Goodness of Fit – Thoughts for Discussion

Richard Lockhart

PHYSTAT Anomalies

May 25, 2022

Conclusions

- I will talk about Goodness-of-Fit generically.
- I won't tell anyone how to do ML.
- I will ask what kind of statistical problem you have.
- I will make a list of ideas that caught my attention.

LHC setup in my words

- Data: sample of N (Poisson) events (recorded as vectors $\boldsymbol{X}_i$).
- Statistical (background) Model: Standard Model plus Detector Model.
- Looking for: other events not predicted by Statistical Model.
- Three statistical attitudes to this problem:
  - This is a two sample problem.
  - This is a goodness-of-fit problem.
  - This is a screening problem.

Two sample problem

- You have a sample of data from the LHC after cuts applied.
- And you have a background sample: Monte Carlo or side-bands.
- Statistical Model has parameters not perfectly known.
- Some estimated within expt, some externally.
- Surely you cannot sample from this model.
- Reason: all events in data have same parameter values; not known.
- Exceptions? Require parameter uncertainty negligible compared to signal.

GOF for statisticians

- Statistical Model: family of densities or intensities, $b(x); x \in \mathcal{X}$, for data:
$$\{b \in \mathcal{B}\}.$$
- Most common case in statistical literature: $\mathcal{B}$ is parametrized:
$$\mathcal{B} = \{b(x;\theta) : \theta \in \Theta_B\}$$
- Goal is to decide if true density is in $\mathcal{B}$.
- Traditional framing: f_0 is true density/intensity. Test null
$$H_0 : f_0(\cdot) = b(\cdot;\theta_0) \text{ some } \theta_0 \in \mathcal{B}$$
versus
$$\text{versus } H_1 : f_0 \notin \mathcal{B}.$$
- Vector θ includes parameters of SM not exactly known.

Commentary

- For anomaly searches high power is very much desired.

Commentary

- For anomaly searches high power is very much desired.
- Especially at correct non SM model of universe.

Commentary

- For anomaly searches high power is very much desired.
- Especially at correct non SM model of universe.
- Fact: most users of GOF tests want null to be right; less incentive for powerful tests.
- Other framings may make more sense:
- Maybe goal of Anomaly detection is “screening”: identify large number of possible anomalies to study in detail at LHC.

Commentary

- For anomaly searches high power is very much desired.
- Especially at correct non SM model of universe.
- Fact: most users of GOF tests want null to be right; less incentive for powerful tests.
- Other framings may make more sense:
- Maybe goal of Anomaly detection is "screening": identify large number of possible anomalies to study in detail at LHC.
- Identify large number of anomalies to justify building different detectors.

One testing strategy: parametric null

- Model predicts mean (expectation) value of $H(\mathbf{X}; t) : t \in \mathcal{T}$ is

$$\mu(t, \theta) = \langle H(\mathbf{X}; t) \rangle .$$

- Study Empirical Discrepancy (here n is expected background total)

$$W_n(t, \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{H(\mathbf{X}_i, t) - \mu(t, \theta)\} .$$

- Build P-value out of distribution of univariate summary of size of W.
- Classic summaries: linear, **quadratic**, supremum.
- Important: null distribution usually depends strongly on $\mathcal{B}$.
- And on true parameter value inside $\mathcal{B}$.

Quadratic Examples

- Empirical Distribution Function (EDF) tests: Anderson-Darling (AD) , Cramér-von Mises (CvM).
- CvM/AD: $H(x,t) = w(t,\theta)1(B(x,\theta) \leq t)$
- In general:

$$\int_t \{w(t,\theta)W_n(t,\theta)\}^2 \, dt$$

or

$$\frac{1}{M}\sum_{j=1}^{M}\{w(t_j,\theta)W_n(t_j,\theta)\}^2$$

evaluated at some estimate of θ_0.
- Get P values? Yes – if you understand θ

Effect of uncertainty in parameters

- Linearization of $H - \mu$ in θ near θ_0:

$$W_n(t,\theta) \approx W_n(t,\theta_0) + \sqrt{N}\,(\theta-\theta_0)^\top \nabla_\theta \mu(t,\theta)\Big|_{\theta_0}.$$

- Approximately Gaussian Process in θ, locally.
- Evaluate at estimate of θ: internal to data, external to data, some of both.
- Use MLE: variability reduced – often a lot.
- Use uncertain estimate from other data: variability increased.
- So increased by systematics, decreased by fitting.
- Maximal decrease by Maximum Likelihood.
- Fit more parameters get smaller statistics.

P-values

- Null limit distribution

$$\sum_{k=1}^{\infty} e_k Z_k^2 = \text{ linear combination of } \chi_1^2$$

- The e_k are eigenvalues of approx covariance function of $W_n(t, \hat{\theta})$.
- Each Z_k is limit of centered scaled sample mean of corresponding eigenfunctions.
- LRT is, for large n, essentially in this class. Smooth tests too.
- IF, you have suitable theory about estimate $\hat{\theta}$, THEN, the e_k can be estimated and P computed / approximated by numerical Fourier inversion (Imhof 1962).
- For maximum likelihood use *sandwich* estimate.
- For externally estimated (systematics) use independence.

Bayes

- If null hypothesis is *NOT* composite then NP lemma can be used.
- Like NP constrain type 1 error rate.
- Maximize average power wrt prior on alternative.
- Strategy following Andrea Wulzer. Model

$$\frac{p(x|w)}{p(x|R)} = exp(f(x,w))$$

- Make $f(x,w)$ GP with covariance. Roeder and Wasserman (1997).
- Localized to $n^{-1/2}$ neighbourhood result is U statistic.
- Power depends on eigenfunctions of covariance.
- Smooth tests are example with finite spectrum.
- Posterior can point, *maybe* to nature of departure.

Conclusions.

- TBD