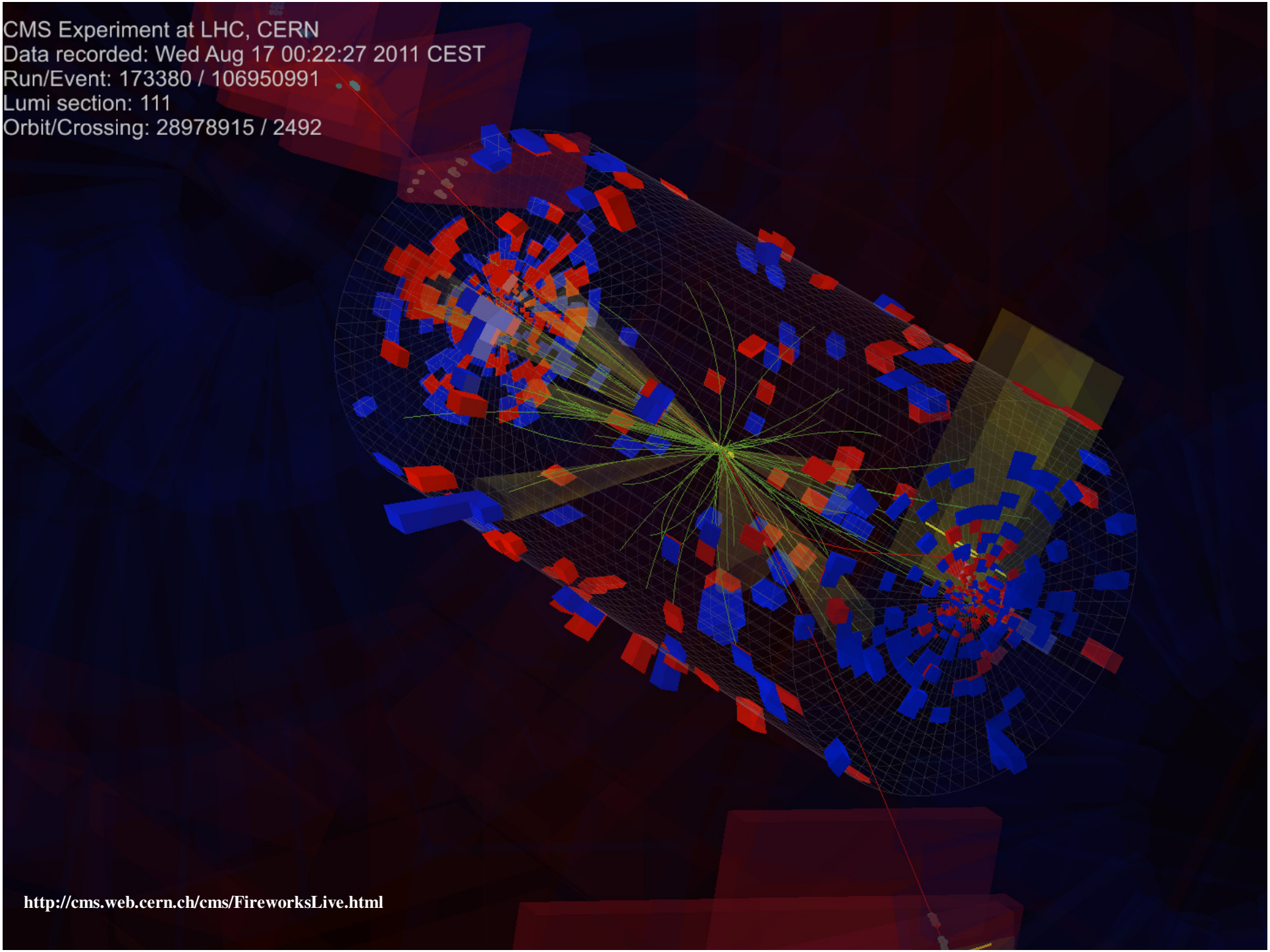
Gibbs sampler for background discrimination in particle physics

ACAT Workshop
Brunel University
September 2011

Federico Colecchia


CMS Experiment at LHC, CERN
Data recorded: Wed Aug 17 00:22:27 2011 CEST
Run/Event: 173380 / 106950991
Lumi section: 111
Orbit/Crossing: 28978915 / 2492
http://cms.web.cern.ch/cms/FireworksLive.html

Motivation

- Background discrimination in particle physics
  - ✓ Events
  - ✓ **Particles**

Motivation

- Background discrimination in particle physics
  - ✓ Events
  - ✓ **Particles**
- Looking at individual particles inside events
  - ✓ Population-based view of particle physics events
  - ✓ Statistical mixture model decomposition: **Gibbs sampler** (GS)

Motivation

- Background discrimination in particle physics
  - ✓ Events
  - ✓ **Particles**
- Looking at individual particles inside events
  - ✓ Population-based view of particle physics events
  - ✓ Statistical mixture model decomposition: **Gibbs sampler** (GS)
- Classification task
  - ✓ Supervised/unsupervised classifiers & **statistical fluctuations**
  - ✓ The Gibbs sampler can be operated in both modes

Motivation

- Background discrimination in particle physics
  - ✓ Events
  - ✓ **Particles**
- Looking at individual particles inside events
  - ✓ Population-based view of particle physics events
  - ✓ Statistical mixture model decomposition: **Gibbs sampler** (GS)
- Classification task
  - ✓ Supervised/unsupervised classifiers & **statistical fluctuations**
  - ✓ The Gibbs sampler can be operated in both modes

Map individual particles to signal/background using PDFs estimated from the data as opposed to high-statistics templates

Motivation

⇨ **POPULATION-BASED MIXTURE MODEL APPROACH**

Motivation

- **POPULATION-BASED MIXTURE MODEL APPROACH**
- **PARTICLE-LEVEL CLASSIFICATION**

Motivation

- **POPULATION-BASED MIXTURE MODEL APPROACH**
- **PARTICLE-LEVEL CLASSIFICATION**
- **DATA-DRIVEN BACKGROUND ESTIMATION**

Statistical model

$\alpha_0 f_0(\eta, p_T) + \alpha_1 f_1(\eta, p_T)$ with $\alpha_0 + \alpha_1 = 1$

bkg PDF sig PDF

"MIXTURE WEIGHTS"

GIBBS SAMPLER*

At each iteration:

For each particle i:

1. Calculate the PDFs for bkg/sig (j=0,1) using η^i, p^i_T: f_0, f_1
2. Map particle i to population j=0,1 with probability $f_j/(f_0+f_1)$
3. Re-estimate mixture weights α_j, j=0,1
4. Re-estimate PDFs

* Geman S and Geman D 1984 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (6) 721–41

Statistical model

$\alpha_0 f_0(\eta, p_T) + \alpha_1 f_1(\eta, p_T)$ with $\alpha_0+\alpha_1=1$

bkg PDF sig PDF

"MIXTURE WEIGHTS"

GIBBS SAMPLER*

At each iteration:

For each particle i:

1. Calculate the PDFs for bkg/sig (j=0,1) using η^i, p^i_T: f_0, f_1
2. Map particle i to population j=0,1 with probability $f_j/(f_0+f_1)$
3. Re-estimate mixture weights α_j, j=0,1
4. Re-estimate PDFs

* Geman S and Geman D 1984 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (6) 721–41

Statistical model

$\alpha_0 f_0(\eta, p_T) + \alpha_1 f_1(\eta, p_T)$ with $\alpha_0 + \alpha_1 = 1$

bkg PDF sig PDF

"MIXTURE WEIGHTS"

