a statistical basis
- Separating signal/background events
- Data modeling and fitting
- Setting confidence intervals and limits
- Controlling and treating systematics
- Optimizing analyses
- Test of the SM
- Providing analysis tools

Flavour tagging

- CP violation measurements with neutral B decays need to know the flavour of the B at production
- Information is carried by taggers
  - Charge of the particles accompanying the signal B at production: same side tagging
  - Charge of the $\mu^{\pm}$, $e^{\pm}$ or $K^{\pm}$ from the decay of the opposite B hadron: lepton and kaon tagging
  - Weighted sum of charges of all particles found to be compatible with being from the opposite B decay: vertex charge tagging
- Tagging result of a signal B is a decision on statistical basis with
  - Average tagging efficiency ε, typically ~ 50–60%
  - Average mistag probability $\omega \equiv N^W/(N^W+N^R)$, typically 30-35%
  - Average statistical power $\varepsilon(1-2\omega)^2$, typically 4-10%
- How to maximize statistical power?
  - Require appropriate statistical methods

Statistical issues in flavour tagging

- Neural net used to get event-by-event mistag of each tagger. Performances depend on the way the NNs work and the way they are trained
- Treatment of correlation between vertex charge and other taggers is non-trivial
  - The other taggers may be included in the vertex
  - Compromise between correct handling of correlations and available statistics to get properties of sub-samples
- Hard assignment of particles to vertices causes loss of statistical power. Need to investigate probability-based assignment of particles to vertices

Separating signal/background events

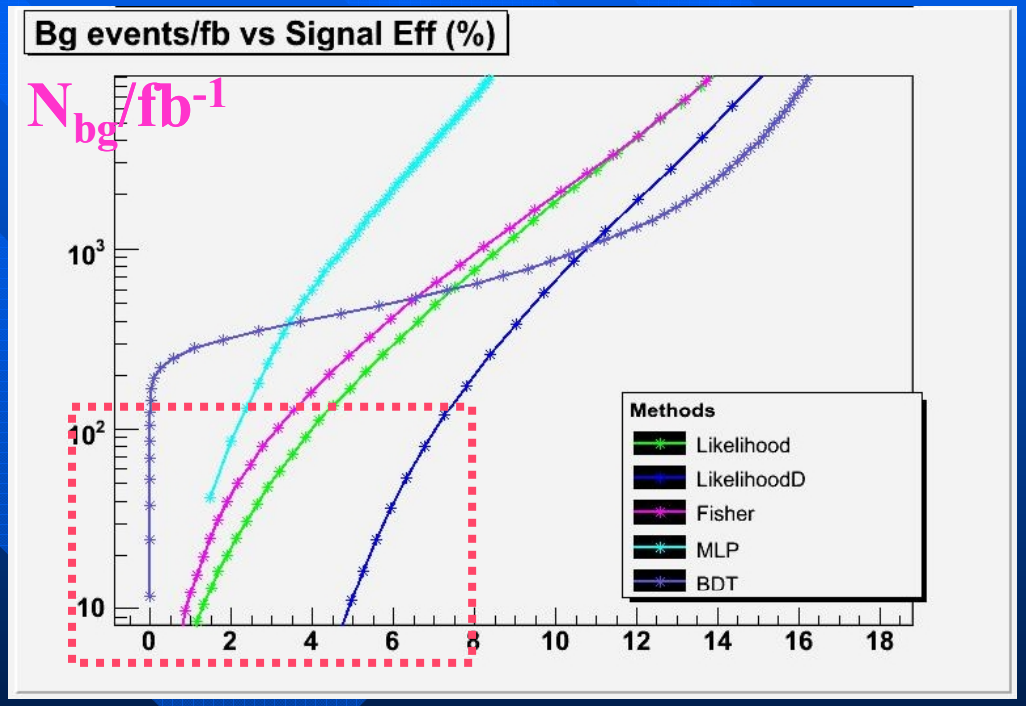
- A demanding task in a hadron machine experiment
  - After trigger the ratio of inclusive bbbar background to signal in a typical channel is at the million level, even bigger for very rare decays
  - In each bbbar event there are not only the two B hadrons but also ~100 tracks from pp interactions
- Information available
  - PID
  - Kinematical: momentum, invariant masses
  - Geometrical: vertex χ^2, event topology
  - More …
- Typically 10-20 variables to look at, each alone with limited separation power
  - Cut-based analysis not optimal for statistical precision
  - Multivariate analysis more powerful but also more difficult for understanding systematics
  - Need trade-off between precision and accuracy

