Handling model uncertainties via informative Goodness-of-Fit

Sara Algeri

School of Statistics, University of Minnesota

Statistical Methods for the Detection, Classification, and Inference of Relativistic Objects,

ICERM Virtual workshop,
November 17, 2020.

How significant is our astrophysical result?

- **Pearson χ^2:** particularly useful when a model for the signal is not known.
- **Likelihood Ratio Test (LRT):** particularly powerful when the signal model is known.

Main Limitations:

- They require that the models for the background and/or the signal are correctly specified.
- When their aren't, they do not really us what's wrong with it nor they say how to fix it...

... basically, they are of no help in learning from our mistakes!

Goal

How can we learn from our mistakes?

The goal of this talk is to propose a (not so new) approach to goodness-of-fit that aims to address precisely this question.

Interestingly, all we need are just two ingredients...

1. A comparison density
2. A smooth model

To warm up a little bit, let's start with the 1D case...

Main reference: *Algeri S. (2020), Physical Reviews D*

The comparison density

Introduced by Parzen in 1979...

Given a random variable X, let F and f be its cdf and pdf respectively. Let G a suitable cdf and let g be the respective pdf. Then, the comparison density between F and G is given by

$$d(u; G, F) = \frac{f\left(G^{-1}(u)\right)}{g\left(G^{-1}(u)\right)} \qquad \text{with } u = G(x), \tag{1}$$

or equivalently

$$d(G(x); G, F) = \frac{f(x)}{g(x)}. \tag{2}$$

We assume that $f = 0$ whenever $g = 0$.

It follows that...

$$f(x) = g(x)\; d(G(x); G, F) \tag{3}$$

Smooth models

Introduced by Neyman in 1937...

If we represent the comparison density through a series of $T_j(x;G)$ orthonormal functions of $G(x)$ then

$$f(x) = g(x) \underbrace{\left\{1 + \sum_{j>0} \theta_j \, T_j(x;G)\right\}}_{d(G(x);G,F)} \tag{4}$$

If we truncate the series at m and we estimate our θ_js...

$$\widehat{f}(x) = g(x)\left\{1 + \sum_{j=1}^{m} \widehat{\theta}_j \, T_j(x;G)\right\} \Rightarrow \textbf{we have estimated a smooth model!} \tag{5}$$

See *Algeri (2020+)* for convenient choices of $T_j(x;G)$.

Estimation

For the moment, let's assume m is fixed to a given number (typically < 10),

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} T_j(x_i; G),$$

for $j = 1, \ldots, m$ and thus

$$\widehat{d}(G(x); G, F) = 1 + \sum_{j=1}^{m} \widehat{\theta}_j\, T_j(x; G)$$

$$\widehat{f}(x) = g(x)\, \widehat{d}(G(x); G, F)$$

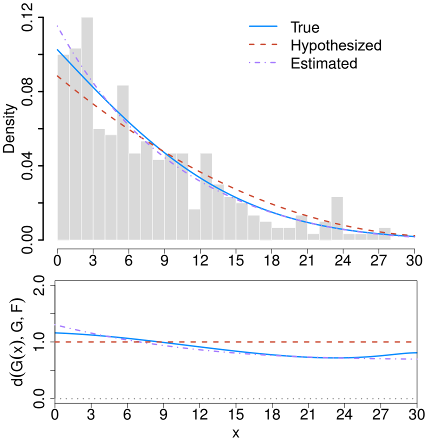


Figure 1. In the upper panel, f is the pdf of $X \sim \text{Normal}_{[0,30]}(-15, 15)$. In this case, $n = 300$ and g is a polynomial pdf. The respective comparison densities are shown in the bottom panels. In both examples $m = 2$.

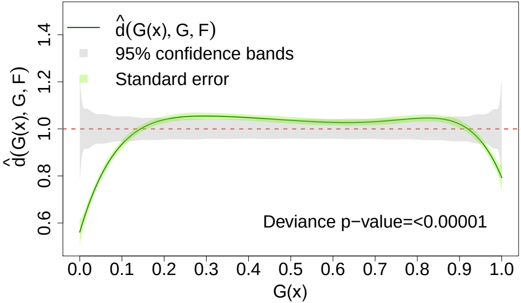
Inference

$$H_0 : \text{d(G(x);G,F)=1} \quad \textbf{vs} \quad H_1 : d(G(x); G, F) \neq 1$$

... but $d(G(x); G, F) = 1 + \sum_{j>0} \theta_j T_j(x; G)$ **remember?**

$$\Rightarrow d(G(x); G, F) = 1 \text{ whenever } \theta_1 = \theta_2 = \cdots = 0 \,!$$

- Deviance test statistic: $D = \sum_{j=1}^{m} \widehat{\theta}_j^2 \xrightarrow{d} \chi^2_m$ (under H_0, as $n \to \infty$)
- Confidence bands:



But what are g and f in practice?

