Bruce Yabsley: Statistical practice at the Belle experiment, and some questions.

- Sometimes-dramatic illustrations of tradeoffs between “doing it right” (or even well-defined) and “getting out a result before it’s obsolete”
- Life is short; pragmatism has a clear role.
- $100 million investments deserve serious analysis… eventually.
- Goal: Educate people so they can judge how “dirty” the “quick” method is; otherwise an appropriate evaluation of the tradeoffs is impossible.

A HARD Problem: circular boundary; points and lines of interest. (transparency). My comments:

- The "answer" should have "P" well-defined:
- *Subjective* degree-of-belief P with prior a mix of delta-function and continuous functions is the path toward a coherent bet
- Confidence regions with frequentist P tell you "if you ran Monte Carlo's using the true values, then you cover…"
- For Bayesian analysis, my personal opinion is that non-subjective priors don't add much, if anything, to a graph of the likelihood function (which in any case is recommended to be published).

(Parodi)

Methods for including Δm_s in CKM Fit

Having $\mathcal{A}$ and $\sigma_{\mathcal{A}}$ for each Δm_s how do we proceed ?

Early days..

$$\chi^2 = \left(\frac{1-A}{\sigma_A}\right)^2$$

Sign of the deviation from 1 is not taken into account !! (osc. $\mathcal{A}$=1, no-osc. $\mathcal{A}$=0).
Moreover $A < 1$ is disfavoured w.r.t. $A > 1$!!

Ciuchini et al.

$$\mathcal{R}(\Delta m_s) = \frac{\mathcal{L}(\Delta m_s)}{\mathcal{L}(\infty)} = e^{\Delta log\mathcal{L}^\infty}$$

$$-2\Delta log\mathcal{L}^\infty = \left(\frac{1-A}{\sigma_A}\right)^2 - \left(\frac{A}{\sigma_A}\right)^2$$

It includes the relative weight between the two hypothesis (osc. $\mathcal{A}$=1, no-osc. $\mathcal{A}$=0)

Hocker et al.

$$\chi^2 = 2 \cdot \left[\mathrm{Erfc}^{-1}\left(\frac{1}{2}\mathrm{Erfc}\left(\frac{1-A}{\sqrt{2}\sigma_A}\right)\right)\right]^2$$

It cures in "ad hoc" way the $\mathcal{A}$>1 problem...

Parodi showed "modified χ^2" doesn't perform well, prefers L ratio.

My Comments

- One should always be skeptical of “modified chi-squares” invented in HEP; a significant burden of proof must be met.
- This problem appears to be a natural candidate for examination with established likelihood-based techniques such as Kendall&Stuart → F-C before trying to patch up chi-square.
- I need to understand better why “A” was chosen as the way to parametrize the problem.

Tutorials, Overviews, Explanations

- Roger Barlow: systematic mistakes/effects/errors.
- Sherry Towers:
  - PDE's
  - Reducing variables in classification
- Harrison Prosper: unity of multi- dimensional methods
- Glen Cowan: Unfolding
- Niels Kjaer: Monte Carlo
- Pekka Sinervo: Signifiance
- Berkan Aslan (G. Zech); Fred James: Goodness of fit
- Tony Vaiculis: Support Vector Machines
- Paul Harrison: Blind Analysis

Let's hope that they write these up for the proceedings!

Sorry I left some out, in particular Pekka!

Conclusions: advice for practitioners

- Thou shalt never say 'systematic error' when thou meanest 'systematic effect'.
- Thou shalt know at all times whether what thou performest is a check for a mistake or an evaluation of an uncertainty
- Thou shalt not incorporate successful check results into thy total systematic error and make thereby a shield behind which to hide thy dodgy result.
- Thou shalt not incorporate failed check results unless thou art truly at thy wits' end
- Thou shalt say what thou doest, and thou shalt be able to justify it out of thine own mouth; not the mouth of thy supervisor, nor thy colleague who did the analysis last time, nor thy local statistics guru, nor thy mate down the pub.