Multivariate analysis

- Essentially about how to construct a best test statistic from many input variables for a hypothesis test
- Neyman-Pearson-Lemma: likelihood ratio is the best
  - The method of representing PDFs by multi-dimensional histograms using Monte Carlo data becomes impractical when the dimension of the PDFs is too big
- The alternatives is to construct estimators to approach the likelihood ratio
  - Decorrelated likelihood method
  - Linear estimator: Fisher's discriminants, …
  - Nonlinear estimator: neural networks, …
  - Boosted decision trees
  - …
- An implementation: TMVA (Toolkit for MultiVariate data Analysis) *(arXiv:physics/0703039)*
- Application in LHCb: SM forbidden $B_s \rightarrow e^{\pm}\mu^{\mp}$ analysis
  - High efficiency has higher priority than small systematics

Application of TMVA in $B_s \rightarrow e^{\pm}\mu^{\mp}$

- Two phases
  - Training
  - Application
- Leave variables with clear separation power outside TMVA and cut on them
- 5 input variables in TMVA: no linear correlation expected
- Winner: decorrelated likelihood method
  - As Neyman-Peasrson Lemma tells us?



Our interest is here

Overtraining with TMVA

- TMVA has a mechanism to monitor overtraining using two independent training and testing samples

```
: ---------------------------------------------------------------------
:
: Testing efficiency compared to training efficiency (overtraining check)
: ---------------------------------------------------------------------
: MVA              Signal efficiency: from test sample (from traing sample)
: Methods:           @B=0.01             @B=0.10             @B=0.30
: ---------------------------------------------------------------------
: LikelihoodD    : 0.586 (0.874)       0.848 (0.925)       0.940 (0.968)
: MLP            : 0.643 (0.617)       0.848 (0.855)       0.955 (0.954)
: BDT            : 0.563 (0.852)       0.845 (0.955)       0.944 (0.974)
: Likelihood     : 0.523 (0.564)       0.803 (0.804)       0.916 (0.918)
: Fisher         : 0.497 (0.504)       0.780 (0.785)       0.917 (0.922)
: ---------------------------------------------------------------------
:
```

- Way to control overtraining in early phase is desirable

Data modeling and fitting

- Maximum likelihood fit is generally used in B physics measurements
- Modeling and fitting made easy with RooFit *(http://rootsourceforge.net)*
- LHCb can benefit from this package and wishes to
  - Have a goodness-of-fit for unbinned maximum likelihood fit implemented
  - Understand how to speed up toy event generation for complicated PDF
  - Understand how to make fit converge if a non-factorizable multi-dimensional PDF has no analytical normalization and can only rely on numerical integration

Confidence intervals/limits

- As in all other experiments, we need to quote confidence intervals/limits for all measurements
- The issue is especially important when working on very rare decays with small signal and huge background
  - significant signals: intervals to establish discrepancy with SM:
  - Insignificant signals: limits to exclude some new physics models
- An example: $B_s \rightarrow \mu^+\mu^-$ sensitivity
  - $BR(B_s \rightarrow \mu^+\mu^-)_{SM} = 3.4 \times 10^{-9}$ but enhanced in some new physics models
  - Exclusion limit: confidence limit for the signal decay branching ratio when the generated *N* events in an Monte Carlo experiment are all background, a measure of sensitivity

Exclusion limit for BR($B_s \rightarrow \mu^+\mu^-$)

- Step 1: construct geometrical, muon-id and invariant mass likelihood ratios between signal and background hypotheses for each event
  - decorrelated likelihood method
- Step 2: divide the 3D space into a number of bins and count events d_i in each bin
  - no cut and N-counting
- Step 3: estimate number of expected background events b_i and signal events s_i (for each assumed branching ratio) in each bin
- Step 4: construct a total likelihood ratio between the signal+background and background-only hypotheses for the whole configuration

$$X = \prod_i \frac{Poisson\left(d_i, \langle d_i \rangle = s_i + b_i\right)}{Poisson\left(d_i, \langle d_i \rangle = b_i\right)}$$

