Matrix Element Analysis for Hadron Collider Data

Igor Volobouev
Texas Tech University
i.volobouev@ttu.edu

PHENO 09, May 11 2009

Matrix Element Analysis

- Build the probability to see an event in the detector:

$$P_{ev}(\mathbf{y}|\mathbf{a}) = \sum_i f_i P_i(\mathbf{y}|\mathbf{a})$$

- Probability for each channel:

$$P_i(\mathbf{y}|\mathbf{a}) = \frac{1}{\sigma_i(\mathbf{a}) A_i(\mathbf{a})} \sum_k \int_X W_k(\mathbf{y}|\mathbf{x},\mathbf{a})\, \epsilon(\mathbf{x},\mathbf{a})\, |M_i(\mathbf{x},\mathbf{a})|^2\, T_i(\mathbf{x},\mathbf{a})\, d\mathbf{x}$$

- $\sigma_i(\mathbf{a})$: cross section
- $A_i(\mathbf{a})$: acceptance
- $\sum_k$: combinatorics
- $\int_X$: phase space
- $W_k(\mathbf{y}|\mathbf{x},\mathbf{a})$: transfer function
- $\epsilon(\mathbf{x},\mathbf{a})$: efficiency
- $|M_i(\mathbf{x},\mathbf{a})|^2$: matrix element
- $T_i(\mathbf{x},\mathbf{a})$: PDFs, flux

$\mathbf{y}$ – observables, $\mathbf{a}$ – parameters (theory/exp), $\mathbf{x}$ – phase space variables

Advantages of the ME Method

- Maximization of the likelihood $L(\mathbf{a}) = \prod P_{ev}(\mathbf{y}|\mathbf{a})$ results in an efficient (in the statistical sense) estimate of $\mathbf{a}$.
- Neyman-Pearson lemma:

$$r_j(\mathbf{y}|\mathbf{a}) = \frac{P_j(\mathbf{y}|\mathbf{a})}{\sum_{i \neq j} f_i P_i(\mathbf{y}|\mathbf{a})}$$

  is the optimal discriminant for channel *j*.
- Unified framework for most HEP analyses. *Can be used to communicate event data and detector descriptions to theorists, and matrix elements to experimentalists.*
- Theory assumptions are incorporated in the most informative manner.
- Natural way to use non-Gaussian transfer functions.

Problems with the ME Method

- The most general case is intractable: dimensionalities of **y** and **x** are too high. Must make dimensionality reduction assumptions.
- Tough even at the hard process level. Example:

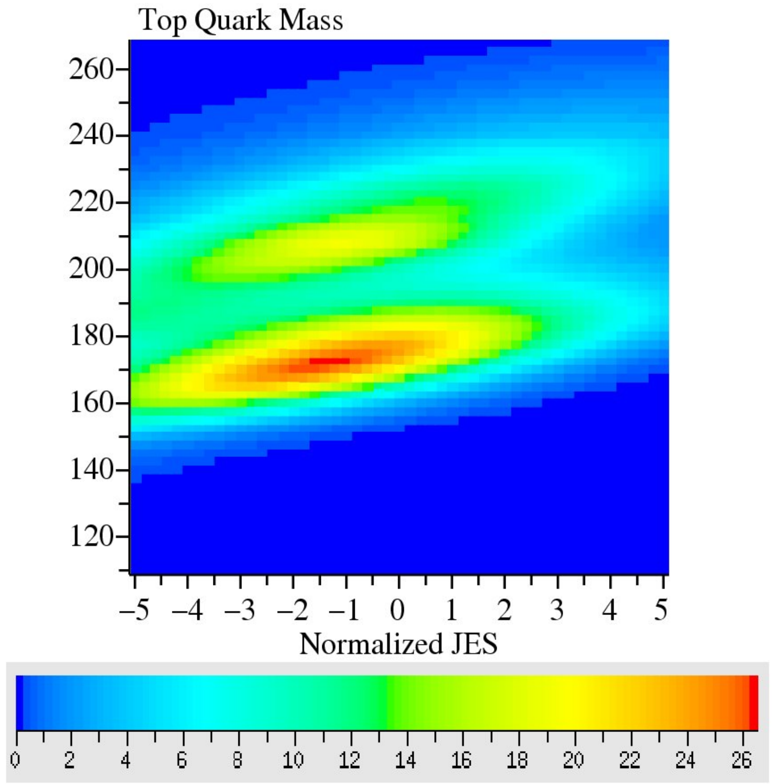
$$p\bar{p} \to t\bar{t} \to W^+bW^-\bar{b} \to \ell\nu + 4\ \mathrm{jets}$$

