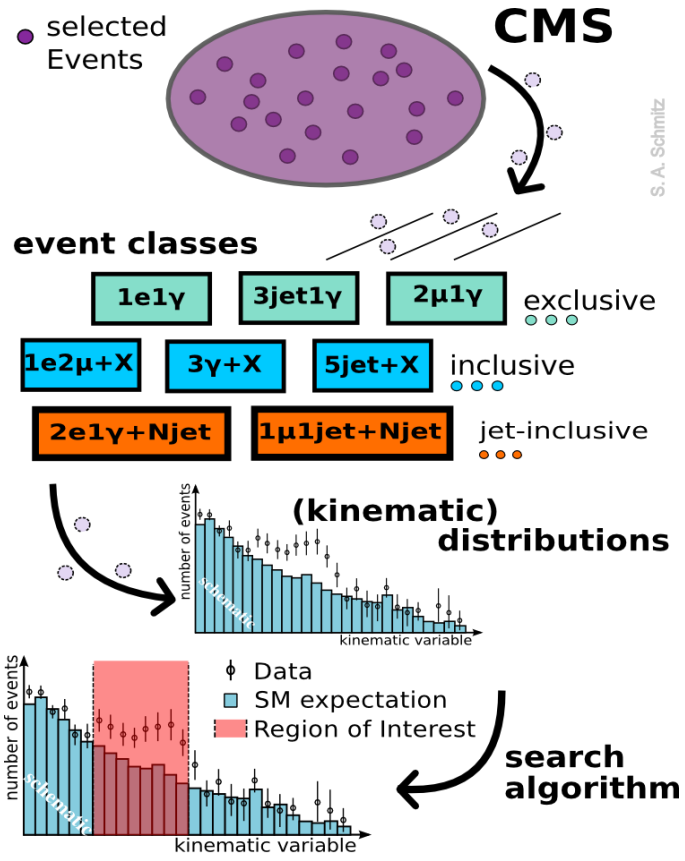
MUSIC COMPUTING PERFORMANCE IMPROVEMENTS: "QUICKSCAN"

Jonas Lieb, 02.06.2015

MUSIC (MODEL UNSPECIFIC SEARCH IN CMS)

- Sort events into **event classes** by their physics object content $(\mu, e, \gamma, \text{jets}, \text{MET})$
- Three distributions of interest: $\sum |\overrightarrow{p_T}|, M_{\text{inv}}, \text{MET}$
- Find most significant region (RoI) in each distribution
- Determine look-elsewhere corrected **p-value** ($\tilde{p}$) for each distribution through pseudo-experiments
- Compare distribution of $\tilde{p}$ from data with MC

