

Goodness of Fit: What Do We Really Want to Know?

I. Narsky

California Institute of Technology, Pasadena, CA 91125, USA

Definitions of the goodness-of-fit problem are discussed. A new method for estimation of the goodness of fit using distance to nearest neighbor is described. Performance of several goodness-of-fit methods is studied for time-dependent CP asymmetry measurements of $\sin(2\beta)$.

1. INTRODUCTION

The goodness-of-fit problem has recently attracted attention from the particle physics community. In modern particle experiments, one often performs an unbinned likelihood fit to data. The experimenter then needs to estimate how accurately the fit function approximates the observed distribution. A number of methods have been used to solve this problem in the past [1], and a number of methods have been recently proposed [2, 3] in the physics literature.

For binned data, one typically applies a χ^2 statistic to estimate the fit quality. Without discussing advantages and flaws of this approach, I would like to stress that the application of the χ^2 statistic is limited. The χ^2 test is neither capable nor expected to detect fit inefficiencies for all possible problems. This is a powerful and versatile tool but it should not be considered as the ultimate solution to every goodness-of-fit problem.

There is no such popular method, an equivalent of the χ^2 test, for unbinned data. The maximum likelihood value (MLV) test has been frequently used in practice but it often fails to provide a reasonable answer to the question at hand: how well are the data modelled by a certain density [4]? It is only natural that goodness-of-fit tests for small data samples are harder to design and less versatile than those for large samples. For small data samples, asymptotic approximations do not hold and the performance of every goodness-of-fit test needs to be studied carefully on realistic examples. Thus, the hope for a versatile unbinned goodness-of-fit procedure expressed by some people at the conference seems somewhat naive.

A more important practical question is how to design a powerful goodness-of-fit test for each individual problem. It is not possible to answer this question unless we specify in more narrow terms the problem that we are trying to solve.

2. WHAT IS A GOODNESS-OF-FIT TEST?

A hypothesis test requires formulation of null and alternative hypotheses. The confidence level, $1 - \alpha_I$,

of the test is then defined as the probability of accepting the null hypothesis given it is true, and the power of the test, $1 - \alpha_{II}$, is defined as the probability of rejecting the null hypothesis given the alternative is true. Above, α_I and α_{II} denote Type I and Type II errors, respectively. An ideal hypothesis test is uniformly most powerful (UMP) because it gives the highest power among all possible tests at the fixed confidence level. In most realistic problems, it is not possible to find a UMP test and one has to consider various tests with acceptable power functions.

There is a long-standing controversy about the connection between hypothesis testing and the goodness-of-fit problem. It can be argued [5] that there can be no alternative hypothesis for the goodness-of-fit test. In this approach, however, the experimenter does not have any criteria for choosing one goodness-of-fit procedure over another. One can design a goodness-of-fit test using first principles, advanced computational methods, rich intuition or black magic. But the practitioner wants to know how well this method will perform in specific situations. To evaluate this performance, one needs to study the power of the proposed method against a few specific alternatives. A certain, perhaps vague, notion of an alternative hypothesis must be adopted for this exercise; hence, a certain, perhaps vague, notion of the alternative hypothesis is typically used to design a goodness-of-fit test.

Consider, for example, testing uniformity on an interval. The alternative is usually perceived as presence of peaks in the data. Suppose we design a procedure that gives the highest goodness-of-fit value for equidistant experimental points. This test will perform well for the chosen alternative. In reality, however, we may need to test exponentiality of the process. For instance, we use a Geiger counter to measure elapsed time between two consecutive events and plot these time intervals next to each other on a straight line. In this case, equidistant data would imply that the process is not as random as we thought, and the designed goodness-of-fit procedure would fail to detect the inconsistency between the data and the model. Tests against highly structured data (e.g., equidistant one-dimensional data) have been, in fact, a subject of statistical research on goodness-of-fit methods.

The question therefore is how to state the alternative hypothesis in a way appropriate for each individ-

ual problem. I emphasize that I am not suggesting to use a directional test for one specific well-defined alternative. The goal is to design an omnibus goodness-of-fit test that discriminates against at least several plausible alternatives.

The null hypothesis is defined as

$$H_0 : X \sim f(x|\theta_0, \eta) , \quad (1)$$

where X is a multivariate random variable, and f is the fit density with a vector of arguments x , vector of parameters θ and vector of nuisance parameters η . The alternative hypothesis is stated in the most general way as

$$H_1 : X \sim g(x) \quad \text{with} \quad g(x) \neq f(x|\theta_0, \eta) . \quad (2)$$

A specific subclass of this alternative hypothesis that is sometimes of interest is expressed as

$$H_1 : X \sim f(x|\theta, \eta) \quad \text{with} \quad \theta \neq \theta_0 . \quad (3)$$

In other words, most usually we would like to test the fit function against different shapes (2). For the test (3), we assume that the shape of the fit function is correctly modelled and we only need to cross-check the value of the parameter.

