Extrapolation Technique Pitfalls in Asymmetry Measurements at Colliders

Katrina Colletti

Texas A&M University
Physics Department
Master's Defense

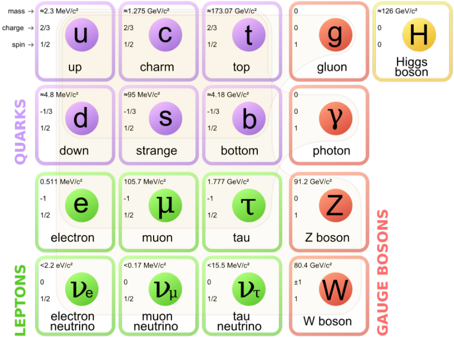
August 11, 2016

Overview[1]

1. Background, Motivation, and Goals
2. Monte Carlo Study
3. Closed Form Statistical Validation
4. Closed Form Numerical Validation
5. Conclusions

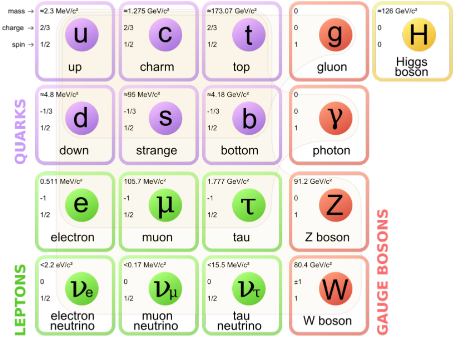
[1] K. Colletti, Z. Hong, D. Toback, and J. S. Wilson, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 830 (2016) 176.

Background, Motivation, and Goals: The Standard Model



- The Standard Model describes the fundamental particles and how they interact with each other
- 4 Fundamental Forces
- Bosons:
  - Gauge bosons (force carriers/mediators): Gluons, W, $Z^{\pm}$, Photon
  - Scalar: Higgs
- 6 Leptons
- 6 Quarks

Background, Motivation, and Goals: The Standard Model

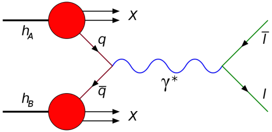


- Agrees extraordinarily with experiment, but not complete
  - A Grand Unified Theory/Gravitons?
  - Matter/Anti-Matter Asymmetry in the Universe/CP Violation?
  - Dark Matter?
  - Why Three Generations?
- *Want* to find disagreements in experiments (new physics?)
- Measuring asymmetries and comparing to SM predictions may lead us to discovering some of these answers

Background, Motivation, and Goals: Asymmetries at Colliders

- Measurements of asymmetries have long been studied at colliders (i.e. the Tevatron and the LHC)
- Can sensitively probe weak properties of particles (i.e. the effective weak mixing angle) through the collisions that take place

Drell-Yan Process in a Hadron Collision:



- In a Drell–Yan process:
  - quark/anti-quark from colliding hadrons annihilate
  - create either virtual γ or Z boson which decays into lepton/anti-lepton pair
- These processes can produce forward–backward asymmetry

Background, Motivation, and Goals: Asymmetries at Colliders

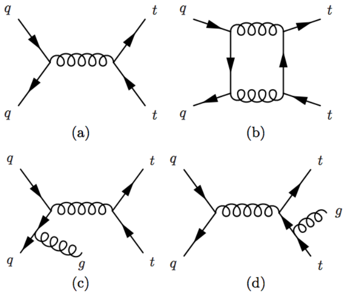


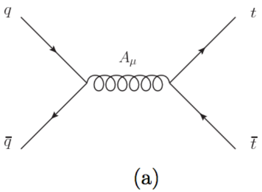
FIG. 1 Sample Feynman diagrams contributing to $q\bar{q} \rightarrow t\bar{t}$ at leading order (LO) (a) and NLO in QCD (b,c,d).

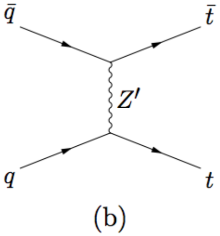
http://arxiv.org/pdf/1406.1798v3.pdf

- SM NLO predicts an asymmetry in the $t\bar{t}$ production A^{FB}
- Due to interference among diagrams, and large EW corrections and QCD corrections of order α_s^3 terms which are odd under the interchange of t and $\bar{t}$

Background, Motivation, and Goals: Asymmetries at Colliders

Example Feynman Diagrams of $t\bar{t}$ production via hypothetical BSM particles: axigluons (a), and Z' bosons (b).