  D(**y**) = 19, D(**x**) = 24. Complex integrand structure.
- Extremely CPU-intensive (~1 hour per event, at least 10^5 events must be processed for simple analyses).
- No general-purpose software, little experience in the HEP community

ME Method: Recent History

- Practical use at the energy frontier demonstrated for the first time by the D0 Run I measurement of the top quark mass. Published in *Nature* in 2004, *8 years* after the run ended.
- Has been applied in a few “high impact” data analyses at Tevatron: Higgs searches, top quark mass measurement, W helicity in top decays, electroweak production of the top quark
- Usually the best method (among several analyses working with the same amount of data), but applications remain rare

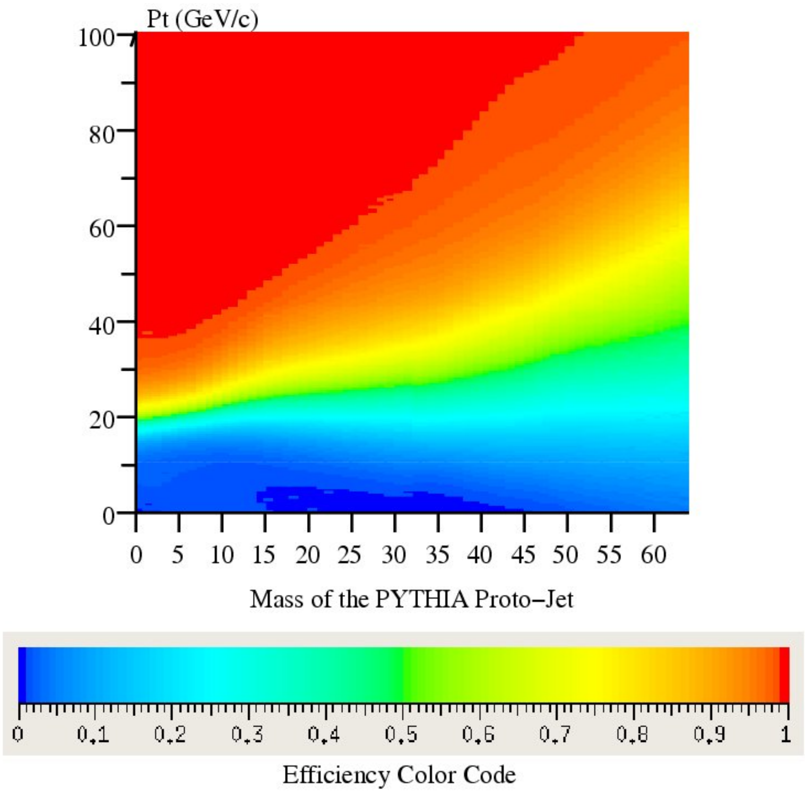
State of the Art



Example likelihood curve obtained using 19-dimensional QMC integration

- The ME method has never been applied in practice in full detail
- Typical simplifying assumptions: leading order ME, incoming partons have no transverse momentum, perfect resolution for some kinematic measurements, neglected efficiencies, ignored backgrounds, *etc*.
- The most advanced implementation at present time: "MTM3" top quark mass measurement by CDF (world's best single measurement with better than 1% precision): $M_t = 172.1 \pm 1.6$ GeV/c^2