If a statistic $S(x)$ is used to judge the fit quality, the goodness-of-fit is given by

$$1 - \alpha_I = \int_{f_S(s) > f_S(s_0)} f_S(s) ds , \quad (4)$$

where s_0 is the value of the statistic observed in the experiment, and $f_S(s)$ is the distribution of the statistic under the null hypothesis.

In practice, the vector of parameter estimates θ_0 is usually extracted from an unbinned maximum likelihood (ML) fit to the data: $\theta_0 = \hat{\theta}(x)$. In this case, the goodness-of-fit statistic must be independent of, or at most weakly correlated to, the ML estimator of the parameter: $\rho(S(x), \hat{\theta}(x)) \approx 0$, where ρ is the correlation coefficient computed under the null hypothesis. If $S(x)$ and $\hat{\theta}(x)$ are strongly correlated, the goodness-of-fit test is redundant.

The ML estimator itself is usually a powerful tool for discrimination against the alternative (3). In this case, the statistic $S(x) \neq \hat{\theta}(x)$ can be treated as an independent cross-check of the parameters θ_0 .

The nuisance parameters η should affect our judgment about the fit quality as little possible. Discussion of methods for handling nuisance parameters is beyond the scope of this note.

3. DISTANCE TO NEAREST NEIGHBOR TEST

The idea of using Euclidian distance between nearest observed experimental points as a goodness-of-fit measure is not new. Clark and Evans [6] used an average distance between nearest neighbors to test two-dimensional populations of various plants for uniformity. Later they extended this formalism to a higher number of dimensions. Diggle [7] proposed to use an entire distribution of ordered distances to nearest neighbors and apply Kolmogorov-Smirnov or Cramer-von Mises tests to evaluate consistency between experimentally observed and expected densities. Ripley [8] introduced a function, $K(t)$, which represents a number of points within distance t of an arbitrary point of the process; he used the maximal deviation between expected and observed $K(t)$ as a goodness-of-fit measure. Bickel and Breiman [9] introduced a goodness-of-fit test based on the distribution of the variable $\exp(-Nf(x_i)V(x_i))$, where $f(x_i)$ is the expected density at the observed point x_i , $V(x_i)$ is the volume of a nearest neighbor sphere centered at x_i and N is the total number of observed points. An approach closely related to distance-to-nearest-neighbor tests is two-sample comparison based on counts of nearest neighbors that belong to the same sample [10]. These methods received substantial attention from the science community and have been applied to numerous practical problems, mostly in ecology and medical research. Ref. [11] offers a survey of distance-to-nearest-neighbor methods.

The goodness-of-fit test [3] uses a bivariate distribution of the minimal and maximal distances to nearest neighbors. First, one transforms the fit function defined in an n -dimensional space of observables to a uniform density in an n -dimensional unit cube. Then one finds smallest and largest clusters of nearest neighbors whose linear size maximally deviates from the average cluster size predicted from uniformity. The cluster size is defined as an average distance from the central point of the cluster to m nearest neighbors. If the experimenter has no prior knowledge of the optimal number m of nearest neighbors included in the goodness-of-fit estimation, one can try all possible clusters $2 \leq m \leq N$, where N is the total number of observed experimental points. The probability of observing the smallest and largest clusters of this size gives an estimate of the goodness of fit and the locations of the clusters can be used to point out potential problems with data modelling.

This method is a good choice for detection of well-localized irregularities, e.g., unusual peaks in the data. Consider, for example, fitting a normal peak on top of the smooth background, as shown in Fig. 3. The likelihood function is sensitive to the mean and width of the normal component. Hence, if the experimenter is mostly interested in how accurately these parameters

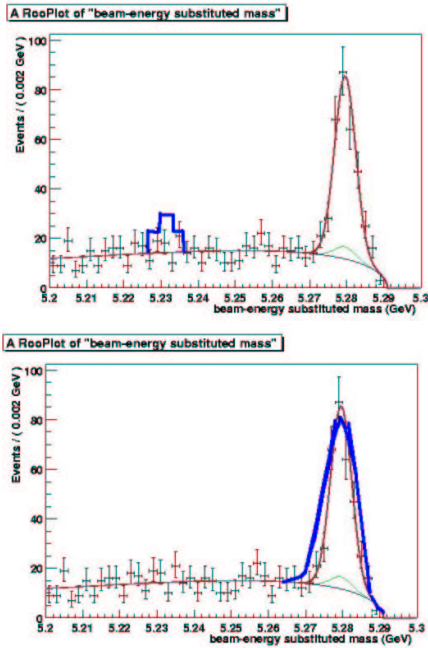


Figure 1: Fits to the sum of a normal signal density and smooth background. Possible deviations of data from the fit function are shown with a thick line. The data exhibit an unusual peak in the background tail (top), and the normal peak in the data is wider than the normal component in the fit function (bottom).

are modelled, the likelihood function can be used to address this question. At the same time, the likelihood shows little sensitivity to the bump in the smooth background tail of the distribution. If the experimenter wants to be aware of such irregularities, the method [3] is a good choice.