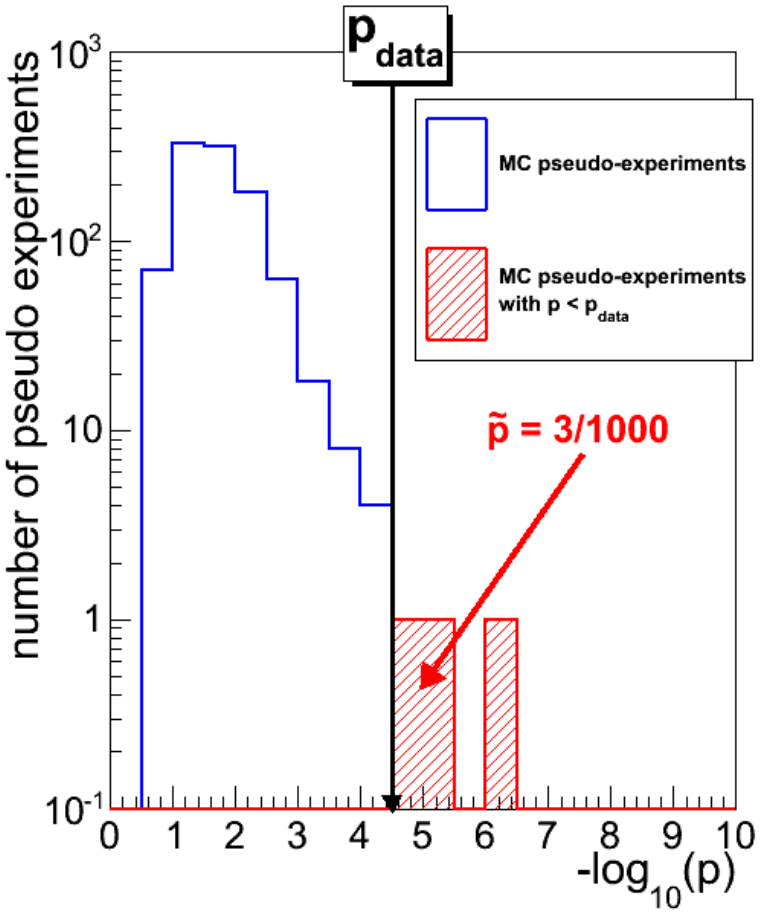


SCANNING

- Construct **connected bin regions** from histogram
- Calculate **p-value** for each region:
  - $p_{\text{data}} = \begin{cases} \sum_{N=N_{\text{data}}}^{\infty} C \cdot \int_0^{\infty} \mathrm{d}\theta \, \exp\left(-\frac{(\theta - N_{SM})^2}{2\,\sigma_{SM}^2}\right) \frac{e^{-\theta}\theta^N}{N!}, & \text{if } N_{\text{data}} \geq N_{\text{SM}} \\ \sum_{N=0}^{N_{\text{data}}} C \cdot \int_0^{\infty} \mathrm{d}\theta \, \exp\left(-\frac{(\theta - N_{SM})^2}{2\,\sigma_{SM}^2}\right) \frac{e^{-\theta}\theta^N}{N!}, & \text{if } N_{\text{data}} < N_{\text{SM}} \end{cases}$
- Find **most significant region (smallest p-value)** for each histogram

CALCULATION OF $\tilde{p}$

- Needed to account for "**look-elsewhere-effect**"
- Repeat scanning with **pseudo-experiments,** each mean is shifted within its Standard Model uncertainty



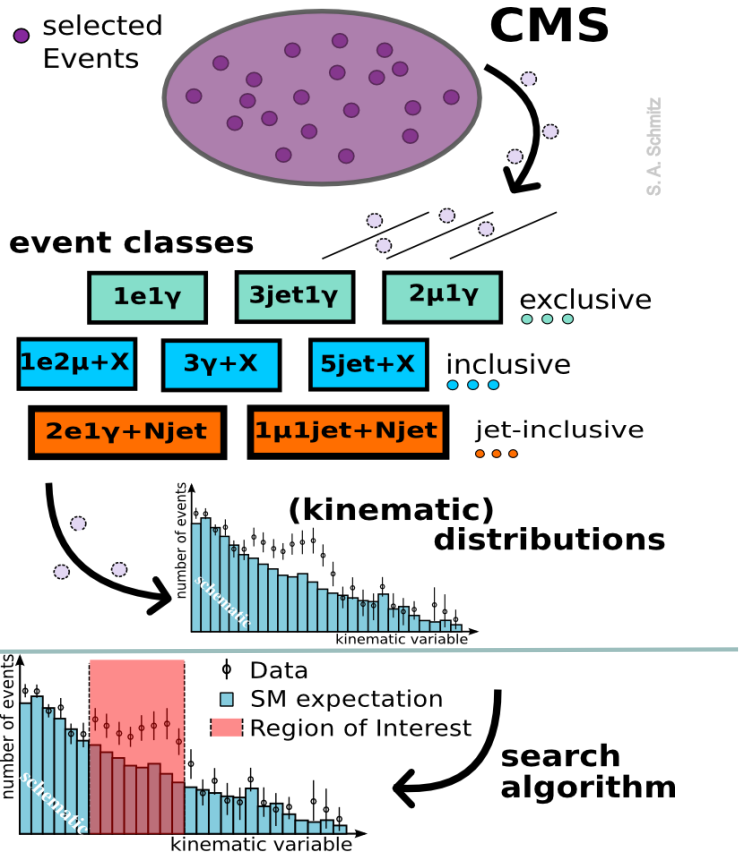
$$\tilde{p} = \frac{\text{number of pseudo experiments with } p_{pseudo} < p_{data}}{\text{total number of pseudo experiments}}$$

QUICKSCAN

- Problem: the p-value is evaluated many times, its calculation is time consuming
- Mitigation: **preselect interesting regions** using a less computation intense algorithm
- Select a certain number of candidate regions, with the maximum

$$\chi = \frac{|N_{obs} - N_{MC}|}{\sigma_{MC}}$$

- This estimator does not consider effects depending on the absolute number of events → "vertical" binning by magnitude
- To select the most significant region, calculate the p-value integral only for the Quickscan candidates
- Two parameters:
  - **number of candidates per vertical bin**
  - **magnitude bin size**



