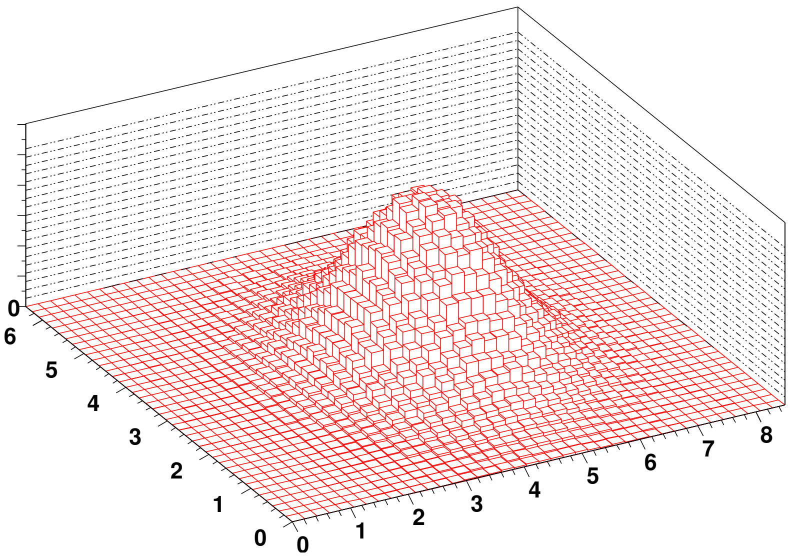
Multivariate Analysis Techniques for Final State Reconstruction

B. Koblitz for the H1 Collaboration, HCP 2002 Conference, 4. Oct 2002



Overview:

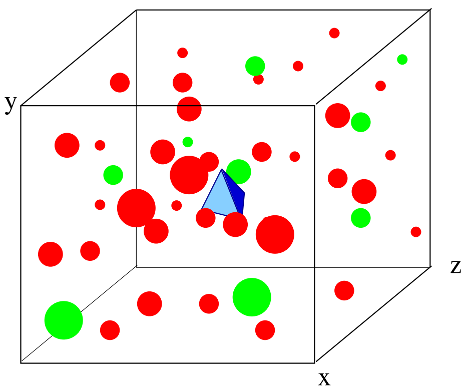
- Multivariate Methods: Kernel PDE's, Neural Nets
- The PDE-RS Method
- Properties of PDE-RS
- Example: Instantons in DIS
- Conclusions

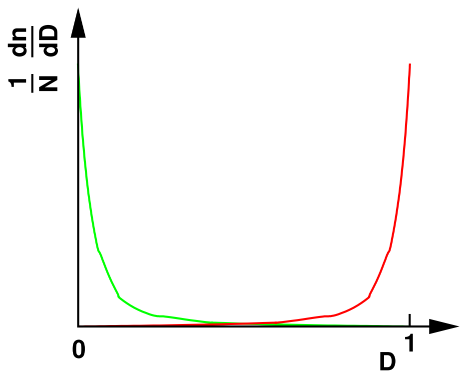
Final State Reconstruction & Event Classification

At hadron colliders, often huge background hides signal:

Problem: Classification as Signal/Background according to measured properties $(x_1, \dots, x_n)$ of event

Solution: Signal propability p_s given by ratio of signal/background event densities at place $(x_1, \dots, x_n)$ in phase space $\Rightarrow$ Modelled by Monte Carlo Generator





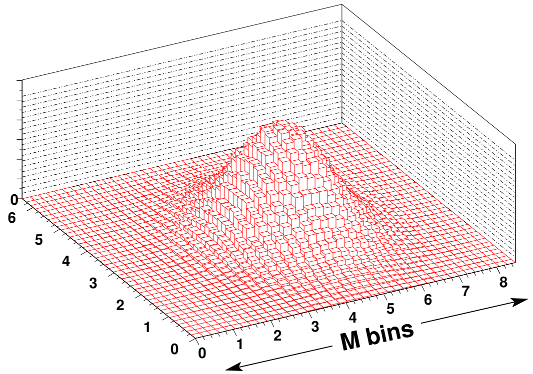
Discriminant $D(x_1, \dots, x_n) = \frac{\rho_S}{\rho_S + \rho_B} \approx p_S$

But: Estimating the n-dim phase space density (Probability Density Estimation, PDE) is often rather difficult

In following some PDE's explained.

