LIP-STUDENTS-20-19

Measurement of B meson production in pp collisions at 5 TeV

Maria Faria[1,a]

[1] *Instituto Superior Técnico, Lisboa, Portugal*
[2] *Massachusetts Institute of Technology, Cambridge, USA*

Project supervisors: N. Leonardo[1], Z. Shi[2]

October 2020

Abstract. We perform a measurement of the B production differential cross section in proton-proton collisions at a center of mass energy of 5.02 TeV. Multivariate (machine learning) algorithms are employed in the data selection, and likelihood methods are used to extract the B signal from data. The detector efficiency is determined from simulation which in turn is validated through detailed comparisons with data. A study of the sources of systematic uncertainty for the cross section is performed. This measurement contributes to furthering our understanding of b-quark and hadron production, and coupled with similar measurements in PbPb collisions provides insight on the nature of the QGP medium.

Keywords: LHC, CMS, b quark, differential cross section, quark gluon plasma

1 Introduction

At our energy scale, quarks and gluons are confined inside hadrons, such as protons and neutrons. However, under extreme conditions of temperature and pressure, a new state of matter is formed, the so-called quark-gluon plasma (QGP), in which quarks and gluons are asymptotically free. This QGP is believed to have existed in the very early universe and can also be re-created at the LHC, in high-energy collisions of heavy ions, such as PbPb collisions.

In this work, we present a preliminary measurement of the B meson differential cross section using the decay $B_s^0 \rightarrow J/\Psi\ \phi$ in proton-proton (pp) collisions at $\sqrt{s}$ = 5.02 TeV, using data collected by the CMS experiment. This differential cross-section measurement is by itself of great significance, and it complements studies performed at other LHC energies, allowing one to study the $\sqrt{s}$ dependence of the cross section. It is also an ingredient necessary to calculate the nuclear modification factor (R_{AA}), a quantity with which we can gain more insight about the QGP, and that can be derived from the results here obtained in pp collisions along with similar studies in PbPb collisions [1].

1.1 The CMS detector

The CMS detector (Compact Muon Solenoid) is a general purpose detector at the LHC, at CERN. It has a cylindrical shape with a superconducting solenoid which provides a magnetic field of 3.8T. It has a central region where the collisions occur, followed by a silicon tracker which tracks the passage of charged particles (which curve in opposite directions for particles with opposite charge), then an electromagnetic calorimeter, where photons and electrons typically lay their energy in the form of energy clusters (showers) and a hadronic calorimeter, where hadrons deposit their energy. Continuing outwards, there are the muon chambers, with up to four stations of gas-ionization muon detectors installed outside the solenoid and sandwiched between the layers of the steel return yoke. A more complete description of the CMS detector can be found elsewhere [3].

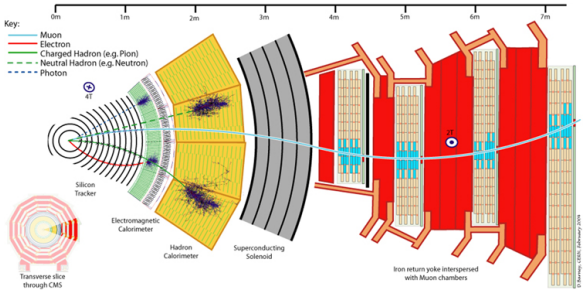


Figure 1: Schematic transverse view of a slice of the CMS detector.

For the purpose of our analysis, the more relevant subdetectors are the innermost and outermost layers of the CMS detector. The latter, the muon chambers, provide identification of the final-state muons, thus further allowing to select the events of interest in real time (trigger). The silicon tracker provides precise measurements of the charged-particles trajectories (muons and kaons), and allows the identification of the secondary decay vertex, the distinctive experimental signature of b-quark hadrons.