It really depends on the astrophysical problem...

For instance...

- If we are trying to assess if the background model is correct:

$$\underbrace{f(x)}_{\text{true (unknown) bkg}} = \underbrace{g(x)}_{\text{postulated bkg}} \underbrace{d(G(x); G, F)}_{\text{comparison density}} \tag{6}$$

- If we are trying to detect the signal of a new source:

$$\underbrace{f(x)}_{\substack{\text{truth (may or may not}\\ \text{contain the signal)}}} = \underbrace{g(x)}_{\substack{\text{true (or estimated)}\\ \text{bkg model}}} \underbrace{d(G(x); G, F)}_{\text{comparison density}} \tag{7}$$

Ok, but what if we are dealing with multidimensional distributions?

Main reference *Algeri S. (2020+), arXiv:2009.00503*

From a more technical perspective...

Methodological tasks

- Extend the concept of comparison density to more than 1D
- Select the “best” estimator for it
- Adjust the inference for post-selection

The joint comparison density

In 1D we considered the density of $u = G(x)$, what do we do now...

Rosenblatt's transform

$$\boldsymbol{u} = (u_1, \ldots, u_p) = \left(G_1(x_1), \ldots, G_p(x_p|x_{p-1}, \ldots, x_1)\right) = \boldsymbol{G}(\boldsymbol{x}),$$

the inverse is $\boldsymbol{G}^{-1}(\boldsymbol{u}) = \boldsymbol{x}$.

The joint comparison density

Let f and F be, respectively, the probability function and the cdf of the random vector $\boldsymbol{X} \in \mathbb{R}^p$. Let g and G be its hypothesized pdf and cdf,

$$d(\boldsymbol{u}; G, F) = \frac{f\left(\boldsymbol{G}^{-1}(\boldsymbol{u})\right)}{g\left(\boldsymbol{G}^{-1}(\boldsymbol{u})\right)} \tag{8}$$

Notice that in (8), $\boldsymbol{G}(\boldsymbol{x}) \neq G(\boldsymbol{x})$.

Let's use a convenient representation for it

Let $\{S_j(\boldsymbol{u})\}_{j\geq 0}$ be an orthonormal (tensor) basis on $L_2[0,1]^p$ with $S_0(\boldsymbol{u}) = 1$. Then,

$$d(\boldsymbol{u}; G, F) = 1 + \sum_{j>0} \theta_j \, S_j(\boldsymbol{u}) \qquad \text{with } \boldsymbol{u} \in [0,1]^p. \tag{9}$$

with

$$\theta_j = \int_{[0,1]^p} S_j(\boldsymbol{u}) \, d(\boldsymbol{u}; G, F) \, \mathrm{d}\boldsymbol{u}. \tag{10}$$

Ok, but how do we estimate this expansion in this setting?

- How do we actually estimate all of the above?
- How many (and which) θ_j coefficients should we use?

Estimation and model selection

i. Choose a sufficiently large value $m_{\max}$ (e.g., $m_{\max} = 20$).

ii. Estimate θ_j, $j = 1, \ldots, m_{\max}$, via

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} S_j(\boldsymbol{u}_i) \qquad \text{with } \boldsymbol{u}_i = \boldsymbol{G}(\boldsymbol{x}_i) \text{ (our Rosenblatt transform).}$$

iii. Sort the $\widehat{\theta}_j$s in decreasing magnitude i.e., $\widehat{\theta}^2_{(1)} \geq \widehat{\theta}^2_{(2)} \geq \cdots \geq \widehat{\theta}^2_{(m_{\max})}$

