



An alternative approach for Standard Model background estimation in the dijet resonance search with CMS at 13 TeV.

Dimitris Karasavvas,

Niki Saoulidou,

Eirini Tziaferi

University of Athens, Greece



- Analysis strategy
- Non-parametric fit in the Control Region.
- Prediction in the Signal Region.
- Proposal & Summary

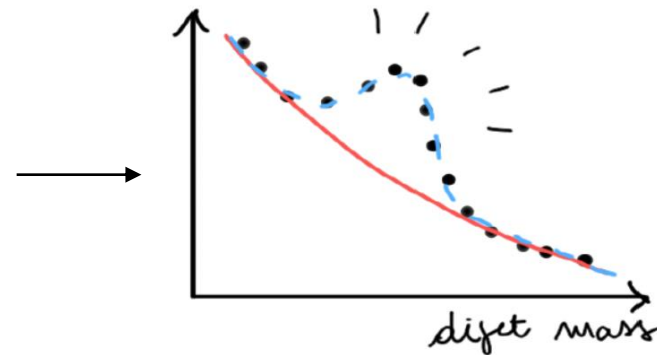
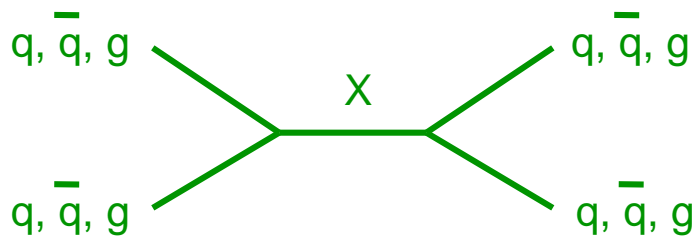


Motivation



Search:

Dijet Resonances



Powerful: LHC in run 2 is a dijet resonance factory at a new energy scale

Broad: search for many sources of new physics in a single simple search

String resonances from string theory

Excited quarks from theories of quark compositeness

W', Z' and **scalar diquarks** from grand unified theory

Gravitons from the Randall-Sundrum model of extra dimensions

Axiguons, Colorons and **Color Octet Scalars** from other new models

Model Independent: Search results are applicable to any model of narrow qq , qg , or gg resonances.



Analysis Strategy



Main Analysis Strategy:

- Fit the data spectrum in the signal region with a smooth empirical parametrization to describe QCD background and use MC templates for the signal models.

Complementary Analysis Strategy under development :

- Define a control region, New Physics depleted, and with similar kinematical characteristics as the signal region, in order to perform additional quality checks **[part of the analysis already]**, and predict the QCD background in the signal region as follows, using the simulation to estimate the following ratio [R_{ext}] :

$$R_{\text{ext}} = M_{jj}^{\text{SR}} / M_{jj}^{\text{CR}}$$

$$\text{Prediction}^{\text{SR}} = R_{\text{ext}} \times M_{jj}^{\text{CR}}$$



Main Analysis Strategy advantages:

- Used for many years, well tested and documented, does not rely on simulation.
- Fit the data spectrum in the signal region with a smooth empirical parametrization to describe QCD background and use MC templates for the signal models

Complementary Analysis Strategy advantages :

- It does not assume a model for the shape of the QCD background in the signal region since it derives it from data in the control region.
- It is potentially less biased with respect to signal-template fitting.

Kernel Estimation : Intro

- One of the most common practical duties of a particle physicist, is to analyze various distributions from a set of data $\{t_i\}$.
- From this analysis one would wish to estimate the probability density function (pdf) from which the data were drawn.

$$P(a \leq X \leq b) = \int_a^b f(x) dx .$$

- In the following slides, **we will discuss a certain non-parametric estimation technique for the parent pdf, called **kernel estimation**.**

Non-parametric techniques

The goal of non-parametric methods is to remove the model dependence of our estimation, in a way that it arises purely from the data. They are defined as **asymptotically local** meaning, *the influence of a data point t_i at a point x of the density should vanish asymptotically.*

Some non-parametric estimates are the following:

- The histogram (sensitive to the binlength and origin choices)
- The Averaged Shifted Histogram (ASH)

(origin-independent, reduced binning effects of traditional histograms)

- Kernel Estimation (origin, bin independent)

Fixed Kernel Estimation (1)



Let us first consider the univariate case, the kernel estimation in its general form is given by:

$$\hat{f}_0(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-t_i}{h}\right), \quad (1)$$

where, $\mathbf{K}(\mathbf{x})$ is some known distribution (Gaussian, Quartic, etc), t_i represent the data and \mathbf{h} is the smoothing parameter.

We will consider ,for simplicity, the case where $K(x)$ is the normal distribution since one can prove that other choices give differences less than 0,5% in the result.

$$\mathbf{G}_{\mu,\sigma}(x) = \frac{1}{\sigma} \mathbf{K}\left(\frac{x-\mu}{\sigma}\right)$$



Fixed Kernel Estimation (2)



Assuming that $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, one can see the kernel estimation as follows:

We assign a Gaussian distribution in each data point t_i with a $\sigma=h$, and we take the mean value of the n distributions as our estimation.

The role of h is to set a scale for our kernels (strength of the locality).

In the f_0 estimation the h is constant, and is chosen as the one that minimizes the **mean integrated squared root error** of the estimation when $n \rightarrow \infty$, in the case of normal distributed data:

$$h^* = \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-1/5}, \text{ where } \sigma \text{ the standard deviation of the data.}$$

Adaptive Kernel Estimation

One might object to the choice of h^* , since σ is a global quantity and non-parametric estimates are supposed to be local. To avoid this, we restrain ourselves from defining a constant h , and in its place we use:

$h_i = h / \sqrt{f(t_i)}$, where $f(t)$ is the parent distribution of our data. This choice actually reflects the fact that in high density regions we can use narrow kernels, but in low density ones we need wide kernels to smooth out statistical fluctuations.