Probability Density Estimation: Histograms and Kernels

Histogramming:

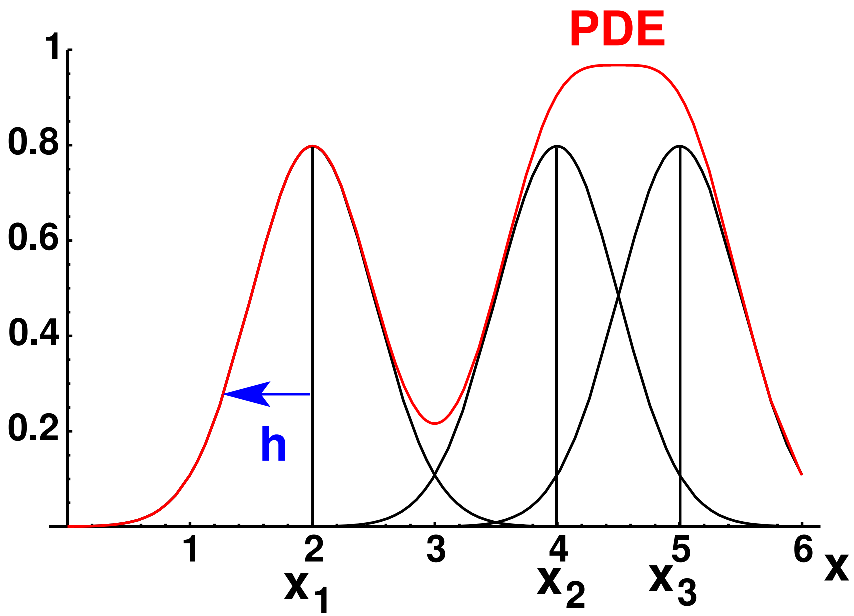


Bins $\sim M^d$, known as "curse of dimensionality"

Kernel Based Methods:

PDE sum of kernel functions with scale h:

$$\tilde{p}(\mathbf{x}) = \frac{1}{N\,h^d} \sum_i K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$



E.g. Gaussian Kernel around MC-points:

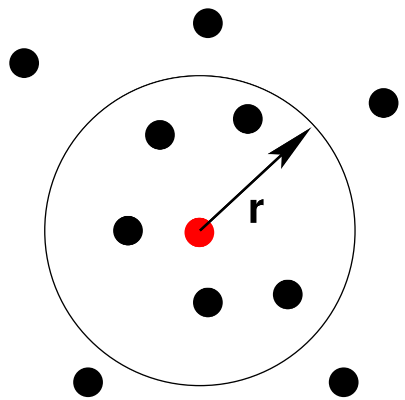
$$K(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}||\mathbf{x}||^2\right)$$

This PDE used by D0 for m_{top}. Slow for large N.

Probability Density Estimation: k Nearest Neighbours

k nearest neighbour method:

expand sphere (or box) around **x** until *k* neighbours of one class are found
⇒ event of this class

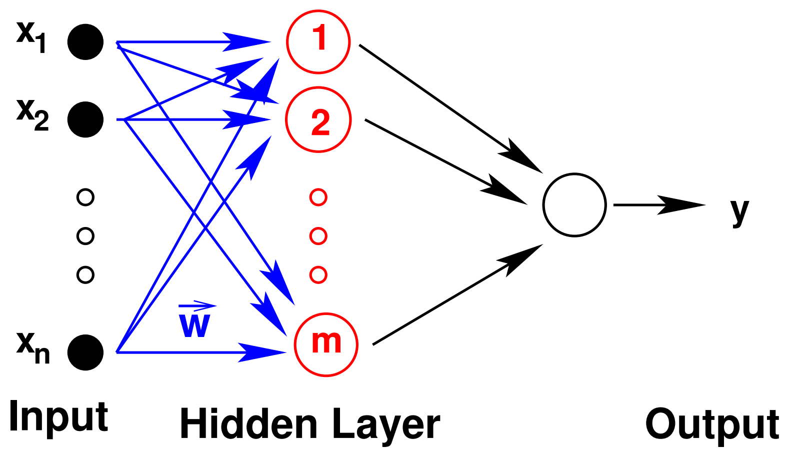


Method similar to kernel methods with hypercubes.

Scale adapted locally to problem
⇒ Better suited for low statistics!

Probability Density Estimation: Neural Networks

Fits PD by adjusting weights in a training phase.



Error function minimised (y_i target values "classes"):

$$R(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} |G(\mathbf{x}_i, \mathbf{w}) - y_i|^2$$

Normally used in HEP: Feed forward nets. Errors backpropagated from output.

Extrapolation out of MC phase space.

Error estimate difficult (Bayesian Networks)

Used for Higgs search at Tevatron run II.

A PDE based on Range Searching (PDE-RS)

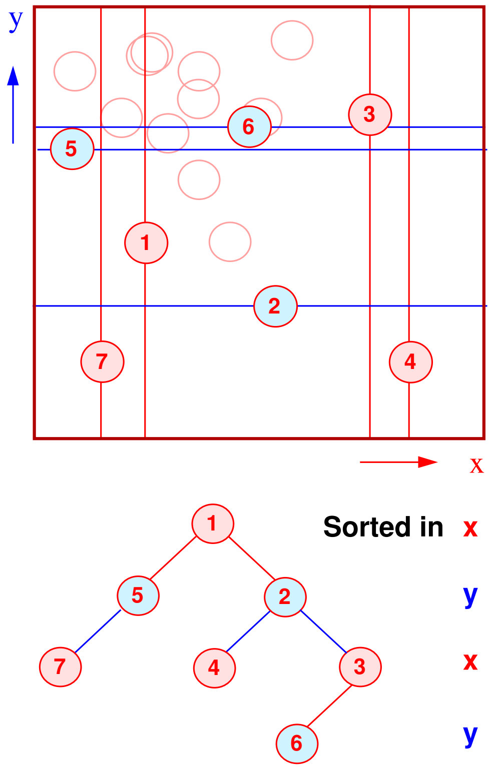
Given sufficient statistics all methods approximate true PD.