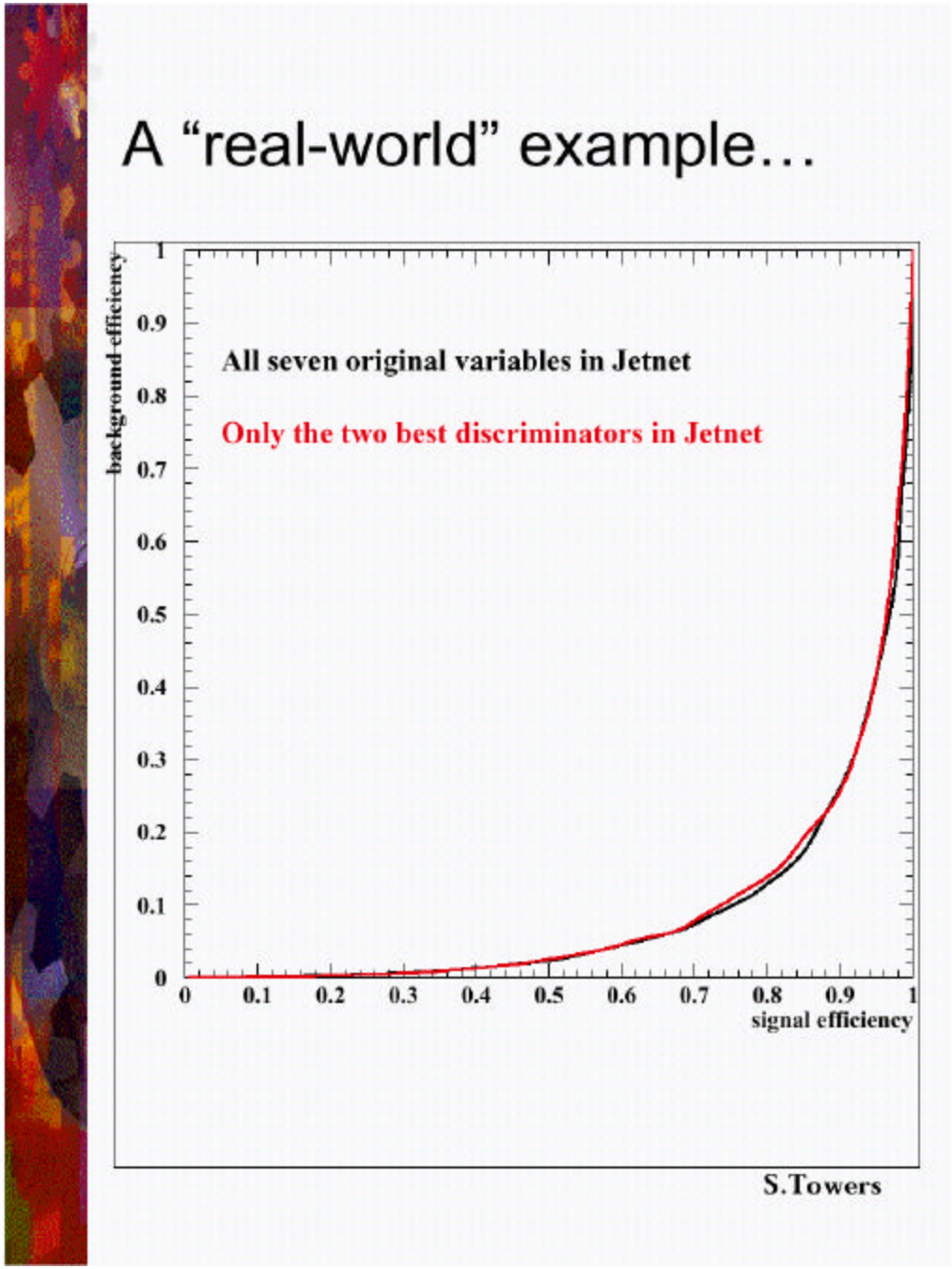
Do these, and thou shalt flourish, and thine analysis likewise.

