HF jets analysis

13.04.2020 ALICE@IFJ meeting

Sebastian Bysiak

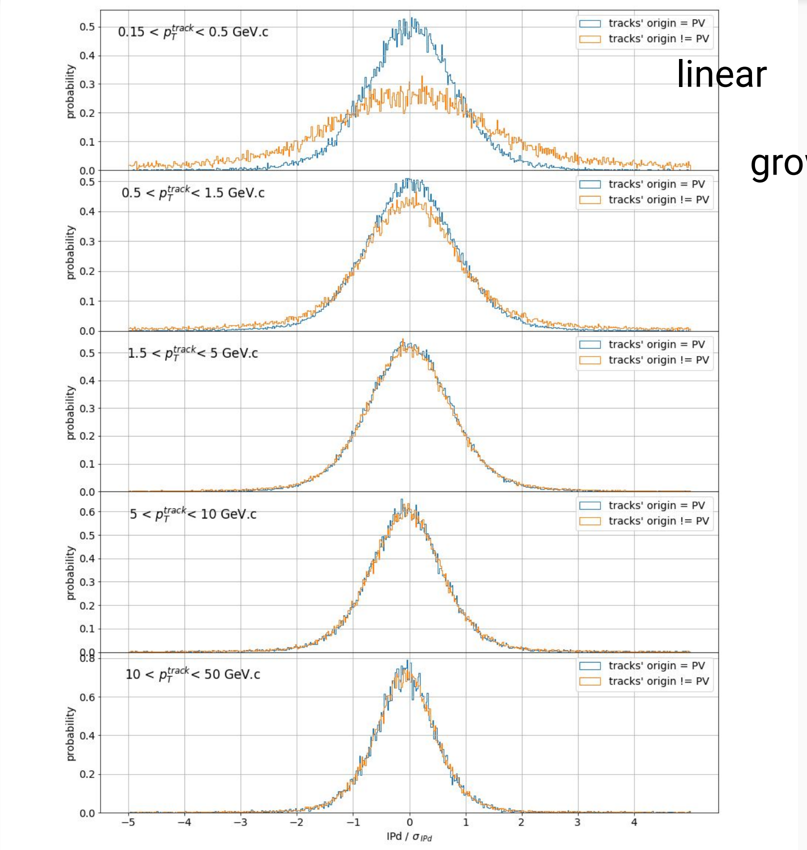
Outline

1. Progress in HF-jets analysis
2. Questions & issues
3. Plans for next week

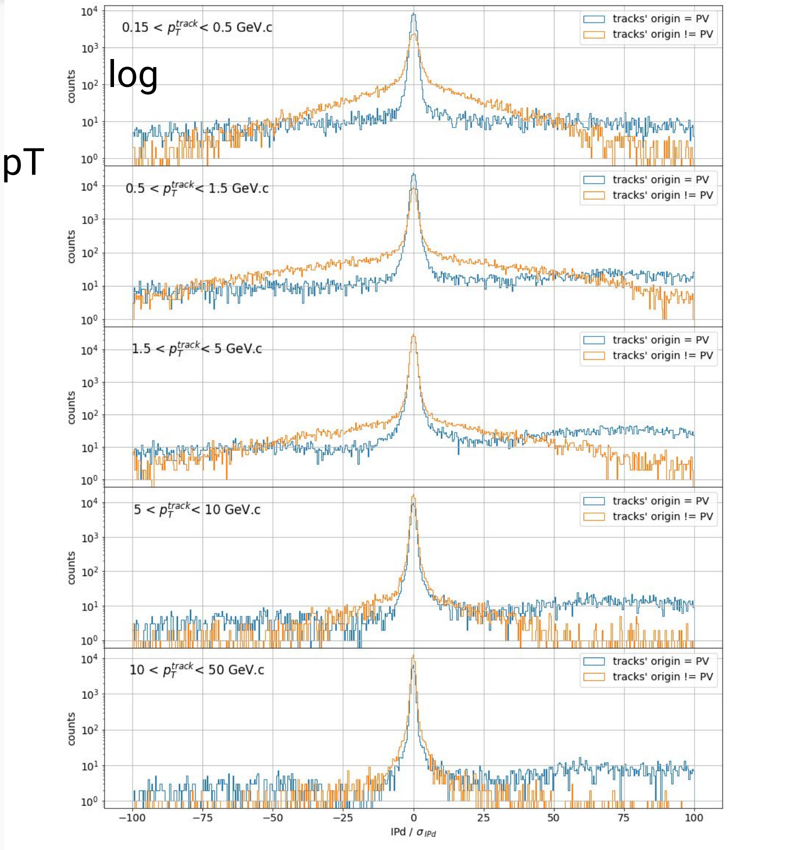
QA data re-merging

- LHC18r - remerging of 59 out of 100 runs completed (accessible on EOS)
- LHC18q - pass3 ongoing on ALICE level, 119 out of 144 started

