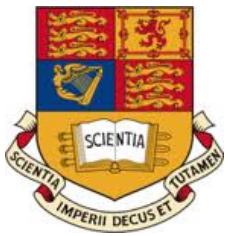
How Good Are Your Fits?

LHCb_Collaboration::Mike_Williams

Imperial College London

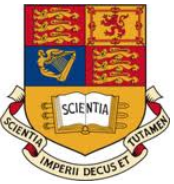
&

MIT



ATHOS 2012

June 21st, 2012



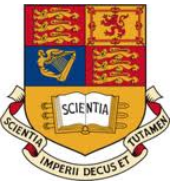
Introduction

A common situation in physics is that an unbinned maximum likelihood fit to the data has been performed to obtain estimators for any unknown parameters in a PDF (f_0).

The question is then "How well does the fit PDF describe the data?" Unfortunately, $\mathcal{L}$ does not provide any information that would help answer this question (beware of claims to the contrary!).

How do we answer this question? Typically we (somehow) determine the "p-value". The p-value is the probability that, if f_0 is the true PDF (f), a repeat of the experiment would have lesser agreement with the data than what we observe in our experiment*.

*The p-value is not the probability that $f_0 = f$. If $f_0 = f$, then the p-value distribution is uniform on $(0, 1)$ for an ensemble of data sets sampled from f.



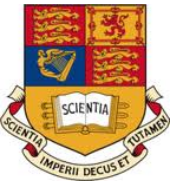
Binned Tests

If the data is dense enough, it can be binned. One would then be tempted to use the χ^2 test to get a p-value . . . but be careful!

If we started by binning the data and determined the estimators for all unknown parameters in f_0 by minimizing the χ^2, then the distribution of the χ^2 statistic ($g(\chi^2)$) is known if $f_0 = f$. Thus, we can determine the p-value from, *e.g.*, `TMath::Prob(chi2,nbins-npars)`.

The problem here is that we started by getting the estimators by maximizing $\mathcal{L}$. This means that we don't know g; all we know is that $\chi^2(n_b - n_p) \leq \chi^2 \leq \chi^2(n_b)$.

In other words, had we explicitly minimized χ^2 then its value would have been $\leq$ the one we obtain from the PDF that maximizes $\mathcal{L}$ (by definition). This is something to be aware of and a mistake that is often made in physics analyses.



Unbinned Tests

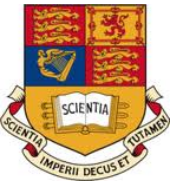
In many analyses, binning just isn't a viable option (high dimensions and/or low stats). Contrary to popular (physics) belief, there are many methods in the statistics literature that deal with these situations.

I put the available methods into 5 categories:

- mixed-sample methods;
- point-to-point dissimilarity methods;
- distance to nearest-neighbor methods;
- local-density methods;
- kernel-based methods.

Clearly, I don't have time to discuss them all*; in this (very short) talk I'll just discuss one method from the point-to-point dissimilarity category.

*See M. Williams, JINST 5, P09004 (2010) [arXiv:1006.3019] for a full review.



Point-to-Point Dissimilarity Methods

If the parent PDF, $f(\vec{x})$, of the data were known, then the GOF of a test PDF, $f_0(\vec{x})$, could be obtained using the statistic

$$T = \frac{1}{2} \int \left(f(\vec{x}) - f_0(\vec{x})\right)^2 d\vec{x}.$$

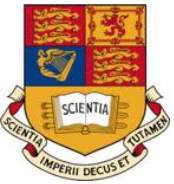
Since f is not known, T cannot be calculated. Of course, if f were known there would also be no reason to perform a fit.

A more general expression involves correlating the difference b/t f and f_0 at different points in the multivariate space using a weighting function:

$$T = \frac{1}{2} \int \int \left(f(\vec{x}) - f_0(\vec{x})\right) \left(f(\vec{x}') - f_0(\vec{x}')\right) \psi(|\vec{x} - \vec{x}'|) d\vec{x} d\vec{x}'.$$

N.b. the 1^{st} expression is the case $\psi(|\vec{x} - \vec{x}'|) = \delta(|\vec{x} - \vec{x}'|)$.

This quantity can be calculated w/o knowing f (using the data)!



Point-to-Point Dissimilarity Methods

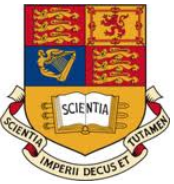
Using the data and a MC data set sampled from f_0, T can be written as

$$T = \frac{1}{n_d(n_d-1)} \sum_{j>i} \psi(|\vec{x}_i - \vec{x}_j|) + \frac{1}{n_{mc}(n_{mc}-1)} \sum_{j>i} \psi(|\vec{y}_i - \vec{y}_j|) - \frac{1}{n_d n_{mc}} \sum_{i,j} \psi(|\vec{x}_i - \vec{y}_j|).$$

Values of the weighting function found in the statistical literature:

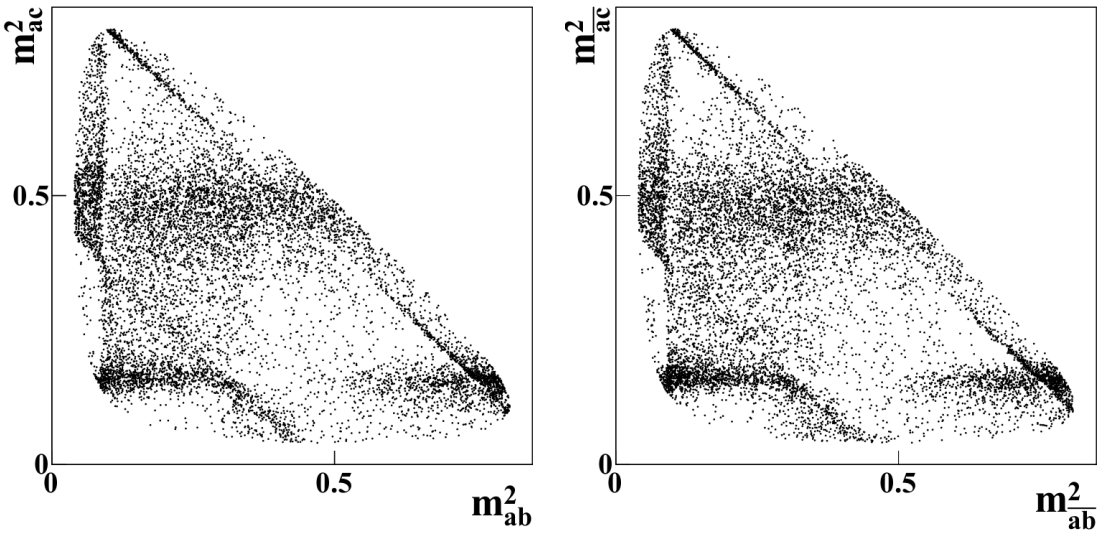
- $\psi(z) = z^2$ [C.M. Cuadras and J. Fortiana, various works (1997-2003)]
- $\psi(z) = z$ [L. Baringhaus and C. Franz, J. Multivariate Anal. **88** (2004) 190-206]
- $\psi(z) = \frac{1}{z}$, $-\log z$ or $e^{-z^2/2\sigma^2}$
  [B. Aslan and G. Zech, Stat. Comp. Simul. **75**, Issue 2 (2004) 109-119]

A&Z note that for $\psi(z) = \frac{1}{z}$ T is the electrostatic energy of two charge distributions w/ opposite sign (which is minimized if $f = f_0$); hence, they called it the "energy test".



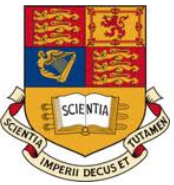
Example: CPV in Dalitz Plots

Dalitz model with a (very!) small amount of *CP* Violation



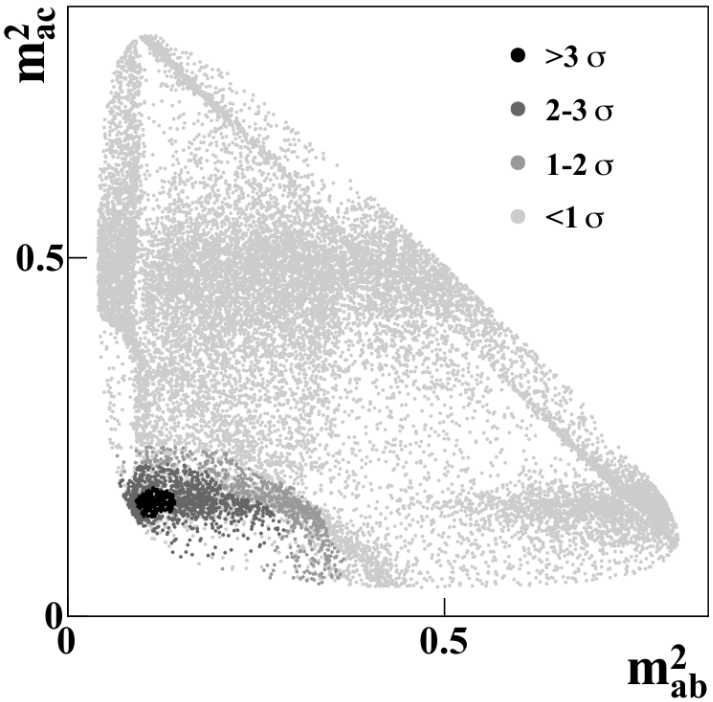
We can use the energy test to tell us if these two data sets are consistent with sharing the same parent distribution (model-independent method for observing *CP* violation!)*.

*M. Williams, PRD 84 054015 (2011) [arXiv:1105.5338]

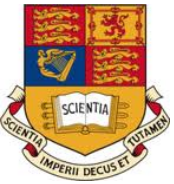


Example: CPV in Dalitz Plots

At the 95% confidence level, the no CPV hypothesis is rejected $(3 \pm 2)\%$ by the χ^2 test but $(52 \pm 5)\%$ by the energy test! The χ^2 test is blind to the CPV here but the energy test is powerful enough to detect it.

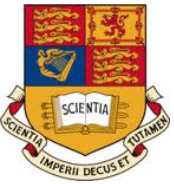


This test can also be used to perform non-parametric regression:
M. Williams, JINST 6, P10003 (2011) [arXiv:1107.2285].



Etc.

- Choose your test before you look at the data (or, at least, before you calculate the p-value)! You can always find a test that will reject the true PDF and another that will except (almost any) alternative PDF.
- Resolution in this test (and any other MC based one) can be handled when generating the MC. Alternatively, resolution can be put into the PDF directly.
- A small p-value could mean either you don't understand the physics or the detector or both! The PDF is a product of the two and no statistical test (using just the data and MC) can determine which is the culprit.



Summary

- For binned analyses, the χ^2 test provides a simple way of obtaining the GOF (but be careful as there are a few common mistakes that can bite you).
- For unbinned analyses, there are many GOF tests on the market. Many are very powerful (some are not!). Some are slow; some are fast; *etc*. Choose the right tool for the job.
- It's always a good idea to validate a GOF method for your analysis in MC. Any true GOF method produces a uniform p-value distribution when the true PDF is used.