Boosting C statistics in Astronomy: High-order Asymptotics for Goodness-of-Fit Test

Xiaoli Li [1] Yang Chen [2] Xiaoli-Li Meng [3] David A. van Dyk [4] Massimiliano Bonamente [5] Vinay L. Kashyap [6]

[1]The University of Chicago [2]University of Michigan, Ann Arbor [3]Harvard University [4]Imperial College London [5]University of Alabama in Huntsville [6]Center for Astrophysics | Harvard & Smithsonian

Introduction

- The **C statistic** is a likelihood ratio statistic widely used for goodness-of-fit assessments in high-energy (astro)physics with **Poisson count data**. Although the C statistic enjoys convenient theoretical properties in certain cases, it is **routinely applied based on unwarranted assumptions**, giving misleading findings.
- In this project, we provide a suite of **new principled user-friendly methods** for computing **well-calibrated** p**-values** and are ready for immediate deployment in the astrophysics data analysis pipeline with practical guidance.
- We present a **comprehensive study** of the theoretical properties of C statistics and evaluate various related goodness-of-fit algorithms, emphasizing **low-count scenarios**. The **superiority of our method** is substantiated with both theoretical and numerical results.

Background and Motivation

Problem Setup

- Model:

$$N_i \mid \boldsymbol{\theta} \overset{\text{indep.}}{\sim} \text{Poisson}(s_i(\boldsymbol{\theta})) \quad \text{for} \quad i = 1, \ldots, n,$$

$$s_i(\boldsymbol{\theta}) = \sum_{j=1}^{J} R(\tilde{E}_j, i) A(\tilde{E}_j) g(\tilde{E}_j, \boldsymbol{\theta})[E_{j+1} - E_j] + B_i,$$

where $\boldsymbol{\theta}$ is the parameter and g is a smooth function of $\boldsymbol{\theta}$, e.g. **Powerlaw** model: $g(E, \boldsymbol{\theta}) = K \cdot e^{-N_H \cdot \sigma(E)} \cdot E^{-\Gamma}$ with $\boldsymbol{\theta} = \{K, \Gamma\}$.

- Hypothesis:

$$H_0 : N_i \mid \boldsymbol{\theta} \overset{\text{indep.}}{\sim} \text{Poisson}(s_i(\boldsymbol{\theta})) \text{ with } \boldsymbol{\theta} \in \mathbb{R}^d,$$

$$H_1 : N_i \overset{\text{indep.}}{\sim} \text{Poisson}(s_i) \text{ with } (s_1, \ldots, s_n) \in \mathbb{R}_+^n.$$

- C Statistics as a Likelihood Ratio Statistics:

$$C_n(\hat{\boldsymbol{\theta}}) = 2 \sum_{i=1}^{n} \left[s_i(\hat{\boldsymbol{\theta}}) - N_i \log s_i(\hat{\boldsymbol{\theta}}) - N_i + N_i \log N_i \right].$$

Motivation and Methodology

1. Based on a **flawed application of Wilks' theorem**, researchers have been using likelihood ratio test with χ^2 approximation (**LR-χ^2 test**) for decades. It is **commonly adopted** by numerous scientific platforms and packages, including the current standard default astrophysical data processing package `sherpa`.
2. A naive method is to use the **marginal asymptotic normality** of C statistics with **uncorrected moments** which calculated by assuming $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}$ (**Naive Plug-in Z-test**). The moments can be computed based on bootstrap estimation, empirical polynomial approximation proposed by astronomers or high-order approximation for the unconditional moments given by us.
3. Motivated by Wilks' Theorem and the fact that C statistics is a Likelihood Ratio Statistics, we derived the **conditional asymptotic normality** of C statistics and gave **computable high-order approximation** for the conditional moments. Based on these results, we propose a original new test named **Corrected Z-test**.

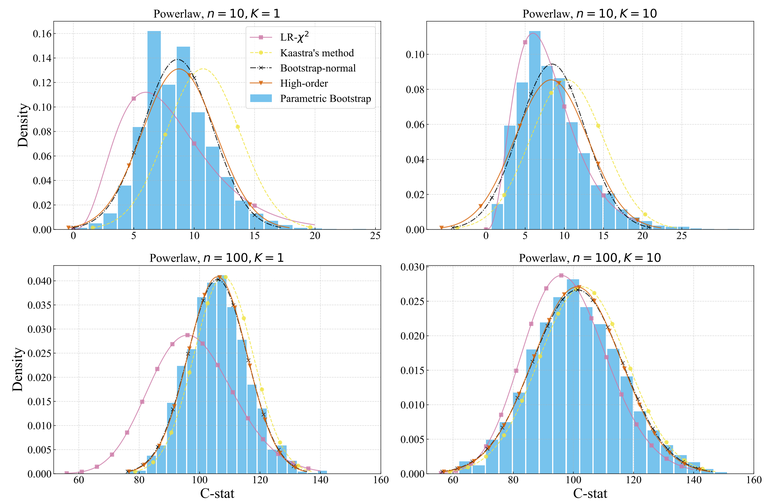

Figure 1. Histograms of null distributions of $C_n(\hat{\boldsymbol{\theta}})$ given by different algorithms in Powerlaw model with $\Gamma = 1$ fixed.