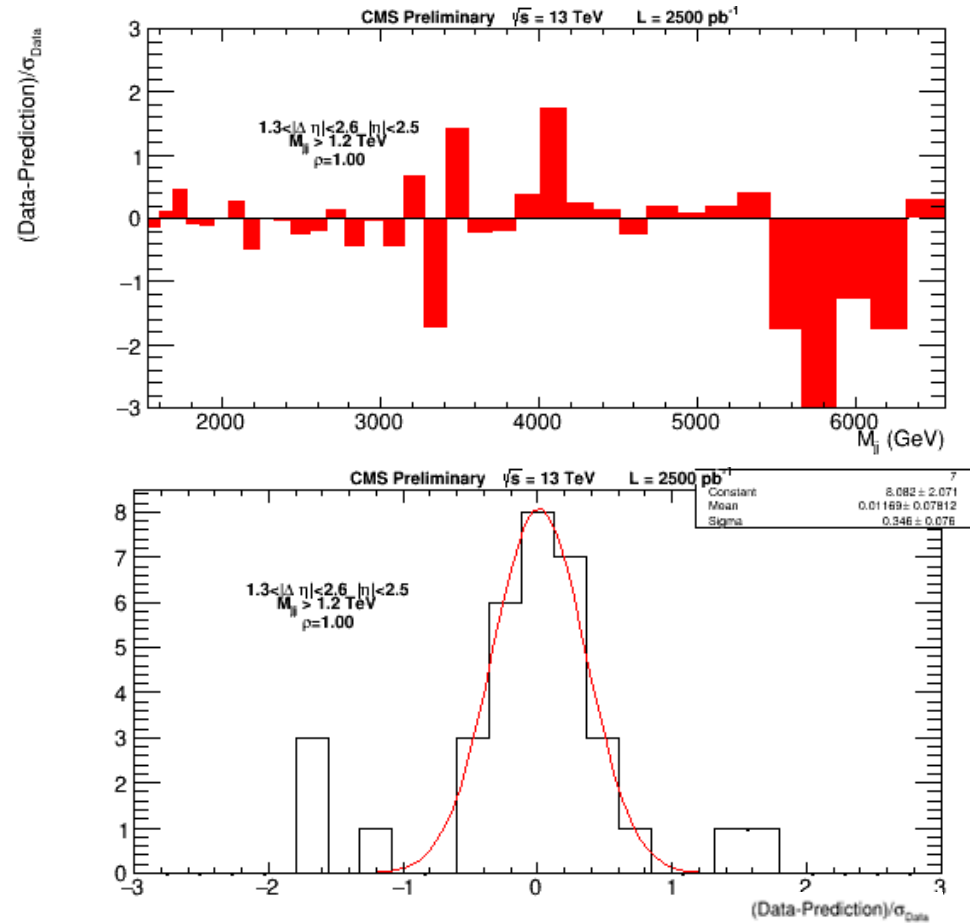
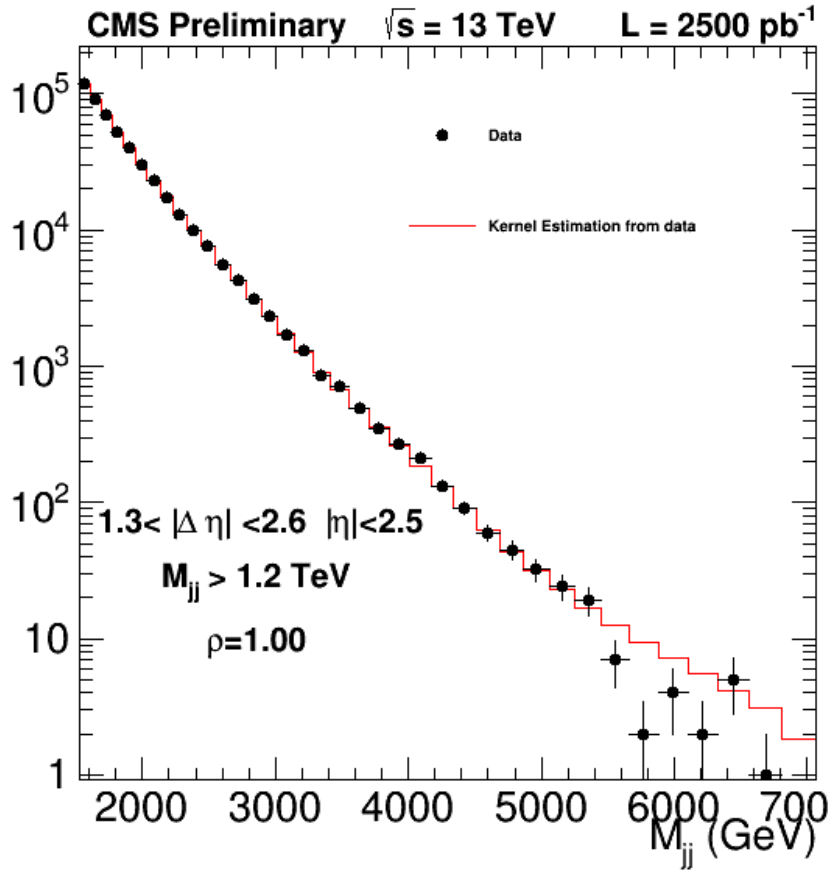
Technically we have 2 problems:

- we don't know the parent distribution of our data
- Optimal choice of h_i is still to be specified

To solve this issue we calculate the h_i 's using the fixed kernel estimation f_0 . With this assumption, we get the optimal choice as:

$$h_i^* = \rho \left(\frac{4}{3}\right)^{1/5} \sqrt{\frac{\sigma}{f_0(t_i)}} n^{-1/5}, \text{ where } \rho = \sqrt{\frac{\sigma_{local}}{\sigma}} \text{ (typical value for } \rho \text{ is unity)}$$

Control Region : non-parametric fit



- Performed a study for various ρ . Showing the $\rho = 1$ case here, and the ρ scans in the backup.
- Smoothing seems to be quite successful judging from the pulls.

