The existence and uniqueness of diffraction at the LHC

Diffraction 2014, Primošten, Croatia, 10-16.9.2014

Mikael Mieskolainen

Helsinki Institute of Physics (HIP)
University of Helsinki
mikael.mieskolainen@cern.ch

15.9.2014

Intro

Introduction to a new Bayesian multivariate analysis framework for the analysis of high energy soft diffraction

Classification analysis $\sim$ cross section estimation of the inelastic *pp*-scattering processes

$$\sigma_{inel} \triangleq \sigma_{SDL} + \sigma_{SDR} + \sigma_{DD} + \sigma_{ND} + (\sigma_{CD}) \tag{1}$$

Probabilistic classification analysis also *disintegrates* differential inclusive distributions such as $dN_{ch}/d\eta$, $dE/d\eta$, $dp_T/d\eta$, $f(p_T)$, $f(\Delta\eta)$ into their corresponding classes (e.g. $dp_T^{(DD)}/d\eta$ etc.)

Recap - definitions of soft diffraction

- s-channel Good-Walker image where diffraction is understood as an elastic or quasi-elastic scattering (absorption) of the eigenstates of proton wave-function
- t-channel vacuum object exchange (Pomeron, Regge pole (or branch cut?) in complex ang.mom. plane). No color flow. In hard diffraction BFKL/QCD image of Pomeron as a gluonic ladder exchange

Several unknowns

What are the eigenstates of soft diffraction ($|t| \lesssim 0.5 \ldots 2$ GeV^2), how to treat low-mass dissociation, QCD image of soft diffraction, transition from soft to hard diffraction, transition between diffraction and non-diffraction (MPI/underlying event), the unified view between s- and t-channels...

Traditional large rapidity gap (LRG) analysis

The de-facto kinematical signature of diffraction (coherence)

Search for a gap of $\Delta\eta \geq 3$ units (same as $\xi = 1 - p_z^f/p_z^i = M_X^2/s \leq 0.05$) by requiring no tracks or energy deposit over some experimental threshold in the given η-interval.

Ill-posedness of the large rapidity gaps

LRG can be **destroyed** e.g. by spectator parton re-scatterings or by experimental reasons like calorimeter noise, fake tracks etc. or **created** artificially by having too high p_T-thresholds!

Due to random QCD fluctuations (which create "exponentially suppressed" LRGs), there is a background coming from non-diffractive events

High mass double diffractive events can in principle overlap in rapidity η-space $\Rightarrow$ experimental signature similar with non-diffractive events

No roman pots means that single and double diffraction are experimentally severely non-uniquely distinguishable (distinquishable only in particular mass ranges)

Multivariate classification analysis

Use maximally all information embedded in the event topology, i.e. based on the fact that the gaps do not carry all information about the events!

Instead of requiring LRGs, vectorize tracking (N_{ch}, p_T) (and calorimetry) information of an event over the available η-span into a continuous random vector $\mathbf{X} \in \mathbb{R}^d$

Estimate event-by-event the probabilities of different processes

$$\textit{Posterior} \propto \textit{Density} \times \textit{Prior} \tag{2}$$

Now, assume there is a function $f_{\mathbf{X}} : \mathbb{R}^d \to [0, \infty)$ such that there exists probability

$$P(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}, \tag{3}$$

where $A \subset \mathbb{R}^d$ is a domain with physically interesting event vector values. The function $f_{\mathbf{X}}$ is known as the total *density* function.

Posterior $\propto$ Density $\times$ Prior

Golden rule idea behind Bayesian inference: Update your prior knowledge with the new measurement

$$P(C = j|\mathbf{X} = \mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|j)P(j)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_j(\mathbf{x})P_j}{\sum_{j'=1}^{|\mathcal{C}|} f_{j'}(\mathbf{x})P_{j'}} \tag{4}$$

Densities f_j

with $j = 1, \ldots, |\mathcal{C}|$, ($\mathcal{C}$ is a discrete set of scattering processes) encapsulate the theoretical input about differential cross sections (e.g. triple Pomeron $1/M_X^2$) + hadronization phase (e.g. Lund string) and experimental detector effects (calorimeter response, track reconstruction efficiency...) (GEANT)

Priors P_j

encapsulate the theoretical integrated cross sections, e.g. single diffraction $P_{SD} \propto \int\int dM_X^2 dt \frac{d^2\sigma_{SD}}{dM_X^2 dt}$ (MC) $\times$ triggering efficiency (geometrical acceptance) (GEANT)