The transformation to uniformity is a common technique used in goodness-of-fit methods. One should be kept in mind, however, that a multivariate transformation to uniformity is not necessarily unique. If a different uniformity transformation is chosen, one can obtain a different goodness-of-fit value. One solution to this ambiguity is discussed in Ref. [3].

4. COMPARISON OF GOODNESS-OF-FIT METHODS FOR $\sin(2\beta)$ MEASUREMENTS

I apply several methods summarized in Table I to unbinned ML fits in time-dependent CP asymmetry measurements and compare their power functions at fixed confidence levels. The fit function in $\sin(2\beta)$

Table I Statistics that can be applied to an unbinned ML fit in a $\sin(2\beta)$ measurement. For convenience, $\sin(2\beta)$ is replaced in all formulas with θ . Here, the random variable u is obtained by uniformity transformation $u = \int_{-t}^{t} f(t|\theta_0) dt$, b_j is a Legendre polynomial of order j , F_n is the experimental cumulative density function (CDF), F is the CDF under the null hypothesis, the upper bar denotes averaging, and n is the total number of events observed. The lifetime τ was allowed to vary in the fit.

Method	Formula
likelihood ratio	$L(\theta_0 x)/L(\hat{\theta} x)$
ML estimator	$\hat{\theta}$
score statistic	$(\partial L(\theta x)/\partial\theta) _{\theta=\theta_0}$
MLV	$L(\theta_0 x)$
Neyman with $K = 1, 2, 3$	$\sum_{j=1}^K \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n b_j(u_i) \right)^2$
Kolmogorov-Smirnov	$\max_{i=1,2,\dots,n} F_n(u_i) - F(u_i) $
Watson U^2	$\int_0^1 (F_n(u) - u)^2 du - \left(\bar{u} - \frac{1}{2} \right)^2$
Anderson-Darling	$\int_0^1 \frac{(F_n(u) - u)^2}{u(1-u)} du$
min distance to nearest neighbor	$d_{\min}(t < 0)/d_{\min}(t > 0)$
max distance to nearest neighbor	$d_{\max}(t < 0)/d_{\max}(t > 0)$

measurements is given by [12]

$$f(t|\sin(2\beta)) = \frac{1}{2\tau} \exp\left(\frac{-|t|}{\tau}\right) \left[1 + \sin(2\beta) \sin(\Delta mt) \right], \quad (5)$$

where positive and negative times t correspond to B tags of opposite flavors in the range $-8 \text{ ps} \leq t \leq 8 \text{ ps}$, $\sin(2\beta)$ is a measure of the asymmetry, $\tau = 1.542 \pm 0.016 \text{ ps}$ is the lifetime of the B meson [13], and $\Delta m = 0.489 \pm 0.009 \text{ ps}^{-1}$ [13] is the $B\bar{B}$ mass mixing. For toy Monte Carlo experiments, I generate samples with 100 events, a typical sample size in a BaBar analysis of $B \rightarrow J/\psi K_S$ decays in which the J/ψ is reconstructed in hadronic final states [12]. I ignore smearing due to detector resolution and background contributions to the data.

I test the null hypothesis $H_0 : \sin(2\beta) = 0.78$ against $H_1 : \sin(2\beta) \neq 0.78$ and plot correlations $\rho(S(x), \hat{\theta}(x))$ for each listed statistic under the null hypothesis, as well as power functions estimated from Monte Carlo samples generated with different values of $\sin(2\beta)$: 0, 0.5, 0.7, and 1. The power functions for $\sin(2\beta) = 0.5$ and the correlations are shown in Fig. 4. While the top three methods in Table I provide good separation between values of $\sin(2\beta)$, they cannot be used as independent cross-checks of the parameter because of the strong correlation with the ML estimator. The MLV method shows a relatively strong correlation with the ML estimator and a relatively poor power function. The three Neyman smooth tests, as well as Kolmogorov-Smirnov, Watson and Anderson-Darling

