Fitting w/ Finite Monte Carlo

D. Jason Koskinen

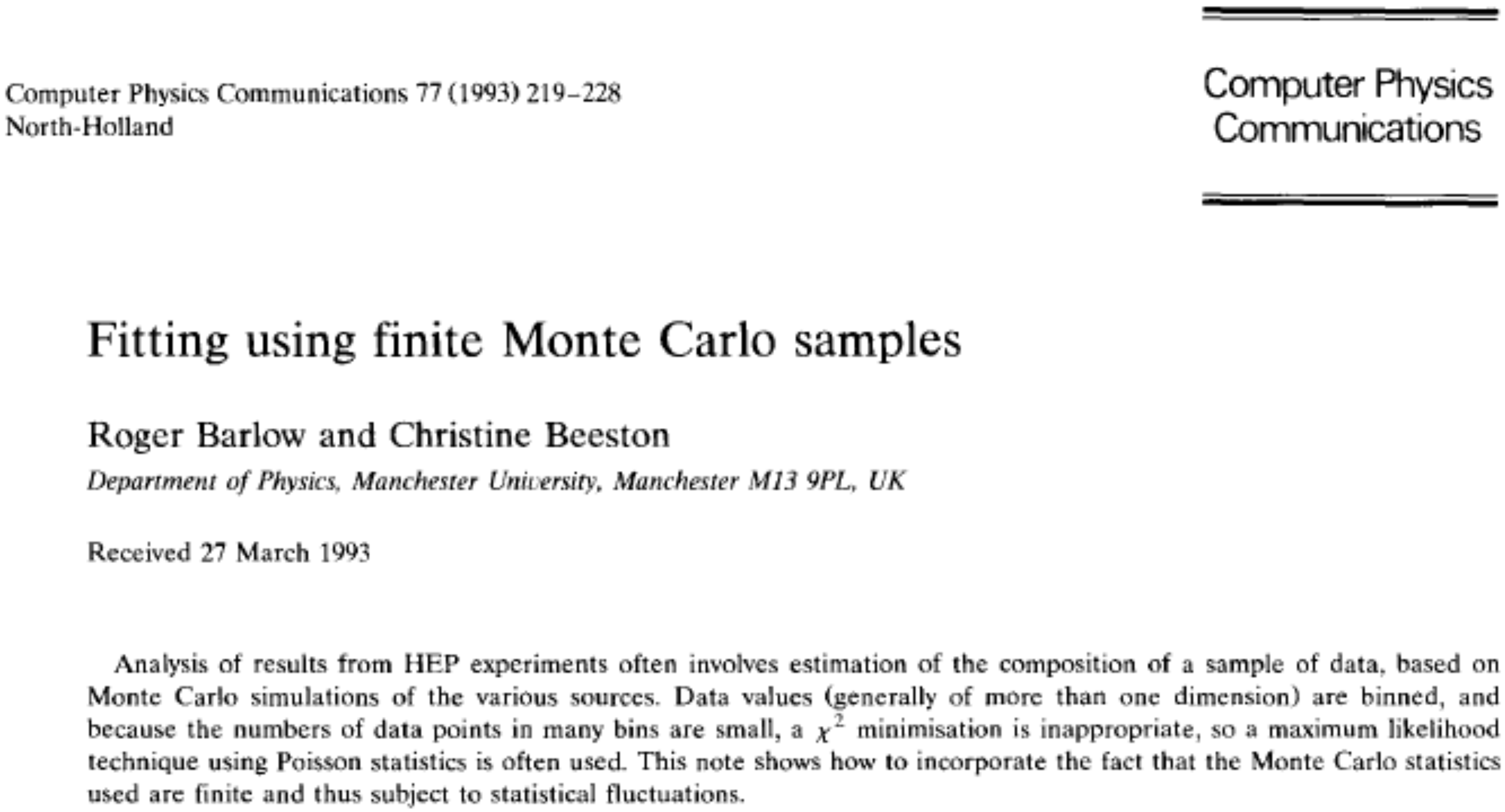
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics

Feb - Apr 2018

Photo by Howard Jackman

Journal Article



Computer Physics Communications 77 (1993) 219–228
North-Holland

Computer Physics
Communications

Fitting using finite Monte Carlo samples

Roger Barlow and Christine Beeston

Department of Physics, Manchester University, Manchester M13 9PL, UK

Received 27 March 1993

Analysis of results from HEP experiments often involves estimation of the composition of a sample of data, based on Monte Carlo simulations of the various sources. Data values (generally of more than one dimension) are binned, and because the numbers of data points in many bins are small, a χ^2 minimisation is inappropriate, so a maximum likelihood technique using Poisson statistics is often used. This note shows how to incorporate the fact that the Monte Carlo statistics used are finite and thus subject to statistical fluctuations.