High background suppression $\Rightarrow$ large statistics needed

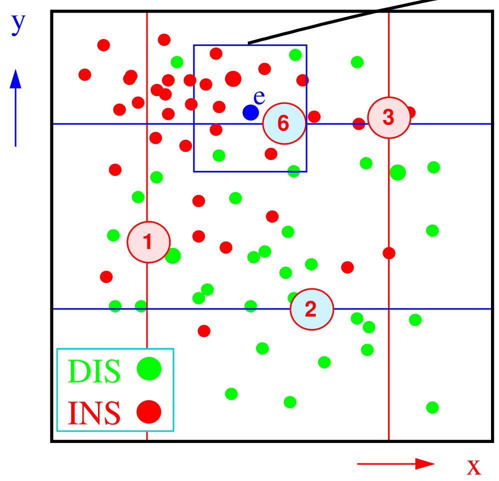
Need fast algorithm to study physical aspects: Systematics, Model dependencies

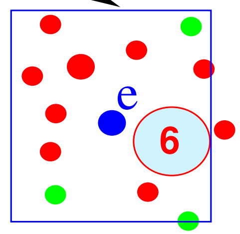
PDE-RS counts MC-events in small box $\triangleq$ Kernel-Method with hypercube

Step 1: Storing event in b–tree



Step 2: finding events in box





$$D = \tilde{p} = \frac{n_S}{n_S + n_B}$$

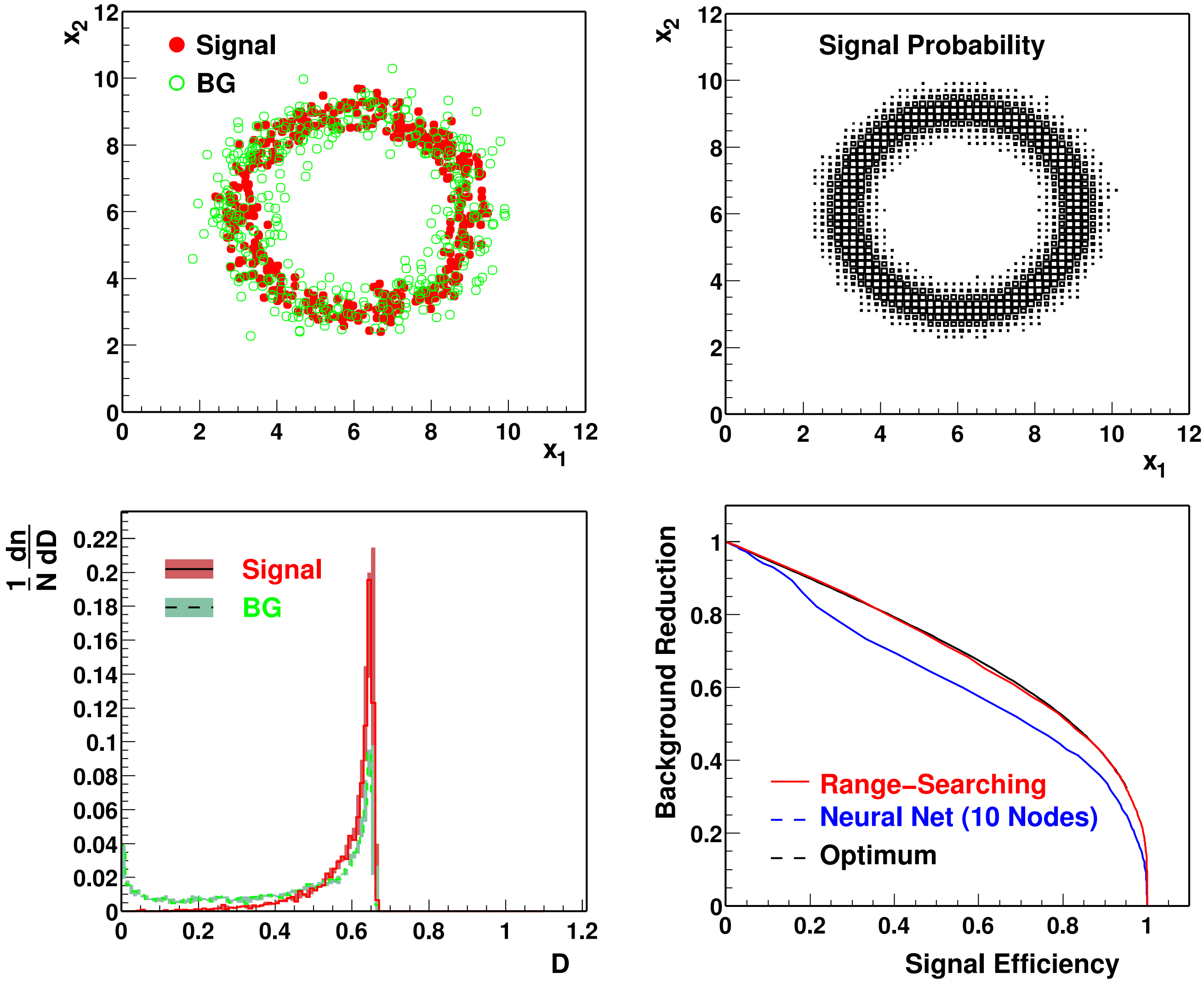
Error on D known!