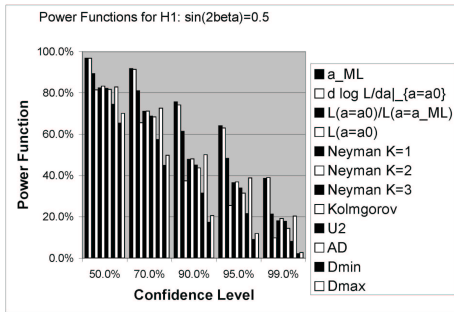
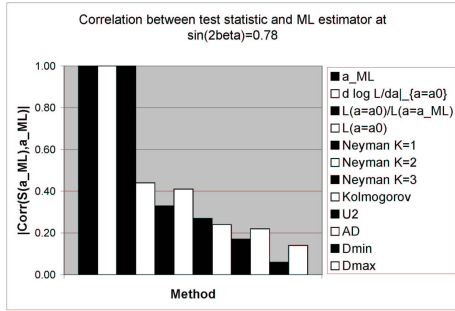


Figure 2: Correlations between the ML estimator of $\sin(2\beta)$ and the chosen statistic (top). Power functions of the hypothesis test $H_0 : \sin(2\beta) = 0.78$ against $H_1 : \sin(2\beta) \neq 0.78$ versus confidence level (bottom) at $\sin(2\beta) = 0.5$.

tests, show small correlation to the ML estimator and decent power functions; these tests perform competitively among each other. Yet Kolmogorov-Smirnov and Anderson-Darling tests produce somewhat better combinations of the small correlation and large power function and should be preferred over others. The distance-to-nearest-neighbor test was designed for detection of well-localized regularities and hence was not expected to give a high power function for the hypothesis test discussed here.

This exercise alone is insufficient to conclude that the two recommended tests are in fact the best omnibus tests for fits of $\sin(2\beta)$. One would have to extend this study to include other alternative densities, e.g., specific background shapes, that can distort the experimental data.

5. SUMMARY

An acceptable goodness-of-fit test is defined as an omnibus test that discriminates against at least sev-

eral plausible alternatives. Numerous distance-to-nearest-neighbor methods for goodness-of-fit estimation have been described in the statistics literature and should be tested in HEP practice. The distance-to-nearest-neighbor test based on minimal and maximal distances should be used for detection of well-localized irregularities in the data. Correlation coefficients and power functions for several statistics are compared for fits of $\sin(2\beta)$ in CP asymmetry measurements for one specific alternative.

Acknowledgments

I wish to thank the organizing committee of PHYS-TAT2003 for their effort. Thanks to Bob Cousins for useful comments on this note.

Work partially supported by Department of Energy under Grant DE-FG03-92-ER40701.

References

- [1] R. D’Agostino and M. Stephens, “Goodness-of-Fit Techniques”, Marcel Decker, Inc., 1986; J. Rayner and D. Best, “Smooth Tests of Goodness of Fit”, Oxford Univ. Press, 1989.
- [2] B. Aslan and G. Zech, “A new class of binning free, multivariate goodness-of-fit tests: The energy tests”, hep-ex/0203010, 2002.
- [3] I. Narsky, “Estimation of Goodness-of-Fit in Multidimensional Analysis Using Distance to Nearest Neighbor”, physics/0306171, 2003.
- [4] J. Heinrich, “Can the likelihood function be used to measure goodness of fit?”, CDF/MEMO/BOTTOM/CDFR/5639, Fermilab; also in these Proceedings.
- [5] See, for example, F. James’ talk at <http://www-conf.slac.stanford.edu/physstat2003/talks/james/james-slac.slides.ps>
- [6] P.J. Clark and F.C. Evans, “Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations”, Ecology **35-4**, 445 (1954); “Generalization of a Nearest Neighbor Measure of Dispersion for Use in K Dimensions”, Ecology **60-2**, 316 (1979).
- [7] P. Diggle, “On Parameter Estimation and Goodness-of-Fit Testing for Spatial Point Patterns”, Biometrics **35**, 87 (1979).
- [8] B.D. Ripley, “Modelling Spatial Patterns”, J. of the Royal Stat. Soc. **B 39-2**, 172 (1977).
- [9] P.J. Bickel and L. Breiman, “Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test”, Ann. of Probability **11**, 185 (1983).
- [10] J.H. Friedman and L.C. Rafsky, “Multivariate Generalizations of the Wald-Wolfowitz and

Smirnov Two-Sample Tests”, Ann. of Statistics **7**, 697 (1979); M.F. Schilling, “Multivariate Two-Sample Tests Based on Nearest Neighbors”, J. of the Amer. Stat. Assoc. **81**, 799 (1986); J. Cuzick and R. Edwards, “Spatial Clustering in Inhomogeneous Populations”, J. of the Royal Stat. Soc. **B 52-1**, 73 (1990).

[11] P.M. Dixon, “Nearest Neighbor Methods”,

<http://www.stat.iastate.edu/preprint/articles/2001-19.pdf>

[12] See, for example, BaBar Collaboration, “Measurement of $\sin^2\beta$ using Hadronic J/psi Decays”, hep-ex/0309039, 2003.

[13] Phys. Rev. **D 66**, Review of Particle Physics, 2002.