1.2 Cross section

A quantity of interest in particle physics is the cross section for a given production process. The differential cross section per transverse momentum is given by:

$$\frac{d\sigma}{dp_T} = \frac{1}{\epsilon B L}\frac{N_s}{\Delta p_T}\ . \tag{1}$$

where ϵ is the detector efficiency, B is the decay branching fraction $B = (31.3 \pm 2.3) \times 10^{-6}$ [4], L is the total integrated luminosity of the data set, $L = 302.3 pb^{-1}$, and N_s is the

[a] e-mail: maria.faria@tecnico.ulisboa.pt

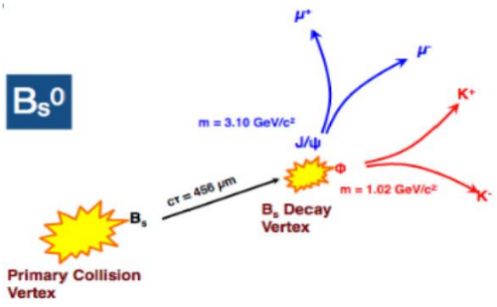


Figure 2: Illustration of the B_s^0 meson decay topology.

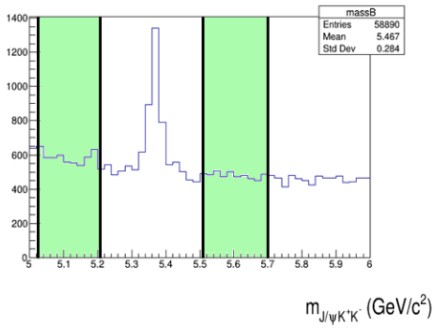


Figure 3: Invariant mass distribution prior to ML selection. The vertical bands (green) indicate the sideband regions used for selecting background events used in training.

raw signal yield extracted from the B_s^0 invariant mass fit, as described in Sect. 3.

In order to probe the properties of the QGP, the so-called nuclear modification factor (R_{AA}) is determined. The R_{AA} is a ratio between the cross sections obtained in heavy ion (PbPb) and proton-proton (pp) collisions, scaled by the expected number of binary collisions:

$$R_{AA} = \frac{1}{\langle N_{coll} \rangle} \frac{(d\sigma/dp_T)_{PbPb}}{(d\sigma/dp_T)_{pp}} \,. \tag{2}$$

As the QGP is expected to be formed in PbPb collisions, as opposed to pp collisions, the R_{AA} quantifies the effect of the QGP medium on the b quark fragmentation compared to vacuum. A study of the differential cross section in PbPb collisions has already been performed in [1, 5].

The structure of this work is as follows. In Sect. 2 the criteria used for identifying the B_s^0 meson signal in the pp data is presented, allowing the signal yield to be extracted in Sect. 3 from the invariant mass spectrum. Simulation is employed for determining the detector efficiency and its validation is carried out in Sect. 5, using the methods detailed in Sect. 4. Finally, the differential cross section measurement is presented in Sect. 6.

2 Dataset and selection

The dataset employed in this analysis was collected by CMS in pp collisions in a dedicated LHC Run in 2017 at an energy $\sqrt{s} = 5.02\,TeV$. The trigger algorithm required events to contain two muon candidates. Offline, the muon pair is combined with a pair of charged tracks to reconstruct the signal decay, as presented in Sect. 2.1. A multivariate classification method, based on machine learning (ML) techniques, is then used in order to separate signal from background, the details of which will be given in Sect. 2.2.

2.1 Decay channel and discriminating variables

The decay channel used in our analysis is $B_s^0 \to J/\Psi\phi$, followed by the decays $J/\Psi \to \mu^+\mu^-$ and $\phi \to K^+K^-$, and is pictorially illustrated in Fig. 2. The B_s^0 meson, composed by a quark s and an anti-quark b, is formed at the pp collision point (the primary vertex, PV), and it decays at a separate point (the secondary vertex, SV). The latter is reconstructed from a vertex-fit to the trajectories of the 4-particle ($\mu^+\mu^-K^+K^-$) final state. The kaon candidates are taken as charged particles reconstructed in the silicon detector, and no further charged hadron identification is performed; they will as such be referred to as *tracks* hereafter.

The B_s^0 meson despite being unstable has relatively a large lifetime, and consequently travels a measurable distance in the detector, between the PV and the SV ($c\tau$ = 456nm, which in the lab frame corresponds to a few mm). Because the J/Ψ and the ϕ have very short lifetimes, they are practically formed and decay at the same point, the SV.

