EVALUATING QUALITY OF FIT IN UNBINNED MAXIMUM LIKELIHOOD FITTING

K. Kinoshita
University of Cincinnati, Cincinnati, Ohio USA
Belle Collaboration

Abstract
The unbinned maximimum likelihood fitting method, used in many current analyses including Belle's measurements of sin2ϕ_1 and lifetimes, maximizes the use of available information to obtain the shape of a distribution in the face of limited statistics. However, a significant difficulty is that there has been no method for evaluating goodness-of-fit for the result. We examine some issues surrounding this question and conclude that an unbinned goodness-of-fit test may be possible.

1 INTRODUCTION

The measurement of sin2ϕ_1 at Belle[1] includes a procedure of fitting the distribution of events in a measured quantity (proper time difference $\Delta\tau$) to a function whose shape is dependent on sin2ϕ_1. If the number of events is sufficient, the distribution may be binned and fitted via the least squares or binned maximum likelihood methods, and the quality of results obtained by different methods is approximately equivalent. For the first measurements, however, the number of events is limited, and to most effectively use all available information, the unbinned maximum likelihood method (*UMxL*) has been used.

One troublesome aspect of fitting with this method has been the lack of any prescription for evaluating goodness-of-fit, to determine whether the data are statistically consistent with the fitted shape. A poor fit could indicate that the fitting function is inappropriate or inaccurate, or, if there is background, that it is not correctly estimated. For any result to be credible, it is important to know that the fits have reasonable confidence. For binned methods, such as least-squares, the chisquare value gives a straightforward measure of confidence. To date, no analogous method has been found for *UMxL*, and it has been necessary to use other means, such as reverting to binning or generating "toy Monte Carlo" distributions, to make this determination. We report here some preliminary findings from a search for a formal means of evaluating goodness of fit within *UMxL*.

2 OUTLINE OF THE *UMxL* METHOD

We begin with an experiment that has collected N events in which the quantity x, which may be multidimensional, is measured for each event. The set $\{x_i\}$ is fitted to a normalized probability distribution function (p.d.f.), $f(x;\alpha)$, where the unknown parameter α may also be multidimensional, by maximizing the likelihood,

$$\mathcal{L}(\alpha) = \prod_i f(x_i;\alpha),$$

with respect to α. α_{max} is defined to be the value of α that maximizes $\mathcal{L}(\alpha)$ in a given data set. It is equivalent, and somewhat simpler, to maximize

$$\ln\mathcal{L}(\alpha) \equiv \lambda(\alpha) = \sum_i^N \ln f(x_i;\alpha)$$

The procedure is described in Ref. [2].

3 ENSEMBLE DISTRIBUTIONS OF λ

In the method of least squares, the χ^2 value is a measure of the magnitude of fluctuations of a distribution from the fitted function. There is a well-established prescription for goodness-of-fit whereby one can determine whether the value is statistically consistent with the fitted distribution. Analogously, one might look at the distribution in $\lambda(\alpha_{max})$ over an ensemble of experiments to determine fit quality. As we will show, however, the $\lambda(\alpha_{max})$ value itself is not useful. Nevertheless, it is instructive to explore this avenue, as it may provide insight into the reasons and suggest promising directions.

One would first like to determine how the $\lambda(\alpha_{max})$ are distributed over an ensemble of experiments with N events each, where the true probability density is $f(x;\bar{\alpha})$. The distribution is presumably peaked, with a well-defined mean value and finite width due to statistical fluctuations. We first consider a distribution with fixed parameters, where the distribution is straightforwardly obtainable, and then examine the effect of allowing parameter variations, where the picture is considerably different.