GIBBS SAMPLER*

At each iteration:

For each particle i:

1. Calculate the PDFs for bkg/sig (j=0,1) using η^i, p^i_T: f_0, f_1
2. Map particle i to population j=0,1 with probability $f_j/(f_0+f_1)$
3. Re-estimate mixture weights α_j, j=0,1
4. Re-estimate PDFs

* Geman S and Geman D 1984 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (6) 721–41

Statistical model

$\alpha_0 f_0(\eta, p_T) + \alpha_1 f_1(\eta, p_T)$ with $\alpha_0 + \alpha_1 = 1$

bkg PDF sig PDF

"MIXTURE WEIGHTS"

GIBBS SAMPLER*

At each iteration:

For each particle i:

1. Calculate the PDFs for bkg/sig (j=0,1) using η^i, p^i_T: f_0, f_1
2. Map particle i to population j=0,1 with probability $f_j/(f_0+f_1)$
3. Re-estimate mixture weights α_j, j=0,1
4. Re-estimate PDFs

* Geman S and Geman D 1984 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (6) 721–41

Monte Carlo study

- Pythia 8140:

  $gg \rightarrow t\bar{t}$ signal + 7 Minimum Bias interactions

  In line with pileup levels at the LHC as of July 2011

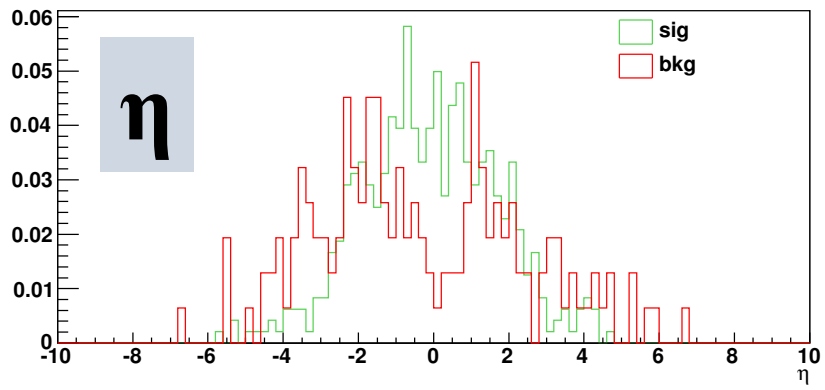
- Particle kinematics: η, p_T

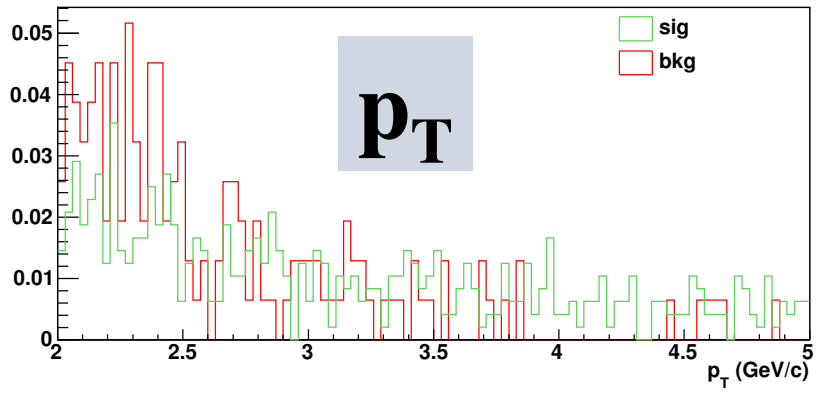
  2 GeV/c < p_T < 5 GeV/c – neglect correlations

