PHYSTAT2011

Unfolding: Introduction

Louis Lyons

Imperial College, London

CERN

20th January 2011

The problem

Given a one-dimensional histogram for a particular variable, obtained in a detector with known experimental resolution, can we estimate the distribution that would have been obtained if the detector did not introduce any smearing? And we need to provide a covariance matrix for our unsmeared distribution.

{Some obvious generalisations of this}

Programme

9.10 Victor Panaretos "A statistician's view"
10.00 Volker Blobel "Unfolding methods for particle physics"
10.40 Coffee, and Poster Session
11.10 Guenter Zech "Regularization and error assignment in unfolding"
11.40 Vato Kartvelishvili "Unfolding with SVD"
12.10 Katharina Bierwagen 'Bayesian Unfolding'
12.20 Comparison on HEP methods
12.45 Lunch

2.00 Kerstin Tackmann 'SVD-based unfolding: implementation and experience'
2.20 Michael Schmelling 'Regularization by control of resolution function'
2.40 Bogdan Malaescu 'Iterative dynamically stabilized method of data unfolding'
3.05 Hans Dembinski 'ARU - towards optimal unfolding of data distributions'
3.30 Tim Adye 'Unfolding algorithms and tests using RooUnfold'
3.55 Coffee
4.25 Matthias Weber 'CMS unfolding'
4.55 Georgios Choudalakis 'ATLAS unfolding'
5.25 Jan Fiete Grosse-Oetringhaus 'ALICE unfolding'
5.45 Summary (Victor Panaretos) + Discussion

Non-expert comments on Unfolding

- Why unfold?
- Bin-by-bin correction factors
- Choice of bin-size
- How good is your method?
- Error estimates

Why Unfold?

If possible, **don’t Unfold data**, but smear theory

When not possible?

a) Compare data with data from different experiment

b) Tune MC by fitting QCD parameters to data

c) Future theories

Also provides more useful result (but complicated correlated errors)

Why Unfold?

If possible, don't Unfold data, but smear theory

When not possible?

a) Compare data with data from different experiment

b) Tune MC by fitting QCD parameters to data

c) ~~Future theories~~ Provide smearing matrix, so that future theorists can smear their theories

Also provides nicer picture to look at (but complicated correlated errors)

Matrix method

$$d_i = \Sigma M_{ij} t_j$$

Assume M_{ij} known with small statistical error from MC {Estimate effect of M_{ij} bias from incorrect resolution}

Max likelihood solution can have large bin-to-bin oscillations. Effect not serious for wide enough bins

Bin-by-bin correction factors

Use MC to find how exptl resolution makes:

'Truth' → Observed data   Beware MC statistical and systematic errors

$t_i = C_i * d_i$ (Contrast $t_i = \Sigma M^{-1}_{ij} * d_j$ for matrix method)

Problems:

1) $C_i = 0.1$, $d_i = 100 \pm 10$ → $t_i = 10 \pm 1$??

   i.e. Error too small. But this error estimate is incorrect

2) C_i depends on assumed distribution of t, which we are trying to find.

   Would iterative approach help?

   (For small bin-size, matrix method is less sensitive to distribution of t)

3) No bin-to-bin estimates of correlations

4) Sum of estimated truth ≠ Sum of observed data (Matrix method O.K.)

CONCLUSION: Do not use (Cf Cousins)

Correction Factors: a trivial example

	Bin 1	Bin 2
Truth	**800**	**200**
Mean observed	760	240

Smearing matrix

		True 1	True 2
Observed	1	0.9	0.2
	2	0.1	0.8

As t_1	As t_2	Re d_1	Re d_2	CF_1	CF_2	Est t_1	Est t_2	Sum
1000	0	900	100	1.11	0	844	0	844
800	200	760	240	1.05	0.83	800	200	1000
760	240	732	268	1.04	0.90	789	215	1004
667	333	667	333	1.00	1.00	760	240	1000
500	500	550	450	0.91	1.11	691	267	958
200	800	340	660	0.59	1.21	447	291	738
0	1000	200	800	0	1.25	0	300	300

As t_i Assumed # in bin i

De d_i Resulting expected # in bin i

CF_i Correction factor for bin i = (As t) / (De d)

Est t_i Estimated true # in bin i

Sum Est t_1 + Est t_2

Est t_1 = 760 * CF_1 (truth = 800)

Est t_2 = 240 * CF_2 (truth = 200)

Bin size?

Not too small

M_{ij} has large off-diagonal elements

Not too large

Lose sensitivity

M_{ij} depends on true distribution

Bin width for unfolded distribution also depends on that for data, and on exptl resolution.

Recommendation for optimum?

Regularisation

Damps out oscillations at price of (small?) bias

Recipe for optimal regularisation? (Depends on……….)

How to judge which method is 'best'?

Bob Cousins' "bottom line test":

Compare 2 theories with data via χ^2 ($\chi^2{}_1$ and $\chi^2{}_2$)

a) by smearing theories (well-defined)

b) by unsmearing data (use your favourite method)

Do $\Delta\chi^2 = \chi^2{}_1 - \chi^2{}_2$ for a) and for b) agree?

Can unfolded dist have σ < √n?

Regularisation can produce this.

Cf Straight line fitting to data → fitted uncertainties smaller than measured ones

Estimate errors by MC or bootstrap replications of data

Looking forward to all the talks, and especially to hearing what Statisticians (especially **Victor Panaretos**) can tell us about the subject.

Looking forward to all the talks, and especially to hearing what Statisticians (especially **Victor Panaretos**) can tell us about the subject.

Maybe even consensus on ‘best approach’?

Uncertainties for bin-by-bin correction factors

Start with 100±10 events in bin

Correction factor determined by MC to be 0.1, with negligible statistical error (but worry about systematics)

Corrected number = 10±1 ?

Related to:

Have N observations, which divide into 2 categories as n_a and n_b

e.g. 100 cosmic ray showers, with 96 induced by protons and 4 by heavier nuclei.

Can think of

N as Poisson distributed and, for given N, n_a and n_b as correlated Binomial;

or, completely equivalently,

n_a and n_b as uncorrelated Poissons.

10±1 ignores Binomial fluctuations. Must be 10±3 or worse (from Poisson for n_a)

And, of course, worry about systematics for correction factors