MAGNITUDE BINNING
EXAMPLE: 5 CANDIDATES, BASE 10

MC-Events	0.1-1	1-10	10-100
Candidates	1. Region(122,125) **0.5 MC events** $\chi = 20.9$ 2. Region(238,313) **0.9 MC events** $\chi = 17.1$ 3. Region(192,206) **0.3 MC events** $\chi = 12.3$	1. Region(73,78) **3 MC events** $\chi = 11.3$ 2. Region(82,93) **9 MC events** $\chi = 8.7$ 3. Region(35,43) **7 MC events** $\chi = 8.6$	1. Region(2,5) **89 MC events** $\chi = 25.9$ 2. Region(8,13) **21 MC events** $\chi = 0.5$ 3. Region(19,21) **23 MC events** $\chi = 0.2$

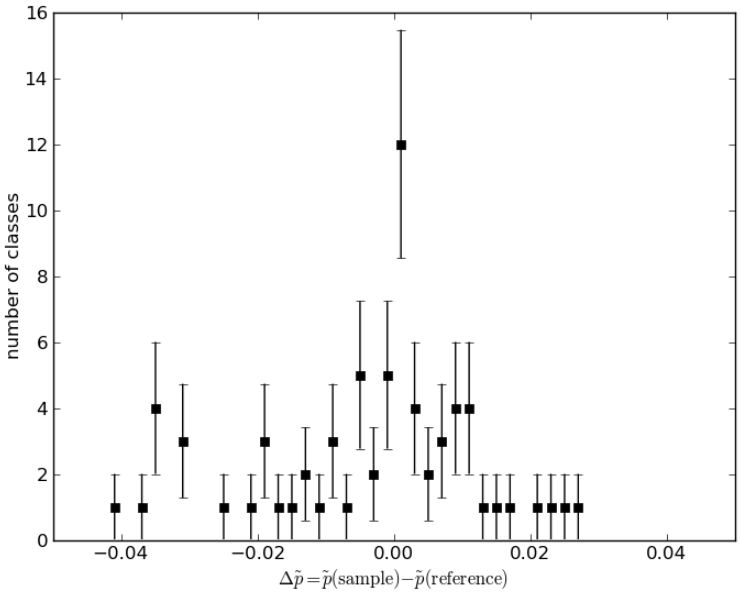
Example numbers

Magnitude bins 0.1-1, 1-10 and 10-100 are shown.

For each magnitude bin, the real p-value is calculated.

PARAMETER OPTIMIZATION

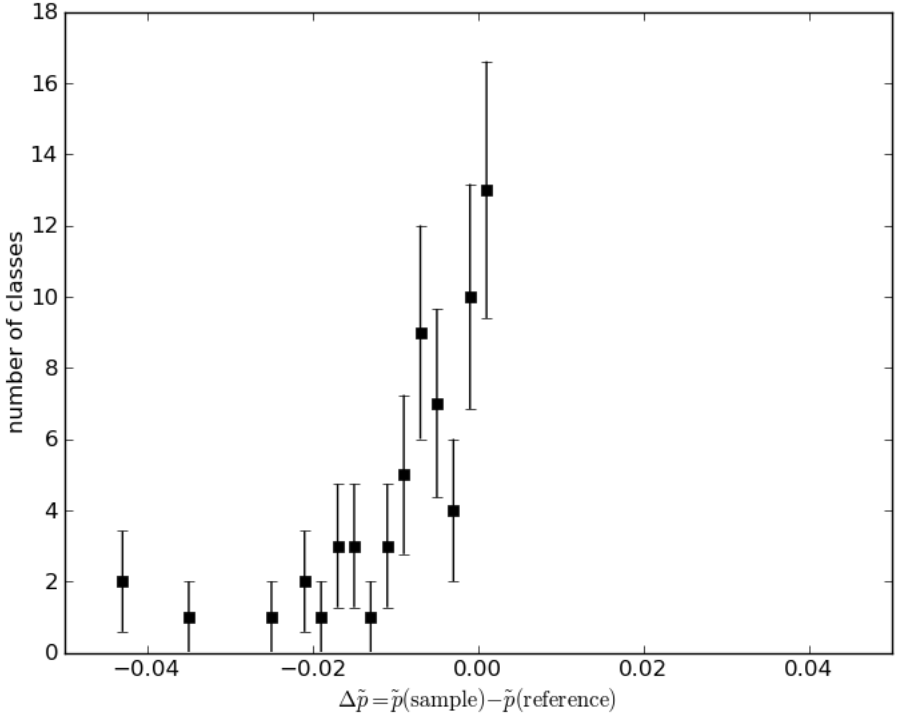
- Optimization of the parameters is performed by measuring their effect on two metrics:
  - **Runtime / Speed-up** $= \frac{T_{classic}}{T_{quickscan}}$
  - **Deviation of $\tilde{p}$:**
    $\Delta\tilde{p} = \tilde{p}(\text{quickscan}) - \tilde{p}(\text{classical}) \ (\leq 0)$
- Working on a **subset**: 2012 data, exclusive classes only, max. 2 jets, dicing exactly 1000 rounds
- **Status quo:**
  - Runtime ~ 3h 30min
  - Random $\Delta\tilde{p}$ spread through dicing: 50% of data has $|\Delta\tilde{p}| \leq 0.008$