3.1 Fixed Distributions

If no parameters are varied in the fitting function $f(x;\bar{\alpha})$, then the likelihood is well-defined. The value of $\lambda(\bar{\alpha})/N$ is the average value of $\ln f(x;\bar{\alpha})$ over the dataset $\{x_i\}$. If the true p.d.f. is also $f(x;\bar{\alpha})$, and N is very large, $\lambda(\bar{\alpha})/N$ approaches the mean $\ln f(x;\bar{\alpha})$ over the distribution $f(x;\bar{\alpha})$, or

$$\lim_{N\to\infty} \lambda(\bar{\alpha}) = N \int dx f(x;\bar{\alpha}) \ln f(x;\bar{\alpha}) \equiv N\hat{\lambda}(\bar{\alpha})$$

which is equal to the ensemble mean for finite N. The statistical variance of $\lambda(\bar{\alpha})$ is[3]

$$\begin{aligned}\lim_{N\to\infty} V[\lambda(\bar{\alpha})] &= N\{\int dx f(x;\bar{\alpha})[\ln f(x;\bar{\alpha})]^2 - \hat{\lambda}^2(\bar{\alpha})\} \\ &\equiv N\hat{\sigma}^2_\lambda(\bar{\alpha})\end{aligned}$$

It can be shown that the distribution in $\lambda(\bar{\alpha})/N$ is Gaussian for large N. Distributons in $\lambda(\bar{\alpha})/N$ for other p.d.f.'s fitted to $f(x;\bar{\alpha})$ may be found in a similar way. One may then construct a test statistic $X^2 \equiv [\lambda(\bar{\alpha}) - N\hat{\lambda}(\bar{\alpha})]^2/N\hat{\sigma}^2_\lambda(\bar{\alpha})$ where large values imply incompatibility between the hypothesis and data. A small value of X^2 signals consistency with the hypothesis but is not as strong as a true goodness-of-fit test in that it can not rule out all alternative hypotheses. What it is able to do is to place limits on specific classes of alternatives, which for many applications may be considered sufficient.

3.2 Distributions with free parameters

More usually, α is allowed to vary, and the best estimate for α, α_{max}, is that for which $\lambda(\alpha)$ is maximum. The ensemble distribution in $\lambda(\alpha_{max})$ is related to but different from that in $\lambda(\bar{\alpha})$. For each experiment, $\lambda(\alpha_{max}) \geq \lambda(\bar{\alpha})$ so the distribution must shift in the positive direction. If the shifts could be characterized, one might expect to be able to extract the goodness-of-fit by comparing $\lambda(\alpha_{max})$ to the expected mean and width. However, a recent study by J. Heinrich[4] indicates that the maximum log likelihood value may be equal to the expected value for the fitted value of α_{max}:

$$\lambda(\alpha_{max}) = N\hat{\lambda}(\alpha_{max}). \quad (Heinrich's\ conjecture) \tag{1}$$

He studied two specific p.d.f.'s, with Monte Carlo and analytically. If this is observation holds more generally, as Heinrich speculates, the "confidence level" is always 100% and the likelihood value contains no information about the goodness of fit. We first examine this conjecture and then return to the question of the ensemble distribution of $\lambda(\alpha_{max})$.

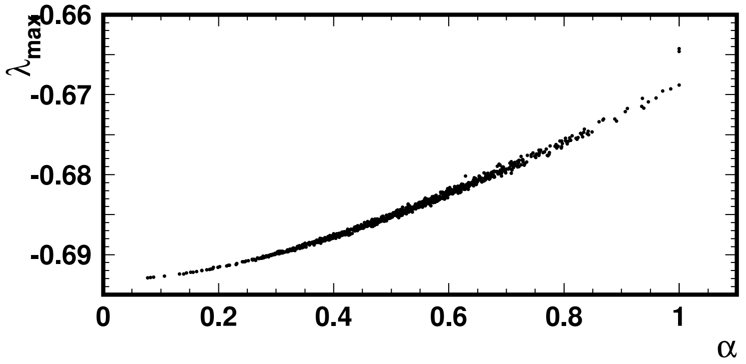

Fig. 1: Distribution of $\lambda(\alpha_{max})/N$ vs. α_{max} for $f(x;\alpha) = \frac{1+\alpha x^2}{2(1+\alpha/3)}$, $\bar{\alpha} = 0.5$, and $N = 1000$.