iv. Select the largest m for which $AIC(m)$ is maximum

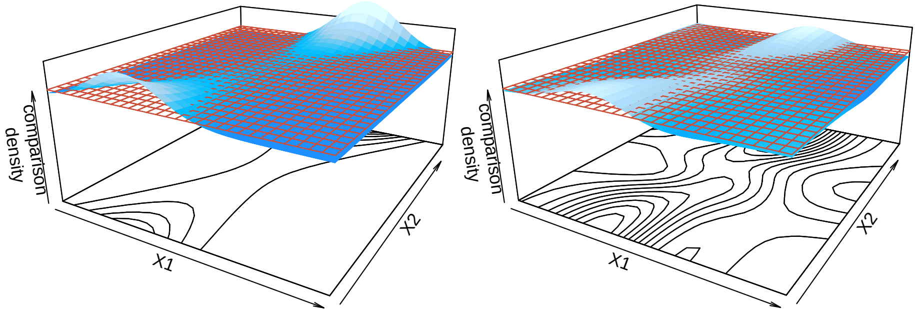
$$AIC(m) = \sum_{(j)=1}^{m} \widehat{\theta}^2_{(j)} - \frac{2m}{n} \qquad \text{(Mukhopadhyay, 2017)}$$

v. Set the remaining "nonsignificant" coefficients equal to zero.

Estimated (joint) comparison density

$$\widehat{d}(\boldsymbol{u}; G, F) = 1 + \sum_{(j)=1}^{m} \widehat{\theta}_{(j)}\, S_{(j)}(\boldsymbol{u})$$

An illustrative example...



Left: True comparison density of a random vector (X_1, X_2) distributed as a mixture of two, overlapping truncated bivariate Gaussians when the postulated model is assumed to be a bivariate truncated normal.**Right:** Estimated comparison density obtained using $m = 9$ out of $m_{\text{max}} = 19$ coefficients selected via the AIC criterion.

Post-selection Inference

Similarly to the 1D case we would like to test:

$$H_0 : \boxed{d(\boldsymbol{u}; G, F) = 1} \quad \textbf{vs} \quad H_1 : d(\boldsymbol{u}; G, F) \neq 1$$

we know that $d(\boldsymbol{u}; G, F) = 1$ whenever $\theta_1 = \theta_2 = \cdots = 0$

WARNING!

When performing model selection the estimators is affected by it. As a result,

$$D = \sum_{(j)=1}^{m} \widehat{\theta}^2_{(j)} \overset{d}{\not\to} \chi^2_m \quad \text{(not even under } H_0 \text{ and/or as } n \to \infty\text{).}$$

A possible post-selection adjustment

$$\text{p-value} = P(\chi^2_{m_{\max}} > D_{obs})$$

with D_{obs} being the value of D observed on the data.

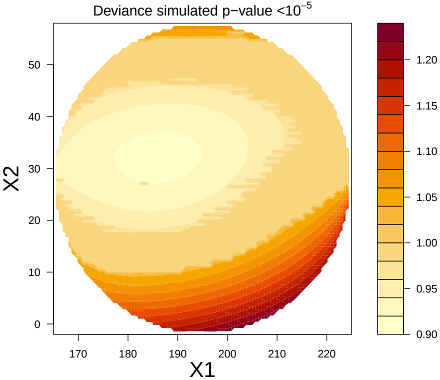
An example on background calibration

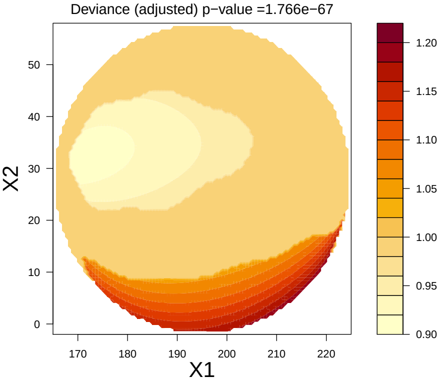
We consider a realistic simulations from the Fermi Large Area Telescope.

- **Goal:** Assess the impact of the (unknown) instrumental error affecting the hypothesized uniform distribution.
- **Region of interest** A circular a disc in the sky of 30° radius and centered at (195RA,28DEC).
- **Data:** 68,658 events from cosmic background.

The impact of the instrumental error

Despite we are not going to see how to construct confidence bands, it is worth mentioning that the theory behind it is essentially the same of that of the look-elsewhere effect in multiple dimensions (e.g. Vitells and Gross, 2011).





$$\widehat{d}(G_1(x_1), G_2(x_2|x_1); G, F) = 1 + 0.02\,T_1(G_1(x_1)) - 0.04\,T_1(G_2(x_2|x_1)) + 0.041\,T_2(G_2(x_2|x_1)).$$

All of this is nice, but what if we have more than 2 or 3 dimensions?