- Input data set: **481 signal particles + 155 background particles**

  ✓ Percentage of background particles ~24%

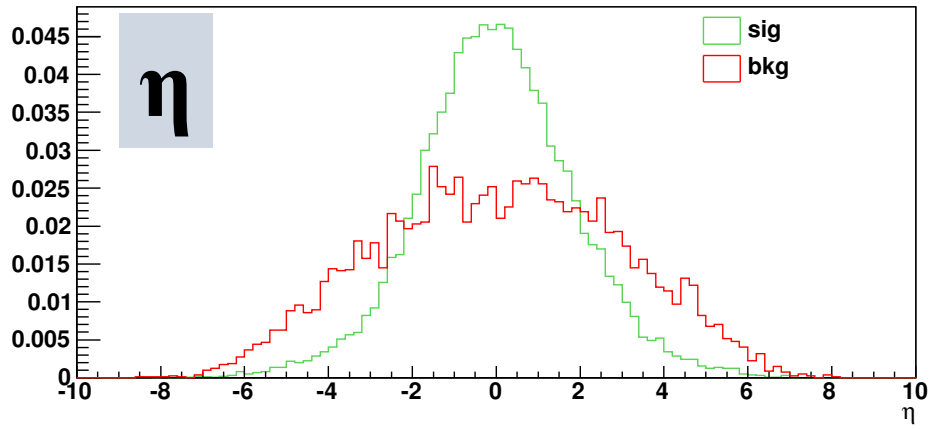
Input data set: 636 particles

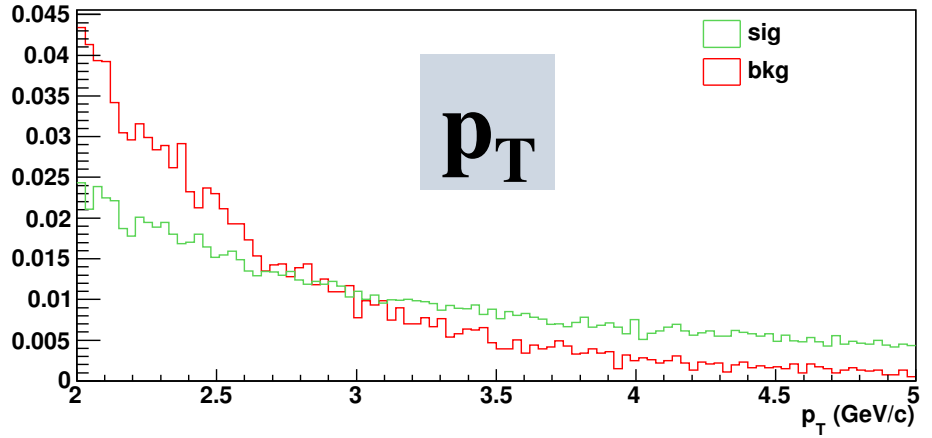


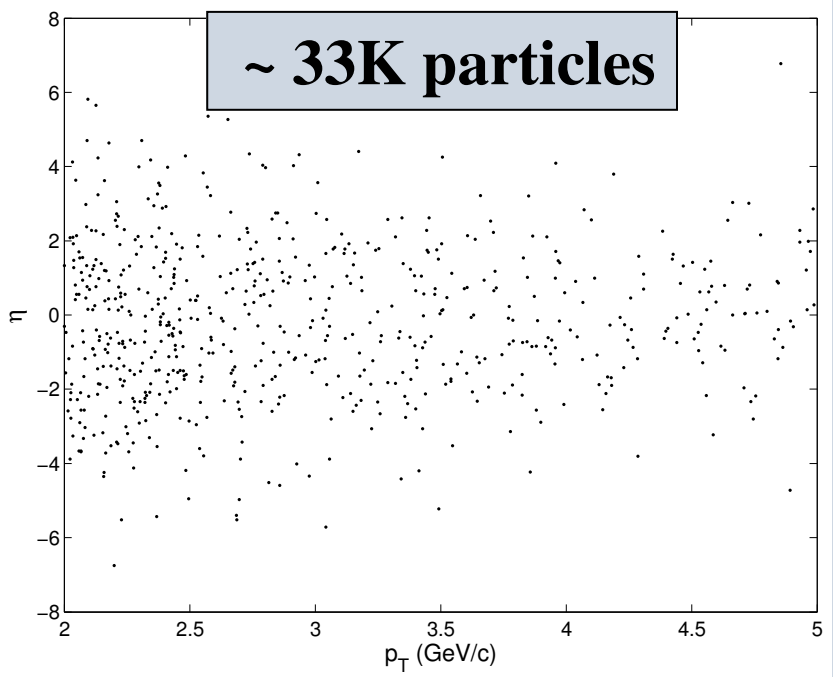


- 481 signal particles
- 155 background particles

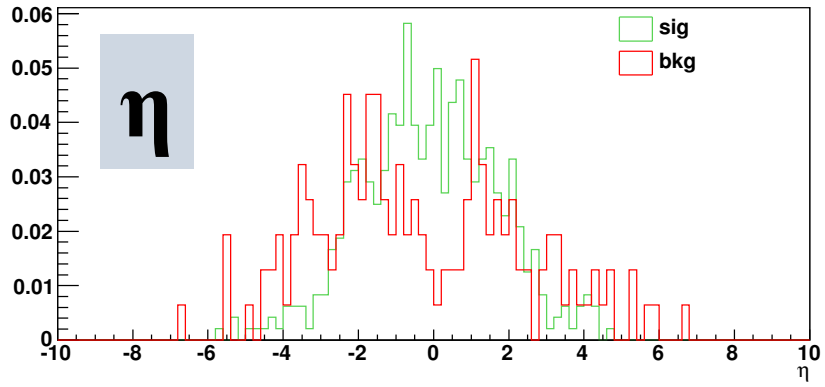
High-statistics control sample

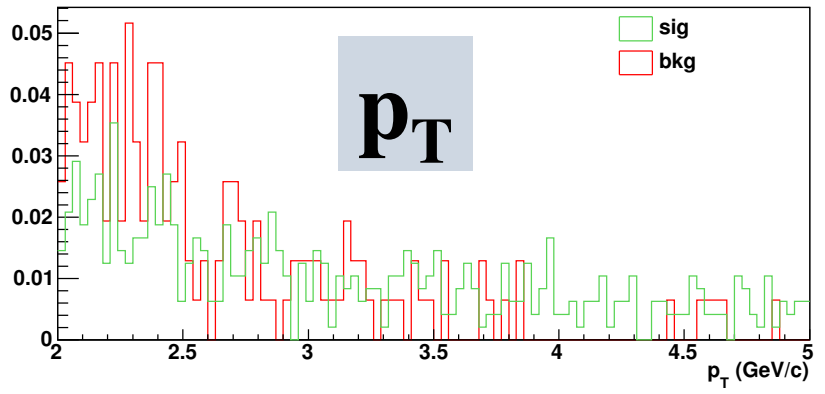






Input data set: 636 particles

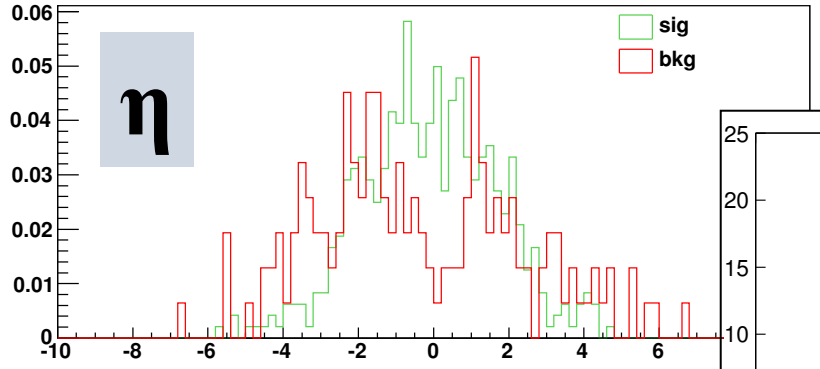


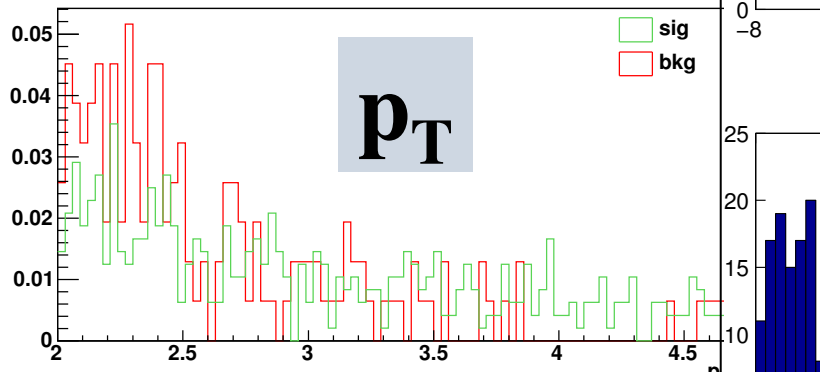


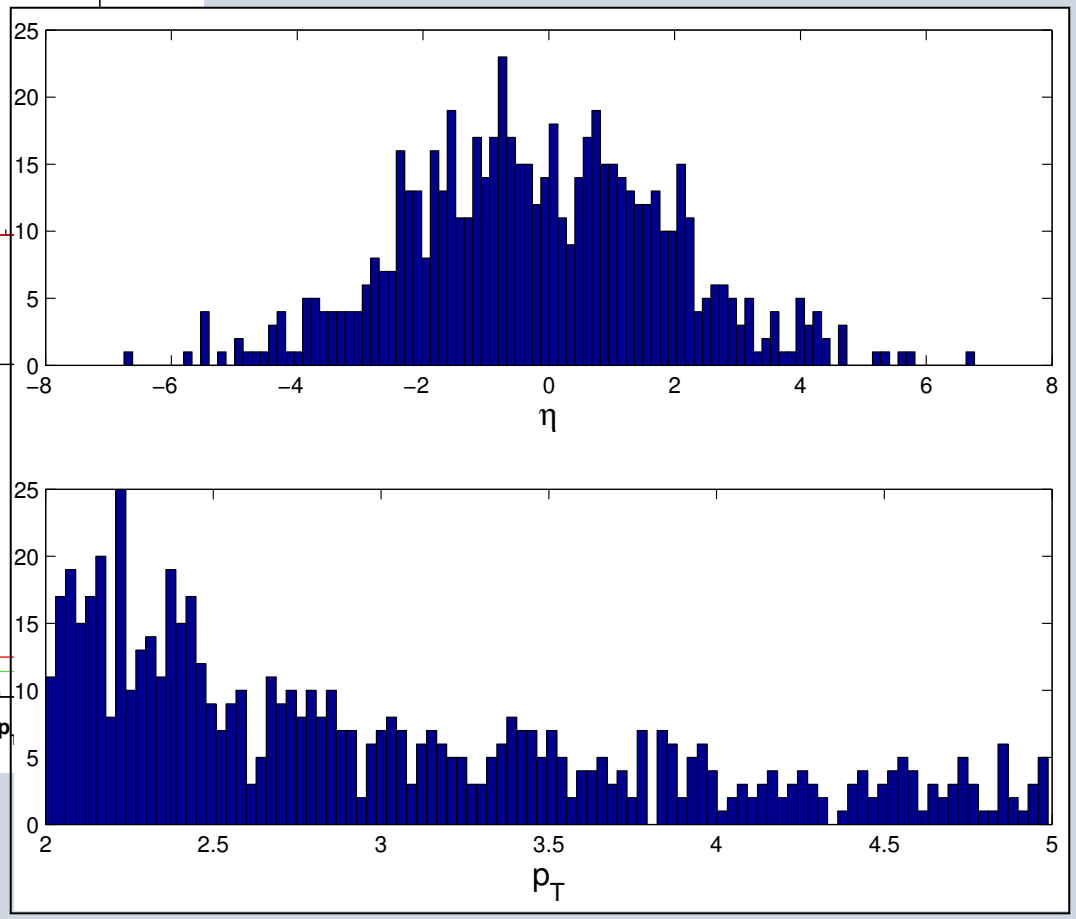
- 481 signal particles
- 155 background particles