The most relevant quantities used for selecting the signal candidates and distinguishing them from background processes include the following:

- Bmass, invariant mass of the B_s^0 meson candidates;
- Bpt. transverse momentum of the B_s^0 meson candidates;
- By. rapidity of the B_s^0 meson candidates;
- BtrkPt / Bmupt, transverse momentum of the tracks / muons;
- BtrkEta / Bmueta, pseudorapidity of the tracks / muons;
- Balpha, 3D opening angle between the B_s^0 3-momentum and the PV to SV vector;
- Bdtheta, 2D (projection in the plane perpendicular to the beam axis) opening angle between the B_s^0 3-momentum and the PV to SV vector;
- BsvpvDistance, 3D distance between PV and SV;
- Bd0, 2D distance between PV and SV.

2.2 Machine Learning

The variables described in Sect. 2.1 are fed into an algorithm for signal versus background discrimination. Two classifiers are tested: Genetic Algorithm (CutsGA) and Boosted Decision Tree (BDT). In the former, rectangular cuts are applied to the data, whereas in the latter, a multi-dimensional correlated selection is performed.

The algorithms are trained using the following labeled data: background events from the mass sidebands, which

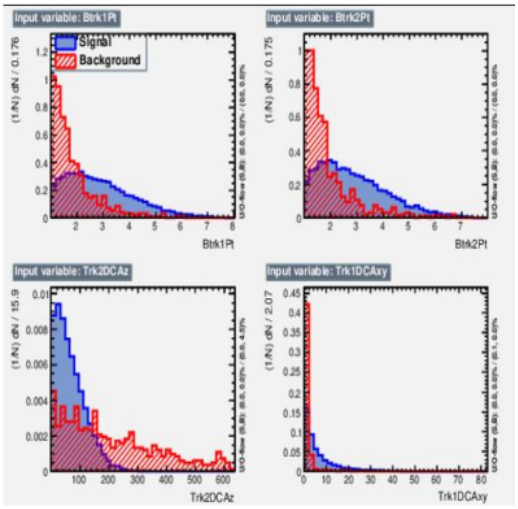


Figure 4: Normalized distributions for signal (blue) and background (red), for selected discriminating variables, that serve as feature input to ML training.

are depicted in green in Fig. 3, and signal events from Monte Carlo simulation (MC). As the background composition varies with the event kinematics, the training is performed independently for the separate p_T ranges, that will be employed in the differential cross-section measurement.

The normalized distributions of some of the discriminating variables are presented in Fig. 4, from both signal and background samples, in the $p_T(B)$ range [10,15] GeV. The correlation matrix for the signal is depicted in Fig. 5.

The performance of the classifiers may be determined from the background rejection versus signal efficiency graph (ROC curve), which is presented in Fig. 6. The BDT displays a favourable performance, which can be inferred as well from the largest AUC (area under the ROC curve).

The optimal working point is the one that maximizes a suitable figure of merit (FOM) for a given analysis. Here we consider the signal yield significance, $FOM = S/\sqrt{S+B}$, where S and B are the number of signal and background events, respectively, that are selected. The significance is shown as a function of the BDT score threshold in Fig. 7 (left). The maximum FOM of 14 if obtained for a BDT score of 0.01, for which the signal efficiency is 0.79, as depicted in Fig. 7 (right).

A comparison of the B_s^0 candidates mass distributions before and after selection is presented in Fig. 8.

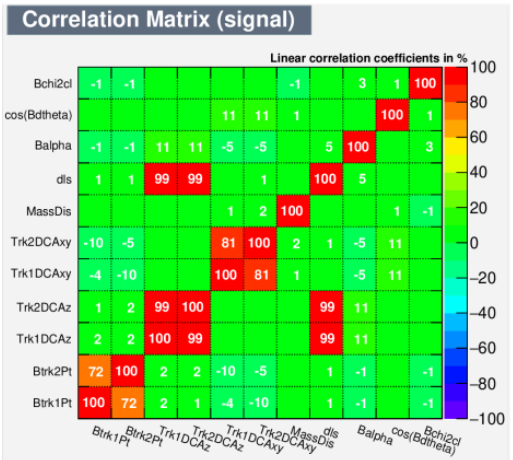


Figure 5: Correlation matrix for the signal.

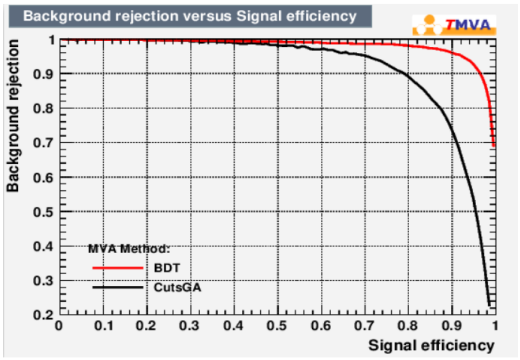


Figure 6: Classifier performance.