Hard classifier (cut on the output distribution)

$g : \mathbb{R}^d \to \mathcal{C}$

These can be seen as mappings

$$g : \mathbf{x} \mapsto \{1, 2, \ldots, |\mathcal{C}|\}. \tag{5}$$

Decision rule mappings g define *decision regions* as

$$\mathcal{R}_j = \{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) = (C = j)\}, \tag{6}$$

and thus $\mathcal{R}_j$ is the region in $\mathbb{R}^d$ where the posterior of class j is the highest. These decision regions can be defined by affine hyperplanes or in general, by nonlinear manifolds (or surfaces).

Bayes' minimum error classifier, optimal in Bayesian sense, does the *hard classification* according to

$$g^\star(\mathbf{x}) = \underset{j=1,\ldots,|\mathcal{C}|}{\arg\max}\, P(j|\mathbf{x}) = \underset{j=1,\ldots,|\mathcal{C}|}{\arg\max}\, f_j(\mathbf{x}) P_j, \tag{7}$$

What is the mathematical cost function to optimize?

- $S/\sqrt{S+B}$ in a case of traditional cross-section measurement (well understood background, i.e. $\langle B \rangle$ known) etc. assumptions

- $S/\sqrt{B}$ when one wants to maximize significance (search for new resonances etc.)

- Here instead, optimize the total classification accuracy, i.e. try to achieve Bayes error rate. Theoretically, this lower bound for classification error is given by

$$e(g^\star) = 1 - \sum_{j=1}^{|\mathcal{C}|} \int_{\mathcal{R}_j} f_j(\mathbf{x}) P_j \, d\mathbf{x}, \tag{8}$$

which is always non-zero for a problem with overlapping class densities.

A concrete algorithm - MLR-ℓ_1

Estimate posteriori probabilities directly, instead of modelling class densities f_j. A so-called "discriminative" approach.

Multinomial Logistic Regression with ℓ_1-norm regularization, gives posteriori probabilities through inner products $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^d$ between MC trained weights $\mathbf{w}_j$ and the event vector $\mathbf{x}$

$$P(C = j | \mathbf{X} = \mathbf{x}; \mathbf{w}) = \frac{\exp(\langle \mathbf{w}_j, \mathbf{x} \rangle) P_j}{\sum_{i=1}^{|\mathcal{C}|} \exp(\langle \mathbf{w}_i, \mathbf{x} \rangle) P_i}. \tag{9}$$

"Training" is done with uniform class fractions, and thus we use explicit priors P_j above. Exponential function guarantees the probabilistic output. Sparsity ℓ_1-regularization allows some interesting physical interpretations.

Note! By slight abuse of notation $\mathbf{w} := [\mathbf{w}_1^T, \ldots, \mathbf{w}_{|\mathcal{C}|}^T]^T$

Concave (− convex) cost function

Convexity guarantees a unique solution, i.e. no problems with local extrema

Formally, conditional ML estimates are obtained by maximizing concave cost function $l : \mathbb{R}^{d|\mathcal{C}|} \to \mathbb{R}$

$$l(\mathbf{w}) = \sum_{j=1}^{n} \ln\ P(\mathbf{y}_j|\mathbf{x}_j, \mathbf{w}) = \sum_{j=1}^{n} \left(\sum_{i=1}^{|\mathcal{C}|} \mathbf{y}_j^{(i)} \langle \mathbf{w}_i, \mathbf{x}_j \rangle - \ln\ \sum_{i=1}^{|\mathcal{C}|} \exp(\langle \mathbf{w}_i, \mathbf{x}_j \rangle) \right), \tag{10}$$

where n is the number of (MC) training vectors, $\mathbf{y}_j \in \{0,1\}^{|\mathcal{C}|}$ encodes class targets (SD,DD,ND etc.).

