Studies MC statistical error

Stefano Lacaprara

INFN Padova

NO meeting,
PD, 12 June 2014

Intro

- The template we are uing for the final fit on AZh analysis are affected by uncertaintiy due to the limited MC statistics;
- we thought that this uncertaintiy were taken into account automatically by the `combine` tool, but it is not the case;
- the only statistical uncertaintiy used is that on data.
- We must fin a way to include a systematic uncertaintiy due to the limited MC statistics.
- How?
  - One option is to let each bin of the template vary intependently and add a nuisance parameter for each of these bins.
  - The problem is that we now have 500 bins in mA, 6 in BDT, and 5 templates, summing up to 15000 bins, so this is not really a viable solution;
- But before: which is our statistical uncertainty on MC? Answer in the following slides.

Selections

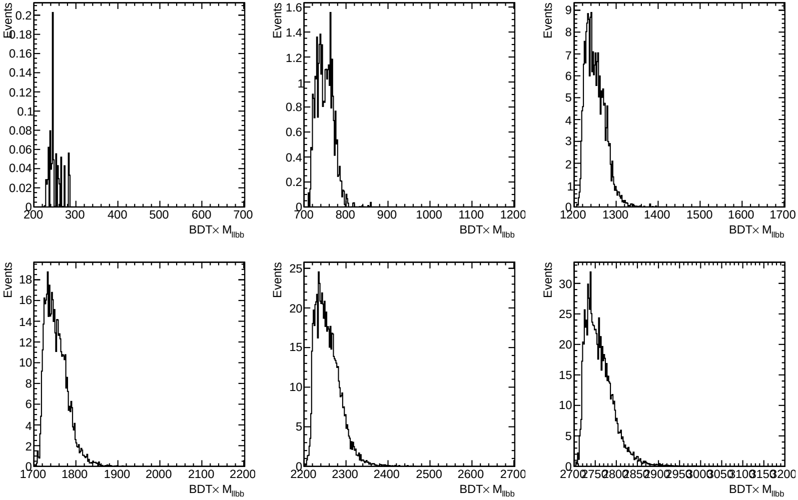
Following the previous study on significance vs BDT bins see slides, in the following I will use these selections:

- $0.6 < BDT < 1.0$;
- 6 bins in BDT;
- selection on M_{bb}: $90 < M_{bb} < 140\ GeV$.

And consider the three mass regions separately (Low, Intermediate, High)