Prediction in the Signal Region

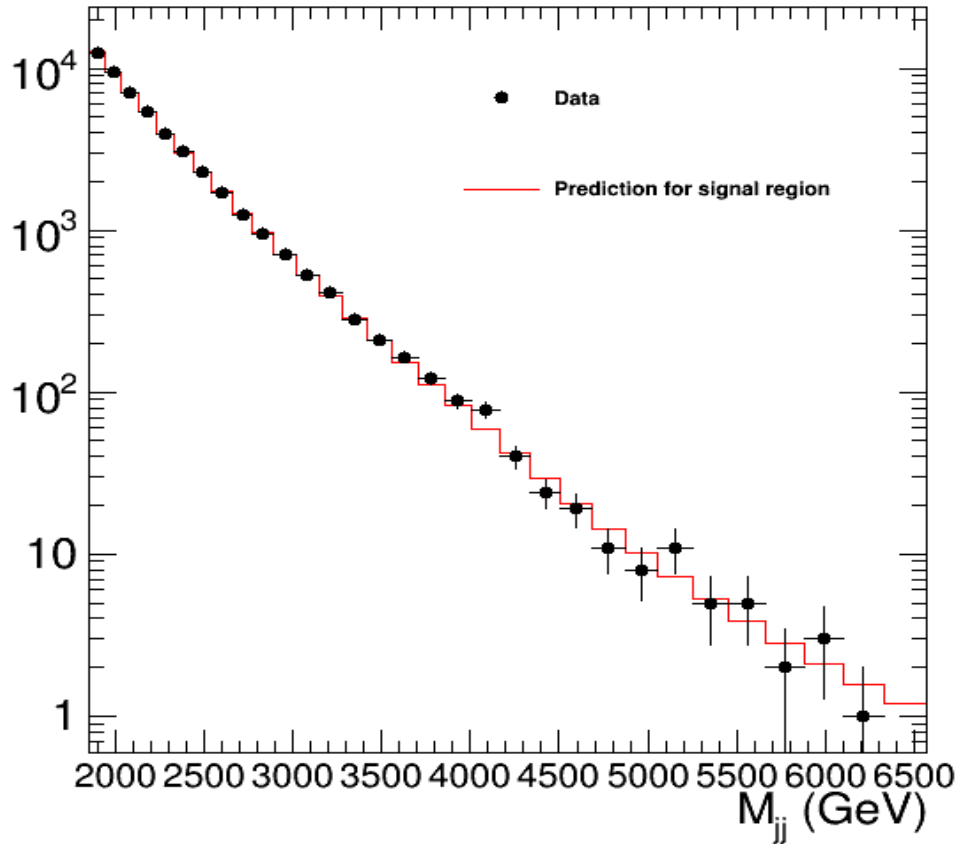
- Statistics in signal region is too low to trust Kernel estimation directly.
 - We take the background estimation using the Ratio Method:
- Use MC simulation, or data (fitted/smoothed), to calculate the following ratio:

$$R_{\text{ext}} = M_{jj}^{\text{SR}} / M_{jj}^{\text{CR}}$$

- The Prediction for the Signal region is calculated through the Kernel estimation on the Control Region:

$$\text{Prediction}^{\text{SR}} = R_{\text{ext}} \times M_{jj}^{\text{CR}}$$

Signal region : Prediction using R from smoothed data ratio



- Data and prediction agree nicely for all M_{jj} values.



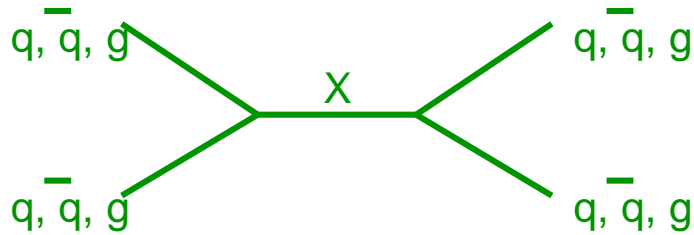
- We have performed non-parametric fit (smoothing) of the data in the control region with encouraging results.
- We have produced the prediction in the signal region using the data in the control region:
 - Using the ratio from simulation produces a structure at $M_{jj} < 2.8$ TeV
- Results so far indicate that this method can work well and with low systematic uncertainties for both narrow and wide resonances.
- We performed the analysis, fitting the data ratio itself.