Woe be unto the person who crosseth Roger!

Summary

- PDE methods are as powerful as neural networks, and offer an interesting alternative
- Very few parameters, easy to use, easy to understand, and yield unbinned estimate of PDF that user can examine in the multidimensional parameter space!

Sherry Towers



Wow! Several questions come to my mind…

[In general case, variables deletion is safer than variable addition. –M.G.]

Harrison Prosper

- Thumbnail sketch of some methods of interest:
  - Fisher Linear Discriminant
  - Principal Components Analysis
  - Independent Component Analysis
  - Self-Organizing Map
  - Grid Search
  - Probability Density Estimation
  - Neural Networks
  - Support Vector Machines
- Said these all are attempts to solve the *single* classification problem whose solution is the Bayes discriminator
  $D(x) = P(S|x)/P(B|x) = (L(S)/L(B))\ (P(S)/P(B))$
  … = Neyman-Pearson when $P(S)=P(B)$
- Multivariate analysis is hard: important to use all the information used by D(x) (which might be lost, e.g., by marginalization). Appears that there is no single optimal approximation.

Outline of topics

N.J. Kjaer (I)

- Introduction
- What is Monte Carlo?
- How do HEP MC work?
- MC weights and reweighting
- Reweighting analyses
- Frequentist resampling
- Systematic errors
- Unifying everything
- Conclusions and outlook

Outline of talk

- The Monte Carlo paradigm
- The Kalman filter
- What is optimal?
- The Minimum Variance Bounds
- The "actions" and constraints
- The optimization
  - Analysis corrections
  - Analysis parameter estimation
  - Combining analyses
- Optimizing globally
- Correlated measurements
- Robustness
- Systematic errors
- Conclusions and outlook

Conclusions and outlook

- k2-filter: Consistent and complete implementation of Kalman filters in MC analyses
- k2-filter optimal likelihood analyses with MC
- MC both integrator and differentiator
- Both Bayesian and Frequentist's
- Applicable to nearly all measurements
- Slightly different implementation for searches
- Both smart and intelligent at the same time?

Lots of experience and food for thought!

Conclusions

- no test statistic is better (concerning all alternatives) than its competitors
- there is no general theory that tells us how to choose a test
- for univariate distributions there are more powerful tests than the χ^2 test
- Neyman, Anderson and Kolmogorov are sensitive to a shift of the mean
- Watson and Kuiper are sensitive to a change in variance rather than in mean
- Energy test (log as corr. function) detects long range distortions
- Energy test (Gaussian as corr. function) useful to detect short range distortions
- Gaussian energy test is superior to χ^2 test (no arbitrary binning, more powerful)
- Energy test can be extended to the multivariate case
  (powerful for testing multivariate normality)

Closing Comments

- BA brings particle physics into line with best practice from other branches o science.
- More a formalisation of good experimental practice, than a radical new idea.
- An analysis which is not blind, is not necessarily a wrong analysis
- An analysis which is blind is not necessarily a right analysis
- The field has its fair-share of embarrassing wrong results
- Even the chance of experimenters' bias reduces our confidence in our results.
- If we can reduce risks of bias, why not do so?

I have now been in three experiments in which blind analysis was done, including the one led by Bill Molzon, referred to by Paul. In my experience it can slow down the analysis considerably, but it is worth it. I would add a few comments: if it's a new experiment, looking at 10% of the data is useful and shouldn't bias things badly. Sometimes cuts must be changed after opening the box. I think a reasonable criterion is: it's OK to add or change a cut if you would look foolish to outsiders if you did NOT change it. One example that happened to me was that we opened the box and found an event with zero's (not pedestals) in the ADCs or TDCs. It would have been silly to keep those events because of an abstract BA principle. Finally, experiments such as BaBar are doing analyses which are blind not only to the signal region but also to the control region used to estimate the background. This even further avoids a bias which might lead one to underestimate the background.