- However, some measurements made were found to disagree significantly with SM NLO predictions, which indicated a smoking gun for possible new physics[a]

- BSM scenarios can help account for the observed asymmetry, for example axigluons (s-channel) and Z' bosons (t-channel)[b]

[a] J. H. Kuhn and G. Rodrigo, Phys. Rev. D 59 (1999) 054017

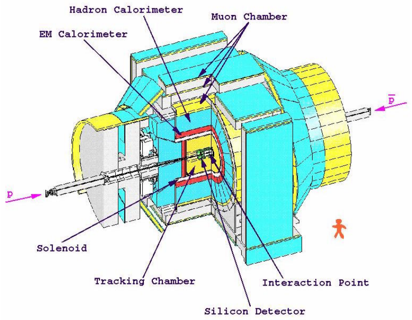
[b] Hong, Ziqing. "Forward–Backward Asymmetry of Top Quark Pair Tevatron." Thesis. Texas A&M University, 2015. Print.

Background, Motivation, and Goals: Measuring Asymmetries

- In general, we define an asymmetry with the partial cross sections, σ_1 and σ_2, over two complementary kinematic or geometric regions:

$$A \equiv \frac{\sigma_1 - \sigma_2}{\sigma_1 + \sigma_2}$$

CDF Detector:



- Most experimental techniques measure an asymmetry in a restricted region (due to geometric constraints of the detector), A^{visible}
- Therefore extrapolating to the inclusive (physical) asymmetry $A^{\text{inclusive}}$ is then necessary

Background, Motivation, and Goals: Extrapolating Asymmetries

- Various extrapolation techniques exist
  - Constant additive factor: $C = A^{\text{inclusive}} - A^{\text{visible}}$
  - Constant multiplicative factor: $R = A^{\text{visible}} / A^{\text{inclusive}}$
  - Matrix unfolding for when the situation is complicated

- Using a constant multiplicative factor, R can be advantageous for certain analyses
  - It is used at the Tevatron for $t\bar{t}$ leptonic asymmetry measurements[a]
  - Appears not to vary significantly with the absolute value of the expected inclusive asymmetry

[a] V. M. Abazov, et al., D0 Collaboration, Phys. Rev. D 88 (2013) 112002
T. Aaltonen, et al., CDF Collaboration, Phys. Rev. D 88 (2013) 072003
T. Aaltonen, et al., CDF Collaboration, Phys. Rev. Lett. 111 (2013) 182002
T. Aaltonen, et al., CDF Collaboration, Phys. Rev. Lett. 113 (2014) 042001
T. Aaltonen, et al., CDF Collaboration, Phys. Rev. D 89 (2014) 072001

Background, Motivation, and Goals: Extrapolating Asymmetries

- Monte Carlo (MC) simulation is typically used to estimate R
- Unfortunately only a single pseudo-experiment (due to the computationally expensive nature of the cross-sections) is typically run and used to make the estimation of R
- Understanding the uncertainty on that estimation is important and can lead to misleading results if not properly accounted for
- Since measurements that disagree with SM NLO predictions are of particular interest (as they can be a signal of new physics), guaranteeing measurement techniques are accurate and error measurements are correct is an important task

Background, Motivation, and Goals

- We expect to need higher statistics as the simulated asymmetry gets smaller, but we need to be able to quantify how much statistics we need for a given asymmetry to be confident in measurements

- This study aims to:
  - point out an important pitfall that analyzers can fall into when using this particular technique,
  - understand what causes the pitfall, and quantify how one can confidently avoid it

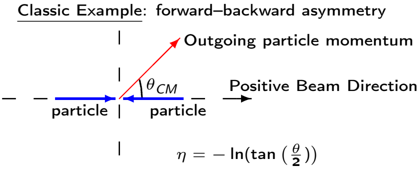
- Specifically, we want to understand two things:
  - Whether we can confidently and reliably use a constant R to perform the extrapolation, and
  - What the required MC sample size is to be able to reliably estimate R for a given asymmetry value