Exp. backgound in each BDT bin

Low Mass

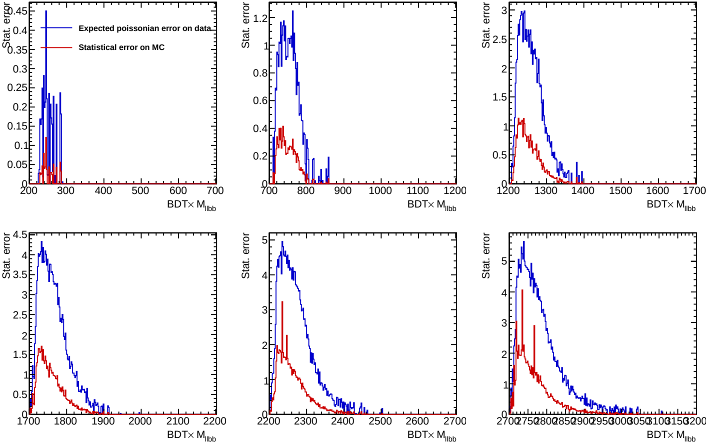


Statistical error in each BDT bin: Data vs MC

Low Mass

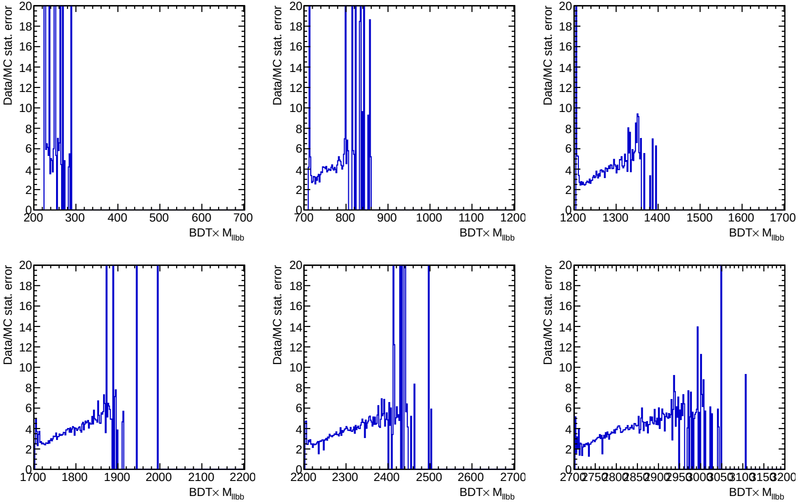
BLUE expected error on Data (poissonian)

RED: statistical error on MC



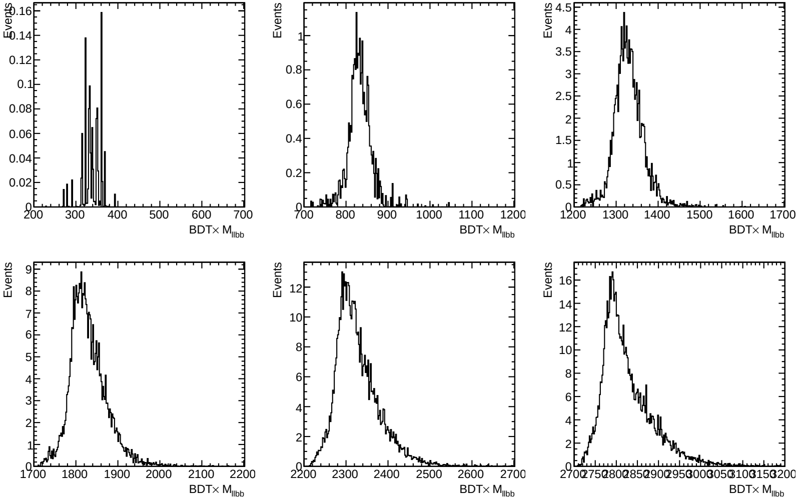
Statistical error in each BDT bin: Data/MC