Resampling

- Bootstrap, jackknife, etc., came up several times. Outside of lattice QCD, I haven't heard the words very often.
- As I recall Efron himself (with Ken Hayes, et al.), did do a bootstrap on the tau 1-prong paradox some years ago. (It didn't solve that particular problem.)

Kay Kinoshita

evaluating quality of fit in Unbinned Maximum Likelihood Fitting

K. Kinoshita
University of Cincinnati
Belle Collaboration

- Statistical distribution of λ - zero free parameters
- impact of free parameters
- some speculations

Summary

Goodness-of-fit for UMxL

- sorry, not possible with λ_{max} alone

Other measures of fit quality

Desirable, especially for multiparameter fitting

- steps toward definition of λ_{max} distribution for general PDF
- speculation - exploit info in $\{\lambda_i(\alpha_{max})\}$

I remind you that the chi-square tests that we use (Gaussian and binned Poisson) can be derived starting from the likelihood ratio theorem. (For a review, see the paper I wrote with Steve Baker, referenced by the PDG RPP.) It's the *ratio* which gives the chi-square distribution of the test statistic (asymptotoically). So when trying to construct a g.o.f. statistic from L, try to find a likelihood ratio. I don't know any way to do this for unbinned likelihood. (And Fred has a way to convince me it's impossible.)

Confidence Limits and Their Errors

Rajendran Raja

So, naturalists observe, a flea
Has smaller fleas that on him prey;
And these have smaller still to bite 'em;
And so proceed ad infinitum.

Jonathan Swift

- Consensus: Idea is trying to get at something we think could be useful, but the explication needs a little work. One of the goals of the last few years has been to clean up our vocabulary to be more consistent with the statistics literature.
- The sampling distribution of any statistic (function of the data) is well-defined and can be illuminating to look at. F-C suggest first-moment of limit, and Giunti has looked at second moment (both metric-dependent).
- We need to understand better his point about combining the errors.

Studies of Intervals

- Byron Roe and Michael Woodroofe: Mini-Boone
- Jan Conrad: Coverage with Systematics
- Rolke and Lopez: Bias correction via double-bootstrap
- Giunti and Laveder: the “power” of confidence intervals
- Punzi: Strong Confidence Intervals
- Giovanni Signorelli et al: Strong C.I. And systematics

I am sorry that due to lack of time (preparation time, plus I am running over in my talk), I won't be able to comment on these talks. Please take a look at them!

A Few Words About Feldman, et al.

- Pros told us it was the “standard method” and eventually we found it in K&S. [transparency] Related to composite Neyman-Pearson test.
- It is well-defined for any problem for which you know the P(data|parameters) and the ensemble.
- Nasty multi-humped likelihood functions are not a problem.
- It gives confidence intervals, with all the good and bad that implies.
- K&S recommended approximate treatment of nuisance parameters; nowadays one can do a little better.

Application to Neutrino Oscillations

- The nu section of the F-C paper gives technical details, but the application is completely determined by the LR ordering introduced in the earlier in the paper.
- In our impenetrable words:

The acceptance region for each point in the $\sin^2(2\theta) - \Delta m^2$ plane is calculated by performing a Monte Carlo simulation of the results of a large number of experiments for the given set of unknown physical parameters and the known neutrino flux of the actual experiment. For each experiment, $\Delta\chi^2$ is calculated according to the prescription of either Eq. 5.5 or 5.6. The single number that is needed for each point in the $\sin^2(2\theta) - \Delta m^2$ plane is $\Delta\chi^2_c(\sin^2(2\theta), \Delta m^2)$, such that α of the simulated experiments have $\Delta\chi^2 < \Delta\chi^2_c$. After the data are analyzed, $\Delta\chi^2$ for the data and each point in the $\sin^2(2\theta) - \Delta m^2$ plane, i.e. $\Delta\chi^2(N|\sin^2(2\theta), \Delta m^2)$, is compared to $\Delta\chi^2_c$ and the acceptance region is all points such that

$$\Delta\chi^2(N|\sin^2(2\theta), \Delta m^2) < \Delta\chi^2_c(\sin^2(2\theta), \Delta m^2). \qquad (5.7)$$

- Here Roe said what he suggests if mini-Boone sees a signal, which appears to be the same. (In talking to him that is what I inferred.) He proposes R-W II for limit.