Backup

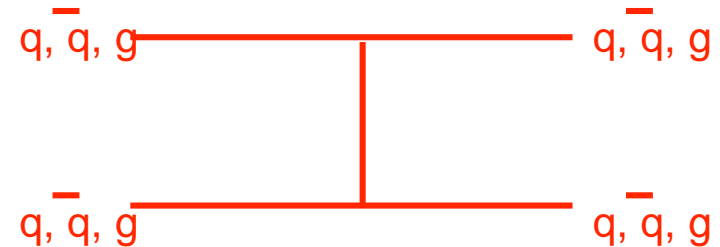


Analysis overview

Resonance Signal



QCD Background



- **Reconstructed objects**

- Particle Flow jets
- No b-jet, V-jet, or H-jet tagging needed for inclusive search

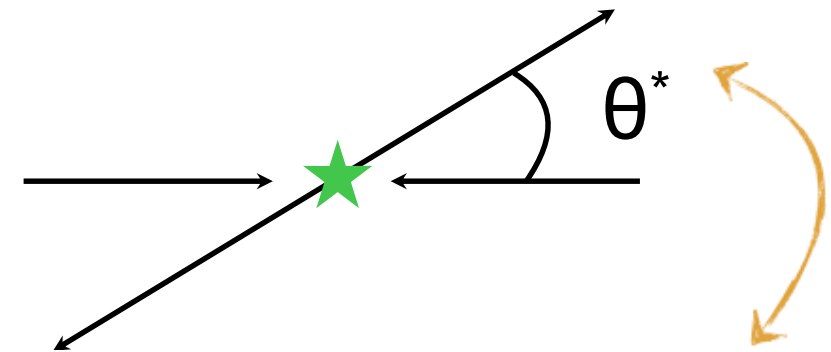
- **Physics observables**

Use two leading jets:

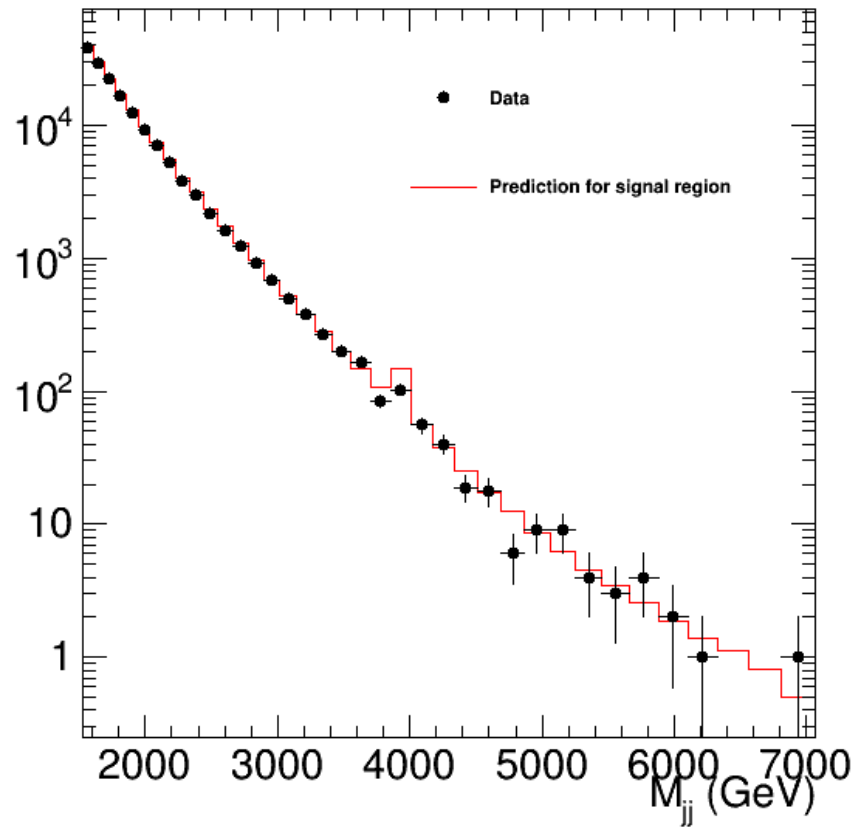
$M(jj) \rightarrow$ Resonance Mass

$\Delta\eta(jj) \rightarrow$ Resonance Spin

(X rest frame)

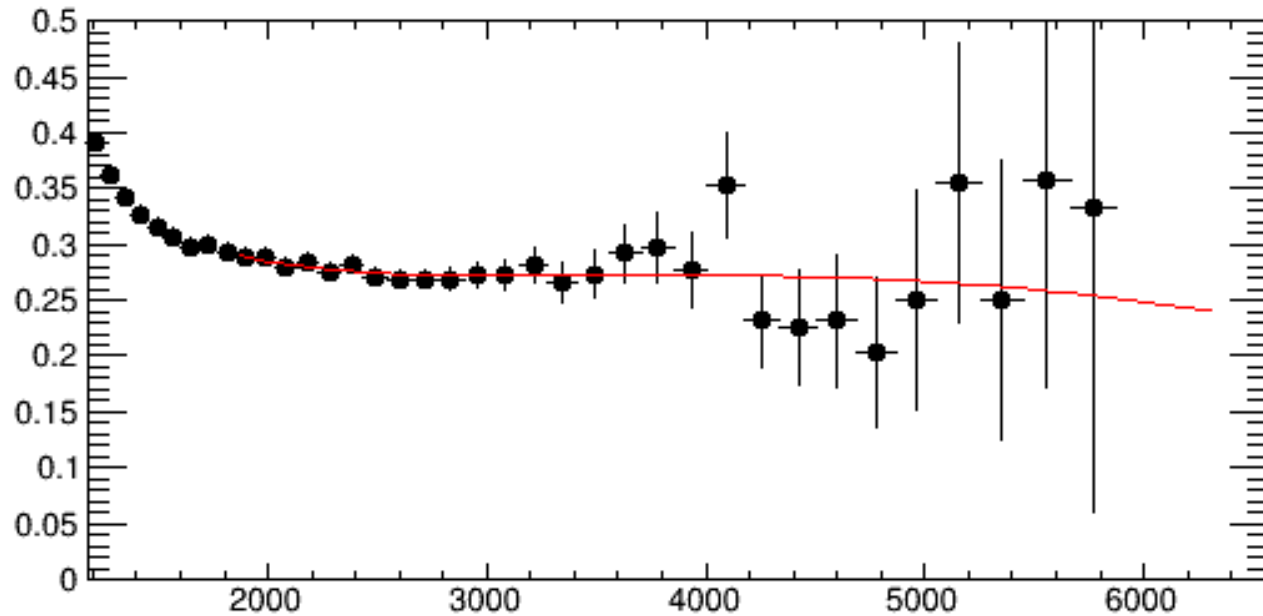


$$\Delta\eta_{12} = |\eta_{jet1} - \eta_{jet2}| = \ln \frac{1 + |\cos\theta^*|}{1 - |\cos\theta^*|}$$