Low Mass



Exp. backgound in each BDT bin

Int Mass

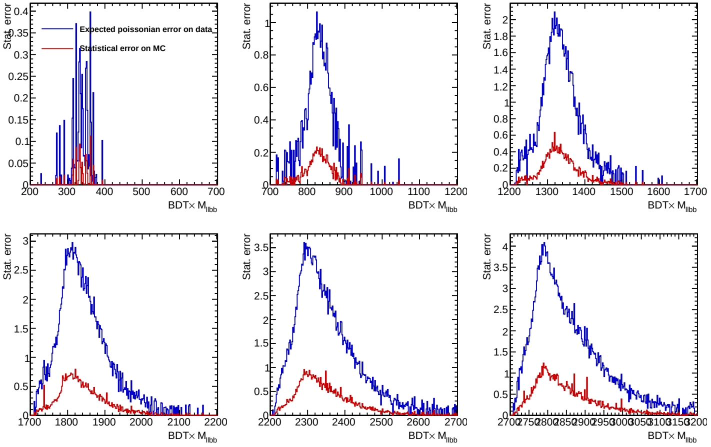


Statistical error in each BDT bin: Data vs MC

Int Mass

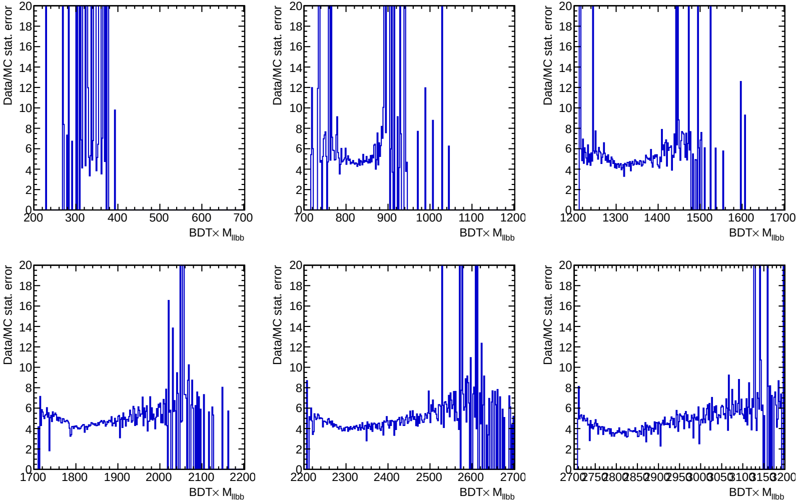
BLUE expected error on Data (poissonian)