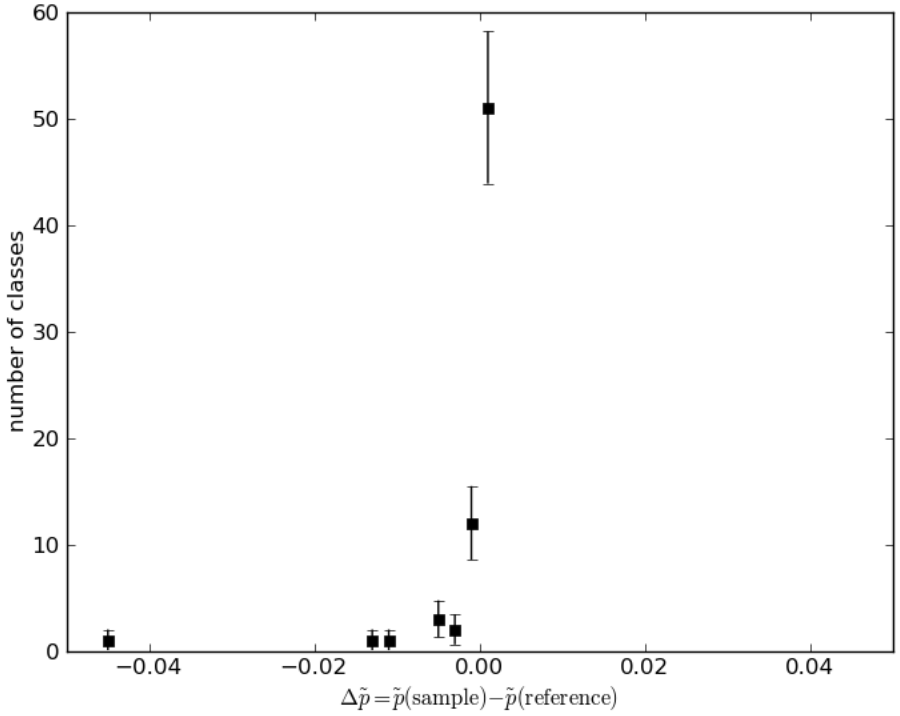
w/o Quickscan vs. w/o Quickscan

SELECTED RESULTS

Worst case: 1 candidate, magnitude bin base 20.
Runtime: 25 minutes
50% has $|\Delta\tilde{p}| \leq 0.007$