doi:10.1016/0010-4655(93)90005-W

The Setup for a Binned Likelihood

- Using a binned data set with i bins from j multiple sources, e.g. signal and background (j=1,2), the predicted number of events in bin i (f_i) is

$$f_i = \sum_{j=1}^{m} p_j a_{ji}$$

a_{ji} is the number of Monte Carlo events from source j in bin i and p_j is the source strength, i.e. fraction

- When comparing to similar binned data (d_i) we can maximize the following log-likelihood which takes into account the poisson fluctuations from bins w/ small numbers of data (d_i)

$$\ln L = \sum_{i=1}^{n} d_i \ln f_i - f_i$$

The Problem

- While the likelihood takes into account fluctuations of the data (d_i) it does not include that the Monte Carlo (a_{ji}) may be of finite statistics and affected by poisson fluctuations
    - Monte Carlo generation can be slow
    - One of the sources could be a rare process
    - Choice of binning is fine in signal dominated region, but due to excellent background rejection the other regions are minimally populated
- A solution is to include the contribution of a poisson fluctuation on a_{ji}:

$$\ln L = \overbrace{\sum_{i=1}^{n} d_i \ln f_i - f_i}^{\text{likelihood for } d_i} + \overbrace{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ji} \ln A_{ji} - A_{ji}}^{\text{likelihood for } a_{ji}}$$

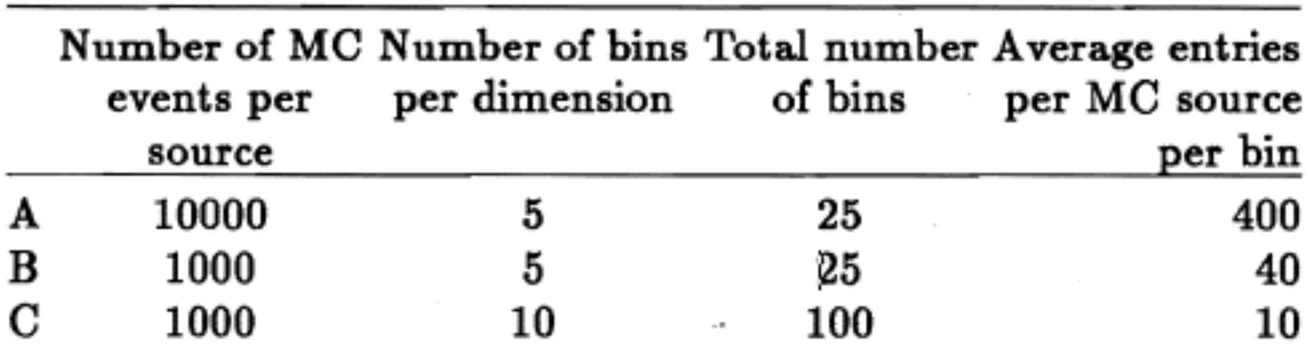
The Goal

- Maximize the LLH by differentiating and setting the derivative equal to zero and solving iteratively or maximize the LLH numerically, e.g. multiply by -1 and use a minimizer
  - For p_j and A_{ji}
  - Maximizing over the actual LLH consists of m*(n+1) unknowns
  - The values of p_j are what we want, because those will be the fractions of sources (signal(s) and background(s)) in the data sample
- The following plots show two sources with a ratio of p_1=1/3 and p_2=2/3 and the following distributions