RED: statistical error on MC

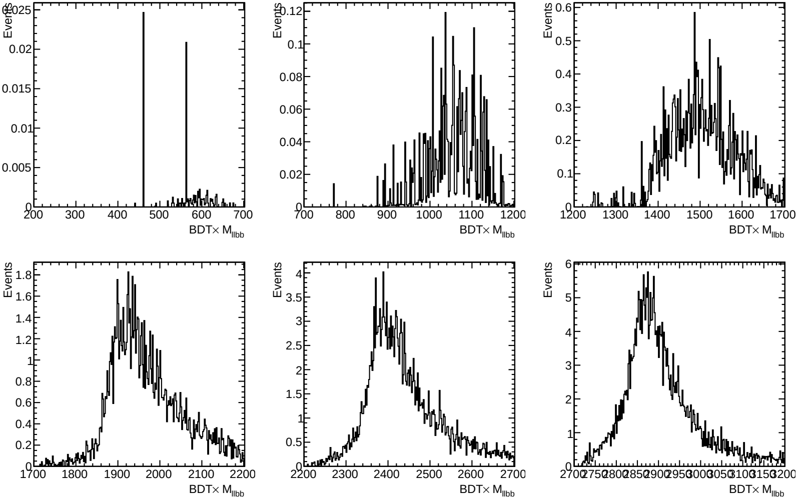


Statistical error in each BDT bin: Data/MC

Int Mass



Exp. backgound in each BDT bin
High Mass

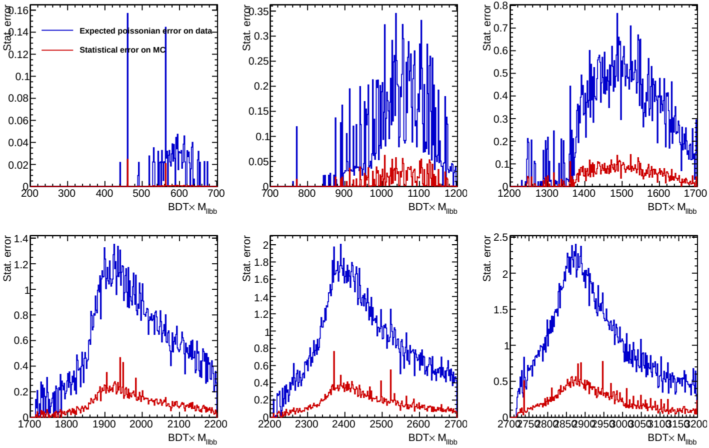


Statistical error in each BDT bin: Data vs MC

High Mass

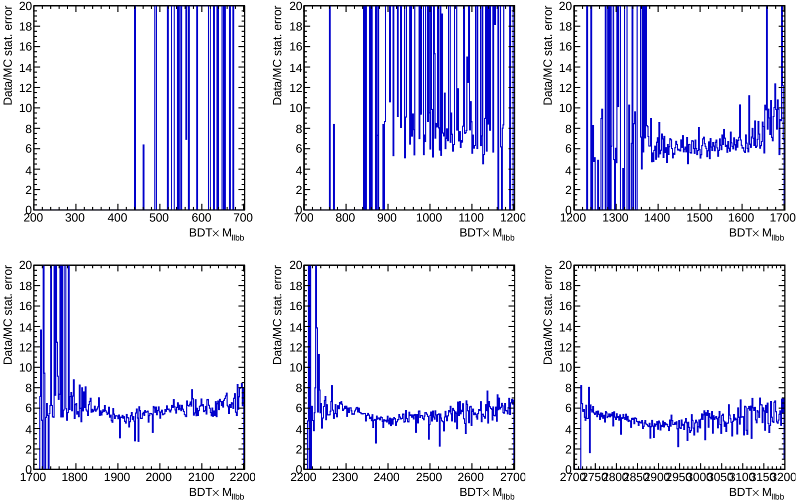
BLUE expected error on Data (poissonian)

RED: statistical error on MC



Statistical error in each BDT bin: Data/MC

High Mass



Estimating the MC stat error syst by toy-MC

- From the expected background estimate the effect of MC statistics by toy-MC;
- for each template of background, generate a new one, by varing each bin by a random number, according to a Gaussian distribuion, with mean value the expected one, and sigma the expected error;
- the error comes from the MC statistics and takes into account the weights distribution;
  - event-by-event the weight can vary
  - different sample, with different equivalent integrated luminosity;
  - weight due to PU reweight, Triggers, CSV reshaping, etc . . . ;
- build a new set of datacards with the new toy templates;
- copute the expected limit and compare with the original one;
- repeat N times

Expected limit with toys

m_A [GeV]	225	250	275	300	325	350	400	500	600
$BDT \in [0.6, 1]$, 6 bins, and $M_{bb} \notin [90, 140]$									
Original	17.812	18.188	16.562	13.312	10.156	8.125	4.8281	2.3984	1.5508
Toy playing with MC stat									
Toy 1	16.438	17.953	15.938	12.688	9.9062	7.9062	4.9219	2.3984	1.6016
Toy 2	17.812	18.062	16.562	13.312	10.219	8.0938	4.9219	2.5078	1.6016
Toy 3	17.188	17.812	15.812	13.062	10.031	7.7812	4.9844	2.4609	1.5508
Toy 4	17.562	18.062	16.062	13.188	9.9062	7.7188	4.8906	2.4609	1.6016
Toy 5	17.062	18.562	16.062	13.562	9.9688	8.2812	4.8281	2.3828	1.5703
Toy 6	17.688	18.062	15.812	12.812	9.8438	7.6562	5.0156	2.3359	1.5312
Toy 7	17.188	18.438	16.438	13.438	9.7812	8.2188	4.8594	2.4297	1.6172
Toy 8	17.188	18.188	16.062	13.438	9.9062	8.1562	4.7656	2.3672	1.5703
Toy 9	17.438	17.938	16.188	13.312	9.7812	7.9062	4.9219	2.4766	1.5625
Toy 10	16.812	17.906	16.375	13.188	9.7188	8.0938	5.0469	2.3984	1.543
Average and RMS of toys									
Average	17.238	18.098	16.131	13.200	9.906	7.981	4.916	2.422	1.575
RMS	0.413	0.238	0.258	0.279	0.144	0.218	0.086	0.054	0.029

Expected limit with correlated MC stat

m_A [GeV]	225	250	275	300	325	350	400	500	600
$BDT \in [0.6, 1]$, 6 bins, and $M_{bb} \notin [90, 140]$									
Original	17.8	18.2	16.6	13.3	10.2	8.12	4.83	2.4	1.55
with correlated MC stat systematics									
with MC stat	18.2	18.3	16.7	13.6	10.3	8.22	4.95	2.51	1.6

Expected limit with un-correlated MC stat

- Consider only bins with S/B above a given threshold;
- Alternative: consider the N most significant bins (S/B);
- For each backgournd source:
  - if template for that background not empty:
  - move that bin Up and Down by the MC statistical error;
  - leave all other bins unchanged;
- Add a `shape` systematic for each of these bins;
- Try different thresholds or N to see the effect.

