Data driven QCD background estimates in HH/HS searches at CMS

Matej Roguljić, Ruđer Bošković Institute,
on behalf of the CMS Collaboration

Higgs Pairs Workshop 2022
01.06.2022.

Supported by
HrZZ
Croatian Science Foundation

Motivation

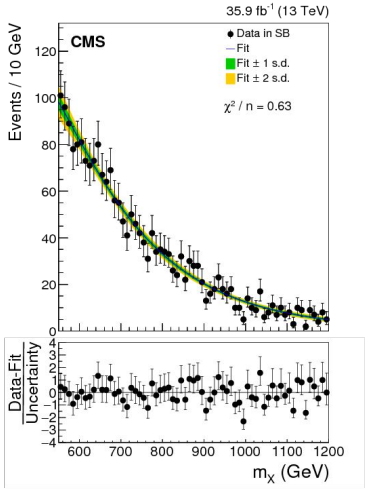
- Major background in HH searches using hadronic final states often are the QCD multijet events
- Simulation of such events is not precise enough
  - Multijet events modelled at LO
  - Possible lack of statistics

Data-driven estimation methods

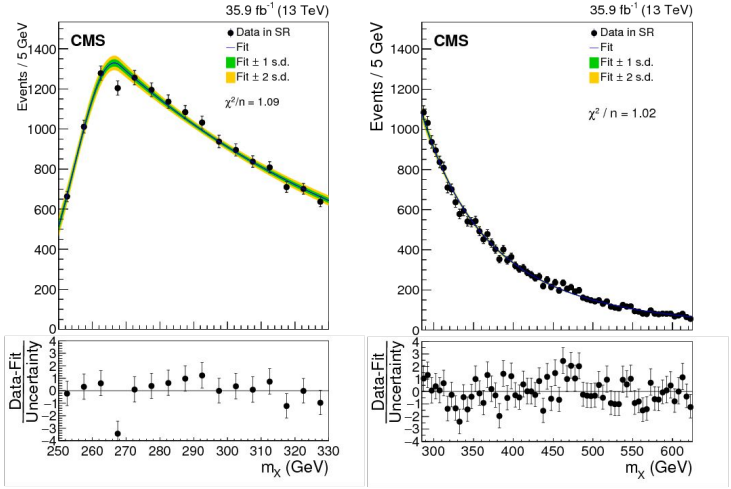
- Used in many of the CMS HH searches presented at the workshop
- **Presenting an overview of the data-driven estimation methods of the multijet background with examples**

Parametric functions

- Directly model the shape with a function
- Can be used when searching for a resonance on a smoothly falling background
  - Turn-on effects may be problematic

