Some Statistical Issues in the LHCb Experiment

Yuehong Xie
The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK

Abstract

This paper describes statistical issues that are of particular importance and interest to the LHCb experiment in probing new physics beyond the Standard Model through study of CP violation and rare phenomena in B decays. A wish list for statistical methods and tools that will help LHCb to exploit its full physics potential is given at the end.

1 Introduction

The LHCb experiment is a dedicated B physics experiment at the Large Hadron Collider (LHC). Its physics aim is to study CP violation and rare phenomena in B decays with very high precision in order to test the Standard Model (SM) in the quark flavour sector and to look for physics beyond the SM. Different from the ATLAS and CMS experiments, which will explore the high energy frontier to search for new physics particles directly produced in proton-proton collisions at the LHC, the LHCb experiment will pursue precision measurements to understand the quantum effects of possible virtual new particles appearing in loop diagrams. LHCb will need to put enormous efforts to understand how to deal with background, control systematic uncertainties and incorporate theoretical errors. Improving signal significance is also very important, especially for measurements of very rare decay processes.

This paper discusses the major issues in LHCb physics analysis that require special statistical treatments. These are illustrated using examples of analysis. In addition a list of statistical methods and tools that LHCb wishes to develop, improve or understand better is given.

2 The physics of the LHCb experiment

In the SM, quark-flavour mixing and CP violation are fully described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix with four independent parameters. The task of flavour physics is to determine these parameters and more importantly to check the validity of the CKM mechanism. LHCb will take two routes to achieve this task. LHCb will make many measurements to over-constrain the CKM matrix. These will provide stringent tests of the SM. Any inconsistency will mean that some new source of flavour mixing and CP violation must exist. LHCb will also study Flavour-Changing-Neutral-Current (FCNC) loop decays. The FCNC decays are forbidden at tree level in the SM, therefore new physics may have significant effects in these processes. Comparing asymmetries or rates in these decays with their SM expectations will be a sensitive test of the SM. Any established discrepancy will indicate new FCNC couplings beyond CKM mixing.

3 A statistical view of the LHCb experiment

Statistics plays an important role in quantifying the level of consistency/inconsistency between physics models and experimental data. In the language of statistics, LHCb will perform hypothesis testing. The null hypothesis is that "the SM is valid at the energy scale relevant to B meson decays". No alternative hypothesis is explicitly given, but rejecting the null hypothesis implies new physics is needed to explain the data. What LHCb needs to do for a hypothesis test of the SM is

– Identify a test statistic, i.e. an observable, in flavour physics which has high power to separate the SM and potential new physics models;

- Measure the test statistic from data;
- Evaluate the tail probability of the null hypothesis, that is, the p-value;
- If the p-value is judged too small, reject the null hypothesis and look for a possible alternative;
- Otherwise go for another observable and repeat the test.

4 Application of statistics in LHCb physics analysis

Statistical methods and concepts are used in almost every aspect of B physics experiments, ranging from pattern recognition to averaging measurements. Since many of these issues have been widely discussed in the B physics community, this paper focuses on those aspects of LHCb data analysis which can potentially benefit a lot from improvement of statistical methods.

4.1 B flavour tagging

For most CP measurements with neutral B decays it is necessary to know the flavour of the B meson at production. This can be inferred from the information carried by the following tagging categories:

- Same side tagger: charge of the particle accompanying the signal B at production;
- Opposite side tagger: charge of muon, electron or kaon from the decay of the opposite side B hadron;
- Vertex charge tagger: the weighted sum of the charges of all particles found to be compatible with being from the opposite side B decay.

The tagging result is a decision made on a statistical basis combining all available taggers. The figures of merit of tagging is the effective tagging power $\epsilon(1-2\omega)^2$, where ϵ is the tagging efficiency and ω is the mistag probability. The current estimates using simulated data are $\epsilon \sim 50-60\%$, $\omega \sim 30-35\%$ and $\epsilon(1-2\omega)^2 \sim 4-10\%$. Since the statistical errors of the CP asymmetries in neutral B decays decrease linearly with the square root of the tagging power, it is crucial to maximize the tagging power using appropriate statistical methods.