Let $\omega = \ln L$, ...

Find minimum of $-\omega$ (MINUIT). Then run a grid of nearby points and find the value of L_{test} such that 95% of the time the L found would be higher than L_{test}. Plot out region(s) of $\phi, \Delta m^2$ where L_{test} corresponds to the L obtained in the experiment.

Dean Karlen's Proposal to Evaluate Credibility of Confidence Intervals

- Yesterday evening, generally interested-to-favorable reaction
- I am an outlier: I think it will only encourage unthinking "easy" use of Bayes, with more flat (i.e., not degree of belief) priors.
- We evaluate Bayesian intervals with serious frequentist methods.
- Why not evaluate confidence intervals with serious Bayesian methods? One metric-dependent prior constituteth not a sensitivity analysis.
- Who was it who said "How do you know that the outlier isn't right?"

Alex Read's Beautiful Talk on CL_S

- Behavior compared to LR Ordering (F-C) is understood and lucidly explained. Application to neutrino oscillations!
- Please see his talk: I couldn't read the file in time for this talk.
- My comment: The non-standard conditioning (inequality, not ancillary statistic) of Zech and Roe&W and Read leads to problems with lower end of confidence intervals (see Cousins PRD Comment). Alex recognized this.
- Therefore, Alex now advocates CLS only for limits and in case of signal, he now would use LR Ordering.

Michael Goldstein

- A real pleasure to have you here!
- Since subjective Bayes is rarely used in HEP, but is "known" to be the "coherent" version, it has been very enlightening:
- **"Sensitivity Analysis is at the heart of scientific Bayesianism"**
  - How skeptical would the community as a whole have to be in order not to be convinced.
  - What prior gives P(hypothesis) > 0.5
  - What prior gives P(hypothesis) > 0.99, etc
- There's a split among Bayesians; M.G. is in the group that sees no virtue in objective ("arbitrary") priors (except as one of many examples of possible prior beliefs in a sensitivity analysis).

Michael Goldstein (cont.)

- Procedures should obey the likelihood principle. Frequentist methods don't obey it: fundamental flaw.
- Bayesian methods are hard to do right, but they are the only way to attack certain hard problems.
- Bayes Linear Methodology: addresses expectations rather than whole pdf's.
- HEP problems: appear to map onto a very similar set of abstract problems.

I would add:

- (Coherent) Subjective priors behave like real probabilities under transformations, unlike, e.g., flat priors.
- M.G. represents only one school of Bayesian stats, but I don't think you will find a school advocating uniform prior for a Poisson mean.
- M.G. portrays Bayesian methods as *hard*, but worth the effort. This should be stressed in HEP, where the hard part (subjective prior) is dodged, and the math is (indeed) easily cranked out (without backwards thinking) to give an "answer" that I think is without much content unless evaluated by frequentist standards.
- I think M.G.'s point about sensitivity analysis has to be taken to heart in HEP, whether one uses objective or subjective priors.

Educate Your Colleagues!

- The area under the likelihood function is meaningless.
- Mode of a probability density is metric-dependent, as are shortest intervals.
- A confidence interval is a statement about P(data | parameters), not P(parameters | data)
- Don’t confuse confidence intervals (statements about parameter) with goodness of fit (statement about model itself).
- P(non-SM physics | data) *requires* a prior; you won’t get it from frequentist statistics.
- The argument for coherence of Bayesian P is based on P = *subjective* degree of belief.

Thanks again!

Have a safe trip home.