With regularization, this is in an augmented functional form

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} L(w) = \arg\max_{\mathbf{w}} \left[l(\mathbf{w}) + \log p(\mathbf{w}) \right], \tag{11}$$

and the regularization (prior) distribution is here $p(\mathbf{w}) \propto \exp(-\lambda \|w\|_{\ell_1})$

Training the algorithm

The optimization rule of the ℓ_1-regularized cost function is given by maximizing [1]

$$\mathbf{w}^T \left(\nabla(l(\hat{\mathbf{w}}^{(k)}) - \mathbf{B}\hat{\mathbf{w}}^{(k)} \right) + \frac{1}{2}\mathbf{w}^T(\mathbf{B} - \lambda\Lambda^{(k)})\mathbf{w}, \tag{12}$$

where $\Lambda^{(k)} = \text{diag}\left(|\hat{w}_1^{(k)}|^{-1}, \ldots, |\hat{w}_{d(|\mathcal{C}|-1)}^{(k)}|^{-1} \right)$ and the training data is in $\mathbf{B} = -\frac{1}{2}[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{|\mathcal{C}|}] \otimes \sum_{j=1}^{n} \mathbf{x}_j\mathbf{x}_j^T$ ($\otimes$ is the Kronecker tensor product).

Iterative algorithm

The iterative steps $1, 2, \ldots, k, k+1$ of the training/optimization algorithm are given by

$$\hat{\mathbf{w}}^{(k+1)} = \left(\mathbf{B} - \lambda\Lambda^{(k)} \right)^{-1} \left(\mathbf{B}\hat{\mathbf{w}}^{(k)} - \nabla l(\hat{\mathbf{w}}^{(k)}) \right), \tag{13}$$

[1] B. Krishnapuram et al. Sparse Multinomial Logistic Regression, 2005.