In LHCb neural network methods are employed to get event-by-event mistag probability of each tagger, the performance of which depends on the way the neural networks are constructed and trained. We expect some room for improvement here. Combining different tagging categories is non-trivial when these are correlated. For example, if the opposite side tagger and the vertex charge tagger use the same particle, correct handling of this correlation requires splitting the data sample into sub-samples depending on whether there is a particle used by the opposite side tagger and the vertex charge tagger or not. LHCb is developing new techniques to optimize the procedure of combining taggers. A possibility for tagging improvement is to investigate using better methods to assign particles to vertices. The tagging algorithm needs determine if a particle originates from a primary vertex or from a tagging B vertex. It may lead to a wrong tagging decision if for example a charged lepton from a primary vertex is mistakenly regarded as being from the tagging B hadron or loss of tagging efficiency if a charged lepton from the tagging B hadron is mistakenly treated as being from a primary vertex. In both cases, the effective tagging power is compromised. We have already investigated various methods to minimize this loss of tagging power, but room for further improvement is still possible.

4.2 Separating signal and background events

Separating signal and background events is a demanding task in LHCb for two reasons: after trigger the ratio of inclusive $b\bar{b}$ background events to signal events is at the order of one million to one in a typical decay channel and can be even larger for a very rare decay channel such as $B_s \to \mu^+\mu^-$; in each $b\bar{b}$ event there are not only the two B hadron decays but also about 50 tracks from proton-proton interactions.

The following information can be used for signal and background separation

- Particle identification;
- Kinematic information such as particle momenta and invariant masses of particle combinations;
- Geometrical information such as the secondary vertex χ^2 and event topology.

There are typically 10-20 variables to look at in an analysis, each alone with limited discrimination power. Therefore, a cut-based analysis method is usually not optimum in terms of statistical precision. Multivariate analysis methods can be more powerful for signal and background separation but this involves more complexity to understand the systematic issues. In a real analysis one needs to find a trade-off between better statistical precision and smaller systematic uncertainty.

A multivariate analysis entails constructing a best test statistic from multiple input variables for a hypothesis test. In principle the Neyman-Pearson lemma tells us the likelihood ratio is the best choice (for simple hypotheses). A straightforward application is to represent the probability density functions (PDFs) of signal and background by multi-dimensional histograms which can be obtained from Monte Carlo simulation. The likelihood ratio then can be computed as the ratio of the two PDFs. However, this procedure becomes impractical when the dimensions of the PDFs are too large.

There are alternative methods which construct estimators to approach the likelihood ratio under certain conditions. Examples include decorrelated likelihood classifier, linear estimators such as Fisher's discriminants, nonlinear estimator such as neural networks and Boosted decision trees. The Toolkit for MultiVariate data Analysis (TMVA) [1], an integrated part of the Root framework, hosts a variety of these multivariate algorithms and provides many techniques that are useful in LHCb data analysis.

The TMVA package has been applied in a simulation study of the decay channel $B_s \to e^{\pm}\mu^{\mp}$ [3], which is forbidden in the SM and therefore requires high selection efficiency and low background level. In this study, variables showing very clear separation between signal and background are directly cut on first. TMVA is used to deal with the less powerful variables. This effectively reduces the complexity in understanding systematics. In this particular case no non-linear correlations between these are expected. A sample of simulated data is used to train several classifiers and an independent data sample is used to evaluate their performances. The results are shown in Fig. 1. Just as the Neyman-Pearson lemma implies, the decorrelated likelihood method, denoted as LikelihoodD in Fig. 1, gives the highest signal efficiency for the same number of background events. This is not necessarily the case in other more complicated analyses with non-linear correlations between the input variables. There are indications that the current TMVA mechanism to monitor over-training of classifiers using independent samples for training and testing may not be sufficient. A way to control over-training in the training phase is desirable.

4.3 Setting confidence limits in case of a small signal

As in all HEP experiments, we need to quote confidence intervals/limits for experimental measurements. This issue is especially important when working on very rare decays with small signals and large background. Here we use the analysis of $B_s \to \mu^+\mu^-$ [3] as an example to illustrate the statistical procedure which LHCb adopts for determination of the experimental sensitivity in very rare decay channels. We know that $B_s \to \mu^+\mu^-$ is highly suppressed in the SM and its branching ratio is expected to be around 3.4×10^{-9}. This can be greatly enhanced in new physics models. We use the average exclusion limit as a measure of the experimental sensitivity, which is defined as the average upper limit that would be obtained from an ensemble of experiments with the expected background and no true signal [4]. We evaluate the average exclusion limit by generating toy experiments with only background events and using the "N-counting" method described below to set an upper limit for each toy experiment. The "N-counting" method includes the following steps:

- Construct geometrical, muon-ID and $\mu^+\mu^-$ invariant mass likelihood ratios between signal and

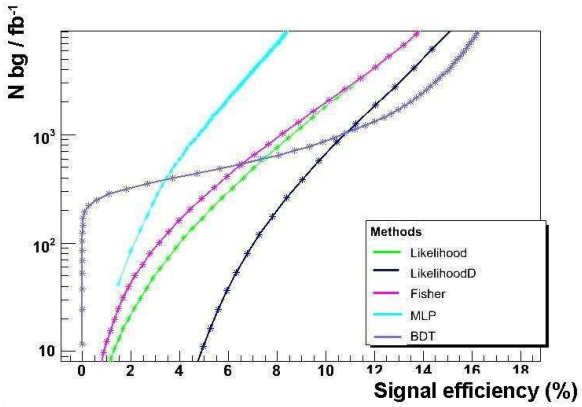

Fig. 1: Number of retained background events per fb^{-1} as a function of signal efficiency (%) for various multivariate methods.

background hypotheses for each event, where the decorrelated likelihood method is used for the geometrical likelihood ratio;

- Divide the 3-dimensional space of the three likelihood ratios into a number of bins and count the number of events in each bin, denoted as d_i;
- Estimate the number of expected background events b_i and signal events s_i for each examined branching ratio in each bin;
- Construct a total likelihood ratio between the signal+background and background-only hypotheses for the whole experiment

$$X = \prod_i \frac{P(d_i, < d_i >= s_i + b_i)}{P(d_i, < d_i >= b_i)}, \tag{1}$$

where $P(x, < x >)$ denotes the Poisson probability of a variable x with the average value $< x >$;
- Evaluate the p-value of the signal+background hypothesis: $probability(X < X_{observed}; S + B)$ and that of the background-only hypothesis: $probability(X > X_{observed}; B)$;
- Form a statistic called CL_s [5] from the ratio

$$CL_s = \frac{p\text{ -value of signal plus background hypothesis}}{1 - (p\text{ -value of background hypothesis})}; \tag{2}$$

- Make a statistical statement: if $CL_s(BR) < \alpha$, then the branching ratio BR is excluded at $1 - \alpha$ confidence level.

The comparison of the N-counting method and a simple counting method is shown in Fig. 2. It can be clearly seen that the N-counting method requires less data to reach the same average exclusion limit at 10% confidence level. While the CL_s test statistic has some advantages over the likelihood ratio X and the CL_s limit is easy to compute, the way the confidence level is set is known to be conservative [6]. The normal procedure requires the p-value of the signal+background hypothesis, not the CL_s, to be smaller than α in order to exclude the signal+background hypothesis at $1 - \alpha$ confidence level.

It should be noted that the significance of a measured result is given by the p-value of the background-only hypothesis, which should not be confused with the p-value of the signal+background hypothesis or the CL_s value.

4.4 Analysis tools for data modelling and fitting

The maximum likelihood fit method is generally used in B physics experiments. This is largely facilitated by the data modelling and fitting package RooFit [7]. The RooFit package has been widely used in LHCb

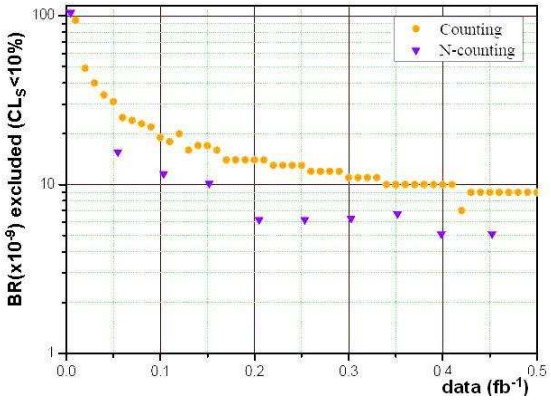

Fig. 2: $B_s \to \mu^+\mu^-$ average exclusion limit (10^{-9}) at 90% confidence level as a function of the integrated luminosity (fb^{-1}) for the N-counting method and the simple counting method as a comparison.

sensitivity studies. The experience shows that LHCb can better benefit from this package if:

- The event generation for complicated PDFs can be made faster;
- We learn how to make fits converge that employ non-factorizable multi-dimensional PDFs that have no analytical normalization and can only be numerically integrated.