Input data set: 636 particles

- 481 signal particles
- 155 background particles







Preferred strategy

SUPERVISED GIBBS SAMPLER

- **Initialisation**
  - PDFs ~ high-statistics control sample
  - $\alpha_0 = \alpha_1 = 0.5$
- **Main loop – α_j floating, PDFs not updated**

Preferred strategy

SUPERVISED GIBBS SAMPLER

- **Initialisation**
  - PDFs ~ high-statistics control sample
  - $\alpha_0 = \alpha_1 = 0.5$
- **Main loop – α_j floating, PDFs not updated**

⬇

α_j **MIXTURE WEIGHTS**

Preferred strategy

SUPERVISED GIBBS SAMPLER

- ❑ **Initialisation**
  - PDFs ~ high-statistics control sample
  - $\alpha_0 = \alpha_1 = 0.5$
- ❑ **Main loop – α_j floating, PDFs not updated**

⬇

α_j **MIXTURE WEIGHTS**

⬇

UNSUPERVISED GIBBS SAMPLER

- ❑ **Initialisation**
  - PDFs ~ high-statistics control sample
  - α_j set to estimate from previous run
- ❑ **Main loop – α_j fixed, PDFs updated**

⬇

PDF SHAPES

Estimating PDF shapes

- At each iteration, particles are mapped to sig/bkg based on current PDF estimates

Estimating PDF shapes

- At each iteration, particles are mapped to sig/bkg based on current PDF estimates
- Unsupervised GS: Obtain new estimates of PDFs via **histogram regularisation**

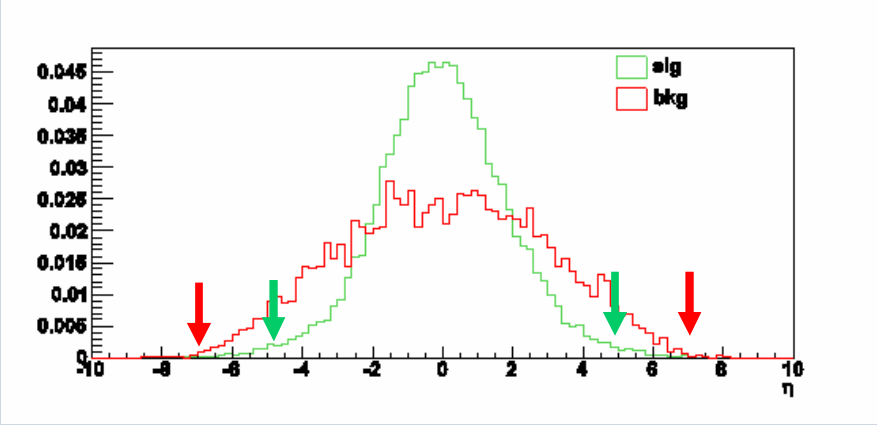
Estimating PDF shapes

- At each iteration, particles are mapped to sig/bkg based on current PDF estimates
- Unsupervised GS: Obtain new estimates of PDFs via **histogram regularisation**
- Use a priori information from control sample to **suppress spurious oscillations**

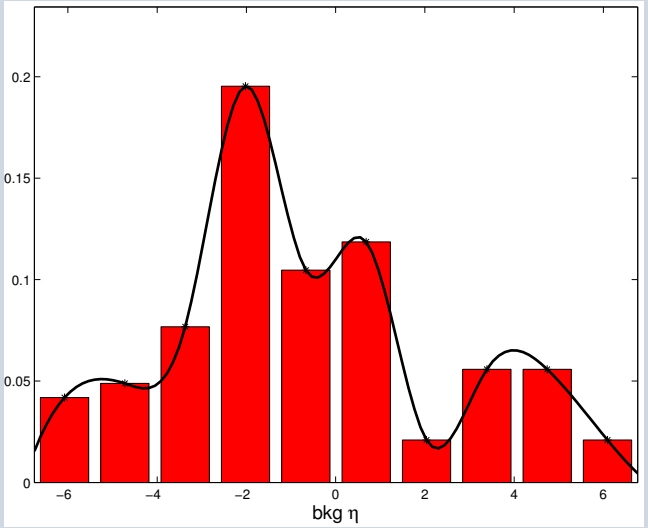
Estimating PDF shapes

- At each iteration, particles are mapped to sig/bkg based on current PDF estimates
- Unsupervised GS: Obtain new estimates of PDFs via **histogram regularisation**
- Use a priori information from control sample to **suppress spurious oscillations**
- Spline interpolation with boundary conditions