Exclusion limit for BR($B_s \rightarrow \mu^+\mu^-$)

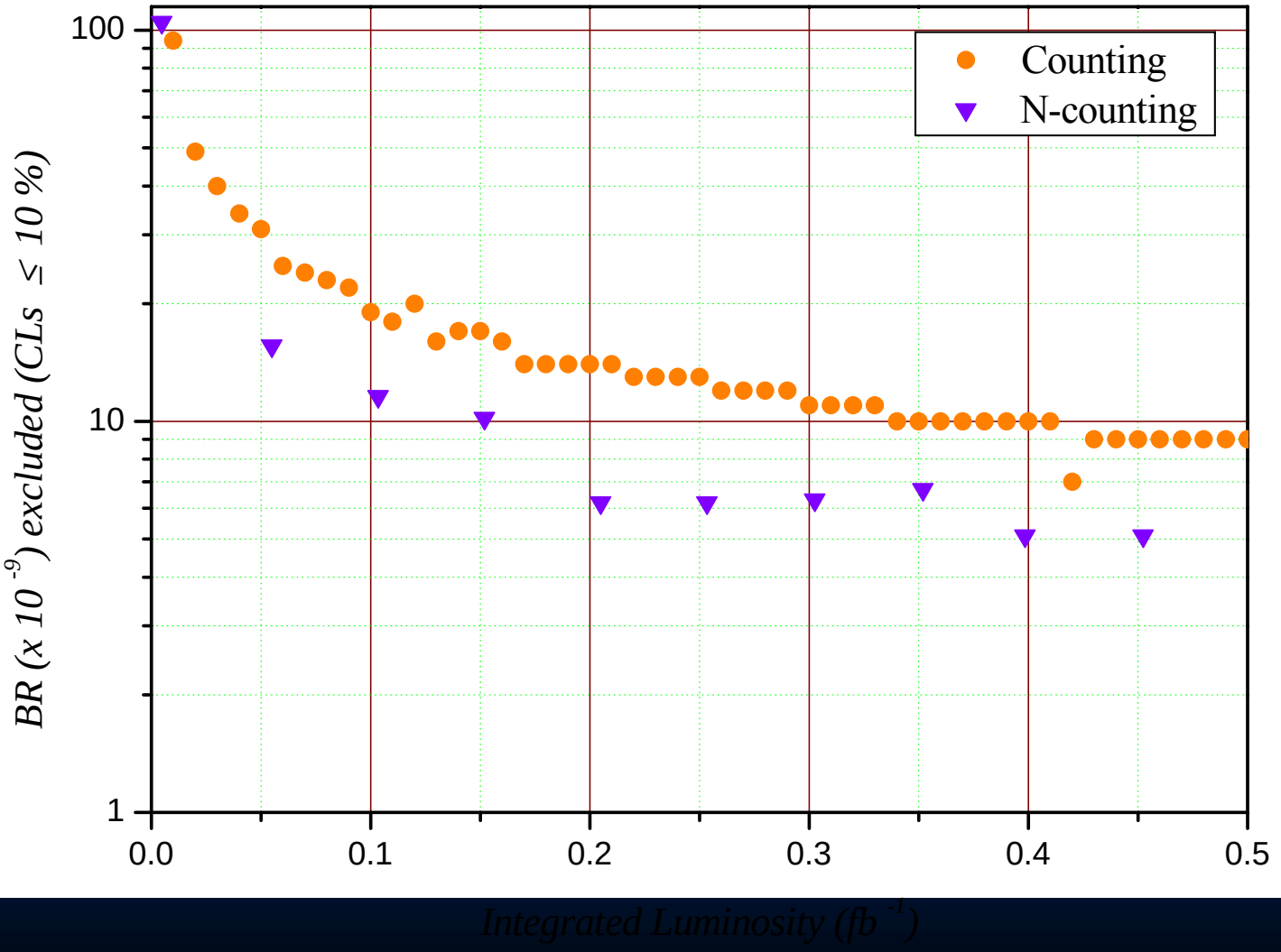
- Step 5: evaluate the *p*-value of each hypothesis
  - Signal+background: *probability*($X<X_{obs}$)
  - Background-only: *probability*($X>X_{obs}$)
- Step 6: compute CL_s *(Thomas Junk, CERN-EP/99-041)*

$$CL_s = \frac{p\text{-value of signal plus background hypothesis}}{1-\left(p\text{-value of hypotheis of background only}\right)}$$

