Measurement of the tZq Differential Cross-section with the ATLAS Detector at the LHC

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Nilima Akolkar
aus
Vadodara, India

Bonn 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. John Smith
2. Gutachterin: Prof. Dr. Anne Jones

Tag der Promotion:
Erscheinungsjahr:

Contents

Todo list

CHAPTER 1

Statistical Methods and Analysis Setup

In particle physics, the abundance of available data for exploration has dramatically increased with the development of powerful particle colliders. Be it estimating a parameter of the SM or finding evidence of new physics, particle physicists heavily rely on statistical techniques to churn out reliable information from observed data. This chapter provides details on the statistical methods used in this thesis. Basic statistical techniques including parameter estimation and method of maximum likelihood are outlined in Section 1.1. The concept of unfolding is introduced and the techniques used in this thesis are explained in detail. This chapter also discusses the method of profile likelihood unfolding, which is the core technique used in this analysis.

1.1 Statistical Inference

The main goal of a statistical analysis is to infer key information from the observed data. Commonly used approaches of inference are frequentist and Bayesian which are based on the interpretation of probability. In statistics, probability is simply the "chance" of an event ocurring. In the frequentist approach, probability is interpreted as the frequency of an event occurring, whereas in the Bayesian approach, probability is the extent of belief in the occurrence of an event [1].

Statistical problems can be categorised into two types: parameter estimation and hypothesis testing. In parameter estimation, the goal is to determine the "best" possible value of a certain physical quantity or a parameter of a certain mathematical model. It is important to note that the parameter value is always accompanied by an error estimate which quantifies the accuracy of a measurement. A measurement may not be perfect, however the extent of imperfection is concealed in the error values. Therefore, error values play an important role in the interpretation of experimental results.

In hypothesis testing, the main goal is to check if a theory is consistent with the observed data. For this case, the answer is not a numerical value, instead, it is a statement implying how confident we are with the consistency of a theory based on the observed data. In reality, hypothesis testing and parameter estimation are not totally independent of each other. Some problems require estimating a parameter in order to test a hypothesis while in some cases, a parameter is estimated assuming a hypothesis is correct [2]. In this analysis, we focus on parameter estimation. In frequentist statistics, method of least squares and method of maximum likelihood are mainly used for parameter estimation. In this thesis, the maximum likelihood method is used which is discussed in the next section.

Method of maximum likelihood

Statistical models are mathematical expressions that relate observations to underlying theories. These models are characterised by a set of parameters. The possibilities of different outcomes, given a mathematical model, is described by probabilities which is a key concept in statistics. Another important concept which connects observations to parameters, is the likelihood. Likelihood is a measure of how likely a set of parameters can describe the actual observed data. Here, the observed data are fixed, while the parameters are unknown, making this approach conceptually opposite to the definition of probabilities.

A likelihood function for independent observations $x_1, x_2, ..., x_n$ and parameter θ can be expressed as a product of probabilities as follows:

$$L(\theta) = L(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i, \theta) \tag{1.1}$$

We try to find the value of θ that makes the probability of the actual observed outcomes as large as possible. In other words, we search for parameter values that maximise the likelihood function and therefore, this is called the method of maximum likelihood. In practice, it is easier to maximise the logarithm of the likelihood which is the sum of probabilities.

The method of maximum likelihood depends on the observed data. The efficiency of parameter determination will depend on the extent of reliability on observed data. In particle physics, this fact imposes a certain level of expectations from particle detectors. Even though our detectors are well-developed and efficient, there are some limitations that lead to distortion of observations. Physicists often use mathematical techniques to remove the distortions caused by detectors. These techniques fall under the category of *unfolding*.

1.2 Unfolding

In any experiment, the accuracy of measurement highly depends on performance of the apparatus being used. In particle physics, an ideal detector provides the original and complete information of collisions without any loss or distortion of information. However, in reality, the data received from a detector is distorted due to effects such as limited acceptance and finite resolution of the detector. Unfolding is a technique to estimate the original fine structure of data devoid of any detector effects.