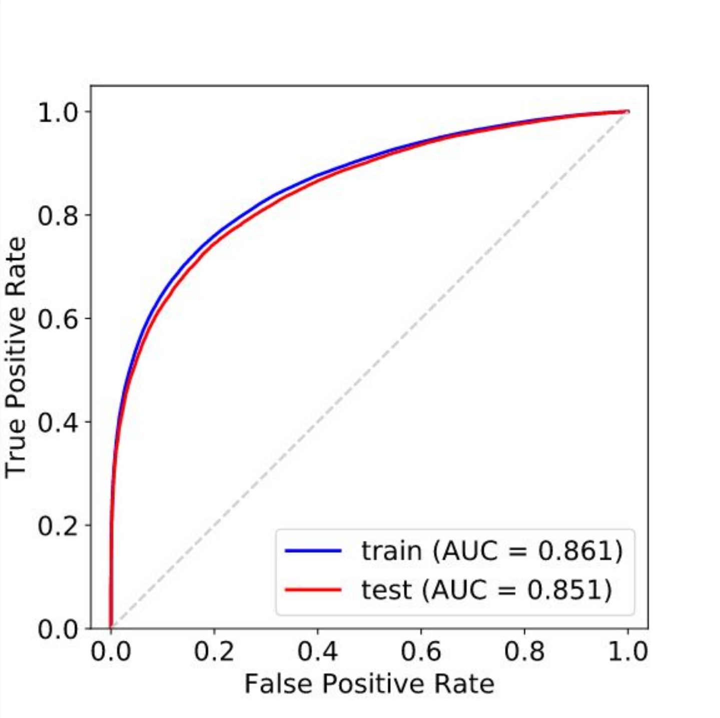
IPd Nsigma distribution



growing track pT



metrics



- for *bc-vs-udsg* we focus on far left part of the plot (False Positive Rate ~ 10^-3 -- 10^-1)
- ROC AUC is most closely correlated with signal eff. for bckg mistag rate = 1e-1

bc-vs-udsg model trained on SV features

"cuts1" = Chi2<10 ; "cuts2" = Chi2<10 and Dispersion < 0.03 and SigmaLxy<0.1

sorting feature + cuts	N_SV = 3				N_SV = 10			
	ROC AUC	signal eff. for bckg mistag rate:			ROC AUC	signal eff. for bckg mistag rate:		
		1e-3	1e-2	1e-1		1e-3	1e-2	1e-1
Chi2	0.7849	23.3	34.5	53.1	0.8092	**25.4**	37.5	57.7
Dispersion	0.7953	**25.2**	**36.3**	54.6	0.8143	**25.6**	**38.5**	58.7
Lxy	**0.8267**	19.9	35.4	60.3	**0.8336**	23.8	**38.8**	**62.0**
LxyNsigma	**0.8285**	19.0	35.3	**61.1**	**0.8331**	23.9	37.8	**61.7**
LxyNsigma+cuts1	0.8218	22.3	**36.1**	60.2	0.8262	24.2	37.0	60.4
LxyNsigma+cuts2	0.8200	21.4	**36.1**	59.6	0.8256	23.3	37.5	60.3

- for N_SV = 3:
  - sorting by Chi2/Dispersion yields highest efficiency for low bckg threshold choice (1e-3)
  - sorting by Lxy/LxyNsigma gives highest efficiency for less strict threshold (1e-1)
- the differences are reduced when we increase N_SV from 3 to 10
- adding quality cuts while sorting by LxyNsigma brings the results closer to sorting by Chi2/Dispersion and is more beneficial for lower lower N_SV as for larger it decreases efficiency

bc-vs-udsg model trained on tracks features

sorting + cuts	N_tracks = 3				N_tracks = 10			
	ROC AUC	signal eff. for bckg mistag rate:			ROC AUC	signal eff. for bckg mistag rate:		
		1e-3	1e-2	1e-1		1e-3	1e-2	1e-1
IPd/cov	0.7950	13.8	30.3	55.2	0.8335	20.1	38.4	62.1
IPdNsigma	0.7935	15.1	29.6	54.6	0.8311	22.7	40.0	62.3
IPdNsigmaAbs	**0.8294**	15.9	34.4	**60.4**	**0.8429**	22.9	41.0	**64.2**
Pt	0.8147	**20.1**	**36.2**	**60.0**	**0.8413**	**23.7**	**41.6**	**64.4**

- previous mistake: sorting by IPd/covIPd instead of IPd/sigma(IPd) gives a little bit worse results than correct one
- but it's always a better idea to sort by its absolute value
- sorting by pT gives similarly good results, especially for lower N_tracks
  worth trying in PbPb when contamination from soft particles with large DCA will be much greater

b-vs-c model trained on SV features

sorting + cuts	N_tracks = 3	N_tracks = 10
	ROC AUC	ROC AUC
Chi2	0.8096	0.8273
Dispersion	0.8164	0.8318
Lxy	0.8211	0.8351
LxyNsigma	0.8310	0.8393
LxyNsigma+cuts1	0.8247	0.8328
LxyNsigma+cuts2	0.8252	0.8330

sorting + cuts	N_tracks = 3	N_tracks = 10
	ROC AUC	ROC AUC
IPd/cov	0.8056	0.8441
IPdNsigma	0.8013	0.8402
IPdNsigmaAbs	0.8172	0.8416
Pt	0.8082	0.8416

- 0.01 higher ROC AUC ~ 2.5% higher eff. of *b-jets* for *c-jets* mistag. rate = 10%
- for SV: sorting by LxyNsigma gives best results
- for tracks: sorting by IPdNsigmaAbs gives highest results
- the differences are reduced when we increase number of objects included

Plans for next week (discussion)

- check IPd Nsigma distribution:
  - plot for data
  - plot IPd & sigma instead of ratios
  - check distributions e.g. phi-eta for tails, take into account asymmetry
- invent new cuts: hint = where the data diverges from MC
- investigate experiments in detail:
  - control plots created for all these experiments, stability, artifacts
  - features used by models
- compare data-MC discrepancy on column level w/ & w/o cuts
- create best possible model with:
  - covIPd and unshuffled-pt fixed
  - added jet shapes/substructure observables e.g. mean/median pT, momentum dispersion, angularity etc
  - check possibility to include SV representative to their distribution, like highest, average and lowest LxyNsigma instead of list of 3 or 10
- apply on data and show critical distributions: Lxy & IP (after loose cuts on SV quality)