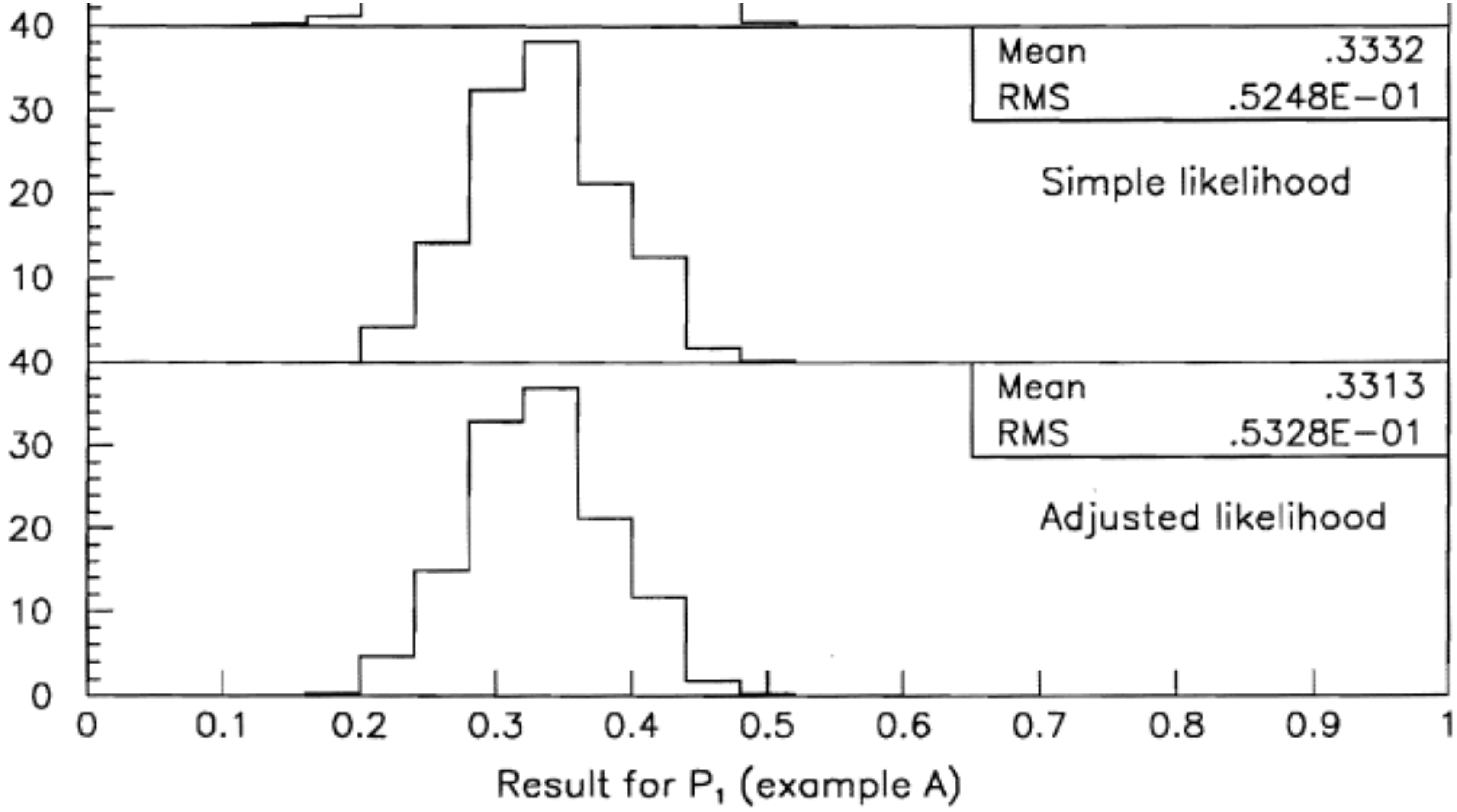
$$F_1(x_1, x_2) \propto \frac{x_1 + x_2}{2} \qquad F_2(x_1, x_2) \propto \frac{1 + x_1 - x_2}{2} \qquad \begin{array}{c} 0 \leq x_1 \leq 1 \\ 0 \leq x_2 \leq 1 \end{array}$$

Results - A

- For various numbers of MC generated events, the maximum likelihood approach for the simple likelihood compared to the adjusted likelihood
- For numerous entries the two are similar

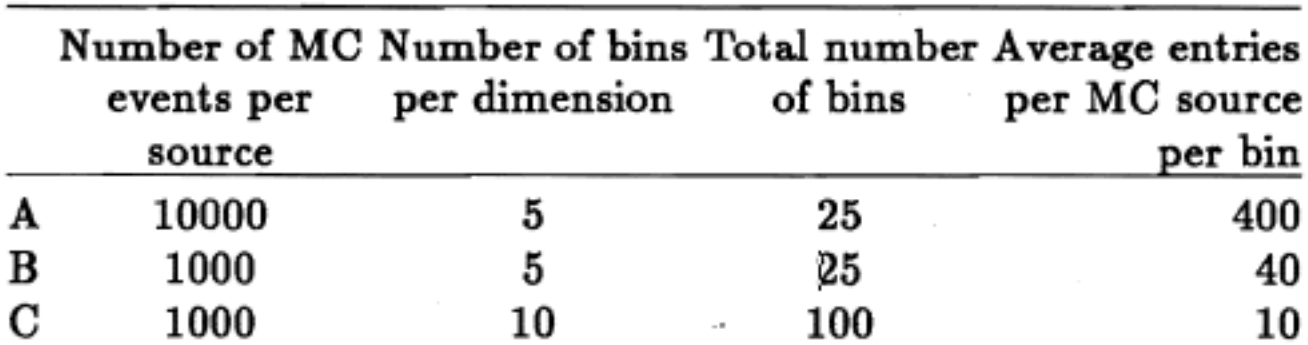


	Number of MC events per source	Number of bins per dimension	Total number of bins	Average entries per MC source per bin
A	10000	5	25	400
B	1000	5	25	40
C	1000	10	100	10