Regularization paths

ℓ_1-regularization induces rapidity gaps as a limit when $\lambda \to \infty$

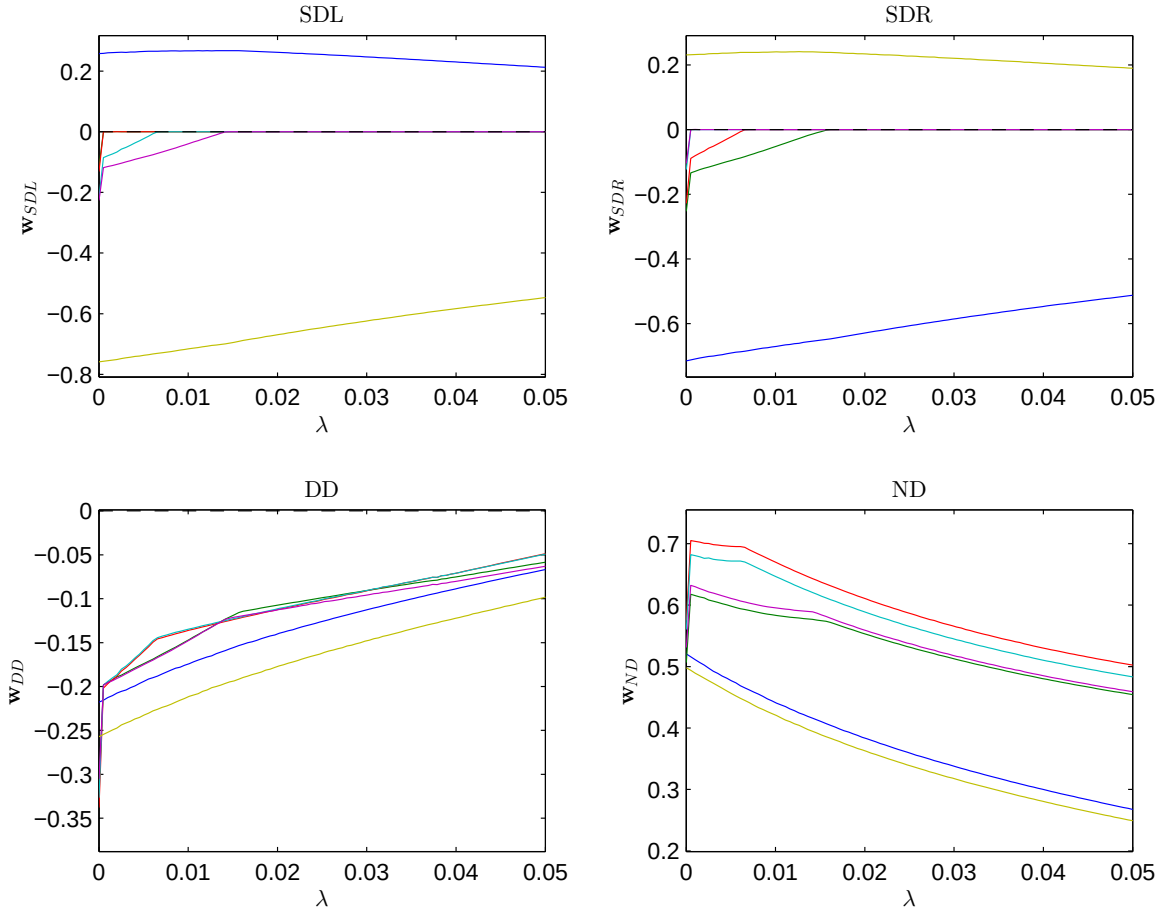

Figure : On y-axis the coefficients of $\mathbf{w}_j$ in order: $w_i :=$ (blue, green, red, light blue, purple, yellow), with binning $\mathbf{d}_\eta = (-3.6, -1.8, -0.9, 0, 0.9, 1.8, 3.6)$, such that $\eta_{\text{min,max}}(w_i) \in [d_i, d_{i+1}]$. Variables are calorimeter deposits integrated over ϕ.

Efficiency-Purity inversion

Important post-processing step due to highly non-diagonal confusion matrix. This step is also needed in every LRG based analysis!

Define the so-called confusion matrix (with indicator function $h_I(j;k) = 1$, if $j = k$, and 0 otherwise) as

$$[A]_{ij} \triangleq \mathbb{E}_{\mathbf{x}|C=i}\left[h_I(g(\mathbf{x});j)\right] = P(g(\mathbf{x}) = j|C = i), \tag{14}$$

which gives the conditional probability of classifying an event vector originating from the i-th class to the j-th class.

1. Class-by-class (bin-by-bin) correction factors (substractive and/or multiplicative)
2. Confusion matrix A regularized inversion (unfolding)
3. **Use event-by-event posteriori probabilities**, the most data-driven method of these!

Hard classification confusion matrix

The structure of this matrix depends on which priors were used to calculate it!

Table : Row normalized confusion matrix (4×4) estimate, with class efficiencies ϵ_j and purities π_j, and total classification accuracy given by PYTHIA 6.x (with CDF experiment GEANT4 simulation) and MLR-ℓ_1 as a hard classifier.

	SDL	SDR	DD	ND	ϵ_j
SDL	0.24	0.02	0.35	0.39	0.24
SDR	0.02	0.23	0.37	0.39	0.23
DD	0.13	0.13	0.43	0.31	0.43
ND	0.00	0.00	0.02	0.98	0.98
π_j	0.48	0.47	0.41	0.90	*Acc* 0.82

Non-diffractive (ND) class dictates the structure of confusion matrix, events are leaking into ND category but not vice versa!

Cross-sections via probabilities

"Soft classification", which is a mixture estimation problem.

It is well-known that conditional expectation values obey the so-called *iterated expectation* relation

$$\mathbb{E}[h(\mathbf{X},\mathbf{Y})] = \mathbb{E}[\mathbb{E}[h(\mathbf{X},\mathbf{Y})|\mathbf{Y}]] = \mathbb{E}[\mathbb{E}[h(\mathbf{X},\mathbf{Y})|\mathbf{X}]], \tag{15}$$

where $\mathbf{X},\mathbf{Y}$ are random vectors and $h(\mathbf{X},\mathbf{Y})$ some arbitrary function of those.

Using this, some previous definitions (and the indicator function h_I), one can show that integrating (summing) posteriori probabilities over an event sample size of n results in

$$\frac{\sigma_k}{\sigma_{inel}} \cong \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h_I(C;k)|\mathbf{X}=\mathbf{x}_i] \tag{16}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{C}|} h_I(j;k)P(j|\mathbf{x}_i) \quad \square \tag{17}$$

Different *philosophies* to choose priors

Bayesian vs Frequentist inference...

1. **Fully Bayesian**, induce distributions for the priors P_j using domain knowledge, and maybe some physical constraints as unitarity, Regge factorization, some symmetry etc. and finally integrate over posteriors $\rightarrow$ a simple sampling algorithm needed $\rightsquigarrow$ Bayesian credibility intervals a natural side-effect
2. **Point priors**, semi-Bayesian where one uses e.g. priors from the given MC model $\rightarrow$ Just use the Bayes' formula shown earlier, no computational complications
3. **Maximum Marginal Likelihood**, $\arg\max_{\{P_j\}} \prod_{i=1}^{n} \sum_{j=1}^{|\mathcal{C}|} f_j(\mathbf{x}_i) P_j$ i.e. maximize the denominator (evidence) in the Bayes' formula over the number of n events. No closed form solution, but iterative $\circlearrowleft$ Expectation Maximization (EM) algorithm can be derived (classic frequentist mixture density problem)