Finding events in box around $e(x,y)$ analogous filling of b-tree.

Search time $\sim \log N$

Performance of PDE-RS method

High statistics example (500,000 events), highly correlated!



Computing time: PDE-RS: 1 min

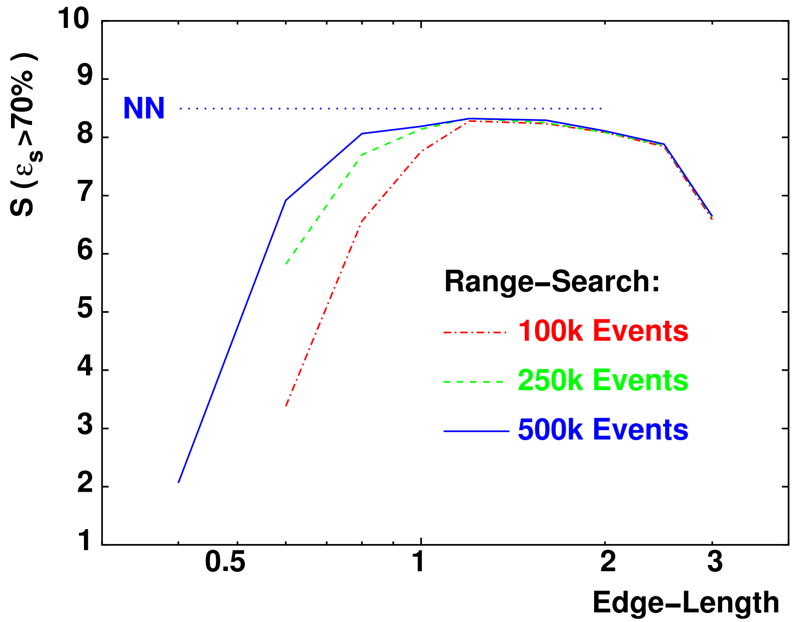
Neural Net: ≈ 2 h

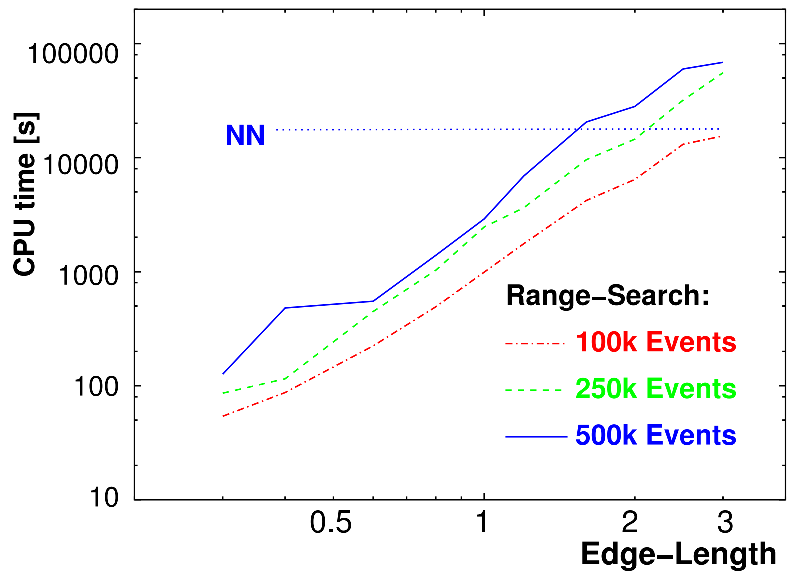
Similar performance of Net only after a day's training.

Properties of PDE-RS method: 5 dim example

Only parameters of PDE-RS: size of box $\Rightarrow$ Fix relative scales of box.