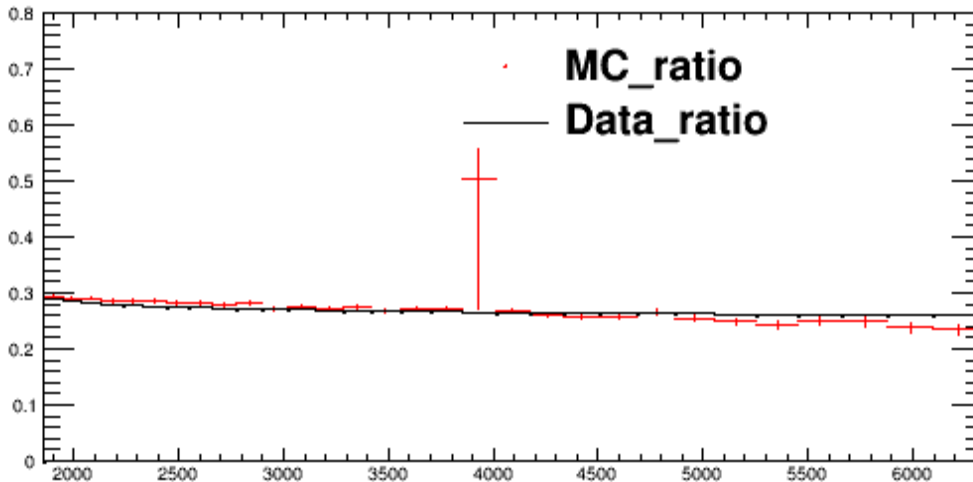
- Data and prediction agree nicely for $M_{jj} > 2.8 - 3.0$ TeV. Below 2.8 TeV we observe large pulls.
- The same behaviour is also observed comparing R between data and simulation (see next slides).

Ratio of Signal/Control



- For low M_{ij} Data ratio can be used.
- Data ratio requires smoothing to avoid statistical fluctuations.
- Agreement of data ratio with simulation needs to be checked.

Ratio of Signal/Control

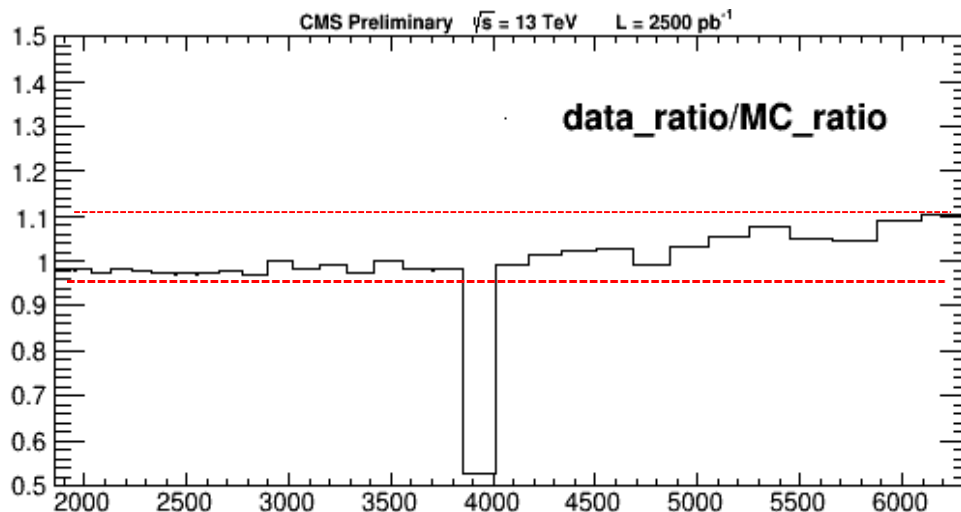


- Data ratio agrees with simulation to within +/-5% for $M_{jj} > 1.5$ TeV (below we have trigger turn on effects in data) .