Pairwise posteriori probability distributions

Distributions below demonstrate the non-unique signature of real events

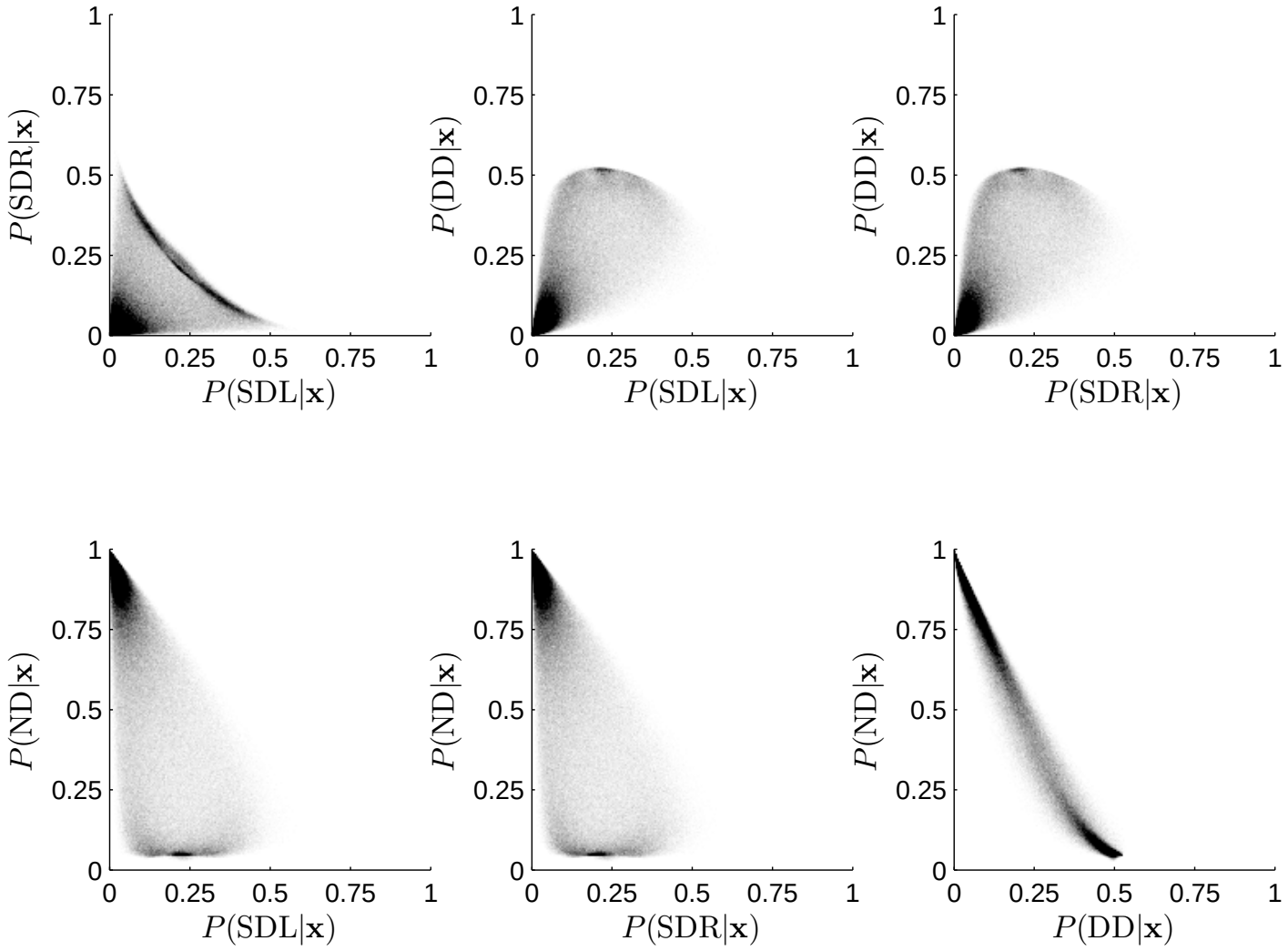

Figure : CDF $\sqrt{s} = 1.96$ TeV 0-bias data, MLR-ℓ_1 algorithm, PYTHIA 6.x MC.