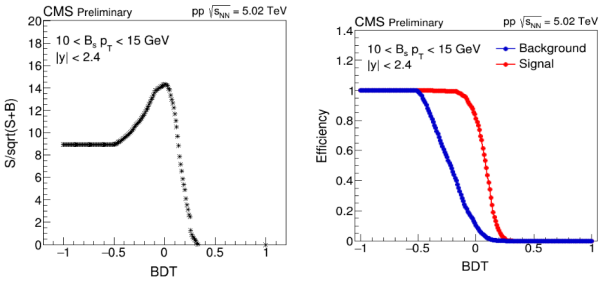


Figure 7: BDT optimal working point for p_T [10,15] GeV.

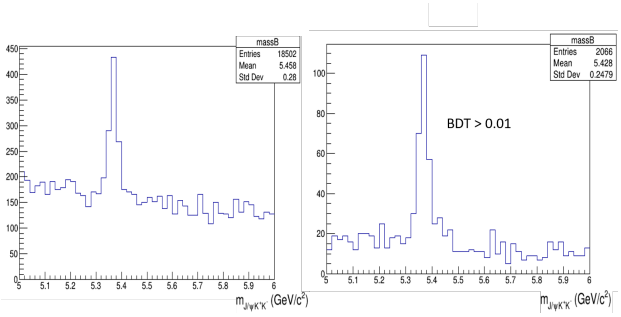
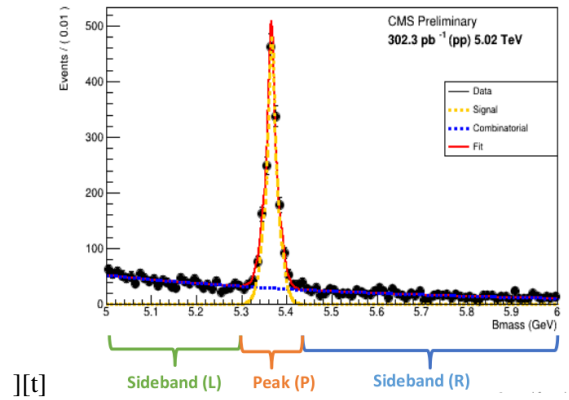


Figure 8: B_s^0 candidates before (left) and after (right) selection.

3 Likelihood method

The Extended Unbinned Maximum Likelihood (EUML) mehod is used to fit the B_s^0 invariant mass distribution. The



][t]

Figure 9: Fit to the invariant mass of the B_s^0 candidates. The results of the fit, and both signal and background components, are overlaid to the data.

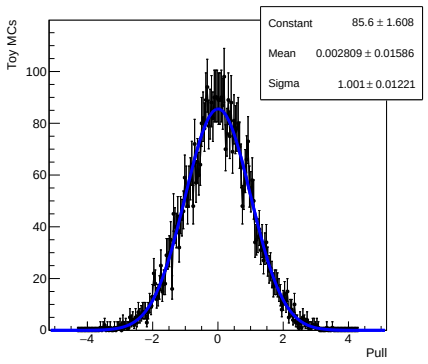


Figure 10: Pull distribution obtained by generating and fitting 5000 toy MCs. A fit to the pull distribution shows its compatibility with a unit Gaussian (overlaid blue line).

results of the fit are presented in Fig. 9. The signal component is described by the sum of two Gaussian functions with the same mean:

$$P_S = \alpha\, Gauss(\mu, \sigma_1) + (1-\alpha)\, Gauss(\mu, \sigma_2)\,, \quad (3)$$

and a combinatorial background component described by an exponential function:

$$P_{CB} = Exp(\lambda)\,. \quad (4)$$

For each B meson candidate with mass m_i we can define the quantity $l(m_i)$ as the sum of these two probability density functions (PDF) multiplied by their respective normalizations (yields):

$$l(m_i) = N_s\, P_S(m_i; \mu, \sigma_1, \sigma_2) + N_{CB}\, P_{CB}(m_i; \lambda)\,. \quad (5)$$

The likelihood function is then given by the product over all observed candidates:

$$L(m_i, \vec{\lambda}) = \prod_{i=1}^{N_{obs}} l(m_i) \times \frac{e^{-N}\, N^{N_{obs}}}{N_{obs}!}\,. \quad (6)$$

The last term constrains the sum of the signal and background candidates, N, to follow a Poisson distribution.