4.5 Controlling systematic errors

Systematic errors arise from incorrect modelling of the detector and/or background effects. Delicate statistical methods are needed to acquire knowledge of these effects from real data and to model them. An example is the efficiency as a function of the proper decay time t and phase space position Ω, denoted as $\varepsilon(t,\Omega)$. Correct modelling of $\varepsilon(t,\Omega)$ is very important for time-dependent and/or an angular analysis. Here we discuss a technique [8] to absorb the effect of $\varepsilon(t,\Omega)$ into a normalization factor. Generally the PDF describing a signal decay has the form

$$p(t,\Omega;A) = \frac{\sum_j h_j(A) f_j(t) g_j(\Omega)\varepsilon(t,\Omega)}{\sum_j h_j(A) \int_t \int_\Omega f_j(t) g_j(\Omega)\varepsilon(t,\Omega) dt d\Omega}, \tag{3}$$

where $h_j(A)$, $f_j(t)$ and $g_j(\Omega)$ are functions that depend only on the physical parameters A, decay time t or phase space position Ω respectively. The likelihood of all signal events is

$$L = \prod_i l_i = \prod_i p(t_i,\Omega_i;A). \tag{4}$$

Varying parameters A to maximize L requires evaluating

$$\frac{d\ln l_i}{dA} = \frac{d}{dA}\ln\frac{\sum_j h_j(A) f_j(t) g_j(\Omega)\varepsilon(t,\Omega)}{\sum_j h_j(A) \int_t \int_\Omega f_j(t) g_j(\Omega)\varepsilon(t,\Omega) dt d\Omega} \equiv \frac{d}{dA}\ln\frac{\sum_j h_j(A) f_j(t) g_j(\Omega)}{\sum_j h_j(A)\Phi_j}, \tag{5}$$

where we have defined $\Phi_j \equiv \int_t \int_\Omega f_j(t) g_j(\Omega)\varepsilon(t,\Omega) dt d\Omega$. These factors Φ_j are independent of the physical parameters A and therefore can be obtained from Monte Carlo simulation before fitting. Note the acceptance function $\varepsilon(t,\Omega)$ in the numerator of Eq. 3 drops out in the last step of Eq. 5 because it only contributes a constant to the log-likelihood function. There is no need to know the explicit form of $\varepsilon(t,\Omega)$ in the maximum likelihood fitting. In addition to the general problem of a lack of goodness-of-fit in a unbinned likelihood fit, a lack of knowledge of $\varepsilon(t,\Omega)$ makes it difficult to check the fit quality by

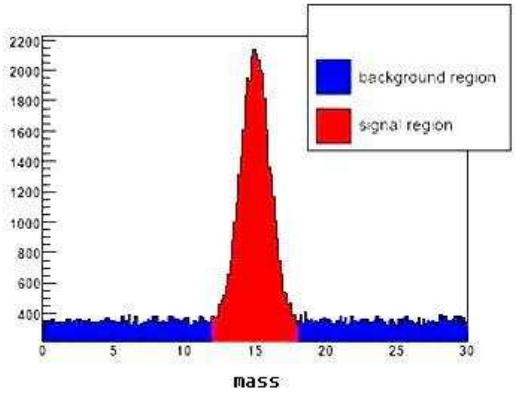

Fig. 3: Schematic view of signal and sideband regions defined with invariant mass.

comparing measured distributions and fitted projections. A solution must be found to ensure the fit result is reliable. This is still under investigation.

When background events are taken into account the total PDF becomes

$$p_{tot} = f \cdot p_{sig} + (1 - f) \cdot p_{bkg}, \tag{6}$$

where f is the fraction of signal events in the data sample and p_{sig}, p_{bkg} denotes signal and background PDF respectively. The total PDF no longer has the form of Eq. (3). Therefore the normalization trick of Eq. (5) cannot be employed.