Monte Carlo Study

- To simplify the discussion we integrate over all variables except one, x, so we can define the visible and inclusive asymmetries as:

$$\sigma_1^{\text{visible}} = \int_0^{x^{\text{visible}}} dx \frac{d\sigma}{dx} \quad \text{and } \sigma_2^{\text{visible}} = \int_{-x^{\text{visible}}}^{0} dx \frac{d\sigma}{dx}$$

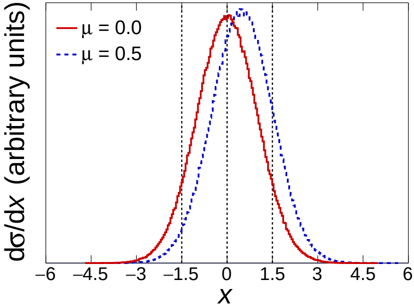
$$\sigma_1^{\text{inclusive}} = \int_0^{\infty} dx \frac{d\sigma}{dx} \quad \text{and } \sigma_2^{\text{inclusive}} = \int_{-\infty}^{0} dx \frac{d\sigma}{dx}$$



- For example, we can say we integrate over all variables except the pseudo-rapidity, η, of a particle, which gives rise to a forward–backward asymmetry

Monte Carlo Study

- We use a simple single Gaussian differential cross section model[2] with a mean, $\mu \propto A$, and unit width
- Below, we show a single pseudo-experiment (PE) for two different values of μ, each with number of events $N = 10^6$:

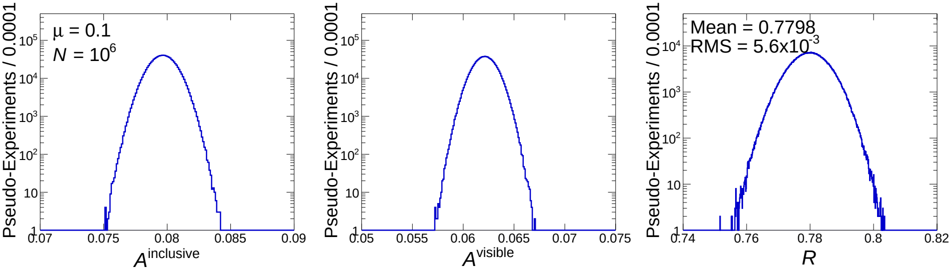


  - We choose $x^{\text{visible}} = 1.5$ which is close to typical values seen in $t\bar{t}$ measurements at the Tevatron
  - As a benchmark, $\mu = 0.1$ corresponds to $A^{\text{inclusive}} \approx 8\%$ which is also typically seen

[2] It has been shown that the leptonic differential cross section is well approximated as the sum of two Gaussians with a common mean, and the multiplicative extrapolation works in this case.
Z. Hong, R. Edgar, S. Henry, D. Toback, J. S. Wilson, and D. Amidei, Phys. Rev. D 90 (2014) 014040.

Monte Carlo Study

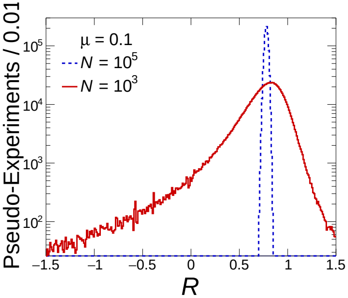
- Measure $A^{\text{inclusive}}$, A^{visible}, and R for each of a large number of PE's
- With many PEs (N_{PE}), we get distributions for $A^{\text{inclusive}}$, A^{visible}, and R:



- This looks like it should work well – R has a small RMS and looks very Gaussian
- But what happens to the R distribution as we vary N and μ?
- With large enough sample size, measurements of R are very accurate