- Step 7: make statistical statement:

  If CL_s(BR) < α, the assumed BR is excluded at 1-α confidence level

$BR(B_s \rightarrow \mu^+\mu^-)$ exclusion limit results



Exclusion limit for BR($B_s \rightarrow \mu^+\mu^-$)

- The combination of various techniques is shown to work better than the cut-and-count method
- "Statistics Review" of PDG2006 claims the CL_s limit is conservative and the confidence level is under-estimated
  - The usual procedure requires the *p*-value, not the CL_s, of a hypothesis to be smaller than α to exclude it at $1-\alpha$ confidence level
- What is the common understanding of how to set better limits in circumstance of insignificant signal?

Controlling systematics

- Systematics arise from incorrect modeling of detector and/or background effects
- Need delicate statistical methods to acquire these effects from real data and model them
- Example: two methods to deal with efficiency as a function of decay time $\varepsilon(t)$ in time-dependent analysis
  - Per-event $\varepsilon_i(t)$: not covered in this talk
  - Normalization trick: Described by Stéphane T'Jampens

https://oraweb.slac.stanford.edu/pls/slacquery/BABAR_DOCUMENTS.DetailedIndex?P_BP_ID=3629 (French thesis)

Normalization trick

- Factorized Signal PDF
  - A: physical parameters
  - t: decay time
  - Ω: position in phase space
- Likelihood
- Maximization

$$p(t,\Omega;A)=\frac{\sum_i h_i(A) f_i(t) g_i(\Omega)\varepsilon(t,\Omega)}{\sum_i h_i(A)\int f_i(t) g_i(\Omega)\varepsilon(t,\Omega)dtd\Omega}$$

$$L=\prod_j l_j=\prod_j p(t_j,\Omega_j;A)$$

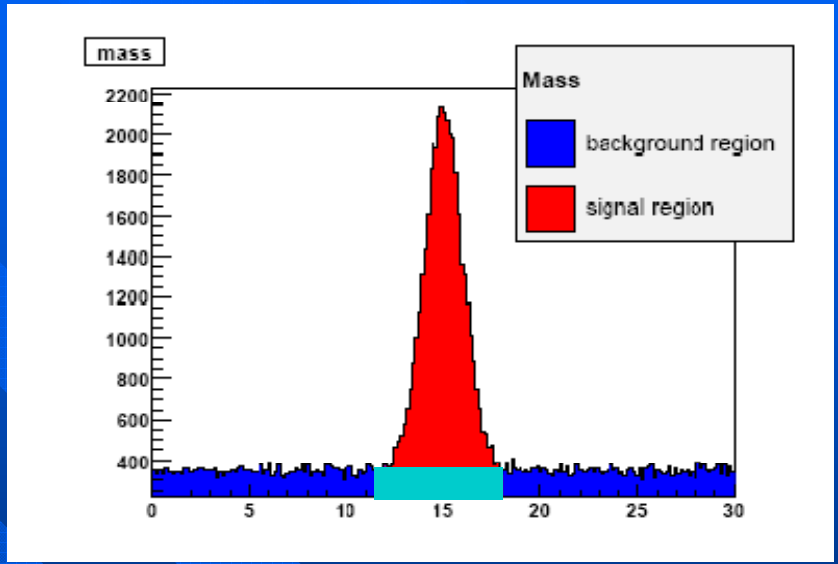
$$\frac{d\ln l_j}{dA}=\frac{d}{dA}\left[\ln\left(\frac{h_i(A) f_i(t_j) g(\Omega_j)\varepsilon}{h_i(A)\int f_i(t) g_i(\Omega)\varepsilon(t,\Omega)dtd\Omega}\right)\right]=\frac{d}{dA}\left[\ln\left(\frac{h_i(A) f_i(t_j) g(\Omega_j)}{h_i(A)\,\Phi_i}\right)\right]$$