3.21 Heinrich's conjecture

It is convenient to rewrite the parametrized p.d.f. as $f(x;a) = n(\alpha)e^{-h(x;\alpha)}$ and the set of measurements, $\{x_i\}$, as a "measured p.d.f.," $g(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i)$, so that

$$\lambda(\alpha) = N\int dx\ g(x)\ln f(x;\alpha) = N\int dx\ g(\ln n - h) = N(\ln n - \langle h\rangle) \tag{2}$$

where $\langle h\rangle \equiv \int dx\ gh$ is the "expectation value" of the function h over the measured p.d.f., g. At the maximum, the first derivatives $\frac{\partial\lambda}{\partial\alpha_i}$ are zero:

$$\frac{\partial\lambda}{\partial\alpha_i} = 0 = N(\frac{\partial\ln n}{\partial\alpha_i} - \langle\frac{\partial h}{\partial\alpha_i}\rangle).$$

One can see that, once α_{max} is determined, the values of $\langle\frac{\partial h}{\partial\alpha_i}\rangle_{\alpha_{max}}$ are also fully defined. To establish the $\lambda(\alpha_{max})$ value, one needs in addition the value of $\langle h\rangle_{\alpha_{max}}$.

Looking at the two p.d.f.'s examined by Heinrich, it is now clear that they are special cases where $\langle h\rangle_{\alpha_{max}}$ is always fully determined by $\langle\frac{\partial h}{\partial\alpha_i}\rangle_{\alpha_{max}}$ $(i.e., h = \sum_i k_i(\alpha)\frac{\partial h}{\partial\alpha_i})$:

- $f(x;\alpha) = \frac{1}{\alpha}e^{-x/\alpha}$: here, $h = x/\alpha$, so $\frac{\partial h}{\partial\alpha} = -h/\alpha$, and $\langle\frac{\partial h}{\partial\alpha}\rangle_{\alpha_{max}} = -\langle h\rangle_{\alpha_{max}}/\alpha_{max}$.
- $f(x;\alpha_1,\alpha_2) = \frac{1}{\sqrt{2\pi}\alpha_2}exp(\frac{-(x-\alpha_1)^2}{2\alpha_2^2})$: here the relevant constraint is $\frac{\partial h}{\partial\alpha_2} = -2h/\alpha_2$.

In other words, $\lambda(\alpha_{max})$ is to be found with no further input from data:

$$\lambda(\alpha_{max}) = N(\ln n(\alpha_{max}) - \sum_i k_i(\alpha_{max})\frac{\partial\ln n}{\partial\alpha_i}|_{\alpha_{max}}).$$

While these are special cases where the relationship between $\langle h\rangle_{\alpha_{max}}$ and $\langle\frac{\partial h}{\partial\alpha_i}\rangle_{\alpha_{max}}$ has no dependence on data, this derivation shows that many cases may have partial correlations. For example, taking

$$f(x;\alpha) = \frac{1+\alpha x^2}{2(1+\alpha/3)}$$

where $|x| < 1$, it is clear that h and its derivatives have different dependences on x. However, if h and $\frac{\partial h(x;\alpha)}{\partial\alpha}$ are displayed as power series expansions:

$$\langle h(x;\alpha)\rangle = -\langle\ln(1+\alpha x^2)\rangle = \sum_{i=1}^{\infty}\frac{(-1)^i\langle(\alpha x^2)^i\rangle}{i} = -\alpha\langle x^2\rangle + \frac{\alpha^2\langle x^4\rangle}{2} + \ ...$$

$$\langle\frac{\partial h(x;\alpha)}{\partial\alpha}\rangle = \frac{1}{\alpha}\sum_{i=1}^{\infty}(-1)^i\langle(\alpha x^2)^i\rangle = \frac{1}{\alpha}[-\alpha\langle x^2\rangle + \alpha^2\langle x^4\rangle + \ ....]$$