The method is called extended (E) since it takes into account the Poisson term, unbinned (U) since it uses the mass m_i of each candidate without making use of bins and maximum (M) since it finds the set of parameters $\vec{\lambda}$ that maximize the likelihood function. The results obtained with the EUML method and their respective statistical uncertainties are summarized in Table 1. Our parameter of interest is the signal yield N_s.

In order to validate the fit, 5000 pseudo-experiments (toy MCs) were generated and the results of the fit were used to obtain a signal yield N_i and its respective uncertainty σ_i to each pseudo-data sample. This allows us to form the *pull* as:

$$Pull = \frac{N_i - N_s}{\sigma_i}\,, \quad (7)$$

whose distribution is plotted in Fig. 10.

The fit is unbiased if it gives the correct value of N_s i.e. if N_i is statistically close to N_s, which translates into a pull distribution given by a unit Gaussian (centered around zero with σ close to 1). A Gaussian fit to the pull distribution yields $\mu = 0.0028 \pm 0.1586$ and $\sigma = 1.001 \pm 0.012$ showing that the fit does not have indeed any significant biased.

4 Signal extraction

In order to separate the signal component from the background, two methods are used: sideband subtraction and sPlot. Here we present a description of both methods and some of the results obtained.

4.1 Sideband subtraction

In the sideband subtraction method, the mass distribution is divided into three regions: a peak region and two sideband regions (left and right); as is illustrated in Fig. 9. The signal distribution of a certain variable V, such as the rapidity of the B meson (Fig. 11), is then calculated in the following way:

$$V_{signal} = V_{peak} - r \times V_{sideband}\,, \quad (8)$$

where the factor r is a ratio between the integrals of the background distribution over the peak region (P) and over

Table 1: Results of the EUML method.

Coefficients	Value±Statistical Uncertainty
α	0.776 ± 0.046
λ	-1.6065 ± 0.0726
μ	5.36677 ± 0.00041
σ_1	0.01775 ± 0.00091
σ_2	0.00539 ± 0.00073
N_s	1409 ± 42
N_{CB}	2582 ± 54

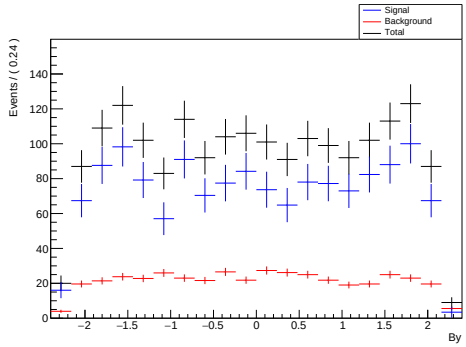


Figure 11: Sideband subtraction result for the B_s^0 rapidity.

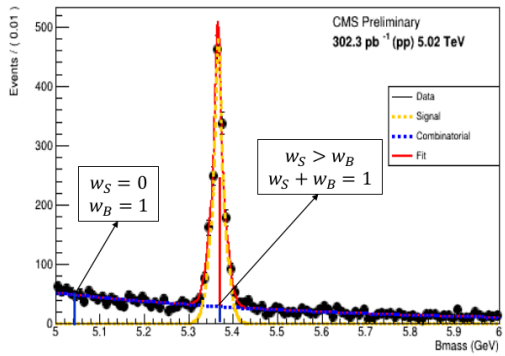


Figure 12: Illustration of the sPlot method.

the sum of left and right sideband regions (L+R):

$$r = \frac{P}{L+R}\,, \tag{9}$$

and gives a measure of how much background there is in the peak region.

4.2 sPlot

The sPlot method uses the result obtained with the EUML fit. For each candidate, it attributes two weights: the probability of being signal (w_S) and the probability of being background (w_B), as is illustrated in Fig. 12.

In the sideband regions, the probability of any point belonging to the signal is zero whereas in the peak region each point has a nonzero w_S and w_B such that $w_S + w_B = 1$. For illustration, w_S corresponds to a ratio between the height of the red bar and the sum of the heights of the red and blue bars; w_B corresponds to a ratio between the height of the blue bar and this sum.