Theoretical Guarantee

1) Under H_0 and mild conditions, as $n \to \infty$, conditional on $\hat{\boldsymbol{\theta}}$ we have

$$\frac{C_n(\hat{\boldsymbol{\theta}}) - \mathbb{E}[C_n(\hat{\boldsymbol{\theta}}) \mid \hat{\boldsymbol{\theta}}]}{\sqrt{\text{Var}[C_n(\hat{\boldsymbol{\theta}}) \mid \hat{\boldsymbol{\theta}}]}} \xrightarrow{D} N(0, 1).$$

2) Under H_0 and mild conditions, we have

$$\mathbb{E}(C_n(\hat{\boldsymbol{\theta}}) \mid \hat{\boldsymbol{\theta}}) = \hat{\kappa}_1^{(\cdot)} - \frac{1}{2} 1^\top \hat{X}^\top \hat{\Sigma} \hat{X} (\hat{X}^\top \hat{W} \hat{X})^{-1} 1 + O(n^{-1/2}),$$

and

$$\text{Var}(C_n(\hat{\boldsymbol{\theta}}) \mid \hat{\boldsymbol{\theta}}) = \hat{\kappa}_2^{(\cdot)} - \hat{\kappa}_{11}^\top \hat{X} (\hat{X}^\top \hat{W} \hat{X})^{-1} \hat{X}^\top \hat{\kappa}_{11} + O(1).$$

Key Findings

- **Dense Data** ($n > 10, s_i > 1$): Corrected Z-test and Parametric Bootstrap work well. In contrast, regardless of n, the p-value given by the LR-χ^2 test is far from uniform distribution unless s_i is uniformly sufficiently large.
- **Extensive but Sparse Data** ($\sum_{i=1}^n s_i > 10$, $n > 10$ but $s_i \leq 1$): The p-value given by Corrected Z-test is approximately uniformly distributed, while the LR-χ^2 test and the bootstrap test fail in this case.
- **Scarce Data** ($\sum_{i=1}^n s_i \leq 10$): Other more robust methods, such as calibrated double bootstrap, are expected to be applied, though due to the scarcity of data, no valid test is expected to enjoy high power in this setting.
- **Computational Cost**: The Corrected Z-test is far more efficient than bootstrap methods, including robust variants such as the double bootstrap.

Numerical Experiments

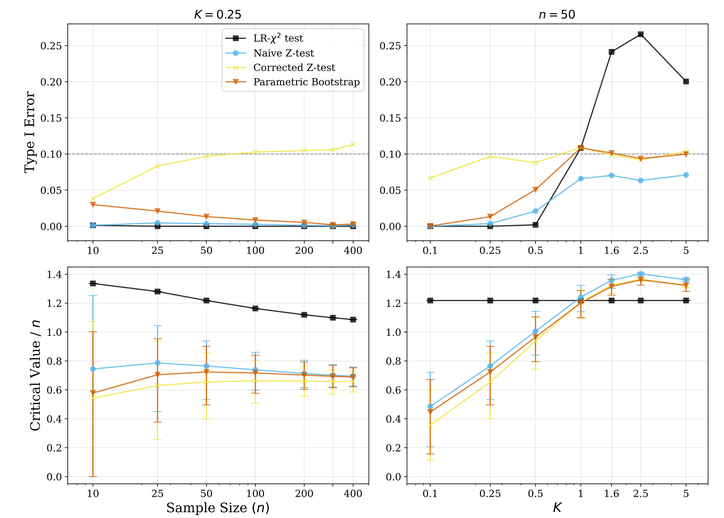

Figure 2. Average performance of four methods when n and K varies. The true models and null models are the Powerlaw model with $\Gamma = 1$ fixed. The dashed line in the first row is the nominal Type I error rate. Tests with Type I error rates close to the nominal rate and small critical values are preferred. The simulation shows the overall strong performance of the Corrected Z-test.

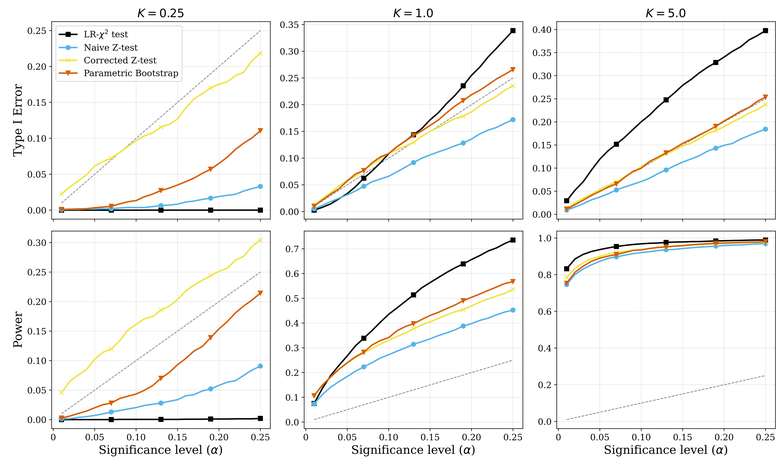

Figure 3. Comparisons of Type I Errors and Powers under different significance levels α. The null models are the Powerlaw model with $n = 50$ and $\Gamma = 1$ fixed. And the true models are the Spectral-Line model. Ideally, the power should be as large as possible while maintaining the Type I Error below α. Overall, the Corrected Z-test is best calibrated in terms of Type I error rate and power.

Contact: xiaolil@uchicago.edu