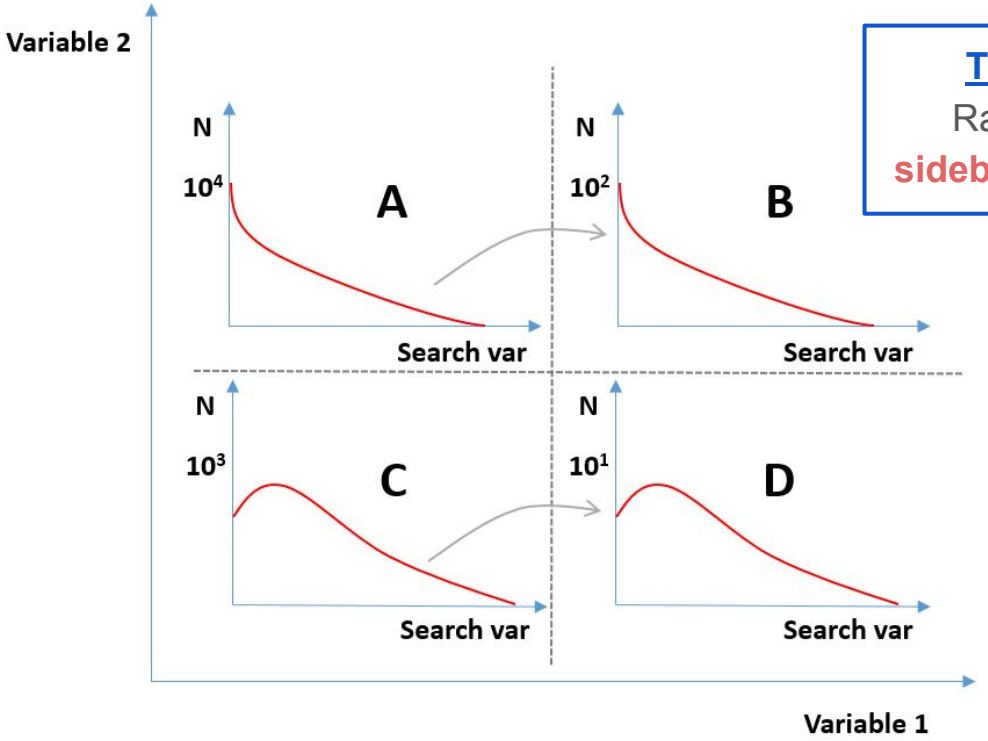


- Resonant HH→4b search, 2016 data (HIG-17-009)
  - Functional forms chosen in studies performed before unblinding, using **control regions**
    - Signal-free regions with kinematic properties similar to events in signal regions



ABCD method

- Define four regions using selection on two independent variables
  - One of them (D) is the signal region
- Estimate shape in D by application of a **transfer factor** to C



Transfer factor

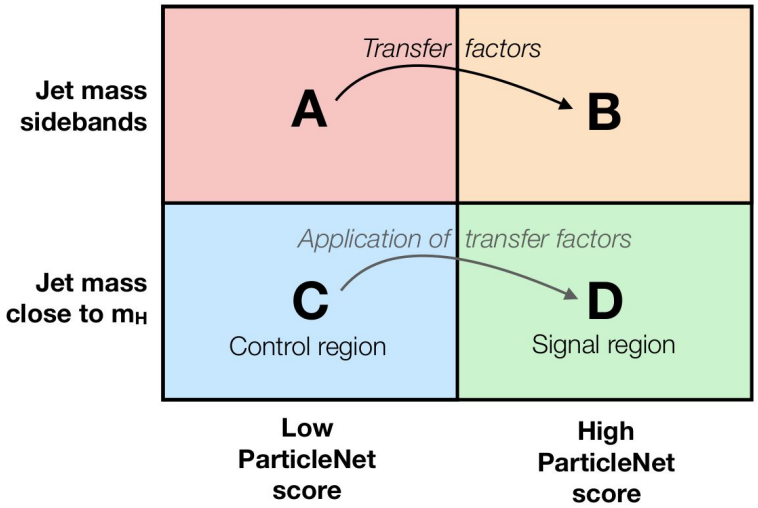
Ratio of events in **sideband regions** (A,B)

Method requirements

1. Similar QCD shapes for C and D in the search variable
2. Transfer factors A→B and C→D are the same
3. A, B and C enriched in QCD events → reliable QCD estimate from data

ABCD method example

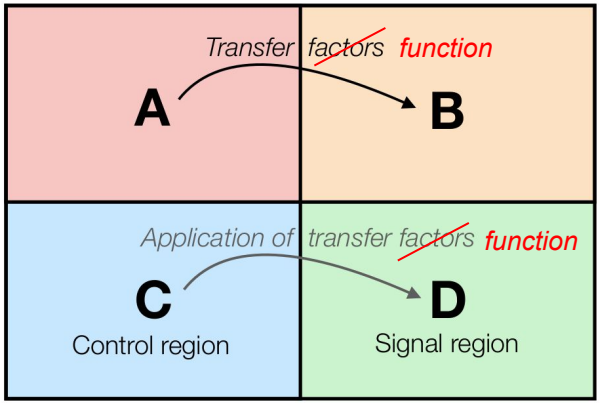
- Non-resonant boosted HH→4b search, (B2G-21-001)
  - Two high-p_T, large-area jets with $m \cong m_H$
    - Search variable is the invariant mass of the two jets
  - High ParticleNet tagger scores indicate that jets come from H→bb
  - ParticleNet tagger ensures that the A→B and C→D transfer factors are the same
    - Tagging performance does not depend on jet mass



Transfer functions ($R_{P/F}$)

- Extension of ABCD when QCD shapes (in search variable) in C and D differ
  - Transfer factors → Transfer functions (also called pass-to-fail ratios, $R_{P/F}$)
- $R_{P/F}$ can be measured between A and B and applied to C
  - Dependence may be measured on the search variable, but also on other variable(s)

- Resonant boosted $X \rightarrow \phi\phi \rightarrow 4b$ search, (B2G-20-003)
  - $R_{P/F}$ modelled as a product of two functions
    - Subleading jet's p_T
    - Subleading jet's mass
  - Measured in **sideband regions**, prior to the fit in the signal regions