The result of applying this method to the B_s^0 transverse momentum (Bpt) is presented in Fig. 13.

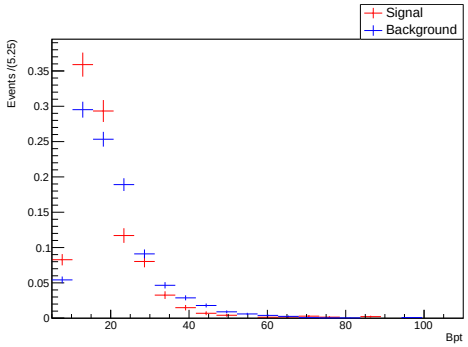


Figure 13: sPlot result for the B_s^0 transverse momentum.

5 MC validation

One of the ingredients in the calculation of the cross section in Eq. 1 is the detector efficiency ϵ, which is determined from simulations of the signal process, referred to as Monte Carlo (MC). The MC needs to be validated by comparing its distributions to those obtained from the data, and evaluating whether or not it provides a good description.

A comparison between the two methods described in Sect. 4 and the MC results can be seen in Fig. 14. The sideband subtraction and sPlot methods are found to agree very well. Some level of disagreement between the data and MC is however dound for some variables, as can be seen in Fig. 15.

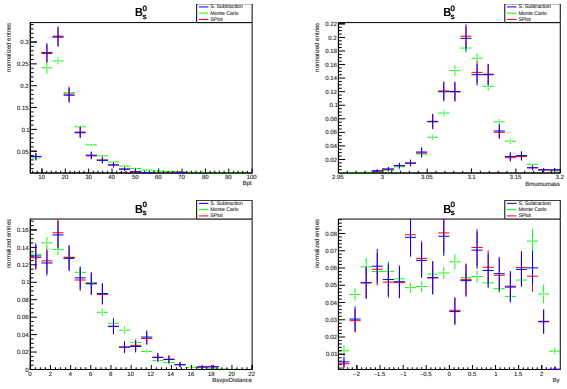


Figure 14: Comparison between the Sideband Subtraction and sPlot methods (data) with MC.

Because the sPlot method is more robust since it uses the result of the full likelihood fit, we use it to compute the ratios between data and MC that can be seen in the bottom panels of Fig. 15.

We can re-weight the MC using the data/MC ratios of a certain variable (e.g. p_T of the B meson) and check how this affects the data-MC agreement of the other variables. The comparison between data and MC after having re-weighted the MC with the Bpt ratios can be seen in Fig. 16.

For the re-weighted variable, Bpt, data and MC distributions now overlap, which is by construction. We can see

some improvement in the data-MC agreement for the p_T of the tracks. However, this agreement remains fairly the same for the opening angle and the pseudo-rapidity of the tracks. This stems from the fact that the p_T of the tracks is more correlated with the p_T of the B meson than the other variables.

The remaining disagreements are used to compute the systematic uncertainty associated to the detector efficiency determination, as will be discussed in the next section.

6 Differential cross section and systematic uncertainties

In order to compute the differential B production cross section in pp collisions, according to Eq. 1, two quantities need to be calculated: the normalized signal yield ($N_s/\Delta p_T$) and the detector efficiency (ϵ). An account of the procedures used to calculate them will be given respectively in Sects. 6.1 and 6.2. The preliminary results are presented in Sect. 6.3.

6.1 Normalized signal yield

In order to compute the normalized signal yield ($N_s/\Delta p_T$), we split the dataset into 4 different p_T regions: 5-10, 10-15, 15-20, 20-50 GeV. For each bin we then extracted the signal yield N_s from the likelihood fit, as discussed in Sect. 3. The ordinate of the points in Fig. 17 is obtained by dividing the raw signal yield (N_s) by the p_T bin width. The abscissa is the p_T mean evaluated with the sPlot method presented in Sect. 4.2.