Expected limit with un-correlated MC stat

m_A [GeV]		225	250	275	300	325	350	400	500	600
$BDT \in [0.6, 1]$, 6 bins, and $M_{bb} \notin [90, 140]$										
Original		17.812	18.188	16.562	13.312	10.156	8.125	4.9219	2.4922	1.5859
$S/B >$	N_{bins}	with MC stat systematics								
0.1	298	17.812	18.188	16.562	13.438	10.156	8.1562	4.9219	0.0005	0.0005
0.01	328	18.812	18.906	17.062	13.812	10.594	8.4688	5.0469	0.0005	0.0005
0.001	1465	20.188	19.812	17.812	14.062	10.719	8.5312	5.0469	0.0005	0.0005
highest S/B	N_{bins}	with MC stat systematics								
	10	18.688	18.312	16.688	13.438	10.156	8.125	4.9219	2.4922	1.5859
	20	19.688	19.25	17.062	13.688	10.281	8.1562	4.9844	2.5078	1.5938
	30	20.062	19.562	17.438	13.812	10.406	8.2812	5.0312	2.5234	1.6016
	50	20.188	19.812	17.688	14.062	10.656	8.4062	5.0469	2.5234	1.6016
	80	20.188	19.812	17.812	14.062	10.719	8.4688	5.0469	2.5234	1.6016

Expected limit with un-correlated MC stat

Full shape syst (new)

m_A [GeV]	225	250	275	300	325	350	400	500	600
Alberto's BDT and Mbb									
Original	16.438	17.438	15.438	12.812	9.9062	7.8438	4.8281	2.4766	1.7109
N_{bins}	with MC-stat systematics								
10	17.438	17.906	15.688	12.938	10	7.8438	4.8906	2.4844	1.7109
20	18.062	18.438	16.062	13.188	10.156	8.0312	4.9219	2.5078	1.7109
30	18.312	18.688	16.312	13.312	10.281	8.0938	4.9531	2.5078	1.7266
50	18.375	18.812	16.438	13.562	10.344	8.1562	4.9844	2.5078	1.7266

- Using full Alberto's shape systematics;
- Using all BDT bins (from 0 to 1);

Effect of using less bins

m_A $[GeV]$	225	250	275	300	325	350	400	500	600
$BDT \in [0.6, 1]$, 6 bins, and $M_{bb} \notin [90, 140]$									
Original	17.8	18.2	16.6	13.3	10.2	8.12	4.83	2.4	1.55
Rebinning High Mass only									
rebin 2	17.8	18.2	16.6	13.3	10.2	8.12	4.92	2.49	1.59
rebin 4	17.8	18.2	16.6	13.3	10.2	8.12	4.95	2.54	1.62
rebin 5	17.8	18.2	16.6	13.3	10.2	8.12	5.08	2.55	1.63
rebin 10	17.8	18.2	16.6	13.3	10.2	8.12	5.61	2.65	1.66
rebin 20	17.8	18.2	16.6	13.3	10.2	8.12	6.2	2.87	1.77

Conclusion

- In general, the statistical error on MC template is about 5 times smaller than the corresponding poissonian error on MC
- The first BDT bin (I guess starting from the end backward?) has very little statistics, leading to very large fluctuation on tempaltes
  - We should definitively consider to merge the last two bins
  - or have a variable size binning on BDT
  - Reducing the number of bins for high mass makes very little difference.
- Toy show a small effect due to MC statistic
- Considering only the N most significant bin (S/B), the limits varies, with a saturation when $\approx 30 - 50$ bins are considered