Control sample



Unsupervised GS
@ given iteration



Results: mixture weights

RESULTS
Supervised GS

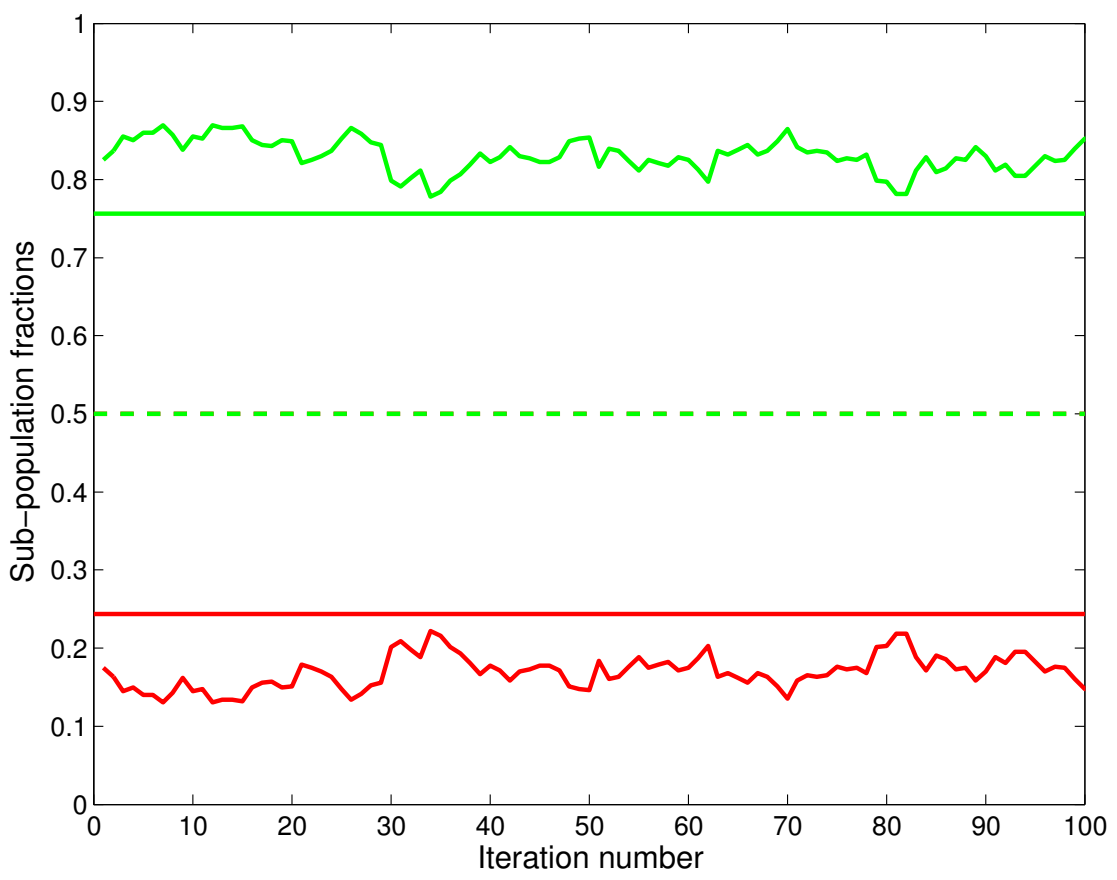
bkg

bkg

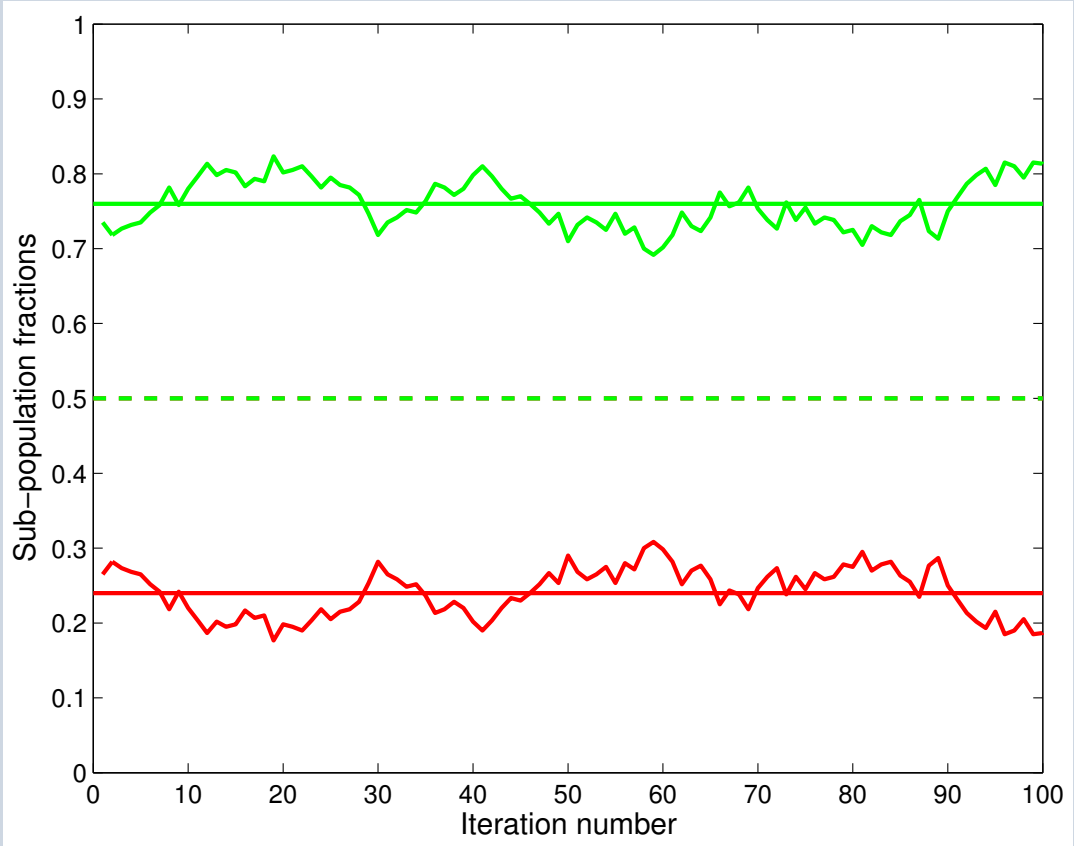
SYSTEMATICS:

- PDFs from control sample
- No contribution from histogram regularisation
- Will need to be studied in order to optimise performance

