S1 Coursework Comparison of Statistical Methods: Multi-Dimensional Likelihood Fitting vs. sWeights

ps2012

December 16, 2024

Abstract

This report compares the statistical power of two approaches for analyzing multi-dimensional probability distributions: a multi-dimensional likelihood fit and a weighted fit utilizing sWeights. Through mathematical derivations, numerical simulations, and parametric bootstrapping, we explore their advantages, drawbacks, and applications. All code and analysis are documented and reproducible.

Contents

1 Introduction

The goal of this study is to assess the performance of two statistical methods: the multi-dimensional likelihood fit and the weighted fit exploiting sWeights. This involves mathematical proofs, visualization of distributions, parameter estimation, and a thorough comparison of the two approaches under various conditions. The Crystal Ball distribution serves as a central component of the analysis, modeling the signal probability density function (PDF).

2 Mathematical Proof of Normalization Constant N

We prove that the normalization constant N^{-1} for the Crystal Ball PDF can be written as:

$$N^{-1} = \sigma \left[\frac{m}{\beta(m-1)} e^{-\beta^2/2} + \sqrt{2\pi}\Phi(\beta) \right]. \tag{1}$$

The derivation includes step-by-step justification, utilizing the properties of Gaussian and power-law tails.

3 Definition and Normalization of PDFs

This section describes the signal and background models:

- Signal PDF in X: Crystal Ball distribution, truncated in $[0, 5]$.
- Signal PDF in Y: Exponential decay with parameter λ.
- Background PDFs: Uniform in X and truncated Gaussian in Y.

Normalization of these PDFs is verified numerically through integration.

4 Visualization of Distributions

Figures in this section include:

- 1D projections of signal, background, and total PDFs in X and Y.
- A 2D plot of the joint probability density function.

These visualizations confirm the correctness of the models.

5 Sampling and Maximum Likelihood Fitting

5.1 Sampling from the Joint PDF

A high-statistics sample of $100,000$ events is generated using the defined PDFs. Numerical methods and the 'timeit' library are used to measure execution time for:

- Sampling $100,000$ events.
- Performing extended maximum likelihood fits.

5.2 Parameter Estimation

The extended maximum likelihood fit estimates the nine parameters of the model, including uncertainties.

6 Parametric Bootstrapping

Using parametric bootstrapping, we:

- Generate ensembles of samples of varying sizes (500, 1000, 2500, 5000, 10000).
- Assess bias and uncertainty in the decay constant λ as a function of sample size.

The results highlight trends in statistical precision and potential biases.

7 sWeights Analysis

The sWeights method is applied by fitting only the X variable and projecting the signal density in Y. Weighted samples are used to estimate λ. The bias and uncertainty are compared with those from parametric bootstrapping.

8 Comparison of Methods

8.1 Advantages and Drawbacks

- **Multi-dimensional likelihood fit**: High statistical power but computationally intensive.
- **sWeights**: Efficient for partial models but prone to biases if weights are not accurately estimated.

8.2 Preferred Scenarios

Recommendations are made for when each method is appropriate, based on the complexity of the problem and computational constraints.

9 Conclusion

This study demonstrates the trade-offs between multi-dimensional likelihood fitting and sWeights. While the former offers higher precision, the latter can be advantageous in high-dimensional problems with limited computational resources. Future work could explore hybrid approaches to leverage the strengths of both methods.

References

- Lecture notes and references cited in the coursework.
- Relevant documentation for libraries used (e.g., scipy, iminuit).