Toy example with multivariate gaussians. Separation $S = \varepsilon_S/\varepsilon_B$





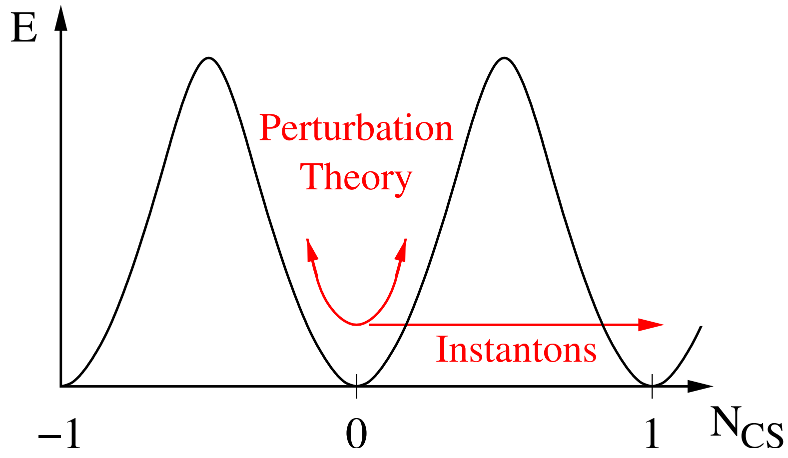
Computing time $\approx$ independent of N: N up $\Rightarrow$ h down

Two analyses done with PDE-RS: Isolated leptons (Zeus), Instantons (H1).

What are Instantons?

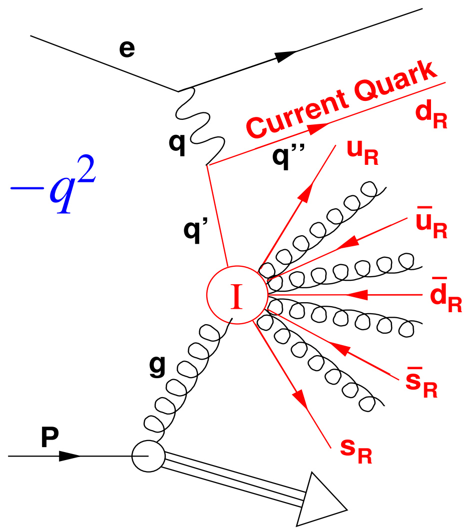
Instantons are:

- Tunneling transitions between topologically different ground states
- Quantum number violating (QCD: Chirality, QFD: $B+L$)
- Part of the Standard Model



Instantons in DIS at HERA:

$Q^2 = -q^2$



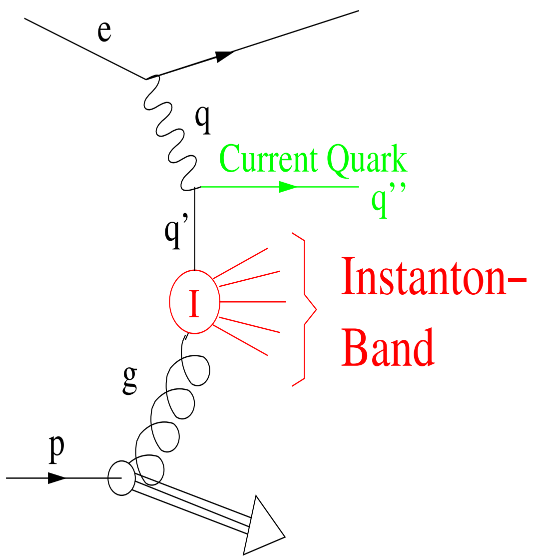
Virtuality of in-coming quark:

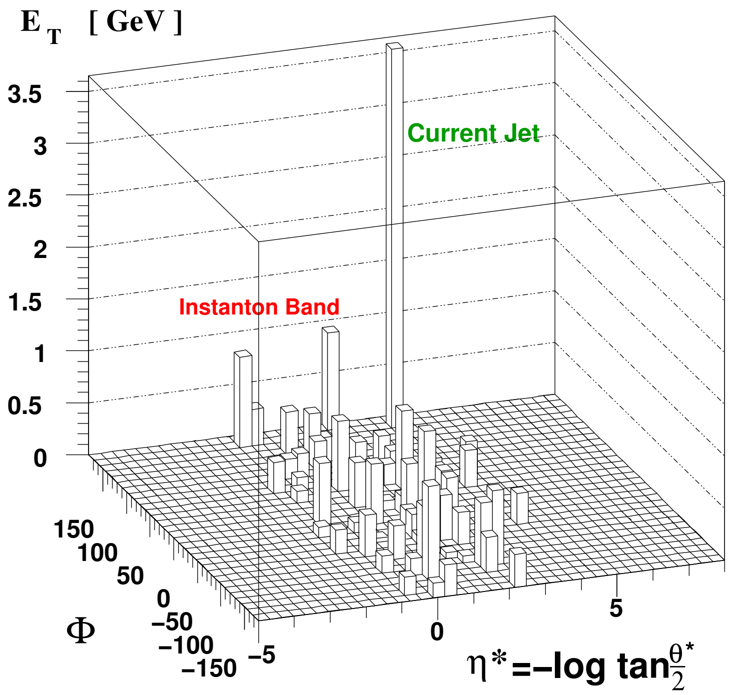
$$Q'^2 = -q'^2$$

An experimental discovery would be a novel manifestation of the non-perturbative properties of QCD.