- Integrated factor Φ_i obtained using MC simulation
- No need for specific shape of $\varepsilon(t,\Omega)$

Fitting with background

- Unable to know the "ideal" background distributions w/o detector effect from physical law
- Solution: use pseudo-log-likelihood to avoid need of background distributions

  Phy.Rew. D71(2005) 032005



$$\ln L = \sum_{i=1}^{N_{sig}} \ln p_{sig}(t_i, \Omega_i, A) - \frac{N_{sb}}{N_b} \sum_{j=1}^{N_b} p_{sig}(t_j, \Omega_j, A)$$

N_{sig}/N_b: number of events in the signal/background region

N_{sb}: number of expected background events in the signal region

Fitting with background

- Errors return by Minuit are too optimistic
- The true variance of a fit parameter is given by

$$Var(s) = \sigma_s^2 \left(1 + \left(1 + \frac{N_{sb}}{N_b} \right) \frac{\sigma_s^2}{\sigma_b^2} \right)$$

(Private communication with Joe Boudreau)

- Three items
  - One from signal events in signal region
  - One from background events in signal region
  - One from events in background region, vanishes as $N_b \rightarrow \infty$
- No need for background PDF

Estimating systematics

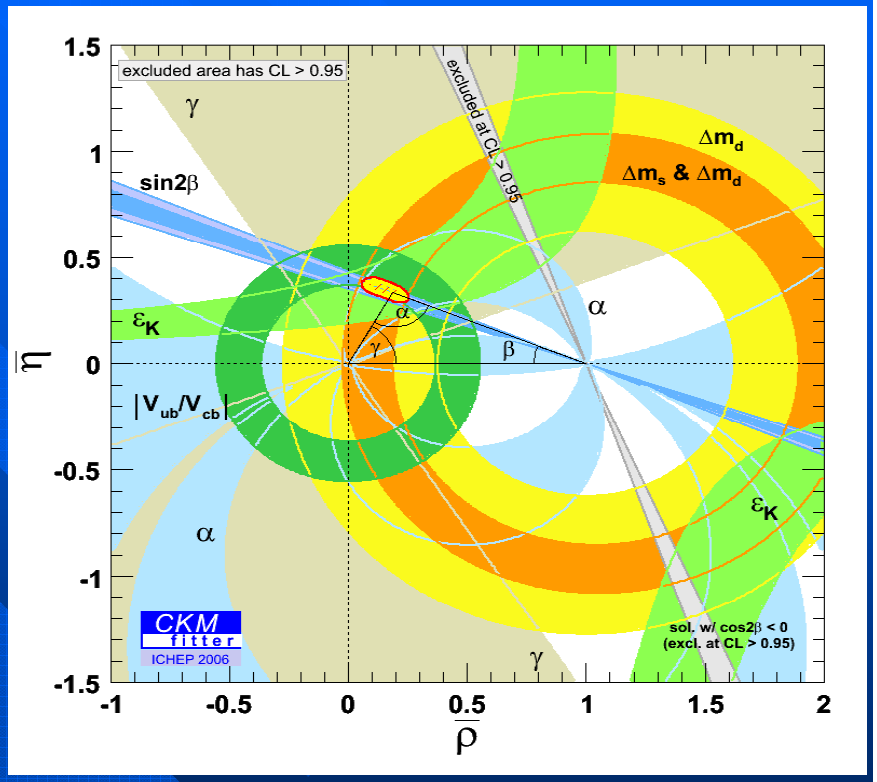
- Every effort has been made to model the detector/background effect correctly
- However, have to use some assumptions
  - E.g., background events in sideband and signal region have same properties
- Also require good agreement of MC data and real data
  - Not everything can be obtained from real data
- What is the proper procedure to estimate systematic errors in the treatment of detector/background effects so that at least different people will come to consistent estimates of the systematic uncertainties if the same analysis method is used?
  - What is the rule to set e.g. "1-σ systematic error"? Vary what quantities to obtain it? By how much?
  - What is the statistical meaning of "1-σ systematic error"?