MIXTURE WEIGHTS



Validation on toy Monte Carlo



***Supervised GS* run on toy Monte Carlo**

Subpopulation PDFs given by truth

Results: PDFs

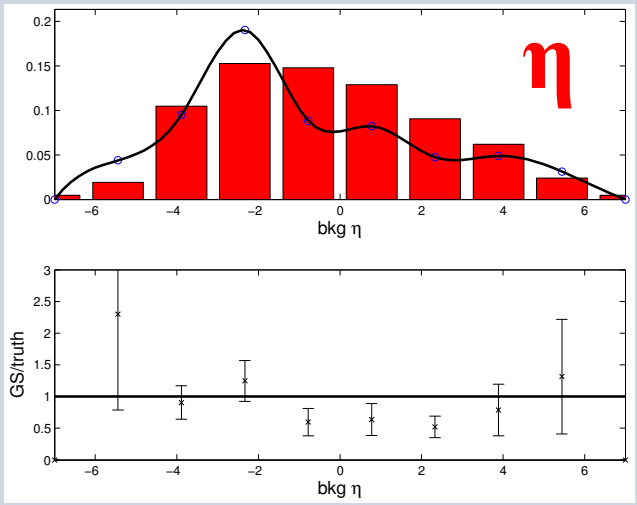
RESULTS
Unsupervised GS

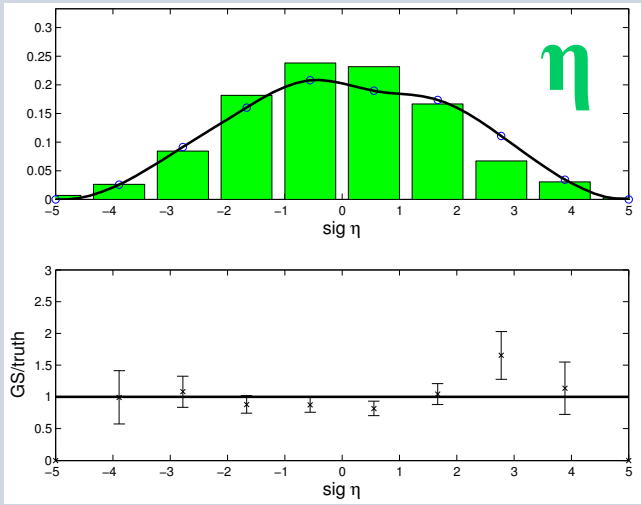
bkg

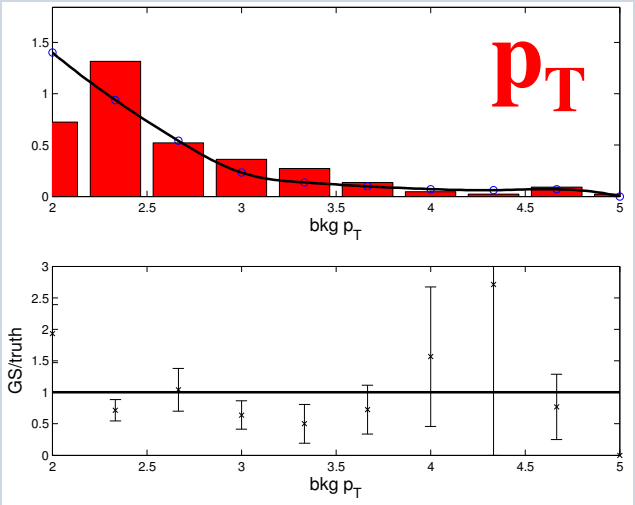
bkg

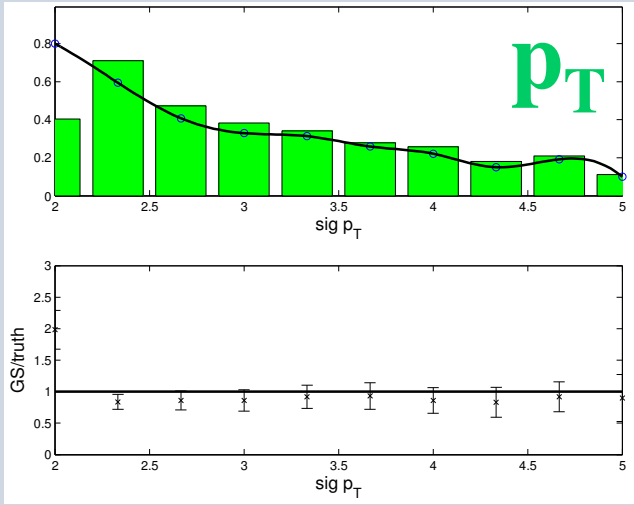
SYSTEMATICS:

Regularisation









Results: PDFs

RESULTS
Unsupervised GS

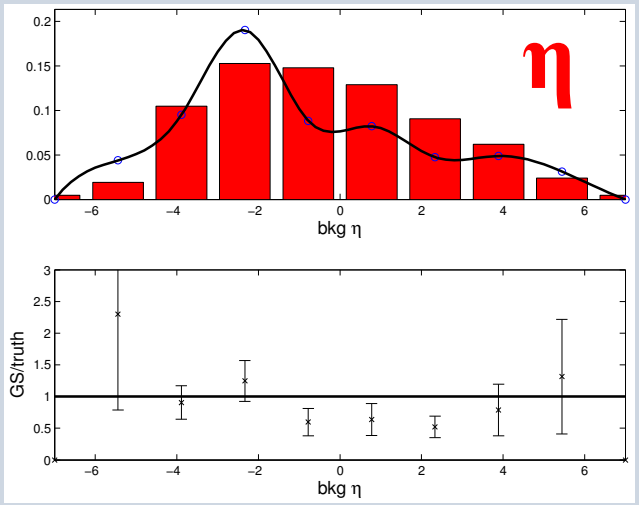
bkg

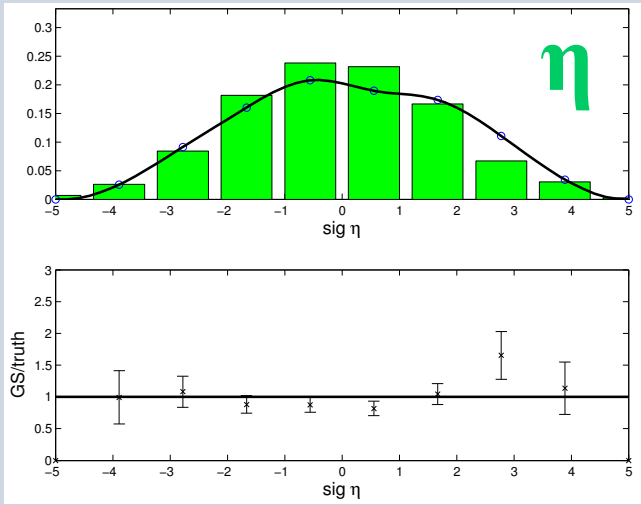
bkg

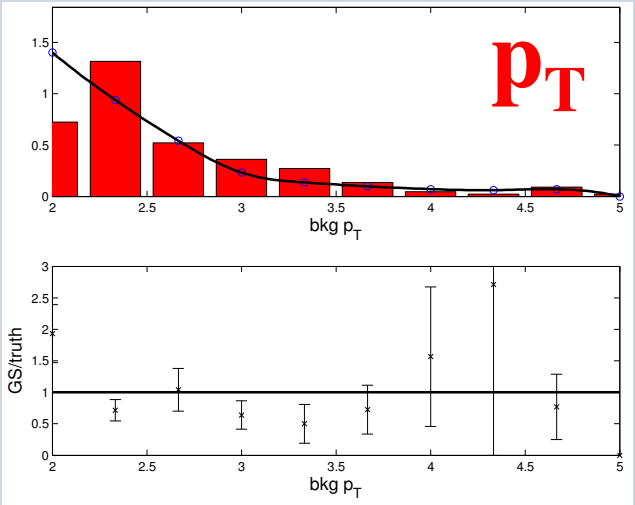
SYSTEMATICS:

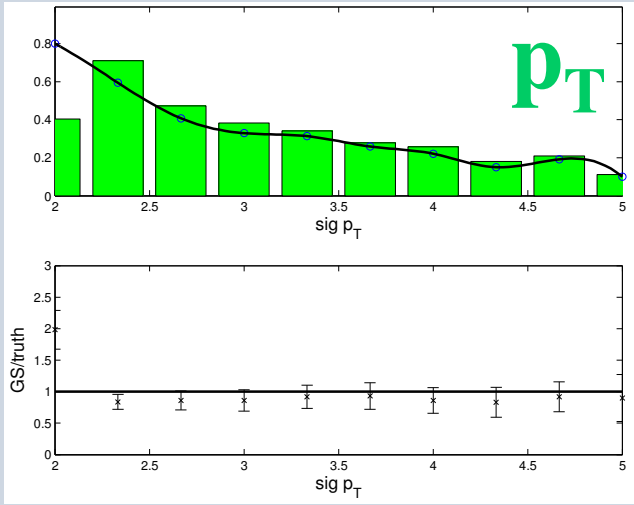
Regularisation

SYSTEMATICS TO BE EVALUATED FOR SPECIFIC ANALYSIS



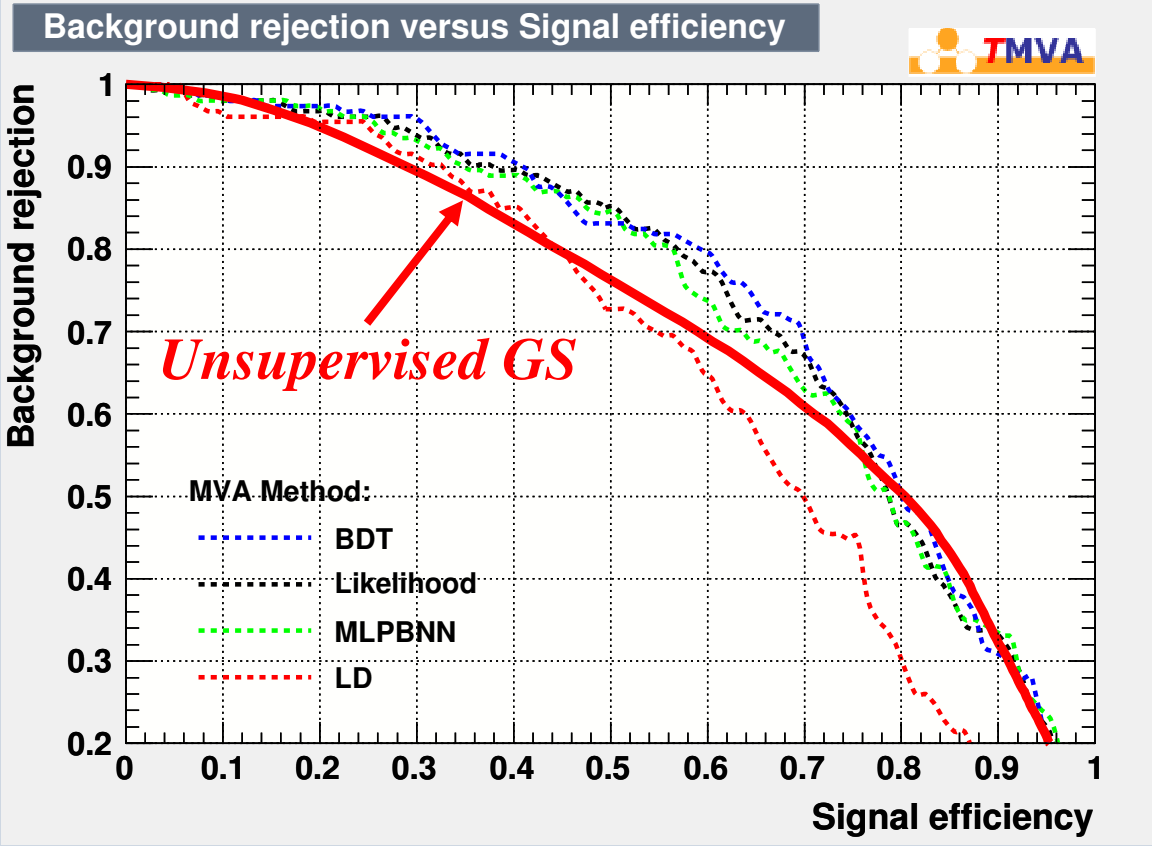






Classification performance

TMVA* V04-01-00



- BDT, Likelihood, NN, LD
- Trained on control sample
- Run on input data set

*Hoecker A *et al* 2007 arXiv:physics/0703039

Conclusions

- Population-based view of particle physics events
  - Focus on individual particles inside events

Conclusions

- Population-based view of particle physics events
  - Focus on individual particles inside events
- Gibbs sampler for mixture model decomposition
  - Hybrid supervised/unsupervised

Conclusions

- Population-based view of particle physics events
  - Focus on individual particles inside events
- Gibbs sampler for mixture model decomposition
  - Hybrid supervised/unsupervised
- Proof of concept on Monte Carlo data
  - Map individual particles to signal/background using PDFs estimated directly from the data
  - Illustrated one possible combined use of supervised/unsupervised GS: strategy may have to be optimised for the specific analysis

Outlook

- **FUTURE WORK:**
  Further investigate bias on mixture weights to improve performance

Outlook

- **FUTURE WORK:**
  Further investigate bias on mixture weights to improve performance

- **PROSPECTIVE GOAL:**
  Tools for intensive offline analysis of individual interesting events at the LHC

Acknowledgments

- **UCL HEP GROUP**
  Prof. Jonathan Butterworth

- Department of Astronomy and Theoretical Physics,
  Lund University, Sweden

BACKUP SLIDES

Revised statistical model

- Assuming an explicit functional form for sig/bkg PDFs leads to bias*
- Estimate PDF shapes by means of **histogram regularisation**, as opposed to estimating PDF parameters given a functional form

$$\alpha_0 f_0(\eta, p_T|\underline{\theta}_0) + \alpha_1 f_1(\eta, p_T|\underline{\theta}_1) \text{ with } \alpha_0+\alpha_1=1$$

$$\alpha_0 f_0(\eta, p_T) + \alpha_1 f_1(\eta, p_T) \text{ with } \alpha_0+\alpha_1=1$$

- 1D histograms for the sake of simplicity – limited kinematic region

$$f_j(\eta, p_T) = g_j(\eta)\, h_j(p_T),\ j=0,1$$

*FC, arXiv:1010.6051

Supervised/unsupervised GS

- PDF shapes initialised using high-statistics sample
- Initialise mixture weights as 50:50

GIBBS SAMPLER

At each iteration:

For each particle i:

1. **Calculate the PDFs for bkg/sig (j=0,1) using η^i, p^i_T: f_0, f_1**
2. **Map particle i to population j=0,1 with probability $f_j/(f_0+f_1)$**
3. **Re-estimate mixture weights α_j, j=0,1**
4. **Re-estimate PDFs**

Supervised/unsupervised GS

- PDF shapes initialised using high-statistics sample
- Initialise mixture weights as 50:50

GIBBS SAMPLER

At each iteration:

For each particle i:

1. Calculate the PDFs for bkg/sig (j=0,1) using η^i, p^i_T: f_0, f_1

2. Map particle i to population j=0,1 with probability $f_j/(f_0+f_1)$

3. Re-estimate mixture weights α_j, j=0,1

4. Re-estimate PDFs

Supervised/unsupervised GS

- PDF shapes initialised using high-statistics sample
- Initialise mixture weights as 50:50

GIBBS SAMPLER

At each iteration:

For each particle i:

1. Calculate the PDFs for bkg/sig (j=0,1) using η^i, p^i_T: f_0, f_1

2. Map particle i to population j=0,1 with probability $f_j/(f_0+f_1)$

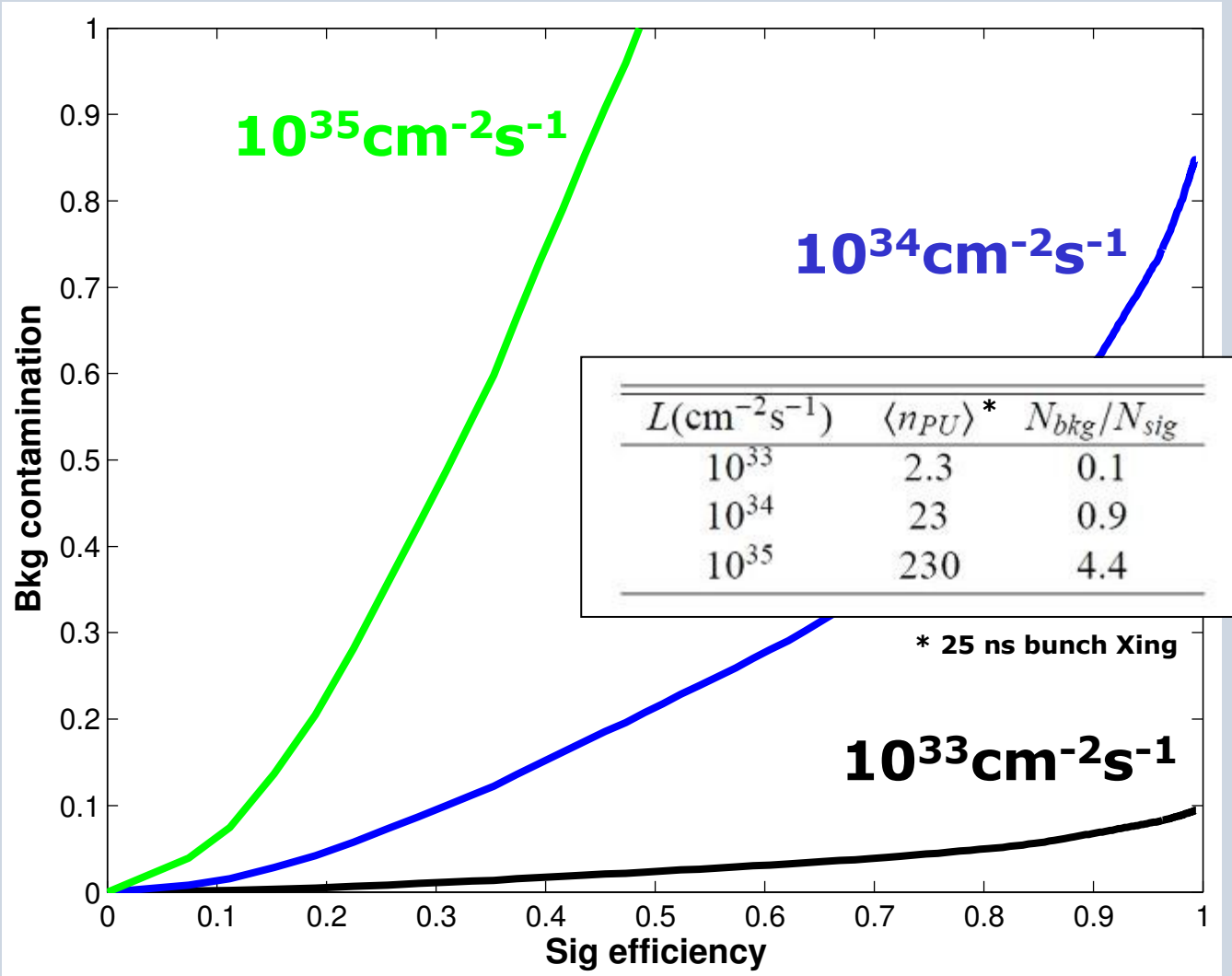
3. Re-estimate mixture weights α_j, j=0,1