The statistical uncertainty in the signal yield N_s comes directly from the likelihood fit. The systematic uncertainty is obtained by comparing the signal yield obtained with the nominal model (signal: 2 Gaussians with the same mean; background: exponential) used in the EUML fit, with the 3 following alternative model descriptions:

- Bkg-Poly: the CB background PDF is a 1st order polynomial;
- Fit-Range: the left sideband is excluded from the fit;

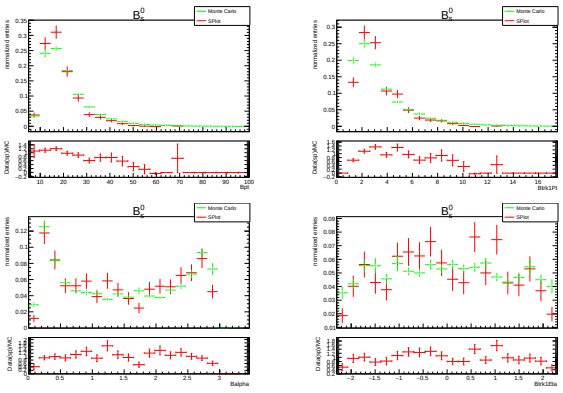


Figure 15: Comparison between the signal distributions obtained from data with the sPlot method and the MC. The bottom panels show the ratios between data and MC.

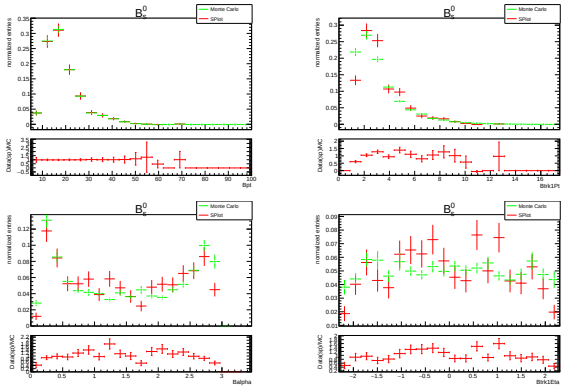


Figure 16: Comparison between the signal obtained from data with the sPlot method and the MC, after having reweighted the MC with the *Bpt* correcting ratios from Fig. 15.

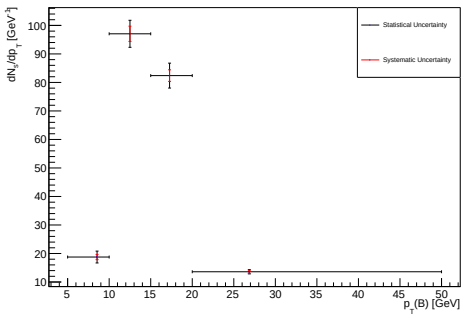


Figure 17: Normalized signal yield for p_T bins 5-10, 10-15, 15-20, 20-50 GeV.

- Signal-1Gauss: the signal PDF is only 1 Gaussian.

The total systematic uncertainty in N_s is obtained by combining the different sources, assumed to be uncorrelated:

$$\sigma_{syst\ yield} = \sqrt{\sigma^2_{bgk-poly} + \sigma^2_{fit-range} + \sigma^2_{signal-1gauss}}\,. \quad (10)$$

6.2 Efficiency

The detector efficiency is determined from MC simulations and measures how much signal is not reconstructed or is rejected by the selection cuts we apply in our analysis. Two MC samples are used: one without any cuts and one with the selection cuts applied.

For performing the differential measurement, the efficiency needs to be determined in the same p_T bins used earlier for yield extraction, in Sect. 6.1. This is done by taking the ratio of the the B meson transverse momentum distributions, with and without selection cuts applied,

$$\epsilon = \frac{(p_T)_{\text{after cuts}}}{(p_T)_{\text{before cuts}}}\,. \quad (11)$$

The expression in Eq. 11 gives the nominal efficiency (ϵ^0), that enters in Eq. 1 and is plotted in Fig. 18 (left), as a function of each p_T bin.

The efficiency determination in Eq. 11 relies on the MC accurately describing the data. Possible mismatches, investigated in Sect. 5, result accordingly in a systematic uncertainty. This uncertainty is determined by recomputing the efficiency (ϵ^1) using the re-weighted MC simulation. The weights correspond to the data/MC ratios obtained with the sPlot method (Sect. 4.2) for the distribution of the BDT score described in Sect. 2.2. The relative systematic uncertainty in the efficiency, shown in Fig. 19, is given by $\Delta = \frac{\epsilon^1 - \epsilon^0}{\epsilon^0}$.

6.3 Cross section results

With all the quantities already calculated, the differential cross section can then be obtained using Eq. 1 and the result is presented in Fig. 20.