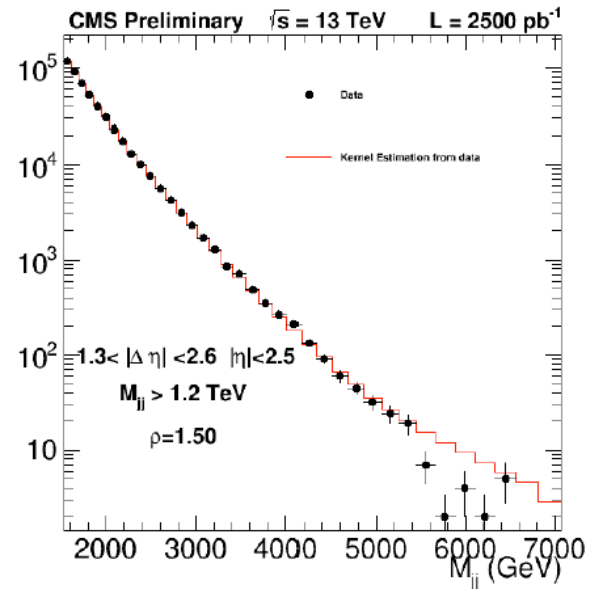
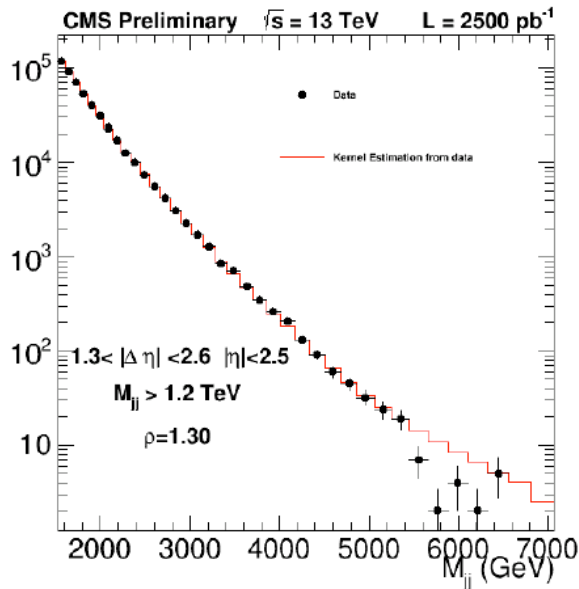
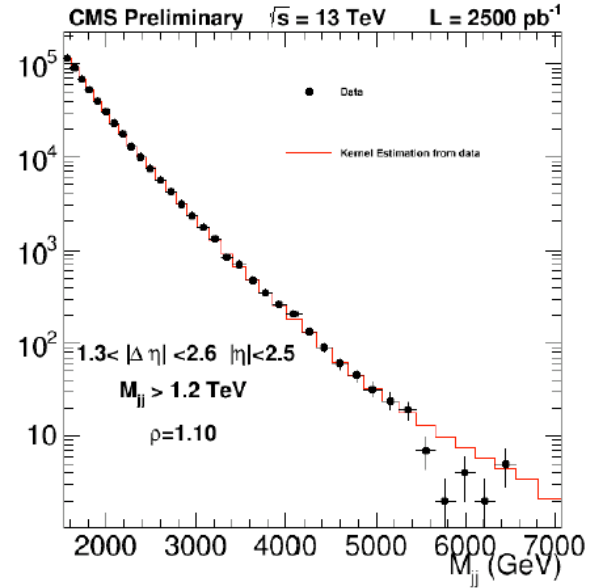
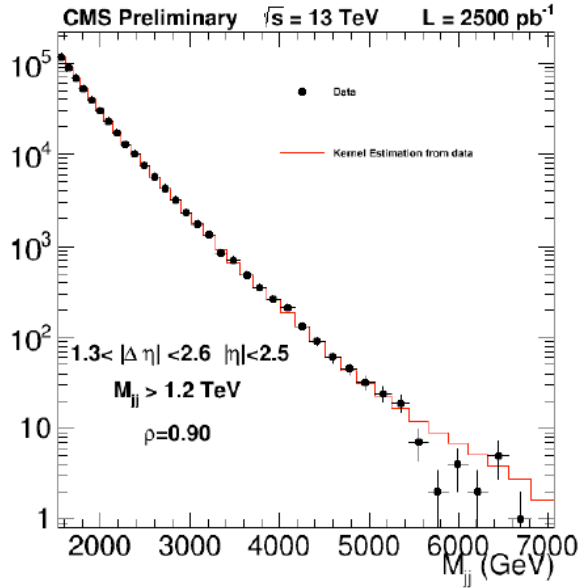
effects in data) .

- Data and simulation agree even better for $M_{jj} > 3$ TeV.