There are certain conventions regarding the entities used in unfolding problems within the particle physics community. The distribution of a physical observable obtained from data recorded by the detector is called detector-level distribution. On the other hand, the truth-level distribution represents data that we should have obtained with an ideal detector. This is generated from MC simulations without applying any detector simulations. In unfolding, we also make use of simulated distribution including detector simulation. Since this so-called reconstructed-level distribution is supposed to mimic actual observed data, it is used to validate an unfolding method. Detector-level and reconstructed-level distributions are conceptually the same when discussing unfolding related quantities. Technically, unfolding is a method of estimating the truth-level by correcting detector effects present at detector- or reconstructed-level.

The reconstructed distribution $\vec{x}$ is related to the truth-level distribution $\vec{y}$ by a response matrix R as shown in Eq. (1.2). Here $\vec{b}$ represents backgrounds. In particle physics problems, data is generally

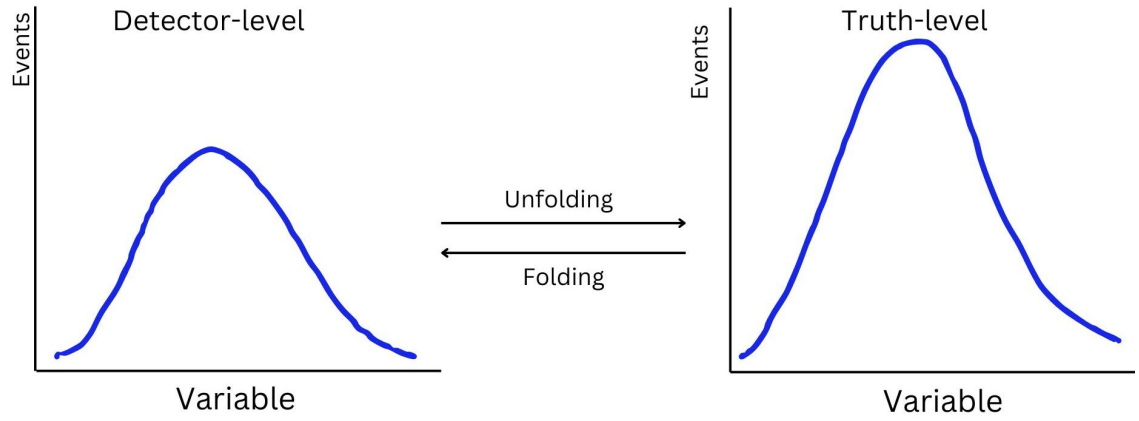

Figure 1.1: An illustration visualising the method of unfolding. The left diagram is an illustration of a histogram obtained from the detector information, called detector-level and the right illustration shows a histogram that is expected form an ideal detector, called truth-level. The difference in the two diagrams depicts the smearing caused by the detector. The procedure to obtain the truth-level information based on the detector-level information is called as unfolding. The reverse procedure is called folding.

organised into histograms with finite bins. In this case the above given relation is given by Eq. (1.3).

$$\vec{x} = R \cdot \vec{y} + \vec{b} \tag{1.2}$$

$$x_i = \sum_{\mathrm{j}\,=\,1}^{\mathrm{bins}} \mathcal{R}_{ij} y_j + \vec{b} \tag{1.3}$$

A response matrix which quantifies the detector effects is computed from simulated data. The quantities used to construct the response matrix are migration matrix and two correction factors namely acceptance and efficiency. A schematic showing reconstructed-level and fiducial truth-level volumes is given in Fig. 1.2. Per-bin acceptance adjusts the number of reconstructed events by the fraction of events that are also present at the fiducial truth-level. It is defined as the ratio of events present at the reconstruction- and truth-levels to the total number of events at the reconstruction-level. Acceptance gives an idea of how well the reconstructed data corresponds the true data. The corrected number of events at the reconstructed-level is given as,