Properties of Instanton Induced Events

E_T-Map of QCDINS[A. Ringwald, F.Schrempp] Event (hadronic CMS ($\vec{q}+\vec{p}=0$)):





Event properties:

- Hard jet $(\rightarrow Q'^2)$
- Densely populated band in η, with high E_T, isotropic in its rest-frame (→ Shape variables)

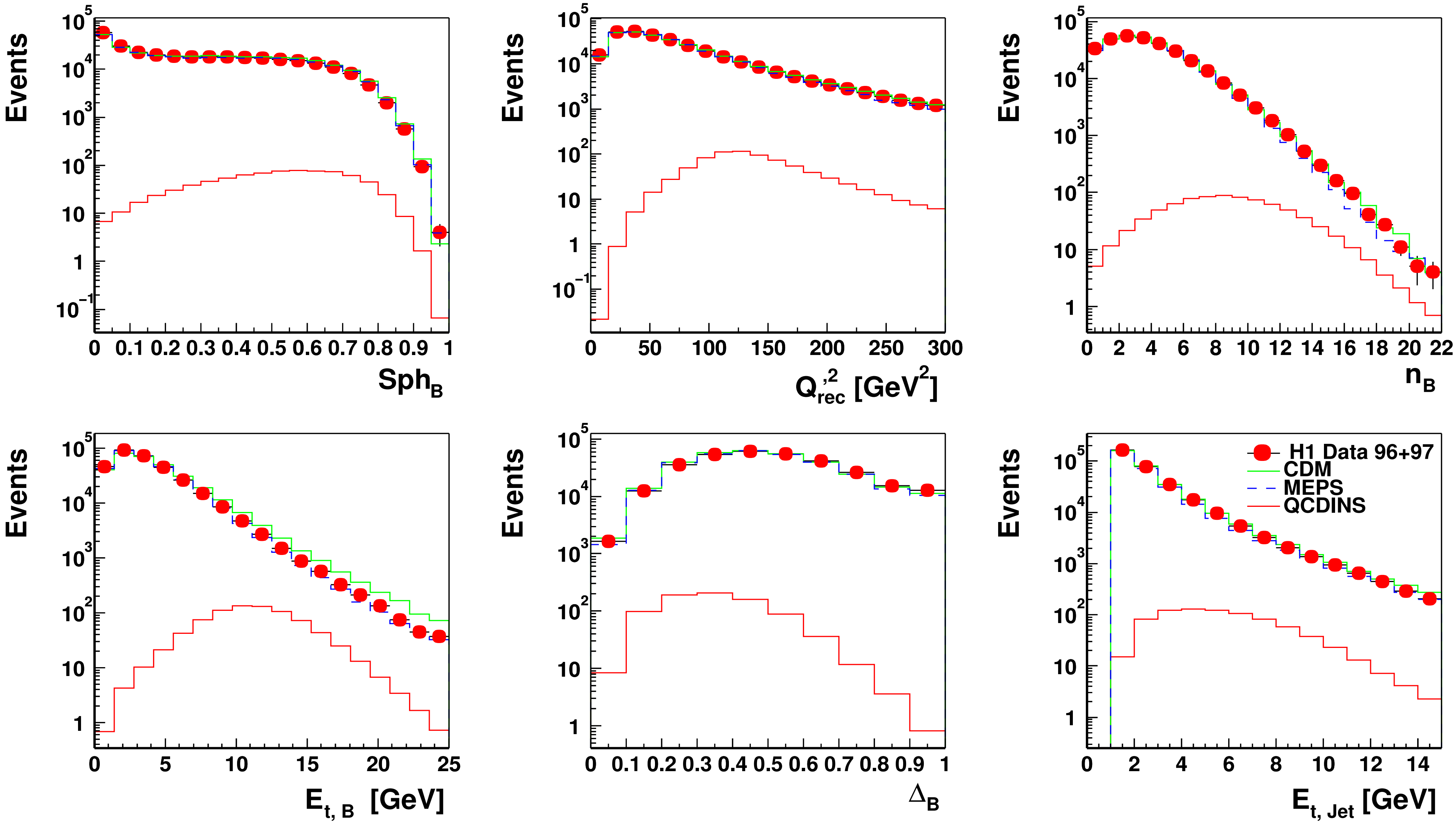
Predicted Cross Section sizeable: [A. Ringwald, F.Schrempp]

$$\sigma^{(I)}_{HERA} = 29.2^{+9.9}_{-8.1}\,\text{pb} \qquad \sigma_{DIS} \approx 3000\,\text{pb}$$

$$(x > 10^{-3},\ 0.1 < y < 0.9,\ Q^2 > 113\,\text{GeV}^2\ )$$

⇒ Search is difficult: Many variables and need for high separation power!