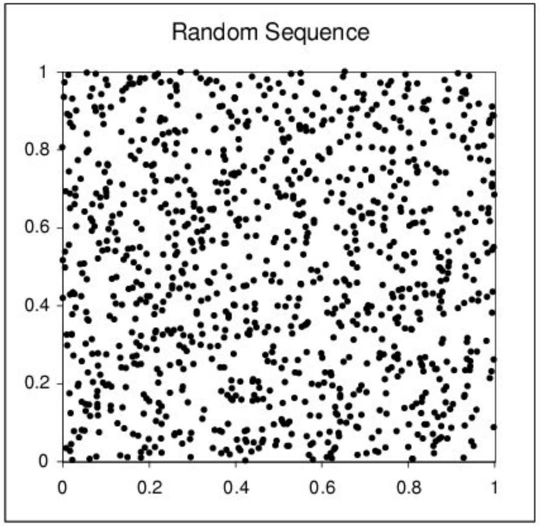
Analysis Features

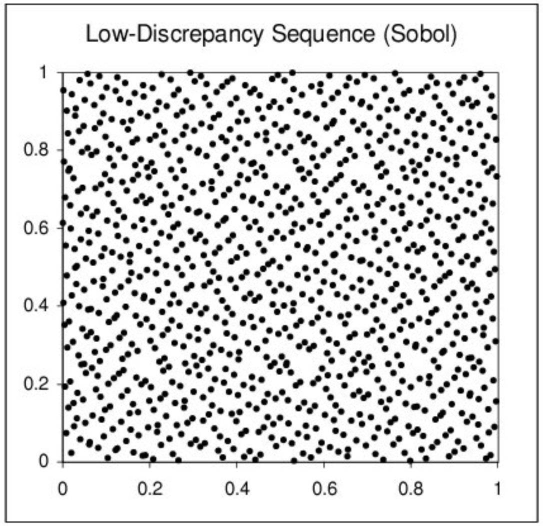


CDF jet reconstruction efficiency, built from MC using a local logistic regression model

- Factorize detector response using "physics objects" (leptons, jets, *b* tags). Correct on top of this.
- Subsume "soft QCD" effects into the transfer functions
- Use nonparametric statistical techniques for modeling transfer functions and efficiencies (let the computer do its job)
- Efficient phase space sampling scheme is built "by hand"
- Quasi-Monte Carlo is used for integration

Quasi-Monte Carlo





- Not Monte Carlo at all: based on the concept of deterministic “low-discrepancy sequences”
- Discrepancy is a measure of uniformity of a set of points. Several types of discrepancies have been studied. The most widely used is the “star discrepancy” which is a multivariate generalization of the Kolmogorov-Smirnov distance w.r.t. the uniform distribution:

$$D^*(P) = \sup_{B \in J^*} \left| \frac{1}{N} \sum_{i=1}^{N} c_B(\vec{x}_i) - S(B) \right|$$

P is the point set

*J** is the family of all subintervals of the d-dimensional unit cube of the form $\prod_{k=1}^{d} [0, u_k)$

c_B is the characteristic function of subinterval *B* (1 if the point is in *B*, 0 otherwise)

S(*B*) is the hypervolume of *B*

Quasi-Monte Carlo Convergence

- Koksma-Hlawka inequality: quasi-MC integration converges at least as fast as $O(\log(N)^d/N)$ for "well-behaved" functions. Compare with $O(1/\sqrt{N})$ convergence for standard MC.
- Note that this inequality provides a *deterministic upper bound* on the convergence rate. The *actual* convergence rate is usually better.
- QMC convergence for various types of integrands is still an active research area in applied mathematics
- Little studied in the HEP context (but see Kleiss and Lazopoulos, Comp. Phys. Commun., vol 175, pp 93-115, 2006).

ME Method: from Art to Technology

- For somebody new to the subject, the initial investment of effort is extremely high
- Proposed solution: build an experiment-independent, user-friendly framework for ME analyses and place it in the public domain
  - Best data analysis tool for the most interesting processes in the SM and beyond
  - Standardized interfaces will provide a bridge between HEP experimentalists and phenomenologists. We will be able to exchange detector models, matrix elements, and ***data*** in a meaningful way.
  - Assist with all the standard tasks: generation of the transfer functions, running integration code, obtaining results and uncertainties, *etc*
  - Base for the future phase space sampling research
- An enabling technology for the future of HEP
- Looking for collaborators!!!