$$N_i^{\mathrm{reco}} = x_i * a_i \tag{1.4}$$

Per-bin efficiency adjusts the number of truth-level events by the fraction of truth-level events that are also found at the reconstructed-level. It is defined as the ratio of a number of events present at both the reconstruction-level and the truth-level to the total number of events at the fiducial truth-level. This correction factor gives an idea of the detector's efficiency to reconstruct true events. The corrected number of events at the fiducial truth-level is given as,

$$N_j^{\mathrm{fid}} = y_j * \epsilon_j \tag{1.5}$$

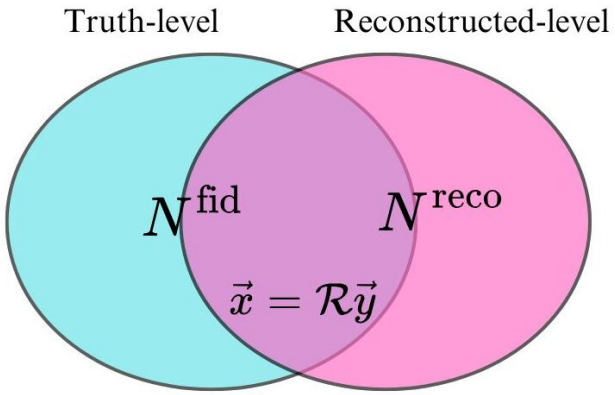

Figure 1.2: A schematic picture showing reconstructed-level volume in blue and truth-level volume in pink.

The migration matrix describes the bin-to-bin migrations between truth-level and reconstruction-level histograms. For instance, M_{ij} represents the fraction of events found in bin i at the reconstruction-level while being created in bin j at the truth-level. A migration matrix with maximum diagonal components indicates that most events are reconstructed in the same bin in which they were generated. Equation (1.3) can be re-written as,

$$N_i^{\text{reco}} = \frac{a_i}{\epsilon_j} * \sum_{\text{j}\,=\,1}^{\text{bins}} M_{ij} \cdot N_j^{\text{fid}} \tag{1.6}$$

Mathematically, the idea of unfolding is to solve Eq. (1.2) for given $\mathcal{R}$, x and b. The resultant values of y can be interpreted as determined true number of events at the truth-level. One would notice a simple way to find the estimators by inverting the response matrix as shown in Eq. (1.7).

$$\vec{y} = \mathcal{R}^{-1}(\vec{x} - \vec{b}) \tag{1.7}$$

Although matrix inversion method is easy to implement, it is a strategy that one should avoid because of its limitations: in some situations, the response matrix is non-invertible then Eq. (1.7) becomes ill-posed. Even though inversion is possible, there are possible statistical fluctuations in the observed data that may cause negative entries in the inverse matrix. This leads to negative number of events in the unfolded distribution which is unrealistic. When a response matrix acts on a true spectrum, it distorts any fine structure present at the truth-level. Despite that, some residue of this fine structure still remains in the reconstructed spectrum [3]. The inverted matrix, acting on the measured data, assumes its statistical fluctuations are the residual fine structure and restores it. In this way, statistical fluctuations are amplified in the unfolded distribution [4] which is undesirable.

In order to overcome the limitations of matrix inversion, various unfolding methods are developed.

Bibliography

[1] H. B. Prosper, *Practical Statistics for Particle Physicists*, arXiv e-prints, arXiv:1504.00945 (2015) arXiv:1504.00945, arXiv: 1504.00945 [stat.ME] (cit. on p. 1).

[2] L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, 1986 (cit. on p. 1).

[3] G. Cowan, *A survey of unfolding methods for particle physics*, Oxford University Press, 1998 (cit. on p. 4).

[4] Spanò, Francesco, *Unfolding in particle physics: a window on solving inverse problems*, EPJ Web of Conferences **55** (2013) 03002, URL: https://doi.org/10.1051/epjconf/20135503002 (cit. on p. 4).

List of Figures

List of Tables