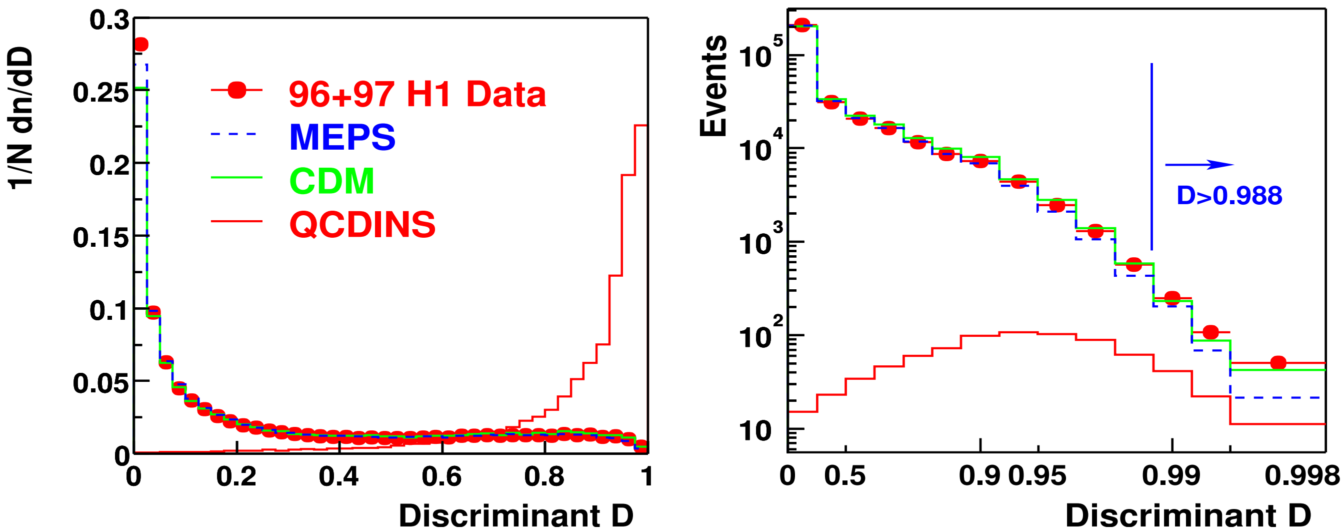
Observables (logarithmic Scale)



Good description by MC's within 5 – 10%

Results of Search

Discriminant optimized for MC: max. $S(\varepsilon_S \geq 10\%)$:



Used cut: $D > 0.988$

$\Rightarrow \varepsilon_{\text{INS}} = 10.0\%$, $S_{\text{CDM}} = 106$, $S_{\text{MEPS}} = 126$

Events after cut:

DATA	CDM	MEPS
410	354^{+40}_{-26}	299^{+25}_{-38}

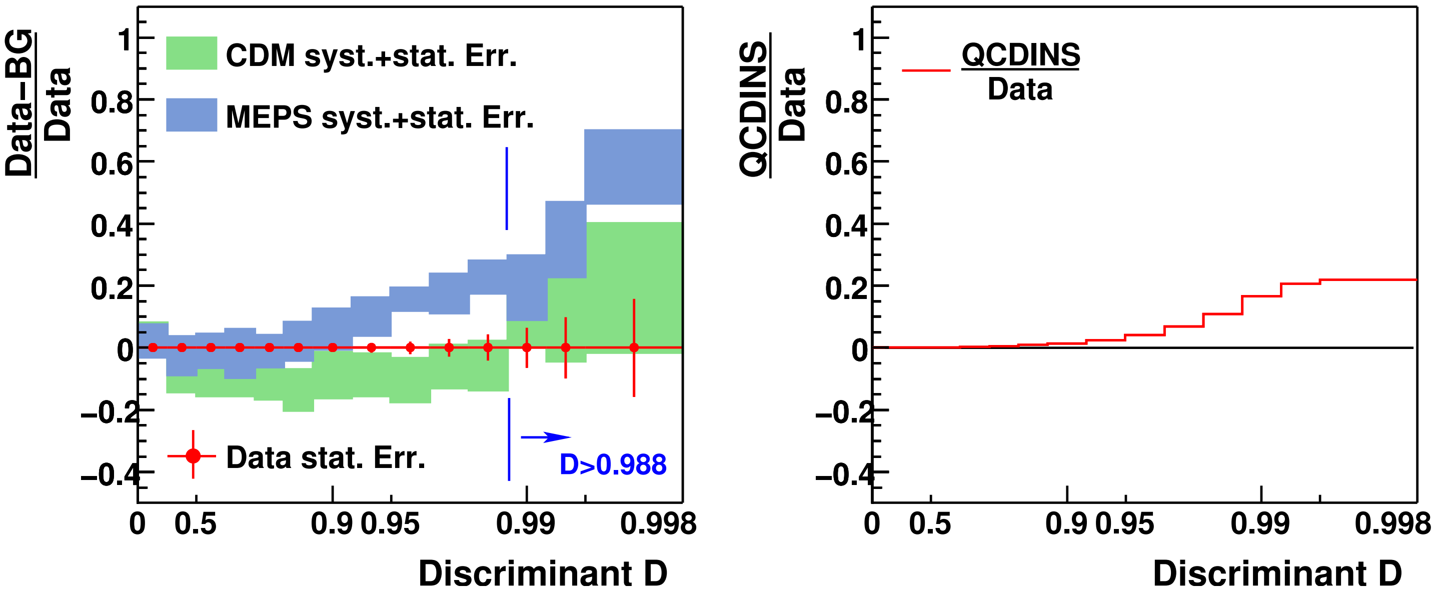
Error contains all statistical and experimental systematic uncertainties.