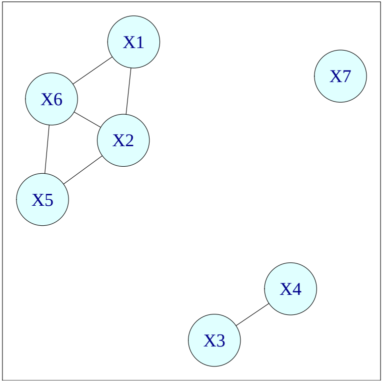
Indeed, the real question here is...

What do we want to know about our $p > 3$ dimensions?

I will discuss these aspect only briefly, but feel free to ask for more details during the discussion (or check out Sections 5 and 6 of *Algeri S. (2020+), arXiv:2009.00503*).

Are we interested in testing for independence?

Suppose we don't know anything about the dependence structure, we can learn about it with a simple plot...



each edge corresponds to a test of hypothesis where we reject the hypothesis of independence between the two variables connected by the edge. If two variables are not connected by an edge it means we have independence.

But what if we do have a G model I want to test...

Variable	**True** (F)	**Hypothesized** (G)	**Correct**
$X_6 \| X_1, X_2, X_5$	$\text{Poi}\left[e^{0.03x_1+0.02x_2 +0.01x_2^2 +0.02x_5}\right]$	$\text{Poi}\left[e^{0.03x_1+0.02x_2+0.02x_5}\right]$	No
X_1, X_2, X_5	$N\left[\begin{pmatrix}10\\15\\11\end{pmatrix}, \begin{pmatrix}4 & 0.5 & 0\\0.5 & 3 & 1\\0 & 1 & 5\end{pmatrix}\right]$	$N\left[\begin{pmatrix}10\\15\\11\end{pmatrix}, \begin{pmatrix}4 & 0.5 & 0\\0.5 & 3 & 1\\0 & 1 & 5\end{pmatrix}\right]$	Yes
$X_4 \| X_3$	$\text{Exponential}\left(\frac{1}{x_3}\right)$	$\text{Exponential}\left(\frac{1}{x_3}\right)$	Yes
X_3	$\text{Exponential}(1)$	$\text{Exponential}(0.9)$	No
X_7	T_3	$\text{Cauchy}(0,1)$	No

Are we interested in learning about the sources of mismodelling?

Suppose we want to know if our G is indeed correct and if it isn't, we want to know where the problems are. We can do that with a simple table...

Random vector	**df**	**(Adjusted) p-value**
$(X_1, X_2, X_5, X_6, X_7)$	16383	$< 10^{-130}$
(X_1, X_2, X_5, X_6)	256	$< 10^{-130}$
(X_1, X_2, X_5)	63	1
(X_3, X_4)	15	$1.799 \cdot 10^{-11}$
X_3	3	$3.801 \cdot 10^{-6}$
X_7	3	$3.732 \cdot 10^{-119}$

Each row corresponds to a test of hypothesis to asses the validity of the distribution we have specified for the random vectors in the first column.

Conclusion

What does informative Goodness-Of-Fit (iGOF) allow us to do?

- We can perform goodness-of-fit for both univariate and multivariate data distributions.
- If $p \leq 3$ we can visualize where and how mismodelling occurs.
- If $p \geq 2$ we can learn the underlying dependence structure.
- If $p \geq 1$ we can identify the sources of mismodelling.

References

- Algeri, S. (2020+). Informative goodness-of-fit for multivariate distributions. *Under review. arXiv:2009.00503*
- Algeri, S., 2020. Detecting new signals under background mismodelling. *Physical Reviews D.*
- Mukhopadhyay, S., 2017. Large-scale mode identification and data-driven sciences. *Electronic Journal of Statistics*.
- Neyman, J., 1937. Smooth test for goodness of fit. *Scandinavian Actuarial Journal.*
- Parzen, E., 1979. Nonparametric Statistical Data Modeling. *Journal of the American Statistical Association.*
- Vitells, O. and Gross, E., 2011. Estimating the significance of a signal in a multi-dimensional search. *Astroparticle Physics.*

Thank you for your time.