Maximal Information Analysis: I - Various Wayne State Plots

G. Bonvicini
Wayne State University, Detroit, MI 48201, USA

Data analysis using all moments of the likelihood $L(\alpha)$ is presented. The relevant plots for various data fitting situations are presented. The goodness of fit parameter (currently the χ^2) is redefined as the isoprobability level in a multidimensional space. Fundamental properties of statistical analysis are described for the first time.

In 1987 I co-wrote a paper that reanalyzed narrow resonance data in e^+e^- collisions[1].The analysis was made necessary by the widespread use in the community of incorrectly calculated radiative corrections. Those resonance data were obtained from fits of the experimental resonance, a bell-shaped curve with three free parameters. Somehow large, shape-like changes in the radiative corrections were re-absorbed by the normalization parameter Γ_{ee}, producing a large bias in that quantity, while producing acceptable χ^2 results. At the end of the study the world average of Γ_{ee} for the various Υ resonances changed by up to three standard deviations. During our reanalysis it was noticed that the only trace of a biased fit was in the abnormally large uncertainty in the Γ_{ee} error, when data were compared with a toy Monte Carlo.

In 1992 I analyzed in detail various 17-keV neutrino experiments[2]. At the time a majority of experiments found null results, but five found results consistent with heavy neutrino mixing ($\sin^2\theta \sim 0.8\%$). In that analysis, it was pointed out that most experiments, including some of those which turned out to be "correct", had abnormally large uncertainties in the $\sin^2\theta$ error, indicating that there were biases, and therefore that the limits could be broader than published (thus, the discrepancy between the groups of experiments would significantly decrease).

In 1997 Jean Dubosq analyzed the end point of the decay $\tau \to 5\pi\nu$ in the CLEO data[3], following a similar analysis by ALEPH[4], for the purpose of extracting the neutrino mass. While CLEO had nearly 40 times the statistics of ALEPH, and a better energy and mass resolution, the ALEPH limit was a factor of 1.5 better. The limit is taken, roughly, as the average μ plus twice the error σ, and an anomalously low σ will produce an underestimated limit.

It is not a surprise that the σ is sensitive to bias (and therefore be included somehow in the definition of goodness-of-fit). The Cramer-Frechet-Rao limit[5] states as much. The variance (defined as σ^2) when bias is absent is (one parameter only)

$$\sigma_0^2 = \Sigma_i((\frac{\partial y_i}{\partial \alpha})^{-1}\delta_{yi})^2 \quad (1)$$

and is changed by a factor

$$\sigma^2 = \sigma_0^2(1+\frac{\partial b}{\partial \alpha})^2 \quad (2)$$

in the presence of a bias b.

With this work the usage of higher moments is introduced, and several, apparently previously unobserved, fundamental properties of statistical analysis are discussed. We consider the likelihood function which is the product of the probabilities for each data point y_i (in a 1-dimensional plot, the ordinate), given the fit parameter(s) α,

$$L(\alpha) = \Pi_i P(y_i|\alpha).$$

The data points are to be fitted with a function $f(x_i,\alpha)$, (the x variable(s) are, in a 1-dimensional plot, the abscissa). The probability $P(\alpha)$, proportional to the likelihood, is also defined, so that

$$P(\alpha) = \frac{L(\alpha)}{\int L(\alpha)d\alpha}.$$

In the cases discussed below, where population histograms are to be analyzed, one uses the binned likelihood with Poissonian statistics

$$L(\alpha) = \Pi_i \frac{e^{-f(x_i,\alpha)}f(x_i,\alpha)^{y_i}}{y_i!}.$$

If the fitting function is truly unbiased, or so close to an unbiased fit that it can be considered unbiased, then only a very restricted region of Hilbert space will be occupied by fits which were generated by the statistics described by the fitting function. The new, generalized, goodness of fit parameter is

$$G = \int P(A|N)P(\alpha)d\alpha. \quad (3)$$

N is the total number of events. In case of a continuous measurement, $N \to \delta$, the set of errors associated with the set of data points (that is, each data point is $y_i \pm \delta_i$). Because of the way G is constructed, the statistics to be used in allocating events to bins is multinomial. In the longer version of this paper, it is shown that the two statistics are equivalent. The procedure to determine G is a two-step procedure, first one determines the likelihood, and then one generates the likelihood-dependent plots described below and finds G.

A detailed discussion of the motivation for the choice of G as a goodness-of-fit statistic will be included in a longer version of this paper. Three points

are made here. The first is that goodness of fit is an internal consistency check for the hypothesis H that the data y were generated by $f(x, \alpha)$. It seems appropriate that all moments of the likelihood be used. Second, suitable combination of the moments describe completely the likelihood in Hilbert space. If one assumes that all the information is contained in the likelihood, then the method retrieves maximal information about H.

