Analysis of Diffraction Beyond Large Rapidity Gaps

Diffractive and Electromagnetic Processes at High Energies, WE-Heraeus Summerschool, Heidelberg, 2-6.9.2013

Mikael Mieskolainen

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

2.9.2013

Intro

Demonstration of some mathematical methods and ideas for analysis of high energy diffraction

Especially classification analysis of main $pp(\bar{p})$-scattering process classes, here defined as

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

Probabilistic classification analysis disintegrates also differential measurements such as $dN/d\eta$, $dE/d\eta$ into their corresponding classes (such as $dN_{DD}/d\eta$ etc.)

Recap - definition of 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 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)...

Traditional 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 threshold in the given η-interval.

However, gap can be destroyed e.g. by spectator parton re-scatterings or by purely experimental reasons (calorimeter noise etc.)

The gap survival probabilities S^2 are process dependent, but in general often estimated to be $\langle S^2 \rangle \lesssim 0.1$

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

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

Multivariate classification analysis

Using multidimensional information embedded in the event topology

Instead of requiring LRGs, vectorize tracking (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

$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior} \tag{2}$$

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 a *density* function (likelihood). One must note that the value $f_{\mathbf{X}}(\mathbf{x})$ is not a probability, but the integral over Ω must be equal to one.

Posterior $\propto$ Likelihood $\times$ Prior

In this talk likelihood synonymous to density, which is not always the case!

Likelihoods 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 (kinematics $\times$ dynamics) $+$ hadronization phase (MC) and include 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

$$g : \mathbf{x} \mapsto \{1, 2, \ldots, |\mathcal{C}|\}. \tag{4}$$

Decision rule mappings g define *decision regions* as

$$\mathcal{R}_j = \{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) = (C = j)\}, \tag{5}$$

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

$$g^\star(\mathbf{x}) = \underset{j=1,\ldots,|\mathcal{C}|}{\arg\max}\, P(j|\mathbf{x}) = \underset{j=1,\ldots,|\mathcal{C}|}{\arg\max}\, f_j(\mathbf{x}) P_j, \tag{6}$$

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

$$e(g^\star) = 1 - \sum_{j=1}^{|\mathcal{C}|} \int_{\mathcal{R}_j} f_j(\mathbf{x}) P_j \, d\mathbf{x}, \tag{7}$$

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

A concrete algorithm - MLR-ℓ_1

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 $\mathbf{x}$

$$P(C = j | \mathbf{X} = \mathbf{x}; \mathbf{w}) = \frac{\exp(\langle \mathbf{w}_j, \mathbf{x} \rangle) P_j}{\sum_{i=1}^{|\mathcal{C}|} \exp(\langle \mathbf{w}_i, \mathbf{x} \rangle) P_i}. \tag{8}$$

"Training" is done with uniform class fractions, and thus we use explicit priors P_j above. Exponential function guarantees the probabilistic output. Sparsity regularization allows mathematical variable selection.

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

Concave (-convex) cost function

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

$$l(\mathbf{w}) = \sum_{j=1}^{n} \ln P(\mathbf{y}_j|\mathbf{x}_j, \mathbf{w}) = \sum_{j=1}^{n} \left(\sum_{i=1}^{|\mathcal{C}|} \mathbf{y}_j^{(i)} \langle \mathbf{w}_i, \mathbf{x}_j \rangle - \ln \sum_{i=1}^{|\mathcal{C}|} \exp(\langle \mathbf{w}_i, \mathbf{x}_j \rangle) \right), \tag{9}$$

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

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} L(w) = \arg\max_{\mathbf{w}} \left[l(\mathbf{w}) + \log p(\mathbf{w}) \right], \tag{10}$$

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]

$$\mathbf{w}^T \left(\nabla(l(\hat{\mathbf{w}}^{(k)}) - \mathbf{B}\hat{\mathbf{w}}^{(k)} \right) + \frac{1}{2}\mathbf{w}^T(\mathbf{B} - \lambda\Lambda^{(k)})\mathbf{w}, \tag{11}$$

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 Kronecker tensor product).

Final training step

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

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

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

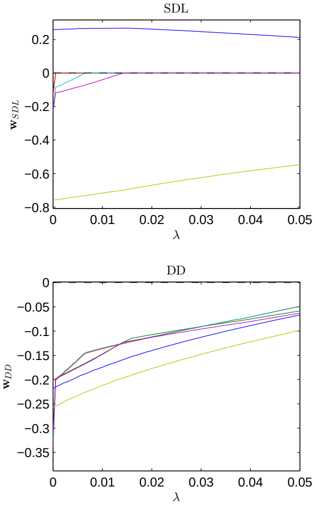

Figure : Regularization λ-paths using MLR-ℓ_1 with the MC training sample. On y-axis the coefficients of $\mathbf{w}_j$ in order: $w_i :=$ (blue, green, red, light blue, purple, yellow), with discrete binning $\mathbf{d}_\eta = (-3.64, -1.78, -0.88, 0, 0.88, 1.78, 3.64)$, such that $\eta_{\min,\max}(w_i) \in [d_i, d_{i+1}]$. Variables are calorimeter deposits integrated over ϕ.

Efficiency-Purity inversion

Important post-processing step due to highly non-diagonal confusion matrix!

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

$$[A]_{ij} \triangleq \mathbb{E}_{\mathbf{x}|C=i}\left[h_I(g(\mathbf{x});j)\right] = P(g(\mathbf{x}) = j|C = i), \tag{13}$$

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
2. Confusion matrix A regularized inversion (unfolding)
3. **Use event-by-event posteriori probabilities**, the most data-driven method of these!