Analysis optimization

- What is the target function to optimize in order to obtain the best statistical precision for CP measurements?
  - Signal events are not with equal weights
  - Events with smaller mistag probability and better time resolution contribute more

SM test: CKM fit

- Different measurements of the sides and angles of the triangle should be consistent
  - UT Fit: Bayesian
  - CKM Fitter: frequentist
  - Which one better serves this purpose?
- Questions
  - How to deal with theoretical uncertainties? Do they have frequentist properties?
  - How do we know if an inconsistency is due to NP or underestimated systematics and theoretical uncertainties?



Should give global χ^2 as measure of agreement with SM

SM test: rare decays?

- LHCb will measure many rare decays
- Individually they are all good probes of NP
- How to combine them to get best sensitivity?
- Not a easy job
  - The SM relations between these quantities are not explicitly given
  - SM prediction for each of them is with big uncertainty
- Need a lot of work in physics
  - Understand the correlations between the SM predictions from physics and quantify the correlations using an error matrix
- Also some thinking in statistics
  - Construct a test statistic using the measurements and their SM predictions, latter (or both) with correlated errors

Analysis tools

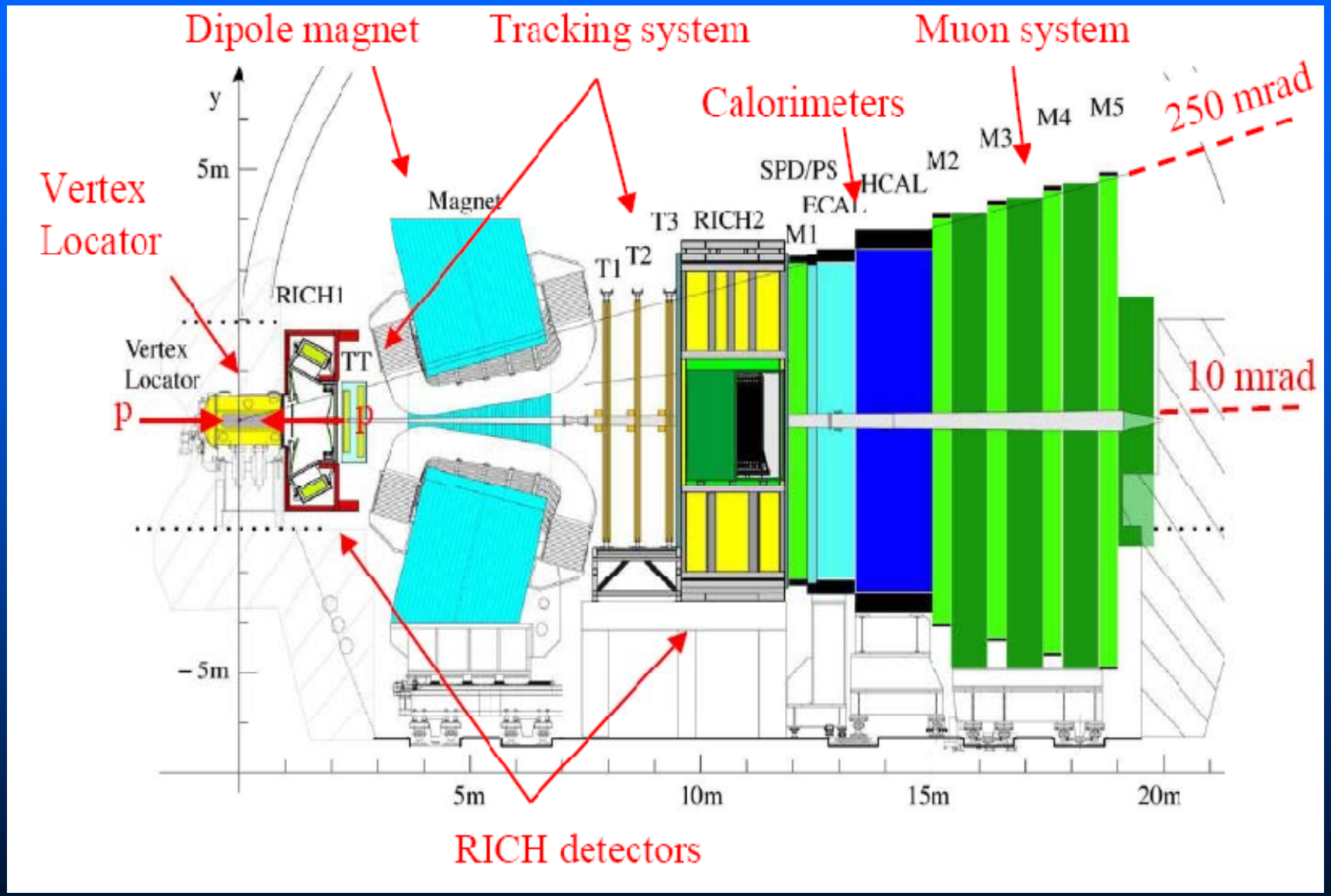
- Tools we have used
  - Data storage and general processing / Root
  - Minimization / Minuit
  - Data modeling and fitting / RooFit
  - Multivariate analysis / TMVA
  - Data unfolding / sPlot
  - Neural networks
- Tools we may need
  - Frequentist limit setting tool
  - Bayesian analysis tool
  - Reliable numerical multi-dimensional integrator