4. Re-estimate PDFs

- At each iteration, particles are mapped to sig/bkg
  - ✓ Re-estimate mixture weights
  - ✓ **Re-estimate PDFs: switched off with unsupervised GS**

Expected performance @ LHC

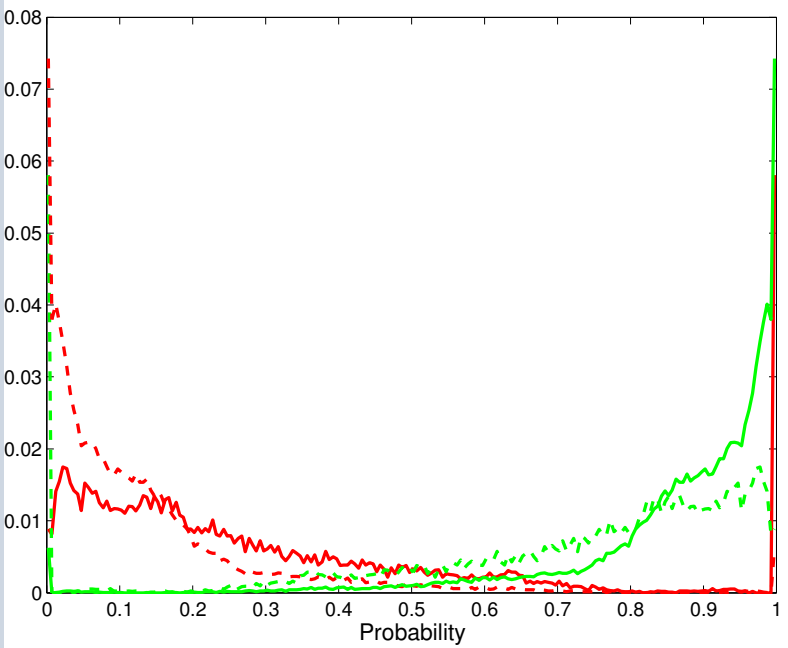
 # misclassified bkg particles / # sig particles



$L(\text{cm}^{-2}\text{s}^{-1})$	$\langle n_{PU} \rangle$*	N_{bkg}/N_{sig}
10^{33}	2.3	0.1
10^{34}	23	0.9
10^{35}	230	4.4

* 25 ns bunch Xing

Results: probabilities (*uGS*)



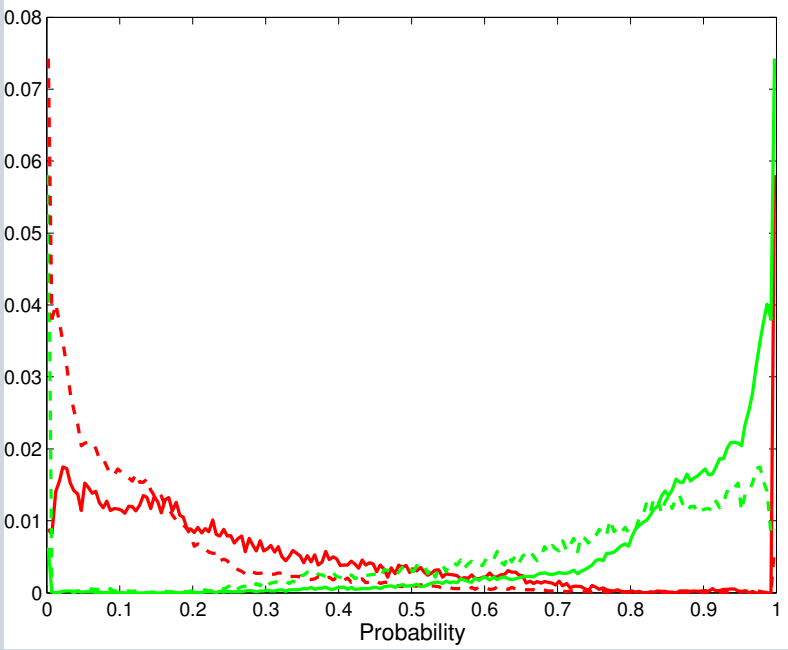
$P_{sig->sig}$

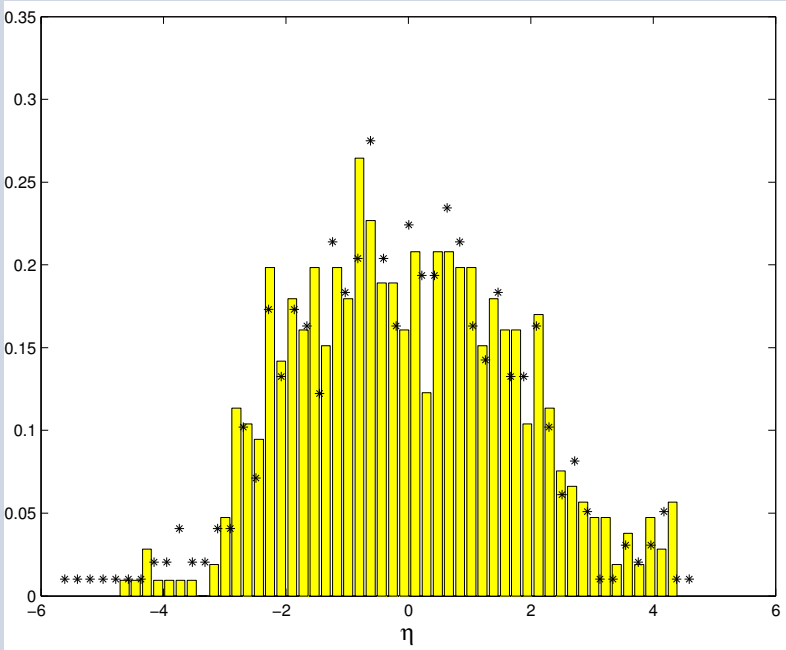
$P_{bkg->sig}$

$P_{bkg->bkg}$

$P_{sig->bkg}$

Results: probabilities (*uGS*)





$P_{sig->sig}$

$P_{bkg->sig}$

$P_{bkg->bkg}$

$P_{sig->bkg}$

Points: particles with $P_{sig}>0.5$

Bars: truth