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

One can see how non-diffractive (ND) class dictates the structure of confusion matrix (due to large cross section)!

Cross-sections via probabilities

"Soft classification"

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

$$\mathbb{E}[h(\mathbf{X},\mathbf{Y})] = \mathbb{E}[\mathbb{E}[h(\mathbf{X},\mathbf{Y})|\mathbf{Y}]] = \mathbb{E}[\mathbb{E}[h(\mathbf{X},\mathbf{Y})|\mathbf{X}]], \tag{14}$$

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

$$\frac{\sigma_k}{\sigma_{inel}} \cong \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[h_I(C;k)|\mathbf{X}=\mathbf{x}_i] \tag{15}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|\mathcal{C}|} h_I(j;k)P(j|\mathbf{x}_i) \quad \square \tag{16}$$

Pairwise posteriori probability distributions

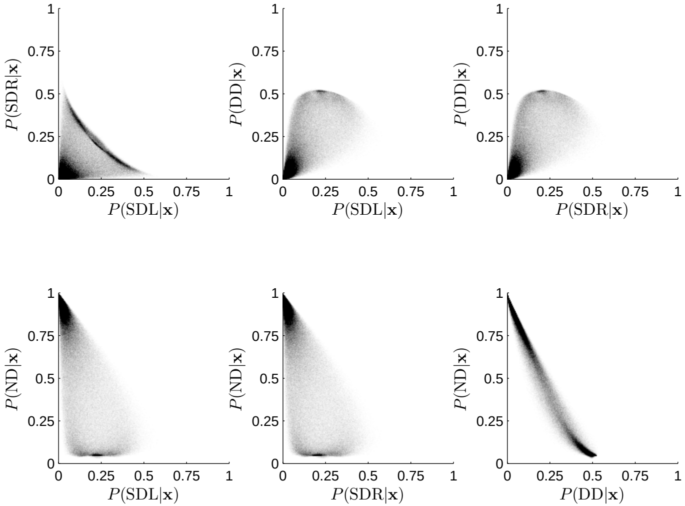

Figure : CDF $\sqrt{s} = 1.96$ TeV 0-bias data, MLR-ℓ_1 algorithm, PYTHIA 6.x MC.

Example of event-by-event probabilistic weighting

Boosted Decision Tree (BDT) is used without any efficiency-purity inversion

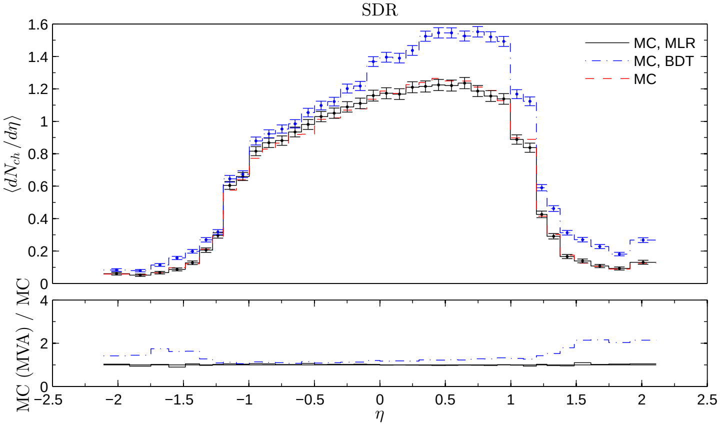

Figure : Monte Carlo vs. multivariate algorithm output with MC input, PYTHIA 6.x MC, CDF experiment GEANT4 chain.

Other multivariate analysis approaches

Multivariate Regression

$$f : \mathbb{R}^d \to \mathbb{R}^n \tag{17}$$

where $n \in \mathbb{N}$, often $n = 1$ (scalar quantity).

Can be used in principle to estimate event-by-event e.g. diffractive mass(es) M_X^2, 4-momentum transfer squared t (or impact parameter b), even if the LRG is destroyed or no leading proton (Roman pot) measurement is available

Modern algorithms to do this are e.g. Gaussian Processes (GP) based (infinite dimensional extensions of 1-hidden layer Neural Nets)

Binary combinatorial analysis

A simplified, integrated "toy" approach to classification

Generator level detector combinatorics 2^D (here $D = 4$) simulation using PYTHIA 8.x (MBR) $\sqrt{s} = 8$ TeV and step-function p_T acceptances for TOTEM T1,T2-detectors. This combinatorics can be seen as a two valued special case of a real valued vector space with replacement $\mathbb{R}^D \rightarrow \{0,1\}^D$.

Table : First five signatures out of $2^4 = 16$ possible, class fractions are f_j.

ID	T2-	T1-	T1+	T2+	f_{ND}	f_{SDL}	f_{SDR}	f_{DD}	f_{CD}	σ_i (mb)
0	0	0	0	0	0.00	0.38	0.39	0.17	0.06	3.4417
1	0	0	0	1	0.00	0.00	0.65	0.31	0.03	1.2377
2	0	0	1	0	0.03	0.00	0.46	0.27	0.24	0.4832
3	0	0	1	1	0.04	0.00	0.57	0.36	0.03	3.9924
4	0	1	0	0	0.03	0.46	0.00	0.27	0.24	0.4797

$\vdots$

Conclusions

One should do **both**, traditional LRG based analysis and (probabilistic) multivariate classification!

Probabilistic multivariate approach can naturally handle the **non-unique** experimental signature between diffraction / non-diffraction.

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

Multivariate methods allow testing the MC models against data in a mathematically consistent way.