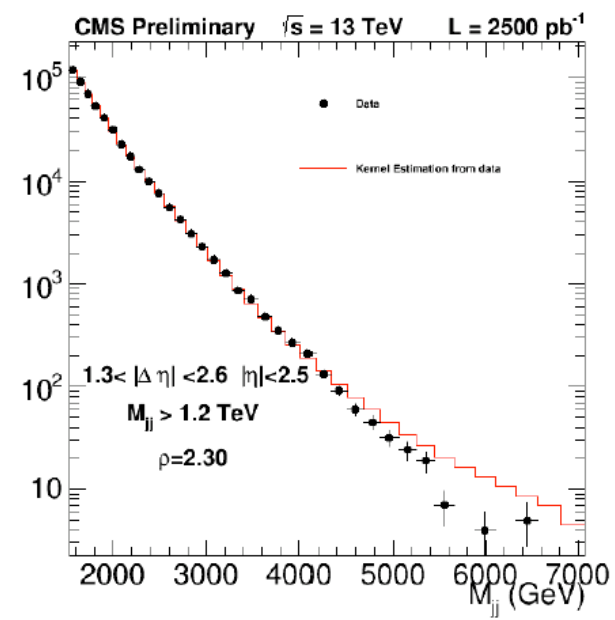
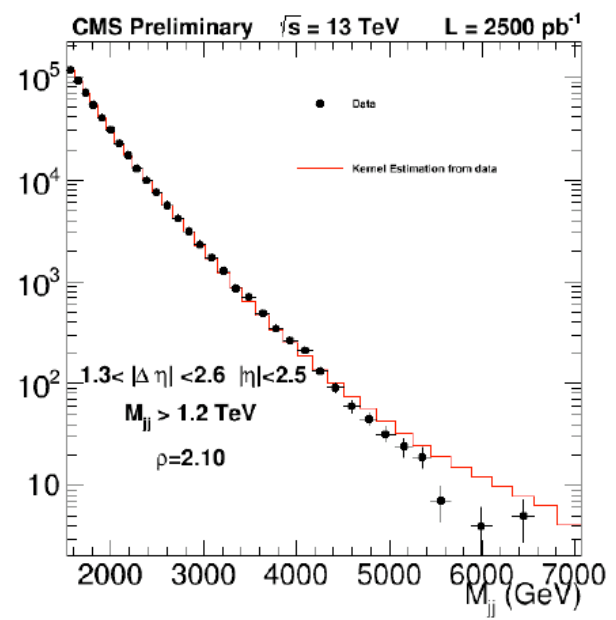
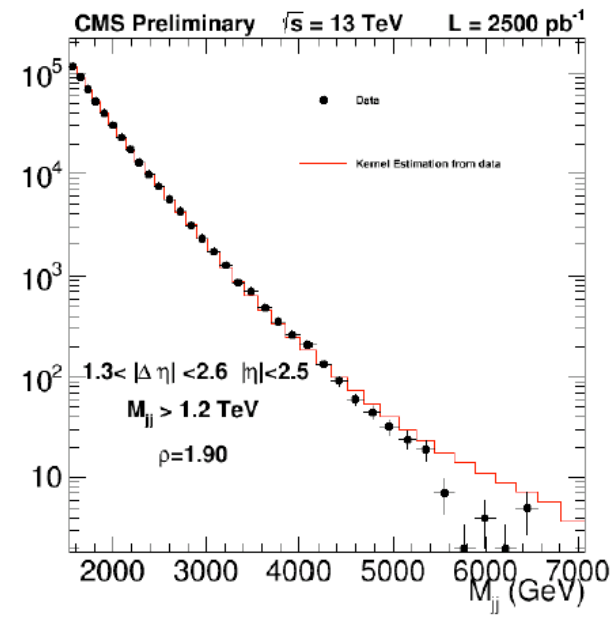
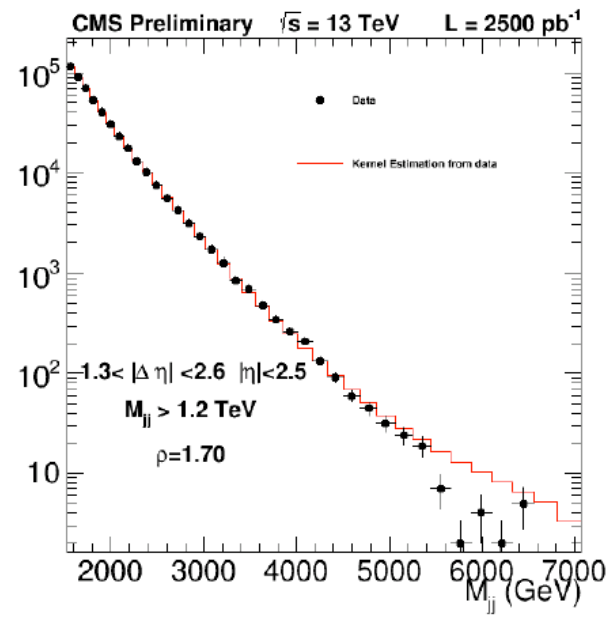
- The agreement with the old JEC can be see in the backup.



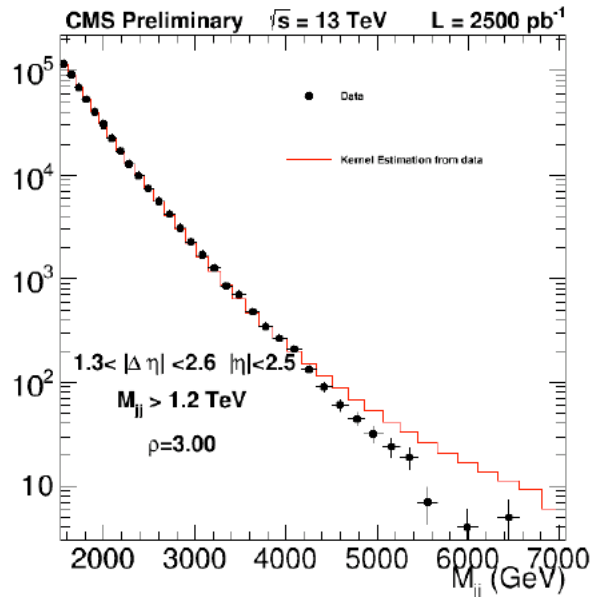
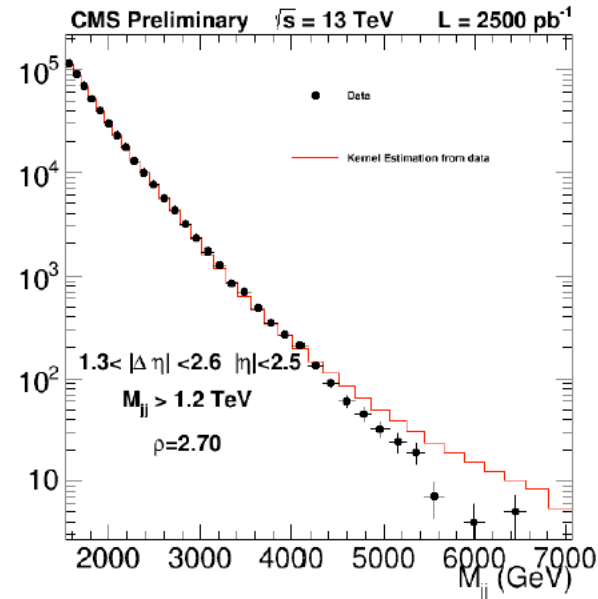
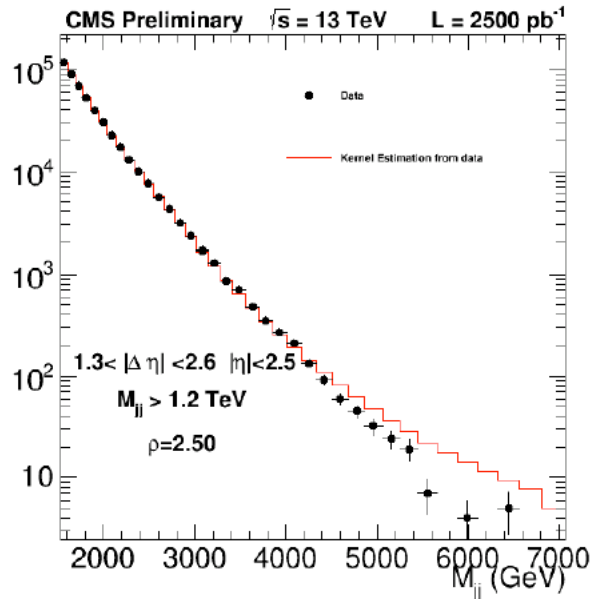
Kernel Estimation vs ρ (1)