Nearly perfect reconstruction via probabilistic weighting

In figure Boosted Decision Tree (BDT) as a hard classifier, which generates biased reconstruction

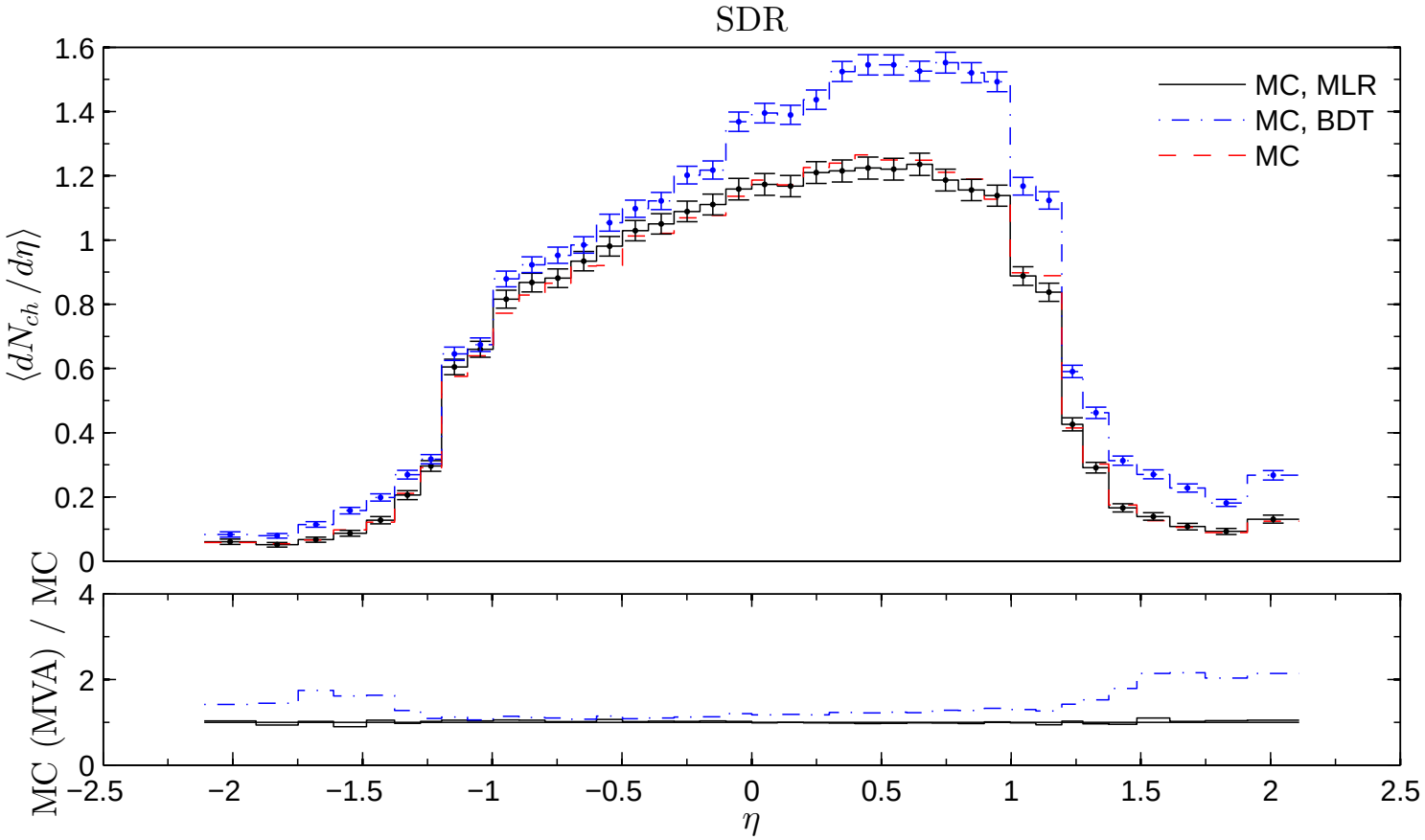

Figure : Monte Carlo vs. multivariate algorithm output with MC input, PYTHIA 6.x MC, CDF experiment GEANT4 chain.

Conclusions

Probabilistic multivariate approach can naturally handle the **non-unique** experimental signature between diffraction / non-diffraction and deals *optimally* with experimental limitations such as p_T-thresholds.

But one should do **both**, traditional LRG based analysis and (probabilistic) multivariate estimation!

By comparing results of these two kind of measurements, one could obtain e.g. estimates of gap survival S^2 values.

The problem of soft diffraction is so severely ill-posed that it requires the best statistical inference approach there's available.