Monte Carlo Study:
Pathological Case to Examine the Low Statistics Simulation

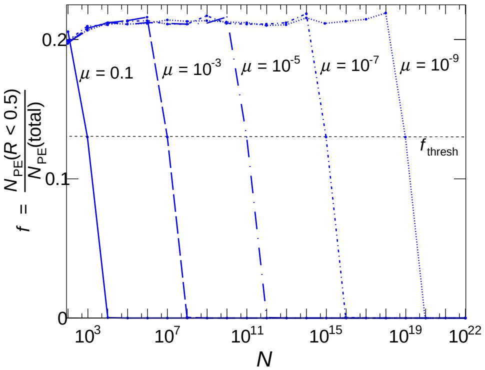
- We now study the R distributions that arise for a fixed value of μ, but with large and small values of N
- This corresponds to high statistics/reliable measurements and low statistics/unreliable measurements respectively
- As N decreases, measurement of R becomes unreliable, and may no longer correctly reproduce $A^{\text{inclusive}}$ from A^{visible}



- Blue data represents a reliable measurement of R with a well understood uncertainty
- Red data represents an unreliable/pathological measurement
- This transition is observed for all values of μ

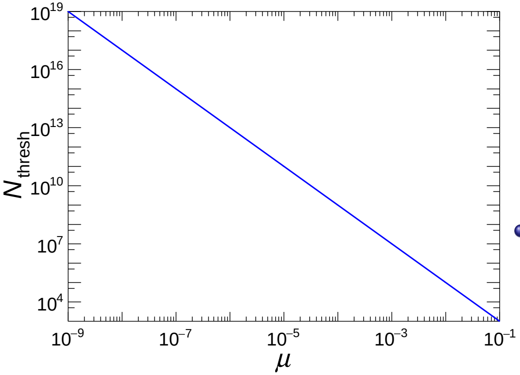
Monte Carlo Study: Quantifying the Transition for Varying μ

- How many events, N_{thresh}, are needed to give a reliable measurement of R?



- We define f as the fraction of pseudo-experiments with $R < 0.5$
- This *should be* many σ from the mean, so we require $f \approx 0$
- To examine/quantify the behavior for reliable measurements, we define a threshold value, f_{thresh}, and examine the relationship between N_{thresh} and μ

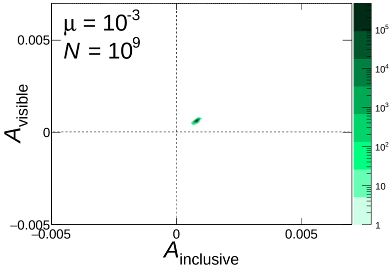
Monte Carlo Study: Results



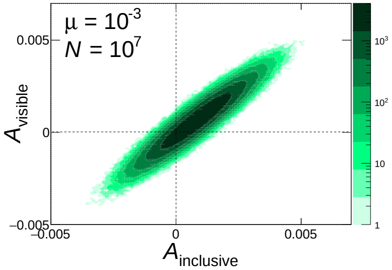
- We find that it can take a much-larger-than-expected sample size to reliably measure R, especially for very small μ (or equivalently A)
    - N_{thresh} rises as $\frac{1}{\mu^2}$ $\left(\text{or } \frac{1}{A^2}\right)$
- We also find that when N is large enough for reliable measurements, R is measured to be close to constant for all values of μ

Closed Form Statistical Validation: Examining Why MC Methods Break Down for Small N

"Enough" Events in Simulation – Reliable Measurements



"Not Enough" Events in Simulation – Unreliable Measurements



- Require $A^{inclusive}$ (denominator of R) to be greater than *at least* 1 σ away from 0

Closed Form Statistical Validation: Number of Events Required for Reliable Measurements

- We use statistics to determine how many events, N_{thresh}, are required for the mean value of $A^{\text{inclusive}}$ to be at least 1 σ away from 0
- In equation form, this condition can be written as:

$$A^{\text{inclusive}} \geq \sigma_{A^{\text{inclusive}}}$$

- We are able to find N_{thresh} as a function of μ for our single Gaussian model (calculation in backup slides):

$$N_{\text{thresh}} \geq 2 \cdot \frac{\left(1 + \text{erf}\left(\frac{\mu}{\sqrt{2}}\right)\right)}{\text{erf}\left(\frac{\mu}{\sqrt{2}}\right)^2}$$

- Some limiting cases:
  - As $\mu \to 0$, $N_{\text{thresh}} \to \infty$
  - $\text{erf}\left(\frac{\mu}{\sqrt{2}}\right) \approx \sqrt{\frac{2}{\pi}}\mu$ for small μ, so we find that $N_{\text{thresh}} \propto \frac{1}{\mu^2}$ which is precisely what we just saw in the MC study

Closed Form Numerical Validation: Is R constant?