The statistical uncertainty comes solely from the signal yield contribution and quantifies the statistical precision allowed by the data. The systematic uncertainty has 4 different sources: integrated luminosity, branching fraction, efficiency and signal yield; and is given by:

$$\sigma_{syst} = \sqrt{\sigma^2_{\text{lumi}} + \sigma^2_{\text{branch}} + \sigma^2_{\text{yield-syst}} + \sigma^2_{\text{eff-syst}}}\,. \qquad (12)$$

Table 2: Cross-section statistical and systematic uncertainties for each p_T bin. Relative errors are shown.

p_T (GeV)	yield-syst	eff-syst	total-syst	total-stat
5-10	0.049	0.028	0.114	0.111
10-15	0.028	0.007	0.103	0.049
15-20	0.025	0.007	0.102	0.053
20-50	0.038	0.0006	0.106	0.055

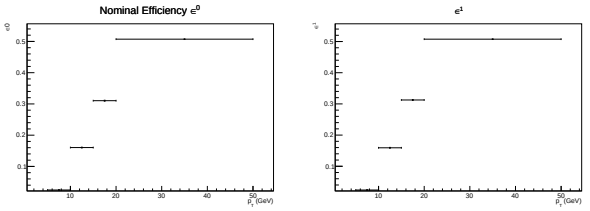


Figure 18: Nominal efficiency (left) and efficiency variation calculated with the re-weighted MC using the BDT score correcting factors (right).

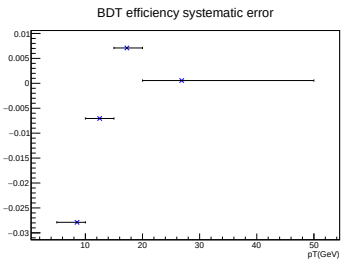


Figure 19: Systematic uncertainty on the efficiency.

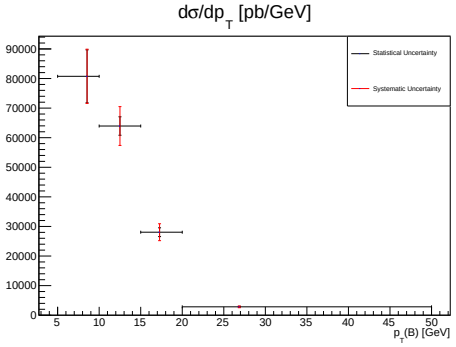


Figure 20: Preliminary results on the B^0_s production

A summary of the differential cross section uncertainties is presented in Table 2. The statistical uncertainty is bigger in lower p_T bins which reflects the lower population of those regions in data. The relative systematic uncertainty in the integrated luminosity and in the branching fraction are, respectively, 0.023 and 0.096, taken from [4]. The total systematic uncertainty is larger than the statistical uncertainty for all bins. The dominant sources of systematic uncertainty come from the branching ratio and the signal yield.

7 Conclusions

We presented preliminary results on the B production differential cross section in pp collisions at a center of mass energy of 5.02 TeV. The signal yield was obtained using the Extended Unbinned Maximum Likelihood (EUML) fit. Two methods were used to extract signal from data: sideband subtraction and sPlot. The latter is more robust, since it uses the result of the EUML fit and was for that reason used to compute the data/MC ratios. These ratios were then used to re-weight the MC and to compute the systematic uncertainty on the detector efficiency. The detector efficiency was determined as a ratio of the p_T distributions of the MC sample before and after the selection cuts are applied.

These results can be used to test theory calculations of the B hadron cross section and its kinematic and energy dependencies. Along with future similar measurements of other B mesons, it allows the study of the b-quark fragmentation in pp collisions. Further studies will also compare the B production cross section in pp and PbPb collisions and calculate the nuclear modification factor, paving the way for a better understanding of the properties of the QGP.

Acknowledgements

I'm grateful to Nuno Leonardo and Zhaozhong Shi for the time and support given and to the work of Alexandra Pardal, João Gonçalves and Júlia Silva that served as inspiration. I further thank Artur Semião, my colleague in this internship, for the work he did in the selection optimization.

References

[1] J. Gonçalves, A. Pardal, LIP-STUDENTS-2019-08 (2019)
[2] PDF, Phys. Rev. D 98, 030001 (2018)
[3] CMS, JINST 3 (2008) S08004
[4] PDG, PTEP 083C01 (2020)
[5] CMS Coll., CMS-PAS-HIN-19-011 (2020)