One solution to this problem is to use a pseudo-log-likelihood method [9], which avoids the use of any background PDF. Instead of maximizing the usual likelihood defined using the total PDF in Eq. (6), one maximizes the pseudo-log-likelihood defined using only the signal PDF

$$\ln L_{pseudo} = \sum_{i=1}^{N_{sig}} \ln(p_{sig}(t_i, \Omega_i; A)) - \frac{N_{sb}}{N_b} \sum_{j=1}^{N_b} \ln(p_{sig}(t_j, \Omega_j; A)), \tag{7}$$

where $N_{sig/b}$ is the total number of events in the signal/sideband region and N_{sb} is the number of expected background events in the signal region. The signal and sideband regions can be defined in terms of invariant mass as shown in Fig. 3. While the minimization of the negative pseudo-log-likelihood leads to a unbiased estimate of the physical parameters A, the errors returned by Minuit at the end of the minimization are generally under-estimated. There are efforts underway to derive a formalism to give the correct estimates of parameter errors [10]. Useful discussions on the topic of background-subtraction during this workshop can be found in [11]. Every effort needs to be made to model the detector and background effect correctly in order to minimize systematic errors. In addition, a proper statistical procedure should be established and strictly followed when estimating systematic errors so that the arbitariness in assigning systematic uncertainties can be minimized.

4.6 CKM fit and other global fits

An area in B physics where statistical analysis plays a major role is the CKM fit. The goal of the CKM fit is to test if different measurements of the sides and angles of the Unitarity Triangle are consistent or not. Currently there are two groups working on this using the B factory measurements: the UTFit group [12] which uses a Bayesian method and the CKMFitter group [13] which employs a frequentist approach. While the two methods are very different from each other, the basic conclusions made by the two groups are the same: no inconsistency between measurements is found so far. Once LHCb starts to produce precision measurements, a more stringent test of the CKM mechanism can be done. At that time it might

be necessary to consider improving the statistical treatments of these existing tools to make best use of the high precision measurements at LHCb. For example, we may want to know how better to incorporate theoretical uncertainties and how to tell if an inconsistency is due to new physics or to under-estimated systematic uncertainties or under-estimated theoretical uncertainties.

LHCb will also measure many rare decay channels. Individually they test the SM and probe new physics in one way or another. We may want to perform a global fit in the SM in order to achieve better sensitivity to new physics. This is not an easy job as the SM relations between the measured quantities in the rare decay channels are not precisely known and the SM predictions of these quantities are usually subject to sizable uncertainties. A lot of analysis efforts are first needed to understand the SM and make better predictions. In terms of statistical methods, some thinking is required to understand how to construct a test statistic using the measurements in rare decay channels and their SM predictions and taking into account the uncertainties of these predictions and the correlation between the predicted errors due to common theoretical sources.

5 LHCb's statistical wish list

Having discussed the aspects of the LHCb experiment that will need a careful statistical analysis, we can give a specific list of topics that LHCb wishes to develop or to improve:

- A well supported tool for data modelling and fitting that can handle general multi-dimensional problems;
- A multivariate analysis tool that is capable of dealing with multiple discriminating variables with non-linear correlations and has a reliable mechanism to monitor and control over-training of classifiers;
- Better understanding of how to treat systematic and theoretical uncertainties;
- New statistical methods to improve flavour tagging;
- Better understanding of how to set confidence limits in case of insignificant signals;
- Recommendation on statistical procedures in data analysis.

References

[1] A. Hocker *et al.*, CERN-OPEN-2007-007.

[2] W. Bonivento and N. Serra, CERN-LHCb-2007-028.

[3] D. Martinez *et al.*, CERN-LHCb-2007-033.

[4] G. J. Feldman and R.D. Cousins, Phys. Rev. D57, 3873 (1998).

[5] T. Junk, CERN-EP/99-041.

[6] W.-M. Yao *et al.*, *Review of Particle Physics*, Journal of Physics G 33, 1 (2006).

[7] W. Verkerkear and D. Kirkby, Xiv:physics/0306116.

[8] S. T. Jampens's thesis, available at https://oraweb.slac.stanford.edu/pls/slacquery /BABAR_DOCUMENTS.DetailedIndex?P_BP_ID=3629.

[9] B. Aubert, Phys. Rev. D71, 032005 (2005).

[10] Private communication with Joe Boudreau at CDF.

[11] J. Linnemann and A. J. Smith, these proceedings.

[12] M. Bona *et al.*, UTfit Collaboration, J. High Energy Phys. 0610, 081 (2006), updated results available at http://www.utfit.org/.

[13] J. Charles *et al.*, CKMfitter Group, Eur. Phys. J. C41, 1 (2005), and updated results available at http://www.slac.stanford.edu/xorg/ckmfitter/.