- Examining the behavior of R as a function of μ analytically is straightforward for the single Gaussian model
  - We set $\sigma = 1.0$ and use the visible region $|x| < 1.5$
- For large values of μ (i.e 0.1), R rises by 0.04% relative to $R(\mu = 0)$

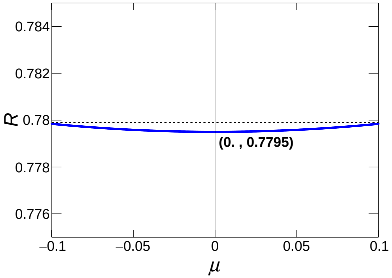
Recall:

$$A^{inclusive} = \frac{\sigma_1^{\text{inclusive}} - \sigma_2^{\text{inclusive}}}{\sigma_1^{\text{inclusive}} + \sigma_2^{\text{inclusive}}}$$

$$A^{visible} = \frac{\sigma_1^{\text{visible}} - \sigma_2^{\text{visible}}}{\sigma_1^{\text{visible}} + \sigma_2^{\text{visible}}}$$

$$R = \frac{A^{\text{visible}}}{A^{\text{inclusive}}}$$

where $\sigma \equiv \text{Gaus}(\mu, \sigma = 1.0)$



Conclusions

- We have studied the multiplicative extrapolation of A^{visible} to $A^{\text{inclusive}}$ for the single Gaussian model, and while a custom study would be needed for any non-Gaussian physics distribution, we have observed that a linear extrapolation can be used in this and other similar cases
- While MC methods work reliably (even for small A), they can require much larger sample sizes than expected, rising as $\frac{1}{A^2}$
- Our results have the potential to be applied to many different asymmetry measurements in collider experiments, and have already been useful at the Tevatron for the $t\bar{t}$ forward-backward asymmetry

Backups: The Cauchy Distribution[3]

- A distribution of the ratio of two independent Gaussian variables
- Mean and RMS are actually undefined; though mode and median are well defined
- $A^{\text{inclusive}}$ and A^{visible} are approximately Gaussian, thus as the mean of the $A^{\text{inclusive}}$ distribution approaches 0, R begins approximating a Cauchy distribution

[3] A. Papoulis, "Probability, Random Variables, and Stochastic Processes", 2nd ed., New York: McGraw-Hill, 1984

Backups: The Statistical Solution Calculation

We need enough statistics such that $A^{inclusive}$, the denominator of R, is more than 1 sigma away from 0 (we will set it to be k, where k will be determined later). In other words, we want to know how many events it takes in a pseudo-experiment to ensure the mean of the full asymmetry will be k standard-deviations away from zero.

To do this we start with the equation

$$\sigma_{A^{inclusive}} = \frac{A^{inclusive}}{k} \tag{1}$$

where $\sigma_{A^{inclusive}}$ is the variation (or uncertainty) of the measured value of $A^{inclusive}$. We will find both $\sigma_{A^{inclusive}}$ and $A^{inclusive}$ as functions of N and μ and substitute them into Eq. 1 to get the functional relation between N and μ for "good statistics".

Backups: The Statistical Solution Calculation

We begin with our definition of asymmetry, where $N_+ \equiv \sigma_1^{\text{inclusive}}$ and $N_- \equiv \sigma_2^{\text{inclusive}}$ as on Slide 5, and thus $N = N_+ + N_-$ is the total number of events in the original Gaussian distribution. Using this information:

$$A^{inclusive} = \frac{N_+ - N_-}{N_+ + N_-} = \frac{2N_+ - N}{N}. \tag{2}$$

We note that since our distributions are Gaussian, we can write N_+ in terms of N and μ, with the relation given by

$$\begin{aligned} N_+ &= \frac{N}{\sqrt{2\pi}} \int_0^\infty \mathrm{dx}\ e^{-(x-\mu)^2/2} \\ &= \frac{N}{2}\left(\mathrm{erf}\left(\frac{\mu}{\sqrt{2}}\right) + 1\right) \end{aligned} \tag{3}$$