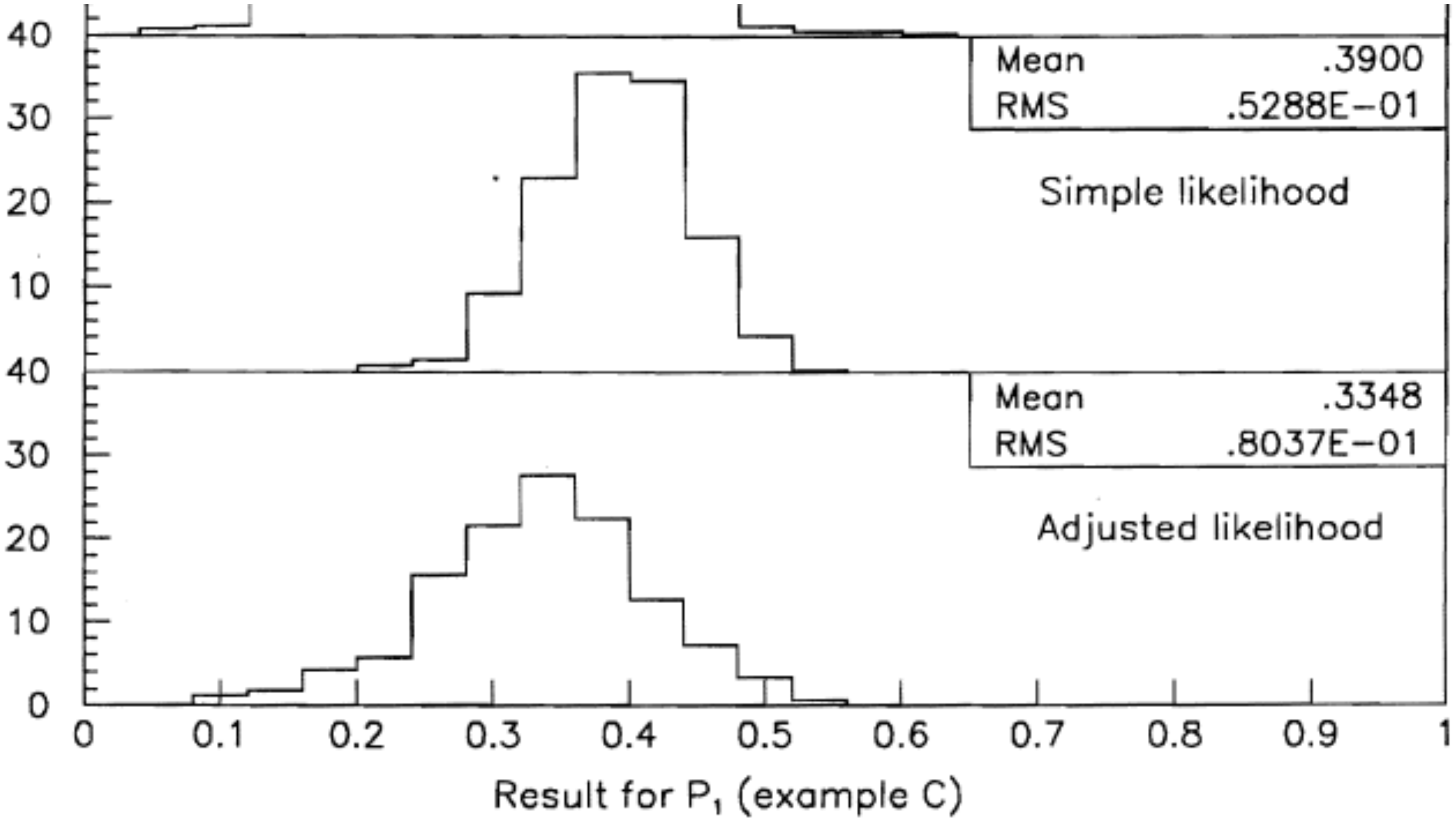


Results - C

- For various numbers of MC generated events, the maximum likelihood approach for the simple likelihood compared to the adjusted likelihood
- For a low number of events the simple likelihood is biased whereas the adjusted likelihood is unbiased



	Number of MC events per source	Number of bins per dimension	Total number of bins	Average entries per MC source per bin
A	10000	5	25	400
B	1000	5	25	40
C	1000	10	100	10



Conclusion

- Modifying the binned likelihood to account for poisson fluctuations in the data and Monte Carlo produces an unbiased estimator for low-statistics data and/or Monte Carlos samples
  - Works for Monte Carlo events which include weights, provided that the events in the same bin have similar weights (for some unknown definition of 'similar')
  - Special considerations are needed for a sample with no events from an MC source in the region of highest signal strength.