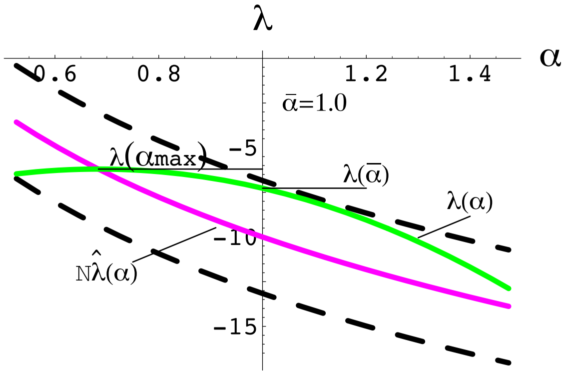

Fig. 2: Relation of $\lambda(\alpha_{max})$ to $\lambda(\bar{\alpha})$ for $f(x;\alpha) = \frac{1}{\alpha}e^{-x/\alpha}$, $N = 10$, and $\bar{\alpha} = 1.0$. The magenta curve is the expected mean, $N\hat{\lambda}(\alpha)$, the dashed curves are the $1-\sigma$ envelope, $N\hat{\lambda}(\alpha) \pm \sqrt{N\sigma_{\hat{\lambda}}^2}$, and the green curve is $\lambda(\alpha)$ for one experiment, with the indicated value of $\lambda(\bar{\alpha})$.

it is clear that there are shared terms. If the true distribution follows the same parametrization, the statistical fluctuations about the remaining uncorrelated part result in small fluctuations of $\lambda(\alpha_{max})$ about the expected value. Figure 1 shows the two-dimensional distribution of α_{max} and $\lambda(\alpha_{max})$. In this case the fluctuations are small compared to the total statistical spread of $\lambda(\alpha_{max})$, so that Heinrich's conjecture holds approximately. The magnitude of the fluctuations relative to the overall spread in $\lambda(\alpha_{max})$ depends on the functional form of the p.d.f.. Distributions that differ greatly from the parameterized form may have $\lambda(\alpha_{max})$ values that diverge from $N\hat{\lambda}(\alpha)$, but no quantitative study has yet been made. In any case, the lack of a universal recipe would seem to be a serious impediment to developing a goodness-of-fit test.

3.22 Distribution in $\lambda(\alpha_{max})$

While we have concluded that $\lambda(\alpha_{max})$ by itself cannot contain goodness-of-fit information, there is still some motivation for studying its ensemble distribution. For example, if one has a combined fit over different sets or types of data, the values of $\lambda(\alpha_{max})$ in subsets of the data may be compared with the expectation, as described in Section 3.1, to determine whether they are individually and collectively consistent with the overall result. In addition, a formal method of deriving the distribution would provide a quick means of gleaning information equivalent to that produced by the cumbersome "toy Monte Carlo" method, facilitating a more thorough exploration of statistical and fitting issues than would otherwise be practical. Knowledge of the ensemble distribution may also allow for a better determination of confidence limits on parameter values. We present here a very preliminary result on this topic. If it can be assumed that $\lambda(\alpha_{max}) = N\hat{\lambda}(\alpha_{max})$[Eq. (1)] is a good approximation, then for a given experiment, $\lambda(\alpha_{max})$ may be found from the $\lambda(\bar{\alpha})$ value by noting that $\lambda(\alpha)$, which is an inverted parabola in the region of interest, intersects $N\hat{\lambda}(\alpha)$ at $\alpha = \alpha_{max}$ with zero slope. A value for the second derivative of $\lambda(\alpha)$ can be found from Eq. (2) with $g(x) = f(x;\alpha_{max})$. Figure 2 illustrates this reasoning. If the average change in $|\alpha_{max} - \bar{\alpha}|$ is of the order of the error on α_{max}, σ_α, then the average $\lambda(\alpha_{max}) - \lambda(\bar{\alpha})$ is (by definition) of order 0.5, which suggests that for each parameter the shift beween the mean $\lambda(\bar{\alpha})$ and the mean $\lambda(\alpha_{max})$ is of order 0.5. A simple numerical test for several cases with one parameter, shown in Table 1, appears to support this hypothesis. We expect to report a more rigorous discussion in the near future.

A corollary of this finding is that the process of maximizing $\lambda(\alpha)$ causes the mean α_{max} to be shifted from the true value. A shift of 0.5 in λ is $0.5/\sqrt{N\hat{\sigma}_{\hat{\lambda}}^2}$ times the statistical error, and one would expect α_{max} to also be shifted by the corresponding amount, in the direction of increasing $\hat{\lambda}(\alpha)$.