Backups: The Statistical Solution Calculation

Plugging this in to Eq. 2 and reducing, we get

$$
\begin{aligned}
A^{inclusive} &= \frac{\cancel{2}\frac{\cancel{N}}{\cancel{2}}\left(\text{erf}\left(\frac{\mu}{\sqrt{2}}\right)+\cancel{1}\right)-\cancel{N}}{\cancel{N}} \\
&= \text{erf}\left(\frac{\mu}{\sqrt{2}}\right) \qquad (4)
\end{aligned}
$$

We next find $\sigma_{A^{inclusive}}$ by beginning with the definition given in Bevington (92) applied to our problem,

$$\sigma_{A^{inclusive}} = \left(\frac{\partial A^{inclusive}}{\partial N_+}\right)\sigma_{N_+} + \left(\frac{\partial A^{inclusive}}{\partial N}\right)\sigma_N. \qquad (5)$$

Taking a simple derivative of $A^{inclusive}$ from Eq. 2 gives us

$$\left(\frac{\partial A^{inclusive}}{\partial N_+}\right) = \frac{2}{N} \qquad (6)$$

Backups:
The Statistical Solution Calculation

To be consistent with the previous study, we fix N and allow N_+ to vary. This means that $\sigma_N = 0$, and from simple statistics

$$\sigma_{N_+} = \sqrt{N_+} \tag{7}$$

Plugging Eqs. 6 and 7 into Eq. 5, we get

$$\sigma_{A^{inclusive}} = \frac{2}{N} \cdot \sqrt{N_+}. \tag{8}$$

Plugging Eq. 3 into this, we get

$$\begin{aligned}\sigma_{A^{inclusive}} &= \frac{2}{N} \cdot \sqrt{\frac{N}{2}\left(\operatorname{erf}\left(\frac{\mu}{\sqrt{2}}\right) + 1\right)} \\ &= \sqrt{\frac{2}{N}} \cdot \sqrt{\left(1 + \operatorname{erf}\left(\frac{\mu}{\sqrt{2}}\right)\right)} \end{aligned} \tag{9}$$

Backups: The Statistical Solution Calculation

Finally, plugging Eqs. 4 and 9 back into Eq. 1 gives us

$$\sqrt{\frac{2}{N}} \cdot \sqrt{\left(1 + \mathrm{erf}\left(\frac{\mu}{\sqrt{2}}\right)\right)} = \frac{\mathrm{erf}\left(\frac{\mu}{\sqrt{2}}\right)}{k}, \tag{10}$$

and solving for N, we get

$$N = \frac{2k^2\left(1 + \mathrm{erf}\left(\frac{\mu}{\sqrt{2}}\right)\right)}{\mathrm{erf}\left(\frac{\mu}{\sqrt{2}}\right)^2} \tag{11}$$

This is, as we set out to solve for, the number of events it takes per pseudo-experiment to ensure the mean of the full asymmetry will be k standard-deviations away from zero, and thus give good statistics. Discussion of the implication of this result is included in the main slides.

Backups: Closed Form Numerical Solution Gaussian Functions

$$A^{inclusive} = \frac{\frac{1}{\sqrt{2\pi}\sigma}\int_0^{\infty} dx\left[\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) - \exp\left(-\frac{(-x-\mu)^2}{2\sigma^2}\right)\right]}{\frac{1}{\sqrt{2\pi}\sigma}\int_0^{\infty} dx\left[\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(-x-\mu)^2}{2\sigma^2}\right)\right]}$$

$$A^{visible} = \frac{\frac{1}{\sqrt{2\pi}\sigma}\int_0^{1.5} dx\left[\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) - \exp\left(-\frac{(-x-\mu)^2}{2\sigma^2}\right)\right]}{\frac{1}{\sqrt{2\pi}\sigma}\int_0^{1.5} dx\left[\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(-x-\mu)^2}{2\sigma^2}\right)\right]}$$

$$R = \frac{A^{visible}}{A^{inclusive}}$$