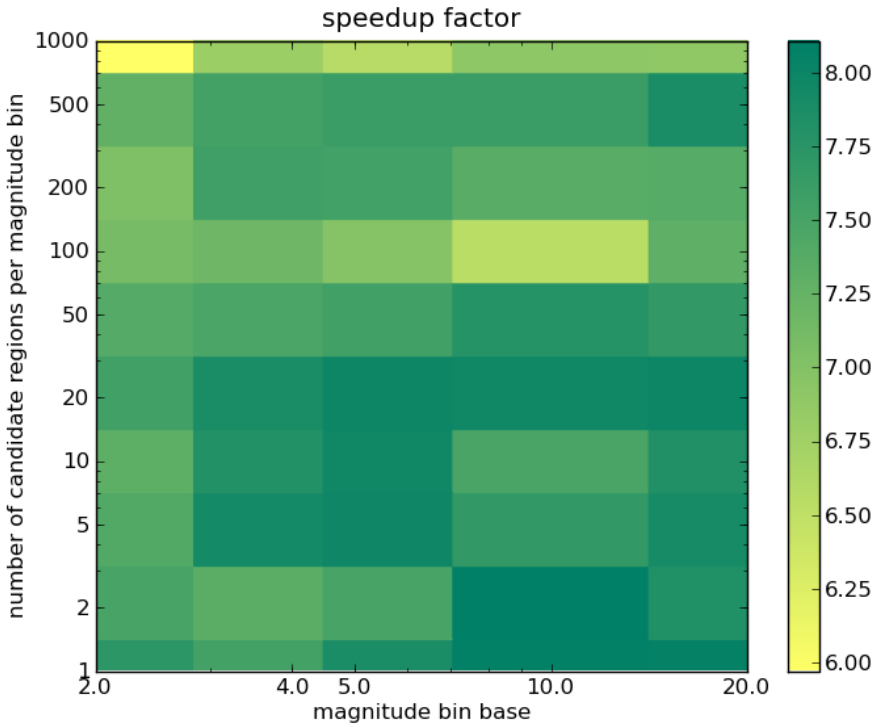


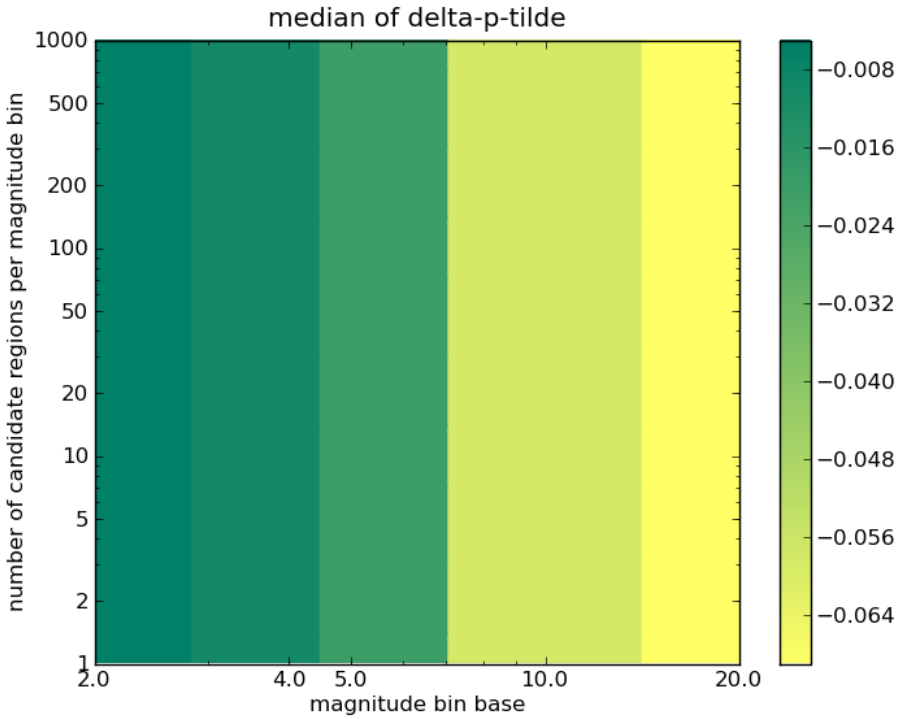
Best case: 1000 candidates, magnitude bin base 2.
Runtime: 34 minutes
More than 50% has $|\Delta\tilde{p}| = 0$



Quickscan vs. w/o Quickscan

RESULTS FOR ALL PARAMETERS





RESULTS

- Quickscan seems to work
- Best physics results can be achieved by choosing narrow vertical bins
- Speed-up currently between 6 and 8 times, giving the same results
- Number of candidate regions per magnitude bin does not influence the physics result ($\Delta\tilde{p}$)

OUTLOOK

- Goal: **further performance improvements,** quantify runtime
- Take a closer look at MUSiC's **parallelization** implementation
- Determine optimal parameters and run on complete data

BACKUP: PARALLELIZATION

- Multiprocessing: Server (MISMaster), Clients (1 dicer, multiple scanners)
- Communication between processes via pipes using a custom text-based protocol
- Tight communication needed for correlated dicing