Wish-list

- A well supported tool for data modeling and fitting, which can handle general multi-dimensional problems numerically
- Better understanding of how to do multivariate analysis
- Better understanding of how to treat systematics and theoretical uncertainties in SM test
- New statistical methods to control systematics using real data
- New statistical methods to improve flavour tagging
- Better understanding of how to set confidence limits in case of insignificant signal
- Statisticians' recommendation on statistical procedures in data analysis

And most importantly a successful LHC(b)!

spare slides


Dipole magnet
Tracking system
Muon system
Calorimeters
Vertex Locator
y
5m
RICH1
Vertex Locator
TT
Magnet
T1
T2
T3
RICH2
M1
SPD/PS
HCAL
ECAL
M2
M3
M4
M5
250 mrad
10 mrad
p
p
– 5m
5m
10m
15m
20m
RICH detectors

Application to $B_s \to e^{\pm} \mu^{\mp}$: training

Various available variables:

we have just linear *decorrelation*

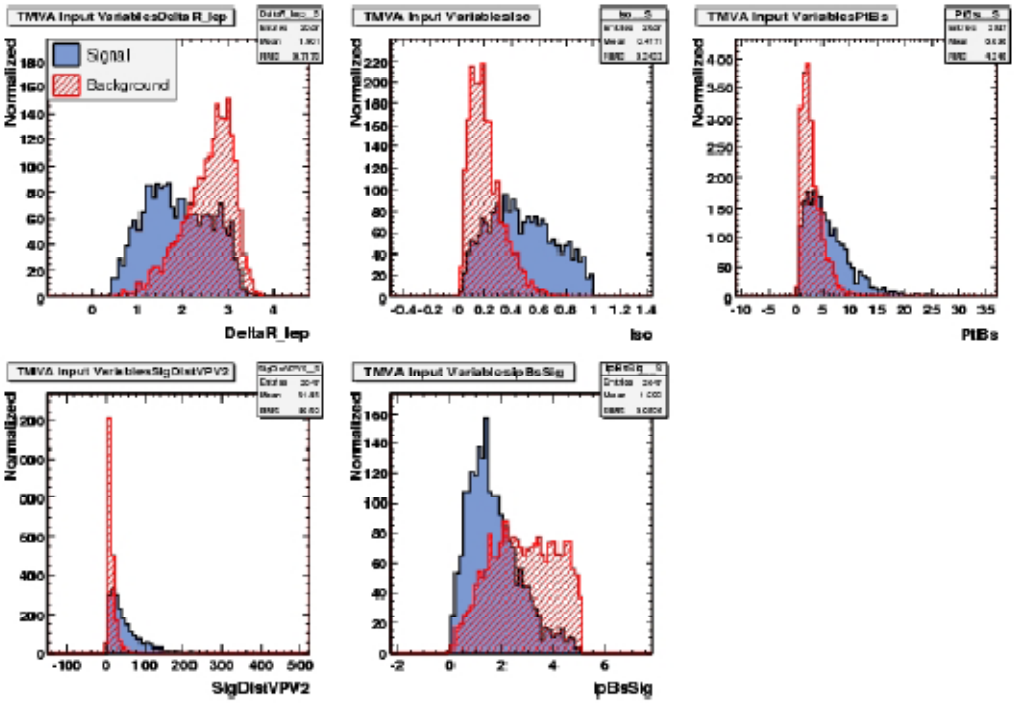
→ we've discarded non-linearly correlated variables