MEPS model sees some excess in data.

But: No significant excess due to background uncertainties.

Study of discriminant

Discriminant shows transition to I-region:



MEPS describes data for $D < 0.9$, below data in I-region, excess comparable to I-prediction

CDM describes data for $D \approx 0$, but not $0.2 < D < 0.9$, above $D > 0.95$ agreement

Large background uncertainties.

Based on these results H1 has set limits on Instanton production.

Conclusions

- Hadron Colliders need sophisticated background reduction techniques!
- Several PDE methods exist: Kernel based, Neural Nets
- PDE-RS fast method based on range searching: $t \sim \log N$
  $\Rightarrow$ suited for high statistics
  $\Rightarrow$ allows high background reduction
  $\Rightarrow$ allows study of physics aspects of problem
- H1 and Zeus have presented analyses based on PDE-RS

Observables

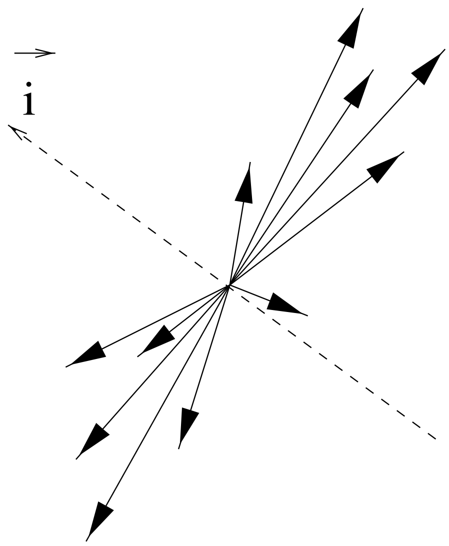
Variables for classification:

1. Virtuality of quark entering I-Subprozess : Q'^2_{rec}
2. Number of charged particles in band: n_{B}
3. Sphericity of band $\mathrm{Sph_B}$

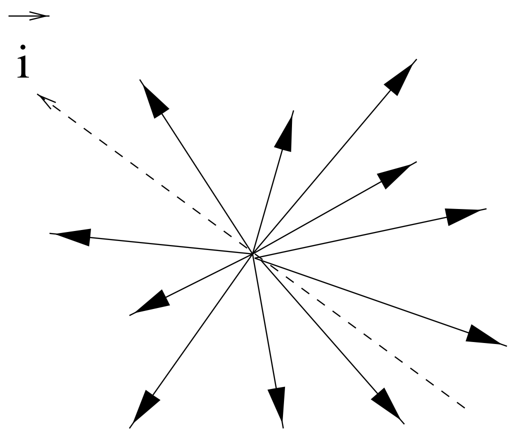
Control variables:

4. $E_{\mathrm{t,Jet}}$ of current jet
5. Total transverse energy of band $E_{\mathrm{t,B}}$
6. Anisotropy of band $\Delta_{\mathrm{B}} = (E_{\mathrm{in,B}} - E_{\mathrm{out,B}})/E_{\mathrm{in,B}}$

$$E_{\mathrm{out}} = \min_{\vec{i}} \sum_{n\ \mathrm{Hadr.}} \left|\vec{p}_n \cdot \vec{i}\right| \qquad E_{\mathrm{in}} = \max_{\vec{i}} \sum_{n\ \mathrm{Hadr.}} \left|\vec{p}_n \cdot \vec{i}\right|$$



$\Delta_{\mathrm{B}} \approx 1$



$\Delta_{\mathrm{B}} \approx 0$

All observables im hCMS except Δ_{B} and $\mathrm{Sph_B}$.

DIS Event-Selection in H1-Detector

Events collected in years 96/97: HERA $27.5\,\mathrm{GeV} \overset{e^+}{\rightarrow}\overset{p}{\leftarrow} 820\,\mathrm{GeV}$, $\mathcal{L} = 21.1\,\mathrm{pb}^{-1}$,

375000 Events selected

Phase space:

$\theta_{\mathrm{el}} > 156^\circ$, $0.1 < y < 0.6$, $x > 10^{-3}$

Simulated I-Event in H1-Detector: