On smooth tests of goodness-of-fit for astrophysical searches under high background

Presenter: Xiangyu Zhang, University of Minnesota

with Sara Algeri, Vinay Kashyap, Margarita Karovska Neily, Charlie Geyer

February 21, 2024

Goodness-of-Fit problem

- Let X be a continuous random variable with support $\mathcal{X} \subseteq \mathbb{R}$ with unknown distribution function P and density p. The goodness-of-fit (GOF) problem aim to assess if P belongs to a family of continuous distribution functions $G_{\boldsymbol{\beta}}$, with PDF $g_{\boldsymbol{\beta}}$, where $\boldsymbol{\beta} \in B \subseteq R^p$. Formally, this corresponds to the hypothesis

$$H_0 : p = g_{\boldsymbol{\beta}}, \text{ for some } \boldsymbol{\beta} \in B \quad \text{versus} \quad H_1 : p \neq g_{\boldsymbol{\beta}}, \text{ for all } \boldsymbol{\beta} \in B. \tag{1}$$

- Smooth tests also have solutions for GOF problem when X is discrete or $\mathcal{X} \subseteq \mathbb{R}^d$. To reduce the technicality of the talk, we will not focus on them today.

Orthonormal expansions of the density ratio

- Assuming that P is absolutely continuous with respect to $G_{\boldsymbol{\beta}}$ ($P \ll G_{\boldsymbol{\beta}}$). If the density ratio $p/g_{\boldsymbol{\beta}} \in L^2(\mathcal{X}, G_{\boldsymbol{\beta}})$, then it can be expanded via a series of orthonormal basis functions $\{\psi_j\}_{j\in\mathbb{N}} \in L^2([0,1], G_{\boldsymbol{\beta}})$

$$\frac{p(x)}{g_{\boldsymbol{\beta}}(x)} = \theta_{0\boldsymbol{\beta}} + \sum_{j=1}^{\infty} \theta_{j\boldsymbol{\beta}} \psi_j\Big(G_{\boldsymbol{\beta}}(x)\Big) = 1 + \sum_{j=1}^{\infty} \theta_{j\boldsymbol{\beta}} h_{j\boldsymbol{\beta}}(x), \quad \text{for all } x \in \mathcal{X}, \tag{2}$$

where $\theta_{0\boldsymbol{\beta}} = 1$ and we denote $\psi_j(G_{\boldsymbol{\beta}}(x))$ as $h_{j\boldsymbol{\beta}}(x)$. The coefficients satisfy

$$\theta_{j\boldsymbol{\beta}} = \int_{\mathcal{X}} h_{j\boldsymbol{\beta}}(x) \frac{p(x)}{g_{\boldsymbol{\beta}}(x)} dG_{\boldsymbol{\beta}}(x) = \int_{\mathcal{X}} h_{j\boldsymbol{\beta}}(x) dP(x). \tag{3}$$

Smooth Models and Smooth Tests

- Given a point of truncation m, a *smooth model* for the true density $p(x)$ is

$$p_m(x) = g_{\beta}(x)\left[1 + \sum_{j=1}^{m} \theta_{j\beta} h_{j\beta}(x)\right], \tag{4}$$

  where the last term is the truncation of the expansion in (2) at order m.

- By the smooth model, we have the test

$$\begin{aligned} &H_0 : \text{for some } \boldsymbol{\beta} \in B,\ \theta_{1\boldsymbol{\beta}} = \ldots = \theta_{m\boldsymbol{\beta}} = 0 \quad \text{versus} \\ &H_1 : \text{for all } \boldsymbol{\beta} \in B, \text{ there exists at least one } j, \text{ such that } \theta_{j\boldsymbol{\beta}} \neq 0. \end{aligned} \tag{5}$$

  and this is commonly referred to as the "smooth test".

- Compared with the classical GOF tests, smooth tests concentrate the power towards a finite number of possible directions specified by $\{h_{1\boldsymbol{\beta}}, \ldots, h_{m\boldsymbol{\beta}}\}$.

Smooth estimator and test statistics

- A particularly appealing feature of smooth tests is that, when the null model is rejected, they naturally correct for it. This correction is called the *smooth estimator*

$$\widehat{p}_m(x) = g_\beta(x)\left[1 + \sum_{j=1}^{m} \widehat{\theta}_{j\beta} h_{j\beta}(x)\right], \tag{6}$$

where

$$\widehat{\theta}_{j\beta} = \int_{\mathcal{X}} h_{j\beta}(x) dP_n(x) = \frac{1}{n}\sum_{i=1}^{n} h_{j\beta}(x_i). \tag{7}$$

where $P_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$ and $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

- Assuming m is given and β is known, the score statistics can be written as

$$S_m = n\sum_{j=1}^{m} \widehat{\theta}^2_{j\beta}.$$

Under H_0 and as $n \to \infty$, $S_m \xrightarrow{d} \chi^2_m$.

Real data example: RT Cru

- **Background:** RT Cru is a symbiotic system where a high-mass white dwarf accretes from the wind of an M5 III red giant companion. RT Cru exhibits variability features like aperiodic flickering at timescales of a few ks, and a strong correlation of spectral intensities with overall brightness. The question that arises then is what the origin of this variability could be.

- Based on the observed X-ray spectrum produced by RT Cru, the origin can be modeled as an intrinsic change in the soft thermal emission component as well as changes in a continuum component due to intervening absorption. The presence of spectral lines during increases in soft flux, especially if they are the dominant contributors to soft emission, would support the former scenario, while the lack of such lines would favor the latter scenario.

- **Problem:** The Chandra/LETGS+HRC-S data was originally obtained to settle this question, but the analysis was limited because of the relatively high instrument background encountered.

Real data example: RT Cru (cont.)

- **Solution (First-step):** Perform the smooth tests to the background model using the background-only data. Use the smooth estimator as the corrected background distribution if rejected.

- **Solution (Second-step):** Perform the smooth tests to the corrected background model using the data containing potential emission lines. If the tests get rejected, we claim the existence of such emission lines, otherwise, we set upper limits on the intensity of the expected signals.

Efficient score functions

- Let $\boldsymbol{u}_{\boldsymbol{\beta}}$ be the score function of the postulated distribution $G_{\boldsymbol{\beta}}$ and let $\Gamma_{\boldsymbol{\beta}}$ be the Fisher information matrix. Then, the normalized score function $\boldsymbol{b}_{\boldsymbol{\beta}}$ is

$$\boldsymbol{b}_{\boldsymbol{\beta}}(x) = \Gamma_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{u}_{\boldsymbol{\beta}}(x) = \left[b_{\boldsymbol{\beta}_1}(x), \ldots, b_{\boldsymbol{\beta}_p}(x)\right]^T, \quad \text{for all } x \in \mathcal{X}. \tag{8}$$

- Define the efficient score functions, $\{\widetilde{h}_{j\boldsymbol{\beta}}\}_{j=1}^m$, as

$$\widetilde{h}_{j\boldsymbol{\beta}}(x) = h_{j\boldsymbol{\beta}}(x) - \sum_{k=1}^{p} \langle h_{j\boldsymbol{\beta}}, b_{\beta_k}\rangle_{G_{\boldsymbol{\beta}}} b_{\beta_k}(x), \qquad j = 1, \ldots, m. \tag{9}$$

for all $x \in \mathcal{X}$, where $\langle h_{j\boldsymbol{\beta}}, b_{\beta_k}\rangle_{G_{\boldsymbol{\beta}}} = \int_{\mathcal{X}} h_{j\boldsymbol{\beta}}(x) b_{\beta_k}(x) dG_{\boldsymbol{\beta}}(x)$.

Generalized score statistic

- Suppose $\widehat{\boldsymbol{\beta}}_n$ be the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ and define $\widehat{\boldsymbol{V}}$ as

$$\widehat{\boldsymbol{V}} = [v_{G,n}(\widetilde{h}_{1\widehat{\boldsymbol{\beta}}_n}), ..., v_{G,n}(\widetilde{h}_{m\widehat{\boldsymbol{\beta}}_n})]^T, \tag{10}$$

where $v_{G,n}(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}(x_i)$. The generalized score statistic is:

$$T_m = \widehat{\boldsymbol{V}}^T \Sigma_{\widehat{\boldsymbol{V}}}^{-} \widehat{\boldsymbol{V}}, \tag{11}$$

where $\Sigma_{\widehat{\boldsymbol{V}}}$ is the covariance matrix of $\widehat{\boldsymbol{V}}$, with elements $(\Sigma_{\widehat{V}})_{ij} = \left\langle \widetilde{h}_{i\widehat{\boldsymbol{\beta}}_n}, \widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n} \right\rangle_{G_{\widehat{\boldsymbol{\beta}}_n}}$.

- Under H_0 and as $n \to \infty$, $T_m \xrightarrow{d} \chi_r^2$, where r is the rank of $\Sigma_{\widehat{\boldsymbol{V}}}$.
- This result can be extened for any $\sqrt{n}$-consistent estimator $(\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = O_p(1))$.

Dara-driven order selection

- The power of the test depends on how well the true distribution is approximated by the m-dimensional smooth model.
- The determination of the m can be seen as a model selection problem to find the best nonparametric density estimates of the true distribution. For instance, Kallenberg and Ledwina (1997) propose the following BIC-type selection criteria

i. As the first step, choose a suitably large value M (usually 10).

ii. Then, obtain the MLE $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ and calculate $v_{G,n}(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n})$ for all $j = 1, ..., M$.

iii. Finally, choose the smallest m that maximizes

$$BIC(m) = \sum_{j=1}^{m} v_{G,n}^2(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}) - m \log n. \tag{12}$$

Post-selection Inferences

- Traditional inference is typically constructed assuming the model under study has been selected independently from the data. However, the order selection is data-driven. The limiting distributions of test statistics are strongly affected by the additional source of variability associated with the selection process.
- We may consider the data splitting or suitable post-selection adjustments for the p-values. But those either need extra sample sizes or are conservative.
- Bootstrap allows for the selection process to be repeated for each bootstrap replicate, which appropriately accounts for the randomness associated with the selection process.

Limitations of classical data-driven smooth tests

- Even without the order selection, the convergence of the generalized score statistic T_m to its limiting distribution is slow. For instance, Klar (2000) demonstrated that sample sizes as large as $10,000$ are required to in testing normal distribution to achieve a satisfactory approximation of the asymptotic null distribution.

- In practice, p-values and critical values are recommended to be determined using parametric bootstrap procedures (Thas, 2010). However, parametric bootstrap procedures can be computationally inefficient due to the complexity involved in
  1. samplings from the postulated distributions,
  2. performing likelihood or score function evaluations,
  3. estimating the parameters and test statistics,

  which makes the procedures infeasible.

Projected parametric bootstrap

- Repetitive estimations of unknown parameter $\boldsymbol{\beta}$ within each bootstrap replicate can be costly. Moreover, the generalized score statistic requires the re-estimation of the efficient score functions, $\{\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}\}_{j=1}^m$, and their covariance matrix.
- Suppose the parametric bootstrap samples from $G_{\widehat{\boldsymbol{\beta}}_n}$ are denoted as $x_{1,boot}, \ldots,$ $x_{n,boot}$, and let $\widehat{\boldsymbol{\beta}}_{boot}$ be the parameter estimated based on the bootstrap samples.
- We have proven

$$v_{G,n}^{boot}(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_{boot}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{h}_{j\widehat{\boldsymbol{\beta}}_{boot}}(x_{i,boot}), \tag{13}$$

  is asymptotically the same as the ones that we use $\widehat{\boldsymbol{\beta}}_n$ instead of $\widehat{\boldsymbol{\beta}}_{boot}$, i.e.,

$$v_{G,n}^{boot}(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}(x_{i,boot}). \tag{14}$$

- Therefore, test statistics that are continuous functionals of them also have the same limiting distribution.

Theorem

Suppose there exists a neighborhood N of $\boldsymbol{\beta}^*$, such that

1. *for almost all $x \in \mathcal{X}$ (w.r.t. the probability measure $G_{\boldsymbol{\beta}^*}$), the function $\boldsymbol{\beta} \mapsto \widetilde{h}_{j\boldsymbol{\beta}}(x)$ is continuously differentiable in the neighborhood of $\boldsymbol{\beta}^*$, for all $j = 1, \ldots, M$, with its gradient denoted as $\nabla_{\boldsymbol{\beta}} \widetilde{h}_{j\boldsymbol{\beta}}(x)$;*
2. *the components of the functions $\nabla_{\boldsymbol{\beta}} \widetilde{h}_{j\boldsymbol{\beta}}(x)$ for all $\boldsymbol{\beta} \in N$ are bounded by a $G_{\boldsymbol{\beta}^*}$-integrable function $M(x)$ for almost all $x \in \mathcal{X}$;*

then, for each deterministic sequence $\delta_n = O(n^{-1/2})$,

$$v_{G,n}(\widetilde{h}_{j\boldsymbol{\beta}}) = v_{G,n}(\widetilde{h}_{j\boldsymbol{\beta}^*}) + R_n(\boldsymbol{\beta}), \quad \text{where} \sup_{\boldsymbol{\beta} \in N : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \delta_n} R_n(\boldsymbol{\beta}) \xrightarrow{p} 0. \tag{15}$$

Corollary

Assume that the regularity conditions of Theorem 4.1 of Babu and Rao (2004) and the assumptions of the Theorem above are satisfied. Then,

$$v_{G,n}^{boot}(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}) = v_{G,n}^{boot}(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_{boot}}) + o_p(1) = v_{G,n}(\widetilde{h}_{j\widehat{\boldsymbol{\beta}}_n}) + o_p(1). \tag{16}$$

Motivation and main idea of the K-2 transformation

- **Motivation:** Not only be inefficient in re-estimation of MLEs and test statistics, but also difficult to generate samples, or cannot easily evaluate its likelihood or score functions.
- **Main idea of the K-2 transformation:** produce new K-2 transformed test statistics whose limiting distribution under the complicated postulated distribution is the same as some statistics under a simple reference distribution.
- Moreover, this method achieves asymptotically distribution-freeness, thus requiring only a single simulation from the reference distribution when testing for various hypothesized distributions.

Extensions of current work

- **Extension to binned data:** Many real-world problems in physics and astronomy depend on binned data. In my future work, I will also extend all the methods described to address the binned data regime. The modeling framework will incorporate the current work by Algeri S. and Khmaladze E.V..

Main references

- Algeri, S. (2020). Detecting new signals under background mismodeling. Physical Review D, 101(1), 015003.
- Algeri, S. (2022). K-2 rotated goodness-of-fit for multivariate data. Physical Review D, 105(3), 035030.
- Algeri, S., & Zhang, X. (2022). Exhaustive Goodness of Fit Via Smoothed Inference and Graphics. Journal of Computational and Graphical Statistics, 31(2), 378-389.
- Khmaladze, E. (2016). Unitary transformations, empirical processes and distribution free testing. Bernoulli, 22(1), 563-588.
- Zhang, X., Algeri, S., Kashyap, V., & Karovska, M. (2023). A novel approach to detect line emission under high background in high-resolution X-ray spectra. Monthly Notices of the Royal Astronomical Society, 521(1), 969-983.
- Zhang, X., Algeri, S., Geyer, C. J. (2024+). Asymptotically distribution-free smooth tests without χ^2.

Other references

- Chen, C.-F., Hart, J. D., and Wang, S. (2001). Bootstrapping the order selection test. Journal of Nonparametric Statistics, 13(6):851–882.
- J. Neyman. (1937). Smooth test for goodness of fit, Scandinavian Actuarial Journal, 1937:3-4, 149-199.
- Inglot, T. and Ledwina, T. (1996). Asymptotic optimality of data-driven Neyman's tests. Ann. Statist. 24 1982–2019.
- Inglot, T., Kallenberg, W., Ledwina, T. (1997). Data driven smooth tests for composite hypotheses. Ann. Statist. 25 (3) 1222 - 1250.
- Kallenberg, W. C. M. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. Journal of the American Statistical Association, 92(439):1094–1104.
- Klar, B. (2000). Diagnostic smooth tests of fit. Metrika, 52(3):237–252.
- Mukhopadhyay, S. (2017). Large-scale mode identification and data-driven sciences. Electronic Journal of Statistics, 11(1):215 – 240.
- Thas, O. (2010). Comparing distributions, volume 233. Springer.

Thanks!

Motivation of Smooth Tests From a Theoretical Perspective

- Neuhaus (1976) and Milbrodt (1990) shows that
  - only very few of deviations from g_{β} with the KS, CVM, AD statistics are of reasonable local asymptotic power[1].
  - only one direction with the highest local asymptotic power.
  - "good" directions correspond to very smooth departures from the postulated model.

- Smooth test considers exatly the smooth departures from the postulated model.

[1] Local asymptotic power: the power of a test under a sequence of distributions in the alternative hypothesis, as a way of approximating the finite-sample power function of a test

Efficient Score Functions

- Apply the Taylor expansion at $\boldsymbol{\beta}$ around $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta} = 0$, we eventually arrive at

$$\frac{p(x)}{g_{\boldsymbol{\beta}}(x)} = \frac{p(x)}{g_{\boldsymbol{\beta}^*}(x)} + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^t \boldsymbol{u}_{\boldsymbol{\beta}}(x) + \boldsymbol{\theta}^T \boldsymbol{h}_{\boldsymbol{\beta}}(x).$$

  This approximation demonstrates that the comparison density lives in a subspace which is spanned by the m-dimensional $\boldsymbol{h}$, but also by the score functions $\boldsymbol{u}_{\boldsymbol{\beta}}$ of the nuisance parameter $\boldsymbol{\beta}$, and the latter actually spans the d-dimensional subspace of comparison densities that are consistent with the null hypothesis.
- Not all of the spanned m-dimensional subspace is relevant for the alternative. It is therefore more efficient to transform h so that it spans a m-dimensional subspace that is exclusively relevant for the alternative,

Order selection test

- The so-called order selection test employs a test statistic that directly involves order selection, thereby removing the need to handle order selection and the associated post-selection inferences issues. In the context of smooth tests, Aerts et al. (1999) introduced the order selection test statistic as

$$\widetilde{T}_m = \max_{1 \le m \le M} \frac{T_m}{m}, \tag{17}$$

  where T_m is the generalized score test statistic.

- Under certain regularity conditions, the asymptotic null distribution T of the test statistic $\widetilde{T}_m$ is given by

$$P(T \le x) = \exp\left[-\sum_{s=1}^{\infty} \frac{P\left(\chi_s^2 > sx\right)}{s}\right]. \tag{18}$$