$f(x;\alpha)$	x range	$\bar{\alpha}$	N	mean $\lambda(\bar{\alpha})$	mean $\lambda(\alpha_{max})$	$\Delta\lambda$
$\frac{1}{\alpha}e^{-x/\alpha}$	$[0,\infty]$	1.0	10	-16.93 ± 0.01	-16.42 ± 0.01	0.51 ± 0.01
$\frac{1}{\alpha}e^{-x/\alpha}$	$[0,\infty]$	1.0	100	-169.5 ± 0.1	-169.0 ± 0.1	0.5 ± 0.1
$\frac{1+\alpha x^2}{2(1+\alpha/3)}$	$[-1,+1]$	0.5	1000	-685.1 ± 0.1	-684.6 ± 0.1	0.5 ± 0.1

Table 1. Difference between mean $\lambda(\alpha_{max})$ and $\lambda(\bar{\alpha})$ for three numerical Monte Carlo experiments.

4 ADDITIONAL INFORMATION FROM DATA

Thus far in this search for a goodness-of-fit test for *UMxL*, it has been demonstrated that there is a correlation between the maximum likelihood value and the value of fitted parameters in the *UMxL* method of parameter estimation and that the degree of correlation depends on the shape of the parameterized function. While by not binning the data the resolution of the measured quantity x is used to best advantage in estimating α, the information that comes out of the fit, α_{max} and $\lambda(\alpha_{max})$, depends on only a few averages over the dataset, $\langle h\rangle_{\alpha_{max}}$ and $\langle\frac{\partial h}{\partial\alpha_i}\rangle_{\alpha_{max}}$. One would think that more information can be extracted from the data, and that some of it may be relevant to goodness-of-fit.

Another way to understand the relationship of $\lambda(\alpha_{max})$ to the data can be seen from equation (1), which shows that $\lambda(\alpha)/N$ is simply the average of $\ln f(x_i;\alpha)$ over the dataset $\{x_i\}$. If the true p.d.f. is $f(x;\alpha_{max})$, then the distribution of $\ln f(x_i;\alpha_{max}) \equiv \tilde{\lambda}_i$ in the limit of large N is

$$\Lambda(\tilde{\lambda}) = \int dx\ f(x;\alpha_{max})\delta(\tilde{\lambda} - \ln f(x;\alpha_{max})).$$

$\lambda(\alpha_{max})/N$ is then the first moment of Λ, and its ensemble distribution approaches a Gaussian for large N, according to the Central Limit Theorem. In principle, higher moments can provide additional information about the data. Or, more simply, one might test goodness-of-fit by performing a Kolmogorov-Smirnov test on $\{\tilde{\lambda}_i\}$. If x is multidimensional this is an attractive possiblity, since there exists thus far no satisfactory multidimensional goodness-of-fit test. This will be investigated in the near future.

5 CONCLUSION

In searching for an unbinned goodness-of-fit test for *UMxL*, we have examined the information content of $\lambda(\alpha_{max})$, starting with the ensemble distribution for a p.d.f. with fixed parameters and following the effect of varying parameters to maximize $\lambda(\alpha)$. The behavior of $\lambda(\alpha_{max})$ in this process reveals that a conjecture by J. Heinrich is approximately true and leads to the reasons why the $\lambda(\alpha_{max})$ value itself provides little or no information about fit quality. At the same time, it suggests some paths that may lead to a workable goodness-of-fit test. A preliminary calculation of the mean of the ensemble distribution of $\lambda(\alpha_{max})$ reveals a shift due to maximization which results in a systematic shift of the fitted parameter(s) from the true values with a well-defined magnitude and direction.

ACKNOWLEDGEMENTS

I thank the conference organizers for their hospitality and for facilitating many lively discussions.

References

[1] A. Abashian *et al.*, *Physical Review Letters* **86**, 2509 (2001); K. Abe *et al.*, *Physical Review Letters* **87**, 091802 (2001).

[2] D. E. Groom *et al.*, *European Physical Journal* **C15**, 1 (2000).

[3] W.T. Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics* (North-Holland Publishing Co., Amsterdam, 1971), p. 271.

[4] J. Heinrich, "Can the Likelihood Function Value Be Used to Measure Goodness-of-Fit?" /CDF/MEMO/BOTTOM/CDFR/5639 (unpublished).