Third and most important, the convolution over the experimentally obtained $P(\alpha)$ is done to take into account the knowledge of the true value of α. Equally prepared experiments will have slightly different plots because of the different results μ. It is important to show that the plots maintain their power when α_{true} is varied.

The set A of statistical quantities depends on the type of fit (how many parameters, and how many of them nuisance) and also on N (the smaller the statistics, the more important the skewness parameter M_3 will be). In the case of one parameter fits discussed below

$$A = (\chi^2, \sigma, (M_3)). \tag{4}$$

These quantities (plus the estimator μ) are used below and are defined as

$$\chi^2 = \ln(L_{max}), \tag{5}$$

$$\mu = \int P(\alpha)\alpha d\alpha, \tag{6}$$

$$\sigma = (\int P(\alpha)\alpha^2 d\alpha - \mu^2)^{1/2}, \tag{7}$$

$$M_3 = (\int P(\alpha)\alpha^3 d\alpha - 3\sigma^2\mu + 2\mu^3)^{1/3}. \tag{8}$$

The extra set of parentheses around M_3 indicates that it might, or might not, be of use. If the error is truly gaussian (or the statistics is truly very large), then M_3 will generally be devoid of information. It is the asymmetric nature of Poissonian statistics that generates a significant M_3. Two sample fitting functions $f(x, \alpha)$ are listed in Table 1. They are used to introduce the properties of the two-dimensional subspaces. Ten data (y) points, each corresponding to a bin centered at $x = 0.05, 0.15, ..., 0.95$, were generated by toy Monte Carlo with total number of events, summed over the ten bins, $N = 500$, and with the true parameter $\alpha_{true} = 1.0$. One such generation is called an "experiment". The number of experiments for each function was 3×10^5. The experiments were then fitted with the same function (by construction, an unbiased fit) and the quantities in Eqs.(5-8) recorded for plotting.

The various plots, obtained by plotting any two of the quantities in Eqs.(5-8), for a sufficient number of experiments, are called the Wayne State plots (plots containing μ as one of the axes are generally not usable in a real life experiment. They are useful at this stage to elucidate some hidden properties of statistical analysis). The population of the plots is equal to the number of experiments attempted in the simulation, in this case 3×10^5. Once a plot is produced, the real experimental result needs to be compared against the plot to evaluate the internal consistency of the analysis.

Table I Functions used to produce the plots. The integral is over the bin width, and the sum over the ten bins described in the text equals one.

Name	**Function**
f1	$\frac{2}{\alpha+2}\int_{min}^{max}(1-x)^{\alpha/2}$
f2	$\frac{1}{1-e^{-\alpha}}\int_{min}^{max} e^{-\alpha x}$

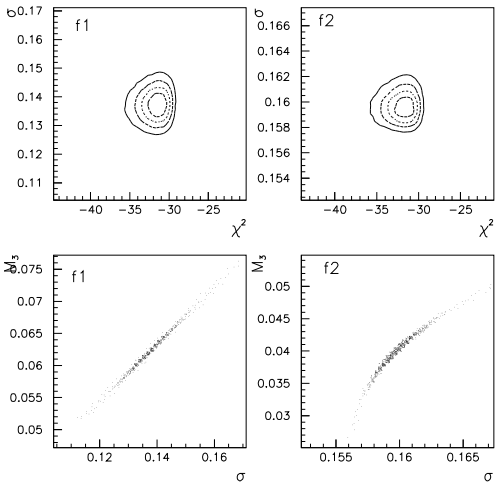

Figure 1: Top rows: first (χ^2, σ) plot for f1 and f2. Bottom row: second (σ, M_3) plot for f1 and f2.

The fit was done by raster scan over a large α interval, then over a more finely segmented, narrower interval within 10σ around the best-fit lattice point. As each point was probed, cumulative quantities useful for the determination of moments were computed.

In Fig. 1, top row, the first Wayne State plot (the (χ^2, σ) plot) are shown. In the limit of a meaningless M_3, this plot recovers maximal information about H. The contour levels shown are isoprobability curves, or rather, iso$-G$ curves. The confidence level for a given fit can be taken to be the fraction of fits lying outside that curve.

There are two interesting properties of the first plot. Along the σ axis, the first plot is narrow. Along the χ^2 axis, the plot is very broad. Generally, σ is far more sensitive to bias than the χ^2, and if one insists on using a single goodness of fit raw parameter, it should be σ and not χ^2. In the case of multi-parameter fits, the σ retains its parameter specific function (the goodness of fit for *that* parameter), whereas the χ^2 could be

dominated by nuisance parameters.