Input variables:

- Leptons DeltaR $= \sqrt{\Delta\phi^2 + \Delta\eta^2}$
- Isolation $Iso = p_T \dfrac{(B_s^0)}{p_T(B_s^0) + \sum_i (p_T)_i}$
- $p_T(B_s^0)$
- Significance of vertices dist. $= \dfrac{|\vec{r}_{PV} - \vec{r}_{BV}|}{\sigma}$
- $IP(B_s^0 \ w.r.t. PV)/\sigma$

Note: we trained our methods with a background sample composed of the different channels weighted on their respective cross-sections

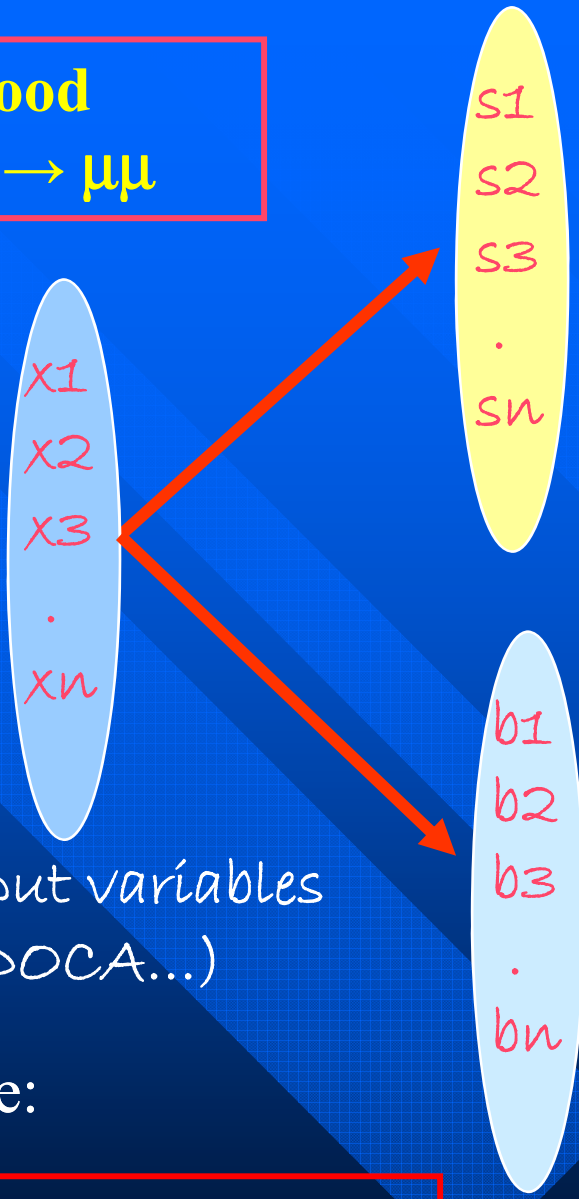
Input Variables distributions



Decorrelated likelihood method used for $B_s \rightarrow \mu\mu$

- Decorrelate the input variables for signal and background separately

- A very similar method is described by Dean Karlen, *Computers in Physics Vol 12, N.4, Jul/Aug 1998*



→ n variables for signal independent and Gaussian distributed

$$\rightarrow \chi^2_S = \Sigma\, s_i^2$$

→ same for background

$$\rightarrow \chi^2_B = \Sigma\, b_i^2$$

Discriminating variable:

$$-2\Delta\ln(L_{sig}/L_{bg}) = \chi^2_S - \chi^2_B$$

Lessons learned with TMVA

- Overtraining is a general problem in this kind of analysis when the training sample has low statistics
- TMVA splits the sample into two
  - One for training and one for test and evaluation
- Would it make more sense to use three independent samples?
  - Sample A for training
  - Sample B for deciding when to stop training by looking at the performance difference between A and B
  - Sample C for evaluation of performance
  - This is because correlation may have been introduced between A and B when B is used to decide when to stop training