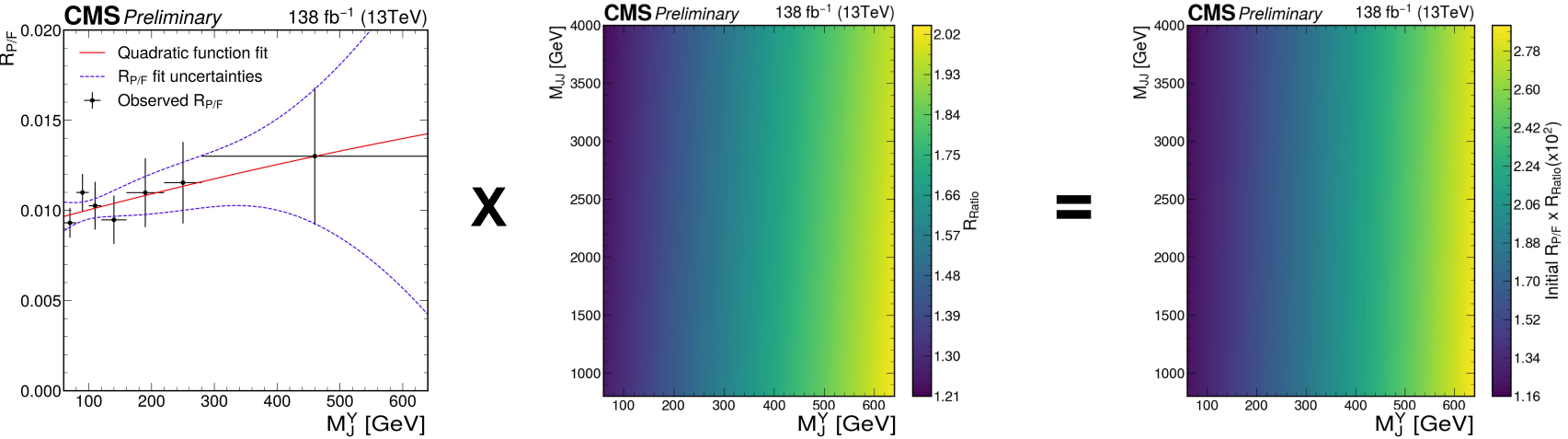
C(A) and D(B) also called "fail" and "pass" regions, usually referring to tagging category

R_{Ratio} method

$$R_{Ratio} = \frac{R_{P/F}^{true}}{R_{P/F}^{init}}$$

Determined during fitting

- Allows additional morphing of the $R_{P/F}$
- Resonant boosted X →HH→4b search, (B2G-20-004) measures the $R_{P/F}^{init}$ for the signal regions using **simulation**
  - R_{Ratio} accounts for the differences between simulation and data
- Resonant boosted X →YH→4b search, (B2G-21-003) measures the $R_{P/F}^{init}$ using **sideband regions** in data
  - R_{Ratio} accounts for the differences between sideband and signal regions



Usage of the methods

- QCD estimation methods in searches shown at HH workshop
- Varying level of systematic unc. on bkg. ≈1-30%
  - Large unc. usually from low number of events in “C” region
  - Usually not the dominating unc.
- Region naming in the methods **may change** from search to search, but only two important things to note:
  1. In which region is the transfer factor (function) measured?
  2. To which region is it applied?

Search	QCD estimation method	Covered by
X→HH→4b resolved HIG-17-009	Parametric functions	Santeri and Marco
X→HH→4b boosted HIG-20-004	R_{Ratio}	Santeri and Marco
X→YH→4b boosted B2G-21-003	R_{Ratio}	Santeri and Marco
X→YH→bbττ HIG-20-014	Transfer functions	Ralf and Tatjana
HH→4b resolved HIG-20-005	Transfer functions	Daniel and Rafael
HH→4b boosted B2G-22-003	ABCD	Daniel and Rafael
HH→bbττ HIG-20-010	ABCD	Francesco and Yanlin

Summary

- We can get reliable QCD background estimate using data-driven methods
  - Shown methods can be used to predict other types of background as well

- Parametric functions possible in some cases
- Transfer function methods more often used
  - Rely on inverting selection and obtaining shape in QCD dominated regions in data
  - Methods evolve based on the use case