Kernel Estimation vs ρ (2)

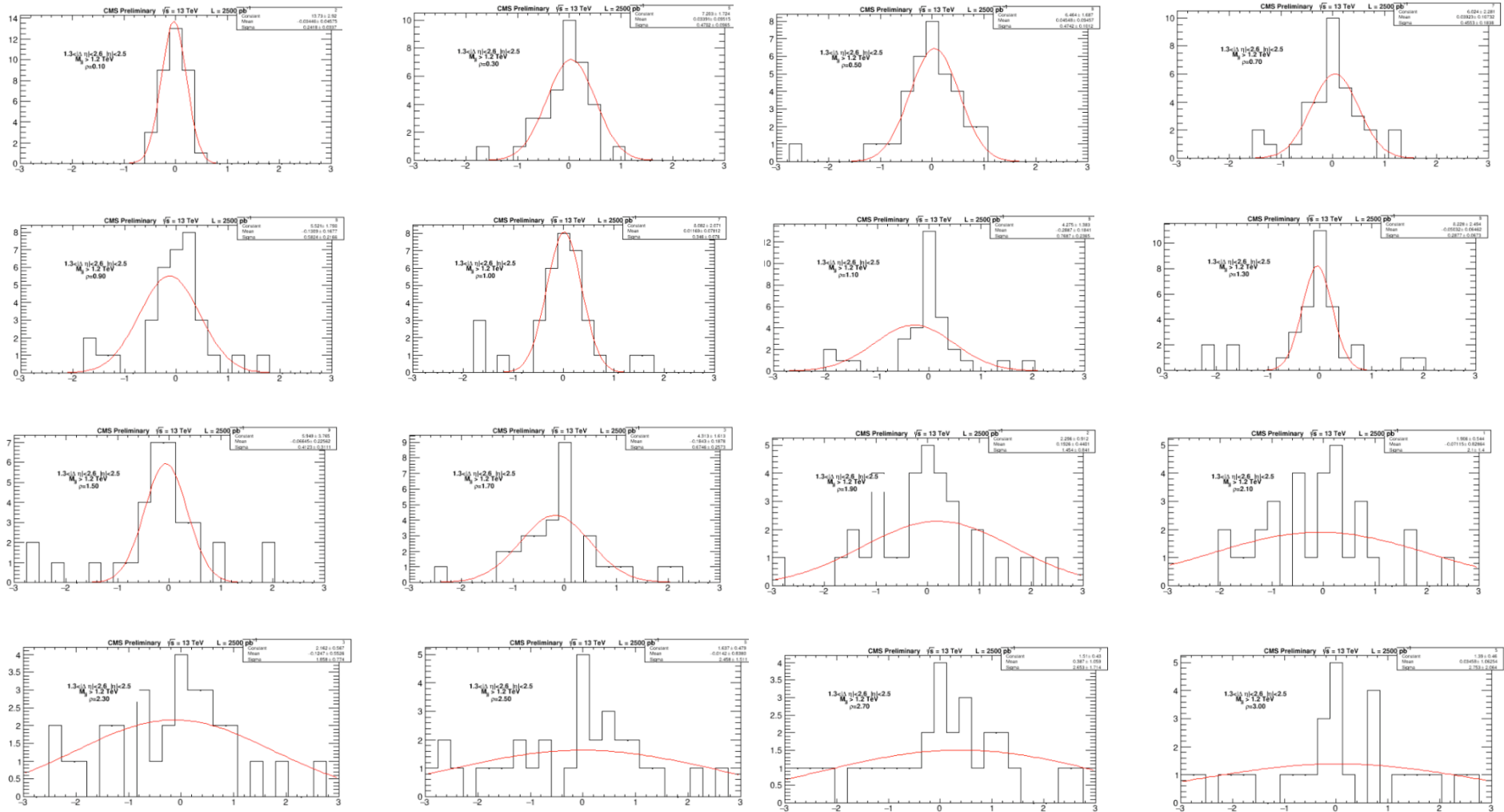


Kernel Estimation vs ρ (3)



For $\rho \ll 1$ the kernel estimation follows very closely the data statistical fluctuations, for $\rho \gg 1$ the kernel estimation over-smooths the spectrum

Kernel Estimation vs ρ : pulls



For $\rho \ll 1$ the kernel estimation follows very closely the data statistical fluctuations, for $\rho \gg 1$ the kernel estimation over-smooths the spectrum