The apparently very low correlation between σ and χ^2 is a general property of the first plot, except in the limit of very low statistics where a correlation exists (sorry, no space for these figures). Its meaning is that both of these quantities are needed to assess the quality of a fit, because they are generally independent and both have some sort of sensitivity to bias.

Fig. 1, bottom row, shows the two (σ, M_3) plots. It is unclear whether these should be called the second plot. First, these plots will be meaningless in case of truly gaussian errors, second, in the case of a fit with one true parameter and one nuisance parameter (the next simplest case of interest), the (σ, ρ) plot, with ρ the correlation coefficient, is more important than this plot.

Fig.1, bottom row, is shown because of the extreme correlation between the two quantities. It is unclear at this point how useful this might be, but a fit that falls only slightly off the strip is bound to be biased. Much more important than that, the nearly complete correlation shows the rapid onset of information replication. G can be defined in a space with limited dimensionality, and considering further moments adds nothing to G. This is important in view of the fairly complex software that will have to be developed to make full use of the method described here. If one uses only the first plot, the fraction of recovered information can be estimated as the global correlation coefficient between the first plot quantities and M_3.

Fig. 2, top row, shows the two (μ, σ) plots. There is clearly extreme correlation between μ and σ. The Particle Data Book model of "lottery winning experiments" (experiments with comparable or inferior statistics and/or resolution, which manage to obtain superior results or limits) can be put to the test here. Assume that ten equal, unbiased experiments be performed, and then the "lottery winning" one be chosen as the best estimate. From the plots one can see that in more than 50% of the cases the very worst experiment will be chosen and the best one (the one closest to the true value) will be picked with about 0.1% probability.

The second reason to show Fig. 2, top row, is to point out a further positive property of using σ as part of the goodness of fit parameter. σ is very sensitive to purely statistical fluctuations, which can bias the final result just as much as uncorrected bias. The σ value may provide further constraints on μ in those cases where the bias is known to be very low.

Fig. 2, bottom row, shows the first two plots, for the first function of Table 1, when one varies α_{true} by 1.5σ. There is little variation in the first plot, whereas in the second plot the plot the variation happens to be along the strip described by the main plot. The plots have clearly semi-invariant properties.

In conclusion, with the arrival of maximal information analysis, statistics will undergo a profound trans-

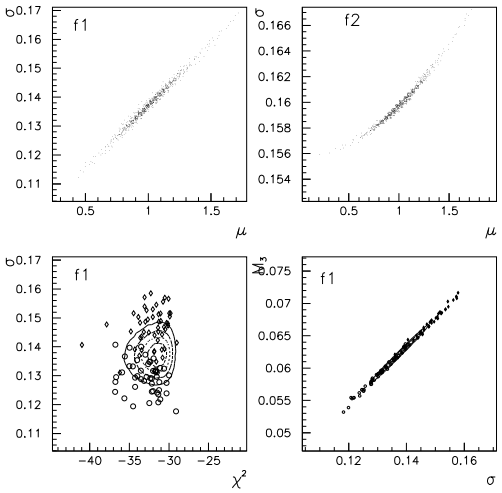

Figure 2: Top row: (μ, σ) plot for f1 and f2. Bottom row: first and second plot for f1 (50 experiments only), when α_{true} is varied. Solid squares: α_{true} =1.0; empty circles: $\alpha_{true} = 1.0 - 1.5\sigma$; empty diamonds: $\alpha_{true} = 1.0 + 1.5\sigma$.

formation. No longer bound by rules which are over 80 years old, we can actually extract much of the information available in a given fit. The method proposed here is valid at low and high statistics, for gaussian and non-gaussian errors, for single and multiple parameter fits, and independent of the definition of likelihood. While the phenomenology of maximal information analysis is fairly clear, the technology is going to be challenging. Certain applications (e.g., which set of nuisance parameters to use in a given fit) require relatively fast software which does not exist yet.

I thank F. Porter for many useful suggestions.

References

[1] J. Alexander *et al.*, Nucl. Phys. B320: 45, 1989.
[2] G. Bonvicini, Z. Phys. A345: 97-117, 1993
[3] R. Ammar *et al.*, (CLEO Collaboration), Phys. Lett. B431: 209-218, 1998
[4] D. Buskulic *et al.*, (ALEPH Collaboration), Phys. Lett. B349:585-596,1995
[5] H. Cramer, Mathematical Methods of Statistics, Princeton Univ. Press, New Jersey (1958).